Space Telescope
Science Institute
Symposium Series

14

A Decade of *Hubble Space Telescope* Science

Edited by
M. Livio, K. Noll and M. Stiavelli

CAMBRIDGE

This page intentionally left blank

SPACE TELESCOPE SCIENCE INSTITUTE

SYMPOSIUM SERIES: 14
Series Editor S. Michael Fall, Space Telescope Science Institute

A DECADE OF *HUBBLE SPACE TELESCOPE* SCIENCE

The *Hubble Space Telescope* has made some of the most dramatic discoveries in the history of astronomy. From its vantage point 600 km above the Earth, *Hubble* is able to capture images and spectra that would be difficult or impossible to obtain from the ground. This volume represents some of the most important scientific achievements of the *Hubble Space Telescope* in its first decade of operation. Written by world experts, the book covers topics ranging from our own solar system to cosmology. Chapters describe cutting-edge discoveries in the study of Mars and Jupiter, of stellar birth and death, of star clusters, of the interstellar medium, of our own Milky Way galaxy and of other galaxies, of supermassive black holes, and of the determination of cosmological parameters, including the age and ultimate fate of our universe. This is an indispensable collection of review articles for researchers and graduate students.

SPACE
TELESCOPE
SCIENCE
INSTITUTE

Other titles in the Space Telescope Science Institute Series.

1 Stellar Populations
Edited by C. A. Norman, A. Renzini and M. Tosi 1987 0 521 33380 6

2 Quasar Absorption Lines
Edited by C. Blades, C. A. Norman and D. Turnshek 1988 0 521 34561 8

3 The Formation and Evolution of Planetary Systems
Edited by H. A. Weaver and L. Danly 1989 0 521 36633 X

4 Clusters of Galaxies
Edited by W. R. Oegerle, M. J. Fitchet and L. Danly 1990 0 521 38462 1

5 Massive Stars in Starbursts
Edited by C. Leitherer, N. R. Walborn, T. M. Heckman and C. A. Norman
1991 0 521 40465 7

6 Astrophysical Jets
Edited by D. Burgarella, M. Livio and C. P. O'Dea 1993 0 521 44221 4

7 Extragalactic Background Radiation
Edited by D. Calzetti, M. Livio and P. Madau 1995 0 521 49558 X

8 The Analysis of Emission Lines
Edited by R. E. Williams and M. Livio 1995 0 521 48081 7

9 The Collision of Comet Shoemaker-Levy 9 and Jupiter
Edited by K. S. Noll, H. A. Weaver and P. D. Feldman
1996 0 521 56192 2

10 The Extragalactic Distance Scale
Edited by M. Livio, M. Donahue and N. Panagia 1997 0 521 59164 2

11 The Hubble Deep Field
Edited by M. Livio, S. M. Fall and P. Madau 1998 0 521 63097 5

12 Unsolved Problems in Stellar Evolution
Edited by M. Livio 2000 0 521 78091 8

13 Supernovae and Gamma-Ray Bursts
Edited by M. Livio, N. Panagia and K. Sahu 2001 0 521 79141 3

A decade of *Hubble Space Telescope* science

Proceedings of the
Space Telescope Science Institute Symposium,
held in Baltimore, Maryland
April 11–14, 2000

Edited by
MARIO LIVIO
Space Telescope Science Institute, Baltimore, MD 21218, USA

KEITH NOLL
Space Telescope Science Institute, Baltimore, MD 21218, USA

MASSIMO STIAVELLI
Space Telescope Science Institute, Baltimore, MD 21218, USA

Published for the Space Telescope Science Institute

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org

Information on this title: www.cambridge.org/9780521824590

First published in print format 2003

ISBN-13 978-0-511-07075-4 eBook (EBL)
ISBN-10 0-511-07075-6 eBook (EBL)

ISBN-13 978-0-521-82459-0 hardback
ISBN-10 0-521-82459-1 hardback

Contents

Participants

Agol, Eric	The Johns Hopkins University
Aloisi, Alessandra	Space Telescope Science Institute
Anderson, Jay	University of California–Berkeley
Andrews, Thomas	
Avera, Randy	NASA/FAA
Bagenal,Fran	University of Colorado
Bahcall, John	Institute for Advanced Study
Bally, John	University of Colorado, CASA
Beckwith, Steven	Space Telescope Science Institute
Bell, James	Cornell University
Biretta, John	Space Telescope Science Institute
Bohlin, Ralph	Space Telescope Science Institute
Bond, Nicholas	Pennsylvania State University
Boyer, Robert	Space Telescope Science Institute
Brown, Robert	Space Telescope Science Institute
Bruhweiler, Fred	Catholic University
Cacciari, Carla	Observatory of Bologna, Italy
Caraveo, Patricia	Istituto di Fisica Cosmica
Carpenter, Kenneth	NASA/Goddard Space Flight Center
Chae, Kyu-Hyun	University of Pittsburgh
Charlton, Jane	Pennsylvania State University
Catzichristou, Eleni	NASA/Goddard Space Flight Center
Christian, Carol	Space Telescope Science Institute
Chu, Ming	The Chinese University of Hong Kong
Clarke, John	University of Michigan
Corbin, Michael	University of Arizona
Courtin, Régis	DESPA-CNRS/Observatoire de Paris
Della Valle, Massimo	Osservatorio Astronomico di Arcetri, Italy
Di Benedetto, G. Paolo	C.N.R.–Istituto di Fisica Cosmica
Diaz-Miller, Rosie	Space Telescope Science Institute
Dolphin, Andrew	KPNO/NOAO
Donahue, Megan	Space Telescope Science Institute
Dressler, Alan	Carnegie
Duerbeck, Hilmar	VUB Brussels
Duncan, Douglas	University of Chicago
Feldman, Paul	The Johns Hopkins University
Ferguson, Harry	Space Telescope Science Institute
Franchini, Mariagrazia	Trieste Astronomical Observatory
Freedman, Wendy	Carnegie Observatories
Fruchter, Andrew	Space Telescope Science Institute
Ganguly, Rajib	Pennsylvania State University
Giavalisco, Mauro	Space Telescope Science Institute
Godon, Patrick	Space Telescope Science Institute
Gonzaga, Shireen	Space Telescope Science Institute
Greyber, Howard	
Griffiths, Richard	Carnegie Mellon Univ.
Groth, Edward	Princeton University
Grunsfeld, John	NASA

Guglielmetti, Fabrizia	Space Telescope Science Institute
Gull, Theodore	NASA, Goddard Space Flight Center
Hammel, Heidi	Space Science Institute (Boulder CO)
Harnett, Kevin	NASA/Goddard Space Flight Center
Harrington, J. Patrick	University of Maryland
Harris, Gretchen	University of Waterloo
Harris, William	McMaster University
Hasan, Hashima	NASA Headquarters
Heap, Sara	NASA/Goddard Space Flight Center
Heaton, Hal	The Johns Hopkins University–Applied Physics Lab.
Heindel, Larry	
Hester, Jeff	Arizona State University
Holberg, Jay	University of Arizona
Hynes, Robert	University of Southhampton
Imhoff, Catherine	Space Telescope Science Institute
Jangren, Anna	Pennsylvania State University
Jannuzi, Buell	NOAO
Jeletic, James	NASA/Goddard Space Flight Center
Koekemoer, Anton	Space Telescope Science Institute
Koenigsberger, Gloria	UNAM-Inst. Astronomia
Kukula, Marek	University of Edinburgh
Kulkarni, Varsha	Clemson University
Lanzetta, Kenneth	University of New York
Lauer, Tod R.	NOAO
Leckrone, David	NASA/Goddard Space Flight Center
Lehner, Nicolas	The Queen's University of Belfast
Leitherer, Claus	Space Telescope Science Institute
Livio, Mario	Space Telescope Science Institute
Lucas, Ray	Space Telescope Science Institute
Lundqvist, Peter	Stockholm Observatory
Macchetto, Duccio	Space Telescope Science Institute
Macri, Lucas	Harvard-Smithsonian Center for Astrophysics
Malagnini, Maria Lucia	Astronomy Department of Trieste University
Martel, André	The Johns Hopkins University
McCray, Richard	University of Colorado
McGrath, Melissa	Space Telescope Science Institute
Mignani, Roberto	Space Telescope European Coordinating Facility
Mitalas, Romas	University of Western Ontario
Morossi, Carlo	Trieste Astronomical Observatory
Mould, Jeremy	Mt. Stromlo Observatory
Muxlow, Thomas	Jodrell Bank Observatory
Nicollier, Claude	NASA
Niedner, Malcolm	NASA/Goddard Space Flight Center
Noll, Keith	Space Telescope Science Institute
Norman, Colin	Space Telescope Science Institute
O'Dell, Robert	Rice University
Oey, Sally	Space Telescope Science Institute
Ogilvie, Gordon	Institute of Astronomy, Cambridge
Patriarchi, Patrizio	Caismi-C.N.R., Firenze
Pearson, Kevin	University of St. Andrews

Perinotto, Mario	University of Firenze, Italy
Perlmutter, Saul	University of California–Berkeley
Pringle, James	Institute of Astronomy, Cambridge
Pun, Chun-Shing Jason	NASA/Goddard Space Flight Center
Rhodes, Jason	
Rieke, Marcia	Steward Observatory
Riess, Adam	Space Telescope Science Institute
Röser, Hermann-Josef	Max-Plank-Institut für Astronomie
Ruzmaikina, Tamara	University of Arizona
Sahu, Kailash	Space Telescope Science Institute
Savage, Blair	University of Wisconsin–Madison
Scarlata, Cladia	Università di Padova
Schneider, Glenn	Steward Observatory
Schreier, Ethan	Space Telescope Science Institute
Schultz, Alfred	CSC/Space Telescope Science Institute
Seab, C. Gregory	University of New Orleans
Seitter, Waltraut	Münster University
Shaw, Richard	Space Telescope Science Institute
Silverstone, Murray	Steward Observatory
Simpson, David	NASA/Goddard Space Flight Center
Smartt, Stephen	Institute of Astronomy, Cambridge
Smett, Alain	NASA/Goddard Space Flight Center
Stecher, Theodore	NASA/Goddard Space Flight Center
Sterken, Chris	University of Brussels
Stiavelli, Massimo	Space Telescope Science Institute
Tammann, Gustav	Basel Astronimisches Institut
Thompson, Rodger	University of Arizona
Tolstoy, Eline	European Southern Observatory
Treu, Tommaso	Space Telescope Science Institute
Trivedi, Pranjal	Cambridge University
Tyson, Anthony	Lucent Technologies
Verner, Katya	University of Kentucky
Vesperini, Enrico	University of Massachusetts–Amherst
Voit, Mark	Space Telescope Science Institute
Weaver, Harold	The Johns Hopkins University
Whitmore, Brad	Space Telescope Science Institute
Williams, Bob	Space Telescope Science Institute
Williger, Gerard	NASA/Goddard Space Flight Center
Wilson, Andrew	University of Maryland
Wiseman, Jennifer	The Johns Hopkins University
Zonak, Stephanie	Pennsylvania State University

Preface

The Space Telescope Science Institute Symposium on "A Decade of HST Science" took place during 11–14 April 2000.

There is no doubt that the *Hubble Space Telescope* (*HST*) in its first decade of operation has had a profound impact on astronomical research. But *HST* did much more than that. It literally brought a glimpse of the wonders of the universe into millions of homes worldwide, thereby inspiring an unprecedented public curiosity and interest in science.

HST has seen farther and sharper than any optical/UV/IR telescope before it. Unlike astronomical experiments that were dedicated to a single, very specific goal, *HST*'s achievements are generally not of the type of singular discoveries. More often, *HST* has taken what were existing hints and suspicions from ground-based observations and has turned them into certainty.

In other cases, the level of detail that *HST* has provided forced theorists to re-think previous broad-brush models, and to construct new ones that would be consistent with the superior emerging data. In a few instances, the availability of *HST*'s razor-sharp vision at critical events provided unique insights into individual phenomena.

These proceedings represent a part of the invited talks that were presented at the symposium, in order of presentation. We thank the contributing authors for preparing their papers.

We thank Sharon Toolan of ST ScI for her help in preparing this volume for publication.

Mario Livio, Keith Noll, Massimo Stiavelli
Space Telescope Science Institute
Baltimore, Maryland
April, 2000

HST studies of Mars

By JAMES F. BELL III

Department of Astronomy, Cornell University, 402 Space Sciences Building, Ithaca, NY 14853-6801

HST observed Mars during all 5 oppositions between 1990 and 1999, providing unique new observations of the planet's atmosphere and surface during seasons which are typically poorly-observed telescopically and in wavelength regions or at spatial scales that are not at all observed by spacecraft. *HST* observations also filled a crucial gap in synoptic observations of Mars prior to 1998, during a time when no spacecraft were observing the planet. *HST* data have provided important new insights and understanding of the Martian atmosphere, surface, and satellites, and they continue to fulfill important spacecraft mission support functions, including atmospheric aerosol characterization, dust storm monitoring, and instrument cross-calibration.

1. Introduction

Mars has been the subject of intense telescopic observations for centuries (see, for example, reviews by Martin et al. 1992 and Sheehan 1988). Interest in the red planet stems partly from its prominent appearance in the night sky as a bright extended object roughly every 26 months, and also from historic telescopic observations and more recent spacecraft encounters that have revealed many similarities between Mars and the Earth in terms of surface and atmospheric characteristics and climatic histories. While cold and arid today and probably inhospitable to most forms of life, evidence exists indicating that Mars once may have had a much more clement climate, during a postulated "warm and wet" epoch early in solar system history (e.g. Pollack et al. 1987; Carr, 1998).

The postulated similarities between early Mars and early Earth has fueled intense speculation and scientific interest on the question of life: could Mars have been (or still be?) a habitable environment for life to form, exist, and evolve? To answer this and other related questions requires a detailed understanding of both the past and present environment of Mars. While important clues have been provided by more than three decades of spacecraft flybys, orbiters, and landed investigations, these have been sporadic (and expensive!) glimpses of a complex planet, usually sampling only one part of the Martian seasonal cycle or one particular landing site in detail. For example, while the Viking Orbiter and Lander missions successfully observed the planet for more than two Mars years from orbit and from two widely-separated landing sites during the mid- to late-1970s, only one spacecraft successfully arrived at Mars during the 1980s (the Soviet Phobos-2 orbiter, which only operated for a few months), and the next successful missions after that didn't occur until Mars Pathfinder, which landed in 1997, and the Mars Global Surveyor orbiter, which began mapping observations in 1997–98. Even these most recent missions were rather narrowly focused (Pathfinder on the local geology of a particular region; Global Surveyor on systematic high-resolution observations from a 2:00 a.m./2:00 p.m. Sun-synchronous orbit). Sadly, we have all been poignantly reminded of the inherent risks associated with spacecraft exploration of Mars recently, with the loss of both missions sent to Mars in 1999.

So, despite the incredible successes of many of the space missions sent to Mars, there is still clearly a niche for systematic synoptic-scale telescopic observations of the planet. Filling this niche requires several key instrumental characteristics, including: (a) high spatial resolution (less than hundreds of km), to resolve small-scale features on the surface and in the atmosphere; (b) high data fidelity and accurate calibration, to detect weak photometric and/or colorimetric differences on the surface and in the atmosphere; and (c) good temporal sampling, in order to be able to quantify surface and atmospheric changes with season in the current Martian climatic regime. While meeting requirement (c) is possible from many large groundbased telescopes, meeting both requirements (a) and (b) in addition is usually not possible from groundbased platforms. This is because the Earth's atmosphere blurs resolution out to a large fraction of an arcsec even during the best seeing conditions (translating to more than several hundred km on Mars even at an excellent Mars opposition), and telluric water vapor and other species produce time-variable absorption features at key wavelengths that could otherwise be use to detect atmospheric and surface constituents on Mars. Rarely, when requirement (a) or (b) has been met by groundbased observations, the data have revealed that it is possible to detect and quantify variations in surface and atmospheric materials (e.g. Singer et al. 1979; Bell et al. 1990; Merényi et al. 1988; Bell and Crisp, 1993).

HST provides the ability to meet all of the requirements for scientifically-meaningful observations of Mars that complement, rather than duplicate, existing or ongoing spacecraft observational programs. Maximizing the ability of *HST* observations to advance Mars science has been a primary goal of all the observations conducted between *HST* Cycles 0 and 9. In this paper I will describe some of the outstanding unresolved issues in Mars studies, and describe the rationale and justification for the use of *HST* to observe such a close and bright object ($z \ll 1$). Next I will describe the observations of Mars that have been obtained by *HST* between 1990 and 1999. The results and scientific implications of the data will be discussed, broken down into the categories of Martian atmosphere, Martian surface, and Martian satellites. Special attention will be paid to describing how these *HST* measurements have played a role in shaping the acquisition and interpretation of ongoing or planned Mars orbital and landed spacecraft datasets. Finally, I will discuss future opportunities for Mars observations with *HST*, and how new data could continue to expand our understanding of the planet.

2. Outstanding issues in Mars studies

Mars is an enigmatic and fascinating planet. It is the most "Earth-like" of the other planets in the solar system. Evidence from decades of telescopic and spacecraft observations reveal geologic and isotopic evidence for substantial changes in the Martian climate during the early history of the planet. Specifically, degraded/eroded landforms, the presence of dendritic valley networks, and isotopic fractionation indicative of the loss of what was once a more substantial atmospheric indicate that the planet may have experienced a "warm and wet" climate regime (Pollack et al. 1987; Carr, 1999) with temperatures substantially above the melting point of water during the first billion years of its evolution. It is unknown whether this clement period in Martian history was short- or long-lived, though, or whether there were multiple such periods in response to planetary orbital/inclination variations or other external forcing processes. The duration and extent of stable liquid water at or near the Martian surface has important and different implications for the geologic and possibly biologic evolution of Mars, and understanding the history and role of water and its implications for climate and life are now the major drivers in NASA's Mars exploration program. Recent announcements concerning the possible presence of fossilized life forms in a Martian meteorite (McKay et al. 1996) and the possible presence of liquid water very close to the Martian surface (Malin & Edgett 2000), while controversial, underscore the intense public and scientific interest in Mars and the role that Martian studies play in larger exobiologic debates.

But understanding of the past climate conditions on Mars cannot be achieved without first understanding the present climate. Mars today is a cold and arid world with a thin atmosphere (tens of mbar) and (probably) little or no internal geologic activity. Conditions at the surface are influenced by seasonal and interannual cycles of CO_2 condensation and sublimation at the poles, by the exchange of small amounts of water vapor (tens of pr μm) between the atmosphere and the regolith, and by the radiative and physical influence of local- and global-scale atmospheric dust (James et al. 1992). The water, CO_2, and dust cycles have been studied intensely by spacecraft, but only at infrequent intervals or for relatively short periods of time. The historic telescopic record reveals that Mars has a dynamic and changing atmosphere and surface on timescales of decades to even centuries (Martin et al. 1992; McKim et al. 1999), and so understanding of the character of these cycles and of the planet's surface-atmosphere interactions must be teased out of both high-resolution focused measurements and long-timescale synoptic observations.

Some specific issues that remain elusive in our understanding of the current Martian atmosphere include: (1) What is the composition, distribution, and opacity of atmospheric aerosols (silicate dust, water ice, CO_2 ice, other aerosols?) and how does that composition change in response to diurnal and seasonal timescale variations in radiative forcing? (2) What is the radiative influence of airborne dust and/or clouds on the energy balance of the Martian atmosphere, and how do these aerosols tangibly influence the climate (temperature, pressure changes due to volatile condensation, local winds). (3) What are the dominant styles and rates of dust and volatile transport in the Martian atmosphere, and how are they influenced by topography, albedo, and seasonal climate variations? And (4) What is the magnitude of the present variability of the Martian climate (seasonal temperature and pressure extrema, dust storm frequency, atmospheric opacity) on yearly, interannual, and even longer timescales decipherable from the telescopic and geologic record?

Understanding the nature of the present Martian surface also plays a key role in determining the climatic and geologic history of the planet. Surface geologic activity like volcanism, tectonism, and impact processes are obvious manifestations of the geologic evolution of the planet. In some cases, these processes can have important effects on the climate, such as the release of greenhouse gases by volcanic eruptions and the heat flux created during large impact cratering events. Most of our understanding of the detailed surface geology of Mars has come from orbital and landed spacecraft observations during the past few decades. These observations have shown that at increasing spatial resolutions, the detailed geomorphology of Mars looks less and less like that of the Earth (e.g. Malin et al. 1998), reflecting instead the specific style and nature of uniquely Martian geologic processes. This underscores an important point: Mars is truly a different world than the Earth, and while the planet's surface and atmospheric processes are dictated by the same physics and chemistry driving familiar processes here, the timescales and boundary conditions for geologic and climatic activity on Mars are substantially different from those on the Earth.

Some of the specific issues that remain outstanding in our study of the Martian surface include: (1) What is the composition and mineralogy of the surface rocks and soils and the airborne dust, and what do the chemistry and mineralogy reveal about the current and (especially) past Martian climate? On Earth, many kinds of rocks and minerals preserve a record of local climate conditions during their formation. Examples include hydroxides, carbonates, and sulfates, which can sequester atmospheric gases into their crystalline structures; and Fe(III) oxides, which have polymorphs that form within specific temperature, pH, humidity, and $f(O_2)$ conditions. Some of these minerals are stable or metastable under current Martian climatic conditions but not in a perhaps warmer, wetter past, meaning that their detection and characterization could also provide information on the extent and timing of climate change over the course of the planet's history. (2) What are the sources and sinks of volatiles (especially water) and the nature and rate of surface/atmosphere volatile exchange on Mars? This is a key question in Martian climate studies, as noted above, and includes not only the mineralogy issues just discussed, but also the growth and decay of the planet's seasonal H_2O and CO_2 polar ice caps, the diurnal exchange of water vapor between the surface and atmosphere, and the controversial issue of deep or shallow subsurface liquid water on Mars, and its possible role in alteration of surface materials and formation of localized geologic features (e.g. Malin & Edgett 2000). And (3) What are the dominant processes responsible for changing the albedo and overall geology of the surface with time? These include aeolian processes like dust storms on a variety of scales and which have been recorded telescopically for centuries, as well as volcanic, tectonic, and impact processes that modify the surface on much longer timescales but whose effects are preserved in the current topography and landforms of the planet.

3. Why Use *HST*?

But why use *HST* to observe Mars, given the fierce competition for *HST* observing time, the availability of large groundbased telescopes to observe the planet, and, more importantly, the armada of spacecraft that have studied the planet or will visit in the near future? There are five primary reasons:

(*a*) *Spectral coverage in the ultraviolet.* UV astronomical observations in general provide an important component of *HST*'s mission in general because of the inability to observe from the ground at these wavelengths. For Mars, UV observations uniquely enable observations of atmospheric O_3. Ozone, though a trace component of the atmosphere (0.04 to 0.2 ppm), plays a critical role in Martian atmospheric photochemistry and also serves as a tracer of atmospheric water vapor transport. UV measurements also provide diagnostic information on atmospheric aerosols, including the ability to discriminate between H_2O and dust clouds as sources of atmospheric opacity. Finally, a number of iron-bearing mineral species have solid state absorption features at UV wavelengths due to Fe(II) and Fe(III) electronic transitions and Fe–O charge transfer transitions that vary systematically with crystalline structure, providing a way to identify the surface mineralogy from remote spectroscopic observations.

(*b*) *Spectral coverage outside of the Earth's atmospheric "windows."* Extinction caused by water, CO_2, O_2, and other gases and aerosols in the Earth's atmosphere prevents groundbased observations at wavelengths that provide unique information about the Martian surface and atmosphere. These include regions near 1 μm and 2 μm where characterization of subtle shifts in the widths and positions of broad iron-bearing silicate absorption features are hampered by telluric water; narrow and weak Martian water vapor bands that are completely obscured by their stronger telluric counterparts and which can only be observed when Doppler-shifted away from the telluric lines (and thus when the planet has a small apparent angular diameter from Earth); the wings of strong CO_2 bands, which provide information on the Martian atmospheric pressure and temperature profile but which are masked by comparably-strong telluric CO_2 lines, and parts of the near-infrared where both strong and weak bands caused by metal–OH absorptions and structural (bound) H_2O in minerals are masked by Earth's water, CO, and CO_2 bands.

(*c*) *Spatial resolution.* The near diffraction-limited performance of the corrected *HST* optics provides the ability to discern subtle spatial structures on the Martian surface and atmosphere that are undetectable from the ground. Typical groundbased resolution on Mars is 150–300 km during good seeing conditions ($0.''5$) and during the month or so around opposition. Adaptive optics or speckle imaging techniques can be used to improve the groundbased resolution, but many of these techniques break down because Mars is so bright ($M_v \sim 4.5/\mathrm{arcsec}^2$) and none have been shown to yield reliable spectrophotometric measurements of extended sources. WFPC2 on *HST* allows resolutions as fine as $\sim$ 20 km/pixel around opposition, and $\sim$ 50 km/pixel routinely for the $\sim$ half of the Martian year observable within *HST* sun avoidance constraints. These kinds of resolutions are comparable to those obtained by the Mariner flyby spacecraft in the late 1960s, the Viking orbiter global approach observations in the mid-1970s, the Phobos-2 imaging spectroscopy measurements of the late 1980s, and are within a factor of $\sim$ 3 of the Mars Global Surveyor (MGS) Thermal Emission Spectrometer (TES) measurements that are currently being obtained.

(*d*) *Mission Support.* In a sense, *HST* is another NASA "mission" to Mars, providing the ability to obtain both synoptic-scale imaging and spectroscopy and fine-spatial regional investigations of the surface and atmosphere. *HST* measurements fill a crucial gap in spacecraft coverage. Because of the loss of the Mars Observer spacecraft in 1993, between the Phobos-2 mission in 1989 and the MGS mission in 1998 there were no spacecraft observations of Mars. The importance and need for high quality supporting telescopic observations of Mars even in an era of expanded planetary missions was reinforced during 1999 with the failure of both the Mars Climate Orbiter (MCO) and Mars Polar Lander missions. MCO in particular would have provided substantial new multispectral and atmospheric sounding data highly complementary to the types of measurements being performed by *HST*. In the absence of these orbital or landed missions, *HST* has continued to provide both the best monitoring observations and the only new measurements of Mars surface and atmospheric phenomena. The planetary science community has embraced this role for the telescope, and fully expect *HST*'s contributions to continue to include important and unique NASA spacecraft mission support functions as well as new science results.

(*e*) *Public relations and outreach. HST*'s images of Mars are spectacular, and they fulfill important non-scientific NASA goals by providing inspirational and educational information about an object that is both familiar and interesting to the general public. Mars has long fascinated the public because of its Earthlike characteristics and the potential for past or even present life on its surface. Scientists and educators capitalize on this nascent interest and have used *HST* and other mission's images and information about Mars to teach concepts of astronomy, geology atmospheric science, celestial mechanics, and even biology. The education and outreach impact can be assessed by the many K–12 curriculum materials, museum exhibits, ST ScI press releases (11), and general interest newspaper and magazine articles that have been produced based on *HST* Mars images and other data.

4. Observations

Earth to Mars oppositions occur every $\sim$ 26 months, with Mars closest approach distance varying from 56 to 101 million km over a $\sim$ 15 year period because of the eccentricity of the Martian orbit. The inclination of Mars (25°) is similar to the Earth's, and the planets' orbital cycles are phased so that oppositions occur at successively advancing Martian seasons during the $\sim$ 15 year cycle (Figure 1). Mars southern hemisphere summer occurs near the perihelion of its orbit and coincides with the closest Earth to Mars oppositions, providing the best possible spatial resolution. The worst oppositions

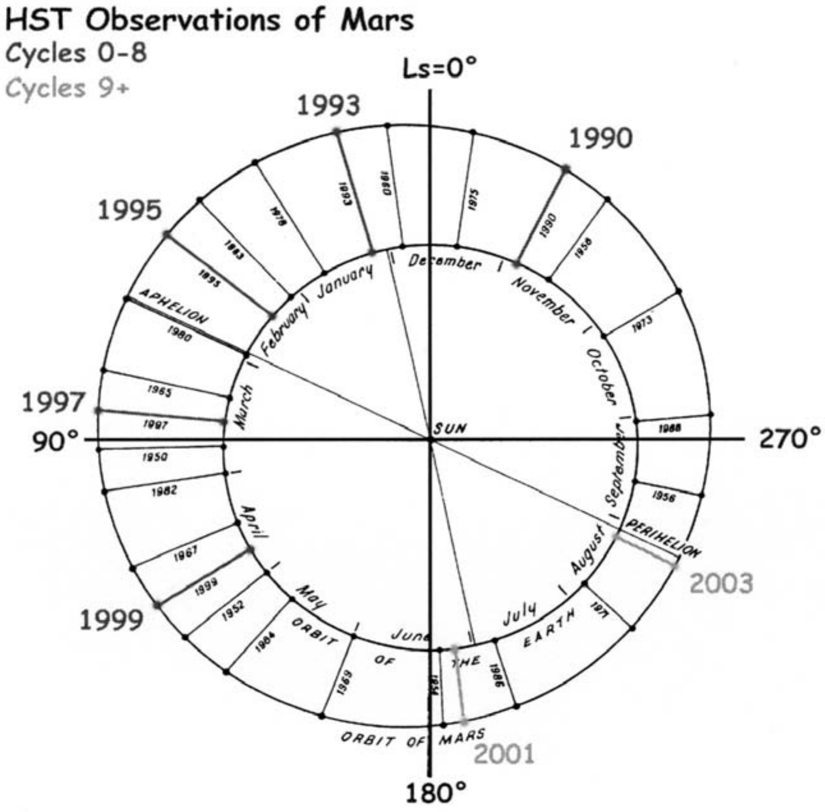


FIGURE 1.

in terms of spatial resolution occur near Martian aphelion, meaning that northern hemisphere summer is the least well studied period of the Martian year.

4.1. *Imaging*

Table 1 summarizes *HST* Imaging Observations during Cycles 0–8 (1990 to 1999). Early Cycle 0–3 WF/PC observations were limited in terms of resolution by the uncorrected spherical aberration of the *HST* primary. In addition, observing time was initially allocated in units of hours, allowing time for usually no more than 6–7 images per visit. As a result, the focus of imaging observations was on UV and blue exposures to study aerosols, and on a modest amount of red and green imaging in order to generate true color images of the planet. A small number of additional near-IR exposures was obtained to try to characterize the spectral behavior of the surface at longer wavelengths. Visit spacings were timed to try to sample the Martian seasonal cycle at intervals corresponding to the timescale of changes observed in the historic telescopic record (e.g. Martin et al. 1992). Occasional visits were conducted on the same Martian sol (a sol is one Martian "day." or 24 hours 37 minutes) in order to search for diurnal variations as well as to construct global maps covering all longitudes.

UT Date[a] YYMMDD	Time,[b] UT	Wavelengths, nm	Diameter, arcsec	SE Lat, deg	SE Lon, deg	Phase, deg	L_s, deg	Resolution,[c] km/pixel	PROGID[d]
HST Cycle 0 data: WF/PC									
901213a	10:34	230, 336, 439, 502, 588, 673, 889	16.6	−11.8	184.8	13.4	348.6	∼60[e]	3103, James
901213b	18:39	230, 336, 439, 502, 588, 673, 889	16.6	−11.8	303.1	13.6	348.8	∼60	3103, James
901214	02:43	230, 336, 439, 502, 588, 673, 889	16.6	−11.9	61.1	13.9	348.9	∼60	3103, James
HST Cycle 1 data: WF/PC									
910102	05:31	413, 502, 673	13.7	−13.1	290.7	25.9	358.7	∼68	3107, James
910207	04:49	230, 336, 413, 673	9.4	−10.0	305.3	36.2	1.6	∼100	3107, James
910320	06:43	413, 673	6.5	−1.8	302.8	37.3	35.3	∼145	3107, James
910514	16:53	230, 336, 413, 502, 673	5.0	11.7	282.2	31.7	59.9	∼188	3107, James
910515a	01:12	413, 673	5.0	11.7	43.6	31.6	60.0	∼188	3107, James
910515b	09:15	413, 673	5.0	11.8	161.1	31.6	60.1	∼188	3107, James
HST Cycle 2 data: WF/PC									
920530	03:40	413, 673	5.4	−22.6	300.4	34.5	259.0	∼175	3763, James
920611	20:56	413, 673	5.4	−20.4	73.7	35.9	267.0	∼175	3763, James
920627	22:57	336, 413, 673	5.8	−16.8	305.8	37.6	277.1	∼162	3763, James
920709	15:13	413, 673	5.8	−13.9	75.4	38.7	284.3	∼162	3763, James
921005	01:03	230, 336, 413, 502, 673	8.3	8.1	98.1	41.9	335.2	∼114	3763, James
921101a	06:00	336, 413, 502, 588, 673, 889	10.1	11.9	274.5	38.5	349.5	∼93	3763, James
921101b	15:13	336, 413, 502, 588, 673, 889	10.1	11.9	49.1	38.4	349.7	∼93	3763, James
HST Cycle 3 data: WF/PC									
930102a	04:19	413, 502, 588, 673, 889, 1042	15.1	8.3	47.5	5.5	20.0	∼62	4771, James
930102b	12:21	336, 413, 502, 588, 673, 889	15.1	8.2	165.1	5.3	20.1	∼62	4771, James
930102c	20:32	413, 502, 588, 673, 889	15.1	8.1	284.9	5.0	20.3	∼62	4771, James
930409	16:27	413, 502, 673	7.2	10.4	59.0	37.0	63.8	∼131	3763, James
930411	06:43	336, 413, 502, 673	7.2	10.7	257.8	37.0	64.5	∼131	3763, James
930616	15:13	413, 502, 673	5.0	22.5	106.7	32.4	93.6	∼189	3763, James

Table 1. HST WF/PC, WFPC2, and NICMOS observations of Mars: 1990–1999

UT Date[a] YYMMDD	Time,[b] UT	Wavelengths, nm	Diameter, arcsec	SE Lat, deg	SE Lon, deg	Phase, deg	L_s, deg	Resolution,[c] km/pixel	PROGID[d]
		HST Cycle 4 data: WFPC2							
940806	07:52	547	5.0	0.1	231.0	32.1	326.0	61.5[f]	5493, James
940823	23:00	255, 336, 410,502, 673	5.2	5.1	287	34.4	335.7	59.1	5493, James
940919	15:15	255, 336, 410,502, 673	5.7	11.8	273	36.5	349.7	54.0	5493, James
941020	11:43	255, 336, 410,502, 673	6.5	17.7	282	38.3	5.3	46.8	5493, James
941118	05:35	255, 336, 410,502, 673	7.8	21.0	275	38.0	19.1	39.3	5493, James
950102	09:56	255, 336, 410,502, 673	11.2	21.8	278	28.0	39.7	27.3	5493, James
950223	13:26	255, 336, LRFs (bad), 1042	13.5	17.6	231	9.7	62.9	22.6	5215, Crisp
950224	17:00	255, 336, 410,502, 673	13.5	17.3	274	10.0	63.1	22.7	5493, James
950225a	01:00	255, 336, 410,502, 673	13.5	17.2	31	10.3	63.6	22.7	5493, James
950225b	09:00	255, 336, 410,502, 673	13.4	17.2	148	10.5	63.7	22.7	5493, James
950225c	20:07	255, 336, LRFs (bad), 1042	13.4	17.4	311	11.4	63.9	22.7	5215, Crisp
950226a	00:58	255, 336, LRFs (bad), 1042	13.4	17.4	22	11.6	64.0	22.7	5215, Crisp
950226b	05:52	255, 336, LRFs (bad), 1042	13.4	17.4	94	11.7	64.1	22.7	5215, Crisp
950408	19:22	255, 336, 410,502, 673	9.8	18.1	281	32.6	81.9	31.4	5493, James
950411	02:05	336	9.6	18.1	0	33.4	83.3	32.0	5215, Crisp
950528	01:46	255, 336, 410,502, 673	6.8	23.0	270	41.7	104.1	45.2	5493, James
		HST Cycle 5 data: WFPC2							
950706a	03:20	255, 336, 410, 502, 673, 740, 860, 953, 1042	5.6	25.7	275	38.9	122.1	55.1	5832, James
950706b	11:23	255, 336, 410, 502, 673, 740, 860, 953, 1042	5.5	25.8	33	38.9	122.2	55.2	5832, James
950711	23:16	255, 336, 410, 502, 673, 740, 860, 953, 1042	5.4	25.8	157	38.1	124.8	54.0	5832, James
950802	21:21	255, 336, 410, 502, 673, 740, 860, 953, 1042	5.0	25.3	274	34.4	135.4	61.0	5832, James
950821	09:21	255, 336, 410, 502, 673, 740, 860, 953, 1042	4.8	23.5	273	30.8	144.6	64.4	5832, James

Table 1. *Continued*

UT Date[a] YYMMDD	Time,[b] UT	Wavelengths, nm	Diameter, arcsec	SE Lat, deg	SE Lon, deg	Phase, deg	L_s, deg	Resolution,[c] km/pixel	PROGID[d]
		HST Cycle 6 data: WFPC2							
960918	20:01	218, 255, 336, 410, 502, 588, 673, 953, 1042	4.6	16.6	161	29.3	11.3	66.7	6741, James
961008	16:10	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	5.0	20.3	272	31.9	20.7	61.3	6741, James
961009	00:13	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	5.0	20.3	30	31.9	20.9	61.3	6741, James
961015	13:54	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	5.1	21.3	169	32.7	24.0	60.1	6741, James
961129	17:58	218, 255, 336, 410, 502, 588, 673, 953, 1042	6.4	24.5	154	36.3	44.5	47.9	6741, James
961230	05:48	218, 255, 336, 410, 502, 588, 673, 953, 1042	8.0	24.0	42	35.6	58.0	38.3	6741, James
970104a	00:15	218, 255, 336, 410, 502, 588, 673, 953, 1042	8.3	23.8	270	35.1	60.1	36.9	6741, James
970104b	17:41	218, 255, 336, 410, 502, 588, 673, 953, 1042	8.3	23.8	167	35.0	60.4	36.9	6741, James
970310a	06:28	255, 336, 433, 467, 554, 763, 835, 893, 953	14.0	22.8	135	6.2	88.6	21.9	6852, Crisp
970310b	11:18	255, 336, 433, 467, 554, 763, 835, 893, 953	14.0	22.8	204	6.1	88.7	21.9	6852, Crisp
970310c	17:46	255, 336, 433, 467, 554, 763, 835, 893, 953	14.0	22.8	299	5.9	88.8	21.9	6852, Crisp
970330a	04:03	255, 336, 410, 467, 502, 547, 588, 631, 673, 1042	14.0	23.4	284	10.7	97.4	21.9	6741, James
970330b	10:30	255, 336, 410, 467, 502, 547, 588, 631, 673, 1042	14.0	23.4	18	10.9	97.6	21.9	6741, James
970330c	12:07	255, 336, 433, 467, 554, 763, 835, 893, 953	14.0	23.4	42	11.0	97.6	21.9	6852, Crisp
970330d	15:21	255, 336, 410, 467, 502, 547, 588, 631, 673, 1042	14.0	23.4	90	11.1	97.7	21.9	6741, James
970330e	22:06	255, 336, 410, 502, 588, 673, 1042	14.0	23.4	288	11.3	97.8	21.9	6741, James
970331	10:42	467, 554, 656, 763, 835, 893, 953, 1042	13.9	23.5	12	12.2	98.3	22.2	6793, Smith
970417	22:09	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	12.7	24.0	30	23.6	105.9	24.1	6741, James
970517	17:09	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	10.2	25.3	43	35.4	119.2	30.1	6741, James
970518a	02:35	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	10.1	25.3	178	35.4	119.7	30.4	6741, James
970518b	09:02	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	10.1	25.3	273	35.4	119.7	30.4	6741, James
970604	01:09	255, 336, 410, 467, 502, 547, 588, 673, 835, 953, 1042	9.1	25.8	357	38.4	126.3	33.7	6793, Smith
970627a	13:50	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	7.6	26.0	323	40.5	139.4	40.3	6741, James
970627b	17:04	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	7.6	26.0	9	40.5	139.4	40.3	6741, James
970627c	20:19	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	7.6	26.0	56	40.5	139.4	40.3	6741, James

Table 1. *Continued*

UT Date[a] YYMMDD	Time,[b] UT	Wavelengths, nm	Diameter, arcsec	SE Lat, deg	SE Lon, deg	Phase, deg	L_s, deg	Resolution,[c] km/pixel	PROGID[d]
		HST Cycle 7 data: WFPC2 and NICMOS							
970709	00:19	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	7.1	25.5	6	40.5	145.4	43.2	7276, James
970710	00:34	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	7.1	25.4	0	40.5	145.9	43.2	7276, James
970711	02:21	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	7.0	25.4	17	40.5	146.4	43.8	7276, James
970715	03:00	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	7.0	25.3	349	40.5	146.9	43.8	6793, Smith
970723	11:30	NICMOS: 950, 970, 1080, 1130, 1450, 1660, 1900, 2120, 2150, 2160, 2370	6.9	25.1	34	40.4	148.4	44.4	7276, James
970729a	13:55	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	6.5	24.0	12	40.0	154.6	47.1	6793, Smith
970729b	17:06	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	6.5	24.0	59	40.0	154.6	43.8	6793, Smith
970729c	20:20	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	6.5	24.0	106	40.0	154.6	43.8	6793, Smith
970812	02:28	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	6.4	23.5	67	39.7	156.7	47.9	7276, James
970829	12:39	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	6.0	21.1	50	38.5	165.2	51.1	7276, James
970901	14:52	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	5.7	18.2	53	37.2	173.4	53.8	7276, James
970912	16:40	410, 502, 673, 1042	5.5	15.4	332	36.0	179.9	55.8	7792, DD
970918	02:10	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	5.4	14.0	62	35.4	182.9	56.8	7276, James
970923	01:04	410, 502, 673, 1042	5.3	12.7	359	34.9	185.8	57.9	7792, DD
970930	08:59	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	5.2	10.6	46	34.0	190.0	59.0	7276, James
971001a	01:08	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	5.2	10.4	282	33.9	190.4	59.0	7276, James
971001b	17:15	255, 336, 410, 502, 588, 673, 763, 835, 953, 1042	5.2	10.4	157	33.9	190.4	59.0	7276, James
971005	00:18	410, 502, 673, 1042	5.2	9.2	233	33.4	192.7	59.0	7792, DD
971009	01:07	410, 502, 673, 1042	5.1	8.0	203	32.9	195.1	60.2	7792, DD
		HST Cycle 8 data: WFPC2							
990427	17:55	255,336,410,502,547,588,631,673,763,835,953,1042	16.2	19.0	17	2.7	130.5	19.4	8152, Bell
990428	00:22	255,336,410,502,547,588,631,673,763,835,953,1042	16.2	19.0	111	2.9	130.6	19.4	8152, Bell
990501	13:47	255,336,410,502,547,588,631,673,763,835,953,1042	16.2	19.5	281	5.9	132.4	19.4	8152, Bell
990506	11:28	255,336,410,502,547,588,631,673,763,835,953,1042	16.1	20.3	205	10.0	134.7	19.5	8152, Bell

[a] Read 940823 as August 23, 1994. a, b, c, etc. indicates first, second, third, etc. set of images obtained on that day.

[b] Time given as the start of the ∼ 25 to 50 minute observing sequence.

[c] Resolution is the maximum spatial resolution at the sub-Earth point for images obtained on the PC or NIC1 chip.

[d] Space Telescope Science Institute Program Identification number and Principal Investigator, for *HST* data archive access.

[e] Cycles 0–3 resolution takes into account blurring of WF/PC images due to uncorrected spherical aberration in *HST* primary.

[f] Cycles 4–8 resolution takes into account COSTAR spherical aberration correction for WFPC2.

Table 1. *Continued*

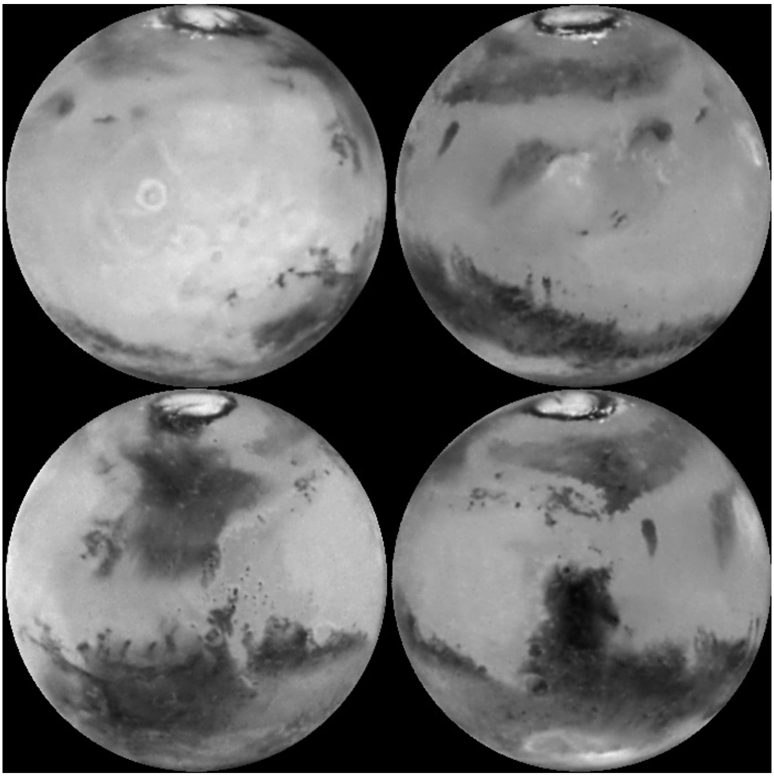


Figure 2.

After Cycle 4 the spherical aberration was corrected by COSTAR, and WFPC2 was able to realize the full diffraction-limited resolution of the system (Figure 2). Also, time allocation was changed to integer numbers of orbits rather than hours, meaning that additional exposures could be obtained during each visit. The important focus on blue/UV imaging was maintained, as was the seasonal sampling, but additional orbit time meant that more near-IR images could be obtained in order to characterize the surface colorimetric, mineralogic, and photometric properties. In additional to the standard filters, several of the WFPC2 linear ramp filters (LRF) were employed beginning in Cycle 5, in order to obtain photometry at key wavelengths diagnostic of broad solid state mineral absorption features.

The only NICMOS Mars observations to date were performed during a single dedicated orbit in July 1997. These observations were designed to sample the Martian near-IR reflectance spectrum at key wavelengths diagnostic of atmospheric CO_2, CO, and H_2O as well as solid state mineral absorptions.

4.2. *Spectroscopy*

Table 2 summarizes *HST* spectroscopic observations during Cycles 0–8. Spectroscopy of Mars in the UV was performed using both FOS and STIS in order to characterize the composition, opacity, and spatial/temporal distribution of atmospheric aerosols. The instrument apertures were frequently positioned across both the limb and terminator in order to maximize atmospheric path length and to detect diurnal variations. FOS "pushbroom" scans and STIS slit-scans were programmed in order to obtain wider spatial coverage as well as measurements spanning specific latitudes/longitudes. STIS long-slit

UT Date YYMMDD	Time, UT	Inst. & Central Wavelength	Diam., arcsec	SE Lat, deg	SE Lon, deg	Phase, deg	L_s, deg	PROGID
			HST Cycle 1 data: FOS					
910102	07:06	FOS, 2650 Å	13.7	−13.1	314.1	25.9	358.7	3107, James
910207	06:27	FOS, 2650 Å	9.4	−10.0	329.2	36.2	16.2	3107, James
			HST Cycle 2 data: FOS					
920627	18:04	FOS, 2650 Å	5.8	−16.9	234.6	37.6	277.0	3763, James
920824	07:45	FOS, 2650 Å	6.8	−1.5	239.5	41.9	311.8	3763, James
			HST Cycle 3 data: FOS					
930102	05:58	FOS, 2650 Å	15.1	8.3	71.7	5.5	20.0	4771, James
930409	14:44	FOS, 2650 Å	7.2	10.4	33.9	37.0	63.8	4771, James
			HST Cycle 4 data: FOS					
950224	18:25	FOS, 2650 Å	13.7	17.5	294.8	10.6	63.5	5493, James
			HST Cycle 6 data: FOS					
960918	18:40	FOS, 2650 Å	4.7	16.8	139.5	29.3	11.3	6741, James
970104	16:18	FOS, 2650 Å	8.3	23.8	144.8	35.0	60.4	6741, James
			HST Cycle 7 data: STIS					
970708	13:28	STIS, 2375 Å	7.2	25.6	208.0	40.6	144.7	7276, James
970724	08:21	STIS, 7751 Å	6.5	24.4	338.0	40.1	152.7	7276, James
970827	10:31	STIS, 2375 Å	5.8	19.1	38.3	37.7	170.9	7276, James
971001	03:18	STIS, 2375 Å	5.4	10.4	311.1	33.9	190.4	7276, James
			HST Cycle 8 data: STIS					
990427	19:53	STIS, 7751 Å	16.2	19.0	45.4	2.8	130.5	8152, Bell
990501	15:46	STIS, 7751 Å	16.2	19.6	310.2	6.0	132.4	8152, Bell
990506	13:26	STIS, 7751 Å	16.1	20.3	232.3	10.1	134.8	8152, Bell
990507	07:14	STIS, 7751 Å	16.1	20.4	132.7	10.7	135.1	8152, Bell

Table 2. *HST* FOS and STIS observations of Mars: 1990–1999
(Table heading abbreviations as in Table 1.)

spectroscopy in the visible was performed in Cycles 7 and 8 using slit scanning to build up 3-dimensional image cubes (spatial × spatial × spectral); due to a commanding error the scan did not work successfully in Cycle 7, but the Cycle 8 scans executed flawlessly.

5. Results

Many scientific results have been published in the peer-reviewed literature from the ∼ 10 years of *HST* Mars observations. This section summarizes the major findings and discusses their implications for both science and NASA mission support. Additional details can be found in the publications cited along with each of the results discussed.

5.1. *Martian atmospheric studies*

5.1.1. *Abundance and spatial distribution of O_3 and H_2O*

Ozone and water are only trace constituents of the Martian atmosphere (0.04 to 0.2 ppm and ∼ 0.03%, respectively; Owen 1992), but their photolysis and recombination play critical roles in Martian atmospheric photochemistry. Specifically, the long-term stability of CO_2 against photolytic breakdown ($CO_2 + h\nu \rightarrow CO + O$) is maintained by the breakdown and recombination of O_3 and H_2O in the Martian atmosphere (e.g. McElroy & Donahue 1972; Clancy & Nair 1996). Ozone has a photochemical lifetime of only a few hours in the Martian atmosphere, but it has been shown by Mariner 9 measurements to undergo large seasonal variations in response to changes in the Mars atmospheric water vapor profile (e.g. Barth et al. 1973). Ozone is difficult to measure on Mars from groundbased or spacecraft techniques, but *HST* can routinely provide excellent data on Mars ozone by UV spectra and imaging within the strong $\sim$ 260 nm O_3 Hartley absorption band. James et al. (1994) used WF/PC UV imaging and an atmospheric scattering model to constrain ozone abundances and atmospheric dust and water ice cloud opacities during late northern winter ($L_s \sim 350°$). Clancy et al. (1996) analyzed FOS spectral scans of Mars from Cycle 4 data (Table 2) to derive ozone column abundances and water ice cloud opacities during the Martian aphelion season. They found twice the ozone abundance as that reported from perihelion IR spectroscopy measurements (Espenak et al. 1991), consistent with photochemical and observational modeling (Clancy & Nair 1996; Clancy et al. 1996) that predicted such a seasonal middle-atmosphere ozone increase associated with lower aphelion atmospheric temperatures and lowering of the altitude of water vapor saturation (see below). This aphelion enhancement of the ozone was observed again in the next Mars year (1996–7) using FOS data (Clancy et al. 1999). *HST* observations currently provide the only way to routinely monitor the variability of ozone, and by inference water vapor, in the Martian atmosphere. As described below, *HST* observations have had a profound impact on current thinking about the stability of the Martian climate.

5.1.2. *Opacity and spatial/temporal variability of atmospheric dust and water ice clouds*

The ability to observe in the UV and to perform accurate flux calibration of *HST* imaging and spectroscopic measurements across the UV to near-IR has enabled substantial progress in the modeling of aerosol opacity in the Martian atmosphere. For example, James et al. (1994) used WF/PC images and a multiple-scattering radiative transfer model to estimate dust and water ice cloud opacities and their radiative influence on the Martian climate. They found low dust opacities ($\tau < 0.1$–0.2) in their December 1990 and May 1992 observations, and water ice cloud opacities near 0.2 for the winter "polar hood" clouds and typically < 0.1 for orographic or early morning limb clouds. Modeling of 1995 WFPC2 images by Wolff et al. (1997) during visibly dusty conditions (Figure 3) allowed quantitative estimates of the dust and water ice cloud opacity as well as refined values of the dust single scattering albedo. They noted the correlation between elevated dust opacities and elevated atmospheric temperatures derived from near-simultaneous microwave imaging, and postulated that the dust activity seen in the 1995 *HST* images likely followed a large regional or possibly even global dust storm which was not noticed by other groundbased observing methods. Most recently, Wolff et al. (1999) analyzed 1992–97 *HST* images to determine dust and water ice cloud opacities and to constrain further the dust single scattering albedo. Their results are consistent with near-simultaneous opacity values derived from the surface by Mars Pathfinder (Smith et al. 1997) as well as by those obtained from MGS (Clancy et al. 2000). The *HST* and MGS opacity values indicate that Mars was generally less dusty during 1992–1997 than during the same seasons during the Viking Lander missions of the late 1970s.

5.1.3. *Changes to the Viking-era climate paradigm?*

Based on a combination of groundbased microwave profiling and *HST* observations, Clancy et al. (1996) proposed a modification in the Viking-era interpretation of the Martian climate. In essence, their analysis indicated that the aphelion season on Mars

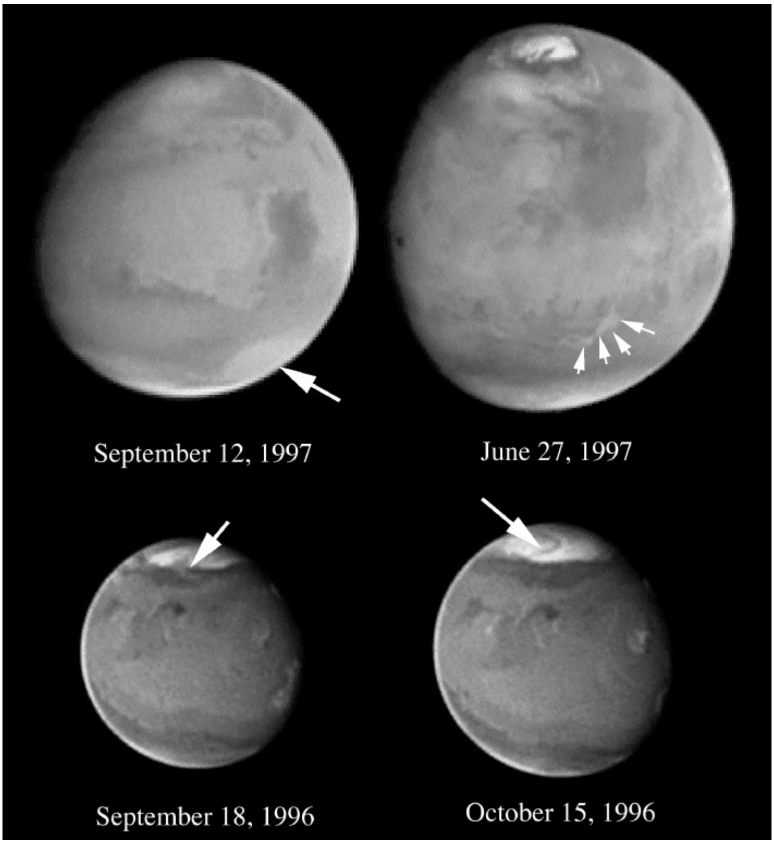


FIGURE 3.

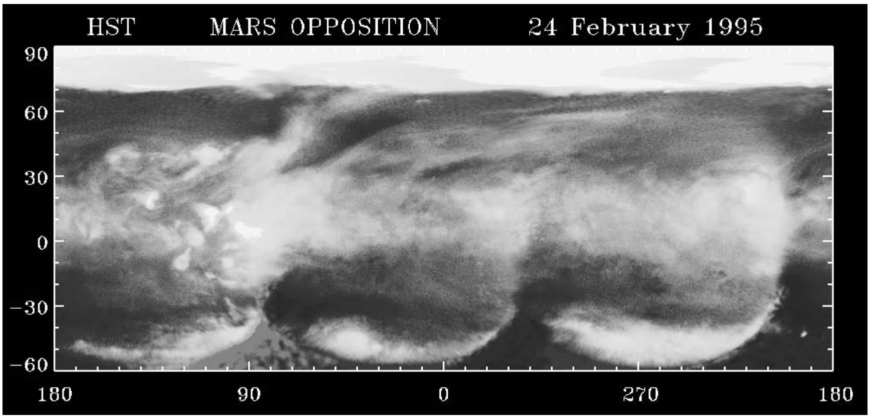


FIGURE 4.

was now colder and less dusty than during the Viking mission. Solar insolation varies by ∼ 40% during the Martian year because of the planet's eccentricity, leading to 30–40 K annual variations in average atmospheric temperatures (in the absence of any other sources of radiative warming such as a large dust opacity). The lower temperatures produce a lower altitude (< 10 km) of water vapor saturation. One directly observable consequence of this different climate is that the aphelion atmosphere of Mars should

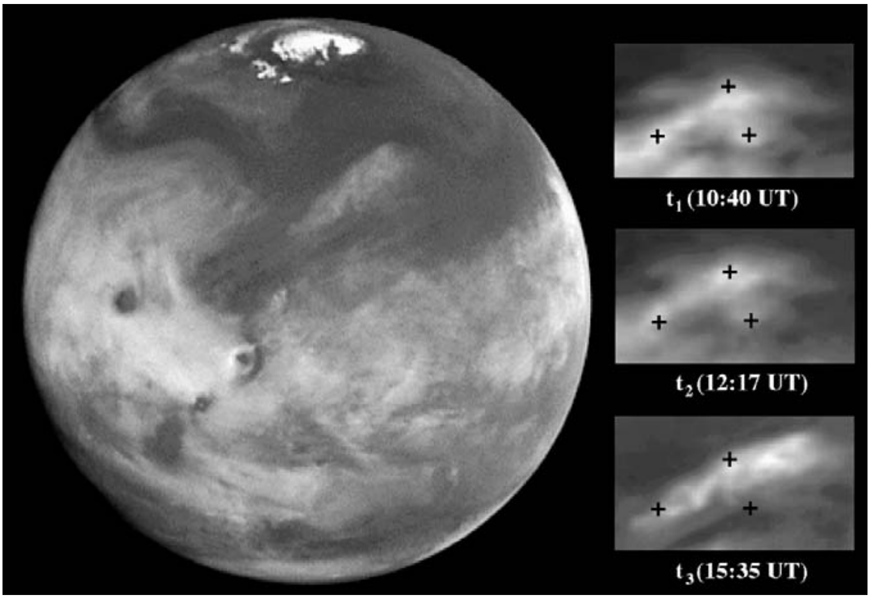


Figure 5.

be extremely cloudy. Until quite recently, it was thought that such cloudiness was not present during the multi-year Viking Orbiter missions of the late 1970s to early 1980s. This fact was addressed by Clancy et al. (1996) with the suggestion that the Viking-era atmosphere was much more dusty than "usual," and that this elevated dust provided a source of radiative heating sufficient to prevent the atmosphere from transitioning to the cooler, cloudier aphelion state. Groundbased optical and IR observations of Mars near aphelion are typically of poor quality because even at opposition the planet is always far from Earth (Figure 1), and so *HST* observations in the mid-1990s provided the first opportunity to test (separately from the microwave data) this "cold and cloudy" vs. "warm and dusty" hypothesis. Figure 4 shows a composite map of 410 nm images from February 1995—an equatorial "belt" of enhanced cloudiness is apparent in these and other *HST* image near aphelion (as well as some older photographic data reported by Slipher 1962), consistent with the Clancy et al. (1996) aphelion climate. A quantitative *HST* study of water ice cloud opacity during the 1990s by Wolff et al. (1999) corroborated the presence of an extensive aphelion equatorial cloud belt for three sequential Martian years. The recent discovery of a bias in the Viking data used to derive atmospheric temperatures (previous ones are 15 to 20 K too warm) (Richardson & Wilson, 2000) and the apparent presence of the aphelion cloud phenomenon in some of the Viking observations (Tamppari et al. 2000) have further highlighted the reality of the Clancy et al. finding. Although much of the original motivation for the Clancy et al. climate model derived from microwave measurements, *HST* observations of the aphelion cloud belt provided a critical piece of evidence that could not be ignored.

5.1.4. *Atmospheric circulation*

The advent of WFPC2 observations during Cycle 4 provided the ability to obtain spectacular UV and blue images of fine-scale structure in the Martian atmosphere (Figure 5), thus allowing cloud feature tracking to derive wind speed and wind direction data and to compare them with predictions from General Circulation Models (GCMs) for specific seasons. Mischna et al. (1998) analyzed Cycle 6 WFPC2 UV/blue images and identified

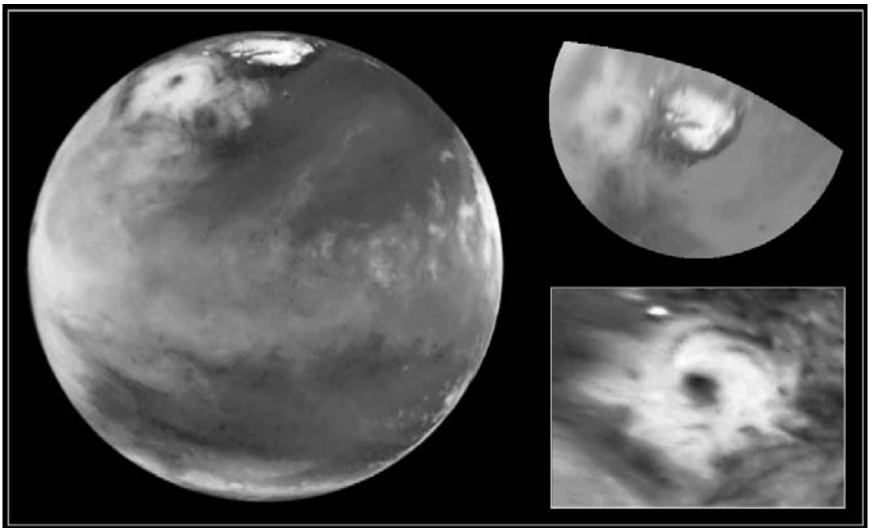


Figure 6.

specific cloud features that could be tracked over a 5 hour time span between images. They found average wind speeds of $\sim 30 \pm 10$ m/sec for the features studied. Both the speeds and the determined wind directions of the clouds were found to be consistent with GCM predictions for early northern summer midlatitudes (Haberle et al. 1993). More recently, Gierasch & Bell (2000) report the detection of a huge cyclonic storm system in Cycle 8 WFPC2 imaging during northern summer (Figure 6). This spiral storm is similar in form to those seen during summer at high northern latitudes by the Viking Orbiter mission (Gierasch et al. 1979) except that it was much larger. Groundbased amateur observers (M. Minami, pers. comm., 1999) report possible evidence for the existence of the storm at least several days before the *HST* observations, but additional *HST* imaging just 6 hours after the discovery observations show it to have rapidly dissipated. MGS Mars Orbiter Camera (MOC) or laser altimeter observations were not obtained at the same time as the *HST* data, but MOC data taken a few days later show evidence for similar short-lived spiral structures in the north polar region (M. Malin, pers. comm., 1999). The *HST*, MOC, and Viking data combined indicate that spiral water ice storms are a common feature of the northern summer polar latitudes. Gierasch and Bell (2000) postulate that these systems may be fueled by the strong temperature gradient between the cold polar atmosphere and the warmer surface of the dark circumpolar sand dunes, much like "polar low" storms on the Earth are fueled by oceanic temperature gradients.

5.2. *Martian surface studies*

5.2.1. *Oxidation state of Fe in the surface materials*

Groundbased telescopic observations of Mars at visible wavelengths have previously revealed that the surface has a steep red spectral slope caused by the presence of oxidized surface minerals with Fe(III) solid state absorptions in the blue to UV (e.g. Adams & McCord 1969; Singer et al. 1979; Bell et al. 1990). Evidence for Fe(II) absorption features caused by relatively unoxidized surface materials, primarily in low albedo regions, has also been presented in these same studies from measurements in the near-infrared, especially near 1.0 μm. The region near 1.0 μm includes absorption from both Fe(II) and Fe(III) solid state mineral features, however, and so inferences regarding the oxidation state of low albedo regions of Mars must properly separate the effects of both ferric and ferrous

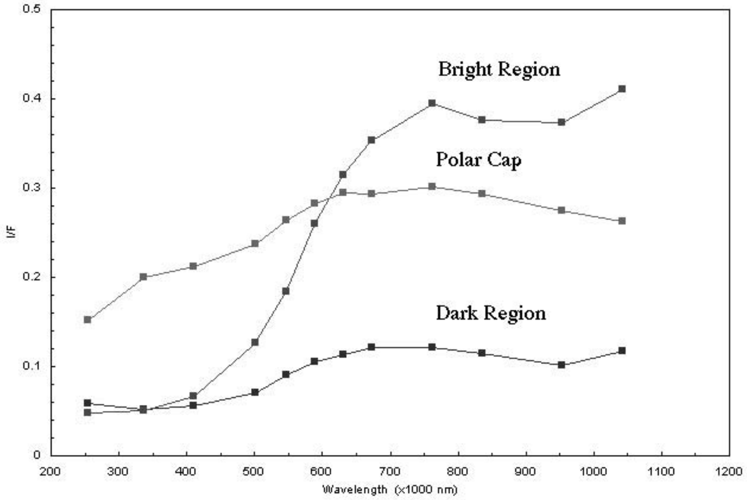


FIGURE 7.

mineral bands in the near-IR (e.g. Morris et al. 1995). Such unmixing is confounded in groundbased studies by the presence of strong telluric absorption near 0.95 μm which can hamper or destroy the ability to proper model the spectral reflectance across the 1 μm band. *HST* measurements are free of telluric contamination and thus provide the ability to properly measure and model this important spectral region. Examples of WFPC2 spectra assembled from multiple narrowband images are shown in Figure 7 (Bell et al. 1999). These data, combined with *HST* multispectral imaging from earlier Cycles (James et al. 1996; Bell et al. 1997), has confirmed that most low albedo, classical “dark” regions of Mars contain a substantial Fe(II) component diagnostic of relatively unweathered and unoxidized volcanic material like pyroxene, mixed with and/or partially covered by small amounts of a much more heavily oxidized Fe(III) component diagnostic of poorly crystalline or nanophase iron oxide. However, the *HST* studies also reveal the existence of anomalous dark regions that appear dominated by ferric rather than ferrous minerals, and may represent the discovery of a new class of surface materials with spectral properties consistent with *hydrated* ferric oxides like goethite (α-FeOOH) or ferrihydrite ($\sim$ 5 $Fe_2O_3 \cdot 9H_2O$) rather than anhydrous ferric oxides like hematite (α-Fe_2O_3) (Bell, 1992; Murchie et al. 1993, 2000; Bell & Morris 2000). If confirmed by additional telescopic and spacecraft investigations, this discovery may indicate regions on Mars where liquid water was extensively involved in the weathering and alteration of surface materials. While OH- or H_2O-bearing minerals should be unstable and should dehydroxylate under present Martian climatic conditions (very low p(H_2O), high UV flux), they could exist metastably because of extremely low temperatures and/or burial within a mixed regolith (e.g. Burns, 1993, Yen et al. 1999).

5.2.2. *Global-scale spectral unit mapping*

The multispectral properties of surface materials on Mars provide diagnostic information on their composition and physical nature. During the Viking Orbiter investigation in the mid 1970s, imaging observations as the spacecraft approached Mars were used to discriminate distinct surface units based on their 3-color (blue, green, red) spectral properties (Soderblom et al. 1978; McCord et al. 1982). Differences in color were related to potential differences in mineralogy, oxidation state, particle size, and/or degree of surface

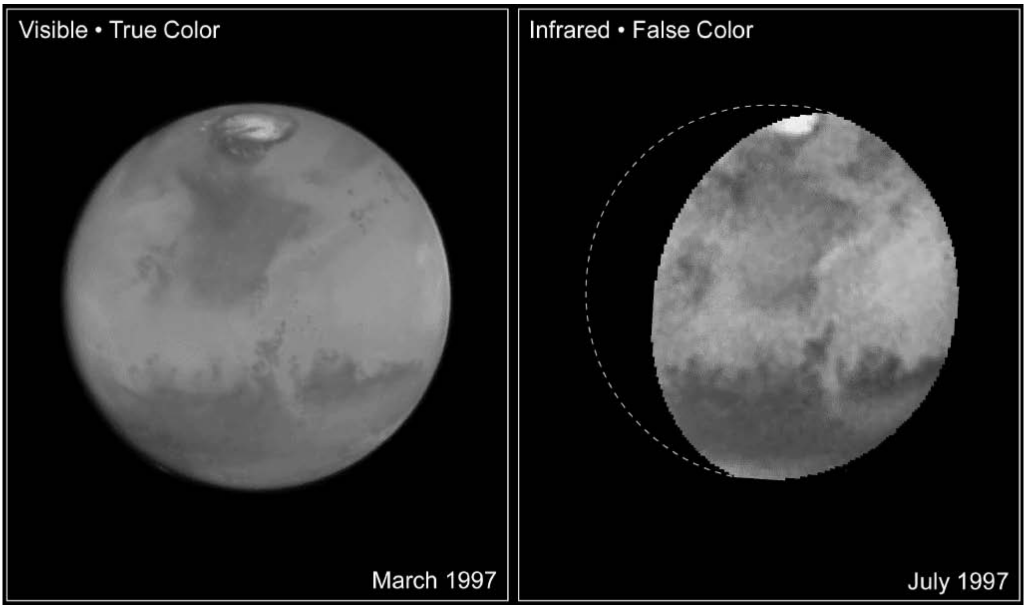


Figure 8.

compaction (fine dust vs. cemented duricrust). *HST* provides unique new data that can extend the Viking-era spectral unit study to higher latitudes and longer wavelengths at a comparable spatial scale (tens of km/pixel). Initial analyses of eastern hemisphere spectral units by Bell et al. (1997) showed broad consistency with the Viking-derived spectral units, but was augmented by the ability to discriminate further by using WFPC2's near-IR filters to reveal additional reflectance trends not identifiable in previous (or current) spacecraft datasets.

5.2.3. *Unique surface mineralogic deposits*

The combination of WFPC2 and NICMOS imaging observations of Mars provides the ability to cover broad regions of the solar reflectance spectrum at spatial scales unobtainable from groundbased telescopes and in spectral regions not being measured by spacecraft, and thus to identify unique spectroscopic signatures from surface materials. Two examples have been reported from preliminary analyses of Cycle 7 and 9 Mars images. First, NICMOS observations during Cycle 7 provided evidence for the presence of spatial variations in the distribution of H_2O- or OH-bearing minerals on the surface (Bell et al. 1998). The Martian surface has long been known to contain a small amount (1 to 3%) of such hydrated minerals, although their specific mineralogic identity has not been determined (e.g. Houck et al. 1973; Pimentel et al. 1974). The NICMOS data, showing variations in the strength of an H_2O vibrational overtone absorption near 1450 nm (Figure 8), have not revealed the identity of the hydrated mineral(s) on the surface, but they have revealed spatial variations in the abundance and/or composition of the hydrated material, especially within the classical dark regions. Continuing analyses are searching for correlations between these variations and other geologic or spectroscopic properties of the surface, in the hopes of identifying the specific mineralogy. Second, WFPC2 Cycle 6 observations in the near-IR revealed the presence of extremely strong absorption near 953 nm isolated to a dark ring of material surrounding the north polar cap (Bell et al. 1996, 1997; Figure 9). This dark material is known to include large expanses of sand dunes and other aeolian features based on Viking Orbiter imaging. The *HST* data have been

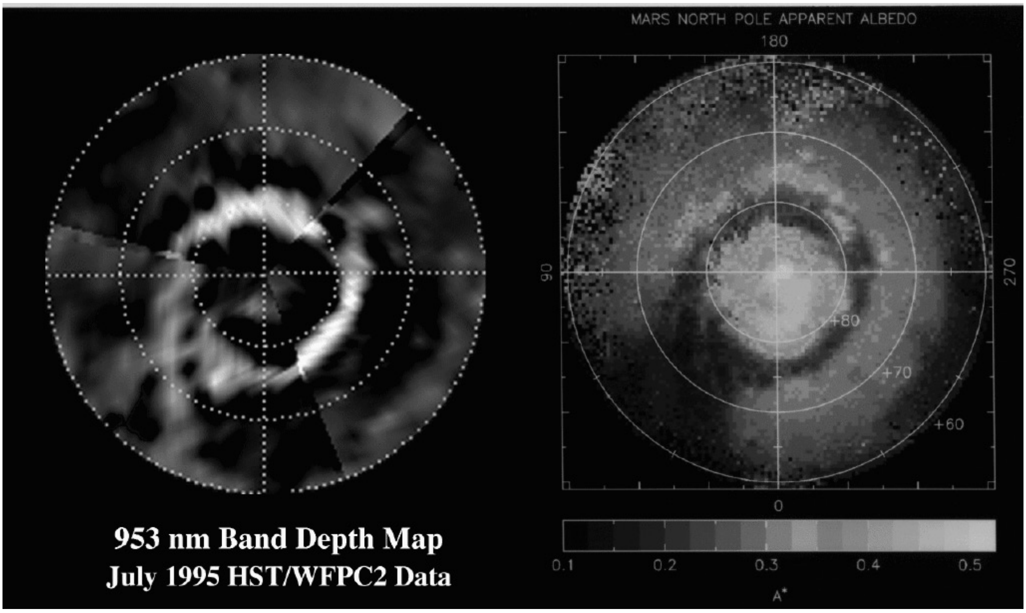


Figure 9.

interpreted to indicate the presence of extremely "fresh" or unoxidized deposits of the igneous mineral pyroxene. The pyroxene signature in the north polar sand sea appears broadly similar to that observed from groundbased telescopes and *HST* in other dark regions of the planet (e.g. Singer et al. 1979; Bell et al. 1997), but this geologic region has either a higher pyroxene abundance, a larger pyroxene grain size, and/or a different composition that leads to a stronger band depth. One proposed model accounts for the stronger bands through a comet-like "sublimation gardening" effect, whereby grains are continually being jostled and freed of dust coatings by the diurnal and seasonal exchange of CO_2 between the polar surface and atmosphere (Bell et al. 1997). This and other models are being tested and refined using higher resolution and extended wavelength coverage of the same region of the planet obtained by WFPC2 during Cycle 8 (Bell et al. 2000).

5.2.4. *Surface albedo changes on seasonal and interannual timescales*

Mars has exhibited changes in surface albedo markings throughout the history of telescopic observations. These variations in albedo markings are caused by aeolian deposition or removal of bright, red, heavily oxidized dust over regions of intrinsically lower albedo. When seasonal or interannual wind circulation patterns change, dust that was preferentially deposited after local or global-scale dust storms can be swept clean, exposing a darker surface. Telescopic and spacecraft observations have shown that the dust particles are only a few microns in size on average (e.g. Pollack et al. 195; Ockert-Bell et al. 1997), and thus they are easily transportable by winds of a few tens of meters per second in the thin Martian atmosphere. Laboratory studies have shown that only a few tens of microns or less of dust are required to optically brighten a low albedo surface (Wells et al. 1984; Johnson et al. 2000). Therefore, changes in the surface albedo distribution can be used to identify changes in global atmospheric circulation patterns, and serve as a proxy for understanding the variability of the current Martian climate. *HST* has provided new insight in this area by revealing surface albedo changes over a $\sim$ 10 year period and at an unprecedented spatial scale. The most prominent reported change is the

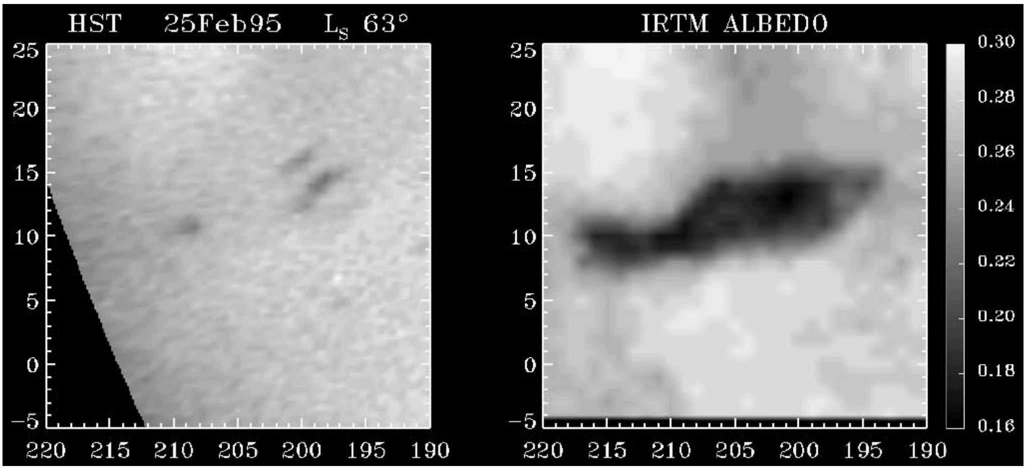


FIGURE 10.

near-complete disappearance of a California-sized low albedo region called Cerberus in the planet's eastern hemisphere (Lee et al. 1996), between the end of the Viking mission in the early 1980s and *HST* observations in the 1990s (Figure 10). Apparently, changing atmospheric circulation patterns during this time resulted in the preferential deposition of windblown dust on top of this feature. Groundbased thermal IR observations of this part of the surface, combined with the *HST* observations, indicate a layer of dust perhaps as thick as 1–2 mm deposited onto the surface over a 5 to 10 year period (Moersch, 1998). The implied dust deposition rate ($\sim 400\ \mu$m/yr) is at the very high end of that inferred from other types of observations, and may indicate the presence of substantial regional or global-scale dust storm activity during the post-Viking but pre-*HST* period.

5.2.5. *Growth and recession of the polar caps*

The most obvious and measurable manifestation of the changing Martian seasons is the waxing and waning of the planet's seasonal polar ice caps. Each fall and winter, nearly 25% of the atmospheric mass condenses out at the poles to form meters-thick deposits of CO_2 frost. Each spring and summer, the CO_2 sublimes back into the atmosphere, exposing in the north a residual water ice polar cap, and in the south either bare surface some years or a small H_2O+CO_2 residual ice cap. Variations in the growth and decay rates of the seasonal polar caps have been observed for centuries telescopically, and these variations reveal short-term fluctuations in the current Martian climate (James et al. 1992). *HST* provides the ability to monitor polar cap recession at scales comparable to previous Mariner 9 and Viking Orbiter imaging, and thus to make detailed comparisons of year-to-year climate variability based on small-scale feature variations. Both James et al. (1996) and Cantor et al. (1998) have analyzed *HST* images of Mars polar cap retreats during the 1990s and have found evidence for differences between the behavior of the cap in the 1990s and that reported from previous epochs. The source of the changes is thought to be related to variations in annual dust storm variability, which can substantially change the seasonal variation of the diurnal atmospheric temperature profile.

5.2.6. *High resolution imaging spectroscopy*

Most recently, *HST* and STIS have been used to obtain imaging spectroscopic data of Mars in the visible to near-IR and at high spatial and spectral resolution (Bell et al.

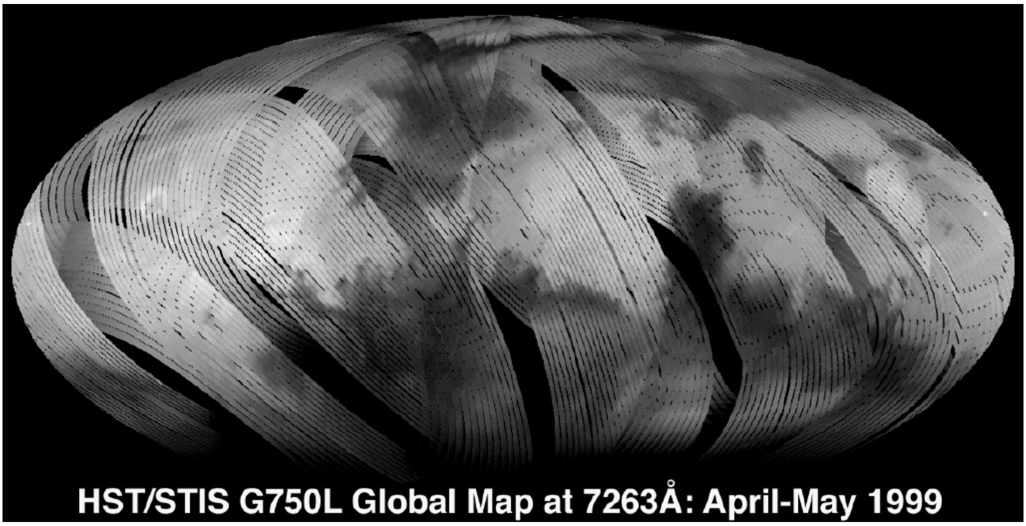


FIGURE 11.

1998, 1999). These data were obtained via a challenging series of *HST* measurements that were obtained by scanning the STIS long slit across the planet and obtaining individual spatial x spectral slit images at each scan position. By merging the individual slit images and mapping them onto the planet, a cartographically-registered 3-dimensional cube (spatial × spatial × spectral) was created covering more than 80% of the planet between ~ 5000 to 10,000 Å at ~ 5 Å resolution (Figure 11). Initial analyses of these data (Bell & Wolff 2000) reveals evidence for subtle variations in Fe(II) and Fe(III) solid state mineral absorption features that cannot be seen in the coarsely-sampled WFPC2 spectra or in groundbased spectra that include telluric absorption in the near-IR.

5.3. *Martian satellites*

HST has also been used to observe the small Martian satellites Phobos and Deimos, whose UV and visible spectral properties have been the subject of some controversy because of difficulties in completely removing contaminating Mars background flux. Observations and analyses by Zellner & Wells (1994) using FOS and Cantor et al. (1999) using WFPC2 clearly separated the satellites from background Mars flux and confirm that Phobos does not have a spectrum that matches a simple carbonaceous meteorite analog material. The leading hemisphere of Deimos is much redder than the trailing hemisphere. While the Deimos leading hemisphere resembles the spectra of D-type asteroids (Zellner & Wells 1994), there are no simple meteorite analogs that match its spectrum.

6. Future prospects

Figure 1 reveals that the first few oppositions of the 21st century provide excellent opportunities for *HST* studies of Mars. Earth-Mars closest approach distances during 2001 and 2003 are the lowest possible during the 15–17 year opposition cycle, and in fact the planet is closer to Earth during the 2003 opposition (0.372 AU) than it has been for the past several thousand years (and it will not get this close to Earth again until 2287) (Goffin & Meeus 1978). This close alignment will provide the opportunity for unprecedented telescopic spatial resolution at opposition (10–15 km/pixel), and very high spatial resolution over a large fraction of the Martian seasonal cycle away from opposition. In addition, the equatorial plane of Mars is near the plane of the ecliptic during the June 2001 opposition, providing the opportunity for extremely low phase angle studies of the Martian satellites as well as a possible search for a Martian ring (Hamilton 1996).

In addition to these orbital/geometric prospects, improvements and upgrades to *HST* systems and instruments will provide additional opportunities for Mars studies in upcoming years. These include the re-commissioning of NICMOS, the installation and availability of the ACS and WFC3, enhancements to *HST*'s pointing and guiding capabilities, and improvements and enhancements to both the WFPC2 and STIS data reduction and calibration pipelines.

7. Conclusions

HST observations of Mars continue to provide important new information about the planet's surface and atmosphere over unique wavelength intervals and spatial/temporal scales that are highly complementary to ongoing and planned Mars spacecraft investigations. New discoveries and insights have been made by *HST* concerning ozone, dust, and water vapor opacity and distribution and their radiative effects on the atmospheric energy balance; on evidence for recent "climate change" between the ~ 20 years between Viking and *HST* observations of Mars; on the dynamics of dust and water vapor transport in the Martian atmosphere and their relationship to GCM predictions; on the oxidation state of the surface, spatial distribution of surface mineralogy, and their implications for past climatic variations; and on changes in the surface albedo distribution on interannual and decadal timescales related to variations in aeolian dust cover and seasonal CO_2 and H_2O ice condensation and sublimation. Highly favorable Earth-Mars viewing conditions over the first few oppositions of the 21st century plus planned telescope and instrument upgrades and improvements will provide continued opportunities for significant and unique contributions to Mars studies using *HST*.

The results described here are based on the work of many PIs and Co-Is on GO and GTO programs during the past decade. I am especially grateful to Mike Wolff (SSI) for his review of an earlier draft of this manuscript. I also thank P. James (U. Toledo); D. Crisp and J. Trauger (JPL); P. Smith (LPL), R. T. Clancy (SSI), S. Lee (LASP), M. Mischna (Penn State), and the rest of the WF/PC and WF/PC2 science teams for sharing their results and insights. *HST* observations of Mars often place substantial demands on both spacecraft and personnel resources. The hard work and dedication of many people have helped to make these observations successful over the past decade. Specifically, those of us involved in *HST* Mars science thank the following people for critical support: S. Baggett, C. Cox, J. Christensen, J. Krist, Z. Levay, A. Lubenow, M. McGrath, K. Noll, K. Rudloff, E. Smith, R. Villard, and D. Weaver (ST ScI); R. Evans and K. Stapelfeldt (JPL); B. Cantor (U. Toledo); B. Cichy and T. Daley (Cornell); H. Ford and D. Golimowski (JHU); R. Singer (LPL); and D. Schroeder (NOAO). Finally, I would like to dedicate this work to the late Leonard J. Martin of Lowell Observatory, a Co-I on many *HST* Mars GO programs and expert on the geography and climatology of the red planet. His insights and telescopic experience were invaluable in the planning and analysis of the data described here, and his enthusiasm for the beautiful images produced by *HST* was inspirational. This research has been supported by grants from the Space Telescope Science Institute (ST ScI GO-08152.01-97A) and the NASA Planetary Geology and Geophysics Program (NAG5-4333).

REFERENCES

Adams, J. B. & McCord, T. B. 1969 *J. Geophys. Res.*, **74**, 4851.

Barth, C. A., Hord, C. W., Stewart, A. I., Lane, A. L., Duck, M. L., & Anderson, G. P. 1973 *Science*, **179**, 795.

Bell III, J. F. & Morris, R. V. 1999 in *Lunar and Planet. Sci. XXX*, Abstract # 1751. LPI.

Bell III, J. F., Cloutis, E. A., Morris, R. V., Wolff, M. J., & Gordon, K. D. 2000 *Eos Trans. AGU*, in press.

Bell III, J. F., McCord, T. B., & Owensby, P. D. 1990 *J. Geophys. Res.*, **95**, 14447.

Bell III, J. F., Wolff, M. J., James, P. B., Clancy, R. T., Lee, S. W., & Martin, L. J. 1997 *J. Geophys. Res*, **102**, 9109,

Bell III, J. F., Wolff, M. J., Comstock, R., & James, P. B. 1998 *B.A.A.S.*, **30**, 1054.

Bell III, J. F. & Wolff, M. J. 2000 in *Lunar and Planet. Sci. XXXI*, Abstract # 1223. LPI.

Bell III, J. F., Wolff, M. J. Glotch, T., Lee, S. W., Martin, P. D. James, P. B. Ravine, M. & Clancy, R. T. 1999 *Eos, Trans. AGU 80*, F627.

Bell III, J. F., Thomas, P. C., Wolff, M. J., Lee, S. W., & James, P. B. 1997 *LPSC XXVIII*, 87.

Bell III, J. F., Thomas, P. C., Wolff, M. J., Lee, S. W., & James, P. B. 1996 *EOS, Trans. A.G.U. 77*, F431.

Bell III, J. F. 1992 *Icarus*, **100**, 575.

Burns, R. G. 1993 *Geochim. Cosmochim. Acta*, **57**, 4555.

Cantor, B. A., Wolff, M. J., Thomas, P. C., James, P. B., & Jensen, G. 1999 *Icarus*, **142**, 414.

Cantor, B. A., Wolff, M. J., James, P. B., & Higgs, E. 1998 *Icarus*, **136**, 175.

Carr, M. H. 1999 *Water on Mars*. Oxford Univ. Press.

Clancy, R. T. & Nair, H. 1996 *J. Geophys. Res.*, **101**, 12785.

Clancy, R. T., Grossman, A. W., Wolff, M. J., James, P. B., Rudy, D. J., Billawala, Y. N., Sandor, B. J., Lee, S. W., & Muhleman, D. O. 1996 *Icarus*, **122**, 36.

Clancy, R. T., Wolff, M. J., & James, P. B. 1999 *Icarus*, **138**, 49.

Clancy, R. T., Sandor, B. J., Wolff, M. J., Christensen, P. R., Smith, M. D., Conrath, B. J., & Wilson, R. J. 2000 *J. Geophys. Res.*, **105**, 9553.

Espenak, F., Mumma, M. J., Kostiuk, T., & Zipoy, D. 1991 *Icarus*, **92**, 252.

Flammarion, C. 1964 *The Flammarion Book of Astronomy*.Simon & Schuster.

Gierasch, P. J. & Bell III, J. F. 2000, manuscript in preparation.

Gierasch, P. J., Thomas, P., French, R., & Veverka, J. 1979 *Geophys. Res. Lett.*, **6**, 405.

Goffin, E. & Meeus, J. 1978 *Sky & Telescope*, **56**, 106.

Haberle, R. M., Pollack, J. B., Barnes, J. R., Zurek, R. W., Leovy, C. B., Murphy, J. R., Lee, H., & Schaeffer, J. 1993 *J. Geophys. Res.*, **98**, 3093.

Hamilton, D. P. 1996 *Icarus*, **119**, 152.

Houck J. R., Pollack J. B., Sagan, C., Schaack D., & Decker, J. 1973 *Icarus*, **18**, 470.

James, P. B., Bell III, J. F., Clancy, R. T., Lee, S. W., Martin, L. J., & Wolff, M. J. 1996 *J. Geophys. Res.*, **101**, 18,883.

James, P. B., Kieffer, H. H., & Paige, D.A. 1992 in *Mars* (eds. H. Kieffer et al.), pp. 934-968, Univ. Arizona Press.

Johnson, J. R. & Grundy, W. M. 2000 in *Lunar Planet. Sci. Conf. 31*, Abstract 1724.

Lee, S. W., Wolff, M. J., James, P. B., Clancy, R. T., Bell III, J. F., & Martin, L. J. 1996 *B.A.A.S.*, **28**, 1061.

Malin, M. C., et al. (15 others) 1998 *Science*, **279**, 1681.

Malin, M. C. & Edgett, K. 2000 *Science*, **288**, 2330.

Martin, L. J., James, P. B., Dollfus, A., Iwasaki, K., & Beish, J. D. 1992 in *Mars* (eds. H. H. Kieffer, et al.), pp. 34-70. Univ. of Ariz. Press.

McCord, T. B., Singer, R. B., Hawke B. R., Adams, J. B., Evans D. L., Head J. W., Mouginis-Mark, P. J., Pieters, C. M., Huguenin, R. L., & Zisk, S. H. 1982 *J. Geophys. Res.*, **87**, 10129.

McElroy, M. B. & Donahue, T. M. 1972 *Science*, **177**, 986.

McKay, D. S., Gibson, E. K., Thomas-Keprta, K. L., Vali, H., Romanek, C. S., Clemett, S. J., Chiller, X. D. F., Maechling, C. R., & Zare, R. N. 1996 *Science*, **273**, 924.

McKim, R. 1999 *History of Mars Dust Storms.* British Astron. Assoc.

Mischna, M., Bell III, J. F., James, P. B. & Crisp, D. 1998 *Geophys. Res. Lett.*, **25**, 611.

Moersch, J. E. 1998 *Ph.D. thesis*, Cornell University.

Morris, R. V., Golden, D. C., Bell III, J. F., & Lauer, Jr., H. V. 1995 *J. Geophys. Res.*, **100**, 5319.

Murchie, S., Mustard, J., Bishop, J., Head, J., Pieters, C., & Erard, S. 1993 *Icarus*, **105**, 454.

Murchie, S., et al. 2000 *Icarus*, in press.

Ockert-Bell, M. E., Bell III, J. F., McKay, C. P., Pollack, J. B., & Forget, F. 1997

Owen, T. 1992 in *Mars* (eds. H. H. Kieffer, et al.), pp. 818–834, Univ. of Ariz. Press.

Paige, D. A., Bachman, J. E., & Keegan, K. D. 1994 *J. Geophys. Res.*, **99**, 25,959.

Pimentel G. C., Forney, P. B., & Herr, K. C. 1974 *J. Geophys. Res.*, **79**, 1623.

Pollack, J. B., Kasting, J. F., Richardson, S. M., & Poliakoff, K. 1987 *Icarus*, **71**, 203.

Pollack, J. B., Ockert-Bell, M. E., & Shepard, M. K. 1995 *J. Geophys. Res.*, **100**, 5235.

Richardson, M. I. & Wilson, R. J. 2000 *American Astronomical Society, DPS meeting #32*, #50.05.

Singer R. B., McCord, T. B., Clark, R. N., Adams, J. B., & Huguenin, R. L. 1979 *J. Geophys. Res.*, **84**, 8415.

Slipher E. C. 1962 *Mars: The Photographic Story.* Sky Publishing Corp.

Smith, P. H., Bell III, J. F., Bridges, N. T., Britt, D. T., Gaddis, L., Greeley, R., Keller, H. U., Herkenhoff, K. E., Jaumann, R., Johnson, J. R., Kirk, R. L., Lemmon, M., Maki, J. N., Malin, M. C. Murchie, S. L. Oberst, J., Parker, T. J., Reid, R. J., Soderblom, L. A., Stoker, C., Sullivan, R., Thomas, N., Tomasko, M. G., & Wegryn, E. 1997 *Science*, **278**, 1758.

Soderblom, L. A., Edwards, K., Eliason, E. M., Sanchez, E. M., & Charette, M. P. 1978 *Icarus*, **34**, 446.

Tamppari, L. K., Zurek, R. W., & Paige, D. A. 2000 *J. Geophys. Res.*, **105**, 4087.

Wells, E. N., Veverka, J., & Thomas, P. 1984 *Icarus*, **58**, 331.

Wolff, M. J., Lee, S. W., Clancy, R. T., Martin, L. J., Bell III, J. F. & James, P. B. 1997 *J. Geophys. Res.*, **102**, 1679.

Wolff, M. J., Bell III, J. F., James, P. B., Clancy, R. T., & Lee, S. W. 1999 *J. Geophys. Res.*, **104**, 9027.

Yen, A. S., Murray, B., Rossman, G. R., Grunthaner, F. J. 1999 *J. Geophys. Res.*, **104**, 27,031.

Zellner, B. & Wells, E. N. 1994 in Lunar Planet. Sci. Conf. 25th, p. 1541.

HST images of Jupiter's UV aurora

By JOHN T. CLARKE

Space Physics Research Laboratory, University of Michigan, Ann Arbor, MI 48109-2143

One of the brightest and most variable UV emissions in the solar system comes from Jupiter's UV aurora. The auroras have been imaged with each camera on *HST*, starting with the pre-COSTAR FOC and continuing with increasing sensitivity to the present with STIS. This paper presents a short overview of the scientific results on Jupiter's aurora obtained from *HST* UV images and spectra, plus a short discussion of Saturn's aurora.

1. The Earth's aurora: Present understanding

With a long history of ground-based and spacecraft measurements, we now have some understanding of the physics of the Earth's auroral processes. A general picture of the nature of auroral activity on the Earth has evolved, without a complete understanding of the many details. In general, auroral emissions are produced by high energy charged particles precipitating into the Earth's upper atmosphere from the magnetosphere (the region of space where the motions of particles are governed by the Earth's magnetic field). It is well established that the Earth's auroral activity is related to solar activity, and more specifically to conditions in the solar wind reaching the Earth. The precipitating charged particles are accelerated to high energies in the Earth's magnetosphere, with some acceleration occurring in the magnetotail region and some occurring by field-aligned potentials in the topside ionosphere. The auroral emissions then strongly modify the Earth's auroral ionosphere with large amounts of ionization and Joule heating. The Earth's auroral oval is known to maintain a pattern fixed with respect to the Earth-Sun line (i.e. along the noon-midnight meridian). The aurora normally exhibits oval-shaped patterns centered on each magnetic pole, while the Earth rotates under these ovals. While the general orientation of the ovals may be fixed, large variations in auroral intensity (as well as the diameters of the auroral ovals) occur over time. The fixed ovals are due to the acceleration of particles in the interaction of the solar wind with the Earth's magnetic field. When auroral storms occur, they typically begin near local midnight with a brightening of one section of the auroral oval. The brightening will then extend along the oval, while at the same time the oval increases its radius, resulting in auroral displays that can be observed as far south as the southern US states.

The general processes are sketched as a cartoon in Figure 1. Auroral storms on the Earth are triggered when the solar wind magnetic field turns southward, and large storms are generally produced when a coronal mass ejection on the Sun directs a high speed stream in the solar wind toward the Earth. This orientation permits the solar wind magnetic field to connect directly with the Earth's magnetic field, so that the electrons and ions in the solar wind cross field lines as they pass down the tail region. The resulting drifts of electrons and ions toward the evening and morning sides of the tail, respectively, produce an electric potential across the magnetotail. This electric potential is what drives the aurora. The much higher electrical conductivity along field lines leads to the motion of high energy charged particles along field lines into the Earth's ionosphere, where the cross-field conductivity is sufficient for the circuit to be closed. The process works much more efficiently when the solar wind magnetic field is pointed southward.

This discussion smooths over many details of the Earth's aurora, but should be sufficient to compare and contrast aurora on the Earth with the aurora on Jupiter.

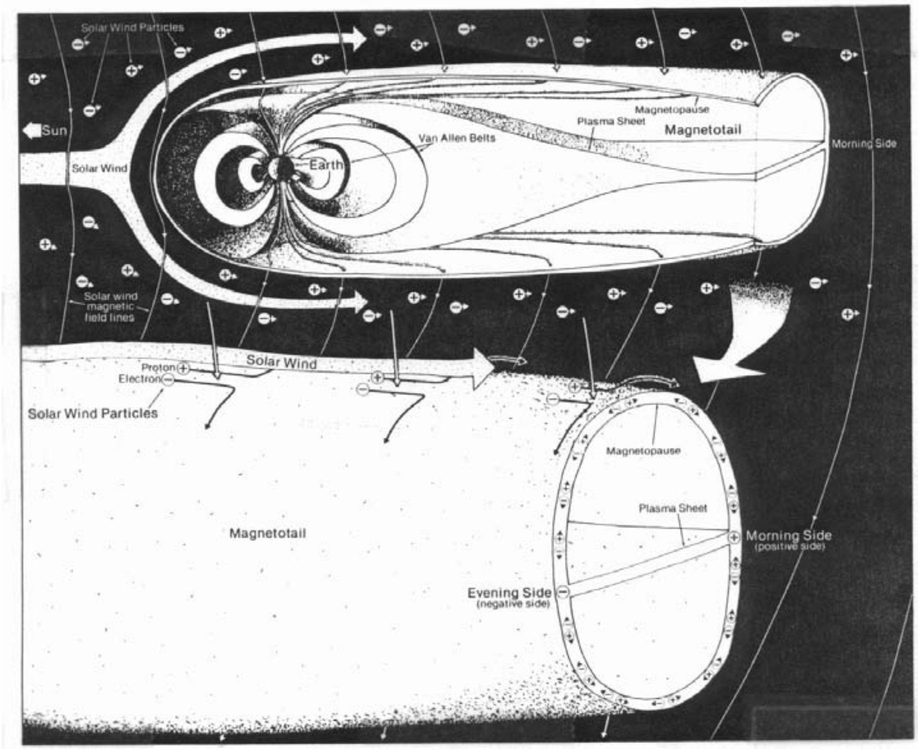


FIGURE 1. Cartoon schematic of the interaction of the solar wind with the Earth's magnetosphere, from Akasofu 1979.

2. Jupiter's aurora: Post-Voyager picture

Jupiter's auroras were first observed by the Voyager 1 ultraviolet spectrometer in 1979 (Broadfoot et al. 1979), and within months were also observed and a long term study begun by the *IUE* (Clarke et al. 1980). The *Voyager 1 & 2* encounters and subsequent extended *IUE* observations led to a "standard" accepted picture of Jupiter's magnetosphere and the production of the polar auroras, which will be outlined below. Based on early radio observations, it had been thought since the 1960s that Jupiter possessed a strong magnetic field of 1–10 Gauss strength, and that this field was strongly non-dipolar. The Pioneer and Voyager missions provided the first *in situ* measurements of the strength and geometry of Jupiter's magnetic field and the plasma properties of the Jovian magnetosphere. The *Voyager* missions also discovered Io's volcanoes, and found the full effects of the plasma torus on Jupiter's magnetosphere.

The resulting picture of Jupiter's magnetosphere was one filled with plasma predominantly from the Io torus, with the ions dominated by S and O species. It was believed that this plasma drifted slowly outward from the torus, and that some particles were accelerated to high energies by unknown processes. These particles then drifted more rapidly back inwards, and were no longer detected inside roughly 10–15 Jovian radii (R_J). The aurora were therefore thought to be produced by precipitating charged particles pitch angle scattered into the loss cone in the range 10–15 R_J, originating in the plasma torus but accelerated farther out in the magnetosphere. It was known that the plasma pressure was sufficiently large in Jupiter's magnetosphere so that the corotating plasma current strongly distorted the local magnetic field outside roughly 6–8 R_J, with the region of corotating plasma in the current sheet extending out to 20–30 R_J. Continuous currents from the magnetosphere to Jupiter's auroral ionosphere were thought to dissipate up to 10^{14} Watts of power in the very bright aurora. This auroral energy was known to be 20–50 times greater than the solar UV radiation absorbed globally in the upper atmosphere, so that the aurora would drive the upper atmosphere on a global basis. The ultimate source of energy for all these processes on Jupiter was Jupiter's rotation, which enforced pickup and corotation on the plasma via the magnetic field. This is in contrast to the situation at Earth, where the solar wind is the main source of energy for the aurora.

The interaction of Io with Jupiter's magnetic field is of particular interest. Io is electrically conducting by virtue of its ionosphere, with Jupiter's magnetic field and the corotating plasma torus sweeping past at a speed exceeding Io's orbital motion by 56 km s^{-1}. Following early decametric observations, a continuous electric current linking Io with Jupiter's ionosphere was proposed (Figure 2), driven by Io acting as a unipolar inductor with a 400 kilovolt potential across its diameter radially away from Jupiter (Goldreich 1969). The *Voyager 1* spacecraft passed about 20,000 km south of Io, and found the local magnetic field and plasma flow distorted by a 3×10^6 Ampere field-aligned current (Acuna et al. 1981) along Io's magnetic flux tube. The existence of the plasma torus along Io's orbit implied that the field-aligned current would be carried by Alfvén waves propagating at a speed determined by the local plasma density (Belcher 1987). The measured torus plasma density suggested that the Alfvén waves carrying the current should return from Jupiter's ionosphere after Io had passed beyond those magnetic field lines, so that the circuit would not maintain a direct current structure. Io's magnetic "footprint" auroral emission would be produced at the point where the circuit is closed by currents in and out of Jupiter's upper atmosphere. Jupiter's magnetic field also picks up ions and electrons from Io, distributing them into a corotating torus-shaped plasma region about Jupiter. Jupiter's rotation with an inclined and asymmetric magnetic field causes the torus to move north and south with respect to Io, thereby varying the current path length through the torus with longitude (in the opposite sense north and south), and the field strength (and corresponding electric potential) at Io varies by 20% with longitude. Considerable variability was therefore expected in the production of auroral emissions at Io's magnetic footprint on Jupiter, which would be diagnostic of the interaction of Io with Jupiter's magnetic field and the plasma torus.

3. Background on observations of Jupiter

The early detection of decametric radio emissions modulated by the orbital location of Io, discussed above, indicated that Jupiter had a magnetic field and an electromagnetic interaction with Io. Interpreting the highest frequency detected as an electron gyrofrequency implied a maximum field strength of 10 Gauss. Jupiter's ionosphere was first detected by radio occultation during the *Pioneer 10* flyby in 1974 (Kliore et al. 1974), and ground-based telescopic observations first detected the plasma torus in the mid 1970s (Brown & Yung 1976). It was during the *Voyager 1* flyby in 1979 that Jupiter's UV auroral were first observed (Broadfoot et al. 1979), and during this and the *Voyager 2* encounter the long aperture of the UVS was used to map the equatorward extent of the UV auroral emissions. These maps indicated that auroral emissions first appeared when the end of the aperture covered the expected latitude of the magnetic mapping of the plasma torus into Jupiter's atmosphere, seemingly implicating the plasma torus as the source of auroral particles. *IUE* observations provided important information on the spatial and temporal variations of the UV aurora from Jupiter's north and south polar regions (Clarke et al. 1980, Livengood et al. 1992, Harris et al. 1996). These in-

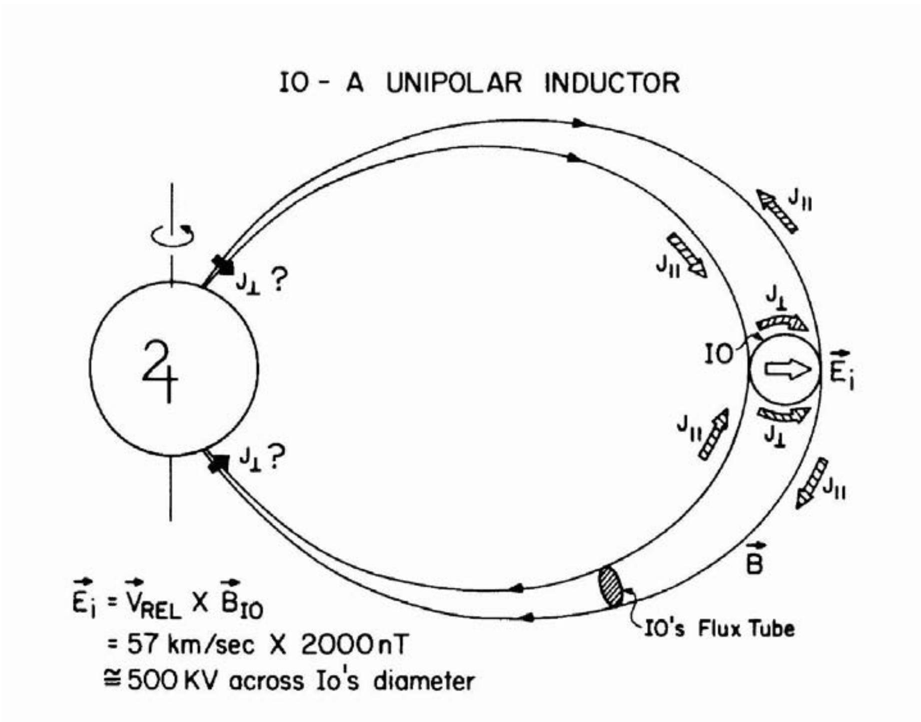


FIGURE 2. Schematic drawing of the electrodynamic interaction of Io with Jupiter's magnetic field, as it was perceived in the post-Voyager epoch (Acuna et al. 1983).

cluded mapping the auroral intensity and "color ratio" (an indication of the atmospheric absorption of departing auroral emissions) with Jovian longitude and time, as well as spectra of Io's airglow and the plasma torus. The UV auroral emissions are similar to the spectrum of electron excitation of H_2 in the laboratory, including H_2 Lyman ($B\,^1\Sigma_u{}^+$–$X\,^1\Sigma_g{}^+$) and Werner ($C\,^1\Pi_u$–$X\,^1\Sigma_g{}^+$) band series plus H Lyα line emission. In addition, near-IR thermal emissions from wavelengths of 2.1 and 3.4 microns have been detected from H_3^+ ions in Jupiter's auroral ionosphere (Drossart et al. 1989), which act to radiatively cool the auroral atmosphere. The IR emissions are not directly excited, as the UV emissions are, and therefore represent the auroral energy dissipation integrated over some period of time rather than a direct link to magnetospheric particle precipitation. The $H_3{}^+$ emissions do exhibit a similar morphology to the UV emissions, although this has only been tested with lower resolution ground-based images, and to date without simultnaeous imaging from the ground and *HST*.

4. FOC, WFPC 2, STIS UV imaging properties

Since the launch of *HST*, three cameras in four imaging modes have been used to obtain images of the UV aurora from Jupiter, counting the modes as the pre- and post-COSTAR FOC, the WFPC2, and the STIS. This section will quickly review the observational strategies used with these cameras, and the relative properties of the cameras in observations of diffuse UV emissions.

Several challenges exist for any camera on *HST* to image the bright UV aurora on Jupiter (or any other planet). These include blocking the longer wavelength light from the sunlit sides of the outer planets, which is 6 orders of magnitude greater than the

Property	*FOC*	*WFPC2*	*STIS*
Point Spread Function (FWHM):	0.05″	0.10″	0.08″
Typical Exposure:	700–900 sec	400–600 sec	100 sec
Limiting Sensitivity:	50 kRayleigh	10 kRayleigh	1 kRayleigh
Field of View:	14″	80″	25″

Table 1. UV imaging properties of the *HST* cameras

UV emission, having sufficient sensitivity to image the diffuse auroral emissions in pixels less than 0.1″ in size, and sufficient dynamic range to record the wide range of auroral brightnesses. Sensitivity and red leaks have been the main limitations in each case. Both FOC and WFPC2 have visible-sensitive detectors, so that filters have to be used to block visible wavelengths, and these filters also limit the UV sensitivity. FOC observations employed two UV filters in series, while the WFPC2 camera has new technology Wood's (alkali metal) filters developed for that instrument, with very efficient red light rejection. STIS was the first camera able to avoid this problem by using a solar-blind detector with a UV photocathode, leading to much higher sensitivity than previously possible. Levels of background and noise in each camera are also important, but this paper will not go into detail on these issues. Suffice it to say that the sensitivity values listed in Table 1 are derived from measured standard deviations in cross-cuts through auroral images with the specified exposure times and reduced by pipeline routines, without any additional deconvolution, smoothing, or red leak image scaling and subtraction.

Jupiter's observed auroral emissions range from the order of 1 kRayleigh (1 kR = 10^9 photons/sec from a 1 cm^2 column of the planet's atmosphere into 4π steradians) to a few MRayleighs. As seen from the table, the sensitivities of auroral images have improved remarkably over time. The FOC images detected only the brightest auroral emissions with exposures of 700–900 sec, WFPC2 images detected the majority of the auroral emissions in 400–600 sec, and STIS images can detect basically all of the auroral features in 100 sec exposures. A subtle feature of the increasing sensitivity is that it also leads to higher effective spatial resolution. While the FOC had the tightest point spread function (PSF), the extended exposures led to blurring of the images from Jupiter's rapid rotation. Jupiter's rotation is approximately $1^0/100$ sec, which differentially blurs the images at different locations on the planet. Exposures of 100 sec lead to essentially no rotational blurring on scales of 0.1″, which corresponds to 290–420 km at Jupiter's distance depending on the Earth-Jupiter distance when *HST* observations are possible.

More detailed plots of the spectral response and red leaks of the three cameras are plotted in Figure 3. These can be compared with the GHRS spectra of auroral emissions shown Figure 7. In addition to these properties, WFPC2 images have the advantage of a sufficiently large field of view to image both north and south poles in each exposure. This makes it possible to explore the extent to which emissions at one pole are accompanied by conjugate emissions at the other pole. Finally, STIS is able to take time-tagged image data on Jupiter, but just barely. Use of the SrF$_2$ filter limits the overall count rate to the range of 20,000–30,000 counts/sec from Jupiter, in which case time-tagged data can be recorded for 200–300 sec before experiencing dropouts from the filled buffer. Unfiltered images including the H Lyα line cannot be taken in time-tag mode due to the high count rate. In practice, time-tagged filtered images can be binned into 10 sec "frames" as a good trade-off between having reasonable sensitivity and high time resolution, producing the first high time resolution movies of Jupiter's UV aurora.

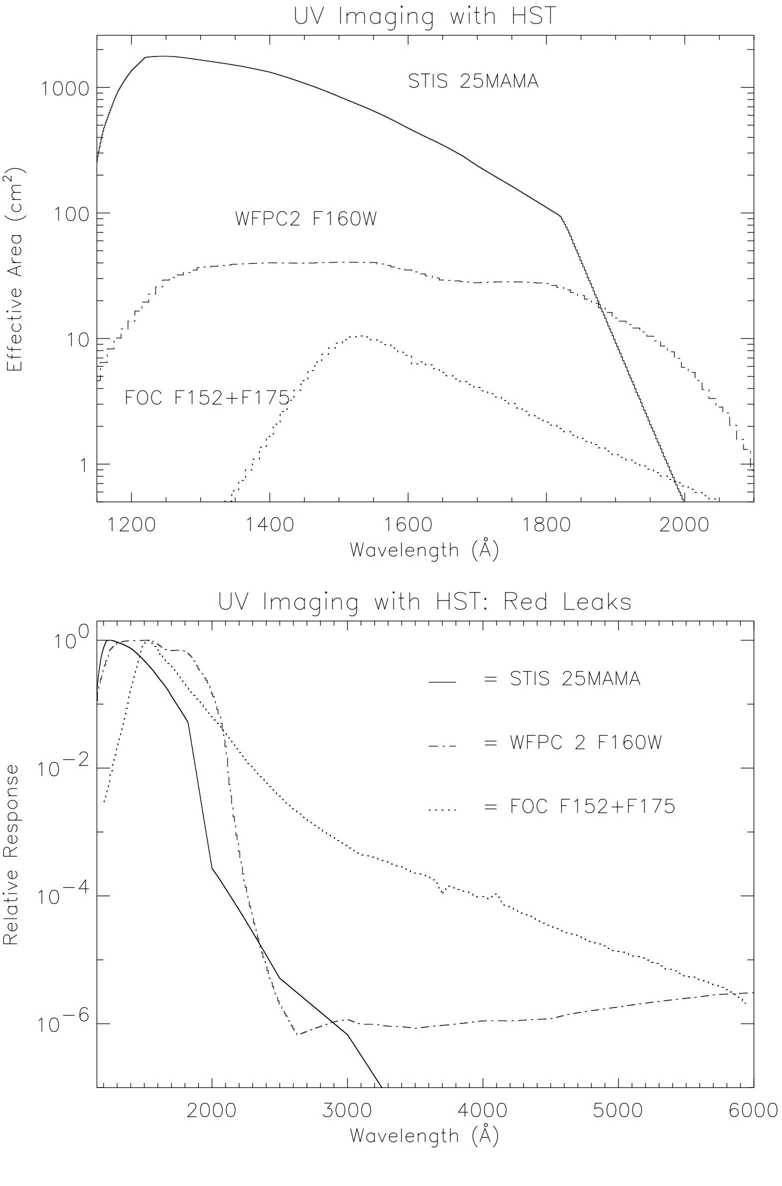


Figure 3. Plots showing the calculated effective area and wavelength responses of the three cameras on *HST*.

5. Summary of FOC results

FOC images of Jupiter's aurora were obtained both pre-COSTAR (Dols et al. 1992, Caldwell et al. 1992, Gérard et al. 1993a, Gérard et al. 1993b) and post-COSTAR (Prangé et al. 1998. Examples are shown in Figures 5 and 6. The FOC images were the first true images (as opposed to long-slit spectra from UVS and limited imaging at 5″ resolution with IUE) of Jupiter's aurora, and they showed for the first time the spatial structure of the emissions at high resolution. While the addition of the two COSTAR mirrors decreased the efficiency by approximately a factor of two at H Lyα, the detection of discrete auroral features was improved overall by the improved Strehl ratio of the corrected PSF.

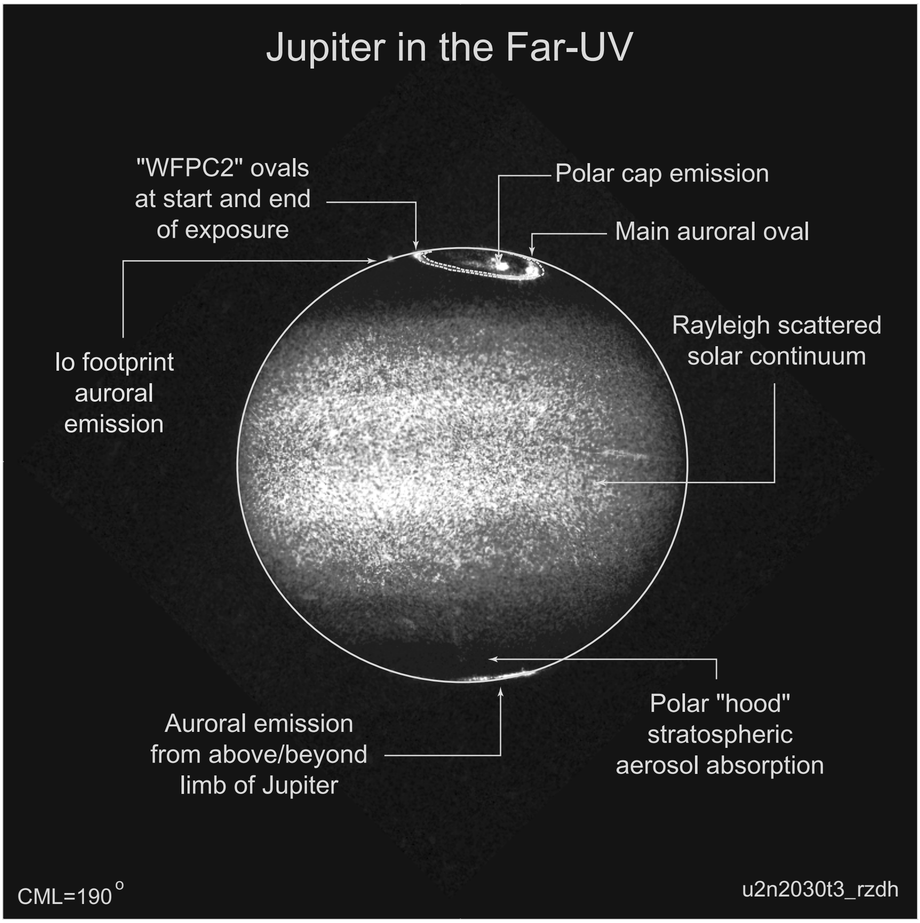

Jupiter in the Far-UV
"WFPC2" ovals at start and end of exposure
Polar cap emission
Main auroral oval
Io footprint auroral emission
Rayleigh scattered solar continuum
Auroral emission from above/beyond limb of Jupiter
Polar "hood" stratospheric aerosol absorption
CML=190°
u2n2030t3_rzdh

FIGURE 4. Example of one WFPC2 UV image of Jupiter, with main features of the planet and auroral emissions indicated.

The FOC images first revealed several important properties of the UV auroral distribution and its variations with time. First, fits of the observed latitude of the auroral ovals were compared with the O6 model for Jupiter's magnetic field, and these suggested that the ovals were better fit by mapping to an equatorial distance of approximately 30 R_J than to either the Io plasma torus or the 10–20 R_J distance suggested by the Voyager data. The magnetic field model was later corrected based on observed locations of the Io footprint (see next section), and the revised model and observations of auroral footprints associated with Europa and Ganymede confirmed and extended this result. The images showed that the northern oval tended to be quite narrowly confined in latitude in the morning sector, and much more diffuse in the afternoon sector, at times breaking up into multiple arcs extending into the polar cap. This pattern appeared to remain fixed in local time, while there was observational bias from a concentration of images taken when the north auroral oval was pointed toward the Earth, resulting in limited longitude coverage (generally true of all observations from the Earth). Initial measurements were also made of the UV color ratio of the aurora through a selection of different filters. Another important feature first seen in FOC images was a dawn storm (Gérard et al. 1994), which appeared along the dawn side of the northern oval, and was much brighter than any previously detected auroral emission. Io's auroral footprint was discovered in

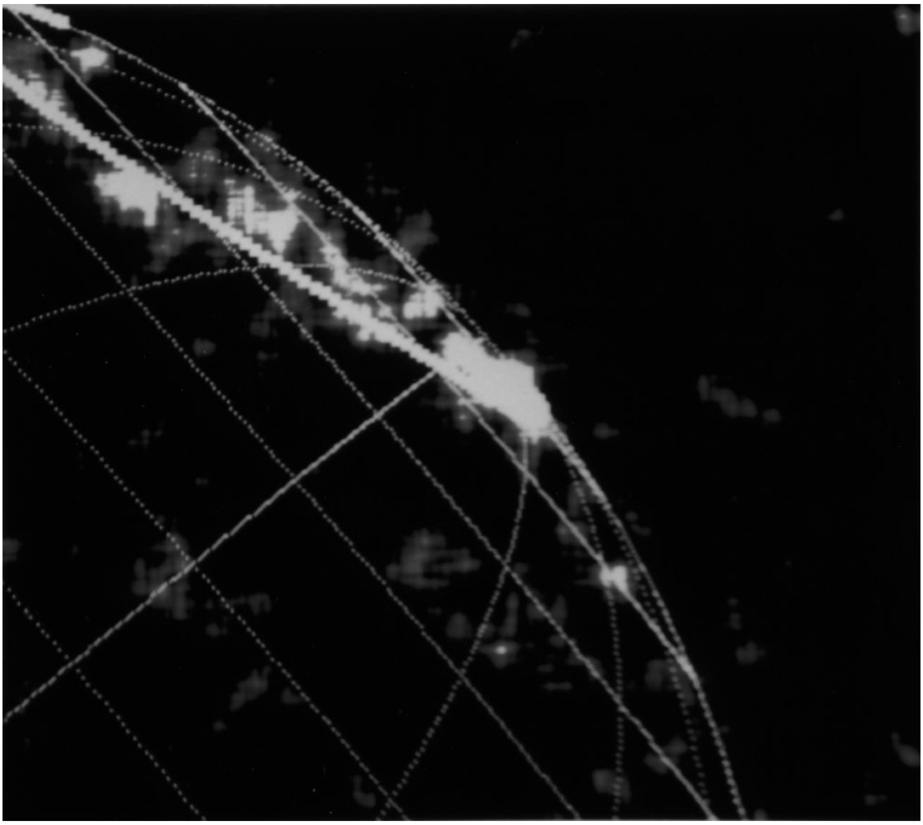


FIGURE 5. UV image of Jupiter's north aurora with pre-COSTAR optics, from Gérard et al. 1993a.

near-IR H_3^+ images (Connerney et al. 1993). A feature interpreted as the Io footprint was also detected in one of four post-COSTAR FOC images when its expected location appeared within the field of view (Prangé et al. 1996) in July 1994, suggesting intrinsic variations in its intensity. This feature was initially interpreted as being quite narrow, suggesting an interaction limited to the near-Io environment. Later images with WFPC2 and STIS at higher sensitivity would reveal a very highly extended region of auroral emission, extending to large distances from Io mainly in the downstream direction.

6. Summary of GHRS results

While the concentration of this review is on images of Jupiter's UV aurora, much progress in understanding the physical processes active in the aurora has been gained through UV spectra obtained by *HST*. The majority of published spectra to date were obtained by the GHRS, while new results from the long aperture STIS spectra are expected to appear soon in the journals. For the diffuse auroral emissions, the large science aperture (LSA) was nearly always used due to the much smaller area and count rates with the small aperture. The LSA area of $2''$(or $1.74''$post-COSTAR) covered areas of several thousand km on Jupiter, sufficient to isolate distinct features of the aurora. The low resolution GHRS spectra show clearly the spectrum of electron impact on H_2 (Figure 7a). The first set of papers reporting GHRS spectra of the aurora presented spectra at 0.05 nm resolution for diffuse emissions filling the LSA. This is sufficient to separate emis-

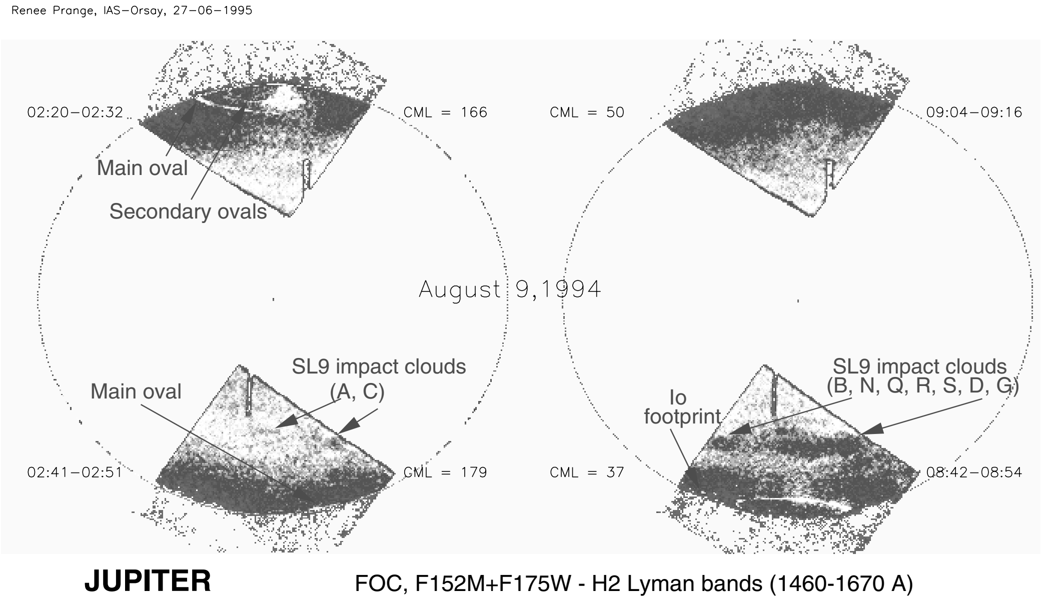
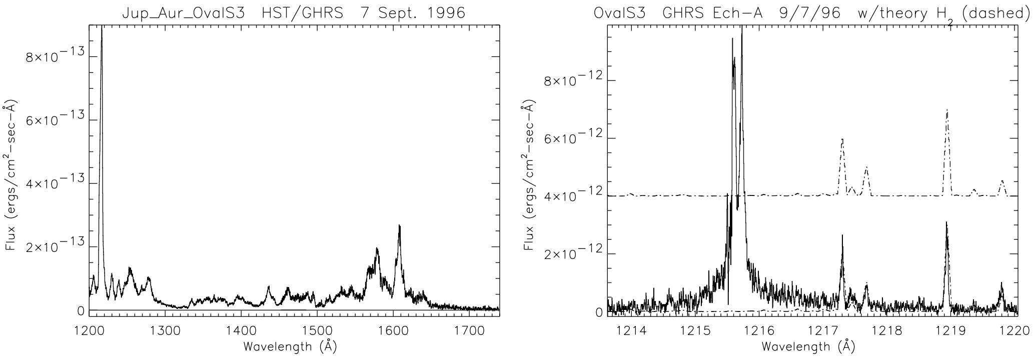


FIGURE 6. UV images of Jupiter's north and south aurora taken with the post-COSTAR FOC, from Prangé et al. 1998.



(a) (b)

FIGURE 7. *HST* GHRS spectra of Jupiter's aurora in September 1996, taken with the G140L low resolution grating (a) and with the Ech-A echelle grating (b). The same location was observed for both spectra, a bright region in the southern main oval near the evening limb. Plots are adapted from Dols et al. 2000.

sion features in the Lyman and Werner band systems, and determine the ro-vibrational temperature of the emitting H_2 in Jupiter's atmosphere. Four separate papers (Clarke et al. 1994, Kim et al. 1995, Liu et al. 1996, Trafton et al. 1994) fit models of the emission spectrum with GHRS spectra of various regions in the northern aurora, deriving temperatures in a broad range of 200–900 K. This may represent the range in altitude of the auroral emitting layer, and possibly also local variations due to auroral heating.

The medium resolution spectra also revealed extended emission wings on the H Lyα line, indicating a population of fast H atoms in the auroral atmosphere moving with velocities up to and exceeding the escape speed from Jupiter's gravity (Ajello et al. 1995, Clarke et al. 1994, Bisikalo et al. 1995). Echelle spectra of the auroral H Lyα emission line at 0.007 nm resolution further (Figure 7b) show the self-absorbed character of this internal emission source (Prangé et al. 1997b) as well as the extended emission wings. Finally, a more recent study of low resolution and echelle spectra of discrete emission features, and comparison with a detailed model for electron energy degradation and radiative transfer, has been presented (Dols et al. 2000). This indicates that the main oval and Io footprint have similar auroral color ratios, indicating similar altitudes of the auroral emission layer and thereby similar energies of the incident electrons. The polar cap emissions, by contrast, have highly variable and very different color ratios, indicating variable and both more and less energetic incident particles than the other auroral emissions.

7. Summary of WFPC2 results

The first WFPC2 images of Jupiter's UV aurora were obtained in May 1994 as part of the WFPC2 science team GTO program. Shortly thereafter (July 1994), the comet Shoemaker-Levy 9 impacts on Jupiter occurred, leading to a concentrated series of images to record these events. Several features of the aurora were reported from these early images (Clarke et al. 1995, Prangé et al. 1995), Clarke et al. 1996, Ballester et al. 1996. Auroral emissions were always detected from both poles at all longitudes, due to the height of the auroral curtain above the limbs. The aurora were separated into three emission regions: the main oval, the diffuse emissions inside the polar cap, and Io's magnetic footprints. Conjugate emissions were seen in the north and south in every case where the observing geometry permitted the emissions to be seen from both poles. Reference ovals were established based on the locations of the main oval emissions in some of the early images (not statistical averages, but reference locations for comparison of different images). It was apparent from the first images in May 1994 that the main oval features were rotating with the planet, in contrast with the sun-planet fixed pattern seen on the Earth. Particularly in the northern aurora, emissions from one range of Jovian longitudes (where the polar cap was generally filled with emission) appeared to consistently move from the polar regions toward the equator as the region rotated with the planet from local morning to local afternoon (the "equatorward surge"). Dawn storms were also observed, similar to the earlier one observed in FOC images (Figure 8), and these were demonstrated to remain fixed in magnetic local time over a period of several hours as the planet (and other features in the main oval) rotated past. The process producing the dawn storms is not understood, but it is believed to be related to the diurnal pattern of dynamical motions in Jupiter's magnetosphere. The solar wind pressure distorts the middle and outer regions of the magnetosphere, enforcing a contraction on the morning side and expansion on the evening side. These motions cover great distances in a fraction of the 10 hour Jovian rotation period, and are generally expected to lead to instabilities in the trapped corotating plasma. In this sense Jupiter's auroral processes are driven in part by the solar wind, but simply by its pressure rather than by reconnection with Jupiter's magnetic field in a more Earth-like process. This picture supports the idea that Jupiter's rotation is the ultimate source of energy for the aurora, not the solar wind.

Auroral emissions were also detected from the Io footprints in all WFPC2 images where Io's magnetic footprint was facing the Earth, and these were generally located 4–6^0 equatorward of the main oval. This indicated that the main oval mapped along Jupiter's magnetic field to much greater distances than the 6 R_J distance of Io. High resolution measurements of the locations of the Io footprint emissions also led to an

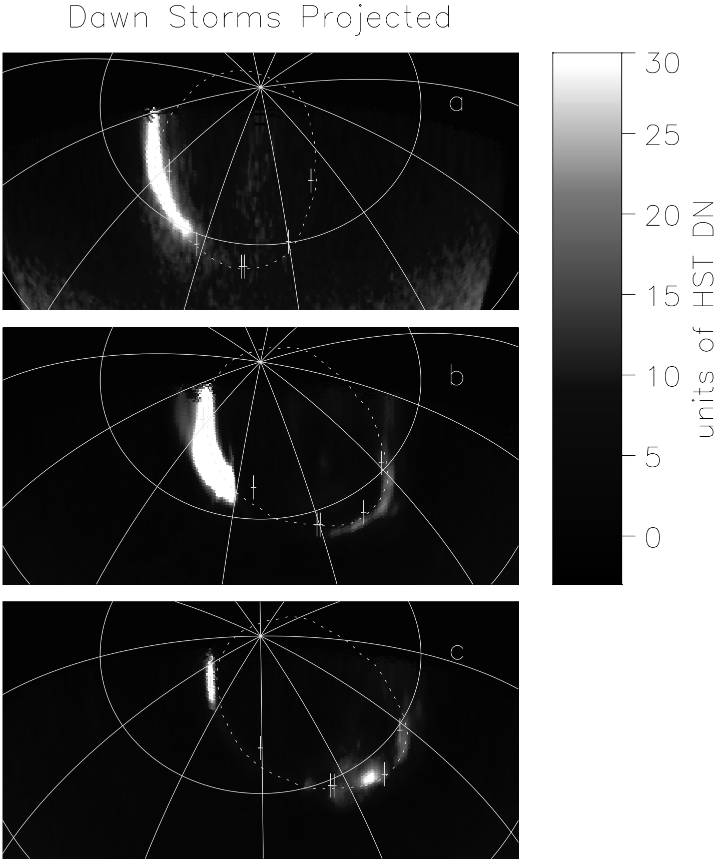


FIGURE 8. Projected UV images of dawn storm auroras on Jupiter. Top image is from FOC, lower two are from WFPC2. Each image has been re-projected as would be seen looking down from above the central meridian at 60^0 latitude, and the WFPC2 reference oval is overplotted as a dashed line. Figure from Clarke et al. 1998.

improvement in knowledge of Jupiter's magnetic field geometry. An updated field model was constructed by forcing field lines from Io's orbit to pass through the observed latitudes of the footprints. The longitudes were left unconstrained due to uncertainties in both the magnitude of the Alfvén wing near Io and the travel time of Alfvén waves from Io to Jupiter. The change in magnetic field geometry as a result of fitting to flyby data from the *Pioneer* and *Voyager* spacecraft, in addition to fitting WFPC2 and near-IR H_3^+ measured locations of the Io footprint (the VIP4 model, Connerney et al. 1998) is shown in Figure 9. Finally, the auroral color ratio of the various features was determined by the difference of clear and filtered images, which effectively block the shortest wavelength H and H_2 emissions. The main oval and Io footprint emissions have very similar color ratios, but the polar cap emissions appear generally more absorbed, implying higher energy incident particles and/or a relatively higher altitude extent of the UV-absorbing hydrocarbons in the polar atmosphere.

WFPC2 auroral images have supported other observations. UV auroral images were obtained during the comet S/L 9 impacts, revealing aurora at unusually low latitudes in the north conjugate to the K impact site (Clarke et al. 1995, Bauske et al. 1999, Figure 10) and pulsating southern auroral emissions apparently produced by the P2 fragment as it approached Jupiter (Prangé et al. 1995). FUV/EUV auroral spectra of the north aurora

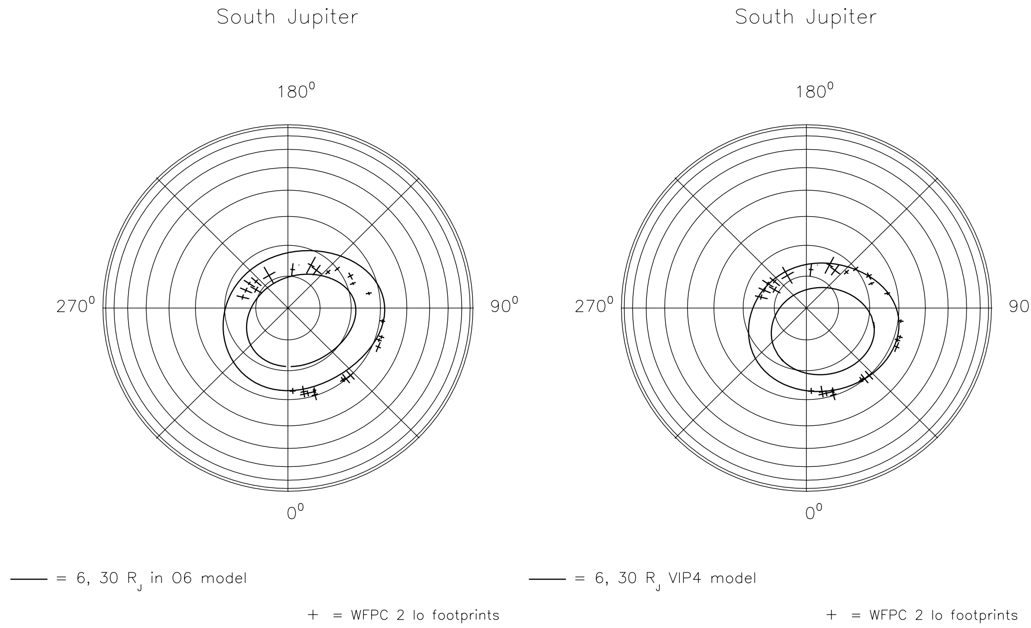


FIGURE 9. Polar projection of Jupiter showing the observed locations of auroral emissions from Io's magnetic footprint. Overplotted are the 6 R_J auroral ovals from the earlier (O6) and modified (VIP4) magnetic field models, showing the improved fits to the footprint locations. Figure from Clarke et al. 1998.

were obtained on the second *HUT* flight in March 1995 simultaneously with WFPC2 images (Morrissey et al. 1997), permitting a detailed spectral study of a known auroral brightness and distribution. In addition to the auroral emissions, the WFPC2 images recorded the distribution of the UV-absorbing aerosols comprising Jupiter's polar hoods. A long term study of these distributions has shown striking wave patterns in the polar stratospheres where the aerosols are confined, analogous to the circumpolar vortices on the Earth, and evidence for auroral-aligned features more directly related to the production of these absorbers (Vincent et al. 1999).

8. Summary of initial STIS results

The advent of STIS on *HST* has provided much higher sensitivity images of Jupiter's UV aurora than previously possible, thanks to its solar-blind photon-counting detector. Initial concerns about bright-light protection prevented images of Jupiter from being acquired until September 1997, and time-tagged images were first taken in December 1997. The first images were taken back to back with WFPC2 images for a direct comparison of the capabilities of the two cameras, as shown in Figure 10. They have been scaled to the same maximum values for the average brightness of the main ovals, and equally stretched with a double log function. This over-stretches the WFPC2 images, showing much of the low level noise, but it is necessary to see the full dynamic range of emissions in the STIS images. The STIS images are 100 sec exposures, and the WFPC 2 images (taken one *HST* orbit later for both north and south poles) are 600 sec exposures. The higher spatial resolution in the shorter exposures is apparent in the STIS images, providing more detail in the auroral structures. While all the main features of the aurora

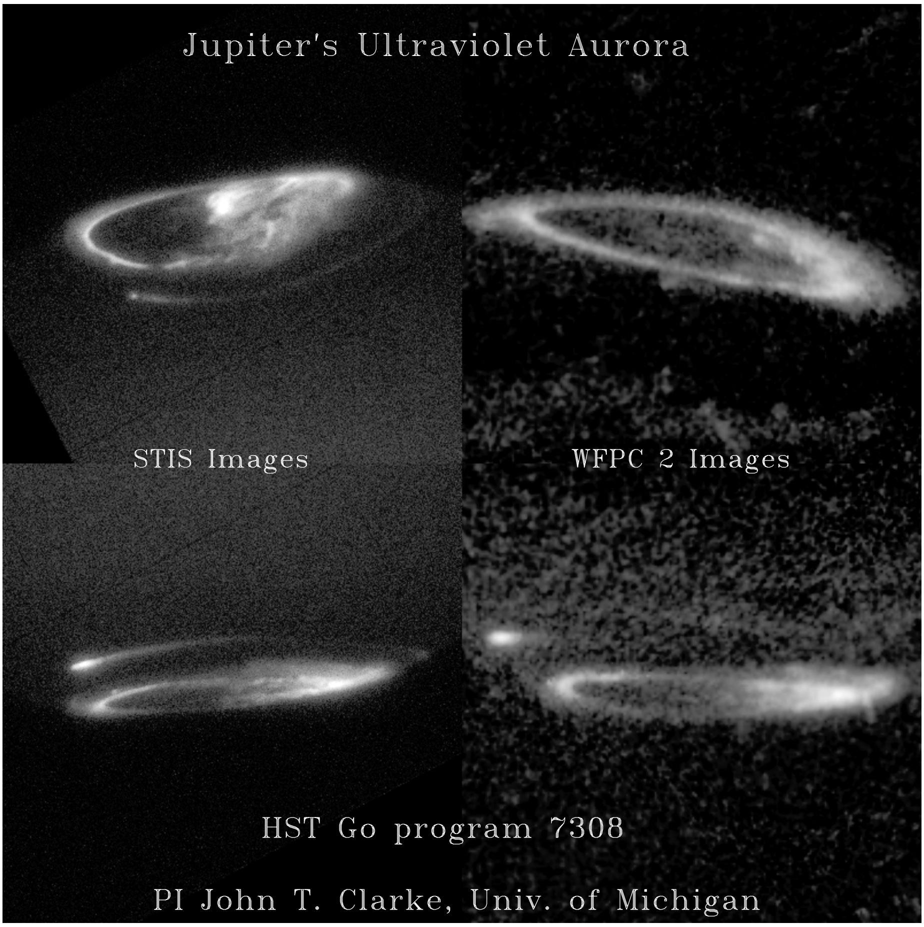

Jupiter's Ultraviolet Aurora
STIS Images
WFPC 2 Images
HST Go program 7308
PI John T. Clarke, Univ. of Michigan

FIGURE 10. UV images of Jupiter's aurora from September 1997, showing the STIS and WFPC2 images of the north and south aurora obtained close in time for a direct comparison of the capabilities of these two cameras, from Clarke et al. 1999.

are also seen in the WFPC 2 images, the higher STIS sensitivity reveals the extended structure in the auroral emissions from Io's magnetic footprints, plus the first detections of Ganymede's auroral footprints just equatorward of the main ovals on the morning side (Clarke et al. 1999). The extended comet-like trails of auroral emissions from Io's footprints extend in the downstream direction of the plasma flow past Io. The physical processes producing these emissions, acting for several hours after Io has passed by, are not understood, but are thought to reflect the residual interaction of the torus plasma with Io. Subsequent imaging of Jupiter's aurora with STIS has provided further measurements of the magnetic footprints of Io, Europa, and Ganymede. The Ganymede footprints always appear equatorward of the main oval, while the Io footprints and their trails map closely to the VIP4 model 6 R_J field line latitudes corresponding to the plasma torus. We have now demonstrated that the main oval maps to a distance of approximately 20–30 R_J from Jupiter, since it is outside the orbit of Ganymede but within the region of plasma corotation. The auroral footprints of Europa and Ganymede are unresolved in the STIS images, and an order of magnitude fainter than the Io footprint emissions. In each case, the emissions have been determined to be associated with each satellite by

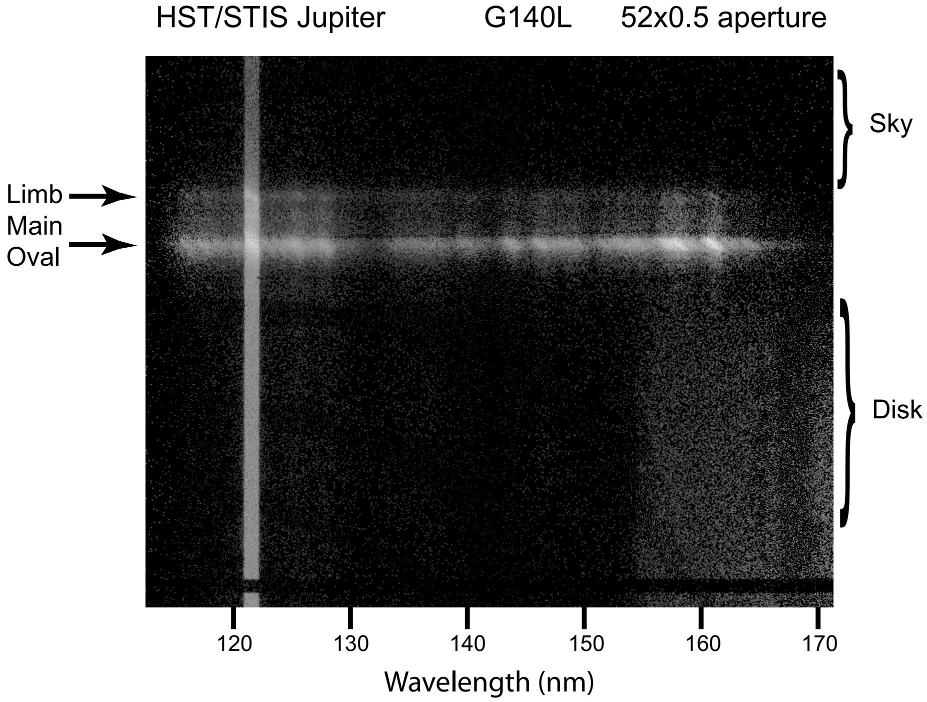


Figure 11. Example of spectral/spatial format in STIS data using sky aperture and G140L low resolution grating to observe Jupiter's polar regions.

observing them to remain under the magnetic footprint of the satellite while all other auroral features rotated past in a time series of images.

STIS also provides long aperture spectra with imaging along the aperture, and one example of these data is shown in Figure 11. The 25″ long aperture overlays simultaneously the sky, several distinct emitting regions within the auroral oval, and the disk of Jupiter with its spectrum of reflected sunlight rising at the longer wavelengths. Such data provide measurements of the auroral color ratio continuously across the auroral emission regions, and show the extent of the auroral emission curtain above the planet limb. Time-tagged images have also been obtained, with one example in Figure 12 showing 10 individual 30 sec "frames" in a 300 sec exposure from July 1998. In general, the main oval and satellite footprint emissions vary little in brightness on this time scale, but the polar cap emissions can vary rapidly from frame to frame, as seen from the bright flare of emission in the afternoon polar cap at the end of the series in Figure 12. These auroral flares are presently under study to determine their characteristics and the physical processes that may produce them in the polar cap, but not in the other auroral emission regions.

9. Coordinated missions and future Jupiter observations

WFPC2, GHRS, and STIS observations of Jupiter's aurora have been made during the Galileo orbiter tour of the Jovian magnetosphere, in concert with both *Galileo* UVS observations of the aurora (Ajello et al. 1998) and measurements by the particles and fields experiments on the *Galileo* orbiter (Clarke et al. 1998). The long aperture spectra obtained with the *Galileo* UVS have used WFPC2 images taken close in time, or average images taken at other times with similar observing geometry, to determine the likely auroral emission distribution within their apertures. Galileo CCD visible wavelength images have also been obtained of the aurora on Jupiter's night side (Vasavada et al. 1999). These images show that the main oval remains broken and diffuse in latitude in the pre-midnight sector, returning to a narrow latitude distribution in the pre-dawn sector, which is consistent with the Uv emission pattern observed on the day side by *HST*. A future opportunity exists as the *Cassini* spacecraft approaches Jupiter at the present time (Fall 2000). Coordinated observing campaigns are planned between *Cassini* solar wind measurements and STIS auroral images in December 2000, and coordinated observations of the day side aurora with STIS and night side aurora with the *Cassini* UVIS in January 2001. These observing campaigns have the potential to clarify the relations between Jupiter's auroral activity and the solar wind dynamic pressure, field direction, etc., as well as to add to our understanding of the relationship between the day side and night side auroral distributions and variations with time. In addition, the measurements of the solar wind conditions near Jupiter in conjunction with measurements near the Earth will provide tests of extrapolations from solar wind conditions at 1 AU to those near Jupiter. These studies will support future work to discover the extent to which Jovian auroral processes are tied to conditions in the solar wind.

10. *HST* images of Saturn's aurora

While most *HST* observations of planetary aurora have been of Jupiter, successful observations of Saturn's UV aurora have also been obtained. Prior to *HST* observations, indirect evidence for Saturn's UV aurora was first obtained from the *Pioneer 11* UV instrument (Judge et al. 1980), and polar brightenings of the planetary H Lyα emission were detected by *IUE* (Clarke et al. 1980, McGrath & Clarke 1992). The *Voyager 1* UVS detected UV emission from a polar aurora, localizing the latitude of the oval to near 80^0 (Broadfoot et al. 1981). The *Voyager* data also showed a correlation between the UV auroral brightness and Saturn's kilometric radio emmission. Such a longitudinal asymmetry is surprising in view of the strongly dipolar and aligned magnetic field of Saturn. FOC images of Saturn's UV aurora (Gérard et al. 1995) showed the auroral oval latitude consistent with the 80^0 latitude found in the *Voyager* data. FOC observations also showed the polar hood of UV absorbers (Gérard et al. 1995, Ben Jaffel et al. 1995, and evidence was presented for an absorption feature consistent with the location of the auroral emission. WFPC2 images showed the northern and southern aurora at higher signal to noise, and a model was presented for the auroral emission, fitting the distribution from the images (Trauger et al. 1998a). The best-fit latitude of the oval is 78^0, and the images showed a bright emission at local dawn similar to the dawn storms seen on Jupiter. No evidence was found for auroral emissions from the magnetic footprints of any satellites, including Titan. All emissions appeared to remain fixed in local time as the planet rotated, in contrast to the situation at Jupiter, but perhaps biased by the presence of the dawn storm. STIS images have also been obtained of Saturn's aurora (Figure 13), with sufficient sensitivity to begin to observe fine structure in the auroral ovals. These images showed bright spots along the main southern oval, and extended H Lyα emission above the limb of the planet (Trauger et al. 1998b). As Saturn's southern pole turns to face more directly toward the Earth, the observing geometry for the southern aurora is improving considerably, and upcoming STIS UV images should soon reveal the full extent of the auroral distribution on Saturn.

Finally, no successful *HST* detection of UV auroral emission from Uranus has been reported to date. This may be difficult to achieve in the present epoch. The changing orbital phase of Uranus now places the weak-field northern auroral zone, where most of

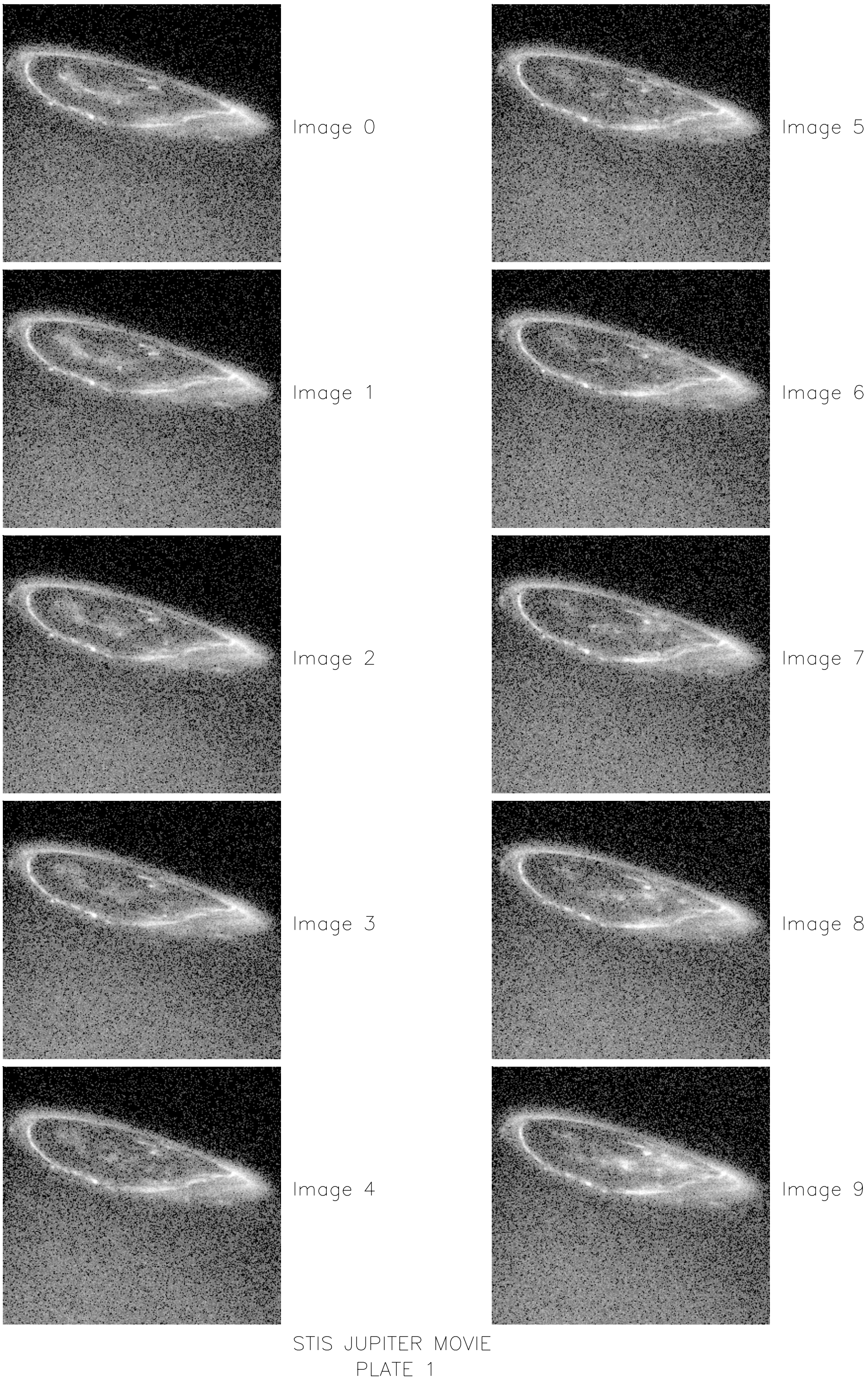


FIGURE 12. Frames of 30 sec exposures from STIS time-tag image data in July 1998 of Jupiter's north aurora with the SrF_2 filter. The polar cap emissions in the afternoon sector are seen to produce a sharp flare of emission in the last two frames.

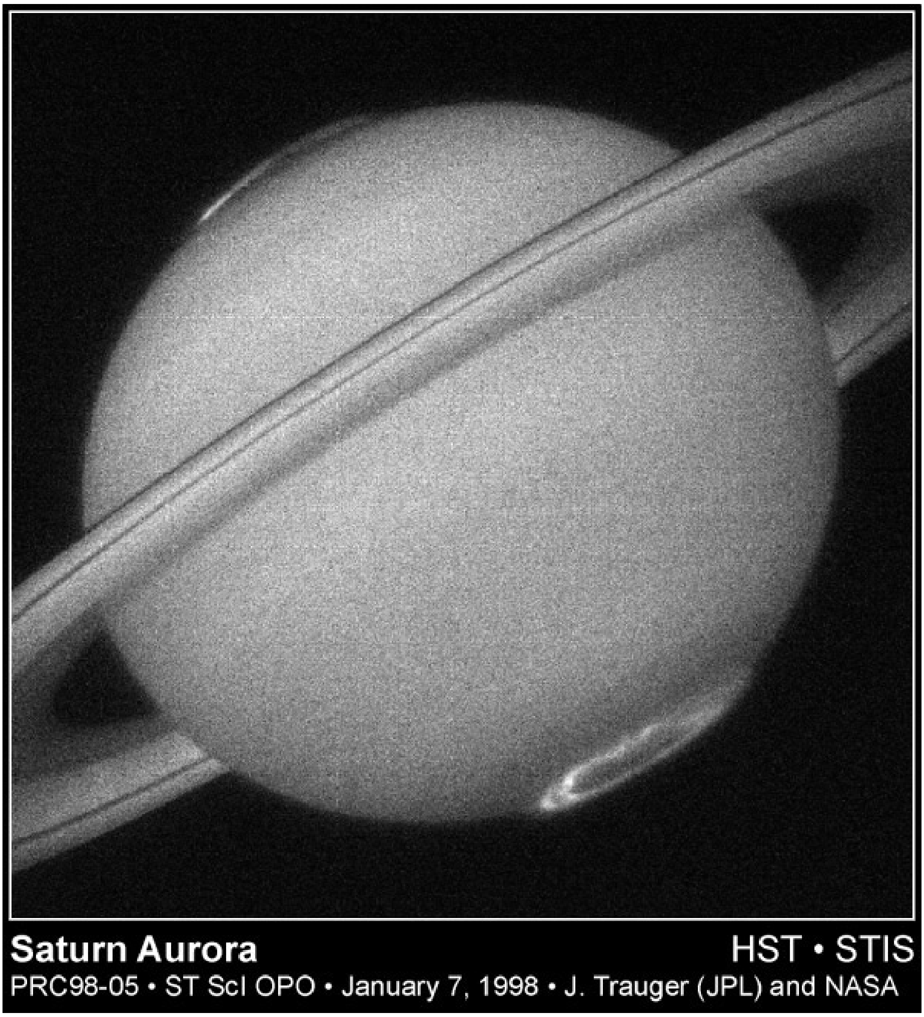

Saturn Aurora
HST • STIS
PRC98-05 • ST ScI OPO • January 7, 1998 • J. Trauger (JPL) and NASA

FIGURE 13. STIS UV image of Saturn's aurora, with colors produced from the combination of unfiltered (red) and SrF_2 filtered (blue) images, from Trauger et al. 1998b.

the emission is produced, on the far side of the planet for most of each Uranus rotation, and the rotational ephemeris of the planet is not known with sufficient accuracy to predict when to perform these observations.

I wish to acknowledge numerous grants over the years from ST ScI to the University of Michigan, most recently grant GO-08171.01-97A from the Space Telescope Science Institute to the University of Michigan. I wish also to thank D. Grodent and J. Caldwell for critical readings of the manuscript.

REFERENCES

ACUNA, M. H., NEUBAUER, F. M., & NESS, N. F. 1981 *J. Geophys. Res.* **86**, 8513.

ACUNA, M. H., BEHANNON, K. W., & CONNERNEY, J. E. P. 1983 in *Physics of the Jovian Magnetosphere*, p. 1. Cambridge University Press.

AJELLO, J. M., KANIK, I., AHMED, S., & CLARKE, J. T. 1995 *J. Geophys. Res.* **100, E12**, 26411.

Ajello, J., et al. 1998 *J. Geophys. Res.* **103, E9**, 20125.

Akasofu, S.-I. 1981 *Alaska Geographic* **6**, 2.

Ballester, G. E. & the WFPC2 Science Team 1996 *Science* **274**, 409.

Bauske, R., Combi, M. R., & Clarke, J. T. 1999 *Icarus* **142**, 106.

Belcher, J. W. 1987 *Science* **238**, 170.

Ben Jaffel, L., Leers, V., & Sandel, B. R. 1995 *Science* **269**, 951.

Bisikalo, D. V., Shmeatovich, V. I., Gérard, J.-C., Gladstone, G. R., & Waite, J. H. 1995 *J. Geophys. Res.* **101, E9**, 21157.

Broadfoot, A. L., et al. 1979 *Science*, **204**, 979.

Broadfoot, A. L., et al. 1981 *Science* **233**, 74.

Brown, R. A. & Yung, Y. L. 1976 in *Jupiter*, (ed. N. Gehrels), p. 1102. Univ. of Arizona Press.

Caldwell, J., Turgeon, B., & Hua, X. M. 1992 *Science* **257**, 1512.

Clarke, J. T., Moos, H. W., Atreya, S., & Lane, L. 1980 *ApJ* **241**, L179.

Clarke, J. T., Moos, H. W., Atreya, S., & Lane, L. 1981 *Nature* **290**, 226.

Clarke, J. T., BenJaffel, L., Vidal-Madjar, A., Gladstone, R., Waite, H., Prangé, R., Gérard, J. C. & Ajello, J. 1994 *ApJ* **430**, L73.

Clarke, J. T., et al. 1995 *Science* **267**, 1302.

Clarke, J. T. & the WFPC2 science team 1996 *Science* **274**, 404.

Clarke, J. T., Ballester, G., Trauger, J., Ajello, J., Pryor, W., Tobiska, K., Connerney, J., Gladstone, R., Waite, H., Ben Jaffel, L., & Gérard, J. C. 1998 *J. Geophys. Res.* **103, E9**, 20217.

Clarke, J. T., et al. 1999 *EOS, Trans. AGU* **80**, 46, F622.

Connerney, J. E. P., Baron, R., Satoh, T., & Owen, T. 1993 *Science* **262**, 1035.

Connerney, J. E. P., Acuna, M., Ness, N., & Satoh, T. 1998 *J. Geophys. Res.* **103, A6**, 11929.

Dols, V., Gérard, J. C., Paresce, F., Prangé, R., & Vidal-Madjar, A. 1992 *Geophys. Res. Lett.* **19**, 1803.

Dols, V., Gérard, J. C., Clarke, J. T., Gustin, J., & Grodent, D. 2000 *Icarus* **147**, 251.

Dougherty, M. K., Dunlop, M., Prangé, R., & Rego, D. 1998 *Planet. Space Sci.* **46**, 531.

Drossart, P., et al. 1989 *Nature* **340**, 539.

Gérard, J. C., Dols, V., Paresce, F., & Prangé, R. 1993 *J. Geophys. Res.* **98**, 18793.

Gérard, J. C., Dols, V., Prangé, R., & Paresce, F. 1993 *Planet. Space Sci.* **42**, 905.

Gérard, J. C., Grodent, D., Dols, V., Prangé, R., Rego, D., Waite, H., Gladstone, R., Ben Jaffel, L., & Ballester, G. 1994 *Science* **266**, 1675.

Gérard, J.C., Dols, V., Grodent, D., Waite, H., Gladstone, R., & Prangé, R. 1995 *Geophys. Res. Lett.* **22**, 2685.

Goldreich, P. & Lynden-Bell, D. 1969 *ApJ* **156**, 59.

Harris, W. M., Clarke, J. T., McGrath, M. A., & Ballester, G. E. 1996 *Icarus* **124**, 350.

Judge, D. L., Wu, F. M., & Carlson, R. W. 1980 *Science* **207**, 431.

Kim, Y. H., Caldwell, J., & Fox, J. L. 1995 *ApJ* **447**, 906.

Kliore, A., Cain, D. L., Fjeldbo, G., & Seidel, B. L. 1974 *Science* **183**, 323.

Liu, W. & Dalgarno, A. 1996 *ApJ* **467**, 446.

Livengood, T. A., Moos, H. W., Ballester, G. E., and Prangé, R. M. 1992 *Icarus* **97**, 26.

McGrath, M. A. & Clarke, J. T. 1992 *J. Geophys. Res.* **97**, 13,691.

Morrissey, P. F., Feldman, P., Clarke, J., Wolfven, B., Strobel, D., Durrance, S., & Trauger, J. 1997 *ApJ*, **476**, L918.

Prangé, R., Engle, I., Clarke, J., Dunlop, M., Ballester, G., Ip, W., Maurice, S., & Trauger, J. 1995 *Science* **267**, 1317.

Prangé, R., Rego, D., Southwood, D., Zarka, P., Miller, S. & Ip, W. 1996 *Nature* **379**, 323.

Prangé, R., Maurice, S., Harris, W., Rego, D., & Livengood, T. 1997 *J. Geophys. Res.* **102**, 9289.

Prangé, R., Rego, D., Pallier, L., Emerich, D., Ben Jaffel, L., Ajello, J., Clarke, J., & Ballester, G. 1997 *ApJ* **484**, L169.

Prangé, R., Rego, D., Pallier, D., Connerney, J., Zarka, P., & Quiennec, J. 1998 *J. Geophys. Res.* **103, E9**, 20195.

Trafton, L. M., Gérard, J.-C., Munhoven, G., & Waite, J. H. 1994 *ApJ* **421**, 816.

Trauger, J. T. & the WFPC2 science team 1998 *J. Geophys. Res.* **103, E9**, 20237.

Trauger, J. T. & the WFPC2 science team 1998 *BAAS* **30**, 1097.

Vasavada, A. R., Bouchez, H., Ingersoll, A. P., Little, B., & Anger, C. D. 1999 *J. Geophys. Res.* **104, E11**, 27133.

Vincent, M. B. & the WFPC2 science team 1999 *Icarus* **143**, 205.

Star formation

By JOHN BALLY

Center for Astrophysics and Space Astronomy, CASA, Campus Box 389, University of Colorado, Boulder, CO 80309

The angular resolution of *HST* has provided stunning images of star forming regions, circumstellar disks, protostellar jets, and outflows from young stars. *HST* has resolved the cooling layers behind shocks, and enabled the determination of outflow proper motions on time scales less than the post-shock cooling time. Observations of the best studied region of star formation, the Orion Nebula, has produced many surprises. *HST*'s superior resolution led to the identification of many new outflow systems based on their proper motions, the discovery of dozens of microjets from young stars, and the detection of wide-angle wind-wind collision fronts. *HST* has also produced spectacular images of circumstellar disks which have led to a rethinking of some aspects of planet formation. It now appears that most stars in the sky are born in environments similar to the Orion Nebula where within a few hundred thousand years after their formation, proto-planetary disks are subjected to the intense radiation fields of nearby massive stars. As a result, Orion's disks are rapidly evaporating. But at the same time their dust grains appear to be growing. Multi-wavelength images indicate that most of the solid mass in these disks may already be in large grains, possibly larger than a millimeter in size. The formation frequency of planets and the architectures of planetary systems will be determined by the competition between grain growth and photo-evaporation.

1. Introduction

Stars are the fundamental building blocks of the Universe. They play a role in astrophysics similar to that of atoms in chemistry. Most of the visible light in the Universe is produced by individual stars and all visible large scale structures such as star clusters, galaxies, and galaxy clusters are built from them. Stars produce the chemical elements. During their brief but brilliant lives, high mass stars convert hydrogen and helium inherited from the Big Bang into the other elements found on the periodic table, including those that make life possible. Their radiation, stellar winds, and terminal explosions energize and determine the state of the interstellar medium. In death, they enrich the interstellar gas with their thermonuclear fusion products. While the short lived massive stars produce the elements, the long lived low mass stars provide the stable environments which make life possible on planets like Earth.

Star formation is the astrophysical process which determines the fate of baryons in the Universe. By forming a spectrum of stellar masses, star formation determines the mix of long lived low luminosity stars and short lived high luminosity ones. Thus, star formation ultimately determines the rate of chemical enrichment in the Universe, the luminosity of baryonic matter, and the rate at which baryons are removed from the interstellar gas to be locked up in stars that can live nearly a Hubble time or longer. Thus, star formation is one of the key processes which determines how galaxies form and evolve. Indeed, a detailed understanding of star formation is required to follow the evolution of baryonic matter from the early Universe to the present.

By the middle of the second half of the twentieth century we developed a good understanding of the structure, evolution, and death of stars. Star formation, however, remained poorly understood until the discovery of molecular clouds in the 1970s and the development of infrared and millimeter-wave techniques that could probe the physical conditions inside the highly obscured environments in which most stars form. By the early 1990s, when the Hubble Space Telescope was launched, theoretical and observational advances led to a general outline of the star formation process.

The unprecedented angular resolution of *HST* has revolutionized our understanding of several crucial aspects of star formation. *HST* has obtained stunning images of a large variety of star forming regions, including dense clouds seen in silhouette such as the beautiful elephant trunks projecting into the interiors of ionized nebulae (the pillars of M16; Hester et al. 1996). *HST* obtained spectacular images of the outflows and jets powered by forming stars. It has been used to resolve the structures of radiative shocks, measure their proper motions, and investigate the evolution of the flows. *HST* has produced dramatic high resolution images of circumstellar and proto-planetary disks. The analyses of these images indicates that the first steps of planet formation may be occurring in some of these systems.

1.1. *How do stars form?*

Stars form from the gravitational collapse of dense cores in molecular clouds. The collapse can be slow if the cloud is supported by magnetic fields and the gas must diffuse through the field. Or, it can be rapid if gravity overwhelms the support pressure. Since cloud cores tend to be much more massive than typical stars, they must fragment as they contract. Indeed observations show that most stars form as parts of multiple star systems or dense stellar groups. However, since the efficiency of star formation (defined as the ratio the total mass of stars formed divided by the total initial gas mass) is low, typically around a few percent, most of these clusters will tend to fly apart once the remaining gas left over from star formation is dissipated.

Star forming cloud cores or fragments have orders of magnitude more specific angular momentum than even the most rapidly spinning stars. Even in the absence of any initial spin, tidal forces in a turbulent but fragmenting cloud core can lead to the stochastic generation of angular momentum. Although the very low angular momentum material may directly fall into the center of a star forming core, most of the mass will collapse onto a spinning protostellar disk. It is generally believed that magnetic fields provide the viscosity responsible for the outward transport of angular momentum and accretion onto the central protostar at rates of order $\dot{\mathrm{M}} \sim 10^{-5}$ $\mathrm{M}_\odot$ yr^{-1} (cf. Stahler 2000).

Low mass protostars are fully convective and tend to spin fast ($P \sim 1$ to 5 days). Thus, a strong stellar dynamo is believed to rapidly generate a stellar magnetosphere whose dipole component can disrupt the surrounding accretion disk at 5 to 10 stellar radii. Material from the disk is lifted by the strong dipolar B-field and channeled to high stellar latitudes in a so-called funnel flow. The rotation rate of the star becomes slaved to the Keplerian rotation of the disk at the point where it is disrupted (Königl 1991). As parcels of gas are accreted, the star must spin-up slightly, driving the field through the inner edge of the accretion disk at slightly faster than Keplerian velocities. Super-Keplerian plasma may be launched onto field lines where centrifugal forces drive it and the entrained field lines to infinity (cf. Shu et al. 1994a, 1994b, 1995, 2000). Thus, accretion may be associated with mass loss. In addition to this so-called 'X-wind,' the star may drive a normal stellar wind. The disk may also launch its own mangeto-centrifugal wind (cf. Königl & Ruden 1993; Königl & Pudritz 2000). In this model, the outflow, which moves along open field lines, removes the excess angular momentum released by accretion onto the star through the rigid co-rotating magnetosphere.

2. Proto-Planetary disks

The direct imaging of proto-planetary disks is one of the "Holy Grails" of NASA's Origins program. Amazingly, *HST* has achieved this goal. Ten years ago most astronomers would have argued that such disks are likely to be so embedded within the remnants of their parent cloud cores as to be only observable at infrared to millimeter wavelengths. Yet *HST* has produced direct images of dozens of disks around young stars in dark cloud environments, in HII regions such as the Orion Nebula, and even surrounding main-sequence stars.

Observations of disks in the Orion Nebula indicate the intriguing possibility that their solids have coagulated into grains with sizes at least as large as several μm, and possibly larger than a few millimeters. Thus, these disks may have already evolved through the first steps of planet formation. However, at the same time, the intense radiation field of the Trapezium stars is rapidly destroying these disks.

2.1. *Disks in dark clouds and around older stars*

The Taurus molecular cloud complex (d $\approx$ 140 pc) contains some of the nearest young stars. The *HST* imaging surveys of young stars in Taurus has led to a rich harvest of stunning images of disks and circumstellar environments with a resolution of order 15 AU (Padgett et al. 1999; Stapelfeldt et al. 1998a, 1998b; Krist et al. 1999; Burrows et al. 1996; Stapelfeldt et al. 1995). Many of these protostars produce extended reflection nebulae within their envelopes. If sufficiently edge-on, their circumstellar disks can be seen in silhouette against these nebulae (e.g. IRAS 04302+2247, HK Tau, and IRAS 04248+2612). In some cases, spectacular jets can be seen to emerge from the central parts of these disks (e.g. HH 30, DG Tau B, and Haro 6−5B). In many objects where ground based images show a point source, *HST* reveals a bright and compact reflection nebula with the central completely obscured (cf. HL Tau—Stapelfeldt et al. 1995). These reflection nebulae are often highly variable in both their intensity and structure. Either moving blobs of opaque material near the central star or local variations in the light output from the stellar surface produce moving and time-variable illumination patterns (Stapelfeldt et al. 1999).

HST has also produced stunning images of a variety of more evolved circumstellar disks surrounding more mature stars with ages ranging from 10^6 to 10^8 years. Examples include the dust rings and disk surrounding HR 4796A (Schneider 1999), HD 141569 (Weinberger 1999), and AB Auriga (Grady et al. 1999). In the latter case, the STIS coronographic images show a bewilderingly complex surface structure consisting of spiral arcs, blobs, and gaps. The nearly edge-on disk surrounding the star Beta Pictoris has also been extensively observed with *HST* (cf. Kalas et al. 2000). Detailed direct imaging has revealed complex warping and brightness asymmetries in the outer parts of the disk which can be interpreted as indirect evidence for orbiting planets.

2.2. *Evaporating disks in HII regions*

The Orion Nebula never fails to amaze. One of the most stunning results of *HST* is the discovery of evaporating circumstellar disks (proplyds) surrounding over a hundred young low mass stars embedded within the Orion Nebula (O'Dell, Wen, & Hu 1993; O'Dell & Wong 1996; McCaughrean & O'Dell 1996; Bally et al. 1998, 2000). The *HST* images show that up to 80% of the stars near the Trapezium, and more than 50% in the outer parts of the nebula are surrounded by tails pointing directly away from θ^1 Ori C or one of the other high mass stars in the region.

All stars associated with externally ionized tails and nebulosity are believed to contain evaporating circumstellar disks. In the initial pre-aberration corrected WFPC1 observations of O'Dell, Wen, & Hu (1993), a nearly edge-on disk could be seen in the center of one of the larger objects, *HST* 10 (182−413). Furthermore, these authors found one disk, *HST* 16 (183−405), seen in silhouette against the background nebular light. *HST* 16 does not have a bright ionized skin or tail, and thus must lie sufficiently far in the foreground so that the ionizing radiation of the Trapezium fails to produce an ionized skin brighter than the nebular background.

Much sharper and deeper follow-up observations with WFPC2 have produced direct images of over 40 disks in the Orion Nebula. McCaughrean & O'Dell (1996) found six disks seen in silhouette against the background nebular light that are not surrounded by ionized skins. Disks are also seen near the heads of over 30 of the largest proplyds having ionized skins and tails (Bally et al. 1998; Bally, O'Dell, & McCaughrean 2000). The latter authors also found additional disks seen only in silhouette, increasing the total number of dark disks found so far to 15.

The inner portion of the Orion Nebula that has been surveyed so far with *HST* contains about 300 young stars of which about 150 are associated with extended circumstellar structure (O'Dell & Wong 1996). The extended Orion Nebula cluster of low mass stars contains about 2,000 members with ages ranging from about 10^5 to 10^6 years (Hillenbrand 1997; Hillenbrand, & Hartmann, 1998). Thus, many more proplyds are likely to be found in future observations.

Johnstone et al. (1998) present a model of these externally irradiated disks. Soft-UV ($912 < \lambda < 2000$ Å), which penetrates the ionization front produced by Lyman continuum ($\lambda < 900$ Å) photons shines on the disk surface and heats it to a temperature of order 10^3 Kelvin, producing a photo-dissociation or photon-dominated region (PDR). The heated gas layer expands at about the sound speed, $c_s \approx 3$ km s^{-1}. However, this expanded disk corona is bound to the central star out to a radius roughly given by $r_G \approx GM_*/c_s^2$ which ranges from about 10 to 100 AU for 0.1–1 M$_\odot$ stars. Beyond this radius, the disk corona expands as a low velocity neutral wind and it can be shown that this wind shields the disk from Lyman continuum photons. However, spherical divergence guarantees that the outer parts of the wind do become ionized. Typically, an ionization front forms at roughly 2 to 3 disk outer radii. It is this ionization front that makes the bright proplyds so conspicuous. The penetration depth of the soft-UV radiation and the mass loss rate from the proplyd is self-regulated by dust entrained in the neutral flow. For normal ISM grain properties and gas-to-dust ratio, the soft-UV is attenuated within a hydrogen column of order N(H) $\sim 10^{21}$ cm^{-2}, and the typical photo-ablated mass loss rate from the disk is predicted to be around $\dot{\mathrm{M}} \sim 10^{-7}$ M$_\odot$ yr^{-1}.

Observations have shown that Orion's proplyds are evaporating with mass loss rates of order 10^{-7} to 10^{-6} M$_\odot$ yr^{-1} (Henney & O'Dell 2000). The 1.3 mm continuum measurements (Bally, Testi, & Sargent 1998) imply an upper bound on the mass of order 10^{-2} M$_\odot$ *assuming normal interstellar grain properties*. Thus, these disks can survive for only about 10^4 to 10^5 years. The large number of stars with proplyd characteristics implies that photo-evaporation started no more than 10^5 years ago, and possibly more recently.

Photo-erosion causes the disks to shrink. As the disk outer radius approaches r_G, the mass loss rate declines, the ionization front moves in towards the disk, and eventually reaches its surface. Then, the sound speed at the disk surface increases to about 11 km s^{-1} and r_G declines by about an order of magnitude. Then, disk erosion via direct ionization can continue until the disk radius reaches the new value of r_G corresponding to the higher sound speed ($r_G \approx 1$ to 10 AU for 0.1–1 M$_\odot$ stars). The ionized thick-disk corona then becomes confined by the central star's gravity.

2.3. *Evidence for large grains in proplyds*

Unlike any other type of interstellar dust, Orion's proto-planetary disks are grey. Recent *HST* results have shown that the translucent outer portions of these disks do not redden background light (Throop, Bally, & Esposito 2000; see Figure 1). The neutral (or grey) colors of the translucent disk edges imply that the absorbing particles are at least several times larger than the longest wavelength at which the disk colors were measured. The implication is that the typical particles responsible for extinction are large compared to the wavelength of light.

More careful modeling supports this conclusion. A 3-dimensional disk model is projected onto the plane of the sky, convolved with the Hubble Space Telescope point-spread-function, and compared to the *HST* data. This procedure is repeated hundreds of thousands of times in a Monte Carlo simulation in which the grain properties and size distributions are varied to map out the acceptable range of values. This method also shows that large grain distributions fit the observations well.

Furthermore, *HST*'s infrared camera NICMOS failed to see the central stars in several of Orion's edge-on disks. This implies large extinction. For normal ISM dust, the failure to detect the central stars in *HST* 10 (182−413) and the giant disk 114−426, implies $A_V > 60$ magnitudes (McCaughrean et al. 1998; Chen et al. 1998). Thus, there must be a very large amount of circumstellar material which ought to be detectable at millimeter and sub-millimeter wavelengths.

But, searches for 1.3 mm wavelength continuum emission from these (presumably most massive) circumstellar disks in Orion have only produced upper limits on the dust emission. At 1.3 mm, the continuum fluxes from the proplyds are below 20 mJy (Bally, Testi, & Sargent 1998). Assuming normal interstellar grain properties, the maximum amount of dust in the proplyd *HST* 10 (Figure 2) which is consistent with the radio data is nearly an order of magnitude lower than the minimum amount of dust required to explain the > 60 magnitudes of visual extinction needed to hide the central stars. One way to resolve this conflict is to assume that the majority of gains are larger than 1.3 mm so that most of the dust mass is not probed by the radio observations.

Combining the NICMOS result with the OVRO limits implies that a large fraction of the mass in these disks may be locked-up in particles larger than the OVRO wavelength of 1.3 mm. Thus, the observations imply the existence of millimeter sized (or larger) particles in the proplyds.

Throop (2000) has constructed a model of grain evolution in which he considers grain growth via particle sticking and ice mantle formation, and grain destruction via photo-evaporation and collisions. These models combine standard disk evolution models with photo-evaporation. The models predict that for the conditions typical of the Solar Nebula, grains can easily grow to sizes of order centimeters to meters within the roughly 10^5 years available prior to irradiation by the massive Trapezium stars. The outer edges of these disks are predicted to be sharply truncated by the photo-evaporative flow from the photo-dissociation region. The observations are consistent with such disk truncation.

2.4. *Consequences for the origins and architectures of planetary systems*

These results have profound implications for planet formation in disks embedded in HII regions. In Orion, the disks are evaporating. At they same time, their solids appear to be coagulating into larger bodies. The outcome of the competition between grain growth and mass loss will determine whether planets will eventually form around these stars or not. This competition may impact whether planets are common or rare in the Universe.

The volatile (H, He, CO, etc.) and small grain components of these disks are lost at rates of order 10^{-7} $M_\odot$ yr^{-1}. Thus, these components will disappear in 10^4 to 10^5

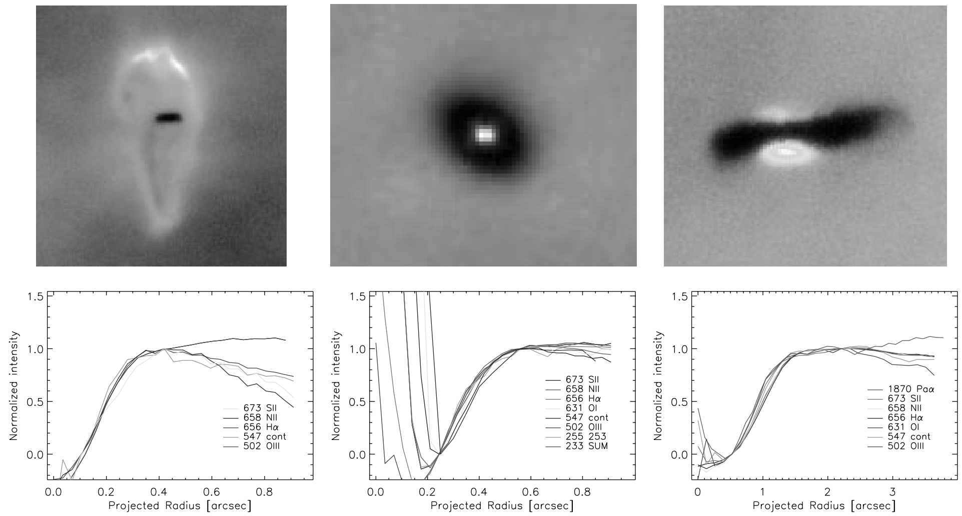


FIGURE 1. Images (top row) and radial brightness profiles (bottom row) of three proplyds; *HST* 10 (left), *HST* 16 (middle), and 114−426 (right). The various line types show the radial brightness profiles at wavelengths (in nm) indicated in the lower right portion of each panel. Taken from Throop (2000) and Throop, Bally, & Esposito (2000).

years after UV exposure begins. But, as the above results imply, enough material may already be locked into large bodies so that rocky planets may eventually form around these stars. However, the resulting planetary systems are likely to lack gas giants such as Jupiter, unless such giants form prior to the onset of irradiation. Thus, the amount of time available to form giant planets is only about 10^5 to 10^6 years, the difference between the ages of the oldest stars in the Orion Nebula cluster and the photo-ionization age of the nebula. Therefore, gas giant planets must either form very fast by a process such as gravitational instability in the disk, or they will be absent from planetary systems formed in Orion-like environments.

A critical issue for the formation frequency of planets around stars is the fraction of stars that form in irradiated (Orion-like) environments. Preliminary estimates show that the majority of young stars are likely to be born in such environments.

3. Where do most stars form?

Star formation within the Solar vicinity (within 600 pc of the Sun) provides a unique environment in which to seek an answer. Our cosmic backyard is the only place in the Universe where we have some understanding of the locations and velocities of stars in six phase-space dimensions. We can determine the positions on the sky, distances, radial velocities, and proper motions for many stars and associated clouds. Furthermore, we have a reasonably complete census of all star forming regions within the Solar vicinity.

Using the best estimates for the Galactic star formation rate (about 3 $M_\odot$ yr^{-1}), and a standard IMF, about 20,000 to 50,000 stars were born within 500 pc of the Sun within the past 10 Myr. The vast majority of star formation occurs in OB associations because the mass spectrum of molecular clouds has a power law index of about −1.6 which implies that most of the molecular mass of the Galaxy is contained in the largest objects, the giant molecular clouds (GMCs). Near the Sun, GMCs have masses of order 10^5 $M_\odot$ and tend to form OB associations. The Solar vicinity contains three major actively forming

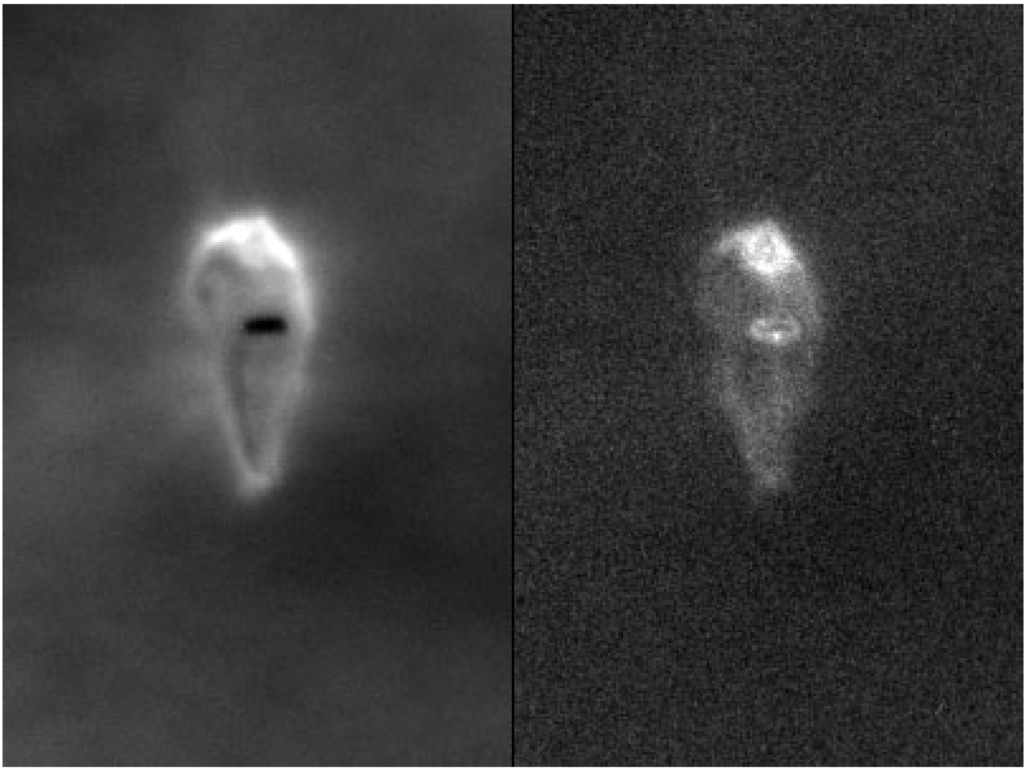


FIGURE 2. Two *HST* images of the edge-on disk 182−413 (*HST* 10) which is embedded within a bright proplyd. The disk is seen in silhouette in both Hα (left) and in the λ = 6300 Å [O I] line (right). In [O I], however, the disk surface is bright on both sides. This emission traces the heated layer at the base of the neutral photo-ablation flow (Störzer & Hollenbach 1998, 1999). A faint filament of [O I] emission extends along the disk axis. This feature may be a microjet (see Bally, O'Dell, & McCaughrean 2000 for details). The central star remains obscured even at 2μm (Chen et al. 1998).

OB associations; Orion OB1, Per OB2, and the Sco-Cen OB association which are less than 10^7 years old (Blaauw 1991; Brown et al. 1994, 1995).

Some star formation does occur within smaller dark cloud complexes. The nearest such dark cloud star forming regions lie within 200 pc of the Sun. These include the Taurus, Corona Australis, ρ-Ophiucus, and Lupus clouds (the latter three regions are on the outskirts of the Sco-Cen OB association and may be associated with it). Within these regions, young stars are shielded from the harsh radiation fields of massive stars. The Solar vicinity contains about 10 dark cloud complexes such as Taurus which have spawned T associations consisting of only low mass stars. If these clouds are typical, the total number of stars born in dark clouds within 600 pc of the Sun within a 10 Myr interval is only a few thousand. Thus, over 90% of all stars are likely to be born in OB associations.

Within Orion-like giant molecular cloud environments, the majority of stars are born within a few parsecs of massive stars. In the Orion A cloud, the Orion Nebula region (OMC1, 2, and 3) contains at least 2,000 young stars (Hillenbrand & Hartmann 1998). Most of these are eventually exposed to powerful radiation fields. In contrast, the entire Orion A molecular cloud south of the Nebula (Bally et al. 1987; 1991) contains only several hundred stars. From our current knowledge of Orion, we can infer that about

80–90% of the YSOs in an OB association are born in dense clusters near massive stars. The majority of stars in a typical OB association are produced in dozens of star forming events similar to the one which spawned the Orion Nebula and its young cluster.

Star counts and cluster modeling of the YSOs in the Orion Nebula indicates that within the central 0.05 pc region, the volume density of young stars exceeds 10^5 stars per cubic parsec (Henney & Arthur 1998). This is more than 10^6 times the density of field stars around the Sun. Therefore, the nearest neighbor distances between young stars in the core of the Orion Nebula cluster is only a few thousand AU, comparable to the observed and expected dimensions of proto-planetary disks. Despite their high stellar density, such ultra-dense clusters of young stars are usually *not* gravitationally bound and will therefore rapidly disperse soon after their formation. But, stars within such groups are exposed to harsh UV radiation, and subjected to violent stellar dynamical interactions during the first few million years of their lives. Thus, a preliminary census of young stars near the Sun implies that the majority of stars are born in highly clustered environments in proximity to massive stars.

3.1. *The formation of bound clusters*

HST has obtained images of some spectacular regions of clustered star formation. Examples include NGC 3603 (Brandner et al. 2000) and 30 Doradus in the LMC (Walborn et al. 1999a, 1999b; Rubio et al. 1998) that contain between 30 and 200 O stars. Unlike most clusters near the Sun, which appear to dissolve rapidly once a few O stars form, these regions may survive as gravitationally bound open clusters. A key difference between these bound cluster forming events and the more common unbound cluster forming clouds is that the molecular clouds producing the former have line-widths larger than the sound speed in ionized gas. In clouds such as Orion, where the cloud line-width is only a few km s^{-1}, ionization can dissipate the star forming gas. However, if the cloud escape speed is larger than the sound speed in a photo-ionized plasma, O stars may not be as disruptive, and star formation can continue despite the birth of massive stars until a supernova explosion shatters the cloud. In such an environment star formation may have higher efficiency, converting a sufficient fraction of the initial gas mass into stars so that when the gas is dispersed, the cluster remains bound by its own gravity.

Recent NICMOS images of young clusters in the vicinity of the Galactic center have also revealed spectacular examples of clustered star formation occurring in large line-width GMCs (Figer et al. 1999). But in this environment, the Galactic differential rotation and associated tidal field may rip apart the resulting clusters.

Globular cluster forming events may be even more spectacular. No such events are seen within the Local Group of galaxies (unless 30 Dor becomes such a cluster as some have suggested). But, it is possible that some of the star forming regions in colliding galaxies such as the Antennae (Gilbert et al. 2000) and NGC 1275 (Brodie et al. 1998) may be spawning globular clusters. Perhaps the conditions for globular cluster formation require that the parent cloud be so massive and dense that star formation can proceed despite one or more supernova explosions.

3.2. *Hazards to planet formation*

Observations have shown that most stars are born in multiple star systems and/or in highly over-dense but short-lived clusters that fly apart soon after birth (e.g. Testi & Sargent 1998). Furthermore, numerical modeling has demonstrated that rotating clouds can shed their excess angular momentum by fragmenting into non-hierarchical multiple protostellar groups orbiting each other, or into dense swarms of stars. Thus, stars are almost never born alone even in relatively isolated dark clouds.

Multiplicity and clustering introduce stochastic processes into protostellar evolution and planet formation because non-hierarchical multiple systems are dynamically unstable (cf. Valtonen & Mikkola 1991). They exhibit chaotic orbital evolution which within something like 10 to 100 orbits leads to disruption of the system. The most common end result is that a non-hierarchical triple star system expels its lowest mass member, leaving behind a tightly bound binary consisting of the two most massive members. Though such interactions may at first sight appear to be rare, they may be common in young star systems (Reipurth et al. 1999; Reipurth 2000). They may play a central role in determining the masses of stars, the fraction of single and multiple stars in the sky, and may seriously impact the formation frequency of planetary systems. Three body interactions may not only be common, they may be responsible for the observed traits of young stars and binaries (Reipurth 2000).

In dense clusters and multiple systems a number of destructive forces curtail the collapse process and may disrupt proto-planetary disks. These include three-body interactions that can destroy nascent proto-planetary disks, disk + star/disk + star interactions in eccentric binaries and dense groups that shear, truncate, and disrupt disks, and, as discussed above, UV induced photo-ablation by nearby massive stars.

The high degree of clustering and multiplicity of the vast majority of protostars has profound implications for the formation rate of exo-planets. Observations and theory are currently driving a shift in our understanding of star formation and a new paradigm is in the making. Three body interactions, clustering, and intense radiation fields introduce a highly stochastic element into our view of steady early stellar evolution. This paradigm shift is likely to greatly impact our understanding of the evolution of circumstellar disks, the formation rates, and expected architectures of planetary systems.

4. Outflows from young stars

Outflows are signposts of stellar birth. The most common tracers of outflow activity are the millimeter wavelength lines of common molecules such as CO (Lada 1985; Bachiller 1996), the near IR lines of shock excited H_2 and [Fe II] (Eislöffel et al. 2000), and Herbig-Haro (HH) objects, which are shock excited visual wavelength nebulae powered by young stars. HH objects are most easily traced by their Hα and forbidden line emission, especially the $\lambda\lambda$ 6717/6731 Å [S II] doublet.

Outflows are ubiquitous, large, and have a profound impact on their surroundings. Proto-stellar outflows are ideal laboratories in which to study the properties of astrophysical jets in general. HH jets are near enough to the Sun so that with the angular resolution of *HST*, proper motions and flow evolution can be observed within a time span of a few months.

4.1. *Outflows from low mass stars*

The first HH objects were recognized around 1950. By the 1970s it became clear that these enigmatic nebulae found in or near dark clouds in star forming regions were radiative shocks (for a recent review, see Hartigan et al. 2000a). By the mid 1980s, several HH objects were found to consist of highly collimated jets consisting of dozens of closely spaced shocks. Nearly 500 HH objects have been discovered (Reipurth 1999). Several dozen consist of highly collimated jets while the rest are either bow shaped or highly irregular and complex objects. Many HH flows are also associated with near-IR H_2 and [Fe II] emission as well as millimeter wavelength CO emission which is probably entrained from the surrounding molecular cloud (Chernin & Masson 1995; Chernin et al. 1994).

Recently, is was recognized that outflows can propagate many parsec from their sources. Dozens of parsec scale flows have been identified (Bally & Devine 1994; Reipurth, Bally, & Devine 1997; Devine et al. 1997; Reipurth, Devine, & Bally 1998) and many molecular clouds are pockmarked with cavities that were produced by flows that entrained surrounding gas and punched out of their parent clouds (Bally et al. 1987; Bally et al. 1999). Giant outflows may be a major source of turbulent motions in molecular clouds. Furthermore, their terminal shocks can dissociate molecules and may contribute to the chemical rejuvenation of molecular clouds. Outflows may be one of the key mechanisms by which star formation is self-regulated. Many well-known Herbig-Haro jets such as HH 1/2, HH 34 (Heathcote et al. 1996), and HH 111 (Reipurth et al. 1996) are merely the inner parts of such giant flows.

Most Herbig-Haro objects trace radiative shocks where fast outflow components overtake slower moving and older ejecta. Thus, many HH objects trace internal shocks within an outflow lobe rather than interactions of ejecta with pristine interstellar material. The characteristic thickness of the radiating layer is given by the cooling length $\mathrm{d_{cool}} \approx 5 \times 10^{14} \mathrm{n_{100}^{-1} V_7^4}$ (cm) where n_{100} is the density in units of 100 cm^{-3} and V_7 is the shock speed in units of 10^7 cm s^{-1}. This corresponds to 30 AU, or about 0.1″ at a distance of 300 pc. Since the nearest HH objects are located a bit more than 100 pc from the Sun, the sub-arcsecond resolution of *HST* is required to distinguish the shocks from the post-shock cooling layers.

Figure 3 shows an *HST* image of the prototypical HH 1/2 outflow system in Orion. HH 1 and 2, the first HH objects to be recognized (Herbig 1951; Haro 1952), consist of a pair of oppositely oriented bow shocks separated by about 2.5′ (0.34 pc projected). The source of this bipolar flow consists of a highly embedded multiple star system that has been resolved in the radio continuum (Rodríguez et al. 1999) but remains invisible below 3 μm. A high velocity jet emerges from the opaque cloud core several arc seconds north of the source. The ratios of proper motions and radial velocities indicate that the jet is inclined 10° with respect to the plane of the sky and moving with speeds ranging from 200 to 380 km s^{-1} (Herbig & Jones 1981; Eislöffel, Mundt, & Böhm 1994). As it emerges from the obscuring cloud core, the jet first becomes visible in the 1.64 μ [Fe II] and 2.12 μm H_2 lines, then a few arc seconds downstream in [S II]. It gradually fades in these tracers about 15″ from the source. Reipurth et al. (2000) present high resolution NICMOS images of the HH 1 jet which points directly at the bright bow shock HH 1. Presumably, the obscuring cloud hides the counter jet aimed towards HH 2 located to the south. The HH 1 and 2 bow shocks are also bright in the near-IR lines of H_2 (Davis, Eislöffel, & Ray 1994). *HST* images of HH 1 and 2 show extremely complex substructure (Hester, Stapelfeldt, & Scowen 1998), indicating the onset of instabilities in the post-shock region. The source, jet, and the HH 1/2 bow shocks are surrounded by a low velocity ($<$ 10 km s^{-1}) molecular outflow visible in CO (Moro-Martín et al. 1999). Ogura (1995) found that the HH 1 and 2 bow shocks are merely the innermost and brightest components in a parsec-scale flow that can be traced to a giant bow shock HH 401 towards the northwest and HH 402 towards the southeast. The projected separation between this pair of bow shocks is 5.9 pc.

4.2. *Proper motions of Herbig-Haro Objects*

The high resolution of *HST* has enabled the determination of accurate proper motions from images obtained over a time interval of only a few years. The short time interval is comparable to the cooling time, which makes it possible to separate the brightening and fading of different parcels of gas from true motions. Several patterns are apparent in the proper motion data. The highest velocities are observed in the jet and along the flow axis

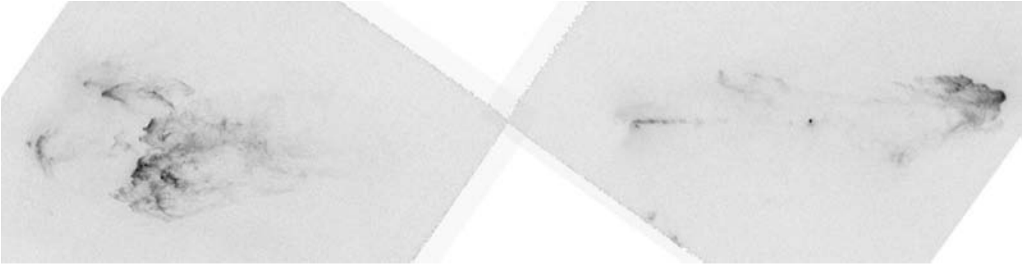


FIGURE 3. A narrow band *HST* image of the HH 1/2 outflow in [S II]. HH 1 is on the right and HH 2 is on the left. The driving source is embedded several arc seconds to the left of the jet (see Hester, Stapelfeldt, & Scowen 1999 for details).

that it defines. In the HH 1 and 2 bow shocks, the highest speeds are measured at the tips of the bows which lie along the axis defined by the jet. In these bows, the average flow velocity declines with increasing distance from the flow axis. Both bow shocks show very complex sub-structure indicating that instabilities have caused the fluid to fragment into dense clumps. Within the envelope of the large bow, smaller bow shocks surround these clumps. Some of these small-scale bows face forward while others face backwards. There is a general trend that the forward facing bow shocks have large proper motions with projected velocities of order 200 to 350 km s^{-1} while the reverse bow shocks move with speeds well under 200 km s^{-1}.

These trends can be explained by models in which the shocks trace internal working surfaces within the outflow where fast ejecta overtakes slower material. The flow pattern orthogonal to the jet axis suggests that a velocity variable jet is propagating through a more slowly moving wide angle flow. It is not clear from the existing data whether this slower wide angle flow represents a separate outflow component, or material entrained from the ambient medium by the passage of previous shocks powered by the axial jet. Instabilities, possibly related to the rapid cooling of shock heated gas, have produced clumps and a much lower density interclump fluid. In one scenario of shock evolution, the denser clumps of one fluid may penetrate into the less dense fluid of the other to produce both forward and reverse facing bow shocks. Alternatively, the present shock is propagating into the clumpy debris left behind by the passage of a previous shock.

4.3. *Properties of Herbig-Haro shocks*

Clumpy structure in the medium into which a shock is propagating is also evident in the detailed *HST* images of HH 29 in L1551 (Devine et al. 2000). Located at a distance of 140 pc, the L1551 dark cloud in Taurus contains the nearest bright interstellar shocks in the sky. This 40 $M_\odot$ cloud has produced over a dozen low mass stars, many of which are binaries. These young stars are crowded together in a region only a few tenths of a parsec in diameter. At least six of these young stars are actively driving jets and outflows into the surrounding medium. The most luminous source, L1551 IRS5 is a $0.3''$ separation binary (Rodríguez et al. 1998) and the *HST* images show that each member powers its own jet (Fridlund & Liseau 1998). The two brightest HH objects in L1551, HH 28 and 29, were at first thought to originate from this proto-binary. However, Devine, Reipurth, & Bally (1998) have recently found that the embedded source L1551NE (which is also probably a binary) is a much more likely driving source.

Recent *HST* observations (Devine et al. 2000) reveal the structure of HH 29 in unprecedented detail. The actual shock front is traced by a Balmer filament (Hα emission and no forbidden lines) while the complex post-shock cooling layer emits both in Hα and

[S II]. It appears that a cluster of 10 to 100 AU scale clumps of dense slowly moving gas are being overtaken by a faster ($\sim$ 200 km s^{-1}) lower density fluid. The clumps must have formed prior to the passage of the currently visible shock. It is possible that the slowly moving lumpy fluid is the cool remnant debris left behind by a shock that has long since faded from view.

In addition to IRS 5 and L1551NE, the L1551 cloud contains the famous HH 30 jet (Burrows et al. 1996) with its nearly edge-on circumstellar disk and the binary star XZ Tauri which is driving an expanding bubble into its surroundings (Krist et al. 1999). Finally, *HST* has shown that at visual wavelengths the star HL Tau is not a star at all, but a very bright and compact reflection nebula. The star itself remains hidden inside a thick molecular torus (Stapelfeldt et al. 1995). The L1551 cloud demonstrates that even in relatively isolated regions of star formation in dark clouds, young stars are born in small and dense clusters with star densities hundreds to thousands of times that of the field. Their outflows and jets frequently overlap on the plane of the sky. Near HH 29, three, and possibly four outflows criss-cross along our line of sight, including the twin jets from L1551 IRS5, the jet from L1551NE that drives HH 29, and possibly the outer region of the flow energized by HH 30.

HST has produced a series of remarkable images of protostellar jets. These include HH 1/2 (Hester et al. 1998), HH 46/47 (Heathcote et al. 1996), HH 111 (Reipurth et al. 1996), and HH 34. All have been imaged at least twice with *HST* at intervals of a few years and detailed proper motion vector fields are being determined. These images show dozens of internal working surfaces whose individual post-shock cooling zones overlap to form the nearly continuous body of the jet. The spacing of these internal shocks increase with increasing distance from the source until growing gaps separate them and the jet ceases to be a continuous luminous fluid. Discrete and well separated bow shocks are often found downstream from these jets.

The complex internal structures of these flows provides a fossil record of the flow velocity variations and subtle ejection direction orientation changes. The proper motion data make it clear that the knots in jets are internal working surfaces within the body of the flow. They are *not* standing shocks or Kelvin-Helmholtz instabilities. These knots move with the fluid and trace either the shocks or the post-shock cooling layers. The increasing gap size between successive shocks implies that large variations in the outflow speed occur over long time intervals (centuries to millennia) while small velocity variations occur much more frequently (years to decades). Thus the overall spectrum of flow velocity variations must behave like a 1/f noise process.

4.4. *Herbig-Haro jets and stellar multiplicity*

Near infrared observations with NICMOS have shown that a very large fraction of the sources that drive HH flows are multiple. One sample of Herbig-Haro energy sources (Reipurth et al. 2000) shows that at least 87% of HH energy source are doubles or higher order multiples. This is the highest stellar multiplicity fraction of any set of stars observed to date. These observations have prompted Reipurth (2000) to propose that the most powerful episodes of protostellar mass loss are triggered by dynamical three body interactions in multiple star systems. In this model periastron passage can cause major disk accretion events which fuel mass loss from the system.

The NICMOS and radio continuum VLA observations of the HH 111 source region provide evidence for such interactions. The NICMOS images (Reipurth et al. 2000) show the presence of two stars in the source region of HH 111. However, neither star is centered within the highly flattened envelope that lies orthogonal to the base of the HH 111 jet. While the brighter source lies on one side of this circumbinary disk, the fainter member is displaced far to the other side. Furthermore, the VLA radio data provides evidence that the brighter star is itself a binary. The radio source drives a pair of thermal radio jets nearly orthogonal to each other. One radio jet drives the bright HH 111 flow while the other drives the much fainter and less continuous HH 121 outflow propagating nearly at right angles to HH 111. Reipurth et al. (1999) propose that the HH 111 cloud core produced a non-hierarchical triple system which is inherently unstable. Within the last 20,000 years, a close stellar encounter in this system resulted in a re-configuration into a hierarchical configuration consisting of a tightly bound binary and a single star that was ejected. As a result, the binary is itself recoiling and all three stars are departing the cloud core. If this scenario is correct, then mass accretion has been brought to a halt by a three-body disruption of this unstable star system. Reipurth (2000) proposes that such three body interactions may play a role in determining the mass spectrum of stars.

4.5. *Irradiated jets and microjets*

As discussed above, most stars in the sky form in OB associations. Within OB associations, most stars form in dense transient clusters such as the Orion Nebula cluster and the somewhat older σ Orionis group which is 2 to 4 million years old (Walter et al. 2000). Thus, the discovery of jets from four separate low mass stars near σ Ori (Reipurth et al. 1998) was a surprise since the main protostellar outflow phase was thought to last $< 10^5$ years. These jets are powered by visible stars located far (more than several parsecs) from any molecular gas. Furthermore, the jets are externally ionized and rendered visible by the radiation field of the massive star σ Ori. If the age estimates for the σ Ori group are correct, then jet production by young stars can last for millions of years. A curious feature of these irradiated jets is that they are predominantly one-sided with the beam aimed towards σ Ori being about 10 times fainter than the beam facing away. Spectra show that this brightness asymmetry may be related to an underlying kinematic asymmetry. The fainter counter beam to HH 444 has a radial velocity that is about 2 to 3 times larger than the brighter beam. There is kinematic evidence that the slower beam has been decelerated by entraining material from the environment on the shaded side of a circumstellar disk presumed to exist around the source star. Thus, the brightness asymmetries in irradiated jets may be a consequence of increased mass loading on the shadowed side of the disk (Bally & Reipurth 2000).

Nearly two dozen externally irradiated microjets have also been identified within the Orion Nebula on *HST* images (Bally, O'Dell, & McCaughrean 2000). Figure 4 shows an example of a one-sided irradiated microjet emerging from a proplyd in the Orion Nebula. The identification of these flows required the high angular resolution of *HST* since many are only about $0.1''$ wide and are therefore lost against the nebular background in most ground-based images. Most of the Orion irradiated jets are powered by stars embedded within proplyds. Some high velocity features detected in high resolution spectra of the Orion Nebula have turned out to be irradiated jets crossing the spectrograph entrance slit.

The physical properties of irradiated jets, such as their densities, temperatures, velocity fields, and spatial structure can be readily determined from standard recombination line theory. In non-irradiated HH objects, the determination of densities and other physical parameters is very difficult since the emission lines are produced in shocks which require a complete non-linear shock model to analyze. However, mass loss rates for irradiated jets can be directly estimated from their velocity and electron density derived from the Hα surface brightness, emission measure, and the jet beam width. This method is easier to apply than the more robust [S II] doublet ratio method since the Hα line is typically about an order of magnitude brighter than the [S II] line in a photo-ionized plasma. The

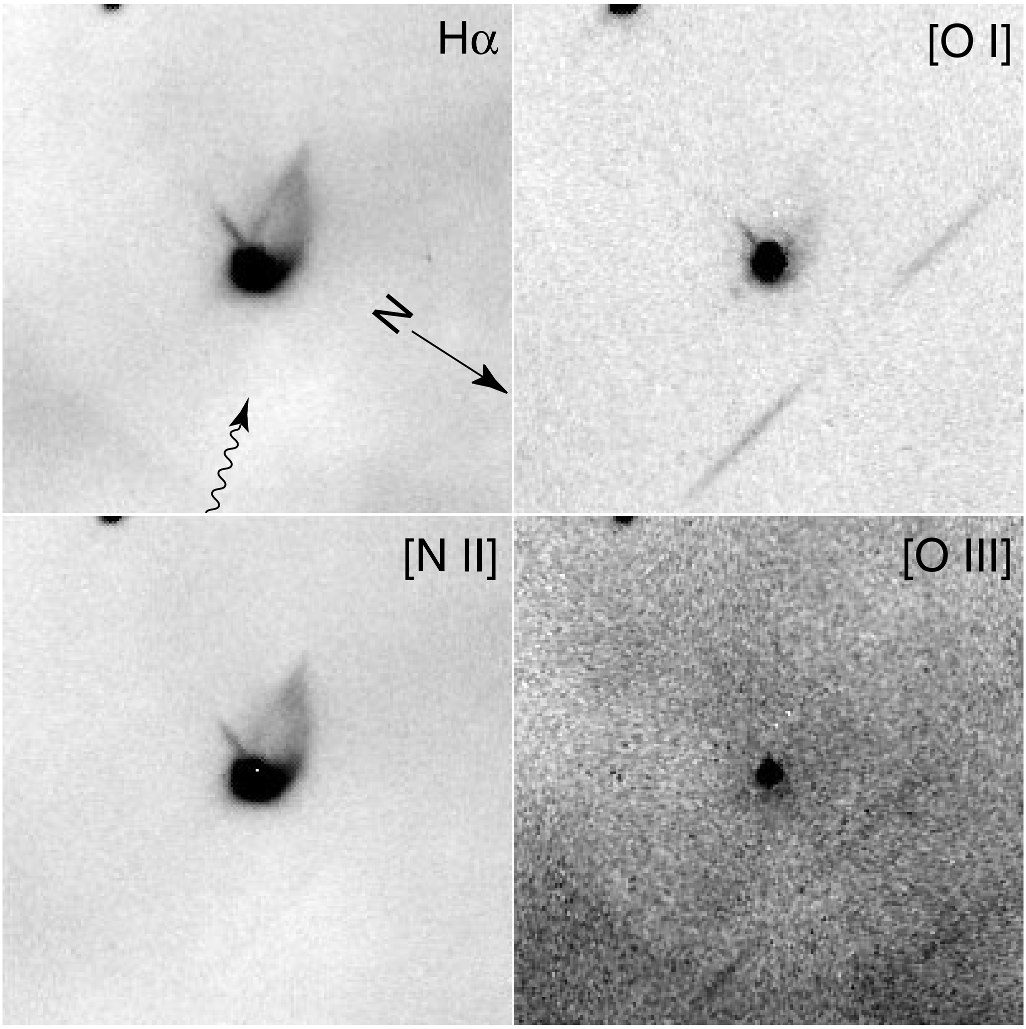


FIGURE 4. A mosaic of four narrow band *HST* images of the one-sided jet, HH 527, emerging from the proplyd 282−458 in the Orion Nebula. (see Bally, O'Dell, & McCaughrean 2000 for details).

typical mass loss rates of the Orion Nebula irradiated jets are about $\dot{M} \approx 10^{-9}$ $M_\odot$ yr^{-1}. The σ Ori jets (HH 444 through 447) have about an order of magnitude larger mass loss rates. Thus, these irradiated jets have mass loss rates one to two orders of magnitude lower than the spectacular Herbig-Haro jets such as HH 34, HH 46/47, or HH 111.

4.6. *Wind-wind collision fronts*

HST has also found evidence for wide angle winds blown by low mass stars in the Orion Nebula. Gull & Sofia (1979) found a parabolic arc of emission facing the bright core of the nebula surrounding the young star LL Ori. *HST* images (Bally, O'Dell, & McCaughrean 2000) show the LL Ori bow shock in unprecedented detail (Figure 5). The images show a chain of small high proper motion knots and small bows moving parallel to the surface of the large bow seen on the ground based images. Furthermore, 9 other stars in the *HST* mosaic of the nebula show smaller and fainter parabolic arcs facing the inner nebula. Six additional arcs have been found on new ground based images (Bally & Reipurth 2000).

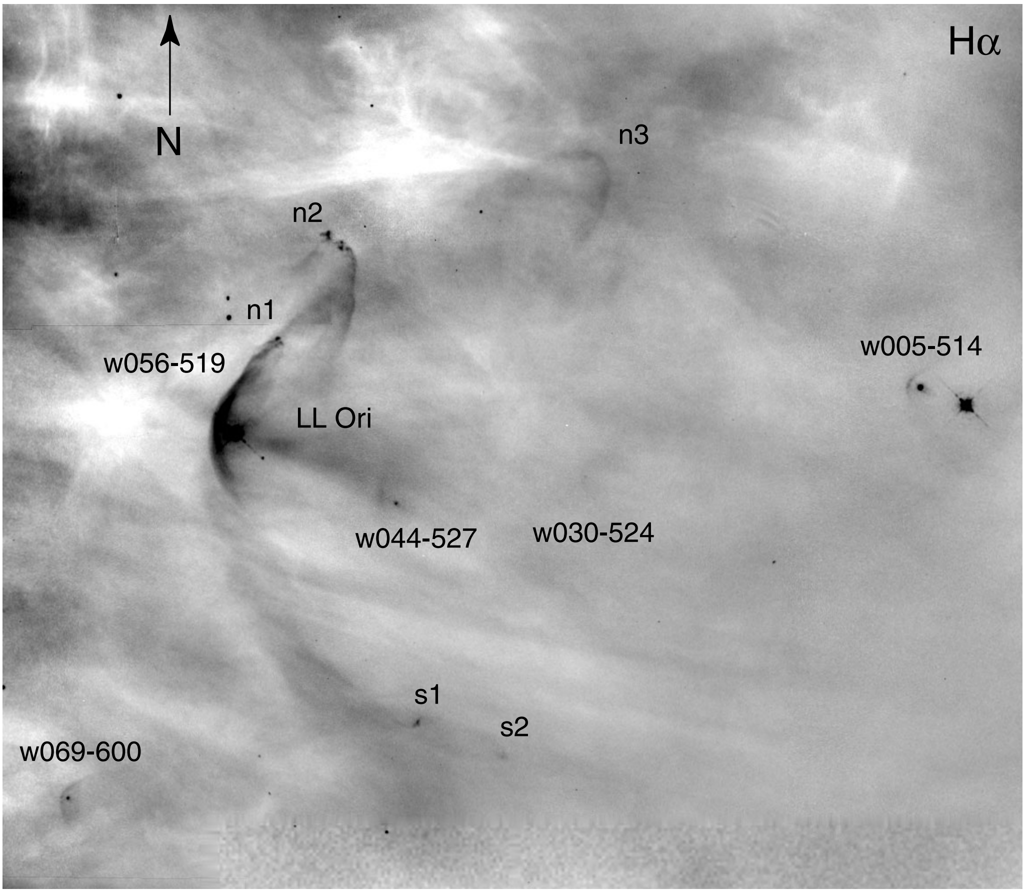


FIGURE 5. A narrow band *HST* image of the LL Ori bow shock in Hα. Several other similar but smaller bow shocks are labeled with a coordinate based number following a letter 'w' which stands for 'wind.' The knots n1, n2, n2, s1, and s2 appear to be condensations in the LL Ori bow shock. The proper motions of n1, and n2 have been measured with *HST* and indicate motions away from the vicinity of LL Ori (see Bally, O'Dell, & McCaughrean 2000 for details).

Thus, 16 LL Ori type shocks have now been recognized in the Orion Nebula. The LL Ori shocks appear to be produced by the collision of a wide-angle stellar wind having a mass loss rate of order $\dot{\mathrm{M}} \approx 10^{-9}$ to 10^{-8} $\mathrm{M}_\odot$ yr^{-1} and a velocity of about 500 km s^{-1} with a mildly supersonic ($\sim$ 20 km s^{-1}) outflow of plasma from the core of the nebula.

4.7. *Flows from massive stars*

The outflows discussed so far are powered by low mass protostars. But, even more powerful and spectacular flows are produced by nascent massive stars. Since massive stars are relatively rare, evolve onto the main sequence rapidly, and tend to form in highly obscured regions, only a few forming massive protostars have been observed with *HST*.

The BNKL complex of infrared sources located less than a parsec behind the Orion Nebula has a luminosity of 10^5 $\mathrm{L}_\odot$. The ultra-compact radio source I (Menten et al. 1995) is the source of a powerful wide angle outflow which produces both high velocity bipolar CO lobes and a spectacular set of near infrared fingers of H_2 emission (Allen & Burton 1994; McCaughrean & Mac Low 1997). The core of the outflow has been imaged with NICMOS and shows dozens of sub-arc second bow shocks in a spray of ejecta that resembles a 4th of July fireworks display (Stolovy et al. 1998). The tips of some of the northern fingers of H_2 emission are bursting into the low extinction neutral layer at the back of the Orion Nebula where they power a cluster of bright Herbig-Haro objects (HH 205 to 210). Proper motions (Jones & Walker 1985; Bally, O'Dell, & McCaughrean 2000) indicate a dynamical age of order 10^3 years for this 10^{48} to 10^{49} erg eruption. Unlike the highly collimated flows powered by many low mass stars, the BNKL outflow is very poorly collimated. The spray of ejecta, the associated CO outflow lobes, and the two oppositely directed cones of millimeter wavelength SiO maser emission lying within 1″ of source I (Greenhill et al. 1998) indicate an opening angle of order a radian or more. Additional masers trace what appears to be a high density disk of material expanding orthogonal to the outflow axis. The eruption responsible for the Orion BNKL outflow released nearly 0.1% of the kinetic energy produced by a supernova explosion. The ejecta appear to obey a Hubble law of expansion, indicating that most of this energy was released on a time scale short compared to the age of the flow.

There are fundamental differences in the physics of high and low mass star formation. Observations show that young high mass stars tend to be only found in clusters (Stahler et al. 2000). For massive stars, the contraction time to the main sequence is shorter than the time required to accumulate the star's mass for reasonable accretion rates. Thus, massive stars do not pass through an extended pre-main sequence phase. Furthermore, while radiation pressure can be neglected in low mass star formation, it plays a major role in the birth of massive stars (Wolfire & Cassinelli 1987). Massive protostars may grow more rapidly because they can accrete from their surroundings at a faster rate than lower mass objects. However, by the time such an object reaches a mass of more than about 20 $M_\odot$, radiation pressure can halt accretion and even blow away the infalling envelope. To grow further, such massive stars may accumulate additional mass by merging with lower mass protostars in the host cluster (Bonnell, Bate, & Zinnecker 1998). Although stellar collisions are extremely rare in the field, in star forming regions, stellar cross-sections are greatly increased by their circumstellar disks and protostellar envelopes. Therefore, close encounters in a high density proto-cluster may readily lead to further growth by stellar merging. Thus, O stars may be the analogs of cD galaxies found in the centers of rich galaxy clusters. Massive stars may also form by coalescence by means of 'protostellar cannibalism.' The merging of a 10 $M_\odot$ star with a 1 $M_\odot$ protostar and disk releases roughly 3×10^{48} erg of gravitational potential energy, comparable to what is needed to drive the BNKL outflow in Orion.

The Orion A molecular cloud has given birth to nearly a dozen massive stars. In addition to the four Trapezium stars that light up the Orion Nebula (two of which are eclipsing binaries), and the high mass protostars in the BNKL core, the region contains yet another region spawning moderate to high mass stars. This region is known as OMC1-S and is located about 90″ south of the BNKL core. A sub-mm continuum peak and a cluster of highly embedded infrared sources first drew attention to this region. Recent proper motion measurements of features in the Orion Nebula with *HST* (Bally, O'Dell, & McCaughrean 2000) identified no less than six major outflow systems bursting from the background molecular cloud associated with OMC1-S. This major region of moderate to high mass star formation has been all but ignored due to its proximity to the other more spectacular objects in Orion Nebula region.

HST has observed shocks associated with the outflows produced by a number of other relatively nearby regions of massive star formation. These include the bright HH objects HH 80/81 located where a highly collimated jet from a 10^4 $L_\odot$ protostar breaks out of its parent molecular cloud (Heathcote, Reipurth, & Raga 1988). HH 80/81 holds the speed record for Herbig-Haro flows with velocities extending up to 1,300 km s^{-1}. Indeed, Heathcote, Reipurth, and Raga (1998) found an extended filamentary nebula which they model as an energy conserving bubble powered by fast shocks associated with HH 80/81. *HST* has also observed HH 168, a bright and fragmented shock complex powered by a forming B star in the Cepheus A star forming region (Hartigan et al. 2000b). Perhaps the most surprising *HST* result is the detection of a bipolar Lyman α jet with STIS, HH 409, emerging from the B star AE Auriga (Grady et al. 1999).

4.8. *The impact of outflows*

Outflows and jets from young stars are ideal laboratories in which to study the hydrodynamics of supersonic flows in the presence of strong cooling and weak magnetic fields. These objects are close enough to the Sun that their time evolution can be directly measured using instruments such as *HST*. Furthermore, flows provide fossil records of the mass loss histories of their source stars. But in addition to being interesting objects in their own right, protostellar outflows play a fundamental role in determining the properties of the surrounding molecular cloud and in the self-regulation of star formation.

Outflows churn their host clouds. They create parsec-scale cavities surrounded by swept-up shells of accelerated molecular or atomic gas. Their shocks dissociated molecules and produce UV radiation which can alter the physical and chemical state of the medium. Their complex non-linear evolution is a major source of turbulent motions in the dense and cold phase of the interstellar medium surrounding star forming regions.

5. Conclusions

HST has made major contributions to our understanding of proto-planetary disks, provided the first hints of evolution towards proto-planets, and has shown us that there may be many astrophysical hazards to planet formation. Since most stars in the sky appear to form in OB associations, they are sooner or later irradiated by strong UV radiation fields. Many may lose their disks to photo-erosion. Even in dark cloud environments, most stars are born in compact groups and multiple systems. Dynamical interactions resulting from the rearrangement of unstable non-hierachical configurations to more stable hierarchical ones can eject member stars and destroy disks. Thus, we are beginning to obtain constraints on the formation rate of planets and on the types of exo-planetary system architectures that may exist. For example, in Orion Nebula-like environments, giant planets may not form at all unless they for very fast.

HST has revolutionized our understanding of the jets and outflows from young stars that excite Herbig-Haro objects. The structure and motions of these flows have been investigated with exquisite detail. We can see time evolution of the shocks and their proper motions can be measured in a time interval short compared to the post-shock cooling time. *HST* has revealed many new flows and new phenomena in star forming regions.

The *Next Generation Space Telescope* (*NGST*) will build on the rich legacy of *HST*. By operating in the 1 to 30 μm region, NGST will penetrate deep into clouds. It will image the environments of protostars and measure their properties anywhere in the Galaxy. Not only will it provide high angular resolution, but thousands of times lower backgrounds will permit unprecedented deep imaging of outflows and their shocks in spectral lines such as H_2, [Fe II], and Brackett α. *NGST* will detect protostellar outflows anywhere in the Galaxy. I expect that *NGST* will bring about a new revolution in our understanding of star formation.

REFERENCES

Allen, D. A. & Burton, M. G. 1993 *Nature*, **363**, 54.

Bally, J. & Devine, D. 1994 *ApJ*, **428**, L65.

Bally, J., Langer, W. D., & Liu, W. 1991 *ApJ*, **383**, 645.

Bally, J., Langer, W. D., Stark, A. A., & Wilson, R. W. 1987 *ApJ*, **313**, L45.

Bally, J., O'Dell, C. R., & McCaughrean, M. J. 2000 *AJ*, **119**, 2919.

Bally, J. & Reipurth, B. 2001 *ApJ*, **546**, 299.

Bally, J., Reipurth, B., Lada, C. J., & Billawala, Y. 1999 *AJ*, **117**, 410.

Bally, J., Sutherland, R. S., Devine, D., & Johnstone, D. 1998 *AJ*, **116**, 293.

Bally, J., Testi, L., Sargent, A., & Carlstrom, J. 1998 *AJ*, **116**, 854.

Blaauw, A. 1991 in *The Physics of Star Formation and Early Stellar Evolution* (eds. C. J. Lada & N. D. Kylafis), p. 125.

Bonnell, I. A., Bate, M. R., & Zinnecker, H. 1998 *MNRAS*, **295**, 93.

Brandner, W., et al. 2000 *AJ*, **119**, 292.

Brodie, J. P., Schroder, L. L., Huchra, J. P., Phillips, A. C., Kissler-Patig, M., & Forbes, D. A. 1998 *AJ*, **116**, 691.

Brown, A. G. A., de Geus, E. J., & de Zeeuw, P. T. 1994 *A&A*, **289**, 101.

Brown, A. G. A., Hartmann, D., & Burton, W. B. 1995 *A&A*, **300**, 903.

Burrows, C. J., et al. 1996 *ApJ*, **473**, 437.

Chen, H., Bally, J., O'Dell, C. R., McCaughrean, M. J., Thompson, R. I., Rieke, M., Schneider, G., & Young, E. T. 1998 *ApJ*, **492**, L173.

Chernin, L., Masson, C., Gouveia dal Pino, E. M., & Benz, W. 1994 *ApJ*, **426**, 204.

Chernin, L. M. & Masson, C. R. 1995 *ApJ*, **455**, 182.

Davis, C. J., Eislöffel, J., & Ray, T. P. 1994 *ApJ*, **426**, L93

Devine, D., Reipurth, B., Bally, J., & Heathcote, S. 1997 *AJ*, **114**, 2095.

Devine, D., Reipurth, B., & Bally, J. 1999 *AJ*, **118**, 972.

Devine, D., Bally, J., Reipurth, B., Stocke, J., & Morse, J. 2000 *AJ*, **117**, 2931.

Eislöffel, J., Mundt, R., & Böhm, K. H. 1994 *AJ*, **108**, 104.

Eislöffel, J., Mundt, R., Ray, T. P., Rodríguez, L. F. 2000 in *Protostars and Planets IV* (eds. V. Mannings, A. P. Boss, & S. S. Russell), p. 815. University of Arizona Press.

Figer, D. F., Kim, S. S., Morris, M., Serabyn, E., Rich, R. M., & McLean, I. S. 1999 *ApJ*, **525**, 750.

Fridlund, C. V. M. & Liseau, R. 1998 *ApJ*, **499**, L75.

Gilbert, A. M., et al. 2000 *ApJ*, **533**, L57.

Grady, C. A., Woodgate, B., Bruhweiler, F. C., Boggess, A., Plait, P., Lindler, D. J., Clampin, M., & Kalas, P. 1999 *ApJ*, **523**, L151.

Greenhill, L. J., Gwinn, C. R., Schwartz, C., Moran, J. M., & Diamond, P. J. 1998 *Nature*, **396**, 650.

Gull, T. R. & Sofia, S. 1979 *ApJ*, **230**, 782.

Haro, G. 1952 *ApJ*, **115**, 572.

Hartigan, P., Bally, J., Reipurth, B., & Morse, J. A. 2000 in *Protostars and Planets IV* (eds. V. Mannings, A. P. Boss, & S. S. Russell), p. 841. University of Arizona Press.

Hartigan, P., Morse, J., Palunas, P., Bally, J., & Devine, D. 2000 *AJ*, **119**, 1872.

Heathcote, S., Morse, J. A., Hartigan, P., Reipurth, B., Schwartz, R. D., Bally, J., & Stone, J. M. 1996 *AJ*, **112**, 1141.

Heathcote, S., Reipurth, B., & Raga, A. C. 1998 *AJ*, **116**, 1940.

Henney, W. J. & Arthur, S. J. 1998 *AJ*, **116**, 322.

Henney, W. J. & O'Dell, C. R. 1999 *AJ*, **118**, 2350.

Herbig, G. 1951 *ApJ*, **113**, 697.

Herbig, G. & Jones, B. F. 1981 *AJ*, **86**, 1232.

Hester, J. J., et al. 1996 *AJ*, **111**, 2349.

Hester, J. J., Stapelfeldt, K. R., & Scowen, P. A. 1998 *AJ*, **116**, 372.

Hillenbrand, L. A. 1997 *AJ*, **113**, 1733.

Hillenbrand, L. A. & Hartmann, L. W. 1998 *ApJ*, **492**, 540.

Johnstone, D., Hollenbach, D., & Bally, J. 1998 *ApJ*, **499**, 758.

Jones, B. F. & Walker, M. F. 1985 *AJ*, **90**, 1320.

Kalas, P., Larwood, J., Smith, B. A., & Schultz, A. 2000 *ApJ*, **530**, L133.

Königl, A. 1991 *ApJ*, **370**, L39

Königl, A. & Ruden, S. P. 1993 in *Protostars and Planets III* (eds. E. H. Levy & J. I. Lunine), p. 641. University of Arizona Press.

Königl, A. & Pudritz, R. E. 2000 in *Protostars and Planets IV* (eds. V. Mannings, A. P. Boss, & S. S. Russell), p. 759. University of Arizona Press.

Krist, J. E., et al. 1999 *ApJ*, **515**, L35.

Lada, C. J. 1985 *ARAA*, **23**, 267.

McCaughrean, M. J. & Mac Low, M.-M. 1997 *AJ*, **113**, 391.

McCaughrean, M. J. & O'Dell, C. R. 1996 *AJ*, **111**, 1977.

McCaughrean, M. J., Chen, H., Bally, J., Erickson, E., Thompson, R., Rieke, M., Schneider, G., Stolovy, S., & Young, E. 1998 *ApJ*, **429**, L157.

Menten, K. M. & Reid, M. J. 1995 *ApJ*, **445**, L157.

Moro-Martín, A., Cernicharo, J., Noriega-Crespo, A., & Martín-Pintado, J. 1999 *ApJ*, **520**, L111.

O'Dell, C. R., Wen, Z., & Hu, X. 1993b *ApJ*, **410**, 696.

O'Dell, C. R. & Wong, S. K. 1996 *AJ*, **111**, 846.

Ogura, K. 1995 *ApJ*, **450**, L23.

Padgett, D. L., Brandner, W., Stapelfeldt, K. R., Strom, S. E., Terebey, S., & Koerner, D. 1999 *AJ*, **117**, 1490.

Reipurth, B., Hartigan, P., Heathcote, S., Morse, J. A., & Bally, J. 1997 *AJ*, **114**, 757.

Reipurth, B., Bally, J., & Devine, D. 1997 *ApJ*, **114**, 2708.

Reipurth, B., Bally, D., Fesen, R., & Devine, D. 1998 *Nature*, **396**, 343.

Reipurth, B., Devine, D., & Bally, J. 1998 *AJ*, **116**, 1396.

Reipurth, B., Yu, K. C., Rodíguez, L. F., Heathcote, S., & Bally, J. 1999 *AA*, **352**, L86.

Reipurth, B. 1999 *A General Catalog of Herbig-Haro Objects*, 2. edition, http://casa.colorado.edu/hhcat/.

Reipurth, B., Yu, K. C., Heathcote, S., Bally, J., & Rodríguez, L. F. 2000 *AJ*, **120**, 1449.

Rodríguez, L. F., Ho, P. T. P., Torrelles, J. M., Curiel, S., & Cantó, J. 1990 *ApJ*, **352**, 645.

Rubio, M., Barbá, R. H., Walborn, N. R., Probst, R. G., García, J., & Roth, M. R. 1998 *AJ*, **116**, 1708.

Schneider, G., et al. 1999 *ApJ*, **513**, L127.

Shu, F. H., Najita, J., Ostriker, E., Wilkin, F., Ruden, S., & Lizano, S. 1994a *ApJ*, **429**, 781.

Shu, F. H., Najita, J., Ruden, S., & Lizano, S. 1994b *ApJ*, **429**, 797.

Shu, F. H., Najita, J., Ostriker, E. C., & Shang, H. 1995 *ApJ*, **455**, L155.

Shu, F. H., Najita, J. R., Shang, H., & Li, Z.-H. 2000 in *Protostars and Planets IV* (eds. V. Mannings, A. P. Boss, & S. S. Russell), p. 789. University of Arizona Press.

Stahler, S. W., Palla, F., & Ho, P. T. P. 2000 in *Protostars and Planets IV* (eds. V. Mannings, A. P. Boss, & S. S. Russell), p. 327. University of Arizona Press.

Stapelfeldt, K. R., et al. 1995 *ApJ*, **449**, 888.

Stapelfeldt, K. R., Krist, J. E., Menard, F., Bouvier, J., Padgett, D. L., & Burrows, C. J. 1998a *ApJ*, **502**, L65.

Stapelfeldt, K. R., et al. 1998b *ApJ*, **508**, 736.

Stapelfeldt, K. R., et al. 1999, *ApJ*, **516**, L95.

Stolovy, S. R., et al. 1998 *ApJ*, **492**, L151.

Störzer, H. & Hollenbach, D. 1998 *ApJ*, **502**, L71.

Störzer, H. & Hollenbach, D. 1999 *ApJ*, **515**, 669.

Testi, L. & Sargent, A. 1998 *ApJ*, **508**, L91.

Throop, H. 2000 *Ph.D. Thesis*, University of Colorado, Boulder.

Throop, H., Bally, J., & Esposito, L. 2000, in preparation.

Valtonen, M. & Mikkola, S. 1991 *ARAA*, **29**, 9.

Walborn, N. R., Barbá, R. H., Brandner, W., Rubio, M., Grebel, E. K., & Probst, R. G. 1999a *AJ*, **117**, 225.

Walborn, N. R., Drissen, L., Parker, J. W., Saha, A., MacKenty, J. W., & White, R. L. 1999b *AJ*, **118**, 1684.

Weinberger, A. J., Becklin, E. E., Schneider, G., Smith, B. A., Lowrance, P. J., Silverstone, M. D., Zuckerman, B., & Terrile, R. J. 1999 *ApJ*, **525**, L53.

Wolfire, M. G. & Cassinelli, J. P. 1987 *ApJ*, **319**, 850.

SN1987A: The birth of a supernova remnant

By RICHARD McCRAY

JILA, University of Colorado and National Institute of Science and Technology,
Boulder, CO 80309-0440

Supernova 1987A has been a prime target for the Hubble Space Telescope since its launch, and it will remain so throughout the lifetime of *HST*. Here I review the observations of SN1987A, paying particular attention to the rapidly developing impact of the blast wave with the circumstellar matter as observed by *HST* and the *Chandra Observatory*.

1. Introduction

If there was ever a match made in heaven, it is the combination of SN1987A and the *Hubble Space Telescope*. Although the *HST* was not available to witness the first three years after outburst, it has been the primary instrument to observe SN1987A since then.

SN1987A in the Large Magellanic Cloud is the brightest supernova to be observed since SN1604 (Kepler), the first to be observed in every band of the electromagnetic spectrum, and the first to be detected through its initial burst of neutrinos. Although the bolometric luminosity of SN1987A today is $\approx 10^{-6}$ of its value at maximum light ($L_{\max} \approx 2.5 \times 10^8$ L$_\odot$), it will remain bright enough to be observed for many decades in the radio, infrared, optical, UV, and X-ray bands.

SN1987A is classified as a Type II supernova (SNeII) by virtue of the strong hydrogen lines in its spectrum. It was atypical of SNeII in that its light curve did not reach maximum until three months after outburst and its maximum luminosity was about 1/10 the mean maximum luminosity of SNeII. These differences can be attributed to the fact that the star that exploded was a blue giant, unlike the progenitors of most SNeII, which we believe to be red giants.

The burst of neutrinos observed from SN1987A proved beyond doubt that its explosion followed the collapse of the core of the star, but subsequent observations have shown no evidence of the neutron star or black hole that we expect to find at the center of the debris.

The expanding gaseous debris of SN1987A cooled rapidly after the explosion (McCray 1993). By 4 months, the debris had become transparent at optical and infrared wavelengths and its spectrum was dominated by emission lines. By 3 years, its temperature was less than 2,000 K throughout and the heavy elements had formed molecules and dust. With a present temperature < 100 K, the inner debris is perhaps the coldest optically emitting source known to astronomers. It is glowing because the atoms (primarily hydrogen) are excited by nonthermal electrons and positrons produced by the decay of radioactive elements, primarily ^{44}Ti. *HST* images of SN1987A (Figure 1) show that this inner debris is slightly elongated in the NS direction, and that it is expanding with transverse velocity $\sim 2,800$ km s^{-1}. The irregular shape of the optical image is most likely a consequence of the irregular distribution of dust within the inner debris.

Perhaps the most outstanding mystery of SN1987A is the absence of any evidence (except the neutrino flash) of a compact object at its center. The central object must have a luminosity $\lesssim 300$ L$_\odot$; otherwise we would have detected it by now.

The next most exciting mystery of SN1987A is the remarkable system of circumstellar rings shown in Figure 1. Evidently, they were ejected by the supernova progenitor some 20,000 years before it exploded. But how do we account for their morphology? At the

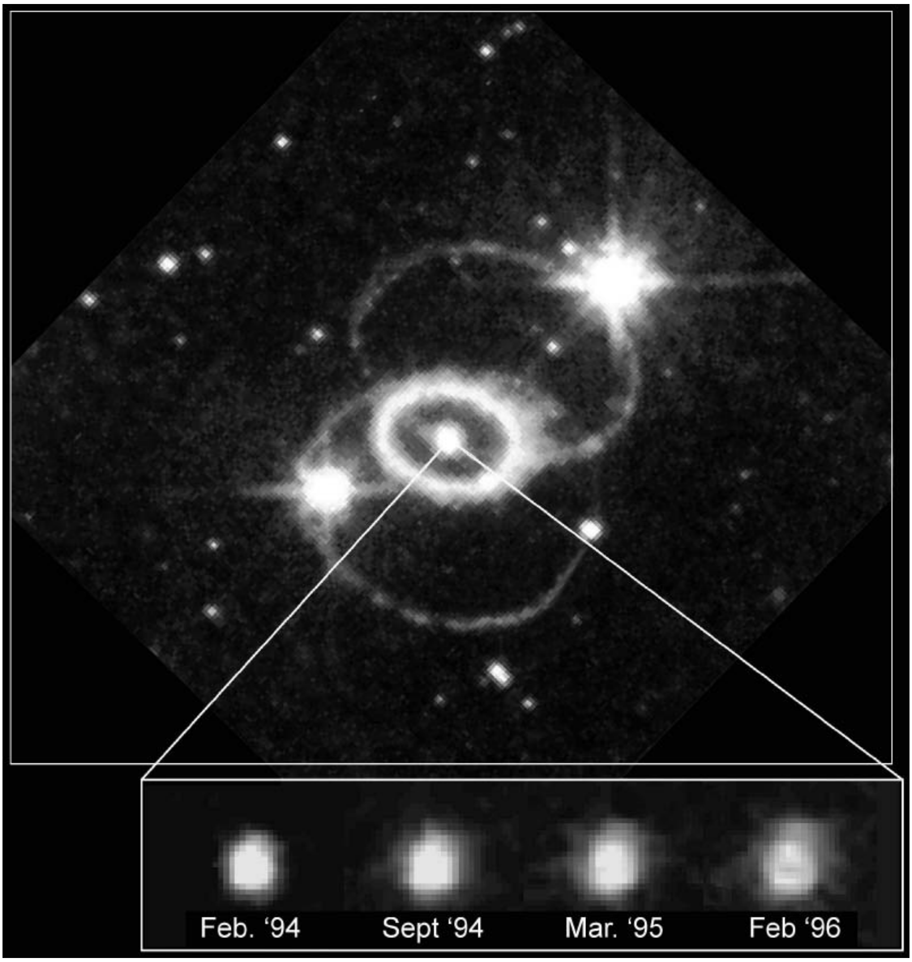


FIGURE 1. *HST* image of SN1987A and its circumstellar rings. The inset at the bottom shows the evolution of the glowing center of the supernova debris (Pun et al. 1997).

moment, there is no satisfactory explanation. But, as I shall describe in this chapter, events beginning now will give us a new window on the supernova's past.

The supernova blast wave is now beginning to strike the inner ring. This impact marks the birth of a supernova remnant, defined as the stage when the supernova light is dominated by the impact of the supernova debris with circumstellar matter. It will be a spectacular event. During the coming decade, the remnant SNR1987A will brighten by orders of magnitude at wavelengths ranging from radio to X-ray. *HST* will continue to be our most powerful tool to observe this unique event. But it will not be the only one. With the *Chandra Observatory*, we have already obtained our first images and spectra of the X-ray source. Large ground-based telescopes equipped with adaptive optics have already begun to provide excellent images and spectra at optical and near-infrared wavelengths. Future observatories, such as the *Space Infrared Telescope Facility* and the *Atacama Large Millimeter Array*, will provide data at other wavelength bands to complement the *HST* observations. The combination of these data will give us a unique opportunity to probe the rich range of physical phenomena associated with astrophysical shocks and to learn about the death throes of a supernova progenitor.

Source	Collapse	Radioactivity	Expansion
Definition	$\sim \dfrac{G\ \mathrm{M}_\odot^2}{R_{N*}}$	^{56}Ni $\rightarrow^{56}$ Co $\rightarrow^{56}$ Fe (0.07 $\mathrm{M}_\odot$)	$\displaystyle\int_{\text{debris}} \frac{1}{2}V^2 dM$
Emerges as:	Neutrinos ($kT \sim 4$ MeV)	O, IR (+ X, γ)	X-rays (+ R, IR, O, UV)
Energy [ergs]	10^{53}	10^{49}	10^{51}
Timescale	$\sim$ 10 seconds	$\sim$ 1 year	$\sim$ 100–1000 years

Table 1. SN1987A Energetics

2. Energetics

Before describing the *HST* observations of SN1987A, it might be useful to review its energy sources. These are summarized in Table I.

As Table I shows, SN1987A has three different sources of energy, each of which emerges as a different kind of radiation and with a different timescale. The greatest is the collapse energy itself, which emerges as a neutrino burst lasting a few seconds. The energy provided by radioactive decay of newly synthesized elements is primarily responsible for the optical display. Most of this energy emerged within the first year after outburst, primarily in optical and infrared emission lines and continuum from relatively cool ($T \lesssim 5,000$ K) gas. Note that the radioactive energy is relatively small, $\sim 10^{-4}$ of the collapse energy.

The kinetic energy of the expanding debris can be inferred from observations of the spectrum during the first three months after explosion. Astronomers infer the density and velocity of gas crossing the photosphere from the strengths and widths of hydrogen lines in the photospheric spectrum. By tracking the development of the spectrum as the photosphere moved to the center of the debris, astronomers can measure the integral defining the kinetic energy. Doing so, they find that $\sim 10^{-2}$ of the collapse energy has been converted into kinetic energy of the expanding debris. Why this fraction is typically 10^{-2} and not, say, 10^{-1} or 10^{-3}, is one of the unsolved problems of supernova theory.

This kinetic energy will be converted into radiation when the supernova debris strikes circumstellar matter. When this happens, two shocks always develop: the blast wave, which overtakes the circumstellar matter; and the reverse shock, which is driven inwards (in a Lagrangean sense) through the expanding debris. The gas trapped between these two shocks is typically raised to temperatures in the range 10^6–10^8 K and will radiate most of its thermal energy as X-rays with a spectrum dominated by emission lines in the range 0.3–10 keV.

Most of the kinetic energy of the debris will not be converted into thermal energy of shocked gas until the blast wave has overtaken a circumstellar mass comparable to that of the debris itself, $\sim$ 10–20 $\mathrm{M}_\odot$. Typically, that takes many centuries, and as a result, most galactic supernova remnants (e.g. Cas A) reach their peak X-ray luminosities after a few centuries and fade thereafter.

As I discuss below, we believe that SN1987A is surrounded by a few $\mathrm{M}_\odot$ of circumstellar matter within a distance of parsec or two. Thus, a significant fraction of the kinetic energy of the debris will be converted into thermal energy within a few decades as the supernova blast wave overtakes this matter.

3. The circumstellar rings

The first evidence for circumstellar matter around SN1987A appeared a few months after outburst in the form of narrow optical and ultraviolet emission lines seen with the *International Ultraviolet Explorer* (Fransson et al. 1989). Even before astronomers could image this matter, they could infer that:

- the gas was nearly stationary (from the linewidths);
- it was probably ejected by the supernova progenitor (because the abundance of nitrogen was elevated);
- it was ionized by soft X-rays from the supernova flash (from emission lines of N V $\lambda\lambda 1239, 1243$ and other highly ionized elements in the spectrum);
- it was located at a distance of about a light year from the supernova (from the rise time of the light curve of these lines); and
- the gas had atomic density $\sim 3 \times 10^3 - 3 \times 10^4$ cm^{-3} (from the fading timescale of the narrow lines).

The triple ring system was first seen in images obtained by the ESO *NTT* telescope (Wampler et al. 1990), but the evidence of the outer loops was not compelling until astronomers obtained an image with the *HST* WFPC-2 (Burrows et al. 1995). By measuring the Doppler shifts of the emission lines, Crotts & Heathcote (1991) found that the inner ring is expanding with a radial velocity ≈ 10 km s^{-1}. Dividing the radius of the inner ring (0.67 lt-year) by this velocity gives a kinematic timescale $\approx 20,000$ years since the gas in the ring was ejected, assuming constant velocity expansion. The more distant outer loops are expanding more rapidly, consistent with the notion that they were ejected at the same time as the inner ring.

The rings observed by *HST* may be only the tip of the iceberg. They are glowing by virtue of the ionization and heating caused by the flash of EUV and soft X-rays emitted by the supernova during the first few hours after outburst. But calculations (Ensman & Burrows 1992) show that this flash was a feeble one. The glowing gas that we see in the triple ring system is probably only the ionized inner skin of a much greater mass of unseen gas that the supernova flash failed to ionize. For example, the inner ring has a glowing mass of only about ~ 0.04 M$_\odot$, just about what one would expect such a flash to produce.

In fact, ground-based observations of optical light echoes during the first few years after outburst provided clear evidence of a much greater mass of circumstellar gas within several light years of the supernova that did not become ionized (Wampler et al. 1990; Crotts, Kunkel, & Heathcote 1995). The echoes were caused by scattering of the optical light from the supernova by dust grains in this gas. They became invisible about five years after outburst.

What accounts for this circumstellar matter and the morphology of the rings? My hunch is that the supernova progenitor was originally a close binary system, and that the two stars merged some 20,000 years ago. The inner ring might be the inner rim of a circumstellar disk that was expelled during the merger, perhaps as a stream of gas that spiraled out from the outer Lagrangean (L2) point of the binary system. Then, during the subsequent 20,000 years before the supernova event, ionizing photons and stellar wind from the merged blue giant star eroded a huge hole in the disk. Finally, the supernova flash ionized the inner rim of the disk, creating the inner ring that we see today.

The binary hypothesis provides a natural explanation of the bipolar symmetry of the system, and may also explain why the progenitor of SN1987A was a blue giant rather than a red giant (Podsiadlowski 1992). But we still lack a satisfactory explanation for

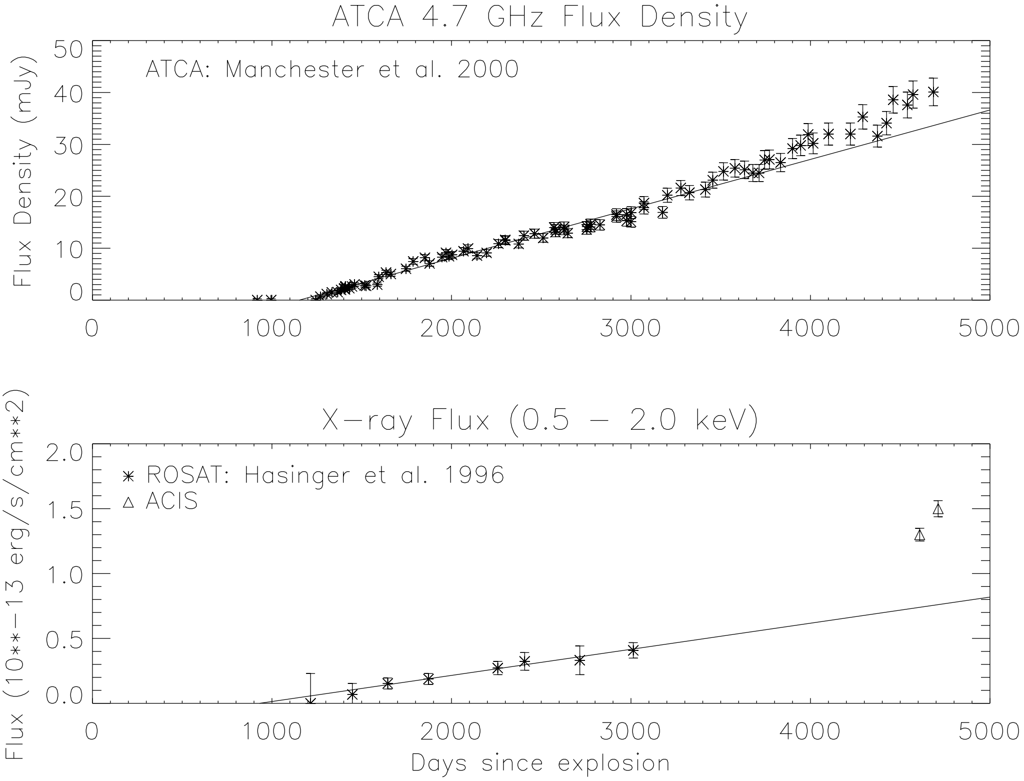


FIGURE 2. Radio (upper) and X-ray (lower) light curves.

the outer loops. If we could only see the invisible circumstellar matter that lies beyond the loops, we might have a chance of reconstructing the mass ejection episode.

Fortunately, SN1987A will give us another chance. When the supernova blast wave hits the inner ring, the ensuing radiation will cast a new light on the circumstellar matter. As I describe below, this event is now underway.

4. The crash begins

The first evidence that the supernova debris was beginning to interact with circumstellar matter came from radio and X-ray observations. As Figure 2 shows, SN1987A became a detectable source of radio and soft X-ray emission about 1200 days after the explosion and has been brightening steadily in both bands ever since. Shortly afterwards, astronomers imaged the radio source with the *Australia Telescope Compact Array* (*ATCA*) and found that the radio source was an elliptical annulus inside the inner circumstellar ring observed by *HST* (Figure 6). From subsequent observations, they found that the annulus was expanding with a velocity $\sim 3,500$ km s^{-1}.

Chevalier (1992) recognized that the radio emission most likely arose from relativistic electrons accelerated by shocks formed inside the inner ring where the supernova debris struck relatively low density ($n \sim 100$ cm^{-3}) circumstellar matter, and that the X-ray emission probably came from the shocked circumstellar matter and supernova debris. Subsequently, Chevalier & Dwarkadas (1995) suggested a model for the circumstellar matter, in which the inner circumstellar ring is the waist of an hourglass-shaped bipolar nebula. The low density circumstellar matter is a thick layer of photoionized gas that lines the interior of the bipolar nebula. The inner boundary of this layer is determined by balance of the pressure of the hot bubble of shocked stellar wind gas and that of the photoionized layer. In the equatorial plane, the inner boundary of this layer is located at about half the radius of the inner ring. According to this model, the appearance of X-ray and radio emission at $\sim$ 1200 days marks the time when the blast wave first enters the photoionized layer.

5. The reverse shock

Following Chevalier & Dwarkadas (1995), Borkowski, Blondin, & McCray (1997a) developed a more detailed model to account for the X-ray emission observed from SN1987A. They used a 2-D hydro code to simulate the impact of the outer atmosphere of the supernova with an idealized model for the photoionized layer. They found a good fit to the *ROSAT* observations with a model in which the thickness of the photoionized layer was about half the radius of the inner ring and the layer had atomic density $n_0 \approx 150$ cm^{-3}.

With the same model, we found to our delight that Lyα and Hα emitted by hydrogen atoms crossing the reverse shock should be detectable with the STIS. Then, in May 1997, only three months after our predictions appeared in the *ApJ* (Borkowski et al. 1997a), the first STIS observations of SN1987A were made, and broad ($\Delta V \approx \pm 12{,}000$ km s^{-1}) Lyα emission lines were detected (Sonneborn et al. 1998). Within the observational uncertainties, the flux was exactly as predicted.

One might at first be surprised that such a theoretical prediction of the Lyα flux would be on the mark, given that it was derived from a hydrodynamical model based on very uncertain assumptions about the density distribution of circumstellar gas. But, on further reflection it is not so surprising because the key parameter of the hydrodynamical model, the density of the circumstellar gas, was adjusted to fit the observed X-ray flux. Since the intensity of Lyα is derived from the same hydrodynamical model, the ratio of Lyα to the X-ray flux is determined by the ratio of cross sections for atomic processes, independent of the details of the hydrodynamics.

The broad Lyα and Hα emission lines are not produced by recombination. (The emission measure of the shocked gas is far too low to produce detectable Lyα and Hα by recombination.) Instead, the lines are produced by neutral hydrogen atoms in the supernova debris as they cross the reverse shock and are excited by collisions with electrons and protons in the shocked gas. Since the cross sections for excitation of the $n \geq 2$ levels of hydrogen are nearly equal to the cross sections for impact ionization, about one Lyα photon is produced for each hydrogen atom that crosses the shock. Thus, the observed flux of broad Lyα is a direct measure of the flux of hydrogen atoms that cross the shock. Moreover, since the outer supernova envelope is expected to be nearly neutral, the observed flux is a measure of the mass flux across the shock.

The fact that the Lyα and Hα lines are produced by excitation at the reverse shock gives us a powerful tool to map this shock. Since any hydrogen in the supernova debris is freely expanding, its line-of-sight velocity, $V_{\|} = z/t$, where z is its depth measured from the mid-plane of the debris and t is the time since the supernova explosion. Therefore, the Doppler shift of the Lyα line will be directly proportional to the depth of the reverse shock: $\Delta\lambda/\lambda_0 = z/ct$. Thus, by mapping the Lyα or Hα emission with STIS, we can generate a 3-dimensional image of the reverse shock.

Figure 3 illustrates this procedure. Panel **a** shows the location of the slit superposed on an image of the inner circumstellar ring, with the near (**N**) side of the tilted ring on the lower left. Panel **b** shows the actual STIS spectrum of Lyα from this observation. The slit is black due to geocoronal Lyα emission. The bright blue-shifted streak of Lyα

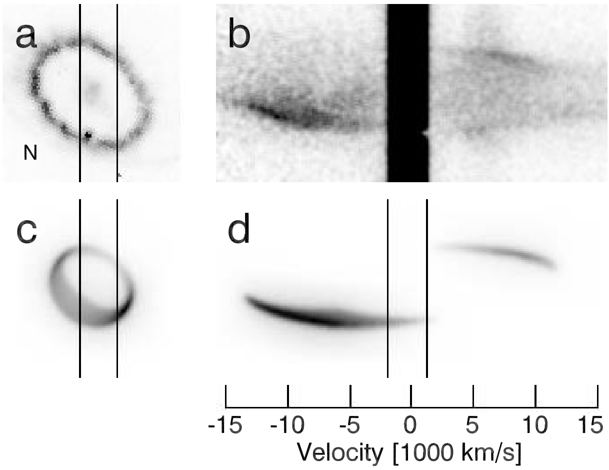


FIGURE 3. STIS spectrum of Lyα emission from the reverse shock (Michael et al. 1998).

extending to the left of the lower end of the slit comes from hydrogen atoms crossing the near side of the reverse shock, while the fainter red-shifted streak at the upper end of the slit comes from the far side of the reverse shock.

From this and similar observations with other slit locations we have constructed a map of the reverse shock surface, shown in panel **c**. Note that the emitting surface is an annulus that lies inside the inner circumstellar ring. Presumably, the reverse shock in the polar directions lies at a greater distance from the supernova, where the flux of atoms in the supernova debris is too low to produce detectable emission. Panel **d** is a model of the STIS Lyα spectrum that would be expected from hydrogen atoms crossing the shock surface illustrated in panel **c**. By comparing such model spectra with the actual spectra (e.g. panel **b**), we may refine our model of the shock surface.

Note that the broad Lyα emission is much brighter on the near (blue-shifted) side of the debris than on the far side, and so is the reconstructed shock surface. There is one obvious reason why this should be so: the blue-shifted side of the reverse shock is nearer to us by several light-months, and so we see the emission from the near side as it was several months later than that from the far side. Since the flux of atoms across the reverse shock is increasing, the near side should be brighter. But this explanation fails quantitatively. The observed asymmetry is several times greater than can be explained by light-travel time delays, and must be attributed to real asymmetry in the supernova debris. As we shall see, observations at radio and X-ray wavelengths also provide compelling evidence for asymmetry of the supernova debris.

6. The hot spots

One can estimate the time that the blast wave should strike the inner circumstellar ring from Chevalier's (1982) self-similar solutions for the hydrodynamics of a freely expanding stellar atmosphere striking a circumstellar medium. Stellar atmosphere models give a good fit to the spectrum of SN1987A during the early photospheric phase with a stellar atmosphere having a power-law density law $\rho(r,t) = At^{-3}(r/t)^{-9}$ (Eastman & Kirshner 1989; Schmütz et al. 1990). If this atmosphere strikes a circumstellar medium having a uniform density n_0, the blast wave will propagate according to the law $R_B(t) \propto A^{1/9} n_0^{-1/9} t^{2/3}$. For such a model, the time of impact can be estimated from the

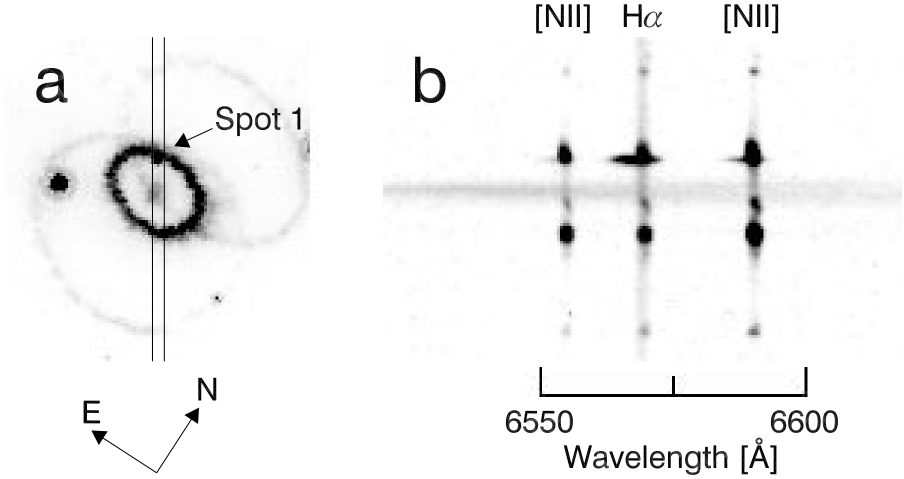


FIGURE 4. Spectrum of Spot 1. (a) Slit orientation on image of SN1987A's triple ring system; (b) Section of STIS G750M spectrum.

equation $t \propto A^{-1/6} n_0^{1/6}$. The coefficient, A, of the supernova atmosphere density profile can determined from the fit to the photospheric spectrum, leaving the density, n_0, of the gas between the supernova and the circumstellar ring as the main source of uncertainty. With various assumptions about the density distribution of this gas, predictions of the time of first contact ranged from 2003 (Luo & McCray 1991) to 1999 ± 3 (Luo, McCray, & Slavin 1994) to 2005 ± 3 (Chevalier & Dwarkadas 1995).

In April 1997, Sonneborn et al. (1998) obtained the first STIS spectrum of SN1987A with the 2×2 arcsecond aperture. Images of the circumstellar ring were seen in several optical emission lines. No Doppler velocity spreading was evident in the ring images except at one point, located at $P.A. = 29°$ (E of N), which we now call "Spot 1," where a Doppler-broadened streak was seen in Hα and other optical lines.

Figure 4 (Michael et al. 2000) shows a portion of a more recent (March 1998) STIS spectrum of Spot 1, where one can see vertical pairs of bright spots corresponding to emission from the stationary ring at Hα and [NII]$\lambda\lambda 6548, 6584$ (one also sees three more fainter spots at each wavelength where the outer loops cross the slit, and a broad horizontal streak at the center due to the Hα emission from the rapidly expanding inner debris). The emission lines are broadened (with $FWHM \approx 250$ km s^{-1}) and blue-shifted (with $\Delta V \approx -80$ km s^{-1}) at the location of Spot 1, which is located slightly inside the stationary ring.

Spot 1 evidently marks the location where the supernova blast wave first touches the dense circumstellar ring. When a blast wave propagating with velocity $V_b \approx 4,000$ km s^{-1} through circumstellar matter with density $n_0 \approx 150$ cm^{-3} encounters the ring, having density $n_r \approx 10^4$ cm^{-3}, one would expect the transmitted shock to propagate into the ring with $V_r \approx (n_0/n_r)^{1/2} V_b \approx 500$ km s^{-1} if it enters at normal incidence, and more slowly if it enters at oblique incidence. Since Spot 1 is evidently a protrusion, a range of incidence angles, and hence of transmitted shock velocities and directions, can be expected. Obviously, the line profiles will be sensitive to the geometry of the protrusion. Since the protrusion is on the near side of the ring and is being crushed by the entering shocks, most of the emission will be blue-shifted, as is observed. But part of the emission is red-shifted because it comes from oblique shocks entering the far side of the protrusion.

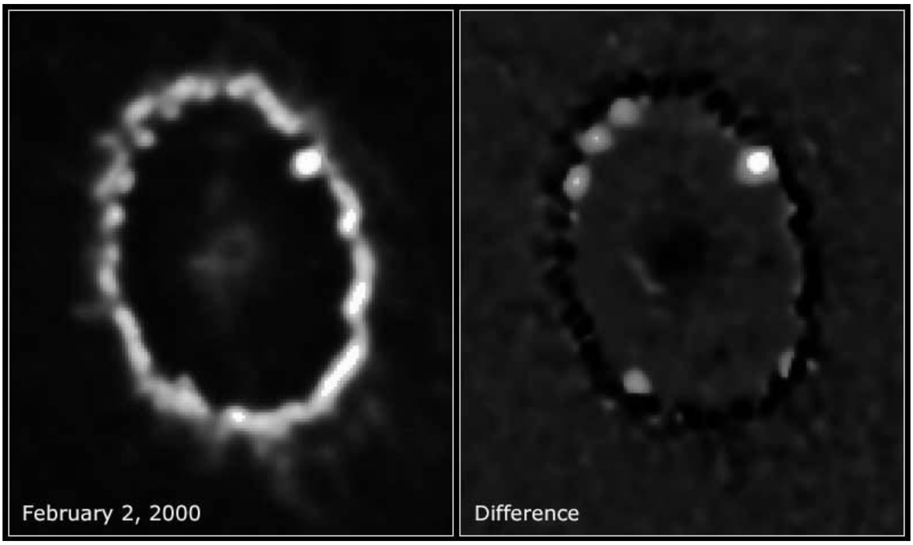


FIGURE 5. Hot spots on the inner ring. The right panel shows the difference between WFPC observations taken on 2 February 2000 and 21 February 1996, showing that the hot spots have brightened while the rest of the ring has faded.

Looking back to previous WFPC images, Garnavich et al. (2000) found that Spot 1 had begun to brighten as early as 1996. It has continued to brighten steadily, with a current doubling time scale of about one year. As of February 2000, Spot 1 had a flux $\approx 7\%$ of the rest of the inner circumstellar ring.

Since the detection of Spot 1, several new spots have appeared, of which four (at $P.A. \approx 91°$, $106°$, $123°$ and $230°$, are evident in Figure 5 (Garnavich et al. 2000; Lawrence et al. 2000). Clearly, the blast wave is beginning to overtake the inner circumstellar ring in several places.

The emission line spectrum of Spot 1 resembles that of a radiative shock, in which the shocked gas has had time to cool from its post-shock temperature $T_1 \approx 1.6 \times 10^5 [V_r/(100 \text{ km s}^{-1})]^2$ K to a final temperature $T_f \approx 10^4$ K or less. As the shocked gas cools, it is compressed by a density ratio $n_f/n_r \approx (T_1/T_f) \approx 160[V_r/(100 \text{ km s}^{-1})]^2 [T_f/(10^4\text{K})]^{-1}$. We see evidence of this compression in the observed ratios of forbidden lines, such as [NII]$\lambda\lambda 6548, 6584$ and [SII]$\lambda\lambda 6717, 6731$, from which we infer electron densities in the range $n_e \sim 10^6$ cm^{-3} using standard nebular diagnostics.

The fact that the shocked gas in Spot 1 was able to cool and form a radiative layer within a few years sets a lower limit, $n_r \gtrsim 10^4$ cm^{-3}, on the density of unshocked gas in the protrusion. Given that limit, we can estimate an upper limit on the emitting surface area of Spot 1, from which we infer that Spot 1 should have an actual size no greater than about one pixel on WFPC2. This result is consistent with the imaging observations.

The cooling timescale of shocked gas is sensitive to the postshock temperature, hence shock velocity. For $n_r = 10^4$ cm^{-3}, shocks faster than 250 km s^{-1} will not be able to radiate and form a cooling layer within a few years. It is quite possible that such fast non-radiative shocks are present in the protrusions but are invisible in optical and UV line emission. For example, I estimated above that a blast wave entering the protrusion at normal incidence might have velocity ~ 500 km s^{-1}. We would still see the line emission from the slower oblique shocks on the sides of the protrusion, however.

We have attempted to model the observed emission line spectrum of Spot 1 with a radiative shock code kindly provided by John Raymond. Up to now, our efforts have

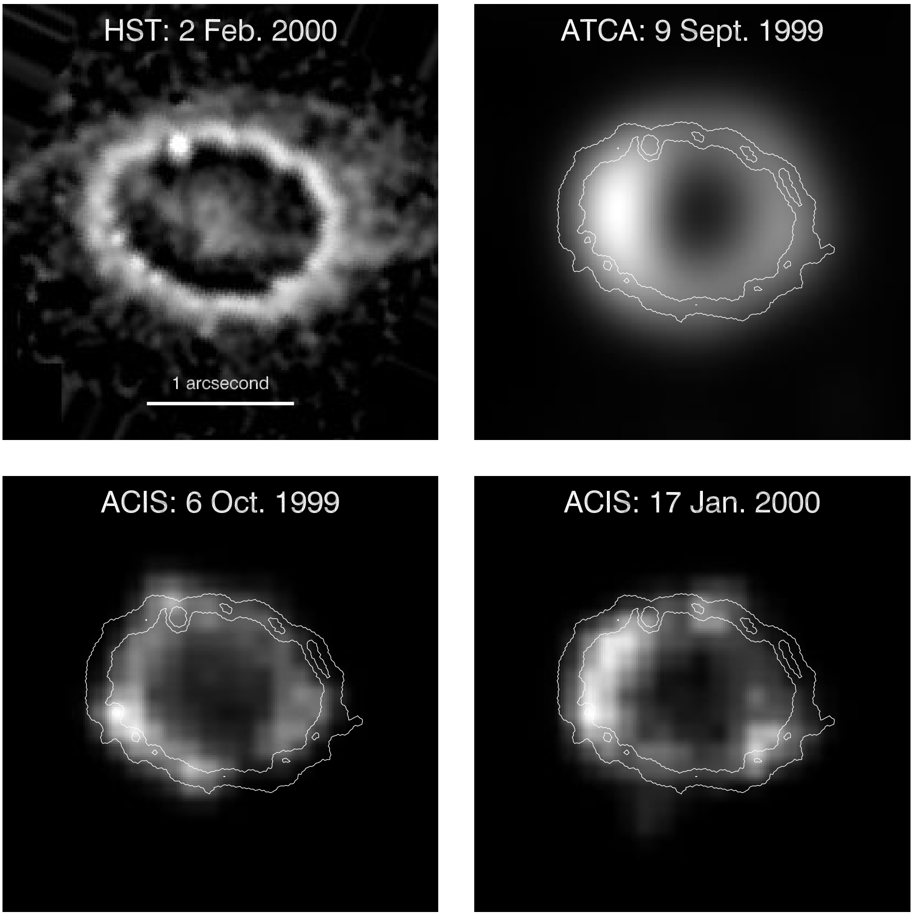


FIGURE 6. Optical, radio and X-ray images of SNR1987A.

met with only partial success. This is perhaps not surprising, given the complexity of the hydrodynamics. It is known, for example, that radiative shocks are subject to violent thermal instabilities (e.g. Innes, Giddings & Falle 1987), which we have not included in our initial attempts to model the shock emission.

7. The X-ray source

As I have already mentioned in §4, we believe that the X-ray emission from SNR1987A seen by *ROSAT* (Figure 2) comes from the hot shocked gas trapped between the supernova blast wave and the reverse shock. But, with its $10''$ angular resolution, *ROSAT* was unable to image this emission; nor was *ROSAT* able to obtain a spectrum.

Very recently, Burrows et al. (2000) used the new *Chandra Observatory* to advance our knowledge of the image and spectrum of X-rays from SNR1987A. Figure 6 is a montage of images of SNR1987A, showing: the *HST* optical image (a); the *ATCA* radio image from the (b); and the X-ray images observed by *Chandra* on 6 October 1999 (c) and 17 January 2000 (d). The optical image (a) is replicated as contour lines on images (b), (c), and (d). We see immediately that the *ATCA* radio source (b) is an annulus that is somewhat smaller than the optical ring and is much brighter on the **E** side than on the **W**.

The two *Chandra* images appear different, but we are not sure whether this difference represents an actual change of the X-ray source or is an artifact of the limited photon statistics of the observations and the deconvolution procedure we used to achieve the maximum possible angular resolution. We can be sure, however, that the X-ray source, like the radio source, is an annulus that lies mostly within the optical ring, and that it is brighter on the **E** side than on the **W**.

The X-ray images of SNR1987A are consistent with the model by Borkowski et al. (1997a), except that they obviously do not have the cylindrical symmetry assumed in that model. The fact that the X-ray and radio images have roughly the same morphologies suggests that the relativistic electrons presumed responsible for the non-thermal radio emission have energy density proportional to that in the X-ray emitting gas and reside in roughly the same volume.

The fact that the X-ray and radio images are both brighter on the **E** side than on the **W** could be explained by a model in which either: (a) the circumstellar gas inside the inner ring had greater density toward the **E**; or (b) the outer supernova debris had greater density toward the **E**. But the fact that most of the hot spots are found on the **E** side favors the latter hypothesis. If the circumstellar gas had greater density toward the **E** side and the supernova debris were symmetric, the blast wave would have propagated further toward the **W** side, and the hot spots would have appeared there first.

This conclusion is also supported by observations of Hα and Lyα emission from the reverse shock (§5), which show that the flux of mass across the reverse shock is greater on the **W** side.

These observations highlight a new puzzle about SN1987A: why was the explosion so asymmetric? We might explain a lack of spherical symmetry by rapid rotation of the progenitor, but how do we explain a lack of azimuthal symmetry?

With the grating spectrometer on *Chandra*, Burrows et al. (2000) also obtained a spectrum of the X-rays from SNR1987A, shown in Figure 7. It is dominated by emission lines from helium- and hydrogen-like ions of O, Ne, Mg, and Si, as well as a complex of Fe-L lines near 1 keV, as predicted by Borkowski et al. (1997a). The characteristic electron temperature inferred from the spectrum, $kT_e \sim 3$ keV, is much less than the proton temperature, $kT_p \sim 30$ keV for a blast wave propagating with $V_b \approx 4,000$ km s^{-1}. This result was expected because Coulomb collisions are too slow to raise the electron temperature to equilibrium with the ions.

The *Chandra* observations show that the current X-ray flux from SNR1987A is about twice the value that would be estimated by extrapolating the *ROSAT* light curve to January 2000 (Figure 2). The X-ray flux is expected to increase by another factor $\sim 10^2$ during the coming decade as the blast wave overtakes the inner circumstellar ring (Borkowski et al. 1997b). Are we already beginning to see the X-ray emission from the shocked ring? Further imaging (with *Chandra*) and spectroscopic (with *XMM*) observations will tell.

8. The future

SNR1987A has been tremendous fun so far, but the best is yet to come. During the next ten years, the blast wave will overtake the entire circumstellar ring. More hot spots will appear, brighten, and eventually merge until the entire ring is blazing brighter than Spot 1. We expect that the Hα flux from the entire ring will increase to $F_{\mathrm{H}\alpha} \gtrsim 3 \times 10^{-12}$ ergs cm^{-2} s^{-1}, or $\gtrsim 30$ times brighter than it is today and that the flux of ultraviolet lines will be even greater (Luo et al. 1994).

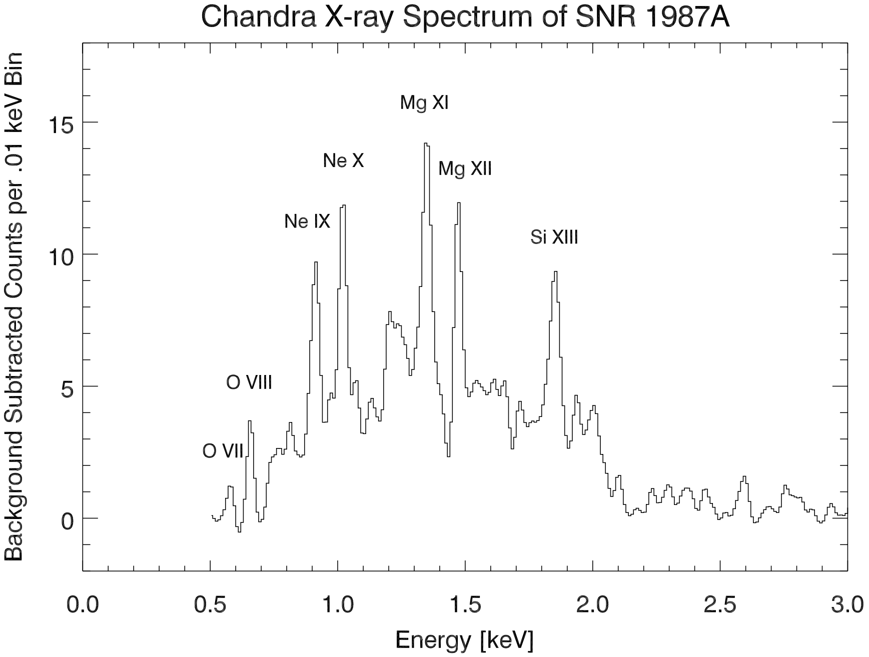


FIGURE 7. X-ray spectrum of SNR1987A.

As we have already begun to see, observations at many wavelength bands are needed to tell the entire story of the birth of SNR1987A. Fortunately, powerful new telescopes and technologies are becoming available just in time to witness this event.

Large ground-based telescopes equipped with adaptive optics will provide excellent optical and infrared spectra of the hot spots. We need to observe profiles of several emission lines at high resolution in order to unravel the complex hydrodynamics of the hot spots. These telescopes also offer the exciting possibility to image the source in infrared coronal lines of highly ionized elements (e.g. [Si IX] 2.58, 3.92 μm, [Si X] 1.43 μm) that may be too faint to see with *HST*. Observations in such lines will complement X-ray observations to measure the physical conditions in the very hot shocked gas.

The circumstellar rings of SN1987A almost certainly contain dust, and so does the envelope of SN1987A (McCray 1993). When dusty gas is shocked to temperatures $\gtrsim 10^6$ K, its emissivity at far infrared wavelengths ($\gtrsim 10$ μm) exceeds its X-ray emissivity by factors $\sim 10^2$–10^3 (Dwek & Arendt 1992). Therefore, we expect that SNR1987A will be brightest at far infrared wavelengths, with luminosity $\sim 10^2$–10^3 times its X-ray luminosity. This makes SNR1987A a prime target for *SIRTF*. Together with the *Chandra* and *XMM* observations of the X-ray spectra, the *SIRTF* observations will give us a unique opportunity to investigate the destruction of dust grains in shocked gas.

The observations with the *ATCA* have given us our first glimpse of shock acceleration of relativistic electrons in real time, but the angular resolution of *ATCA* is not quite good enough to allow a detailed correlation of the radio image with the optical and X-ray images. This will become possible several years from now when the *Atacama Large Millimeter Array* (*ALMA*) is completed. Such observations will give us a unique opportunity to test our theories of relativistic particle acceleration by shocks.

Finally, I'll take this opportunity to do some shameless special pleading to continue intensive observations of SNR1987A with *HST*. Most likely, we won't have a comparable opportunity to observe the birth of a supernova remnant during the lifetime of *HST*—indeed, during our lifetimes. The opportunities to observe SNR1987A with other new facilities do not challenge the preeminence of *HST*; they enhance it.

Of course, we should continue to map the emission of fast Lyα and Hα from the reverse shock with STIS. Such observations give us a three-dimensional image of the flow of the supernova debris across the reverse shock, providing the highest resolution map of the asymmetric supernova debris. We expect this emission to brighten rapidly, doubling on a timescale $\sim$ 1 year. Most exciting, such observations will give us an opportunity to map the distribution of nucleosynthesis products in the supernova debris. We know that the debris has a heterogeneous composition. The early emergence of gamma rays from SN1987A showed that some of the newly synthesized ^{56}Co (and probably also clumps of oxygen and other elements) were mixed fairly far out into the supernova envelope by instabilities following the explosion (McCray 1993). When such clumps cross the reverse shock, the fast Hα and Lyα lines will vanish at those locations, to be replaced by lines of other elements. If we keep watching with STIS, we should see this happen during the coming decade.

Likewise, we should continue to monitor the development of the hot spots with *HST*, using WFPC2 to observe images and STIS to observe spectra. Optical and infrared spectra obtained with ground-based telescopes will be of limited value unless they are complemented by *HST* observations to tell us which spots are producing which lines. Moreover, only *HST* can observe UV lines such as N IV] $\lambda\lambda1483, 1486$ and N V $\lambda\lambda1239, 1243$, the ratio of which are sensitive functions of shock velocity. We need to measure these ratios as a function of Doppler shift to untangle the complex hydrodynamics of the shocks entering the spots. Spot 1 is now becoming bright enough to do that with *HST*.

The shocks in the hot spots are surely producing ionizing radiation, roughly half of which will propagate ahead of the shock and ionize heretofore invisible material in the rings. The effects of this precursor ionization will soon become evident in the form of narrow cores in the emission lines from the vicinity of the hot spots.

In §3 I pointed out that the circumstellar rings of SN1987A represent only the inner skin of a much greater mass of circumstellar matter, and that we obtained only a fleeting glimpse of this matter through ground-based observations of light echoes. The clues to the origin of the circumstellar ring system lie in the distribution and velocity of this matter, if only we could see it clearly. Fortunately, SNR1987A will give us another chance. Although it will take several decades before the blast wave reaches the outer rings, the impact with the inner ring will eventually produce enough ionizing radiation to cause the unseen matter to become an emission nebula. Luo et al. (1994) have estimated that the fluence of ionizing radiation from the impact will equal the initial ionizing flash of the supernova within a few years after the ring reaches maximum brightness. I expect that the circumstellar nebula of SNR1987A will be in full flower within a decade. In this way, SN1987A will be illuminating its own past.

Many of the results reported here are the fruits of my happy collaborations with John Blondin, Kazik Borkowski, Dave Burrows, Steve Holt, Una Hwang, Eli Michael, Jason Pun, George Sonneborn, Svet Zhekov, and the Supernova INtensive Study team led by Bob Kirshner. My own work on SN1987A is currently supported by grants from NASA (ATP NGT 5-80) and the Space Telescope Science Institute (GO-08243.03-97A).

REFERENCES

Borkowski, K. J., Blondin, J. M., & McCray, R. 1997a *ApJ* **476**, L31.

Borkowski, K. J., Blondin, J. M., & McCray, R. 1997b *ApJ* **477**, 281.

Burrows, C. J., et al. 1995 *ApJ* **452**, 680.

Burrows, D. N., et al. 2000 *ApJ*, **343**, L149.

Chevalier, R. A. 1982 *ApJ* **258**, 790.

Chevalier, R. A. 1992 *Nature* **355**, 617.

Chevalier, R. A. & Dwarkadas, V. I. 1995 *ApJ* **452**, L45.

Crotts, A. & Heathcote, S. R. 1991 *Nature* **350**, 683.

Crotts, A. P. S., Kunkel, W. E., & Heathcote, S. R. 1995 *ApJ* **438**, 724.

Dwek, E. & Arendt, R. G. 1992 *ARA&A* **30**, 11.

Eastman, R. G. & Kirshner, R. P. 1989 *ApJ* **347**, 771.

Ensman, L. & Burrows, A. 1992 *ApJ* **393**, 742.

Fransson, C., et al. 1989 *ApJ* **336**, 429.

Garnavich, P., Kirshner, R., & Challis, P. 2000 *IAU Circ. 7360*

Hasinger, G., Aschenbach, B., & Trümper, J. 1996 *A&A* **312**, L9.

Innes, D. E., Giddings, J. R., & Falle, S. A. E. G. 1987 *MNRAS* **226**, 67.

Lawrence, S. S., et al. 2000 *ApJ*, **537**, L123.

Luo, D. & McCray, R. 1991 *ApJ* **372**, 194.

Luo, D., McCray, R., & Slavin, J. 1994 *ApJ* **430**, 264.

McCray, R. 1993 *ARA&A* **31**, 175.

Michael, E., et al. 1998 *ApJ* **509**, L117.

Michael, E., et al. 2000 *ApJ*, **542**, L53.

Podsiadlowski, P. 1992 *PASP* **104**, 717.

Pun, C. S. J. 1997 *http://oposite.stsci.edu/pubinfo/PR/97/03.html/*

Schmütz, W., et al. 1990 *ApJ* **355**, 255.

Sonneborn, G., et al. 1998 *ApJ* **492**, L139.

Wampler, J., et al. 1990 *ApJ* **362**, L13.

Globular clusters: The view from *HST*

By WILLIAM E. HARRIS

Department of Physics & Astronomy, McMaster University, Hamilton ON L8S 4M1 Canada

Globular clusters represent only a small fraction of the total mass in their parent galaxy, but provide a vast array of tests for stellar physics, dynamics, and galaxy formation. This review discusses the prominent accomplishments of *HST*-based programs:

(a) The definition of precise fiducial sequences in the HR diagram, extending down to the hydrogen-burning limit,

(b) Discovery of the upper white dwarf cooling sequence in several clusters,

(c) Discovery of a highly consistent IMF on the lower main sequence,

(d) Definitive age measurements for the oldest clusters in the outermost halo of the Milky Way, the Magellanic Clouds, and the dwarf elliptical satellites of the Milky Way,

(e) Elucidation of the innermost structure of M15 and other core-collapsed clusters,

(f) Discovery of surprisingly large "anomalous" populations of stars within dense cluster cores: extended blue horizontal-branch stars, blue stragglers, and others,

(g) The first reliable color-magnitude studies for globular clusters in M31, M33, and other outlying Local Group members,

(h) Discovery of massive young clusters in starburst galaxies with ages as small as 1 Myr,

(i) Measurement of metallicity distribution functions among globular cluster systems in many giant E galaxies—bimodality is common, but details differ strongly, and

(j) Deep imaging of cluster luminosity distributions in gE galaxies in Virgo, Fornax, and other Abell clusters as distant as Coma.

The review concludes with brief speculations about future directions in which *HST* and (later) *NGST* might contribute to globular cluster studies.

As scientists, we are only as good as our ideas and our tools. There have been few times indeed in the history of astronomy when a single major tool has made the impact of *HST*. But while this magnificent observatory has helped us to lay out new paths that scarcely existed a decade ago, we have also used it to radically transform more traditional territory.

One of these areas—star clusters and stellar populations—spans the entire 20th century in the astrophysics literature. Few branches of astronomy have such a long history as the study of globular clusters, or have re-invented themselves so many times because of technological advances and innovations. *HST* is the latest and certainly among the most influential of these revolutions. A flood of observational results from it has generated yet another transformation in our understanding of these dense, populous, and endlessly fascinating stellar systems.

The key advantages conferred by *HST*—unequaled spatial resolution, photometric depth, access to the ultraviolet, and unique time-series observations—are the same as for any other field, though the first two in the list have been the major players here more than in other areas. For these reasons, almost all of the major contributions of *HST* to globular cluster studies have entered the literature only after 1995—that is, starting with Cycle 4 and the installation of COSTAR and WFPC2. This brief overview starts from within the Milky Way and works outward to the current limits of distance to which we can probe globular cluster systems. No claim is made to completeness, but it isolates what I will suggest to be the main directions in which *HST* has influenced the discipline.

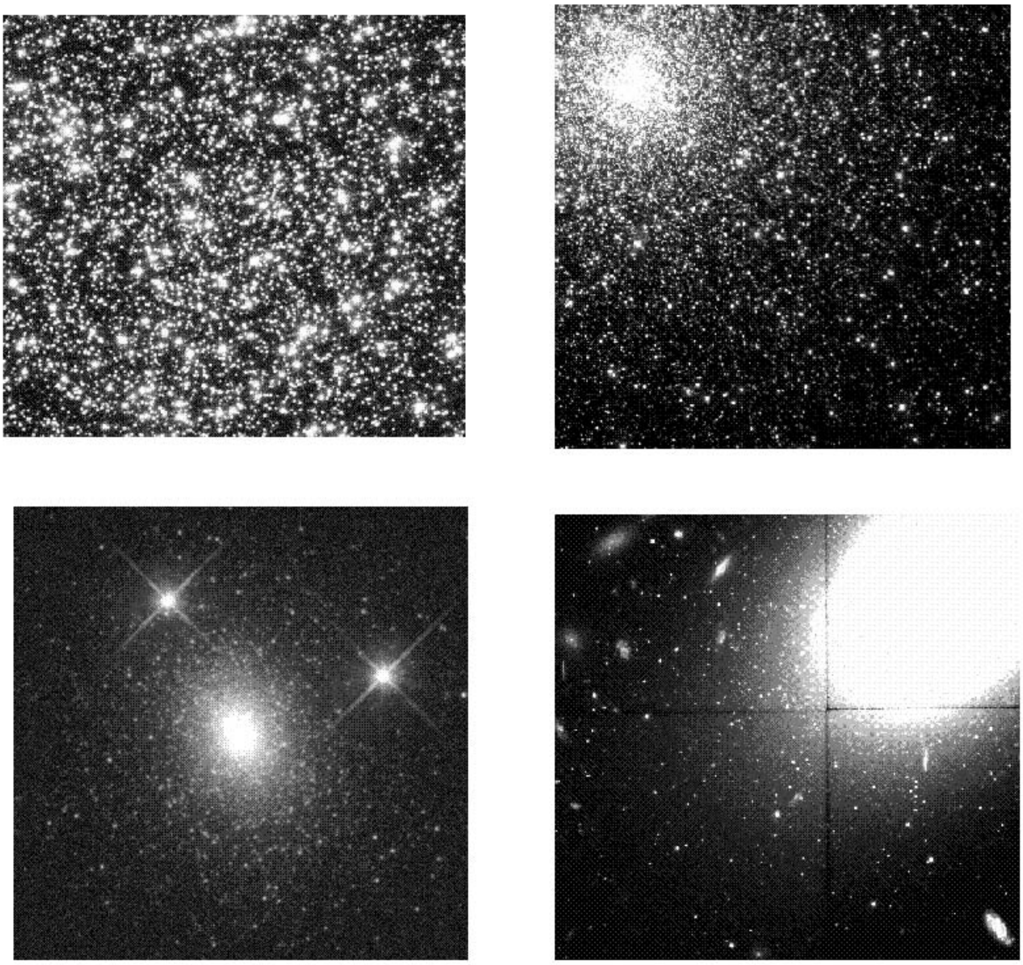


FIGURE 1. Globular clusters, as seen by *HST* from four distance perspectives: (a) *Upper left*: the center of 47 Tucanae ($d = 4.5$ kpc), showing the dense population of giants and main sequence stars (R. Saffer and D. Zurek, NASA/St ScI Press Release PRC97-35). (b) *Upper right*: NGC 6093 ($d = 10$ kpc; Hubble Heritage image). (c) *Lower left*: Cluster G1 = Mayall II, the most luminous globular cluster in the halo of M31 at $d = 0.75$ Mpc (M. Rich, K. Mighell, J. Neill, and W. Freedman, NASA/St ScI Press Release PRC96-11). (d) *Lower right*: the Coma giant elliptical IC 4051 ($d = 100$ Mpc), around which the globular clusters show up as a sprinkling of faint starlike images (Baum et al. 1997; Woodworth & Harris 2000).

1. Stellar populations in globular clusters

1.1. *The limits of the color-magnitude diagram*

The color-magnitude diagram (CMD) for globular clusters forms the classic testbed for studying the evolution and mass distribution of low-mass stars. Before the launch of *HST*, a whole series of unanswered and quite fundamental questions about such stars were on hand: What does the main sequence look like down to the hydrogen-burning limit? What is the IMF (initial mass function) for stars down to the same limit, and does it depend strongly on metallicity or other factors? Would the position of the white dwarf sequence confirm the expectations of standard stellar structure theory? What trends are exhibited by "anomalous" types of stars such as blue stragglers and extended horizontal-

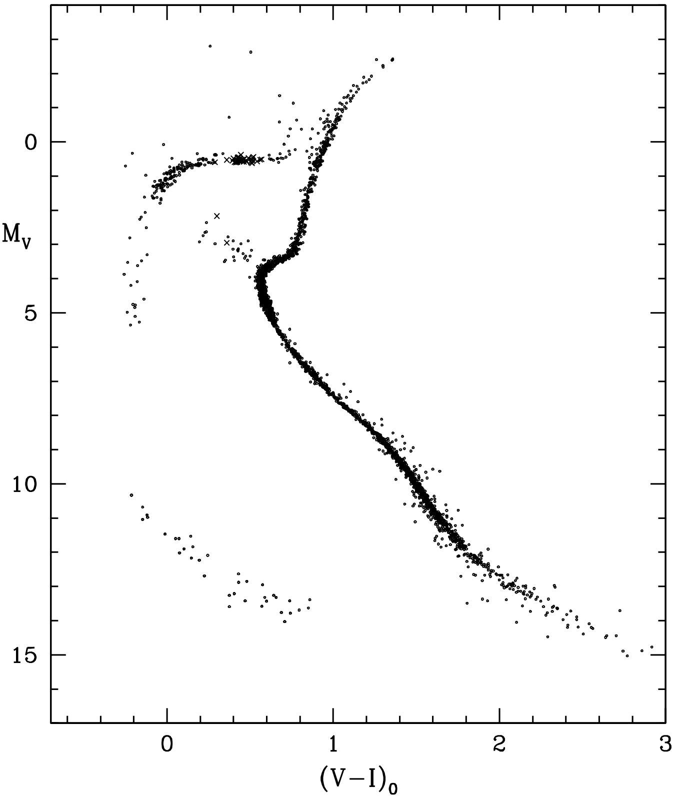


Figure 2. Composite color-magnitude diagram for a fiducial "metal-poor" ([Fe/H] $\simeq -2$) globular cluster. Here data are combined from NGC 6397 (King et al. 1997), M55 (Piotto 2000), M68 (Walker 1994), NGC 2419 (Harris et al. 1997), and M3 (Ferraro et al. 1997b). Each point represents a real star taken from one of the above datasets, but they have been deliberately and rather arbitrarily selected to delineate the various evolutionary stages in the CM diagram as clearly as possible. *Small crosses* denote variable stars (RR Lyraes and two SX Phe variables).

branch stars? Are there distinctive features in the CMD distribution that were previously unknown?

HST programs have provided stimulating responses to all these questions. The present "state of the art" in our definition of the full CMD array for globular clusters can be seen in Figure 2, which is a composite of data from five different clusters of low metallicity ([Fe/H] $\simeq -2$) and is deliberately constructed to show the entire range of globular cluster stars as we now understand them. The precise definition of the principal sequences, as well as the total span in luminosity—covering more than 18 magnitudes from the brightest giants to the faintest dwarfs—far exceeds anything from the pre-*HST* era.†

† It should be emphasized that Fig. 2, as a composite from several sources, does *not* correctly represent any single cluster, nor does it show the true relative numbers of stars in very different sections of the CMD. For example, the true numbers of HB and red giants relative to the unevolved main sequence stars are far smaller in any real cluster than indicated here. The diagram is intended only to display the locations of the fiducial sequences. Some "anomalous" regions such as the blue stragglers and extreme blue extension of the horizontal branch have been somewhat deliberately overemphasized. Finally, in combining data from the clusters listed, I have taken arbitrary liberties in removing a few stars lying far off any of the sequences which almost certainly represent residual field contamination.

The genuine level of progress achieved by the *HST*-based photometry can be appreciated further by contrasting Fig. 2 with the very earliest CMDs of globular clusters, first measured 90 years ago by Shapley and his collaborators. (For an excellent and representative example, see the diagram for M3 by Shapley & Davis 1920; it was presented in the form of a 'spreadsheet', with numbers of stars labeled in bins of magnitude and 'color class'; the upper red giant branch and most of the horizontal branch can clearly be seen). Another of the two or three truly major turning points in the history of this subject took place just half a century ago, when the long-sought *main sequence turnoff* point (MSTO) was reached for the first time (Arp, Baum, & Sandage 1953; the cluster concerned was the prototypical extreme low-metallicity object M92). This event also represented an extremely rare instance in which both observers and theorists converged on the same target at the same time: it was no accident that stellar models were being constructed specifically to try to understand the puzzles posed by Population II stars and their old, metal-poor red giant branch. The first models which could claim basic success in matching the globular cluster turnoff region and the transition to the convective red giant structure—in other words, the first "isochrone fits" in the sense we use them today—suggested that the age of M92, and thus the halo of the Galaxy, was 3–4 Gyr (Sandage & Schwarzschild 1952). The biggest error in these fits was a too-bright $M_{bol} \simeq -0.5$ adopted luminosity scale for the RR Lyraes, which set the zeropoint for the main sequence—again, an issue which persists today albeit at a lower level (see, for example, Carretta et al. 2000 for a comprehensive recent discussion of the globular cluster distance scale issues).

Ever since that critical point in history, the mutual stimulation between stellar structure theory and globular cluster photometry has constantly been active, and has taken a new leap in the *HST* era. The delineation of the faintest sections of the zero-age main sequence (ZAMS) all the way to its lower limit has unquestionably encouraged the development of low-mass stellar models over a range of metallicities (e.g. Cassisi et al. 2000) which follow the observed ZAMS shape in remarkable detail. Similarly, new predictions now exist for the shape of the white dwarf cooling curve as it extends below even the faintest observational limits that we now have (Figure 3).

1.2. *The quest for the white dwarf sequence*

One of the original goals in the scientific case for the launch of *HST* was to discover the very deepest and faintest parts of the CMD that had never been seen from the ground—the extension of the ZAMS all the way to the hydrogen burning limit, and the cooling sequence of the white dwarfs. But even with *HST* and WFPC2, these dim stars—some of them barely able to generate a slow trickle of hydrogen fusion, while the others are leaking away their tiny remaining reservoirs of internal heat—can clearly be seen only through long exposures of just the nearest globular clusters (those within a few kiloparsecs of the Sun).

For the white dwarf sequence especially, the challenge was not simply one of limiting magnitude. The WDs are not present in the same numbers as the ubiquitous main-sequence stars, and obtaining many of them within the limited WFPC2 field of view requires selecting target fields uncomfortably close to the cluster center, with the attendant problems of crowding and scattered light from the other types of cluster stars, all of which are much brighter (Figure 4).

Despite these hurdles, clear successes have been achieved. Unambiguous *sequences* of stars delineating the white dwarf cooling curve were published almost simultaneously for three such clusters: M4 (Richer et al. 1995), NGC 6397 (Cool et al. 1996), and NGC 6752 (Renzini et al. 1996), all as a result of Cycle 4 programs. Several other clusters were soon

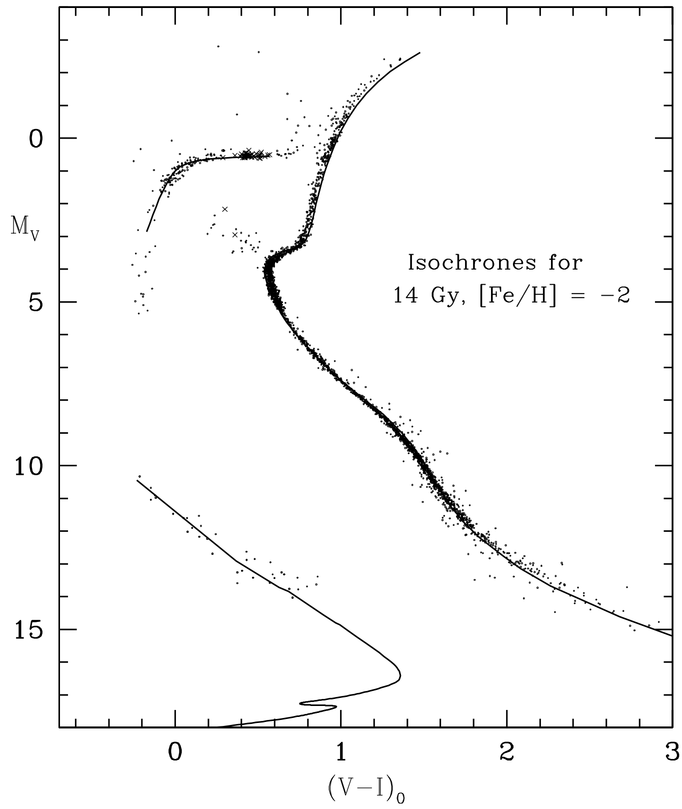


FIGURE 3. Composite isochrone fits to the fiducial CMD of Figure 2. The adopted age (14 Gyr) and metallicity ([Fe/H] = −2) is the same for all parts of the sequences. The ZAMS model line fainter than $M_V \simeq 8$ is from Cassisi et al. (2000), while the upper main sequence region, turnoff, and red giant branch models are from Harris et al. (1997) and the horizontal branch from VandenBerg et al. (2000). The white dwarf evolutionary track is from isochrones provided by Fontaine (2000, private communication).

added to the list (ω Cen, 47 Tuc, M15, and others). It was a significant triumph for both stellar structure modeling and the observational calibration of the ZAMS and ZAHB distance scales that these sequences appeared at the luminosity and color levels at which they were expected for white dwarf masses of 0.5–0.6 $M_\odot$ (Figure 5).

For the lowest-temperature white dwarfs, the stellar atmospheres become cool enough that H_2 opacity dominates, and the flux is redistributed in such a way that the $(V-I)$ color indices become progressively bluer again as the stars become still fainter (Hansen 1999). For the old globular clusters, this *white dwarf turnback point* (WDTB) should appear about three magnitudes fainter than the limits of currently published *HST*-based CMDs (see Figure 4) and thus presents a formidable challenge for future photometry. Nevertheless, globular clusters are old enough that the first white dwarfs produced in them—12 Gyr ago or more—should have passed the WDTB point and now lie on the lower branch of the cooling sequence. This final frontier of the CMD will provide an intriguing test of the basic theory for white dwarfs that have evolved nearly to extinction.

1.3. *The faint main sequence and the IMF*

Stars that are just as faint as the white dwarfs, and equally interesting for a different set of reasons, are those at the bottom end of the main sequence. In the pre-*HST* era, in-

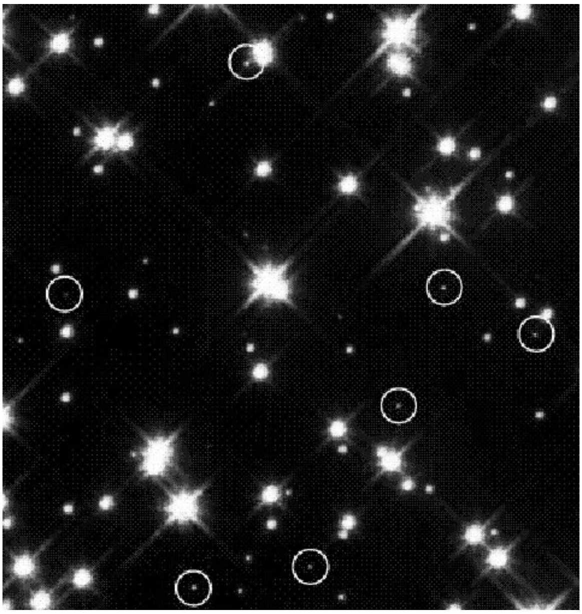


FIGURE 4. A portion of the field within M4 (Richer et al. 1995; from H. Richer and H. Bond, NASA/ST ScI Press Release PR95-32) in which the white dwarf sequence was first identified. Some of the more than 100 WDs found in this cluster are shown as the faint blue stars in the circles; most of the other objects are main-sequence stars.

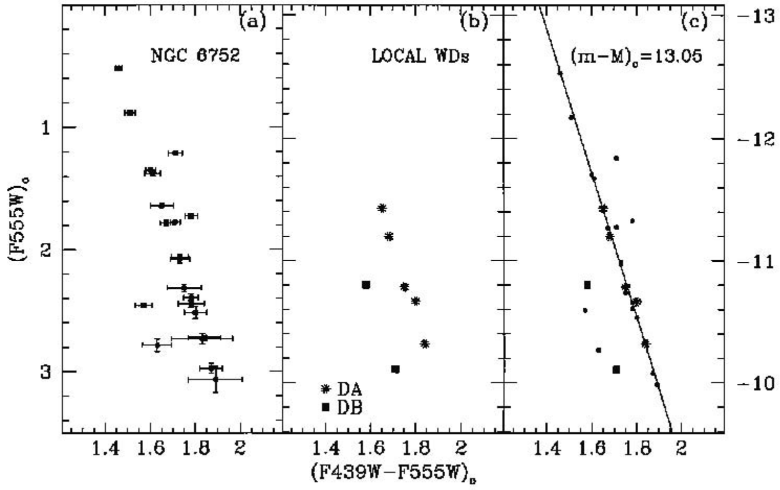


FIGURE 5. The white dwarf sequence in NGC 6752 (from Renzini et al. 1996), superimposed on theoretical cooling curves for 0.5 and 0.6 $M_\odot$.

vestigations of main sequence luminosity functions (LF) suggested that cluster-to-cluster differences in the LF might exist, possibly correlated with cluster metallicity or Galactocentric location (e.g. Capaccioli et al. 1991, 1993; Djorgovski et al. 1993). However, these studies were always limited by the inability of ground-based photometry to reach further down the main sequence than $\sim 0.3\ M_\odot$, or to survey all parts of the clusters including their populous core regions. The conversion of the LF to the more fundamental initial

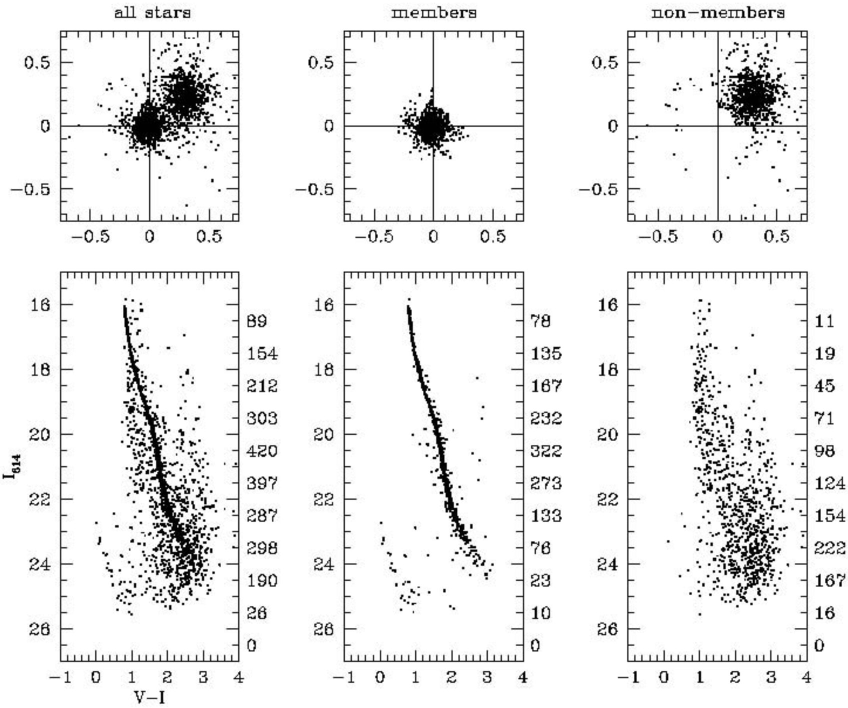


FIGURE 6. WFPC2 photometry of the lower main sequence for NGC 6397, from King et al. (1998). The top panels show the proper motion measurements for all stars in the field (at left), then their division into cluster members (center) and field stars (right). The corresponding color-magnitude arrays in $(I, V-I)$ are shown below each panel, with numbers of stars in each magnitude bin shown on the right border. The use of proper motion provides an extremely effective filter against field contamination.

mass function (IMF), through an assumed mass-luminosity relation and modeling of the dynamical history of the cluster, also left important uncertainties in the discussion.

This situation changed dramatically with the advent of *HST*. A variety of photometric studies reaching absolute limits $M_I = 10$ or more on the main sequence (e.g. von Hippel et al. 1996; Ferraro et al. 1997a; DeMarchi & Paresce 1997, among many others) strongly suggested that the LF in several clusters exhibited a consistent, sharp turndown near $M_I \simeq 9$. Converting the LF into the *mass* function and properly accounting for metallicity meant that this turndown corresponded to a nearly uniform mass of 0.25 $M_\odot$ in all clusters. Below this point, the IMF was nearly flat in number of stars per unit mass interval. Still deeper data now available for a total of a dozen globular clusters (Paresce & De Marchi 2000) confirm the uniformity of this turnover, and appear to put to rest much of the earlier debate over strong cluster-to-cluster differences. Correlations with extreme differences in Galactocentric distance, however, remain to be explored fully.

The deepest surveys of all—just as for the white dwarf sequence—have been performed on the very nearby NGC 6397 (e.g. King et al. 1998; De Marchi et al. 2000). Using deep WFPC2 images of cluster fields separated by a mere three-year baseline, King et al. (1998) have employed proper motion measurements to separate out cluster members from foreground and background field stars, tremendously reducing the field contamination (Figure 6). With this remarkably clean CMD in hand, we can trace the faint end of the main sequence usefully to a point not far above the hydrogen-burning limit: in the

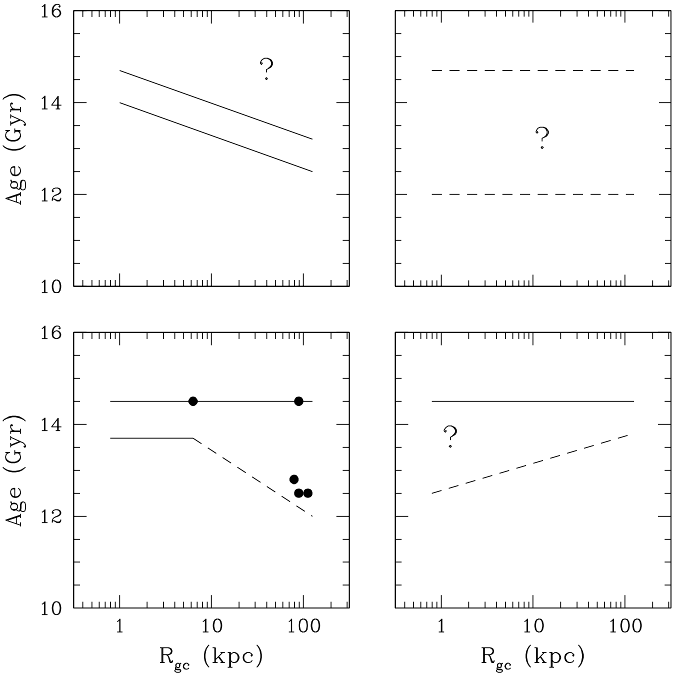


FIGURE 7. Various theoretical scenarios for the age distribution of globular clusters as a function of Galactocentric distance. (The age scale arbitrarily assumes the oldest clusters began at $\tau \simeq 14.5$ Gyr.) See text for explanation. The datapoints in the lower left panel give schematic recent results for some of the outer-halo clusters.

resulting mass function, the IMF stays nearly flat for $0.25\ M_\odot \gtrsim M \gtrsim 0.10\ M_\odot$ until at last we see a strong hint of a turndown before the ultimate limits of the data at $\sim 0.09\ M_\odot$.

For $M \lesssim 0.1\ M_\odot$, the luminosity and temperature of the star decline so steeply with decreasing mass (compare again Fig. 4) that deeper measurements of the IMF in *any* cluster will be extremely difficult to achieve. Nevertheless, the studies already in hand represent a major accomplishment. They strongly suggest that the IMF for a "generic" globular cluster can be rather simply represented by an approximate Salpeter MF for $M \gtrsim 0.3\ M_\odot$ and a flat MF over 0.3–0.1 $M_\odot$. Large amounts of mass in a near-invisible substellar (brown-dwarf) component seem unlikely.

2. The age profile of the Galactic halo

A long-standing astrophysical issue has been to determine the ages of the globular clusters and to relate them to the star-forming history of our Galaxy. Since they are found at all galactocentric distances from the Galaxy's inner bulge out to remote satellites at 100 kpc or more, the globular clusters provide the best chance we have to determine the *age distribution function* τ for the Galactic halo.

In principle, τ could vary with Galactocentric location R_{gc}, metallicity [Fe/H], or other factors; in the absence of data, there would be no shortage of theoretical possibilities for the distribution of τ (Figure 7). Perhaps, as in the early picture of Eggen, Lynden-Bell, & Sandage (1962; ELS) the Galactic halo built rather rapidly from the inside out in a single major star-forming epoch (Fig. 7a). In the simplest versions of this model, we would expect to see a narrow range of cluster ages and a tight correlation with R_{gc}. Or, as in the opposite extreme of Searle & Zinn (1978; Fig. 7b), perhaps the entire halo was built piecemeal in a longer and more disconnected series of events involving star formation in dwarf-galaxy-sized protogalactic clouds, which then amalgamated to form the larger Galaxy; in that case, there might be a much larger range of cluster ages and little correlation with R_{gc}. Tangible evidence that the halo has grown by accretion in this way can be found in the Sagittarius dwarf now being tidally damaged (e.g. Ibata et al. 1995, 1997) and in the moving groups of halo stars that have been isolated in recent surveys (Majewski et al. 1996; Majewski 1999). Hybrid combinations are, of course, possible: one version mentioned frequently in the recent literature, beginning with Searle & Zinn, postulates that the inner ($R_{gc} \lesssim R_{\odot} \simeq 8$ kpc) halo could have formed rapidly as in ELS but that the outer halo accreted or underwent star formation over progressively longer times (Fig. 7c). Alternately, we could even imagine—based on the many examples of merger events between disk galaxies that are going on today and could have been more frequent in the past—that gas accretion into the Galactic bulge, with attendant star and cluster formation, could have gone on for many Gyr and that the cluster age range in the *inner* halo could be largest (Fig. 7d).

The expectations of these various schematic models are different enough that it should be possible for the actual data to rule out at least some of them. In principle, the observational task is straightforward: (a) obtain a CMD for each cluster which defines the main sequence, turnoff region, giant branch, and horizontal branch precisely. (b) Measure the cluster metallicity [Fe/H] and (if possible) the relative abundances of C,N,O, and other key elements. (c) Decide on the cluster foreground reddening and (somehow) its distance, perhaps through a calibration of the horizontal branch or RR Lyrae luminosity, or through subdwarf parallaxes normalized to the metallicity of the cluster. Then, (d) fit appropriately constructed model isochrones transformed to the observational plane, and find the best-fit age.

Every step in this prescription has generated controversy and a lengthy history in the literature. Furthermore, even the first step (which is where *HST* observations come in) is fraught with problems of its own. In the distance regime $R_{gc} \lesssim 5$ kpc, most of the target clusters are strongly affected by differential reddening from dust clouds along the line of sight in the Galactic disk, and contamination from Galactic bulge stars. Furthermore, a large fraction of the field stars are at nearly the same metallicity, and nearly the same distance, as the cluster stars. At the opposite extreme ($\sim$ 40–120 kpc, out to the remote limits of the halo), field contamination or differential reddening are unimportant, but image crowding and sheer distance are important. The most secure zone is still the intermediate halo ($\sim$ 5–40 kpc) in which ground-based photometry from large telescopes can provide good results.

The *outermost halo* has only a handful of clusters but can provide critical leverage on the model interpretations. To date, deep main sequence photometry has been obtained for four of these, all at $R_{gc} \sim 100$ kpc (NGC 2419, Pal 3, 4, and Eridanus), with two remaining ones (Pal 14, AM1) under study. The prominent results so far are that the biggest and most metal-poor of these (NGC 2419; Harris et al. 1997) has the same age to within ±0.5 Gyr as the comparison "template" low-metallicity cluster M92; while the three sparse clusters (Pal 3, 4, Eri; Stetson et al. 1999) collectively are about 1.5–2 Gyr younger than the template clusters M3 or M5 of similar metallicity. (In turn, M3 and M5 might be slightly younger than the more metal-poor M92.)

The eventual goal of this process is of course to establish absolute cluster ages, but the age scale depends on detailed concerns in the stellar models that continue to be debated in the literature (e.g. Carretta et al. 2000; Chaboyer et al. 1996, 1998; VandenBerg 1999; VandenBerg et al. 1996, among many others), with best estimates for the *oldest*, most metal-poor clusters ranging from 12 to 14 Gyr. Instead, to make progress on un-

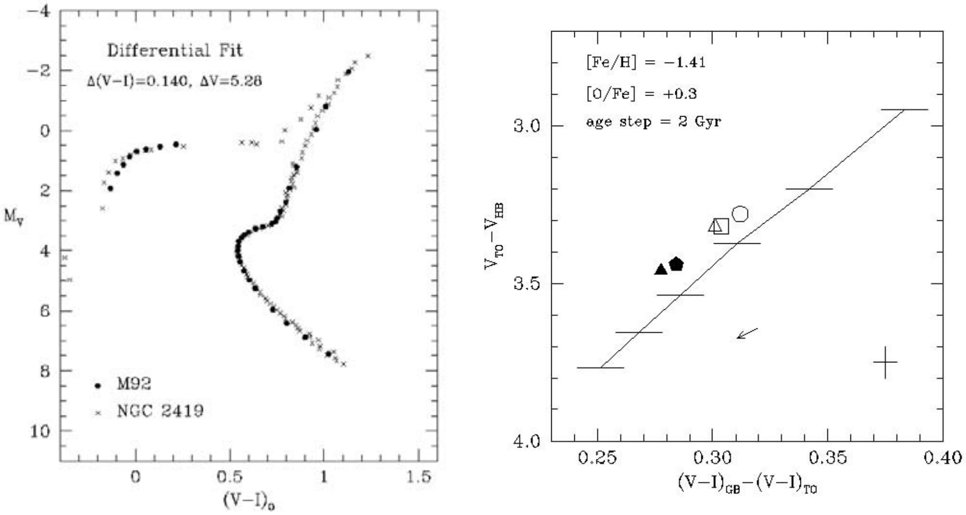


FIGURE 8. *Left panel:* CMD match between the low-metallicity clusters M92 and NGC 2419 (Harris et al. 1997), displayed as mean points in $V, (V - I)$. Both clusters appear to be the same age to within 0.5 Gyr, under the assumption that their chemical compositions are similar. *Right panel:* Parametric match ($\Delta V, \Delta C$) from Stetson et al. (1999) between M3 and M5 (solid symbols) and Pal 3, 4, Eridanus (open symbols). For smaller ages, the magnitude difference ΔV decreases while the color difference ΔC increases; the theoretical model line is marked in 2-Gyr increments.

derstanding the shape of the age distribution $\tau(R_{gc}, \text{[Fe/H]})$, attention has turned over the past decade to measuring *relative* ages by matching CMDs of one cluster to another (VandenBerg et al. 1990; Sarajedini & Demarque 1990). Two principal age indicators are used: the magnitude difference $\Delta V(MSTO - HB)$ between the horizontal branch and main sequence turnoff; and the color difference ΔC between the turnoff and the base of the giant branch. Two ways to display these parameters are shown in Figure 8.

The clear indication from these studies, and others now in the literature which cover the intermediate and outer halo, is that the lowest-metallicity clusters ([Fe/H] $\lesssim -2$) have identical ages (to within the current limits of precision of about 0.5 Gyr) *everywhere in the halo*. Thus, star formation began throughout the Galactic protohalo (or equivalently, throughout all the fragmentary clouds that would eventually amalgamate) at almost the same time. Comparing the metal-rich and metal-poor cluster ages is trickier, since here we must rely on the absolute accuracy of the stellar models rather than differential comparison, but the rather clear signal that has emerged from all the data over the past decade is that the main system of halo clusters took about 2 Gyr to form. In addition, a *few* small clusters (IC 4499, Pal 12, Rup 106, Ter 7, Arp 2) have been found with ages as much as 5 Gyr below than the average of the main system (Stetson et al. 1989; Buonanno et al. 1994; Ferraro et al. 1995; Rosenberg et al. 1998).

It remains to ask what has developed for the Galactic bulge clusters. The most commonly used metal-rich template, 47 Tucanae, has an age $\sim$ 12–14 Gyr (depending on the details of the adopted composition and stellar models) which is not obviously different from the outer clusters (e.g. Hesser et al. 1987; Santiago et al. 1996). But many of the clusters much further in to the Galactic center have significantly different metallicities, structures, orbital characteristics, etc. and it is not at all clear that they would necessarily have the same ages too. An analysis of NGC 6352 (a Herculean effort by Fullton et al. 1995 with the original WF/PC camera) yielded a close match to 47 Tuc to within perhaps 1 Gyr. A WPFC2 analysis of NGC 6528 and 6553, both inner bulge clusters, showed that these too were classically "old" and "... strongly support the notion of fast formation and chemical enrichment of the bulge" (Ortolani et al. 1995). More recently, it has been suggested that some other bulge objects (such as Terzan 1; see Ortolani et al. 1999) might indeed be younger by something near 2 Gyr. All of these studies have been severely hampered by the obstacles of differential reddening and field contamination mentioned above, and it is likely that no age comparisons more precise than ±2 Gyr *at best* have yet been possible there. An alternate approach would be to do the photometry in the near infrared, which would greatly alleviate the crushing handicap of differential reddening and thus enable the definition of precise main sequences. However, a compensating loss is that the isochrones in *JHK* are not as sensitive to age differences as they are in optical color indices.

These clusters, embedded deep in the starfields of the Galactic bulge and lurking behind the dust clouds scattered through the disk, have stubbornly resisted four decades' worth of attempts to determine precise ages. But without them, the full story for the early star formation history of the Galaxy remains untold in its entirety.

3. Core structures and stellar populations

Another of the original high expectations for *HST* was to resolve the stellar populations of globular clusters in their innermost cores, which had been seen in only the sketchiest terms through ground-based imaging. Keen interest accompanied the discovery in the late 1980s of clusters in a wide range of dynamically evolved states, including many that had experienced core collapse and strong mass segregation, and the cameras aboard *HST* afforded the first clear and comprehensive view of these objects (see again Fig. 1a).

The single cluster which has been subjected to the most intensive scrutiny is unquestionably M15, which is relatively nearby, populous, old, metal-poor, and core-collapsed. A series of *HST*-based photometric studies following on from pioneering ground-based programs have shown that the power-law rise in surface density that is characteristic of post-core-collapsed (PCC) clusters continues inward in M15 to the smallest radii yet observable (e.g. Yanny et al. 1994; Guhathakurta et al. 1996; and references cited). At the distance of M15, this innermost radial zone at $r \sim 0''.3$ is equivalent to about 33,000 AU, or the size of the Oort cloud (Figure 9).

A long-standing question has been whether or not such clusters might contain some central massive object (such as a single black hole, or a cluster of neutron stars). Further information can be obtained through the stellar radial velocities, and any radial trends of the velocity dispersion profile or mean rotation speed. Results obtained with adaptive optics at the CFHT (Gebhardt et al. 2000) show that an isotropic velocity dispersion model with $(M/L)_V = 1.7$ matches the core velocities acceptably, but the effect of a central condensed mass such as a black hole (if any) would only show up definitively at radii $r \lesssim 2''$. Interestingly, the stars in the inner $3''.4$ (1.7 pc) show a net rotation $v(\mathrm{rot}) = 10$ km s^{-1}, and all the data combined are consistent with "... a central mass concentration equal to 2500 M$_\odot$." A STIS program by van der Marel et al. (2000) now in progress will add more data interior to this radius and should be able to detect any central pointlike mass larger than ~ 1000 M$_\odot$.

Although the new data on core *structures* were eagerly, and correctly, anticipated, unexpected results lay in store for the *stellar populations*. One of the first results from a *BV* imaging survey of cluster centers showed, to everyone's surprise, that there were metal-rich clusters with strong *blue* horizontal-branch components (NGC 6388, 6441; Rich et al. 1997; also see Buonanno et al. 1997). These high-T_e HB stars are presum-

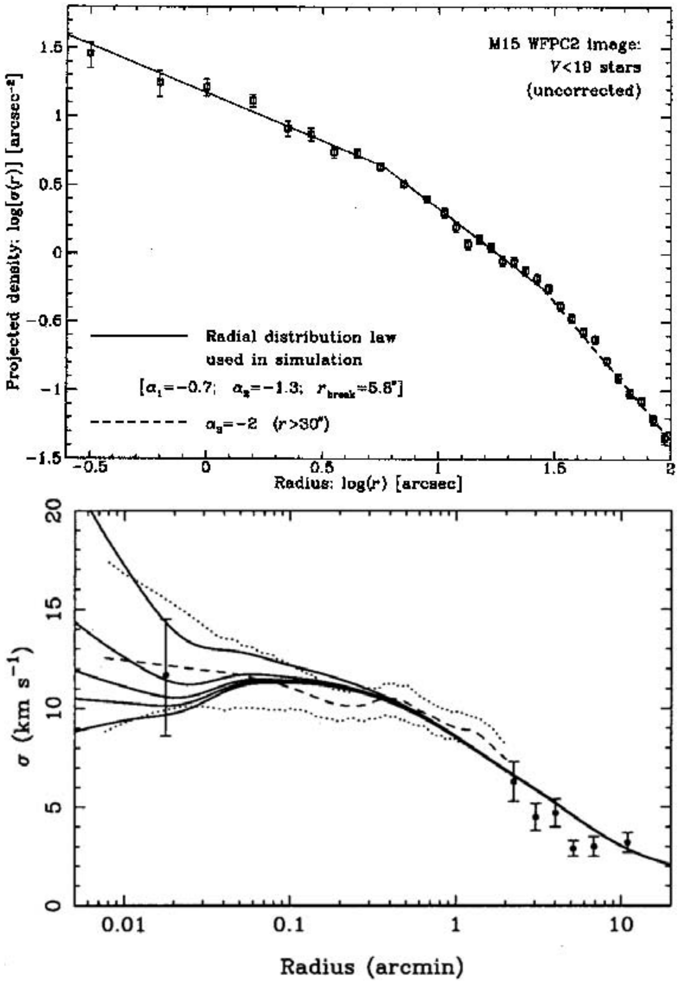


FIGURE 9. *Upper panel*: Projected radial density profile for M15, from Guhathakurta et al. (1996). The power-law rise of the core density continues inward to the smallest observed radius, equivalent to $r = 33,000$ AU. *Lower panel*: Velocity dispersion profile for M15 from high resolution ground-based measurements, from Gebhardt et al. (2000). The five model curves show the expected trend for isotropic models containing a central black hole of (from lowest curve upward) 0, 500, 1000, 2000, and 6000 $M_{\odot}$.

ably ones in which almost all the surface hydrogen envelope has been removed, leaving only a low-mass surface envelope above the helium-burning core. Such stars are almost completely absent in the outskirts of the same clusters, but are found in large numbers within the dense core-collapsed centers. In other metal-rich clusters with less dense cores (47 Tuc, NGC 5927) they are not to be found. What has happened to drive extra mass loss or envelope stripping from the giants in these unusually dense environments?

The compact cores of many clusters have also turned out to be preferred habitats for other types of unusual stars mixed in with their more normal neighbors. Binary stars should be much more common in the cluster cores, since they are expected to form through dynamical evolution and core contraction (e.g. Hut et al. 1992), and it came as no surprise that *HST* imaging programs have shown that core binary populations do indeed follow the predicted trends (see, e.g. the results of Edmonds et al. 1996 for

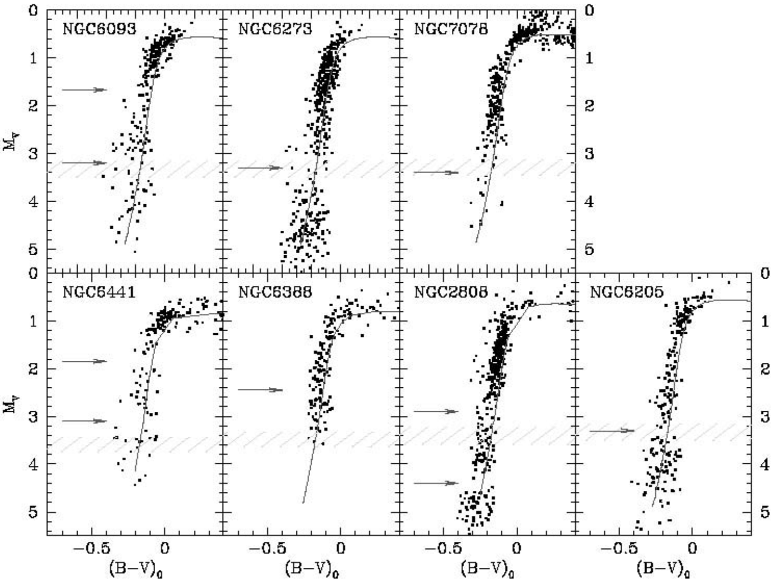


FIGURE 10. Color-magnitude diagrams for the extended blue horizontal branches of seven clusters, from Piotto et al. (1999). Metal-poor clusters are in the top row, while more metal-rich ones are in the bottom row. Arrows mark distinctive gaps in the EBHB sequence. In all seven, a gap appears at a temperature corresponding to 0.535 $M_\odot$.

47 Tuc). However, the extreme blue HB (EBHB) stars mentioned above have now been found routinely in many low-metallicity clusters as well (e.g. Sosin et al. 1997; Ferraro et al. 1998; Piotto et al. 1999), with the additional effect of clear 'gaps' at certain places along the BHB sequence (all of them far hotter than the RR Lyrae region). The positions of these features may be quite similar from cluster to cluster (Figure 10). Such gaps were suspected in pre-*HST* work, but the much larger HB populations accessible from the cluster cores have reinforced their identification. Contemporary BHB stellar models indicate that one of these gaps falls consistently at a mass of 0.535 $M_\odot$ regardless of metallicity. We still await convincing evolutionary interpretations of this puzzle.

Blue stragglers—which lie nearly along the main sequence above the cluster turnoff point, and can be the result of either stellar collisions or amalgamation of close binaries—were already well known types of cluster stars from ground-based photometry, but have turned up in sometimes-astonishing numbers in the cores (e.g. Guhathakurta et al. 1998; Ferraro et al. 1999). In the extreme case of M80, the number of blue stragglers is similar to the number of HB stars in the core; the sheer size of the population, its lengthy extension up the main sequence, and the high concentration of the M80 core suggest that stellar collisions may be the dominant channel of formation there. When we go on to add the presence of other core phenomena such as millisecond pulsars, novae, dwarf novae, and other peculiar stars (Shara et al. 1996a,b; Edmonds et al. 1997; Gilliland et al. 1998; Lyne 1995; Lyne et al. 1996; Robinson et al. 1995; Phinney 1993; Hut et al. 1992; and references cited), it is clear that we still have much to learn about how stars in crowded and dynamically evolved environments can strongly affect each other in individual detail.

Lastly, the cores of selected nearby clusters can be employed as the basis for surveys and time-series studies of new kinds. In a recent Major Program, Gilliland et al. (2000) are analyzing 636 sequenced exposures of 47 Tuc containing more than 30,000 stars. The primary aim is to search for the eclipse signatures of Jovian planets around the lower main sequence stars, but their database will almost certainly yield a rich harvest of many kinds of variables and eclipsing binaries.

4. Globular clusters in local group dwarf galaxies

Old star clusters are to be found in virtually all large galaxies. But exactly how old? Are the ones in other galaxies to be considered near-twins of the classic Milky Way halo clusters, or merely close cousins? The implications for galaxy formation studies are obvious: did the small galaxies that are still isolated dwarfs *today* (as opposed to ones that might have been absorbed at early times by larger protogalaxies) begin their own star formation at the same epoch as did the larger galaxies?

Once again, *HST* imaging programs have completely transformed this subject, bringing the stellar content of all Local Group galaxies clearly within reach. There are four satellite galaxies (Sagittarius, Fornax, and the Magellanic Clouds) which have old globular clusters *and* which are within $\lesssim 200$ kpc, close enough that rigorous differential age comparisons can be made with the Milky Way clusters from main-sequence fitting. In both Sagittarius and Fornax, a handful of clusters have each been identified as belonging to the small host galaxy. All are "old" in a generic sense, but only some recent painstaking photometry has been able to circumvent the combined problems of distance, heavy field crowding, and contamination, to produce precise age measurements.

In Fornax, the CMDs (Buonanno et al. 1998, 1999) establish that all but one of its clusters are identical in age with the low-metallicity template M92 to within a Gyr or so. The remaining one (cluster 4) may be up to 3 Gyr younger. The total history of star formation within Fornax is a much more extended one, with identifiable field stars as young as < 0.5 Gyr and the majority of stars belonging to a few-Gyr population (e.g. Buonanno et al. 1999; Saviane et al. 2000). The very old halo which the clusters represent is thus only a small fraction of this tiny galaxy as a whole. A somewhat similar history seems to have happened within Sagittarius (Ibata et al. 1995, 1997). Three clusters (M54, Arp 2, Ter 8) are classically metal-poor and old at $\gtrsim 13$ Gyr, while the fourth (Ter 7, also the most metal-rich) is about 8 Gyr old; meanwhile, the stellar component of Sagittarius is mostly near ~ 10 Gyr but with traces extending down to less than 1 Gyr (see Layden & Sarajedini 2000 for a complete discussion with references). In both Fornax and Sagittarius, it is intriguing that *very* few field-halo stars seem to have remained from the first early burst that created the old clusters, with most of the field-star formation belonging to bursts a few Gyr later—yet it was only the first burst which formed massive clusters.

The Large and Small Magellanic Clouds hold a rich variety of star clusters over all ages and sizes, and we can do nothing more here than touch on results for the few oldest ones. Differential CMD fitting of the oldest LMC clusters with the usual Milky Way templates such as M92 and M3 has revealed several with clearly similar CMDs and ages in the standard 12–14 Gyr range (see particularly Johnson et al. 1999 and Olsen et al. 1998; also Brocato et al. 1996; Mighell et al. 1996; and Bica et al. 1998). The SMC by contrast may have no clusters quite this old; its most prominent candidate, NGC 121, is a rich cluster with RR Lyraes but is probably 2 Gyr younger than Milky Way halo clusters (Shara et al. 1998). Both the SMC and LMC have several massive clusters in the age range of 5 to 10 Gyr, a regime not well represented at all in the Milky Way and thus quite useful for comparisons of moderately metal-poor isochrone models (e.g. Mighell et al. 1998; Sarajedini 1998).

A useful parametric comparison (Johnson et al. 1999; Olsen et al. 1998) in addition to the CMD fitting is the standard graph of horizontal-branch morphology vs. metallicity

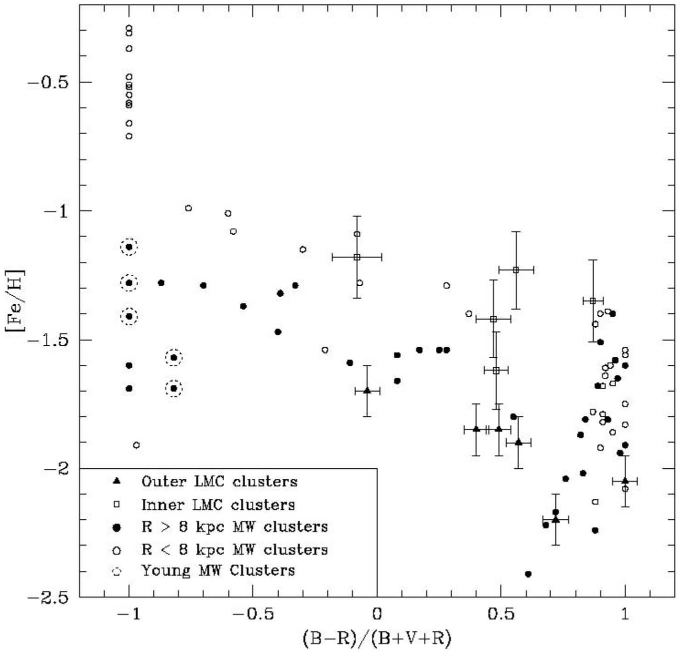


FIGURE 11. Globular cluster metallicity [Fe/H] plotted against horizontal-branch type, from Johnson et al. (1999). Clusters with red HBs are on the left side, blue HBs on the right. Old LMC clusters are the symbols with error bars; the others are Milky Way clusters in the inner halo (solid dots) or outer halo (open dots).

(Figure 11). Here again, close similarities hold between the LMC and Milky Way halo, and are consistent with an interpretation which has the LMC beginning its first star-forming period within a Gyr of the Milky Way and its other dwarf satellites (cf. the papers cited above for more extensive discussion).

The analyses summarized above point to an evolutionary picture in which most or all of the systems in and around the Milky Way *began cluster formation almost simultaneously* about 13 to 14 Gyr ago. The Milky Way and its satellites now occupy a region of space ∼ 300 kpc across, though its comoving volume would have been much smaller at those early times. Were those first star-forming events forcibly “synchronized” by local events perhaps involving gas cloud collisions and shocks? Or is it much more natural to assume that cluster formation within dwarf-sized gas clouds would simply have begun everywhere in the universe not long after the epoch of recombination, so that they would automatically be in step with one another to within a typical ∼ 1 Gyr scatter? If the latter view is true, then the maximum age of our local globular clusters should indeed place a very stringent limit on the true age of the universe.

5. M31 and beyond

More than half the classically old globular clusters in the Local Group reside in the ∼ 750−kpc distant M31, a significantly bigger galaxy than the Milky Way. From ground-based indicators—luminosities, integrated colors and spectra, and even attempts at color-magnitude diagrams—the M31 halo clusters by and large resemble our own. The definitive tests of their stellar content and evolutionary history, however, lie in obtaining deep CMDs. Even with *HST*, it has not been possible (yet!) to reach the unevolved main

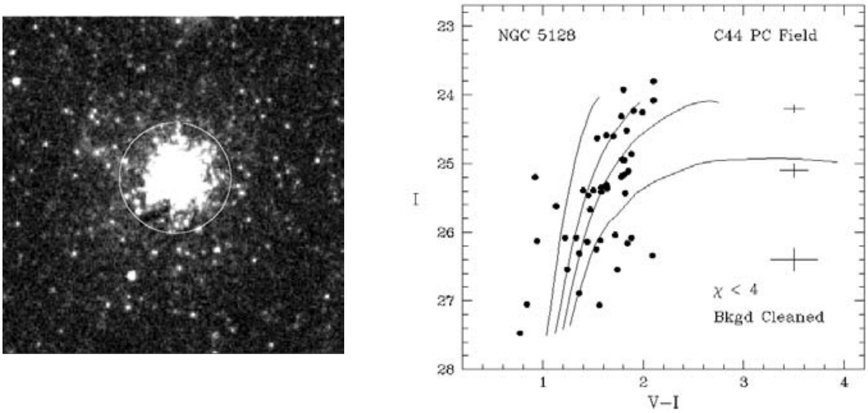


FIGURE 12. *Left panel*: WFPC2 Planetary Camera image of globular cluster C44 in the halo of NGC 5128, from G. Harris et al. (1998). The box shown is $9''$ across. *Right panel*: Observed color-magnitude diagram for N5128-C44. Out of more than 100,000 stars in the cluster, only $\simeq 40$ are uncrowded enough and luminous enough for photometric measurement at *HST* resolution. The four lines are standard giant-branch sequences for the Milky Way clusters M15, NGC 1851, 47 Tuc, and NGC 6553. C44 appears to be of moderately low metallicity.

sequences of these clusters, but photometry available so far (Rich et al. 1996; Fusi Pecci et al. 1996; Holland et al. 1997) with limiting magnitudes $M_V \sim +2$ shows that their giant-branch and horizontal-branch morphologies exhibit the "standard" (Milky Way) correlation with metallicity and thus, presumably, age. Deeper and more precise photometry will be possible in the coming Cycles.

The third largest Local Group galaxy, M33, has perhaps two dozen halo clusters which have been classified as conventionally "old" on the basis of integrated colors and spectra. However, a first CMD reconnaissance of some of these with *HST* (Sarajedini et al. 1998) reveals an intriguing range of RGB and HB morphologies suggestive of an age range as large as 5 Gyr or more. The M33 halo seems to have been remarkably slow to get started, particularly in comparison with the far smaller dwarf ellipticals.

As distance increases, information fades. The most remote cluster for which a direct CMD has been obtained is one in the outer halo of the giant elliptical NGC 5128 (G. Harris et al. 1998; see Figure 12). At its distance of 4 Mpc from the Milky Way, 99.95% of this cluster's stars are hopelessly crowded or too faint for individual measurement even at the best *HST* camera resolution. However, the few dozen red giants that are measurable roughly delineate an RGB locus with normal characteristics. This study can best be viewed as a trial run for much better things to come, but it gives the first *direct*, star-by-star evidence that the halo clusters in a giant elliptical galaxy are at least roughly similar to those in the very different Local Group members. In coming *HST* Cycles with newer cameras, it should be possible to explore the NGC 5128 halo stars and clusters down to the horizontal-branch level.

6. Young globular clusters: Starbursts and mergers

Sometimes the things right in front of us are the hardest to recognize for what they are. Our own Galaxy presents such a clear-cut division between star clusters in the halo (old, massive, mostly metal-poor) and the disk (young, small, and near-Solar metallicity) that for decades it was conventional to think of them as fundamentally different objects. The separate subfields that grew up around open clusters and globular clusters rarely

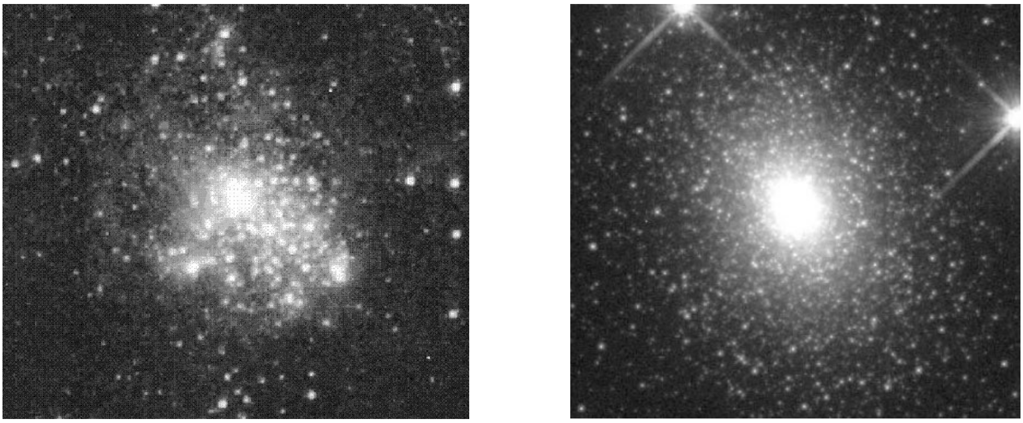


FIGURE 13. *Left panel*: The 4-Myr-old cluster NGC 2403-II (Drissen et al. 1999): a “young globular cluster.” *Right panel*: cluster Mayall II = G1, the brightest old-halo star cluster in M31 (from Fig. 1): an “ancient super star cluster.”

communicated with one another, and astronomers studying globular clusters connected even less with those studying gaseous nebulae and star formation. Questions regarding globular cluster formation tended to be addressed only through speculative theoretical models which invoked special, early-universe conditions.

Today, little remains of that artificial and restrictive paradigm. One of the truly major accomplishments of the *HST* imaging programs has been, at long last, to open up the rich range of star cluster properties in all galaxies, and to expose our Milky-Way-based prejudices for what they were. In the 3-space of cluster age τ, metallicity [Fe/H], and mass M, it is possible to find clusters from *some* galaxy that occupy every corner of that coordinate space.

Most astronomers in this field would now agree that it has become routinely easy to identify objects which fully qualify as “young globular clusters”; that is, star clusters in the mass range $\sim 10^4$–10^6 $M_\odot$ which differ only in age from the familiar old-halo globulars. The most obvious and best-studied example—in fact, the clear prototype for such objects—is likely to be the central cluster R136 in the 30 Doradus complex. Although this extraordinarily compact and luminous system has long been known, the spatial resolution of *HST* was needed to demonstrate beyond doubt that it was indeed a young star cluster. Its CMD (Hunter et al. 1995; Massey & Hunter 1998) reveals as many as ~ 60 O3-type stars, the biggest of which may reach 120 $M_\odot$. The cluster as a whole appears to have a normal IMF at least down to 3 $M_\odot$; accounting for fainter stars, it may have a total mass $\sim 4 \times 10^4$ $M_\odot$. In structure and size it easily qualifies as a small globular cluster seen at an age of just 4 Myr (Hunter et al. 1996). The vastly larger 30 Doradus gas complex surrounding R136 is highly ionized and stirred up by the outpouring energy from this central engine. We see here a clear suggestion that a *large* reservoir of gas, a tiny fraction of which gets compressed into a small volume a few parsecs across, is needed to form a globular cluster. Only a small part of the total gas in the 30 Dor region has gone into this single massive star cluster, though the whole complex also contains other much smaller clusters (Walborn et al. 1999).

Within the Milky Way, a smaller analog may be found in NGC 3603, a similarly young cluster of a few thousand $M_\odot$ still surrounded by the tangled nebula from which it emerged (Brandner et al. 2000). Objects like R136 and NGC 3603, but even more massive, were tentatively identified long ago from ground-based imaging of starburst galaxies such as NGC 1569, NGC 1705, or M82 (see O'Connell et al. 1994, 1995; De Marchi et al. 1997 for *HST* data). The term "super star clusters" was coined to refer to these young objects (van den Bergh 1971), though the connection with witnessing the actual process of globular cluster formation (an obvious one, with the usual benefit of hindsight) was not firmly made till much later. It now seems plain that globular clusters and super star clusters are one and the same (Figure 13), and that if we want to understand how globular clusters formed, a highly informative approach is to scrutinize the gas clouds within which massive clusters are forming today.

The *HST* cameras have made it possible to investigate these intriguing objects in considerable detail, and to reveal many more sites where ultra-recent star and cluster formation is happening on large scales. Such events are taking place not only in small starburst irregulars, but in many other circumstances where *large amounts of gas* have been collected together and compressed into $\gtrsim 10^8$ $M_{\odot}$-sized GMCs. They appear in very gas-rich single disk galaxies such as NGC 2403 (Drissen et al. 1999); disk/disk mergers such as NGC 3256 (Zepf et al. 1999) and the exhaustively studied Antennae (Whitmore et al. 1999; see also Whitmore's review in this volume); or giant ellipticals within which massive HI gas infall is taking place, such as the Perseus cD NGC 1275 (Holtzman et al. 1992; Carlson et al. 1998).

With amazing rapidity, examples have been found which penetrate close to the ultimate age limit $\tau \rightarrow 0$. For example, in the starburst dwarf NGC 5253, Calzetti et al. (1997) isolated half a dozen young globulars likely to be all less than 10 Myr old; the youngest, in the centermost starburst region, may be ~ 1 Myr old and is apparently still heavily shrouded in dust and ionized gas (Turner et al. 2000). This object, which the authors claim to hold several *thousand* O stars, may still be in transition from its "protoglobular" gaseous state. A similar situation is found in another starburst dwarf NGC 2366 (Drissen et al. 2000), where at least one cluster (N2363-A) may be $\lesssim 1$ Myr old. It is natural to speculate that active dwarf galaxies like this might closely resemble the protogalactic fragments ($\equiv$ supergiant molecular clouds) out of which larger galaxies are postulated to have assembled (Searle & Zinn 1978; Harris & Pudritz 1994).

This rewarding lode of new data has been directly responsible for stimulating new and more empirically based models for cluster formation (e.g. McLaughlin & Pudritz 1996; Elmegreen & Efremov 1997). Much has been found out about the luminosity distribution of young clusters, which follows a rough power-law form $n(L)dL \sim L^{-\alpha}dL$ strongly resembling the mass spectrum of giant molecular clouds (Harris & Pudritz 1994). The exponent α is near 2 independent of environment, but may progressively steepen from ~ 1.7 at low masses to $\gtrsim 2.3$ at the high-mass end (see Whitmore et al. 1999). Both the mean value of α and its gradual steepening can be understood quantitatively within a model framework whereby protocluster gas clouds build up by collisional accretion (McLaughlin & Pudritz 1996). However, much remains to be understood about the later evolution of the mass distribution function due to progressive dynamical destruction of clusters at various masses, and the effects of mergers and accretion of satellites (see, e.g. Zhang & Fall 1999; Murali & Weinberg 1997; Gnedin & Ostriker 1997; Vesperini 1998; Vesperini & Heggie 1997; Côté et al. 2000 for more extensive discussion of these issues).

7. Giant ellipticals: Another direction for galaxy archaeology

Even the most thorough analyses of the Milky Way and other Local Group galaxies do not exhaust what there is to learn about the classic old-halo star clusters. Vastly more populous globular cluster systems reside in giant E galaxies; and their clusters display a surprising range of characteristics far beyond what we see in disk galaxies or dwarfs

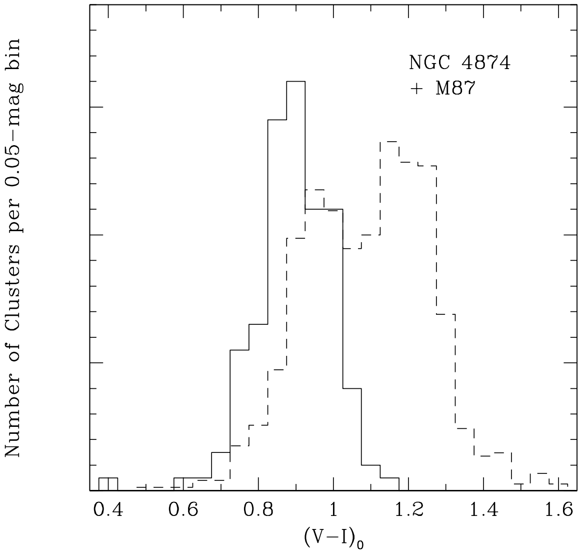


FIGURE 14. The metallicity distribution functions for two cD galaxies, M87 in Virgo (dashed line, from Kundu et al. 1999), and NGC 4874 in Coma (solid line, from Harris et al. 2000). M87 has the classic bimodal form discovered in many large spirals and ellipticals, while puzzlingly NGC 4874 is missing the metal-richer component.

(for extensive reviews, see Harris 1991, 1999, 2000; Ashman & Zepf 1998). But almost all gEs are at distances which make them a stiff challenge for ground-based photometry or spectroscopy.

While the ground-based 8- and 10-meter-class telescopes have enabled a new round of spectroscopic studies of globular cluster systems (radial velocities, chemical compositions, dynamics), *HST* has exerted deep influences once again through its imaging capabilities. Any shortlist of highlights would surely include the following:

(a) Two extremely deep photometric studies of the Virgo giant M87 (Whitmore et al. 1995; Kundu et al. 1999) showed definitively that the luminosity function of its globular cluster system (GCLF) has the same log-normal form (number of clusters per unit magnitude) expected from classic studies of the Local Group members. Coupled with considerable previous evidence that the bright end of the GCLF (i.e. above the turnover point at $M_V \simeq -7.5$) is very much the same in all large galaxies, this work added much weight to the view that the GCLF is a near-universal phenomenon.

(b) Many ellipticals in the Virgo/Fornax distance regime and slightly beyond have now been imaged deeply enough to construct populous and nearly contamination-free samples of globular clusters. The metallicity distribution function (MDF), another informative relic of the galaxy's early history, can be assembled from the two-color photometry (to this stage, usually $(V-I)$ from WFPC2). Notably comprehensive and precise studies of M87 and NGC 4472 in Virgo have been published (Puzia et al. 1999; Kundu et al. 1999), while statistical analyses of many more E and S0 galaxies have been carried out by Neilsen & Tsvetanov (1999), Gerhard & Kissler-Patig (1999), and Kundu & Whitmore (1999). These galaxies differ surprisingly in the form of their MDFs, but half or more show distinct bimodality indicative of two major stages of cluster formation (Figure 14). These studies have helped fuel a highly active and progressive debate on the relative merits of *in situ* formation, satellite accretion, or major mergers (e.g. Côté et al. 2000;

Forbes et al. 1997; Harris et al. 1998, 2000; Schweizer et al. 1996; Zepf et al. 1999; Sharples et al. 1998, among many others).

(c) For the first time, the globular cluster systems in the 100-Mpc-distant Coma cluster have come within easy reach—an important step since Coma represents a much richer and denser environment of galaxies than anything at closer distances, and thus probably also a different evolutionary history. A series of studies of the Coma giants (Baum et al. 1995, 1997; Woodworth & Harris 2000; Kavelaars et al. 2000; Harris et al. 2000) is beginning to reveal combinations of characteristics that were not seen in any previous cases. The GCLFs in these Coma ellipticals have been studied down to their turnover points, allowing them to be used for distance determination and a new estimate of the Hubble constant (Kavelaars et al. 2000). Under the restricted assumption that the GCLFs in *giant ellipticals* are similar everywhere (and in particular, between Virgo and Coma), the current results suggest $H_0 = 69 \pm 9$, a result highly consistent with recent determinations through several other standard candles.

8. What lies ahead?

From our vantage point well into the *HST* era, some old questions are still on the table, and many new ones have been put there. Some of these may find answers as soon as the next Cycle of *HST* programs is underway, or as far downstream as the era of NGST:

- What does the faintest part of the white dwarf cooling sequence look like, and how will it help us constrain both white dwarf physics and cluster ages?
- What are the interactive processes which produce the many anomalous types of stars residing in the central regions of core-collapsed clusters?
- What is the "missing" stellar physics which governs mass loss and the mass distribution for cluster stars in the giant and horizontal-branch stages?
- Do any globular clusters have central black holes? If so, how did they arise? What other types of objects reside in the innermost centers of core-collapsed clusters?
- What kinds of planets reside around globular cluster stars, especially in low-metallicity clusters?
- What are the ages of the Galactic bulge clusters?
- Typically 1% of the gas in a star-forming GMC finds its way into a few condensed star clusters. What governs this mass fraction? Does it differ strongly under different conditions? How does a protocluster cloud make the transition into a bound star cluster over the space of ~ 1 Myr? How does this depend on mass and composition?
- Is the mass distribution function at formation a near-universal phenomenon for globular clusters, regardless of host galaxy type?
- What drives the perplexing range of metallicity distributions and specific frequencies in giant E galaxies? Can we "read" these quantities to deduce what fraction of such a galaxy was assembled from satellite accretions or gas-rich mergers?

It is not hard to predict that globular clusters will continue to provide superb laboratories for stellar physics, as well as testbeds for modeling the early evolution of galaxies.

REFERENCES

Arp, H. C. Baum., W. A., & Sandage, A. R. 1953 *AJ* **58**, 4.

Ashman, K. M. & Zepf, S. E. 1998, *Globular Cluster Systems*. Cambridge University Press.

Baum, W. A., et al. 1995 *AJ* **110**, 2537.

Baum, W. A., et al. 1997 *AJ* **113**, 1483.

Brocato, E., Castellani, V., Ferraro, F. R., Piersimoni, A. M., & Testa, V. 1996 *MNRAS* **282**, 614.

Bica, E., Geisler, D., Dottori, H., Clariá, J. J., Piatti, A. E., & Santos, J. F. C. Jr. 1998 *AJ* **116**, 723.

Brandner, W., et al. 2000 *AJ* **119**, 292.

Buonanno, R., Corsi, C. E., Bellazini, M., Ferraro, F. R., & Fusi Pecci, F. 1997, AJ, 113, 706

Buonanno, R., Corsi, C. E., Castellani, M., Marconi, G., Fusi Pecci, F., & Zinn, R. 1999 *AJ* **118**, 1671.

Buonanno, R., Corsi, C. E., Fusi Pecci, F., Fahlman, G. G., & Richer, H. B. 1994 *ApJ* **430**, L121.

Buonanno, R., Corsi, C. E., Zinn, R., Fusi Pecci, F., Hardy, E., & Suntzeff, N. 1998 *ApJ* **501**, L33.

Calzetti, D., et al. 1997 *AJ* **114**, 1834.

Capaccioli, M., Ortolani, S., & Piotto, G. 1991 *AAp* **244**, 298.

Capaccioli, M., Piotto, G., & Stiavelli, M. 1993 *MNRAS* **261**, 819.

Carlson, M. N., et al. 1998 *AJ* **115**, 1778.

Carretta, E., Gratton, R. G., Clementini, G., & Fusi Pecci, F. 2000 *ApJ* **533**, 215.

Cassisi, S., Castellani, V., Ciarcelluti, P., Piotto, G., & Zoccali, M. 2000 *MNRAS*, **315**, 679.

Chaboyer, B., Demarque, P., Kernan, P. J., & Krauss, L. M. 1998 *ApJ* **494**, 96.

Chaboyer, B., Demarque, P., & Sarajedini, A. 1996 *ApJ* **459**, 558.

Cool, A. M., Piotto, G., & King, I. R. 1996 *ApJ* **468**, 655.

Côté, P., Marzke, R. O., West, M. J., & Minniti, D. 2000 *ApJ* **533**, 869.

De Marchi, G., Clampin, M., Greggio, L., Leitherer, C., & Nota, A. 1997 *ApJ* **479**, L27.

De Marchi, G. & Paresce, F. 1997 *ApJ* **476**, L19.

De Marchi, G., Paresce, F., & Pulone, L. 2000 *ApJ* **530**, 342.

Djorgovski, S., Piotto, G., & Capaccioli, M. 1993 *AJ* **105**, 2148.

Drissen, L., Roy, J.-R., Moffat, A. F. J., & Shara, M. M. 1999 *AJ*, **117**, 1249.

Drissen, L., Roy, J.-R., Robert, C., Devost, D., & Doyon, R. 2000 *AJ* **119**, 688.

Edmonds, P. D., et al. 1996 *ApJ* **468**, 241.

Edmonds, P. D., et al. 1997 *Bull.AAS* **191**, 44.15.

Eggen, O. J., Lynden-Bell, D., & Sandage, A. R. 1962 *ApJ* **136**, 735 (ELS).

Elmegreen, B. G. & Efremov, Y. N. 1997 *ApJ* **480**, 235.

Ferraro, F. R., Carretta, E., Bragaglia, A., Renzini, A., & Ortolani, S. 1997a *MNRAS* **286**, 1012.

Ferraro, I., Ferraro, F. R., Fusi Peccci, F., Corsi, C. E., & Buonanno, R. 1995 *MNRAS* **275**, 1057.

Ferraro, F. R., Paltrinieri, B., Fusi Pecci, F., Rood, R. T., & Dorman, B. 1998 *ApJ* **500**, 311.

Ferraro, F. R., Paltrinieri, B., Rood, R. T., & Dorman, B. 1999 *ApJ* **522**, 983.

Ferraro, F. R., et al. 1997b *AAp* **320**, 757.

Forbes, D. A., Brodie, J. P., & Grillmair, C. J. 1997 *AJ* **113**, 1652.

Fullton, L. K., et al. 1995 *AJ* **110**, 652.

Gebhardt, K. & Kissler-Patig, M. 1999 *AJ* **118**, 1526.

Gilliland, R. L., et al. 1998 *ApJ* **507**, 818.

Gilliland, R. L., et al. 2000 *Bull.AAS* **196**, 02.02.

Gnedin, O. Y. & Ostriker, J. P. 1997 *ApJ* **474**, 223.

Guhathakurta, P., Guhathakurta, P., Schneider, D. P., & Bahcall, J. N. 1996, AJ, 111, 267

Hansen, B. M. S. 1999 *ApJ* **520**, 680.

Harris, G. L. H., Poole, G. B., & Harris, W. E. 1998 *AJ* **118**, 2866.

Harris, W. E. 1991 *ARAA* **29**, 543.

Harris, W. E. 1999. In *Globular Clusters, Tenth Canary Islands Winter School* (eds. C. Martinez Roger, I. Perez Fournon, & F. Sanchez), p. 325. Cambridge University Press.

Harris, W. E. 2001. In *Star Clusters, 28th Saas-Fee Advanced Course for Astrophysics and Astronomy*. p.223. Springer-Verlag.

Harris, W. E., Kavelaars, J. J., Hanes, D. A., Hesser, J. E., & Pritchet, C. J. 2000 *ApJ* **533**, 137.

Harris, W. E., et al. 1997 *AJ* **114**, 1030.

Harris, W. E., Harris, G. L. H., & McLaughlin, D. E. 1998 *AJ* **115**, 1801.

Harris, W. E. & Pudritz, R. E. 1994 *ApJ* **429**, 177.

Holtzman, J. A., et al. 1992 *AJ* **103**, 691.

Hunter, D. A., O'Neil, E. J. Jr., Lynds, R., Shaya, E. J., Groth, E. J., & Holtzman, J. A. 1996 *ApJ* **459**, L27.

Hunter, D. A., Shaya, E. J., Holtzman, J. A., Light, R. M., O'Neil, E. J. Jr., & Lynds, R. 1995 *ApJ* **448**, 179.

Hut, P., et al. 1992 *PASP* **104**, 981.

Ibata, R. A., Gilmore, G., & Irwin, M. J. 1995 *MNRAS* **277**, 781.

Ibata, R. A., Wyse, R. F. G., Gilmore, G., Irwin, M. J., & Suntzeff, N. B. 1997 *AJ* **113**, 634.

Johnson, J. A., Bolte, M., Stetson, P. B., Hesser, J. E., & Somerville, R. S. 1999 *ApJ* **527**, 199.

Kavelaars, J. J., Harris, W. E., Hanes, D. A., Pritchet, C. J., & Hesser, J. E. 2000 *ApJ* **533**, 125.

King, I. R., Anderson, J., Cool, A. M., & Piotto, G. 1998 *ApJ* **492**, L37.

Kundu, A. & Whitmore, B. C. 1999 *Bull.AAS* **194**, 35.04.

Kundu, A., Whitmore, B. C., Sparks, W. B., Macchetto, F. D., Zepf, S. E., & Ashman, K. M. 1999 *ApJ* **513**, 733.

Layden, A. C. & Sarajedini, A. 2000 *AJ* **119**, 1760.

Lyne, A. G. 1995. In *ASP Conf. Ser. 72* (eds. A. Fruchter, M. Tavani, & D. Backer), p. 35. ASP.

Lyne, A. G., Manchester, R. N., & D'Amico, N. 1996 *ApJ* **460**, L41.

Majewski, S. R. 1999 *ApSS* **265**, 115.

Majewski, S. R., Munn, J. A., & Hawley, S. L. 1996 *ApJ* **459**, L73.

Massey, P. & Hunter, D. A. 1998 *ApJ* **493**, 595.

McLaughlin, D. E. & Pudritz, R. E. 1996 *ApJ* **457**, 578.

Mighell, K. J., Rich, R. M., Shara, M., & Fall, S. M. 1996 *AJ* **111**, 2314.

Mighell, K. J., Sarajedini, A., & French, R. S. 1998 *AJ* **116**, 2395.

Murali, C. & Weinberg, M. D. 1997 *MNRAS* **288**, 767.

Neilsen, E. H. Jr. & Tsvetanov, Z. I. 1999 *ApJ* **515**, L13.

O'Connell, R. W., Gallagher, J. S., & Hunter, D. A. 1994 *ApJ* **433**, 65.

O'Connell, R. W., Gallagher, J. S., Hunter, D. A., & Colley, W. N. 1995 *ApJ* **446**, L1.

Olsen, K. A. G., et al. 1998 *MNRAS* **300**, 665.

Ortolani, S., et al. 1995 *Nature* **377**, 701.

Ortolani, S., Barbuy, B., Bica, E., Renzini, A., Marconi, G., & Gilmozzi, R. 1999 *AAp* **350**, 840.

Paresce, F. & De Marchi, G. 2000 *ApJ* **534**, 870.

Phinney, E. S. 1993. In *ASP Conf. Ser. 50* (eds. S. G. Djorgovski & G. Meylan), p. 141. ASP.

Piotto, G. 2000, private communication.

Piotto, G., et al. 1999 *AJ* **118**, 1727.

Puzia, T. H., Kissler-Patig, M., Brodie, J. P., & Huchra, J. P. 1999 *AJ* **118**, 2734.

Renzini, A., et al. 1996 *ApJ* **465**, L23.

Rich, R. M., Mighell, K. J., Freedman, W. L., & Neill, J. D. 1996 *AJ* **111**, 768.

Rich, R. M., et al. 1997 *ApJ* **484**, L25.

Richer, H. B., et al. 1995 *ApJ* **451**, L17.

Robinson, C., Lyne, A. G., Manchester, R. N., Bailes, M., D'Amico, N., & Johnston, S. 1995 *MNRAS* **274**, 547.

Rosenberg, A., Saviane, I., Piotto, G., & Held, E. V. 1998 *AAp* **339**, 61.

Sandage, A. R. & Schwarzschild, M. 1952 *ApJ* **116**, 463.

Sarajedini, A. 1998 *AJ* **116**, 738.

Sarajedini, A., Geisler, D., Harding, P., & Schommer, R. 1998 *ApJ* **508**, L37.

Sarajedini, A. & Demarque, P. 1990 *ApJ* **365**, 219.

Saviane, I., Held, E. V., & Bertelli, G. 2000 *AAp* **355**, 56.

Schweizer, F., Miller, B. W., Whitmore, B. C., & Fall, S. M. 1996 *AJ* **112**, 1839.

Searle, L. & Zinn, R. 1978 *ApJ* **225**, 357.

Shapley, H. & Davis, H. N. 1920 *ApJ* **51**, 140.

Shara, M. M., Bergeron, L. E., Gilliland, R. L., Saha, A., & Petro, L. 1996a *ApJ* **471**, 804.

Shara, M. M., Zurek, D. R., & Rich, R. M. 1996b *ApJ* **473**, L35.

Shara, M. M., Fall, S. M., Rich, R. M., & Zurek, D. 1998 *ApJ* **508**, 570.

Sharples, R. M., et al. 1998 *AJ* **115**, 2337.

Sosin, C., et al. 1997 *ApJ* **480**, L35.

Stetson, P. B., et al. 1999 *AJ* **117**, 247.

Stetson, P. B., Hesser, J. E., Smith, G. H., VandenBerg, D. A., & Bolte, M. 1989 *AJ* **97**, 1360.

Turner, J. L., Beck, S. C., & Ho, P. T. P. 2000 *ApJ* **532**, L109.

van den Bergh, S. 1971 *AAp* **12**, 474.

VandenBerg, D. A. 1999. In *The Galactic Halo*, ASP Conf. Ser. 165 (eds. B. Gibson, T. Axelrod, & M. Putnam), p. 46. ASP.

VandenBerg, D. A., Bolte, M., & Stetson, P. B. 1990 *AJ* **100**, 445.

VandenBerg, D. A., Stetson, P. B., & Bolte, M. 1996 *ARAA* **34**, 461.

VandenBerg, D. A., Swenson, F. J., Roger, F. J., Iglesias, C. A., & Alexander, D. R. 2000 *ApJ* **532**, 430.

Vesperini, E. 1998 *MNRAS* **299**, 1019.

Vesperini, E. & Heggie, D. C. 1997 *MNRAS* **289**, 898.

von Hippel, T., Gilmore, G., Tanvir, N., Robinson, D., & Jones, D. H. P. 1996 *AJ* **112**, 192.

Walborn, N. R., Barbá, R. H., Brandner, W., Rubio, M., Grebel, E. K., & Probst, R. G. 1999 *AJ* **117**, 225.

Walker, A. 1990 *AJ* **100**, 1532.

Walker, A. 1994 *AJ* **108**, 555.

Whitmore, B. C., Sparks, W. B., Lucas, R. A., Macchetto, F. D., & Biretta, J. A. 1995 *ApJ* **454**, L73.

Whitmore, B. C., Zhang, Q., Leitherer, C., Fall, S. M., Schweizer, F., & Miller, B. W. 1999 *AJ* **118**, 1551.

Woodworth, S. C. & Harris, W. E. 2000 *AJ*, **119**, 2699.

Yanny, B., Guhathakurta, P., Bahcall, J. N., & Schneider, D. P. 1994 *AJ*, **107**, 1745.

Zepf, S. E., Ashman, K. M., English, J., Freeman, K. C., & Sharples, R. M. 1999 *AJ*, **118**, 752.

Zhang, Q. & Fall, S. M. 1999 *ApJ* **527**, L81.

Ultraviolet absorption line studies of the Galactic interstellar medium with the Goddard High Resolution Spectrograph

By BLAIR D. SAVAGE

Department of Astronomy, University of Wisconsin, Madison, WI 53706

The high spectral resolution and high signal to noise capabilities of the Goddard High Resolution Spectrograph (GHRS) have permitted very accurate measurements of the gas phase abundances and physical conditions in interstellar clouds found in the Galactic disk and low halo and of the matter in several Galactic high velocity clouds. The interstellar gas phase abundances provide important clues about the composition of dust grain mantles and cores, and about the origins of intermediate and high velocity gas in the Galactic disk and halo. The processes that circulate gas from the disk into the low halo do not destroy dust grain cores. The gas in Complex C in the direction of Mrk 290 has a metallicity of 0.089 ± 0.024 solar, which implies the accretion of low metallicity gas by the Milky Way at a rate per unit area sufficient to solve the long standing Galactic G-dwarf problem. GHRS studies of interstellar Si IV, C IV, and N V absorption toward stars and AGNs have yielded measures of the 3 to 5 kpc extension of hot gas into the halo of the Milky Way. The GHRS results coupled with new measurements from the *Far-Ultraviolet Spectroscopic Explorer* (*FUSE*) satellite of O VI absorption by hot halo gas permit a study of the physical conditions in the hot Galactic Corona originally envisioned by Lyman Spitzer in his classic 1956 paper "On a Possible Interstellar Galactic Corona." The Space Telescope Imaging Spectrograph (STIS) with its high resolution and high multiplexing efficiency promises to provide the UV spectroscopic observations required to extend the studies begun with the GHRS into denser and more distant regions of the Galactic ISM.

1. Introduction

The high resolution ultraviolet spectrographs aboard the *Hubble Space Telescope* (*HST*) have allowed astronomers to make significant progress in obtaining accurate information about elemental abundances and physical conditions in the Galactic interstellar medium (ISM). This is because most atoms and molecules have their resonance absorption lines (originating from the ground state) in the ultraviolet at wavelengths below the atmospheric cutoff near 3000 Å. Therefore, studies of the cool, warm, and hot phases of the ISM greatly benefit from access to spectroscopic facilities on orbiting observatories such as the *HST*. Before the launch of the *HST*, most UV absorption line observations of the ISM were either from the *Copernicus* satellite which operated from 1972 to 1980 or from the *International Ultraviolet Explorer* (*IUE*) satellite which operated from 1978 to 1996. Both satellites provided important information on abundances and physical conditions in interstellar clouds (see Spitzer & Jenkins 1975; Jenkins 1987; de Boer, Jura & Shull 1987) but lacked the spectral resolution for detailed studies of individual interstellar clouds and lacked the very faint object capability for probing gas in the outermost regions of the Milky Way or in very dense interstellar clouds.

In this presentation of selected examples of scientific results from the *HST* based on high resolution UV absorption line observations, I have chosen to emphasize:

(*a*) High precision measures of elemental abundances in diffuse interstellar clouds in the Galactic disk and halo.

(*b*) The implications of elemental gas phase abundance studies for the composition of dust grain cores.

(*c*) The gas phase abundances in Galactic high velocity clouds.

(*d*) Studies of the distribution and properties of hot interstellar gas extending away from the Galactic plane into the halo.

Because of space limitations there are many interesting topics in UV interstellar absorption line spectroscopy I was unable to discuss. For example, *HST* has provided important information about the distribution, kinematics and ionization state of gas in the Local ISM (for a review see Linsky & Wood 1998), while properties of the small scale structure of the ISM have been probed by Lauroesch et al. (1998). The ionization conditions and the presence of dust in the warm ionized medium have been studied by Howk & Savage (1999). *HST* ISM studies have also yielded important information about isotopic abundances. Studies of D/H in the LISM are important for an ultimate understanding the effects of astration following the production of D in the big bang (see Linksy et al. 1995). Measures of the isotopes of C and O in CO have provided insights about interstellar molecular fractionation (see Lambert et al. 1994).

Several general reviews of *HST* ISM absorption line results in the literature include Cardelli (1994), Savage & Sembach (1996a), Sembach (1998), and Meyer (1999). A recent discussion of the implications of *HST* abundance studies for the nature of interstellar dust and the intrinsic composition of the ISM is found in Mathis (1999).

2. Diagnostic power of ultraviolet absorption line spectroscopy

Optical absorption line observations of interstellar gas from ground based telescopes are limited to several molecules and ions from elements of relatively low cosmic abundance. In contrast, Table 1 provides a list of the very large number of atomic and molecular species detected with the *HST* through absorption line studies extending over the 1150 to 3200 Å region. The UV window allows the study of absorption by abundant atoms such as C, N, O, Mg, Si, and Fe in a number of ionization states, including those found in cool neutral gas (C I, C II, N I, O I, etc.) and those found in the hot interstellar medium (C IV and N V). Observations of adjacent ionization states such as: C I-II; Mg I-II; Si I-II-III-IV; S I-II-III; and P I-II-III are valuable for determining physical and ionization conditions in the ISM.

Studies of rare isotopes (i.e. D) and elements of low cosmic abundance (B, Ga, Ge, As, Se, Kr, Pb, Sn, Te, Tl, and Pb) are possible at UV wavelengths. Understanding the gas phase abundances of these species offers the possibility of gaining information about primordial nucleosynthesis and subsequent abstration processes (Linsky et al. 1995) and about the enrichment of the interstellar gas with heavy elements created through slow and rapid neutron capture processes (Cardelli et al. 1993).

The UV also enables sensitive searches for a number of interstellar molecules that provide information about interstellar chemical processes. Examples of molecules with lines in the accessible to the *HST* include: CO, CO^+, C_2, CN^+, CS, CH_2, N_2, NO, NO^+, OH, H_2O, MgH^+, SiO, and HCl. Molecules detected with the *HST* are listed at the bottom of Table 1. Studies of abundant molecules such as CO are valuable in probing interstellar isotopic abundances, the role of chemical fractionation, and differences in photodestruction rates. The electronic transitions for the most abundant interstellar molecule, H_2 (in the ground $v'' = 0$ vibration level), occur at wavelengths $\lambda < 1110$ Å which are inaccessible to the *HST*. However, the *Far Ultraviolet Spectroscopic Explorer* (*FUSE*) satellite is now providing information about interstellar H_2 on a regular basis (Shull et al. 2000).

Atoms[a]	Z[b]	IP(eV)[c]	IP(eV)[c]
(1150 < λ < 3200 Å)		(I to II)	(II to III)
H I	1	13.60	...
D I	1	13.60	...
B II	5	8.30	25.15
C I, C I*, C I**, **C II**, C II*, C IV	6	11.26	24.38
N I, N V	7	14.53	29.60
O I, O I*	8	13.62	35.12
Mg I, **Mg II**	12	7.65	15.04
Al II, Al III	13	5.99	18.83
Si I, **Si II**, Si II*, Si III, Si IV	14	8.15	16.35
P I, **P II**, P III	15	10.49	19.73
S I, **S II**, S III	16	10.36	23.33
Cl I	17	12.97	23.81
Cr II	24	6.77	16.50
Mn II	25	7.44	15.64
Fe II	26	7.87	16.18
Co II	27	7.86	17.06
Ni II	28	7.64	18.17
Cu II	29	7.73	20.29
Zn II	30	9.39	17.96
Ga II	31	6.00	20.51
Ge II	32	7.90	15.93
As II	33	9.81	18.63
Se II	34	9.75	21.19
Kr I	36	14.00	24.36
Sn II	50	7.34	14.63
Tl II	81	6.11	20.43
Pb II	82	7.42	15.03
Molecules			
H_2(v = 3), OH, ^{12}CO, ^{13}CO, C^{17}O, C^{18}O, C_2, HCl			

[a] The dominant ions found in neutral hydrogen regions are boldfaced. Very little ionizing radiation with E > 13.6 eV exists in H I regions and the dominant ions are determined by whether the first ionization potential IP(I to II) is less than or greater than 13.6 eV. An exception is chlorine for which Cl I often is the dominant ion in regions containing H I and H_2 because exchange reactions involving H_2 are responsible for establishing the ionization equilibrium.

[b] Atomic number.

[c] First and second ionization potentials in eV from Moore (1970) are listed.

Table 1. Species detected in the ISM with the GHRS

3. Goddard High Resolution Spectrograph and the Space Telescope Imaging Spectrograph

The GHRS was the primary first generation instrument aboard the *HST* for absorption line studies of the Galactic interstellar gas at UV wavelengths from 1150 to 3200 Å (Brandt et al. 1994; Heap et al. 1995). The GHRS contained first order diffraction gratings for low ($\lambda/\Delta\lambda \approx 2000$; $\Delta v = 150$ km s^{-1}) and intermediate resolution ($\lambda/\Delta\lambda \approx 20,000$; $\Delta v = 15$ km s^{-1}) spectroscopy and an echelle grating (in combination with two cross-dispersers) for high resolution spectroscopy ($\lambda/\Delta\lambda \approx 85,000$; $\Delta v = 3.5$ km s^{-1}).

The GHRS 512 channel Digicon detectors were capable of very high signal-to-noise spectroscopy. The combination of high resolution and high signal-to-noise permitted searches for elements with low cosmic abundances and studies of very weak interstellar features, which are important for accurate abundance measurements.

The windows and photocathodes of the GHRS Digicon detectors introduced low amplitude fixed pattern noise. However, this structure could be removed reliably by obtaining multiple spectra with different detector alignments and solving for the noise pattern in the resulting data (Fitzpatrick & Spitzer 1994; Cardelli & Ebbets 1994). Through this process, Meyer et al. (1994) and Lambert et al. (1994) have obtained spectra with signal-to-noise ratios as large as 1200.

The Space Telescope Imaging Spectrograph (STIS) was installed into *HST* in 1997 February during the second *HST* servicing mission (Woodgate et al. 1998; Kimble et al. 1998). STIS was designed to replace and extend the spectroscopic capabilities of the GHRS and the Faint Object Spectrograph (FOS) which were removed during the mission. With its large format two-dimensional array detectors STIS provides a very large spectroscopic multiplexing advantage over the GHRS and FOS. Operating in its medium and high-resolution echelle modes, the STIS yields spectra with 20–35 times greater simultaneous wavelength coverage than in the corresponding GHRS modes. The STIS also has produced UV spectra with spectral resolutions approximately twice that of the GHRS. With a low dark count rate over the 1150 to 1800 Å spectral region and a demonstrated capability to produce high S/N spectra, STIS is an outstanding instrument for UV interstellar absorption line studies of Galactic and extragalactic sources. However, only few papers relating to absorption line spectroscopy of the Galactic ISM have so far appeared in the refereed literature (Walborn et al. 1998; Jenkins et al. 1998; Howk et al. 2000). This is undoubtedly the result of the approximately two year period during which observations with the Near-Infrared Camera and Multiobject Spectrometer (NICMOS) were given preference over those involving other *HST* instruments. An added complication is created by the tremendous spectroscopic multiplex advantage of STIS compared to the earlier spectrographs. The scientific information content of the STIS intermediate and high resolution spectra is enormous. Therefore, detailed ISM sight line analysis projects represent a major analysis and interpretative challenge. Since very few STIS ISM absorption line papers have appeared in the referred literature, most of the discussions to follow cover the ISM scientific results from the GHRS.

4. Abundance notation system

We adopt an notation system where N(X) refers to the total column density (atoms cm^{-2}) of species X in a H I region. N(X) will usually be reliably approximated by either N(X I) or N(X II) depending on whether the ionization potential of the neutral atom is greater or less than 13.6 eV. For hydrogen, N(H) = N(H I) + 2N(H$_2$), where N(H$_2$) = Σ N(H$_2$)$_J$ ≈ N(H$_2$)$_0$+ N(H$_2$)$_1$, since most of the molecular hydrogen in interstellar space is in two lowest (J = 0 or 1) rotational levels in diffuse clouds (Savage et al. 1977).

The gas-phase abundance of species X with respect to hydrogen is given by (X/H)$_g$ = N(X)/N(H), where the subscript "g" refers to gas. The normalized gas-phase abundance with respect to cosmic abundances is (X/H)$_g$/(X/H)$_c$, where the subscript "c" refers to cosmic. We adopt the standard logarithmic notation system used in stellar astrophysics, [X/H] = log(X/H)$_g$–log(X/H)$_c$.

In the ISM literature the linear depletion δ(X) of species X is defined as δ(X) = (X/H)$_g$/(X/H)$_c$, and the logarithmic depletion D(X) of species X is D(X) = [X/H]. The linear and logarithmic depletions are the linear and logarithmic gas-phase abundances of a species referenced to cosmic or solar abundances. If δ(X) = 1.0 element X has an interstellar gas-phase abundance equal to its cosmic abundance. An element is said to

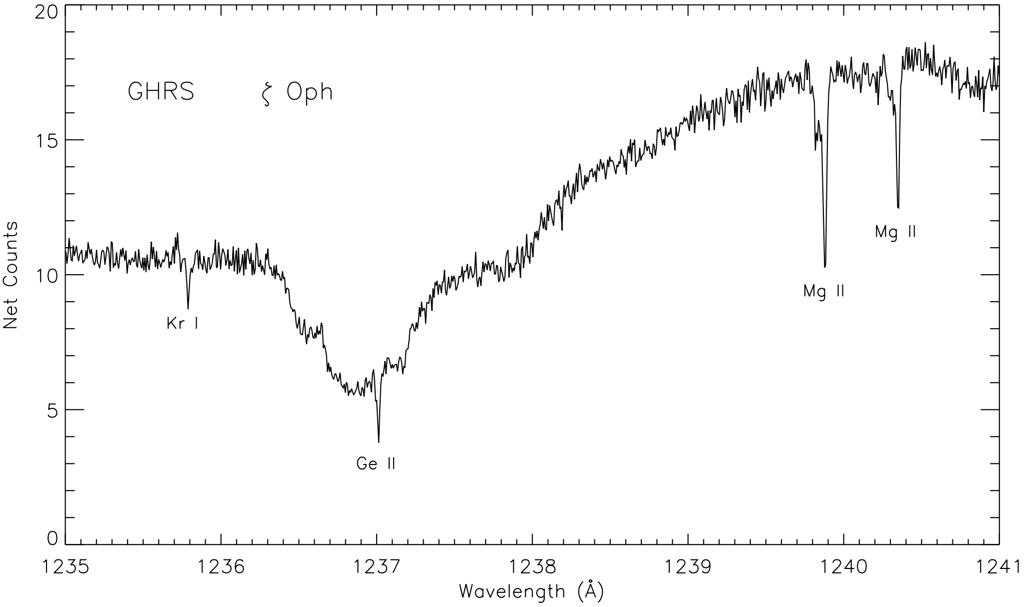


FIGURE 1. GHRS echelle mode spectrum of ζ Oph from 1235 to 1241 Å. The observed count rate is plotted against heliocentric wavelength in Å. This 6 Å region of the UV spectrum reveals interstellar absorption from Kr I λ1235.838, Ge II λ1237.059, and Mg II $\lambda\lambda$1239.925, 1240.395. The narrow ISM lines are easy to measure when observed against the smooth continuum of a rapidly rotating hot star. The GHRS spectral resolution of 3.6 km^{-1} is adequate to resolve the two velocity component structure seen in each of the interstellar lines of Mg II. An expanded view of the Mg II and Kr I and Ge II absorption along with profiles for other species is shown in Figure 2.

be depleted if $\delta(X) < 1.0$. Lightly depleted elements have $0.3 < \delta(X) < 1.0$, while highly depleted elements can have $\delta(X)$ as small as 0.001. Highly depleted elements have large "depletion factors," where the term depletion factor is $1/\delta(X)$. It is often assumed that the amount of a given species missing from the gas is contained in the interstellar dust. Therefore, we also define a linear dust-phase abundance, $(X/H)_d = (X/H)_c–(X/H)_g$. For a highly depleted element [small $(X/H)_g$] the dust-phase abundance is essentially equal to the cosmic abundance, $(X/H)_c$.

5. Gas phase abundances in diffuse clouds

Figures 1 and 2 provide examples of GHRS high resolution data illustrating the spectroscopic richness of the UV wavelength region and a few of the important capabilities of the GHRS for elemental abundance studies in the ISM. Interstellar absorption line observations are shown for the bright rapidly rotating O9.5 V star ζ Ophiuchi ($l = 6.3°$, $b = +23.6°$, $d = 140$ pc, E(B-V) = 0.32). The line of sight to ζ Oph samples gas in the local ISM, in a warm neutral cloud, in a relatively dense diffuse cloud, and in the H II regions surrounding ζ Oph. Figure 1 displays the GHRS 500 channel Digicon spectrum covering the 1235 to 1241 Å region of the spectrum. The broad features are stellar photospheric and wind absorption lines broadened by rotation and mass outflow. The various narrow lines trace ions found in the diffuse clouds along the line of sight. In this 6 Å wavelength interval absorption lines of Kr I λ1235.838, Ge II λ1237.059, and Mg II $\lambda\lambda$1239.925, 1240.395 Å are recorded. All three ions are in the dominant ionization state of each species in neutral hydrogen regions. In the case of the Mg II doublet lines, the absorption is strong enough to reveal the principal absorption cloud with a heliocentric velocity of -15 km s^{-1} along with a weaker absorbing cloud at -27 km s^{-1}. The high resolution of the GHRS permits a study of the elemental abundances and physical conditions in each cloud.

Figure 2 displays continuum normalized absorption line profiles for selected ISM absorption lines in the direction of ζ Oph plotted on a heliocentric velocity basis. The left panel shows a set of weak lines of elements that are lightly depleted (N I, O I, Cu II), moderately depleted (Mg II, Mn II), and highly depleted (Fe II, Ni II, Cr II). Note how the warm neutral cloud component at -27 km s^{-1} becomes increasingly visible as the gas phase abundance pattern changes and the warm and cool diffuse cloud line strengths become more similar. This is the result of substantially different gas phase abundances for the interior of the colder dense diffuse cloud compared to the warmer neutral medium along the line of sight. The bottom right panel shows a similar case for two stronger lines (Zn II and Fe II). The upper right panel shows a high S/N observation of the C II] intersystem line at 2325.403 Å. This weak line is valuable for determining reliable column densities of C in the ISM. The middle right panel shows several examples of weak line detections of heavy ($Z > 30$) elements toward ζ Oph.

From observations like those shown in Figures 1 and 2 it has been possible to accurately determine elemental gas phase abundances in cool diffuse clouds, in the warm neutral medium, and in warm neutral clouds found in the halo. The two best studied lines of sight are toward ζ Oph (see compilation of references in Savage & Sembach 1996a) and μ Columbae (Howk, Savage & Fabian 1999; Brandt et al. 1999). The line of sight to the O 9.5 V star μ col (l $= 237.3°$, $b = -27.1°$, $d = 400$ pc and E(B-V) $= 0.02$) mostly samples gas in the lower density warm neutral medium that fills some of the space between the denser diffuse clouds.

In Figure 3 we present the cool diffuse cloud abundance results for ζ Oph plotted in the familiar form of gas-phase abundance versus condensation temperature. The GHRS data (filled squares) are supplemented by *Copernicus* and ground-based observations (filled circles) for a few elements. In the cool cloud C, N, O, S, Ar, Kr, and some heavy elements have depletion factors of less than 3. P, Zn, and Ge have slightly larger depletion factors. Ca, Ti, V, Cr, Fe, Co, and Ni have depletion factors in excess of 100. This is the most complete set of elemental abundances available for any interstellar cloud. The depletion pattern exhibited by this cloud (with elements having larger condensation temperatures generally showing greater depletion factors) was first studied by Field (1974) who noted that the abundance deficiencies generally correlate with the condensation temperatures derived for particles condensing out of the gas-phase in cool stellar atmospheres.

Efforts to obtain reliable abundances for the abundant elements (C, N and O) in the ISM have taken advantage of the fact that each of the dominant ionization states of these elements has one or more relatively low oscillator strength (f-value) line in the wavelength region accessible to the *HST*. For example, the lines of N I, O I, and C II illustrated in Figure 2 are strong enough to be well detetected given the high S/N capabilities of the *HST* spectrographs but weak enough to lie on or near the linear part of the curve of growth thereby permitting very accurate measures of ISM column densities. Studies of the gas phase abundances of C, N and O with the *HST* toward 10 to 15 stars have recently been reviewed by Meyer (1999). The abundances of O are particularly interesting since they suggest that the intrinsic ISM composition (gas + dust) may be ~ 0.7 times solar, a result that appears consistent with measures of current epoch abundances in B stars which have recently formed from interstellar gas. Measures of the abundance of interstellar N by the GHRS imply an ISM value of ~ 0.8 times solar but uncertainties in the solar N abundance and in the f-values for the UV lines limits the significance of

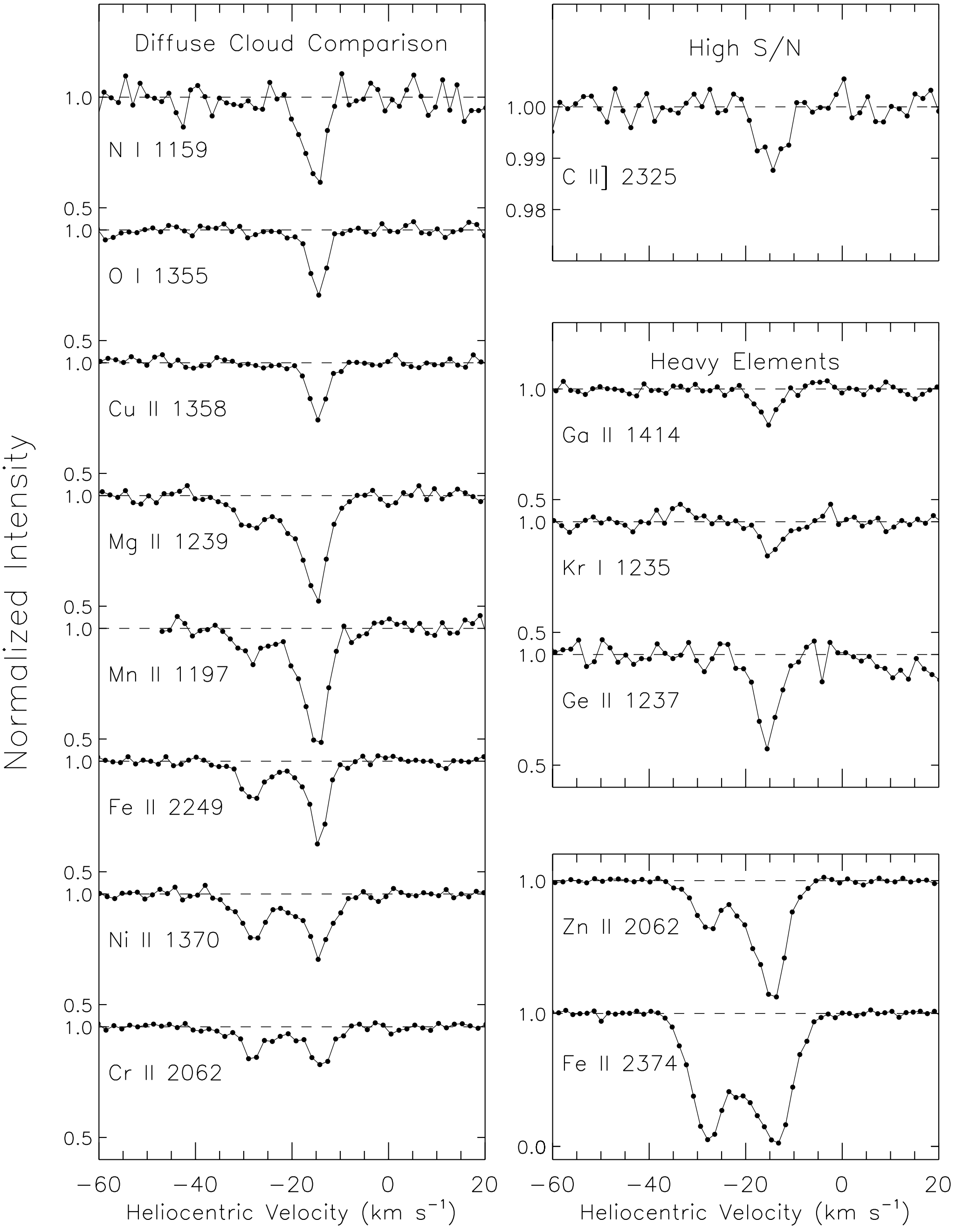


FIGURE 2. Continuum normalized profiles for selected interstellar lines in the direction of ζ Oph. The left panel shows a series of weak lines of elements that are lightly depleted (N I, O I, Cu II), moderately depleted (Mg II, Mn II), and highly depleted (Fe II, Ni II, Cr II). The bottom right panel shows two stronger lines for a lightly depleted (Zn II) and highly depleted species (Fe II). The upper right panel shows a very high S/N observation of the important C II] intersystem line at 2325.403 Å. This weak unsaturated line is valuable for determining reliable column densities of C in the ISM. The middle right panel shows several examples of weak line detections of heavy ($Z > 30$) elements toward ζ Oph. Note that the vertical scales in all panels are not identical (this figure is from Savage & Sembach 1996a).

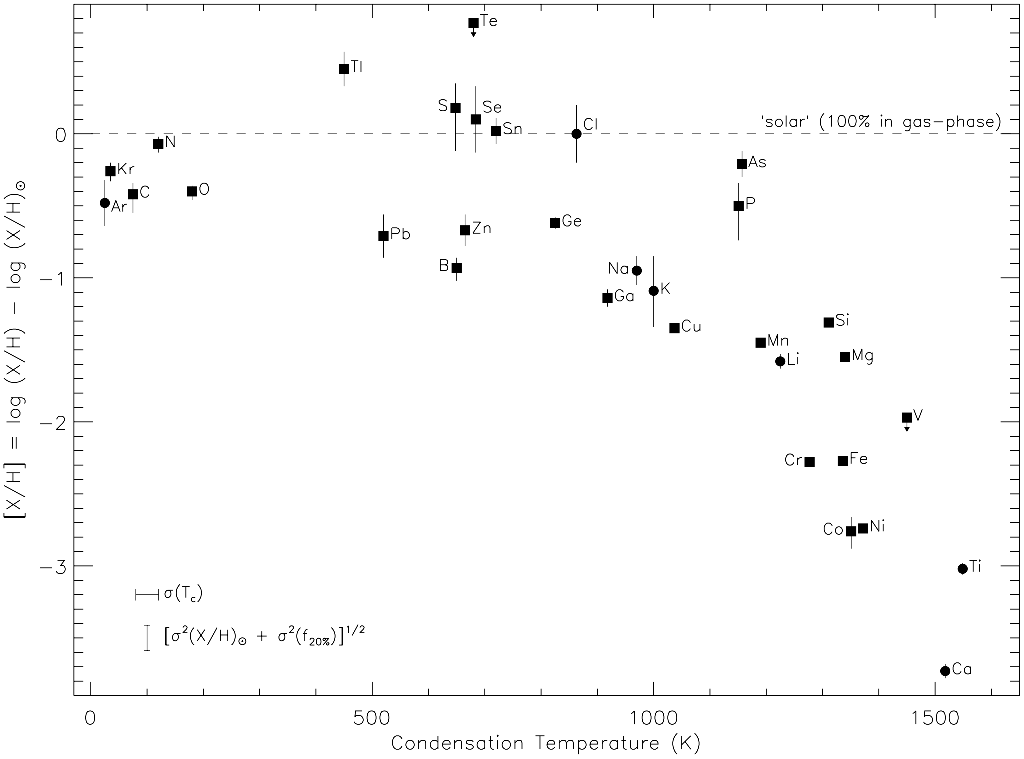


FIGURE 3. Gas-phase abundance versus condensation temperature for the cool, diffuse interstellar cloud toward ζ Oph. The condensation temperature is the temperature at which 50% of an element is expected to be removed from the gas phase due to incorporation into particulate matter. GHRS data points referenced to solar abundances are shown as filled squares. *Copernicus* and optical data points are indicated with filled circles. The error bars on all points represent measurement errors only and do not account for f-value uncertainties. The 1σ errors in condensation temperature (±20 K) and solar reference abundances combined with f-value uncertainties (±0.04 dex) are shown in the lower left corner of the plot (this figure is from Savage & Sembach 1996a).

this result. Since N is not likely to be incorporated into interstellar dust, efforts to better determine the solar N abundance and to improve the f-values of the UV lines are highly desirable. Similarly, measures of Kr through the Kr I λ1235.838 line observed toward 10 stars by Cardelli & Meyer (1997) yield an ISM gas phase Kr abundance of 0.6 times solar. Since it is a noble gas, Kr is not expected to be incorporated into interstellar dust. Unfortunately this represents another case where uncertainites in the solar abundance of Kr limit the significance of the result.

Several species yield ISM gas phase abundances close to solar. For the gas in the carefully studied warm neutral cloud toward μ Col, Howk et al. (1999) found solar gas phase abundances for Zn, P, and S. Their study included an allowance for possible ionization corrections when relating observations of Zn II, P II, and S II to H I. Similarly, observations of interstellar Sn II toward stars observed by Sofia, Meyer & Cardelli (1999) yields [Sn/H] = +0.05 for nine lines of sight where the fractional abundance of H_2 is < 0.1 and where the dust depletion effects should be small. The Sn observations are suggestive of s-process enrichment of the ISM by low-to-intermediate mass asymptotic branch stars (Sofia et al. 1999).

ISM gas phase abundances for the refractory elements are generally found to be higher in the lower density warm neutral clouds than in the cold diffuse clouds such as that toward ζ Oph. Measuring these abundance changes is important for understanding the cycling of atoms from the solid phase to the gas phase in interstellar clouds as the result of gas phase accretion processes and the destruction of grains in interstellar shocks. An overview of these changes is found in the next section which emphasizes the behavior when moving from clouds in the disk to clouds in the Galactic halo.

6. Changes in the gas phase abundances from disk clouds to halo clouds

The interstellar species suitable for halo gas abundance studies accessible by the GHRS over the wavelength region 1150 to 3200 Å are Mg II, Si II, Fe II, S II, Ni II, Cr II, Mn II, and Zn II. These ions are the dominant ionization states in neutral hydrogen regions since these elements have first ionization potentials < 13.6 eV and second ionization potentials > 13.6 eV. We assume the ionization corrections are small. The value of N(H I) is from the Lyman α absorption line, 21 cm emission measures, or estimated from N(Zn II) assuming Zn is not depleted and Zn/H is given by the cosmic reference abundance. The errors associated with these ionization assumptions are discussed by Sembach & Savage (1996) and appear to be approximately 0.05 to 0.1 dex.

Although lines of O I, C II, and N I are also observable by the GHRS, the absorption lines are either too strong and saturated or too weak and not easily detected in halo gas clouds with N(H I) typically less than 2×10^{20} atoms cm^{-2}.

An overall summary of elemental abundance results for clouds in the Galactic disk and halo is provided in Figure 4 and Table 2. Figure 4 from Sembach & Savage (1996) shows [X/Zn] for different interstellar paths as indicated by the legend on the figure. The paths include: (1) the cool disk clouds in the direction of ζ Oph and ξ Per; (2) the warm disk clouds toward HD 93521, μ Col, and ζ Oph; (3) the warm disk and warm halo clouds toward HD 18100 and HD 167756; and (4) the warm halo clouds toward HD 116852, HD 149881, HD 93521, μ Col, and the QSO 3C 273. Since Zn is generally undepleted in the warm neutral gas, it is expected that [X/Zn] $\sim$[X/H]. The sources of these various measurements are as follows: ζ Oph (Savage et al. 1992); ξ Per (Cardelli et al. 1991); HD 93521 (Spitzer & Fitzpatrick 1993); μ Col (Sofia, Savage & Cardelli 1993); HD 18100 (Savage & Sembach 1996b); HD 167756 (Cardelli et al. 1995); HD 116852 (Sembach & Savage 1996); HD 149881 (Spitzer & Fitzpatrick (1995); QSO 3C 273 (Savage et al. 1993).

The abundance trends illustrated in Figure 4 and summarized in Table 2 are informative. There is a general progression toward increasing gas phase abundances of the depleted elements from the cool disk clouds, to the warm neutral gas of the disk, to the warm neutral clouds of the halo. The variation in the gas phase abundances among the different halo clouds is small. For example, the seven data points for [Fe/H] in the warm halo cloud gas range from -0.58 to -0.69 dex (see Table 2). This small variation is quite surprising since the various clouds studied span halo gas in the solar neighborhood, under the Sagittarius spiral arm, and the Norma spiral arm. There evidently is no systematic dependence of these gas phase halo abundances on Galactocentric distance from $R_g \sim 7$ to ~ 10 kpc.

Cloud Type	Mg[a,b]	Si[b]	S[b]	Mn[b]	Cr[b]	Fe[b]	Ni[b]
Halo	(<−0.01,−0.29)	(−0.09,−0.47)	(−0.23,+0.16)	(−0.47,−0.72)	(−0.38,−0.63)	(−0.58,−0.69)	(−0.77,−0.91)
Disk+ Halo	(−0.32,−0.35)	(−0.23,−0.28)	(+0.03)	(−0.66)	(−0.72,−0.88)	(−0.80,−1.04)	(−1.15)
Warm Disk	(−046,−0.63)	(−0.35,−0.51)	(−0.03,+0.14)	(−0.85,−0.99)	(−1.04,−1.15)	(−1.19,−1.24)	(−1.44,−1.48)
Cool Disk	(−0.97,−1.29)	(−1.31)	(∼0.00)	(−1.32,−1.45)	(−2.08,−2.28)	(−2.09,−2.27)	(−2.46,−2.74)

[a] The values listed for each element represent the range of [X/H] found by Sembach & Savage (1996) for each type of cloud. For some sight lines in the halo, disk + halo, and warm disk categories we assumed [X/H] ≈ [X/Zn] ≈ [X/S] since no H I or H_2 estimates were available for the individual clouds studied. Zn and S are nearly undepleted in such environments. For cases where only one value is listed, the element was measured for only one sight line.

[b] The Mg abundances listed in Sembach & Savage (1996) have been increased by 0.27 dex to reflect the 0.27 dex average decrease in the Mg II $\lambda\lambda$1239.925 and 1240.395 f-values between those adopted by Sembach & Savage and the current recommendation by Morton (2000) based on the theoretical calculations of Theodosiou & Federman (1999).

TABLE 2. Diffuse cloud gas-phase abundance summary

Over the period 1994 to 1999 the recommended f-values for the important Mg II UV multiplet at 1239.925 and 1240.395 Å have changed considerably based on studies of Sofia et al. (1994), Fitzpatrick (1997), Theodosiou & Federman (1999), and Sofia, Fabian, & Howk (2000). The abundance results for Mg II plotted in Figure are based on the Sofia et al. (1994) f-values of f(1239.925) = 12.5×10^{-4}, and f(1240.395) = 6.25×10^{-4}. The most recent compilation of Morton (2000) adopts the f-values of Theodosiou & Federman (1999) of f(1239.925) = 6.32×10^{-4} and f(1240.395) = 3.56×10^{-4}. With the newer f-values the Mg II gas phase abundances plotted in Figure 4 will increase by a factor of approximately 1.87 or 0.27 dex. The gas phase abundance values given in Tables 2 and 3 include the effects of this 0.27 dex adjustment.

7. Implications for the composition of interstellar grain cores

The upper envelope of abundances for the halo clouds shown in Figure 4 implies that the refractory elements are depleted (deficient) from the gas and there is very little variation in the depletion level from cloud to cloud over widely separated regions of the Galaxy (i.e. halo gas in the solar vicinity and under the Sagittarius and Norma spiral arms).

The lack of abundance variation coupled with the observed level of depletion supports the idea that the grains in halo clouds have been eroded down to grain cores which are difficult to destroy. The processes that move gas from the disk to the halo apparently are not violent enough to completely destroy the dust (Sembach & Savage 1996).

We explore the composition of these grain cores in Table 3 where we list values of $10^6(\mathrm{X/H})_\mathrm{c}$, $10^6(\mathrm{X/H})_\mathrm{g}$, and $10^6(\mathrm{X/H})_\mathrm{d}$ where c, g, and d represent cosmic, gas, and dust, respectively. In the case of Mg the listed values include a correction for the 0.27 dex revision in the Mg II weak line f-values discussed earlier. The dust phase abundance of a particular element is obtained by assuming $10^6(\mathrm{X/H})_\mathrm{d} = 10^6(\mathrm{X/H})_\mathrm{c} – 10^6(\mathrm{X/H})_\mathrm{g}$. The assumption of cosmic (current epoch) disk abundances is probably valid if the gas of the low halo actually is supplied from the disk through a process that circulates the matter on a relatively short time scale ($<\sim 10^9$ years). For the cosmic abundance we use solar system meteoric abundances from Anders & Grevesse (1989). To explore the possibility the Sun is overabundant by about 0.2 dex compared to current epoch Population I abundances we list in the footnote to the table the effect of reducing the solar system reference abundances by 0.2 dex for all the elements.

Assuming solar abundances are the valid reference abundances, the numbers in Table 3 imply that $\sim$ 45, 44, and 78% of the Mg, Si, and Fe in the halo clouds reside in dust grains. Since these are the most abundant heavy elements (other than C, N and O for which we have no information for halo cloud gas or dust), it is of interest to investigate the implication of this particular mixture of elements for the composition of the dust grain cores. The existence of silicate grains is well established in the ISM of the Galactic disk from the 9.7 and 18 μm SiO stretch and bend features observed along high extinction sight lines (Roche & Aitken 1985) with strengths that imply that most of the Si must exist in grains (Draine & Lee 1984). It is believed these silicate grains may be in the form of (Mg, Fe)$\mathrm{SiO_3}$ and/or (Mg, Fe)$_2\mathrm{SiO_4}$ which are pyroxenes and olivines, respectively. For pure pyroxenes the expected value of (Mg+Fe)/Si is 1.0. For pure olivines the expected value is 2.0. From Table 3 we see the observed value of (Mg+Fe)/Si is (17+25)/16 = 2.6 for solar reference abundances and (3+13)/2.9 = 5.5 for reference abundances 0.2 dex smaller than solar (see the notes to Table 3). In either case the value of the ratio is substantially larger than the expected range from 1.0 to 2.0 for pure silicate grains. In addition to being present in silicates, the Mg and Fe in these halo clouds must exist

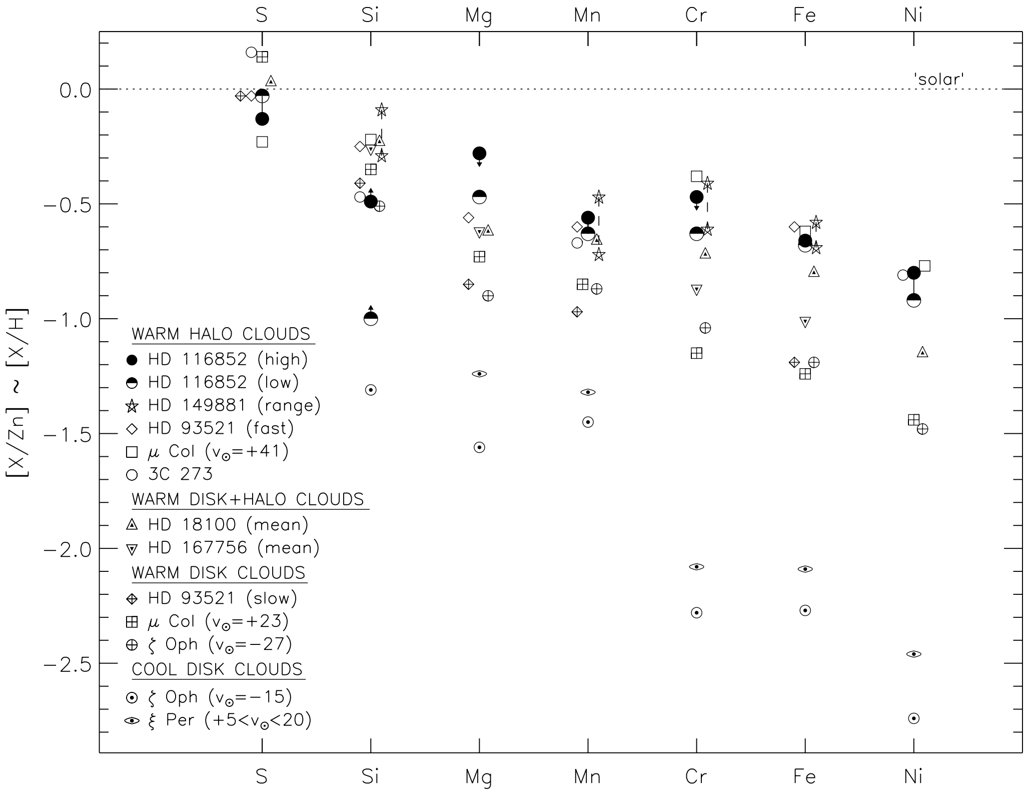


FIGURE 4. A comparison of gas phase abundances in clouds found in the Galactic disk and halo. [X/Zn] = log(X/Zn) – log(X/Zn)$_{\odot}$ is plotted for S, Si, Mg, Mn, Cr, Fe and Ni. Since Zn is normally only slightly depleted in the warm gas of the ISM [X/Zn] should be ∼[X/H]. The legend identifies each sight line. Note the trend of increasing gas phase abundances in going from cool disk clouds, to warm disk clouds, and to warm halo clouds. The variation in the gas phase abundances for the warm halo clouds is remarkably small given that the sight lines sample a number of directions in the solar vicinity and includes gas in the halo under the Sagittarius and Norma spiral arms The abundance results shown here are summarized in Table 2 (this figure is from Sembach & Savage 1996).

in some other type of dust grain core. Likely possibilites include various oxides such as MgO, Fe_2O_3, and Fe_3O_4 (Nuth & Hecht 1990; Fadeyev 1988) and pure Fe grains. However, since Fe grains are so easy to destroy, the presence of Mg and Fe in oxides seems more likely.

8. Gas phase abundances in high velocity clouds

The high velocity clouds (HVCs) of the Milky Way are clouds detected in 21 cm H I emission with $|v_{\rm LSR}| > 100$ km s^{-1} whose radial velocity is inconsistent with Galactic rotation. The properties of HVCs are reviewed by Wakker & van Woerden (1997). The origin of the HVCs is poorly understood but it is clear that abundance and distance measurements are crucial for understanding the role the HVCs play in Galactic phenomena. We discuss the HVCs as a topic separate from that of gas in the low galactic halo since it appears that many of the HVCs are quite distant. For example, the HVC known as the Magellanic Stream which is a ∼ 180° by 20° band of H I emission extending from the Magellanic Clouds to the south Galactic pole and beyond is almost certainly asso-

	Mg	Si	Fe	S	Ni	Cr	Mn
$10^6(X/H)_c$	38	36	32	19	1.8	0.48	0.34
$10^6(X/H)_g$	21[a]	20	7.4	17	0.26	0.15	0.083
$10^6(X/H)_d$[b]	17	16	25	2.0	1.5	0.33	0.26

[a] The abundance of Mg has been adjusted to reflect the revison in the f-values for the important Mg II 1239.925 and 1240.29 Å doublet adopted by Morton (2000).

[b] We assume $(X/H)_d = (X/H)_c - (X/H)_g$. For the cosmic reference abundance, we assume meteoritic abundances from Anders & Grevesse (1989). If the true current epoch reference abundances are 0.2 dex smaller for all the elements, the seven values of $10^6(X/H)_d$ from left to right become: 3, 2.9, 13, 0.0, 0.87, 0.15, 0.13.

Table 3. Gas and dust phase abundances for halo clouds

ciated with the Magellanic clouds and therefore represents phenomena occurring in the outermost regions of the Milky Way.

The *HST* has been extremely important in providing the first reliable information about abundances in the HVCs. For recent reviews see Savage & Sembach (1996a) and Wakker & van Woerden (1997). The highest quality data for abundances in HVCs are for the paths to NGC 3783 ($l = 287.5^o$, $b = 23.0^o$) and Mrk 290 ($l = 91.5^o$, $b = 47.9^o$). The path to NGC 3783 likely samples Magellanic Stream gas (see below) while the path to Mrk 290 samples gas in HVC Complex C.

For NGC 3783, Lu et al. (1994, 1998) detected the HVC at +240 km s^{-1} in the absorption lines of S II, Si II, and Fe II. This HVC is cataloged as HVC 287.5+22.5+240. Sulfur is the most valuable element for deriving an abundance for this HVC since the observed lines of S II are not saturated and sulfur is not readily depleted onto interstellar dust. Furthermore, with an ionization potential of 23 eV, S II is expected to be the dominant state of ionization in the HVC, which has $N(H\ I) = 8.0 \times 10^{19}$ cm^{-2} based on Australia Telescope Compact Array interferometer observations obtained at an angular resolution of $1'$ (Wakker et al. 1999b). The observed Si II line is saturated and only provides a lower limit to the abundance of Si in the HVC. However, the S II and Fe II observations are of high quality and yield S II/H I = 0.25±0.07 time solar and Fe II/H I = 0.033 ± 0.006 times solar. Lu et al. (1998) showed that for this particular HVC with its large value of N(H I), it is unlikely that the S II and Fe II abundances obtained above are subjected to large ionization corrections. Therefore the observed value of N(S II)/N(Fe II) implies that S/Fe in the HVC is 7.6 ± 2.2 times higher than the solar value and S/H is 0.25 ± 0.07 time the solar value. Supersolar S/Fe ratios are often found in Galactic ISM clouds because S is essentially unaffected by the presence of dust and Fe is easily incorporated into dust. The HVC in the direction of NGC 3783 therefore likely contains dust and has an intrinsic metallicity (as measured by S/H) of 0.25±0.07 times solar. The metallicity level and the amount of Fe depletion are very similar to those found in the interstellar gas of the Magellanic Clouds. These similiarities coupled with the position of the HVC on the sky suggested to Lu et al. (1998) that the HVC originated from the Magellanic Clouds. In particular it appears that HVC 287.5+22.5+240 may represent gas in the "leading arm" that was predictied in tidal models of the Magellanic Stream by Gardiner & Noguchi (1996). The metallicity in this particular HVC appears to be too high to be consistent with the Blitz et al. (1999) model which suggests that many of the HVCs represent gas left over from the formation of the Local Group.

Mrk 290 lies in the direction of Complex C which is a large HVC spanning ~ 20 by 120 degrees on the sky and moving toward the Sun at ~ 150 km s^{-1}. S II absorption in the direction of Mrk 290 was observed with the GHRS by Wakker et al. (1999a). Combining those observations with Westerbork Synthesis Radio Telescope measures of N(H I) at a angular resolution of $1'$ and Hα emission observations using the Wisconsin Hα Mapper (WHAM) at 1^{o} resolution, they conclude that S/H in Complex C is 0.089 ± 0.024 times the solar value. Only a lower limit (5 kpc) to the distance of Complex C is available which implies the structure is at least 3 kpc above the Galactic plane. The mass of Complex C is estimated to be $6.2 \times 10^{6}(\mathrm{D}/10\ \mathrm{kpc})^{2}\ \mathrm{M}_{\odot}$ where $\mathrm{M}_{\odot}$ is the solar mass and D is the distance to Complex C. The associated mass inflow rate per unit area is estimated to be 0.002–$0.004(\mathrm{D}/10\ \mathrm{kpc})^{-1}\ \mathrm{M}_{\odot}\ \mathrm{yr}^{1}\ \mathrm{kpc}^{-2}$. To solve the long standing G-dwarf problem, models of Galactic chemical evolution require an inflow of gas with 0.1 solar abundances at a current rate of approximately $0.004\ \mathrm{M}_{\odot}\ \mathrm{yr}^{-1}\ \mathrm{kpc}^{-2}$ over the entire Galaxy. The G-dwarf problem refers to the observational fact that the metallicities of most long-lived stars near the sun lie in a relatively narrow range. To explain this result the models of Galactic chemical evolution have required the continuous inflow of of low metallicity gas to dilute the enrichment arising from the production of heavy elements in stars. Up until now there has been no direct evidence for such an inflow. However, if the inflow of low metallicity gas similar to that found in Complex C is a common phenomena over the Milky Way, it would appear that the G-dwarf problem may have been solved.

9. The extension of highly ionized gas into the Milky Way halo

The *HST* has been an important facility for studying the hot gas in the ISM as traced by the highly ionized atoms Si IV, C IV and N V. If these ions are created under conditions of collisional ionization equilibrium in a hot gas, they peak in abundance at temperatures of approximately (0.8, 1.0 , and 2.0) $\times 10^{5}$ K, respectively. However, in the case of Si IV and C IV which can be produced by photons more energetic than 33 and 47 eV it is possible that photoionization may also be responsible for some of these species found in the ISM. The possibility that hot gas might extend away from the plane of the Galaxy in a hot halo was first proposed by Spitzer (1956) in his classic paper "On a Possible Interstellar Galactic Corona." His paper noted the existence of cool clouds in the low halo of the galaxy detected in optical absorption line measurements of required the pressure support that might be provided by a hot unseen phase of the gas. Spitzer's paper contained the theoretical prediction of the hot phase of the interstellar gas and included discussions of the ionization, heating, cooling, and kinematical processes likely affecting the gas. Spitzer proposed several ways the gas might be detected. An insightful quote from his paper is: "The ultimate lines of N V and C IV, at about 1240 and 1550 Å, respectively, might be observable. It would appear in principle an interstellar corona could be detected and analyzed by means of spectroscopic measures from an satellite." Observations with the *HST*, the observatory Lyman Spitzer helped create, have yielded fundamental information about Spitzer's Galactic Corona.

Two methods have been used to estimate the density distribution of highly ionized gas away from the Galactic plane. In one method N(X) $\sin|b|$ is compared to $|z|$ for many sight lines and the scale height of the gas is estimated by finding the value of $|z|$ where N(X) $\sin|b|$ ceases to increase beyond an amount allowable by a simple gas distribution, such as an exponential layer. In the other method, individual absorption line profiles are assumed to be influenced by the effects of Galactic rotation and an estimate of the scale height of the gas follows from an analysis of the line profile shape given assumptions about the nature of the Galactic rotation curve at large $|z|$.

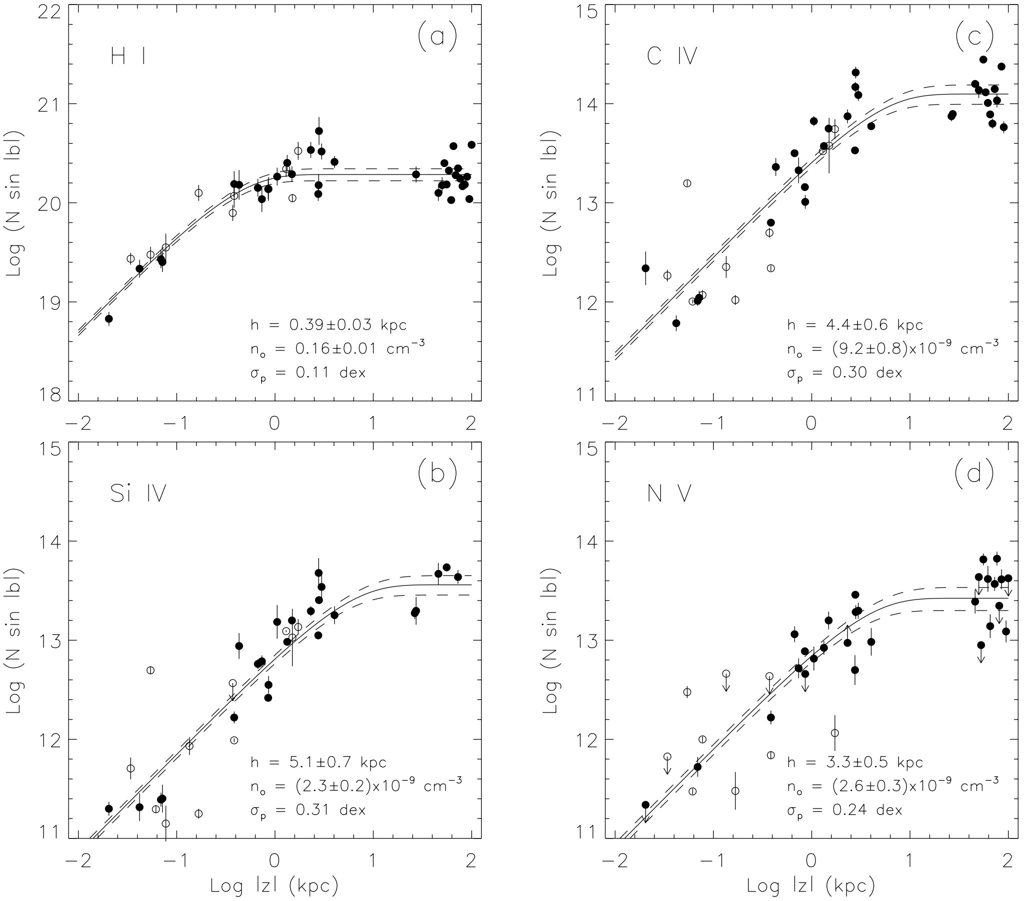


FIGURE 5. Estimates of the distribution of highly ionized gas away from the galactic plane are obtained from plots of projected column density, N(ion)sin $|b|$, versus vertical distance, $|z|$, to stars and objects beyond the galaxy. (a) N sin $|b|$ versus $|z|$ for H I. The solid line indicates the best fit curve through the data for a simple exponential density distribution with vertical scale height h and mid-plane density n_o. The fits allowed by the 1σ errors in the values of n_o and h are shown as dashed lines. The model allows for the irregular distribution of the absorbing gas. (b) Same as (a), except for Si IV. (c) Same as (a), except for C IV. (d) same as (a) except for N V (this figure is from Savage et al. 1997).

Values of N(ion) sin $|b|$ versus $|z|$ are shown for H I, Si IV, C IV, and N V in Figure 5 for a group of stars and extragalactic objects observed by *IUE* and the *HST* (see Savage et al. 1997). The major improvement in the evaluation of the $|z|$ distribution of the gas from these measurements over earlier estimates of the scale heights of the high ions is the significant increase in the number of quality measurements for C IV and N V along extragalactic sight lines provided by the GHRS. For all four species shown in Figure 4 there is a clear increase of N(ion) sin $|b|$ with increasing $|z|$. However, the values of N(ion) sin $|b|$ for Si IV, C IV and N V continue to increase to larger $|z|$ than those for H I, indicating that the high ions are considerably more extended than H I.

To estimate the stratification of H I, Si IV, C IV, and N V away from the Galactic plane for the observations shown in Figure 5, we assume the gas distribution in the $|z|$-direction is described by a simple exponential layer that is intrinsically patchy or inhomogeneous. The logarithmic uncertainty of each measurement is assumed to have two uncertainties that add in quadrature; one is from the observational uncertainty, σ_o, and the other is from the intrinsic patchiness of the interstellar gas, σ_p. The model parameters, which include the exponential scale height (h) and the mid-plane density (n_o), are determined by minimizing the χ^2 of the fit of the model to the data. The value of σ_p is adjusted during the fitting process to produce an acceptable reduced χ^2 of 1.0. The resulting best fit curves shown in Figure 5 imply scale heights for the highly ionized gas of h(Si IV) = 5.1 ± 0.7 kpc, h(C IV) = 4.4 ± 0.6 kpc, and h(N V) = 3.3 ± 0.5 kpc. In contrast the neutral gas has h(H I) = 0.39 ± 0.03 kpc. In the case of N V, a somewhat larger and more uncertain result is obtained, h(N V) = 3.9 ± 1.4 kpc, whcn properly allowing for the fact that a number of the N V column density measurements are simply upper limits.

A determination of the scale heights for the ions of Si IV and C IV from a rotational analysis of the line profiles assuming co-rotation of disk and halo gas yields results roughly consistent with that obtained from the N $\sin|b|$ versus $|z|$ analyis method shown in Figure 5. However at this time the rotational analysis has only been applied to a relatively small number of extragalactic sources.

The study of highly ionized gas in Spitzer's Galactic Corona will be greatly accelerated now that the Far Ultraviolet Spectroscopic Explorer (FUSE) satellite is routinely producing 20 km s^{-1} resolution spectra of AGNs and Galactic stars in the wavelength region from 912 to 1187 Å (Moos et al. 2000). This region of the spectrum contains the extremely important doublet absorption lines of O VI at 1031.926 and 1037.617 Å. With the large abundance of oxygen and the high ionization potential for the conversion of O V to O VI of 113 eV, O VI represents the best tracer of hot interstellar gas available to astronomy. Interestingly, the first *FUSE* results relating to the Galactic Corona imply an O VI Galactic scale height of h(O VI) = 2.7 ± 0.4 kpc based on *FUSE* observation of 11 AGNs (Savage et al. 2000). The smaller scale height of O VI than C IV appears to be supported by an observed factor of ~ 4 increase in the C IV to O VI column density ratios obtained toward objects in the disk of the Galaxy compared to the halo. Several possible explanations for these scale height differences are discussed by Savage et al. (2000) and Spitzer (1996).

Much of the new information on elemental abundances and physical conditions in the Galactic ISM discussed here was obtained by the GHRS on the *HST*. We thank the many individuals who contributed to the construction and operation this observational capability. BDS acknowledges support from NASA under grant NAG5-1852.

REFERENCES

Anders, E. & Grevesse, N. 1989 *Geochim. Cosmochim. Acta* **53**, 197.

Blitz, L., Spergel, D., Teuben, P., Hartmann, D., & Burton, W. B. 1999 *ApJ* **514**, 818.

Brandt, J. C., Heap, S. R., et al. 1994 *PASP* **106**, 890.

Brandt, J. C., et al. 1999 *AJ* **117**, 400.

Cardelli, J. A. 1994 *Science* **265**, 209.

Cardelli, J. A. & Ebbets, D. C. 1994. In *Calibrating Hubble Space Telescope* (eds. J. C. Blades & S. J. Osmer), p. 322. Space Telescope Science Institute.

Cardelli, J. A., Federman, S. R., Lambert, D. L., Theodosiou, C. E. 1993 *ApJ* **416**, L41.

Cardelli, J. A., Sembach, K. R., & Savage, B. D. 1995 *ApJ* **440**, 241.

de Boer, K. S., Jura, M. A., & Shull, J. M. 1987. In *Exploring the Universe with the IUE Satellite* (ed. Y. Kondo), p. 533. Reidel.

Draine, B. T. & Lee, H. M. 1984 *ApJ* **285**, 89.

Fadeyev, Y. 1988. In *Atmospheric Diagnostics of Stellar Evolution* (ed. K, Nomoto), p. 174. Springer-Verlag.

Field, G. 1974 *ApJ* **187**, 453.

Fitzpatrick, E. L. 1997 *ApJ* **482**, L199.

Fitzpatrick, E. L. & Spitzer, L. 1994 *ApJ* **427**, 232.

Gardiner, L. T. & Noguchi, M. 1996 *MNRAS* **278**, 191.

Grevesse, N. & Noels, A. 1993. In *Origin of the Elements* (eds. N. Prantzos, E. Vangioni-Flam, & M. Casse), p. 15. Cambridge Univ. Press.

Howk, J. C. & Savage, B. D. 1999 *ApJ* **517**, 746.

Howk, J. C., Savage, B. D., & Fabian, D. 1999 *ApJ* **525**, 253.

Howk, J. C., Sembach, K. R., & Savage, B. D. 2000 *ApJ*, **543**, 278.

Heap, S. R., Brandt, J. C., et al. 1995 *PASP* **107**, 871.

Jenkins, E. B. 1987. In *Interstellar Processes* (eds. D. J. Hollenbach & H. A. Thronson), p. 533. Reidel.

Jenkins, E. B., et al. 1998 *ApJ* **492**, L147.

Kimble, R. A., et al. 1998 *ApJ* **492**, L83.

Lambert, D. L., Sheffer, Y., Gilliland, R. L., & Federman, S. R. 1994 *ApJ* **420**, 756.

Lauroesch, J. T., Meyer, D. M., Watson, J. K., & Blades, J. C. 1998 *ApJ* **507**, L89.

Linsky, J. L., Diplas, A., Wood, B. E., Brown, A., Ayres, T. R., & Savage, B. D. 1995 *ApJ* **451**, 335.

Linsky, J. L. & Wood, B. E. 1998. In *ASP Conf. Ser. Vol. 143, Scientific Impact of the Goddard High Resolution Spectrograph* (eds. J. C. Brandt, T. B. Ake, & C. C. Petersen), p. 197. ASP.

Lu, L., Savage, B. D., & Sembach, K. R. 1994 *ApJ* **426**, 563.

Lu, L., Savage, B. D., Sembach, K. R., Wakker, B. P., Sargent, W. L. W., & Oosterloo, T. A. 1998 *AJ* **115**, 162.

Mathis, J. S. 1999. In *Chemical Evolution from Zero to High Redshift, ESO Astrophysics Symposium* (eds. J. R. Walsh & M. R. Rosa), p. 54. Springer.

Meyer, D. M., Jura, M. J., Hawkins, I., & Cardelli, J. A. 1994 *ApJ* **437**, L59.

Meyer, D. 1999. In *Chemical Evolution from Zero to High Redshift, ESO Astrophysics Symposium* (eds. J. R. Walsh & M. R. Rosa), p. 44. Springer.

Moore, C. E. 1970. *Ionization Potentials and Ionization Limits Derived from the Analysis of Optical Spectra*, Rep. No. NSRDS-NBS34. U.S. Dept. of Commerce.

Moos, H. W., et al. 2000 *ApJ*, **538**, L1.

Morton, D. C. 2000 *ApJS*, **130**, 403.

Nuth, J. A. & Hecht, J. H. 1990 *Astrophys. Space Sci.* **163**, 79.

Roche, P. F. & Aitken, D. K. 1985 *MNRAS* **215**, 35.

Savage, B. D., Bohlin, R. C., Drake, J. F., & Budich, W. 1977 *ApJ* **216**, 291.

Savage, B. D., Cardelli, J. A., & Sofia, U. J. 1992 *ApJ* **401**, 706.

Savage, B. D., Lu, L., Weymann, R., Morris, S. L., & Gilliland, R. L. 1993 *ApJ* **404**, 134.

Savage, B. D. & Sembach, K. S. 1996a *ARAA* **34**, 279.

Savage, B. D. & Sembach, K. S. 1996b *ApJ* **470**, 893.

Savage, B. D., Sembach, K. S., & Lu, L. 1997 *AJ* **113**, 2158.

Savage, B. D., Sembach, K. S., et al. 2000 *ApJ*, **538**, L27.

Sembach, K. S. 1998. In *ASP Conf. Ser. Vol. 143, Scientific Impact of the Goddard High Resolution Spectrograph* (eds. J. C. Brandt, T. B. Ake, & C. C. Petersen), p. 181. ASP.

Sembach, K. S. & Savage, B. D. 1996 *ApJ* **457**, 211.

Shull, J. M., et al. 2000 *ApJ*, **538**, L73.

Sofia, U. J., Cardelli, J. A., & Savage, B. D. 1994 *ApJ* **430**, 650.

Sofia, U. J., Fabian, D., & Howk, J. C. 2000 *ApJ*, **531**, 384.

Sofia, U. J., Meyer, D. M., & Cardelli, J. A. 1999 *ApJ* **522**, L137.

Sofia, U. J., Savage, B. D., & Cardelli, J. A. 1993 *ApJ* **413**, 251.

Spitzer, L. 1956 *ApJ* **124**, 20.

Spitzer, L. 1996 *ApJ* **458**, L29.

Spitzer, L. & Fitzpatrick, E. L. 1993 *ApJ* **409**, 299.

Spitzer, L. & Fitzpatrick, E. L. 1995 *ApJ* **445**, 196.

Spitzer, L. & Jenkins, E. B. 1975 *ARAA* **13**, 133.

Theodosiou, C. E. & Federman, S. R. 1999 *ApJ* **527**, 470.

Walborn, N. R. 1998 *ApJ* **492**, L169.

Wakker, B. P., Howk, J. C., et al. 1999a *Nature* **402**, 388.

Wakker, B. P., Savage, B. D., Oosterloo, T. A., & Putman, M. 1999b. In *ASP Conf. Ser. 166, Stromlo Workshop on High Velocity Clouds* (eds. B. R. Gibson & M. E. Putman), p. 302. ASP.

Wakker, B. P. & van Woerden, H. 1997 *ARAA* **35**, 217.

Woodgate, B., et al. 1998 *PASP* **110**, 1183.

HST's view of the center of the Milky Way galaxy

By MARCIA J. RIEKE

Steward Observatory, University of Arizona, Tucson, AZ 85721

The Galactic Center has been the subject of a variety of *HST* observing programs, mainly since the installation of NICMOS. The observational strengths of NICMOS lie with its sensitivity and very stable point spread function which enables a variety of studies including sensitive searches for variable sources and accurate colors across the 1 to 2.5 μm region. The emission line filters in NICMOS enable studies of the interstellar medium and a search for [SiVI] emission as a 'smoking gun' for gas clouds near a black hole powered accretion disk.

1. Introduction

The center of the Milky Way is of course the closest galaxy nucleus and is a natural area to choose to study in detail. The discovery of a peculiar radio source, SgrA*, and the subsequent demonstration that it is a black hole has only heightened interest in the center. Figure 1 shows a contour plot at 1.04 μm compared to a NICMOS image at 1.45 μm which clearly shows why the Galactic Center requires use of infrared instrument like NICMOS with $A_V \sim 30$ while $A_K \sim 3.3$.

The Galactic Center has been studied with *HST* from the first observing cycle using WFPC proposed in an era where the nature of many of the stars was not understood, and the existence of a cluster in very close proximity to the black hole, SgrA*, was unknown. these early observations were to have provided a long time baseline for computing proper motions to be used in conjunction with NICMOS data taken much later—the WFPC image in Figure 1 was produced by one of these programs.

Our current understanding of the Galactic Center leads to many reasons for observing the Center with *HST*. The ability to study processes in the environment of a massive black hole is a prime driver as is studying the nuclear stellar population. We can probe the neighborhood of a massive black hole in detail to look for evidence indicating whether the black hole modifies the stars or the interstellar medium. Ground based spectroscopy (e.g. Haller et al. 1996, Blum et al. 1999, Krabbe et al. 1995) has revealed a number of post-main sequence stars including both red and blue supergiants. However, it has been difficult to pin down the age of this collection of stars. Using the ability of NICMOS to measure colors accurately over the entire 1 to 2.5 μm offers the possibility of finding the tip of the main sequence associated with the luminous post-main sequence stars and hence the age of the stellar group. Other ground based studies have determined the mass of the black hole, SgrA*, as 2.5×10^6 $M_\odot$ using kinematic data as well as proper motions. NICMOS offers the ability to measure proper motions of fainter sources and over a wider field than has been done from the ground. When these studies reach their full potential with the revival of NICMOS with the installation of the cryocooler, much more will be known about anisotropies in stellar orbits and similar issues.

Because of the stability of NICMOS both from a photometric standpoint and from the standpoint of the point spread function, it is an ideal instrument for searching for variable sources. SgrA* is an obvious candidate for such a study as is the stellar population generally. Searching for any flux at all from SgrA* is another important goal which benefits from the high strehl ratio and stability of the NICMOS point spread function.

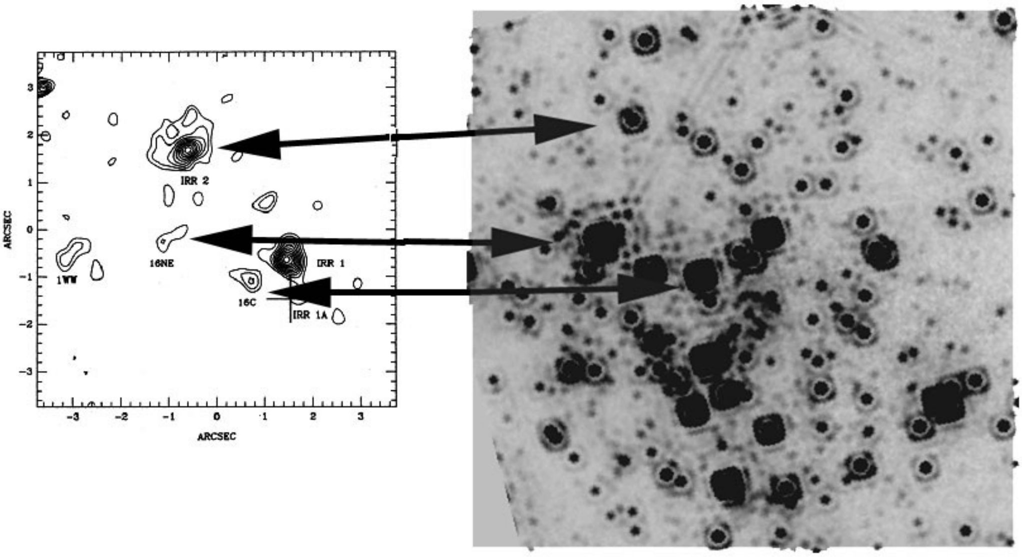


FIGURE 1. The contour plot on the left is from Liu et al. 1993 and was taken at 1.04 μm with WFPC. The right hand panel shows a NICMOS image at 1.45 μm and illustrates how devastating the effects of extinction are. The arrows connect the the same three stars.

Last, NICMOS is equipped with a variety of narrow band filters sensitive to emission lines. The suite of available lines includes [FeII] at 1.644 μm, Pα at 1.875 μm, [SiVI] at 1.965 μm and H_2(1-0) at 2.122 μm. These lines are diagnostics for shocks as well as the strength and temperature of the ionizing radiation field. The coronal line of silicon is especially interesting as it has been detected in Seyfert galaxy nuclei and would be a "smoking gun" for any accretion disk around SgrA*. Pα is also very important because it is 30 times stronger than Brγ, the only other hydrogen line that has been observed much at the Galactic Center. It plays the role that Hα does in extragalactic studies of unobscured galactic nuclei.

2. Search for variability

The entire program of Galactic Center observations acquired so far with NICMOS allows a range of time scales to be probed. Stolovy et al. 1999 used two orbits to look for variations on minute scales that would be indicative of either orbital motions or gravitational lensing of clumps in even a very small accretion disk. Stolovy et al. report that no variability to a level of $\sim$ 0.5 μJy is seen in any position within a few arc secs of SgrA* an scales of a minute to an hour.

Figure 2 shows four epochs of data spanning seven months (March 1998, July 1998, September 1998, and October 1998) covering the central $\sim$ 5.5 arcsec of the center at 1.45 μm. A short dash indicates the location of a variable source on two of the epochs—this source was very faint during July, 1998. The NICMOS data set is being studied for variability on scales as long as 16 months (Leistra et al., in preparation) and reveal what fraction of the population is variable on such time scales with amplitudes as small as 0.5 μ Jy/pixel. The only time scale not well-studied by NICMOS is that of days, precisely the scale on which the eclipsing variability of IRS16SW was detected by Ott et al. 1999 with a period of 9.72 days. Of particular note in Figure 2 is the repeatability

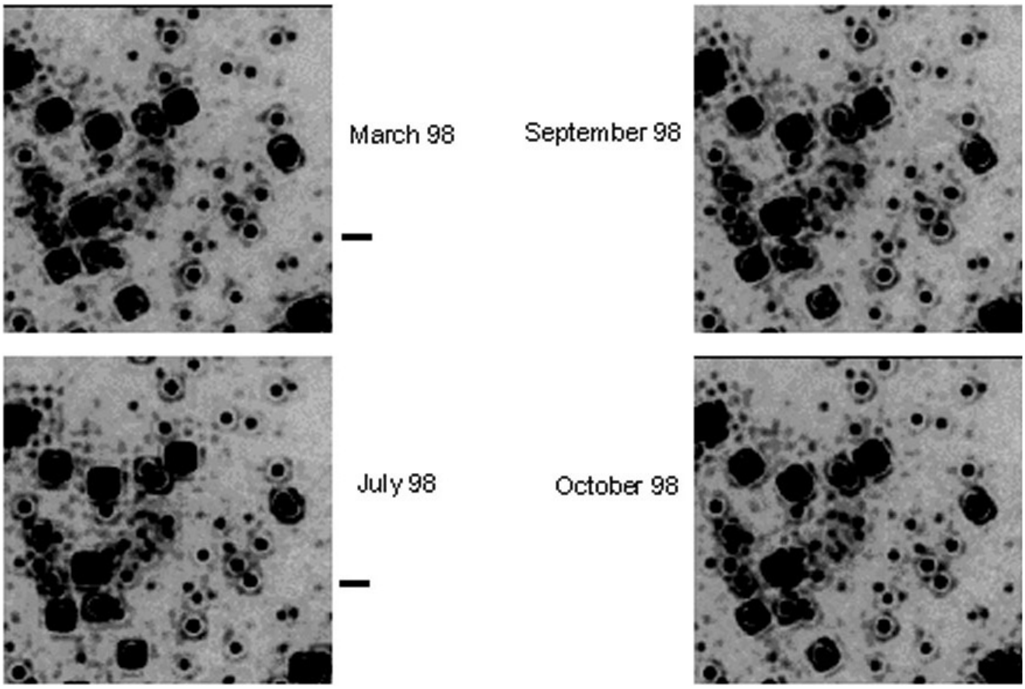


FIGURE 2. The four panels show ~ 5.5 arcsec of the Galactic Center at 1.45 μm using camera 1 on NICMOS. A short dash is positioned to the right of a variable source which was particularly faint in July, 1998.

of the appearance of the stellar cluster surrounding SgrA* located at the center of the images.

3. Emission line imaging

As mentioned in the introduction, NICMOS is equipped with a variety of emission line filters. Programs to study all the lines mentioned in the introduction were executed with preliminary results on [FeII] and [SiVI] described in Stolovy et al. Since [SiVI] requires higher excitation than can be provided by stars, not surprisingly it has only been observed from Seyfert galaxy nuclei. The NICMOS imaging in this line sets an upper limit of $< 1.2 \times 10^{-16}$ erg/sec/cm^2 which translates to a luminosity of $< 4.8 \times 10^{31}$ erg/sec. By comparison, the Circinus galaxy which harbors a black hole only slightly more massive, 4×10^6 M$_\odot$ than that at the Galactic Center has a detected [SiVI] flux of 1.45×10^{-13} erg/sec/cm^2 or a dereddended luminosity of 2.6×10^{38} erg/sec. If the emission from the Galactic Center were concentrated in a few pixels, it could suffer an additional 60 magnitudes of extinction as compared to that over the general Galactic Center region and still be detectable in the NICMOS image. In fact, the Circinus emission comes from an area ~ 10 pc in size (Maiolino et al. 2000) so the allowed extra extinction is much less, but nonetheless, no evidence is found for high temperature excitation by an accretion disk.

Similarly, no Pα emission is seen which can be unambiguously associated with SgrA* as illustrated in Figure 3. This image, taken with NICMOS camera 2 and covering the central 19 arc sec, reveals several interesting structures with both extended Pα and Pα associated with stars visible. The "mini-cavity" is seen to be a region cleared out by a

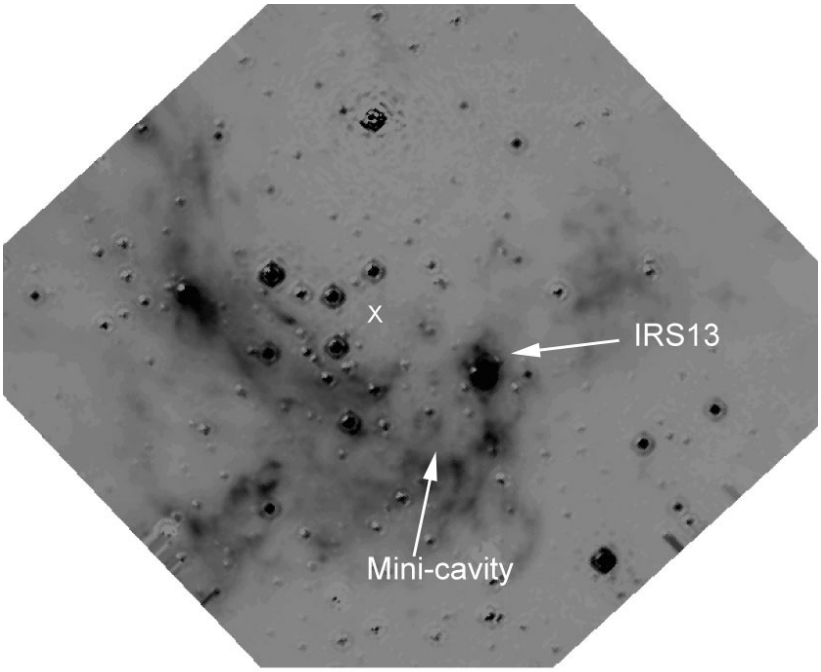


FIGURE 3. The central 19 arcsec in a continuum-subtracted image in the light of Pα. An 'X' indicates the position of SgrA*. The image has been rotated so that north is up and east is to the right.

star whose NICMOS colors and luminosity strongly suggest that it is a B supergiant. The Pα emission has a morphology very similar to the 1.3 cm emission as displayed in Zhao & Goss 1999. The Pα emission also exhibits some similarity to the 10 μm emission which is produced by warm dust as seen in Cotera et al. 1999. A larger scale comparison of the Pα emission with the radio emission is being prepared by Scoville and Stolovy (private communication) which shows that the Pα is not strongly distorted by obscuration as compared to the radio emission with a few stars (for example, IRS13) displaying Pα emission from presumably circumstellar shells. The NICMOS data are the first time that any transition outside of radio wavelengths could be studied with such high spatial resolution over the entire central arc minute.

4. Flux from the black hole?

A long standing issue in Galactic Center research is the lack of emission from any accretion disk surrounding SgrA*. This object which is the black hole based on its lack of proper motion (Backer & Sramek 1999), and on its lying at the location of the large mass concentration deduced from proper motions of surrounding stars (Ghez et al. 1998, Genzel et al. 2000) which is in turn at the dynamical center of the Galaxy has until recently only been detected at radio wavelengths. Its radio properties make a unique source in the Galaxy because of its small size (Lo et al. 1999), polarization properties (Bower 2000), high brightness temperature, non-thermal spectrum and variability (Wright & Backer 1993). Its luminosity is relatively low at only 30 $L_{\odot}$. The recent report of a detection at x-ray wavelengths at a lower than expected flux level using Chandra (Baganoff et al. 2000) has only added to the mystery. Figure 4 adopted from Quataert & Narayan 1999 shows how difficult it is to produce the observed radio flux without producing more emission at

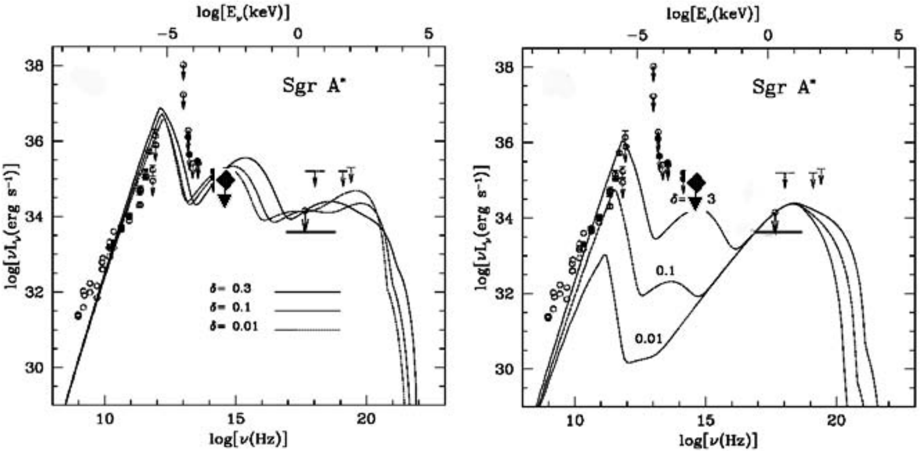


FIGURE 4. Two suites of ADAF models for SgrA* from Quataert & Narayan 1999 are shown along with the NICMOS limit at 1.6 μm from Stolovy et al. 1999 shown as a diamond and the Chandra detection by Baganoff et al. as quoted in Quataert & Gruzinov 2000 is shown as a bar.

shorter wavelengths than has been observed. Previous ground based observations using various techniques to sharpen the image have been hampered in trying to set a near-infrared flux limit for SgrA* because of the at least 100 times brighter sources lying less than an arc second away whose PSF wings limit the flux level than can be probed and because of the cluster of stars lying within 0.5″ of SgrA* (Menten et al. 1997).

NICMOS offers the opportunity to probe to fainter flux levels. The first attempt to set a limit on the flux of SgrA* using NICMOS is reported in Stolovy et al. 1999 at 1.6μm resulted in a limit of 18 mJy as shown in Figure 4. This limit used only one epoch of the NICMOS data with the best possible limit yet to come from a complete analysis of the images taken periodically at 1.45 μm. A limit at least 3 times better than the 18 mJy reported by Stolovy et al. seems possible. In any event, the near-infrared limits coupled with the x-ray detection and the possible submillimeter polarization detection reported by Aitken et al. 2000 are pushing the advection dominated accretion flow (ADAF) and other models to explain the low flux from SgrA*.

5. Star formation at the Galactic Center

Abundant evidence for star formation right at the center has accumulated over the past 20 years with the discovery of both red and blue post-main sequence stars. IRS7 is a well known red supergiant with a progenitor mass of at least 20 M$_\odot$. Other stars such as some of the IRS16 objects are intrinsically blue with characteristics of Wolf-Rayet stars or Luminous Blue Variables and have even higher masses and younger ages than IRS7 (Najarro et al. 1997, Tamblyn et al. 1996). However, because all of these stars are post-main sequence objects, accurate determination of the age of the luminous stellar population has not been possible. The presence of likely AGB stars (Blum et al. 1999) also suggests that several episodes of star formation have occurred. The classical method of determining the age of a stellar cluster is to observe the location of the main sequence turn off in an HR diagram which can be translated directly into the age. This method requires accurate color or temperature information and luminosities. Because of the heavy and somewhat variable extinction across the face of the Galactic Center, enough color information both to deredden the stars and to assign spectral types is needed, and is the

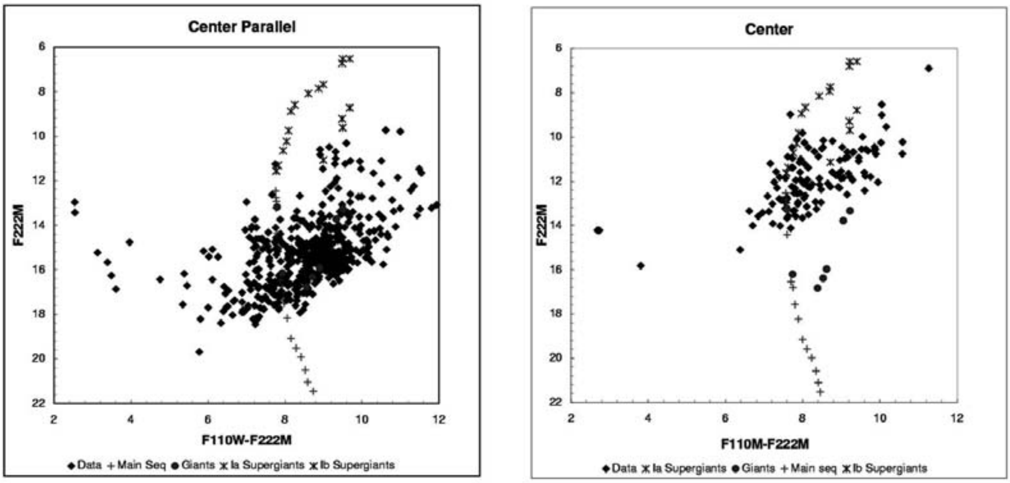


FIGURE 5. Color magnitude diagrams (not dereddened) for the 19″ × 19″ at the Center and the same size field lying ∼ 90″ away.

information that NICMOS can provide and which is unavailable from ground based data at the needed shorter wavelengths.

The approximate colors and magnitudes of the main sequence tip stars can be estimated from the properties of the observed post-main sequence stars. An age of 5 to 10 million years implies that the tip of the main sequence has a spectral type between O6 and B0. With an extinction at V of 30 magnitudes and a distance of 8 kpcs, such stars will have magnitudes at 2.22 μm of 13 to 14. At 1.6 μm and at 1.1 μm they will be 1.9 and 7.9 magnitudes fainter respectively. The faintness at 1.1 μm has made the search for these stars very difficult from the ground. Figure 5 shows the observed color-magnitude diagrams for the NICMOS observations of both the center and a parallel field lying about 90″ away from the center. Immediately apparent is the reduced luminosity of the parallel field. Examination of the average colors in this field show that the average extinction is the same as at the center so the stellar population must be fainter. This is no surprise and has been known for a long time. It is also clear that the tip of the main sequence has likely been observed at the center, but deeper images at 1.1 μm are needed to make this result secure and to provide sufficiently accurate colors that the spectral types and extinctions can be determined well enough. Use of the entire data set including measurements at 1.1, 1.45, 1.6, 1.9, and 2.22 μm will need to be used.

Of particular interest are the colors of the stars closest to SgrA* first discovered by Eckart et al. 1996. Table 1 lists the colors of some of these stars as well as two members of the IRS16 complex. Note that F110M–F160W is similar to but redder than J–H while F160W–F222M is close to H–K. Remarkably, the stars closest to SgrA* have colors at least as blue as the IRS16 stars which are known to be very hot ($T_{\rm eff} \sim 30000°$). These colors confirm on a star by star basis the spectroscopy of the combined light of these stars in Eckart et al. 1999 and Figer et al. 2000. The luminosities of the these stars are in the range expected for the tip of the main sequence, but to have many such stars clumped together so close to the SgrA* location suggests that some other process such as stellar collisions in the region of highest stellar densities may be operating (see Bailey & Davies 1999 for estimates of the rates).

Name	F110M–F160W	F160W–F222M	F222M
S2	$6.43 \pm .09$	$2.31 \pm .14$	$13.66 \pm .13$
S8	$6.32 \pm .09$	$1.93 \pm .20$	$14.81 \pm .19$
S5	$6.36 \pm .09$	$2.27 \pm .10$	$14.20 \pm .09$
S11	$7.08 \pm .08$	$2.31 \pm .09$	$13.53 \pm .09$
IRS16NE	$6.46 \pm .03$	$2.45 \pm .02$	$8.85 \pm .02$
IRS16NW	$6.30 \pm .02$	$2.49 \pm .04$	$10.20 \pm .04$

Table 1. Colors and magnitudes of stars near SgrA*

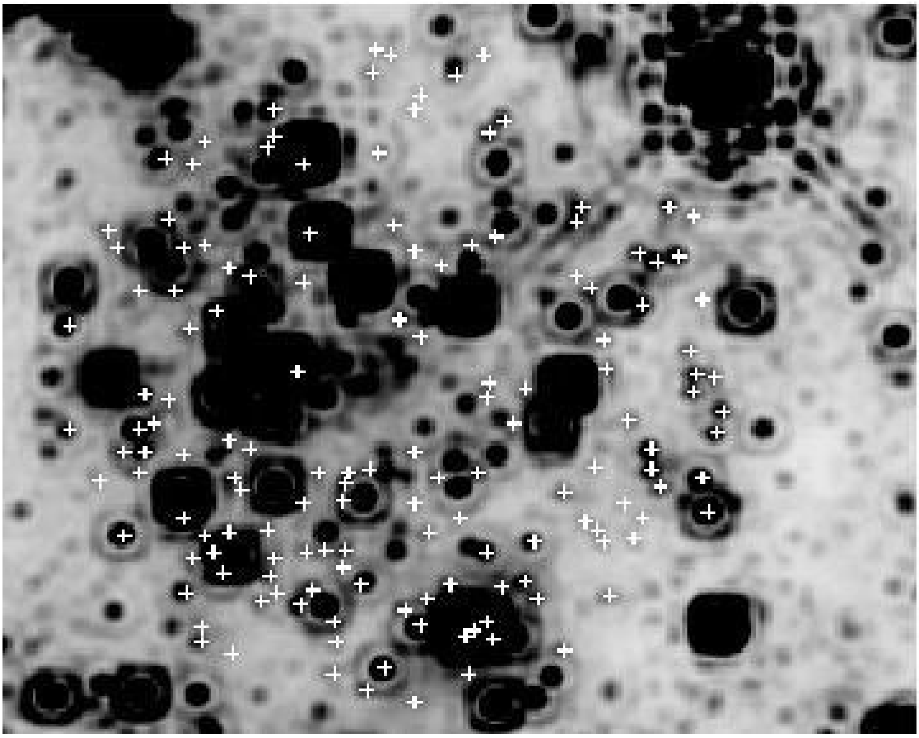


Figure 6. The Center at 2.2 μm with stars within $4''$ of SgrA* that do not appear in the tabulation of sources in Genzel et al. 2000 marked with a '+.' North is in the direction of the upper right corner.

5.1. *Construction of complete samples*

Another important aspect of using NICMOS on the Galactic Center which derives from the stability of its point spread function is the ability to construct complete samples over wider areas than has been done from the ground. Figure 6 shows the center of the NICMOS 2.2 μm image with stars marked within $4''$ of SgrA* if they do not appear in the list compiled by Genzel et al. 2000. A remarkable number of even relatively bright stars have not been previously tabulated. Sample completeness is crucial when searching for evidence of a central cusp as has been illustrated in Alexander 1999. This is another area where NICMOS should eventually provide a definitive answer about the existence or lack of a stellar cusp around the black hole.

5.2. *Star formation in the vicinity of the Galactic Center*

Figer et al. 1999 report NICMOS observations of two massive stellar clusters lying about 30 pc from the Center. These two clusters, the Arches and the Quintuplet, have ages of only a few million years and total cluster masses of at least 6300 $M_\odot$. These clusters are easier to study than the Center itself because of fewer old stars from the bulge (or the center of the disk) contaminating the data. Figer et al. show that the initial mass function here may be flatter (weighted towards more massive stars) than elsewhere in the galaxy. This is reminiscent of the stellar populations in starburst galaxies (e.g. Rieke et al. 1993) and may be indicative of the population of the Center itself.

6. Summary of the NICMOS view of the Galactic Center

NICMOS subjected the Galactic Center region to a variety of observational tests. No evidence for an accretion disk around SgrA* has been found either in emission line data or in terms of variability. The emission line data do reveal streamers, bullets, and a bubble cleared out by a B supergiant star. NICMOS is also providing hints that the stars closest to SgrA* may be modified via collisions but this assertion needs further proof including a NICMOS test of the existence or lack of a stellar cups that is needed if the collision rate is to be significant. NICMOS data also support the presence of recent star formation at the Center.

When NICMOS is revived with the installation of a cryocooler, some of these results such as the F110M colors of the stars can be placed on a firmer footing. Having a longer time baseline of data will enable searches for micro-lensing events and well as proper motions over a larger area than can be done from the ground which will assist in tests for velocity isotropy at the center.

The author thanks the NICMOS Team at Space Telescope Science Institute for enabling the observations presented here. Susan Stolovy and Nick Scoville are thanked for sharing data in advance of publication. G. Rieke and T. Alexander are thanked for useful conversations about the Galactic Center.

REFERENCES

Aitken, D. K., Greaves, J. S., Crysostomou, A., Holland, W. S., Hough, J. H., Pierce-Price, D., & Richer, J. S. 2000 *ApJ* **534**, L173.

Alexander, T. 1999 *ApJ* **527**, 835.

Backer, D. C. & Sramek, R. S. 1999. In *The Central Parsecs of the Galaxy* (eds. H. Falcke, A. Cotera, W. Duschl, F. Melia, & M. Rieke), p. 67. ASP.

Baganoff, F., Angelini, L., Bautz, M., Brandt, N., Cui, W., Doty, J., Feigelson, E., Garmire, G. Kallman, T., Maeda, Y., Morris, M., Nishikida, K., Pravdo, S., Ricker, G. & Townsley, L. 2000 *BAAS* **31**, 62.01.

Bailey, V. C. & Davies, M. B. 1999 *MNRAS* **308**, 257.

Bower, G. 2000 *Galactic Center Newsletter* **11**, 4.

Blum, R. D., Ramierz, S. V., & Sellgren, K. 1999. In *The Central Parsecs of the Galaxy* (eds. H. Falcke, A. Cotera, W. Duschl, F. Melia, & M. Rieke), p. 291. ASP.

Cotera, A., Morris, M., Ghez, A., Becklin, E. E., Tanner, A., Werner, M., & Stolovy, S. 1999. In *The Central Parsecs of the Galaxy* (eds. H. Falcke, A. Cotera, W. Duschl, F. Melia, & M. Rieke), p. 240. ASP.

Eckart, A., Ott, T., & Genzel, R. 1999 *A&A* **352**, 22.

Figer, D., Becklin, E., McLean, I., Gilbert, A., Graham, J., Larkin, J., Levenson, N., Teplitz, H., Wilcox, M., & Morris, M. 2000 *ApJ* **533**, L49.

Figer, D., Kim, S., Morris, M., Serabyn, E., Rich, R. M., & McLean, I. 1999 *ApJ* **525**, 750.

Genzel, R., Pichon, C., Eckart, A., Gerhard, O., & Ott, T. 2000 *MNRAS*, **317**, 348.

Ghez, A., Klein, B. L., Morris, M., & Becklin, E. 1998 *ApJ* **509**, 678.

Haller, J., Rieke, M., Rieke, G., Tamblyn, P., Close, L. & Melia, F. 1996 *ApJ* **456**, 194.

Krabbe, A., Genzel, R., Eckart, A., Najarro, F., Lutz, D., Cameron, M., Kroker, H., Tacconi-Garman, L. E., Thatte, N., Weitzel, L., Drapatz, S., Geballe, T., Strenberg, A., & Kudritzki, R. 1995 *ApJ* **447**, L95.

Liu, T., Becklin, E., Henry, J., & Simons, D. 1993 *AJ* **106**, 1484.

Lo, K., Shen, Z.-Q., Zhao, J.-H., & Ho, P. 1999. In *The Central Parsecs of the Galaxy* (eds. H. Falcke, A. Cotera, W. Duschl, F. Melia, & M. Rieke), p. 72. ASP.

Maiolino, R., Alonso-Herrero, A., Anders, S., Quillen, A., Rieke, M. J., Rieke, G. H., & Tacconi-Garman, L. E. 2000 *ApJ* **531**, 219.

Menten, K., Reid, M., Eckart, A., & Genzel, R. 1997 *ApJ* **475**, 111.

Najarro, F., Krabbe, A., Genzel, R., Lutz, D., Kudritzki, R., & Hillier, D. 1997 *A&A* **325**, 700.

Ott, T., Eckart, A. & Genzel, R. 1999. In *The Central Parsecs of the Galaxy* (eds. H. Falcke, A. Cotera, W. Duschl, F. Melia, & M. Rieke), p. 310. ASP.

Quataert, E. & Gruzinov, A. 2000 *ApJ* **545**, 842.

Quataert, E. & Narayan, R. 1999 *ApJ* **520**, 298.

Rieke, G., Loken, K., Rieke, M., & Tamblyn, P. 1993 *ApJ* **412**, 99.

Stolovy, S., McCarthy, D., Melia, F., Rieke, G., Rieke, M., & Yusef-Zadeh, F. 1999. In *The Central Parsecs of the Galaxy* (eds. H. Falcke, A. Cotera, W. Duschl, F. Melia, & M. Rieke), p. 39. ASP.

Tamblyn, P., Rieke, G. H., Hanson, M. M., Close, L., McCarthy, D., & Rieke, M. 1996 *ApJ* **456**, 206.

Writh, M. & Backer, D. 1993 *ApJ* **417**, 560.

Zhao, J.-H. & Goss, M. 1999. In *The Central Parsecs of the Galaxy* (eds. H. Falcke, A. Cotera, W. Duschl, F. Melia, & M. Rieke), p. 224. ASP.

Stellar populations in dwarf galaxies: A review of the contribution of *HST* to our understanding of the nearby universe

By ELINE TOLSTOY†

European Southern Observatory, Karl-Schwarzschild Strasse 2, Garching bei München, Germany

This review aims to give an overview of the contribution of the *Hubble Space Telescope* to our understanding of the detailed properties of Local Group dwarf galaxies and their older stellar populations. The exquisite stable high spatial resolution combined with photometric accuracy of images from the *Hubble Space Telescope* have allowed us to probe further back into the history of star formation of a large variety of different galaxy types with widely differing star formation properties. It has allowed us to extend our studies out to the edges of the Local Group and beyond with greater accuracy than ever before. We have learned several important things about dwarf galaxy evolution from these studies. Firstly we have found that no two galaxies have identical star formation histories; some galaxies may superficially look the same today, but they have invariably followed different paths to this point. Now that we have managed to probe deep into the star formation history of dwarf irregular galaxies in the Local Group it is obvious that there are a number of similarities with the global properties of dwarf elliptical/spheroidal type galaxies, which were previously thought to be quite distinct. The elliptical/spheroidals tend to have one or more discrete episodes of star formation through-out their history and dwarf irregulars are characterized by quasi-continuous star-formation. The previous strong dichotomy between these two classes has been weakened by these new results and may stem from the differences in the environment in which these similar mass galaxies were born into or have inhabited for most of their lives. The more detailed is our understanding of star formation processes and their effect on galaxy evolution in the nearby Universe the better we will understand the results from studies of the integrated light of galaxies in the high-redshift Universe.

1. Introduction

This review is a survey of the *Hubble Space Telescope* (*HST*) studies of resolved stellar populations of nearby galaxies that have determined accurate global star formation histories going back several Gyr. The determination of such a detailed the star formation history depends upon the ability to accurately photometer many Gyr old stars. This was possible from ground based imaging for our nearest companions, the dwarf spheroidals, and the Magellanic Clouds. *HST*, with it's outstanding combination of lower sky brightness, high resolution, and a stable and constant point spread function, has allowed us to probe further out into the Local Group and even beyond this to look at a much more diverse sample of galaxies than previously possible, and it has also allowed us renewed insights into our nearest neighbors.

I am leaving out any discussion of the considerable body of literature on the *HST* observations of the stellar populations of the Magellanic Clouds (e.g. Holtzman et al. 1999; Panagia et al. 2000) which technically do not count as dwarf galaxies (e.g. Tammann 1994; Binggeli 1994) but are often assumed to be so. Because these galaxies are so large on the sky it is difficult for *HST* by itself to gain a perspective of the global star-formation

† Present address: UK Gemini Support Group, University of Oxford, Keble Road, Oxford, UK; email: etolstoy@astro.ox.ac.uk

properties, although attempts are being made (e.g. Smecker-Hane et al. 1999). There are also numerous *HST* studies of stellar populations at the small scale of individual star clusters and H II regions in the Clouds (e.g. Da Costa 1999; Massey 1999; Hunter 1999). Studies of the stellar populations of the Magellanic Clouds could easily take up an entire review, not to say a conference, in their own right. I have also neglected considerable work on resolved stellar populations of dwarfs in the UV (e.g. Brown et al. 2000; Cole et al. 1998), because the detailed connection to quantifying a star formation history is unclear. This is also true of studies of star clusters around nearby dwarf galaxies (e.g. Da Costa 1999; Hodge et al. 1999; Mighell, Sarajedini & French 1998).

1.1. *Dwarf Galaxy types*

Classification is a difficult and often emotive business, and I do not attempt to seriously address this issue, except to make it easier for me to refer to the sample of galaxies in the Local Group in broad terms, and with my particular interest in their star-forming properties. So, bearing this caveat in mind let me introduce the four different classes of dwarf galaxies that should cover anything out there:

Dwarf Irregular (dI) galaxies are arguably the most common type of galaxy in the Universe (cf. Ellis 1997), they are not clustered around larger galaxies, but appear to have a fairly random distribution through out the Local Group, and indeed in the Universe. They are usually loosely structured late-type gas rich systems with varying levels of star formation occurring in a haphazard manner across the galaxy. The velocity field of the HI gas in these systems can be dominated by random motions rather than rotation (e.g. Lo, Sargent & Young 1993, but see Skillman 1996), for the fainter dIs (e.g. Leo A; $v_{rot} \sim 5$ km s^{-1}), but for the more massive dIs (e.g. NGC 6822, Sextans A) solid body rotation is clearly seen, with amplitudes of 30–40 km s^{-1}.

Blue Compact Dwarf (BCD) galaxies are gas rich systems dominated by a region of extremely active star formation, and resembling the massive HII regions which can be found in larger galaxies. They are thought to be forming stars at a rate which they can only maintain for a short period (e.g. Searle, Sargent & Bagnuolo 1973). This type of galaxy may be a dI undergoing a period of particularly active star formation (e.g. Tolstoy 1998a). Within the Local Group, IC 10 is a fairly good approximation of what we expect a BCD to look like (see van den Bergh 2000; Hunter et al.; in preparation), and perhaps also IC 5152. The more distant BCDs could easily be embedded in larger low surface brightness galaxies, which are easier to see within the Local Group. There are several, classical, examples of BCDs just beyond the Local Group (e.g. NGC 1569 & VII Zw403).

Dwarf Elliptical (dE) galaxies are basically low luminosity Elliptical galaxies, with smooth surface brightness distributions (e.g. Ferguson & Binggeli 1994). They are typically dominated by an old stellar population, but as Baade already noticed in 1951, they are subject to the same extreme variations of stellar population as other dwarf galaxy types. Baade (1951) found that the archetypal dEs NGC 185 and NGC 205 contain B stars along with gas and dust. Recent detailed analysis suggests that several epochs of star formation over long time scales are needed to explain the characteristics of dE stellar populations (e.g. Ferguson & Binggeli 1994; Han et al. 1997). Most of the bright dEs ($M_B < -16$) appear to have nuclei, and there is some evidence that these are dynamically separate super-massive star clusters (e.g. M 54 in Sagittarius, Ibata et al. 1994; and NGC 205, Carter & Sadler 1990). dEs are strongly clustered with the largest galaxies, and four of the five dEs in the Local Group are found in proximity to M 31.

Dwarf Spheroidal (dSph) galaxies are basically low-surface brightness, non-nucleated dEs. Many argue that the dSph are merely the low luminosity tail of the dE galaxy class (e.g. Ferguson & Binggeli 1994), and the fact that they are *often* clustered around bigger

name	type	M_V	Σ_0 mag	M_{tot}	M_{tot}/M_{HI}	[Fe/H]
			arcsec^{-2}	10^6 $M_\odot$		dex
Spiral galaxies:						
M 31	Sb	−21.2	10.8	2×10^6	0.002	+0.2
Milky Way	Sbc	−20.9:		10^6	0.004	0.
M 33	Sc	−18.9	10.7	10^5	0.02	−0.2
Irregular galaxies:						
NGC 3109	Irr	−15.7	23.6±0.2	6550	0.11	−1.2 ± 0.2[a]
LMC	Irr	−18.1	20.7	6000	0.5	−0.7
SMC	Irr	−16.2	22.1	2000	0.25	−1.0
Dwarf Ellipticals:						
M 32	dE2	−16.7	<11.5:	2120	<0.001	−1.1±0.2
NGC 205	dE5	−16.6	20.4±0.4	740	0.001	−0.8±0.1
NGC 185	dE3	−15.5	20.1±0.4	130	0.001	−1.2±0.15
NGC 147	dE4	−15.5	21.6±0.2	110	<0.001	−1.1±0.2
Sagittarius	dE7	−13.4	25.4±0.3	500:	<0.0001	−1.0±0.2
Dwarf Irregulars:						
NGC 6822	dIrr	−15.2	21.4±0.2	1640	0.08	−0.7 ± 0.2[a]
IC 10	Irr	−15.7	22.1±0.4	1580	0.10	−0.7±0.15[a]
IC 1613	dIrr	−14.7	22.8±0.3	795	0.07	−1.1±0.2[a]
IC 5152	dIrr	−14.8		400	0.15	−0.6±0.2[a]
Sextans B	dIrr	−14.2		885	0.05	−1.1±0.3[a]
Sextans A	dIrr	−14.6	23.5±0.3	395	0.20	−1.4±0.2[a]
WLM	dIrr	−14.5	20.4±0.05	150	0.40	−1.1±0.2[a]
Phoenix	dIrr/dE	−10.1		33	0.006:	−1.9±0.1
Pegasus	dIrr	−12.9		58	0.09	−1.0±0.14[a]
LGS 3	dIrr/dE	−10.5	24.7±0.2	13	0.03	−1.8±0.3
Leo A	dIrr	−11.4		11	0.72	−1.6±0.15[a]
SagDIG	dIrr	−12.3	24.4±0.3	9.6	0.92	−1.5±0.3[a]
DDO 210	dIrr	−10.0	23.0±0.3	5.4	0.35	−1.9±0.12
EGB 0427+63	dIrr	−12.6	23.9±0.1			−1.4±0.1[a]
Dwarf Spheroidals:						
Fornax	dE3	−13.2	23.4±0.3	68	<0.001	−1.3±0.2
Ursa Minor	dE5	−8.9	25.5±0.5	23	<0.002	−2.2±0.1
Draco	dE3	−8.8	25.3±0.5	22	<0.001	−2.1±0.15
Leo I	dE3	−11.9	22.4±0.3	22	<0.001	−1.5±0.4
Sextans	dE4	−9.5	26.2±0.5	19	<0.001	−1.7±0.2
Carina	dE4	−9.3	23.9±0.4	13	<0.001	−2.0±0.2
Sculptor	dE	−11.1	23.7±0.4	6	<0.004	−1.8±0.1
Antlia	dE3	−10.8	24.3±0.2	12	0.08	−1.8±0.25
Tucana	dE5	−9.6	25.1±0.06			−1.7±0.15
Cetus	dE4	−10.1	25.1±0.1			−1.9±0.2
Leo II	dE0	−9.6	24.0±0.3	10	<0.001	−1.9±0.1
And I	dE0	−11.9	24.9±0.01			−1.5±0.2
And II	dE3	−11.1	24.8±0.05			−1.5±0.3
And III	dE6	−10.3	25.3±0.05			−2.0±0.2
And V	dE3	−9.1	24.8±0.2			−1.6±0.2
And VI (Peg dSph)	dE3	−11.3	24.3±0.05			−1.6±0.2
And VII (Cass dSph)	dE3	−12.0	23.5±0.05			−1.4±0.3

[a] these values were converted from [O/H] measurements, assuming constant [Fe/O] = 0.

Table 1. The Local Group

galaxies such as our Galaxy, M 31, and perhaps also NGC 3109 tends to support this. However it is also possible that at least a fraction of this class fit into a common evolutionary scenario with dIs and BCDs, their past star-formation having been dominated by bursts (e.g. Carina). There are a number of so called *transition* objects, such as Phoenix, Antlia and LGS 3 which are hard to fit unambiguously into either the dE or dI class.

They have very little or no present-day star formation, and yet they have associated HI gas, so it seems a fair assumption that they will form stars again before too long, and will then clearly belong to the dI class.

1.2. *The properties of Local Group Galaxies*

Our local neighborhood, the Local Group, is arguably as representative a piece of the Universe as any (e.g. van den Bergh 2000). What we learn about the properties of star formation and galaxy evolution here can justifiably be extrapolated to explain what we see in the distant, early Universe. That which is on our doorstep provides us with the chance to properly understand the dominant physical processes in great detail.

As with galaxy classification, there are different selection criteria which vary the number of dwarf and irregular galaxies included in a census of Local Group members. I have chosen to follow dynamical arguments (e.g. Irwin 1998), which tends to allow a slightly more distant limit to the distance from the center than van den Bergh (2000). There is no straight forward boundary between the Local Group and the Sculptor, and many of the galaxies in this region could "belong to either," by dynamical arguments. Listed in Table 1 are some basic properties of the galaxies I have assumed to be members of the Local Group. For each galaxy there is listed the absolute V magnitude (M_V), the central surface brightness in V (Σ_0), the total mass (M_{tot}), and the fraction of this total in HI gas, when I could find them. Also listed is an estimate of the mean metallicity of each galaxy, given as [Fe/H], for the sake of uniformity. Where the metallicity was "converted" from [O/H], this is noted. The rest of the values are predominantly based upon the color of the red giant branch, and thus ought to be treated as lower limits. A lot of detail is glossed over in presenting a uniform "metallicity" for a range of galaxy type, as here (cf. Skillman 1998 for much more detail on this complex topic).

Most of the data in Table 1 came from Mateo (1998), which contains a very complete collection of what is known about Local Group dwarf galaxies. Mateo also provides explanations of the error bars, and where these data originally come from. The purpose of Table 1 is largely illustrative and I would recommend anyone who would like to use the values in this table to look them up in Mateo, and even better still in the original references provided there. I also recommend the detailed descriptions of the individual galaxies in van den Bergh (2000). The data for the more massive galaxies in the Local Group: the Milky Way, M31, M33 and the LMC/SMC are extracted from various standard sources, and not meant to be definitive, merely to illustrate the scale sizes between dwarf galaxies and their large neighbors in the Local Group.

The mass of the Local Group is dominated by three large spiral galaxies, namely our Galaxy, M 31 and M 33. However, the largest population *by number* are the dwarf type galaxies (see Table 1). The Local Group contains several examples of each class of dwarf galaxy, as defined above. All dwarf galaxy types could plausibly come from the same type of progenitor but for reasons of differences in initial dark matter content, or environment or chance encounters with other galaxies follow different evolutionary paths which result in the different present-day properties (e.g. Ferrara & Tolstoy 2000; Binggeli 1994; Davies & Phillips 1988).

The total mass of a galaxy has a critical impact on the ability of that galaxy to form stars. It determines how effectively gas can be compressed to increase the efficiency of star formation and also how hard it is for supernova explosions to disrupt or even rid the system of gas delaying or preventing future star formation (e.g. Mac Low & Ferrara 1999; De Young & Heckman 1994; Dekel & Silk 1986). The lowest mass dwarf galaxies are often at the hairy edge of being able to retain their gas whilst forming stars, and many clearly have lost the battle. For the lowest mass galaxies any small perturbation

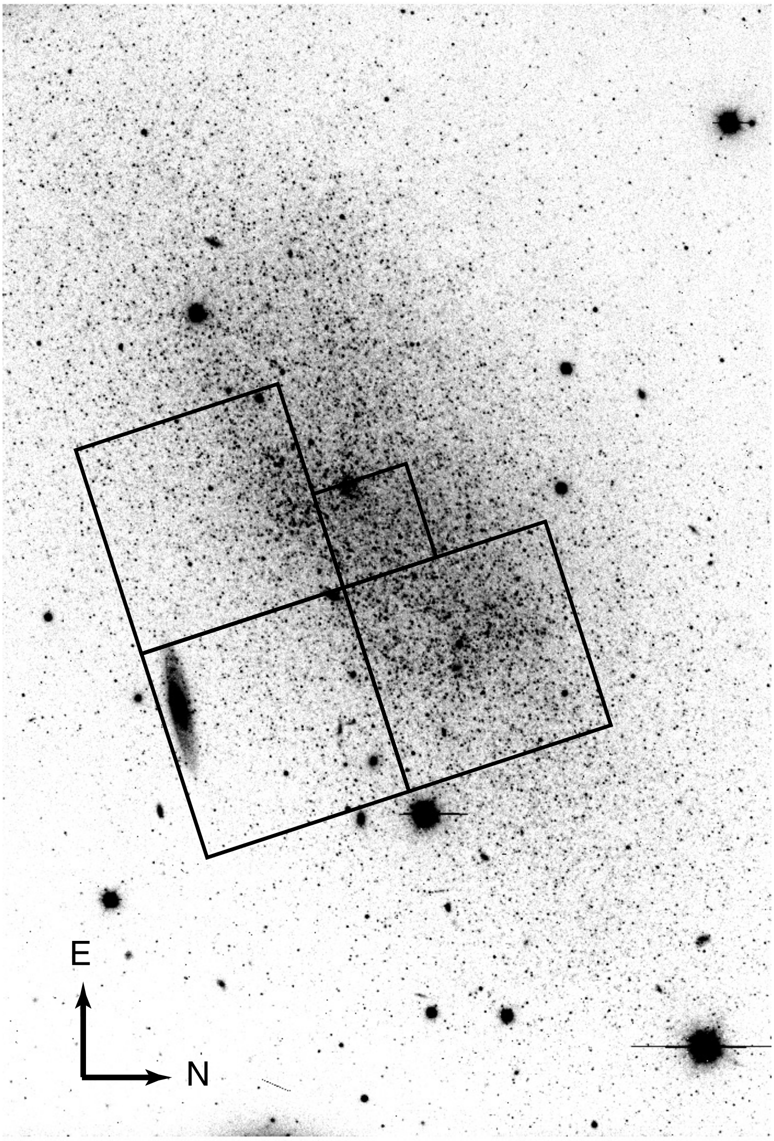


Figure 1. This is a WIYN 3.5m telescope image of the Pegasus dwarf taken in 0.″6 seeing, from Gallagher et al. (1998). This is a good example of a low surface brightness dI in the Local Group. It shows the bright core where star formation is occurring, and the more extended dE-like main body of the galaxy. The field of view shown here is 4.′8 × 6.′7 on a side, which corresponds to 1 × 1.5 kpc at the distance of Pegasus (760 kpc). The location of the WFPC2 pointing is also shown.

can have a dramatic impact on the evolution of these systems, and it is this sensitivity to initial conditions and random events which may explain some of the dispersion of the values in Table 1 (see Ferrara & Tolstoy 2000).

One of the few obvious correlations between dwarf galaxy properties is that between absolute magnitude (M_V) and global metallicity (e.g. Skillman, Kennicutt & Hodge 1989). If a reasonably uniform mass to light (M/L) ratio can be assumed, then this can be interpreted as a mass-metallicity relation. Those galaxies which fall significantly off the relation (e.g. extreme BCDs) can be assumed, with very good reasons, to have a significantly different M/L. Indeed some scatter in the M/L of average galaxies might well account for the scatter seen in the mass-metallicity relation (e.g. Ferrara & Tolstoy 2000). A large uncertainty in proving or disproving this is the lack of reliable measurements of total (or dynamical) masses of dwarf galaxies.

Basically, the global star-formation properties of a galaxy appear to be dominated by the total mass. This may perhaps vary when galaxies interact and merge, but it is difficult to disentangle the effects of merging two small galaxies to form a bigger one and the temporary boost to the star formation rate from the merger. These two effects might balance each other—because taken at face value the luminosity/mass-metallicity relation seems to say that a galaxy knows how big it is from the earliest times. If a lot of dwarfs were added together they would retain the global metallicity of the original pieces, whilst increasing in luminosity. This simple arithmetic could also argue against merging having a significant impact upon Local Group galaxies, at least in recent times.

2. Why study dwarf galaxies?

As demonstrated in Table 1 Local Group dwarf galaxies have a wide range of different properties. They span a large *mean* metallicity range, down the lowest seen anywhere. They also exhibit a range of gas fractions, and density, from no gas all the way to gas dominated. They are also to be seen in a range of proximities to other systems of varying mass. Thus a study of the dwarf galaxy members of the Local Group allows us to study star formation over a large range of initial conditions. When the star formation properties of the Local group galaxies are looked at together (e.g. Mateo 1998; Da Costa 1998; Grebel 1998), the only global statement that can be made with no fear of contradiction is that no two are exactly alike.

The smaller dwarf galaxies effectively have a single-cell mode of global star formation, which in principle ought to be more straight forward to comprehend than larger galaxies where different star formation regions can apparently be unaware of each other or interfere strongly with each other, or anything between these extremes. This, including spiral density waves, bars, jets and other dynamical effects can create severe complications to the straight forward interpretation of the relationship between gas and stars and star formation. Although, as can be seen by the results presented in this review, even these small galaxies are capable of a high degree of complexity, so "straight forward" is a very relative statement.

Dwarf galaxies have the added benefit for *HST* observations, that they are small enough, at the modest distances typical for Local Group members that a significant fraction of the surface area of a dwarf can typically fit into the WFPC2 field of view (see Figure 1).

The Local Group contains galaxies whose star formation histories should be typical of galaxy group members, and thus of star formation throughout the Universe. It must therefore include remnants of the epoch ∼ 5–8 Gyr before the present when actively star forming galaxies produced the faint blue galaxy population seen at intermediate redshifts. We can directly measure star formation histories of nearby galaxies back to the era of faint blue galaxies with sufficiently deep and accurate imaging and using established quantitative techniques for analyzing color-magnitude diagrams (e.g. Tosi et al. 1991; Tolstoy 1996; Aparicio et al. 1996).

Consistent with the properties of faint blue galaxies in redshift surveys, dwarf galaxies appear to have erratic star formation rates, and they can host bright, short lived *bursts* of star formation which could make these currently dim and inconspicuous galaxies dominate the luminosity of the Local Group for short periods of time. Because dwarfs are so

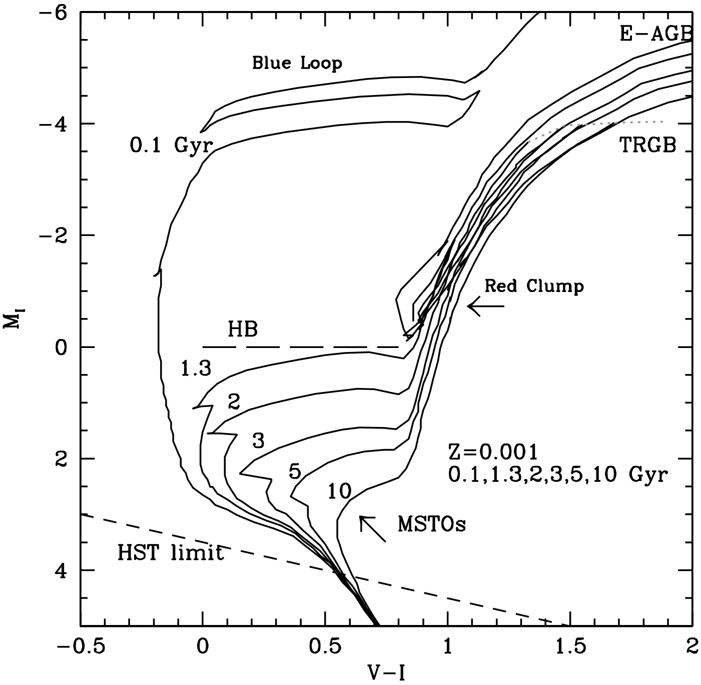


FIGURE 2. Here are plotted theoretical isochrones (from Bertelli et al. 1994) for a single metallicity ($Z = 0.001$, or [Fe/H] $= -1.3$) and a range of ages, as marked in Gyr at the MSTOs to illustrate the CMD features discussed in §3. Also plotted is the sensitivity limit that *HST* can reach for all the galaxies in the Local Group with a modest investment of time.

numerous, it only requires each galaxy to burst once or twice in its life time before the dwarfs in the Local group are effectively always visible in redshift surveys. This means that these small galaxies could be dominating redshift surveys in the intermediate redshift range (e.g. Lilly et al. 1996). The ubiquitous faint blue galaxies seen in deep imaging and spectroscopic surveys at intermediate redshifts could be a population of dwarf galaxies.

3. Color-Magnitude Diagram analysis: How to study dwarf galaxies

The study of resolved stellar populations provides a powerful tool to follow galaxy evolution consistently and directly in terms of physical parameters such as age (star formation history), chemical composition and enrichment history, initial mass function (IMF), environment, and dynamical history of the system. Photometry of individual stars in at least two filters and the interpretation of Color-Magnitude Diagram (CMD) morphology gives the least ambiguous and most accurate information about variations in star formation within a galaxy back to the oldest stars (see Figure 2). Some of the physical parameters that affect a CMD are strongly correlated, such as metallicity and age, since successive generations of star formation may be progressively enriched in the heavier elements. Careful, detailed CMD analysis is a proven, uniquely powerful approach (e.g. Tosi et al 1991; Tolstoy & Saha 1996; Aparicio et al. 1996; Mighell 1997; Dohm-Palmer et al. 1997, 1998; Gallagher et al. 1998; Tolstoy et al. 1998; Tolstoy 1998a) that benefits enormously from the high spatial resolution of *HST*.

3.1. *Useful features in a Color-Magnitude diagram*

Stellar evolution theory provides a number of clear predictions, based on relatively well understood physics, of features expected in CMDs for different age and metallicity stellar populations. There are a number of clear indicators of varying star formation rates at different times which can be combined to obtain a very accurate picture of the entire star formation history of a galaxy (see Figures 2 and 3). Here I provide a brief description of each of the separate indicators, in order of preference. The indicators are thus presented in an order which broadly represents the ease with which age and metallicity information can be extracted.

3.1.1. *Main Sequence Turnoffs (MSTOs)*

The Main Sequence is a well understood mass-luminosity-lifetime relation, which allows us to extract (relatively) unambiguous information about the star formation rate with time over the lifetime of a galaxy. With exposures (going down to $M_V \sim +4$) of the resolved stellar populations in nearby galaxies we can obtain the *unambiguous age information that comes from the luminosity of MSTOs* back to the oldest ages. The MSTOs do not overlap each other and hence provide the most direct, accurate information about the star formation history of a galaxy (see Gallart et al. 1999). MSTOs can clearly distinguish between bursting star formation and quiescent star formation. The age resolution that is possible does vary, becoming coarser going back in time, and can also be affected by metallicity evolution. Our ability to disentangle the variations in star formation rate depends upon the intensity of the past variations and how long ago they occurred and which filters are used for observation.

3.1.2. *The Core-Helium Burning Blue Loop Stars (BLs)*

Stars of low metallicity and intermediate mass go on extensive “Blue Loop” (BL) excursions after they ignite He in their core. Stars in the BL phase are several magnitudes brighter than when on the main sequence ($M_V \lesssim -1$). The shape of these “loops” are a strong function of metallicity and age. They thus provide a more luminous opportunity to accurately determine the age and metallicity of the young stellar population (in the range, $\sim$ 1 Gyr old) in nearby low metallicity galaxies The luminosity of a BL star is fixed for a given age, and thus subsequent generations of BL stars do not over-lie each other, and can be used to trace *spatial* variations in recent star formation over a galaxy (e.g. Dohm-Palmer et al. 1997b). The lower the metallicity of the galaxy, the older will be the oldest BLs and the further back in time an accurate spatially resolved star formation history can easily be determined.

3.1.3. *The Red Giant Branch (RGB)*

The RGB is a bright evolved phase of stellar evolution, where the star is burning H in a shell around its He core. It is characterized by a fairly constant maximum (or tip) luminosity (at $M_I = -4.$), and stars are distributed all the way down to $M_I \sim +2$). Metallicity is the most important effect in determining the width of the RGB in color, especially for ages > 2 Gyr. However, correlations between age and metallicity can mask a metallicity spread, as $\sim$ 4 Gyr of age difference can produce the same effect as 0.1 dex of metallicity difference, in (V−I). For a *given* metallicity the RGB blue and red limits are given by the age spread of the stars populating it (ages $\gtrsim$1 Gyr), because as a stellar population ages the RGB moves to the red. However increasing the metallicity of a stellar population will also make the RGB redder, and thus produce the same effect as aging. This is the (in)famous age-metallicity degeneracy problem. The result is that if there is metallicity evolution within a galaxy, it is impossible to uniquely disentangle effects due to age and metallicity on the basis of the optical colors of the RGB alone.

3.1.4. *The Red Clump/Horizontal Branch (RC/HB)*

Red Clump (RC) stars ($M_V \sim +0.5$), low-mass analogues of the BL stars, and their lower mass cousins the Horizontal Branch (HB) stars ($M_V \sim 0.$) are core He-burning stars, and they don't obey a simple mass-luminosity law, as their core mass is mostly independent of their total mass. Their luminosity and color varies depending upon age, metallicity and mass loss (Caputo, Castellani & degl'Innocenti 1995). The extent in luminosity of the RC can be used to estimate the age of the population that produced it (see Tolstoy 1998b). This age measure is *independent of absolute magnitude and hence distance.*

The classical RC and RGB appear in a population at about the same time (after $\sim$ 0.9–1.5 Gyr, depending on model details), where the RGB are the progenitors of the RC stars. The lifetime of a star on the RGB, $t_{\rm RGB}$, is a strongly decreasing function of $M_{\rm star}$, but the lifetime in the RC, $t_{\rm RC}$ is roughly constant. Hence the ratio, $t_{\rm RC}$ / $t_{\rm RGB}$, is a decreasing function of the age of the dominant stellar population in a galaxy, and the ratio of the numbers of stars in the RC, and the HB to the number of RGB is sensitive to the star formation history of the galaxy (e.g. Cole 1999; Gallagher et al. 1998; Tolstoy et al. 1998; Han et al. 1997). Thus, the higher the ratio, N(RC)/N(RGB), the younger the dominant stellar population in a galaxy.

The presence of a large HB population on the other hand (high N(HB)/N(RGB) or even N(HB)/N(MS)), is caused by a predominantly much older ($>$ 10 Gyr) stellar population in a galaxy. The HB is the brightest unambiguous indicator of very lowest mass (hence oldest) stellar populations in a galaxy, it is however impossible to use it to infer star formation rates at these ancient epochs, because of the "second parameter effect" (e.g. Fusi Pecci & Bellazzini 1997), which decouples the HB lifetimes from initial conditions, is well known, but not yet understood, from globular cluster studies.

3.1.5. *The Extended Asymptotic Giant Branch (EAGB)*

Extended Asymptotic Giant Branch (EAGB) stars are very bright, red evolved stars ($M_V > -4.$ and typically $V - I > 1.5$). The temperature and color of the EAGB stars in a galaxy are determined by the age and metallicity of the population they represent. However there remain a number of uncertainties in the comparison between the models and the data (e.g. Gallagher et al. 1998; Lynds et al. 1998). It is necessary that more work is done to enable a better calibration of these very bright indicators of past star formation events. The future of this field probably lies in infra-red observations of these stars.

3.2. *Monte-Carlo simulations of CMDs*

One of the most impressive advances in interpreting observed CMDs in terms of a detailed star formation history has come from the technique of re-creating an observed CMD from model stellar evolution tracks by means of Monte-Carlo simulations. This technique has the advantage that it can account for the many uncertainties which plague our understanding of a CMD in what is arguably the most physically realistic manner. This approach was pioneered by Tosi and co-workers in Bologna (e.g. Tosi et al. 1991), and has since been used and adapted as the standard method of CMD analysis (e.g. Tolstoy & Saha 1996; Aparicio et al. 1996; Dolphin 1997; Hernandez et al. 1999).

The main uncertainties in the interpretation of CMDs of nearby galaxies come from: estimates of the distance of the galaxy; the foreground and internal extinction, both the

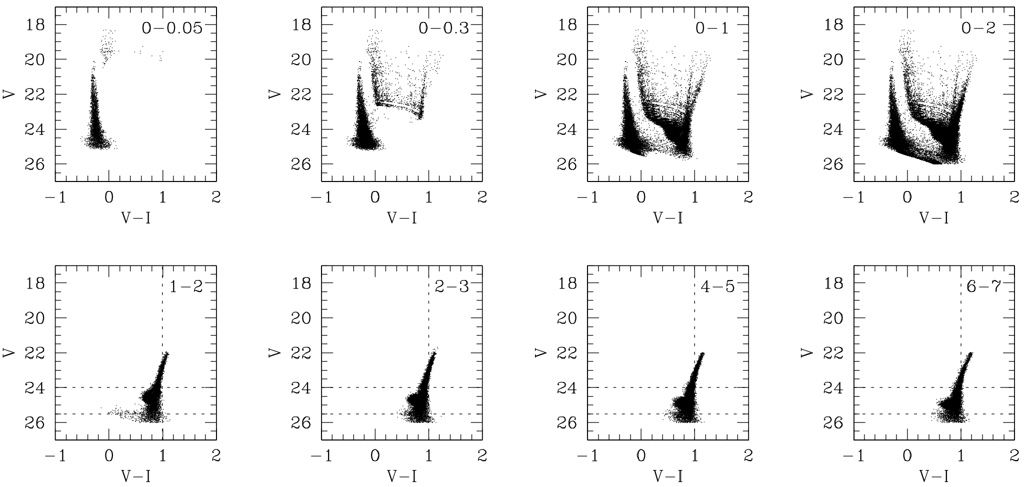


FIGURE 3. Model CMDs, from Monte-Carlo simulations, assuming a standard IMF. based upon theoretical stellar evolution tracks at metallicity, $Z = 0.0004$ (or [Fe/H] $= -1.7$), from Fagotto et al. (1994). The periods of star-formation are marked in the top right-hand corner of each figure in Gyr. Each period is assumed to have a constant star formation rate. All the models are made for the sensitivity of WFPC2 over two orbits (one in F555W, one in F814W) to a resolved stellar population at a distance modulus, (m-M)$_o$ $= 24.2$ (or 700 kpc), which is about the mean for dwarf galaxies in the Local Group.

absolute values, and the patchiness can be uncertain; metallicity, and how this might vary in time within the galaxy; the initial mass function, what it is and if it might vary; the fraction of binary stars in the CMD; photometric errors, or the accuracy with which measurements can be made; incompleteness, which is a measure of how the number of stars detected per resolution element affects both the determination of photometric errors and the number of stars of different luminosities that will be hidden behind and in the wings of brighter neighbors; and last but not least and perhaps most difficult of all the uncertainties in the theoretical models of stellar evolution. One problem for the modelers is to find useful data sets to compare with models. Globular clusters are excellent test data for checking low metallicity and very old models, and open clusters are mostly quite metal rich. Dwarf galaxies tend to be dominated by intermediate and even young metal-poor stellar populations, and so it is hard to find fiducials. The result is that we are forced to try and understand what is happening with an uncertain star formation history and uncertain model effects all at the same time. Any one of the uncertainties listed here could (and have) produced (long) papers in their own right. They can have profound effects on the star formation history determined from a CMD.

So, clearly a great deal of care has to be taken not to over-interpret CMDs, because having so many factors which affect the modeling means that there is a lot of parameter space to explore, and the effect of all the uncertainties is to smear out features which can result in a very shallow minimum to any χ^2 estimate or Likelihood function to find the best model. Each change of initial assumptions requires generation of a complete new set of models and goodness of fit assessment. This can be computationally intensive and, in the end, it can be difficult to find a unique solution. Best solutions are only ever one out many that are possible. It is thus important to tie in star formation rate variations to distinct and well determined features in a CMD, and to justify all variations that are seen. This is especially true in older populations where very small and subtle changes can

Dwarf Ellipticals:	
M 32	Grillmair et al. 1996
NGC 205	Jones et al. 1996
NGC 185	Geisler et al. in preparation
NGC 147	Han et al. 1997
Sagittarius	Mighell et al. 1997
Dwarf Irregulars:	
NGC 6822	Wyder et al. 2000
IC 1613	Cole et al. 1999; Dolphin et al. 2000
IC 5152	snap-shot in archive
IC 10	Hunter et al. in preparation
Sextans B	no data
Sextans A	Dohm-Palmer et al. 1997a,b
WLM	Dolphin 2000; Rejkuba et al. 2000
Phoenix	Holtzman et al. 2000
Pegasus	Gallagher et al. 1998
LGS 3	Miller et al. in preparation
Leo A	Tolstoy et al. 1998
SagDIG	no data
DDO 210	no data
EGB 0427+63	no data
Dwarf Spheroidals:	
Fornax	Buonanno et al. 1999
Ursa Minor	Mighell & Burke 1999; Feltzing et al. 1999; Hernandez et al. 2000
Draco	Grillmair et al. 1998
Leo I	Gallart et al. 1999; Hernandez et al. 2000
Sextans	no data
Carina	Mighell 1997; Hernandez et al. 2000
Sculptor	Monkiewicz et al. 1999
Antlia	no data
Tucana	Seitzer et al. in preparation
Cetus	no data
Leo II	Mighell & Rich 1996; Hernandez et al. 2000
And I	Da Costa et al. 1996
And II	Da Costa et al. 2000
And III	Da Costa et al. in preparation
And V	Armandroff et al. in preparation
And VI	Armandroff et al. in preparation
And VII	snap-shot in archive

Table 2. *HST* CMDs of Local Group dwarf galaxies

have dramatic impacts on assumed age and metallicity variations (cf. Tolstoy & Saha 1996; Tolstoy et al. 1998), and see Figure 3.

In Figure 3, I have created a series of model CMDs (see Tolstoy 1996), which, for a constant metallicity, show the variations that age can give to a CMD distribution. In the younger populations this is obvious, but as older populations dominate it takes very accurate photometry to distinguish between dramatically different ages of stellar populations. If metallicity is involved this becomes even more tricky because age and metallicity can produce very similar effects on an RGB (see §3.1.3). The CMDs in Figure 3 are made assuming a galaxy at 700 kpc, and only one orbit of integration time in each filter. This means there is no MSTOs older than about 800 Myr, which explains why

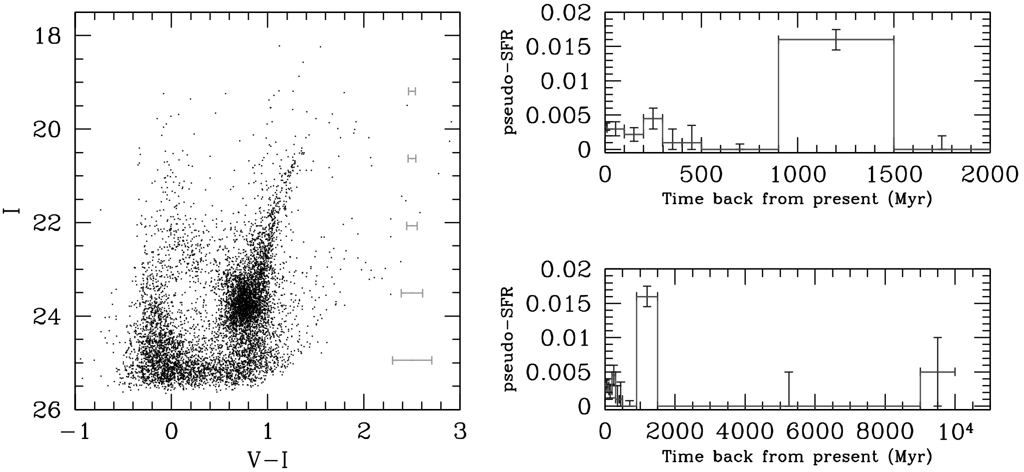


FIGURE 4. A Leo A (I, V−I) CMD from two orbits of WFPC2 observations (one in filter F555w and one in F814W), from Tolstoy et al. (1998). Also shown is a possible star formation history from detailed modeling of the CMD, and consistent with the known details of the stellar population, such as the luminosity spread of the RC; the width of the RGB; and the presence of BLs. The error bars shown are not in any specific sense statistical, they merely give an indication of how flexible the star formation rate at any time can be before significantly affecting the good match of this model to the data (see Tolstoy et al. for details).

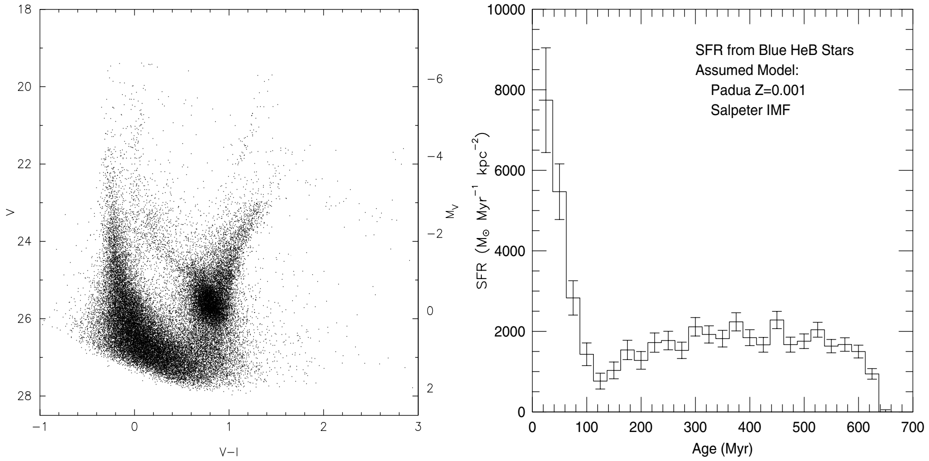


FIGURE 5. On the right-hand side is the new (V, V−I) CMD of Sextans A from Dohm-Palmer et al. 2000, in preparation. It consists of 8 orbits in F555W and 16 in F814W, and shows the presence of the RC and goes further down the Main Sequence than the previous 2-orbit CMD of Dohm-Palmer et al. (1997). On the left-hand side is the recent star formation history of Sextans A determined from the BL stars by Dohm-Palmer et al. (1997).

there is such similarity between CMDs which have differing proportions of population older than about 800 Myr.

Despite this dramatic list of problems, surprisingly detailed and robust results have come out from the analysis of CMDs. In the next section I will illustrate some of the most impressive CMDs and their analyses which have come from *HST* data for all the different classes of dwarf galaxy, as defined in §1.1.

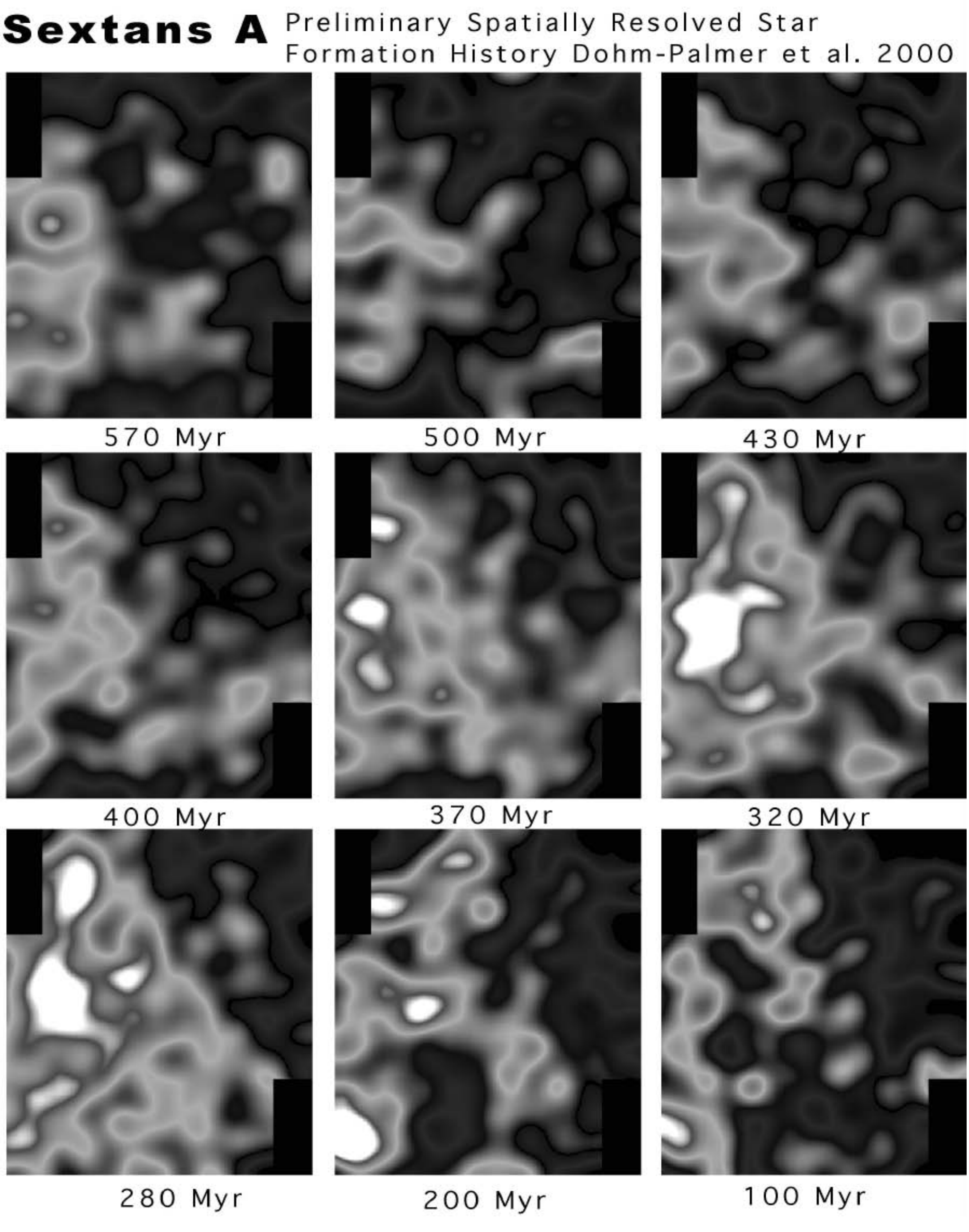


FIGURE 6. Here we see a gallery of movie still frames created by Robbie Dohm-Palmer (cf. Dohm-Palmer et al. 1997), which show how the intensity of star formation varies spatially with time across Sextans A. These are preliminary results, which combine the two WFPC2 pointings on Sextans A so that the spatial coverage of nearly the whole luminous center of the galaxy is achieved. The time separation between each frame is not uniform, and was chosen to high-light interesting features. See Dohm-Palmer et al. 1997; 2000 for more details. The newer data is on top, and despite the presence of young blue stars and H-α, this does not seem to be representative of a global increased star formation rate in this part of the galaxy, unlike the lower area which has vigorous star formation over the last several hundred million years.

4. *HST* observations of Local Group dwarf galaxies

There have been a number of spectacular results from *HST* imaging of the resolved stellar populations in Local Group dwarf galaxies (see Table 2). In the next sections I provide examples of some of the most beautiful *HST* CMDs for a selection of dwarf galaxies in the Local Group, and just beyond:

4.1. *Leo A*

Leo A (DDO 69) is a gas-rich dI galaxy, with an extremely low HII region abundance ($\sim$ 3% solar, van Zee, Skillman & Haynes 1999). The interpretation of a CMD from two orbits of WFPC2 data (Tolstoy et al. 1998; see Figure 4) is based upon extremely low metallicity ($Z = 0.0004$; or $[\mathrm{Fe/H}] = -1.7$) theoretical stellar evolution models (Fagotto et al. 1994; see also Figure 3), which suggest that this galaxy is predominantly young, i.e. $<$ 2 Gyr old. A major episode of star formation 900–1500 Gyr ago can explain the RC luminosity and also fits in with the interpretation of the number of anomalous Cepheid variable stars seen in this galaxy. The presence of an older, underlying globular cluster age stellar population could not be ruled out with these data, however, using the currently available stellar evolution models, it would appear that such an older population is limited to no more than 10% of the total star formation to have occurred in the center of this galaxy. Theoretical models of the chemical evolution of dwarf galaxies by Ferrara & Tolstoy (2000) imply that, even though this galaxy is extremely metal-poor, an underlying older stellar population is required to build up the current metallicity. Perhaps this older population resides in an outer halo. Of course, neither the chemical evolution models nor the existing CMDs can distinguish between an old population which formed in a large burst, or more sedate and roughly constant rate through-out a longer time.

4.2. *Sextans A*

Sextans A (DDO 75) is a gas-rich dI galaxy with a low metal abundance ($[\mathrm{Fe/H}] \sim -1.4$), and active star formation, which is located on the periphery of the Local Group (1.4 Mpc away). The initial WFPC2 CMD of Sextans A, based on two orbits of telescope time, shows several clearly separated populations that align well with stellar evolution model predictions for a low metallicity system (see Dohm-Palmer et al. 1997a,b). This was the first time a BL sequence had been so definitively identified in a CMD (see the CMD in Figure 5). The star formation history from the main sequence and BL stars (in Figure 5) was determined by Dohm-Palmer et al. for the last 600 Myr using theoretical stellar evolution models. The spatial distribution of the BL stars was then used to determine the spatial variation of the star formation across Sextans A with time (see Figure 6). Figure 6 is a preliminary result including new WFPC2 data which covers the whole galaxy (Dohm-Palmer et al. 2000, in preparation). The modeling concludes that in the past 50 Myr, Sextans A has had an average star formation rate that is $\sim$ 10 times that of the average over the history of the galaxy. This current activity is highly concentrated in a young region in the South-East roughly 25 pc across. This coincides with the brightest HII regions and the highest column density of HI. Between the ages of 100 and 600 Myr ago, the star formation has been roughly constant at slightly above the average value. There are regions (200–300 pc across) with a factor of $\sim$ 5 enhancement in star formation rate with a duration of 100–200 Myr.

4.3. *IC 1613*

IC 1613 is a Magellanic type dI galaxy, with young stars of SMC-like metallicity ($\sim$ 10% solar), which was first resolved into stars by Baade (1928), it was later used by Baade (1963) to illustrate the archetypal "Baade's sheet" of underlying RGB (population II)

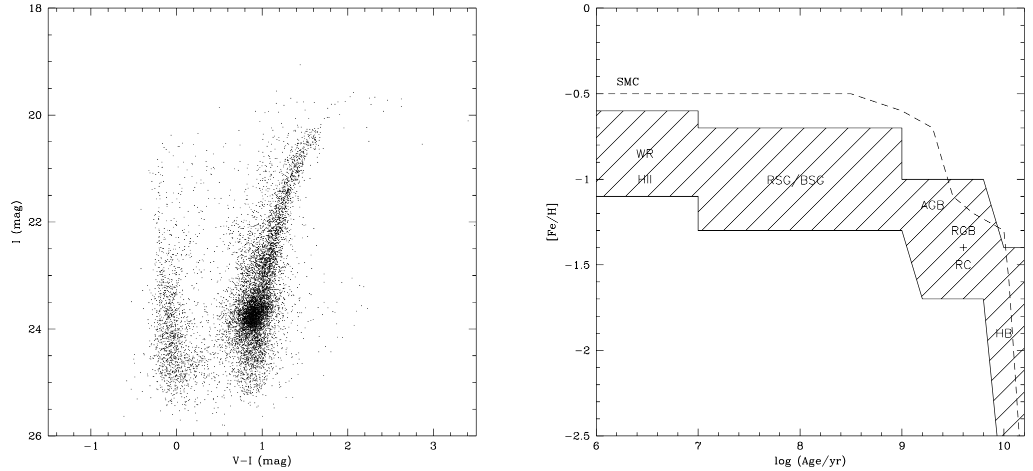


FIGURE 7. The IC 1613 (I, V−I) CMD from 8 orbits of WFPC2 observations (four in filter F555w and four in F814W) of a field right in the center of IC1613, away from any bright HII regions, from Cole et al. (1999). This was a fairly crowded field and fairly brutal cuts in S/N were made for this CMD. Also shown is a possible age-metallicity relation for IC 1613, also from Cole et al. Each region of the figure is labeled with the CMD feature that constrains the metallicity to lie within the shaded region. Abbreviations are given in the text; and in addition WR denotes the Wolf-Rayet star; HII denotes the H II regions; RSG and BSG are red and blue super giants. The dotted line shows the mean age-metallicity relation for the SMC.

stars, which have now been found to exist in most, if not all Local Group dwarfs (e.g. Sandage 1971; Hodge 1986; Saha 1995). IC 1613 is to date the only Local Group dI (excluding the Magellanic Clouds), in which RR Lyrae variables have been detected (Saha et al. 1992; Dolphin et al. 2000).

This spatially extended galaxy has had two pointings with WFPC2, one in the central region (Cole et al. 1999, see Figure 7) and one in the outskirts (Dolphin et al. 2000, Figure 8). The two CMDs, despite being in quite different environments within the galaxy, look remarkably similar.

The main-sequence luminosity function provides evidence for a roughly constant star formation rate of ($\sim 3.5 \times 10^{-4}$ M$_\odot$ yr^{-1} across the central WFPC2 field of view (0.22 kpc^2)) during the past $\sim$ 250–350 Myr, and going back to $\sim$ 1 Gyr in the outer field (Tolstoy et al. in preparation). Structure in the BL function implies that the star formation rate was $\sim$ 50% higher 400–900 Myr ago than today. The blue HB was also detected in both IC 1613 pointings, again showing that the ancient stars in this galaxy are uniformly distributed in the inner and outer regions of IC 1613. It was already known that IC 1613 contains a population of RR Lyrae variable stars (Saha et al. 1992) in its outer halo, and hence an ancient stellar population, but this is the first time that a deep enough CMD was made to detect the HB. From the different populations identified and modeled, an approximate age-metallicity relation for IC 1613 was determined (see Figure 7), which appears, like the present day metallicity, to be similar to that of the SMC.

The natural *HST* orbital cadence of 90 minutes is ideal for the detection of short period variable stars, such as RR Lyrae. As the second WFPC2 pointing in the outskirts of IC 1613 consisted of 8 orbits in F555W and 16 orbits in F814W this data set contains just enough distinct observations to be able to identify and classify short period variables in the WFPC2 field of view (Dolphin et al. 2000). These are plotted on top of the deeper CMD of the outer region in Figure 8. There are 13 RR Lyraes, which unambiguously

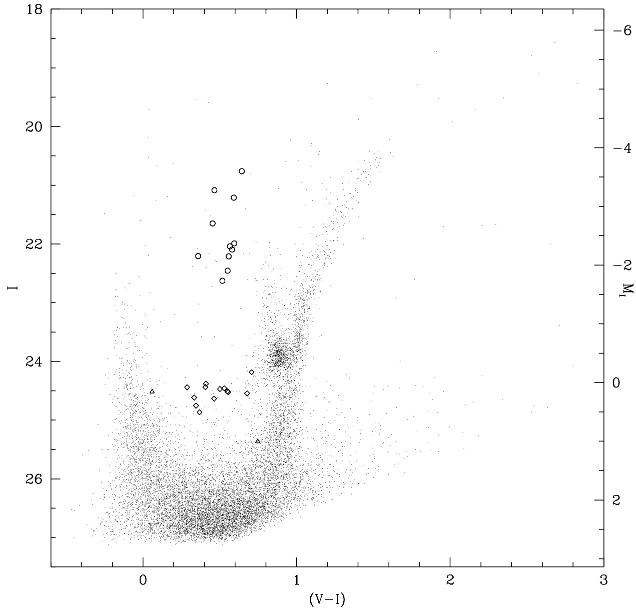


FIGURE 8. The WFPC2 (I, V−I) CMD of the outer field of IC 1613, showing the variable stars found there (from Dolphin et al. 2000), coming from 24 orbits, 16 in F814W and 8 in F555W. The circles represent the Cepheids, the diamonds the RR Lyraes, and the triangles the two possible eclipsing binaries.

mark out the presence of the modestly populated HB, and also 11 short-period Cepheids, which are indicators of an intermediate-age population in IC 1613.

4.4. *WLM*

WLM (Wolf-Lundmark-Melote; DDO 221) is another large Local Group Magellanic dI galaxy, with approximately SMC-like present-day metallicity. It is the only dI galaxy in the Local Group with a bona-fide globular cluster (see Hodge et al. 1999).

The *HST* pointing on WLM in September 1998 must rate as an extremely efficient one. The WFPC2 PC chip contained the globular cluster, and was used for a study of this unique object (Hodge et al. 1999). The remaining 3 WF chips contained stars from the field population of WLM, and were used by Dolphin (2000) to study the star formation history of the field population of this galaxy. The RC is clearly detected (see Figure 9), it is less clear if there is an HB present. The star formation history and the corresponding metal enrichment history which Dolphin determined from the WFPC2 CMD are also shown in Figure 9. The models are most reliable for the young and intermediate-age populations. Finally, as part of a targeted STIS parallel survey program, STIS images were also taken in parallel to the WFPC2 primary observations in both broad band filters available (Clear [CL] and Long Pass [LP], see Gardner et al. 1998), at a position about $4'$ away from the WFPC2 position. The STIS CCD in combination with these extremely broad filters means the images go fainter than the WFPC2, however the field of view is also a lot smaller ($\sim 25'' \times 50''$). It had been shown theoretically that these (very broad) STIS filters can be usefully transformed to an effective V and I (Gregg & Minniti 1997), and this has been confirmed by the results of Rejkuba et al. 2000 (see Figure 10) using these parallel data from WLM. They clearly detect a RC (as Dolphin 2000, in the inner WFPC2 field of view does), but unlike Dolphin they clearly detect a HB. This is due

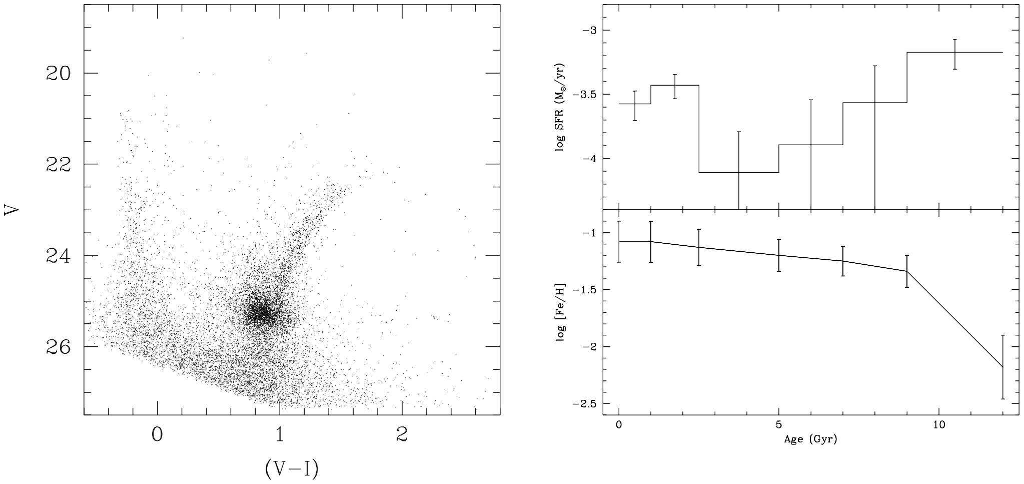


FIGURE 9. The WFPC2 (V, V−I) CMD of WLM, the combined results from photometry on the 3 WF chips. The data were taken over 4 orbits, 2 in both F814W and F555W, from Dolphin (2000). Also shown are the star formation history and the chemical evolution history that Dolphin derives from these data.

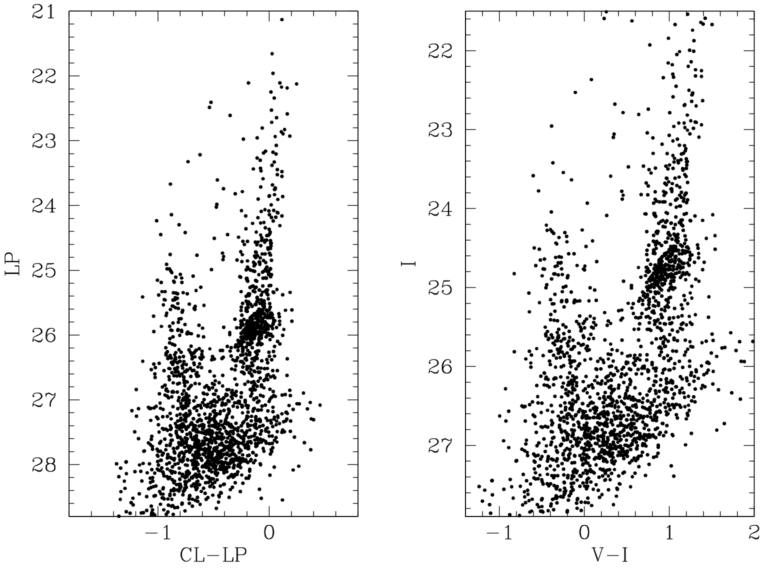


FIGURE 10. Here are the STIS/CCD imaging results for WLM from Rejkuba et al. (2000). On the left is the CMD in the raw STIS photometric system (LP, CL–LP) and on the right is the CMD after conversion to a standard photometric system (I, V−I). These data were taken in parallel to the WFPC2 data shown in Figure 9.

to the increased sensitivity of the STIS CCD; the Dolphin WFPC2 CMD barely reaches the HB magnitude, whereas the STIS CMD clearly goes beyond.

4.5. *Leo I*

Leo I is a nearby (250 kpc) relatively isolated dSph. It has been poorly studied from the ground because of the proximity of the 1st magnitude star, Regulus. However WFPC2 has produced one of its most detailed CMDs from Leo I (Gallart et al. 1999; see Figure 11). The CMD goes down to an absolute magnitude, M_V = +4.5 on the Main Sequence.

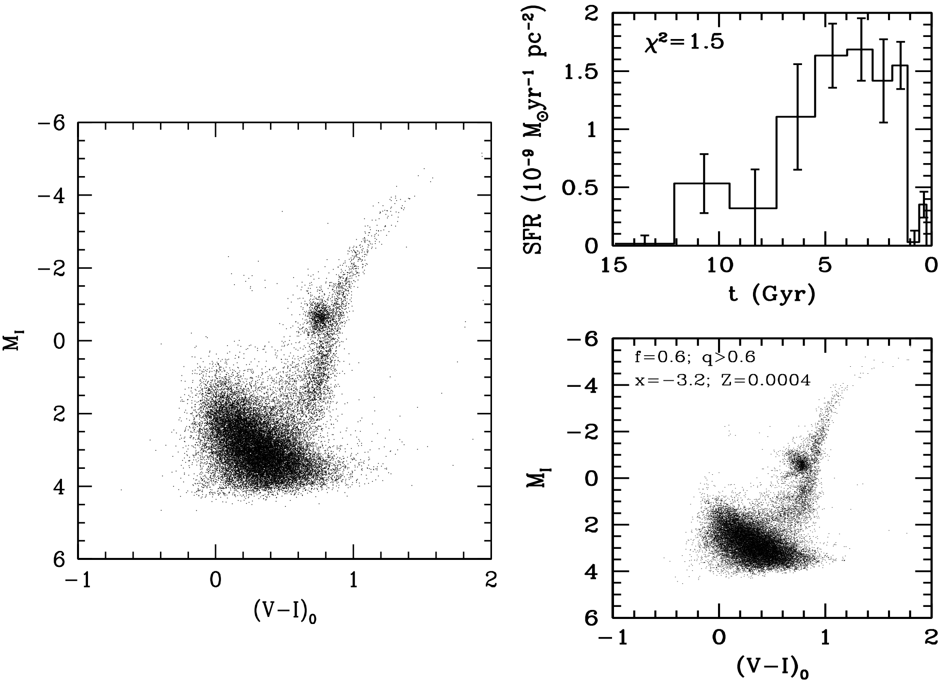


FIGURE 11. The WFPC2 (I, V−I) CMD of Leo I is shown here, assuming a distance modulus of (m-M) = 22.18, from Gallart et al. (1999). These data come from 6 orbits of exposure time, 3 in both F555W and F814W. Also shown is the star formation history that Gallart et al. found to be the best match to the data, and the resulting model CMD form this history. They come from a model with metallicity $Z = 0.0004$ ([Fe/H] = −1.7), an IMF slope, $x = -3.2$, and a carefully determined binary fraction, see Gallart et al. for more details.

This depth of CMD allows a very accurate star formation history to be determined from MSTOs (Gallart et al. 1999, see Figure 11). They found that most (70–80%) of star formation activity in Leo I occurred between 1 and 7 Gyr ago. A fairly uniform star formation rate dropped dramatically about a Gyr ago, and around 300 Myr it seems to have stopped altogether. It is not clear from these results whether or not Leo I contains an ancient (> 10 Gyr old) stellar population. It does not have an obvious HB, and it is one of the few dSph not to have a detected RR Lyrae population but a very large population of anomalous Cepheids (Lee et al. 1993). This is consistent with the Gallart et al. models which find that this galaxy is dominated by an intermediate-age, metal poor, stellar population.

4.6. *Andromeda II*

Andromeda II (And II) is a dSph companion to M 31. The WFPC2 CMD (Da Costa et al. 2000; see Figure 12) shows a predominantly red HB (like in most other dSph). In And II there is no evidence for a radial gradient, unlike, And I (Da Costa et al. 1996), or NGC 147 (Han et al. 1997). In And II Da Costa et al. also detect probable RR Lyrae variable stars, although the number of images taken of And II are not sufficient to accurately classify the variables found, but those identified with the colors of the HB can safely be assumed to be RR Lyrae variables. To interpret the And II CMD in terms of a star formation history Da Costa et al. have used a combination of standard Galactic globular cluster CMDs scaled to reproduce the And II mean abundance and abundance

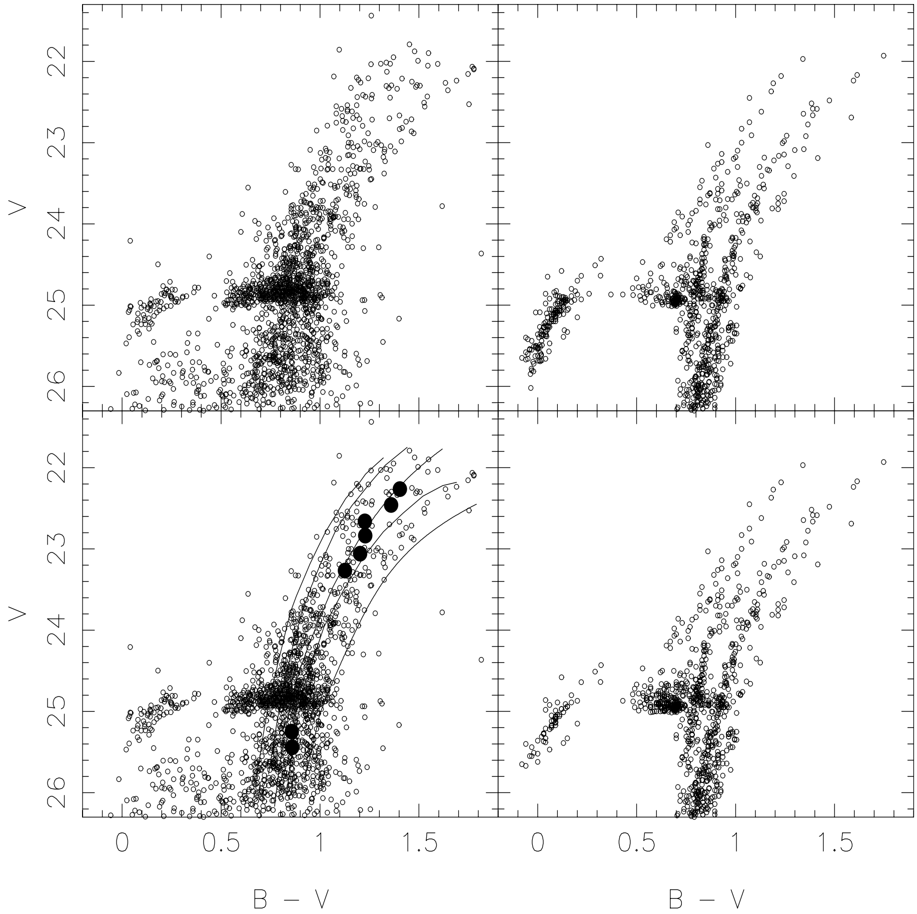


FIGURE 12. Here are presented, on the left-hand side, the WFPC2 CMDs for 11 orbits of exposure time on Andromeda II transformed to the standard (V, B−V) magnitude system, from Da Costa et al. (2000). The candidate variables have been excluded from these plots. The lower left-hand CMD is identical to the upper, but shown superposed with the giant branches of the standard globular clusters M68 ([Fe/H] = −2.09), M55 ([Fe/H] = −1.82), NGC 6752 ([Fe/H] = −1.54), NGC 362 ([Fe/H] = −1.28), and 47 Tuc ([Fe/H] = −0.71). The filled symbols give the mean And II RGB colors in ±0.1 V magnitude bins. On the right-hand side, the upper panel is a composite CMD made up from observed CMDs of the Galactic globular clusters M 55 ([Fe/H] = −1.82), NGC 1851 ([Fe/H] = −1.29), and 47 Tuc ([Fe/H] = −0.71), with the relative star numbers (44/45/11), scaled to reflect the And II abundance distribution. This CMD clearly has relatively more blue HB stars than And II has. In the lower right-hand panel all the NGC 1851 blue HB and 40% of the M 55 blue HB stars have been replaced with red HB stars from NGC 362 (80%) and 47 Tuc (20%), and the HB morphology of the "model" CMD is now a better match to that which is observed. See Da Costa et al. for many more details.

dispersion, to interpret the observed HB morphology (see Figure 12). They find that at least 50% of the total stellar population must be younger than the age of the globular clusters. This inference is strengthened by the small number of upper-AGB carbon stars, and the relatively faint luminosities ($M_{\rm bol} \sim -4.1$) of these stars. These upper-AGB carbon stars have assumed ages of around 6–9 Gyr, whilst the existence of blue HB and RR Lyrae variable stars argues for the presence of an old (> 10 Gyr) population. Thus, And II must have had an extended epoch of star formation like many of the Galactic dSphs. The RGB colors yield a mean abundance of $<$[Fe/H] $>= -1.49\pm0.11$ and a

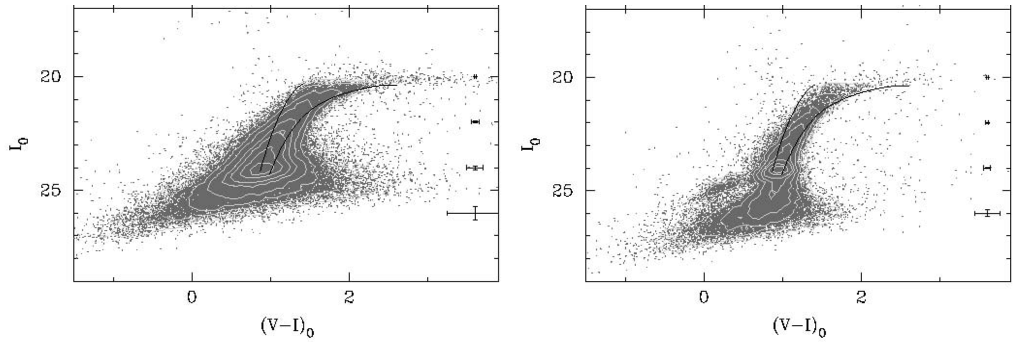


FIGURE 13. Here are shown the WFPC2 (I,V−I) CMDs for two fields in NGC 147, from Han et al. (1997), which come from 9 orbits of telescope time. On the left-hand side is the CMD for the inner most field of NGC 147, which does very close into the extremely crowded heart of this small galaxy. This CMD contains nearly 80,000 photometered stars. This means that the relative numbers of stars is also indicated by density contours overlying the points. Also over-plotted are the average error bars on the photometry at given magnitudes, and also Galactic globular cluster fiducials for M 15 and 47 Tuc. This is then done identically on the right-hand side for the outer field of NGC 147, which contains only about half the number of stars in the central field. It also contains a much more distinctive HB, which is not only a function of the reduced crowding in the outer field. See Han et al. for more details.

surprisingly large internal abundance spread, of about 0.36 dex. It is not possible to model the abundance distribution in And II with single component simple chemical enrichment model. However, a simple model with a dominant "metal-poor" ([Fe/H] = −1.6) and a "metal-rich" ([Fe/H] = −0.95) component appears to produce the best match (see Figure 12).

4.7. *NGC 147*

NGC 147 is a dE galaxy associated with M 31. There are WFPC2 CMDs at two positions in this galaxy, at inner and outer positions (Han et al. 1997, see Figure 13). There are significant differences between the inner and outer field stellar populations (as can be seen at a glance of Figure 13), and these cannot be explained by differences in crowding properties of these fields, even though the inner field is extremely crowded. The RGB suggests a metallicity of [Fe/H] = −0.9 in the inner, central field and [Fe/H] = −1.0 in the outer one, and the outer field shows a weak tendency of increasing metallicity with galactocentric radius. The RGB also shows evidence of a metallicity dispersion in NGC 147, with a larger dispersion closer to the center of the galaxy. The age of most of the stars in the RGB is assumed to be >5 Gyr. The small population of EAGB stars does show the presence of an intermediate-age population (a few Gyr old), which seems to be larger towards the center of the galaxy, contrary to the bulk of the older stars. The HB stars are more populous towards the outer part of the galaxy. Again, consistent with an age (and metallicity) gradient within NGC 147. The absence of any main sequence stars shows that any star formation completely ceased at least a Gyr ago.

4.8. *VII Zw403*

VII Zw403 (UGC 6456) is definitely not in the Local Group, but at a distance of ~ 4.5 Mpc, it is most likely an isolated member at the far side of the M 81 group. The WFPC2 CMD (see Figure 14, from Lynds et al. 1998) is the best example of a resolved BCD. Also shown in Figure 14, is a possible star formation history determined from a quantitative analysis of the CMD from Lynds et al. Another study of the same *HST* observations of VII Zw 403 is presented by Schulte-Ladbeck et al. (1999a), and they get

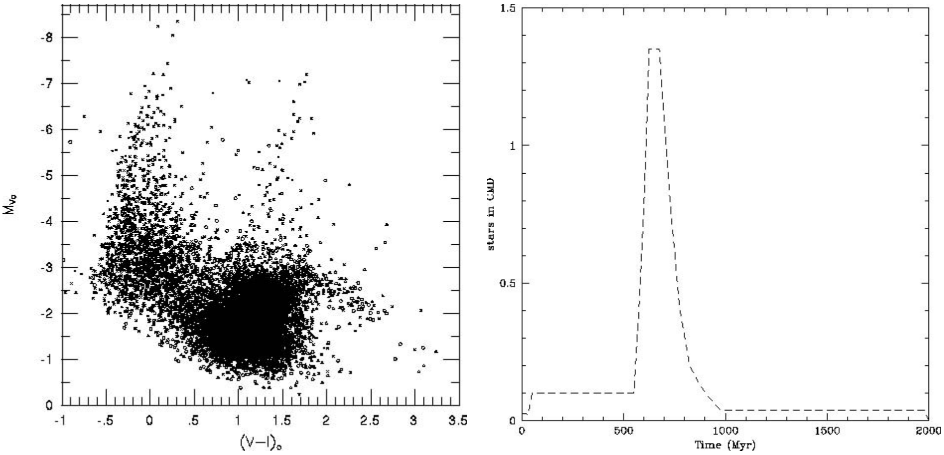


FIGURE 14. Here is the WFPC2 (V,V−I) CMD for the BCD VII Zw403, taken from Lynds et al. (1998). This CMD is based on four orbits of telescope time (2 per filter), and also shown is a possible star formation history based on a careful modeling of the CMD. See Lynds et al. for more details.

similar results. The *HST* CMD of this relatively distant galaxy is directly comparable to ground based observations of closer dwarf galaxies. The similarity between Figure 14 and the ground-based CMD of NGC 6822 of Gallart et al. (1994) is quite startling, especially in the properties of the EAGB. Clearly this is a similarity which needs further study. NGC 6822 is close enough to calibrate the EAGB versus star formation history determined from the detection of older MSTOs.

VII Zw403 is also one of the first galaxies to have a NICMOS, IR CMD (Schulte-Ladbeck et al. 1999b), which detects a large number of red super-giant and AGB stars, and reaches the tip of the RGB in J and H. In principle extending the color baseline out from the optical to the IR offers advantages for separating out different stellar phases from one another (e.g. Bertelli et al. 1994). However, without high S/N IR data the photometric errors significantly limit the improvement.

5. New Results from a New Telescope

In the future, the large, ground based telescopes can play a vital complementary role to the *HST*. With their relatively wide fields of view and larger apertures, ground based telescopes will become the instruments of choice for imaging the extended and less crowded halo populations of the nearby galaxies. Large telescopes on the ground are also ideal for spectroscopic follow-up on individual stars in a CMD to determine abundances and the internal dynamics of nearby galaxies. To show that competition from new ground based facilities is coming along fast, in Figure 15 I show CMDs for three Local Group galaxies (and a calibration globular cluster) from the VLT, and the FORS1 imaging/spectrograph. These data were taken in excellent seeing conditions in August 1999, and have been published in preliminary form by Tolstoy et al. (2000). The galaxies Cetus, Aquarius (DDO 210), and Phoenix were selected because they are relatively nearby, open-structured dI/dSph systems. Because of the excellent seeing conditions we were able to obtain very deep exposures covering the densest central regions of these galaxies, without our images becoming prohibitively crowded. From these images we have made very accurate CMDs of the resolved stellar population down below the magnitude of the HB region. In this way we have made the first detection of RC and/or HB populations

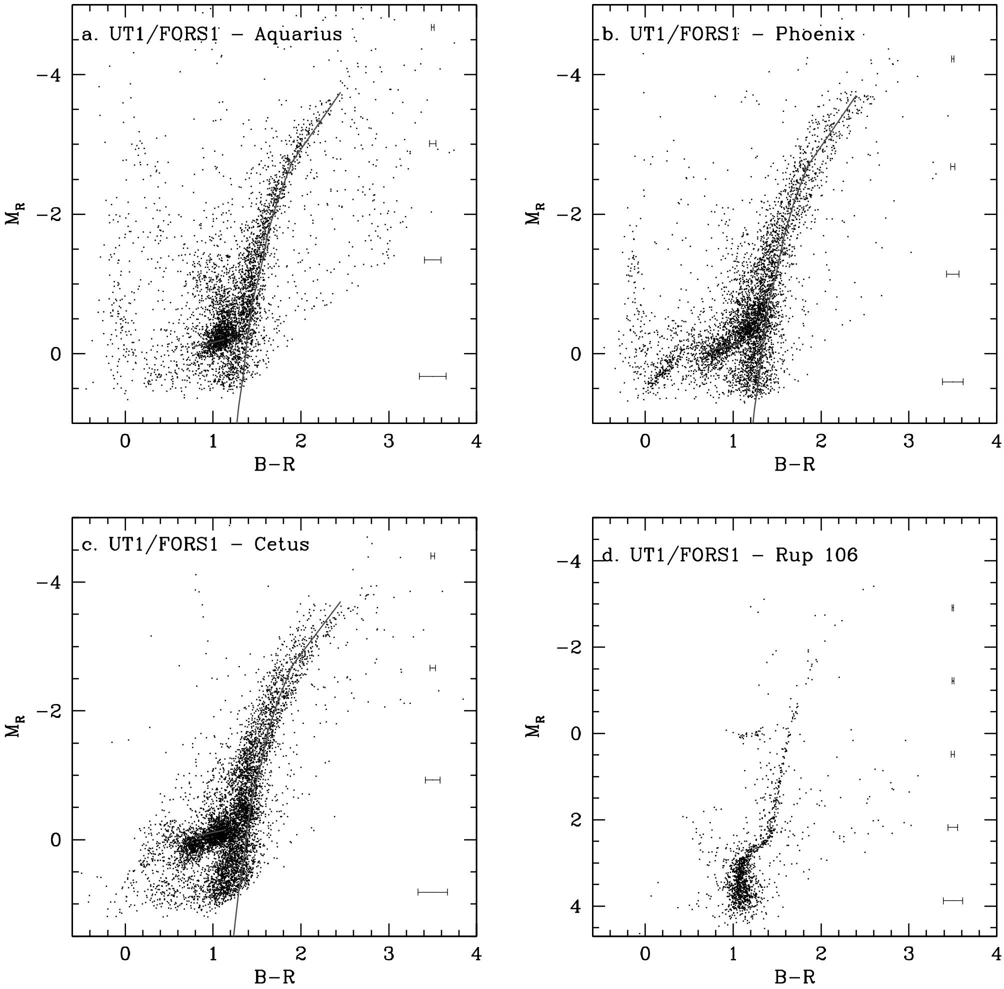


FIGURE 15. Here are four VLT/UT1 (R, B−R) CMDs for three Local Group galaxies (DDO 210, Cetus and Phoenix) and one youngish (12–13 Gyr old) globular cluster (Ruprecht 106), from Tolstoy et al. (2000). DDO 210 and Cetus had exposure times of 3000 sec in R and 3600 sec in B; roughly equivalent to a total of 3 orbits of *HST*. Phoenix received only 1600 sec in R and 1800 sec in B, and Ruprecht 106 30 sec in R and 80 sec in B. Representative error bars are plotted for each data set, and the fiducial RGB and HB from Ruprecht 106 data are over-plotted on each of the galaxy CMDs.

in these galaxies, which reveal the presence of intermediate and old stellar populations. In the case of Phoenix we detect a distinct and populous blue HB, which indicates the presence of quite a number of stars > 10 Gyr old. These results further strengthen evidence that most, if not all, galaxies no matter how small or metal poor contain some old stars. Another striking feature of our results is the marked difference between the Color-Magnitude diagrams of each galaxy, despite the apparent similarity of their global morphologies, luminosities and metallicities. For the purposes of accurately interpreting our results we have also made observations in the same filters of a Galactic globular cluster, Ruprecht 106, which has a metallicity similar to the mean of the observed dwarf galaxies.

6. Conclusions

A survey of the resolved stellar populations of all the galaxies in our Local Group provides a uniform picture of the global star formation properties of galaxies with a wide variety of mass, metallicity, gas content, etc., and makes a sample that ought to reflect the star formation history of the Universe and give results which can be compared to high redshift survey results (e.g. Steidel et al. 1999). Initial comparisons suggest these different approaches do not yield the same results (Tolstoy 1998b; Fukugita et al. 1998), but the errors are large due to the lack of *detailed* star formation histories of nearby galaxies. The CMDs presented here whilst beautiful and dramatic represent the tip of the iceberg for the Local Group. Most of the observations reported here consisted of 1 or 2 orbits of integration time per filter. To really complete a detailed census of the nearby resolved stellar populations we need to go as deep as the sensitivity limit given in Figure 2 for all Local Group galaxies, to detect the oldest MSTOs. This includes the large Local Group galaxies, such as M 31, which along with our Galaxy represent the dominant mode of star formation in the Local Group. With data like this we will know the star formation history of the Local Group, going back to the earliest times. We have also shown that data from ground based telescopes in excellent seeing with active optics can compete with *HST* images, and indeed make an excellent complement, by being more blue sensitive, and having a larger field of view. It is only with the deepest exposures of the most crowded regions that *HST* is still the undisputed winner, because it is hard to gain enough exposure time in excellent stable seeing conditions on the ground, as no where under the Earth's atmosphere can ideal conditions be guaranteed. It looks promising that Adaptive Optics on large ground based telescopes will one day rival *HST* supremacy, but this is an endeavor that will be restricted to infra-red wavelengths for the foreseeable future. *HST* must lead the way to extend detailed star formation studies past the Galaxy and its immediate satellites.

I would like to thank those who generously provided me with postscript files of their figures, to my specifications, to use in this article, namely: Gary Da Costa; Carme Gallart; Marina Rejkuba; Andrew Cole; Robbie Dohm-Palmer; Deidre Hunter; Andy Dolphin and Jay Gallagher. I would also like to thank Evan Skillman, Jay Gallagher and Andrew Cole for useful discussions about many of the details of form and content presented here. Last but not least I would like to thank the organizers of the *HST* Symposium for inviting me; it was an honor and a pleasure to put together and present this talk.

REFERENCES

Aparicio, A., Gallart, C., Chiosi, C., & Bertelli, G. 1996 *ApJ* **469**, L97.

Baade, W. 1928 *Astronomische Nachrichten* **234**, 407.

Baade, W. 1951 In *Publ. Obs. U. Michigan*, p. 7.

Baade W. 1963 In *The Evolution of Stars and Galaxies* (ed. C. Payne-Gaposchkin), p. 15. Harvard University Press.

Bertelli, G., Bressan, A., Chiosi, C., Fagotto, F., & Nasi, E. 1994 *A&A Supp.* **106**, 275.

Binggeli, B. 1994 In *Dwarf Galaxies* (eds. G. Meylan & P. Prugniel), p. 13. ESO.

Brown, T. M., Bowers, C. W., Kimble, R. A., Sweigart, A. V., & Ferguson, H. C. 2000 *ApJ* **532**, 308.

Buonanno, R., Corsi, C. E., Castellani, M., Marconi, G., Fusi Pecci, F., & Zinn, R. 1999 *AJ* **118**, 1671.

Caputo, F., Castellani, V., & degl'Innocenti, S. 1995 *A&A* **304**, 365.

Carter, D. & Sadler, E. M. 1990 *MNRAS* **245**, 12.

Cole, A. A. 1999 *Ph.D. Thesis*, University of Wisconsin.

Cole, A. A., Gallagher, J. S., Mould, J. R., Clarke, J. T., Trauger, J. T., Watson, A. M., & WFPC2 IDT 1998 *ApJ* **505**, 230.

Cole, A. A., Tolstoy, E., Gallagher, J. S., Hoessel, J. G., Mould, J. R., Holtzman, J. A., Saha, A., & WFPC2 IDT 1999 *AJ* **118**, 1657.

Da Costa, G. S. 1998 In *Stellar Astrophysics for the Local Group* (eds. A. Aparicio, et al.), p. 351. Cambridge University Press.

Da Costa, G. S. 1999 In *New Views of the Magellanic Clouds* (Eds. Y.-H. Chu, et al.), p. 397, IAU Symposium 190. ASP.

Da Costa, G. S., Armandroff, T. E., Caldwell, N., & Seitzer, P. 1996 *AJ* **112**, 2576.

Da Costa, G. S., Armandroff, T. E., Caldwell, N., & Seitzer, P. 2000 *AJ* **119**, 705.

Davies, J. & Phillips, S. 1998 *MNRAS* **233**, 553.

Dekel, A. & Silk, J. 1986 *ApJ* **303**, 39.

De Young, D. S. & Heckman, T. 1994 *ApJ* **431**, 598.

Dohm-Palmer, R. C., Skillman, E. D., Saha, A., Tolstoy, E., Mateo, M., Gallagher, J. S., Hoessel, J., Chiosi, C., & Dufour, R. J. 1997a *AJ* **114**, 2527.

Dohm-Palmer, R. C., et al. 1997b *AJ* **114**, 2514.

Dohm-Palmer, R. C., et al. 1998 *AJ* **116**, 1227.

Dolphin, A. 1997 *New Astronomy* **2**, 397.

Dolphin, A. 2000 *ApJ* **531**, 804.

Ellis, R. 1997 *Ann. Rev. Astron. & Astrophys.* **35**, 389.

Fagotto, F., Bressan, A., Bertelli, G., & Chiosi, C. 1994 *A&A Supp.* **104**, 365.

Feltzing, S., Gilmore, G., & Wyse, R. F. G. 1999 *ApJ* **516**, L17.

Ferguson, H. C. & Binggeli, B. 1994 *Astron. & Astrophys. Rev.* **6**, 67.

Ferrara, A. & Tolstoy, E. 2000 *MNRAS* **313**, 291.

Fukugita, M., Hogan, C. J., & Peebles, P. J. E. 1998 *ApJ* **503**, 518.

Fusi Pecci, F. & Bellazzini, M. 1997 In *Third Conference on Faint Blue Stars* (Eds. A. G. Davis Philip, J. W. Liebert, R. A. Saffer, & D. S. Hayes), p. 255. L. Davis Press.

Gallagher, J. S., Tolstoy, E., Dohm-Palmer, R. C., Skillman, E. D., Cole, A. A., Hoessel, J. G., Saha, A., & Mateo M. 1998 *AJ* **115**, 1869.

Gallart, C., Aparicio, A., Chiosi, C., Bertelli, G., & Vilchez, J. M. 1994 *ApJ* **425**, L9.

Gallart, C., Freedman, W.L., Aparicio, A., Bertelli, G., & Chiosi, C. 1999 *AJ* **118**, 2245.

Gardner, J. P., et al. 1998 *ApJ* **492**, 99.

Grebel, E. 1998 In *The Stellar Content of Local Group Galaxies* (eds. P. Whitelock & R. Cannon) p. 17, IAU Symposium 192. ASP.

Gregg, M. & Minniti, D. 1997 *PASP* **109**, 1062.

Grillmair, C. J., et al. 1996 *AJ* **115**, 144.

Grillmair, C. J., et al. 1998 *AJ* **115**, 144.

Han, M., Hoessel, J. G., Gallagher, J. S., Holtzman, J., Stetson, P. B., & WFPC2 IDT 1997 *AJ* **113**, 1001.

Hernandez, X., Valls-Gabaud, D., & Gilmore, G. 1999 *MNRAS* **304**, 705.

Hernandez, X., Gilmore, G., & Valls-Gabaud, D. 2000 *MNRAS* **317**, 831.

Hodge, P. W. 1986 *Ann. Rev. Astron. & Astrophys.*, **27**, 139.

Hodge, P. W., Dolphin, A. E., Smith, T. R., & Mateo, M. 1999 *ApJ* **521**, 577.

Holtzman, J. A., Gallagher, J. S., Cole, A. A., & WFPC2 IDT 1999 *AJ* **118**, 2262.

Holtzman, J. A., Smith, G. H., & Grillmair, C. J. 2000 *AJ* **120**, 3060; astro-ph/0008468.

Hunter, D. A. 1999, see Da Costa, 1999, p. 217.

Ibata, R. A., Gilmore, G., & Irwin, M. J. 1994 *Nature* **370**, 194.

Irwin, M. 1998 see Grebel 1998, p. 409.

Jones, D. H., et al. 1996 *ApJ* **466**, 742.

Lee, M. G., Freedman, W., Mateo, M., et al. 1993 *AJ* **106**, 1420.

Lilly, S. J., Le Fevre, O., Hammer, F. & Crampton, D. 1996 *ApJ* **460**, L1.

Lo, K. Y., Sargent, W. L. W., & Young, K. *AJ* **106**, 507.

Lynds, R., Tolstoy, E., O'Neil, E., & Hunter, D. A. 1998 *AJ* **116**, 146.

Mac Low, M.-M. & Ferrara, A. 1999 *ApJ* **513**, 142.

Massey, P. 1999, see Da Costa, 1999, p. 173.

Mateo, M. 1998 *Ann. Rev. Astron. & Astrophys.* **36**, 435.

Mighell, K. J. & Rich, M. 1996 *AJ* **111**, 777.

Mighell, K. J. 1997 *AJ* **114**, 1458.

Mighell, K. J., et al. 1997 *BAAS* **190**, 3505.

Mighell, K. J., Sarajedini, A., & French, R. S. 1998 *AJ* **116**, 2395.

Mighell, K. J. & Burke, C. J. 1999 *AJ* **118**, 366.

Monkiewicz, J., et al. 1999 *PASP* **111**, 1392.

Panagia, N., Romaniello, M., Scuderi, S., & Krishner, R. P. 2000 *ApJ* **539**, 197.

Rejkuba, M., Minniti, D., Gregg, M. D., Zijlstra, A. A., Alonso, M. V. & Goudfrooij, P. 2000 *AJ* **120**, 801.

Saha, A. 1995 In *Stellar Populations* (eds. P. C. van der Kruit & G. Gilmore) IAU Symp. 164, p. 175. Kluwer Academic Publisher.

Saha, A., Freedman, W. L., Hoessel, J. G., & Mossman, A. E. 1992 *AJ* **104**, 1072.

Sandage, A. R. 1971 In *Ponitifica Academia Scientarium Scripta Varia*, (ed. D. J. K. O'Connell), p. 601. Elsevir.

Schulte-Ladbeck, R. E., Hopp, U., Crone, M. M., & Greggio, L. 1999a *ApJ* **525**, 709.

Schulte-Ladbeck, R. E., Hopp, U., Greggio, L., & Crone, M. M. 1999b *AJ* **118**, 2705.

Searle, L., Sargent, W. L. W., & Bagnuolo, W. G. 1973 *ApJ* **179**, 427.

Skillman, E. D. 1996 In *The Minnesota Lectures on Extragalactic Neutral Hydrogen* (ed. E. D. Skillman), ASP Conf. Ser. 106. p. 208. ASP.

Skillman, E. D. 1998, see Da Costa, 1998, p. 457

Skillman, E. D., Kennicutt, R. C., & Hodge, P. W. 1989 *ApJ* **347**, 875.

Smecker-Hane, T. A., et al. 1999, see Da Costa, 1999, p. 343.

Steidel, C. C., et al. 1999 *ApJ* **519**, 1.

Tammann, G. 1994 *Dwarf Galaxies*, see Binggeli, 1994, p. 3.

Tolstoy, E. 1996 *ApJ* **462**, 684.

Tolstoy, E. 1998a, see Grebel, 1998, p. 218.

Tolstoy, E. 1998b, In *Dwarf Galaxies & Cosmology* (eds. Thuan et al.), p. 171. Editions Frontieres.

Tolstoy, E. & Saha, A. 1996 *ApJ* **462**, 672.

Tolstoy, E., Gallagher, J. S., Cole, A. A., Hoessel, J., Saha, A., Dohm-Palmer, R. C., Skillman, E. D., Mateo, M., Hurley-Keller, D. 1998 *AJ* **116**, 1244.

Tolstoy, E., Gallagher, J., Greggio, L., Tosi, M., de Marchi, G., Romaniello, M., Minniti, D. & Zijlstra, A. 2000 *The ESO Messenger*, **99**, 16.

Tosi, M. M., Greggio, L., Marconi, G., Focardi, P. 1991 *AJ* **102**, 951.

van den Bergh, S. 2000 *The Galaxies of the Local Group*, Cambridge University Press.

van Zee, L., Skillman, E., & Haynes, M. 1999 *BAAS* **194**, 0504.

Wyder, T. K., Hodge, P. W., & Zucker, Daniel B. 2000 *PASP* **112**, 1162.

The formation of star clusters

By BRADLEY C. WHITMORE

Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD, 21218

The ability of *HST* to resolve objects ten times smaller than possible from the ground has rejuvenated the study of young star clusters. A recurrent morphological theme found in nearby resolved systems is the observation of young (typically 1–10 Myr), massive (10^3–10^4 $M_\odot$), compact ($\rho \approx 10^5$ $M_\odot$ pc^{-3}) clusters which have evacuated the gas and dust from a spherical region around themselves. New stars are being triggered into formation along the edges of the envelopes, with pillars (similar to the Eagle Nebula) of molecular gas streaming away from the regions of star formation. The prototype for these objects is 30 Doradus (Figures 1 and 2). Another major theme has been the discovery of large numbers of young (typically 1–500 Myr), massive (10^3–10^8 $M_\odot$), compact star clusters in merging, starbursting, and even some barred and spiral galaxies. The brightest of these clusters have all the attributes expected of protoglobular clusters, hence allowing us to study the formation of globular clusters in the local universe rather than trying to ascertain how they formed ≈ 14 Gyr ago. The prototype is the Antennae Galaxy (Figures 3 and 4).

1. Introduction

1.1. *Hubble's first six months*

The discovery of spherical aberration in the summer of 1990 raised serious questions about the ability of *HST* to do the unique science it was designed for, and caused general consternation throughout the astronomical community. Early *HST* observations of compact star clusters in 30 Doradus and NGC 1275 played pivotal roles in demonstrating that *HST*, even in its crippled state, could produce stunning results that were impossible to obtain from the ground. This provided a much needed shot in the arm for the Hubble project. In addition, image deconvolution techniques developed to reconstruct the 30 Doradus images showed that most of the compromised resolution could be restored for bright objects.

Before the *HST* image of 30 Doradus came out in the summer of 1990, several papers had argued that R136 (the central object in the 30 Doradus cluster) was a single star with a mass of several thousand solar masses (e.g. Cassinelli, Mathis & Savage 1981). By the time *HST* was launched, speckle observations by Weigert & Baier (1985) had resolved the central region into roughly a dozen individual stars. However, concerns about possible artifacts introduced by the speckle technique, and limitations imposed by the small field of view, limited the impact of the results. As the saying goes, "a picture is worth a thousand words," and it was the spectacular direct images obtained with the WFPC1 on *HST* that first made it clear just how rich this cluster really was. For example, Hunter et al. (1995) identified over 3500 stars in the central $\approx 8''$ alone.

HST was built to make new discoveries, and one of the first was the discovery of young compact clusters in NGC 1275, the central galaxy in the Perseus cluster (Holtzman et al. 1992). This demonstrated that *HST* had no peers when it came to detecting compact, point-like objects against a bright background. Objects that were impossible to see from the ground suddenly became visible. Holtzman et al. (1992) also made the rather daring assertion that the young clusters were *protoglobular* clusters formed by a merger event. This was the catalyst for what has become a major cottage industry for Hubble, and provides one of the primary topics for this review.

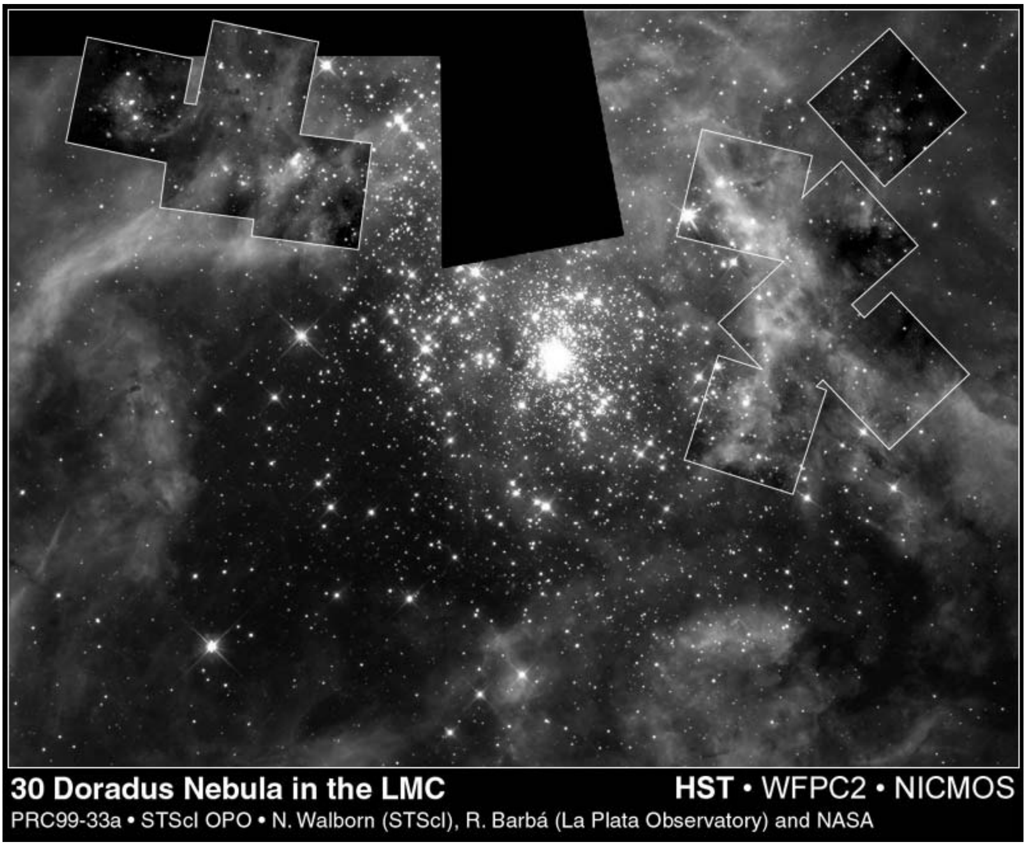

30 Doradus Nebula in the LMC
HST • WFPC2 • NICMOS
PRC99-33a • STScI OPO • N. Walborn (STScI), R. Barbá (La Plata Observatory) and NASA

Figure 1. Combined WFPC2 (background) and NICMOS (square inserts) image of 30 Doradus from a 1999 *HST* press release by Walborn and Barba. The central star cluster is R136.

1.2. *Motivation*

There are three basic reasons why the *HST* observations of young star clusters caught the attention of the astronomical community.

1) They provide insight into the mechanisms of star formation, the most fundamental process in astronomy. While we have very detailed models of the structure and evolution of stars (e.g. isocrones in the HR diagram), we have only sketchy ideas of how stars form to begin with. Some of the basic questions that remain to be solved are: Is the initial stellar mass function the same in all environments? Can star formation trigger subsequent star formation? Are all stars formed in groups and clusters?

An obvious approach to solving these questions is to go to where lots of stars are forming, such as merging and starbursting galaxies. When we do this we find that a large fraction of the star formation is in the form of massive, compact star clusters. Understanding how these clusters form should go a long ways toward understanding star formation in general.

2) They allow us to study the formation and evolution of globular star clusters in the local universe rather than trying to ascertain how they formed ≈ 14 billion years ago. An analogy is often made between the study of old globular clusters and the study of fossils on the earth, since both provide an evolutionary record. Given the opportunity, I

believe many paleontologists would switch fields if they could go to where the dinosaurs are still living.

3) They are relevant to the question of whether spiral galaxies can merge to form elliptical galaxies. One of the primary objections to this hypothesis was raised by van den Bergh (1990), who pointed out that spirals have fewer globular clusters per unit luminosity than ellipticals. It now appears that globular clusters can be formed by mergers of gas rich systems (§3.1), providing a natural explanation for this difference.

1.3. *Nomenclature*

For historical reasons, a wide variety of names are currently being used to describe what are physically similar objects. Researchers studying merging galaxies often use names such as “protoglobular cluster” or “young globular clusters.” This is because the focus has been on the question of whether merging galaxies can form globular clusters, in response to van den Bergh’s (1990) criticism of the merger hypothesis for the origin of elliptical galaxies. On the other hand, researchers studying nearby starburst galaxies generally use the term “super star clusters,” a term that was first introduced by Arp & Sandage (1985) when referring to the dominant cluster in the starburst dwarf galaxy NGC 1569. Other names in use include “blue populous clusters” (in the LMC) and “young massive clusters” (in normal spirals; e.g. Larsen & Richtler 1999).

None of these names really ring true, however, which is why one has not become the standard. While there is good evidence that some of the brightest compact clusters become globular clusters, it is obvious that they do not all become globular clusters, since the specific globular cluster frequency in merger remnants would then be too high. The main objection to the term “super star cluster” is that while they may be very luminous for a short period of time, their masses, which are a more fundamental property, are not “super.” They are instead similar to normal globular clusters or open clusters. In essence, there are probably no major physical differences between these various clusters; we are simply seeing young globular clusters, open clusters, and associations at different stages of their evolution or in different environments than we are use to seeing them in the Milky Way.

The defining properties for most of the objects discussed in this review are that they are young, compact, and the brightest are ultra luminous in comparison to old globular clusters. I will simply call them young compact star clusters.

1.4. *Goals*

I have three goals for this review. The first is to highlight the primary contributions that *HST* has made to the study of the formation of young compact star clusters (§2 and 3). There will be an obvious bias towards observations of young compact clusters in merging galaxies, since that is what I know the most about, but I will also highlight some of the major results for nearby clusters such as 30 Doradus. The reader is referred to reviews in this volume by Harris (old globular clusters), Bally (star formation), and Leitherer (starbursts) for related discussions. The second goal is to provide a compilation of the literature on young unresolved compact clusters, as discussed in §4. The third goal is to use the compilation to examine some of the demographics of young cluster formation (§4).

2. Nearby resolved star clusters

The techniques for studying star clusters that can be resolved into individual stars are much different than the techniques available for studying more distant clusters. For

galaxy	N	M_V(bright)	$R_{\rm eff}$	mass	age	α
			(pc)	($M_\odot$)	(Myr)	
near Gal. Center	2	—	—	10^4	3	—
LMC	8	−11.3	2.6	—	3	—
M82	100	−14.5	3.5	—	100	—
HE 2−10	76	−12.7	3	10^3–10^5	—	−1.7
ESO 338−IG04	112	−15.5	—	10^3–10^7	10–10,000	—
NGC 1569	7	−13.9	2.2	—	15	—
NGC 5253	6	−11.1	—	10^6	2.5	—
NGC 1705	36	−13.7	3.4	—	15	—
NGC 1741	314	−15	—	10^4–10^6	4	−1.85
ESO 565−11	700	−13.4	—	—	4–6	−2.2
NGC 1097	88	−13.8	2.5	—	—	—
NGC 4038/39	800	−15.8	4	10^3–10^7	1–500	−2.1
NGC 3256	1000	−15	5–10	—	—	−1.8
NGC 3256 tail	50	−10	—	—	—	—
NGC 3597	700	−13.2	—	—	—	−2.0
NGC 7252	500	−16.2	5	10^4–10^8	600	−1.8
NGC 1275	800	−15.8	—	10^4–10^8	100–1,000	−1.9
NGC 3921	102	−14	<5	—	500	−2.1

Notes to Table 1: See Table 2 for references. N is the number of clusters (to $M_V \approx -9$ mag in most cases). M_V(bright) is the M_V magnitude for the brightest cluster. $R_{\rm eff}$ is the average effective radius for the clusters. α is the power-law index for the luminosity function.

Table 1. Properties of young compact star clusters (approximately in order of distance)

example, we can determine the stellar luminosity and mass functions, the color-magnitude diagram for the stars, and the radial density profile. For the more distant clusters some of the typical tools are the specific globular cluster frequency, the cluster luminosity function, and color-color diagrams. Before *HST*, the dividing line between these two regimes was essentially the edge of the Milky Way galaxy. One of the promises of *HST* was that it would push this dividing line roughly ten times farther away, allowing us to study the clusters in the Magellanic Clouds, and to a lesser extent the nearby galaxies in the Local Group, with the same techniques we have been using for the clusters in the Milky Way. This has indeed been the case. In this section we describe three of the primary *HST* results for the nearby resolved clusters. Table 1 lists the number, luminosity of the brightest cluster, size, mass, age, and power law index for the luminosity function for a variety of young star clusters that are discussed in this review.

2.1. *Young clusters near the center of the Milky Way*

The Near-infrared Camera and Multi-object Spectrometer (NICMOS) was used by Figer et al. (1999) to penetrate the dust toward the center of the Milky Way in order to study two remarkable young clusters near the Galactic Center. Based on turnoffs in the color-magnitude diagrams they estimate that the Arches cluster is 2 ± 1 Myr old and the Quintuplet cluster is 4 ± 1 Myr old. Based on number counts and an extrapolation to 1 $M_\odot$, they estimate the masses of the clusters are $\approx 10^4$ $M_\odot$, and the densities are $\approx 10^5$ $M_\odot$ pc^{-3}.

The existence of these clusters was somewhat unexpected; one would think that the strong tidal shear produced by the central black hole, the high velocity dispersion, and the strong magnetic field would make this a hostile environment for forming young massive clusters. However, given that starburst galaxies are able to support prodigious rates of star formation near their centers, perhaps we should not be surprised that star clusters can also form near the Galactic Center.

Portegies-Zwart (2000) performed n-body simulations which indicate that the two clusters should dissolve after 10–60 million years in the tidal field of the Galaxy. They also point out that the stellar density near the center of the Galaxy is so high that the clusters would only be distinguishable for a short fraction of their existence, as low as 5% in some models. Based on this result they conclude that the Galactic Center may be hiding between 10 and 40 clusters which are similar to the Arches and Quintuplet clusters, but slightly older and less compact. Similar simulations by Kim et al. (1999) suggest that very massive stars play an important role in the evolution of these clusters because relaxation and mass segregation times are comparable to or even smaller than the lifetimes of the stars.

These clusters are reminiscent of the young clusters found in the central 1.5 kpc of the merger NGC 7252. Miller et al. (1997) found ≈ 40 such clusters, all less than ≈ 10 Myr years old based on the (U−B) vs. (V−I) diagram. Hence, it looks like the centers of galaxies may actually be good places to make star clusters, but few if any will survive very long. This suggests that a sizeable fraction of the field star population may have been formed in clusters.

2.2. *The Initial Stellar Mass Function (IMF)*

The light from young star clusters is dominated by ultraluminous O and B stars. However, if these were the only stars the cluster would not be stable since the massive stars are destined to go supernova, returning their mass to the interstellar medium. Hence, the identification of low mass stars is critical for the survival of the compact star clusters. In addition, the question of whether the IMF is universal, or is instead a function of the environment, is a question which is currently being hotly discussed (see reviews by Scalo 1998, Larson 1999, and Elmegreen 1999). I will briefly comment on a few examples where *HST* observations of young clusters are relevant to this debate.

The initial stellar mass function is generally parameterized as a power law with an index Γ. The canonical value of Γ for field stars in the Milky Way is -1.35 (Salpeter 1955). There is a lively debate over the question of how universal Γ is, and in particular, whether Γ is the same in clusters and in the field. Several authors have argued that the formation of high mass stars may be favored in starburst regions, which would result in a lower value of Γ.

Initial measurements of the IMF in 30 Doradus were made by Malumuth & Heap (1994; found over 800 stars in the inner $8''$ using the WFPC1), and Hunter et al. (1995; found over 3500 stars in the inner $8''$ using WFPC2). More recently, Massey & Hunter (1998) measured the IMF down to 2.8 $M_\odot$, and find that despite the largest number of high mass and luminosity stars ever seen, the IMF is completely normal. This implies that the IMF is the same over several orders of magnitudes in density, from field stars to starburst clusters like 30 Doradus. The large number of O stars is simply a result of the youth (≈ 2 Myr) and richness of the cluster. A more recent study of WFPC2 images by Sirianni et al. (2000) detects stars in 30 Doradus roughly 1 magnitude deeper than Massey and Hunter, and argues that these are pre-main sequence stars in the mass range 0.6–3 $M_\odot$. They construct the IMF in the range 1.35–6.5 $M_\odot$ and find a flattening below ≈ 2 $M_\odot$. While there are several examples of a flat IMF for stars less massive than 1 $M_\odot$ (see Larson 1999), this is the first example of a flattening above this point. However, Brandl et al. (1996, 2000) find no evidence for a truncation down to at least 1 $M_\odot$.

Another possible example of a relatively flat IMF, but this time at the high mass end, are the results of Figer et al. (1999) for the Arches cluster, one of the two massive

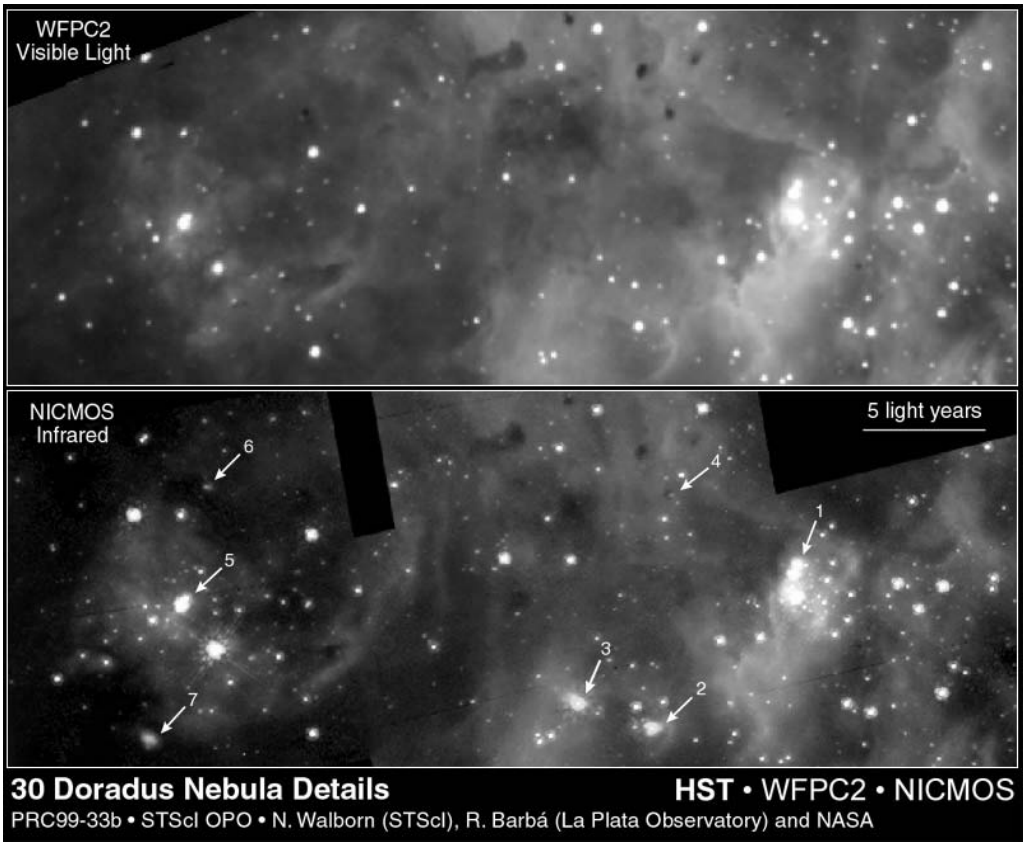

WFPC2
Visible Light
NICMOS
Infrared
5 light years
1
2
3
4
5
6
7
30 Doradus Nebula Details
HST • WFPC2 • NICMOS
PRC99-33b • STScI OPO • N. Walborn (STScI), R. Barbá (La Plata Observatory) and NASA

Figure 2. WFPC2 and NICMOS images in an outlying region of 30 Doradus (i.e. the upper left square inserts in Figure 1) from a 1999 *HST* press release by Walborn & Barba. Note the young stars still embedded in dust which are only visible on the NICMOS image (e.g. objects identified as 2, 3, 6 and 7). Also notice the pillar-like dust feature around object 1 pointing away from the central star cluster R136 (see Figure 1 for context).

young clusters near the Galactic Center. They find $\Gamma = -0.7$ when fitting over the range 6–125 $M_\odot$ for stars in an annulus from $3''$ to $7.5''$. However, incompletion caused by crowding make this a difficult measurement at the faint end of the mass function. For example, Figer et al. note incompletion fractions of 50% for stars up to 35 $M_\odot$ in the inner $3''$. In addition, they find a very flat value of $\Gamma = -0.2$ in the region $2.5''$ to $4.5''$, which may be due to truncation of the faint end by crowding. If we stick to the brighter end of the mass function from 16 to 125 $M_\odot$, and only use stars in an annulus from $4.5''$ to $7.5''$, $\Gamma \approx -1.1$, still low, but not too different from the Salpeter IMF.

Hence, while there is tentative evidence for deviations in the IMF for some environments, the near uniformity over an extremely wide range of environment is quite remarkable. Larson (1999) concludes "this large body of direct evidence does not yet demonstrate convincingly any variability of the IMF, although the uncertainties are still large. Some indirect evidence based on the photometric properties of more distant and exotic systems suggests that the IMF may vary in extreme circumstances, possibly being top-heavy in starbursts and high redshift galaxies." We also note that low-mass stars are clearly forming in R136 (e.g. down to 0.6 $M_\odot$; Sirianni et al. 2000), and in other young nearby compact clusters such as NGC 3603 (i.e. no evidence for a flattening down to

1 $M_{\odot}$, Eisenhauer et al. 1998, or down to 0.1 $M_{\odot}$, Brandl et al. 2000). Hence concerns that the young clusters will be unstable due to the absence of low mass stars appear to be unfounded.

2.3. *Triggered star formation*

30 Doradus has been called a "Starburst Rosetta," since it is the nearest example of a young massive starburst cluster and hence can be studied in unique ways that are not possible for its more distant counterparts. Until recently, 30 Doradus was believed to be a well evolved H II region, with no current star formation going on. Early indications that this was not true were the discovery of an H_2O maser (Whiteoak 1983) and four luminous IR protostars (Hyland et al. 1992). More recently, Walborn & Blades (1997) used optical spectral classification of 106 OB stars to show the presence of five distinct stellar populations in 30 Doradus, with ages ranging from ≈ 1 to ≈ 10 Myr. Hence, the simple models of either continuous or single-burst star formation appear to be incorrect. It now appears that a starburst in one area can trigger the formation of stars in a nearby region.

Corroborating evidence for this picture has been obtained by Walborn et al. (1999a,b) using NICMOS, as shown in Figure 1. A roughly spherical shell of "second generation" star formation, including a host of newly discovered IR sources, can be seen around the central R136 concentration (Figure 2). Several massive dust pillars are found streaming away from R136, similar to the famous *HST* image of the Eagle Nebula. At the heads of these pillars are the sites where active star formation is currently being triggered. The molecular gas revealed by the dust provides the raw material for the star formation. Scowen et al. (1998) shows that these pillars are indeed very similar to the pillars in the Eagle Nebula. This same picture of a bright compact central cluster which has evacuated a roughly spherical nebular envelope around it can be seen in a number of the *HST* press releases (e.g. NGC 604, N11 in the LMC, NGC 4214).

3. Distant unresolved star clusters

In this section we move further out to where it becomes difficult to resolve the individual stars, except in extraordinarily extended clusters such as knot S in NGC 4038/4039 (the "Antennae Galaxies," Whitmore et al. 1999). Historically, most of the early *HST* observations of compact star clusters were in merging galaxies, followed shortly by similar observations in starburst galaxies. More recently, young compact clusters have also been found in other environments, including barred galaxies, tidal tails, and normal spiral galaxies.

3.1. *Young compact star clusters in merging galaxies*

The primary question for many of the early *HST* studies of merging galaxies was whether globular clusters were being formed. This possibility was proposed by Schweizer (1987) and Burstein (1987), primarily to address van den Bergh's (1990) objection to the merger model based on the higher specific frequency of globular clusters in elliptical galaxies. Ashman & Zepf (1992) and Zepf & Ashman (1993) further developed these ideas, and made predictions about the bimodality of the metallicity histogram of globular clusters that should result if most ellipticals are formed by merging spirals.

A hint that young globular clusters might be formed in mergers was provided by Schweizer's (1982) observations of six unresolved bluish knots in the merger remnant NGC 7252. However, with so few objects he could not be sure they were not simply field stars. Lutz (1991) observed roughly a dozen blue, point-like objects in the merger

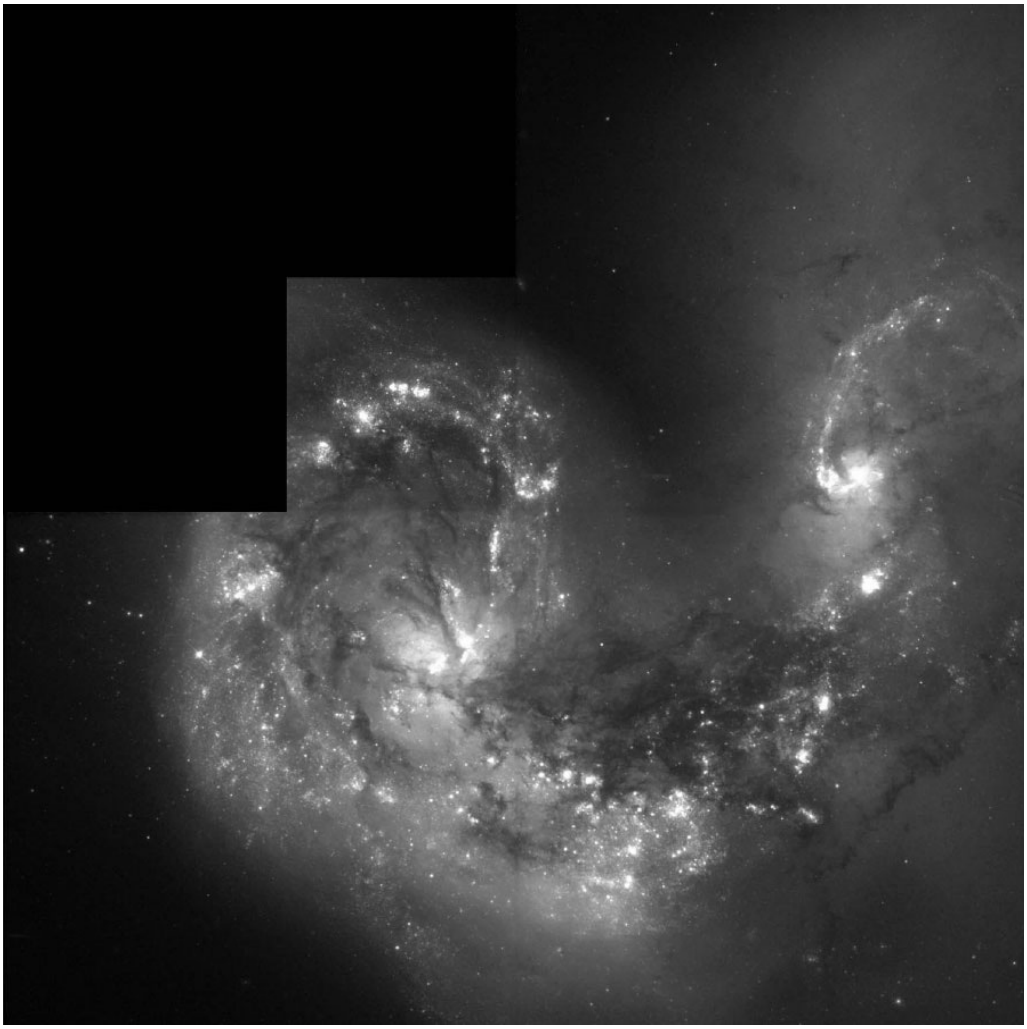


FIGURE 3. Image of the Antennae Galaxies (NGC 4038/4039) from Whitmore et al. (1999).

remnant NGC 3597, but was not able to resolve the objects and hence could not be certain they were not associations or giant H II regions.

As briefly discussed in §1, the *HST* observations of NGC 1275 by Holtzman et al. (1992) provided the original breakthrough and was the primary catalyst in this field. They discovered a population of about 60 blue compact clusters, and suggested that they were protoglobular clusters which formed $\leq$ 300 Myr years ago during a merger of NGC 1275 with another galaxy. Unfortunately, NGC 1275, the central cooling-flow galaxy in the Perseus cluster, is such a peculiar galaxy that it is not clear which of its peculiarities is responsible for the formation of the young clusters (e.g. see Richer et al. 1993, who suggested that the cooling flows are responsible for the formation of the clusters).

Whitmore et al. (1993), using WFPC1 observations of the prototypical merger remnant NGC 7252 (Toomre 1977), found a population of about 40 blue point-like objects with luminosities and colors nearly identical to those found in NGC 1275. Unlike NGC 1275, with all its peculiarities, NGC 7252 is an isolated galaxy which therefore provided a much

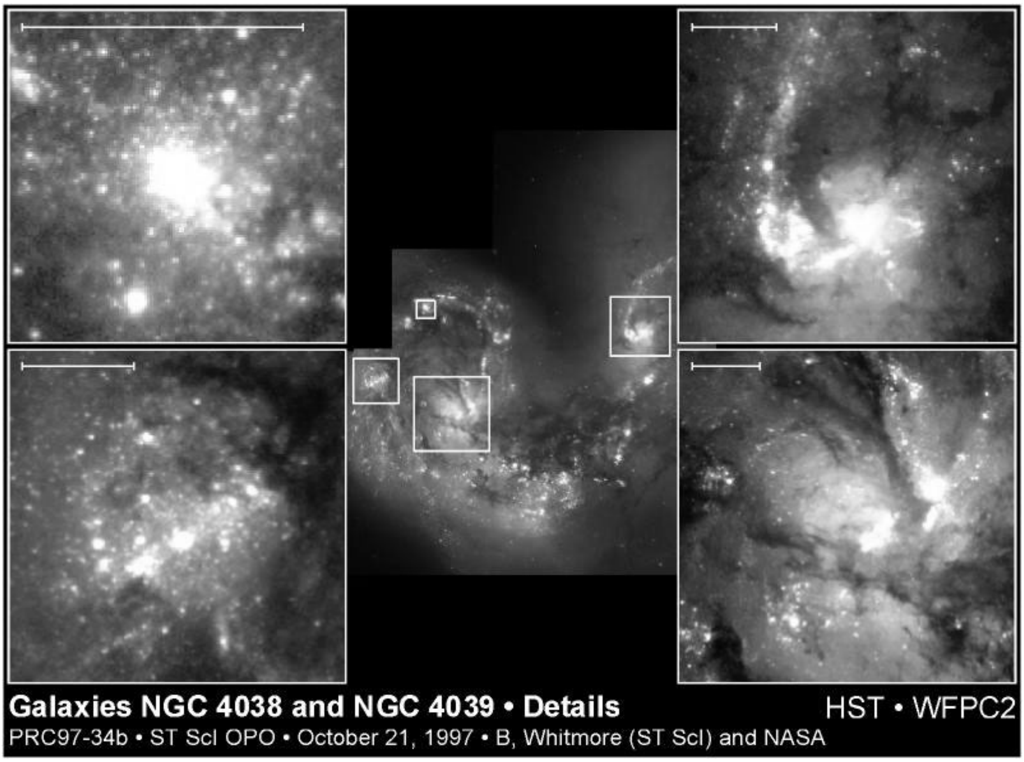

Galaxies NGC 4038 and NGC 4039 • Details
HST • WFPC2
PRC97-34b • ST ScI OPO • October 21, 1997 • B, Whitmore (ST ScI) and NASA

FIGURE 4. Blowup of two of the brightest clusters in the Antennae (*left*) and the central regions of the two galaxies (*right*) from Whitmore et al. (1999).

cleaner connection between the formation of young star clusters and merging galaxies. Whitmore & Schweizer (1995) followed this up with pre-refurbishment observations of another prototypical merger, NGC 4038/4039 (see Figures 3 and 4). Over 700 young star clusters were found in this galaxy. Subsequent observations of both these galaxies using WFPC2 (NGC 7252—Miller et al. 1997; NGC 4038/4039—Whitmore et al. 1999) have increased the numbers of cluster candidates tenfold.

Roughly 30 different gas-rich mergers have now been observed with *HST*, as summarized in Table 2. In all cases young massive compact clusters have been observed, *the brightest of which have the luminosities, colors, sizes, masses, distributions and spectra that we would expect for globular clusters with ages in the range 1 to 500 Myr*. A few of the key observations are described below, but the reader is referred to the papers listed in Table 2 for the details.

3.1.1. *Luminosities and colors*

The luminosities of young globular clusters with ages ≈ 10 Myr should be ≈ 5–6 magnitudes brighter than classical old globular clusters, according to the Bruzual & Charlot (1996) models. The models also predict that young globular clusters should be ≈ 1.0–1.2 magnitude bluer in (V−I). Figure 5 shows that this is indeed the case. It also shows how the luminosities and colors of the clusters can be used to age date the clusters. NGC 4038/4039 is clearly the youngest merger remnant, with the mean age of the clusters ≈ 30 Myr and many clusters only a few Myr old. The clusters in NGC 3921 and NGC 7252 are roughly 500 Myr old while NGC 3610 appears to be a 4 ± 2 Gyr merger remnant (Whitmore et al. 1997) which may provide the missing link between young

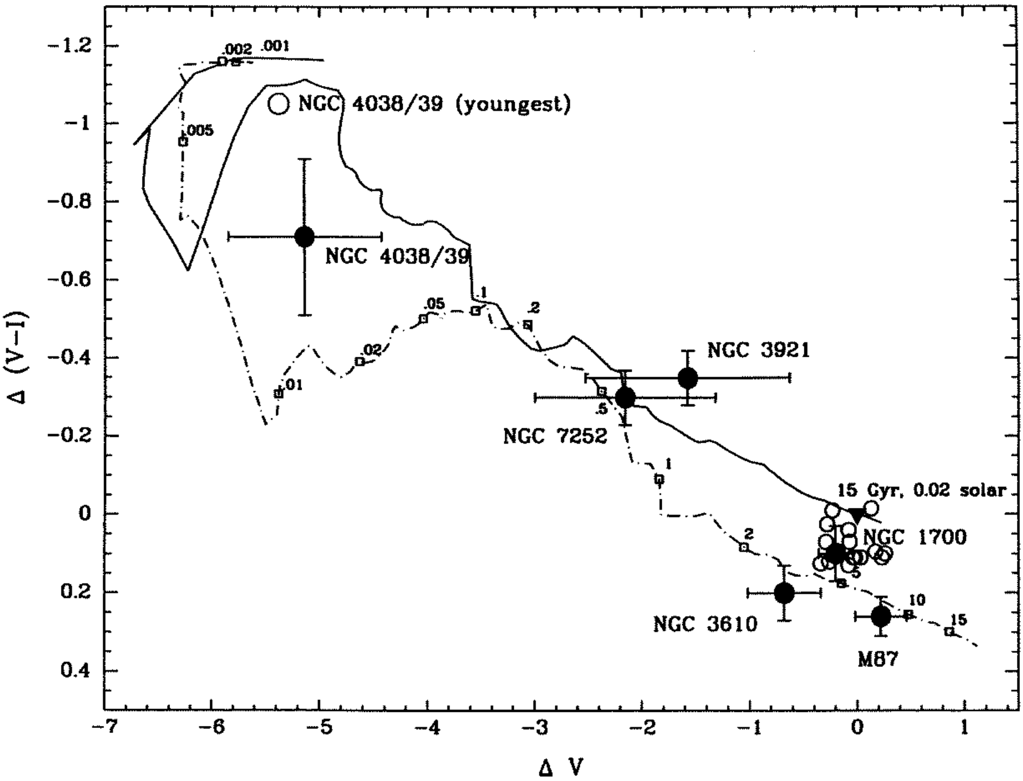


FIGURE 5. Plot of the evolution in luminosity (Δ V) and in color (Δ (V–I)) of star clusters, based on the Bruzual-Charlot (1996) tracks for a metal-poor population (solid line) and a solar metallicity population (dashed-dot line). The values are normalized to an old, metal-poor population (filled triangle). Ages in Gyr for the solar metallicity track are marked with squares. See the original article for further details (Whitmore et al. 1997).

mergers and old ellipticals. The case is less certain for NGC 1700, although Brown et al. (2000) have recently claimed that this galaxy also has a population of metal-rich clusters that are 3 ± 1 Gyr old.

3.1.2. *Sizes*

The ability to measure sizes using *HST* has been critical to the identification of the young clusters in mergers as candidate globular clusters. Ground based observations, such as those of Lutz (1991), were inconclusive, since they were not able to resolve the clusters to determine whether they were associations or H II regions, with $R_{\rm eff} \approx 100$ pc, or compact clusters similar to the globular clusters in the Milky Way, with $R_{\rm eff} \approx 3$ pc (van den Bergh, 1996). Early *HST* observations using WFPC1 indicated that the clusters were compact, with $R_{\rm eff} \approx 10$ pc (Whitmore & Schweizer, 1995). However, van den Bergh (1995) argued that this was too large, and the clusters were more likely to be open clusters. Meurer et al. (1995) found that the compact clusters he was studying in very nearby starburst dwarfs were smaller, with $R_{\rm eff} \approx 2$ pc. He suggested that the apparently larger values in the Antennae were due to poorer resolution and crowding.

Recent observations using the WFPC2 (corrected for spherical aberration) have removed this concern. Several authors have recently measured $R_{\rm eff}$ for young clusters in mergers in the range 3–6 pc (i.e. NGC 3921—Schweizer et al. 1996, NGC 7252—Miller et al. 1997, NGC 3610—Whitmore et al. 1997, and NGC 1275—Carlson et al. 1998). Perhaps the best case is for the Antennae galaxies as measured by Whitmore et al. (1999), since this is the nearest of the prototypical mergers and the observations were made with subpixel dithering which improves the resolution still further. They find the median effective radii for the clusters is $R_{\rm eff} = 4 \pm 1$ pc, similar to or slightly larger than those of globular clusters in the Milky Way.

3.1.3. *Ages*

Ages for the clusters have been estimated in a variety of manners. Figure 5 demonstrates how the luminosities and colors can be used to estimate the ages, as already discussed in §3.1.1. More precise age estimates are possible with more colors, and provide an independent means of solving for the age and the reddening caused by dust. For example, Whitmore et al. (1999) use UBVI photometry and reddening-free Q parameters to determine ages for the clusters in the Antennae (Figure 6). They find evidence for four populations of clusters, ranging in age from < 5 Myr to 500 Myr. They also isolate a population of old globular clusters in this galaxy. *Hence, it appears that we can study the entire evolution of globular clusters in this single galaxy.* This is consistent with the simulations of Mihos, Bothun, & Richstone (1993) who find that the merger process takes several hundred million years to complete, hence producing clusters with a wide range of ages.

Hα can be used in two ways to estimate the ages of the younger clusters. The existence of Hα emission itself indicates that a cluster is $\lesssim 10$ Myr, since the O and B stars required to ionize the gas only live for this long (e.g. see the Leitherer & Heckman 1995 models). The second method is to use the size of the Hα ring around a cluster. Whitmore et al. estimate that the clusters in the western loop of NGC 4038 are 5–10 Myr, since many of them have rings with diameters of $\approx$ 100–500 pc and measured expansion velocities $\approx$ 25–30 km s^{-1} (Whitmore et al. 1999). The clusters in the overlap region appear to be < 5 Myr old, since the rings are smaller or non-existent in this region.

The most accurate method of estimating ages is to obtain spectra. Zepf et al. (1995) obtained spectra of the brightest cluster in NGC 1275 which showed strong Balmer absorption lines, typical of A stars. They estimate ages of 500 Myr for the clusters, although ages from 100–900 Myr cannot be ruled out. Schweizer & Seitzer (1998) obtained UV-to-visual spectra of eight cluster candidates in NGC 7252. Six of the clusters have ages in the range 400–600 Myr, roughly consistent with the mean photometric age estimate of 650 Myr from Miller et al. (1997). One cluster turned out to be an emission-line object with an age estimate of < 10 Myr, indicating that cluster formation is still going on at a low level even in the outer parts of the galaxy. Whitmore et al. (1999) obtained GHRS spectra of two clusters in the Antennae with age estimates of 3 ± 1 Myr and 7 ± 1 Myr, in good agreement with the estimates based on the UBVI colors and the Hα morphology.

The youngest clusters appear to be very red objects, which Whitmore & Schweizer (1995) suggested were only now emerging from their dust cocoons. Several of these have recently been identified as strong IR sources (Vigroux et al. 1996, Mirabel et al. 1998, Wilson et al. 2000, Gilbert et al. 2000, Mengel et al. 2000). In fact, the brightest IR source in the Antennae is one of these very red objects (WS80), rather than the nucleus of one of the two galaxies. Wilson et al. (2000) find three separate molecular clouds around WS80 within a region of 1 kpc^2, and suggest that cloud-cloud collisions may play an important role in cluster formation. However, the lack of similar morphologies for the other very red objects suggest that this may not be the universal mechanism.

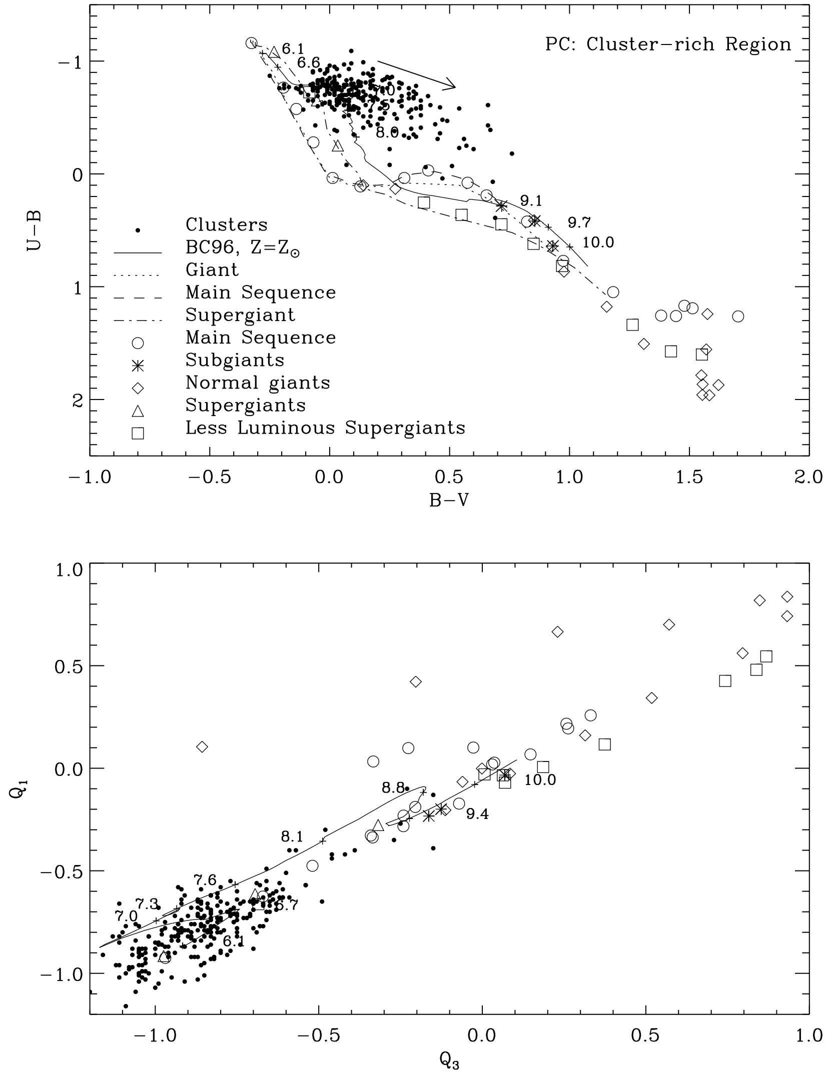


FIGURE 6. Color-color diagram and reddening-free Q parameter diagram for clusters in the Antennae. The numbers on the plots are the values of log(age). See Whitmore et al. (1999) for details.

3.1.4. *Mass*

Mass estimates of young compact clusters have been made in two ways. The first is based on the luminosity and color of the clusters using stellar population models such as Bruzual & Charlot (1996). These estimates generally range from 10^3 to 10^7 $M_\odot$ (see Tables 1 and 2), in good agreement with old globular clusters with a mean of 2×10^5 $M_\odot$ (Mandushev, Spassiva & Staneva (1991). A more direct method of determining the mass is to measure the velocity dispersion of the stars in the clusters. Observations have been obtained for nine clusters so far (two in NGC 1705 and one in NGC 1569 by Ho & Fillipenko 1996; two in M82 by Smith & Gallagher 2000, 4 in NGC 4038/39 by Mengel et al. 2000). The dispersions range from 10–20 km s^{-1} and the masses range from 1×10^5 to 4×10^6 $M_\odot$, in good agreement with values for the more massive old globular clusters in the Milky Way. The size and dispersion measurements also show that the crossing times for the clusters are ≈ 1 Myr. Hence, even the younger clusters have survived many crossing times. The older clusters in NGC 7252, NGC 3921, NGC 4038/4039, and NGC 1275 (≈ 500 Myr; see Tables 1 and 2), have survived for several hundred crossing times and appear to be quite stable. Their densities are ≈ 10^5 $M_\odot$ pc^{-3}, similar to old globular clusters, hence these clusters will almost certainly last for tens of Gyr.

3.1.5. *The luminosity function*

To first order, the luminosity functions of young compact clusters in merging galaxies are power laws of the form $\phi(L)dL \propto L^\alpha dL$, with index $\alpha \approx -2$ (Table 1). Harris & Pudritz (1994) have pointed out that the mass function for giant molecular clouds is also a power law with a similar index. Hence, all that may be necessary is a triggering mechanism to get the molecular clouds to collapse and form star clusters. Jog & Solomon (1992) have suggested that merger induced starbursts can raise the ambient pressure in the ISM and trigger the implosion of the molecular clouds. Elmegreen & Efremov (1997) agree that high-pressure environments are needed to trigger the star formation and suggest other mechanisms might be the high background virial density (e.g. in dwarf galaxies), turbulent compression, or large-scale shocks (in interacting galaxies).

The power law index for the young clusters is markedly different than the Gaussian profile found for old globular clusters (e.g. Figure 3 of Zhang & Fall 1999). However, various destruction mechanisms (e.g. 2-body evaporation, bulge and disk shocking, dynamical friction, stellar mass loss) should modify the distribution with time. Two-body evaporation appears to be the strongest amongst these mechanisms, destroying the fainter more diffuse clusters first, and in certain conditions leaving a peaked distribution similar to what is seen for old globular clusters (e.g. Fall & Zhang, 2001). This is similar to young clusters in the Milky Way with the OB associations typically only lasting tens of Myr, and open clusters lasting hundreds of Myr. Other examples of clusters which are apparently dissolving are the Arches and Quintuplet clusters near the Galactic Center, since no older clusters are seen in their vicinity, and the ≈ 40 clusters in the inner $6''$ of NGC 7252, which all have ages less than about 10 Myr (Miller et al. 1997). Finally, the number of young clusters in the Antennae galaxies is so large that it requires most of the clusters to dissolve or the value of S_N will be too high when it settles down to become an elliptical (Whitmore et al. 1999).

Fritze-v.Alvensleben (1999), following a similar line of reasoning to Meurer (1995), has attempted to determine the mass function for the clusters in the Antennae using the color information from the WFPC1 observations by Whitmore & Schweizer (1995). She concludes that the original mass function is a Gaussian which gets spread out in time to form the power law luminosity function we observe today. However, their analysis does not take into account the fact that the cutoff in the observed luminosity function is due to incompletion at the faint end (see Zhang & Fall 1999 for a discussion). When convolved with uncertain age estimate used to convert from luminosity to mass (e.g. due to reddening from dust and the availability of only V−I colors), the resulting distribution will artificially appear to be roughly Gaussian. A more complete treatment by Zhang & Fall (1999), using UBVI colors based on WFPC2 observations by Whitmore et al. (1999) and corrections for reddening and incompletion, concludes that the mass function is roughly a power law.

There is some evidence that the luminosity function for the young clusters is not a perfect power law, but is steeper for bright magnitudes (NGC 4038/4039—Whitmore et al. 1999, NGC 3256—Zepf et al. 1999). In the Antennae, Whitmore et al. (1999) find that the cluster luminosity function appears to have a bend at $M_V \approx -10.4$ (≈ -11.4 after making a correction for extinction). For absolute magnitudes brighter than $M_V \approx$

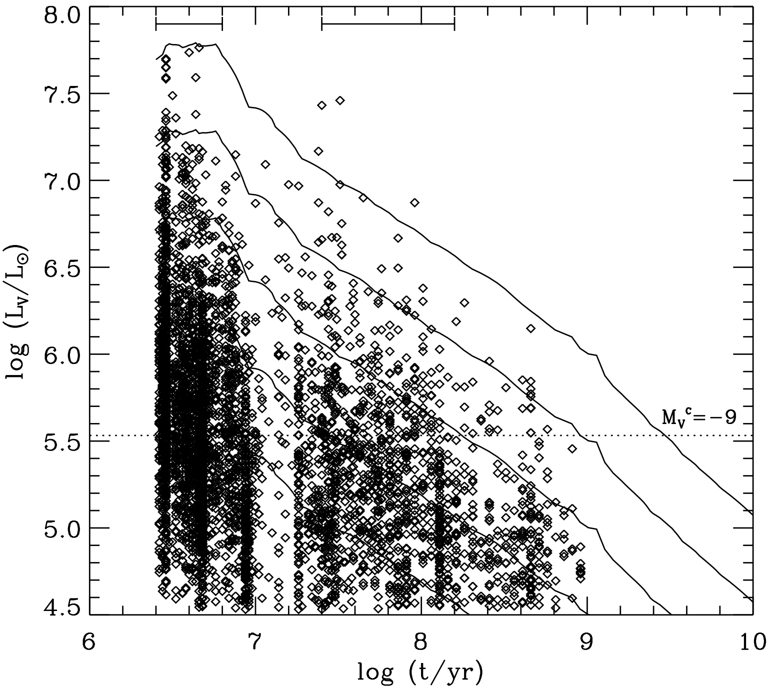


FIGURE 7. Luminosity of cluster candidates in the Antennae as a function of their ages (from Zhang & Fall, 1999, Fig. 2). The lines represent the Bruzual-Charlot (1996) tracks with $\log(M/M_\odot) = 6.0$ (*top*), 5.5, 5.0, 4.5, and 4.0. See Zhang & Fall for further details.

-10.4 the power law is steep and has an exponent of $\alpha = -2.6 \pm 0.2$, while for the range $-10.4 < M_V < -8.0$ the power law is flatter, with $\alpha = -1.7 \pm 0.2$. Assuming a typical age of 10 Myr for the clusters, and 1 mag of extinction, the apparent bend in the LF corresponds to a mass $\approx 1 \times 10^5$ M$_\odot$, only slightly lower than the characteristic mass of globular clusters in the Milky Way. A similar bend may be present in the mass function derived by Zhang & Fall (1999). The bend may be a precursor to what will become the peak of the globular cluster luminosity function.

3.2. *Young compact star clusters in starburst galaxies*

Young compact star clusters are also found in many starburst galaxies, but in much smaller numbers than the merging galaxies. Meurer et al. (1992), using ground-based observations, found a population of young clusters in the nearby starburst dwarf galaxy NGC 1705, the brightest of which was an unresolved off-center nucleus which they proposed as a young (13 Myr) globular-like cluster with a mass $\approx 1.5 \times 10^6$ M$_\odot$. Meurer et al. (1995) followed this up with an extensive study of nine starburst galaxies obtained with the Faint Object Camera on *HST*. All nine of the galaxies contained young compact star clusters. On average, 20% of the UV light from the galaxies comes from the clusters. The brightest clusters are preferentially found near the centers of the galaxies. They find the sizes are similar to Galactic globular clusters and the luminosity function has an index ≈ -2. Hence, the clusters found in the starburst galaxies appear to be similar to the clusters found in merging galaxies.

Several other authors find similar examples in other starburst galaxies, as listed in Table 3. Conti & Vacca (1994) observed 19 "knots" in the Wolf-Rayet galaxy He 2−10, each with a luminosity, mass, and size similar to Galactic globular clusters. Other early observations include those of O'Connell et al. (1994) for NGC 1569 and NGC 1705, Hunter et al. (1994) for NGC 1140, and Watson et al. (1996) for NGC 253.

The case of M82, the prototypical starburst dwarf galaxy, is especially interesting. O'Connell et al. (1995) find a complex of over 100 compact, luminous "super star clusters" concentrated in the inner 100 pc of the galaxy shining though a relatively dust free region. The brightest cluster has $M_V = -13.2$ while the mean M_V is -11.6 mag. Since most of this galaxy is embedded in dust the total number of young clusters is likely to be several times this value. It is quite possible that the starburst in M82 was triggered by a tidal interaction with its larger neighbor, M81, hence it is not clear whether M82 (or several other starburst galaxies with evidence for interactions) should be in this section or in the previous section on interacting galaxies. De Gris, O'Connell, & Gallagher (2001) studied a region farther from the center of M82 where active star formation is not occuring. They estimate the ages of the clusters in this region at 20–100 Myr. They find that the objects in the outer regions have sizes in the range $2.3 < R_{\rm eff} < 8.4$ pc. While the lower value is similar to the sizes of galactic globular clusters, the higher value is more typical of open clusters. They also find that the brightest clusters have $M_V \approx -10$ mag, and most are in the range -5 to -7 mag. Hence, most of these clusters are too faint to become globular clusters since they will fade several magnitudes as the stars evolve. It appears that this region is not able to form the true "super star clusters" seen near the center of M82 and in other merger and starburst systems.

The lesson appears to be that luminous young star clusters are found whenever there is vigorous star formation, whether it be in mergers or starburst galaxies. Since the ultraluminous IRAS sources are essentially all mergers (Sanders et al. 1988), it is not surprising that mergers show the largest populations of young star clusters.

3.2.1. *Young compact star clusters in barred galaxies*

Barth et al. (1995) found young clusters in the circumnuclear star-forming rings around the barred spiral galaxies NGC 1097 and NGC 6951. The clusters are compact, with $R_{\rm eff} \approx 2.5$ pc in NGC 1097 and ≤ 4 pc in NGC 6951. The brightest cluster has M_V(uncorrected for extinction) $= -12.6$. They estimate an intrinsic M_V in the range -14 to -15, since the clusters are on the outer edges of prominent dust lanes. Hence, these clusters appear to be quite similar to clusters in merging and starbursting galaxies. Buta et al. (2000) have done a careful analysis of young compact clusters in the nuclear ring of NGC 1326, an early-type barred spiral in the Fornax cluster. They find 269 candidate clusters with ages in the range 5 to 200 Myr, but no clusters older than this. The older clusters still lie within the ring, with no evidence of migration. The luminosity function has an index of -2.1, similar to the other compact clusters discussed in this review, but the brightest clusters are fainter than the brightest clusters found in mergers or starburst galaxies, with no M_V (uncorrected) brighter than -11. The authors conclude that this galaxy lacks any true super star clusters, and suggest that super star clusters are not a universal property of star-forming rings. It is interesting to note that while the strong Lindblad resonance in this galaxy can produce clusters with a typical power-law luminosity function, it apparently cannot form the brightest clusters which are the best candidates for protoglobular clusters.

3.3. *Young star clusters in spiral galaxies*

Larsen & Richtler (1999) carried out a systematic ground-based search for young massive clusters in 21 nearby non-interacting spiral galaxies and found young massive star clusters in about one quarter of the galaxies. In a followup paper (Larsen & Richtler 2000), they add a variety of other galaxies to the sample from the literature, including merging and

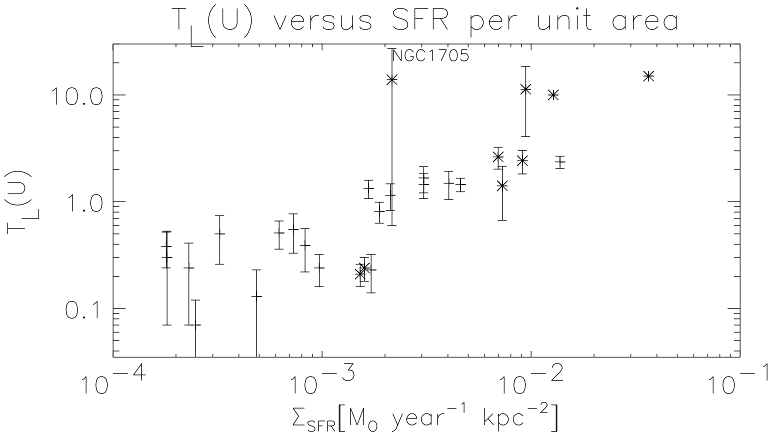


FIGURE 8. Plot of the specific luminosity for the clusters in the U band vs. the star formation rate per unit area for a sample of spiral, starburst, and merging galaxies (from Larsen & Richtler et al. 2000, Fig. 6).

starbursting galaxies, in order to test what conditions are most advantageous for making large numbers of massive clusters. They define the *specific cluster frequency* (not to be confused with the specific globular cluster frequency, S_N, see Harris 1991) as the fraction of light in clusters to the fraction of light in the total galaxy: $T_L = 100 \times L_{\rm clusters}/L_{\rm galaxy}$.

They prefer to make the measurement in U which is most sensitive to young clusters. Their primary result is that T_L(U) is well correlated with the star formation rate per unit area (Figure 8). Galaxies with very active star formation form proportionally more of their stars in clusters than in the field, with some merger and starburst galaxies devoting as much as 15–20% of their luminosity to clusters. Note that this is precisely what is needed to increase the specific globular cluster frequency, a concern voiced by Harris (1999). Larsen & Richtler (2000) also argue that "The cluster formation efficiency seems to depend on the SFR in a continuous way, rather than being related to any particularly violent mode of star formation."

Closer to home, Chandar et al. (1999) have used WFPC2 observations to study the young compact clusters in M33. They finds 44 young clusters with ages ≤ 100 Myr and masses in the range 6×10^2 to 2×10^4 $M_\odot$. Hence, M33 appears to be able to make many young compact clusters, but few if any with the masses of regular globular clusters.

3.3.1. *Young compact star clusters in tidal tails*

Knierman et al. (2000) have examined six tidal tails in four prototypical mergers (NGC 3256, NGC 3921, NGC 4038/4039 and NGC 7252). They find that only one of the tails (the western tail of NGC 3256) currently has a large number of young compact clusters (i.e. ≈ 50 clusters with the brightest having $M_V \approx -10$ mag). It is not clear whether the clusters were formed when the tidal tail was ripped from the galaxy or are currently forming. Some of the other tails appear to have only a few young clusters, (e.g. NGC 7252 and NGC 3921) while others (e.g. NGC 4038/39) appear to have essentially no clusters in the tails. Hence, it appears that there is a wide range in the number of clusters in tidal tails, perhaps due to differences in how the tails were generated (e.g. gas-rich versus gas poor, deep penetrating orbit versus quiescently being pulled out from the outer regions of the galaxy, etc.). Other studies including observations of young compact star clusters in tidal tails include Lee, Kim & Geisler (1997), Tyson et al. (1998), and Gallagher et al. (2000).

4. A compilation of the literature and a discussion of broader issues

Tables 2, 3, and 4 provide a compilation of *HST* (and occasionally key ground-based) observations of young compact star clusters in merging, starbursting, and miscellaneous other galaxies.

Based on the discussion in §2 and 3, and the compilation in Tables 1, 2, 3, and 4, it is clear that luminous young compact star clusters are produced in a wide variety of environments, but in much greater number in mergers and starburst galaxies, systems where vigorous star formation is occurring. A similar conclusion was reached by van den Bergh (2000), who comments, "Presently available data strongly suggest that the specific cluster forming frequency is highest during violent bursts of star formation." In addition, the most luminous clusters are formed in the regions with the most violent star formation.

An important question is whether this is a statistical effect due to the lower number of clusters in galaxies with low star formation, or whether it is physically more difficult to form massive clusters in relatively quiescent systems (i.e. is there a cutoff at the high end of the luminosity function for quiescent galaxies?). The situation may be analogous to the upper IMF in 30 Doradus. It was presumed that the large number of very luminous stars indicated that conditions in 30 Doradus were especially conducive for making high mass stars. However, Massey & Hunter (1998) find that the IMF is normal; that the large number of massive stars is simply due to the tremendous number of stars in the system and the young age of the cluster.

The most straightforward approach to answering this question would be to look at the mass function of the clusters for a variety of galaxies. Unfortunately, this is quite difficult given the large amounts of dust and the dimming caused by stellar evolution. The only galaxy where this has been attempted in detail is the Antennae (Zhang & Fall 1999). However, we can attempt to make the comparison using the luminosity function, as shown in Figure 9 for 8 galaxies. We find that all the galaxies have luminosity functions with similar slopes, with an average power law index $\alpha = -1.93 \pm 0.06$ (uncertainty in the mean; the scatter is 0.18). The primary difference is the normalization of the luminosity function, with NGC 3256 and NGC 4038/39 having large numbers of clusters while NGC 3921 and HE 2−10 have relatively few clusters. There is no obvious trend for a cutoff at high luminosity for the more quiescent galaxies, suggesting a universal luminosity function is a reasonable approximation.

Such an approach is oversimplified for a number of reasons, primary amongst them being that the luminosity of the clusters vary with time. For example, a single-age burst population will evolve to the right in Figure 9, making it difficult to determine whether the luminosity function is lower because of evolution or due to a smaller number of clusters originally forming. Other difficulties with this simple model are that it assumes similar star formation histories for the various galaxies (e.g. continuous rather than sporadic bursts at different times), and ignores the fact that the faint end will probably undergo rapid evolution as the faint clusters dissolve. Nevertheless, to first order the luminosity functions appear to be remarkably similar in form.

Another approach which allows us to increase the sample at the expense of more scatter for any particular galaxy is to plot the magnitude of the brightest cluster vs. the number of clusters in the galaxy, as shown in Figure 10. This figure uses the groundbreaking survey of Larsen & Richtler (2000), with the additions of some new points for merging and starbursting galaxies drawn from the papers in Tables 2, 3, and 4. We find a clear trend between the number of clusters observed and the magnitude of the brightest cluster. The solid line is the fit to the data (excluding NGC 1569) with a slope $= -2.3 \pm 0.2$.

Reference	Brief Description
Schweizer (1982)	NGC 7252 (ground-based, 6 knots, stat. significant?)
Lutz (1991)	NGC 3597 (ground-based, ≈ 10 knots, lacked resolution)
Holtzman et al. (1992)	NGC 1275 (proposed "protoglobular clusters," $n = 60$)
Whitmore et al. (1993)	NGC 7252 (prototypical merger, $n = 40$)
Crabtree et al. (1994)	NGC 7727 (ground-based)
Whitmore & Schweizer (1995)	NGC 4038/4039 ($n = 700$, Antennae galaxies)
Zepf et al. (1995)	NGC 1275 (ground-based spectra, .1–1 Gyr)
Borne et al. (1996)	Cartwheel galaxy (clusters in rings)
Holtzman et al. (1996)	NGC 3597, NGC 6052 (mergers, not cooling flows)
Schweizer et al. (1996)	NGC 3921 (102 candidate globulars, 49 "associations")
Hilker & Kissler-Patig (1996)	NGC 5018 (several hundred Myr to 6 Gyr)
Miller et al. (1997)	NGC 7252 ($n = 499$, 3 pop., < 10 Myr for R $< 6''$)
Whitmore et al. (1997)	NGC 1700, NGC 3610 (missing link with ellipticals?)
Schweizer & Seitzer (1998)	NGC 7252 (spectra, $n = 8$, ages, metallicities)
Brodie et al. (1998)	NGC 1275 (ground-based spectra, age ≈ 450 Myr)
Carlson et al. (1998)	NGC 1275 ($n = 3000$, mix of red and blue clusters)
Johnson et al. (1998)	NGC 1741 (starburst, interacting, Hickson group)
Stiavelli et al. (1998)	NGC 454 (5–10 Myr, effects of emission on photometry)
Dinshaw et al. (1999)	NGC 6090 ($n = 4$, NICMOS observations)
Zepf et al. (1999)	NGC 3256 ($n = 1000$, 15–20% of U light, break in LF?)
Whitmore et al. (1999)	NGC 4038/4039 ($n = 800$ to 8000, break in LF?)
Gallagher et al. (2000a)	Stephan's Quintet ($n = 150$, galaxies and tidal tails)
Alonso-Herrero et al. (2000)	Arp 299 (ULIRG, $n = 40$)
Forbes & Hau (2000)	NGC 3597 (ground-based, K band, $\alpha = -2$)
Johnson & Conti (2000)	HCG 31 (several in Hickson Compact Group 31)
Gilbert et al. (2000)	NGC 4038/39 (IR spectra, ages, masses)
Mengel et al. (2001)	NGC 4038/39 (IR spectra, ages, masses)

Table 2. Observations of interacting galaxies with young star clusters

Reference	Brief Description
Arp & Sandage (1985)	NGC 1569 (ground-based, coined "super star clusters")
Kennicutt & Chu (1988)	LMC (cores of H II regions may be globular clusters)
Meurer et al. (1992)	NGC 1705 (ground-based, 10^6 M$_\odot$)
Conti & Vaca (1994)	He 2-10 (Wolf-Rayet galaxy, 1–10 Myr, 10^5–10^6 M$_\odot$)
Hunter et al. (1994)	NGC 1140 ($n = 7$, merger?, 3–15 Myr)
O'Connell et al. (1994)	NGC 1569, 1705 ($n = 3$, $R_{\rm eff} \approx 3$ pc, density $\gg$ R136)
Meurer et al. (1995)	9 starbursts (20% of UV from clusters, $\alpha = -2$)
O'Connell et al. (1995)	M82 ($n \approx 100$, $R_{\rm eff} = 3.5$ pc, near center)
Watson et al. (1996)	NGC 253 ($n = 4$, brightest = -15 mag and 1.5×10^6 M$_\odot$)
Leitherer et al. (1996)	NGC 4214 (FOC and FOS, $n = 200$, 4–5 Myr)
de Marchi et al. (1997)	NGC 1569 (1569A is superposition of two clusters)
Ho & Fillppenko (1997)	NGC 1705, NGC 1569 (spectra, velocity disp., 3.3×10^5 M$_\odot$)
Calzetti et al. (1997)	NGC 5253 (BCG, $n = 6$, 2.5 Myr, $\approx 10^6$ M$_\odot$)
Östlin et al. (1998)	ESO338−IG04 (BCG)
De Grijs et al. (2001)	M82 (outer region, 20–100 Myr, fainter than -10 mag)
Gallagher et al. (2000)	NGC 7673, NGC 3310, Haro I (clumps of SSCs)
Johnson et al. (2000)	He 2−10 (Wolf-Rayet galaxy, WFPC2, H$_\alpha$, GHRS)
Smith et al. (2000)	M82 (ground-based spectra, 60 Myr, 2×10^6 M$_\odot$)
Östlin (2000)	Mrk 930, ESO185−IG13, ESO350−IG38 (BCGs)
Meurer (2000)	NGC 3310 (0–few 100 Myr, continuous formation)

Table 3. Observations of starburst galaxies with young star clusters

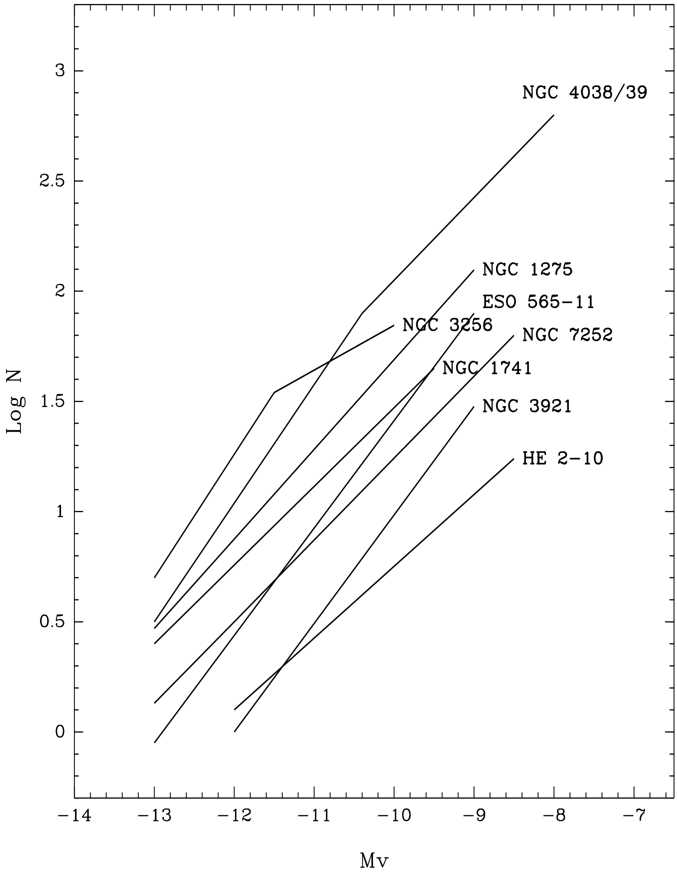


FIGURE 9. Approximate luminosity functions for galaxies in Table 1, normalized to have 0.25 mag bins.

The dotted line shows the trend expected if there is a universal luminosity function with $\alpha = -2$ and the increase in the luminosity is simply due to a larger sample of clusters (i.e. the slope is -2.5). Again, to first order it appears that a universal luminosity function can explain the data, even with the large scatter expected from low number statistics, non-uniform databases, differences in selection criteria, and differences in cutoff magnitudes (only those with cutoffs ≈ -9 have been included).

However, this may not be the whole picture. It is easy to think of examples that do not appear to fit this picture. For example, the brightest clusters in NGC 1569 and NGC 1705 are 2–3 magnitudes brighter than the second brightest cluster in the galaxy (O'Connell, Gallagher & Hunter 1994, see Figure 9 of Meurer et al. 1995 which provides a graphical representation for NGC 1705), suggesting something special is happening in these clusters. In addition, many galaxies are currently forming a few very young compact clusters (e.g. the central regions of the Milky Way and NGC 7252) which, assuming a steady formation rate over a long period of time, implies that a few very massive clusters should eventually form, based on the statistics. These appear to be missing. However, it is possible that the young clusters are forming in regions that are not conducive to the long-term survival of the clusters, such as near the center of the galaxy (Figer et al. 1999), or in spiral arms where frequent encounters with giant molecular clouds may disrupt young clusters. Conversely, conditions appear to be globally conducive to forming clusters in

Reference	Brief Description
Barth et al. (1995)	NGC 1097, NGC 6951 (barred, 2–3 pc, $n = 88$ and 24)
Holtzman (1996)	Abell 496, 1795, 2029, 2597 (not related to cooling flows)
Lee et al. (1997)	UGC 7636 (dwarf near NGC 4472, $n = 18$, tidal tail)
Carollo et al. (1997)	35 spirals (M_V vs. R_e diagram)
Bresolin et al. (1998)	OB associations and populous clusters in seven spirals
Tyson et al. (1998)	NGC 5548 (in tidal tail of Seyfert galaxy)
Buta et al. (1999)	ESO 565-11 (barred, $n = 700$, 4–6 Myr, $\alpha = -2.2$)
Chandar et al. (1999)	44 clusters in M33 (6×10^2–10^5 $M_\odot$)
Buta et al. (2000)	NGC 1326 (barred, $n = 269$, 5–200 Myr)
Chandar et al. (2000)	4 clusters in NGC 6822, spectroscopy, ages
Larsen & Richtler (2000)	21 spirals (ground-based spirals, also mergers & starbursts)
Gallagher et al. (2000)	Stephans' Quintet (Hickson compact group)
Knierman et al. (2000)	Tidal tails in 4 mergers (not all have clusters)

TABLE 4. Observations of other galaxies with young star clusters

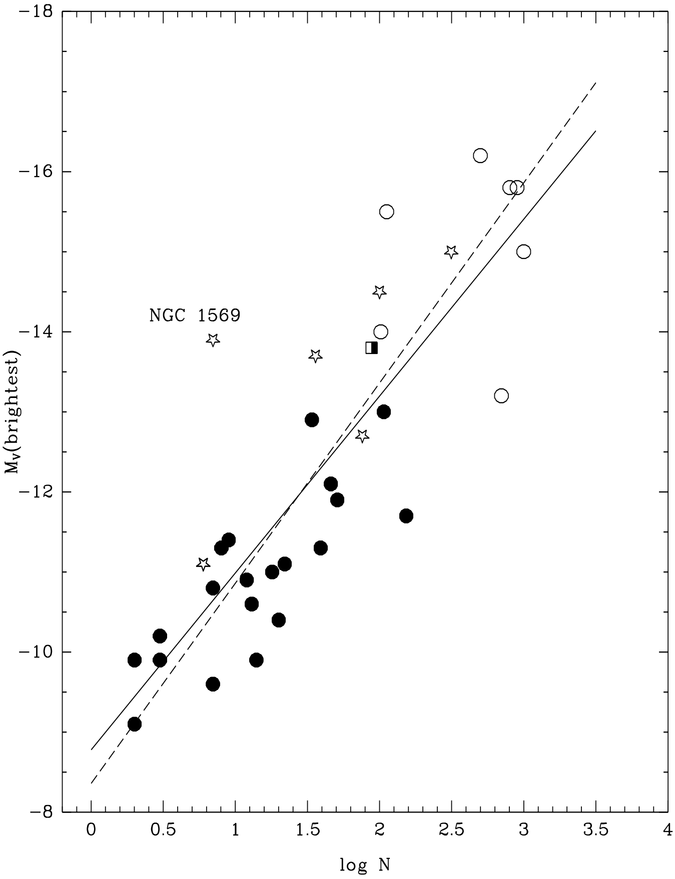


FIGURE 10. Plot of the magnitude of the brightest cluster vs. the log of the number of clusters. Filled circles are spiral galaxies from Larsen & Richtler (2000), open circles are mergers, stars are starbursting galaxies, and the half filled square is a barred galaxy (Table 1). The solid line is a best fit (excluding NGC 1569) while the dashed line is the prediction from a universal power law luminosity function with index $\alpha = 2$. See text for more details.

merging galaxies, where violent relaxation will populate all available orbits. These clusters are more likely to survive.

5. Current and future questions

The *HST* observations of star clusters have answered many questions, but typical of any active field of science, they have introduced even more new questions. Here are some of the fundamental questions that should be addressed over the coming decade.

5.1. *Will some of the young compact clusters survive to become classical globular clusters, and if so, how many?*

Historically, astronomers have been approaching this question from two different directions. Looking at resolved clusters, Kennicutt and Chu (1988) concluded "that populous clusters may be forming in giant H II regions, but only a small fraction of giant H II regions are likely to contains such clusters." The prototype for this idea is 30 Doradus. For the more distant galaxies where the clusters are not well resolved, most of the original motivation came from trying to understand the specific frequency of globular clusters in elliptical galaxies (Schweizer 1982, Burstein, 1982, Ashman & Zepf 1992). As pointed out in §3.1.5, we actually need the vast majority of clusters to dissolve or we end up with specific globular cluster frequencies that are too high.

It now seems well established that some of the young compact clusters will survive to form globular clusters. For example, in NGC 7252 and NGC 3921 the clusters are already 500 Myr (several hundred crossing times), have the distribution expected of globular clusters, and have the same masses and densities of classical globular clusters. In addition, intermediate-age clusters with ages ≈ 3 Gyr are found in dynamically young ellipticals (e.g. Whitmore et al. 1997, Goudfrooij et al. 2000). The remaining question is *how many* of the clusters will survive and become old globular clusters. In particular, is this how the red (metal-rich) population of globular clusters found in elliptical galaxies are formed, as Ashman & Zepf (1992) propose, or is this just a minor trace population? For example, Schweizer et al. (1996) concludes that the total number of globular clusters in NGC 3921 has increased by only 40%. While sizeable, this may not be enough to explain the increase in S_N in ellipticals unless the typical elliptical has several major mergers in its lifetime. In addition, a clear prediction from Ashman & Zepf is that the ratio of red (metal-rich) to blue clusters should increase for high S_N galaxies. However, Forbes, Brodie, & Grillmair (1997) find that the number of red clusters does not increase with S_N. It is possible that this is because their sample is dominated by ellipticals in clusters of galaxies, where other mechanisms might also be operating (e.g. stripping the globular clusters out of nearby dwarf galaxies; see the Harris review in this volume). In summary, it appears that many of the brighter young compact clusters will become classical globular clusters, but the jury is still out on whether this is the cause of the increase in S_N for elliptical galaxies.

5.2. *What fraction of stars are formed in clusters?*

Since only a subsample of the young clusters are likely to survive, an obvious question is whether most of the field stars in a galaxy are originally formed in clusters. In the Milky Way, approximately 0.1% of the stars are currently in globular clusters. However, in some starbursting and merging galaxies the fraction of light from the clusters is as high as 20% (Meurer et al. 1995). Even in these young star forming regions many of the field stars are from clusters that have already dissolved, hence the true percentage of stars that were originally in clusters is even higher, and might conceivably be ≈ 100%.

HST observations may allow us to answer this question by determining the rate at which clusters dissolve. For example, if we were to assume that the Antennae has been making clusters at the same rate for the past 200 Myr (a rather uncertain assumption to say the least), we could use Figure 2 from Zhang & Fall (reproduced as Figure 7) to show that for every 20 clusters originally formed, only about one will survive to an age of ≈ 100 Myr (i.e. there are roughly the same number of clusters in the 0–10 Myr age bin as in the 20–200 Myr age bin). While this very crude calculation is probably not justifiable for a single galaxy, which we may be catching during the peak of cluster formation, once a larger sample becomes available this would be a reasonable approach.

An intriguing result is the finding that the luminosity function in the Ursa Major dwarf spheroidal galaxy and the globular cluster M15 are essentially indistinguishable (Wyse et al. 1999), even though the densities differ by three orders of magnitude. It is tempting to suggest that perhaps most of the stars are formed in groups and clusters, and that the field stars are simply the remnants of the fainter, less dense clusters which have dissolved.

5.3. *What fraction of star formation is triggered by other star formation?*

There appears to be a variety of ways to form stars (e.g. gravitational instabilities, shocks between colliding clouds of gas, enhanced pressure of the ISM, etc.). As discussed in §2.3, *HST* observations suggest that star formation can also be triggered by nearby bursts of star formation (e.g. around 30 Doradus; see Figures 1 and 2). An interesting question is what fraction of all stars have been formed this way? At any one time only a relatively small fraction of star formation appears to be triggered (e.g. the clusters around 30 Doradus are relatively modest compared to 30 Doradus itself). However, in principle this is a self-propagating process which may continue over a much longer period of time, hence it is possible that overall a relatively large fraction of star formation is triggered. The fact that much of the triggered star formation is still embedded in dust clouds makes it difficult to obtain a complete census. New observations with the NICMOS + cryocooler, and the IR channel of the WFC3, will help answer the question of how important this mechanism is to the total production of stars in a galaxy.

5.4. *Is a massive open cluster the same as a low-mass globular cluster?*

The fact that the luminosity functions for young clusters are power laws begs the question of whether there is anything fundamentally different between a massive open cluster and a low-mass globular cluster. Are we looking at a continuum, or a bimodal distribution with fundamentally different formation mechanisms for open clusters and globular clusters? It seems possible that the distinction between the two is artificial, and is due to the fact that we live in a galaxy which had an initial burst of star formation 14 Gyr ago but no major bursts since then. The only clusters that have survived from the initial burst are, by necessity, massive and compact. These we call globular clusters. In the present epoch, the star formation rate is percolating at a much lower rate and we are only able to see the spectrum of clusters from associations to open cluster. We do not see young globular clusters for several possible reasons. First, they should only form very rarely, since the star formation rate is so low (see §4). Second, we would probably call them open clusters anyway, since we are not used to calling anything young a globular cluster. Indeed, there is overlap in the masses of open and globular clusters. Candidate open cluster/globular clusters might include M67 (5 Gyr), Be 17 and Lynga 7, which according to Phelps et al. (1994), may be as old as the youngest globular clusters.

5.5. *Can we develop a unified picture of cluster formation that explains all this?*

While we are making good progress understanding many pieces of the puzzle, how it all fits together is a much tougher question. Is it possible to develop a universal model that provides a framework for understanding cluster formation both near (e.g. the classic picture of associations, open clusters, and globular clusters in the Milky Way), intermediate ("populous" star clusters in the LMC and "super star clusters" in nearby dwarf starbursts), and far (the young compact clusters in mergers and starbursts); for spiral and elliptical galaxies; for the initial collapse of many galaxies ≈ 14 Gyr ago and mergers of galaxies in the local universe; and for violent starbursts as well as "quiescent" galaxies? Some of the ideas discussed in this review lead to the following working hypothesis, portions of which several groups are pursuing; in particular Elmegreen & Efremov (1997), Vesperini (1998), and Fall & Zhang (2001).

The luminosity functions for young clusters (e.g. Whitmore & Schweizer 1995), molecular clouds (Harris & Pudritz 1994) and H II regions (Elmegreen & Efremov 1997) are all power laws with index ≈ -2. Hence, we start with a universal power law luminosity (mass) function. The total number of clusters is normalized depending on how active the star formation is (e.g. Larsen & Richtler 2000, §4). This explains the existence of large numbers of bright clusters in mergers, since they have the most active star formation. The physics of how the cluster formation is triggered is still uncertain, but several possible mechanisms have been proposed(e.g. Jog & Solomon 1992, Kumai et al. 1993, Elmegreen & Efremov 1997). The power law evolves due to both internal (e.g. evaporation, stellar mass loss) and external (e.g. tidal stress, triggered star formation) influences (e.g. Fall & Zhang 2001), with most of the faint and/or diffuse clusters dissolving, just as they do in the Milky way. This model would then need to be convolved with models of galactic evolution (i.e. a combination of initial collapse, hierarchical merging, and internal galactic dynamics), and stellar evolution (dimming and reddening of the starlight) to produce the wide variety of cluster demographics we see in galaxies today.

5.6. *What is the limiting redshift for observing young globular clusters with HST and NGST?*

The current limiting redshift for observing young globular clusters with characteristics similar to the brightest young clusters in the Antennae is $Z \approx 0.5$, using the WFPC2 on *HST*. It will be possible to do slightly better (i.e. $Z \approx 0.8$), with the Advanced Camera for Surveys since it has a quantum efficiency which is roughly 3 times better than WFPC2. However, the real breakthrough will be *NGST*, where Burgarella & Chapelon (1998) estimate that it will be possible to observe young globular clusters out to $Z \approx 9$, if they exist. Since globular clusters appear to be the oldest fossils we observe in galaxies it is quite possible that the first objects we will see emerging from the "dark ages" will be young compact star clusters, similar to what we are seeing in nearby galaxies.

REFERENCES

Arp, H. & Sandage, A. 1985 *AJ* **90**, 1163.

Ashman, K. M. & Zepf, S. E. 1992 *ApJ* **384**, 50.

Alonso-Herero, A., Rieke, G. H., Rieke, M. J., & Scoville, N. Z. 2000 *ApJ* **532**, 845.

Barth, A. J., Ho, L. C., Filippenko, A. V., & Sargent, W. L. W. 1995 *AJ* **110**, 1009.

Borne, K. D., et al. 1996, private communication.

Brandl, B., et al. 2000. In *From Darkness to Light* (eds. T. Montmerle & P. Andre). ASP Conf. Ser.; astro-ph/0007298.

Brandl, B., Bertoldi, S. F., Eckart, A., Genzel, R., Drapatz, S., Hofman, R., Lowe, M., & Quirrenbach, A. 1996 *ApJ* **466**, 254.

Brandl, B., Brandner, W., Eisenhauer, F., Moffat, A. F. J., Palla, F., & Zinnecker, H. 1999 *A&A* **352**, L69.

Bresolin, F., et al. 1998 *AJ* **116**, 119.

Brodie, J. P., Schroder, L. L., Huchra, J. P., Phillips, A. C., Kissler-Patig, M., & Forbes, D. F. 1998 *AJ* **116**, 691.

Brown, R. J. N., Forbes, D. A., Kissler-Patig, M., & Brodie, J. P. 2000 *MNRAS* **317**, 406.

Bruzual, A. G. & Charlot, S. 1996, unpublished.

Burgarella, D. & Chapelon, S. 1998. In *Science with the NGST* (eds. E. P. Smith & A. Koratkar) ASP Series Vol. 133, p. 227. ASP.

Burstein, D. 1987. In *Nearly Normal Galaxies* (ed. S. Faber), p. 47. Springer.

Buta, R., Crocker, D. A., & Byrd, G. G. 1999 *AJ* **118**, 2071.

Buta, R., Treuthardt, P. M., Byrd, G. G., & Crocker, D. A. 2000 *AJ* **120**, 1289.

Calzetti, D., Meurer, G. R., Bohlin, R. C., Garnett, D. R., Kinney, A. L., Leitherer, C., Storchi-Bergmann, T. 1997 *AJ* **114**, 1834.

Carlson, M. N. & the WFPC2 team 1998 *AJ* **115**, 1778.

Carollo, C. M., Stiavelli, M., de Zeeuw, P. T., & Mack, J. 1997 *AJ* **114**, 2366.

Cassinelli, J. P., Mathhis, J. S., & Savage, B. D. 1981 *Sci* **212**, 1497.

Chandar, R., Bianchi, L., & Ford, H. C. 2000 *AJ* **120**, 3088.

Chandar, R., Bianchi, L., Ford, H. C., & Salasnich, B. 1999 *PASP* **111**, 794.

Conti, P. S. & Vacca, W. D. 1994 *ApJ* **423**, L97.

Crabtree, D. R. & Smecker-Hane, T. A. 1994 *BAAS* **26**, 1494.

de Grijs, R., O'Connell, R. W., & Gallagher, J. S. 2001 *AJ* **121**, 768.

de Marchi, G., Clampin, M., Greggio, L., Leitherer, C., Nota, A. & Tosi, M. 1997 *ApJ* **479**, L27.

Dinshaw, N., Evans, A. S., Epps, H., Scoville, N. Z., & Rieke, M. 1999 *ApJ* **525**, 702.

Eisenhauer, F., Quirrenbach, A., Zinnecker, H., & Genzal, R. 1998, *ApJ* **498**, 278.

Elmegreen, B. G. 1999. In *The Evolution of Galaxies on Cosmological Timescales* (eds. J. E. Beckman & T. J. Mahoney) ASP Conference Series, San Francisco. ASP.

Elmegreen, B. G. & Efremov, Y. N. 1997 *ApJ* **480**, 235.

Fall, S. M. & Zhang, Q. 2001 *ApJ*, submitted.

Figer, D. F., Kim, S. S., Morris, M., Serabyn, E., Rich, R. M., & McLean, I. S. 1999 *ApJ* **525**, 750.

Forbes, D., Brodie, J. P., & Grillmair, C. J. 1997 *AJ* **113**, 1652.

Forbes, D. A. & Hau, G. K. T. 2000 *MNRAS* **312**, 703.

Fritze-von Alvensleben, U. 1999 *A&A* **342**, L25.

Gallagher, J. S., Homeier, C. J., & Conselice, C. J. 2000a. In *Massive Stellar Clusters* (eds. A. Lancon & C. Boily), p. 258. ASP Conf. Ser.

Gallagher, S. C., Hunsberger, S. D., Charlton, J. C., Zaritsky, D. 2000b. In *Massive Star Clusters* (eds. A. Lancon & C. M. Boily), p. 247. ASP Conf. Ser.

Gilbert, A., Graham, J. R., McLean, I. S., Becklin, E. E., Larkin, J., Wilcox, M. K., Figer, D. F., Levenson, N. A., & Teplitz, H. I. 2000. In *Massive Stellar Clusters*, (eds. A. Lancon & C. M. Boily), p. 101. ASP Conf. Ser.

Goudfrooij, P., Mack, J., Kissler-Patig, M., Meylan, G., & Minniti, D. 2000 *MNRAS* **322**, 643.

Harris, W. 1991 *ARA&A* **29**, 543.

Harris, W. 1999, private communication.

Harris, W. E. & Pudritz, R. E. 1994 *ApJ* **429**, 177.

Hilker, M. & Kissler-Patig, M. 1996 *A&A* **314**, 357.

Ho, L. C. & Filippenko, A. V. 1997 *ApJ* **472**, 600.

Holtzman, J. A., et al. (the WFPC2 team) 1996 *AJ* **112**, 416.

Holtzman, J. A., et al. (the WFPC team) 1992 *AJ* **103**, 691.

Hunter, D. A., O'Connell, R. W., & Gallagher, J. S. 1994 *AJ* **108**, 84.

Hunter, D. A., Shaya, E. J., Holtzman, J. A., Light, R. M., O'Neil, E. J., & Lynds, R. 1995 *ApJ* **448**, 179.

Hyland, A. R., Straw, S., Jones, T. J., & Gatley, I. 1992 *MNRAS* **257**, 391.

Jog, C. & Solomon, P. M. 1992 *ApJ* **387**, 152.

Johnson, K. E. & Conti, P. 2000 *AJ* **119**, 2146.

Johnson, K. E., Vacca, W. D., Leitherer, C., Conti, P., & Lipsey, S. J. 1998 *AJ* **117**, 1708.

Kennicutt, R. C. & Chu, Y.-H. 1988 *AJ* **95**, 720.

Kim, S. S., Morris, M., & Lee, H. M. 1999 *ApJ* **525**, 228.

Knierman, K., Hunsberger, S. C., Gallagher, S. D., Charlton, J. C., Whitmore, B. C., Kundu, A., Hibbard, J. E., Zaritsky, D. 2000, astro-ph/0009196.

Kumai, Y., Hashi, Y., & Fujimoto, M. 1993 *ApJ* **404**, 144.

Larsen, S. S. & Richtler, T. 1999 *A&A* **345**, 59.

Larsen, S. S. & Richtler, T. 2000 *A&A* **354**, 836.

Larson, R. 1999. In *Star Formation 1999* (ed. T. Nakamoto), p. 336. Nobeyama Radio Observatory.

Lee, M. G., Kim, E., & Geisler, D. 1997 *AJ* **114**, 1824.

Leitherer, C. & Heckman, T. M. 1995 *ApJS* **99**, 173.

Leitherer, C., Vacca, W. D., Conti, P. S., Filippenko, A. V., Robert, C., Sargent, W. L. W. 1996 *AJ* **465**, 717.

Lutz, D. 1991 *A&A* **245**, 31.

Malumuth, E. M. & Heap, S. R. 1994 *AJ* **107**, 1054.

Mandushev, G., Spassova, N., & Staneva, A. 1991 *A&A* **252**, 94.

Massey, P. & Hunter, D. 1998 *ApJ* **493**, 180.

Mengel, S., Lehnert, M. D., Thatte, N., Tacconi-Garman, L. E., & Genzel, R. 2001 *ApJ* **550**, 280.

Meurer, G. R. 2000. In *Massive Stellar Clusters* (eds. A. Lancon & C. M. Boily), p. 81. ASP Conf. Ser.

Meurer, G. R., Freeman, K. C., Dopita, M. A., & Cacciari, C. 1992 *AJ* **103**, 60.

Meurer, G. R., Heckman, T. M., Leitherer, C., Kinney, A., Robert, C., & Garnett, D. R. 1995 *AJ* **110**, 2665.

Mihos, J. C., Bothun, G. D., & Richstone D. O. 1993 *ApJ* **418**, 82.

Miller, B. W., Whitmore, B. C., Schweizer, F., & Fall, S. M. 1997 *AJ* **114**, 2381.

Mirabel, I. F., et al. 1998 *A&A* **333**, L1.

O'Connell, R. W., Gallagher, J. S., & Hunter, D. A. 1994 *ApJ* **433**, 65.

O'Connell, R. W., Gallagher, J. S., Hunter, D. A., & Colley, W. N. 1995 *ApJ* **446**, L1.

Östlin, G., Bergvall, N., & Ronnback, J. 1998 *A&A* **335**, 85.

Östlin, G. 2000. In *Massive Stellar Clusters* (eds. A. Lancon & C. M. Boily), p. 63. ASP Conf. Ser.

Phelps, R. L., Janes, K. A., & Montgomery, K. A. 1994 *AJ* **107**, 1079.

Portegies-Zwart, S. F. 2001 *ApJ* **546**, 101.

Richer, H. B., Crabtree, D. R., Fabian, A. C., & Lin, D. N. C. 1993 *AJ* **105**, 877.

Salpeter, E. E. 1955 *ApJ* **121**, 161.

Sanders, D. B., et al. 1988 *ApJ* **325**, 74.

Scalo, J. 1998. In *The Stellar Initial Mass Function* (eds. G. Gilmore & D. Howell), p. 201. ASP Conf. Ser.

Schweizer, F. 1982 *ApJ* **252**, 455.

Schweizer, F. 1987. In *Nearly Normal Galaxies* (ed. S. Faber), p. 18. Springer.

Schweizer, F., Miller, B., Whitmore, B. C., & Fall, S. M. 1996 *AJ* **112**, 1839.

Schweizer, F. & Seitzer, P. 1998 *AJ* **116**, 2206.

Scowen, P. A., et al. 1998 *AJ* **116**, 163.

Sirianni, M., Nota, A., Leitherer, C., De Marchi, G. D., & Clampin, M. 2000 *AJ* **533**, 203.

Smith, L. & Gallagher, J. S. 2000. In *Massive Stellar Clusters* (eds. A. Lancon & C. M. Boily), p. 90. ASP Conf. Ser.

Stiavelli, M., Panagia, N., Carollo, M., Romaniello, M., Heyer, I., & Gonzaga, S. 1998 *ApJ* **492**, L135.

Toomre, A. 1977. In *The Evolution of Galaxies and Stellar Populations* (eds. B. M. Tinsley & R. B. Larson), p. 401. Yale.

Tyson J. A., et al. 1998 *AJ* **116**, 102.

van den Bergh, S. 1995 *Nature* **374**, 215.

van den Bergh, S. 1996 *AJ* **112**, 2634.

van den Bergh, S. 1990. In *Dynamics and Interactions of Galaxies* (ed. R. Wielen), p. 492. Springer.

van den Bergh, S. 2000 *PASP* **112**, 932.

Vesperini, E. 1998 *MNRAS* **299**, 1019.

Vigroux, L., et al. 1996 *A&A* **315**, L93.

Walborn, N. R. & Blades, J. C. 1997 *ApJ* **112**, 457.

Walborn, N. R., Drissen, L., Parker, J. W., Saha, A., MacKenty, J. W., & White, R. L. 1999a *AJ* **118**, 1684.

Walborn, N. R., Barba, R. H., Brandner, W., Rubio, M., Grebel, E., & Probst, R. 1999b *AJ* **117**, 225.

Watson, A., et al. (WFPC2 team) 1996 *AJ* **112**, 534.

Weigelt, G. & Baier, G. 1985 *A&A* **150**, L18.

Whiteoak, J. B., et al. 1983 *MNRAS* **205**, 275.

Whitmore, B. C., Miller, B. W., Schweizer, F., & Fall, S. M. 1997 *AJ* **114**, 1797.

Whitmore, B. C. & Schweizer, F. 1995 *AJ* **109**, 960.

Whitmore, B. C., Schweizer, F., Leitherer, C., Borne, K., & Robert, C. 1993 *AJ* **106**, 1354.

Whitmore, B. C., Zhang, Q., Leitherer, C., Fall, S. M., Schweizer, F., & Miller, B. W. 1999 *AJ* **118**, 1551.

Wilson, C. D., Scoville, N., Madden, S. C., & Charmandaris, V. 2000 *ApJ* **542**, 120.

Wyse, R. F. G., Gilmore, G., Feltzing, S., & Houdashelt, M. 1999. In *The Hi-Redshift Universe* (eds. A. Bunker, & W. van Breugel), p. 181. ASP Conf. Ser.

Zepf, S. E. & Ashman, K. M. 1993 *MNRAS* **264**, 611.

Zepf, S. E., Ashman, K. M., English, J., Freeman, K. C., & Sharples, R. M. 1999 *AJ* **118**, 752.

Zepf, S. E., Carter, D., Sharples, R. M., & Ashman, K. M. 1995 *ApJ* **445**, L19.

Zhang, Q. & Fall, S. M. 1999 *ApJ* **527**, L81.

Starburst galaxies observed with the *Hubble Space Telescope*

By CLAUS LEITHERER

Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218 †

The contributions of the *Hubble Space Telescope* to our understanding of starburst galaxies are reviewed. Over the past decade, *HST*'s imagers and spectrographs have returned high-quality data from the far-ultraviolet to the near-infrared at unprecedented spatial resolution. A representative set of *HST* key observations is used to address several relevant issues: Where are starbursts found? What is their stellar content? How do they evolve with time? How do the stars and the interstellar medium interact? The review concludes with a list of science highlights and a forecast for the second decade.

1. Overview

Almost exactly 10 years ago ST ScI hosted its annual symposium entitled *Massive Stars in Starbursts* (Leitherer et al. 1991). Those were the weeks immediately prior to *HST*'s launch, and the conference organizers felt it appropriate to have a meeting on the subject of starbursts because *HST* had the potential for significant contributions. Starbursts are *compact* (10°–10^3 pc), *young* ($\sim 10^6$–10^8 yr) sites of star formation, often with high *dust* obscuration. These properties make starbursts ideal targets for *HST*, given its superior spatial resolution, ultraviolet (UV) sensitivity, and (later-on) infrared (IR) capabilities.

As we all know, the high hopes were not immediately fulfilled, and it was not until after the First Servicing Mission that *HST* lived up to the expectations. Nevertheless, it is worth noting that *HST*'s first scientifically useful image after the launch was a WF/PC exposure of the central region of 30 Doradus in the Large Magellanic Cloud (see p. xi of Leitherer et al. 1991)—a star-formation complex which is considered the "Starburst Rosetta" by Walborn (1991).

Over the first 10 years of its life, and in particular after the installation of Costar, *HST* has made significant contributions to the field of starbursts which have helped address fundamental issues such as: the birthplace and environment of starbursts, the stellar content of starbursts, the temporal and spatial evolution of starburst, and the effects of starbursts on their environment. I will address these points in this review from an observational point of view, guided by the relevant *HST* data. Of course the selection reflects my personal perspective, and space limitations do not allow me to discuss other, similarly important, material.

2. Starburst hosts

Starbursts are a mixed bag. There is a continuum of objects between sites of massive-star formation like W51 in the Galaxy (Goldader & Wynn-Williams 1994) and the nucleus of the nearby dwarf galaxy M82 (Rieke et al. 1980). The latter is considered a proto-typical starburst galaxy whereas the former is usual taken as a "normal" star-formation region. Since starbursts are selected on the basis of their high levels of ionizing and non-ionizing UV radiation (observed directly or via reprocessed recombination-line or thermal

† Operated by AURA for NASA under contract NAS5-26555

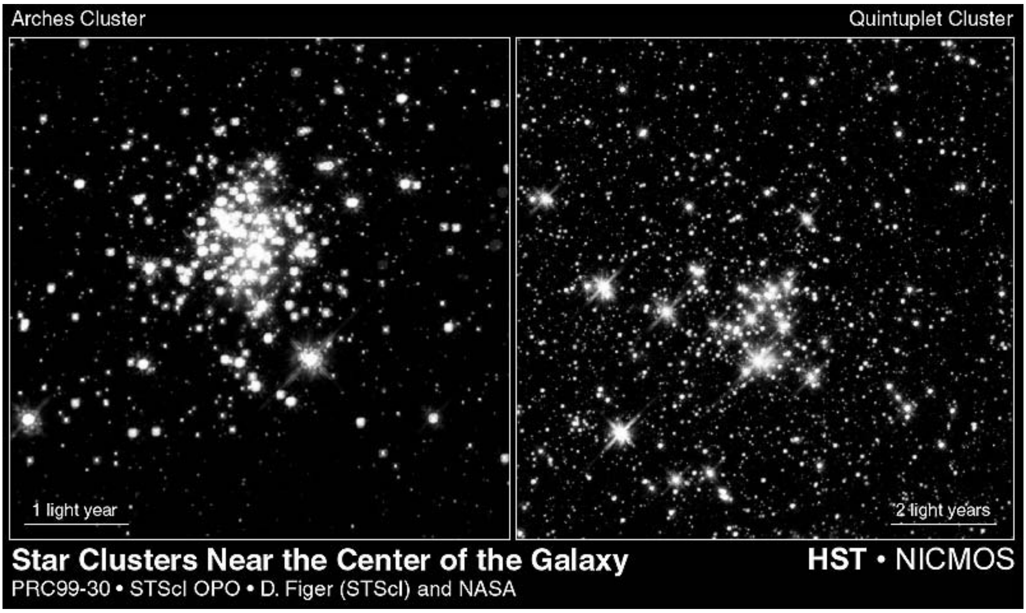

Arches Cluster
Quintuplet Cluster
1 light year
2 light years
Star Clusters Near the Center of the Galaxy
HST • NICMOS
PRC99-30 • STScI OPO • D. Figer (STScI) and NASA

FIGURE 1. Composite *HST* NICMOS F110W, F160W, and F205W images of the Arches and Quintuplet clusters in the Galactic Center region (Figer et al. 1999).

dust emission), I will use an empirical definition: Starbursts have star-formation rates high enough that a statistically significant number of stars form which produce UV radiation. Such stars have masses between 10 and 100 $M_\odot$. Therefore an equivalent definition is to require the upper mass function to be well enough populated that stochastic effects are not important. A sufficiently populated mass function leads to total starburst masses of at least about 10^5 $M_\odot$.

Following the definition of Terlevich (1997), we can extend the definition from a starburst to a starburst galaxy: In a *starburst galaxy* the entire luminosity is due to the starburst itself ($L^{\rm Burst} \approx L^{\rm Galaxy}$); if the starburst luminosity is substantial but smaller than the host galaxy luminosity ($L^{\rm Burst} < L^{\rm Galaxy}$), a *starburst region* in a galaxy is observed; if $L^{\rm Burst} \ll L^{\rm Galaxy}$ for any individual star forming region, the object is classified as a *star-forming galaxy*.

With these definitions, the central region of our Galaxy (Genzel & Eckart 1998) is at the low-mass (and by implication, the low-luminosity) end of the starburst scale. The region has been known for some time to be the site of massive star formation but it was *HST*, together with the largest ground-based telescopes, which provided the first census of the massive-star population. The Galactic Center region is of particular interest due to its proximity of 7.5 kpc, permitting a close-up view of a metal-rich, dust-shrouded small starburst. It can serve as a training ground for calibration of methods to be applied to distant, dust-obscured starburst galaxies.

The strong gravitational field in the Galactic Center is predicted to lead to a rapid evaporation of newly formed star clusters (Kim et al. 1999), consistent with the observed absence of clusters older than tens of Myr. Survival times of young star clusters are of relevance to the interpretation of the cluster luminosity function of merging galaxies where the evolutionary link between newly formed and old globular clusters has not yet been established (Zhang & Fall 1999).

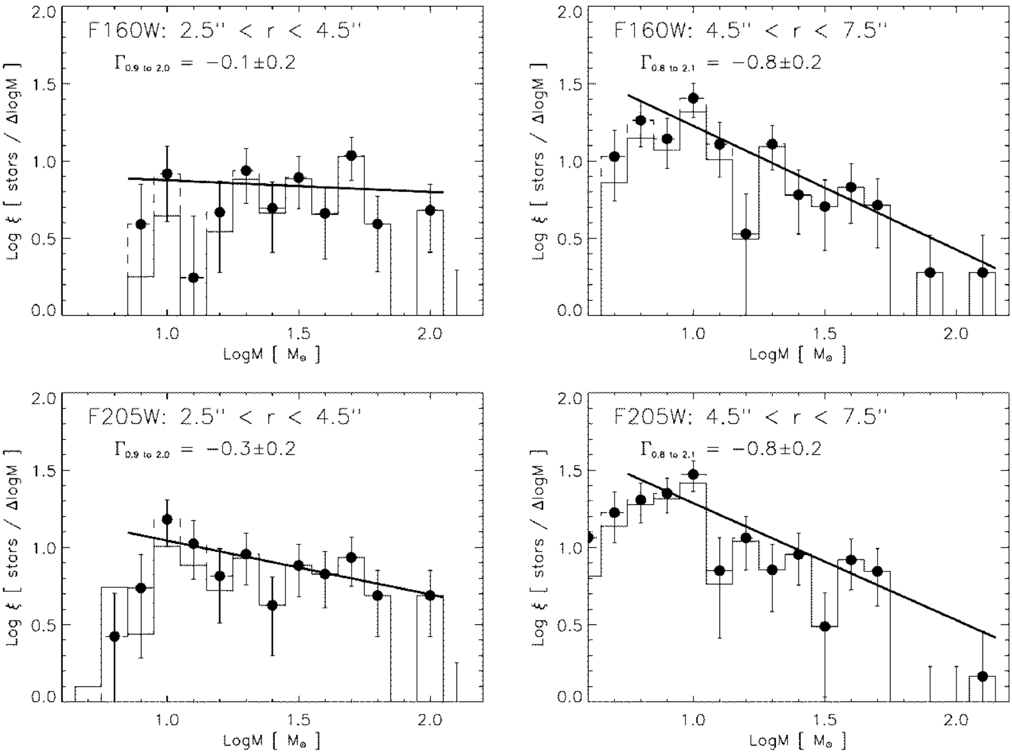


FIGURE 2. IMF of the Arches cluster derived from the F160W (top) and F205W (bottom) photometry of Fig. 1. Left: inner region; right: outer region (Figer et al. 1999).

Evidence for a high metallicity in the center region is overwhelming but individual studies are still discordant. The observed Galactic oxygen abundance gradient of -0.07 dex kpc^{-1} (Smart & Rolleston 1997) suggests $Z \approx 3$ Z$_\odot$ for the Galactic Center. This agrees with an Fe abundance of about 3 Z$_\odot$ found for the luminous, hot "Pistol Star" by Najarro et al. (1999). In contrast, Ramírez et al. (2000) analyzed several red supergiants close to the Galactic Center. They found an average Fe abundance of -0.09 dex, i.e. close to the solar value. The reason for the discrepancy is not yet understood. It may indicate a real abundance spread, or just be caused by systematic errors in either one of the atmosphere analyses.

Figer et al. (1999) performed *HST* NICMOS near-IR imaging of the Arches and Quintuplet clusters, two very young clusters near the Galactic Center. Their composite images are reproduced in Fig. 1. Using crowded-field photometry, the luminosity function could be derived. Subsequently, evolution models allowed conversion to a mass function (Fig. 2). Figer et al. identified main-sequence stars with initial masses well below 10 M$_\odot$ and derived a slope of the initial mass function (IMF) that suggests an excess of massive stars in the cluster center relative to the periphery. The ages of the two clusters are 2–4 Myr.

The Galactic Center can be contrasted with Arp 220, an IR-luminous galaxy at $d = 77$ Mpc, 10^4 times more distant than the Arches and Quintuplet clusters. At that distance, $1''$ corresponds to 400 pc. (For comparison, the smallest structures resolved in the Galactic Center by NICMOS are about 0.004 pc.) A multi-band NICMOS image taken by Scoville et al. (1998) is in Fig. 3. The image clearly resolves the twin nucleus which has resulted from a recent merger. Numerous luminous super star clusters have formed during the recent starburst episode. The total bolometric luminosity of Arp 220

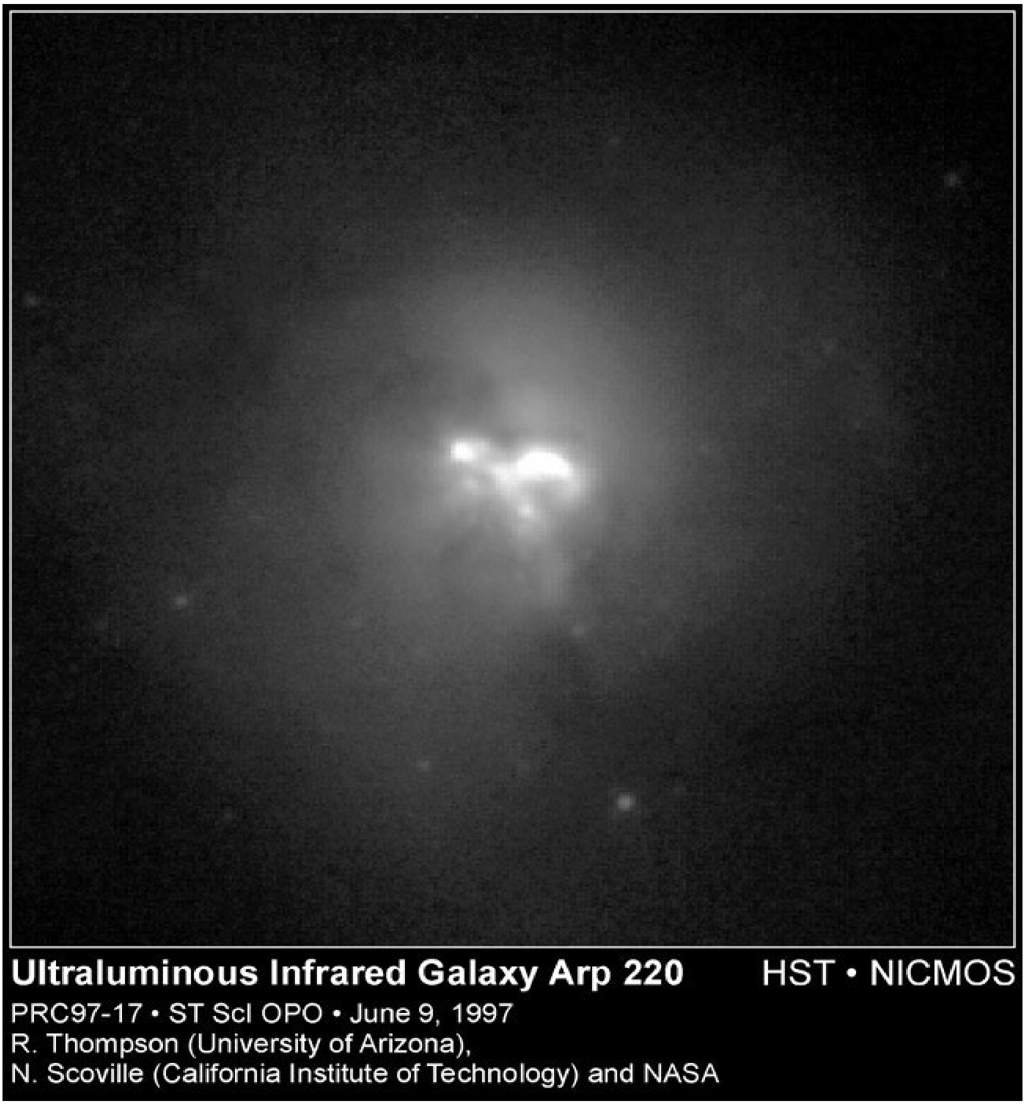

Ultraluminous Infrared Galaxy Arp 220
HST • NICMOS
PRC97-17 • ST ScI OPO • June 9, 1997
R. Thompson (University of Arizona),
N. Scoville (California Institute of Technology) and NASA

Figure 3. Composite *HST* NICMOS F110W, F160W, and F222M image of Arp 220. Field size is 19″ or 7.5 kpc (Scoville et al. 1998).

can be accounted for by star formation alone (Smith et al. 1998). This leaves open the significance of an active galactic nucleus (AGN), which may be contributing as well.

The relative importance of the starburst versus the AGN in Arp 220 is not clear. It is one of *HST*'s major scientific achievements to demonstrate that at least in some galaxies hosting an AGN a central starburst can be dominant in the UV to near-IR energy output and can be significant bolometrically as well. The presence of starbursts in active galaxies was suggested before (e.g. Terlevich 1992) but observational proof has remained elusive: young massive stars have few strong spectral features in the optical and near-IR so that their presence is easily hidden by a strong non-stellar continuum and by emission lines. The situation changes in the UV, where hot stars have unique, broad Si IV λ1400 and C IV λ1550 features. *HST*'s UV sensitivity, combined with its spatial and spectral resolution allowed the detection of *undiluted* stellar lines, and therefore proof that stars

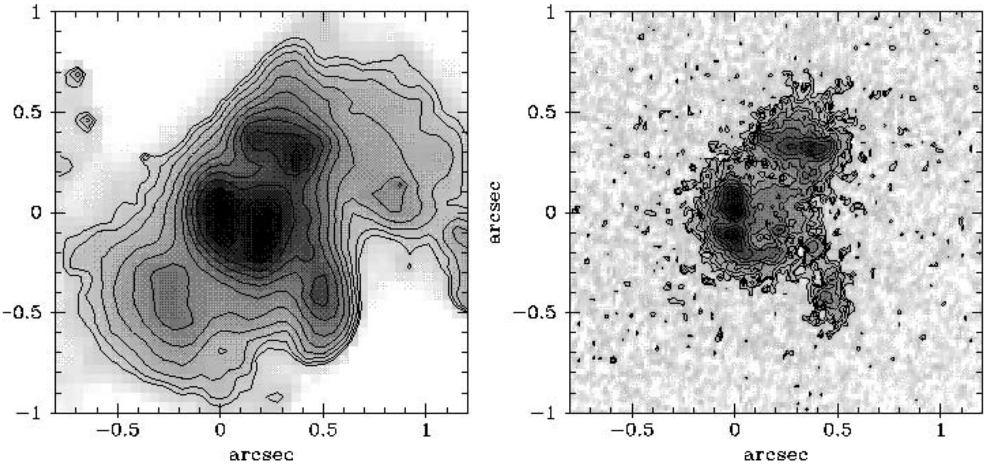


FIGURE 4. Central 2″ × 2″ (620 × 620 pc) of NGC 7130 in optical (left) and UV (right) light (González Delgado et al. 1998).

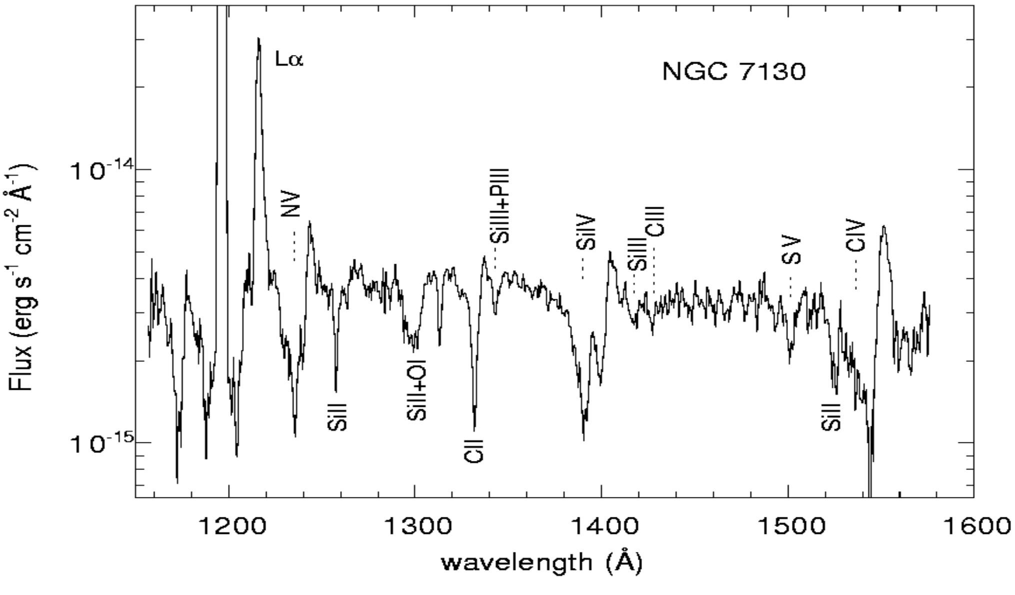


FIGURE 5. UV spectrum of the nucleus of NGC 7130 obtained through the 1.7″ × 1.7″ square aperture of the GHRS on *HST*. The strongest features are the stellar-wind lines of Si IV λ1400and C IV λ1550 (González Delgado et al. 1998).

dominate the continuum in some active galaxies (González Delgado et al. 1998; Maoz et al. 1998).

HST WFPC2 and FOC 2200 Å imaging of a sample of UV-bright Seyfert 2 galaxies was done by González Delgado et al. (1998). Their images of NGC 7130 are shown in Fig. 4. The nucleus, which was previously thought to be point-like, displays a complex morphology, suggestive of a circumnuclear starburst ring. The true nature of the emitted light becomes even clearer from the UV spectrum (Fig. 5). A comparison with a UV spectrum of a genuine starburst region (cf. Fig. 7) convincingly demonstrates that the

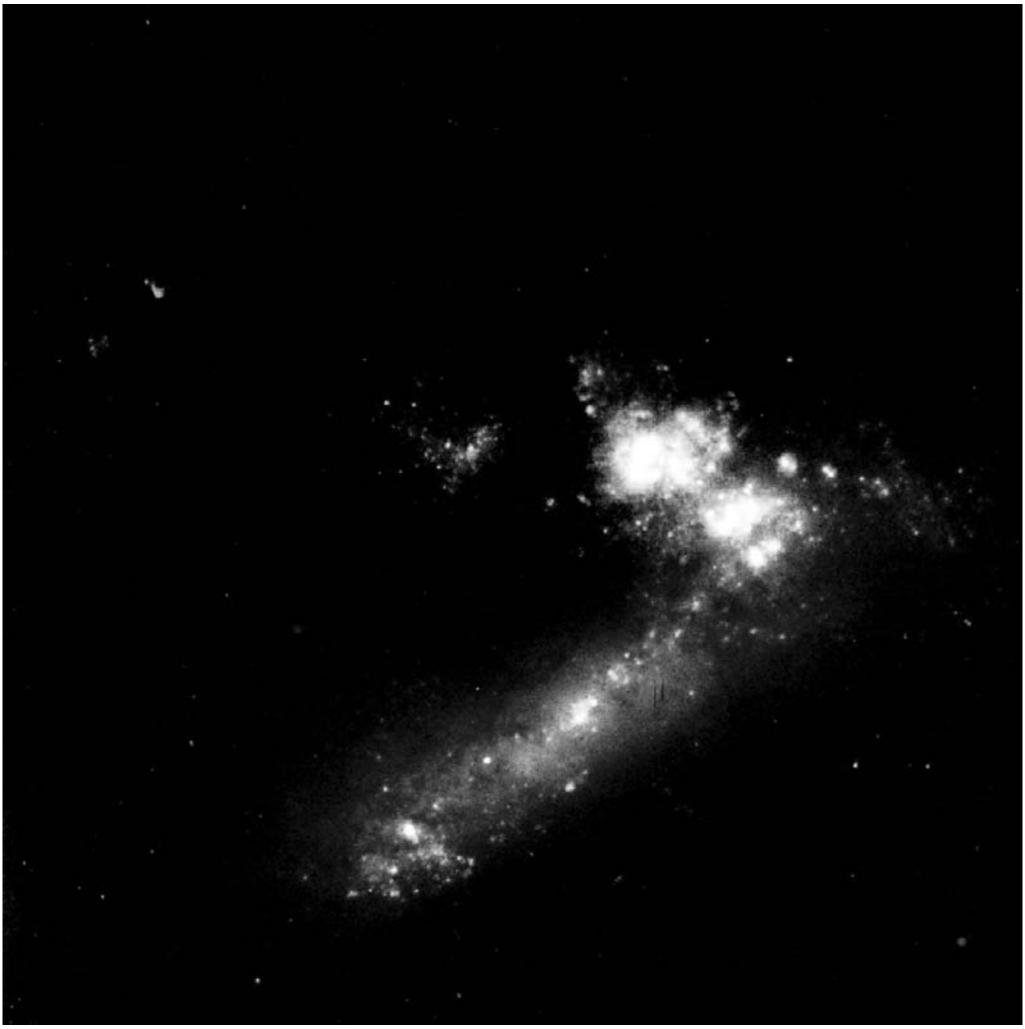


FIGURE 6. WFPC2 F439W, F555W, and F814W composite image of NGC 1741. Field size is $36'' \times 36''$ or 8.7×8.7 kpc (Johnson et al. 1999).

UV light (and the optical to near-IR light as well) comes from stars—as opposed to AGN emission. Clearly these results must not be generalized; the sample was deliberately chosen to maximize the detection probability of hot stars. Nevertheless, the *HST* data are convincing evidence for the ubiquity of starbursts in AGN and their energetic significance in some cases.

After showing examples of nuclear starbursts, of metal-rich, dust-obscured starburst galaxies, and of starbursts in the vicinity of AGNs, I am turning to a more typical case in the local universe. Most known nearby starbursts are hosted by gas-rich galaxies at somewhat subsolar metallicity and luminosity around and below that of our Galaxy. NGC 1741 is a well-studied example (Fig. 6). It is a member of the Hickson compact group 31 and presumably owes its current level of star-formation activity to interaction with one or more companions (Iglesias-Páramo & Vílchez 1997; Johnson et al. 1999; Johnson & Conti 2000). The dominant starburst (recognizable as the bright twin cluster in Fig. 6) is approximately 100 times as luminous as the 30 Doradus region and has an age of 4–5 Myr.

Massive-star formation in starbursts is an important star formation mode. Heckman (1997) estimates that about a quarter of all massive stars in the local universe form in a starburst hosted by galaxies of the types discussed in this section.

3. Stellar content

Even the closest starburst galaxies are at distances > 1 Mpc, making studies of individual stars difficult, if not infeasible. Therefore IMF studies must often rely on less certain, indirect techniques, rather than on a stellar census. The central region of 30 Doradus is one of the few notable exceptions. Although it barely deserves the name "starburst" due to its relatively low mass of order 10^5 $M_\odot$, it is by far the closest, *unobscured* example of a region resembling a starburst galaxy nucleus. *HST* observations have permitted IMF determinations from the least to the most massive stars formed.

WF/PC (Malumuth & Heap 1994) and WFPC2 (Hunter et al. 1995) imagery can fully resolve the central region into stars so that crowded-field photometry techniques can be applied. The derived IMF is close to the traditional Salpeter (1955) slope. For masses higher than about 30 $M_\odot$, optical (and even UV) photometry becomes degenerate with respect to stellar mass, and spectroscopy is required. Massey & Hunter (1998) took advantage of *HST*'s superb capabilities for crowded-field *spectroscopy* and obtained a complete spectroscopic census of the 65 most massive stars in R136, the center of 30 Dor. Their work allowed the extension of the IMF up to $\sim$100 $M_\odot$. 30 Dor is sufficiently massive that the upper IMF is well populated. As a result, the IMF determination of Massey & Hunter has the best number statistics in the 50 to 100 $M_\odot$ range of any observed individual stellar cluster. The high-mass IMF slope turns out to be remarkably similar to a Salpeter IMF.

The low-mass end of the IMF in 30 Dor was studied by Sirianni et al. (2000). They pushed *HST* WFPC2 exposures to the limits and detected stars down to about 1 $M_\odot$. The detection of stars in this mass range is significant since low-mass star formation could be suppressed in the vicinity of massive stars with their prodigious output of ionizing radiation and stellar winds. Sirianni et al. found a flattening of the IMF around 2 $M_\odot$. The overall shape of the mass spectrum at low masses resembles that of the solar neighborhood (Kroupa, Tout, & Gilmore 1993). It is not clear if the low-mass end of the IMF in 30 Dor applies to other starburst regions as well. The low-mass end of the IMF (below $\sim$5 $M_\odot$) in starburst galaxies is inferred from the observed mass-to-light ratio. Dynamical masses of starburst nuclei derived from rotation curves are relatively low, suggesting an absence of stars below 3–5 $M_\odot$ (Rieke 1991; Joseph 1999). Velocity dispersion measurements in individual starburst clusters are an alternative method for a mass estimate. Results obtained for the super star clusters in NGC 1569 and NGC 1705 indicate stars down to about 1–3 $M_\odot$ (Ho & Filippenko 1996). However, systematic uncertainties exist, such as virialization and equipartition.

IMF determinations in starbursts beyond the Local Group must rely on an integrated light analysis. The various techniques are summarized by Leitherer (1998). In this review I will focus on an approach which was made possible with *HST*'s UV capabilities: analysis of ultraviolet line profiles from hot stars in the wavelength region between 1200 Å and 2000 Å. This region is dominated by stellar-wind lines of, e.g. C IV λ1550 and Si IV λ1400, which are the strongest features of hot stars in a young population (e.g. Robert, Leitherer, & Heckman 1993; Leitherer, Robert & Heckman 1995; de Mello, Leitherer, & Heckman 2000). In contrast, the optical and IR spectral regions show few, if any,

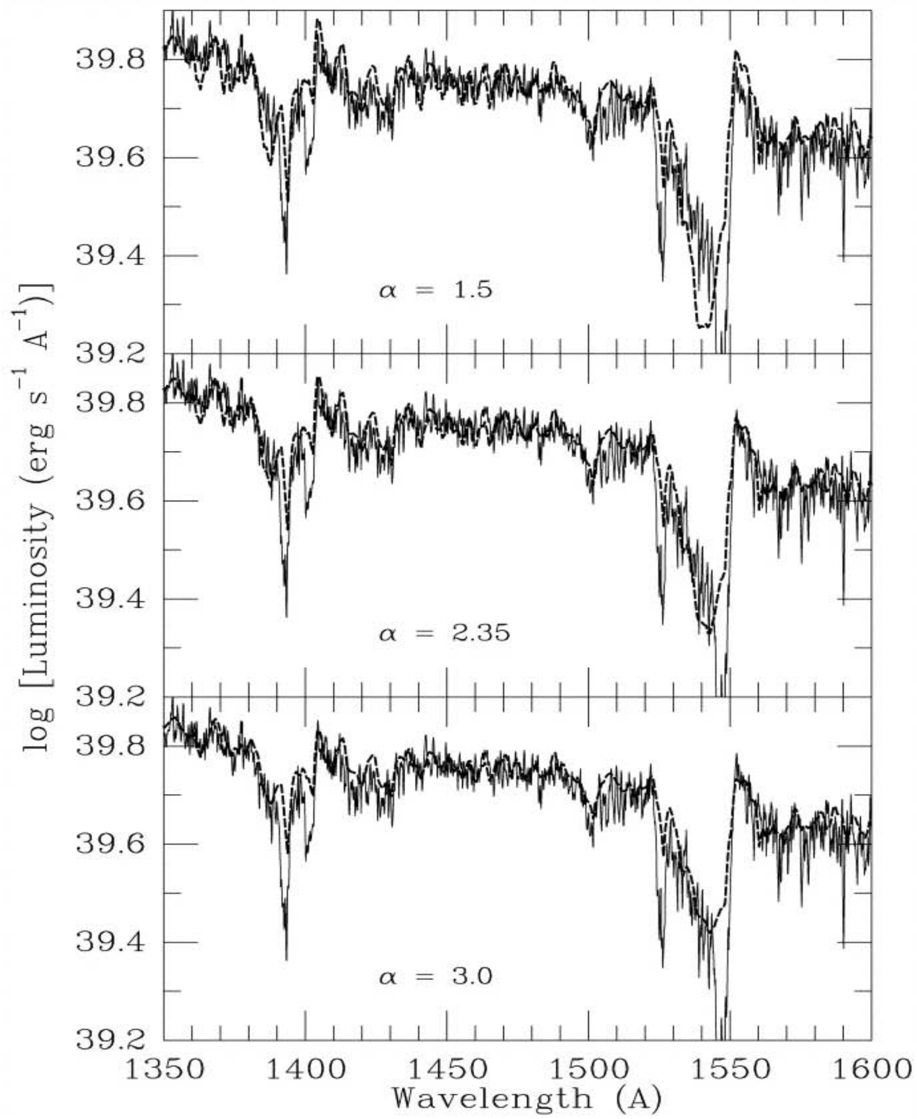


FIGURE 7. Comparison between the observed GHRS UV spectrum of NGC 1741 around Si IV λ1400 and C IV λ1550 (solid lines) and synthetic spectra for three IMF slopes (dashed lines). Upper panel: $\alpha = 1.5$; middle panel: $\alpha = 2.35$ (Salpeter slope); lower panel: $\alpha = 3.0$ (Johnson et al. 1999).

spectral signatures of hot stars, both due to blending by nebular emission and the general weakness of hot-star features longward of 3000 Å. Hot-star winds are radiatively driven, with radiative momentum being transferred into kinetic momentum via absorption in metal lines, like those observed in the satellite-UV. Since the stellar far-UV radiation field depends on the proportion of the most massive, ionizing stars, changes in the IMF and/or the age of the population can be measured as changes in the line profiles. Fig. 7 shows an example of this technique. The GHRS was centered on the southern (lower right in Fig. 6) of the two giant star clusters in NGC 1741. The spectrum was modeled with three choices of the IMF slope. A flatter slope results in a larger number of massive stars

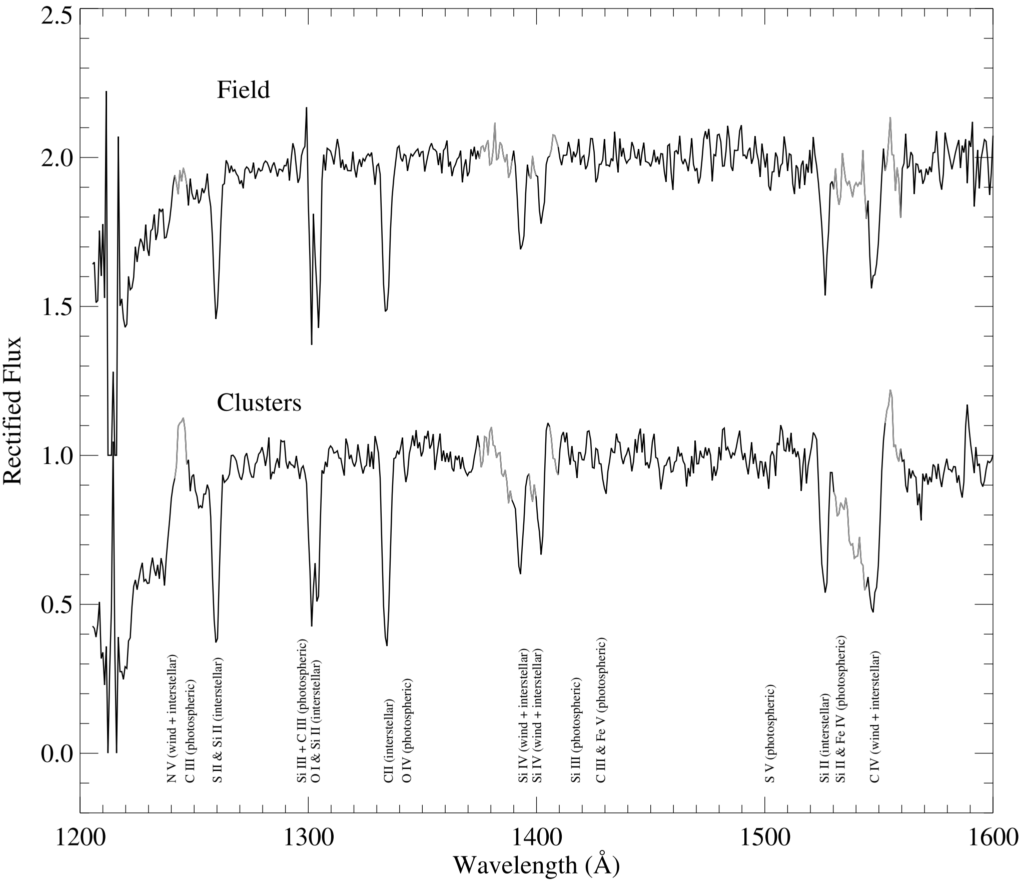


FIGURE 8. STIS UV spectra of cluster G (upper)and of the diffuse field (lower) in NGC 5253 (Tremonti et al. 2000).

and stronger line profiles. The data suggest a slope between $\alpha = 2.35$ and 3.0, which is again close to the Salpeter slope.

Application of this technique to other starburst galaxies leads to similar results: the IMF is remarkably homogeneous, with an average slope close to Salpeter's classical value. This holds over a range of galaxy parameters, in particular metallicity. The surveyed galaxies span the metallicity range from roughly $Z_\odot$ to 0.1 $Z_\odot$, with no systematic trend of the IMF properties. This is consistent with the absence of such a trend in young star clusters and OB associations in the Galaxy and the Magellanic Clouds (Massey 1998).

A caveat remains: The galaxies studied were all UV-selected and are *on average* not IR-bright. Extension of this method to IR-bright galaxies is not currently feasible due to sensitivity limitations of UV detectors. The IMF of IR-luminous galaxies may indeed be different: the hardness of the ionizing radiation field suggests a truncation of the IMF well below 100 $M_\odot$ in these objects (e.g. Goldader et al. 1997). Alternatively, absorption of stellar UV photons by dust could be important even in the near-IR, and observations in the mid-IR would be required. ISO mid-IR spectroscopy does indeed indicate a harder radiation field than inferred from the near-IR (Rigopoulou et al. 1999).

Almost all spectroscopic studies of starburst galaxies naturally focus on the brightest galaxy region, which is the nucleus or another dominant star cluster. Meurer et al. (1995) performed an imaging survey at 2200 Å of a sample of starburst galaxies using *HST*'s

FOC. On average, $\sim$ 20% of the light at 2200 Å comes from a population of compact star clusters, the remaining 80% are diffuse light spread all across the galaxy. The proportion of clusters versus field changes to approximately 5% versus 95% in the optical (e.g. Johnson et al. 1999), most likely the result of different population ages sampled in the optical versus the UV. A more detailed discussion of the overall properties of the star clusters and their significance for galaxy evolution can be found in B. Whitmore's contribution to this volume.

The source of the diffuse UV light could be either unresolved stellar light, or dust-scattered light from the UV-bright cluster stars. Even relatively small amounts of dust inside young star clusters can produce a strong diffuse component, observable either directly in the UV or as thermal dust emission in the far-IR. Examples are 30 Dor (Cheng et al. 1992) or the Galactic reflection nebula IC 435 (Calzetti et al. 1995).

What is the situation in starburst galaxies? In at least one case, NGC 5253, the diffuse component has been shown to be unresolved stellar light by Tremonti et al. (2000). STIS long-slit spectra of a starburst cluster and of the intercluster field are reproduced in Fig. 8. The two spectra are clearly different, therefore scattered light is not important but two distinct stellar populations are observed. The cluster spectrum is characteristic of a single population having an age of a few Myr and with massive stars up to $\sim$100 $M_\odot$. In contrast, the field spectrum has weak Si IV λ1400 and C IV λ1550, suggesting a deficit of very massive stars. This could be an IMF difference between the cluster and field population. Such a difference is observed in the Magellanic Clouds as well (Massey et al. 1995).

4. Evolution of starbursts

I will focus on three topics: the triggering and onset of a starburst, the propagation of star formation, and the termination of a starburst.

The triggering mechanism in starbursts is far from being fully understood. Starbursts can be triggered by a variety of mechanisms, like galaxy-galaxy interactions, merging, secular evolution of bars, and tidal shear in the solid body rotation region (Kennicutt et al. 1987; Sanders et al. 1988; Norman, Sellwood, & Hasan 1996). The general picture is that during any of this processes, the gas in the galaxies is compressed, and while it dissipates energy, moves inward and triggers star formation (e.g. Friedli & Benz 1995; Mihos & Hernquist 1996; Hibbard 1997).

Ultraluminous infrared galaxies (ULIRG) have bolometric luminosities above 10^{12} $L_\odot$. All of their far-IR flux can be accounted for by starburst activity, with some yet unknown contribution from an AGN. ULIRGs are particularly suited to study the relationship between interaction and starbursts since most ULIRGs are found in interacting and/or merging systems (Sanders & Mirabel 1996; Sanders 1997). Borne et al. (2000) surveyed a sample of ULIRGs in the I-band with WFPC2. They identified a significant subsample showing evidence for multiple mergers. The evidence comes from multiple remnants in the galaxy cores and from the fact the some galaxies are found in dense groups of interacting galaxies. This raises the possibility that the progenitors of ULIRGs may be classical, weakly interacting compact groups of galaxies and that evolution progresses from compact groups to pairs to ULIRGs to elliptical galaxies.

An example of a case study of an IR-luminous galaxy is in Fig. 9. Dinshaw et al. (1999) performed NICMOS F110W, F160W, and F222M imaging of NGC 6090, a luminous ($L = 3 \times 10^{11}$ $L_\odot$) starburst merger at a distance of 120 Mpc. The NICMOS images are centered on the two nuclei of the merger and reveal the spiral structure of the eastern galaxy and the amorphous nature of the western galaxy. Bright knots and clusters are

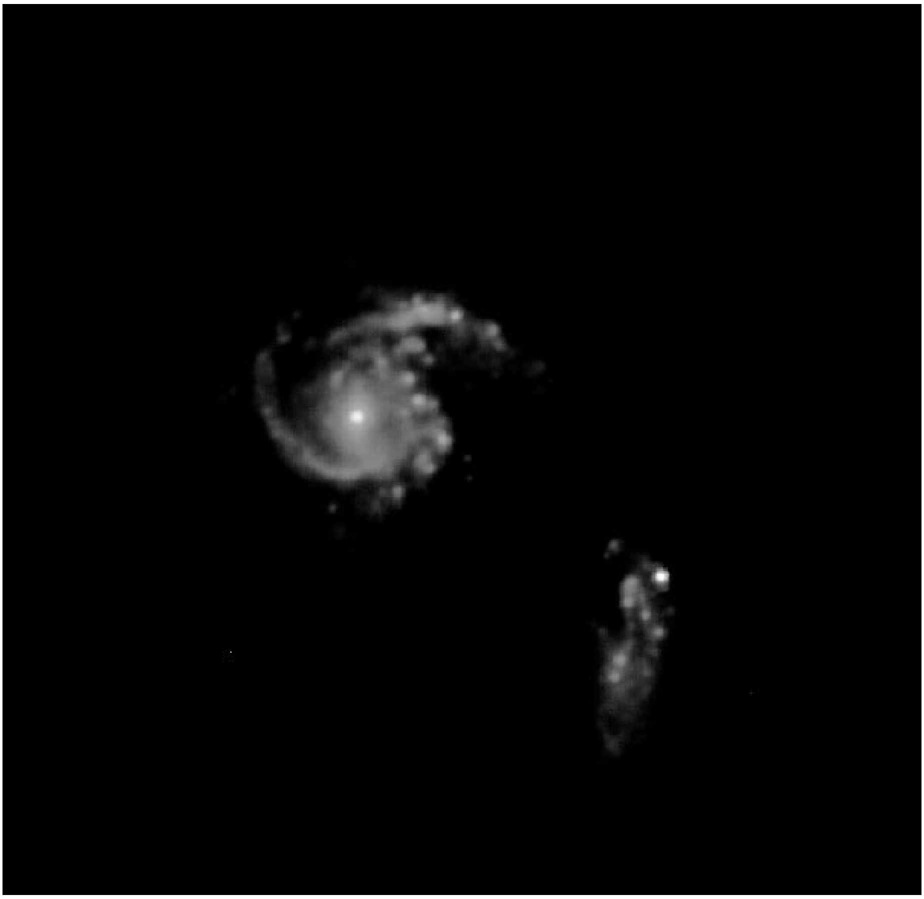


Figure 9. Multiwavelength (F110W, F160W, F222M) image of NGC 6090. Field size is 15″. North is up and east to the left (Dinshaw et al. 1999).

visible in the region overlapping the merging galaxies. The knots overlap with a region where molecular gas was detected by Bryant & Scoville (1999). Much of the present star formation is occurring outside the nuclear region of NGC 6090. This is similar to what was found in other luminous IR galaxies with multiple nuclei (NGC 6240: Bryant & Scoville 1999; VV 114: Frayer et al. 1999).

Even *HST* cannot provide us with the spatial resolution necessary to study the "microphysics" of the star formation process in a starburst galaxy. Again, 30 Doradus is a Rosetta Stone. Walborn et al. (1999) documented an extensive next generation of star formation in the periphery of the central R136 region. Very likely, this second generation was triggered by the R136 cluster itself. Many new IR sources, including multiple systems, clusters, and nebular structures, are found. Fig. 10 shows NICMOS H, K, and narrow-band H_2 (F212N) images of several compact nebulosities. A jet-like structure is prominent in the H_2 image. The R136 region hosts numerous examples of IR-bright knots, which turn out to be groups of massive, early-type stars embedded in nebulosity. The most spectacular and brightest knot resides at the top of a massive dust pillar oriented directly toward R136. Other knots have pc-scale jet structures associated with them. The structures consist of detached, non-stellar IR sources aligned on either side of the stellar system. They could be impact points of a highly collimated, bipolar jet on the

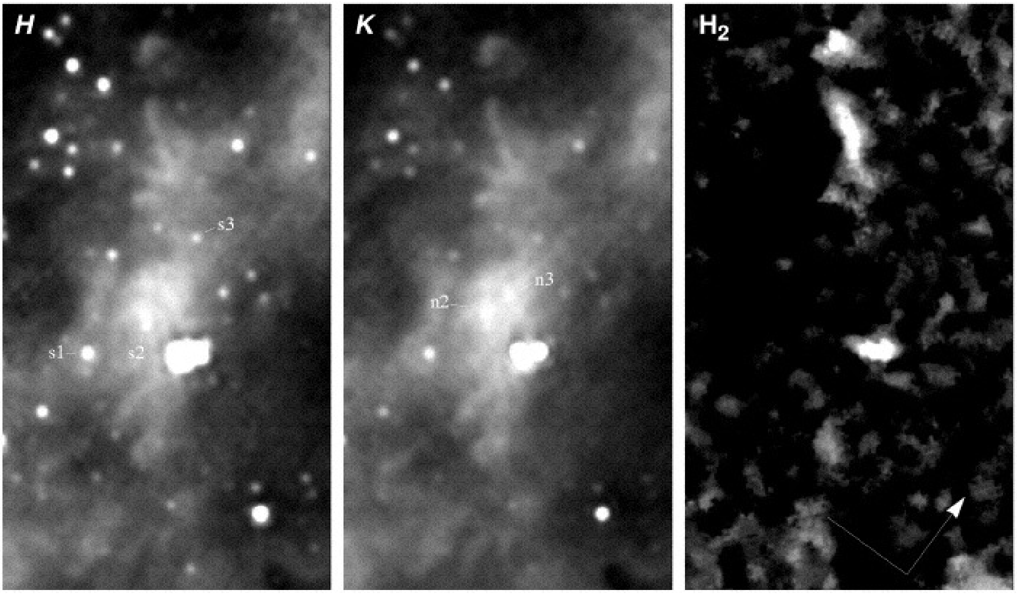


FIGURE 10. The 30 Doradus region imaged with NICMOS in the F160W (left), F205W (middle), and F212N (right) filters. The F212N image is continuum subtracted. The scale bars indicate N, E and are 0.5 pc in length (Walborn et al. 1999).

surrounding dark clouds. These outflows from young massive stars in 30 Dor are the first such detections outside our Galaxy. Their morphologies are strikingly similar to those seen in WFPC2 images of the Orion nebula (O'Dell & Zheng 1994) and M16 (Hester et al. 1996). These results establish the 30 Doradus nebula as a prime region in which to investigate the formation and very early evolution of massive stars and multiple systems. Star formation in 30 Doradus is not a continuous process; it occurs in multiple, instantaneous events, each possibly triggering (and terminating) the other. If 30 Doradus were viewed from larger distance, as starburst galaxies are, the decreased spatial resolution would mimic a star-formation region continuously forming stars over at least 10 Myr. Much to our frustration, 30 Dor serves as a reminder that the physical scales associated with the star-formation process in starburst galaxies may be well below *HST*'s resolution limit.

At some point the starburst terminates. This typically occurs within less than about 10^8 yr. A hard upper limit to the starburst duration can be derived from the timescale of the exhaustion of the gas reservoir to form stars. This argument can even be used to *define* a starburst in terms of the exhaustion timescale versus the Hubble time (Weedman 1987). For a specific example consider a luminous starburst galaxy with a star-formation rate of 100 $M_\odot$ yr^{-1} (see Heckman, Armus, & Miley 1990). After $\sim 10^8$ yr, more than 10^{10} $M_\odot$ of stars have been generated, which starts exceeding the total HI masses in typical $L_\star$ galaxies. The previous estimate strongly depends on the formation rate of low-mass stars, a quantity which is difficult to determine observationally. If such low-mass stars do not form (see Section 3), the time to exhaust the gas reservoir can be increased by a factor of several. Even so, it is easy to argue that a closed-box model for a starburst is too simplistic for an estimate of the available gas reservoir. Both infall and outflow should be taken into account.

An *HST* GHRS spectrum of the proto-typical Wolf-Rayet starburst galaxy He 2−10 is shown in Fig. 11. In this figure the observed spectrum of He 2−10 is compared to a

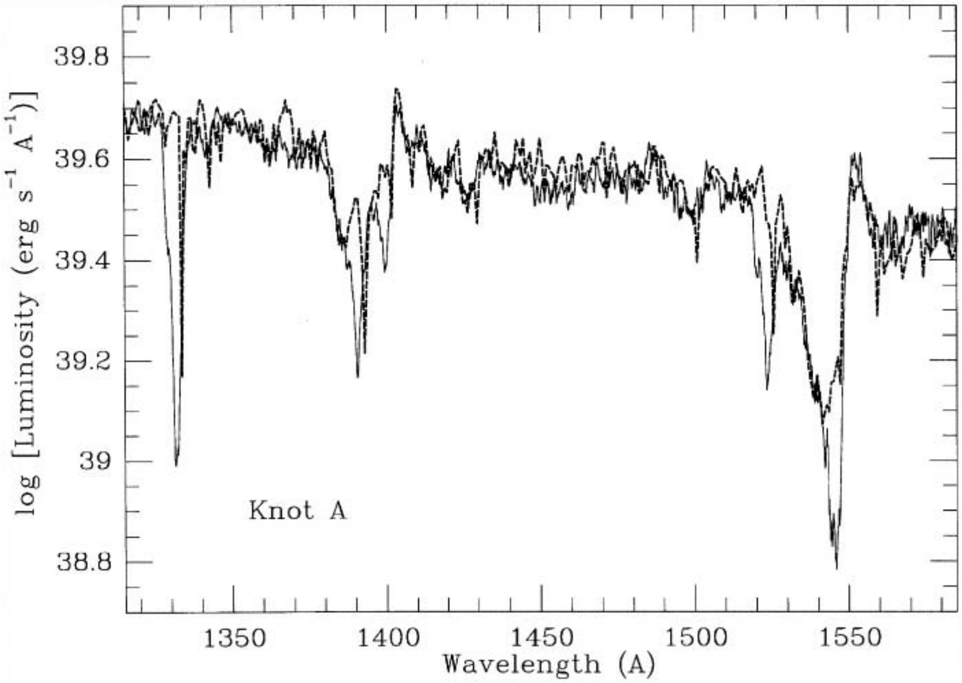


FIGURE 11. *HST* GHRS spectrum He 2−10. The dashed spectrum is a model fit to the *broad* stellar lines. The *narrow* interstellar lines are offset by ∼400 km s^{-1} (Johnson et al. 2000).

synthetic model for a starburst age of 4 Myr. Apparently all *stellar* photospheric and wind lines are reproduced extremely well by the model. Compare, for instance, the blue wings of the Si IV λ1400 and C IV λ1550 wind lines, the standard indicators for massive stars. On the other hand, C II λ1335, Si IIλ1526, and the deep, narrow absorption components of Si IV λ1400 and C IV λ1550 are much broader and more blue-shifted in the observations than in the model. The lines are purely interstellar, arising in the interstellar medium in and around the starburst. The sharp interstellar lines seen in the model spectra can be used to measure the outflow velocity in He 2−10, suggesting a bulk motion of at least −360 km s^{-1}. The energy source are almost certainly winds and supernovae. They are capable of initiating large-scale outflows of interstellar gas via so-called galactic superwinds (Chevalier & Clegg 1985). Johnson et al. (2000) estimate that the mass-loss rate of the interstellar medium is quite similar to the star-formation rate in He 2−10. Taken at face value, this suggests that the available gas reservoir will not only be depleted by the star-formation process but, more importantly, by removal of interstellar material. Starbursts may determine their own fate by their prodigious release of kinetic energy into the interstellar medium.

5. Effects of star formation on the environment

The impact of multiple stellar-wind and supernova events on the interstellar medium is seen via gaseous shells, bubbles, and outflows. The mechanical energy release of a hot star over its lifetime and of a supernova explosion both are of order 10^{51} erg. Initially the wind and supernova material is in a brief phase of free expansion until a sufficiently large mass of interstellar gas has been swept up. This phase is too short to be of observational significance. The fast stellar and supernova winds interact with the swept up material,

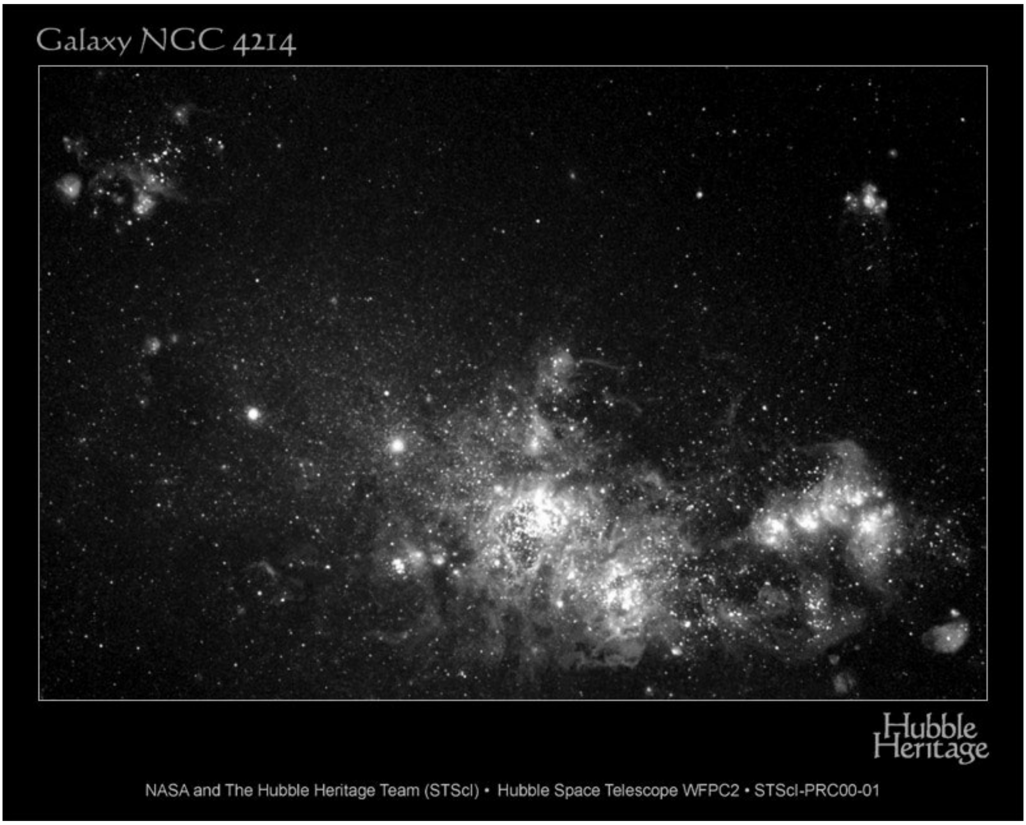

Galaxy NGC 4214
Hubble Heritage
NASA and The Hubble Heritage Team (STScI) • Hubble Space Telescope WFPC2 • STScI-PRC00-01

Figure 12. WFPC2 Hα + [O III] + optical continuum image of NGC 4214 (MacKenty et al. 2000).

producing a hot cavity, surrounded by a cool shell of interstellar material. Such shells are commonly observed around regions of high-mass star formation (e.g. Oey 1999). Depending on the mass of the central stellar cluster, their radii, expansion velocities, and ages are of order 100 pc, 25 km s^{-1}, and 10 Myr, respectively.

The nearby ($d = 4.1$ Mpc) irregular galaxy NGC 4214 host numerous spectacular wind-blown shells (Fig. 12). MacKenty et al. (2000) discuss Hα and [O III] narrow-band images of NGC 4214, obtained with the WFPC2 onboard *HST*. The *HST* images resolve features down to physical scales of 2–5 pc, revealing several young (< 10 Myr) star forming complexes of various ionized gas morphologies (compact knots, complete or fragmentary shells) and sizes (10–200 pc). The morphologies are suggestive of evolutionary trends: The youngest, smaller, filled regions that presumably are those just emerging from dense star forming clouds, tend to be of high excitation and are highly obscured. Evolved, larger shell-like regions have lower excitation and are less extincted due of the action of stellar winds and supernovae. Evidence for induced star formation is found, which has led to a two-stage starburst. This is similar to the sequential star formation seen in greater detail in 30 Doradus, hinting at what the morphologies of starburst galaxies at even larger distance would look like. NGC 4214 might well be a lower luminosity counterpart of some of the star-forming galaxies seen at cosmological redshift. Its spectral morphology in the ultraviolet (Leitherer et al. 1996) is strikingly similar to that of Lyman-break galaxies. Comparison of the first available spectra of bona fide star-forming galaxies at redshift

$\sim$ 3 with an *HST* FOS spectrum of the dominant cluster in NGC 4214 convincingly demonstrated their similar stellar content (Steidel et al. 1996).

The starburst in NGC 4214 has an age of about 10 Myr or less. The majority of the newly formed stars has not had time to evolve into supernovae, and the stellar energy release is still in its early evolutionary phase. The dynamical evolution of a starburst-driven outflow on a galactic scale has been extensively discussed (e.g. Suchkov et al. 1994, 1996; MacLow 1996). Initially, the deposition of mechanical energy by supernovae and stellar winds results in an over-pressured cavity of hot gas inside the starburst. This is the phase we currently observe in NGC 4214 and corresponds to the "classical" wind-blown bubble discussed above.

This hot cavity will continue to expand and sweep up more ambient material. If the ambient medium is stratified (like a disk), the bubble will expand most rapidly in the direction of the maximum pressure gradient, usually along the minor axis of the galaxy. After the bubble size reaches several disk vertical scale heights, the expansion will accelerate, and it is believed that Raleigh-Taylor instabilities will then lead to the fragmentation of the bubble's outer wall. This allows the hot gas to "blow out" of the disk and into the galactic halo in the form of a weakly collimated bipolar outflow.

Emission of Lyman-α radiation appears to be immediately related to galaxy outflows. The ionizing radiation from the newly formed young stars should lead to prominent Lyman-α emission due to recombination of hydrogen in the interstellar medium. Long ago, Partridge & Peebles (1967) suggested the Lyman-α line as an important spectral signature in young galaxies at high redshift since the expected Lyman-α luminosity amounts to a few percent of the total galaxy luminosity. Major observational efforts were undertaken to search for Lyman-α emission from faint galaxies at high redshift (Djorgovski & Thompson 1992). While *some* star-forming galaxies with Lyman-α emission were found (e.g. Keel et al. 1999; Kudritzki et al. 2000), their number is by far lower than expected from the cosmic star-formation history and line formation purely by recombination.

The assumption of Lyman-α being a pure recombination line in a gaseous medium may be too simple. Meier & Terlevich (1981), Hartmann et al. (1988), Neufeld (1990), and Charlot & Fall (1993) considered the effects of dust on Lyman-α. They found that dust scattering and absorption can be very efficient in removing Lyman-α photons from the line of sight to the observer, leading to much lower line strengths. Additionally, Lyman-α photons produced in galaxies suffer a large number of resonant scatterings in neutral atomic hydrogen. Depending on the aspect angle of the galaxy as seen from the observer, this may lead to a decrease of the Lyman-α equivalent width.

HST GHRS spectroscopy of eight gas-rich irregular galaxies by Kunth et al. (1998) indicates yet another, and most likely the dominant parameter governing Lyman-α emission: *outflows*. Kunth et al. found Lyman-α emission with blueshifted absorption in four galaxies (see Fig. 13). In these objects the O I and Si II absorption lines are also blueshifted, suggesting an outflow of the neutral gas with velocities of up to 200 km s^{-1}. The other four galaxies show broad damped Lyman-α absorption profiles centered on the wavelength of the ionized gas. The eight galaxies in Fig. 13 span a metallicity range of more than a factor of 10: IRAS 0833+6517 has solar abundance, and I Zw 18 is extremely metal-poor ($\sim 1/50$ Z$_\odot$). There is no correlation between metal abundance and Lyman-α emission strength. The velocity structure of the neutral gas in these galaxies is the driving factor that determines the detectability of Lyman-α in emission. Relatively small column densities of *static* neutral gas with even very small dust content would destroy the Lyman-α photons. The situation changes dramatically when most of the neutral gas is velocity-shifted relative to the ionized regions because resonant scattering by neutral hydrogen will be most efficient at wavelengths < 1216 Å, allowing the Lyman-α photons

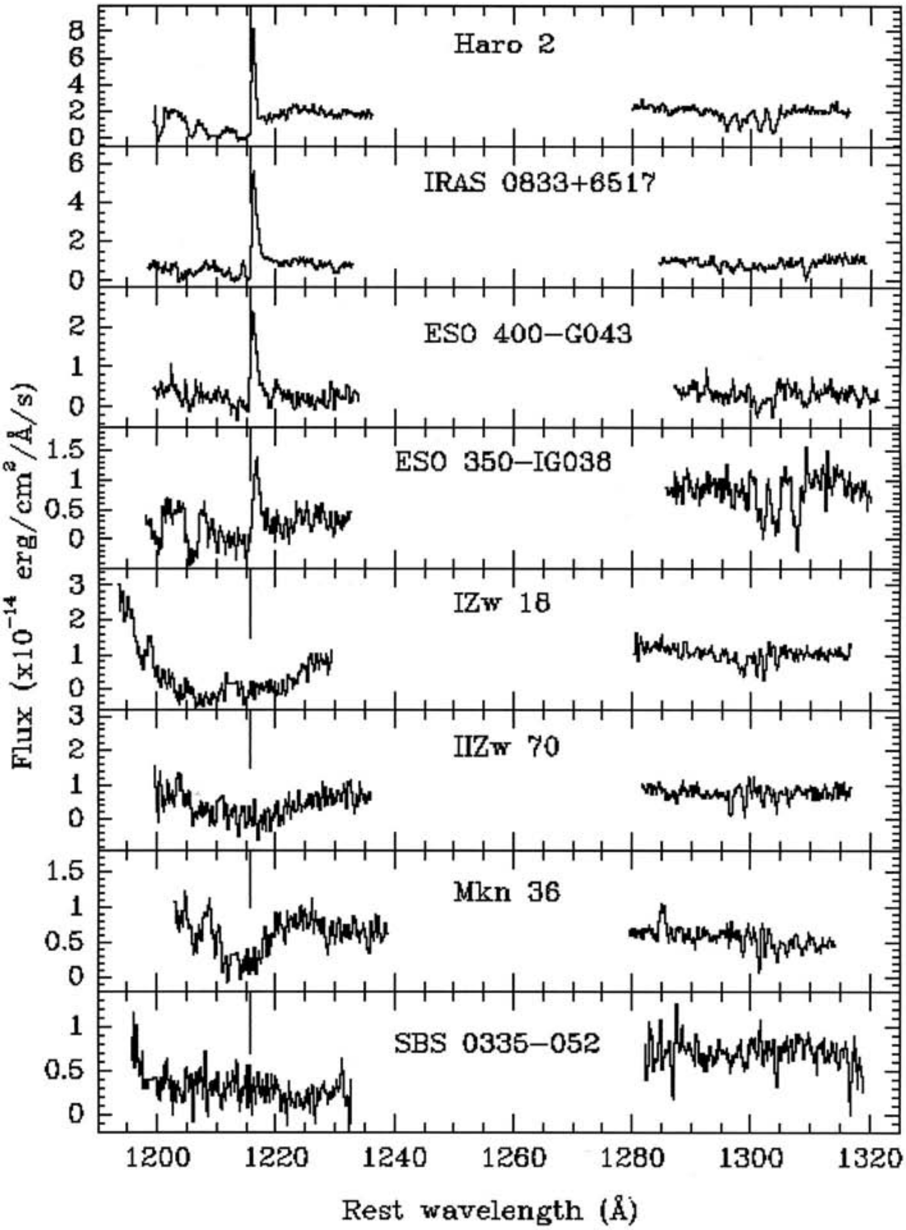


FIGURE 13. *HST* GHRS spectra of eight starburst galaxies around Lyman-α (Kunth et al. 1998).

to escape, a suggestion supported by recent models of Tenorio-Tagle et al. (1999). The implication is that feedback from the massive stars via ionization and the creation of superbubbles and galactic scale outflows leads to the large variety of Lyman-α profiles. The escape of Lyman-α photons depends critically on the column density of the neutral gas and dust, the morphology of the supershells, and the kinematics of the galactic wind. Since these effect can be highly stochastic, theoretical predictions for the Lyman-α

strength are quite uncertain. Therefore attempts to derive star-formation rates at high redshift from Lyman-α emission searches are quite challenging.

6. Science highlights: Past and future

HST observations of starburst galaxies have advanced the subject in several key areas. As expected from its instrumental capabilities, *HST* made the greatest impact with high S/N ultraviolet spectroscopy and with UV to near-IR imaging at high spatial resolution. Highlights are:

- The first complete stellar census of a metal-rich, young star cluster in the Galactic Center.
- The documentation of the fine-structure in nearby starbursts down to physical scales associated with star formation.
- The universality of the upper IMF in UV-selected starburst galaxies with a broad range of physical properties.
- The detailed study of the morphology of starburst galaxies and the importance of cluster formation as a star-formation mode.
- The detection of starbursts near active galactic nuclei, and the demonstration of their energetic significance.
- Detailed studies of the feedback between star formation and the interstellar medium and the associated cosmogonic and cosmological impact.

There are good reasons to predict even brighter prospects for the next decade of *HST* observations. The upcoming *HST* instruments ACS, WFPC3, and COS will have vastly higher photon collection efficiencies. This will allow target selection driven by astrophysical requirements rather than exposure duration constraints. With NICMOS being restored, highly spatially resolved imaging can be extended to the IR. The resulting panchromatic imaging will reveal the physics of dust-enshrouded starburst galaxies. Finally, the ever growing *HST* Archive and Large observing programs will establish unbiased surveys of starburst galaxy properties.

I wish to thank my current and previous local collaborators for all their contributions during the first decade of *HST*: Alessandra Aloisi, Daniela Calzetti, Duília de Mello, Daniel Devost, Jeff Goldader, Rosa González, Tim Heckman, Gerhardt Meurer, Carmelle Robert, Daniel Schaerer, and Christy Tremonti. Daniela Calzetti's and Brad Whitmore's comments on an earlier draft of this paper are gratefully acknowledged.

REFERENCES

Borne, K. D., Bushouse, H., Lucas, R. A., & Colina, L. 2000 *ApJ* **529**, L77.

Bryant, P. M. & Scoville, N. Z. 1999 *AJ* **117**, 2632.

Calzetti, D., Bohlin, R. C., Gordon, K. D., Witt, A. N., & Bianchi, L. 1995 *ApJ* **446**, L97.

Charlot, S. & Fall, S. M. 1993 *ApJ* **415**, 580.

Cheng, K.-P., Michalitsianos, A. G., Hintzen, P., Bohlin, R. C., O'Connell, R. W., Cornett, R. H., Roberts, M. S., Smith, A. M., Smith, E. P., & Stecher, T. P. 1992, *ApJ* **395**, 29.

Chevalier, R. A. & Clegg, A. W. 1985 *Nature* **317**, 44.

de Mello, D. F., Leitherer, C., & Heckman, T. M. 2000 *ApJ* **530**, 251.

Dinshaw, N., Evans, A. S., Epps, H., Scoville, N. Z., & Rieke, M. 1999 *ApJ* **525**, 702.

Djorgovski, S. & Thompson, D. J. 1992. In *IAU Symp. 149, Stellar Populations* (eds. B. Barbuy & A. Renzini), p. 337. Kluwer.

Figer, D. F., Kim, S. S., Morris, M., Serabyn, E., Rich, R. M., & McLean, I. S. 1999 *ApJ* **525**, 750.

Frayer, D. T., Ivison, R. J., Smail, I., Yun, M. S., Armus, L. 1999 *AJ* **118**, 139.

Friedli, D. & Benz, W. 1995 *A&A* **301**, 649.

Genzel, R. & Eckart, A. 1998. In *IAU Symp. 184, The Central Regions of the Galaxy and Galaxies* (ed. Y. Sofue), p. 421. Kluwer.

Goldader, J. D., Joseph, R. D., Doyon, R., & Sanders, D. B. 1997 *ApJS* **108**, 449.

Goldader, J. D. & Wynn-Williams, C. G. 1994 *ApJ* **433**, 164.

González Delgado, R. M., Heckman, T. M., Leitherer, C., Meurer, G. R., Krolik, J., Wilson, A. S., Kinney, A. L., & Koratkar, A. 1998 *ApJ* **505**, 174.

Hartmann, L., Huchra, J. P., Geller, M. J., O'Brien, P., & Wilson, R. 1988 *ApJ* **326**, 101.

Heckman, T. M. 1997. In *Star Formation Near and Far* (eds. S. S. Holt & L. G. Mundy), p. 271. AIP.

Heckman, T. M., Armus, L., & Miley, G. K. 1990 *ApJS* **74**, 833.

Hester, J. J., Scowen, P. A., Sankrit, R., Lauer, T. R., Ajhar, E. A., Baum, W. A., Code, A., Currie, D. G., Danielson, G. E., Ewald, S. P., Faber, S. M., Grillmair, C. J., Groth, E. J., Holtzman, J. A., Hunter, D. A., Kristian, J., Light, R. M., Lynds, C. R., Monet, D. G., O'Neil, E. J., Jr., Shaya, E. J., Seidelmann, K. P., & Westphal, J. A. 1996 *AJ* **111**, 2349.

Hibbard, J. E. 1997. In *Star Formation Near and Far* (eds. S. S. Holt & L. G. Mundy), p. 259. AIP.

Ho, L. C. & Filippenko, A. V. 1996 *ApJ* **466**, L83.

Hunter, D. A., Shaya, E. J., Holtzman, J. A., Light, R. M., O'Neil Jr., E. J., & Lynds, R. 1995 *ApJ* **448**, 179.

Iglesias-Páramo, J. & Vílchez, J. M. 1997 *ApJ* **479**, 190.

Johnson, K. E. & Conti, P. S. 2000 *AJ* **119**, 2146.

Johnson, K. E., Leitherer, C., Vacca, W. D., & Conti, P. S. 2000 *AJ*, **120**, 1273.

Johnson, K. E., Vacca, W. D., Leitherer, C., Conti, P. S., & Lipscy, S. J. 1999 *AJ* **117**, 1708.

Joseph, R. D. 1999. In *IAU Symp. 193, Wolf-Rayet Phenomena in Massive Stars and Starburst Galaxies* (eds. K. A. van der Hucht, G. Koenigsberger, & P. R. J. Eenens), p. 568. ASP.

Keel, W. C., Cohen, S. H., Windhorst, R. A., & Waddington, I. 1999 *AJ* **118**, 2547.

Kennicutt, Jr., R. C., Roettiger, K. A., Keel, W. C., van der Hulst, J. M., & Hummel, E. 1987 *AJ* **93**, 1011.

Kim, S. S., Morris, M., & Lee, H. M. 1999 *ApJ* **525**, 228.

Kroupa, P., Tout, C. A., & Gilmore, G. 1993 *MNRAS* **262**, 545.

Kudritzki, R.-P., Méndez, R. H., Feldmeier, J. J., Ciardullo, R., Jacoby, G. H., Freeman, K. C., Arnaboldi, M., Capaccioli, M., Gerhard, O., & Ford, H. C. 2000 *ApJ* **536**, 19.

Kunth, D., Mas-Hesse, J. M., Terlevich, E., Terlevich, R., Lequeux, J., & Fall, S. M. 1998 *A&A* **334**, 11.

Leitherer, C. 1998. In *The Stellar Initial Mass Function*, eds. G. Gilmore & D. Howell, p.61. ASP.

Leitherer, C., Robert, C., & Heckman, T. 1995 *ApJS* **99**, 173.

Leitherer, C., Vacca, W. D., Conti, P. S., Filippenko, A. V., Robert, C., & Sargent, W. L. W. 1996 *ApJ* **465**, 717.

Leitherer, C., Walborn, N. R., Heckman, T. M., & Norman, C. A. 1991 *Massive Stars in Starbursts*. CUP.

MacKenty, J. W., Maíz-Apellániz, J., Pickens, C. E., Norman, C. A., & Walborn, N. R. 2000 *AJ*, **120**, 3007.

Malumuth, E. M. & Heap, S. R. 1994 *AJ* **107**, 1054.

Maoz, D., Koratkar, A., Shields, J. C., Ho, L. C., Filippenko, A. V., & Sternberg, A. 1998 *AJ* **116**, 55.

Massey, P. 1998. In *The Stellar Initial Mass Function* (eds. G. Gilmore & D. Howell), p. 17. ASP.

Massey, P. & Hunter, D. A. 1998 *ApJ* **493**, 180.

Massey, P., Lang, C. C., DeGioia-Eastwood, K., & Garmany, C. D. 1995 *ApJ* **438**, 188.

McKee, C. F. 1996. In *The Interplay Between Massive Star Formation, the ISM and Galaxy Evolution* (eds. D. Kunth, B. Guiderdoni, M. Heydari-Malayeri, & T. X. Thuan), p. 223. Editions Frontieres.

Meier, D. & Terlevich, R. 1981 *ApJ* **246**, L109.

Meurer, G. R., Heckman, T. M., Leitherer, C., Kinney, A., Robert, C., & Garnett, D. R. 1995 *AJ* **110**, 2665.

Mihos, J. & Hernquist, L. 1996 *ApJ* **464**, 641.

Najarro, F., Hillier, D. J., Figer, D. F., & Geballe, T. R. 1999. In *The Central Parsecs of the Galaxy* (eds. H. Falcke, A. Cotera, W. J. Duschl, F. Melia, & M. J. Rieke), p. 340. ASP.

Neufeld, D. A. 1990 *ApJ* **350**, 216.

Norman, C. A., Sellwood, J. A., & Hasan, H. 1996 *ApJ* **462**, 114.

O'Dell, C. R. & Zheng, W. 1994 *ApJ* **436**, 194.

Oey, M. S. 1999. In *IAU Symp. 193, Wolf-Rayet Phenomena in Massive Stars and Starburst Galaxies* (eds. K. A. van der Hucht, G. Koenigsberger, & P. R. J. Eenens), p. 627. ASP.

Partridge, R. & Peebles, P. J. E. 1967 *ApJ* **147**, 868.

Ramírez, S. V., Sellgren, K., Carr, J. S., Balachandran, S. C., Blum, R., Terndrup, D. M., & Steed, A. 2000 *ApJ* **537**, 205.

Rigopoulou, D., Spoon, H. W. W., Genzel, R., Lutz, D., Moorwood, A. F. M., & Tran, Q. D. 1999 *AJ* **118**, 2625.

Rieke, G. H. 1991. In *Massive Stars in Starbursts* (eds. C. Leitherer, N. R. Walborn, T. M. Heckman, & C. A. Norman), CUP.

Rieke, G. H., Lebofsky, M. J., Thompson, R. I., Low, F. J., & Tokunaga, A. T. 1980 *ApJ* **238**, 24.

Robert, C., Leitherer, C., & Heckman, T. M. 1993 *ApJ* **418**, 749.

Salpeter, E. E. 1955 *ApJ* **121**, 161.

Sanders, D. B. 1997. In *Starburst Activity in Galaxies* (eds. J. Franco, R. Terlevich, & A. Serano), *Rev. Mex. Astron. Astrofis. Conf. Ser.* **6**, 42.

Sanders, D. B. & Mirabel, I. F. 1996 *ARA&A* **34**, 749.

Sanders, D. B., Soifer, B. T., Elias, J. H., Madore, B. F., Matthews, K., Neugebauer, G., & Scoville, N. Z. 1988 *ApJ* **325**, 74.

Scoville, N. Z., Evans, A. S., Dinshaw, N., Thompson, R., Rieke, M., Schneider, G., Low, F. J., Hines, D., Stobie, B., Becklin, E., & Epps, H. 1998 *ApJ* **492**, L107.

Sirianni, M., Nota, A., Leitherer, C., De Marchi, G., & Clampin, M. 2000 *ApJ* **533**, 203.

Smart, S. J. & Rolleston, W. R. J. 1997, *ApJ* **481**, L47.

Smith, H. E., Lonsdale, C. J., Lonsdale, C. J., & Diamond, P. J. 1998 *ApJ* **493**, L17.

Steidel, C. C., Giavalisco, M., Pettini, M., Dickinson, M., & Adelberger, K. L. 1996 *ApJ* **462**, L17.

Suchkov, A. A., Balsara, D. S., Heckman, T. M., & Leitherer, C. 1994 *ApJ* **430**, 511.

Suchkov, A. A., Berman, V. G., Heckman, T. M., & Balsara, D. S. 1996 *ApJ* **463**, 528.

Tenorio-Tagle, G., Silich, S. A., Kunth, D., Terlevich, E., & Terlevich, R. 1999 *MNRAS* **309**, 332.

Terlevich, R. 1992. In *Relationships Between Active Galactic Nuclei and Starburst Galaxies* (ed. A. V. Filippenko), p. 133. ASP.

—— 1997. In *Starburst Activity in Galaxies* (eds. J. Franco, R. Terlevich, & A. Serano), *Rev. Mex. Astron. Astrofis. Conf. Ser.* **6**, 1.

Tremonti, C. A., Calzetti, D., Heckman, T. M., & Leitherer, C. 2000 *ApJ* submitted.

Walborn, N. R. 1991. In *Massive Stars in Starbursts* (eds. C. Leitherer, N. R. Walborn, T. M. Heckman, & C. A. Norman), p. 145. CUP.

Walborn, N. R., Barbá, R. H., Brandner, W., Rubio, M., Grebel, E. K., & Probst, R. G. 1999 *AJ* **117**, 225.

Weedman, D. W. 1987. In *Star Formation in Galaxies* (ed. C. J. Lonsdale), p. 351. NASA.

Zhang, Q. & Fall, S. M. 1999 *ApJ* **527**, L81.

Supermassive black holes

By F. D. MACCHETTO†

Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218

One of the important topics of current astrophysical research is the role that supermassive black holes play in shaping the morphology of their host galaxies. There is increasing evidence for the presence of massive black holes at the centers of all galaxies and many efforts are directed at understanding the processes that lead to their formation, the duty cycle for the active phase and the question of the fueling mechanism. Related issues are the epoch of formation of the supermassive black holes, their time evolution and growth and the role they play in the early ionization of the Universe. Considerable observational and theoretical work has been carried out in this field over the last few years and I will review some of the recent key areas of progress.

1. Introduction

It is now widely accepted that quasars (QSOs) and Active Galactic Nuclei (AGN) are powered by accretion onto massive black holes. This has led to extensive theoretical and observational studies to elucidate the properties of the black holes, the characteristics of the accretion mechanisms and the mechanisms responsible for the production and transportation of the energy from the central regions to the extended radio lobes.

However, over the last few years there has been an increasing realization that Massive Dark Objects (MDOs) may actually reside at the centers of *all* galaxies (Ho 1998, Magorrian et al. 1998, Richstone et al. 1998, Gebhardt et al. 2000a, Gebhardt et al. 2000b, Merrit & Ferrarese 2001, van der Marel 1999). The mean mass of these objects, of order $10^{-2.5}$ times the mass of their host galaxies, is consistent with the mass in black holes needed to produce the observed energy density in quasar light if we make reasonable assumptions about the efficiency of quasar energy production (Chokshi & Turner 1992, Blandford 1999). This raises a number of important new questions and has fundamental implications for the role of the black holes in contributing or being responsible for the ionization (or reionization) of the early universe and for their role in the processes leading to the formation of galaxies. Conversely the apparent correlation between the black hole mass and the mass of the spheroidal component in elliptical and spirals points towards a close interaction between the galaxy size and morphology and its central black hole. Models in which elliptical galaxies form from the mergers of disk galaxies whose bulges contain black holes are consistent with the “core fundamental plane, the relation between the central parameters of early-type galaxies. Furthermore it is clear that the dynamical influence of a supermassive black hole can extend far beyond the nucleus if a substantial number of stars are on orbits that carry them into the center. Work by Merritt (1998) has shown that nuclear black holes are important for understanding many of the large-scale properties of galaxies, including the fact that elliptical galaxies come in two, morphologically-distinct families, the absence of bars in most disk galaxies, and the shapes of the spiral galaxy. Since the growth of the black hole mass depends on the global morphology of the host galaxies, the link between black hole and galaxy structure implies a feedback mechanism that determines what fraction of a galaxys mass ends up in the center.

† On assignment from the Astrophysics Division, Space Science Department of the European Space Agency

2. Dynamical evidence for massive black holes

2.1. *Megamasers in NGC 4258*

The best observation showing the presence of a Keplerian disk around a black hole was the VLBI observation of megamasers in the nucleus of the Seyfert 2 galaxy NGC 4258 reported by Miyoshi et al. (1995). These observations reveal individual masing knots revolving at distances ranging from ~ 13 pc to 25 pc around the central object. These data show a near-perfect Keplerian velocity distribution, implying that almost all the mass is located well within the inner radius where the megamasers reside,and they derive a central mass of $\sim 3.6 \times 10^7\,\mathrm{M}_\odot$ within the inner ~ 13 pc.

Given this mass and the fact that NGC 4258 is a relatively low luminosity ($\sim 10^{42}$ erg s^{-1}) object, the emission is sub-Eddington, with $L/L_\mathrm{E} \sim 3\times10^{-4}$. Such sub-Eddington sources are likely to have accretion disk structures, where the accreting gas is optically thin and radiates inefficiently, and the accretion energy that is dissipated viscously, is advected with the accretion flow (see, e.g. Ichimaru 1987, Narayan & Yi 1994, Abramowicz et al. 1995).

2.2. *Kinematic studies using optical emission lines*

The other line of evidence for the presence of black holes in galaxies is the velocity field of the matter emitting closely to the nucleus. Very high spatial resolution observations using the long-slit spectrograph on the *Faint Object Camera* of *HST* were carried out by Macchetto et al. (1997). We observed the ionized gas disk in the emission line of [OII]λ3727 at three different positions separated by 0.2 arcsec, with a spatial sampling of 0.03 arcsec (or ~ 2 pc at the distance of M87), and measured the rotation curve of the inner $\sim 1''$ of the ionized gas disk to a distance as close as $0\overset{''}{.}07$ ($\simeq 5$ pc) to the dynamical center. We modeled the kinematics of the gas under the assumption of the existence of both a central black hole and an extended central mass distribution, taking into account the effects of the instrumental PSF, the intrinsic luminosity distribution of the line, and the finite size of the slit. We found that the central mass must be concentrated within a sphere whose maximum radius is $\simeq 3.5$ pc and showed that both the observed rotation curve and line profiles are best explained by a thin-disk in Keplerian motion. Finally, we proved that the observed motions are due to the presence of a super-massive black-hole and derived a value of $\mathrm{M_{BH}} = (3.2 \pm 0.9) \times 10^9\,\mathrm{M}_\odot$ for its mass.

2.3. *Virial masses*

The virial masses and emission-line region sizes of Active Galactic Nuclei (AGNs) can be measured by “reverberation-mapping” techniques. Wandel, Peterson & Malkan (1999) have compiled a sample of 17 Seyfert 1 and 2 quasars with reliable reverberation and spectroscopic data and used these results to calibrate similar determinations made by photoionization models of the AGN line-emitting regions. Reverberation mapping uses the light travel-time delayed emission-line response to continuum variations to determine the size and kinematics of the emission-line region. The distance of the broad emission-line region (BLR) from the ionizing source is then combined with the velocity dispersion, derived from either the broad Hβ line width or from the variable part of the line profile to estimate the virial mass. When they compare the central masses calculated with the reverberation method to those calculated using a photoionization (Hβ line) model, they find a nearly linear correlation (Table 1). They find that the correlation between the masses is significantly better than the correlation between the corresponding BLR sizes calculated by the two methods, which further supports the conclusion that both methods measure the mass of the central black hole. They also derive the Eddington

Name	$\log R_{\rm ph}$	log lag	$\log M_{\rm ph}$	$\log M_{\rm rev}$	$M_{\rm rev} (l0^7\ {\rm M}_\odot)$	$\log L_{\rm ion}$	$\log \left(\frac{L_{\rm ion}}{L_{\rm Edd}}\right)$
3C 120	0.92	1.64	6.86	7.49	$3.1^{+2.0}_{-1.5}$	45.03	–0.57
3C 390.3	0.89	1.38	8.26	8.59	39.1^{+12}_{-15}	44.51	–2.19
Akn 120	1.07	1.59	7.97	8.29	$19.3^{+4.1}_{-4.6}$	44.92	–1.48
F9	1.13	1.23	8.03	7.94	$8.7^{+2.6}_{-4.5}$	44.21	–1.84
1C 4329A	<0.56	0.15	7.34	<6.86	<0.73	<42.04	<–2.93
Mrk 79	0.81	1.26	7.48	8.02	$10.5^{+4.0}_{-5.7}$	44.26	–1.87
Mrk 110	0.78	1.29	6.46	6.91	$0.80^{+0.29}_{-0.30}$	44.33	–0.69
Mrk 335	0.89	1.23	6.68	6.58	$0.39^{+0.14}_{-0.11}$	44.20	–0.49
Mrk 509	1.08	1.90	7.17	7.98	$9.5^{+1.1}_{-1.1}$	45.54	–0.54
Mrk 590	0.85	1.31	7.00	7.15	$1.4^{+0.3}_{-0.3}$	44.37	–0.89
Mrk 817	0.86	1.19	7.53	7.56	$3.7^{+1.1}_{-0.9}$	44.13	–1.54
NGC 3227	0.16	1.04	6.92	7.69	$4.9^{+2.7}_{-5.0}$	43.82	–1.98
NGC 3783	0.52	0.65	7.05	7.04	$1.1^{+1.1}_{-1.0}$	43.05	–2.10
NGC 4051	<0.14	0.81	5.37	6.15	$0.14^{+0.15}_{-0.09}$	41.57	–0.84
NGC 4151	0.44	0.48	7.35	7.08	$1.2^{+0.8}_{-0.7}$	42.70	–2.49
NGC 5548	0.73	1.26	7.70	7.83	$6.8^{+1.5}_{-1.0}$	44.27	–1.83
NGC 7469	0.90	0.70	6.87	6.88	$0.76^{+0.75}_{-0.76}$	43.14	–1.86
PG 0804+762	1.39	2.00	7.74	8.34	$21.9^{+3.8}_{-4.5}$	45.75	–0.70
PG 0953+414	1.54	2.03	7.83	8.19	$15.5^{+10.8}_{-9.1}$	45.81	–0.49

TABLE 1. Reverberation BLR sizes and central masses compared with photoionization sizes and masses. The last two columns give the ionizing luminosity derived from the lag and the corresponding Eddington ratio (Wandel et al. 1999).

ratio, which for the objects in the sample fall in the range $L_V/L_{\rm Edd} \sim 0.001$–$0.03$ and $L_{\rm ion}/L_{\rm Edd} \approx 0.01$–$0.3$.

2.4. *The black hole mass of a Seyfert galaxy*

In a recent study Winge et al. (1999) have analyzed both ground-based, and *HST/FOC* long-slit spectroscopy at subarcsecond spatial resolution of the narrow-line region (NLR) of NGC 4151. They found that the extended emission gas ($R > 4''$) is in a normal rotation in the galactic plane, a behavior that they were able to trace even across the nuclear region, where the gas is strongly disturbed by the interaction with the radio jet and connects smoothly with the large-scale rotation defined by the neutral gas emission. The *HST* data, at $0\farcs03$ spatial resolution, allow for the first time truly to isolate the kinematic behavior of the individual clouds in the inner narrow-line region. They find that, underlying the perturbations introduced by the radio ejecta, the general velocity field can still be well represented by planar rotation down to a radius of $\sim 0\farcs5$ (30 pc), the distance at which the rotation curve has its turnover. The most striking result that emerges from the analysis is that the galaxy potential derived fitting the rotation curve changes from a "dark halo" at the extended narrow-line region distances to being dominated by the central mass concentration in the NLR, with an almost Keplerian falloff in the $1'' < R < 4''$ interval. The observed velocity of the gas at $0\farcs5$ implies a mass of $M \sim 10^9\,{\rm M}_\odot$ within the inner 60 pc. The presence of a turnover in the rotation curve

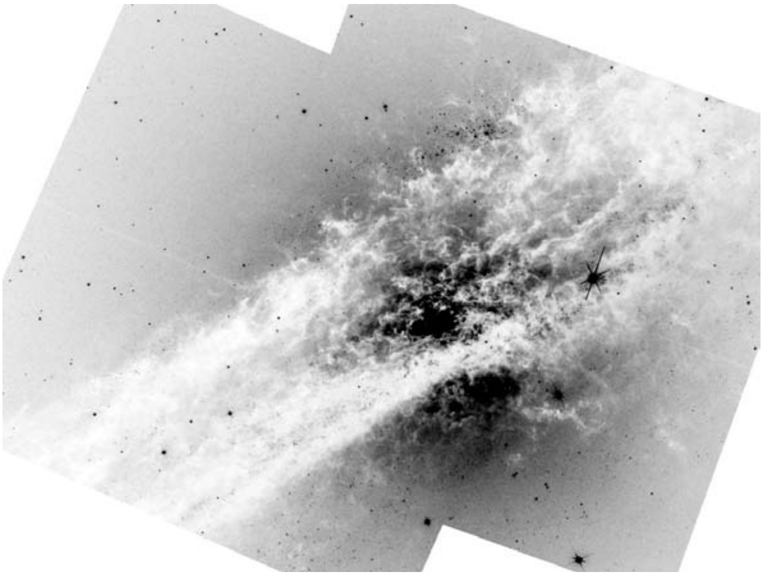


FIGURE 1. Grayscale representation of the Cen A mosaic in the *WFPC2* F814W filter. Surface brightness ranges from 0 (white) to 1.6 in units of 10^{-16} erg s^{-1} cm^{-2} Å^{-1} arcsec^{-2}. Image sizes are $225'' \times 170''$. North is up and East is left (Marconi et al. 2000).

indicates that this central mass concentration is extended. The first measured velocity point (outside the region saturated by the nucleus) would imply an enclosed mass of $\sim 5 \times 10^7\,\mathrm{M}_\odot$ within $R \sim 0\farcs15$ (10 pc), which represents an upper limit to any nuclear point mass.

3. Extended nuclear disks

Observations of a number of extended (a few 100 pc) nuclear disks with the *HST* has provided new evidence and constraints on the mass of the MDOs in early type galaxies.

Ferrarese & Ford (1999) carried out *HST* imaging and spectroscopy of NGC 6251, a giant E2 galaxy and powerful radio source which is at a distance of $\sim$ 106 Mpc. The *WFPC2* images show a well defined dust disk, 730 pc in diameter, whose normal is inclined by 76° to the line of sight. The *FOS* $0\farcs09$ square aperture was used to map the velocity of the gas in the central $0\farcs2$, from the kinematics of the gas they derive a value for the central mass concentration, 4×10^8 to $8 \times 10^8\,\mathrm{M}_\odot$.

Other galaxies studied with *HST* at high spatial resolution include NGC 4261, NGC 4374, NGC 7052 (Ferrarese et al. 1996, Bower et al. 1998, van der Marel & Van den Bosch 1998) and show black hole masses in the range 2–6 $\times 10^8\,\mathrm{M}_\odot$.

3.1. *Cen A*

Centaurus A (NGC 5128) is the closest (3.5 Mpc) giant elliptical galaxy hosting an active galactic nucleus (AGN) and a jet (Fig. 1). The prominent dust lane, which obscures the inner half kiloparsec of the galaxy, with associated gas, young stars and HII regions, is interpreted as the result of a relatively recent merger event between a giant elliptical galaxy and a small, gas rich, disk galaxy (Baade & Minkowski 1954, Graham 1979, Malin et al. 1983).

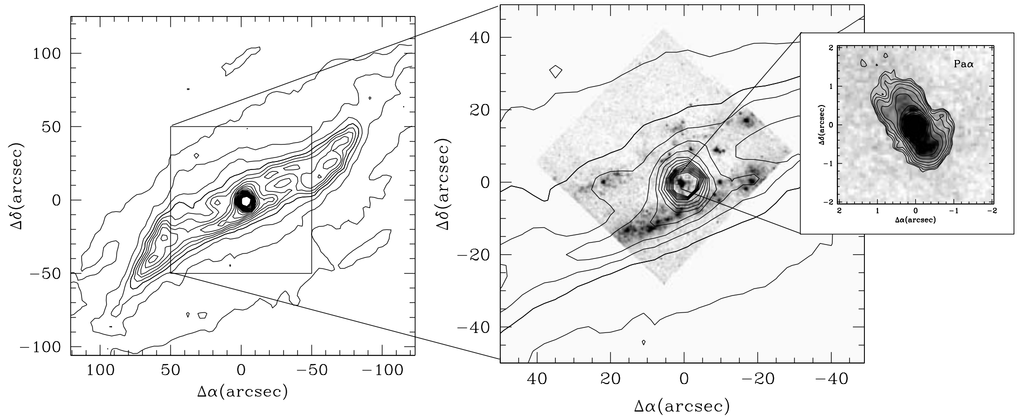


FIGURE 2. (**Left panel**): contours from the ISOCAM image at 7 μm. (**Center panel**): overlay of the ISOCAM contours on the NIC3 Paα image showing the morphological association between the Paα emission and the edges of the putative bar. (**Right panel**): the Paα disk from Paper II. Note that its major axis is perpendicular to the edges of the "bar" (Marconi et al. 2000).

IR and CO observations of the dust lane have been modeled by a thin warped disk (Quillen et al. 1992, Quillen et al. 1993) which dominates ground-based near-IR observations along with the extended galaxy emission (Packham et al. 1996). Earlier R-band imaging polarimetry from *HST* with *WFPC* (Schreier et al. 1996) are also consistent with dichroic polarization from such a disk.

Recent *HST WFC 2* and *NICMOS* observations of Centaurus A have shown that the 20 pc-scale nuclear disk previously detected by NICMOS in Paα (Schreier et al. 1998) has also been detected in the [FeII]λ1.64 μm line which shows a morphology similar to that observed in Paα with an [FeII]/Paα ration typical of low ionization Seyfert galaxies and LINERSs (Fig. 2). Marconi et al. (1999) derive a map of dust extinction, E(B–V), in a $20'' \times 20''$ circumnuclear region and reveal a several arcsecond long dust feature near to but just below the nucleus, oriented in a direction transverse to the large dust lane. This structure may be related to the bar observed with ISO and SCUBA, as reported by Mirabel et al. (1999). They find rows of Paα emission knots along the top and bottom edges of the bar, with they interpret as star formation regions, possibly caused by shocks driven into the gas. The inferred star formation rates are moderately high ($\sim 0.3\,\mathrm{M}_\odot\,\mathrm{yr}^{-1}$). If the bar represents a mechanism for transferring gas in to the center of the galaxy, then the large dust lane across the galaxy, the bar, the knots, and the inner Paα disk all represent aspects of the feeding of the AGN. Gas and dust are supplied by a recent galaxy merger; a several arcminute-scale bar allows the dissipation of angular momentum and infall of gas toward the center of the galaxy; subsequent shocks trigger star formation; and the gas eventually accretes onto the AGN via the 20 pc disk.

By reconstructing the radial light profile of the galaxy to within $0''\!.1$ of the nucleus Marconi et al. (1999) show that Centaurus A has a core profile. Using the models of van der Marel (1999), they estimate a black hole mass of $\sim 10^9\,\mathrm{M}_\odot$, consistent with ground based kinematical measurements (Israel 1998).

4. Statistical properties of AGN and radio galaxies

An important question about AGN hosts is whether there is anything unusual about their morphology, whether they occur only in a certain type of galaxy, or can be found in all galaxies but their active phase lasts for only a fraction of a Hubble time. Furthermore, the morphology of the host may provide important information about the dynamics that funnel accretion fuel into the nucleus. Studies of the environments of quasars can also provide insight into the AGN phenomenon in general, such as the relationship between quasars and radio galaxies. If indeed quasars and radio galaxies are objects differentiated only by viewing angle, then quasars might also be expected to exhibit an alignment effect over the same redshift range as the radio galaxies.

The original classification of radio galaxies by Fanaroff & Riley (1974) is based on a morphological criterion, i.e. edge darkened (FR I) vs. edge brightened (FR II) radio structure. It was later discovered that this dicothomy corresponds to a continuous transition in total radio luminosity (at 178 MHz) which formally occurs at $L_{178} = 2 \times 10^{26}\,\mathrm{W\,Hz^{-1}}$.

4.1. *Host morphology of radio galaxies and quasars*

HST observations of 273 sources in the 3CR catalog were carried out by Martel et al. (1997, 1999). To study the morphology distribution of the radio galaxies in the sample, they selected those at relatively small redshift. This ensures adequate image quality to permit reliable determination of the morphology, and minimizes the effects due to cosmological evolution of either the population of radio galaxies or the nature of their hosts. The result of this study is that more than 80% of the radio sources are found in elliptical galaxies, and the remainder have hosts whose morphologies are difficult to determine.

The 3CR sample is particularly well suited for investigating the relationship between radio galaxies and quasars, and the results have been discussed by Martel et al. (1997, 1999) and Lehnert et al. (1999a) (Figs. 3 and 4).

The study shows that the quasar "fuzz" contributes from <5% to as much as 100% of the total light from the quasar, with a typical value of about 20%. Most of the sources are resolved and show complex morphology with twisted, asymmetric, and/or distorted isophotes and irregular extensions. In almost every case of the quasars with spatially resolved "fuzz," there are similarities between the radio and optical morphologies. A significant fraction ($\sim 25\%$) of the sources show nearby galaxies in projection and $\sim 10\%$ of the sources show obvious signs of interactions with these nearby companions. These results show that the generally complex morphologies of host galaxies of quasars are influenced by the radio emitting plasma and by the presence of nearby companions.

Bahcall et al. (1997) have studied in detail nine radio-loud quasars and found that the hosts are either bright ellipticals or occur in interacting systems (Fig. 5). There is a strong correlation between the radio emission of the quasar and the luminosity of the host galaxy; the radio-loud quasars reside in galaxies that are on average about 1 mag brighter than hosts of the radio-quiet quasars.

Further *HST* observations of radio-loud quasars by Lehnert et al. (1999a), analyzed the spatially-resolved structures around five high-redshift radio-loud quasars.

Comparing the images with high resolution *VLA* radio images they conclude that all of the high redshift quasars are extended in both the rest-frame UV continuum and in Lyα.

The typical integrated magnitude of the host is $V \sim 22 \pm 0.5$, the typical UV luminosity is $\sim 10^{10}\,\mathrm{L_\odot}$, and the Ly$\alpha$ images are also spatially-resolved. The typical luminosity of the extended Lyα is about few $\times 10^{44}\,\mathrm{ergs\,s^{-1}}$; these luminosities require roughly a few percent of the total ionizing radiation of the quasar.

These results show that the generally complex morphologies of host galaxies are influenced by the radio emitting plasma. This manifests itself in the "alignment" between the radio, Lyα, and UV continuum emission, in detailed morphological correspondence in some of the sources which suggests "jet-cloud" interactions, and in the fact that the

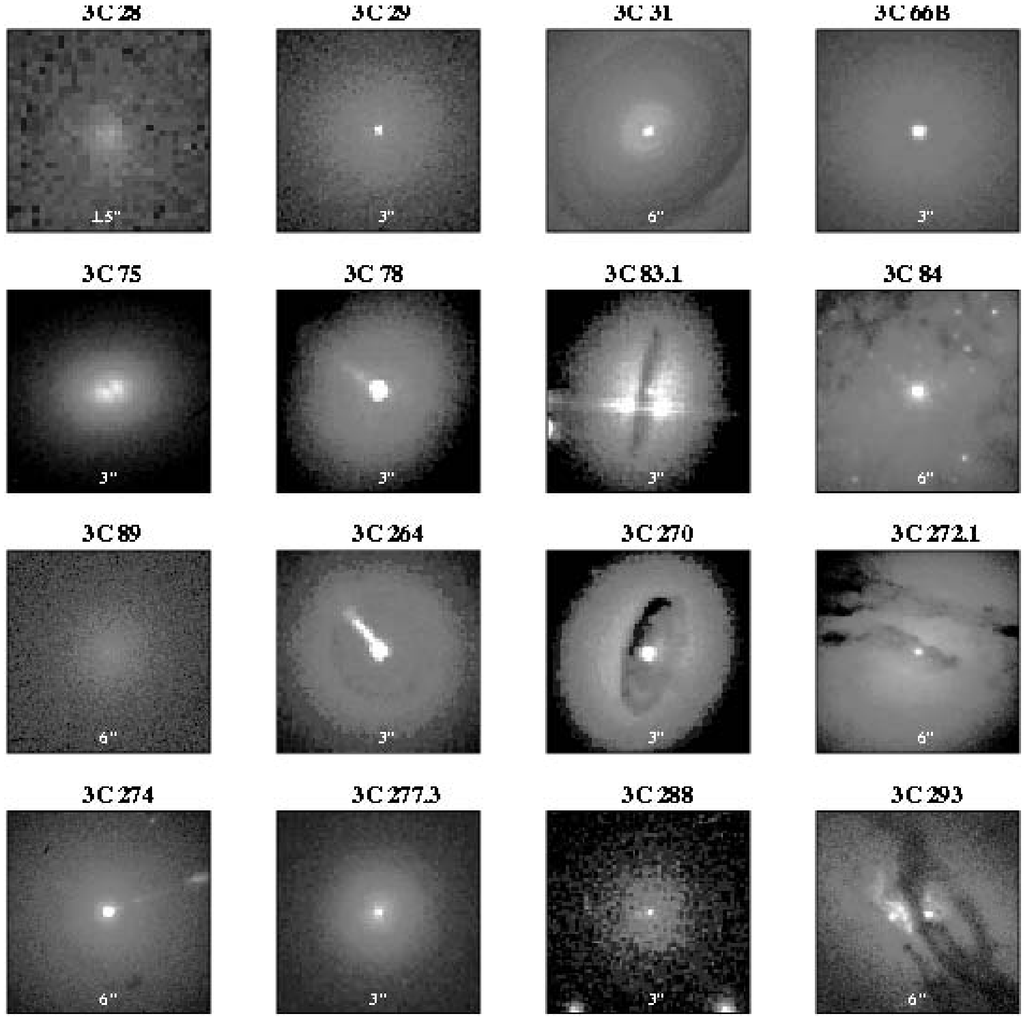


FIGURE 3. *HST/WFPC2* broad band images of a selection of the 3C sources.

brightest radio emission and the side of the radio emission with the shortest projected distance from the nucleus occurs on the same side of the quasar nucleus as the brightest, most significant Lyα emission.

There are few studies of radio-quiet quasars. Some early *HST* observations by Bahcall et al. (1996, 1997) and more detailed observations by Disney et al. (1995), Boyce et al. (1996, 1998), showed that for a total of about 25 objects the parent galaxies can be either ellipticals or spirals. Overall there are six clear examples of strong ongoing gravitational interaction between two or more galaxies and in 19 other cases close companion objects are detected, suggesting recent gravitational interaction.

4.2. *Seyfert morphologies*

The study of fueling processes in AGNs is key to our understanding of the structure and evolution of the central black hole and their host galaxies. Although fuel is readily available in the disk, it needs to overcome the centrifugal barrier to reach the innermost regions in disk and elliptical galaxies. Large-scale non-axisymmetries, such as galactic bars, are thought to be related to starburst activity within the central kpc, which preferentially oc-

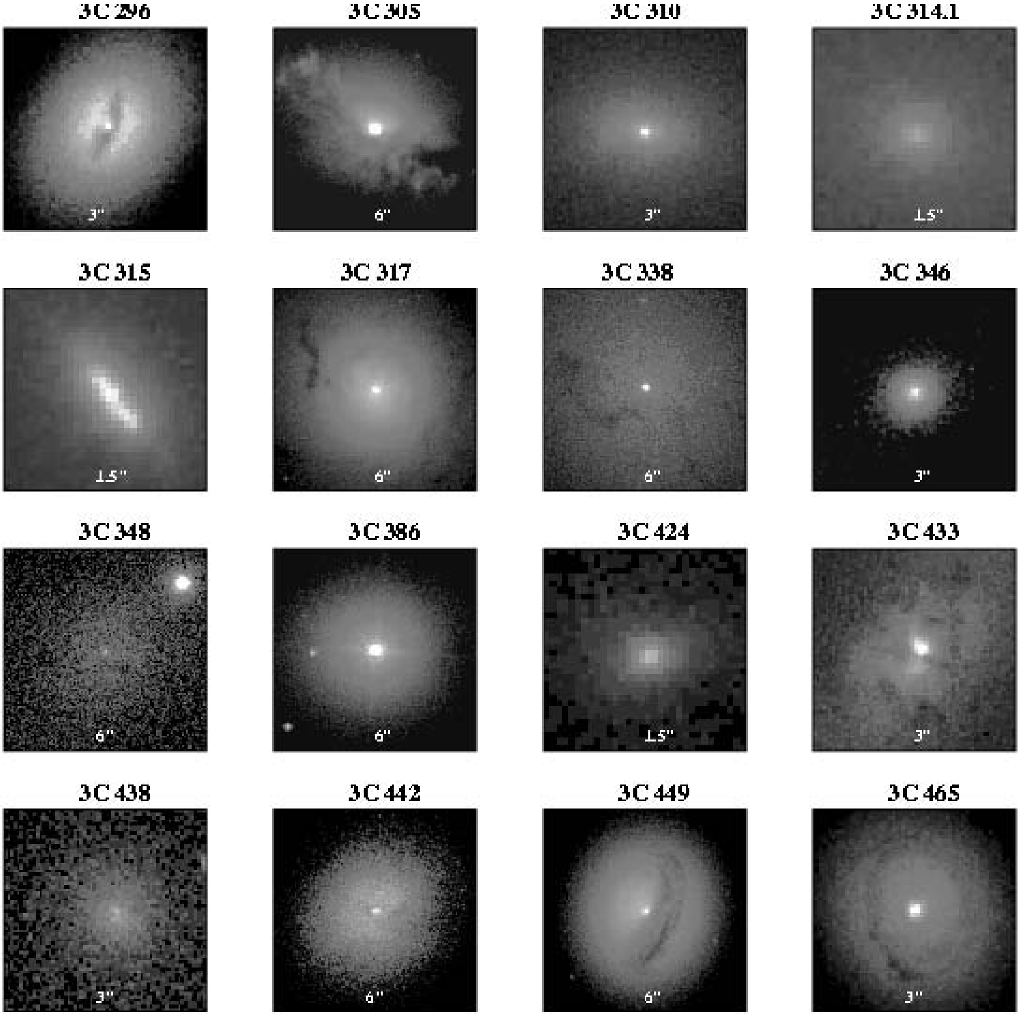


FIGURE 4. *HST/WFPC2* broad band images of a selection of the 3C sources.

curs in barred hosts (e.g. Heckman 1980, Balzano 1983, Devereux 1987, Kennicutt 1994). In a number of early optical surveys, the fueling of Seyfert activity in disk galaxies was linked to non-axisymmetric distortions of galactic gravitational potentials by large-scale stellar bars and tidal interactions (Adams 1977, Simkin et al. 1980, Dahari 1985a). This was supported by the argument that gravitational torques are able to remove the excess angular momentum from gas, which falls inwards, giving rise to different types of activity at the center (Sellwood & Wilkinson 1993, Phinney 1994). However early studies were not successful in showing significantly higher fractions of bars in host galaxies of AGN. Recently Knapen, Shlosman & Peletier (1999) have carried out NIR observations at high spatial resolution, on a sample of 34 non-active galaxies from the CF3 catalogue as well as a sample of 48 AGN from the CFA survey. They find that Seyfert hosts are barred more often than normal galaxies, 79% ± 7.5% barred for the Seyferts, vs. 59% ± 9% for the control sample, which is 2.5σ result.

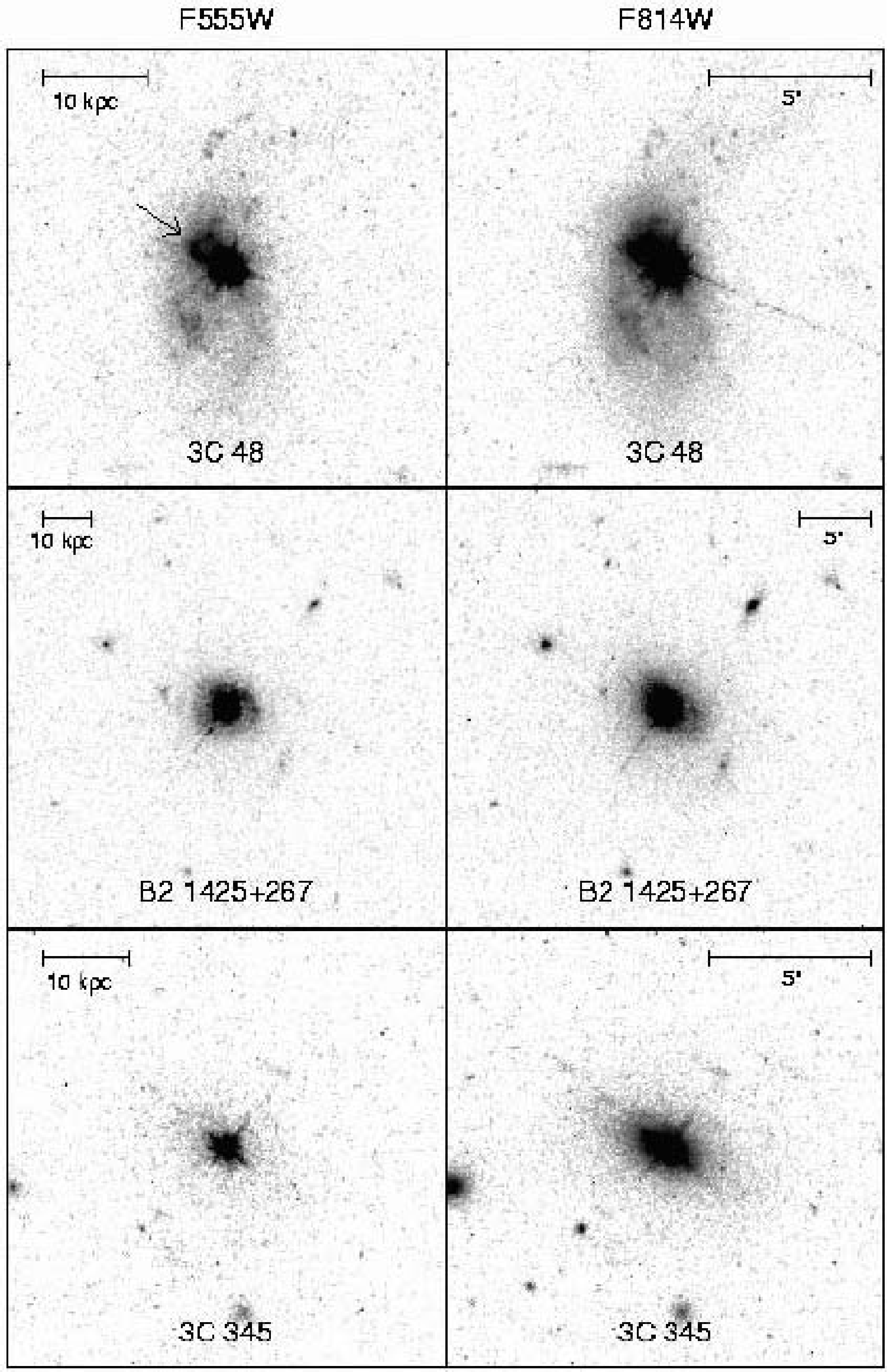


FIGURE 5. The host galaxies of three radio loud quasars (Bahcall et al. 1997).

Their result suggests, but does not prove, that there is an underlying morphological difference between Seyfert and non-Seyfert galaxies, and emphasize the prevalence of barred morphologies in disk galaxies in general, and in active galaxies in particular.

4.3. *Seyfert nuclei*

In the standard paradigm where AGNs are powered by non-spherical accretion onto massive black holes, the AGN's luminosity is proportional to the black hole mass accretion rate, which is about $0.01\,\mathrm{M}_\odot\ \mathrm{year}^{-1}$ for a bright Seyfert nucleus. Strong interactions or mergers with another galaxy are very efficient at funneling large amounts of gas by distorting the galactic potential and disturbing the orbits of gas clouds (Shlosman et al. 1989, Shlosman et al. 1990, Hernquist & Mihos 1995). This fuel is then brought down to several thousand Schwarzschild radii, or 10^{17} cm for a black hole mass of $10^8\,\mathrm{M}_\odot$, at which point viscous processes drive the final accretion onto the black hole.

However, direct observational evidence that galaxy encounters stimulate the luminosity of an AGN has been ambiguous (Adams 1977, Dahari 1985a, Petrosian 1983, Kennicutt & Keel 1984, Dahari 1985b, Bushouse 1986, Fuentes-Williams & Stocke 1988).

Malkan, Gorjiam & Tam (1998) have recently published the results of an *HST* snapshot imaging survey of 256 cores of active galaxies selected from the *Catalog of Quasars and Active Nuclei* by Véron-Cetty & Véron (1986, 1987). Of these, 91 are galaxies with nuclear optical spectra classified as "Seyfert 1," 114 galaxies are classified as "Seyfert 2," and 51 galaxies are classified as "HIIs." This large sample of high-resolution images was used to search for statistical differences in their morphologies.

The Seyfert galaxies do not, on average, resemble the HII galaxies, which have more irregularity and clumpiness associated with their high rates of current star formation. Conversely, none of the HII galaxies have the filaments or wisps photoionized by the active nucleus which are seen in Seyfert 1 and 2 galaxies. Of the Seyfert 1 galaxies, 63% have an unresolved nucleus, 50% of which are saturated, and 6% have such dominant nuclei that they would appear as "naked quasars" at higher redshifts. The presence of an unresolved and/or saturated nucleus is anti-correlated with an intermediate spectroscopic classification (such as Seyfert 1.8 or 1.9) and implies that those Seyfert 1s with weak nuclei in the *HST* images are extinguished and reddened by dust.

The vast majority of the Seyfert 2 galaxies show no central point source. If all Seyfert 2s were to have unresolved continuum sources like those in Seyfert 1s, they would be at least an order of magnitude fainter. In those galaxies without any detectable central point source (37% of the Seyfert 1s; 98% of the Seyfert 2s, and 100% of the HIIs), the central surface brightnesses are statistically similar to those observed in the bulges of normal galaxies.

Seyfert 1s and 2s both show circumnuclear rings in about 10% of the galaxies. Malkan et al. (1998) identified strong inner bars as often in Seyfert 1 galaxies (27%) as in Seyfert 2 galaxies (22%).

The Seyfert 2 galaxies are more likely than Seyfert 1s to show irregular or disturbed dust absorption in their centers as well as galactic dust lanes which pass very near their nuclei, and on average, tend to have latter morphological types than the Seyfert 1s. Thus it appears that the host galaxies of Seyfert 1 and 2 nuclei are *not* intrinsically identical. A galaxy with more nuclear dust and in particular more irregularly distributed dust is more likely to harbor a Seyfert 2 nucleus. This indicates that the higher dust-covering fractions in Seyfert 2s are the reason for their spectroscopic classification: their compact Seyfert 1 nucleus may have been obscured by galactic. This statistical result does not agree with the unified scheme for Seyfert galaxies, thus Malkan et al. (1998) propose that the obscuration which converts an intrinsic Seyfert 1 nucleus into an apparent Seyfert 2 occurs in the host galaxy hundred of parsecs from the nucleus. If so, this obscuration may have no relation to a hypothetical dust torus surrounding the central black hole.

4.4. *Morphologies of FR I radio galaxies*

Significant progress in the understanding of the inner structure of FR I have been obtained thanks to *HST* observations. A newly discovered feature in FR I are faint, nuclear optical components, which might represent the elusive emission associated with the AGN. Their study can be a powerful tool to directly compare the nuclear properties of FR I with those of other AGNs, such as BL Lac objects and powerful radio galaxies.

Chiaberge, Capetti & Celotti (1999) have studied a complete sample of 33 FR I sources from the 3CR observations carried out as part of the *HST* snapshot survey and discussed by Martel et al. (1997, 1999) (objects with $z < 0.1$) and by de Koff et al. (1996) (objects with $0.1 < z < 0.5$). Chiaberge et al. (1999) have shown that an unresolved nuclear source (Central Compact Core, CCC) is present in the great majority of these objects. The CCC emission, found to be strongly connected with the radio core emission, is anisotropic and can be identified with optical synchrotron radiation produced in the inner regions by a relativistic jet. These results are qualitatively consistent with the unifying model in which FR I radio galaxies are misoriented BL Lac objects. However, the analysis of objects with a total radio power of $< 2 \times 10^{26}$ W Hz^{-1}, shows that a CCC is found in all galaxies except three, for which absorption from extended dust structures clearly plays a role. This result casts serious doubts on the presence of obscuring thick tori in FR I as a whole.

The CCC luminosity represents a firm upper limit to any thermal component, and implies an optical luminosity of only $\lesssim 10^{-5}$–10^{-7} times Eddington (for a $10^9\,\mathrm{M}_\odot$ black hole). This limit on the radiative output of accreting matter is independent from but consistent with those inferred from X-ray observations for large elliptical galaxies, thus suggesting that accretion might take place in a low efficiency radiative regime (Fabian & Rees 1995).

The picture which emerges is that the innermost structure of FR I radio galaxies differs in many crucial aspects from that of the other classes of AGN; they lack the substantial BLR, tori and thermal disk emission, which are usually associated with active nuclei. Similar studies of higher luminosity radio galaxies will be clearly crucial to determine if either a continuity between low and high luminosity sources exists or, alternatively, they represent substantially different manifestations of the accretion process onto a supermassive black hole.

5. Demographics of massive black holes

5.1. *Ellipticals and S0 galaxies*

The evidence that massive dark objects (MDOs) are present in the centers of nearby galaxies has been reviewed by Kormendy & Richstone (1995), Bender, Kormendy, & Dehnen (1996), Kormendy et al. (1997), and van der Marel (1998). The MDOs are probably black holes, since star clusters of the required mass and size are difficult to construct and maintain, and since black hole quasar remnants are expected to be common in galaxy centers. Kormendy & Richstone (1995), Gebhardt et al. (2000a,b), and Merrit & Ferrarese (2000) suggest that at least 20% of nearby kinematically hot galaxies (ellipticals and spiral bulges) have MDOs and show a correlation $M_\bullet \simeq 0.003 M_{\mathrm{bulge}}$, where M_{bulge} is the mass of the hot stellar component of the galaxy. For a "bulge" with constant mass-to-light ratio Υ and luminosity L, $M_{\mathrm{bulge}} \equiv \Upsilon L$.

To further probe this correlation Gebhardt et al. (2000a,b) constructed dynamical models for a sample of nearby galaxies with *HST* photometry and ground-based kinematics.

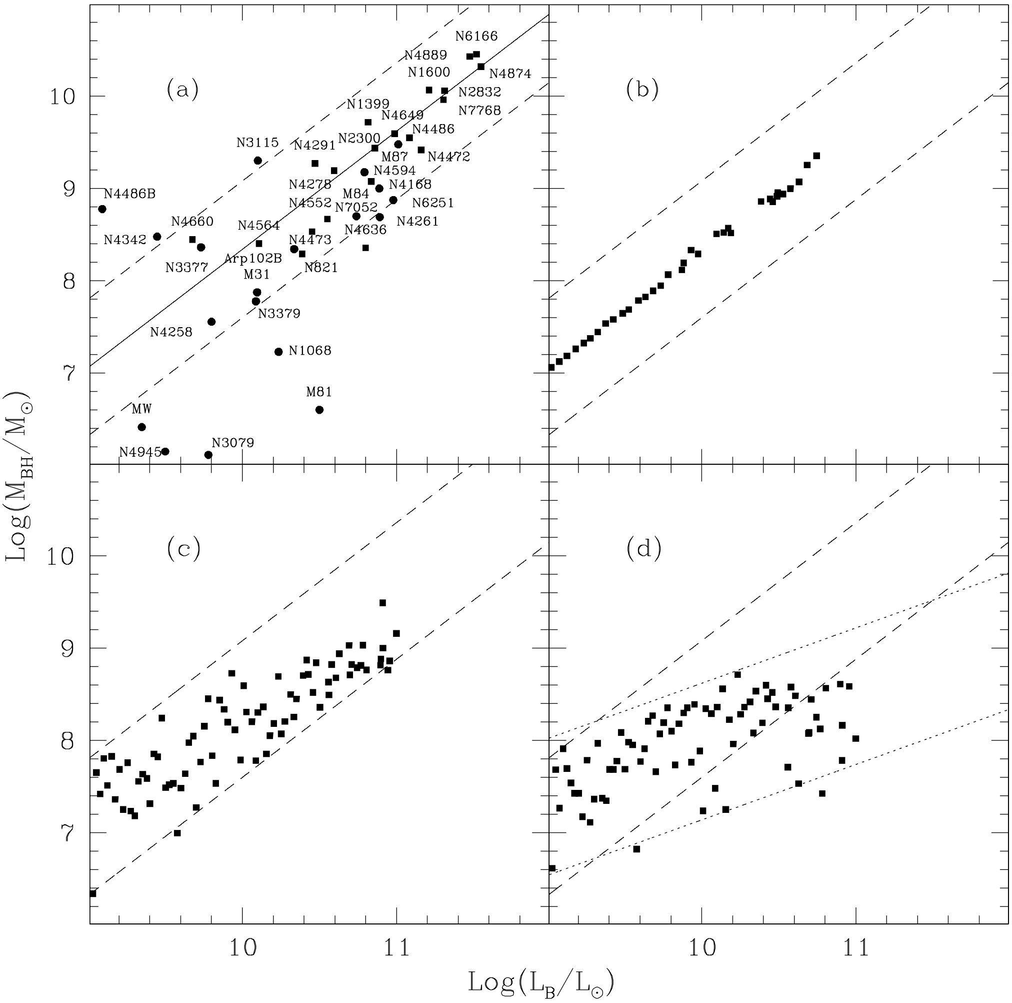


FIGURE 6. (**a**) Observed black hole masses and bulge luminosities for the samples of nearby galaxies compiled by Ho 1998, (*circles*) and Magorrian et al. (1998), (*squares*). The solid line shows the linear least square fit to the data, $\log(M_{\rm BH}\,{\rm M}_\odot) = 1.28\log(L_{\rm B}/{\rm L}_\odot) - 4.46$, while the dashed lines show the $\pm 1\sigma$ deviation ($\sigma = 0.74$). The dashed lines are the same in all panels for comparison. (**b**) Monte Carlo simulations of black hole masses and bulge luminosities with $M_{\rm acc} = 6\times 10^{-3}\Delta M_*$. (**c**) Same as (b) with $M_{\rm acc} = 1.4\times 10^{-3}(1+z)^2\Delta M_*$. (**d**) Same as (b) with $M_{\rm acc} = 10^{-6}(1+z)^2\ M_{\rm halo}\ \exp[-v_c/300\,{\rm km\,s^{-1}})^4]$. The linear least square fit to the results of model (iii) has a slope of 0.6, shown by the two dotted lines (Gebhardt et al. 20000a,b).

The models assume that each galaxy is axisymmetric, with a two-integral distribution function, arbitrary inclination angle, a position-independent stellar mass-to-light ratio Υ, and a central massive dark object (MDO) of arbitrary mass $M_\bullet$. They provide acceptable fits to 32 of the galaxies, and the mass-to-light ratios inferred show a correlation $\Upsilon \propto L^{0.2}$ (Fig. 6).

The result is that virtually every hot galaxy hosts a MDO with a mass ranging from $\sim 10^8\,{\rm M}_\odot$ to $2\times 10^{10}\,{\rm M}_\odot$ and roughly proportional to the mass of the spheroidal stellar component $M_\bullet \sim 0.006 M_{\rm bulge}$. MDO masses are just large enough to match those related to the QSO phenomenon. In fact, the highest bolometric luminosities of

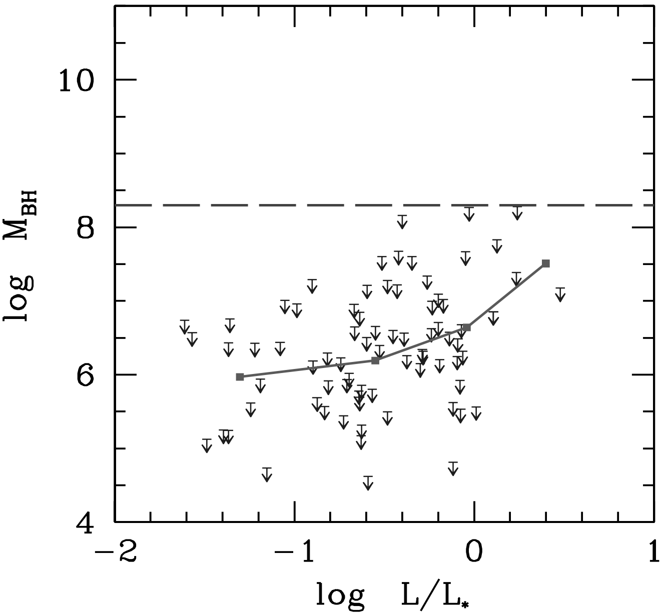


FIGURE 7. The upper limits on the MDO mass as a function of luminosity. The solid line represents the corresponding luminosity-averaged values. Also shown the *minimum* mass for QSO remnants (dashed line) (Sellwood & Moore 1999).

Quasars ($L_{\rm bol} \lesssim 4 \times 10^{48}$ erg/s) imply, under the assumption that they radiate at the Eddington limit, that the underlying black hole masses are comparable with those of the biggest MDOs detected in ellipticals, while the lowest QSO bolometric luminosities $L_{\rm bol} = 10^{46}$ erg/s still imply quite conspicuous black hole masses: $M_{\rm BH} > 2 \times 10^8\,{\rm M}_\odot$.

5.2. *Early and late type spirals*

Salucci et al. (1999) have studied the rotation curves of about one thousand spiral galaxies to investigate whether they could host *relic* black holes. The sample comprises late type spirals with at least one velocity measurement inside 250 pc for 435 objects and inside 350 pc for the remaining $\sim$ 500 objects. This would allow detections of MDOs of mass $M_{\rm MDO} \gtrsim (1-2) \times 10^8\,{\rm M}_\odot$, typical of a black hole powering a QSO and much larger than the ordinary stellar component inside this radius.

The upper limits obtained are shown in Fig. 7: the central objects in spirals are remarkably less massive than those detected in ellipticals: strict upper limits to their mass range are between $10^6\,{\rm M}_\odot$ at $L_{\rm B} \simeq 1/20 L_*$ to $\simeq 10^7\,{\rm M}_\odot$ at $\sim 3L_*$.

Salucci et al. (1999) also analyzed Sa galaxies which are considerably fewer than late type spirals, but in view of their very massive bulge, $M_b > 10^{10}\,{\rm M}_\odot$, they may well be the location of black holes with $M_{\rm BH} \gtrsim 10^8\,{\rm M}_\odot$.

They derived MDO upper limits which range from $5 \times 10^6\,{\rm M}_\odot < M^{\rm u}_{\rm MDO} < 10^{10}\,{\rm M}_\odot$, and are, at a given luminosity, one order of magnitude larger than the upper limits of the late type spirals MDOs. This implies that inside the innermost kpc, early type spirals have a large enough mass to envelop a large MDO and thus comfortably hide the remnant of a bright quasar.

6. Formation and evolution

There is a strong empirical relationship between black holes and their host galaxies, it is therefore important to compare their properties and distribution during the quasar era at $z \sim 3$. Were today's MBHs already fully formed by that time, or was the average MBH smaller in the past and later grew through accretion or mergers to form the present population?

The epoch of maximal activity in the Universe peaked just before the epoch of maximal star formation, and MBHs must have been formed and active before this time to provide the energy to power the quasars. While the rise of luminous quasars follows closely the rise in starbirth, the bright quasars reach their peak at $z \gtrsim 2(t \lesssim 1.6 \times 10^9\,\mathrm{yr})$, and then their number decline about $10^9\,\mathrm{yr}$ before the peak in star formation, which occurs at $z \sim 1.2(t = 2.6 \times 10^9\,\mathrm{yr})$.

This scenario favors models in which the black hole forms before the formation of the densest parts of galaxies. For this reason we can associate the birth of quasars with the spheroid formation, a process which is closely coupled to dense regions that collapse early.

The decline in the number of quasars at $z < 2$ may be caused by several mechanisms including the exhaustion of the available fuel. Galaxy mergers, which are an effective gas transport process, became less frequent as time evolves and involve a lower mean density.

Limits to the total mass of the black hole can occur through different mechanisms. Sellwood & Moore (1999) have suggested that strong bars form in the centers of recently formed galaxies and channel mass inwards to the central black hole which grows until its mass is ~ 0.02 of the mass of the disk. The bar then weakens and infalling mass forms a much more massive bulge which creates an inner Lindblad resonance which suppresses re-formation of a bar. Another mechanism proposed by Merritt (1998) is that the black hole makes the central stellar orbits become chaotic, with the consequence that non-axisymmetric disturbances are smoothed out and the rate of infall of accreting gas falls.

Cattaneo et al. (1999) have studied a very simple model in which both spheroids and supermassive black holes form through mergers of galaxies of comparable masses. They assumed that cooling only forms disk galaxies and that, whenever two galaxies of comparable masses merge, the merging remnant is an elliptical galaxy, a burst of star formation takes place and a fraction of the gas in the merging remnant is accreted by a central supermassive black hole formed by the coalescence of the central black holes in the merging galaxies.

This simple model is consistent with the shape of the quasar luminosity function, but its redshift evolution cannot be explained purely in terms of a decrease in the merging rate and of a decline in the amount of fuel available. To explain the evolution of the space density of bright quasars in the interval $0 < z < 2$, additional assumptions are needed, such as a redshift dependence of the fraction of available gas accreted or of the accretion time-scale.

In another scenario, proposed by Silk & Rees (1998) and by Haehnelt et al. (1998), a $\sim 10^6\,\mathrm{M}_\odot$ black hole forms by coherent collapse in the nucleus before most of the bulge gas turns into stars. If the black hole accretes and radiates at the Eddington limit, it can drive a wind with kinetic luminosity ~ 0.1 of the radiative luminosity. This deposits energy into the bulge gas, and will unbind it on a dynamical timescale with the result that the black hole mass will be limited to a value where it is able to shut off its own fuel supply (Blandford 1999).

A bright AGN may also limit infall of gas to form a disk, through Compton heating, radiation pressure on dust or direct interaction with a powerful wind. When the black hole mass and luminosity are large, the weakly bound, infalling gas will be blown away and an elliptical galaxy will be left behind. Only when the black hole mass is small, will a prominent disk develop. In this case, the bulge to disk ratio should correlate with the black hole mass fraction.

In summary, it seems very likely that black holes form first at quite large redshifts, $z \gg 2$ and can grow to their present sizes with standard radiative efficiency, by the time of the main quasar epoch at $t \sim 3\,\text{Gyr}$. There are several plausible mechanisms to limit the growth of the black hole by switching off the fuel supply, all of which need to be need to be studied further.

REFERENCES

Abramowicz, M., Chen, X., Kato, S., Lasota, J. P., & Regev, O. 1995 *ApJ* **438**, L37.

Adams, T. F. 1977 *ApJS* **33**, 19.

Baade, W. & Minkowski, R. 1954 *ApJ* **119**, 215.

Bahcall, J. N., Kirhakos, S., Saxe, D. H, & Schneider, D. P. 1997 *ApJ* **479**, 642.

Bahcall, J. N., Kirhakos, S., & Schneider, D. P. 1996. In *Quasar Hosts* (eds. D. Clements & I. Perez-Fournon). p. 37. Springer-Verlag.

Balzano, V. A. 1983 *ApJ* **268**, 602.

Bender, R., Kormendy, J., Dehnen, W. 1996 *ApJ* **464**, L123.

Blandford, R. D. 1999 In *Galaxy Dynamics* (eds. D. Merritt, M. Jalluri, & J. A. Sellwood). p. 182. ASP.

Bower, G. A., Green, R. F., Danks, A., et al. 1998 *ApJ* **492**, L111.

Boyce, P. J., Disney, M. J., Blades, J. C., Boksenberg, A., Crane, P., Deharveng, J. M., Macchetto, F. D., Mackay, C. D., & Sparks, W. B. 1996 *ApJ* **473**, 760.

Boyce, P. J., Disney, M. J., Blades, J. C., Boksenberg, A., Crane, P., Deharveng, J. M., Macchetto, F. D., Mackay, C. D., & Sparks, W. B. 1998 *MNRAS* **298**, 121.

Bushouse, H. A. 1986 *AJ* **91**, 255.

Cattaneo, A., Haenelt, M. G., & Rees, M. J. 1999 *MNRAS* **308**, 77.

Chiaberge, M., Capetti, A., & Celotti, A. 1999 *A&A* **349**, 77.

Chokshi, A. & Turner, E. L. 1992 *MNRAS* **259**, 421.

Dahari, O. 1985a *AJ* **90**, 1772.

Dahari, O. 1985b *ApJS* **57**, 643.

de Koff, S., Baum, S. A., Sparks, W. B., et al. 1996 *ApJS* **107**, 621.

Devereux, N. A. 1987 *ApJ* **323**, 91.

Disney, M. J., Boyce, P. J., Blades, J. C., Boksenberg, A., Crane, P., Deharveng, J. M., Macchetto, F. D., Mackay, C. D., Sparks, W. B., & Phillips, S. 1995 *Nature* **376**, 150.

Fabian, A. C. & Rees, M. J. 1995 *MNRAS* **277**, L55.

Fanaroff, B. L. & Riley, J. M. 1974 *MNRAS* **167**, 31.

Ferrarese, L. & Ford, H. C. 1999 *ApJ*, **515**, 563.

Ferrarese, L., Ford, H., & Jaffe, W. 1996 *ApJ* **470**, 444.

Fuentes-Williams, T. & Stocke, J. 1988 *AJ* **96**, 1235.

Gebhardt, K., et al. 2000a *ApJ* **539**, L9.

Gebhardt, K., et al. 2000b *ApJ*, **543**, L5, astro-ph/0007123.

Graham, J. A. 1979 *ApJ* **232**, 60.

Haehnelt, M., et al. 1998 *MNRAS* **300**, 817.

Heckman, T. 1980 *A&A* **88**, 365.

Hernquist, L. & Mihos, J. C. 1995 *ApJ* **448**, 41.

Ho, L. C. 1998 In *Observational Evidence for Black Holes in the Universe* (ed. S. K. Chakrabarti). p. 157. Kluwer.

Ichimaru, S. 1987 *ApJ* **214**, 840.

Israel, F. P. 1998 *A&A Rev.* **8**, 237.

Kennicutt, R. C. 1994. In *Mass-Transfer Induced Activity in Galaxies* (ed. I. Shlosman) p. 131. Cambridge Univ. Press.

Kennicutt, R. C. Jr. & Keel, W. C. 1984 *ApJ* **2791**, 5.

Knapen, J. H., Schlosman, I., & Peletier, R. F. 2000 *ApJ* **529**, 93.

Kormendy, J. & Richstone, D. 1995 *ARA&A* **33**, 581.

Kormendy, J., et al. 1997 *ApJ* **482**, L139.

Lehnert, M. D., Miley, G. K., Sparks, W. B., Baum, S. A., Biretta, J., Golombek, D., de Koff, S., Macchetto, F. D., & McCarthy, P. J. 1999 *ApJS* **123**, 351.

Macchetto, F. D., Marconi, A., Axon, D. J., Capetti, A., Sparks, W. B., & Crane, P. 1997 *ApJ* **489**, 579.

Magorrian, J., Tremaine, S., Richstone, D., Bender, R., Bower, G., Dressler, A., Faber, S. M., Gebhardt, K., Green, R., Grillmair, C., Kormendy, J., Lauer, T. R. 1998 *AJ* **115**, 2285.

Malin, D. F., Quinn, P. J., Graham, J. A. 1983 *ApJ* **272**, L5.

Malkan, M. A., Gorjian, V., & Tam, R. 1998 *ApJS* **117**, 25.

Marconi, A., Schreier, E. J., Koekemoer, A., Capetti, A., Axon, D. J., Macchetto, F. D., & Caon, N. 2000 *ApJ* **528**, 276.

Martel, A. R., Baum, S. A., Sparks, W. B., et al. 1997 *BAAS* **192**, 5204.

Martel, A. R., Baum, S. A., Sparks, W. B., Wyckoff, E., Biretta, J. A., Golombek, D., Macchetto, F. D., McCarthy, P. J., de Koff, S., & Miley, G. K. 1999 *ApJS* **122**, 81.

Merritt, D. 1998 *Comm. Ap.* **19**, 1.

Merrit, D. & Ferrarese, L. 2001 *ApJ* **547**, 140, astro-ph/0008310.

Mirabel, I. F., Laurent, O., Sanders, D. B., et al. 1999 *A&A* **341**, 667.

Miyoshi, M., et al. 1995 *Nature* **373**, 127.

Narayan, R. & Yi, I. 1994 *ApJ* **428**, L13.

Packham, C., Hough, J. H., Young, S., et al. 1996 *MNRAS* **278**, 406.

Petrosian, A. R. 1983 *Astrofizika* **18**, 548.

Phinney, E. S. P. 1994. In *Mass-Transfer Induced Activity in Galaxies* (ed. I. Shlosman). p. 1. Cambridge Univ. Press.

Quillen, A. C., de Zeeuw, P.T., Philley, E. S., Phillips, T. G. 1992 *ApJ* **391**, 121.

Quillen, A. C., Graham, J. R., & Frogel, J. A. 1993 *ApJ* **412**, 550.

Richstone, D., Ajhar, E. A., Bender, R., Bower, G., Dressler, A., Faber, S. M., Filippenko, A. V., Gebhardt, K., Green, R., Ho, L. C., Kormendy, J., Lauer, T. R., Magorrian, J., & Tremaine, S. 1998 *Nature* **395**, 14.

Salucci, P., Ratnam, C., Monaco, P., & Danese, L. 1999 *MNRAS* **307**, 637.

Schreier, E. J., Capetti, A., Macchetto, F. D., Sparks, W. B., Ford, H. J. 1996 *ApJ* **459**, 535.

Schreier, E. J., Marconi, A., Axon, D. J., Caon, N., Macchetto, F. D., Capetti, A., Hough, J. H., Young, S., & Packham, C. 1998 *ApJ* **499**, L143.

Sellwood, J. & Moore, E. M. 1999 *ApJ* **510**, 125.

Sellwood, J. A. & Wilkinson, A. 1993 *Rep. Prog. Phys.* **56**, 173.

Shlosman, I., Begelman, M. C., & Frank, J. 1990 *Nature* **345**, 679.

Shlosman, I., Frank, J., & Begelman, M. C. 1989 *Nature* **338**, 45.

Silk, J. I. & Rees, M. J. 1998 *A&A* **331**, L1.

Simkin, S. M., Su, H. J., & Schwarz, M. P. 1980 *ApJ* **237**, 404.

van der Marel, R. P. 1999 *AJ* **117**, 744.

van der Marel, R. P. & Van den Bosch, F. C. 1998 *AJ* **116**, 2220.

Véron-Cetty, M. P. & Véron, P. 1986 *A&AS* **66**, 335.

Véron-Cetty, M. P. & Véron, P. 1987 *ESO Sci. Rep.* **No. 5**.

Wandel, A., Peterson, B. N., & Malkan, M. A. 1999 *ApJ* **526**, 579.

Winge, C., Axon, D. J., Macchetto, F. D., Capetti, A., & Marconi, A. 1999 *ApJ* **519**, 134.

The *HST* Key Project to measure the Hubble Constant

By WENDY L. FREEDMAN,[1]
ROBERT C. KENNICUTT,[2]
JEREMY R. MOULD,[3] AND BARRY F. MADORE[4]

[1]Carnegie Observatories, 813 Santa Barbara St., Pasadena, CA 91101; wendy@ociw.edu

[2]Steward Observatory, University of Arizona, Tucson, AZ 85721

[3]Australian National University, Weston Creek, Canberra, ACT 2611, Australia

[4]NASA's IPAC Extragalactic DB, IPAC 100-22, Caltech, Pasadena, CA 91125

A decade of observing with *HST* also coincides with the completion of the last of the initial three Key Projects for *HST*, the measurement of the Hubble constant, H_0. Here we present the final results of the Hubble Space Telescope (*HST*) Key Project to measure the Hubble constant, summarizing our method, the results and the uncertainties. The Key Project results are based on a Cepheid calibration of several secondary distance methods applied over the range of about 60 to 400 Mpc. Based on the Key Project Cepheid calibration and its application to five secondary methods (type Ia supernovae, the Tully-Fisher relation, surface brightness fluctuations, type II supernovae, and the fundamental plane for elliptical galaxies), a combined value of $H_0 = 72 \pm 8$ km/sec/Mpc is obtained. An age conflict is avoided for current estimates of globular clusters and H_0 if we live in a Λ-dominated (or other form of dark energy) universe.

1. Introduction

When planning *HST*, pinning down H_0 was one of the scientific programs that drove the design and construction of the telescope. Although the original plans for a *Large Space Telescope* were scaled down during the mid-1970s, one of the primary arguments for an aperture of at least 2.4m was to enable the detection of Cepheid variables in the Virgo cluster (Smith 1989), a goal that was achieved within months of the corrective optics being installed in *HST* in December, 1993. This volume, however, is a testament to the variety of research areas which have exploded since the launch of *HST*, some in areas not predicted at all initially.

A decade ago, one of the biggest impediments to determining H_0 was the paucity of Cepheid-calibrating galaxies with which to calibrate the extragalactic distance scale. From the ground, there were only a handful of spiral galaxies for which Cepheid distances were available, useful for distance scale calibration. However, there were no galaxies close enough to measure Cepheid distances for galaxies host to type Ia supernovae, for example. And, although the precision in measuring relative distances to galaxies had improved enormously, with several new, independent secondary methods available, many of these methods could not be calibrated with Cepheids. Hence, the goal of the *HST* Key Project was to discover Cepheids and increase the numbers of calibrating galaxies applicable to a wide range of secondary methods, ultimately measuring H_0 to a level of ±10%, including systematic errors.

Obtaining an accurate value for the Hubble constant has proved an extremely challenging endeavor, a result primarily of the underlying difficulty of establishing accurate distances over cosmologically significant scales. Given the history of systematic errors dominating the accuracy of distance measurements, the approach adopted was to avoid relying on a single method alone, and instead to average over the systematics by calibrating and using a number of different methods. Determining H_0 accurately requires the measurement of distances far enough away that both the small and large-scale motions of galaxies become small compared to the overall Hubble expansion. To extend the distance scale beyond the range of the Cepheids, a number of methods that provide relative distances were chosen. We have used the *HST* Cepheid distances to provide an absolute distance scale for these otherwise independent methods, including the Type Ia supernovae (Gibson et al. 2000), the Tully-Fisher relation (Sakai et al. 2000), the fundamental plane for elliptical galaxies (Kelson et al. 2000), surface-brightness fluctuations (Ferrarese et al. 2000a), and Type II supernovae. This review is based on the results presented in more detail in a final summary of the Key Project by Freedman et al. (2001), which also contains a more extensive reference list. An earlier summary is given by Mould et al. (2000).

2. The H_0 *HST* Key Project

The main aims of the H_0 Key Project were (1) to discover Cepheids in, and determine distances to, a sample of nearby (closer than ∼ 20 Mpc) galaxies, and establish an accurate local distance scale, (2) to determine H_0 by applying the Cepheid calibration to several secondary distance indicators operating further out in the Hubble flow, (3) to intercompare the Cepheid and other distances to provide estimates of the external uncertainties for all of the methods, (4) to conduct tests of the universality of the Cepheid period-luminosity relation, in particular as a function of metal abundance. Finally, a further goal was to measure Cepheid distances to a small number of galaxies in each of the two nearest clusters (Virgo and Fornax) as an independent check on other Hubble constant determinations.

The Key Project involved the dedication and hard work of an enormous number of people with a range of expertise in the Cepheid distance scale, secondary distance methods, and crowded-field photometry. Over the years, the members of the Key Project have included W. L. Freedman, R. Kennicutt, J. R. Mould (co-PIs), F. Bresolin, S. Faber, L. Ferrarese, H. Ford, J. Graham, J. Gunn, M. Han, P. Harding, R. Hill, J. Hoessel, J. Huchra, S. Hughes, G. Illingworth, D. Kelson, L. Macri, B. F. Madore, R. Phelps, D. Rawson, A. Saha, S. Sakai, K. Sebo, N. Silbermann, P. Stetson, and A. Turner.

The excellent image quality of *HST* extends the limit out to which Cepheids can be discovered by a factor of ten from ground-based searches, and the effective search volume by a factor of a thousand. Furthermore, *HST* offers a unique capability in that it can be scheduled optimally and independently of the phase of the Moon, the time of day, or weather, and there are no seeing variations. Before the launch of *HST*, most Cepheid searches were confined to our own Local Group of galaxies, and the very nearest surrounding groups, and the numbers of Cepheid calibrators for various methods was dismally small (5 for the Tully-Fisher relation, one for the surface-brightness fluctuation method, and *no* Cepheid calibrators were available for Type Ia supernovae.

In each nearby target spiral galaxy in the Key Project sample, Cepheid searches were undertaken in regions active in star formation, but low in apparent dust extinction. To the largest extent possible, we avoided high-surface-brightness regions in order to minimize source confusion or crowding. For each galaxy, over a two-month time interval, *HST* images in the visual (V-band, 5550Å), and in the near-infrared (I band, 8140 Å), were made using the corrected Wide Field and Planetary Camera 2 (WFPC2). Two wavelength bands were chosen to enable corrections for dust extinction. The time distribution of the observations was set to follow a power-law, enabling the detection and measurement of Cepheids with a range of periods optimized for minimum aliasing between 10 and 50 days.

For each galaxy observed as part of the Key Project, the Cepheid positions, magnitudes, and periods are available at http://www.ipac.caltech.edu/H0kp/H0KeyProj.html. In addition, photometry for non-variable stars that can be used for photometry comparisons, as well as medianed (non-photometric) images for these galaxies are also available. These images are also archived in NED, and can be accessed on a galaxy-by-galaxy basis from http://nedwww.ipac.caltech.edu.

Since each individual secondary method is likely to be affected by its own (independent) systematic uncertainties, to reach a final overall uncertainty of ±10%, the numbers of calibrating galaxies for a given method were chosen initially so that the final (statistical) uncertainty on the zero point for that method would be only $\sim$ 5%. Cepheid distances were obtained for 18 galaxies. These galaxies lie at distances between 3 and 25 Mpc. *HST* has also been used to measure Cepheid distances to 6 galaxies, targeted specifically to be useful for the calibration of Type Ia supernovae (e.g. Sandage et al. 1996). Finally, an *HST* distance to a single galaxy in the Leo I group, NGC 3368, was measured by Tanvir and collaborators (Tanvir et al. 1999). Subsequently and fortuitously, NGC 3368 was host to a Type Ia supernova, useful for calibrating H_0 (Jha et al. 1999). In addition, recently, SN1999by occurred in NGC 2841, a galaxy for which Cepheid observations have been taken in Cycle 9 (GO-8322).

Each galaxy within the Key Project was analyzed by two independent groups within the team: only at the end of the data reduction process (including the Cepheid selection and distance determinations) were the two groups' results intercompared. This "double-blind" procedure proved extremely valuable, both for catching simple (operator) errors, as well as enabling us to provide a more realistic estimate of the external data reduction errors for each galaxy distance. We also undertook a series of artificial star tests to better quantify the effects of crowding, and to understand the limits in each of these software packages (Ferrarese et al. 2000b). The final distances were obtained by fitting each individual galaxy VI period-luminosity relations to those for the Large Magellanic Cloud (LMC) measured by Udalski et al. (1999), assuming a distance to the LMC of 18.50 ± 0.10 mag (*rms*).

3. Metallicity and the Cepheid Distance Scale

Accurately establishing whether the zero point of the Cepheid period-luminosity relation sensitive to chemical composition has proven to be very challenging, and the issue has not yet been definitively resolved (see Freedman et al. 2001 and references therein). Neither the magnitude of the effect nor its wavelength dependence have yet been firmly established, but the observational and theoretical evidence for an effect is steadily growing. Some recent theoretical models (e.g. see Alibert et al. 1999) suggest that at the VI bandpasses of the H_0 Key Project, the effect of metallicity on the derived distance is small, amounting to only about 0.1 mag over a dex (or a factor of ten in metallicity). Unfortunately, however, the sign of the effect is still uncertain. For example, Caputo, Marconi & Musella (2000) find a slope of 0.27 mag/dex, with the opposite sign. Thus, for the present, calibrating the metallicity effect based on theoretical models alone is not feasible. Considering all of the evidence presently available and the (still considerable) uncertainties, we adopted a metallicity correction to the Key Project distances of -0.2 ± 0.2 mag/dex, approximately the mid-range of current empirical values.

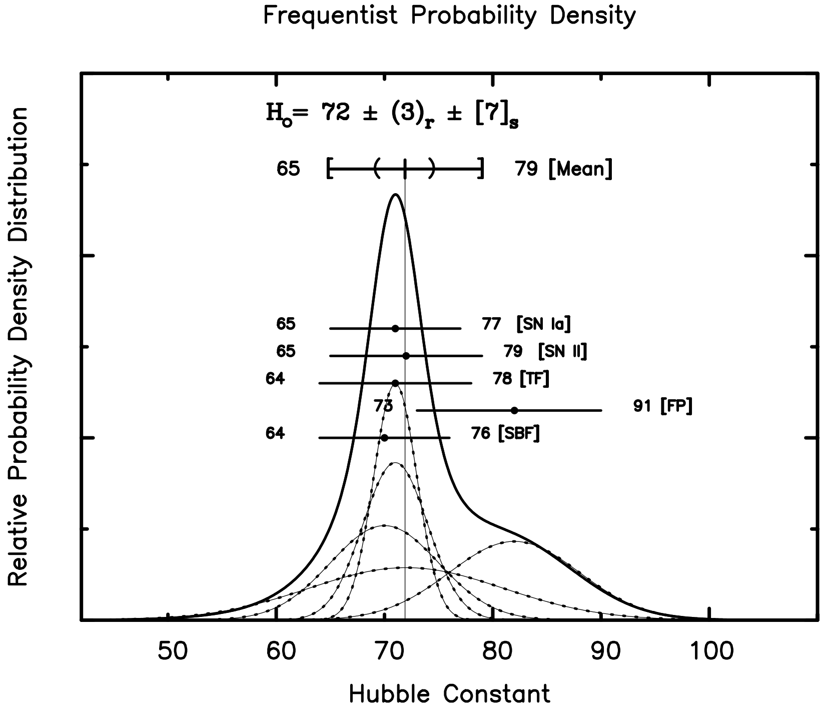


FIGURE 1. Probability distributions for the individual H_0 determinations. Each is represented by a Gaussian of unit area, with a dispersion given by the individual σ values. The cumulative distribution is given by the solid thick line. The median value is $H_0 = 72 \pm 3 \pm 7$ km/sec/Mpc. The *random* uncertainty is defined at the ±34% points of the cumulative distribution.

4. The Hubble Constant

Calibrating 5 secondary methods with Cepheid distances, Freedman et al. (2001) find $H_0 = 72 \pm 3$ (random) ±7 (systematic) km/sec/Mpc. Type Ia supernovae are the secondary method which currently extends out to the greatest distances, ~ 400 Mpc. All of the methods (Types Ia and II supernovae, the Tully-Fisher relation, surface brightness fluctuations, and the fundamental plane) are in extremely good agreement: four of the methods yield a value of H_0 between 70–72 km/sec/Mpc, and the fundamental plane gives $H_0 = 82$ km/sec/Mpc. As described in detail in Freedman et al., the largest remaining sources of error result from (a) uncertainties in the distance to the Large Magellanic Cloud, (b) photometric calibration of the *HST* Wide Field and Planetary Camera 2, (c) metallicity calibration of the Cepheid period-luminosity relation, and (d) cosmic scatter in the density (and therefore, velocity) field that could lead to observed variations in H_0 on very large scales. These systematic uncertainties affect the determination of H_0 for all of the relative distance indicators, and they cannot be reduced by simply combining the results from different methods: they dominate the overall error budget in the determination of H_0.

Figure 1 shows the probability distributions for the individual H_0 determinations. The median value is $H_0 = 72 \pm 3 \pm 7$ km/sec/Mpc. A Bayesian analysis was also done assuming that the priors on H_0 and on the probability of any single measurement being correct are uniform. Here the product of the probability distributions yields $H_0 = 72 \pm 2 \pm 7$ km/sec/Mpc. The formal uncertainty on this result is very small, and simply reflects the fact that four of the values are clustered very closely, while the uncertainties in the

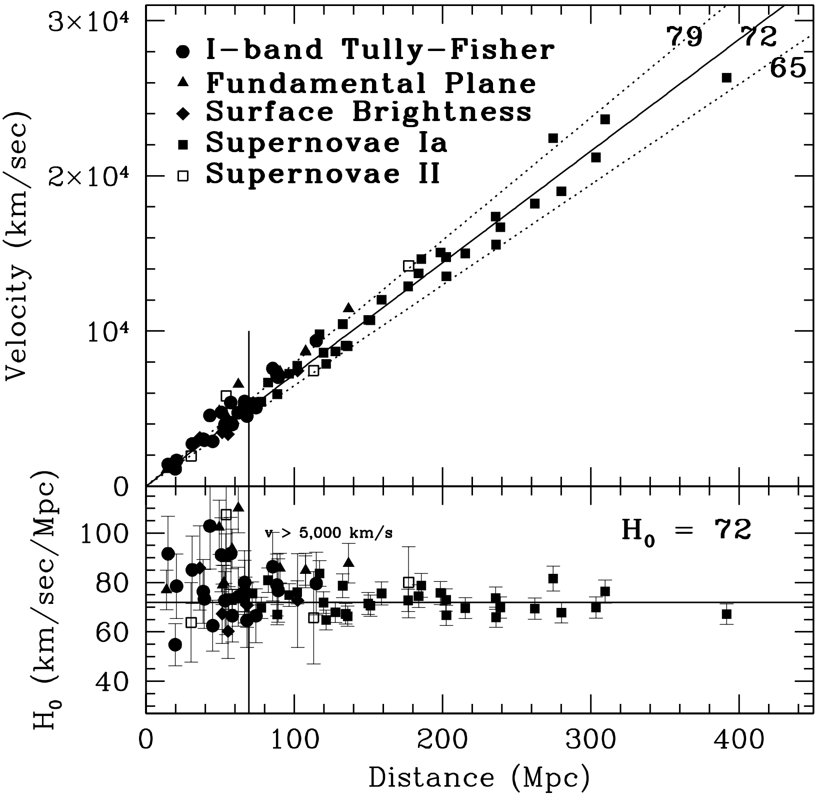


FIGURE 2. Composite Hubble diagram of velocity versus distance for Type Ia supernovae (solid squares), the Tully-Fisher relation (solid circles), surface-brightness fluctuations (solid diamonds), the fundamental plane (solid triangles), and Type II supernovae (open squares). In the bottom panel, the values of H_0 are shown as a function of distance. The Cepheid distances have been corrected for metallicity. The Hubble line plotted in this figure has a slope of 72 km/sec/Mpc, and the adopted distance to the LMC is taken to be 50 kpc.

fundamental method are large. Adjusting for differences in calibration, these results are also in excellent agreement with the weighting based on numerical simulations of the errors by Mould et al. (2000) which yielded 71 ± 6 km/sec/Mpc similar to Madore et al. (1999) giving $H_0 = 72 \pm 5 \pm 7$ km/sec/Mpc based on a smaller subset of available Cepheid calibrators. Figure 2 displays the results graphically in a composite Hubble diagram. The Hubble line plotted in this figure has a slope of 72 km/sec/Mpc.

5. H_0 from methods independent of Cepheids

At present, to within the uncertainties, there is broad agreement in H_0 values for completely independent techniques. Published values of H_0 based on the Sunyaev-Zeldovich (SZ) method have ranged from ~ 40–80 km/sec/Mpc (e.g. Birkinshaw 1999). The most recent two-dimensional interferometry SZ data for well-observed clusters yield $H_0 = 60 \pm 10$ km/sec/Mpc. The systematic uncertainties are still large, but the near-term prospects for this method are improving rapidly as additional clusters are being observed, and higher-resolution X-ray and SZ data are becoming available (e.g. Reese et al. 2000). A second method for measuring H_0 at very large distances, also independent of the need for any local calibration, comes from the measurement of time delays in

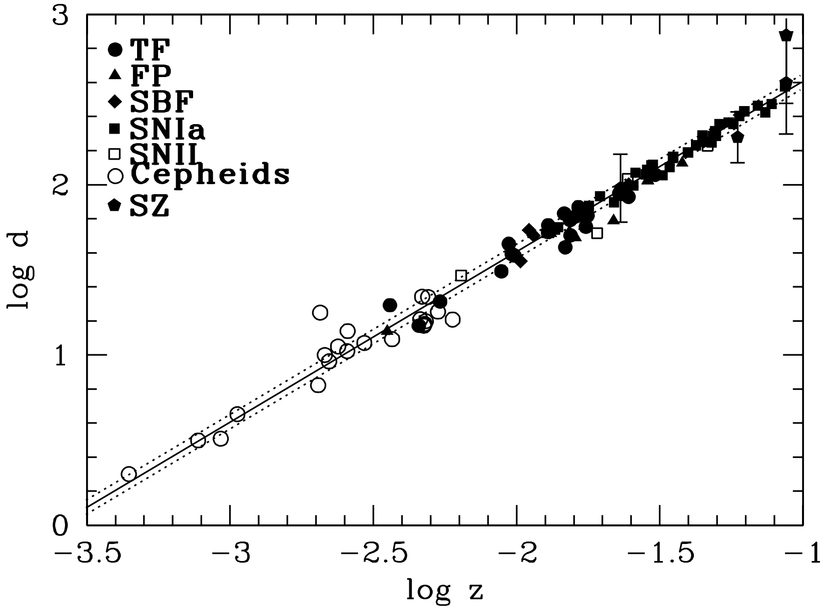


FIGURE 3. Hubble diagram ($\log d$ versus $\log v$) covering over 3 orders of magnitude, including distances obtained locally from Cepheids, from 5 secondary methods, and for 4 clusters with recent Sunyaev-Zel'dovich measurements out to $z \sim 0.1$. At redshifts beyond z of 0.1, other cosmological parameters (the matter density, Ω_m, and the cosmological constant, Ω_Λ) become important.

gravitational lenses. H_0 values based on this technique appear to be converging to the mid-60 km/sec/Mpc range (Williams & Saha 2000). As more lenses with time delays are discovered and monitored, this method also is likely to improve substantially in the near future. A Hubble diagram ($\log d$ versus $\log v$) is plotted in Figure 3.

6. The Expansion Age and implications for cosmology

An accurate determination of the expansion age of the universe requires not only the value of H_0, but also accurate measurements of Ω_m and Ω_Λ. At the time when the Key Project was begun, the strong motivation from inflationary theory for a flat universe, coupled with a strong theoretical preference for $\Omega_\Lambda = 0$, favored the Einstein-de Sitter model (e.g. Kolb & Turner 1990). In addition, the ages of globular cluster stars were estimated at that time to be $\sim$ 15 Gyr (Chaboyer et al. 1996). However, for a value of $H_0 = 72$ H_0, the Einstein-de Sitter model yields a very young expansion age of only 9 ± 1 Gyr, significantly younger than the globular cluster and other age estimates.

In Figure 4 H_0t_0 is plotted as a function of Ω, for a value of $H_0 = 72$ km/sec/Mpc and $t_0 = 12.5$ Gyr. The $\pm$1- and 2-σ limits are plotted for $H_0 = 72$ km/sec/Mpc, $t_0 = 12.5$ Gyr, assuming independent uncertainties of $\pm 10\%$ in each quantity, and adding the uncertainties in quadrature. These data are consistent with either a low-density ($\Omega_m \sim 0.1$) open universe, or a flat universe with $\Omega_m \sim 0.35$, $\Omega_\Lambda = 0.65$; however, with these data alone, it is not possible to discriminate between an open or flat universe.

A non-zero value of the cosmological constant helps to avoid a discrepancy between the expansion age and other age estimates. For $H_0 = 72$ km/sec/Mpc, $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, the expansion age is 13 ± 1 Gyr. This age is consistent to within the uncertainties with recent globular cluster ages, which have been revised downward to 12–13 Gyr based on

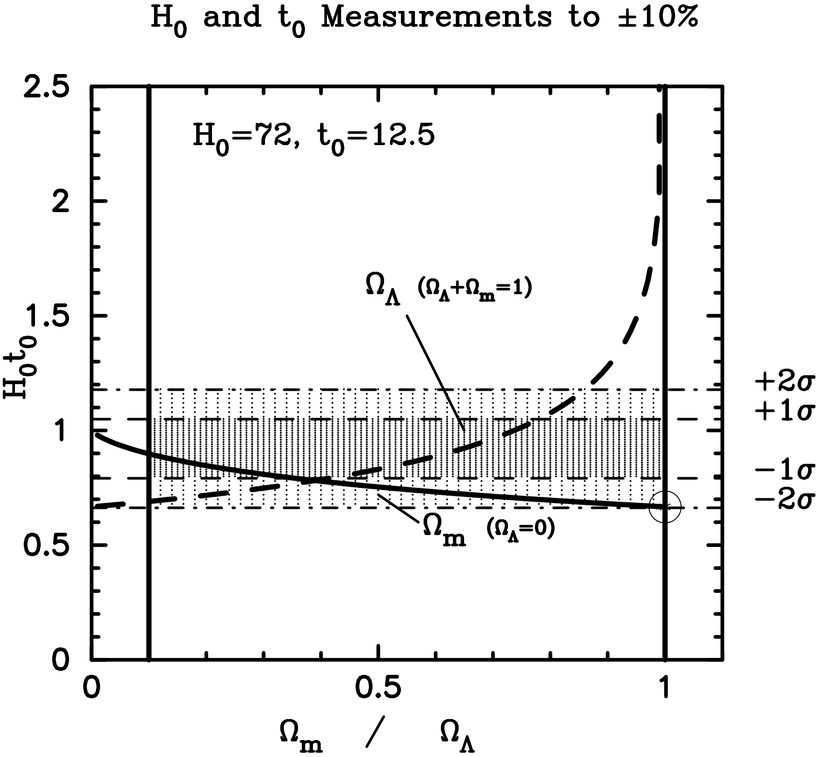


FIGURE 4. Plot of H_0t_0 as a function of Ω. Two curves are shown: the solid curve is for the case where $\Lambda = 0$, and the dashed curve allows for non-zero Λ under the assumption of a flat universe. The open circle at $\Omega_m = 1$, $\Lambda = 0$, represents the Einstein-de Sitter case, and is inconsistent with the current values of H_0 and t_0 only at a $\sim$ 2-σ level.

a new calibration from the Hipparcos satellite (Chaboyer 1998), with the evidence from recent cosmic microwave background anisotropy experiments (de Bernardis et al. 2000 and with recent data from high-redshift supernovae providing evidence for a non-zero cosmological constant (Riess et al. 1998; Perlmutter et al. 1999).

7. Conclusions

The *HST* Key Project to measure the Hubble constant has now been completed. *HST* was used to measure Cepheid distances to 18 nearby spiral galaxies. Calibrating 5 secondary methods with these revised Cepheid distances yields $H_0 = 72 \pm 3$ (random) ± 7 (systematic) km/sec/Mpc, or $H_0 = 72 \pm 8$ km/sec/Mpc, combining the total errors in quadrature. To within existing uncertainties, these results are in good agreement with other completely independent methods for measuring H_0, for example, the Sunyaev-Zeldovich and gravitational lense time delay methods. The largest remaining sources of error result from (a) uncertainties in the distance to the Large Magellanic Cloud, (b) photometric calibration of the *HST* Wide Field and Planetary Camera 2, (c) metallicity calibration of the Cepheid period-luminosity relation, and (d) cosmic scatter in the density (and therefore, velocity) field that could lead to observed variations in H_0 on very large scales. A value of $H_0 = 72$ km/sec/Mpc yields an expansion age of $\sim$ 13 Gyr for a flat universe (consistent with the recent cosmic microwave background anisotropy results) if $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$. Combined with the current best estimates of the ages of globular clusters (~ 12.5 Gyr), our results favor a Λ-dominated universe.

We extend our thanks to the organizing committee for a very stimulating meeting covering such a wide range of topics. It is a pleasure to acknowledge the many individuals at ST ScI and NASA for their help in carrying out this project over the past decade. In particular we thank our *HST* program coordinator throughout our entire (complicated) observing process, Doug Van Orsow, and former ST ScI director Robert Williams. We also fondly remember Marc Aarsonson who initially led our project before his tragic death in 1987.

REFERENCES

Alibert, Y., Baraffe, I., Hauschildt, P., & Allard, F. 1999 *A&A* **344**, 551.

Birkinshaw, M. 1999 *Phys. Rep.*, **310**, 97.

Caputo, F., Marconi, M., Musella, I., & Santolamazza, P. 2000 *A&A* **359**, 1059.

Chaboyer, B., Demarque, P., Kernan, P. J., & Krauss, L. M. 1996 *Science* **271**, 957.

Chaboyer, B., Demarque, P., Kernan, P. J., & Krauss, L. M. 1998 *ApJ* **494**, 96

de Bernardis, P., et al. 2000 *Nature* **404**, 955.

Ferrarese, L., et al. 2000a *ApJ* **529**, 745.

Ferrarese, L., Silbermann, N. A., Mould, J. R., Stetson, P. B., Saha, A., Freedman, W. L., & Kennicutt, R. C. 2000b *PASP*, **112**, 177.

Freedman, W. L., et al. 2001 *ApJ* **553**, 47, astroph/0012376.

Gibson, B. K., et al. 2000 *ApJ* **529**, 723.

Jha, S., et al. 1999 *ApJS* **125**, 73.

Kelson, D. D., et al. 2000 *ApJ* **529**, 768.

Kolb, E. W. & Turner, M. S. 1990 *The Early Universe*. Addison-Wesley.

Madore, B. F., et al. 1999 *ApJ*, **515**, 29.

Mould, J. R., et al. 2000 *ApJ* **529**, 786.

Perlmutter, S., et al. 1999 *ApJ* **517**, 565

Reese, E. D., Mohr, J. J., Carlstrom, J. E., Joy, M., Grego, L., Holder, G. P., Holzapfel, W. L., Hughes, J. P., Patel, S. K., & Donahue, M. 2000 *ApJ*, **533**, 38.

Riess, A. G., et al. 1998 *AJ* **116**, 1009

Sandage, A. R., Saha, A., Tammann, G. A., Labhardt, L., Panagia, N., & Macchetto, F. D. 1996 *ApJ* **460**, 15.

Smith, R. 1989 in *The Space Telescope*, pp. 145–146. Cambridge University Press.

Tanvir, N. R., Ferguson, H. C., & Shanks, T. 1999 *MNRAS* **310**, 175

Udalski, A., Szymanski, M., Kubiak, M., Pietrzynski, G., Soszynski, I, Wozniak, P., & Zebrun, K. 1999 *Acta Astronomica*, **49**, 201.

Williams, L. L. R. & Saha, P. 2000 *AJ*, **119** 39.

H_0 from Type Ia supernovae

By G. A. TAMMANN,[1] A. SANDAGE,[2] AND A. SAHA[3]

[1]Astronomisches Institut der Universität Basel, Venusstrasse 7, CH-4102 Binningen, Switzerland

[2]The Observatories of the Carnegie Institution of Washington, 813 Santa Barbara Street, Pasadena, CA 91101

[3]National Optical Astronomy Observatories, 950 North Cherry Avenue, Tucson, AZ 85726

The Hubble diagrams in B, V, and I of a complete sample of 35 SNe Ia with $(B-V) < 0.06$ and $1200 < v \lesssim 30\,000\ \mathrm{km\,s^{-1}}$ have a scatter of only $0\overset{\mathrm{m}}{.}1$, after small corrections are applied for differences in decline rate Δm_{15} and color $(B-V)$. The tightness of the Hubble diagrams proves blue SNe Ia to be the best "standard candles" known. Their absolute magnitudes $M_{B,V,I}$ are calibrated by eight SNe Ia with Cepheid distances from *HST*. Combining this calibration with the appropriate Hubble diagrams yields a large-scale value of $H_0 = 58.5 \pm 6.3$ at the 90% confidence level.

The Hubble diagram of SNe Ia has so small scatter that it seems feasible to determine Λ "locally," i.e. within $z \lesssim 0.12$, once 100–200 SNe Ia with good photometry will be available. Such a local determination would minimize evolutionary effects and K-term corrections.

Clusters of galaxies have provided useful Hubble diagrams through brightest cluster members, TF distances, and $D_n - \sigma$ and fundamental plane distances, but with significantly more scatter ($\sigma = 0\overset{\mathrm{m}}{.}2 - 0\overset{\mathrm{m}}{.}3$) than SNe Ia. The zeropoint calibration of these Hubble diagrams is an additional problem, which is aggravated by the high weight of any adopted distance of the Virgo cluster and by selection effects of clusters with only few well studied members. If a Virgo cluster modulus of $(m-M) = 31.60 \pm 0.20$ is adopted for calibration—a value which is well secured by Cepheids, SNe Ia, and the TF method, and which also agrees with less definitive distances from the globular cluster luminosity function, novae, and the $D_n - \sigma$ method—one finds $H_0 = 55 \pm 5$. The reasons are explained why some authors have found higher values from clusters.

The determination of H_0 from field galaxies is beset by selection effects of magnitude-limited samples (Malmquist bias). Authors who have properly allowed for bias have consistently obtained $H_0 \approx 55 \pm 5$ within $v \lesssim 5000\ \mathrm{km\,s^{-1}}$ based on the TF and other methods. The calibration rests only on Cepheids, independent of any adopted Virgo cluster modulus.

Giving highest weight to SNe Ia it is concluded that $H_0 = 58 \pm 6$.

1. Introduction

HST has brought an enormous progress in determining extragalactic distances by providing 26 Cepheid distances to late-type galaxies. This is of paramount importance not only for the determination of the Hubble constant H_0, but increasingly also for the physical understanding of individual galaxies whose linear sizes, luminosities, masses, radiation densities etc. depend on distance. Unfortunately the progress is confined to late-type galaxies, the distances of early-type galaxies having profited only indirectly.

An overview of the Cepheid distances from *HST* is given in Table 1. Cepheids carry by far the largest weight for the foundation of the extragalactic distance scale. Their zeropoint is based on an adopted LMC modulus of 18.50, which is confirmed by various distance indicators to within $\pm 0\overset{\mathrm{m}}{.}10$. Cepheid distances are uncontroversial to a large extent with possibly remaining small metallicity effects discussed in Section 2.5.

The 26 galaxies with Cepheid distances from *HST*, all beyond the Local Group, yield a mean value of $H_0 = 65 \pm 4$ [$\mathrm{km\,s^{-1}\,Mpc^{-1}}$]. Yet these galaxies lie within $v = 1200\ \mathrm{km\,s^{-1}}$ which is too local for this determination having any cosmic significance.

No. of Galaxies	Authors	No. of SNe Ia
18 Cepheid distances	Freedman 2001	1
1 Cepheid distance	Tanvir et al. 1995	1
7 Cepheid distances	Saha et al. 1999, 2000a	8

Table 1. Cepheid distances from *HST*

An additional distance indicator is therefore needed which can be calibrated by means of the available Cepheids and which carries the distance scale out to $\gtrsim 10\,000\ \mathrm{km\,s^{-1}}$, i.e. well beyond the influence of peculiar and streaming motions. This distance indicator must be *proven* to be reliable and its intrinsic dispersion must be known in order to control selection effects (cf. Section 4).

Many distance indicators have been proposed, but the proof of their reliability is extremely difficult. The demonstration that they can reproduce a limited number of Cepheid distances, which carry themselves individual errors of up to $\sim 0\overset{\mathrm{m}}{.}2$, is not good enough, and the internal dispersion remains ill defined. Just because various distance indicators are said to agree amongst themselves is no proof that they form a correct distance scale. If they have similar intrinsic dispersion and if each is not corrected for systematic effects of observational bias errors, their agreement is spurious.

The only satisfactory way to prove potential distance indicators to be useful for the determination of $\mathrm{H_0}$ is the demonstration that they define a linear relation of slope 0.2 (corresponding to linear expansion) in the Hubble diagram out to $\gtrsim 10\,000\ \mathrm{km\,s^{-1}}$. The scatter about this line provides in addition the intrinsic dispersion of the method if proper allowance is made for observational errors and for the influence of peculiar motions which, however, at $10\,000\ \mathrm{km\,s^{-1}}$, are negligible for all practical purposes.

This requirement of reliable long-range distance indicators is very well met by supernovae of type Ia (SNe Ia), and they can be calibrated by Cepheids. SNe Ia offer therefore the optimum route to $\mathrm{H_0}$ and are discussed in Section 2. Cluster distances follow in Section 3; they also define a useful Hubble diagram, but their zeropoint calibration is still under debate. The most difficult and least satisfactory way to $\mathrm{H_0}$ by means of distances of field galaxies is discussed in Section 4. The conclusions are given in Section 5.

2. $\mathbf{H_0}$ from SNe Ia

This Section is the result of an *HST* project for the luminosity calibration of SNe Ia that also includes L. Labhardt, F. D. Macchetto, and N. Panagia. The collaboration of J. Christensen, B. R. Parodi, H. Schwengeler, and F. Thim at different stages of the project is acknowledged. We thank also the many collaborators who work behind the scene at the Space Telescope Science Institute for their continued support.

2.1. *The sample*

From a parent population of 67 SNe Ia with known B and V at maximum, $(B_{\max} - V_{\max}) < 0.20$ (In the following $(B - V)$ for short), and $v^{\mathrm{CMB}} \lesssim 30\,000\ \mathrm{km\,s^{-1}}$ (to minimize cosmological effects; galaxies with $V_{220} > 3000\ \mathrm{km\,s^{-1}}$ being corrected for the local $630\ \mathrm{km\,s^{-1}}$ motion with respect to the CMB) a sample of 44 SNe Ia was selected which fulfill the additional conditions: (1) having occurred after 1985 to ensure photometric quality, and (2) $v_{220} > 1200\ \mathrm{km\,s^{-1}}$ (after correction for Virgocentric infall).

The color cut in $(B - V)$ is justified in Fig. 1 where *all* SNe Ia after 1985 are plotted. Their color distribution is sharply peaked. Some of the 10 red objects with $(B-V) > 0.20$

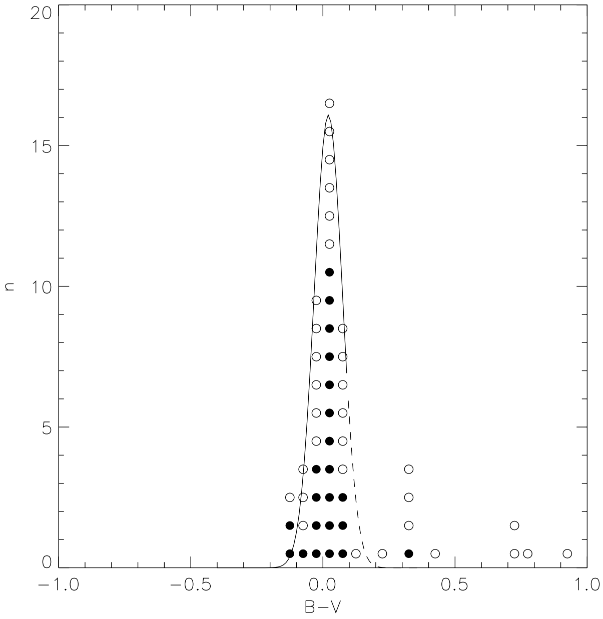


FIGURE 1. The color distribution of all known SNe Ia after 1985 with $v < 30\,000\ \mathrm{km\,s^{-1}}$. Open symbols are for more distant SNe Ia. The binned intervals embrace $\Delta(B{-}V) = 0\overset{\mathrm{m}}{.}05$. A Gaussian fit to all SNe Ia with $(B - V) \leq 0\overset{\mathrm{m}}{.}10$ gives $<B - V>= 0.020, \sigma_{\mathrm{B-V}} = 0\overset{\mathrm{m}}{.}053$.

probably have high internal reddening, others are known to have low expansion velocities (like SN 1986G) or quite peculiar spectra (like SN 1991bg, 1992K). The latter may define special subclasses of SNe Ia and clearly should be excluded from a homogeneous set of SNe Ia. Their exclusion is in any case indicated as long as none of the calibrating SNe Ia is of this type, which is not the case (Section 2.4).

Seven SNe Ia of the blue sample have $0.06 < (B - V) < 0.20$. All seven are underluminous and five of them lie in the inner parts of spiral galaxies. They are therefore suspected to suffer mild absorption in their parent galaxies. Their exclusion is permissible because none of the calibrating SNe Ia is as red.

Two SNe Ia of the blue sample, i.e. SN 1991T (Phillips et al. 1992) and SN 1995ac (Garnavich et al. 1996), had peculiar spectra during their early phases. SN 1991T has long been suspected to be overluminous, but a recent Cepheid distance (Saha et al. 2001) suggests this overluminosity to be only marginal. SN 1995ac lies, however, significantly above the Hubble line and is certainly overluminous. Both objects are excluded here.

The remaining 35 blue SNe Ia constitute our "fiducial sample." They are listed in Parodi et al. (2000). Their spectra, as far as available, are Branch-normal. For 29 SNe Ia also $m_I(\max)$ is known.

2.2. *The color of blue SNe Ia*

The mean color of 16 SNe Ia of the fiducial sample, that have occurred in early-type galaxies, is $<B - V>= -0.013 \pm 0.015$ and is identical with the mean color of 9 *outlying* SNe Ia in spirals. We therefore take this as the *true* mean intrinsic color of SNe Ia. The mean intrinsic color derived by Phillips et al. (1999) from different assumptions is bluer by ~ 0.03. This is irrelevant for the following discussion as long as the calibrating SNe Ia and the remaining objects of the fiducial sample conform with the adopted zeropoint color. (If our adopted mean color was too red, any additional absorption would equally affect the calibrators and the fiducial sample and have no effect on the derived distances). Indeed, after the eight calibrating SNe Ia (Section 2.4) are individually corrected for absorption they exhibit a mean color of $<B-V>= -0.009 \pm 0.015$, i.e. only insignificantly different from our adopted mean intrinsic color. Also the 10 remaining SNe Ia of the fiducial sample, which lie in the inner parts of spirals or whose position within their parent spirals is poorly known, are redder by a negligible amount of only $0^{\rm m}_{.}012 \pm 0^{\rm m}_{.}016$.

The scatter in $(V-I)$ is somewhat larger ($\sigma_{V-I} = 0^{\rm m}_{.}08$) than in $(B-V)$, but here again the mean color of $<V-I>= -0.276 \pm 0.016$ is the same for the SNe Ia in E/S0 galaxies and in spirals as well as for the calibrating SNe Ia.

2.3. *Second-parameter correction*

The 35 SNe Ia of the fiducial sample define very tight Hubble diagrams in B, V, and I for the range $1200 < v \lesssim 30\,000$ km s^{-1} (cf. Parodi et al. 2000; their Fig. 3), the scatter being only $\sigma_B = 0^{\rm m}_{.}21$, $\sigma_V = 0^{\rm m}_{.}18$, and $\sigma_I = 0^{\rm m}_{.}16$. This proves their usefulness as "standard candles."

Although the Hubble diagrams of SNe Ia in B, V, and I are tighter than for any other known objects, they still contain systematic effects. As suspected early on by some authors and pointed out again by Phillips (1993) the peak luminosity of SNe Ia correlates with the decline rate. Phillips introduced Δm_{15}, i.e. the decline in magnitudes during the first 15 days after B maximum, as a measure of the decline rate. Indeed the residuals of the SNe Ia of the fiducial sample correlate with $\Delta m15$ (Fig. 2a). The relation is roughly $\delta M \propto 0.5\Delta m_{15}$ in all three colors, slow decliners being brighter.

In addition the luminosities of SNe Ia correlate with the color $(B-V)$ (Tammann 1982; Tripp 1998). The proportionality factor between the magnitude residual δM and $\Delta(B-V)$ decreases from ~2.6 in B to ~1.2 in I for the fiducial sample, blue SNe Ia being brighter (Fig. 2b). This dependence is significantly shallower than for absorption by standard dust. There are indeed strong reasons (Saha et al. 1999) to believe that the dependence of luminosity on color is an intrinsic effect of SNe Ia.

The "second parameters" Δm_{15} and $(B-V)$ are orthogonal to each other. A simultaneous fit of the residuals δM in function of Δm_{15} *and* $(B-V)$ is therefore indicated. In that case the fiducial sample yields (Parodi et al. 2000)

$$\delta M_B^{\rm corr} = 0.44_{\pm 0.13}\,(\Delta m_{15} - 1.2) + 2.46_{\pm 0.46}\,[(B-V) + 0.01] - 28.40_{\pm 0.16},\ \sigma_B = 0.129 \quad (2.1)$$

$$\delta M_V^{\rm corr} = 0.47_{\pm 0.11}\,(\Delta m_{15} - 1.2) + 1.39_{\pm 0.40}\,[(B-V) + 0.01] - 28.39_{\pm 0.14},\ \sigma_V = 0.129 \quad (2.2)$$

$$\delta M_I^{\rm corr} = 0.40_{\pm 0.13}\,(\Delta m_{15} - 1.2) + 1.21_{\pm 0.43}\,[(B-V) + 0.01] - 28.11_{\pm 0.17},\ \sigma_I = 0.122 \quad (2.3)$$

It is interesting to note that the residuals δM correlate also with the Hubble type (SNe Ia in early-type galaxies being fainter) and marginally so with the radial distance from the galaxy center, but that these dependencies disappear once the apparent magnitudes are corrected for Δm_{15} and $(B-V)$.

The apparent magnitudes $m_{B,V,I}^{\rm corr}$ of the 35 SNe Ia, corrected for Δm_{15} and color by means of equations (2.1)–(2.3), define Hubble diagrams as shown in Fig. 3. *Their tightness is astounding.* The Cerro Tololo collaboration, to whom one owes 70% of the photometry of the fiducial sample, quote a mean observational error of their $m_{\rm max}$-values of $\sim 0^{\rm m}_{.}10$ and of their colors $(B-V)$ of $\sim 0^{\rm m}_{.}05$. This alone would suffice to explain the observed scatter of $\sigma_m = 0^{\rm m}_{.}12 - 0^{\rm m}_{.}13$. An additional error source are the corrections for Galactic absorption which were adopted from Schlegel et al. (1998). In fact, if one excludes the nine SNe Ia with large Galactic absorption corrections ($A_V > 0^{\rm m}_{.}2$) the scatter decreases to $0^{\rm m}_{.}11$ in all three colors. Two important conclusions follow from this. (1) If the total observed scatter of the Hubble diagrams is read vertically as an effect of peculiar motions,

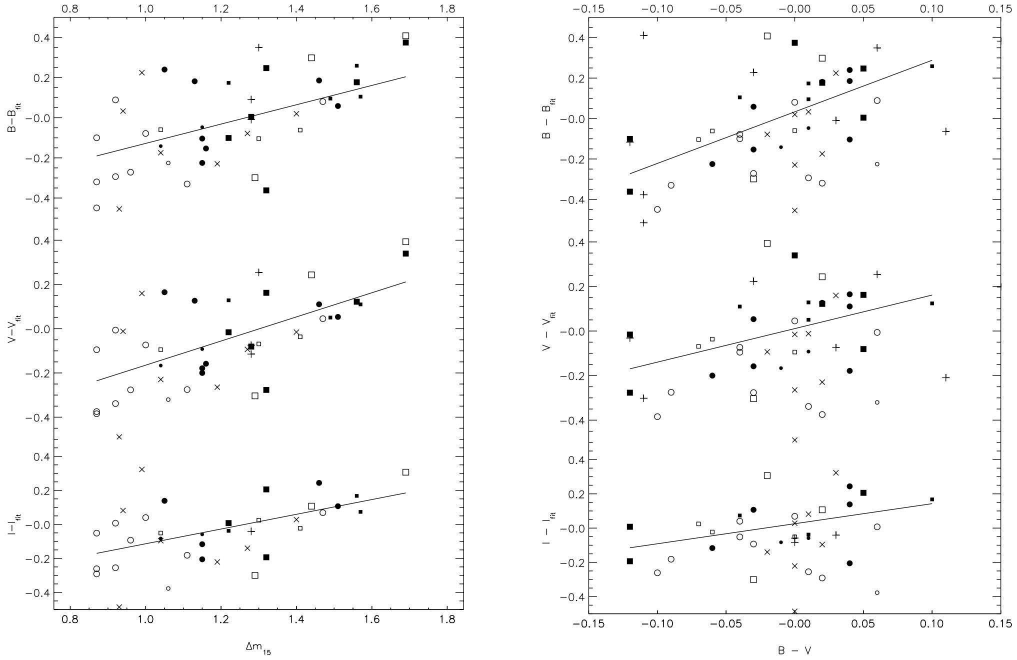


FIGURE 2. *Left panel:* Relative magnitudes (i.e. residuals from the Hubble line) for the SNe Ia of the fiducial sample in function of the decline rate Δm_{15}. Circles are SNe Ia in spirals, squares in E/S0 galaxies. Open symbols are SNe Ia with $1200 < v < 10\,000\,\mathrm{km\,s^{-1}}$, closed symbols are for more distant SNe Ia. Small symbols are SNe Ia whose observations begin eight days after B maximum or later. Neither the SNe Ia before 1985 with known Δm_{15} (shown as crosses) nor the seven blue, but reddened SNe Ia (shown as 'X's) are considered for the weighted least-squares solutions (solid lines). *Right panel:* Relative absolute magnitudes (i.e. residuals from the Hubble line) for the SNe Ia of the fiducial sample in function of their color $(B-V)$. Symbols as in the left panel.

a generous upper limit is set of $\Delta v/v = 0.05$, which holds for the range of $3500 \lesssim v \lesssim 30\,000\,\mathrm{km\,s^{-1}}$. The (all-sky) distance-dependent variation of $\mathrm{H_0}$ must be even smaller. (2) If on the other hand the scatter is read horizontally and if allowance is made for the observational errors of the apparent magnitudes (and for any peculiar motions) one must conclude that the luminosity scatter of blue SNe Ia, once they are homogenized in Δm_{15} and color, is smaller than can be measured at present. With other words, they are extremely powerful standard candles.

It is obvious that if one can determine the absolute magnitude of a few (nearby) SNe Ia this offers what we believe to be a definitive route to determine the large-scale value of $\mathrm{H_0}$.

2.4. *The luminosity calibration of SNe Ia*

As stated before, *HST* has provided Cepheid distances to nine galaxies which have produced 10 SNe Ia. Excluding the spectroscopically peculiar SN 1991T in NGC 4527 (Saha et al. 2001) leaves eight galaxies with nine SNe Ia. They are listed in Table 2.

The weights of the individual values of M in Table 2 are quite different due to the different quality of the SNe Ia light curves and of the Cepheid distances, but they are also relatively strongly affected by the corrections for internal absorption. The estimated individual errors are compounded and carried on to determine the total weights. This procedure is much safer than to exclude single SNe Ia for the one or the other reason.

SN (1)	Galaxy (2)	$(m\text{-}M)^0$ (3)	ref. (4)	M_B^0 (5)	M_V^0 (6)	M_I^0 (7)	Δm_{15} (8)
1895 B	NGC 5253	28.01 (08)	2	−19.54 (22)	...	...	...
1937 C	IC 4182	28.36 (09)	1	−19.56 (15)	−19.54 (17)	...	0.87 (10)
1960 F	NGC 4496A	31.04 (10)	3	−19.56 (18)	−19.62 (22)	...	1.06 (12)
1972 E	NGC 5253	28.61 (08)	2	−19.64 (16)	−19.61 (17)	−19.27 (20)	0.87 (10)
1974 G	NGC 4414	31.46 (17)	4	−19.67 (34)	−19.69 (27)	...	1.11 (06)
1981 B	NGC 4536	31.10 (05)	5	−19.50 (14)	−19.50 (10)	...	1.10 (07)
1989 B	NGC 3627	30.22 (12)	6	−19.47 (18)	−19.42 (16)	−19.21 (14)	1.31 (07)
1990 N	NGC 4639	32.03 (22)	7	−19.39 (26)	−19.41 (24)	−19.14 (23)	1.05 (05)
1998 bu	NGC 3368	30.37 (16)	8	−19.76 (31)	−19.69 (26)	−19.43 (21)	1.08 (05)
mean (straight, excl. SN 1895 B)				−19.57 (04)	−19.56 (04)	−19.26 (06)	1.06 (05)
mean (weighted,excl. SN 1895 B)				−19.55 (07)	−19.53 (06)	−19.25 (09)	1.08 (02)
$M^{\mathrm{corr}}(\Delta m_{15} = 1.2; (B-V) = -0.01)$				−19.48 (07)	−19.47 (06)	−19.19 (09)	1.08 (02)

References—(1) Saha et al. 1994 (2) Saha et al. 1995 (3) Saha et al. 1996b (4) Turner et al. 1998 (5) Saha et al. 1996a (6) Saha et al. 1999 (7) Saha et al. 1997 (8) Tanvir et al. 1995.

TABLE 2. Absolute B, V, and I magnitudes of blue SNe Ia calibrated through Cepheid distances of their parent galaxies

The adopted mean absolute magnitudes in Table 2 agree fortuitously well with explosion models, ejecting about $0.6\,\mathfrak{M}_\odot$ of ^{56}Ni, by Höflich & Khokhlov (1996) for equally blue SNe Ia (cf. also Branch 1998).

The *HST* photometry for the Cepheids in the galaxies listed in Table 2 (except NGC 4414) has been re-analyzed by Gibson et al. (2000). For 114 Cepheids in common with Saha et al. (1994, 1995, 1996a,b, 1997, 1999) they find a brighter photometric zeropoint by $0^{\mathrm{m}}.04 \pm 0^{\mathrm{m}}.02$, which is as satisfactory as can be expected from the subtle photometry with WFPC2. On the other hand a recent check on the zeropoint by A. Saha suggests that it should become fainter by $0^{\mathrm{m}}.02$.

Gibson et al. (2000) have added Cepheids which were reduced only with the photometric ALLFRAME package. Many of the additional Cepheids had also been detected by us, but were discarded because of what we considered to be insuperable problems such as poor light curves or excess crowding. With their additional Cepheids Gibson et al. (2000) have derived a distance modulus of NGC 5253 that is $0^{\mathrm{m}}.39$ smaller than listed in Table 2. Their reduction is unlikely for us because it would imply a very faint tip of the red-giant branch. For the remaining galaxies in Table 2 they suggest a mean decrease of the distance moduli by $0^{\mathrm{m}}.11 \pm 0^{\mathrm{m}}.03$. This would lead to an increase of H_0 by 5–6%. We do not consider this possibility pending independent confirmation of the ALLFRAME Cepheids.

2.5. *The value of H_0*

Combining the weighted mean absolute magnitude M_{BVI}^0 of SNe Ia from Table 2 with the observed Hubble diagrams in B, V, and I of the fiducial sample, corrected only for Galactic absorption, leads immediately to a mean value of $\mathrm{H}_0(BVI) = 58.3 \pm 2.0$ (internal error) (cf. Parodi et al. 2000). The three colors B, V, and I give closely the same results.

However, the calibrating SNe Ia lie necessarily in late-type galaxies (because the parent galaxies must contain Cepheids), and these SNe Ia are therefore expected to be somewhat more luminous than their counterparts of the fiducial sample which lie in galaxies of all Hubble types (cf. Section 2.3). Because of the correlation between Hubble type and decline rate Δm_{15}, the calibrators should have also slower decline rates than average.

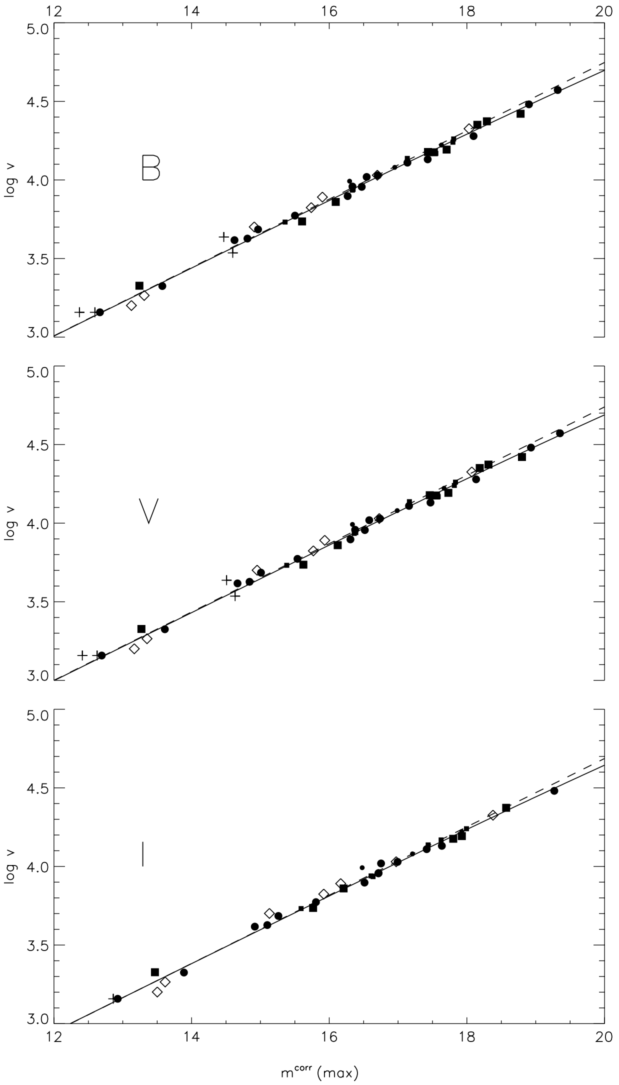


FIGURE 3. The Hubble diagrams in B, V, (and I) for the 35 (29) SNe Ia of the fiducial sample with magnitudes $m^{\rm corr}$, i.e. corrected for decline rate Δm_{15} and color $(B{-}V)$ (equations 2.1–2.3). Circles are SNe Ia in spirals, squares in E/S0 galaxies. Small symbols are SNe Ia whose observations begin eight days after B maximum or later. Solid lines are fits to the data assuming a flat universe with $\Omega_{\rm M} = 0.3$ and $\Omega_\Lambda = 0.7$; dashed lines are linear fits with a forced slope of 0.2 (corresponding approximately to $\Omega_{\rm M} = 1.0$ and $\Omega_\Lambda = 0.0$). Not considered for the fits are the SNe Ia before 1985 and the seven SNe Ia with $0.06 < (B - V) < 0.20$ that are suspected to be reddened. They are shown as diamonds after absorption correction; their inclusion would have nil effect on the fit.

This is indeed the case. Consequently the calibrators and the fiducial sample should be homogenized as to Δm_{15}, i.e. the corrected magnitudes $M^{\rm corr}_{BVI}$ from Table 2 should be compared with the corrected Hubble diagram in Fig. 3. It may be noted that the correction for variations in color $(B - V)$ has here no net effect because the calibrators and the fiducial sample have identical mean colors.

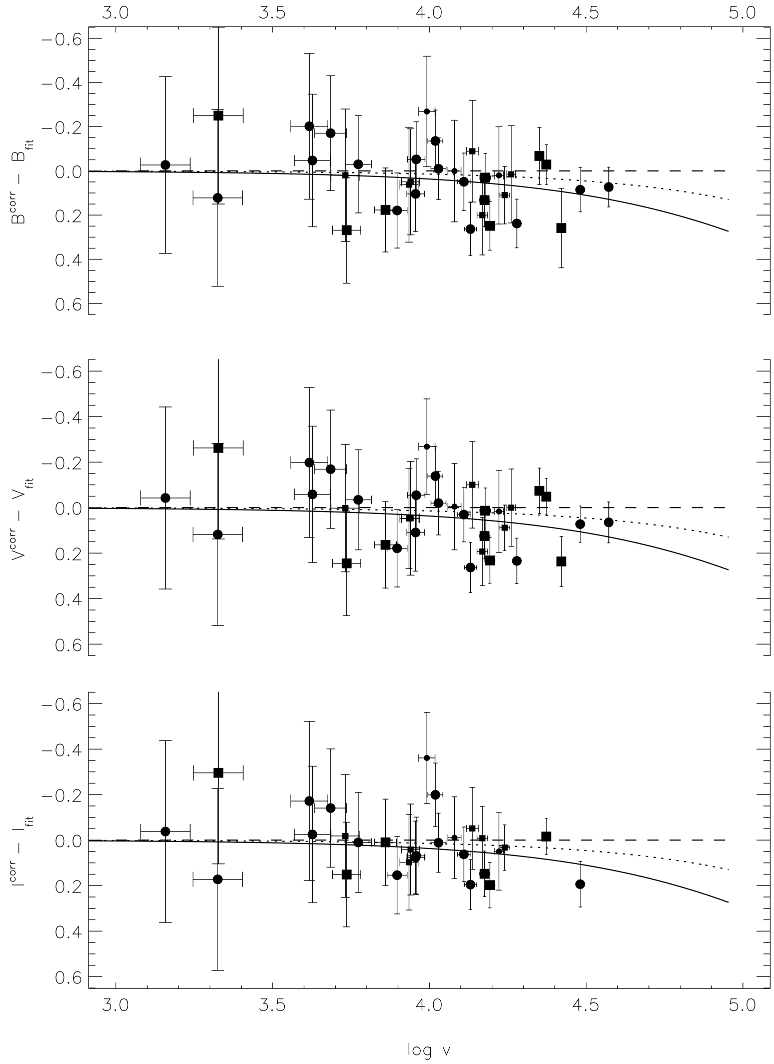


FIGURE 4. Differential Hubble diagrams ($m^{\rm corr} - m_{\rm fit}$) vs. $\log v$ in B, V, (and I) for the 35 (29) SNe Ia of the fiducial sample. Symbols as in Fig. 3. The dashed line is for a flat cosmological model with $\Omega_{\rm M} = 1.0$ and $\Omega_\Lambda = 0.0$; the theoretical apparent magnitudes $m_{\rm fit}$ correspond to this model. The full line is for a flat model with $\Omega_{\rm M} = 0.3$ and $\Omega_\Lambda = 0.7$; the dotted line is for an open universe with $\Omega_{\rm M} = 0.2$ and $\Omega_\Lambda = 0.0$.

An exact comparison should allow for the fact that cosmological effects on the Hubble diagram are non-negligible at $\sim$30 000 km s^{-1}. Three different model universes are therefore fitted to the data as illustrated in a differential Hubble diagram (Fig. 4):

1. A flat Universe with $\Omega_M = 1.0$ ($q_0 = 0.5$; Sandage 1961, 1962). When the magnitudes $m^{\rm corr}_{BVI}$ of the fiducial sample are fitted to the corresponding Hubble line one obtains, after inserting $<M^{\rm corr}_{\rm BVI}>$ of the calibrators, $\mathrm{H}_0(B) = 60.2 \pm 2.1$. The values in V and I are very similar (60.1 and 60.0, respectively).

2. An open Universe with $\Omega_M = 0.2$ ($q_0 = 0.1$; Sandage 1961, 1962). The fit is in this case somewhat better, giving $\mathrm{H}_0(BVI) = 60.2 \pm 2.0$.

3. A flat Universe with $\Omega_M = 0.3$, $\Omega_\Lambda = 0.7$ ($q_0 = -0.55$). This model is favored by *high-redshift* SNe Ia (Perlmutter 1998, Riess et al. 1998, Schmidt et al. 1998, Perlmutter et al. 1999). The Hubble line is given in this case by (cf. Carroll, Press, & Turner 1992)

$$m_{B,V,I} = 5\log\left(\frac{c}{\mathrm{H}_0}(1+z_1)\int_0^{z_1}[(1+z)^2(1+\Omega_M z) - z(2+z)\Omega_\Lambda]^{-1/2}dz\right) + M_{B,V,I} + 25\,. \tag{2.4}$$

This model gives the best fit to the fiducial sample and yields, after insertion of $<M_{\mathrm{BVI}}^{\mathrm{corr}}>$ from Table 2, $\mathrm{H}_0(B) = 61.0 \pm 2.1$, $\mathrm{H}_0(V) = 60.9 \pm 1.8$, and $\mathrm{H}_0(I) = 60.7 \pm 2.6$. A mean value of $\mathrm{H}_0 = 60.9 \pm 1.8$ is adopted for the following discussion.

The three model Universes can be distinguished with the present fiducial sample only at the 1σ-level. But the possibility to determine Λ from rather local SNe Ia is interesting in principle. The advantage would be good multi-color photometry (e.g. in B, V, I), allowing homogenization of the SNe Ia in color and good control of internal absorption, quite small K-corrections, and short look-back times minimizing evolutionary effects. From 100–200 SNe Ia with good observations and $z < 0.12$ one should obtain a significant value of Λ.

2.6. *Discussion*

It should be noted that the second-parameter corrections increase H_0 by only 4.3%, i.e. from 58.3 to 60.9. This is supported by the fact that if one restricts the discussion to SNe Ia in spirals, minimizing in this way the difference between the parent galaxies of the calibrators and of the SNe Ia of the fiducial sample, one obtains $\mathrm{H}_0(BVI) = 59.1$ independent of any second-parameter correction. Alternatively, if one considers the 21 SNe Ia of the fiducial sample with $\Delta m_{15} \geq 1.3$, whose mean decline rate of $<\Delta m_{15}>= 1.08 \pm 0.02$ is the same as that of the calibrators, one obtains $\mathrm{H}_0(BVI) = 59.2$ (Parodi et al. 2000).

Other authors have derived from part of the fiducial sample and from all or some of the calibrators in Table 2 values from $\mathrm{H}_0 = 50 \pm 3$ (Lanoix 1998) to $\mathrm{H}_0 = 72 \pm 4$ (Richtler & Drenkhahn 1999). Intermediate values of $\mathrm{H}_0 = 63 - 64 \pm 2.2$ were found by Suntzeff et al. (1999) and Phillips et al. (1999), however they base their result on a significantly steeper Δm_{15}-luminosity correction than found here. This would lead to an overcorrection of the larger fiducial sample used here. If their correction was applied to the present data, one would obtain $\mathrm{H}_0 = 59.7$ for the SNe Ia with $\Delta m_{15} < 1.2$ (n =17), and $\mathrm{H}_0 = 64.6$ for those with $\Delta m_{15} \geq 1.2$ (n =18). Since seven of the eight calibrators fall into the first category, the lower value must be more nearly correct.

By relying exclusively on the Cepheid distances by Gibson et al. (2000; cf. Section 2.4), Freedman (2001) was able to push the solution of Suntzeff et al. (1999) and Phillips et al. (1999) up to $\mathrm{H}_0 = 68$.

Riess et al. (1998) employed a so-called “Multi Light Curve Shape” (MLCS) method to correct simultaneously for Galactic absorption and the relative SN Ia luminosity. The resulting distance moduli (their Table 10) imply, however, that H_0 depends on their correction parameter Δ, i.e. $<\mathrm{H}_0>= 66.6 \pm 1.1$ for $\Delta < -0.20$ (n =9) and $<\mathrm{H}_0>= 61.8 \pm 1.3$ for $\Delta > -0.20$ (n =18). Jha et al. (1999) employed also the MLCS method to derive $\mathrm{H}_0 = 64 \pm 7$ without listing Δ-values of individual SNe Ia. Tripp & Branch (1999), correcting for Δm_{15} and color $(B - V)$, have obtained $\mathrm{H}_0 = 62\,(\pm 4)$.

The present result of $\mathrm{H}_0 = 60.2 \pm 2.1$ is still affected by external errors. Yet the largest systematic error source of all distance indicators that depend on an adopted mean luminosity (or size), i.e. selection effects against underluminous (or undersized) “twins,” is negligible in the case of blue SNe Ia because of their exceptionally small luminosity dispersion.

External errors of either sign are introduced (cf. Parodi et al. 2000) by the photometric zeropoint of the WFPC-2 photometry ($0\overset{\mathrm{m}}{.}04$), by the adopted slope of the Δm_{15}-luminosity relation ($0\overset{\mathrm{m}}{.}02$), by the velocity correction for Virgocentric infall and by the correction for motion relative to the CMB ($0\overset{\mathrm{m}}{.}02$).

Other errors are asymmetric with a tendency to underestimate distances. The adopted LMC modulus of 18.50 for the zeropoint of the Cepheid PL relation is probably too small by $\sim 0\overset{\mathrm{m}}{.}06$ (e.g. Federspiel, Tammann, & Sandage 1998; Madore & Freedman 1998; Feast 1999; Walker 1999; Gilmozzi & Panagia 1999; Gratton 2000; Sakai, Zwitsky, & Kennicutt 2000). Smaller LMC moduli suggested on the basis of statistical parallaxes of RR Lyrae stars and red giant clump stars depend entirely on the sample selection and on the absence of metallicity and evolutionary effects, respectively. The higher LMC modulus will increase all moduli by $0\overset{\mathrm{m}}{.}06 \pm 0\overset{\mathrm{m}}{.}10$. Incomplete Cepheid sampling near the photometric threshold always tends to yield too short distances (Sandage 1988a; Lanoix, Paturel, & Garnier 1999; Mazumdar & Narasimha 2000). The effect is estimated here to be $0\overset{\mathrm{m}}{.}05 \pm 0\overset{\mathrm{m}}{.}05$. Stanek & Udalski (1999) have proposed that photometric blends of Cepheids in very crowded fields lead to a serious underestimate of the distances. Careful analyses by Saha et al. (2000) and Ferrarese et al. (2000) show the effect to be more modest for the Cepheid distances in Table 2, say $0\overset{\mathrm{m}}{.}03 \pm 0\overset{\mathrm{m}}{.}03$ on average. Absorption corrections of the calibrating SNe Ia and those of the fiducial sample, having identical colors after correction, enter only differentially. However, if one excludes the nine SNe Ia with large Galactic absorption (Section 2.3) the Hubble line of Fig. 3 shifts faintwards by $0\overset{\mathrm{m}}{.}05$ in B. Finally seven SNe Ia were excluded on the suspicion of having some internal absorption (Section 2.1). If they had been included after being corrected for absorption, their effect on the Hubble line would be negligible. If, however, their colors are intrinsic they would shift the Hubble line by $0\overset{\mathrm{m}}{.}02$ towards fainter magnitudes. Combining the absorption errors it is estimated that the distances of the fiducial sample are too small by $0\overset{\mathrm{m}}{.}05 \pm 0\overset{\mathrm{m}}{.}05$ on average. Finally there is much discussion of the metallicity effect on Cepheid distances. Theoretical investigations show this effect to be nearly negligible (Sandage, Bell, & Tripicco 1999; Alibert et al. 1999) Other authors do not even agree on the sign of the correction. Based on Kennicutt et al. (1998) Gibson et al. (2000) have concluded that the Cepheid distances in Table 4 are too small by $0\overset{\mathrm{m}}{.}07$ due to variations of the metallicity. A distance increase of $0\overset{\mathrm{m}}{.}04 \pm 0\overset{\mathrm{m}}{.}10$ is adopted as a compromise.

Adding the various error sources in quadrature leads to a correction factor of 0.96 ± 0.08 which is to be applied to $\mathrm{H}_0 = 60.2 \pm 2.1$. At a 90-percent confidence level one obtains then

$$\mathrm{H}_0 = 58.5 \pm 6.3. \tag{2.5}$$

If only the Cepheid distances of Gibson et al. (2000; cf. Section 2.4) had been used, excluding their unlikely value for NGC 5253, one would have obtained $\mathrm{H}_0 = 61.6 \pm 6.6$.

The standards for the determination of H_0 are now set by SNe Ia. In the next Section it will be asked to what extend these standards can be met by other distance indicators, i.e. how reliable are they as relative distance indicators and how accurate is their zeropoint calibration?

3. $\mathbf{H_0}$ from Cluster Distances

3.1. *The Hubble diagram of clusters*

3.1.1. *Brightest cluster members*

A deep (to $z = 0.45$) and tight ($\sigma_m = 0\overset{\mathrm{m}}{.}32$) Hubble Diagram is that of 1st-ranked E and S0 cluster galaxies corrected for Galactic absorption, aperture effect, K-dimming, Bautz-Morgan effect, and cluster richness (Sandage & Hardy 1973). They define a Hubble line in the range of $1200 < v < 30\,000\ \mathrm{km\,s^{-1}}$ of

$$\log v = 0.2\, m_{V_c} + (1.359 \pm 0.018); \quad \sigma_{m_V} = 0.32; \quad n = 76 \tag{3.1}$$

which is easily transformed into

$$\log \mathrm{H}_0 = 0.2\, M_V(\mathrm{1st}) + (6.359 \pm 0.018)\ . \tag{3.2}$$

Weedman (1976) has established a Hubble diagram using the mean magnitude of the 10 brightest cluster members. The small scatter of $\sigma_{m_{10}} = 0\overset{\mathrm{m}}{.}15$ is not directly comparable with other values because the magnitudes are defined within a *metric* diameter and depend somewhat on redshift.

Another Hubble diagram of 1st-ranked galaxies has been presented by Lauer & Postman (1992). It comprises the complete sample of 114 Abell clusters with $v < 15\,000\ \mathrm{km\,s^{-1}}$. The scatter about the mean Hubble line amounts to $\sigma_m = 0\overset{\mathrm{m}}{.}3$.

3.1.2. *The Tully-Fisher (TF) relation*

Dale et al. (1999) have derived *relative I*-band TF distances of 52 clusters from an average of 8–9 members per cluster. The *mean* cluster distances define a Hubble line with a scatter of only $\sigma_{(m-M)} = 0\overset{\mathrm{m}}{.}12$, i.e. similar to SNe Ia. The clusters are corrected for *differential* selection bias, but the data are still unsuitable to derive absolute distances because with only a few members per cluster the Teerikorpi cluster incompleteness bias must be severe (Section 3.4). The same holds for an older sample of 18 clusters by Giovanelli et al. (1997), for which Sakai et al. (2000) have derived *I*-band TF distances. Sixteen of their clusters with $1500 < v < 9000\ \mathrm{km\,s^{-1}}$ and each with 15 studied galaxies on average define a Hubble line with a scatter of $\sigma_{(m-M)} = 0\overset{\mathrm{m}}{.}18$, which implies a scatter of $\sigma > 0\overset{\mathrm{m}}{.}5$ of the TF modulus of an individual galaxy.

3.1.3. *The $D_n - \sigma$ relation*

Kelson et al. (2000) have derived $\mathrm{D}_n\!-\!\sigma$ and fundamental plane (FP) distances of (only) 11 clusters out to $v \sim 10\,000\ \mathrm{km\,s^{-1}}$. The two resulting sets define Hubble diagrams with a remarkably small scatter of $\sigma_{(m-M)} = 0\overset{\mathrm{m}}{.}19$.

3.1.4. *Combining different Hubble diagrams*

A combination of data of brightest cluster galaxies and $\mathrm{D}_n - \sigma$ measurements by Faber et al. (1989) have been used to derive cluster distances relative to the Virgo cluster (Jerjen & Tammann 1993). When these are augmented by the relative TF distances of clusters by Giovanelli (1997) one obtains a Hubble diagram as shown in Fig. 5. The data scatter by $\sigma_{(m-M)} = 0\overset{\mathrm{m}}{.}20$ about a Hubble line. The latter is found by linear regression and implies

$$\log \mathrm{H}_0 = -0.2\, (m - M)_{\mathrm{Virgo}} + (8.070 \pm 0.007) \tag{3.3}$$

(Federspiel, Tammann, & Sandage 1998). The equation is particularly useful because the calibration of H_0 can simply be accomplished by inserting the Virgo cluster modulus.

It should be noted that the observed scatter of the above Hubble diagrams is almost entirely due to intrinsic scatter of the distance indicators and to measurement errors because the contribution of peculiar velocities is $\lesssim 0\overset{\mathrm{m}}{.}1$ as shown by the Hubble diagram of SNe Ia (Fig. 3).

3.2. *Relative distances of local groups and clusters*

The reliability of other distance indicators, which have failed so far to establish a Hubble diagram, can be tested by comparing their relative distances of local groups or clusters,

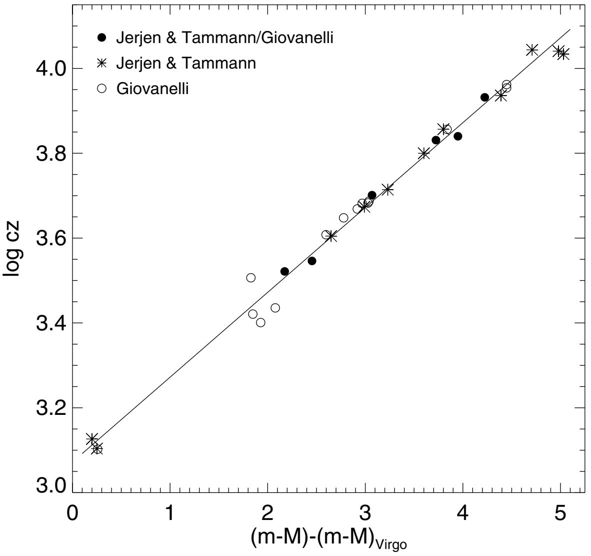


Figure 5. Hubble diagram of 31 clusters with known relative distances. Asterisks are data from Jerjen & Tammann (1993). Open circles are from Giovanelli (1997). Filled circles are the average of data from both sources.

Method (1)	$\Delta(m-M)$ Virgo–Leo I (2)	$\Delta(m-M)$ Fornax–Virgo (3)	$\Delta(m-M)$ Fornax–Leo I (4)	Source (5)
Cepheids, SNe Ia	1.37 ± 0.21	0.05 ± 0.22	1.42 ± 0.10	Tables 4 and 6
$\mathrm{D_n} - \sigma$	0.92 ± 0.32	0.17 ± 0.16	1.09 ± 0.32	Faber et al. 1989
FP	$\cdots$	(0.52 ± 0.17)	$\cdots$	Kelson et al. 2000
TF ($BVRI$)	$(1.95 \pm 0.23)^1$	(-0.40 ± 0.10)	$(1.55)^1$	Schröder 1995
vel. ratio^2	1.29 ± 0.31	0.27 ± 0.30	1.57 ± 0.18	Kraan-Korteweg 1986
mean:	1.25 ± 0.15	0.15 ± 0.12	1.43 ± 0.08	
PNe	0.86 ± 0.09	0.33 ± 0.09	1.19 ± 0.10	Ferrarese et al. 2000
SBF	0.88 ± 0.07	0.40 ± 0.05	1.28 ± 0.06	Ferrarese et al. 2000
	0.93 ± 0.05	0.37 ± 0.04	1.30 ± 0.05	Tonry et al. 2000
GCLF	1.62 ± 0.31	-0.35 ± 0.14	1.27 ± 0.31	Tammann & Sandage 1999
	1.69 ± 0.56	0.14 ± 0.07	1.83 ± 0.56	Ferrarese et al. 2000

[1] From only 3 spirals in the Leo I group (Federspiel 1999)

[2] Assuming $v_{220}(\text{Leo}) = 648$, $v_{220}(\text{Virgo}) = 1179$, $v_{220}(\text{Fornax}) = 1338\ \text{km}\,\text{s}^{-1}$ (Kraan-Korteweg 1986) and allowing for a peculiar velocity of Leo I of $100\ \text{km}\,\text{s}^{-1}$ and of Fornax of $200\ \text{km}\,\text{s}^{-1}$

Table 3. Relative distances of Leo I, Virgo, and Fornax from various distance indicators. (The errors shown are from the quoted sources)

specifically between the Leo I group, the Virgo cluster, and the Fornax cluster. It is a most gentle test because the mean distances from several member galaxies are compared and the problem of the zeropoint calibration does not enter.

In Table 3 the modulus differences Virgo–Leo I, Fornax–Virgo, and Fornax–Leo I are listed as derived from various distance indicators. The sources of these differences are also listed. The first four lines show the differences of the distance indicators which have passed the Hubble diagram test. The fifth line gives the relative moduli as derived from the mean group/cluster velocities. Most of the entries are in statistical agreement with the weighted means in the sixth line.

There are, however, four entries shown in parentheses, which deviate significantly from the adopted means. The result of the FP that the Fornax cluster is more distant by as much as $0\overset{m}{.}5$ than the Virgo cluster is very unlikely. The TF distance of the Leo I group carries little weight being based on only three galaxies, and the small TF distance of the Fornax cluster, i.e. $0\overset{m}{.}40$ nearer than the Virgo cluster, is simply impossible. A compilation (Tammann & Federspiel 1997) of the relative distance Fornax–Virgo, as determined by 30 different authors, suggests Fornax to be slighty more distant than Virgo in agreement with the adopted means in Table 3, but all attempts to determine the TF distance of Fornax have come out with suspiciously small values. The discrepancy is not sufficiently explained by the fact that even a complete sample of the Fornax cluster contains only eight full-size spirals suited for the TF method in addition to 19 small late-type galaxies with large scatter. The discrepancy remains hence enigmatic. The case shows the possibility that even tested distance indicators may fail in individual cases.

In the lower part of Table 3 the relative distances are shown by different authors from planetary nebulae (PNe), the surface brightness fluctuations (SBF) and of the globular cluster luminosity function (GCLF). The modulus differences of the PNe are rather small. In particular the value of Virgo–Leo I is $0\overset{m}{.}39 \pm 0\overset{m}{.}17$ smaller than adopted. This unsatisfactory result of the PNe does not come as a surprise. The method rests on the assumption that the luminosity of the brightest planetary shell, as seen in a bright emission line, is a fixed standard candle and does not depend on sample size (i.e. galaxy luminosity). This is not only against the expectation from any realistic luminosity function, but is also contradicted by observations (Bottinelli et al. 1991; Tammann 1993; Soffner et al. 1996). Moreover the shell luminosity is predicted to depend on metallicity and age (Méndez et al. 1993).

Also the SBFs give a small difference Virgo–Leo I and an almost certainly too large difference Fornax–Virgo. If the quoted small errors are taken at face value the method is incompatible with the adopted relative position Leo I–Virgo–Fornax. In any case the method should be given low weight until its real capabilities are proved beyond doubt.

The distance differences of the GCLFs listed in Table 3 have too large errors and differ too much between different authors, due to different sample selections, that any clear conclusion could be drawn. In the case of the Virgo cluster it yields a perfect distance determination (cf. Table 5); in other cases it gives quite erratic results (Tammann & Sandage 1999).

3.3. *The distance of the Virgo cluster and of other clusters*

Four galaxies with known Cepheid distances lie in the Virgo cluster proper, i.e. within the isopleths and the X-ray contours (cf. Binggeli, Popescu, & Tammann 1993). As can be seen in Table 4 they have widely different distances. Three of the galaxies have been selected from the atlas of Sandage & Bedke (1988) on grounds of their exceptionally good resolution; they are therefore *expected* to lie on the near side of the cluster. The fourth galaxy, NGC 4639, which has a *low* recession velocity and can therefore not be assigned to the background, but must be a dynamical member of the cluster, is more distant by almost $1\overset{m}{.}0$ and must lie on the far side of the extended cluster. The relative position of the four galaxies is fully confirmed by their TF distances. The cluster *center* must lie somewhere between the available Cepheid distances, say at $(m - M) = 31.5 \pm 0.3$. Much of the confusion of the extragalactic distance scale comes from the ill-conceived notion that the three highly resolved Virgo galaxies could reflect the *mean* distance of the cluster.

Virgo				**Fornax**			
Object (1)	$(m-M)^0$ (2)	Ref. (3)	Remarks (4)	Object (5)	$(m-M)^0$ (6)	Ref. (7)	Remarks (8)
			Cepheids				
NGC 4321	31.04	1	highly resolved	NGC 1326A	31.49	7	
NGC 4535	31.10	2	highly resolved	NGC 1365	31.39	8	
NGC 4548	31.04	3	highly resolved	NGC 1425	31.81	9	
. . .							
NGC 4639	32.03(!)	4	normally resolved				
mean:	~31.5			31.56 ± 0.13			
			SNe Ia				
SN 1984A	31.42	5	in NGC 4419	SN 1980N	31.76	5	in NGC 1316
SN 1990N	32.12	5	in NGC 4639	SN 1981D	31.51	5	in NGC 1316
SN 1994D	31.27	5	in NGC 4526	SN 1992A	31.84	5	in NGC 1380
mean:	31.60 ± 0.30			31.70 ± 0.10			
			Tully-Fisher Relation *(complete samples)*				
mean (n=49):	31.65 ± 0.25	6			(31.25 ± 0.10)	10	
overall mean:	31.60 ± 0.20			overall mean:	31.65 ± 0.08		

References—(1) Ferrarese et al. 1996 (2) Macri et al. 1999 (3) Graham et al. 1999 (4) Saha et al. 1997 (5) see text (6) Federspiel et al. 1998; Federspiel 1999 (7) Prosser et al. 1999 (8) Silbermann et al. 1999 (9) Mould et al. 2000 (10) Schröder 1995; Federspiel 1999

TABLE 4. Distances of the Virgo and Fornax clusters

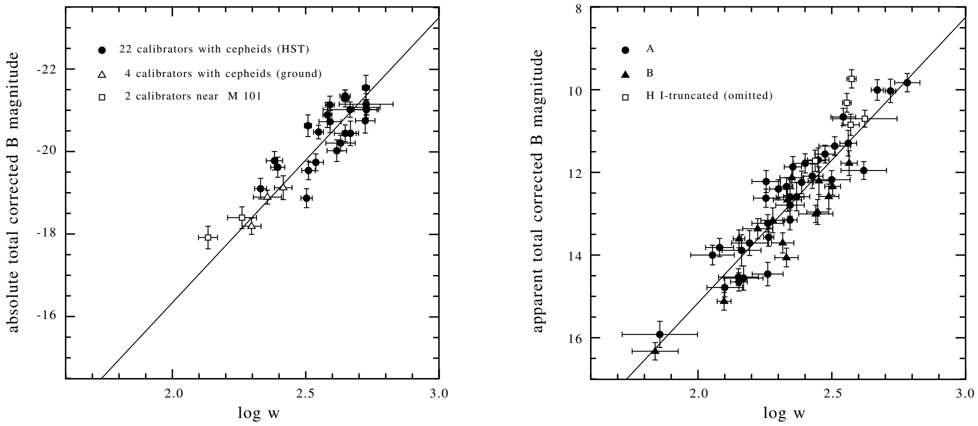


FIGURE 6. *left panel*: Tully-Fisher relation of 28 galaxies with independently known (Cepheid) distances. *right panel*: Tully-Fisher relation for a complete sample of 49 Virgo cluster spirals.

Three well observed SNe Ia have appeared in Virgo cluster members. Their distances in Table 4, corrected for decline rate Δm_{15} and color $(B-V)$, have been calculated from their individual parameters as compiled by Parodi et al. (2000) and from the calibration in Table 2.

The best possible application of the TF method is provided by the Virgo cluster, because a very deep catalog of the cluster (Binggeli, Sandage, & Tammann 1985) allows selection of a *complete* sample of all 49 sufficiently inclined cluster spirals. They define a reliable position and slope of the TF relation in B (Fig. 6b). (It is sometimes argued that I-magnitudes [albeit incomplete!] should be used to minimize the internal-absorption correction, but the advantage is offset by the steeper slope of the TF relation at long wavelengths [Schröder 1995]). Combining these data with the excellent calibration of the TF relation (Fig. 6a), which rests now on 26 galaxies with Cepheid distances and 2 companions of M 101, for which a Cepheid distance is available, yields the TF modulus shown in Table 4.

The distance of the Virgo cluster is such an important milestone for the extragalactic distance scale that the value adopted in Table 4 should be compared with other distance indicators. Three distance indicators of early-type galaxies shall be considered, although their reliability is much less tested. (1) The peak of the luminosity function of globular clusters (LFGC) is interesting because its calibration of the zeropoint rests on the Cepheid distance of M 31 *and* in excellent agreement on the RR Lyrae distance of Galactic globular clusters. (2) Six known novae in Virgo cluster ellipticals can be compared to the novae in M 31 whose *apparent* distance modulus can be derived from Cepheids *or* from Galactic novae (Capaccioli et al. 1989). Alternatively the semi-theoretical zeropoint of novae can be taken from Livio (1997). (3) The $D_{\rm n} - \sigma$ relation can be applied to E/S0 galaxies and to the bulges of S0–Sb galaxies. The mean of the two applications is given here. The zeropoint of the S0–Sb bulges is better determined than that of E/S0 galaxies, because the zeropoint of the latter depends on the assumption that the only two early-type galaxies of the Leo I group lie at the same distance as its spiral members. The results of the three methods are compiled in Table 5 and discussed in more detail by Tammann, Sandage, & Reindl (2000). The resulting mean distance modulus is in excellent agreement with the adopted value in Table 4 and is consistent with the assumption that early-type galaxies and spirals of the Virgo cluster are at the same distance.

Method	$(m-M)_{\mathrm{Virgo}}$	Original Source
Globular Clusters	31.70 ± 0.30	Tammann & Sandage 1999
Novae	31.46 ± 0.40	Pritchet & van den Bergh 1987
$\mathrm{D_n} - \sigma$	31.70 ± 0.15	Dressler 1987, Faber et al. 1998
mean:	31.66 ± 0.17	

TABLE 5. Additional distance determinations of the Virgo cluster

Method (1)	Object (2)	$(m-M)^0$ (3)	Source (4)	Remarks (5)
Cepheids	NGC 3351	30.01 ± 0.15	Graham et al. 1997	
	NGC 3368	30.37 ± 0.16	Tanvir et al. 1995	
	NGC 3627	30.22 ± 0.12	Saha et al. 1999	
SNe Ia	1998bu	30.32 ± 0.15	$m_{BVI}^{\mathrm{corr}}(\max)$ + Table 2	in NGC 3368
mean:		30.23 ± 0.07		

TABLE 6. The distance of the Leo I group

Table 4 shows also the distance modulus of the Fornax cluster as derived from the three galaxies with Cepheids and three SNe Ia. The individual distances show much less scatter than in the case of the Virgo cluster; this is obviously a result of the smaller size and depth of the Fornax cluster.

For comparison with other distance indicators it is useful to have also the Cepheid and SNe Ia distance of the Leo I group. The relevant data are set out in Table 6. The adopted mean modulus is consistent with the GCLF distance (30.08 ± 0.29; Tammann & Sandage 1999) and the surface brightness fluctuation (SBF) distance (30.30 ± 0.06; Ferrarese et al. 2000).

Finally, it is noted that the distance of the Coma cluster relative to the Virgo cluster is well determined. The value $\Delta(m-M)_{\mathrm{Coma-Virgo}} = 3.71 \pm 0.08$ (Tammann & Sandage 1999) from brightest cluster galaxies and the TF and $\mathrm{D}_n - \sigma$ methods is quite uncontroversial. With the Virgo modulus in Table 4 one obtains $(m-M)_{\mathrm{Coma}} = 35.31 \pm 0.22$ in excellent agreement with the independently calibrated GCLF distance from *HST* (Baum et al. 1997; cf. Tammann & Sandage 1999).

3.4. *The Teerikorpi cluster incompleteness bias*

It is not the place here to discuss the Teerikorpi cluster incompleteness bias (Teerikorpi 1997 and references therein; Sandage, Tammann, & Federspiel 1995) in any detail, because it is well known that distance indicators with non-negligible internal scatter yield *too small distances* if applied to *incomplete* cluster samples.

For illustration the complete sample of 49 Virgo spirals as shown in Fig. 6b was cut at different apparent-magnitude limits, and each time the mean TF distance of the remaining sample was calculated. The mean distance of each sample increases monotonically with the depth of the cut. The result is shown in Fig. 7. The true, asymptotic distance is only reached if one samples $\sim 4\overset{\mathrm{m}}{.}0$ into the cluster.

The bias effect explains why Sakai et al. (2000) have derived too high a value of H_0 by directly applying the TF calibration in Fig 6a to 15 clusters of Giovanelli et al. (1997). The shallow cluster samples are far from being complete and are bound to yield too small cluster distances. In fact the authors have derived $<\mathrm{H}_0>= 86$ for the clusters whose distances rest on less than 10 members, while they have found $<\mathrm{H}_0>= 70$ for the

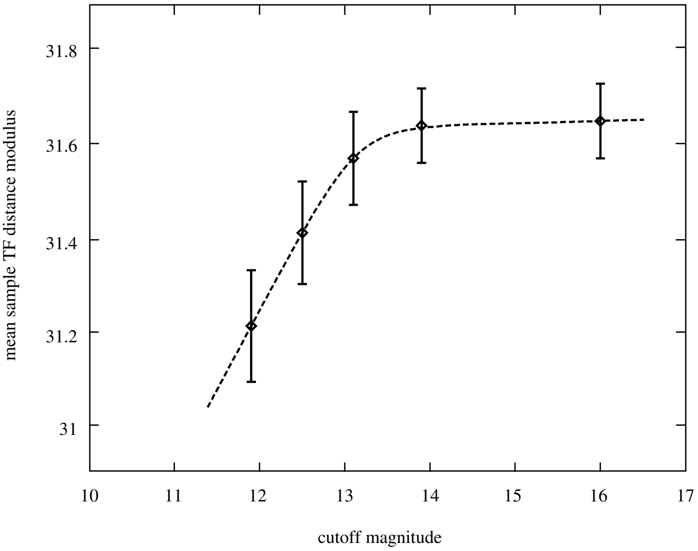


FIGURE 7. Illustration of the Teerikorpi cluster incompleteness bias. If the complete sample of 49 Virgo spirals is cut at levels brighter than $m_B \approx 13^{\mathrm{m}}.5$ one derives too small TF distances. A cut at $m_B = 12^{\mathrm{m}}.0$ introduces a distance modulus error of $0^{\mathrm{m}}.4$ (20 percent in distance).

Cluster	$m_{V_c}(1st)^1$	$(m-M)$	$M_V(1st)$
Virgo	8.21	31.60 ± 0.20	-23.39 ± 0.36
Fornax	8.83	31.65 ± 0.08	-22.82 ± 0.32
Coma	11.58	35.31 ± 0.22	-23.73 ± 0.39
mean:			-23.26 ± 0.20

[1] From Sandage & Hardy (1973). A scatter of $0^{\mathrm{m}}.32$ is allowed for.

TABLE 7. The absolute magnitude of 1st-ranked cluster galaxies.

clusters with 10 to 28 members. *It is clear that more nearly complete samples would give still lower values of* H_0.

3.5. *The value of* H_0 *from clusters*

The distances of the Virgo, Fornax, and Coma clusters in Section 3.3 yield the absolute magnitudes of three 1st-ranked cluster galaxies (Table 7). Inserting the mean value into equation (3.2) gives

$$\mathrm{H}_0 = 51 \pm 7\ . \tag{3.4}$$

Combining alternatively the cluster distances relative to the Virgo cluster as given by equation (3.3) (cf. Fig. 5) with the Virgo cluster modulus of 31.60 ± 0.20 from Table 4 gives

$$\mathrm{H}_0 = 56 \pm 6\ . \tag{3.5}$$

The result shows that once the Virgo cluster distance is fixed the value of H_0 has little leeway.

The relatively large error of the Virgo modulus is due to the important depth effect of the cluster. The four Cepheid and three SNe Ia distances do not suffice to sample the cluster in depth. The situation is aggravated by the fact that the Cepheid distances of three galaxies are biased because they were selected on the grounds of high resolution and hence must lie on the near side of the cluster. If one took unjustifiedly the mean distance of only these three Cepheids as the true cluster distance one would derive $\mathrm{H}_0 = 72$.

Fortunately the SNe Ia, the most reliable distance indicators known, confirm the Virgo cluster distance estimated from Cepheids *and* the important depth of the cluster. Moreover, the now very solid Cepheid calibration of the TF relation finds its most powerful application in the complete sample of Virgo spirals and corroborates the conclusions from Cepheids and SNe Ia.

Kelson et al. (2000) have derived $\mathrm{H}_0 = 75$ and 80 from the Hubble diagram based on $\mathrm{D}_n - \sigma$ and FP data (Section 3.1.3). They have chosen certain distances of Leo I, Virgo, and Fornax for the calibration. Had they used the distances given in Table 4 and 6, they would have found $\mathrm{H}_0 = 60(\pm 6)$ and $65(\pm 6)$, instead. This is a clear demonstration that the $\mathrm{D}_n - \sigma$/FP route to H_0 is in no way an independent method, but depends entirely on distances used as calibrators, and in particular on their distance of the Virgo cluster which we dispute. The power of H_0 from SNe Ia, the calibration of which only depends on Cepheids, bypassing the still controversially discussed Virgo cluster distance, becomes here evident.

Lauer et al. (1998) have tried to calibrate the Hubble diagram of 1st-ranked cluster galaxies of Lauer & Postman (1992; Section 3.1.1) by means of the SBF method. For this purpose they have observed the SBF with *HST* of four 1st-ranked galaxies at $\sim 4300\ \mathrm{km\,s^{-1}}$. On the unproven assumption that the fluctuation magnitude $\overline{m}_I$ of these very particular objects is the same as in local E/S0 galaxies *and* spiral bulges, and adopting a local calibration magnitude $\overline{M}_I$, which in turn is controversial, they have derived distances of these four galaxies which they claim are in agreement with the turnover magnitudes m_I^{T} of the respective GCLFs. However, their calibrating turnover magnitude M_I^{T} *rests entirely on M87*, which is known to have a peculiar bimodal GCLF, *and on an adopted Virgo cluster modulus which is* $0^{\mathrm{m}}.6$ *smaller than shown in Table 4 and 5.* Thus, judging only from GCLFs their proposed value of $\mathrm{H}_0 = 82$ should be reduced by a factor of 1.32 to give $\mathrm{H}_0 = 62$. In any case their procedure appears like a complicated way—particularly in comparison to SNe Ia—to transport the Virgo cluster distance into the expansion field at $v \sim 10\,000\ \mathrm{km\,s^{-1}}$.

4. $\mathbf{H_0}$ from Field Galaxies

The most difficult and least satisfactory determination of H_0 comes from field galaxies. The difficulty comes from selection effects (Malmquist bias); the restricted impact comes from the fact that the method can hardly be carried beyond $\sim 5000\ \mathrm{km\,s^{-1}}$ and hence does not necessarily reflect the large-scale value of H_0.

The problem of selection effects is illustrated in Fig. 8b. 200 galaxies of constant space density and with $v < 5000\ \mathrm{km\,s^{-1}}$ were randomly distributed in space by a Monte Carlo calculation. The "true" distances were expressed in velocities. It has further been assumed that the distance moduli of each galaxy had also been determined by the TF relation with an intrinsic scatter of $\sigma_{(m-M)} = \pm 0^{\mathrm{m}}.4$. The corresponding Hubble diagram in Fig. 8b gives the false impression as if the scatter would increase at larger distances; actually the increasing *number* of distant galaxies produces deviations by ±2 or even ±3 sigma. This makes the Hubble diagram "ugly" as compared to the one of SNe Ia in Fig. 8a (repeated here from Fig. 3 for comparison), but still the mean Hubble line through the points has the correct position *provided the sample is complete out to a given distance limit.* If, however, the sample is cut by an *apparent-magnitude limit* m_{lim} the Hubble line

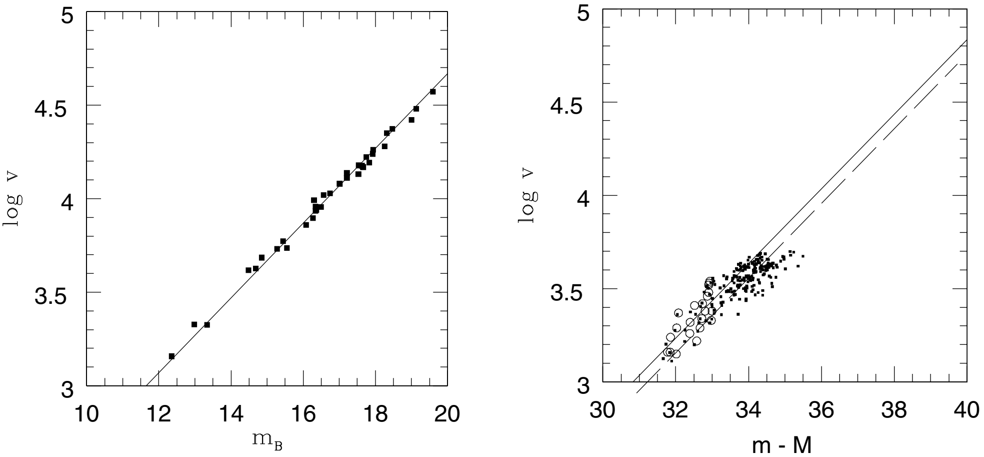


FIGURE 8. a) The Hubble diagram in B of SNe Ia (from Fig. 3; shown here only for comparison). b) The Hubble diagram of 200 galaxies with $v < 5000\,\mathrm{km\,s^{-1}}$. Their velocities were assigned by a Monte Carlo calculation assuming constant space density. For each galaxy it was assumed that its TF distance has been determined within an intrinsic scatter of $0\overset{m}{.}4$. For the calibration of the abscissa it was assumed that all galaxies have a true absolute magnitude of $M = -20.0$ and that $H_0 = 60$. The dashed line is the best fit of the Hubble line for all points. The galaxies with *apparent* magnitude $m < 13.0$ are shown as large symbols. They define a Hubble line (full line) which is $0\overset{m}{.}29$ brighter, corresponding to a too high value of H_0.

shifts upwards, resulting in too high a value of H_0. With the present, realistically chosen parameters ($<M>= -20.0, \sigma_M = 0^m4, H_0 = 60, m_{\rm lim} = 13.0$) the overestimate of H_0 amounts to 14%. Samples which are not even complete to a given apparent-magnitude limit can give much larger systematic errors.

Another illustration of the Malmquist bias is given in the so-called Spaenhauer diagram (Fig. 9), where 500 galaxies were randomly distributed in space out to 42 Mpc using a Monte Carlo routine. The galaxies are assumed to scatter by $\sigma_M = 2\overset{m}{.}0$ about a fixed mean luminosity of $M = -18.0$. If this distance-limited sample is cut by a limit in apparent magnitude a sample originates with very complex statistical properties. In particular the mean luminosity within any distance interval increases with distance. In the lower panel of Fig. 9 the increase amounts to $\Delta M = 2\overset{m}{.}6$ which would overestimate H_0 by a factor of 3.3 if one were to force the local calibration of $M = -18.0$ on the most distant galaxies of the biased sample. For the sake of the argument the intrinsic dispersion of $\sigma_M = 2^m0$ was chosen here to be unrealistically large, but the effect is omnipresent in all apparent-magnitude-limited samples, *it always overestimates* H_0. Only if $\sigma_M \lesssim 0\overset{m}{.}2$, as in the case of SNe Ia, the Malmquist bias becomes negligible. An additional crux of the bias is that the observable scatter within any distance interval *is always smaller than the true intrinsic scatter. This has mislead several authors to assume that σ_M is small and hence to underestimate the importance of bias.*

There is an additional statistical difficulty as to the derivation of H_0 from field galaxies. Frequently H_0 is determined from the arithmetic mean of many values of H_i found from individual field galaxies. The underlying assumption is that the values H_i have a Gaussian distribution which is generally not correct. If the distance errors are symmetric in the moduli $(m-M)$, they are not in linear distance r and consequently cause a skewness of H_i towards high values. An actual example is provided by the bias-corrected TF distances of 155 field galaxies with $v < 1000\,\mathrm{km\,s^{-1}}$. A simple arithmetic mean of their corresponding H_i values, which have a non-Gaussian distribution, would give $<H_i>= 64 \pm 2$, while the

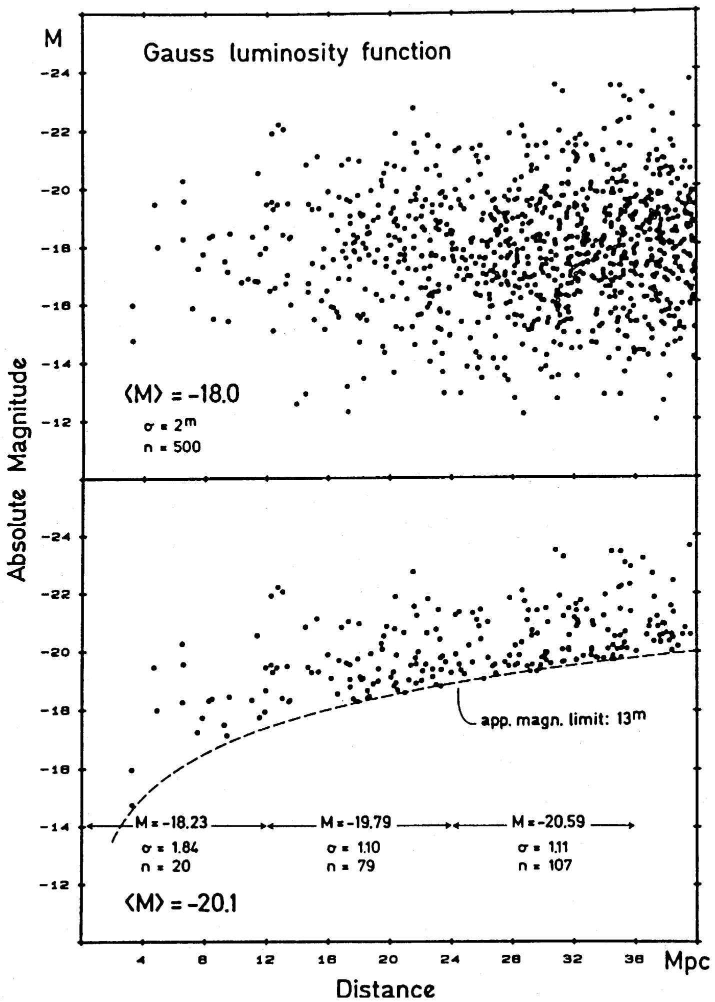


FIGURE 9. *Upper panel:* Monte Carlo calculation of the distance and absolute magnitude of 500 galaxies with constant space density and $r < 42$ Mpc. Their mean absolute magnitude is $<M>= -18.0$ with an intrinsic scatter of $\sigma_M = 2^{\rm m}\!.0$. *Lower panel:* The same distribution cut by an apparent-magnitude limit $m < 13.0$. Note that the mean luminosity of the magnitude-limited sample increases with distance.

median value is 60.0±3. Yet a more nearly correct solution is given by averaging the values $\log \mathrm{H}_i$, which have a perfect Gaussian distribution. This then gives the best estimate of $\mathrm{H}_0 = 58 \pm 2$ (Federspiel 1999).

Unfortunately, the hope that the inverse TF relation (line width versus magnitude) was bias-free (Sandage, Tammann, & Yahil 1979; Schechter 1980) has been shattered (Teerikorpi et al. 1999).

Strategies have been developed to correct for Malmquist bias, particularly in the case of the (direct) TF relation, by e.g. Sandage (1988b, 1995) and Federspiel, Sandage, & Tammann (1994) and similarly by Teerikorpi (1994, 1997), Bottinelli et al. (1986, 1995), and Theureau (2000).

Method (1)	H_0 (2)	Range [$\mathrm{km\,s^{-1}}$] (3)	Source (4)
Luminosity classes	56 ± 5	4000	Sandage 1996a
Morphol. twins of M 31, M 101	50 ± 5	5000	Sandage 1996b
TF (using mag. + diam.)	55 ± 5	5000	Theureau et al. 1997
Galaxy diam.	$50 - 55$	5000	Goodwin et al. 1997
TF	53 ± 5	500	Federspiel 1999
	58 ± 2	1000	Federspiel 1999
Luminosity classes	55 ± 3	4000	Sandage 1999
$D_n - \sigma$	52 ± 8	4000	Federspiel 1999
TF (using mag. + diam.)	55 ± 5	8000	Theureau 2000
Morphol. twins of M 31, M 81	60 ± 10	5000	Paturel et al. 1998
Inverse TF	53 ± 6	8000	Ekholm et al. 1999

TABLE 8. Values of H_0 from field galaxies corrected for Malmquist bias

A selection of H_0 values, which have been derived from magnitude-limited, yet bias-corrected samples of field galaxies have been compiled in Table 8. The conclusion is

$$H_0 = 55 \pm 5 \tag{4.1}$$

from field galaxies, which is valid within various distance ranges up to $\sim 5000\ \mathrm{km\,s^{-1}}$. It should be noted that the different methods use spirals only—except the $D_n - \sigma$ entry—and their calibration rests only on Cepheids, independent of any adopted distance of the Virgo cluster.

Tonry et al. (2000) have derived $H_0 = 77 \pm 4$ from the SBF of 300 field galaxies with $v \lessapprox 4000\ \mathrm{km\,s^{-1}}$ and on the basis of a very specific flow model. However, the usefulness of the SBF method has never been tested beyond doubt; a Hubble diagram of field galaxies and a corresponding determination of the intrinsic scatter of the method and hence of the importance of the Malmquist bias are still missing. The authors state correctly "The level of this Malmquist-like bias in our sample is difficult to quantify ... and we may also be subject to possible biases and selection effects which depend on distance." Moreover, in discussing Table 3 it was noted that the small relative SBF distance Virgo–Leo and the large value Fornax–Virgo are in serious contradiction to the evidence of Cepheids and SNe Ia. It is well possible that the SBF method is still affected by hidden second-parameter effects.

5. Conclusions

The exceptionally tight Hubble diagram of blue SNe Ia out to $\sim 30\,000\ \mathrm{km\,s^{-1}}$ combined with their mean absolute magnitude, which is determined from eight galaxies whose Cepheid distances have been determined with *HST*, offers the ideal instrument for the calibration of the large-scale value of H_0. The result is—after allowance is made for the relatively weak dependence of SNe Ia luminosities on decline rate Δm_{15} and color and for some small systematic effects—$H_0 = 58.5 \pm 6.3$.

The quoted external error is mainly determined by the calibrating Cepheid distances. Although Cepheids are generally considered to be the most reliable distance indicators, certain questions remain as to their zeropoint calibration through LMC and the effect of metallicity variations.

Phillips et al. (1999) and Suntzeff et al. (1999) have adopted a steeper relation between the SNe Ia luminosity and the decline rate Δm_{15}, which is not supported by the present larger sample. If their steep slope is taken at face value H_0 would be increased by 5%. Gibson et al. (2000) have increased the Cepheid distances of Saha et al. and Tanvir et al. in Table 2 by $0\overset{\mathrm{m}}{.}11$ on average on the basis of additional Cepheids of lower weight. This would increase H_0 by another 5 percent. If one choses to cumulate the two effects one could defend a value of $H_0 = 64$ (cf. Freedman 2001).

The calibration of H_0 through the Hubble diagram of clusters and through field galaxies in Section 3 and 4 has given values around $H_0 = 55$, i.e. somewhat smaller, if anything, than the value from SNe Ia. Since the latter offer the most direct way to H_0, they are given the highest weight. A value of

$$H_0 = 58 \pm 6 \tag{5.1}$$

is therefore adopted as the best estimate available at present.

The two most frequent errors in the extragalactic distance scale, which can lead to values of $H_0 > 65$, are too small a Virgo cluster distance and Malmquist bias.

The distance of the Virgo cluster is still sometimes equated to the mean Cepheid distance of three Virgo spirals which are visibly on the near side of the cluster due to their high resolution into stars—in fact they had been preselected on grounds of their high resolution to facilitate work with *HST*. Thus, even if the larger cluster distance of $(m - M)_{\mathrm{Virgo}} = 31.60 \pm 0.20$ (cf. Tables 4 and 5) was not required by a fourth Cepheid distance of a bona fide cluster member (NGC 4639) and in addition by the three SNe Ia and the TF relation, the three Cepheid distances of highly resolved galaxies could set only *a lower limit on the cluster distance*, i.e. $(m - M)_{\mathrm{Virgo}} > 31.05$.

The distance of the Virgo cluster has entered with high or even full weight into the zeropoint calibration of the $D_n - \sigma$, FP, and SBF methods and occasionally even of the turnover magnitude of the GCLF. (The significance of the GCLF, although not yet proven beyond doubt, is just that the calibration rests on the Cepheid distance of M 31 *and* on the RR Lyr distances of Galactic globular clusters [cf. Tammann & Sandage 1999]). If one inserts in these cases too low a Virgo cluster distance, one obtains incorrect values of $H_0 \approx 70$, or even more, by necessity.

Locally calibrated distance indicators with non-negligible scatter can exclusively be applied to distance-limited samples (or volume-limited samples in case of clusters), but only apparent-magnitude-limited samples are available, which frequently are not even complete to any specific magnitude limit. The statistical differences between a distance- and magnitude-limited sample are known for almost eighty years (Malmquist 1920, 1922), but this insight is violated again and again, always leading to too high values of H_0. Malmquist's message is that control of the objects missing out to a fixed distance from a magnitude-limited catalog is equally important as the catalog entries themselves.

The future will see high-weight determinations of H_0 from the Sunyaev-Zeldovich effect, from gravitationally lensed quasars, and from the CMB fluctuations. At present these methods yield values of $50 \lesssim H_0 \lesssim 70$, which is not yet competitive with the solution of SNe Ia. In the case of the CMB fluctuations one can solve for H_0 only in combination with other (poorly known) cosmological parameters. Therefore an *independent*, high-accuracy determination of H_0, to be used as a prior, is as important as ever.

The authors thank Bernd Reindl, Daniel Cerrito, and Hans Schwengeler for their most valuable help in preparing the manuscript. G.A.T. acknowledges financial support of the Swiss National Science Foundation.

REFERENCES

Alibert, Y. et al. 1999 *A&A* **344**, 551.

Baum, W. A., Hammergren, M., Thomsen, B., Groth, E. J., Faber, S. M., Grillmair, C. J., & Ajhar, E. A. 1997 *AJ* **113**, 1483.

Binggeli, B., Popescu, C. C., & Tammann, G. A. 1993 *A&AS* **98**, 275.

Binggeli, B., Sandage, A., & Tammann, G. A. 1985 *AJ* **90**, 1681.

Bottinelli, L., Gouguenheim, L., Paturel, G., & Teerikorpi, P. 1986 *A&A* **156**, 157.

Bottinelli, L., Gouguenheim, L., Paturel, G., & Teerikorpi, P. 1991 *A&A* **252**, 550.

Bottinelli, L., Gouguenheim, L., Paturel, G., & Teerikorpi, P. 1995 *A&A* **296**, 64.

Branch, D. 1998 *ARA&A* **36**, 17.

Carroll, S. M., Press, W. H., & Turner, E. L. 1992 *ARA&A* **30**, 499.

Capaccioli, M., della Valle, M., & D'Onofrio, M. 1989 *AJ* **97**, 1622.

Dale, D. A., Giovanelli, R., Haynes, M. P., Camusano, L. E., & Hardy, E. 1999 *AJ* **118**, 1489.

Dressler, A. 1987 *ApJ* **317**, 1.

Ekholm, T., Teerikorpi, P., Theureau, G., Hanski, M., Paturel, G., Bottinelli, L., & Gouguenheim, L. 1999 *A&A* **347**, 99.

Faber, S. M., et al. 1989 *ApJS* **69**, 763.

Feast, M. W. 1999 *PASP* **111**, 775.

Federspiel, M. 1999 *Ph.D. thesis*, Univ. of Basel.

Federspiel, M., Sandage, A., & Tammann, G. A. 1994 *ApJ* **430**, 29.

Federspiel, M., Tammann, G. A., & Sandage, A. 1998 *ApJ* **495**, 115.

Ferrarese, L., et al. 1996 *ApJ* **464**, 568.

Ferrarese, L., et al. 2000 *ApJ* **529**, 745.

Freedman, W. L. 2001, this volume.

Garnavich, P. M., Riess, A. G., Kirshner, R. P., Challis, P. & Wagner, R. M. 1996 *A&AS* **189**, 4509.

Gibson, B. K., et al. 2000 *ApJ* **529**, 723.

Gilmozzi, R. & Panagia, N. 1999 *Space Telescope Science Institute Preprint Series*, **No. 1319**.

Giovanelli, R. 1997, *private communication.*

Giovanelli, R., et al. 1997 *AJ* **113**, 22.

Goodwin, S. P., Gribbin, J., & Hendry, M. A. 1997 *AJ* **114**, 2212.

Graham, J. A., et al. 1999 *ApJ* **516**, 626.

Gratton, R. 2000. In *XIXth Texas Symposium on Relativistic Astrophysics and Cosmology*, (eds. E. Aubourg, et al., Mini-symposium 13/03).

Höflich, P. H. & Khokhlov, A. 1996 *ApJ* **457**, 500.

Jerjen, H. & Tammann, G. A. 1993 *A&A* **276**, 1.

Jha, S., et al. 1999 *ApJS* **125**, 73.

Kelson, D. D., et al. 2000 *ApJ* **529**, 768.

Kennicutt, R. C., et al. 1998 *ApJ* **498**, 181.

Kraan-Korteweg, R. C. 1986 *A&AS* **66**, 255.

Lanoix, P. 1998 *A&A* **331**, 421.

Lanoix, P., Paturel, G., & Garnier, R. 1999 *ApJ* **517**, 188.

Lauer, T. R. & Postman, M. 1992 *ApJ* **400**, L47.

Lauer, T. R., et al. 1998 *ApJ* **499**, 577.

Livio, M. 1997. In *The Extragalactic Distance Scale* (eds. M. Livio, M. Donahue, & N. Panagia), p. 186. Cambridge Univ. Press.

Madore, B. & Freedman, W. L. 1998 *ApJ* **492**, 110.

Macri, L. M., et al. 1999 *ApJ* **521**, 155.

Malmquist, K. G. 1920 *Lund Medd. Ser. II* **22**, 1.

Malmquist, K. G. 1922 *Lund Medd. Ser. I* **100**, 1.

Mazumdar, A. & Narasimha, D. 2000 *preprint.*

Méndez, R. H., Kudritzki, R. P., Ciardullo, R., & Jacoby, G. H. 1993 *A&A* **275**, 534.

Mould, J. R., et al. 2000 *ApJ* **528**, 655.

Parodi, B. R., Saha, A., Sandage, A., & Tammann, G. A. 2000 *ApJ*, **540**, 634.

Paturel, G., et al. 1998 *A&A* **339**, 671.

Perlmutter, S. 1998. In *Supernovae and Cosmology* (eds. L. Labhardt, B. Binggeli, & R. Buser), p. 75. Astronomisches Institut der Universität Basel.

Perlmutter, S., et al. 1999 *ApJ* **517**, 565.

Phillips, M. M. 1993 *ApJ* **413**, 105.

Phillips, M. M., Wells, L. A., Suntzeff, N. B., Hamuy, M., Leibundgut, B., Kirshner, R. P., & Foltz, C. B. 1992 *AJ* **103**, 1632.

Phillips, M. M., Lira, P., Suntzeff, N. B., Schommer R. A., Hamuy, M., & Maza, J. 1999 *AJ* **118**, 1766.

Prosser, C. F., et al. 1999 *ApJ* **525**, 80.

Pritchet, C. J. & van den Bergh, S. 1987 *ApJ* **318**, 507.

Richtler, T. & Drenkhahn, G. 1999. In *Cosmology and Astrophysics: A collection of critical thoughts* (eds. W. Kundt & C. van de Bruck), Lecture Notes in Physics. Springer; astro-ph/9909117.

Riess, A. G., et al. 1998 *AJ* **116**, 1009.

Saha, A., Labhardt, L., & Prosser, C. 2000b *PASP* **112**, 163.

Saha, A., Labhardt, L., Schwengeler, H., Macchetto, F. D., Panagia, N., Sandage, A., & Tammann, G. A. 1994 *ApJ* **425**, 14.

Saha, A., Sandage, A., Labhardt, L., Schwengeler, H., Tammann, G. A., Panagia, N., & Macchetto, F. D. 1995 *ApJ* **438**, 8.

Saha, A., Sandage, A., Labhardt, L., Tammann, G. A., Macchetto, F. D., & Panagia, N. 1996a *ApJ* **466**, 55.

Saha, A., Sandage, A., Labhardt, L., Tammann, G. A., Macchetto, F. D., & Panagia, N. 1996b, *ApJS* **107**, 693.

Saha, A., Sandage, A., Labhardt, L., Tammann, G. A., Macchetto, F. D., & Panagia, N. 1997 *ApJ* **486**, 1.

Saha, A., Sandage, A., Tammann, G. A., Labhardt, L., Macchetto, F. D., & Panagia, N. 1999, *ApJ* **522**, 802.

Saha, A., Sandage, A., Thim, F., Labhardt, L., Tammann, G. A., Macchetto, F. D., & Panagia, N. 2001 *ApJ*, **551**, 973.

Sakai, S., Zwitsky, D., & Kennicutt, R. C. 2000 *AJ* **119**, 1197.

Sakai, S., et al. 2000 *ApJ* **529**, 698.

Sandage, A. 1961 *ApJ* **133**, 355.

Sandage, A. 1962 *ApJ* **136**, 319.

Sandage, A. 1988 *PASP* **100**, 935.

Sandage, A. 1988 *ApJ* **331**, 583.

Sandage, A. 1995. In *The Deep Universe* (eds. B. Binggeli & R. Buser), p. 210. Springer.

Sandage, A. 1996a *AJ* **111**, 1.

Sandage, A. 1996b *AJ* **111**, 18.

Sandage, A. 1999 *ApJ* **527**, 479.

Sandage, A. & Bedke, J. 1988 *Atlas of Galaxies useful to measure the Cosmological Distance Scale*. NASA.

Sandage, A., Bell, R. A., & Tripicco, M. J. 1999 *ApJ* **522**, 250.

Sandage, A. & Hardy, E. 1973 *ApJ* **183**, 743.

Sandage, A., Tammann, G. A., & Federspiel, M. 1995 *ApJ* **452**, 1.

Sandage, A., Tammann, G. A., & Yahil, A. 1979 *ApJ* **232**, 352.

Schechter, P. L. 1980 *AJ* **85**, 801.

Schlegel, D., Finkbeiner, D., & Davis, M. 1998 *ApJ* **500**, 525.

Schmidt, B., et al. 1998 *ApJ* **507**, 46.

Schröder, A. 1995 *Ph.D. thesis*, Univ. of Basel.

Silbermann, N. A., et al. 1999 *ApJ* **515**, 1.

Soffner, T., Méndez, R., Jacoby, G., Ciardullo, R., Roth, M., & Kudritzki, R. 1996 *A&A* **306**, 9.

Stanek, K. Z. & Udalski, A. 1999, *preprint*; astro-ph/9909346.

Suntzeff, N. B., et al. 1999 *AJ* **117**, 1175.

Tammann, G. A. 1982. In *Supernovae: A Survey of Current Research* (eds. M. J. Rees & R. J. Stoneham), p. 371. Reidel.

Tammann, G. A. 1993. In *IAU Symp. 155, Planetary Nebulae* (eds. R. Weinberger & A. Acker), p. 515. Kluwer.

Tammann, G. A. & Federspiel, M. 1997. In *The Extragalactic Distance Scale* (eds. M. Livio, M. Donahue, & N. Panagia), p. 137. Cambridge Univ. Press.

Tammann, G. A. & Sandage, A. 1999. In *Harmonizing Cosmic Distance Scales in a Post-Hipparcos Era* (eds. D. Egret & A. Heck), p. 204.

Tammann, G. A., Sandage, A., & Reindl, B. 2000. In *XIXth Texas Symposium on Relativistic Astrophysics and Cosmology* (eds. E. Aubourg, et al.), Mini-Symposium 13/11.

Tanvir, N. R., Shanks, T., Ferguson, H. C., & Robinson, D. R. T. 1995, *Nature* **377**, 27.

Teerikorpi, P. 1984 *A&A* **141**, 407.

Teerikorpi, P. 1997 *ARA&A* **35**, 101.

Teerikorpi, P., Ekholm, T., Hanski, M. O., & Theureau, G. 1999 *A&A* **343**, 713.

Theureau, G. 2000. In *XIXth Texas Symposium on Relativistic Astrophysics and Cosmology* (eds. E. Aubourg, et al.), Mini-Symposium 13/12.

Theureau, G., Hanski, M., Ekholm, T., Bottinelli, L., Gouguenheim, L., Paturel, G., & Teerikorpi, P. 1997 *A&A* **322**, 730.

Tonry, J. L., Blakeslee, J. P., Ajhar, E. A., & Dressler, A. 2000 *ApJ* **530**, 625.

Tripp, R. 1998 *A&A* **331**, 815.

Tripp, R. & Branch, D. 1999 *ApJ* **525**, 209.

Turner, A., et al. 1998 *ApJ* **505**, 207.

Walker, A. R. 1999. In *Post-Hipparcos cosmic candles* (eds. A. Heck & F. Caputo), p. 125. Kluwer Academic Publishers.

Weedman, D. 1976 *ApJ* **203**, 6.

Strong gravitational lensing: Cosmology from angles and redshifts

By ANTHONY TYSON

Bell Labs, Lucent Technologies, Murray Hill, NJ 07974

It is rare in astronomy to have a purely physics-based technique for studying the distant universe. Rooted in General Relativity, the image distortion and time delay of light from distant objects caused by foreground gravitational lenses offers such a window on the universe. Using only combinations of measured redshifts, angles, and arrival times of source intensity fluctuations, lensing observations can probe the mass distribution of the lens, the rate of expansion of the universe (the Hubble constant), the acceleration of expansion (dark energy), and the total amount of matter in the universe. The *HST* has made and will continue to make unique contributions to this new window on the universe.

1. Introduction

The universe is not as it seems: distant galaxies and quasars are in the wrong places. Their apparent positions on the sky have moved relative to where they would normally appear, and the culprit is mass-energy. Specifically, a massive object (a star, a galaxy, a cluster of galaxies) will warp space-time around it, causing light rays to bend as they pass by. If a mass concentration lies between us and a distant source, that source will appear in an altered location. The effect is called gravitational lensing, and it also systematically distorts the images of resolved sources like galaxies. Gravitational lensing over vast distances is sensitive to all forms of mass-energy. Because of this, we can use the distant quasars and galaxies as tools to study foreground masses as well as the mass-energy content of the universe. These gravitational lens tools hold unique promise for probing cosmology and the underlying physics of our universe.

1.1. *Dark matter and dark energy*

Since Zwicky studied the dispersion of velocity of galaxies in clusters in the thirties, the existence of a dark component that dominates the universe's mass has been suspected. Over the last two decades, the astronomy community has reached the consensus that dark matter is present at a variety of scales from spiral and elliptical galaxies to super clusters of galaxies. The theory of big bang nucleosynthesis, together with recent measurements of primeval deuterium, imply that only five percent of the mass-energy of the universe is in the form of ordinary matter—baryons. Stars and gas and dust contribute less than half a percent of the mass. Roughly 95% of the mass of our universe is of an unknown form and dark! This dark matter must be made of something different than ordinary matter—non-baryonic. A theoretical favorite, motivated by problems with the standard model of particle physics, is one or more non-baryonic particles produced in the early hot universe. These theoretical candidates for this "cold dark matter" can gravitationally clump on the scales of galaxies: axions or weakly interacting massive particles.

Over the last few years the composition of the universe has become even more puzzling, as the observed luminosity of Type Ia supernovae at high redshift, and other observations, appear to imply an acceleration of the universe's expansion in recent times. In order to explain such an acceleration, we need "dark energy." Even more puzzling is the fine tuning of parameters which appears to be required to explain why the dark energy density today is about the same as that of dark matter, since it evolves differently with time. This may

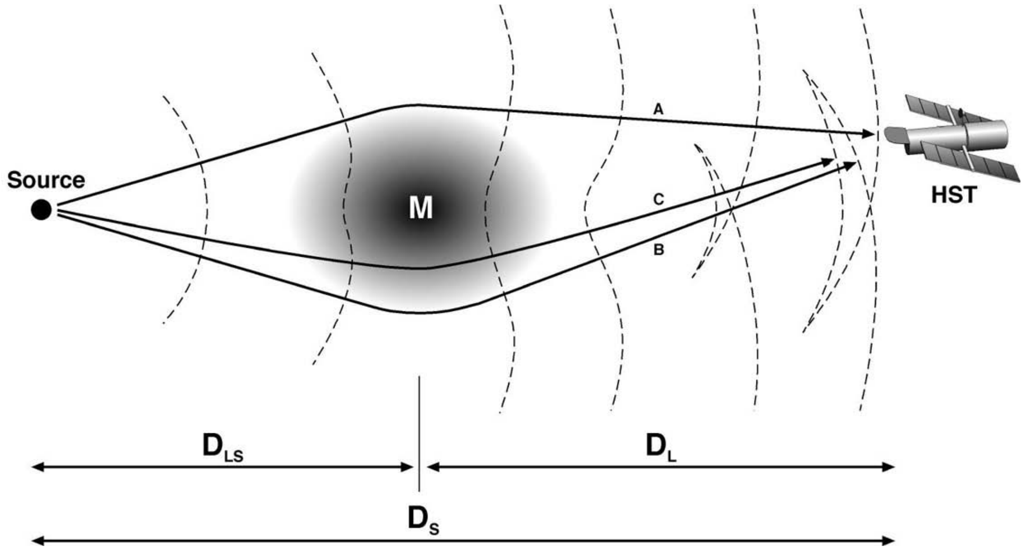


FIGURE 1. Light rays (solid lines) and wavefronts (dashed lines) are shown for light traveling from a source S past a mass M to us. Wavefronts deform and intersect as they pass by the lens mass. Travel time has contributions from the gravitational potential as well as path geometry. In this cartoon the light travel time is shortest for ray A and longest for ray C. We see a separate image of the source for each path whose light travel time is stationary relative to nearby paths. Less symmetric lens mass distributions can give rise to more images. In general, there are an odd number of images, one of which (nearest the center) is demagnified in proportion to the central compactness of the mass. The distances D_L, etc. are angular diameter distances.

imply that our understanding of the underlying physics is incomplete. It is conceivable that what we call dark matter and dark energy arise from some unknown aspect of spacetime geometry.

1.2. *Lensing as a tool*

Gravitational light-bending involves only mass-energy, distances, and angles. This cosmological tool is thus free of the dependency on radiation and ordinary baryonic matter that so plagues traditional cosmological studies. Astronomy traditionally is biased to luminosity—but the gravitational lens tools are opening a new and independent window on the universe of mass-energy.

The *HST* is particularly well suited to exploit so-called "strong gravitational lensing," where the lens mass is so centrally peaked that it causes multiple images of the source. The reason is the high angular resolution of *HST* for optical imaging: many of the effects of strong lensing are on sub-arcsecond scales. To understand how strong lensing works, we consider a wavefront of light from a distant source as it encounters a mass shown in Figure 1. An initially spherical wavefront is perturbed by the lensing mass distribution. These perturbations grow, as waves on a pond, into components traveling in multiple directions. Multiple intersecting wavefronts, each arriving at our telescope at a different time, are created by the lens mass. Fermat's principle tells us that images of the source correspond to rays for which the light travel time is locally stationary; relative to nearby paths the travel time is a minimum, a maximum, or a saddle-point.

If the lens mass has an extent small compared with its distance from us and the source, the perturbation can be understood as an effective (spatially varying) index of refraction

$$n = 1 - \frac{2\Phi}{c^2}$$

where Φ is the gravitational potential of the lens mass distribution. A larger gravitational potential (more mass) increases the light travel time. The travel time also depends on the length of the path (the geometrical time delay). For a given ray the total time delay τ can be written in terms of a 2-dimensional effective potential Ψ and the angular diameter distances (see Fig. 1):

$$\Psi = \frac{D_{LS}}{D_S} \int_{\rm obs}^{\rm source} \frac{2\Phi}{c^2} \frac{dl}{D_L}$$

$$\tau(\vec{\theta}) = \frac{1+z_L}{c} \left(\frac{D_L\ D_S}{D_{LS}}\right) \left[\frac{1}{2}|\vec{\theta} - \vec{\phi}|^2 - \Psi(\vec{\theta})\right]$$

where $\vec{\theta}$ is the apparent position of the image in the sky and $\vec{\phi}$ is the position it would have had with no lens. For a non-singular mass distribution there are in general an odd number of images. Their location on the sky is given by the solution of the "lens equation" derived from Fermat's condition $\delta\tau/\delta\vec{\theta} = 0$:

$$\vec{\theta} - \vec{\phi} = \frac{\partial \Psi}{\partial \vec{\theta}}$$

In strong lensing the mass distribution is sufficiently peaked to give multiple solutions. Typical gravitational light bend angles $\vec{\theta} - \vec{\phi}$ range from one arcsec for a galaxy mass to more than 30 arcsec for the mass in a cluster of galaxies. As we shall see, good angular resolution is required in both cases if we wish to use these natural lenses as tools for detailed reconstruction of the lens mass. In the extreme case of a singular point mass, a source directly behind is lensed into a so-called Einstein ring of angular radius

$$\theta_E = 1.6\ h \left(\frac{M}{10^{12}\ \mathrm{M}_\odot}\right)^{1/2}\ f_H^{-\frac{1}{2}} \quad \text{arcsec}\ ,$$

where h is the Hubble constant in units of 100 km s^{-1} Mpc^{-1} and f_H is a distance for the lens-source system in units of the Hubble length: $f_H = [D_L D_S/D_{LS}](c/H_0)^{-1}$. For lenses and sources at cosmological distances, f_H is of order unity (see Turner et al. 1984). The Einstein ring is a simple example of a critical line in the image plane, corresponding to a point caustic (infinite magnification) in the source plane. For general lens mass distributions, critical lines and caustics are complicated curves; sources inside and outside caustics appear as a different odd number of images.

The various images of a strongly lensed source appear either as direct or flipped images of the source. (Just as in the classical terrestrial mirages, the so-called "Fata Morgana" being the flipped image mirage). Different images of the source are magnified in varying amounts $(\partial\vec{\phi}/\partial\vec{\theta})^{-1}$, and the eigenvalues of this magnification matrix give the amount by which the images are stretched in each orthogonal direction. The determinant of the magnification matrix is positive at maxima and minima of τ (positive parity images) and negative at saddle-points (negative parity "flipped" images). For example, in a modified version of the above point mass case, use a more realistic non-singular mass and also mis-align the source; this produces a central parity-flipped and demagnified image in addition to two magnified direct images lying near the Einstein ring.

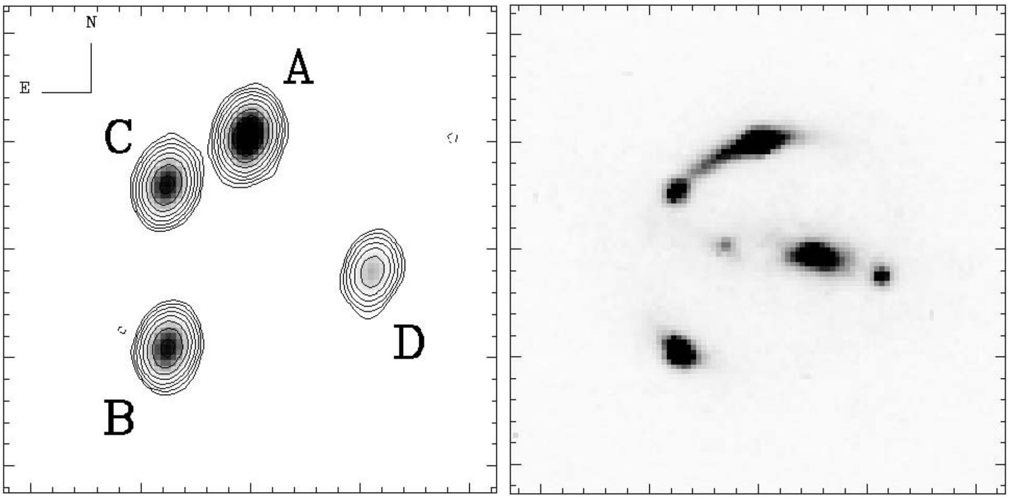


FIGURE 2. The quad lens B1608+656. The quasar is at redshift 1.39. The 8 GHz *VLA* radio image (*left*) and the I-band *HST* image (*right*) are shown. The size of this field is only 4 arcsec; the largest distance between any pair of QSO images is 2 arcsec. The optical image reveals a partial Einstein ring from the stellar light in the quasar's host galaxy, in addition to the two $z = 0.63$ lens galaxies. Courtesy, C. Fassnacht.

2. Quasars lensed by galaxies

The first lensed QSO, the radio source 0957+561, was discovered in 1978 by Walsh, Carswell & Weymann. It appears as two images of a $z = 1.41$ quasar separated on the sky by 6 arcsec, lensed by a $z = 0.36$ elliptical galaxy. Two images of the radio jet are also seen in the radio. A third, demagnified, image is hidden by the glare of the elliptical galaxy. There is also a small cluster of galaxies associated with this lens, contributing to the 6 arcsec image separation. Most multiply lensed QSOs have image separations closer to the 1–2 arcsec expected from a lone foreground galaxy. Thus, ground-based optical surveys are naturally biased against small separation lenses. There are now over 50 quasars and AGNs which are observed to be multiply lensed by intervening lenses. The majority are doubles like 0957+561, but about a third are quadruples. Why is this?

If the lens projected mass distribution departs from circularity, new solutions appear and more images are seen. An elliptical lens mass distribution easily produces four magnified source images and one demagnified image. An example is shown in Figure 2, where the *VLA* radio image is compared with the *HST* optical image.

It was realized early on that lensed QSOs would be a powerful statistical probe of both galaxy mass and cosmology, so unbiased samples are important (Turner, Ostriker & Gott 1984). Recently, arcsec resolution radio surveys have been used to generate an unbiased sample of candidate lensed quasar systems, for spectroscopic follow-up, photometric monitoring, and analysis. Many of the recent lensed quasar systems have been found in this way, and more are coming (CLASS and JVAS surveys).

There is a second problem, however, which may not be as easily avoided. In order to reconstruct the gravitational potential of the lens, the observed lensed images must sample the lens potential at various locations; i.e. there must be many sources distributed over the corresponding angular scale behind the lens. But in the case of the classical single point QSO source, that is not possible; even multiple images of that source merely sample the potential at roughly a single radius. As described below, this has led to the practice of using simplified lens model potentials, some of which are degenerate. In turn, this

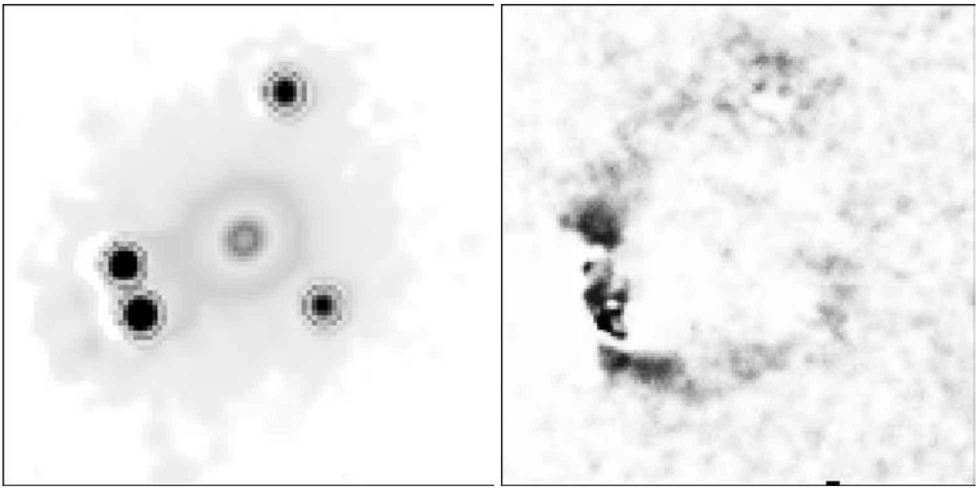


FIGURE 3. The quad lens PG1115+080. The I-band *HST* image is shown on the left. The four magnified $z = 1.72$ quasar images surround the image of the foreground $z = 0.3$ lens elliptical galaxy. A *Subaru* K-band image (0.3 arcsec FWHM), with QSO and lens galaxy images subtracted, is shown on the right. The largest distance between any pair of QSO images is 2 arcsec. The outer parts of the host galaxy, seen in the K imaging, samples a different part of the lens potential. Given morphological structure in the host galaxy and sufficient signal-to-noise ratio, this multiple sampling of the lens can lead to the removal of lens model degeneracies. Credit: C. Impey, et al. (*HST* image); F. Iwamuro, et al. 2000 (*Subaru* image).

gives rise to systematic error. The solution lies in the stellar light from the QSO host. *HST* images in the red and near IR have revealed arc-like lensed sub-images of the host galaxy. An example is shown in Figure 3 (see Schechter, et al. 1997). Since the light from the host is offset from the QSO, analysis of these arc images in deep imaging (revealing lensed features in the host galaxy) together with the QSO multiple images can break the degeneracies in the source model potential.

2.1. *Direct measure of the Hubble constant*

QSOs flicker in luminosity, so cross-correlation of time series photometry can yield a measurement of the time delay between multiple images. Angular diameter distances are proportional to c/H_0 and the redshifts of the source and lens. If we knew the effective potential Ψ (the lens mass distribution), we could use the observed time delays and the lens-predicted delays at the image positions θ_A and θ_B and solve for the only remaining unknown: H_0 (Refsdal 1964). It cannot be over-emphasized that this is a direct measure of H_0, independent of the systematics in the usual distance ladder. This technique bypasses parallaxes, C-M diagrams of star clusters, apparent magnitudes of Cepheids, rotation velocities of galaxies, supernovae—all of which are baryon-based heuristic calibrators. By contrast, lensing is a physics-based measurement, depending only on mass-energy, observed angles, and redshifts. Finally, lensed QSO estimates of H_0 are on a cosmological scale, not the local scale of the distance ladder techniques.

To apply Refsdal's clever technique, one needs extensive monitoring of individual QSO image intensity fluctuations, as well as a robust estimate of the lens mass distribution. In the past, *HST* has been used primarily in the snapshot mode on lensed QSOs in order to obtain accurate astrometry for the multiple images. While this has been valuable, a different program of *VLA* and deep *HST* imaging may be more effective. The most studied lens is 0957+561; many years of photometric monitoring have produced a reliable

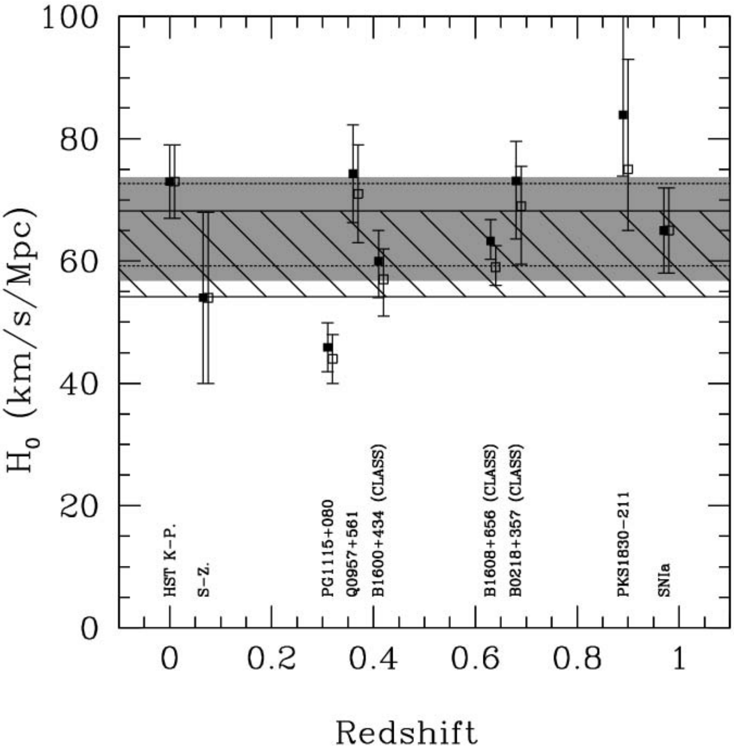


FIGURE 4. Various determinations of the Hubble constant are shown as a function of redshift, ca. 2000. The local ($z = 0$) value is from the *HST* Key Project using Cepheids. Estimates using the Sunyaev-Zeldovitch effect (clusters of galaxies at about $z = 0.1$–0.2) will improve in accuracy by at least a factor of two in the next few years. On the far right is shown the SN1a determination by Riess et al. (1998). The rest of the points are from lensed QSO estimates, which are expected to improve in both number and accuracy. The black points are for a cosmology with $\Omega_m = 0.3$ and $\Omega_\Lambda = 0.7$. The shaded band indicates the 2-σ limits determined from the lens sample (excluding PG1115+080). Error bars are 1-σ, but systematic errors dominate all estimates. Adapted from Koopmans & Fassnacht (1999).

estimate for the time delay. These observations are now being carried out for a number of multiply-lensed quasar systems and there is some hope that with a campaign of intensity monitoring and sufficient *HST* imaging—of a different kind than has traditionally been obtained—a robust global value for the Hubble constant will be obtained. The current data, however, are dominated by systematic error. This may be seen at a glance in Figure 4, where the error bars are purely statistical; the scatter is far larger. In fact, systematic error dominates all current estimates of H_0, independent of technique.

The systematic errors in the current gravitational lens determinations of H_0 are almost completely in the lens model. One must have a good model of the lens projected mass distribution, or the potential Ψ. The relative magnifications and positions of the various source images act as constraints, but generally there are too few constraints. The problem is that in the case of quasars lensed by galaxies there is a single bright source, the quasar. The lens potential is thus sampled at only a few points—where the images of the quasar appear. This is why 2-image lenses like 0957+561 and even many of the 4-image lenses have been so difficult. The mass distribution in the lens is sufficiently complex that it requires more than the several observable parameters for a full description. Bernstein

& Fischer (1999) give an excellent account of the degeneracies in the lens model for the case of 0957+561, and the enlarged error limit on H_0 when these degeneracies are marginalized against.

With sufficient time-delay measurements from flux monitoring, the systematic error is virtually all in the lens model potential. There are two main sources of lens model systematic error: the lens model in the strong lensing region, and the sheet mass degeneracy. Naturally, H_0 is sensitive to the details of the lens potential within the region of the observed lensed images. It is particularly sensitive to the radial profile of the lens potential in this region. Shallow mass profiles lead to lower values of H_0. Several images of the point source quasar, seen around the Einstein radius, do not sample this profile well. Models of the lens potential, for example, often assume what lens galaxy mass distributions should look like.

Generally, this problem may be cured if the source is resolved, or if there are multiple sources which are multiply imaged. In the absence of these deeper imaging data, imposing some kind of external constraint, such as constant mass-to-light ratio (if true), can help break the degeneracy. For example, the estimate for H_0 based on a simple parametric model for the lens potential of PG1115+080 moves from the value shown in Figure 4 to $H_0 = 65 \pm 5$ if a constant M/L constraint is imposed on the lens. Likewise, making use of all time-delay ratios for all pairs of QSO images can further constrain the lens potential since only one time delay is needed to determine the scale factor in the equation for $\tau(\vec{\theta})$. Clearly, a solution will be to obtain deeper data which detect morphological features in the host galaxy. Already, resolved arcs have been detected in several systems using *HST* red or near IR exposures (see Figures 2 and 3). A campaign of deep *HST* IR imaging of selected lensed quasar systems would almost certainly detect multiply imaged knots, resulting in significant reduction of systematic error.

Finally, the sheet mass degeneracy: a uniform sheet of mass of dimensionless surface density κ_s produces an effective potential

$$\Psi_s(\theta) = \kappa_s \theta^2$$

and the quasar image pattern is stretched by the factor $1/(1 - \kappa_s)$, leading to an overestimate of H_0 for a given measured time delay. This degeneracy may be removed by an external measurement of κ_s. Weak lensing (single distorted images of each resolved source) is ubiquitous since there are nearly a million background galaxies per square degree anywhere on the sky, and can yield moderate resolution maps of projected mass over wide areas. The Advanced Camera on *HST* would be ideal for this weak lens reconstruction of the overall projected mass density over a roughly ten arcminute field. Such observations would also help regularize the inner lens mass reconstruction. On a larger scale, weak lens shear measurements utilizing 100,000 galaxies over degree scales can be done at adequate resolution from ground-based telescopes, thus removing the mass sheet degeneracy on all relevant scales when combined with *HST* imaging.

With these systematics removed, there will be no impediment to a precision global measurement of the Hubble constant. Surviving errors in each system will be independent and the increasingly large numbers of lensed quasar systems which are being studied will give rise to ever more accurate H_0 values. There is hope in numbers.

2.2. *Constraining the cosmological constant*

The statistics of multiply imaged sources contains information on the mass distribution in the lenses as well as the underlying cosmology. Zwicky (1937) first estimated the probability of galaxies lensing distant sources (about 10^{-3}). Press & Gunn (1973) studied the probability that sources are multiply imaged in a universe closed by point masses.

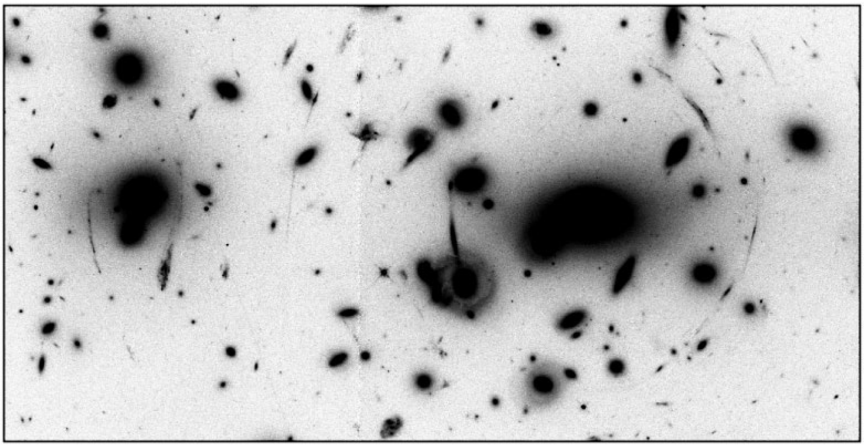


FIGURE 5. A Hubble Space Telescope image of the cluster of galaxies A2218, showing multiple "arcs": lens-distorted images of background galaxies. The mass of the dark matter in the cluster is far greater than the mass of all the stars. This mass bends light rays from distant galaxies, causing a tangential stretching of the images of the background galaxies. Credit: W. Couch, R. Ellis, and NASA.

There are sample selection and observational systematic errors which must be taken into account, and these were addressed by Kochanek (1991). Recently, purely radio selected samples of quasars such as the CLASS and JVAS surveys have avoided these effects (see Falco et al. 1998).

Either massive lenses (large Ω_m in galaxies) or larger distances to the quasar sources (large Ω_Λ) increase the strong lensing probability, so that with some plausible assumptions statistics on multiply lensed quasars may be used to set limits on $\Omega_\Lambda - \Omega_m$. The effects of complicated lenses (and adequately modeling them) on multiple-image statistics prevents high accuracy constraints for moderate values of Ω_Λ, but as Ω_Λ increases above 0.7 the lensing probability rises dramatically. Current upper limits on Ω_Λ of approximately 0.7 have been obtained by this statistical technique (Chiba & Yoshii 1999; Helbig 1999).

3. Clusters of galaxies lensing background galaxies

While it is arguably the most basic cosmological parameter, the Hubble constant is not the most interesting constraint on cosmology provided by strong lensing. The nature of dark matter and dark energy can be probed with lensing. Mass distributions of all sizes lens the background universe, and the larger the mass of the lens the more one potentially learns about the overall cosmic mass distribution. Indeed, for sources at sufficiently high redshift, the effects of lensing involve all forms of mass-energy, including the effects of a cosmological constant or quintessence. Figure 5 shows a good example of gravitational lensing (both strong and weak) by a cluster of galaxies. In cluster lenses the light bending angle is a factor of thirty higher than a typical QSO-galaxy lens, and the field of view must be correspondingly larger.

Cluster masses and their distribution with cosmic time, are very useful tools for cosmology. Comparisons of the cluster mass spectrum, particularly its evolution over cosmic time, with N-body simulations of dark matter structure formation for different cosmological models can be used to distinguish among these models. The average matter density of the universe Ω_m may be estimated by extrapolating observations of cluster luminosi-

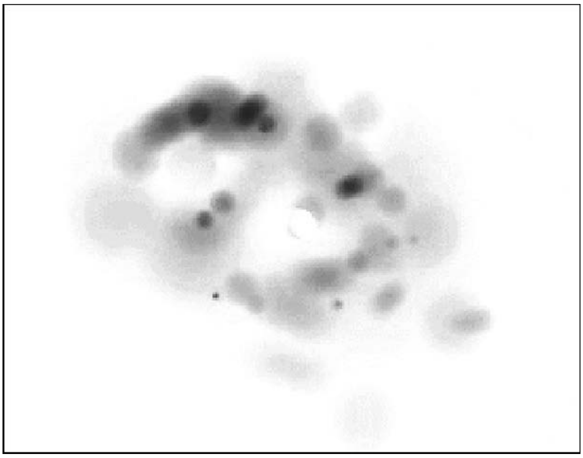


FIGURE 6. A reconstructed image of a star-forming galaxy at $z = 1.7$. The image was made by unlensing the observed gravitationally lensed images (see Figure 7 and discussion). The galaxy is 1 arcsec long.

ties to the rest of the universe and assuming the mass-to-luminosity ratio is the same for clusters and the universe as a whole. Similarly, the gravitational lens measurement of projected mass may be combined with a measurement of the baryonic density (from the SZ effect) plus the determination of the overall baryonic density from deuterium cooked in the big bang, yielding a more secure estimate of Ω_m.

The only direct method (avoiding reliance on baryons) to determine the entire mass distribution of a galaxy cluster is gravitational lensing. The observed distorted images of background galaxies leads to a map of the mass responsible for the distortion. There are various ways to reconstruct the lens mass map, depending on the number of observational constraints. The angular resolution of the mass map is proportional to the density of constraints on the plane of the sky. Strong lensing, with its multiple resolved images of a single source, can offer high mass resolution near the source image positions, but the resolution degrades far from these positions.

Previous work has used the temperature (or equivalently the cluster velocity dispersion) distribution as a proxy for the mass, but these make dynamical assumptions. Exploring mass structure to pin down Ω_m via direct observations of mass would significantly clarify cosmology. Recently, the first steps have been taken toward addressing three fundamental questions: what is the total mass of clusters, how is the mass distributed within clusters, and how do cluster masses evolve over time?

3.1. *High resolution maps of cluster mass*

A single background galaxy (see Figure 6), if placed by chance directly behind the cluster lens, is heavily distorted into one or more long arcs concentric with the mass centroid of the cluster. This type of strong lensing forms the basis of the highest resolution mass maps. In cases where multiple images of a source are created by the lens, the details of the position and distortion of these sub-images are highly sensitive to the projected 2-dimensional mass distribution within the lens. Parametric models of the lens mass can then be used to first unlens the sub-images to get an image of the source, and then ray trace light from this source past the lens in an iterative fit in the image plane for the lens mass parameters.

As an example, Figures 7 and 8 show an *HST* image and a high resolution mass map of the cluster 0024+1654, some three billion light-years distant, based on parametric

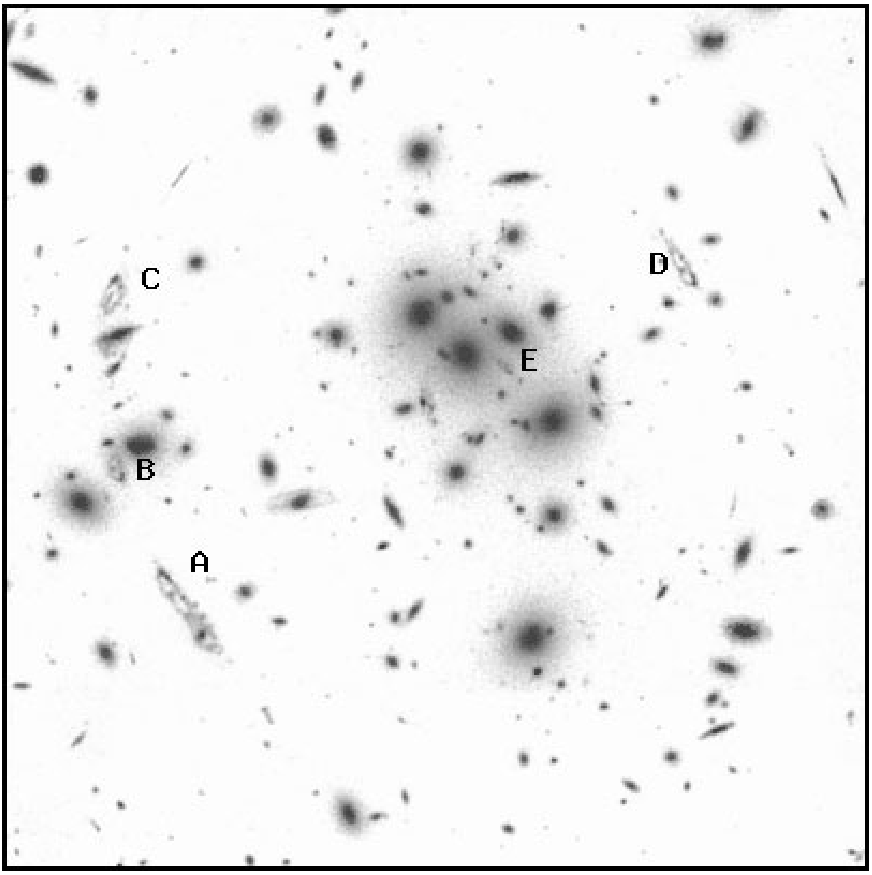


Figure 7. A *HST* image of the redshift 0.4 cluster CL0024+1654. This is one third of the WFPC2 image, showing the 3-arcmin strong-lensing core. Multiple images (A–E) of a $z = 1.7$ background galaxy may be seen. The existence of the only slightly demagnified image E so near the center of mass implies a soft mass core. For a color image, see: http://dls.bell-labs.com/dls/Current0024.jpg/.

inversion of the associated gravitational lens. This lens creates eight well resolved whole or partial sub-images of a background galaxy, seen in deep imaging with the *HST*. Excluding mass concentrations centered on visible galaxies, more than 98% of the remaining mass is represented by a smooth concentration of dark matter centered near the brightest cluster galaxies, with a 50 kpc soft core (using Hubble constant $H_0 = 67$ km s^{-1} Mpc^{-1}).

The dark matter distribution observed in CL0024 is far more smooth, symmetric, and nonsingular than in typical simulated clusters using the CDM model. Inside 160 kpc radius, the rest-frame mass to light ratio is 170 times the solar value, rising with radius. Because galaxies were brighter in the past, this translates to a mass-to-light ratio of 280 now. For scale, we would need around 1400 for an average cluster to extrapolate to a closed universe.

The mass core radius is smaller than most observed X-ray core radii in nearby clusters, suggesting that the X-ray gas may be less relaxed dynamically than the dark matter. However, recent high resolution N-body simulations of dark matter clustering develop mass cores even smaller than the observed mass cores. Perhaps there are other physical mechanisms at work in the cores of clusters, beyond gravitational instability and merging of dark matter halos, which are capable of damping the buildup of the mass at the

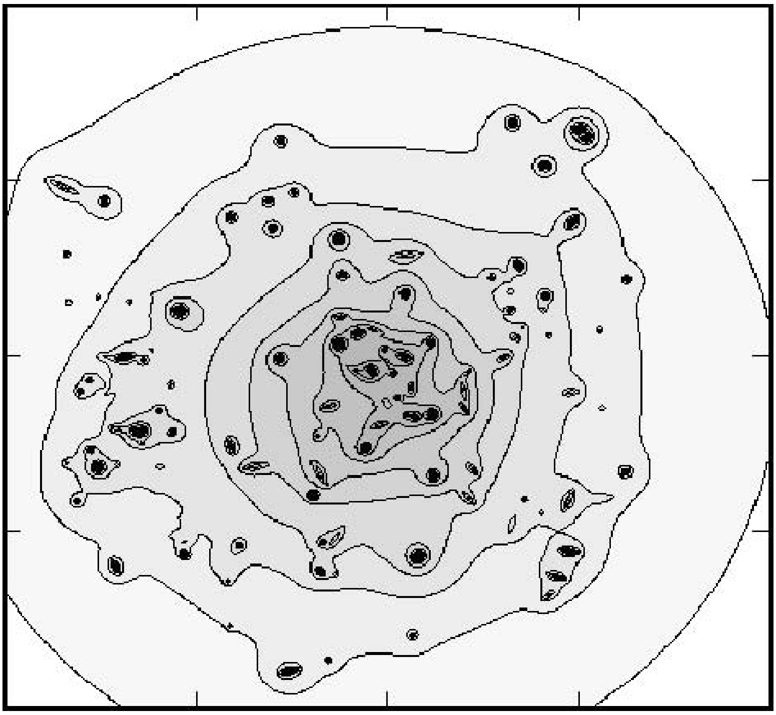


FIGURE 8. A contour map of the projected mass density in CL0024+1654, obtained by fitting the observed distorted images of the source galaxy. The concentrated dark matter halos of individual cluster galaxies can be seen. However, most of the mass in the cluster is in a smooth non-singular distribution. The mass map is 500 kpc wide (Tyson, Kochanski & Dell'Antonio 1998).

center. Due to this disparity between the N-body cold dark matter simulations and the observations in CL0024, as well as similar predictions of more sub-structure than observed in galaxies, some have considered the possibility that dark matter may be self-interacting (Spergel & Steinhardt 2000; Firmani, et al. 2000). The idea is that dark matter, while non-dissipative, may have a non-zero cross section for interaction with itself while having zero interaction with baryonic matter.

On the observational side, more high resolution mass maps of clusters are needed. Cases like 0024+1654, where a morphologically rich source is multiply lensed, are rare. However, deep *HST* imaging of clusters can often reveal many arcs from numerous background galaxies. Since there are over 50 high-redshift galaxies per square arcmin, this is guaranteed. One such example is shown in Figure 9, where over 55 arcs are found near the core of the $z = 0.18$ cluster Abell 1689. Originally studied via weak lensing (5950 low surface brightness blue arclets were found in a 140 arcmin^2 area), these *HST* data provide information on the profile of the mass in the core region sensitive to self-interacting dark matter. The mass centroid is within several arcsec of the smoothed red luminosity centroid. Mass follows smoothed light, outside the core region, with a rest-frame V band mass-to-light ratio of $400 \pm 60\ h\ (\mathrm{M/L}_V)_\odot$.

Like CL0024+1654, this cluster shows a soft mass core: the existence of *radial* arcs like 11 and 12 (see Figure 9) imply a projected mass profile which is soft inside that radius. Steeper mass profiles, of the sort now predicted by all cold dark matter simulations, move such radial arcs into the center and demagnify them. More clusters should be studied with deep multi-band imaging with *HST*. This would permit the use of color-redshifts

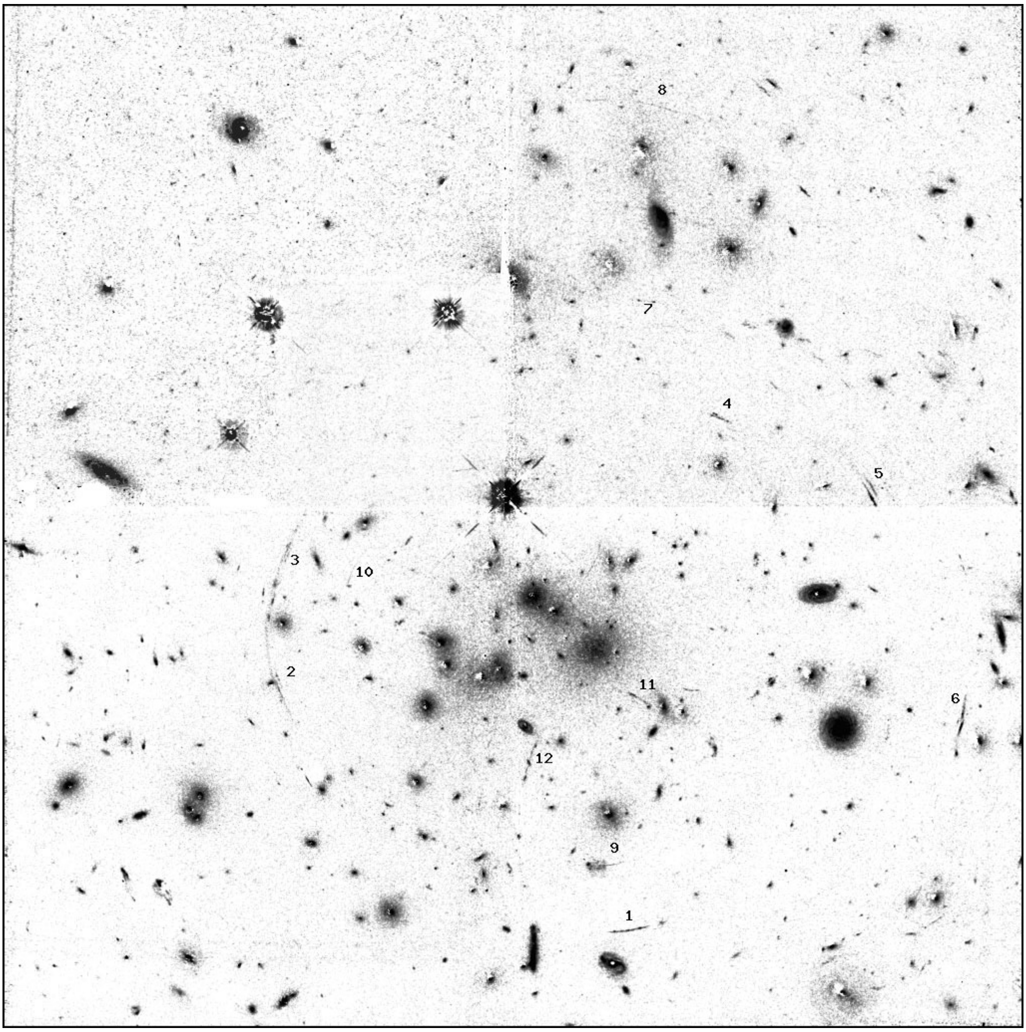


Figure 9. A *HST* image of the redshift 0.19 cluster Abell 1689, made by subtracting a fraction of an I image from the V image. This enhances blue-excess objects in the field, seen in this WFPC2 image covering the 3-arcmin strong-lensing region. Multiple images (arcs) of background galaxies may be seen. Some of the brighter arcs are numbered (1–12). The existence of radial arcs 11 and 12 imply a soft mass core. For a color image, see: http://dls.bell-labs.com/dls/CurrentA1689.jpg/.

for the arc systems, bypassing the most difficult phase of the analysis: sorting out which arcs belong to the same source.

3.2. *Cosmology from the cluster double-source lens*

Sources are lensed differently by the same lens mass, depending on their angular diameter distance. In turn, this depends on the cosmology. This provides yet another tool for probing cosmology via strong gravitational lensing. To use this tool one must first have a lens which is massive (creates multiple strongly lensed arcs from more than one background source) and which has its projected mass mapped at high resolution. Take CL0024 as an example. Several strongly magnified background sources can be seen at lower surface brightness in Figure 7.

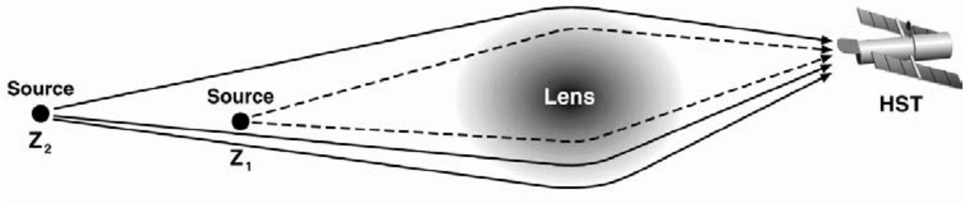


FIGURE 10. The double source lens cosmological test.

These arcs are undoubtedly images of other background source(s). If confirmed via deeper imaging and Keck spectroscopy, this will enable a purely geometrical determination of a simple algebraic combination of the cosmological parameters Ω_M and Ω_Λ. The lens mass and the relative angular diameter distances determine the strength of the lensing event, which is just the Einstein ring angular diameter for reasonably simple lens mass distributions.

Thus, the relative strength (ratio of Einstein ring diameters) of the lensing produced by this same foreground lens acting on sources at different background redshifts is a function only of a dimensionless combination of the angular diameter distances to the lens and two sources. Hence it is a function only of the cosmological parameters Ω_M and Ω_Λ plus the measurable redshifts of the lens and sources. The lens total mass, overall distance scale (H_0) and lens mass distribution essentially "cancel out" in the ratio of lensing strengths and thus provide a direct geometrical and largely model independent cosmological test.

3.3. *First maps of galaxy mass*

N-body simulations of cold dark matter structure formation have reached sufficiently high resolution for comparison with direct observations of mass morphology on 5 kpc scales in clusters of galaxies. Clusters at $z = 0.4$ are expected to have extensive dark matter substructure. Parametric mass reconstruction techniques could resolve mass components as small as $10^9\ h^{-1}\ \mathrm{M}_\odot$ near the projected positions of the arcs. Does light trace mass on all scales or is there dark mass segregation on some scales? The high surface density of galaxies in clusters create an ideal hunting ground for dark galaxies. White & Rees (1978) first pointed out that low mass compact galaxies, because of their low collision cross-section, survive in clusters of galaxies for a Hubble time.

Clusters like CL0024+1654 also represent a unique opportunity for a detailed study of the distribution of mass within known individual cluster galaxies. The giant arcs created by the lensing are projected on top of three bright foreground galaxies. Because the background galaxy's morphology is quite complex, one can measure the relative positions of all of the hot spots in each of the lensed images. This allows a measurement of the mass morphology of the most favorably placed galaxies.

4. Conclusions

I have reviewed some examples, past and future, of contributions to cosmology via strong gravitational lensing using the *HST*. The combination of high angular resolution and optical-IR imaging capability have been enabling. In the future, some ground-based telescopes with adaptive optics may fill some of the need, specifically for sub-arcminute fields. But the *HST* with the Advanced Camera and NICMOS will continue to contribute to this unique probe of cosmology, with its greatest contributions undoubtedly still to come.

It is a pleasure to acknowledge conversations with Chris Fassnacht, Brian McLeod, Paul Schechter, Ed Turner, and David Wittman. The parametric reconstruction of mass in clusters is being done with collaborators Greg Kochanski, Ian Dell'Antonio, and Alex Wissner-Gross.

REFERENCES

Bernstein, G. M. & Fischer, P. 1999 *AJ* **118**, 14.

Chiba, M. & Yoshii, Y. 1999 *ApJ* **510**, 42.

Falco, E. E., Kochanek, C. S., & Munoz, J. A. 1998 *ApJ* **494**, 47.

Firmani, C., D'Onghia, E., Avila-Reese, V., Chincarini, G., & Hernandez, X. 2000 *M.N.R.A.S.* **315**, 29.

Helbig, P. 1999 *A&A* **350**, 1.

Iwamuro, F., et al. 2000 *P.A.S.J.* **52**, 25.

Kochanek, C. S. 1991 *ApJ* **379**, 517.

Koopmans, L. V. E. & Fassnacht, C. D. 1999 *ApJ* **527**, 513.

Press, W. H. & Gunn, J. E. 1973 *ApJ* **185**, 397.

Refsdal, S. 1964 *M.N.R.A.S.* **128**, 307.

Schechter, P., et al. 1997 *ApJ* **475**, 85.

Spergel, D. N. & Steinhardt, P. J. 2000 *Phys. Rev. Lett.* **84**, 3760.

Turner, E. L., Ostriker, J. P., & Gott, J. R. 1984 *ApJ* **284**, 1.

Tyson, J. A., Kochanski, G. P., & Dell'Antonio, I. P. 1998 *ApJ* **498**, 107.

White, S. D. M. & Rees, M. J. 1978 *M.N.R.A.S.* **183**, 341.

Zwicky, F. 1937 *Phys. Rev.* **51**, 679.