Advances in Experimental Medicine and Biology 799

Natalia Maltsev
Andrey Rzhetsky
T. Conrad Gilliam *Editors*

Systems Analysis of Human Multigene Disorders

Springer

Advances in Experimental Medicine and Biology

For further volumes:
http://www.springer.com/series/5584

Natalia Maltsev • Andrey Rzhetsky
T. Conrad Gilliam
Editors

Systems Analysis of Human Multigene Disorders

Springer

Editors
Natalia Maltsev
Department of Human Genetics
Institute for Genomics and Systems
Biology, Computation Institute
The University of Chicago
Chicago, IL, USA

T. Conrad Gilliam
Department of Human Genetics
Institute for Genomics and Systems
Biology, Computation Institute
The University of Chicago
Chicago, IL, USA

Andrey Rzhetsky
Deapartment of Medicine
Department of Human Genetics
Institute for Genomics and Systems
Biology, Computation Institute
The University of Chicago
Chicago, IL, USA

ISSN 0065-2598 ISSN 2214-8019 (electronic)
ISBN 978-1-4614-8777-7 ISBN 978-1-4614-8778-4 (eBook)
DOI 10.1007/978-1-4614-8778-4
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013952921

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Understanding the genetic architecture underlying complex multigene disorders is one of the major goals of human genetics in the upcoming decades. Advances in whole genome sequencing and the success of high-throughput functional genomics allow supplementing conventional reductionist biology with systems-level approaches to human heredity and health as systems of interacting genetic, epigenetic, and environmental factors. This integrative approach holds the promise of unveiling yet unexplored levels of molecular organization and biological complexity. It may also hold the key to deciphering the multigene patterns of disease inheritance. Studies by countless groups have identified genes associated with many rare single gene (Mendelian) developmental disorders, but only limited progress has been made in finding the underlying causes for autism, schizophrenia, diabetes, and predisposition to cardiovascular disease, as they display complex patterns of inheritance and may result from many genetic variations, each contributing only weak effects to the disease phenotype. A major challenge to the detection and analysis of heritable patterns of disease susceptibility is the exponentially expanding search space required to explore all combinations of m genes or m genetic loci. Even the largest studies in human genetics are limited to the observation of several thousand meiotic events (i.e., the number of occasions in which the transmission of a given genetic variant from parents to offspring can be evaluated). Consequently, an exhaustive combinatorial search of even very small sets of multiple genetic loci leads to a huge burden of false-positive signals for every true-positive signal. This is because the number of statistical tests of significance performed on the same data set becomes too large to retain any statistical power. A biologically grounded approach is needed to constrain the plausible combinations of genomic regions that must be tested, drastically reducing the number of statistical tests.

The second obstacle to detecting multigenic inheritance is the need to understand relationships among the set of genes related to a disease and to determine how variations within each gene affect disease susceptibility. The influence of such diverse genetic interactions remains unknown. Genetic variations across multiple

interacting genes may affect the phenotype in a linear (additive) or nonlinear (epistatic) manner. Groups of interacting genes are likely to affect disease phenotypes via as-yet-unknown mixtures of both types of interaction. Furthermore, disease susceptibility may increase incrementally with increasing genetic variation or dichotomously via a threshold effect. Finally, the genetic causes of a given disorder may differ, in whole or in part, in different affected families. Although it is tempting to test the entire spectrum of inheritance models, this is currently impractical. The total number of possible models of inheritance involving m genetically interacting genes grows exponentially with m, further amplifying the exponential growth of the number of distinct gene sets of size m. A biology-grounded plan of prioritizing genetic models by their likelihood and systematic analysis of model space is critically important.

The third obstacle is that it is extremely hard and expensive to design large-scale studies that account for interactions between environmental factors and genetic variation in relation to disease phenotypes; a typical large-scale genetic analysis avoids explicit modeling and/or extensive measuring of environmental factors.

The scientific community has made enormous investments in developing the scientific infrastructure necessary to enable breakthrough discoveries of the primary biological risk factors for common disorders, such as diabetes, autism, susceptibility to cardiovascular diseases, and cancer. These investments have made possible investigations to understand disease-associated risk factors, on a scale unpredictable even a few years ago. Studies such as those based on genome-wide association have become standard and have led to a substantial number of discoveries. Although progress has been made in understanding some of these complex traits, our grasp of the patterns of risk are reduced to simple, short lists of weakly associated, noninteracting genetic variants that explain only a very low percentage of the estimated heritability. Some other challenges in constructing disease risk models are as follows: multigenic models of inheritance are usually ignored; genetic heterogeneity of commonly investigated phenotypes can lead to inefficient studies; and the wealth of information available on the biological system is generally ignored in constructing models of disease risk.

The volume is structured to introduce the major perspectives on intellectual and technological challenges facing systems-level translational medicine.

Chapter 1 addresses the need for the integration of clinical and genomic profiling with preventative healthcare. In recent years it became exceedingly clear that genotypes alone are insufficient to predict health outcomes, since they fail to account for individualized responses to the environment and life history. Integrative genomic approaches incorporating whole genome sequencing, transcriptomics, and epigenomics should be combined with clinical interpretation in the light of the triggers, behaviors, and environment unique to each person. Such integration will allow for an accurate prediction of the disease progression for a particular patient and significant improvement of personalized treatment strategies. The chapter discusses some of the major obstacles to implementation of such an approach, from development of

risk scores through integration of diverse omic data types, to presentation of results in a format that fosters development of personal health action plans.

Chapter 2 provides a comprehensive review of high-throughput data generation technologies employed by high-throughput systems-level biomedical studies with the emphasis on the next generation DNA sequencing platforms (NGS). NGS provides an inexpensive and scalable approach for detection of the molecular changes at the genetic, epigenetic, and transcriptional level. Furthermore, existing and developing single molecule sequencing platforms will soon allow direct RNA and protein measurements, thus increasing the specificity of current assays and making it possible to better characterize "epi-alterations" that occur in the epigenome and epitranscriptome. The authors describe novel approaches for generation and processing of genomic data, the development of the integrative models, and the increasing ubiquity of self-reporting and self-measured genomics and health data.

Chapter 3 addresses the challenges and best practices of high-throughput integrative medicine. Efficient mining of vast and complex data sets for the needs of biomedical research critically depends on seamless integration of clinical, genomic, and experimental information with prior knowledge about genotype–phenotype relationships accumulated in a plethora of publicly available databases. Furthermore, such experimental data should be accessible to a variety of algorithms and analytical pipelines that drive computational analysis and data mining. Translational projects require sophisticated approaches that coordinate and perform various analytical steps involved in extraction of useful knowledge from accumulated clinical and experimental data in an orderly semi-automated manner. The chapter explores cross-cutting requirements from multiple translational projects for data integration, management, and analysis.

Chapter 4 describes the algorithmic approaches for selecting and prioritizing disease candidate genes. The authors review the prioritization criteria and the algorithms along with some use cases that demonstrate how these tools can be used for identifying and ranking human disease candidate genes.

Chapter 5 presents a clinical perspective on systems-level translational research using lung cancer as an example. Lung cancer is no longer considered a single disease entity and is now being subdivided into molecular subtypes with dedicated targeted and chemotherapeutic strategies. The concept of using information from a patient's tumor to make therapeutic and treatment decisions has revolutionized the landscape for cancer care and research in general. Future directions will involve incorporation of molecular characteristics and next generation sequencing into screening strategies to improve early detection, while also having applications for joint treatment decision-making in the clinics with patients and practitioners.

This volume targets the readers who wish to learn about state-of-the-art approaches for systems-level analysis of complex human disorders.

The audience may range from graduate students embarking upon a research project to practicing biologists and clinicians working on systems biology of complex disorders and to bioinformaticians developing advanced databases, analytical tools,

and integrative systems. With its interdisciplinary nature, this volume is expected to find a broad audience in pharmaceutical companies and in various academic departments in biological and medical sciences (such as molecular biology, genomics, systems biology, and clinical departments) and computational sciences and engineering (such as bioinformatics and computational biology, computer science, and biomedical engineering).

We thank all the authors and coauthors who have contributed to this volume. We would like to extend our thanks to Melanie Tucker and Meredith Clinton of Springer US for their help in the preparation of this book.

Chicago, IL

Natalia Maltsev
Andrey Rzhetsky
T. Conrad Gilliam

Contents

Contributors

Gady Agam Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA

Bruce J. Aronow Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA

Sandhya Balasubramanian Department of Human Genetics, University of Chicago, Chicago, IL, USA

Rishi Batra 987400 Nebraska Medical Center, University of Nebraska Medical Center, Omaha, NE, USA

Eduardo Berrocal Department of Human Genetics, University of Chicago, Chicago, IL, USA

Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA

Daniela Börnigen Department of Human Genetics, University of Chicago, Chicago, IL, USA

Toyota Technological Institute at Chicago, Chicago, IL, USA

Apoorva Chawla Department of Medicine, Section of Hematology/Oncology, University of Chicago, Chicago, IL, USA

Bhadrachalam Chitturi Department of Computer Science, Amrita Vishwa Vidyapeetham University, Amritapuri Campus, Kollam, Kerala, India

Utpal Dave Computation Institute, University of Chicago/Argonne National Laboratory, Chicago, IL, USA

Inna Dubchak Genomics Division, Berkley National Laboratory, Walnut Creek, CA, USA

Bo Feng Department of Human Genetics, University of Chicago, Chicago, IL, USA

Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA

Greg Gibson Georgia Institute of Technology, Predictive Health Institute and School of Biology, Atlanta, GA, USA

T. Conrad Gilliam Department of Human Genetics, Institute for Genomics and Systems Biology, Computation Institute, The University of Chicago, Chicago, IL, USA

Thomas Hensing Department of Medicine, Section of Hematology/Oncology, NorthShore University Health System, Evanston, IL, USA

Department of Medicine, Section of Hematology/Oncology, University of Chicago, Chicago, IL, USA

Anil G. Jegga Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA

Natalia Maltsev Department of Human Genetics, Institute for Genomics and Systems Biology, Computation Institute, The University of Chicago, Chicago, IL, USA

Christopher E. Mason Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY, USA

The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY, USA

Sandra G. Porter Digital World Biology, Seattle, WA, USA

Ravi Salgia Department of Medicine, Section of Hematology/Oncology, University of Chicago, Chicago, IL, USA

Todd M. Smith PerkinElmer, Seattle, WA, USA

Dinanath Sulakhe Computation Institute, University of Chicago/Argonne National Laboratory, Chicago, IL, USA

Andrew Taylor Department of Human Genetics, University of Chicago, Chicago, IL, USA

Chao Wu Department of Computer Science, College of Engineering and Applied Science, University of Cincinnati, Cincinnati, OH, USA

Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Bingqing Xie Department of Human Genetics, University of Chicago, Chicago, IL, USA

Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA

Jinbo Xu Toyota Technological Institute at Chicago, Chicago, IL, USA

Cheng Zhu Department of Computer Science, College of Engineering and Applied Science, University of Cincinnati, Cincinnati, OH, USA

Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Chapter 1
Wellness and Health Omics Linked to the Environment: The WHOLE Approach to Personalized Medicine

Greg Gibson

Abstract The WHOLE approach to personalized medicine represents an effort to integrate clinical and genomic profiling jointly into preventative health care and the promotion of wellness. Our premise is that genotypes alone are insufficient to predict health outcomes, since they fail to account for individualized responses to the environment and life history. Instead, integrative genomic approaches incorporating whole genome sequences and transcriptome and epigenome profiles, all combined with extensive clinical data obtained at annual health evaluations, have the potential to provide more informative wellness classification. As with traditional medicine where the physician interprets subclinical signs in light of the person's health history, truly personalized medicine will be founded on algorithms that extract relevant information from genomes but will also require interpretation in light of the triggers, behaviors, and environment that are unique to each person. This chapter discusses some of the major obstacles to implementation, from development of risk scores through integration of diverse omic data types to presentation of results in a format that fosters development of personal health action plans.

It is a truth universally acknowledged that personal genome sequences will be a core component of individualized health care in the coming decades [23]. It is rightly claimed that genomic medicine should be predictive, personalized, preventive, and participatory, meaning that individuals will be encouraged to understand their own health risks and take preemptive measures to avert the onset of disease [5, 25]. Yet expert geneticists are debating whether genotypic predictors are now or will ever be more predictive than family history and clinical indicators [17, 28, 62], and there is reasonable skepticism surrounding causal inference from rare deleterious variants

G. Gibson (✉)
Georgia Institute of Technology, Predictive Health Institute and School of Biology, Atlanta, GA 30332, USA
e-mail: greg.gibson@biology.gatech.edu

N. Maltsev et al. (eds.), *Systems Analysis of Human Multigene Disorders*, Advances in Experimental Medicine and Biology 799, DOI 10.1007/978-1-4614-8778-4_1,

[37, 39]. So while tremendous progress is being made toward routine incorporation of whole genome sequence analysis for rare congenital disorders detected at birth [4, 22, 40, 56], and in cancer diagnosis and prognostics [9, 10, 18, 51], broader application to the complex common diseases that eventually afflict most adults remains to be introduced. The purpose of this chapter is to argue that the gap between hype and reality [53] needs to be addressed on two fronts: recalibration of expectation from prediction to classification and incorporation of functional genomic data into integrative predictive health.

The WHOLE approach encoded in the acronym for Wellness and Health Omics Linked to the Environment also places emphasis on the concept of wellness. Whereas the focus of most western medicine is on curing illness, universal public health strategies should attend more to disease prevention. As one of the leaders of this movement, Dr. Ken Brigham at Emory University remarks [7], the goal of health care should be to assure that "as many of us as possible should age with grace and die with painless dignity of natural causes." Our vision at the Center for Health Discovery and Well Being in Atlanta [6, 47] is that genomic data will be integrated into primary medical care precisely for this purpose, to help people make better lifestyle choices that promote the maintenance of good health.

There are three major challenges we see that need to be confronted, which are discussed successively below. The first is the development of genomic classifiers that explain a sufficient proportion of the variance in disease risk to be informative. In the near future, these will be genotype based, incorporating rare and common variants, clearly utilizing advanced statistical methodologies but also requiring adjustment for population genetic differences [15] and family structure [49]. The second is integration of sequence data with other genomic data types [19, 25, 33], such as transcriptomic, epigenomic, and metabolomic profiles, as well as with relevant clinical and biochemical measures and family history data. Whether or not the environment can be directly incorporated as well is an open question [2], though it can be argued that functional genomic data captures lifetime environmental exposure indirectly. The third challenge is working out how to present all of this data to healthy adults in a manner that is understandable and sufficiently actionable that they will commit to positive health behaviors. To this end I conclude with an outline of one strategy that is likely to involve the training of a new generation of professional genomic counselors.

1.1 Genomic Classifiers

The foundation of genomic classification is always likely to be genotypic. Single nucleotide polymorphisms identified through genome-wide association studies [28] or classical candidate gene approaches can be combined to more accurately discriminate cases and controls than single locus classifiers [61, 63]. The simplest multivariate scores are allelic sums, where the number of alleles that is associated with disease is tallied across all identified loci. For n loci, the score theoretically ranges from zero to $2n$, and the distribution is normal, but it will be skewed as a

function of the allele frequency spectrum. Few individuals will have extreme values, but under a liability threshold model, it is assumed that individuals with scores at the top of the range are at the most elevated risk of disease.

Risk can be modeled as a function of the score as a predictor in the same sense as Framingham risk scores predict likelihood of onset of disease in a given time period [8, 59], or more simply individuals above and below an appropriate value can be classified as high or low risk. For type 2 diabetes, a classifier based on 18 loci established that individuals with the top 1 % of simple allelic sum scores (25 or more risk alleles) have quadruple the risk relative to the bottom 2 % (fewer than 12 risk alleles) and slightly more than double the risk of the general population [30; see also 58]. This measure only marginally improves on the Framingham risk score for diabetes [60] and alone does not approach it for predictive power. However, at least in our CHDWB study the two measures (allelic sum and FRS) are only mildly correlated (unpublished observation), and so it is interesting to ask whether extreme genotype scores may suggest an alternative mode of diabetes risk.

A slightly more sophisticated approach is to weight the allelic scores by the magnitude of their effect. If one allele has a relative risk of 1.4, then it should have twice the impact of one with a relative risk of 1.2. In practice, it is not clear that weighted allelic sums improve on simple ones (Fig. 1.1a), perhaps reflecting the small amount of variance explained by current models built with variants that in general collectively explain no more than 20 % of disease risk. There is also likely to be large error in the estimation of individual allelic effects both due to sampling biases and incomplete LD between tagging SNPs and unknown causal variants. Nevertheless, for type 1 diabetes, a multiplicative allelic model based on 34 loci that collectively explain 60 % of the expected genetic contribution has been introduced [14, 44]. A score with a sensitivity of 80 % is achieved in 18 % of the population even though only less than half of one percent is type 1 diabetic. However, the positive predictive value remains fairly low since the false positive rate still exceeds 90 %. It seems that for rare diseases (less than 1 % of the population), it is unlikely that genotypic measures will ever be predictive in a clinical setting. Nevertheless, as a screening tool, there may be enormous financial and medical value in focusing resources on the highest risk portion of the population and excluding those least at risk from unnecessary surveillance or treatment.

If allele sums are used, it also makes sense to attempt to weight scores by allele frequencies. Two individuals may have the same score, but if one of them has most of the risk attributed to alleles that are not typically the risk allele in the population, whereas the other has the common high-risk variants, then it stands to reason that the former is likely to be at elevated overall relative risk. This is illustrated in Fig. 1.1b. An obvious way to achieve the weighting is to convert relative risks into odds ratios, compute the log sum of those odds, and regenerate a probability of disease [35]. Starting with a baseline risk for the relevant gender, ethnicity, and age group, each successive allele adds to or subtracts from the log odds, which are a function of the allelic effect and frequency.

The immediate problem with this approach is that it is susceptible to variation in allele frequencies among populations. Two people with identical weighted allelic sums may nevertheless have very different relative risks according to whether they are, for example, of African, Asian, or European descent (Fig. 1.1c). Somewhat

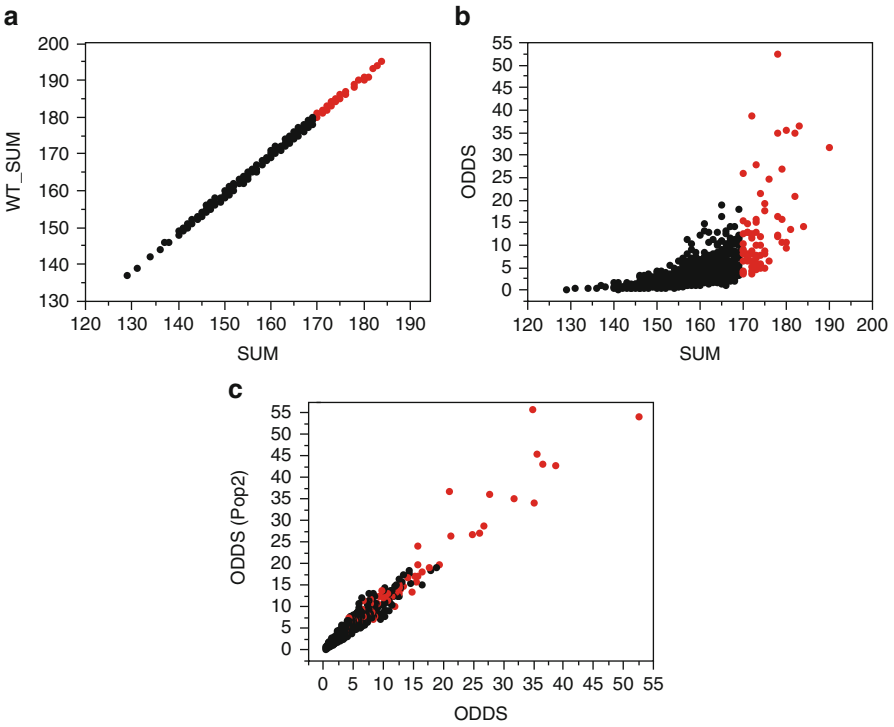


Fig. 1.1 Comparison of risk scores. The three *x*–*y* plots compare risk scores generated by three different methods, applied to a simulated dataset consisting of 200 disease SNPs measured in 1,000 people. The alleles range in risk allele frequency from 0.1 to 0.9 with a bias toward lower frequencies, and effect sizes were drawn from a normal distribution with mean of zero and standard deviation of 0.07. (**a**) Comparison of simple allelic sum score and weighted allelic sum score, showing a modest effect of weighting the sum by the effect size. Red points highlight individuals in the top decile of scores. (**b**) Comparison of simple allelic sum score and probability calculation from odds ratios obtained following the method in Morgan, Chen, and Butte [35] which computes the probability of disease from the summation of log odds ratios that are necessarily conditioned on the allele frequency. Despite increased variance of the score reflecting the multiplicative nature of the risk assessment (due to summation of log odds), the correlation in ranks is strong. (**c**) Comparison of probability scores for the same data as in (**b**) with computations assessed after randomizing the frequencies of one-quarter of the alleles, showing how population structure potentially affects disease risk assessment even where allelic effect sizes are assumed to be constant

paradoxically, heterozygosity at a single contributing locus can either increase or decrease the odds in different ethnicities, according to whether the risk allele is rare or common in either population. Accommodations can be made by deriving separate multi-allelic scores for each ethnicity, but an additional complication arises where admixture (population mixing) exists, which is the norm in contemporary America at least. Perhaps risk scores should be adjusted by the allelic frequencies expected of individuals with the observed mixture of ethnicities, but a case for local ancestry adjustment with phased genomes can be made [54, 55], and then the issue of the

appropriate baseline prevalence arises. It is not yet clear how much of an issue this is, and clearly much more research needs to be done, likely also including attention to geographically structured cultural and environmental modifiers of prevalence.

Finally, predictors and classifiers that do not assume additive effects of GWAS hits are being introduced. Sparse factorization and machine-learning approaches offer very powerful approaches that generate scores, incorporating SNPs that do not have strong univariate associations, or whose effects are conditional on other terms in the model [1, 3, 24, 29]. Often scores are developed purely as mathematical abstractions, though the interpretation is that they incorporate cryptic epistasis (genotype-by-genotype interactions) as well as environment or gender-specific interactions [48]. In these cases, there is always the assumption that the conditions and effects are consistent across populations. Again, it is not yet clear how reliable this assumption is and hence how transitive machine-learning based scores typically will be.

1.2 Integrating Functional Genomic and Clinical Data to Capture Environmental Contributions

Irrespective of the nature of the risk score, the second major challenge is to combine these into an overall personal health profile. A key insight is that the extensive comorbidity of diseases establishes the expectation that genotypic risks should covary [42, 52]. Given risk scores for dozens or even hundreds of diseases, further mathematical manipulations may facilitate gains in prediction or classification accuracy that borrow power from across diseases. At the current stage of development of personal genomic medicine, there is insufficient data to discern robust patterns of covariance, with the exception of autoimmune diseases that share common polymorphisms [32, 46]. So long as individual disease risk scores only capture a minor fraction of the genotypic risk, they are unlikely to capture to true architecture of comorbidity, but presumably this will change as more comprehensive predictors are developed.

In the mean time, Ashley et al. [2] presented a mode of visualization of combined risk that suggests how path analyses might integrate univariate risk scores. This is reproduced in Fig. 1.2b focusing just on a half dozen common disease conditions mostly related to metabolic syndrome. On the left (Fig. 1.2a), the so-called risk-o-grams [2, 13, 16] show how baseline risk for these conditions is modified by a hypothetical individual's genotypic risk. The point estimates should not be over-interpreted, the more important information being contained in the sign and magnitude of the genetic contribution. These are modified by comorbidity and redrawn in the form of the size of the font on the right, where larger circles represent increasingly elevated risk due to the individual's genotypes and the disease interactions. Interrelated disease conditions are connected by directed edges where, for example, the likelihood of developing cardiovascular disease is increased by the person's elevated risk of obesity but decreased by their low hypertension risk. Unfortunately, we do not yet have the tools to estimate the strengths of the connections, and much theoretical work on the optimal multivariate integration strategy remains to be performed.

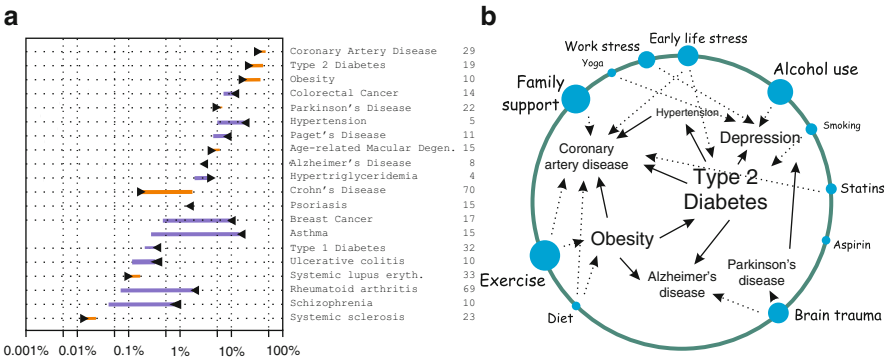


Fig. 1.2 Risk-o-grams. Following Ashley et al. [2], a hypothetical risk-o-gram (**a**) shows how genotypic risk can be used to generate a point estimate of probability of disease conditioned on the population prevalence. The figure shows a hypothetical risk assessment on the log scale for 20 diseases where the black triangles show the prevalence for the individual's gender, ethnicity, and age group, pointing to the right if genotype is predicted to increase risk or left if it decreases risk relative to the population average. The horizontal bars show the degree of genotypic effect, where, for example, Crohn's disease risk is highly elevated, but asthma and breast cancer are reduced. (**b**) These risks need to be combined, recognizing the comorbidity matrix of disease and the influence of environmental factors, including dietary and psychological stressors, exercise patterns and drug usage, and personal history of illness. The modified risk for each condition conditioned on the matrix of influences is represented by the size of the font. Although we are a long way from being able to generate robust assessments, the figure implies that classification into high- and low-risk classes should be feasible in the near future

Just as importantly, the grand circle surrounding the disease prediction network shows that the environment must also be incorporated into computations. In this case, the individual's heavy alcohol usage and lack of exercise also increase their risk of metabolic syndrome, as does a history of early life stress coupled with low family support and high work pressure. It is apparent that they are already taking statins and eating a low-fat diet to offset some of the risk, and regular yoga practice may help qualitatively. A traumatic brain injury suffered in a car accident as a child may have been a trigger that cannot be factored into population-based measures of risk, but it also feeds into likely cognitive decline with age. Again, it is not yet obvious how these environmental risks should be formulated from a statistical perspective. Drug usage can conceivably be incorporated as a cofactor in the computation of individual risk scores, but it is less obvious how to model diet and mental stress, or what the appropriate multivariate framework may be. A further advantage of this visualization is that it readily lends itself to dynamic representation of how lifestyle modifications may reduce the risk of key diseases, as individuals can observe projected changes in risks if they adopt new health behaviors.

Another aspect of the environment that we may endeavor to incorporate is cultural and geographic differentiation. Perusal of the Centers for Disease Control (CDC) database of morbidity (see, e.g., http://www.cdc.gov/cancer/dcpc/data/state.htm for cancer data) shows that most diseases have very different prevalence

according to the location within the United States. An excellent example is the well-known southern stroke belt [11] stretching from Louisiana across Alabama to Georgia and the Carolinas, but cancer incidence and many other diseases vary from region to region. Undoubtedly, rural and urban lifestyles impact disease risk, and we have shown that they also impact peripheral blood gene expression profiles [26, 36], while emerging data also suggests differences in the microbiome [64]. Most readily, this type of information could be incorporated into risk prediction already at the level of baseline prevalence, which might be assessed regionally rather than simply by gender and ethnicity. Of course someone who moves from Manhattan, New York, to Manhattan, Kansas, does not modify their risk overnight, so yet another obstacle to absolute risk prediction lies in assessing the perdurance of lifestyle effects and the impact of life stage. Notably, there is accumulating evidence that early life stress is among the biggest risk factors for a wide range of diseases, particularly in lower socioeconomic strata [21, 34, 41].

Another unresolved issue is to what extent genotype-by-environment interactions need to be taken into account in risk evaluation. There is very little evidence from GWAS that G×E is either prevalent or of sufficient magnitude to be important components of population variance [57], notwithstanding occasional reports, for example, of smoking by nicotinic acetylcholine receptor polymorphism interactions with lung cancer [65] and of arsenic by solute carrier interactions for bladder cancer [27]. This is surprising given the prevalence of both genotypic and environmental effects on gene expression [26]. Supposing that low transcript abundance for a particular gene in a relevant tissue contributes to disease risk, those homozygous for a low expression *cis*-regulatory polymorphism, in an environment where expression is significantly reduced as well, will constitute the most at-risk group. Under a liability model, G×E for disease is plausible, even in the absence of interaction effects between the genotype and gene expression. However, large eQTL effects do not translate into large disease effects measured in case-control GWAS settings. It is possible that genotypic risk score-by-environment interactions will be observed, but such studies are yet to be performed. Furthermore, perhaps the more important mode of interaction is with individualized effects, such as triggers (accidents, transient stresses) that either are not captured in epidemiological surveys or have such high variance that interaction effects do not attain significance in population-scale studies.

All of these considerations add uncertainty to risk assessment and raise the question of whether it might not be better to measure the impact of the environment biochemically. The notion is that a person's individuality results from the longitudinal interaction of their genome with all of the above lifestyle and environmental factors. These influences mediate disease risk ultimately by modifying metabolism and physiology, which in turn are a function of gene expression, which is subject to epigenetic modification. Consequently, measurement of the metabolome, transcriptome, and epigenome (e.g., chromatin methylation) should provide parallel omic information of high relevance to health care [25]. This systems biology approach is much hyped [53], but many would argue that it has yet to provide the clinical or mechanistic insights that have stemmed from genotype and sequence-based genomic medicine. A major limitation of course is that only a few tissues, principally

peripheral blood or sometimes adipose biopsy, are readily available for high-throughput analysis. Blood does reflect immune and metabolic function and possibly mirrors psychological stressors [20, 31], so there is undoubtedly much to be learned from characterization of the sources of variance, and major advances in predictive health can be expected from this approach in the next decade.

Just to briefly highlight two strategies from our own work. First, characterization of extremes of individual transcript abundance detected by either microarray or RNASeq analysis of individuals is in many ways equivalent to rare deleterious coding variant detection from sequencing. We do not yet know how to read regulatory variation directly, but this is unnecessary if it can be directly demonstrated that an RNA (or protein) is not expressed in a particular individual. Association of such differential expression with phenotypes is subject to the same caveats as rare variant association analysis. Second, transcriptional variation is highly structured and characterized by major axes that represent aspects of lymphocyte function such as B and T cell signaling, antiviral responsiveness, and inflammation [45]. This variability is evident in the principal components of peripheral blood gene expression, but also appears in modules and axes of variation that are captured by the expression of biomarker genes [12], or blood informative transcripts. We postulate that the level of activity of gene expression in these axes will be found to correlate with aspects of immune and metabolic health.

1.3 Presenting and Interpreting Genomic Risk for Wellness

The third great challenge is to present genomic indicators of disease risk to healthy individuals in a manner that will help them to make sensible health behavior choices. This is one of the major goals of the emerging discipline of medical informatics. Risk-o-grams (Fig. 1.2a) are an excellent starting point since they present risk both in absolute terms as well as apportioning the genetic contribution relative to the population average. However, they have some obvious drawbacks, not least of which is the overwhelming number of assessments, many of which are for rare conditions or are clinically not actionable. They also fail to convey a sense of the error associated with risk assessments: we are used to the notion that heavy smoking more than doubles your lifetime risk of lung cancer, yet know heavy smokers who never get lung disease and never-smokers who do. Inevitably inappropriate presentation of genetic risks will engender skepticism toward genomic medicine that may undermine the certain benefits that stand to be realized.

For this reason, in the context of wellness, classification is the more appropriate emphasis than prediction. Classification into very-high-, high-, normal-, low-, and very-low-risk levels should help individuals to focus on those aspects of their health that will benefit from close attention. It draws attention away from the myriad statistical issues discussed above, instead promoting joint consideration of genetic and clinical measures. Furthermore, it is consistent with a simplification of risk presentation in health domains that recognize patterns of comorbidity and leverage existing

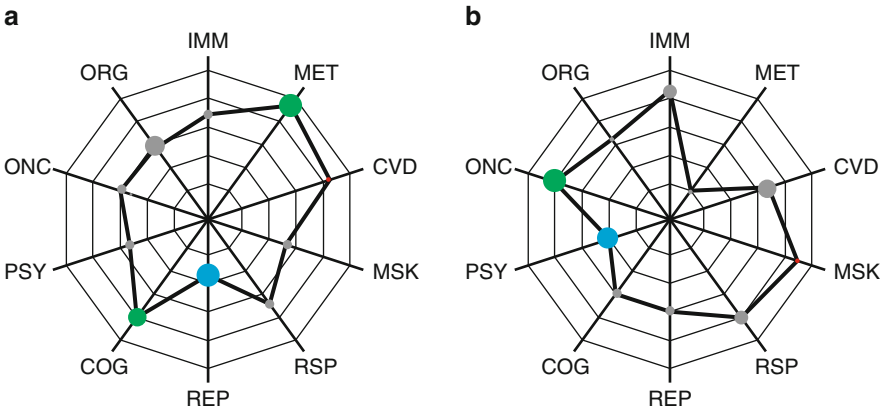


Fig. 1.3 Spider-web plots representing genomic and clinical risk in ten health domains for two hypothetical individuals. Genomic risk scores, generated by combining genotypic and functional genomic evidence, place each person in one of five risk classifications from very high (*outer band*) to very low (*inner band*) in ten health domains (*IMM* immunological, *MET* metabolic, *CVD* cardiovascular, *MSK* musculoskeletal, *RSP* respiratory, *REP* reproductive, *COG* cognitive, *PSY* psychiatric, *ONC* oncological, *ORG* organ failure). Clinical risk assessments generated from comprehensive medical examinations as well as personal and family history of disease are indicated by the size of the dots in each axis. Colors represent discordance between genomic and clinical risk as these situations are likely to be of greatest interest for individuals, alongside concordance for high risk, as they develop health action plans. Details and actual individual examples are described in Patel et al. [43]

modes of health assessment. At the Center for Health Discovery and Well Being, we are promoting the idea that comprehensive clinical evaluation annually, starting in the fourth decade of life, will foster prevention over reaction as it engages individuals in their own health choices [43]. Figure 1.3 suggests one mode of presentation of genomic data that may be incorporated into the preventative medicine framework.

Each radiating axis on the spider-web plots represents one of ten health domains. The bold polygon crosses each axis at a point, representing genomic risk in that domain (points further out mean higher risk), while the size of the circle at that point represents the observed clinical risk and/or evidence for disease. A quick glance at the spider-web plot tells an individual where they have high or low genetic and clinical risk. Areas of continuity between genetics and clinical data are highlighted as green dots. Discontinuities may be even more interesting. Those indicated in red where genetic risk is high but there is no sign of clinical danger (cardiovascular disease for A and musculoskeletal decay for B) suggest situations where the individual may pay close attention despite current good health. By contrast, situations where the genetic risk is low but clinical signs are not hopeful (respiratory disease for the smoker A and psychiatric problems for the socially isolated person B) may suggest that lifestyle changes are likely to have an impact. The main objective of this combined genomic and clinical classification is not to predict disease but to help individuals focus attention on areas where they should concentrate their health-related behaviors and surveillance.

The proposed ten common health domains are as follows:

- Immunological, including autoimmune (type 1 diabetes, multiple sclerosis, SLE, arthritis), inflammatory (especially bowel diseases), and infectious (viral and microbial) disease susceptibility, many of which show comorbidity and all of which should be related to gene expression in various blood cells
- Metabolic syndrome, generally referring to obesity and either hyperlipidemia or high blood glucose, leading to type 2 diabetes, and encompassing impaired insulin production and sensitivity
- Cardiovascular, primarily atherosclerosis and hence related to metabolic dysfunction, but also including cardiomyopathy, arrhythmia, and heightened risk of myocardial infarction or stroke
- Respiratory discomfort, namely, asthma, COPD, and fibrosis, all of which are exacerbated by smoking and call for attention to genotype-by-environment interaction
- Musculoskeletal problems, such as low bone density, chronic back pain, and muscle weakness or wasting, which are a primary cause of reduced quality of life for large percentage of the elderly
- Mental health, manifesting as depression and/or anxiety in an increasingly alarming percentage of adults, but also including schizophrenia, autism spectrum, and attention deficit disorders in adolescents and young adults
- Cognitive decline, whether due to Alzheimer's disease, Parkinson's disease, or generalized senile dementia and expected to become the major public health burden of the twenty-first century
- Cancer risk, assessed from family history and possibly peripheral blood biomarkers
- Organ malfunction, which is unlikely to have a common genomic foundation but collectively loss of eyesight, hearing, and renal and liver function, are a major source of morbidity
- Reproductive health, namely, the capacity to conceive and maintain pregnancy or to produce fertile sperm, but also including endometriosis and other causes of uterine discomfort

Pharmacological variation, for both toxicity and responsiveness to specific drugs, is also an important aspect of genomic health, sometimes having a simple genetic basis (e.g., warfarin [50]) but generally as complex as disease risk [38]. This is not by any means an exhaustive list of disease but is meant to capture the major domains that concern adults as they enter middle age and begin to make lifestyle modifications in response to self-perception of personal health concerns. Genome-wide association studies have been performed for specific diseases in each domain, and thousands of variants are available for generation of risk scores. Similarly, relevant clinical measures can be taken during routine medical checkups or as part of a dedicated wellness program such as the CHDWB and collectively generate risk profiles in these ten domains as well.

An immediate concern is how to collapse disparate genotypic and clinical risk scores into summary measures of risk for the various domains. For clinical measures, z-scores place each person in relative risk categories with those within one

standard deviation of the mean being at intermediate risk, those between 1 and 2 standard deviations at high (or low) risk, and everyone at the extremes at the very-high- or low-risk categories. A similar strategy could be applied to genotypic risk, or thresholds can be established based on the risk score distributions. Geometric means might be used to combine multiple scores, enhancing the relevance of individual high-risk values. My concern here is not with the optimal mode of collapsing but rather to suggest how spider-plot or similar visualization might be interpreted.

After consulting the spider-web plot with a physician or other health professional, the next step would be to examine the contributing risk factors in more detail. Consider the three examples. (1) In the cardiovascular domain, individual B in Fig. 1.3 has intermediate overall risk, but close examination shows that she is discordant for high blood pressure and lower than average genotypic risk of hypertension. This may suggest that some aspect of lifestyle, either high levels of job stress or a high salt diet, is responsible, and the low genetic risk might in some cases provide impetus for the individual to address the root cause. (2) Person A is concordant for obesity and high genetic risk of obesity, both of which produce high scores in the metabolic domain. Rather than accepting this as a fait accompli, with appropriate counseling she may learn that much of the genetic risk is due to neurological factors rather than any deficit in metabolic enzyme function, and this may help him to seek guidance in controlling dietary compulsions. (3) Another individual may be discordant in the organ failure domain for high genetic risk of age-related macular degeneration, but as a 70-year-old with above average eyesight has paid no attention to the possibility that he may soon suffer from loss of vision. Knowing the genetic risk, he will now have regular eye exams and follow emerging guidelines directed at preventing onset of the disease.

As discussed earlier, I envisage that genomic risk assessment will eventually incorporate transcriptional, epigenomic, and metabolic measures. The costs involved will be an obstacle for the foreseeable future, and it is not clear who will pay. It is nevertheless not difficult to see how a few thousand dollars spent on genomic analyses in middle age may save tens or hundreds of thousands of dollars in acute medical care for people approaching retirement age. Employers stand to benefit from reduced absenteeism and elevated productivity, and economic modeling suggests that the savings can be substantial. Scientific demonstration of the clinical efficacy of joint genomic and clinical profiling will likely take thousands of case studies over several years, a daunting challenge, but given the stakes, one that must be taken on.

1.4 Conclusion

Assuming success of the WHOLE paradigm, there will also be a need for training of a new class of health-care professional. A few genetic counseling programs are beginning to provide training in the interpretation of genome sequences. At the CHDWB, we have developed a Certificate program for Health Partners who consult with participants on the interpretation of their clinical profiles and help them to formulate personal health action plans. The combination of advanced genetic

counseling with a health partner is expected to yield genomic counselors, masters level professionals who will work alongside physicians, dieticians, personal trainers, and clinical geneticists to provide people who care to take advantage of the wealth of information implicit in genomic medicine, with a path to health maintenance and extended well-being.

References

1. Abraham G, Kowalczyk A, Zobel J, Inouye M (2012) Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. Genet Epidemiol 2012 Epub ahead of print
2. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE et al (2010) Clinical assessment incorporating a personal genome. Lancet 375:1525–1535
3. Ayers KL, Cordell HJ (2010) SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genet Epidemiol 34:879–891
4. Bick D, Dimmock D (2011) Whole exome and whole genome sequencing. Curr Opin Pediatr 23:594–600
5. Bousquet J, Anto JM, Sterk PJ, Adcock IM, Chung KF et al (2011) Systems medicine and integrated care to combat chronic noncommunicable diseases. Genome Med 3:43
6. Brigham KL (2010) Predictive health: the imminent revolution in health care. J Am Geriatr Soc 58(Suppl 2):S298–S302
7. Brigham KL, Johns MME (2012) Predictive health: how we can reinvent medicine to extend our best years? Basic Books, New York, NY
8. Brindle P, Emberson J, Lampe F, Walker M, Whincup P, Fahey T, Ebrahim S (2003) Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study. BMJ 327:1267
9. Cancer Genome Atlas Network (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474:609–615
10. Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. Nature 487:330–337
11. Casper ML, Wing S, Anda RF, Knowles M, Pollard RA (1995) The shifting stroke belt: changes in the geographic pattern of stroke mortality in the United States, 1962 to 1988. Stroke 26:755–760
12. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C et al (2008) A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity 29:150–164
13. Chen R, Butte AJ (2011) The reference human genome demonstrates high risk of type 1 diabetes and other disorders. Pac Symp Biocomput 2011:231–242
14. Clayton DG (2009) Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS Genet 5:e1000540
15. de Roos AP, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. Genetics 183:1545–1553
16. Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C et al (2011) Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. PLoS Genet 7:e1002280
17. Do CB, Hinds DA, Francke U, Eriksson N (2012) Comparison of family history and SNPs for predicting risk of complex disease. PLoS Genet 8:e1002973
18. Ellis MJ, Ding L, Shen D, Luo J, Suman VJ et al (2012) Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature 486:353–360
19. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F et al (2008) Genetics of gene expression and its effect on disease. Nature 452:423–428
20. Glatt SJ, Tsuang MT, Winn M, Chandler SD, Collins M et al (2012) Blood-based gene expression signatures of infants and toddlers with autism. J Am Acad Child Adolesc Psychiatry 51:934–944

21. Gluckman PD, Hanson MA, Cooper C, Thornburg KL (2008) Effect of *in utero* and early-life conditions on adult health and disease. N Engl J Med 359:61–73
22. Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ (2012) Next generation sequencing in clinical medicine: challenges and lessons for pathology and biomedical informatics. J Pathol Inform 3:40
23. Hamburg MA, Collins FS (2010) The path to personalized medicine. N Engl J Med 363:301–304
24. Han F, Pan W (2010) Powerful multi-marker association tests: unifying genomic distance-based regression and logistic regression. Genet Epidemiol 34:680–688
25. Hood L, Balling R, Auffray C (2012) Revolutionizing medicine in the 21st century through systems approaches. Biotechnol J 7:992–1001
26. Idaghdour Y, Czika W, Shianna KV, Lee SH, Visscher PM et al (2010) Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nat Genet 42:62–67
27. Karagas MR, Andrew AS, Nelson HH, Li Z, Punshon T et al (2012) SLC39A2 and FSIP1 polymorphisms as potential modifiers of arsenic-related bladder cancer. Hum Genet 131:453–461
28. Kooperberg C, LeBlanc M, Obenchain V (2010) Risk prediction using genome-wide association studies. Genet Epidemiol 34:643–652
29. Kruppa J, Ziegler A, König IR (2012) Risk estimation and risk prediction using machine-learning methods. Hum Genet 131:1639–1654
30. Lango H; UK Type 2 Diabetes Genetics Consortium, Palmer CN, Morris AD, Zeggini E et al (2008) Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. Diabetes 57: 3129–3135
31. Le-Niculescu H, Kurian SM, Yehyawi N, Dike C, Patel SD et al (2009) Identifying blood biomarkers for mood disorders using convergent functional genomics. Mol Psychiatry 14:156–174
32. Lettre G, Rioux JD (2008) Autoimmune diseases: insights from genome-wide association studies. Hum Mol Genet 17(R2):R116–R121
33. Lusis AJ, Attie AD, Reue K (2008) Metabolic syndrome: from epidemiology to systems biology. Nat Rev Genet 9:819–830
34. Miller GE, Chen E, Fok AK, Walker H, Lim A, Nicholls EF, Cole S, Kobor MS (2009) Low early-life social class leaves a biological residue manifested by decreased glucocorticoid and increased proinflammatory signaling. Proc Natl Acad Sci U S A 106:14716–14721
35. Morgan AA, Chen R, Butte AJ (2010) Likelihood ratios for genome medicine. Genome Med 2:30
36. Nath AP, Arafat D, Gibson G (2012) Using blood informative transcripts in geographical genomics: impact of lifestyle on gene expression in Fijians. Front Genet 3:243
37. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE et al (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485:242–245
38. Nebert DW, Zhang G, Vesell ES (2008) From human genetics and genomics to pharmacogenetics and pharmacogenomics: past lessons, future directions. Drug Metab Rev 40:187–224
39. Need AC, McEvoy JP, Gennarelli M, Heinzen EL, Ge D et al (2012) Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. Am J Hum Genet 91:303–312
40. Need AC, Shashi V, Hitomi Y, Schoch K, Shianna KV, McDonald MT, Meisler MH, Goldstein DB (2012) Clinical application of exome sequencing in undiagnosed genetic conditions. J Med Genet 49:353–361
41. Pace TWW, Mletzko TC, Alagbe O, Musselman DL, Nemeroff CB, Miller AH, Heim CM (2006) Increased stress-induced inflammatory responses in male patients with major depression and increased early life stress. Am J Psychiatry 163:1630–1633
42. Park S, Yang JS, Kim J, Shin YE, Hwang J, Park J, Jang SK, Kim S (2012) Evolutionary history of human disease genes reveals phenotypic connections and comorbidity among genetic diseases. Sci Rep 2:757
43. Patel CJ, Sivadas A, Tabassum R, Thanawadee P, Zhao J, Arafat D, Chen R, Morgan AA, Martin G, Brigham KL, Butte AJ, Gibson G (2013) Whole genome sequencing in support of wellness and health maintenance. Genome Med 5:58

44. Polychronakos C, Li Q (2011) Understanding type 1 diabetes through genetics: advances and prospects. Nat Rev Genet 12:781–792
45. Preininger M, Arafat D, Kim J, Nath A, Idaghdour Y, Brigham KL, Gibson G (2013) Blood-informative transcripts define nine common axes of peripheral blood gene expression. PLoS Genet 9:e1003362
46. Ramos PS, Criswell LA, Moser KL, Comeau ME, Williams AH et al (2011) A comprehensive analysis of shared loci between systemic lupus erythematosus (SLE) and sixteen autoimmune diseases reveals limited genetic overlap. PLoS Genet 7:e1002406
47. Rask KJ, Brigham KL, Johns MME (2011) Integrating comparative effectiveness research programs into predictive health: a unique role for academic health centers. Acad Med 86:718–723
48. Ritchie ND, Hahn LW, Moore JH (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24:150–157
49. Roberts NJ, Vogelstein JT, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE (2012) The predictive capacity of personal genome sequencing. Sci Transl Med 4:133ra58
50. Rost S, Fregin A, Ivaskevicius V, Conzelmann E, Hörtnagel K et al (2004) Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. Nature 427:537–541
51. Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu YM et al (2011) Personalized oncology through integrative high-throughput sequencing: a pilot study. Sci Transl Med 3:111ra121
52. Rzhetsky A, Wajngurt D, Park N, Zheng T (2007) Probing genetic overlap among complex human phenotypes. Proc Natl Acad Sci U S A 104:11694–11699
53. Salari K, Watkins H, Ashley EA (2012) Personalized medicine: hope or hype? Eur Heart J 33:1564–1570
54. Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in admixed populations. Am J Hum Genet 82:290–303
55. Shriner D, Adeyemo A, Rotimi CN (2011) Joint ancestry and association testing in admixed individuals. PLoS Comput Biol 7:e1002325
56. Talkowski ME, Ordulu Z, Pillalamarri V, Benson CB, Blumenthal I et al (2012) Clinical diagnosis by whole-genome sequencing of a prenatal sample. N Engl J Med 367:2226–2232
57. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90:7–24
58. Weedon MN, McCarthy MI, Hitman G, Walker M, Groves CJ et al (2006) Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. PLoS Med 3:e374
59. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97:1837–1847
60. Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB (2007) Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Arch Intern Med 167:1068–1074
61. Wray NR, Goddard ME (2010) Multi-locus models of genetic risk of disease. Genome Med 2:10
62. Wray NR, Goddard ME, Visscher PM (2008) Prediction of individual genetic risk of complex disease. Curr Opin Genet Dev 18:257–263
63. Wray NR, Yang J, Goddard ME, Visscher PM (2010) The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet 6:e1000864
64. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG et al (2012) Human gut microbiome viewed across age and geography. Nature 486:222–227
65. Yu K, Wacholder S, Wheeler W, Wang Z, Caporaso N, Landi MT, Liang F (2012) A flexible Bayesian model for studying gene-environment interaction. PLoS Genet 8:e1002482

Chapter 2
Characterizing Multi-omic Data in Systems Biology

Christopher E. Mason, Sandra G. Porter, and Todd M. Smith

Abstract In today's biology, studies have shifted to analyzing systems over discrete biochemical reactions and pathways. These studies depend on combining the results from scores of experimental methods that analyze DNA; mRNA; noncoding RNAs, DNA, RNA, and protein interactions; and the nucleotide modifications that form the epigenome into global datasets that represent a diverse array of "omics" data (transcriptional, epigenetic, proteomic, metabolomic). The methods used to collect these data consist of high-throughput data generation platforms that include high-content screening, imaging, flow cytometry, mass spectrometry, and nucleic acid sequencing. Of these, the next-generation DNA sequencing platforms predominate because they provide an inexpensive and scalable way to quickly interrogate the molecular changes at the genetic, epigenetic, and transcriptional level. Furthermore, existing and developing single-molecule sequencing platforms will likely make direct RNA and protein measurements possible, thus increasing the specificity of current assays and making it possible to better characterize "epi-alterations" that occur in the epigenome and epitranscriptome. These diverse data types present us with the largest challenge: how do we develop software systems and algorithms that can integrate these datasets and begin to support a more

C.E. Mason (✉)
Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY, USA

The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY, USA
e-mail: chm2042@med.cornell.edu

S.G. Porter
Digital World Biology, 2442 NW Market St., PMB 160, Seattle, WA 98107, USA
e-mail: Sandra@digitalworldbiology.com

T.M. Smith (✉)
PerkinElmer, 100 W. Harrison St. NT 330, Seattle, WA 98119, USA

Digital World Biology, 2442 NW Market St., PMB 160, Seattle, WA, 98107, USA
e-mail: todd.smith@perkinelmer.com; todd@digitalworldbiology.com

N. Maltsev et al. (eds.), *Systems Analysis of Human Multigene Disorders*, Advances in Experimental Medicine and Biology 799, DOI 10.1007/978-1-4614-8778-4_2,

democratic model where individuals can capture and track their own medical information through biometric devices and personal genome sequencing? Such systems will need to provide the necessary user interactions to work with the trillions of data points needed to make scientific discoveries. Here, we describe novel approaches in the genesis and processing of such data, models to integrate these data, and the increasing ubiquity of self-reporting and self-measured genomics and health data.

2.1 Introduction

The selection pressures of evolution drive the enormous complexity of biological systems and create an elaborate panoply of organisms. However, unlike a well-engineered machine, where iterations of the design can increase its capabilities while simultaneously reducing complexity, improving stability, or decreasing energy costs, a biological system works differently. Biological systems adapt to change and respond to mutation in many ways that add, rather than remove, complexity. Biological systems under stress may increase gene expression noise [1], leverage functionally redundant pathways to tolerate substantial gains and/or losses of genes [2] or entire chromosomal segments [3], or increase mutation rates to expedite the ascertainment of a protective mutation [4].

Due to these myriad complexities, complex biological systems do not fit the model of a Cartesian "clockwork" machine that can be simplified and reduced. Increasingly, biological systems are being recognized through emergent properties that can only occur when nested in such irreducible complexity [5]. Accordingly, a goal of modern biology and medicine is to understand the fundamental relationship between an organism's genome and how its genotype affects the dynamic and complex networks of interacting biochemical processes and networks, known as "systems biology [6]" or "pathway analysis." Once these networks are established, scientists and clinicians then aim to understand how these networks are affected by disease and development and maintained in health.

The increased throughput and decreased cost of massively parallel, next-generation DNA sequencing (NGS) has enabled these technologies to emerge as a primary method for measuring genetic variation, gene expression, promoter activity, DNA structure, interacting RNA molecules, and chemical changes to DNA and RNA that define epi-omic data. These assays are being actively expanded and are providing unprecedented insights into relationships between the genome, the transcriptome, their chemical modifications, and their role in controlling many of the essential steps that define the biological network [7].

However, there is increasing evidence that even a perfect understanding of an organism's genotype and phenotype map (GaP) will be incomplete, due to the additional biology that rests *within*, *on, and around* the organism. For example, every human being contains an "$n+1$" organ beyond the normal catalog of anatomy, which consists of the trillions of commensal and pathogenic microorganisms known as the "microbiome." This microbiomic "organ," which outnumbers human cells in

both transcriptional output [8] and cellular count [9], is literally an entire ecosystem of additional biological complexity that is intimately woven into the health and molecular biology of each person that can directly impact obesity [10], diabetes [11], and response to infection [12].

Moreover, even a simple biochemical reaction—[enzyme] [substrate] <> [product]—can be extremely complex. Though the rate of product formation is mostly a function of the enzyme, substrate, and product concentration, enzyme rates can change during the course of a reaction through allosteric regulation. Indeed, enzyme activity can be affected through many protein modifications, such as phosphorylation and dephosphorylation of serine or threonine sites to activate or deactivate an enzyme [13], or chemical modifications such as myristoylation [12] or glycosylation [14] that target enzymes to specific cellular environments and change the effective concentration. Mutations at the DNA level or edits to RNA sequences [15] can also affect enzyme activity by changing the protein structure or impacting catalytic activity. Further, a protein's concentration is controlled by the concentrations of mRNA, tRNA, and amino acids. These affect the rate of translation via ribosome binding, codon usage, and protein turnover. Thus, even the seemingly simple process of translation involves myriad genes and their products, including DNA and RNA polymerases, tRNA synthetases, initiation factors, DNA and mRNA sequence motifs, noncoding RNA molecules, methylases, and proteases. And in the context of systems biology, a simple, rate-controlled biochemical reaction is the result of several layers of biochemistry.

In addition, new "epi-data" have been demonstrated to play an important role in development [16], cancer subtypes [17], and other diseases [18]. These data include epigenetic changes to nucleic acids such as cytosine methylation and hydroxymethylation and adenosine methylation, posttranslational changes to proteins such as phosphorylation and ubiquitination, as well as newly discovered epitranscriptomic changes such as chemical modifications to RNA that produce methyl-6-adenosine and impact brain development [19]. Also, certain processes such as RNA editing can completely alter a biological message as it moves through a cell or within an organism, indicating that any snapshot of a biological system can only be understood in a limited context. Taken together, these factors add critical layers of regulatory complexity, both within and between different biological molecules that have dramatically altered the central dogma of molecular biology (Fig. 2.1).

The central dogma has changed not only in complexity but also in directionality. At the DNA, RNA, and protein levels, self-replication has been observed, and "backward" or "sideway" directions are now options in the central dogma, as seen in the cases of retroviruses, RNA editing, and prions. The alluring, reductionist approach to biology that explored single genes and their products will no longer suffice, even though these methods have historically yielded valuable insights. Indeed, our new understanding is that molecules in an organism function together in networks and must be studied and modeled as such (Fig. 2.2), spanning many layers of organismal molecular complexity. Yet, this realization of the complexity has exacerbated the data analysis problem. Indeed, only one example has been developed to describe a full model of the inner workings of an organism [20], and that was in a simple, prokaryotic system with only several hundred genes. To create

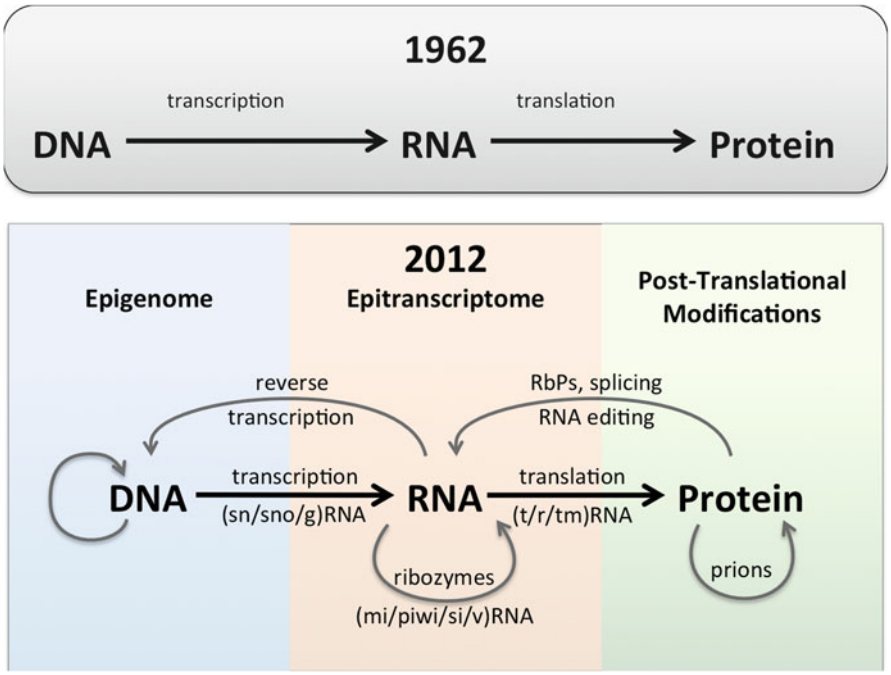


Fig. 2.1 The increasing complexity of the central dogma of molecular biology. (*Top*) The originally proposed central dogma with unidirectional information flow. (*Bottom*) The current view of the central dogma, where information content can flow "backwards" or "sideways" with reverse transcriptases (RT), RNA-binding proteins (RbPs), and RNA editing. Also, information can be copied within each of the three realms: genetic (*blue*) copying such as with transposable elements, transcriptional (*red*) copying with ribozymes and rich levels of RNA regulation using small RNAs (micro, piwi, si, vi RNAs), and proteomic (*green*) copying using prions

complex models for larger organisms, both experimental and computational advances will be needed. On the experimental side, high-throughput assay systems exist that can measure several features of a cell. Mass spectrometry can quantify proteins and protein–protein interactions, along with all the small molecules in the metabolome. High-content imaging and cell sorting systems allow researchers to explore biochemical localization, cell–cell interactions, and tissue morphology with extremely high resolution in multiple dimensions

However, to be effective, researchers need to understand the relationships between the assay results and be able to see different data types merged to present a more complete picture of the underlying biology or etiology of a disease. As data collection systems increase in scale, practical issues related to data management and analysis also emerge. In the following sections, we review the state-of-the-art methods in NGS and DNA sequencing-based assays, discuss solutions for working with the data, and highlight emerging methods for data collection and personalization. From these advances, we observe that, while the digitization of the genome is becoming standardized, there is an urgent need to digitize the phenome and to integrate physiological

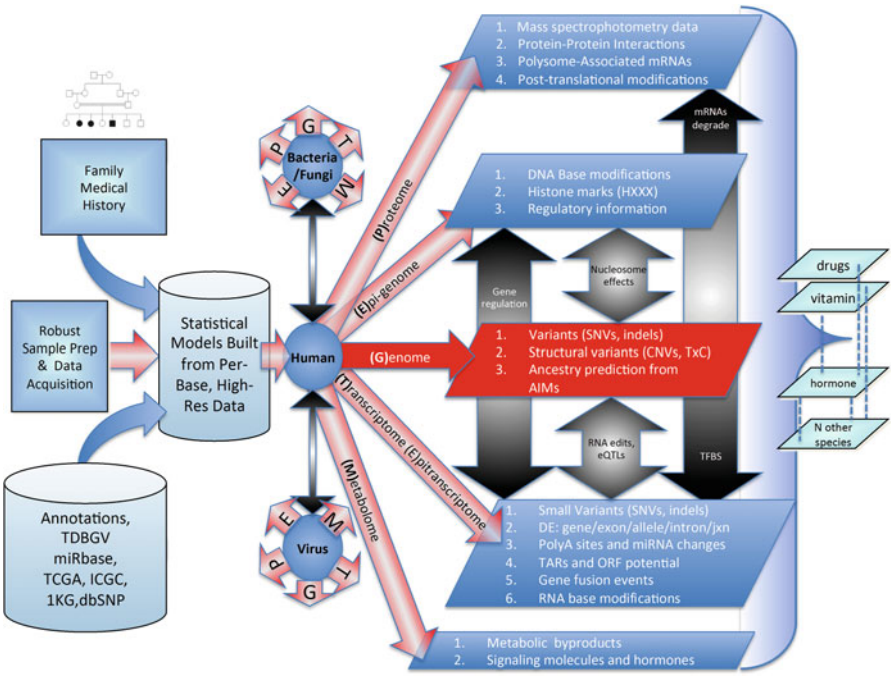


Fig. 2.2 Integrative genomic data sources. Systems biology approaches to biology and medicine now require integration of large amounts of genetic (G), epigenetic (E), transcriptomic (T), proteomic (P), and metabolomic (M) data before, during, and after each assay, which is also ideally applied to each virus and bacteria as well (*P*, *G*, *T arrows*). Ideally, the molecular assays are done with extreme rigor and incorporate knowledge of these assays in the context of patient/sample information (such as pedigree information) as well as public datasets such as The Cancer Genome Atlas (TCGA), dbSNP, dbGaP, miRbase, and others. These data are then collected and reference with respect to the other molecules (*right*) present in a patient, such as vitamins, drugs, diet, and other molecular assays, and modeled into one complete system

data with several layers of molecular data in a genomic context [21, 22]. Completing these steps will improve both research and clinical genomics and enable a patient-driven paradigm of personalized, more accurate molecular medicine.

2.2 Sequencing Technologies

Since 2006, the pace of NGS technological development has accelerated and even surpassed the traditional performance improvement curve, known as Moore's law. Moore's observation, first described for computers, states that digital systems and technologies double in performance characteristics at a 12–18-month rate. DNA sequencing systems followed a similar paradigm until the introduction of massively parallel sequencing (NGS) systems, for which data output doubled on average every

5 months between 2007 and 2010 (http://www.genome.gov/sequencingcosts). Though early NGS technologies relied on fluorescent dye-labeled nucleic acids to measure DNA base incorporation during sequencing, other technologies have now emerged that leverage changes in conductivity or electromagnetic substrates to detect DNA bases (Table 2.1). These technological advances have bifurcated the NGS field into two realms: (1) optical and (2) electrical sequencing, with single-molecule, "third-generation" methods available in both realms.

Many of these nascent technologies have matured since the last detailed review [23]. Electrical sequencing methods have advanced significantly and nanopore- or nano-channel-based systems from Oxford Nanopore, NABsys, Genia, IBM, and Stratos Genomics continue to develop. Because of their potential to directly read native DNA sequences at the single-molecule level, nanopore technologies have generated a high level of excitement. Many nanopore systems a set of protein pores that sit within cell membranes, such as alpha-hemolysin (α-hem), which selectively allows small molecules to pass through its aperture. The pore is attached to a membrane with an application-specific integrated circuit (ASIC) that runs an electrical charge through the nanopore. Since the pore changes its electrical conductivity when it changes shape, and given that the ASIC can run at 32 KHz or higher, thousands of observations per second can be used to detect the presence or movement of a molecule as it passes through the nanopore. Thus, as DNA passes through the pore, the distinctive steric and electrochemical nature of each base reveals a specific conductivity profile, allowing the base to be identified and accomplishing the "sequencing" of the nucleic acid during transit.

In principle, the technology can also distinguish a regular cytosine from a methylated cytosine (mC) or a hydroxyl-methylated cytosine (hmC), both of which are important in development and cancer (above). Moreover, the nanopores should be able to detect these epi-phenotypes in RNA as well, such as methyl-6-adenosine (m6A), which is an "epitranscriptomic" change important in neurodevelopment and obesity [19]. For proteins, it might be possible to use nanopores to identify various modifications or determine if a protein is bound to an aptamer. Thus, considerable excitement has emerged around nanopore-type systems because they could enable the simultaneous detection of genetic/epigenetic information, proteomic alterations, and transcriptomic/epitranscriptomic information; produce extremely long molecular sequences; and be able to use the original non-modified templates which could theoretically be recaptured for other purposes after nanopore-based examination. At present, the estimated error rates for this technology are high, and all of these single-molecule platforms require further maturation before they can be commercialized.

Along similar lines, recent work in nano-channel-based systems have shown promise for extremely long reads of DNA at the single-molecule level, including preliminary instruments available from Nabsys and BioNano Genomics (Table 2.1). These technologies use small microchannels that are barely as wide as one strand of DNA and then used electric currents or electrophoresis to pull the DNA through the channels. Given a set of tags attached to the DNA, it is possible to pull DNA through channels at rates as high as 1 MB/s and create a spatially resolved map of the long strands of DNA. This map depends on the displacement of volume as the DNA with

Table 2.1 Types of high-throughput sequencing technologies

Optical sequencing					
Platform	Instrument	Template preparation	Chemistry	Average length	Longest read
Illumina	HiSeq2500	BridgePCR/cluster	Rev. term., SBS	100	150
Illumina	HiSeq2000	BridgePCR/cluster	Rev. term., SBS	100	150
Illumina	MiSeq	BridgePCR/cluster	Rev. term., SBS	250	300
GnuBio	GnuBio	emPCR	Hyb-assist sequencing	1,000[a]	64,000[a]
Life Technologies	SOLiD 5500	emPCR	Seq. by Lig.	75	100
LaserGen	LaserGen	emPCR	Rev. term., SBS	25[a]	100[a]
Pacific biosciences	RS	Polymerase binding	Real-time	1,800	15,000
454	Titanium	emPCR	PyroSequencing	650	1,100
454	Junior	emPCR	PyroSequencing	400	650
Helicos	Heliscope	Adaptor ligation	Rev. term., SBS	35	57
Intelligent BioSystems	MAX-Seq	Rolony amplification	Two-step SBS (label/unlabell)	2×100	300
Intelligent BioSystems	MINI-20	Rolony amplification	Two-step SBS (label/unlabell)	2×100	300
ZS Genetics	N/A	Atomic lableing	Electron microscope	N/A	N/A
Halcyon Molecular	N/A	N/A	Direct observation of DNA	N/A	N/A
Electical sequencing					
Platform	Instrument	Template preparation	Chemistry	Average length	Longest read
IBM DNA transistor	N/A	None	Microchip nanopore	N/A	N/A
NABsys	N/A	None	Nanochannel	N/A	N/A
Bionanogenomics	N/A	Anneal 7mers	Nanochannel	N/A	N/A
Life Technologies	PGM	emPCR	Semi-conductor	150	300
Life Technologies	Proton	emPCR	Semi-conductor	120	240
Life Technologies	Proton 2	emPCR	Semi-conductor	400[a]	800[a]
Genia	N/A	None	Protein nanopore (a-hemalysin)	N/A	N/A
Oxford nanopore	MinION	None	Protein nanopore	10,000	10,000[a]
Oxford nanopore	GridION 2K	None	Protein nanopore	10,000	500,000[a]
Oxford nanopore	GridION 8K	None	Protein nanopore	10,000	500,000[a]

[a]Values are estimates from companies that have not yet released actual data

a tag passes by a detector, and given a known rate of passage, the coordinates along 100 s of kilobases or even megabases of DNA can be resolved. These tools also have the advantage of being a single-molecule instrument, which will allow a chance to observe epigenetic and epitranscriptomic modifications as well.

Advances in other single-molecule technologies, such as Pacific Biosciences' single-molecule, real-time (SMRT) monitoring system, are already changing the epi-omic landscape. Specifically, the PacBio RS system has demonstrated an ability to detect several DNA modifications, including mC, hmC, and 8-oxo-guanosine [24]. These types of changes are uncovering entirely new layers of regulatory information in bacteria [25] and humans [26], and many modifications of deoxyribonucleic acids, ribonucleic acids, and the DNA backbone are now recognized phenotypes that must be added and considered in biological models.

Preliminary reports of the number of genes with modified DNA [27] and RNA [18, 28] span at least half of the genes in the genome, and as such, these modifications (and their detection) are now an essential component of molecular profiling [29]. The PacBio RS system has been reported to observe RNA modifications such as m6A in RNA using a reverse transcriptase in the zero-mode waveguides (ZMWs) instead of DNA polymerase, creating a unique kinetic signature for modified vs. unmodified adenosines [30]. Thus, the use of single-molecule monitoring systems can, in principle, detect many of these modifications to DNA and RNA.

In summary, these nanopore and single-molecule systems have the potential to examine any aspect of biology. For example, the PacBio system has been used to measure the speed of protein translation [31], and modifications of this protocol could clarify the effects of antibiotics or small molecules on translation efficiency. Also, fixing a field of ZMWs with one protein and then washing fluorescently labeled protein partners over the SMRT cells could examine protein–protein interactions. In nanopores, protein–aptamer matches could be detected as the nanopore changes shape during the interaction, and similar changes could be revealed during small-molecule binding. These rapid methods for high-throughput screening of many molecules' interactions with native biological materials will enable a new field of discovery about the detailed progression of information in a cell and the critical places that can be targeted or modified.

2.3 Sequencing Assays

DNA sequencing platforms generally provide a single kind of output—sequences of letters that represent bases in DNA. As already discussed, certain single-molecule platforms can detect base modifications by measuring additional attributes such as kinetic values or steric differences. A common feature of all modern sequencing platforms is that they collect data in a massively parallel format. Indeed, the combination of miniaturization and increasing parallelization is the primary driver for the phenomenal reduction in data costs experienced from 2007 to 2013. Parallelization also allows us to interrogate molecules at an individual level rather than as an

ensemble, as was the case for capillary-based, or Sanger, sequencing methods, allowing low-frequency base changes to be measured more accurately.

The ability to measure a large repertoire of individual molecules, simultaneously, at high resolution, allows researchers to use general-purpose technologies in unique and novel ways. This power was recognized early on [32], and great debates were fueled concerning the demise of microarray technologies in favor of NGS [33]. While initially there were claims that older technologies would completely disappear, currently there exists an ecosystem where different technological platforms are being used for discrete purposes [34]. If publications are an indicator, "dinosaur" technologies remain active for many kinds of research, due to overall cost and process factors or the need for to compare new results to archived data or data from long-term research projects.

Nonetheless, NGS is now the method of choice for exploring uncharted molecular waters, due to its ability to measure low levels of signal in a high noise background. To date, almost all large-scale genomics and cancer discovery projects have changed to almost exclusively NGS-based methods, including ENCODE, modENCODE, TCGA, ICGC, the NIH's Epigenomics Roadmap, RIKEN, FANTOM, and BLUEPRINT. These massive genomics datasets will generate a per-base mutational map for many tissues, tumors, and cell lines, as well as a complete regulatory map of DNA and histone modifications, while novel NGS-based methods for RNA sequencing (RNA-Seq) will contribute several kinds of global expression maps.

RNA-Seq is notable for the breadth of information that can be acquired in a single experiment [35]. Specifically, mRNA-Seq can measure expression levels for genes and exons, multiple isoforms of expressed genes, mutations, allele-specific expression, RNA editing, intron retention, UTR-length changes, gene fusions, polyadenylation sites, antisense transcription, and novel transcriptionally active regions (TARs). While the overall numbers of transcripts annotated in common gene annotations differ between various datasets and the number of RNAs being edited remains controversial [15, 36, 37], it is now well established that much of the genome is likely expressed [38] and RNA can be edited at many sites. Finally, there are now almost a dozen "flavors" of RNA-Seq methods that can be used, including many that preserve the strand of origin, furthering efforts to understand the interplay between sense and antisense transcription with respect to gene regulation [39].

Like the development of new sequencing technologies for RNA, new assays that use DNA sequencing to assess the various "omes" related to functional genomics are also emerging at rapid rates [32]. In many cases, assays developed in the early days of molecular biology have been scaled up to measure genome-wide features. RNA-Seq grew from microarrays [40] which were inspired by sequencing-based EST (expressed sequence tag) [41] and SAGE (serial analysis of gene expression [42]) methods. Researchers in the ENCODE project have refined several assays based on DNase I hypersensitivity (DHS) and transcription-binding assays (ChIP-Seq) to create novel assays for measuring specific sites in the human genome involved in gene regulation. Indeed, the current catalog of assays from the Epigenomics Roadmap spans dozens of methods, including the examination of 27 histone variants (http://www.roadmapepigenomics.org/).

Massively parallel DNA sequencing is also being used to explore the world of noncoding RNA. Noncoding RNA analysis is an area where creative method development has yielded many new assays and results that have significantly changed our thinking about molecular biology. While noncoding RNA and its role in regulating gene expression through mRNA interactions have been known for several years [43], new assays have demonstrated that far more kinds of RNAs exist than previously known, and they control functional gene expression at several levels—currently there are at least 26 types of RNA and over 100 known RNA modifications (from the RNA modification database) (Table 2.2).

However, a challenge in measuring RNA is that many techniques require an early step where native RNA is often converted to cDNA, then sequenced. This process can miss biological entities because of biases in both cDNA synthesis and the subsequent amplification reactions. These problems may be circumvented through the use of new approaches for directly sequencing single molecules of RNA [44] and other single-molecule methods (above, Table 2.1). When intermediate sample processing steps can be eliminated, assays naturally increase in their sensitivity and specificity because the loss-of-signal artifacts that result from purification steps and noise created through PCR or enzymatic synthesis steps can be reduced. In early demonstrations, single-molecule RNA-Seq has shown that RNAs have extremely divergent 3′UTRs and also that small nucleolar RNAs can also be polyadenylated. Coupled with the ribo-depletion technologies that are now becoming common for RNA-Seq and being used for formalin-fixed, paraffin-embedded (FFPE) tissues, an appreciation of each RNA molecule, and its particular contribution to RNA biology, is now possible for research and clinical samples.

Table 2.2 RNA modifications

Abbreviation	Chemical name
$m^1acp^3\Psi$	1-Methyl-3-(3-amino-3-carboxypropyl) pseudouridine
m^1A	1-Methyladenosine
m^1G	1-Methylguanosine
m^1I	1-Methylinosine
$m^1\Psi$	1-Methylpseudouridine
m^1 Am	1,2′-*O*-Dimethyladenosine
m^1Gm	1,2′-*O*-Dimethylguanosine
m^1Im	1,2′-*O*-Dimethylinosine
m^2A	2-Methyladenosine
ms^2io^6A	2-Methylthio-N^6-(*cis*-hydroxyisopentenyl) adenosine
ms^2hn^6A	2-Methylthio-N^6-hydroxynorvalyl carbamoyladenosine
ms^2i^6A	2-Methylthio-N^6-isopentenyladenosine
ms^2m^6A	2-Methylthio-N^6-methyladenosine
ms^2t^6A	2-Methylthio-N^6-threonyl carbamoyladenosine
s^2Um	2-Thio-2′-*O*-methyluridine
s^2C	2-Thiocytidine

(continued)

Table 2.2 (continued)

Abbreviation	Chemical name
s^2U	2-Thiouridine
Am	2′-*o*-Methyladenosine
Cm	2′-*o*-Methylcytidine
Gm	2′-*O*-Methylguanosine
Im	2′-*o*-Methylinosine
Ψm	2′-*o*-Methylpseudouridine
Um	2′-*o*-Methyluridine
Ar(p)	2′-*o*-Ribosyladenosine (phosphate)
Gr(p)	2′-*o*-Ribosylguanosine (phosphate)
acp^3U	3-(3-Amino-3-carboxypropyl)uridine
m^3C	3-Methylcytidine
$m^3Ψ$	3-Methylpseudouridine
m^3U	3-Methyluridine
m^3Um	3,2′-*O*-Dimethyluridine
imG-14	4-Demethylwyosine
s^4U	4-Thiouridine
chm^5U	5-(Carboxyhydroxymethyl)uridine
$mchm^5$ U	5-(Carboxyhydroxymethyl)uridine methyl ester
inm^5s^2U	5-(Isopentenylaminomethyl)-2-thiouridine
inm^5Um	5-(Isopentenylaminomethyl)-2′-*O-methyluridine*
inm^5U	5-(Isopentenylaminomethyl)uridine
nm^5s^2U	5-Aminomethyl-2-thiouridine
ncm^5Um	5-Carbamoylmethyl-2′-*O*-methyluridine
ncm^5U	5-Carbamoylmethyluridine
$cmnm^5Um$	5-Carboxymethylaminomethyl-2′-*O-methyluridine*
$cmnm^5s^2U$	5-Carboxymethylaminomethyl-2-thiouridine
$cmnm^5U$	5-Carboxymethylaminomethyluridine
cm^5U	5-Carboxymethyluridine
f^5Cm	5-Formyl-2′-*O*-methylcytidine
f^5C	5-Formylcytidine
hm^5C	5-Hydroxymethylcytidine
ho^5U	5-Hydroxyuridine
mcm^5s^2U	5-Methoxycarbonylmethyl-2-thiouridine
mcm^5Um	5-Methoxycarbonylmethyl-2′-*O-methyluridine*
mcm^5U	5-Methoxycarbonylmethyluridine
mo^5U	5-Methoxyuridine
m^5s^2U	5-methyl-2-thiouridine
mnm^5se^2U	5-Methylaminomethyl-2-selenouridine
mnm^5s^2U	5-Methylaminomethyl-2-thiouridine
mnm^5U	5-Methylaminomethyluridine
m^5C	5-Methylcytidine
m^5D	5-Methyldihydrouridine
m^5U	5-Methyluridine
$τm^5s^2U$	5-Taurinomethyl-2-thiouridine
$τm^5U$	5-Taurinomethyluridine

(continued)

Table 2.2 (continued)

Abbreviation	Chemical name
m^5Cm	5,2′-*O*-Dimethylcytidine
m^5Um	5,2′-*O*-Dimethyluridine
$preQ_1$	7-Aminomethyl-7-deazaguanosine
$preQ_0$	7-Cyano-7-deazaguanosine
m^7G	7-Methylguanosine
G^+	Archaeosine
D	Dihydrouridine
oQ	Epoxyqueuosine
galQ	Galactosyl-queuosine
OHyW	Hydroxywybutosine
I	Inosine
imG2	Isowyosine
k^2C	Lysidine
manQ	Mannosyl-queuosine
mimG	Methylwyosine
m^2G	*N*^{2}-Methylguanosine
m^2Gm	*N*^{2},2′-*O*-Dimethylguanosine
$m^{2,7}G$	*N*^{2},7-Dimethylguanosine
$m^{2,7}Gm$	*N*^{2},7,2′-*O*-Trimethylguanosine
$m^2{}_2G$	*N*^{2},*N*^{2}-Dimethylguanosine
$m^2{}_2Gm$	*N*^{2},*N*^{2},2′-*O*-Trimethylguanosine
$m^{2,2,7}G$	*N*^{2},*N*^{2},7-Trimethylguanosine
ac^4Cm	*N*^{4}-Acetyl-2′-*O*-methylcytidine
ac^4C	*N*^{4}-Acetylcytidine
m^4C	*N*^{4}-Methylcytidine
m^4Cm	*N*^{4},2′-*O*-Dimethylcytidine
$m^4{}_2Cm$	*N*^{4},*N*^{4},2′-*O*-Trimethylcytidine
io^6A	*N*^{6}-(*cis*-hydroxyisopentenyl)adenosine
ac^6A	*N*^{6}-Acetyladenosine
g^6A	*N*^{6}-Glycinylcarbamoyladenosine
hn^6A	*N*^{6}-Hyd roxynorvalylcarbamoyladenosine
i^6A	*N*^{6}-Isopentenyladenosine
m^6t^6A	*N*^{6}-Methyl-*N*^{6}-threonylcarbamoyladenosine
m^6A	*N*^{6}-Methyladenosine
t^6A	*N*^{6}-Threonylcarbamoyladenosine
m^6Am	*N*^{6},2′-*O*-Dimethyladenosine
$m^6{}_2A$	*N*^{6},*N*^{6}-Dimethyladenosine
$m^6{}_2Am$	*N*^{6},*N*^{6},2′-*O*-Trimethyladenosine
o_2yW	Peroxywybutosine
Ψ	Pseudouridine
Q	Queuosine
OHyW*	Undermodified hydroxywybutosine
cmo^5U	Uridine 5-oxyacetic acid
$mcmo^5U$	Uridine 5-oxyacetic acid methyl ester
yW	Wybutosine
imG	Wyosine

To summarize, over the past few years, a plethora of assays that measure the nucleotide molecules involved in the functional biology of genomics have emerged. The picture that is developing of gene expression and its regulation is far more complicated than originally thought (Fig. 2.1). Assays continue to be invented that explore new features, and, at the same time, methods like RNA-Seq are becoming more standardized and their variance characterized, with projects such as the Sequencing Quality Control (SeQC—www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject) Consortium, the RNA Genome Alignment Assessment Project (RGASP—www.gencodegenes.org/rgasp/), and the Association of Biomedical Resource Facilities (ABRF) NGS study all underway (www.abrf.org/index.cfm/group.show/NextGeneration Sequencing(NGS).75.htm). As the technologies mature, it is likely their utility and use will become even more ubiquitous.

2.4 Data Analysis

Even though sequencing technology provides abundant ways in which one can measure sequences and base modifications, the deluge of resultant data is quickly becoming the research bottleneck. For example, data storage and other information technology (IT) concerns become dominant issues. Some have even estimated that IT system performance predicted from Moore's law cannot keep pace with NGS throughput [45], meaning that we will eventually need to trim datasets, use non-lossless compression schemes, or even throw away old data.

In addition to the impact on IT systems, the plethora of NGS-based assays has a corresponding complex, tangled "hairball" of application-specific data processing steps that are needed to turn raw data into new knowledge. Just as each NGS application requires different biochemical procedures for sample preparation, each NGS dataset creates specific analytic niches to contextualize the data. Popular NGS websites can be found that list 500 or more algorithms, data views, and software packages (http://seqanswers.com/wiki/Software). A problem that emerges with the emerging plethora of niche-based analyses is that it becomes difficult to merge data from different assays to gain broader insights.

To understand the complexity of data analyses and how data may be merged, an understanding of the general parts of the application-specific data processing steps is also required (Fig. 2.3). NGS produces three general types of data: nucleotide, kinetic, or steric values. For each type, specific data processing steps can be organized into a general framework that consists of four phases: collect, reduce, compare, and discover. Within the NGS community, these four phases are also described as primary, secondary, and tertiary analysis. Tertiary analyses group the compare and discover phases together. Details about the phases are provided below.

Primary data analysis begins when data are collected. Each sequencing platform produces raw data in the form of images, electrical signals, movies, or conductance grids. Software, largely developed by the instrument manufacturers, converts these data to the forms of familiar DNA sequence strings that enter bioinformatics

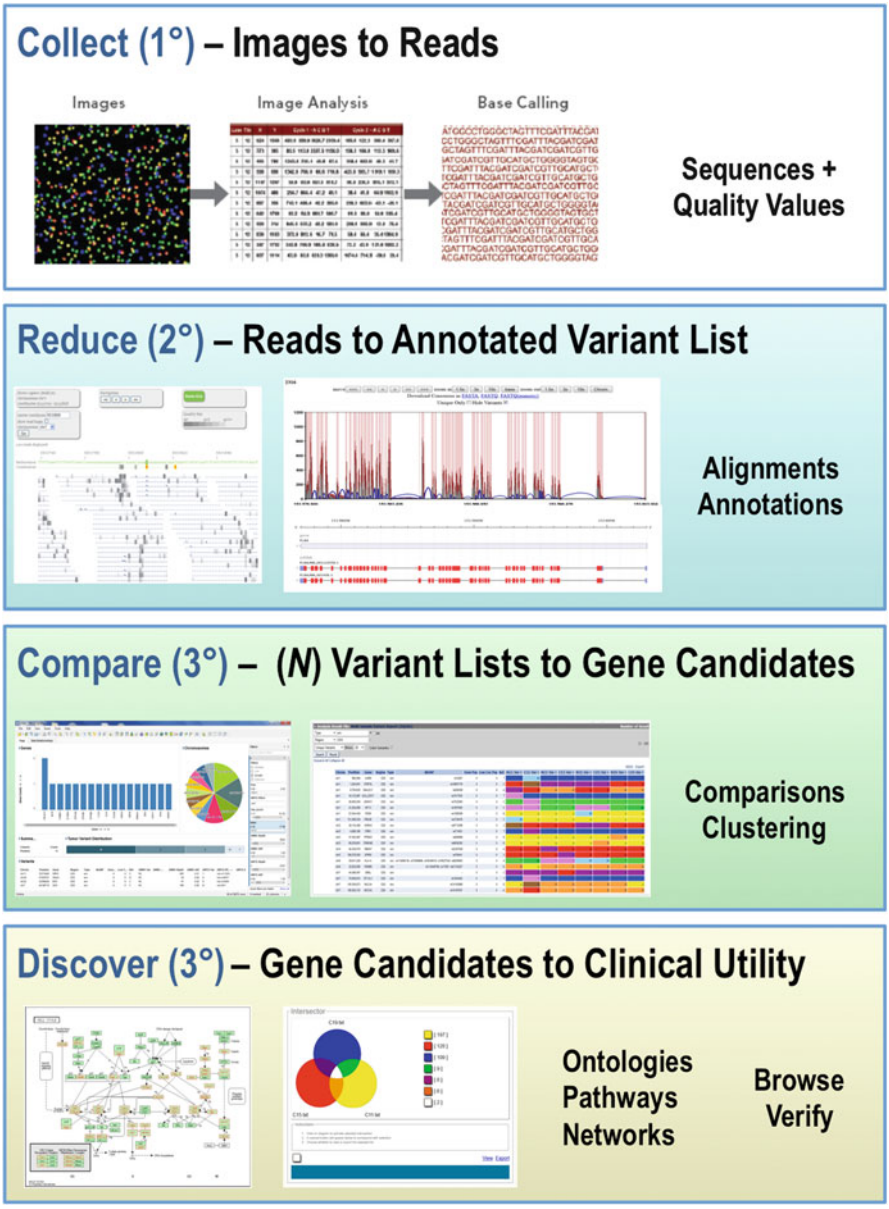


Fig. 2.3 Data analysis steps. Data analysis steps are often grouped into primary, secondary, and tertiary analysis. While most instruments complete the primary or secondary analysis steps, tracking the provenance of those steps is often left to the researcher. Also, to initiate and complete tertiary analyses often requires specialized staff or computational resources, and the outputs from the primary data are often left unconnected. Ideally, this would instead be as above—with each stage aware of, and feeding into, the other. Images are obtained from GeneSifter, ® PerkinElmer Inc.

pipelines for data processing. Quality values, probabilistic measures that describe the "correctness" of the base call, are also produced for each base in the sequence. In some cases, specialized data such as color codes or kinetic data are also produced. In all cases, these "raw data" require further processing to be useful for research.

The next step of analysis is to reduce the large number of sequences to data forms that can be statistically evaluated and compared between samples. One form of data is tables that list chromosome positions and sequence variations relative to a reference, such as a variant call format (VCF) file. These tables may also include other data to describe additional features at a position, such as quality values for the mapped base, or a haplotype. In the case of a whole human genome, between 100 and 200 GB of data may be reduced to less than 10 MB of information. For RNA-Seq, the tables store similar information but also include expression values that are computed from the density of reads mapping to particular positions. The primary activity in the data reduction phase is to align (map) the reads to reference data. Several open-source and proprietary commercial algorithms are used in this process [46]. Each algorithm uses different approaches for identifying matching sequences and all make trade-offs between sensitivity and speed. BWA [47] and Bowtie [48] are popular examples. The specific algorithms used, choice of reference data, and other parameters are defined by the particular application.

For the majority of NGS-based assays, the secondary analysis data are reduced by mapping reads to reference data as described above. However, in some applications, such as determining the sequence of a new organism, sampling genes in the environment (meta-genomics), or creating complete transcript sequences from RNA-Seq data, raw reads are reduced into larger contiguous sequences (contigs) and linked contigs (scaffolds) by comparing the individual reads to one another and merging them through sequence assembly [49]. Like alignment tools, many programs can be used for sequence assembly [50]. Some are tuned for sequences from a particular platform (Newbler, 454); others utilize data from multiple platforms to improve both accuracy and contiguity by combining long reads from lower throughput systems, which have higher error rates, with data from higher-throughput short read systems, with lower error rates, in "hybrid assembly" approaches [51].

From an IT perspective, read mapping and assembly require different computational systems. In read mapping, individual reads can be aligned independently and the data processed in parallel to increase throughput. In computer science terms, mapping is embarrassingly parallel. That is, the mapping rate is simply a function of the number of reads that can be mapped per CPU per time interval multiplied by the total number of CPUs. Random access memory (RAM) requirements are low, and large data processing systems can in principle be built from relatively lost-cost hardware. Although systems that can keep pace with higher-throughput NGS platforms require redundancy, and, high-bandwidth connections between storage and CPUs, and cooling systems, which increase costs significantly. Assembly, on the other hand, requires multiple interdependent alignment steps. These programs utilize dynamic programming, which stores intermediate information in RAM to have reasonable performance and memory requirements scale with the size of the

problem. Large numbers of similar sequences, however, such as repeats or reads from highly expressed genes, can have an unexpected impact on memory requirements. Creating similar kinds of parallel systems for assembly has been challenging; however, some efforts are underway to ameliorate this problem [52].

The final stages of data reduction involve interpreting large tables of aligned reads to produce lists of high-quality information that can include gene expression values, variant data, and other information. These intermediate tables are often larger than the original sequence data, and examining the tables to remove artificial signal and other kinds of noise can exceed the initial computational requirements calculated for read mapping.

Once the tables are created, tertiary analysis can begin. As noted, tertiary analysis has two parts: compare and discover. At this point, the analysis steps can follow as many paths as there are questions about the data. Data can be compared between samples, within samples, across samples, in series over time or with different concentrations of a drug or other variables, and against other datasets to generate lists of genes/pathways/targets that are statistically significant for a particular question. For RNA analysis, many of the same statistical methods employed in microarray analysis are used to identify differential gene expression [53]. Other approaches integrate the data reduction and comparison phases to improve the isoform detection aspect of RNA-Seq [54, 55]. In DNA variant analysis, the work of the Genome Analysis Toolkit (GATK) [56] and Picard (http://picard.sourceforge.net/index.shtml), both of which were born from members of the 1000 Genomes Project (1KGP) [57], has created many well-documented steps for comparing sequence variation between genomes.

For more integrated data analyses, where several assays have been performed to measure genomic features, the tables of quantitative values are converted to signals that can be analyzed in segments to see which assays identify common features. Some of these methods, employed in the ENCODE project, have been used to identify positions in various genes where RNA polymerase stalls, waiting to be activated by other cofactors [58]. Also, other machine-learning methods [59] have shown that certain genomic features predict polymerase behavior better than others. For example, DHS data has been shown to be as effective as all the histone marks combined in identifying an open chromatin state [60], this could not have been known without (1) the large amount of data produced by the ENCODE project and (2) the methods to characterize the data.

Creating new knowledge requires a final phase in the collect, reduce, compare, and discover framework. The comparative phase produces lists of differentially expressed genes, genes with particular variation profiles, or other features that show interesting patterns when samples are compared. The final phase involves working with these lists to explore changes in networks of interactions, see which ontologies might be enriched, and identify affected pathways. These data may also be aggregated with other data in novel ways. In the discovery phase, researchers look at their data in different contexts, sometimes even across species. These contexts are derived from additional information residing in internal and external databases, third-party data, public sites (like the UCSC genome browser), and literature. From a software

development perspective, systems need to provide researchers with easy ways to work with their data and perform iterations of analyses whereby samples are organized and compared in different ways. For this activity, complete systems and packages like the commercially available GeneSifter [61] or open-source packages like GeneHunter [62] are typically used.

Clearly, the compare and discover phases of tertiary analysis benefit from and require access to external data and information. Here the challenges arise in terms of what resources to use, how to access resources, and how to integrate resources into analyses. These questions become daunting because the issues move from pure technical to educational, social, ethical, and policy issues. Just as data throughput rates are increasing, so is the creation of isolated "data silos" of information [63]. Specialized biological databases continue to increase in numbers and content. Nearly one million gene expression datasets are now publically available, and only a few researchers have been able to effectively utilize these resources [64]. In many cases, these databases contain redundant information, and for new ones that may hold unique information, it is not clear how long they will be maintained, or persist, on the Internet. Setting technical issues relating to data formats and their change control aside, integrating such repositories into analyses may also require licensing or other agreements that diverge between commercial and noncommercial entities. As medical sequencing and personalized genomics increase the need for data sharing, these policy and education issues will become more acute.

2.5 The Personalization and Publicizing of Genomes

In human genomics, we are fast approaching a tragedy of the "anti"-commons in genomics, wherein researchers around the world are performing whole genome sequencing (WGS) or whole exome sequencing (WES) and locking the data away in restricted silos. There is little sharing of control datasets or samples among genomics researchers, and the infrastructures that do exist, such as the database of genotypes and phenotypes (dbGAP), discourage use by requiring a lengthy application process to access to the data. Or, in the case of large-scale releases of WES data, one can only get general allele frequencies of mutations, such as the laudable NHLBI Exome Sequencing Project (ESP) dataset. While it contains data from a reasonably large number of exomes, it is only a small portion of the number of exomes or genomes that are being sequenced, for only two of the world's populations.

The reticence in releasing genomic information for patients' or controls' WGS/WES data is partially due to the inherent properties of genetic data. Given any set of genomic data, it is possible to pinpoint the likely region of the world the sample came from [65], and when combined with any other data about the sample's history, it is (in principle) possible to identify the exact individual [66]. Thus, volunteers in genomics studies are now confronted with the possibility that their information may be rediscovered, and their genetic traits could be used against them—the employment and insurance provisions of the Genetic Information Nondiscrimination Act (GINA)

notwithstanding (http://thomas.loc.gov/cgi-bin/bdquery/z?d110:H.R.493). From these concerns, detailed protocols and restraints have emerged for NIH grants and IRB boards to ensure that genetic information is protected.

Despite these concerns, there are several groups who have moved in the opposite direction, using an open-data model for collaboration and sharing of patient data. In particular, PatientsLikeMe, SAGE Bionetworks' Synapse, the Data Enabled Life Sciences Alliance (DELSA), and a variety of genomics companies have all released portions or complete sets of their data for users. For-profit companies like Recombinant, BioFortis, Ingenuity, Ariadni, GeneCo, and NextBio have done so as well. Moreover, individuals are also beginning to collect and collate their own data, using tools like FitBit, the Quantified Self interface, and Do-It-Yourself Genomics (DITGenomics.org). The American College of Medical Genetics (ACGM) has released a position statement that genetic data must remain widely accessible and affordable (http://www.acmg.net/AM/Template.cfm?Section=Home3 &Template=/CM/HTMLDisplay.cfm&ContentID=7367), and the ClinVar database seeks to provide accessible and affordable access [67]. The goal for this data-sharing model is twofold—first, the coordination and collaborative model for sharing controls will give genetic studies much more statistical power for their analyses, and second, these open interfaces allow model builders to compete. Such competition/collaborative models have been applied extraordinarily well in studying proteins and RNA [68], where the users competed to develop better computer algorithms, and both groups learned from each other.

2.6 Conclusions

Looking ahead, if NGS costs continue to decrease at (or near) their current pace, genome sequencing will eventually become a low-cost commodity. When the cost of genome sequencing is low enough, all public spaces could be readily assayed for genomes multiple times a day. A scenario can be imagined where genomic information is used for public surveillance much like the street cameras today. Just as today, we view cameras on streetlights as a "normal" way to prevent crime; we may soon view genomic information in the same way. If and when that time arrives, genomic privacy concerns related to health will become a moot point. Indeed, every time a person walks into and out of any room, copies of his or her genome are left behind for a sequencer to characterize. In this situation, pieces of everyone's personal health information (PHI) become public genomic information (PGI), which can readily combined with video feeds, facial recognition software, and social media posts to track every genome as it moves around the planet in real time. The only sure way to avoid such tracking would be to stay home—hermetically sealed in plastic—since even if you are at home, some of your cells will be released from your home's exhaust.

The ability to identify and track genomes and PGI can be used for purposes both good and bad. Just as the knowledge of nuclear fission can create energy to fuel cities as well as destroy them, the large-scale capture and characterization of the

myriad public human genomes and their tagged subsequent data (transcriptional, proteomic, metabolomic, microbiomic) can be used to enable a better quality of life or a life that is potentially restrained. In a malevolent context, a corrupt government could frame individuals for crimes they never committed, track all citizens and fine them for their molecular details, or revoke their right to reproduce. In an ideal state, a "disease weather map" could be created that tracks outbreaks of infections or organisms in real time, as they emerge, and responds with proper treatments. Moreover, if the genomic backgrounds of the different populations in a city are known, then the populations that carry the highest frequency of disease-susceptible alleles could be prioritized for treatment and even treated with more appropriate and effective drugs. We already do this on an individual level with cytochrome P450 and drug metabolism [69] (warfarin) or VKORC1 and coumarins [70], and these ideas have already been applied throughout a hospital to track drug-resistant bacteria in real time [71]. Thus, expanding these personalized medicine and "allelic response" ideas to a building or city-sized scale is simply a matter of degree. Indeed, it is likely that as these technologies and analytic methods mature even further, the "post-genomics era" maybe aptly named the "ubiquitous genomics era."

Acknowledgments This work was supported by the National Institutes of Health grants 1R01HG006798, 2R44HG005297, and 1R01NS076465. GeneSifter® is a registered trademark of PerkinElmer Inc.

References

1. Charlebois DA, Abdennur N, Kaern M (2011) Gene expression noise facilitates adaptation and drug resistance independently of mutation. Phys Rev Lett 107:218101
2. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40:1166–1174
3. Nóbrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM (2004) Megabase deletions of gene deserts result in viable mice. Nature 431:988–993
4. Levin BR, Cornejo OE (2009) The population and evolutionary dynamics of homologous gene recombination in bacterial populations. PLoS Genet 5:e1000601
5. Wang IM, Zhang B, Yang X, Zhu J, Stepaniants S, Zhang C, Meng Q, Peters M, He Y, Ni C, Slipetz D, Crackower MA, Houshyar H, Tan CM, Asante-Appiah E, O'Neill G, Jane Luo M, Thieringer R, Yuan J, Chiu CS, Yee Lum P, Lamb J, Boie Y, Wilkinson HA, Schadt EE, Dai H, Roberts C (2012) Systems analysis of eleven rodent disease models reveals an inflammatome signature and key drivers. Mol Syst Biol 8:594
6. Koch C (2012) Systems biology. Modular biological complexity. Science 337:531–532
7. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee BK, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Dunham I, Ernst J, Furey TS, Gerstein M, Giardine B, Greven M, Hardison RC, Harris RS, Herrero J, Hoffman MM, Iyer S, Kelllis M, Khatun J, Kheradpour P, Kundaje A, Lassman T, Li Q, Lin X, Marinov GK, Merkel A, Mortazavi A,

Parker SC, Reddy TE, Rozowsky J, Schlesinger F, Thurman RE, Wang J, Ward LD, Whitfield TW, Wilder SP, Wu W, Xi HS, Yip KY, Zhuang J, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M, Pazin MJ, Lowdon RF, Dillon LA, Adams LB, Kelly CJ, Zhang J, Wexler JR, Green ED, Good PJ, Feingold EA, Bernstein BE, Birney E, Crawford GE, Dekker J, Elinitski L, Farnham PJ, Gerstein M, Giddings MC, Gingeras TR, Green ED, Guigó R, Hardison RC, Hubbard TJ, Kellis M, Kent WJ, Lieb JD, Margulies EH, Myers RM, Snyder M, Starnatoyannopoulos JA, Tennebaum SA, Weng Z, White KP, Wold B, Khatun J, Yu Y, Wrobel J, Risk BA, Gunawardena HP, Kuiper HC, Maier CW, Xie L, Chen X, Giddings MC, Bernstein BE, Epstein CB, Shoresh N, Ernst J, Kheradpour P, Mikkelsen TS, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne MJ, Durham T, Ku M, Truong T, Ward LD, Altshuler RC, Eaton ML, Kellis M, Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Batut P, Bell I, Bell K, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena HP, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo OJ, Park E, Preall JB, Presaud K, Ribeca P, Risk BA, Robyr D, Ruan X, Sammeth M, Sandu KS, Schaeffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Hayashizaki Y, Harrow J, Gerstein M, Hubbard TJ, Reymond A, Antonarakis SE, Hannon GJ, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR, Rosenbloom KR, Sloan CA, Learned K, Malladi VS, Wong MC, Barber GP, Cline MS, Dreszer TR, Heitner SG, Karolchik D, Kent WJ, Kirkup VM, Meyer LR, Long JC, Maddren M, Raney BJ, Furey TS, Song L, Grasfeder LL, Giresi PG, Lee BK, Battenhouse A, Sheffield NC, Simon JM, Showers KA, Safi A, London D, Bhinge AA, Shestak C, Schaner MR, Kim SK, Zhang ZZ, Mieczkowski PA, Mieczkowska JO, Liu Z, McDaniell RM, Ni Y, Rashid NU, Kim MJ, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Birney E, Iyer VR, Lieb JD, Crawford GE, Li G, Sandhu KS, Zheng M, Wang P, Luo OJ, Shahab A, Fullwood MJ, Ruan X, Ruan Y, Myers RM, Pauli F, Williams BA, Gertz J, Marinov GK, Reddy TE, Vielmetter J, Partridge EC, Trout D, Varley KE, Gasper C, Bansal A, Pepke S, Jain P, Amrhein H, Bowling KM, Anaya M, Cross MK, King B, Muratet MA, Antoshechkin I, Newberry KM, McCue K, Nesmith AS, Fisher-Aylor KI, Pusey B, DeSalvo G, Parker SL, Balasubramanian S, Davis NS, Meadows SK, Eggleston T, Gunter C, Newberry JS, Levy SE, Absher DM, Mortazavi A, Wong WH, Wold B, Blow MJ, Visel A, Pennachio LA, Elnitski L, Margulies EH, Parker SC, Petrykowska HM, Abyzov A, Aken B, Barrell D, Barson G, Berry A, Bignell A, Boychenko V, Bussotti G, Chrast J, Davidson C, Derrien T, Despacio-Reyes G, Diekhans M, Ezkurdia I, Frankish A, Gilbert J, Gonzalez JM, Griffiths E, Harte R, Hendrix DA, Howald C, Hunt T, Jungreis I, Kay M, Khurana E, Kokocinski F, Leng J, Lin MF, Loveland J, Lu Z, Manthravadi D, Mariotti M, Mudge J, Mukherjee G, Notredame C, Pei B, Rodriguez JM, Saunders G, Sboner A, Searle S, Sisu C, Snow C, Steward C, Tanzer A, Tapanari E, Tress ML, van Baren MJ, Walters N, Washieti S, Wilming L, Zadissa A, Zhengdong Z, Brent M, Haussler D, Kellis M, Valencia A, Gerstein M, Raymond A, Guigó R, Harrow J, Hubbard TJ, Landt SG, Frietze S, Abyzov A, Addleman N, Alexander RP, Auerbach RK, Balasubramanian S, Bettinger K, Bhardwaj N, Boyle AP, Cao AR, Cayting P, Charos A, Cheng Y, Cheng C, Eastman C, Euskirchen G, Fleming JD, Grubert F, Habegger L, Hariharan M, Harmanci A, Iyenger S, Jin VX, Karczewski KJ, Kasowski M, Lacroute P, Lam H, Larnarre-Vincent N, Leng J, Lian J, Lindahl-Allen M, Min R, Miotto B, Monahan H, Moqtaderi Z, Mu XJ, O'Geen H, Ouyang Z, Patacsil D, Pei B, Raha D, Ramirez L, Reed B, Rozowsky J, Sboner A, Shi M, Sisu C, Slifer T, Witt H, Wu L, Xu X, Yan KK, Yang X, Yip KY, Zhang Z, Struhl K, Weissman SM, Gerstein M, Farnham PJ, Snyder M, Tenebaum SA, Penalva LO, Doyle F, Karmakar S, Landt SG, Bhanvadia RR, Choudhury A, Domanus M, Ma L, Moran J, Patacsil D, Slifer T, Victorsen A, Yang X, Snyder M, White KP, Auer T, Centarin L, Eichenlaub M, Gruhl F, Heerman S, Hoeckendorf B, Inoue D, Kellner T, Kirchmaier S, Mueller C, Reinhardt R, Schertel L, Schneider S, Sinn R, Wittbrodt B, Wittbrodt J, Weng Z, Whitfield TW, Wang J, Collins PJ, Aldred SF, Trinklein ND, Partridge EC, Myers RM, Dekker

J, Jain G, Lajoie BR, Sanyal A, Balasundaram G, Bates DL, Byron R, Canfield TK, Diegel MJ, Dunn D, Ebersol AK, Ebersol AK, Frum T, Garg K, Gist E, Hansen RS, Boatman L, Haugen E, Humbert R, Jain G, Johnson AK, Johnson EM, Kutyavin TM, Lajoie BR, Lee K, Lotakis D, Maurano MT, Neph SJ, Neri FV, Nguyen ED, Qu H, Reynolds AP, Roach V, Rynes E, Sabo P, Sanchez ME, Sandstrom RS, Sanyal A, Shafer AO, Stergachis AB, Thomas S, Thurman RE, Vernot B, Vierstra J, Vong S, Wang H, Weaver MA, Yan Y, Zhang M, Akey JA, Bender M, Dorschner MO, Groudine M, MacCoss MJ, Navas P, Stamatoyannopoulos G, Kaul R, Dekker J, Stamatoyannopoulos JA, Dunham I, Beal K, Brazma A, Flicek P, Herrero J, Johnson N, Keefe D, Lukk M, Luscombe NM, Sobral D, Vaquerizas JM, Wilder SP, Batzoglou S, Sidow A, Hussami N, Kyriazopoulou-Panagiotopoulou S, Libbrecht MW, Schaub MA, Kundaje A, Hardison RC, Miller W, Giardine B, Harris RS, Wu W, Bickel PJ, Banfai B, Boley NP, Brown JB, Huang H, Li Q, Li JJ, Noble WS, Bilmes JA, Buske OJ, Hoffman MM, Sahu AO, Kharchenko PV, Park PJ, Baker D, Taylor J, Weng Z, Iyer S, Dong X, Greven M, Lin X, Wang J, Xi HS, Zhuang J, Gerstein M, Alexander RP, Balasubramanian S, Cheng C, Harmanci A, Lochovsky L, Min R, Mu XJ, Rozowsky J, Yan KK, Yip KY, Birney E (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

8. Zhu B, Wang X, Li L (2010) Human gut microbiome: the second genome of human body. Protein Cell 1:718–725
9. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD, Wang J (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65
10. Cho I, Blaser MJ (2012) The human microbiome: at the interface of health and disease. Nat Rev Genet 13:260–270
11. Koren O, Goodrich JK, Cullender TC, Spor A, Laitinen K, Kling Bäckhed H, Gonzalez A, Werner JJ, Angenent LT, Knight R, Bäckhed F, Isolauri E, Salminen S, Ley RE (2012) Host remodeling of the gut microbiome and metabolic changes during pregnancy. Cell 150:470–480
12. Craven M, Egan CE, Dowd SE, McDonough SP, Dogan B, Denkers EY, Bowman D, Scherl EJ, Simpson KW (2012) Inflammation drives dysbiosis and bacterial invasion in murine models of ileal Crohn's disease. PLoS One 7:e41594
13. Hunter T (1987) A thousand and one protein kinases. Cell 50:823–829
14. Petsko GA, Ringe D (2004) Protein structure and function (illustrated ed.). London: New Science Press
15. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG (2011) Widespread RNA and DNA sequence differences in the human transcriptome. Science 333:53–58
16. Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, Diez J, Sanchez-Mut JV, Setien F, Carmona FJ, Puca AA, Sayols S, Pujana MA, Serra-Musach J, Iglesias-Platas I, Formiga F, Fernandez AF, Fraga MF, Heath SC, Valencia A, Gut IG, Wang J, Esteller M (2012) Distinct DNA methylomes of newborns and centenarians. Proc Natl Acad Sci U S A 109:10522–10527
17. Akalin A, Garrett-Bakelman FE, Kormaksson M, Busuttil J, Zhang L, Khrebtukova I, Milne TA, Huang Y, Biswas D, Hess JL, Allis CD, Roeder RG, Valk PJ, Löwenberg B, Delwel R, Fernandez HF, Paietta E, Tallman MS, Schroth GP, Mason CE, Melnick A, Figueroa ME (2012) Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. PLoS Genet 8:e1002781
18. Fernandez AF, Assenov Y, Martin-Subero JI, Balint B, Siebert R, Taniguchi H, Yamamoto H, Hidalgo M, Tan AC, Galm O, Ferrer I, Sanchez-Cespedes M, Villanueva A, Carmona J, Sanchez-Mut JV, Berdasco M, Moreno V, Capella G, Monk D, Ballestar E, Ropero S, Martinez R, Sanchez-Carbayo M, Prosper F, Agirre X, Fraga MF, Graña O, Perez-Jurado L, Mora J, Puig S, Prat J, Badimon L, Puca AA, Meltzer SJ, Lengauer T, Bridgewater J, Bock C, Esteller M (2012) A DNA methylation fingerprint of 1628 human samples. Genome Res 22:407–419

19. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR (2012) Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. Cell 149:1635–1646
20. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, Covert MW (2012) A whole-cell computational model predicts phenotype from genotype. Cell 150:389–401
21. Califano A, Butte AJ, Friend S, Ideker T, Schadt E (2012) Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet 44:841–847
22. Schadt E (2012). Nat Biotechnol 30:769–770
23. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31–46
24. Korlach J, Turner SW (2012) Going beyond five bases in DNA sequencing. Curr Opin Struct Biol 22(3):251–261
25. Bashir A, Bansal V, Bafna V (2010) Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance. BMC Genomics 11:385
26. Song CX, Clark TA, Lu XY, Kislyuk A, Dai Q, Turner SW, He C, Korlach J (2012) Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. Nat Methods 9:75–77
27. Clark TA, Spittle KE, Turner SW, Korlach J (2011) Direct detection and sequencing of damaged DNA bases. Genome Integr 2:10
28. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, Sorek R, Rechavi G (2012) Topology of the human and mouse m6a RNA methylomes revealed by m6a-seq. Nature 485:201–206
29. Saletore Y, Chen-Kiang S, Mason CE (2013) Novel RNA regulatory mechanisms revealed in the epitranscriptome. RNA Biol 10(3):342–346
30. Saletore Y, Meyer K, Korlach J, Vilfan I, Jaffrey S, Mason CE (2012) The birth of the Epitranscriptome: deciphering the function of RNA modifications. Genome Biol 13(10):175
31. Tsai A, Petrov A, Marshall RA, Korlach J, Uemura S, Puglisi JD (2012) Heterogeneous pathways and timing of factor departure during translation initiation. Nature 487:390–393
32. Kahvejian A, Quackenbush J, Thompson JF (2008) What would you do if you could sequence everything? Nat Biotechnol 26:1125–1133
33. Shendure J (2008) The beginning of the end for microarrays? Nat Methods 5:585–587
34. Shendure J, Lieberman Aiden E (2012) The expanding scope of DNA sequencing. Nat Biotechnol 30:1084–1094
35. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517
36. Kleinman CL, Majewski J (2012) Comment on "widespread RNA and DNA sequence differences in the human transcriptome". Science 335:1302, author reply 1302
37. Pickrell JK, Gilad Y, Pritchard JK (2012) Comment on "widespread RNA and DNA sequence differences in the human transcriptome". Science 335:1302, author reply 1302
38. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007) What is a gene, post-encode? History and updated definition. Genome Res 17:669–681
39. Yassour M, Pfiffner J, Levin JZ, Adiconis X, Gnirke A, Nusbaum C, Thompson DA, Friedman N, Regev A (2010) Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. Genome Biol 11:R87
40. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470
41. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656
42. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484–487
43. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391:806–811

44. Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM (2009) Direct RNA sequencing. Nature 461:814–818
45. Stein LD (2010) The case for cloud computing in genome informatics. Genome Biol 11:207
46. Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11:473–483
47. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25:1754–1760
48. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9:357–359
49. Nagarajan N, Pop M (2013) Sequence assembly demystified. Nat Rev Genet 14:157–167
50. Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95:315–327
51. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Phillippy AM (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30(7):693–700
52. Schatz MC, Langmead B, Salzberg SL (2010) Cloud computing and the DNA data race. Nat Biotechnol 28:691–693
53. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80
54. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–578
55. Anders S, Reyes A, Huber W (2012) Detecting differential usage of exons from RNA-seq data. Genome Res 22:2008–2017
56. McKenna AH, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, Depristo M (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303
57. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073
58. Wang C, Tian R, Zhao Q, Xu H, Meyer CA, Li C, Zhang Y, Liu XS (2012) Computational inference of mRNA stability from histone modification and transcriptome profiles. Nucleic Acids Res 40:6414–6423
59. Agius P, Arvey A, Chang W, Noble WS, Leslie C (2010) High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. PLoS Comput Biol 6:e1000916
60. Nimrod G, Szilágyi A, Leslie C, Ben-Tal N (2009) Identification of DNA-binding proteins using structural, electrostatic and evolutionary features. J Mol Biol 387:1040–1053
61. Porter S, Olson NE, Smith T (2009) Analyzing gene expression data from microarray and next-generation DNA sequencing transcriptome profiling assays using GeneSifter analysis edition. Curr Protoc Bioinformatics 7(14):1–35
62. Nyholt DR (2002) GENEHUNTER: your 'one-stop shop' for statistical genetic analysis? Hum Hered 53:2–7
63. Galperin MY, Fernández-Suárez XM (2012) The 2012 nucleic acids research database issue and the online molecular biology database collection. Nucleic Acids Res 40:D1–D8
64. Baker M (2012) Gene data to hit milestone. Nature 487:282–283
65. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. Nature 456:98–101
66. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 4:e1000167

67. Baker M (2012) One-stop shop for disease genes. Nature 491:171
68. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popović Z, Players F (2010) Predicting protein structures with a multiplayer online game. Nature 466:756–760
69. Suarez-Kurtz G, Botton MR (2013) Pharmacogenomics of warfarin in populations of african descent. Br J Clin Pharmacol 75:334–346
70. Chouchane L, Mamtani R, Dallol A, Sheikh JI (2011) Personalized medicine: a patient-centered paradigm. J Transl Med 9:206
71. Snitkin E, Zelazny AM, Thomas PJ, Stock F, NISC Comparative Sequencing Program, Henderson DK, Palmore TN, Segre JA (2012) Tracking a hospital outbreak of carbapenem-resistant klebsiella pneumoniae with whole-genome sequencing. Sci Transl Med 4:148ra116

Chapter 3
High-Throughput Translational Medicine: Challenges and Solutions

Dinanath Sulakhe, Sandhya Balasubramanian, Bingqing Xie, Eduardo Berrocal, Bo Feng, Andrew Taylor, Bhadrachalam Chitturi, Utpal Dave, Gady Agam, Jinbo Xu, Daniela Börnigen, Inna Dubchak, T. Conrad Gilliam, and Natalia Maltsev

Abstract Recent technological advances in genomics now allow producing biological data at unprecedented tera- and petabyte scales. Yet, the extraction of useful knowledge from this voluminous data presents a significant challenge to a scientific community. Efficient mining of vast and complex data sets for the needs of biomedical research critically depends on seamless integration of clinical, genomic, and experimental information with prior knowledge about genotype–phenotype relationships accumulated in a plethora of publicly available databases. Furthermore, such experimental data should be accessible to a variety of algorithms and analytical pipelines that drive computational analysis and data mining.

D. Sulakhe (✉)
Computation Institute, University of Chicago/Argonne National Laboratory, 5735 S Ellis Ave, Chicago, IL 60637, USA
e-mail: sulakhe@mcs.anl.gov

S. Balasubramanian • A. Taylor
Department of Human Genetics, University of Chicago, Chicago, IL, USA

B. Xie • E. Berrocal • B. Feng
Department of Human Genetics, University of Chicago, Chicago, IL, USA

Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA

B. Chitturi
Department of Computer Science, Amrita Vishwa Vidyapeetham University, Amritapuri Campus, Clappana, Kollam, Kerala, India

U. Dave
Computation Institute, University of Chicago/Argonne National Laboratory, Chicago, IL, USA

G. Agam
Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA

J. Xu
Toyota Technological Institute at Chicago, Chicago, IL, USA

N. Maltsev et al. (eds.), *Systems Analysis of Human Multigene Disorders*, Advances in Experimental Medicine and Biology 799, DOI 10.1007/978-1-4614-8778-4_3,

Translational projects require sophisticated approaches that coordinate and perform various analytical steps involved in the extraction of useful knowledge from accumulated clinical and experimental data in an orderly semiautomated manner. It presents a number of challenges such as (1) high-throughput data management involving data transfer, data storage, and access control; (2) scalable computational infrastructure; and (3) analysis of large-scale multidimensional data for the extraction of actionable knowledge.

We present a scalable computational platform based on crosscutting requirements from multiple scientific groups for data integration, management, and analysis. The goal of this integrated platform is to address the challenges and to support the end-to-end analytical needs of various translational projects.

3.1 Introduction

Understanding the genetic architecture underlying complex biological phenomena and heritable multigene disorders is one of the major goals of human genetics in the next decade. Advances in whole genome sequencing and the success of high-throughput functional genomics help to supplement conventional reductionist biology with system-level approaches that allow researchers to study biology and medicine as complex networks of interacting genetic and epigenetic factors in relevant biological contexts. This integrative approach holds the promise of unveiling hitherto unexplored levels of molecular organization and biological complexity. It also holds the key to deciphering the multigene patterns of inheritance that predispose individuals to a wide array of genetic diseases. Studies by countless groups have identified genes associated with many rare single gene (Mendelian) developmental disorders, but only limited progress has been made in finding the underlying causes for autism, schizophrenia, diabetes, and predisposition to cancer and cardiovascular diseases as they display complex patterns of inheritance and may result from many

D. Börnigen
Department of Human Genetics, University of Chicago, Chicago, IL, USA

Toyota Technological Institute at Chicago, Chicago, IL, USA

I. Dubchak
Genomics Division, Berkley National Laboratory, Walnut Creek, CA, USA

T.C. Gilliam
Department of Human Genetics, University of Chicago, Chicago, IL, USA

Computation Institute, University of Chicago/Argonne National Laboratory, Chicago, IL, USA

N. Maltsev (✉)
Department of Human Genetics, University of Chicago, Chicago, IL, USA

Computation Institute, University of Chicago/Argonne National Laboratory,
5735 S Ellis Ave, Chicago, IL 60637, USA
e-mail:Maltsev@uchicago.edu

genetic variations, each contributing only weak effects to the disease phenotype. Identification of causative disease genes or genetic variations within the myriad of susceptibility loci identified in linkage and association studies is difficult because these loci may contain hundreds of genes. Fortunately, recent advances in biological science have provided new perspectives into studies of complex heritable disorders, including (1) high-throughput integrative genomics and informatics, (2) network-based view of human disorders, and (3) emergence of "phenomics" and a notion of interrelatedness of diseases and disease traits. These approaches offer a strategy for system-level exploration of complex clinical phenotypes in relevant biological contexts. They utilize expertise from the fields of genomics, molecular biology, bioinformatics, and clinical studies to develop integrative models of molecular events driving the emergence of cellular and organismal phenotypes. At the basic science level, this research seeks to understand the nascent properties of interacting molecular networks and how they relate to biological complexity. At the applied level, identifying combinations of interacting genes that underlie complex genetic disorders is the practical first step in moving from today's genetic understanding to the era of individualized medicine. Understanding the genetic architecture of common diseases will afford presymptomatic testing of individuals at risk for common disorders, gradually shifting the practice of medicine from a "reactive" science to a "predictive" science. It will also allow state-of-the-art technologies such as high-throughput genetic screening to advance drug discovery and development.

However, the extraction of meaningful information from an avalanche of available biomedical information requires seamless integration of data and services across the analytical workflows. These workflows start from the raw experimental data and include multiple analytical steps leading to the generation of high-confidence hypotheses regarding molecular mechanisms contributing to the phenotypes of interest. Each step of such a pipeline generates additional annotations consumed by the subsequent steps of analysis or displayed to the user to aid in manual investigation of the data. The nature of contemporary biology dictates the need for the use of multiple data sources and distributed analytical services developed by a number of scientific groups. This distributed research paradigm calls for integrated analytical platform to address the end-to-end requirements of translational projects. Such a platform requires advanced computational technologies that will ensure fast and reliable movement of terabytes of data, provide on-demand scalable computational resources, and guarantee security and provenance of every analytical step. The need for a profound integration of data and services was expressed in numerous publications [1–3]. A number of large-scale initiatives were launched to bring together information resources and make them available to the scientific community [4–6].

Here we present an example of a project-driven integrated computational platform that relies on data, expertise, and analytical services provided by a number of scientific groups. The goal of this scalable platform is to support the end-to-end analytical requirements for individual translational projects. As the number of users grows, so grows the network of data and services to meet evolving user requirements. Working with multiple translational projects allowed us to identify crosscutting shared computational and analytical requirements. These projects have converged

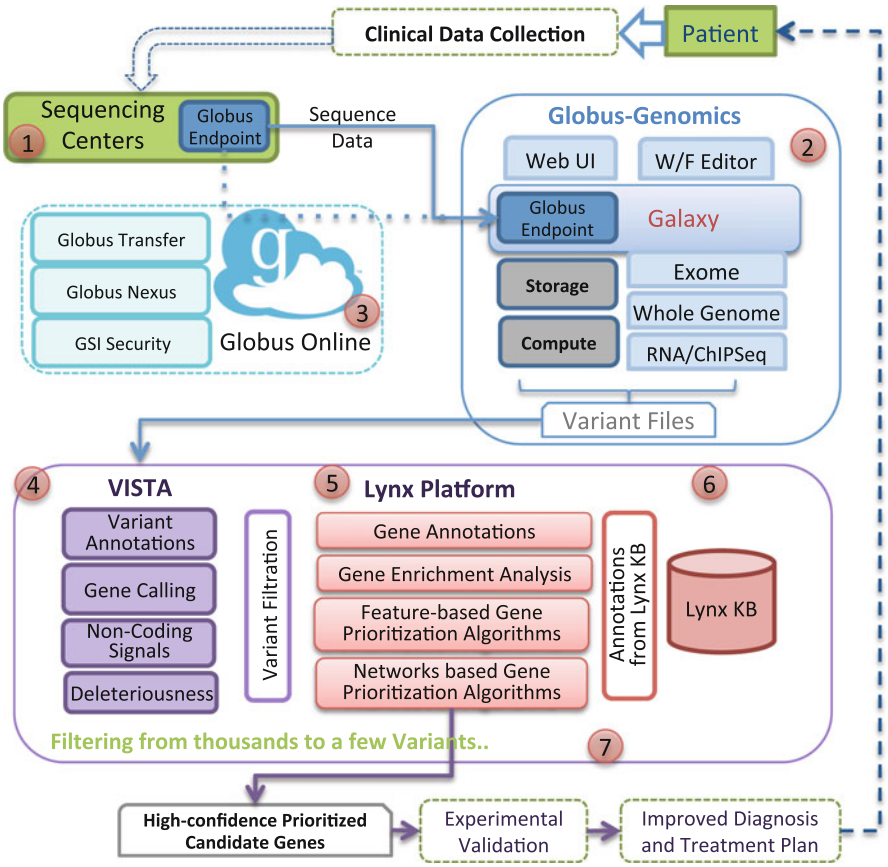


Fig. 3.1 End-to-end analytical pipeline overview

towards well-defined standard steps of analysis of translational data as represented in Fig. 3.1. These requirements also stressed the need for an integrated scalable solution servicing all steps of a translational project including data acquisition, integration, analysis, and presentation to the user. Current disjoint and inefficient processing of high-volume data leads to waste of computational and human resources and reduction of quality of the scientific outcome. Sections below will describe the steps involved in translational data analysis in greater detail (Fig. 3.2).

3.2 Challenges

In less than 5 years, next-generation sequencing (NGS) and other high-throughput genomic technologies have gone from radical to routine (see Chap. 2 of this volume by C. Mason for a detailed review). Drastic reduction in NGS costs and availability,

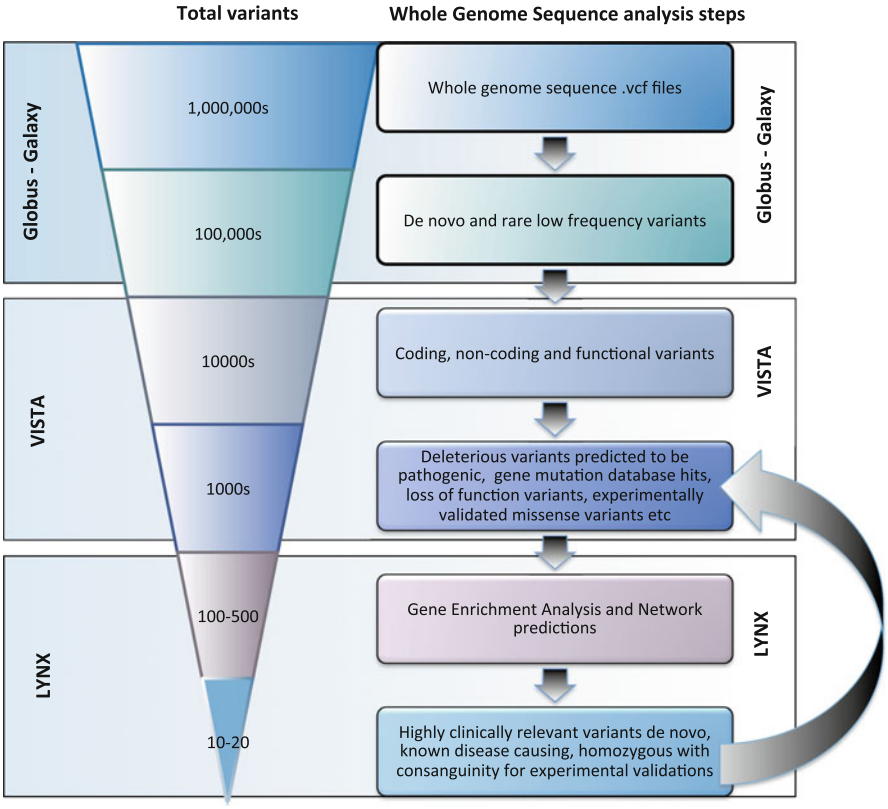


Fig. 3.2 Steps of whole genome sequence analysis

proliferation of commercial sequencing entities, and growing interest in personalized medicine have made whole genome sequencing commonplace. These technologies result in copious amounts of data to process, validate, interpret, and identify candidate variants, potentially contributing to the phenotypes of interest to researchers. Although scientific interests of groups that undertake translational projects differ dramatically, most of them converge on the final goal of their endeavor: identification of high-confidence genetic factors and molecular mechanisms contributing to the phenotype of interest.

An abundance of information generated by translational projects now poses a challenge to scientists and underscores the need for sophisticated tools to mine, integrate, and prioritize massive amounts of information [7]. Success in identification of key genetic factors that play a role in a disease or a complex biological process critically depends on seamless integration of clinical, genomic, and experimental data with prior knowledge about genotype–phenotype relationships accumulated in a plethora of publicly available databases. Such experimental data should be accessible to a variety of algorithms and analytical pipelines that drive

computational analysis and extraction of actionable knowledge [8]. Translational projects require an analytical engine that coordinates and performs various steps involved in extraction of meaningful information from accumulated clinical and experimental data in an orderly semiautomated manner. However, the above trends represent both an opportunity and a challenge for the extraction of useful knowledge from this wealth of data:

1. Challenges of high-throughput data management such as data transfer, data storage, access control, and data management
2. Challenges of scalable computational infrastructure
3. Challenges of analysis of large-scale multidimensional data for the extraction of actionable knowledge for the development of applications in biotechnology and biomedicine

3.2.1 Challenges of High-Throughput Data Management

The computational infrastructure that is required for maintenance, processing, and analysis of massive data sets is usually not available for individual laboratories and groups and is increasingly posing challenges even for large scientific centers [9]. Data transfers from sequencing centers or between collaborating research facilities and storage of the big data have become a substantial issue for translational medicine projects. These are followed by considerable computational challenges involved in memory- and data-intensive analyses. As the cost of sequencing technologies continues to diminish, the amount of raw data available for interpretation will continue to grow exponentially. These big data challenges in translational research demand advanced approaches in data management and scalable computational infrastructure based on state-of-the-art innovations in computer science.

3.2.2 Challenges in Analysis of Large-Scale Multidimensional Data for the Extraction of Actionable Knowledge

The human organism represents one of the most complex biological systems that science will need to unravel and cannot be understood without a systems (non-Cartesian) approach. While interest in the molecular mechanisms that lead to heritable disorders such as autism, diabetes, and predisposition to cardiovascular diseases and cancer is increasingly high, the discovery of the underlying genetic mechanisms has proven elusive despite some recent advances. The development of novel integrative approaches exploring biomedical phenomena on a systems level is critical for understanding molecular mechanisms behind the emergence of human phenotypes in health and disease. Common multigene disorders display complex patterns of inheritance that suggest the cumulative effects from many genetic variations, each

contributing only weakly to the overall disease phenotype [10]. Moreover, cellular components perform their functions through complicated networks of regulatory, metabolic, and protein interactions. This implies that the effects of different genetic variations will propagate within the molecular network, affecting the function of genes and gene products that otherwise carry no defects. Such intricate interdependencies between cellular components create the possibility of functional and causal relationships between apparently distinct disease phenotypes [11].

Studies of unique human disorders (e.g., psychiatric and behavioral) are further complicated by the fact that functional studies of these conditions in animal models are expensive, time consuming, and unsuitable for testing large-scale hypotheses [12, 13]. They may also not accurately recapitulate human phenotypes.

The development of predictive integrative multi-scale models of human disorders based on novel algorithmic approaches and fusion of genomic, clinical, contextual, and experimental information is critically important for further progress of high-throughput biology.

3.3 Typical Steps of Translational Data Analysis

Although translational projects are highly diverse in their scientific interests and approaches, a significant number of them converge on basic steps of analytical flow that are not dependent on a particular scientific question the research team plans to address. These include movement and storage of high-throughput data, identification of genetic variants, annotation and gene enrichment analysis, followed by the prediction of genes and molecular mechanisms contributing to the phenotypes of interest. Below we will investigate these crosscutting approaches in more detail (see Fig. 3.1).

3.3.1 Data Movement and Storage

With the availability and generation of large volumes of sequence data and analytical results, researchers face acute challenges in the areas of efficient and reliable data movement and availability of scalable infrastructure for data storage. One of the initial steps in a translational project involves transfers of large volumes of raw sequence data from sequencing centers to the points of compute or storage within a research lab facility. Typically research labs send batches of samples for sequencing and often times utilize multiple sequencing facilities. As such, significant time and effort is taken to just "ship" the results of sequence data analysis (e.g., exome, whole genome, and RNA-seq) from the sequencing center back to the investigator. Currently, most of the sequencing facilities send the raw analytical results on hard disks using snail mail. It can take a few weeks or even longer to receive these disks [14]. Furthermore, this approach adds additional risks of data corruption or loss of

data due to poor handling of the disks. Research groups are then required to mount these disks on local storage or compute servers to access the sequence data for further analysis. Sharing of the raw sequence data or the results of analysis with other collaborators also poses a challenge. Labs rely on conventional transfer protocols such as scp, ftp, http, which are inefficient in handling movement and sharing of large data. Research labs typically store all of their raw sequence data and results of analysis on local storage servers. We need better data management tools for efficient, reliable, and high-throughput data transfers; simple and intuitive data sharing; and scalable data storage resources to address exponentially growing volumes of data even at small research labs.

3.3.2 Automated Analytical Workflows

Once the sequence data is available, the next step is the analysis of the data using various analytical tools leading towards the identification of genetic variations. In recent years, a number of various standard approaches for execution of analytical pipelines for the high-throughput genomic data have emerged. Depending on the type of sequencing performed (exome, RNA-seq, ChIP-Seq, or whole genome), a combination of various approaches is applied, and various combinations of available tools are used to analyze the data. For example, the GATK best practices variant detection pipeline for exome data [15, 16] serves as a standard approach for the analysis of exome sequences involving various steps such as sequence alignment with a reference genome using tools such as BWA [17], base quality recalibration using the GATK toolkit [16], variant calling using GATK's Unified Genotyper, variant filtration, and variant annotation. Each research lab takes its own approach depending on their research requirements and preferences. One of the major requirements in all these approaches is the ability to define these analytical steps and perform them automatically over and over again for multiple data sets. In many research labs, typically these analytical steps are executed manually using semiautomated scripts (shell, Perl, Python, etc.) on local compute resources such as local clusters or large servers. In some labs, workflow management tools such as Galaxy [18] and Taverna [19] are used; however, handling the analysis of hundreds of genomes and exomes demands scalable solutions that can seamlessly integrate data management tools and also provide data provenance, reproducibility of analyses, and a collaborative and intuitive interface.

3.3.3 Annotation of Genetic Variants

Once genetic variations are identified, the first questions that intrigue investigators are as follows: "What is there? What functional categories are overrepresented in a list of thousands of identified genetic variations? What variations are most likely to contribute to the phenotypes of interest?"

Placing analytical results in a context of preexisting knowledge is essential for answering these questions and formulating an initial hypothesis regarding biological mechanisms potentially involved in a phenotype under consideration even if no prior hypothesis is available.

A wealth of information describing genetics, molecular physiology, and biological phenotypes for a variety of organisms and conditions was accumulated in public databases and private collections. A number of excellent resources have developed comprehensive knowledge bases specializing in a particular data type: genome-centric [EMBL [20], UCSC Genome Browser [21, 22], NCBI genomes [23]], protein centric [24–26], pathway centric [27–31], disease centric [32–35], pharmacogenetic [36, 37], and many others. Several public and private annotation engines are further integrating information from these already integrated resources to provide a one-stop annotation for a variety of data types [38–40].

3.3.3.1 Annotation of Genomic Features

Identification of functional SNPs responsible for variations in specific phenotypes, especially disease or drug response phenotypes of direct medical relevance, is a primary aim of numerous studies in the field of human genetics. Availability of a hypothesis about functional consequences of a sequence variant is important both for understanding of a disease genetic architecture and for the selection of the most informative (potentially causative) SNPs for genotyping. Traditionally, successful disease mapping studies have searched for sequence variants leading to amino acid changes in protein-coding regions where the functional consequences of protein-coding variants are easier to interpret. With the completion of the International HapMap project in 2005 [41, 42], and recent successes of the Encyclopedia of DNA Elements (ENCODE) project [16], it became clear that the polymorphisms with the strongest association with the majority of common disorders and all of their perfect proxies were found in the introns or intergenic regions of a genome, but not in linkage with coding regions. In September 2012, initial results of the ENCODE project were released in a coordinated set of 30 publications in major scientific journals [16, 43]. Published ENCODE data demonstrated that approximately 20 % of noncoding DNA in the human genome is functional, while the function of an additional 60 % is still unknown. Much of the functional noncoding DNA contains regulatory sites controlling the level, location, and chronology of gene expression, therefore making these regions especially interesting for biomedical investigation [44]. Furthermore, the expression of coding genes is apparently controlled by multiple regulatory sites both in proximity and distant from the gene [45]. The ENCODE data describing functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements controlling gene expression and cell behavior is available to a wide scientific community via the web portal at UCSC (http://encodeproject.org/).

3.3.3.2 Evolutionary Conservation

Comparison of DNA sequences from different species is a fundamental approach for identifying functional elements, such as exons or enhancers, as they tend to exhibit significant sequence similarity due to purifying selection [46, 47]. If two species are separated by sufficient evolutionary distance, their genome sequence comparisons will often reveal functional sequences as better conserved than their nonfunctional surroundings. Importance of conservation analysis was noted in the guide for epidemiologists. Comparative genomics, coupled with the increasing availability of sequenced vertebrate genomes, has significantly boosted identification of functional noncoding elements and their conservation across species.

3.3.4 Enrichment Analysis

However, the amount of data available for annotation of experimental results can be overwhelming. This calls for efficient statistical approaches to identify subcategories of interest for the user. Such approaches are implemented in a multitude of servers for enrichment analysis of sequence data reviewed in [48]. Gene enrichment analysis is based on the assumption that high-throughput analysis of multiple samples related to the same biological phenomenon (e.g., cohorts of patients with a particular disease or phenotype) would identify categories of functionally related genes enriched for genetic variations. Such rationale allows looking for groups of functionally related genes rather than individual genes and increases the likelihood of identification of the biological processes most likely contributing to the biological phenomena under investigation. The enrichment can be quantitatively measured by statistical methods. A large number of servers are providing tools for enrichment analysis weighing statistical association of gene lists provided by the user against a variety of biological annotation terms (e.g., GO categories, pathways, phenotypes). The enriched annotation terms associated with the large gene lists produced by high-throughput technologies (e.g., genetic variations or lists of co-expressed genes) will allow generating a working hypothesis regarding functional categories and molecular pathways contributing to a phenotype of interest.

Da Wei Huang et al. [48] classify all enrichment algorithms into three classes:

Class 1: Singular enrichment analysis (SEA) represents the most traditional and a very efficient strategy for analysis of very large gene lists generated by any high-throughput genomic studies. The enrichment *P*-value calculation, i.e., number of genes in the list that hit a given annotation class as compared to pure random chance, can be performed with the aid of well-known statistical methods such as chi-square, Fisher's exact test, binomial probability, and hypergeometric distribution and has been used in [49–51]. However, the SEA-based tools produce a list of terms that can be very large, and interpreting interrelationships between these terms can be prohibitively difficult.

Class 2: Gene set enrichment analysis (GSEA) is based on SEA, but with a distinct algorithm to calculate enrichment *P*-values as compared to SEA (see Subramanian et al. [52] for a detailed description of an algorithm). Limitations of GSEA approach are related to the fact that this method requires a summarized biological value (e.g., fold change) for each of the genome-wide genes as input—a difficult task considering complexity of biological functions.

Class 3: Modular enrichment analysis (MEA) utilizes the basic SEA enrichment calculation and incorporates additional network discovery algorithms by considering the term-to-term relationships. The key advantage of this approach is the introduction of term-to-term relationships that may lead to identification of multidimensional patterns spanning across multiple annotations that uniquely characterize the phenomenon of interest. MEA represents a substantial step forward in the discovery of complex interconnections between biological objects. Table 3.1 provides some of the examples of different classes of gene enrichment tools.

3.3.4.1 Meta-analysis Tools

Comprehensive meta-analyses of high-throughput association experiments typically combine heterogeneous data from genome-wide association (GWA) studies, protein–protein interaction screens, disease similarity, linkage studies, and gene expression experiments into a multilayered evidence network which is used to prioritize the entire protein-coding part of the genome identifying a shortlist of candidate genes (see Chap. III by C. Zhu et al. of this volume, Doncheva NT [53], and Moreau and Tranchevent [54] for a comprehensive review). The most popular meta-analysis tools include Endeavour [55], MetaRanker [56], and ToppGene [57].

Over 30 meta-analysis tools are listed at the Gene Prioritization Portal (available at http://www.esat.kuleuven.be/gpp/) [58]. This portal allows us to compare the available analytical tools based on the supported classes of annotations, as well as the data sources used for analysis, the implemented algorithmic approaches, and the user support capabilities.

3.3.5 Gene Prioritization and Network Reconstruction

Complex phenotype–genotype relations in biological systems are best conceptualized by the development of network-based models describing system's components and functional relationships between them. Such models allow generating predictive hypotheses regarding genetic factors contributing to phenotypes of interest. Although experimental studies provide the direct way for determining which cellular components interact and how, this approach may be prohibitively expensive in terms of time and resources and may not be feasible for some contexts (e.g., human brain). Meaningful "weighted" hypotheses generated by bioinformatics provide valuable

Table 3.1 Examples of different classes of gene enrichment analysis tools

Name	Paper ref	Tools home page	Type of enrichment	Statistical method	Multiple test correction
DAVID	[103]	http://david.abcc.ncifcrf.gov	SEA	Fisher's exact (modified as EASE score)	FDR, Bonferroni, Benjamin
FACT	[104]	http://www.factweb.de/	SEA	Adopt GeneMerge and GO::TermFinder statistical module	Adopt GeneMerge and GO::TermFinder statistical module
FunSpec	[105]	http://funspec.med.utoronto.ca	SEA	Hypergeometric	Bonferroni
GeneMerge	[106]	http://genemerge.bioteam.net/	SEA	Hypergeometric	Bonferroni
GoMiner	[107]	http://discover.nci.nih.gov/gominer	SEA	Fisher's exact	
MAPFinder	[108]	http://www.genmapp.org/	SEA	Z-score, hypergeometric	
Onto-express	[109]	http://vortex.cs.wayne.edu/projects.htm#Onto-Express	SEA	Fisher's exact; hypergeometric; binomial; chi-square	Bonferroni; Sidak; Holm; FDR
Web Gestalt	[110]	http://bioinfo.vanderbilt.edu/webgestalt/	SEA	Hypergeometric	
GeneTrail	[111]	http://genetrail.bioinf.unisb.de/enrichment_analysis.php?js=1&cc=1	GSEA	Hypogeometric; Kolmogorov–Smirnov v-like statistic	Benjamin–Hochberg
GSEA	[52]	http://www.broad.mit.edu/gsea	GSEA	Kolmogorov–Smirnov v-like statistic	
MetaGP	[112]	http://metagp.ism.ac.jp/	GSEA	Z-score	Bonferroni; Benjamin, etc.
Ontologizer	[113]	http://compbio.charite.de/index.php/ontologizer2.html	MEA	Fisher's exact	FWER; Bonferroni; resampling-based Westfall–Young correction
ProfCom	[114]	http://webclu.bio.wzw.tum.de/profcom/index.php	MEA	Greedy heuristics	Monte Carlo simulation
topGO	[115]	http://topgo.bioinf.mpi-inf.mpg.de/index.php	MEA	Fisher's exact	FDR
ENDEAVOUR	[55]	http://homes.esat.kuleuven.be/~bioiuser/endeavour/tool/endeavourweb.php	MEA		
ToppGene	[57]	http://toppgene.cchmc.org/	MEA		FDR; Bonferroni

guidance in planning the experiments and eventually leads to the development of efficient diagnostic and therapeutic strategies.

Gene prioritization aims to identify the most promising genes among a larger pool of candidates through integrative computational analysis of genomic data, as to maximize the yield and biological relevance of further downstream validation experiments and functional studies. In the past, several computational methods and tools have been proposed to rank promising candidate genes based on their probability of being the disease-causing genes using different data types, such as genomic or transcriptomic data. In a recent review, Moreau and Tranchevent [54] discussed a large variety of computational tools for prioritizing candidate genes for boosting disease gene discovery. In their review, the authors describe in detail the typical gene prioritization workflow, starting with generating a list of candidate genes, which may come from linkage, association, or chromosomal aberration studies, and followed by collecting prior knowledge about the disease, such as known disease genes or disease-related keywords or phenotypic descriptions. Then, a gene prioritization algorithm uses this collected prior knowledge and prioritizes the candidate genes using underlying genomic or transcriptomic data, whereas the resulting rank order of the candidate genes resembles the probability of being an actual disease causing gene. The top ranked genes may then be followed up in downstream experimental validations.

In the last couple of years, several reviews have discussed existing gene prioritization methods. One of the first reviews that compared different gene prioritization tools was published by Tiffin et al. [59] who reviewed seven independent computational disease gene prioritization methods and then applied them to the analysis of a set of candidate genes for type 2 diabetes (T2D) and the related trait obesity. As a result, the authors could generate and analyze nine primary candidate genes for T2D genes and five for obesity, whereas two genes were common to these two sets. In a more recent review, Tiffin et al. [60] discussed different approaches for the identification of candidate disease genes, such as using known disease genes as training genes for inferring new disease genes, using protein–protein networks for identifying disease-related genes, or using disease phenotypes for identifying phenotypic-related genes. Recently, Tranchevent et al. (2011) [58] developed the Gene Prioritization Portal, an online resource that was designed to help biologists and geneticists to select the prioritization methods that best correspond to their needs. The portal is frequently updated and currently summarizes 33 gene prioritization tools (as of April 2013) that are freely accessible as web tools. To assess their differences, the authors compare the underlying data sources, the required input data (training data and candidate genes), as well the output they provide as described in [55].

To assess the performance of a gene prioritization method, cross-validation is a widely used statistical benchmark approach for gene prioritization tools, such as Endeavour [7] or GeneWanderer [61]. However, cross-validation on retrospective data may provide performance estimates likely to be overoptimistic because some of the data sources are contaminated with knowledge from the disease gene association. To address this challenge, Börnigen et al. [62] suggested an unbiased

evaluation of gene prioritization tools, comparing the performance of several gene prioritization tools on novel data by selecting recently reported disease gene associations manually from literature and making predictions immediately after publication. By doing this, the authors could guarantee that the novel disease gene association of interest was not yet included in the databases that underlie the gene prioritization tools that would mimic a novel gene discovery, and their results show that all tools had lower AUC values than previously reported. Therefore, the authors believe that developers should take extra care when benchmarking their tools and that this field needs to consolidate through improved benchmarking efforts due to the lack of a ground truth for evaluating the performance of prioritization methods.

As illustrated in [55], gene prioritization methods may rely on a large variety on different data sources, such as literature, protein interactions, experimental conditions, pathway information, or phenotypic relationships. Many of these methods integrate protein or regulatory interactions with additional information, such as experimental conditions (microarray or GWAS), or phenotypic information, for prioritizing genes in a network-based approach. The underlying protein or regulatory interactions can be of different nature, such as:

1. *Directed graphs* representing steps of a particular pathway or a biological process, such as metabolic or regulatory pathways as chains of binary interactions between systems components
2. *Undirected graphs* representing the interactions where the direction of a signal is not known (e.g., protein–protein interactions)
3. *Unordered lists* of genes that are known to have some functional relationships with each other, but information describing binary interactions between the genes in the list is not available (e.g., gene expression data, disease associations)

This information is used by network-based gene prioritization algorithms.

3.4 End-to-End Solution for Translational Genomics

3.4.1 An Overview

In a collaborative effort between the Department of Human Genetics and Computation Institute at the University of Chicago as well as VISTA project at Lawrence Berkeley National Laboratory (LBL), we have developed an integrated *end-to-end solution* to support the extraction of actionable knowledge from the massive data sets generated by translational medicine projects. This integrated computational platform leverages the following advanced computational technologies and bioinformatics resources being developed by the participating groups:

(a) *Globus Genomics—Combining Globus Online* [63]*, Galaxy* [18]*, and cloud-based scalable computational infrastructure* for secure high-performance data transfer, data storage, graphical workflow definition, and high-throughput computations being developed at the Computation Institute at the University

of Chicago. This leading-edge scientific computing and data management infrastructure is a software-as-a-service [64, 65] offering widely used by the global research community and seamlessly integrates with the Galaxy Platform for identification and annotation of genetic variations in raw sequence data.

(b) *VISTA suite of bioinformatics tools* to support identification of coding and non-coding genomic features and evolutionary analysis of genomic intervals. VISTA [66] provides investigators with a unified framework for the alignment of long genomic sequences and whole genome assemblies, interactive visual analysis of alignments along with functional annotation, and many other comparative genomics capabilities. VISTA Browser allows to examine pre-computed pairwise and multiple alignments of whole genome assemblies; Whole Genome rVISTA supports the identification of transcription factor binding sites conserved between species and overrepresented in upstream regions of groups of genes. VISTA RViewer [67], a new addition to the VISTA suite of tools for comparative genomics, calculates a set of important parameters for both coding and noncoding regions and provides investigators with capabilities to compare the intervals and identify those that are likely to be significant in a particular study. The VISTA Enhancer Browser [68] is a central resource for experimentally validated human and mouse noncoding fragments with gene enhancer activity as assessed in transgenic mice. It also links to mouse phenotype and gene expression data, thus allowing for comparisons and prioritization of significant genes and non-coding regions across a region.

(c) *Lynx-integrated knowledge base and bioinformatics workbench* for annotation and extraction of meaningful information from genomic data generated by sequencing projects. Lynx workbench includes tools for gene enrichment and prioritization and reconstruction of network-based disease models. These tools support prediction of high-confidence genetic factors and generation of the testable hypotheses regarding molecular mechanisms underlying neurodevelopmental disorders of interest.

The sections below demonstrate the utility of the proposed approach for the analysis of the genomic data.

3.4.2 Scalable Computational Infrastructure for High-Throughput Genomics

An integrated platform combining advanced data transfer capabilities, graphical workflow tools, and elastic cloud-based infrastructure is the subject of research and development efforts at the Computation Institute at University of Chicago. The resultant platform or solution, called Globus Genomics, attempts to address these set of challenges and is anchored by Globus Online.

Globus Online is an advanced computational platform to support high-throughput sciences. Globus Online addresses a central problem in the emerging world of big data research: moving large quantities of information reliably, efficiently, and securely

among data centers, scientific facilities, research laboratories, and supercomputing sites where data are produced, transformed, stored, and consumed. We leverage advanced Globus components such as:

- *Globus Transfer*—a high-performance and fault-tolerant file transfer service being developed at the Computation Institute to seamlessly move data from sequencing centers to research labs and then to available compute resources
- *Globus Sharing*—a file sharing service that enables researchers to easily share data with their collaborators, wherever the data resides
- *Globus Nexus*—provides user identity, profile, and group management capabilities, allowing users to seamlessly navigate between various components such as Globus Transfers, Globus Store, Globus Galaxy, and the Lynx system
- *Security*—a Grid Security Infrastructure (GSI) [69] for secure communication, authentication, delegation, and single sign-on (SSO) capabilities at various components of a distributed infrastructure

Globus Genomics platform integrates Globus Online with the Galaxy workflow management platform. The resultant solution provides access to a wide variety of biomedical tools for the analysis of sequence data and allows Galaxy to handle massive amounts of data from within the platform. A Globus Genomics platform can be hosted on elastic cloud infrastructure (e.g., Amazon AWS [70]) that allows on-demand scaling of computational and storage resources and supports parallel analysis of multiple exomes/whole genomes for faster results. Galaxy takes away the mundane process of running individual tools one after another and allows researchers to weave their analytical steps in the form of a workflow and repeatedly run it on various data sets. The Globus Genomics solution essentially brings the massive amounts of data to the point of analysis and allows researchers to store their analytical steps.

In addition to solving the two major challenges of data management and analysis, the Globus Genomics platform attempts to address remaining problem areas:

3.4.2.1 Scalability

The Globus Genomics analytical platform uses Amazon Web Services' (AWS) on-demand computational and storage resources. The platform is designed to use Amazon's EC2 to scale the computational requirements of analyzing large numbers of sequencing projects. Galaxy's workflow and job management backend has been modified to use Condor [71] to dynamically submit jobs to Amazon spot instances on demand. It provides virtually unlimited scalability and is capable of analyzing hundreds of genomes in parallel. Researchers who have typically waited weeks or months to get their data analyzed can now get it completed in a matter of hours or days, enabled by on-demand scalable access to compute resources. In addition to the compute resources, using Globus Online endpoints [72] on AWS allows research labs to utilize AWS's storage solutions such Elastic Block Store (EBS), Simple Storage Service, and Amazon Glacier to store their rapidly growing data sets.

Globus Online endpoints allow researchers to seamlessly move their data into AWS storage for temporary staging or permanent archiving.

3.4.2.2 Provenance and Reproducibility

Galaxy addresses an important aspect of research: provenance and reproducibility of data sets and computational steps. It stores detailed metadata on every data transformation during the analysis, in the form of a history, and allows researchers to track provenance information. It allows researchers to extract a workflow from a history of events and thus reproducing the results.

3.4.2.3 Sharing and Collaboration

While data management and data analysis are challenges for each biomedical research group, a scenario that is commonly encountered is sharing and collaborating on the data and analysis. We find many research groups that regularly fund the sequencing projects for patients of interest often share the data and collaborate on the analysis. Globus Online and Galaxy integrated together offers an excellent solution. Globus Online Sharing provides a data sharing service for big data that is typical of scientific collaborations. By setting up Globus Online endpoints by various collaborators, they can easily share the data. Galaxy on the other hand provides a robust solution to share the analytical environment in the form of workflows representing different types of analysis and histories representing various analytical steps used to derive results.

3.4.2.4 Economic Considerations

With the use of on-demand AWS spot instances and storage resources, the Globus Genomics platform provides a very cost-effective model to many research groups for the analysis of their growing volume of data. It eliminates the requirement of purchasing and maintaining costly compute and storage resources especially for medium to small research groups. With AWS, researchers can handle their peak computational requirements and pay only the costs incurred by the actual usage of the resources.

3.4.2.5 Platform-as-a-Service (PaaS) Model

An important feature of the Globus Genomics offering is the platform-as-a-service model that allows users to manage their data movement and complex analysis of the data via simple web interfaces. Most of the solutions available today act as a black box where the type of analysis, tools used, and the parameters used are either

unknown or not possible to control by the users. With Globus Genomics, the researchers are directly involved and are in total control right from moving the data from the sequencing centers to planning their analysis by selecting the tools and parameters via a simple web interface.

Tight correlation between the growing computational needs of application sciences and the ongoing research in computational sciences is an essential factor in meeting big data challenges of contemporary translational medicine.

3.4.3 The Discovery-Based Approach Using Lynx and VISTA Systems

Sometimes the researchers do not have any prior hypothesis regarding molecular mechanisms involved in the disorder or phenotype under investigation. In these cases, the bioinformatics discovery-based approach may aid in the development of a working hypothesis to concentrate investigation on the mechanisms that are the most relevant to the phenomenon under study. Here we will describe two discovery-based approaches offered by our analytical pipeline. These include an extensive annotation of the identified genetic variants with the prior knowledge from VISTA and Lynx knowledge bases and gene enrichment analysis offered by Lynx.

3.4.3.1 Annotations by VISTA and RViewer

The described analytical pipeline uses a web-based RViewer component [67] developed by the VISTA project team for identification of genetic variations residing in genomic regulatory signals of both coding and noncoding regions. RViewer presents several functionalities unique for this server. First, it allows the display and analysis of multiple genomic intervals simultaneously, allowing for interactive visual inspection of the evolutionary conservation of genomic regions and their associated functional features. Second, it provides unique annotations of the coding and noncoding intervals such as experimentally verified enhancers [68], UTRs, and transcription factor binding site clusters [73] (Fig. 3.3). Prediction of genetic variations residing in these functional elements is used by subsequent analyses for enhancing the network-based gene prioritization by the described analytical pipeline. RViewer can be used as a *plugin to make RViewer's functionality available to other applications*.

3.4.3.2 Annotations by Lynx

In the described analytical pipeline, annotations provided by VISTA are then further supplemented with information from the integrated Lynx knowledge base. This resource integrates multiple classes of biomedical data from over 35 public

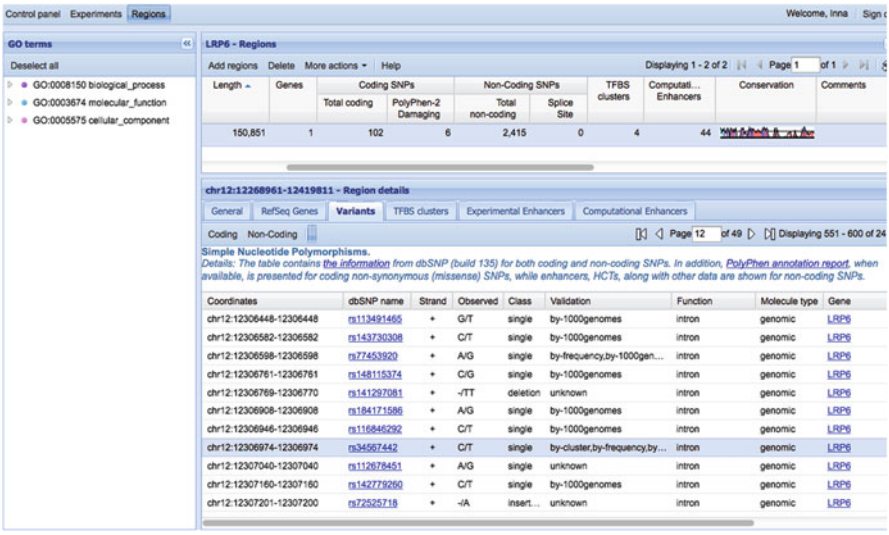


Fig. 3.3 VISTA RViewer displaying regions analyzed in this case study (*top panel*), with annotation of region length, number of RefSeq genes, coding and noncoding SNPs, as well as HCTs and enhancers. A conservation *plot* is displayed in the last column. The region details view (*bottom panel*) displays RefSeq genes contained within a selected region, with links to KEGG and mouse phenotype data

databases and private collections, metadata such as provenance and cross-reference data connecting different sources, as well as clinical and genomic data provided by translational projects into a relational data warehouse (Fig. 3.4). The following public data sources are currently integrated into the Lynx KB: NCBI databases, EMBL [20], UniProt [24], molecular pathways (e.g., Reactome [27], BioCarta [28], KEGG [29]), NCI pathways [30]), phenotypic (OMIM [32], disease ontology [34], phenotype ontology databases [35]) databases and ontologies (GO [74], BioPAX [31], MI- PSI [26], etc.), text mining (GeneWays [75]), and many others. Data in the Lynx KB is modeled, cross-referenced, and stored in a relational database. The integrated knowledge base is used for data annotation and further analysis by algorithms for feature-based gene enrichment and prediction of high-confidence genes and molecular networks contributing to phenotypes of interest. Annotated results analyses are also presented to the user in a graphical or tabular form via web-based interface for interactive analysis [Lynx].

3.4.3.3 Gene Enrichment Analysis

The wealth of information provided by the Lynx and VISTA annotation engines, however, may be overwhelming for the user. Our analytical pipeline offers enrichment analysis tools that allow investigators to identify functional categories

overrepresented in the experimental results, thus providing the clues regarding the potential molecular mechanisms involved in the phenomena under the study. Two SEA algorithms, Bayes factor and P-value estimates, are used in our pipeline for this purpose. The analysis can be performed against the complete human proteome or against particular sets of genes (e.g., genes expressed in a particular tissue of interest: brain, liver, etc.). The following categories of information are supported: gene ontology terms, disease associations, pathway enrichment and enrichment in particular phenotypes or genomic features (e.g., enhancers, TFBS clusters). Lynx enrichment analysis also supports the use of the customized ontologies provided by user (e.g., brain connectivity ontology developed in collaboration with Drs. W. Dobyns and A. Paciorkowski, University of Washington). The results of these analyses will allow us to identify functional groups enriched with genetic variants and narrow the task of identification of genetic variants to the groups of genes relevant to the investigation.

3.4.4 Network-Based Gene Prioritization

Lynx toolkit contains a network-based gene prioritization workbench based on the PINTA tool, which was developed by the Moreau lab [76]. This suite includes five random walk-based algorithms for network propagation, namely, heat kernel diffusion, simple random walk, PageRank with priors, HITS with priors, and K-step Markov model algorithm, using the STRING network [77] as the underlying protein interaction network.

Originally, PINTA was developed to use expression data as input data, but here, we replaced the continuous data (coming from expression data) with binary data using seed genes coming from a set of ranked genes known to be associated with the process or phenotype of interest: a 1 is fed as an input for each seed gene, and a 0 is associated to all non-seed genes. Further, the algorithms were modified to accommodate a variety of weighted data types to be used for gene prioritization including *inter alia* ranked gene to phenotype associations, weighted canonical pathways, gene expression, and NGS data.

3.4.4.1 Identification of Genetic Factors Contributing to the Pathogenesis of Spina Bifida Using the Described Analytical Pipeline

We will illustrate the utility of the network-based hypothesis-based approach on the example of identification of genetic variations contributing to pathogenesis of *spina bifida* (SB).

Spina bifida is the most common congenital birth defect and manifests itself by the incomplete closing of the embryonic neural tube [78, 79]. Some vertebrae overlying the spinal cord are not fully formed and remain unfused and open.

As SB is a multifactorial, complex disorder, whole genome sequencing was performed for four SB patients and four parents (with no SB phenotype) from the same consanguineous family. Our initial hypotheses were based on three mechanisms that have been suggested to be causative for *spina bifida*, namely, (1) compromised folate metabolism and/or transport [80, 81] in mothers or probands, (2) Defects in Wnt planar cell polarity pathway (PCP) [82, 83], and (3) potential mouse mutants with SB [84, 85] for the study in human neural tube defect etiology. In order to identify molecular mechanism(s) causative for each of the patients, all three potential mechanisms were investigated. First, the sets of genes known to be associated with each of these mechanisms (e.g., genes involved in folate biosynthesis or transport) were extracted and ranked according to the strength of their association with the SB phenotype using a method described in https://www.dbdb.urmc.rochester.edu/loe. These lists of ranked genes were used as seeds for the network-based gene prioritization experiments. Network reconstruction/prioritization was performed using heat kernel diffusion algorithm with default parameters. Genes highly ranked by the algorithm (*E*-values <0.05) were extracted and used for network visualization and further investigation.

The reconstructed folate transport and biosynthesis network allowed identifying potentially causal genetic variations in both parents and probands. The network-predicted cubilin gene (CUBN), encoding a receptor for B12 complex with high significance, was found to contain deleterious mutations in both mothers. The gene also had another variant (rs703064; rs56059527) that was homozygous in one of the children and heterozygous in both parents from the consanguineous family under investigation. Vitamin B(12) has been previously shown to play a major role in folate metabolism [86, 87].

Moreover, both the mothers also showed an exonic variant (rs1051266) in the SLC19A1 gene encoding folate placental transporter. Such polymorphisms in vitamin B receptor (CUBN) and in SLC19A1 (RFC) in mothers have been previously shown to result in spina bifida or other neural tube defects [88–90] supporting the above inferences.

The network-based gene prioritization was also performed for the Wnt planar cell polarity seed genes. Annotations of the noncoding regions by VISTA allowed identifying an intronic variant in the LRP6 gene, serving as a co-receptor for a number of WNT genes. Lrp6 has been shown to be interacting with Wnt5a in mouse models in Wnt/beta-catenin pathway [91]. The Lrp6−/− embryos showed neural tube defects, and the use of dietary folic acid supplementation helped reduce the disorder occurrence [92], thus strengthening the connection of the folate and Wnt signaling to the pathogenesis of SB. Network analysis and reconstruction allowed predicting additional disease candidate genes and mechanisms that could be contributing to this neural tube defect in patients.

Hence, the above example showcases the need for an integrated approach of the coding and noncoding signals from both hypothesis-based network prioritization and discovery-based analysis for the identification of candidate genes contributing to complex phenotypes. These genes will be further validated by an iterative process of experimental validations.

3.5 Future Directions

Recently, a number of new directions started to emerge in integrative translational genomics and informatics. These include contextual approach to modeling of biological systems and exploration of their modular organization, comparative phenomics, and the studies of the impact of environmental factors on biological systems (reviewed in Chap. 1 by G. Gibson of this volume).

These distinct but related scientific strategies promise to significantly increase our understanding of the living systems and unveil mechanisms of complex disorders that are still poorly understood.

3.5.1 Contextual and Modular Organization of Biological Systems

Systems biology approach is contextual by definition. It studies an emergent behavior of self-organizing biological systems in relevant biological contexts: organismal, spatial, and temporal. Yet, despite the importance of context, most of the current approaches to modeling and analysis of biological systems disregard contextual information, thus significantly reducing the predictive power of the developed models. The need for the development of comprehensive approaches for the reconstruction of context-specific maps of molecular interactions has been expressed in a number of publications [93–95]. Recently, several successful efforts in this direction were reported in the literature [95, 96].

3.5.2 Comparative Phenomics

In recent years, it became increasingly evident that human diseases are related to each other and share common phenotypic features and pathophysiological mechanisms. Rzhetsky et al. [97] were the first to explore overlap in susceptibility as a way to find common genetic determinants for multifactorial diseases. This study has demonstrated that phenome should be regarded as a network of interrelated diseases and disease traits rather than a list of distinct disease entities. It is now widely accepted by the scientific community that systematic investigation of relationships between different disorders will provide new insights into etiology, pathogenesis, and classification of the diseases and will assist in the development of new therapeutic strategies [98–102].

The described trends, however, require efficient computational infrastructure and algorithmic support to support exploration of biological complexity on these new levels. Seamless integration of the scientific efforts of the researches in the fields of biology, computer science, mathematics, and informatics will ensure the success of these future endeavors.

Acknowledgement This work is supported in part by Mr. and Mrs. Lawrence Hilibrand, the Boler Family Foundation, and NIH/NINDS grant NS050375—The Genetic Basis of Mid-Hindbrain Malformations.

References

1. Ranganathan S, Schönbach C, Kelso J, Rost B, Nathan S, Tan TW (2011) Towards big data science in the decade ahead from ten years of InCoB and the 1st ISCB-Asia Joint Conference. BMC Bioinforma 12(Suppl 13):S1. doi:10.1186/1471-2105-12-S13-S1
2. Chen J, Qian F, Yan W, Shen B (2013) Translational biomedical informatics in the cloud: present and future. Biomed Res Int 2013:658925. doi:10.1155/2013/658925
3. Payne PR, Embi PJ, Sen CK (2009) Translational informatics: enabling high-throughput research paradigms. Physiol Genomics 39(3):131–140. doi:10.1152/physiolgenomics.00050.2009
4. Schuler R, Smith DE, Kumaraguruparan G, Chervenak A, Lewis AD, Hyde DM et al (2012) A flexible, open, decentralized system for digital pathology networks. Stud Health Technol Inform 175:29–38 [Research Support, N.I.H., Extramural]
5. Boyd LB, Hunicke-Smith SP, Stafford GA, Freund ET, Ehlman M, Chandran U, Dennis R, Fernandez AT, Goldstein S, Steffen D, Tycko B, Klemm JD (2011) The caBIG® life science business architecture model. Bioinformatics 27(10):1429–1435. doi:10.1093/bioinformatics/btr141
6. Hillman-Jackson J, Clements D, Blankenberg D, Taylor J, Nekrutenko A, Galaxy Team (2012) Using Galaxy to perform large-scale interactive data analyses. Curr Protoc Bioinformatics; Chapter 10:Unit10.5. doi:10.1002/0471250953.bi1005s38
7. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F et al (2006) Gene prioritization through genomic data fusion. Nat Biotechnol 24(5):537–544 [Research Support, Non-U.S. Gov't]
8. Knaup P et al (2004) Towards clinical bioinformatics: advancing genomic medicine with informatics methods and tools. Methods Inf Med 43(3):302–307
9. Desai AN, Jere A (2012) Next-generation sequencing: ready for the clinics? Clin Genet 81(6):503–510
10. Bill BR, Geschwind DH (2009) Genetic advances in autism: heterogeneity and convergence on shared pathways. Curr Opin Genet Dev 19(3):271–278
11. Iossifov I, Zheng T, Baron M, Gilliam TC, Rzhetsky A (2008) Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. Genome Res 18(7):1150–1162. doi:10.1101/gr.075622.107, Epub 2008 Apr 16. PubMed PMID: 18417725; PubMed Central PMCID: PMC2493404
12. Sarnyai Z, Alsaif M, Bahn S, Ernst A, Guest PC, Hradetzky E, Kluge W, Stelzhammer V, Wesseling H (2011) Behavioral and molecular biomarkers in translational animal models for neuropsychiatric disorders. Int Rev Neurobiol 101:203–238. doi:10.1016/B978-0-12-387718-5.00008-0, Review. PubMed PMID: 22050853
13. de Mooij-van Malsen AJ, Vinkers CH, Peterse DP, Olivier B, Kas MJ (2011) Cross-species behavioural genetics: a starting point for unravelling the neurobiology of human psychiatric disorders. Prog Neuropsychopharmacol Biol Psychiatry 35(6):1383–1390. doi:10.1016/j.pnpbp.2010.10.003, Epub 2010 Oct 16. Review. PubMed PMID: 20955750
14. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP (2010) Computational solutions to large-scale data management and analysis. Nat Rev Genet 11(9):647–657
15. Broad Institute Best Practice Variant Detection. http://gatkforums.broadinstitute.org/discussion/1186/best-practice-variant-detection-with-the-gatk-v4-for-release-2-0
16. McKenna A et al (2010) The genome analysis toolkit: a map reduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303

17. Li H, Durbin R (2010) Fast and accurate long-read alignment with burrows–wheeler transform. Bioinformatics 26(5):589–595
18. Goecks J, Nekrutenko A, Taylor J, Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11(8):R86
19. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P, Bhagat J, Belhajjame K, Bacall F, Hardisty A, Nieva de la Hidalga A, Balcazar Vargas MP, Sufi S, Goble C (2013) The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Res, First published online 2 May 2013. doi:10.1093/nar/gkt328
20. Kulikova T et al (2007) EMBL Nucleotide Sequence Database in 2006. Nucleic Acids Res 35:D16–D20
21. Karolchik D, Hinrichs AS, Kent WJ (2012) The UCSC Genome Browser. Curr Protoc Bioinformatics; Chapter 1:Unit1.4. doi:10.1002/0471250953.bi0104s40. PubMed PMID: 23255150
22. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, Lee BT, Barber GP, Harte RA, Diekhans M, Long JC, Wilder SP, Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ (2013) ENCODE data in the UCSC genome browser: year 5 update. Nucleic Acids Res 41(Database issue): D56–D63. doi:10.1093/nar/gks1172, Epub 2012 Nov 27. PubMed PMID: 23193274; PubMed Central PMCID: PMC3531152
23. NCBI Resource Coordinators (2013) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 41(Database issue):D8–D20. doi:10.1093/nar/gks1189, Epub 2012 Nov 27. PubMed PMID: 23193264; PubMed Central PMCID: PMC3531099
24. UniProt Consortium (2013) Update on activities at the universal protein resource (UniProt) in 2013. Nucleic Acids Res 41(Database issue):D43–D47. doi:10.1093/nar/gks1068, Epub 2012 Nov 17. PubMed PMID: 23161681; PubMed Central PMCID: PMC3531094
25. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35(Database issue):D61–D65, Epub 2006 Nov 27. PubMed PMID: 17130148; PubMed Central PMCID: PMC1716718
26. Hermjakob H et al (2004) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. Nat Biotechnol 22(2):177–183
27. Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L (2007) Reactome: a knowledge base of biologic pathways and processes. Genome Biol 8(3):R39
28. BioCarta Pathways. http://biocarta.com/
29. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Yeast 17(1):48–55
30. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH (2009) PID: the Pathway Interaction Database. Nucleic Acids Res 37:D674–D679
31. BioPAX-Consortium (2006) BioPAX: biological pathways exchange. http://www.biopax.org/
32. Online Mendelian Inheritance in Man (OMIM). http://www.ncbi.nlm.nih.gov/omim/
33. Mottaz A, Yip YL, Ruch P, Veuthey AL (2008) Mapping proteins to disease terminologies: from UniProt to MeSH. BMC Bioinforma 9(Suppl 5):S3
34. Disease ontology. http://diseaseontology.sourceforge.net/
35. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S (2008) The HUMAN Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 83(5):610–615
36. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 39(Database issue):D1035–D1041, PMID: 21059682

37. Davis AP, Wiegers TC, Johnson RJ, Lay JM, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, Murphy CG, Mattingly CJ (2013) Text mining effectively scores and ranks the literature for improving chemical-gene-disease curation at the comparative toxicogenomics database. PLoS One 8(4):e58201. doi:10.1371/journal.pone.0058201
38. Kanehisa M (1997) Linking databases and organisms: GenomeNet resources in Japan. Trends Biochem Sci 22(11):442–444, Review. PubMed PMID: 9397687
39. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH (2010) The NCBI BioSystems database. Nucleic Acids Res 38(Database issue):D492–D496. doi:10.1093/nar/gkp858, Epub 2009 Oct 23. PubMed PMID: 19854944; PubMed Central PMCID: PMC2808896
40. Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL (2008) The vertebrate genome annotation (Vega) database. Nucleic Acids Res 36(Database issue):D753–D760, Epub 2007 Nov 14. PubMed PMID: 18003653; PubMed Central PMCID: PMC2238886
41. Altshuler DM, Gibbs RA, Peltonen L et al (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52–58. doi:10.1038/nature09298, PubMed PMID: 20811451; PubMed Central PMCID: PMC3173859
42. Buchanan CC, Torstenson ES, Bush WS, Ritchie MD (2012) A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. J Am Med Inform Assoc 19(2):289–294. doi:10.1136/amiajnl-2011-000652, PubMed PMID: 22319179; PubMed Central PMCID: PMC3277631
43. Maher B (2012) ENCODE: the human encyclopaedia. Nature 489(7414):46–48, PubMed PMID: 22962707
44. ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ et al (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74. doi:10.1038/nature11247. PubMed PMID: 22955616; PubMed Central PMCID: PMC3439153
45. Pennisi E (2012) Genomics. ENCODE project writes eulogy for junk DNA. Science 337(6099):1159, 1161. doi:10.1126/science.337.6099.1159. PubMed PMID: 22955811
46. Hardison RC (2003) Comparative genomics. PLoS Biol 1(2):E58
47. Cheng JF, Priest JR, Pennacchio LA (2007) Comparative genomics: a tool to functionally annotate human DNA. Methods Mol Biol 366:229–251
48. da Huang W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1):1–13
49. Curtis RK, Oresic M, Vidal-Puig A (2005) Pathways to the analysis of microarray data. Trends Biotechnol 23(8):429–435
50. Khatri P, Draghici S (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 21(18):3587–3595
51. Rivals I, Personnaz L, Taing L, Potier MC (2007) Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23(4):401–407 [Evaluation Studies]
52. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102(43):15545–15550
53. Doncheva NT, Kacprowski T, Albrecht M (2012) Recent approaches to the prioritization of candidate disease genes. Wiley Interdiscip Rev Syst Biol Med 4(5):429–442 [Research Support, Non-U.S. Gov't Review]
54. Moreau Y, Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet 13(8):523–536
55. Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B et al (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res 36(Web Server issue):W377–W384 [Research Support, Non-U.S. Gov't]

56. Pers TH, Dworzyñski P, Thomas CE, Lage K, Brunak S (2013) MetaRanker 2.0: a web server for prioritization of genetic variation data. Nucleic Acids Res 41(Web Server issue): W104–W108
57. Chen J, Bardes EE, Aronow BJ, Jegga AG (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res 37(Web Server issue): W305–W311 [Research Support, N.I.H., Extramural]
58. Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y (2011) A guide to web tools to prioritize candidate genes. Brief Bioinform 12(1):22–32 [Research Support, Non-U.S. Gov't Review]
59. Tiffin N, Adie E, Turner F, Brunner HG, van Driel MA, Oti M et al (2006) Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. Nucleic Acids Res 34(10):3067–3081 [Evaluation Studies Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]
60. Tiffin N, Andrade-Navarro MA, Perez-Iratxeta C (2009) Linking genes to diseases: it's all in the data. Genome Med 1(8):77
61. Kohler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82(4):949–958 [Evaluation Studies Research Support, Non-U.S. Gov't]
62. Börnigen D, Tranchevent LC, Bonachela-Capdevila F, Devriendt K, De Moor B, De Causmaecker P et al (2012) An unbiased evaluation of gene prioritization tools. Bioinformatics 28(23):3081–3088 [Research Support, Non-U.S. Gov't]
63. Foster I (2011) Globus online: accelerating and democratizing science through cloud-based services. IEEE Internet Comput 15:70–73
64. Dubey A, Wagle D (2007) Delivering software as a service. The McKinsey Quarterly 6:1–12
65. Waters B (2005) Software as a service: a look at the customer benefits. J Digit Asset Manag 1(1):32–39
66. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. Nucleic Acids Res 32(Web Server issue):W273–W279
67. Lukashin I, Novichkov P, Boffelli D, Paciorkowski AR, Minovitsky S, Yang S, Dubchak I (2011) VISTA Region Viewer (RViewer)–a computational system for prioritizing genomic intervals for biomedical studies. Bioinformatics 27(18):2595–2597
68. Visel A, Minovitsky S, Dubchak I, Pennacchio LA (2007) VISTA Enhancer Browser–a database of tissue-specific human enhancers. Nucleic Acids Res 35(Database issue):D88–D92
69. Foster I, Kesselman C, Tsudik G, Tuecke SA (1998) Security architecture for computational grids. 5th ACM conference on computer and communications security conference, 1998, San Francisco, CA, USA pp 83–92
70. Amazon Web Services. http://aws.amazon.com
71. Litzkow M, Livny M, Mutka M (1998) Condor – a hunter of idle workstations. Proceedings of the 8th international conference of distributed computing systems, June 1988, San Jose, CA, USA pp 104–111
72. Foster I, Kesselman C, Tuecke S (2001) The Anatomy of the Grid: Enabling Scalable Virtual Organizations. Int J Supercomput Appl 15(3):200–222
73. Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I (2010) Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res 20(5):565–577
74. Gene Ontology Consortium (2006) The gene ontology (GO) project in 2006. Nucleic Acids Res 34:D322–D326
75. Rzhetsky A et al (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Inform 37(1):43–53
76. Nitsch D, Tranchevent LC, Goncalves JP, Vogt JK, Madeira SC, Moreau Y (2011) PINTA: a web server for network-based gene prioritization from expression data. Nucleic Acids Res 39(Web Server issue):W334–W338 [Research Support, Non-U.S. Gov't]

77. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P et al (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39(Database issue):D561–D568 [Research Support, Non-U.S. Gov't]
78. Padmanabhan R (2006) Etiology, pathogenesis and prevention of neural tube defects. Congenital anomalies 46(2):55–67
79. Mitchell LE, Adzick NS, Melchionne J, Pasquariello PS, Sutton LN, Whitehead AS (2004) Spina bifida. Lancet 364(9448):1885–1895. doi:10.1016/S0140-6736(04)17445-X, ISSN 0140-6736
80. Boyles AL, Billups AV, Deak KL, Siegel DG, Mehltretter L, Slifer SH et al (2006) Neural tube defects and folate pathway genes: family-based association tests of gene-gene and gene-environment interactions. Environ Health Perspect 114(10):1547–1552 [Research Support, N.I.H., Extramural]
81. Ross ME (2010) Gene-environment interactions, folate metabolism and the embryonic nervous system. Wiley Interdiscip Rev Syst Biol Med 2(4):471–480
82. Wu G, Huang X, Hua Y, Mu D (2011) Roles of planar cell polarity pathways in the development of neural [correction of neutral] tube defects. J Biomed Sci 18:66
83. Wen S, Zhu H, Lu W, Mitchell LE, Shaw GM, Lammer EJ et al (2010) Planar cell polarity pathway genes and risk for spina bifida. Am J Med Genet A 152A(2):299–304
84. Harris MJ, Juriloff DM (2007) Mouse mutants with neural tube closure defects and their role in understanding human neural tube defects. Birth Defects Res A Clin Mol Teratol 79(3):187–210
85. Harris MJ, Juriloff DM (2010) An update to the list of mouse mutants with neural tube closure defects and advances toward a complete genetic perspective of neural tube closure. Birth Defects Res A Clin Mol Teratol 88(8):653–669
86. Kozyraki R, Fyfe J, Kristiansen M, Gerdes C, Jacobsen C, Cui S et al (1999) The intrinsic factor-vitamin B12 receptor, cubilin, is a high-affinity apolipoprotein A-I receptor facilitating endocytosis of high-density lipoprotein. Nat Med 5(6):656–661 [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.]
87. Wahlstedt-Froberg V, Pettersson T, Aminoff M, Dugue B, Grasbeck R (2003) Proteinuria in cubilin-deficient patients with selective vitamin B12 malabsorption. Pediatr Nephrol 18(5):417–421 [Research Support, Non-U.S. Gov't]
88. Franke B, Vermeulen SH, Steegers-Theunissen RP, Coenen MJ, Schijvenaars MM, Scheffer H et al (2009) An association study of 45 folate-related genes in spina bifida: Involvement of cubilin (CUBN) and tRNA aspartic acid methyltransferase 1 (TRDMT1). Birth Defects Res A Clin Mol Teratol 85(3):216–226 [Research Support, Non-U.S. Gov't]
89. Aminoff M, Carter JE, Chadwick RB, Johnson C, Grasbeck R, Abdelaal MA et al (1999) Mutations in CUBN, encoding the intrinsic factor-vitamin B12 receptor, cubilin, cause hereditary megaloblastic anaemia 1. Nat Genet 21(3):309–313 [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.]
90. Whitehead VM (2006) Acquired and inherited disorders of cobalamin and folate in children. Br J Haematol 134(2):125–136
91. Andersson ER, Bryjova L, Biris K, Yamaguchi TP, Arenas E, Bryja V (2010) Genetic interaction between *Lrp6* and *Wnt5a* during mouse development. Dev Dyn 239:237–245. doi:10.1002/dvdy.22101
92. Gray JD, Nakouzi G, Slowinska-Castaldo B, Dazard J-E, Sunil Rao J, Nadeau JH et al (2010) Functional interactions between the LRP6 WNT co-receptor and folate supplementation. Hum Mol Genet 19(23):4560–4572
93. Lefebvre C, Rieckhof G, Califano A (2012) Reverse-engineering human regulatory networks. Wiley Interdiscip Rev Syst Biol Med 4(4):311–325 [Review]
94. Tkacik G, Walczak AM (2011) Information transmission in genetic regulatory networks: a review. J Phys Condens Matter 23(15):153102 [Review]
95. Kirouac DC, Saez-Rodriguez J, Swantek J, Burke JM, Lauffenburger DA, Sorger PK (2012) Creating and analyzing pathway and protein interaction compendia for modelling signal

transduction networks. BMC Syst Biol 6:29 [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]
96. Guan Y, Gorenshteyn D, Burmeister M, Wong AK, Schimenti JC, Handel MA et al (2012) Tissue-specific functional networks for prioritizing phenotype and disease genes. PLoS Comput Biol 8(9):e1002694 [Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S.]
97. Rzhetsky A, Wajngurt D, Park N, Zheng T (2007) Probing genetic overlap among complex human phenotypes. Proc Natl Acad Sci U S A 104(28):11694–11699
98. Oti M, Brunner HG (2007) The modular nature of genetic diseases. Clin Genet 71(1):1–11 [Research Support, Non-U.S. Gov't Review]
99. Oti M, Huynen MA, Brunner HG (2008) Phenome connections. Trends Genet 24(3):103–106
100. Vidal M, Cusick ME, Barabasi AL (2011) Interactome networks and human disease. Cell 144(6):986–998 [Research Support, N.I.H., Extramural Review]
101. Piro RM, Ala U, Molineris I, Grassi E, Bracco C, Perego GP et al (2011) An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction. Eur J Hum Genet 19(11):1173–1180 [Research Support, Non-U.S. Gov't]
102. Wysocki K, Ritter L (2011) Diseasome: an approach to understanding gene-disease interactions. Annu Rev Nurs Res 29:55–72 [Review]
103. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC et al (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4(5):P3 [Research Support, U.S. Gov't, P.H.S.]
104. Kokocinski F, Delhomme N, Wrobel G, Hummerich L, Toedt G, Lichter P (2005) FACT–a framework for the functional interpretation of high-throughput experiments. BMC Bioinforma 6:161 [Evaluation Studies Research Support, Non-U.S. Gov't]
105. Robinson MD, Grigull J, Mohammad N, Hughes TR (2002) FunSpec: a web-based cluster interpreter for yeast. BMC Bioinforma 3:35 [Research Support, Non-U.S. Gov't]
106. Castillo-Davis CI, Hartl DL (2003) GeneMerge–post-genomic analysis, data mining, and hypothesis testing. Bioinformatics 19(7):891–892
107. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M et al (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 4(4):R28 [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.]
108. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol 4(1):R7 [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.]
109. Khatri P, Draghici S, Ostermeier GC, Krawetz SA (2002) Profiling gene expression using onto-express. Genomics 79(2):266–270 [Evaluation Studies Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.]
110. Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 33(Web Server issue):W741–W748 [Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.]
111. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA et al (2007) GeneTrail–advanced gene set enrichment analysis. Nucleic Acids Res 35(Web Server issue): W186–W192 [Research Support, Non-U.S. Gov't]
112. Gupta P, Yoshida R, Imoto S, Yamaguchi R, Miyano S (2007) Statistical absolute evaluation of gene ontology terms with gene expression data. In: M*f*Éndoiu I, Zelikovsky A (eds) Bioinformatics research and applications. Springer, Berlin, pp 146–157
113. Bauer S, Grossmann S, Vingron M, Robinson PN (2008) Ontologizer 2.0 – a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 24(14): 1650–1651

114. Antonov AV, Schmidt T, Wang Y, Mewes HW (2008) ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. Nucleic Acids Res 36(Web Server issue):W347–W351 [Research Support, Non-U.S. Gov't]
115. Alexa A, Rahnenfuhrer J, Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22(13):1600–1607 [Research Support, Non-U.S. Gov't]
116. Naidoo N, Pawitan Y, Soong R, Cooper DN, Ku CS (2011) Human genetics and genomics a decade after the release of the draft sequence of the human genome. Hum Genomics 5(6):577–622, Review. PubMed PMID: 22155605; PubMed Central PMCID: PMC3525251

Chapter 4
Computational Approaches for Human Disease Gene Prediction and Ranking

Cheng Zhu, Chao Wu, Bruce J. Aronow, and Anil G. Jegga

Abstract While candidate gene association studies continue to be the most practical and frequently employed approach in disease gene investigation for complex disorders, selecting suitable genes to test is a challenge. There are several computational approaches available for selecting and prioritizing disease candidate genes. A majority of these tools are based on guilt-by-association principle where novel disease candidate genes are identified and prioritized based on either functional or topological similarity to known disease genes. In this chapter we review the prioritization criteria and the algorithms along with some use cases that demonstrate how these tools can be used for identifying and ranking human disease candidate genes.

4.1 Introduction

The majority of common diseases, common traits, and pharmacological drug response are genetically intricate, polygenic, multifactorial, and often result from an interaction of genetic, environmental, and physiological factors. Although

Cheng Zhu and Chao Wu contributed equally to this work.

C. Zhu • C. Wu
Department of Computer Science, College of Engineering and Applied Science, University of Cincinnati, Cincinnati, OH, USA

Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave. – MLC 7024, Cincinnati, OH 45229-3039, USA

B.J. Aronow • A.G. Jegga, D.V.M., M.S. (✉)
Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA

Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave. – MLC 7024, Cincinnati, OH 45229-3039, USA
e-mail: Anil.Jegga@cchmc.org

N. Maltsev et al. (eds.), *Systems Analysis of Human Multigene Disorders*, Advances in Experimental Medicine and Biology 799, DOI 10.1007/978-1-4614-8778-4_4,

high-throughput, genome-wide studies like linkage analysis and gene expression profiling are useful for classification and characterization, they often fail to provide sufficient information to identify specific disease causal genes or drug targets. Both of these approaches typically result in the identification of hundreds of potential candidate genes and cannot effectively reduce the number of target genes to a manageable figure for further validation.

4.2 Bioinformatic Tools for Gene Prioritization

Several computational approaches (Table 4.1) have been developed for gene prioritization to overcome the limitations of high-throughput, genome-wide studies like linkage analysis and gene expression profiling, both of which typically result in the identification of hundreds of potential candidate genes [1–3, 8, 10, 16, 59, 61, 62, 65, 76]. See recent reviews [7, 29, 43, 46, 50, 60, 64, 76] for technical and algorithmic details of various gene prioritization tools. While a majority of these tools are based on the assumption that similar phenotypes are caused by genes with similar or related functions [9, 20, 27, 55, 65], they differ by the strategy adopted in calculating similarity and by the data sources utilized [63]. Further, no single source of data can be expected to capture all relevant relations. For example, using coexpression data alone will fail to detect many effects of posttranscriptional modifications, while relying on protein–protein interaction data alone will fail to capture transcriptional regulation. Since these different data types are complementary, they need to be merged not only to improve coverage but to infer stronger relationships through the accumulation of evidence [43]. While this is true, except for Endeavour [3, 63] and ToppGene [9, 10], most of the existing approaches mainly focus on the combination of only a few data sources.

4.2.1 Functional Annotation-Based Approaches

The functional annotation-based candidate disease gene prioritization approaches are usually based on the guilt-by-association principle which asserts that reliable predictions about the disease involvement (“guilt”) of a gene can generally be made if several of its partners (e.g., genes with correlated expression profiles or protein interaction partners or genes involved in same biological process or pathway) share a corresponding “guilty” status (“association”) [43]. Incorporating the prior information or knowledge about a disease is thus critical for this type of approach. One of the fundamental challenges for these approaches is the ability to gather, normalize, and integrate heterogeneous data from multiple sources and keeping them current. There are now several online tools available which make carrying out such analyses intuitively without the need for having programming knowledge or direct support of a bioinformatics expert (see [29, 46, 64] for a list of such Web-based

Table 4.1 List of current bioinformatics approaches and tools to rank human disease candidate genes

Approach	Online availability	Data types used	Training set (Input)
Approaches based on disease gene properties			
DGP [39]	http://cgg.ebi.ac.uk/services/dgp/	Sequence	Not applicable (N/A)
Prospectr [1]	http://www.genetics.med.ed.ac.uk/prospectr/	Sequence	N/A
Approaches using links between genes and phenotypes			
Genes2Diseases [48, 49]	http://www.ogic.ca/projects/g2d_2/	Sequence, gene ontology (GO), literature mining	Phenotype GO terms Known genes
BITOLA [23]	http://www.mf.uni-lj.si/bitola/	Literature mining	Concept
GeneSeeker [68, 69]	http://www.cmbi.ru.nl/GeneSeeker/	Expression, phenotype, literature mining	N/A
GFINDer [41, 42]	http://www.bioinformatics.polimi.it/GFINDer/	Expression, phenotype	N/A
TOM [52]	http://www-micrel.deis.unibo.it/~tom/	Expression, GO	Known genes and/or disease loci
Approaches using functional relatedness between candidate genes			
OMIM phenome map [67]	http://www.cmbi.ru.nl/MimMiner/	Phenotype, sequence, GO, protein interactions	N/A
Suspects [2]	http://www.genetics.med.ed.ac.uk/suspects/	Sequence, expression, GO	Known genes
Prioritizer [15]	http://www.prioritizer.nl/	Expression, GO, protein interactions	Disease loci
Endeavour [3]	http://www.esat.kuleuven.be/endeavour/	Sequence, expression, GO, pathways, literature mining	Known genes
ToppGene [9]	http://toppgene.cchmc.org	Mouse phenotype, expression, GO, pathways, literature mining	Known genes
ToppNet [8]	http://toppgene.cchmc.org	Protein interactions	Known genes

The first column has the source or the name of the tool. The second column shows the URL of the corresponding web application. The third column shows the list of genomic annotation types/features used by each of the methods for candidate gene ranking. The last column has details of the training or the input data, if used (*Note: modified from Kaimal et al.* [29]*, this list is extensive, but not exhaustive; references* [43, 50] *provide an additional list of tools*)

tools). While the usage of multiple heterogeneous data in the ranking makes the functional annotation-based approaches more thorough and less biased global assessment of candidate genes, they still suffer with a bias towards the training set and have some limitations. For instance, by using a training set, it is assumed that the disease genes yet to be discovered will be consistent with what is already known about a disease and/or its genetic basis. This assumption may not always be true. Additionally, since these approaches rely on known gene annotation, they tend to be biased towards selecting better annotated genes. For example, a "true" candidate gene can be missed if it lacks sufficient annotations. Thus, the effectiveness of this approach depends critically on how well the disease under investigation is defined both molecularly and physiologically. Second, it is important to note that the annotations and analyses provided, and the prioritization by these approaches, can only be as accurate as the underlying original sources from which the annotations are retrieved. For instance, only one fifth of the known human genes have pathway or phenotype annotations, and there are still more than 30 % genes whose functions are not well-defined. Third, using an appropriate or "true representative" training set is critical. For instance, in an earlier study, we observed that using larger training sets (>100 genes) decreases the sensitivity and specificity of the prioritization compared to smaller training sets (7–21 genes) [10]. Lastly, almost all of the current disease gene identification and prioritization approaches are coding-gene-centric, while it has been speculated that complex traits result more often from noncoding regulatory variants than from coding sequence variants [32, 35, 40].

4.2.2 Network-Based Approaches

A majority of the current computational disease candidate gene prioritization methods [1–3, 10, 16, 59, 61, 62, 65, 76] rely on functional annotations, gene expression data, or sequence-based features. The coverage of the gene functional annotations, however, is still a limiting factor. Currently, only a fraction of the genome is annotated with pathways and phenotypes [10]. While two thirds of all the genes are annotated by at least one functional annotation, the remaining one third has yet to be annotated. Interestingly, because biological networks have been found to be comparable to communication and social networks [28] through commonalities such as scale-freeness and small-world properties, the algorithms used for social and Web networks should be equally applicable to biological networks.

Recent biotechnological advances (e.g., high-throughput yeast two-hybrid screening) have facilitated generation of proteome-wide protein–protein interaction networks (PPINs) or "protein interactome" maps in model organisms and humans [53, 56]. Additionally, the shift in focus to systems biology in the post-genomic era has generated further interest in these networks and pathways. As a result, PPINs have been increasingly used not only to identify novel disease candidate genes [17, 30, 34, 73, 74] but also for candidate gene prioritization [8, 11, 34, 45, 73]. At the same time, network topology-based analyses hitherto used in social and Web network analyses have been successfully used in the identification and prioritization of disease candidate genes [8, 12, 19, 24, 34, 36, 54, 57, 70, 73]. Broadly, network

topology-based candidate gene ranking approaches can be grouped into two categories: parameter-based and parameter-free methods. The parameter-based methods, such as PageRank with Priors (PRP [8]), Random Walk (RW [34]), and PRIoritizatioN and Complex Elucidation (PRINCE [70]), as the name indicates require additional auxiliary parameters that need to be trained by using available data sets. The PRP, for example, needs a parameter β to control the probability of jumping back to the initial node [8]. Similarly, the PRINCE algorithm uses a parameter to describe the relative importance of prior information [70]. However, selecting optimal parameters is often a challenge, and therefore the more "user-friendly" parameter-free approaches are preferred [24]. Further, most of the parameter-based approaches take into account the global information in the entire network, and thus they typically require extensive computation. For instance, in PRP, scores of all the vertices in the network need to be updated iteratively until they converge. This process tends to be slow and inefficient especially when the network size is large. The parameter-free methods (e.g., interconnectedness or ICN [24]), on the other hand, measure closeness of each candidate gene to known disease genes by taking into account direct link and the shared neighbors between two genes and therefore are relatively less intensive computationally. However, the performance of parameter-free methods was not comparable to those of parameter-based approaches. To address this, we recently developed a novel network-based parameter-free framework for discovering and prioritizing human rare disease candidate genes [75]. Our goals were to (a) enhance prioritizing performance compared to current parameter-free methods and (b) achieve a comparable performance to the parameter-based ones. Using several test cases, we compared the performance of our method (Vertex Similarity (VS)-based approach) to two approaches, one each from parameter-based (PRP) and parameter-free methods (ICN), and also used it to rank the immediate neighbors of known rare disease genes as potential novel candidate genes.

Network-based approaches using protein–protein interaction data while useful have some practical limitations [29]. First, high-throughput protein–protein interaction sets, especially yeast two-hybrid sets, are inherently noisy and may contain several interactions with no biological relevance [18, 26, 37, 66]. Surprisingly, only 5.8 % of the human, fly, and worm yeast two-hybrid interactions have been confirmed by the HPRD (Human Protein Reference Database), a manually curated compilation of protein interactions [47]. Second, the protein interactome tends to be biased towards well-studied proteins. Third, some of the human protein interactome data is derived by extrapolating high-throughput interactions from other species. Even though previous studies have shown that PPINs are conserved across species [25], there is a possibility for species-specific protein interactions. Fourth, two interacting proteins need not lead to similar disease phenotypes when mutated—for instance, they may have redundant or different but overlapping functions, or one may be more dispensable than the other [47]. Additionally, disease proteins may lie at different points in a molecular pathway and not necessarily interact directly. Fifth, disease mutations need not always involve proteins (e.g., telomerase RNA component in congenital autosomal dominant dyskeratosis) [47]. Lastly, most of the network topology-based algorithms were originally developed to identify "important" nodes in networks. Although extended versions of these algorithms are used to prioritize nodes to selected "seeds," they could still be biased towards hubs.

4.3 ToppGene Suite: A One-Stop Portal for Candidate Gene Prioritization Based on Functional Annotations and Protein Interactions Network

In this section, we describe the ToppGene Suite (http://toppgene.cchmc.org) [8–10], a unique, one-stop online assembly of computational software tools that enables biomedical researchers to perform candidate gene prioritization based on (a) functional annotation similarity between training and test set genes (ToppGene) [10], (b) protein interactions network analysis (ToppNet) [8], and (c) identify and rank candidate genes in the training set interactome based on both functional annotations and PPIN analysis (ToppGeNet) [8]. The ToppGene knowledgebase combines 17 gene features available from the public domain. It includes both disease-dependent and disease-independent information in the nature of known disease genes, previous linkage regions, association studies, human and mouse phenotypes, known drug genes, microarray expression results, gene regulatory regions (transcription factor target genes and microRNA targets), protein domains, protein interactions, pathways, biological processes, and literature co-citations.

4.3.1 ToppGene: Functional Annotations-Based Candidate Gene Prioritization

In the first step, ToppGene generates a representative profile of the training genes using as many as 17 features and identifies over-representative terms from the training genes. Each of the test set genes is then compared to this representative profile of the training set, and a similarity score for each of the 17 features is derived and summarized by the 17 similarity scores. Different methods are used for similarity measures of categorical (e.g., GO annotations) and numeric (i.e., gene expression) annotations. For categorical terms, a fuzzy-based similarity measure (see Popescu et al. [51] for additional details) is applied, while for numeric annotation, i.e., the microarray expression values, the similarity score is calculated as the Pearson correlation of the two expression vectors of the two genes. The 17 similarity scores are combined into an overall score using statistical meta-analysis, and a *p-value* of each annotation of a test gene G is derived by random sampling of the whole genome. The *p-value* of the similarity score S_i is defined as:

$$p(S_i) = \frac{count\ of\ genes\ having\ score\ higher\ than\ \mathrm{G}\ in\ the\ random\ sample}{count\ of\ genes\ in\ the\ random\ sample\ containing\ annotation}.$$

To combine the *p-values* from multiple annotations into an overall *p-value*, Fisher's inverse chi-square method, which states that $-2\sum_{i=1}^{n} \log p_i \rightarrow \chi^2(2n)$ (assuming the p_i values come from independent tests) is used. The final similarity score of the test gene is then obtained by 1 minus the combined *p-value*. Additional

details explaining the development of this method along with the validation process and comparison with other approaches have been previously published [9, 10].

4.3.2 ToppNet: Network Analysis-Based Candidate Gene Prioritization

ToppNet gene prioritization is based on the analysis of the protein–protein interaction network. Motivated by the observation that biological networks share many properties with social and Web networks [28], ToppNet uses extended versions of three algorithms from White and Smyth [72]: PageRank with Priors (PRP), HITS with Priors, and K-step Markov. The disease candidate genes (test set) are ranked by estimating their relative importance in the PPIN to known disease-related genes (training set). The PageRank with Priors, based on White and Smyth's PageRank algorithm [72], mimics the random surfer model wherein a random Internet surfer starts from one of a set of root nodes, R, and follows one of the links randomly in each step. In this process, the surfer jumps back to the root nodes at probability β, thus restarting the whole process. Intuitively, the PRP algorithm generates a score that is proportional to the probability of reaching any node in the Web surfing process. This score indicates or measures the relative "closeness" or importance to the root nodes. The second algorithm is HITS with Priors, an extension of HITS (Hyperlink-Induced Topic Search) developed by Jon Kleinberg to rank Web pages. It determines two values for a page: "hubness," representing the value of its links to other pages, and "authority," which estimates the value of the content of the page [33]. Here, too, the surfer starts from one of the root nodes. In the odd steps he/she can either follow a random "out-link" or jump back to a root node, and in the even steps he/she can instead follow an "in-link" or jump back to a root node. As in the case of PRP, HITS with Priors also estimates the relative probability of reaching a node in the network. The third algorithm is the K-Step Markov method which mimics a surfer who starts with one of the root nodes and then follows a random link in each step before returning to the root node (after K steps) and restarts surfing. For additional details readers are referred to our original published study [8].

4.3.3 ToppGeNet: Prioritization of Disease Gene Neighborhood in the Protein Interactome

ToppGeNet allows the user to rank the interacting partners (direct or indirect) of known disease genes for their likelihood of causing a disease. Here, given a training set of known disease genes, the test set is generated by mining the protein interactome and compiling the genes interacting either directly or indirectly (based on user input) with the training set genes. The test set genes can then be ranked using either ToppGene (functional annotation-based method) or ToppNet (PPIN-based method).

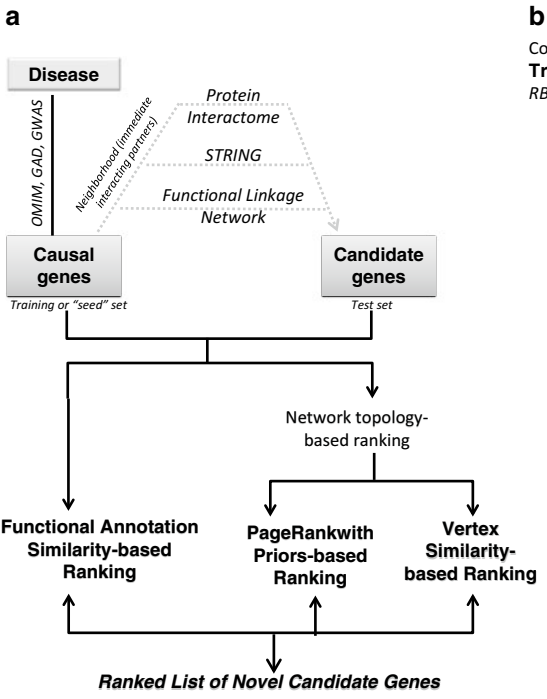
4.4 Case Studies to Demonstrate the Utility of Computational Approaches for Human Disease Gene Prediction and Ranking

In the following sections we present two sets of case studies to demonstrate the utility of computational approaches in discovering and ranking novel candidate genes for human diseases. In an earlier study, Tiffin et al. [61] used some of the computational approaches for disease gene identification and prioritization and concluded that using the methods in concert was more successful in prioritizing candidate genes for disease than when each was used alone. Hence, in the first case study, we select ten diseases and use both functional annotations-based and network-based approaches to identify and rank novel candidate genes for these diseases. We used ToppGene [9] for functional annotation-based ranking, and for network-based ranking we used both parameter [8]- and nonparameter [75]-based approaches (see next section for details). In the second case study, we present two recent examples that demonstrate the power of using bioinformatics techniques with the exome sequencing technologies in identifying novel candidate genes for rare disorders.

4.4.1 Case Study 1: Identifying and Ranking Novel Candidate Genes for Ten Human Diseases

The workflow (Fig. 4.1) described here is based on a simulation of a researcher's approach to selecting and ranking candidate disease genes. In this process, a variety of relevant database sources are mined for compiling both the training and test set genes. Known disease-associated genes for the ten selected diseases (from a recent review [43]) were obtained by combining gene lists from OMIM [21], the Genetic Association Database [4], GWAS [22], and diseases biomarkers from the Comparative Toxicogenomics Database [13] (see Table 4.2 for the list of selected ten diseases and their training sets or known causal genes). The test set or candidate genes to be ranked are compiled mining protein interactome and functional linkage networks. Briefly, for each of the training set genes (known disease causal gene), we extracted their interacting partners (both from the protein interactome and functional networks). The protein interactome data was downloaded from the NCBI (ftp://ftp.ncbi.nih.gov/gene/GeneRIF/interactions.gz), while for functional networks, we used two sources: (a) Functional Linkage Network (FLN) [38] and (b) STRING (score $\geq$ 700) [58]. Thus, for each disease, we compiled three test sets using the three databases.

The test sets were then ranked by three approaches: (a) functional annotations-based ranking (using ToppGene), (b) PageRank with Priors (parameter-dependent network topology-based approach), and (c) Vertex Similarity (parameter-free network topology-based approach). We used the harmonic mean of the individual ranks from the three approaches to obtain the final-ranked list. We repeated the



b

Congenital diaphragmatic hernia

Training Set: *GATA4, HCCS, NR2F2, PDGFRA, RBP1, RBP2, SLIT3, STRA6, WT1*

Top ten test genes from interactome and functional network-based candidate (test set) genes

Top ten test set genes

PPI

Overall Rank	Symbol	Harmonic Mean	ToppGene Rank	PRP Rank	VS Rank
1	*LRAT*	1.364	1	5	1
2	*UBC*	2.864	35	1	53
3	*LINC00152*	5.378	56	25	2
4	*ZFPM2*	5.549	2	97	33
5	*ELAVL1*	5.609	39	2	108
6	*PDGFRB*	6.013	3	44	7
7	*BRMS1L*	7.715	83	23	3
8	*CAPN1*	8.305	89	3	60
9	*NKX2-5*	9.11	5	55	9
10	*ROBO2*	9.694	40	4	29

FLN

Overall Rank	Symbol	Harmonic Mean	ToppGene Rank	PRP Rank	VS Rank
1	*ZFPM2*	1.846	1	8	2
2	*LRAT*	1.946	2	1	24
3	*NR2F6*	2.769	48	16	1
4	*SP1*	3.452	28	2	3
5	*EP300*	4.544	13	3	4
6	*NKX2-5*	6.207	5	12	5
7	*PDGFRB*	6.264	3	29	9
8	*MYOD1*	6.923	12	4	10
9	*NR3C1*	7.75	49	5	6
10	*NR5A1*	9.545	10	14	7

STRING

Overall Rank	Symbol	Harmonic Mean	ToppGene Rank	PRP Rank	VS Rank
1	*LRAT*	1.286	1	1	3
2	*PDGFRB*	1.932	2	19	1
3	*BRMS1*	4.821	90	9	2
4	*ROBO2*	5.163	45	2	17
5	*ZFPM2*	5.26	3	27	5
6	*HDAC1*	7.597	32	3	33
7	*LRP2*	7.689	4	8	66
8	*STAT3*	7.805	23	11	4
9	*NKX2-5*	8	6	24	6
10	*HDAC2*	8.101	34	4	11

Fig. 4.1 Panel (**a**) shows schematic representation of the workflow for identifying and ranking novel disease candidate genes using functional annotation- and network-based approaches. Candidate genes are compiled using both protein interactions and functional associations (Functional Linkage Network and STRING). The candidate genes are ranked using both functional annotations (ToppGene) and network topology (PageRank with Priors and Vertex Similarity-based approaches). The final ranks are generated by taking the harmonic mean of the ranks of a gene from the three methods (ToppGene, PRP, and VS). Panel (**b**) shows the top-ranked genes for congenital diaphragmatic hernia using functional annotation- and network-based approaches. Highlighted genes (*LRAT*, *ZFPM2*, *NKX2-5*, and *PDGFRB*) represent those that have been ranked among top ten by different approaches

same process for two other test sets obtained from functional networks (FLN and STRING). In the final step, we intersected the top ten genes from the three networks (PPIN, FLN, and STRING) to see the intersection. The last column in Table 4.2 shows those genes that are ranked among the top ten in the three networks. For example, in congenital diaphragmatic hernia (CDH), four genes (*LRAT, ZFPM2, NKX2-5*, and *PDGFRB*) were ranked among top ten in all the three networks. Interestingly, the retinol status in newborns is associated with CDH, and genetic analyses in humans suggest a role for retinoid-related genes in the pathogenesis of CDH [6]. *LRAT* (lecithin retinol acyltransferase) ranked among the top mediates cellular uptake of retinol and plays an important regulatory role in cellular vitamin A homeostasis [31]. Similarly, Wat et al. [71] identified three unrelated patients

Table 4.2 Top-ranked novel candidate genes for ten select diseases

Disease name	Known disease-causing genes (training set)	Top-ranked novel candidate genes (using different approaches and data sets)
Congenital diaphragmatic hernia	*GATA4, HCCS, NR2F2, PDGFRA, RBP1, RBP2, SLIT3, STRA6, WT1*	*LRAT, NKX2-5, PDGFRB, ZFPM2*
Bipolar disorder	*ABCA13, BCR, BDNF, BRCA2, COMT, CUX2, DRD4, HTR4, PALB2, SLC6A3, SLC6A4, TRPM2, XBP1*	*ADRB2, BRCA1, DRD2, NTRK2*
Nasopharyngeal carcinoma	*CCND1, CDH13, COX7B2, CTLA-4, CYP2A6, CYP2E1, CYP2F1, ERCC1, FAS, GABBR1, GSTM1, HHATL, HLA-A, HLA-B, HLA-C, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRB1, HLA-E, HLA-F, HP, HSPA1B, IFNA17, IL10, IL12A, IL16, IL18, IL1B, IL8, ITGA9, LOC344967, MDM2, MECOM, MICA, MMP1, MMP2, N4BP2, NAT2, NFKB1, OGG1, PLUNC, PTGS2, RASSF1A, TAP1, TGFB1, TLR10, TLR3, TLR4, TNF, TNFRSF19, TP53, UBAP1, VEGFA, XPC, XRCC1*	*HLA-G, HLA-DPA1, HLA-DRA*
Testicular germ cell tumor	*ATF7IP, BAK1, DMRT1, FGFR3, KIT, KITLG, LTA, SPRY4, STK11, TGFB1, TNF*	*LTB, IFNG*
Crohn's disease	*ATG16L1, C11orf30, CCR6, CDKAL1, FUT2, ICOSLG, IL12B, IL23R, IRGM, ITLN1, JAK2, LRRK2, MST1, MUC19, NKX2-3, NOD2, ORMDL3, PTGER4, PTPN2, PTPN22, STAT3, TNFSF15, ZNF365*	*IL12RB1, IL23A, JAK1, STAT1, STAT5B*
Asthma	*ACE, ADAM33, ADRB2, CC16, CCL11, CCL5, CD14, CMA1, CSF1R, CTLA4, FLG, GPRA, GSTM1, GSTP1, GSTT1, HAVCR1, HLA-DPB1, HLA-DQB1, HLA-DRB1, IL10, IL13, IL18, IL4, IL4R, LTA, LTC4S, NAT2, NOS1, SPINK5, STAT6, TBXA2R, TGFB1, TNF*	*IL1B, HLA-DRA*
Metopic craniosynostosis	*FGFR1, FGFR2, FGFR3, GLI3, TWIST1*	*FGF9, FGF2*
Nonsyndromic cleft lip/palate	*BMP4, IRF6, MSX1, MTR, PVRL1, STOM, SUMO1, TP63*	*MSX2, PAX3*
Arthrogryposis	*MYH3, TNNI2, TNNT3, TPM2, UTRN*	*ACTA1, DMD, TNNC1, TNNC2, TNNT1, TPM1*
Bipolar schizoaffective disorder	*ABCA13, BCR, BDNF, COMT, CUX2, DRD4, GABRR1, HTR4, PALB2, SLC6A3, SLC6A4, TRPM2, XBP1*	*ADRB2, DRD2, ITPR3, SLC6A9*

with CDH who had a heterozygous deletion of chromosome 8q involving *ZFPM2*, which was ranked among the top five in the three networks. It is beyond the scope of this chapter to discuss about the top-ranked genes for all the ten diseases. The supplementary file (Supplementary File 1) shows the complete lists of training and ranked test set genes for the ten select diseases along with the details of rankings from each of the three approaches using three different networks (PPIN, FLN, and STRING).

4.4.2 *Case Study 2: Exome Sequencing and Bioinformatics Applications to Identify Novel Rare Disease Causal Variants*

In the following sections we present two examples from recently published studies [5, 14] where computational approaches for candidate gene ranking were used in concert with exome sequencing to identify novel disease causal variants.

The first example [14] illustrates the potential of combining genomic variant and gene level information to identify and rank novel causal variants of rare diseases. Combining computational gene prediction tools with traditional mapping approaches, Erlich et al. [14] demonstrated how rare disease candidate genes from exome resequencing experiment can be successfully prioritized. In this study, a familial case of hereditary spastic paraparesis (HSP) was analyzed through whole-exome sequencing, and the four largest homozygous regions (containing 44 genes) were identified as potential HSP loci. The authors then applied several filters to narrow down the list further. For instance, a gene was considered as potentially causative if it contains at least one variant that is either under purifying selection or not inherited from the parents or absent in dbSNP or the 1,000 Genomes Project data. Because majority of the known rare disease variants affect coding sequences, the authors also checked if the variant is non-synonymous. After this filtering step, 15 candidate genes were identified and this list was further prioritized using three computational methods (Endeavour [3], ToppGene [9], and Suspects [2]). As a training set, a list of 11 seed genes associated with a pure type of HSP was compiled through literature mining. Interestingly, the top-ranking gene from all the three bioinformatics approaches (each of which uses different types of data and algorithms for prioritization) was *KIF1A*. Subsequent confirmation of *KIF1A* as the causative variant was done using Sanger sequencing.

In the second example, Benitez et al. [5] used disease-network analysis approach as supporting in silico evidence of the role of the adult neuronal ceroid lipofuscinosis (NCL) candidate genes identified by exome sequencing. In this case, the authors used Endeavour [3] and ToppGene [9] to rank the NCL candidate variant genes identified by exome sequencing. Known causal genes of other NCLs along with genes that are associated with phenotypically close disorders were used as training set. Interestingly, the three variants identified by exome sequencing (*PDCD6IP, DNAJC5,* and *LIPJ*) were among the top five genes in the combined analysis using ToppGene and Endeavour, suggesting that they may be functionally or structurally related with NCL encoded genes and constituting true causative variants for adult NCL.

4.5 Final Remarks

The selection of "best" computational approach for identifying and ranking disease candidate genes is not an easy task and depends on several various factors. Since a majority of these approaches are based on guilt-by-association principle, having a

"good" or representative training set is critical. The training set may not necessarily be always a set of known causal genes but can be an implicated pathway or biological process or even a list of symptoms (or phenotype). Additionally, prior knowledge can sometimes be also inferred from related or similar diseases. This similarity can be either similar manifestation or symptoms or similar molecular mechanisms of related or similar diseases. Second, selecting an appropriate approach is also important and frequently depends on the disease type and the molecular mechanism that causes it. For example, using protein–protein interaction data for identifying novel candidates may be useful when a disease is known to be caused by the disruption of a larger protein complex. On the other hand, using a protein interaction network may not be totally justified for a disease known to be caused by aberrant regulatory mechanisms. In such cases, either using gene regulatory networks and/or high-throughput gene expression data may be more apt [50]. Third, since several previous studies have shown that the computational approaches for disease gene ranking are largely complementary [5, 14, 44, 61], we recommend using a combination of at least two different approaches (e.g., functional annotation-based and network topology-based approaches).

References

1. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS (2005) Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics 6:55
2. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics 22(6):773–774
3. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y (2006) Gene prioritization through genomic data fusion. Nat Biotechnol 24(5):537–544
4. Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. Nat Genet 36(5):431–432. doi:10.1038/ng0504-431, ng0504-431 [pii]
5. Benitez BA, Alvarado D, Cai Y, Mayo K, Chakraverty S, Norton J, Morris JC, Sands MS, Goate A, Cruchaga C (2011) Exome-sequencing confirms DNAJC5 mutations as cause of adult neuronal ceroid-lipofuscinosis. PLoS One 6(11):e26741. doi:10.1371/journal.pone.0026741, PONE-D-11-16499 [pii]
6. Beurskens LW, Tibboel D, Lindemans J, Duvekot JJ, Cohen-Overbeek TE, Veenma DC, de Klein A, Greer JJ, Steegers-Theunissen RP (2010) Retinol status of newborn infants is associated with congenital diaphragmatic hernia. Pediatrics 126(4):712–720. doi:10.1542/peds.2010-0521, peds.2010-0521 [pii]
7. Bornigen D, Tranchevent LC, Bonachela-Capdevila F, Devriendt K, De Moor B, De Causmaecker P, Moreau Y (2012) An unbiased evaluation of gene prioritization tools. Bioinformatics 28(23):3081–3088. doi:10.1093/bioinformatics/bts581, bts581 [pii]
8. Chen J, Aronow BJ, Jegga AG (2009) Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics 10:73. doi:1471-2105-10-73, [pii] 10.1186/1471-2105-10-73
9. Chen J, Bardes EE, Aronow BJ, Jegga AG (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res 37(Web Server issue):W305–W311. doi:gkp427, [pii] 10.1093/nar/gkp427
10. Chen J, Xu H, Aronow BJ, Jegga AG (2007) Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinformatics 8(1):392

11. Chen JY, Shen C, Sivachenko AY (2006) Mining Alzheimer disease relevant proteins from integrated protein interactome data. Pac Symp Biocomput 367–378
12. Chen X, Yan GY, Liao XP (2010) A novel candidate disease genes prioritization method based on module partition and rank fusion. OMICS 14(4):337–356. doi:10.1089/omi.2009.0143
13. Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, Mattingly CJ (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic Acids Res 37(Database issue):D786–D792. doi:gkn580, [pii] 10.1093/nar/gkn580
14. Erlich Y, Edvardson S, Hodges E, Zenvirt S, Thekkat P, Shaag A, Dor T, Hannon GJ, Elpeleg O (2011) Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. Genome Res 21(5):658–664. doi:gr.117143.110, [pii] 10.1101/gr.117143.110
15. Franke L, Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78(6):1011–1025
16. Freudenberg J, Propping P (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics 18(Suppl 2):S110–S115
17. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. Nucleic Acids Res 34(19):e130
18. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL Jr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM (2003) A protein interaction map of Drosophila melanogaster. Science (New York, NY) 302(5651):1727–1736. doi:10.1126/science.1090289, 1090289 [pii]
19. Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, Droege A, Lindenberg KS, Knoblich M, Haenig C, Herbst M, Suopanki J, Scherzinger E, Abraham C, Bauer B, Hasenbank R, Fritzsche A, Ludewig AH, Bussow K, Coleman SH, Gutekunst CA, Landwehrmeyer BG, Lehrach H, Wanker EE (2004) A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. Mol Cell 15(6):853–865. doi:10.1016/j.molcel.2004.09.016, S1097276504005453 [pii]
20. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL (2007) The human disease network. Proc Natl Acad Sci U S A 104(21):8685–8690. doi:0701361104, [pii] 10.1073/pnas.0701361104
21. Hamosh A, Scott A, Amberger J, Bocchini C, McKusick V (2005) Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33:D514–D517
22. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106(23):9362–9367. doi:0903103106, [pii]10.1073/pnas.0903103106
23. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM (2005) Using literature-based discovery to identify disease candidate genes. Int J Med Inform 74(2–4):289–298
24. Hsu C, Huang Y, Hsu C, Yang U (2011) Prioritizing disease candidate genes by a gene interconnectedness-based approach. BMC Genomics 12(3):S25
25. Huynen MA, Snel B, van Noort V (2004) Comparative genomics for reliable protein-function prediction from genomic data. Trends Genet 20(8):340–344
26. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98(8):4569–4574. doi:10.1073/pnas.061034498, 061034498 [pii]
27. Jimenez-Sanchez G, Childs B, Valle D (2001) Human disease genes. Nature 409(6822):853–855

28. Junker BH, Koschutzki D, Schreiber F (2006) Exploration of biological network centralities with CentiBiN. BMC Bioinformatics 7:219
29. Kaimal V, Sardana D, Bardes EE, Gudivada RC, Chen J, Jegga AG (2011) Integrative systems biology approaches to identify and prioritize disease and drug candidate genes. Methods Mol Biol 700:241–259. doi:10.1007/978-1-61737-954-3_16
30. Kann MG (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief Bioinform 8(5):333–346
31. Kim YK, Wassef L, Hamberger L, Piantedosi R, Palczewski K, Blaner WS, Quadro L (2008) Retinyl ester formation by lecithin: retinol acyltransferase is a key regulator of retinoid homeostasis in mouse embryogenesis. J Biol Chem 283(9):5611–5621. doi:M708885200, [pii] 10.1074/jbc.M708885200
32. King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science (New York, NY) 188(4184):107–116
33. Kleinberg J (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632
34. Kohler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82(4):949–958. doi:S0002-9297(08)00172-9, [pii] 10.1016/j.ajhg.2008.02.013
35. Korstanje R, Paigen B (2002) From QTL to gene: the harvest begins. Nat Genet 31(3):235–236
36. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N et al (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25(3):309–316
37. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M (2004) A map of the interactome network of the metazoan C. elegans. Science (New York, NY) 303(5657):540–543. doi:10.1126/science.1091403, 1091403 [pii]
38. Linghu B, Snitkin ES, Hu Z, Xia Y, Delisi C (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol 10(9):R91. doi:10.1186/gb-2009-10-9-r91, gb-2009-10-9-r91 [pii]
39. Lopez-Bigas N, Ouzounis CA (2004) Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res 32(10):3108–3114
40. Mackay TF (2001) Quantitative trait loci in Drosophila. Nat Rev 2(1):11–20
41. Masseroli M, Galati O, Pinciroli F (2005) GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. Nucleic Acids Res 33(Web Server issue):W717–W723
42. Masseroli M, Martucci D, Pinciroli F (2004) GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. Nucleic Acids Res 32(Web Server issue):W293–W300
43. Moreau Y, Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev 13(8):523–536. doi:10.1038/nrg3253, nrg3253 [pii]
44. Navlakha S, Kingsford C (2010) The power of protein interaction networks for associating genes with diseases. Bioinformatics 26(8):1057–1063. doi:10.1093/bioinformatics/btq076, btq076 [pii]
45. Ortutay C, Vihinen M (2009) Identification of candidate disease genes by integrating gene ontologies and protein-interaction networks: case study of primary immunodeficiencies. Nucleic Acids Res 37(2):622–628. doi:gkn982, [pii]10.1093/nar/gkn982
46. Oti M, Ballouz S, Wouters MA (2011) Web tools for the prioritization of candidate disease genes. Methods Mol Biol 760:189–206. doi:10.1007/978-1-61779-176-5_12
47. Oti M, Snel B, Huynen MA, Brunner HG (2006) Predicting disease genes using protein-protein interactions. J Med Genet 43(8):691–698

48. Perez-Iratxeta C, Bork P, Andrade MA (2002) Association of genes to genetically inherited diseases using data mining. Nat Genet 31(3):316–319
49. Perez-Iratxeta C, Wjst M, Bork P, Andrade MA (2005) G2D: a tool for mining genes associated with disease. BMC Genet 6:45
50. Piro RM, Di Cunto F (2012) Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS J 279(5):678–696. doi:10.1111/j.1742-4658.2012.08471.x
51. Popescu M, Keller JM, Mitchell JA (2006) Fuzzy measures on the gene ontology for gene product similarity. IEEE/ACM Trans Comput Biol Bioinform 3(3):263–274
52. Rossi S, Masotti D, Nardini C, Bonora E, Romeo G, Macii E, Benini L, Volinia S (2006) TOM: a web-based integrated approach for identification of candidate disease genes. Nucleic Acids Res 34(Web Server issue):W285–W292
53. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437(7062):1173–1178. doi:nature04209, 10.1038/nature04209
54. Sam L, Liu Y, Li J, Friedman C, Lussier YA (2007) Discovery of protein interaction networks shared by diseases. Pac Symp Biocomput 76–87
55. Smith NG, Eyre-Walker A (2003) Human disease genes: patterns and predictions. Gene 318:169–175
56. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122(6):957–968. doi:S0092-8674(05)00866-4, 10.1016/j.cell.2005.08.029
57. Sun PG, Gao L, Han S (2010) Prediction of human disease-related gene clusters by clustering analysis. Int J Biol Sci 7(1):61–73
58. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39(Database issue):D561–D568. doi:10.1093/nar/gkq973, gkq973 [pii]
59. Thornblad TA, Elliott KS, Jowett J, Visscher PM (2007) Prioritization of positional candidate genes using multiple web-based software tools. Twin Res Hum Genet 10(6):861–870
60. Tiffin N (2011) Conceptual thinking for in silico prioritization of candidate disease genes. Methods Mol Biol 760:175–187. doi:10.1007/978-1-61779-176-5_11
61. Tiffin N, Adie E, Turner F, Brunner HG, van Driel MA, Oti M, Lopez-Bigas N, Ouzounis C, Perez-Iratxeta C, Andrade-Navarro MA, Adeyemo A, Patti ME, Semple CA, Hide W (2006) Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. Nucleic Acids Res 34(10):3067–3081
62. Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res 33(5):1544–1552
63. Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B, Aerts S, Moreau Y (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res 36(Web Server issue):W377–W384
64. Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y (2011) A guide to web tools to prioritize candidate genes. Brief Bioinform 12(1):22–32. doi:10.1093/bib/bbq007, bbq007 [pii]
65. Turner FS, Clutterbuck DR, Semple CA (2003) POCUS: mining genomic sequence annotation to predict disease genes. Genome Biol 4(11):R75
66. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T,

Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403(6770): 623–627. doi:10.1038/35001009
67. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA (2006) A text-mining analysis of the human phenome. Eur J Hum Genet 14(5):535–542
68. van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG (2003) A new web-based data mining tool for the identification of candidate genes for human genetic disorders. Eur J Hum Genet 11(1):57–63
69. van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG, Vriend G (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. Nucleic Acids Res 33(Web Server issue):W758–W761
70. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R (2010) Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol 6(1):e1000641. doi:10.1371/journal.pcbi.1000641
71. Wat MJ, Veenma D, Hogue J, Holder AM, Yu Z, Wat JJ, Hanchard N, Shchelochkov OA, Fernandes CJ, Johnson A, Lally KP, Slavotinek A, Danhaive O, Schaible T, Cheung SW, Rauen KA, Tonk VS, Tibboel D, de Klein A, Scott DA (2011) Genomic alterations that contribute to the development of isolated and non-isolated congenital diaphragmatic hernia. J Med Genet 48(5):299–307. doi:10.1136/jmg.2011.089680, 48/5/299 [pii]
72. White S, Smyth P (2003) Algorithms for estimating relative importance in networks. Paper presented at the KDD '03: proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining
73. Wu X, Jiang R, Zhang MQ, Li S (2008) Network-based global inference of human disease genes. Mol Syst Biol 4:189. doi:msb200827, [pii] 10.1038/msb.2008.27
74. Xu J, Li Y (2006) Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics 22(22):2800–2805. doi:btl467, [pii] 10.1093/bioinformatics/btl467
75. Zhu C, Kushwaha A, Berman K, Jegga AG (2012) A vertex similarity-based framework to discover and rank orphan disease-related genes. BMC Syst Biol 6(Suppl 3):S8. doi:10.1186/1752-0509-6-S3-S8, 1752-0509-6-S3-S8 [pii]
76. Zhu M, Zhao S (2007) Candidate gene identification approach: progress and challenges. Int J Biol Sci 3(7):420–427

Chapter 5
A Personalized Treatment for Lung Cancer: Molecular Pathways, Targeted Therapies, and Genomic Characterization

Thomas Hensing, Apoorva Chawla, Rishi Batra, and Ravi Salgia

Abstract Lung cancer is a heterogeneous, complex, and challenging disease to treat. With the arrival of genotyping and genomic profiling, our simple binary division of lung cancer into non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC) is no longer acceptable. In the past decade and with the advent of personalized medicine, multiple advances have been made in understanding the underlying biology and molecular mechanisms of lung cancer. Lung cancer is no longer considered a single disease entity and is now being subdivided into molecular subtypes with dedicated targeted and chemotherapeutic strategies. The concept of using information from a patient's tumor to make therapeutic and treatment decisions has revolutionized the landscape for cancer care and research in general.

Management of non-small-cell lung cancer, in particular, has seen several of these advances, with the understanding of activating mutations in EGFR, fusion genes involving ALK, rearrangements in ROS-1, and ongoing research in targeted therapies for K-RAS and MET. The next era of personalized treatment for lung cancer will involve a comprehensive genomic characterization of adenocarcinoma, squamous-cell carcinoma, and small-cell carcinoma into various subtypes.

T. Hensing, M.D. (✉)
Department of Medicine, Section of Hematology/Oncology, NorthShore University Health System, Kellogg Cancer Center, 2650 Ridge Avenue, Evanston, IL 60201, USA

Department of Medicine, Section of Hematology/Oncology, University of Chicago, 5841 S. Maryland Avenue, MC 2115, Chicago, IL, USA
e-mail: THensing@northshore.org

A. Chawla, M.D. • R. Salgia, M.D., Ph.D.
Department of Medicine, Section of Hematology/Oncology, University of Chicago, 5841 S. Maryland Avenue, MC 2115, Chicago, IL, USA
e-mail: achawla@medicine.bsd.uchicago.edu; rsalgia@medicine.bsd.uchicago.edu

R. Batra, B.S.
987400 Nebraska Medical Center, University of Nebraska Medical Center, Omaha, NE 68198, USA
e-mail: rishi.batra@unmc.edu

N. Maltsev et al. (eds.), *Systems Analysis of Human Multigene Disorders*, Advances in Experimental Medicine and Biology 799, DOI 10.1007/978-1-4614-8778-4_5,

Future directions will involve incorporation of molecular characteristics and next generation sequencing into screening strategies to improve early detection, while also having applications for joint treatment decision making in the clinics with patients and practitioners. Personalization of therapy will involve close collaboration between the laboratory and the clinic. Given the heterogeneity and complexity of lung cancer treatment with respect to histology, tumor stage, and genomic characterization, mind mapping has been developed as one of many tools which can assist physicians in this era of personalized medicine. We attempt to utilize the above tool throughout this chapter, while reviewing lung cancer epidemiology, lung cancer treatment, and the genomic characterization of lung cancer.

Abbreviations

ALK	Anaplastic lymphoma kinase
ASCO	American Society of Clinical Oncology
CI	Confidence interval
c-MET	*N-methyl-N'-nitro-N-nitroso-guanidine* (MNNG) HOS transforming gene
CT	Computed tomography
EGF	Epidermal growth factor
EGFR	Epidermal growth factor receptor
EML4	Echinoderm microtubule-associated protein-like 4
EMT	Epithelial to mesenchymal transition
ERCC1	Excision repair cross-complementation group 1
EZH2	Enhancer of zeste homolog 2
FDA	Food and Drug Administration
FISH	Fluorescence in situ hybridization
GDP	Guanosine diphosphate
GTP	Guanosine triphosphate
HDAC	Histone deacetylase
HGF/SF	Hepatocyte growth factor/scatter factor
HGFR	Hepatocyte growth factor receptor
HR	Hazard ratio
HSP-90	Heat shock protein-90
IGFR1	Insulin-like growth factor receptor 1
IHC	Immunohistochemistry
IPASS	Iressa Pan-Asia Study
LDCT	Low-dose computed tomography
MEK	Mitogen-activated protein kinase kinase
MTD	Maximum tolerated dose
mTOR	Mammalian target of rapamycin
NCCN	National Comprehensive Cancer Network
NGS	Next generation sequencing
NSCLC	Non-small-cell lung cancer
OR	Odds ratio

OS	Overall survival
PCR	Polymerase chain reaction
PET	Positron emission tomography
PFS	Progression-free survival
PI3K	Phosphatidylinositol 3-kinase
ROS-1	Reactive oxygen species-1
RT	Radiation therapy
RTK	Receptor tyrosine kinase
RTOG	Radiation Therapy Oncology Group
SCLC	Small-cell lung cancer
Siah 2	Seven in absentia homolog 2
TKI	Tyrosine kinase inhibitor
TP63	Tumor protein 63
TS	Thymidylate synthase
TTF-1	Thyroid transcription factor
VATS	Video-assisted thoracoscopic surgery
VEGF	Vascular endothelial growth factor

5.1 Background

Lung cancer (both small cell and non-small cell) is the second most common cancer in both men and women (not counting skin cancer). In men, prostate cancer is more common, while in women breast cancer is more common. Lung cancer accounts for about 14 % of all new cancers [1]. Recent data suggests that lung cancer is likely to overtake breast cancer as the main cause of cancer death among European women by the middle of this decade [2].

The American Cancer Society provides estimates for new cancer cases and deaths in the United States. In 2013, an estimated 228,190 new cases of lung cancer will be diagnosed (118,080 in men and 110,110 in women) [3]. There will be an estimated 159,480 deaths from lung cancer (87,260 in men and 72,220 among women), accounting for about 27 % of all cancer deaths [3]. Lung cancer is by far the leading cause of cancer death among both men and women. Each year, more people die of lung cancer than of colon, breast, and prostate cancers combined [4].

Genetic and environmental factors, as well as their interaction, influence the risk of developing lung cancer. Cigarette smoking is well established as one of the most important risk factors for the development of lung cancer, with smokers being 10–20 times more likely to develop lung cancer than nonsmokers [5]. Recent reviews suggest that the hazard ratios for lung cancer mortality are staggering: 17.8 for female smokers and 14.6 for male smokers [6]. The process of tobacco-induced lung carcinogenesis takes place over decades, and hence, the majority of the burden of smoking-related lung cancer occurs in the elderly. Individual risk is affected by factors including patient age, duration of smoking, intensity of smoking, age of initiation, and age of cessation (if present) [7]. Cessation, and reduction of smoking, has been shown to significantly cut the risk of developing lung cancer [8].

The International Agency for Research on Cancer has identified several substances implicated as carcinogens in the development of lung cancer. While asbestos and radon gas are perhaps the best-known occupational lung carcinogens, other agents including beryllium, bis-chloromethyl ether, cadmium, chromium, polycyclic aromatic hydrocarbons, nickel, mustard gas, silica, inorganic arsenic, vinyl chloride, and particulate air matter are prevalent in various geographic areas [9].

Asbestos has been linked to the development of lung cancer and mesothelioma. Historically, asbestos was used in tiles and paints, to thin cement and plastics, as well as for insulation, roofing, fireproofing, and sound absorption. Exposure and inhalation of small asbestos fibers released into the air when such asbestos-containing products are disturbed led to a public health movement whereby several prohibitions and restrictions were placed on asbestos use in the United States since the 1970s [10]. Given the latency period for asbestos-related lung cancer, the impact of this factor continues to this day.

Demographic factors, including race, gender, and socioeconomic factors, have been investigated with regard to effect on susceptibility to development of lung cancer. Though the incidence of lung cancer remains lower among women than men, recent data suggests that the two genders are approaching parity. However, black men have had a significantly higher incidence of lung cancer, when compared to white men. Black women and white women, on the other hand, have had similar incidence rates. Similarly, between 2003 and 2007, mortality rates were higher among black men than white men (87.5 and 68.3 deaths per 100,000 people, respectively) while comparable between black and white women (39.6 and 41.6 deaths per 100,000 people, respectively) [4]. Socioeconomic status is inversely associated with lung cancer risk and may be related to higher smoking prevalence, lower quit rates among those of lower socioeconomic status, less healthy diet, occupational and environmental carcinogens, and exposure to secondhand smoke.

Lung cancer only occurs in a minority of smokers, suggesting that inherited factors and genetic predisposition may play a role. It has been shown that patients with a previous personal or first-degree relative family history of lung cancer have a higher risk of developing the disease [11], though the molecular basis of such familial risk remains to be characterized. Recent studies have revealed distinct gene-environment interactions in smokers and nonsmokers with lung cancer, confirming the importance of multifactorial interaction in risk assessment of lung cancer [12].

As most patients diagnosed with lung cancer already have advanced disease (40 % are stage IV and 30 % are stage III), screening for lung cancer has been attempted in an effort to detect more early-stage cancers and thereby improve overall survival (OS). Earlier randomized controlled trials involving chest radiographs and sputum cytology for lung cancer screening found that these strategies detected slightly more lung cancers, smaller tumors, and more stage I tumors, but the detection of a larger number of early-stage cancers was not accompanied by a reduction in the number of advanced lung cancers or a reduction in lung cancer deaths [13–15].

More recently, with the advent of low-dose computed tomography (LDCT), renewed enthusiasm for lung cancer screening arose. To date, several randomized controlled trials have attempted to determine the effect of LDCT screening on lung cancer mortality. The largest, the National Lung Screening Trial, demonstrated

that among 53,454 participants enrolled, screening resulted in significantly fewer lung cancer deaths (356 vs. 443 deaths; lung cancer-specific mortality, 274 vs. 309 events per 100,000 person-years for LDCT and control groups, respectively; relative risk, 0.80; 95 % CI 0.73–0.93; absolute risk reduction, 0.33 %; $p = 0.004$) [16, 17]. In terms of potential harms of LDCT screening, across all trials and cohorts, approximately 20 % of individuals in each round of screening had positive results requiring some degree of follow-up, while approximately 1 % had lung cancer [16]. There was marked heterogeneity in this finding and in the frequency of follow-up investigations, biopsies, and percentage of surgical procedures performed in patients with benign lesions. The present consensus suggests that LDCT screening likely benefits individuals at an increased risk for lung cancer, though the above must be balanced with the substantial percentage of false-positive results.

5.2 Treatment Considerations

Treatment for NSCLC is multimodal in scope and incorporates information both from patient-associated factors (outlined previously) and from the tumor, including tumor stage, tumor histology, tumor biology, and more recently, molecular biomarkers.

While several groups vary in their definition of molecular biomarkers, the National Cancer Institute, in particular, defines a biomarker as "a biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease. A biomarker may be used to see how well the body responds to a treatment for a disease or condition" [18]. Biomarkers can range from molecular biomarkers to serum/blood biomarkers (e.g., circulating tumor cells or serum miRNA) to radiographic biomarkers. In cancer research and medicine, such biomarkers are used in three primary ways: (a) to help diagnose conditions, as in identifying and confirming a cancer diagnosis (diagnostic biomarker); to forecast the aggressiveness of a disease state or condition in the absence of treatment (prognostic biomarker); and to predict how well a patient may respond to treatment (predictive biomarker). Interestingly, several molecular biomarkers are not only prognostic/predictive but are also actionable with respect to targeted chemotherapeutic agents.

5.3 Tumor Histology

Non-small-cell lung cancer comprises over 70 % of cases of lung cancer. The major subtypes of NSCLC include adenocarcinoma (approximately 45 % of cases), squamous-cell carcinoma (approximately 20 % of cases), and large-cell carcinoma (estimated at 4 % of cases). Accurate histologic determination has become essential in treatment decision making, as potential efficacy and toxicity of chemotherapeutic agents and targeted therapies can differ based on histologic subtype.

For example, in the early development of bevacizumab—a monoclonal antibody directed against the vascular endothelial growth factor (VEGF)—squamous

histology was identified as a predictor of risk for bleeding complications [19]. The subsequent randomized trial which led to the approval for bevacizumab excluded patients with squamous NSCLC. Interestingly, in the phase III Eastern Cooperative Oncology Group trial, severe pulmonary hemorrhage (defined as greater than or equal to grade III) was only seen in 2.3 % of patients with non-squamous histology, which is in contrast to 9.1 % of patients on the phase II trial that did not exclude patients with squamous tumor histology [20].

Prior studies have also shown that non-squamous histology is a predictor of improved survival after treatment with the antifolate pemetrexed. When reviewing prognostic and predictive factors in a randomized phase III trial comparing cisplatin-pemetrexed versus cisplatin-gemcitabine in advanced NSCLC, a prespecified subset analysis demonstrated that median survival was superior on the pemetrexed arm for patients with non-squamous tumors (adenocarcinoma and large-cell carcinoma vs. squamous histology, $n=1{,}000$, 11.8 vs. 10.4 months, respectively; hazard ratio (HR) 0.81; 95 % CI 0.70–0.84, $p=0.05$) [21]. Patients with squamous tumors had inferior survival on the pemetrexed arm compared to cisplatin-gemcitabine (median 9.4 months vs. 10.8 months, HR 1.23, 95 % CI 1.0–1.51, $p=0.05$) [21].

In a subsequent prospective study, maintenance pemetrexed was compared to placebo for patients with advanced NSCLC with stable or responding disease after four cycles of platinum doublet therapy. The study confirmed a significant improvement in overall survival with pemetrexed in patients with non-squamous tumors (median survival 15.5 vs. 10.3 months, pemetrexed vs. placebo, respectively; HR 0.70; 95 % CI 0.56–0.88; $p=0.002$), but there was no benefit of maintenance pemetrexed in patients with squamous tumors [22].

Biomarkers have become incorporated with routine histologic analysis to improve diagnostic accuracy. Antibody-based immunohistochemistry (IHC) staining for thyroid transcription factor (TTF-1) and tumor protein 63 (TP63), as well as other markers, can provide additional confirmation and diagnosis of NSCLC subtype. Clinical trials initially suggested that low thymidylate synthase-mRNA expression was associated with improved time to progression and time to treatment failure on pemetrexed arms, though the results were not statistically significant [23].

5.4 Tumor Stage

Clinical staging of NSCLC involves a thorough yet focused history and physical exam and computed tomography (CT) scan of the chest and upper abdomen. One of the critical early distinctions is to define possible resectability for potential cure. Presurgical planning should continue with combined CT/positron emission tomography (PET) imaging. This strategy with PET-CT and cranial imaging identifies more patients with mediastinal and extra-thoracic disease than conventional staging, thereby sparing more patients from stage-inappropriate surgery [24]. Pathologic mediastinal lymphadenopathy—as defined by short axis size > 1 cm and/or those found to be metabolically active on PET scan—is oftentimes sampled. Mediastinoscopy has remained the gold standard for lymph node sampling of the mediastinum, with reported sensitivity and specificity of 87 and 100 %,

respectively [25]. Alternate methods of transesophageal endoscopic ultrasound-guided fine-needle aspiration and endobronchial ultrasound-guided transbronchial needle aspiration are less invasive means to stage patients and are also proven to be effective [26]. Bronchoscopy is utilized in the preoperative staging of early-stage central and peripheral NSCLC lesions, to better assess the size, or T stage, of the tumor. Brain MRI is recommended for advanced-stage disease to establish the absence/presence of distant metastases. Pulmonary function testing is performed to define medial fitness prior to possible surgical resection.

Staging for NSCLC is based on the tumor-node-metastasis (TNM) system. Staging may be through the use of a clinical staging system or a histopathologic staging system. Clinical stage relies on history, physical exam, laboratory and radiographic evidence, and tissue sampling. Histopathologic staging relies on information from the resected tumor (including such factors as histologic grade, tumor margins, and lymphovascular invasion). The American Joint Commission on Cancer updated their staging guidelines for NSCLC in January 2010, and the above seventh edition is most commonly employed for determination of tumor stage [27].

5.5 Management of Stage I and II NSCLC

Stage I plus stage II disease accounts for only 30 % of patients with NSCLC at diagnosis [28]. If there are no contraindications, the primary strategy for these patients involves surgical resection, which offers the highest chance of cure, taking into account age, pulmonary function, and comorbidity. Patients able to undergo resection are offered anatomic lobectomy with lymph node dissection or sampling, while patients unfit for surgery are considered for more limited resection or radiation therapy (RT). For patients with stage IIB (T3, N0) disease due to chest wall invasion, en bloc resection followed by chemotherapy has demonstrated a 5-year survival rate of approximately 40 %, regardless of adjuvant radiotherapy use [29]. In patients with early-stage NSCLC, video-assisted thoracoscopic surgery (VATS) may be an alternative to open thoracotomy for patients undergoing lobectomy [30]. Recent data has suggested decreased morbidity with VATS, which may enhance compliance with postoperative adjuvant chemotherapy [31]. Some proximal tumors may not be readily resected by lobectomy. In such cases, sleeve resection is preferred over pneumonectomy, whenever possible, based on equivalent results, better preservation of lung function, and avoidance of the complications of pneumonectomy [32].

Patients with completely resected pathologic stage IA and IB disease are entered into active clinical surveillance, and adjuvant chemotherapy is typically not recommended. Adjuvant chemotherapy is often recommended for selected patients with stage II disease; it may have a role for a subset of patients with stage IB disease though this is still debated [33]. Patients with completely resected stage IIA or IIB disease are treated with a cisplatin-based adjuvant therapy regimen [34], while those patients who are incompletely resected are offered re-resection, followed by chemotherapy. Incomplete resection which leaves behind visible tumor (R2) is without therapeutic advantage, as it causes postponement of additional therapy while adding undue morbidity to the patient.

5.6 Management of Stage III NSCLC

Stage III non-small-cell lung cancer comprises a heterogeneous group, which usually requires a combined modality approach. Historically, primary surgical resection was used to treat non-bulky stage IIIA disease. However, due to poor outcomes with primary surgery and the routine practice of pathologic staging of the mediastinum, the role for surgery has lessened. For patients with surgical resection of clinical stage I or II disease, found to have incidental mediastinal involvement during resection via endobronchial ultrasound or mediastinoscopy, recent studies have demonstrated that adjuvant chemotherapy can result in a significant survival benefit [35]. Both the National Comprehensive Cancer Network (NCCN) and the American Society of Clinical Oncology (ASCO) have updated their treatment guidelines for NSCLC to recommend cisplatin-based regimens in patients with resected stage II as well as stage III disease [36]. The role for postoperative RT in resected stage IIIA disease is still debated and should likely be dependent on the extent of nodal involvement (N1 or N2) and use of adjuvant chemotherapy, as suggested by results from the phase III ANITA trial [35, 37]. Specifically, for patients with N2 disease, survival was longer in patients who received postoperative RT than those not given RT. In contrast, for N1 disease, postoperative RT had a negative effect on survival in those given adjuvant chemotherapy, but a positive effect in those not receiving adjuvant chemotherapy [37].

Initially, for patients with unresectable stage III NSCLC, sequential chemotherapy followed by RT was used to avoid overlapping toxicities, and clinical trials established the benefits of this approach compared to RT alone. Subsequent studies comparing sequential versus concurrent chemoradiotherapy have demonstrated a survival benefit for the concurrent approach. The largest, multicenter, randomized phase III trial to demonstrate the above was the Radiation Therapy Oncology Group (RTOG) trial which demonstrated a median survival of 17.0 versus 14.6 months (hazard ratio for death 0.81, 95 % CI 0.663–0.996) [38]. The survival benefit likely relates to early treatment of micrometastatic disease as well as synergism of chemotherapy and radiation therapy to enhance local control. Concurrent chemoradiation has become the preferred approach for unresectable stage IIIA and IIIB disease. Studies have shown that full-dose chemotherapy can be given concurrently with manageable toxicity, and this has become the standard approach.

5.7 Management of Stage IV NSCLC

For patients with stage IV disease, chemotherapy alone is the primary treatment modality. In addition to tumor histology, patient performance status and underlying medical comorbidities are often the primary factors that are considered in treatment decision making. More recently, with routine genetic and molecular tumor typing, activating mutations involving epidermal growth factor receptor (EGFR), fusion

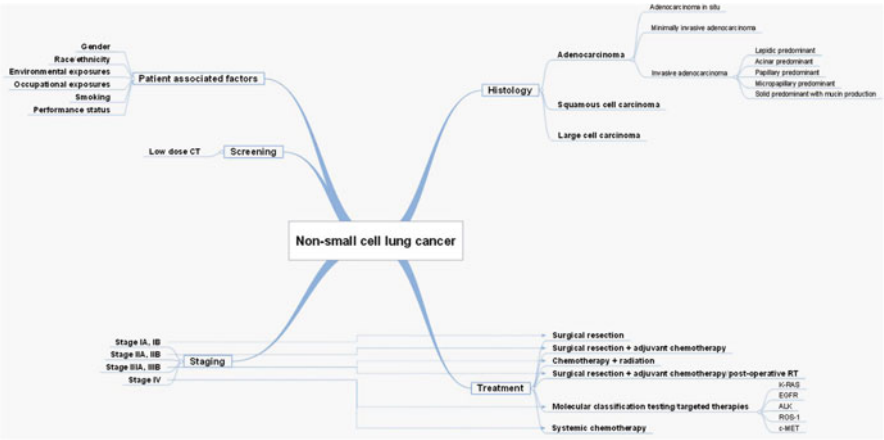


Fig. 5.1 General mind map and treatment overview for non-small-cell lung cancer**.** Treatment for NSCLC is multimodal in scope and incorporates information from patient-associated factors as well as tumor stage, tumor histology, tumor biology, and more recently, molecular biomarkers. Stage and overview of treatment considerations are shown for stage I–stage IV NSCLC. *NSCLC* non-small-cell lung cancer, *CT* computed tomography, *RT* radiation therapy

genes involving anaplastic lymphoma kinase (ALK), rearrangements in reactive oxygen species-1 (ROS-1), and other driver mutations in the *N-methyl-N'-nitro-N-nitroso-guanidine* (MNNG) HOS transforming gene (c-MET) are being routinely tested with the hope of using targeted therapies or offering clinical trial options. For example, erlotinib is approved in the United States as first-line therapy for patients with advanced or metastatic NSCLC whose tumor harbors EGFR mutation. The primary molecular pathways, along with a focus on the landscape of targeted therapies and the genomic characterization of lung cancer, form the basis of the remainder of this chapter. A general mind map and conceptual treatment overview for NSCLC is depicted in Fig. 5.1.

5.8 Molecular Pathways in NSCLC

Over the past 10 years, a significant paradigm shift occurred in understanding that lung cancer is a complex, heterogeneous, and multigene disorder with subsets containing specific genetic alternations that are critical to the growth and survival of these cancers. Molecular subsets of lung adenocarcinoma, in particular, have been well characterized with respect to clinically relevant driver mutations, and it is now estimated that over half of NSCLC tumors with adenocarcinoma histology demonstrate such an oncogenic driver mutation or fusion [39] (Table 5.1).

One of the largest series to date, the Lung Cancer Mutation Consortium has analyzed approximately 800 lung adenocarcinoma samples and identified mutations in 54 % of samples [40]. It is interesting to note that even though it has been

Table 5.1 Molecular subsets of NSCLC tumors with adenocarcinoma histology and clinically relevant mutations/translocations

Mutation/translocation	Lung cancer consortium	Sequist et al.
K-RAS mutation	25 %	24 %
EGFR mutation	23 %	13 %
ALK rearrangement	6 %	5 %
BRAF mutation	3 %	2 %
PIK3CA mutation	3 %	4 %
MET amplification	2 %	–
Her-2 mutation	1 %	< 1 %
MEK-1 mutation	0.4 %	–
N-RAS mutation	0.2 %	1 %
AKT-1 mutation	0 %	–
B-catenin mutation	–	2 %
IDH1 mutation	–	< 1 %

The second column depicts data from the Lung Cancer Mutation Consortium, which utilizes multiplexed assay for *K-RAS, EGFR, HER-2, BRAF, PIK3CA, AKT1*, *MEK1*, and *N-RAS* along with FISH for *ALK* rearrangement and *MET* amplification ($n = 830$, enrollment ongoing) [40]. The third column depicts data from 552 patients with NSCLC tested with multiplexed PCR-based assay (SNaPShot) along with FISH for *ALK* translocation [126]

hypothesized that primary tumor versus metastatic deposits may harbor different genetic alterations, in one detailed study using next generation sequencing platform, there were minimal differences identified [41]. We focus much of the remaining discussion on NSCLC—and adenocarcinoma histology—as the model for considering lung cancer as a multigene disorder.

Targeting these various molecular pathways and driver mutations in patient-selected subsets offers the possibility of improving efficacy and reducing toxicity. Selecting patients for such targeted therapies will likely improve with the development of specific gene signatures which refine prognosis and also guide therapeutic decision making. We focus on the EGFR pathway as the historical prototype for molecular targeting in lung cancer while discussing the molecular pathways and current therapeutic approaches involving K-RAS-, ALK-, ROS-1-, and MET-driven tumors.

5.9 EGFR Pathway

The epidermal growth factor receptor is a cell-surface receptor which is activated in more than half of patients with NSCLC. This activation can result from protein overexpression, increased gene copy number, or genetic mutation [42]. All members of the receptor family have intrinsic tyrosine kinase activity, with the notable exception of Her-3. The epidermal growth factor receptor family is involved in a plethora of cellular responses, including proliferation, suppression of apoptosis, cell motility, invasion, and angiogenesis. The receptors exist as inactive monomers.

However, the binding of extracellular growth factors, such as epidermal growth factor (EGF) and other EGF-like growth factors, including transforming growth factor alpha, results in receptor dimerization. Dimerization results in autophosphorylation of tyrosine residues in the activation loop of the EGFR kinase domain, which activates downstream and intracellular signaling cascades. Docking and partner proteins trigger two primary pathways—namely, the RAS-RAF-MEK-ERK pathway as well as the PI3K-AKT-mTOR pathway—which induce cell proliferation, survival, angiogenesis, and metastasis. The JAK-STAT survival pathway is also activated through EGFR receptor dimerization.

EGFR mutant non-small-cell lung cancer was first recognized in 2004 as a distinct, clinically relevant molecular subset of lung cancer. More recently, it has become a paradigm for understanding and treating oncogenic-driven carcinomas in general.

5.10 EGFR-Directed Therapies

The initial anti-EGFR therapies in the 1990s were directed at the wild-type receptor, which was overexpressed in many epithelial cancer types. Investigators initially noted that the first EGFR tyrosine kinase inhibitor (TKI), gefitinib, demonstrated objective radiographic responses in 10 of 100 patients with NSCLC who had received several prior lines of chemotherapy [43]. In phase II studies, it was determined that responders tended to be patients with adenocarcinoma histology, East Asian ethnicity, never smokers, and female gender [44]. However, these simplistic criteria can no longer be considered the sine qua non for patient selection for molecular marker testing.

In 2004, EGFR kinase domain mutations were first described. The presence of activating somatic mutations in the EGFR kinase domain was shown to be associated with the clinical characteristics of responding patients and increased sensitivity to the EGFR TKIs, gefitinib and erlotinib [45, 46]. Activating mutations in EGFR occur in exons encoding the kinase domain (exons 18–21). The most common activating mutations in EGFR are a point mutation in exon 21, which substitutes an arginine for a leucine (L858R), and a small in-frame deletion in exon 19 that removes four amino acids (LREA) between residues 747–750 of the EGFR polypeptide. Together, these two genetic changes account for ~90 % of TKI-sensitive mutations that are observed in EGFR mutant tumors.

EGFR mutant tumors are seen most commonly in patients with adenocarcinoma histology and are associated with better prognosis than EGFR wild-type tumors [47]. In the past 5–10 years, numerous prospective trials for patients with advanced NSCLC and activating EGFR mutations have confirmed the benefit of EGFR TKIs in EGFR mutant lung cancer. Trials have been performed broadly, in East Asia, the United States, and Europe, with both gefitinib and erlotinib, with radiographic response rates ranging from 55 to 91 % and progression-free survival (PFS) ranging from 7.7 to 12.3 months [48].

In 2009, the landmark randomized phase III Iressa Pan-Asia Study (IPASS) showed that an EGFR TKI is superior to chemotherapy as an initial treatment for chemotherapy-naïve EGFR mutant metastatic lung cancer [49]. The IPASS trial was conducted in Asia and involved clinically selected patients with limited tobacco exposure and advanced adenocarcinoma of the lung who had not received previous chemotherapy. Patients were randomized to receive gefitinib or standard chemotherapy with carboplatin and paclitaxel. Although there was no overall difference in PFS between the two groups, the PFS of patients with EGFR mutant tumors was significantly longer among those who received gefitinib than among those who received carboplatin-paclitaxel (hazard ratio for progression or death 0.48, 95 % CI 0.36–0.64, $p<0.001$) [49]. In two randomized trials conducted in Japan, gefitinib was compared with two combinations of platinum and taxane agents as first-line therapy for patients with advanced NSCLC who carried activating EGFR mutations. Gefitinib significantly improved median PFS, as compared with carboplatin-paclitaxel (10.8 vs. 5.4 months; HR 0.30; 95 % CI 0.22–0.41, $p<0.001$), and as compared with cisplatin and docetaxel (9.2 vs. 6.3 months; HR 0.49; 95 % CI 0.34–0.71, $p<0.001$) [50, 51]. Erlotinib has been shown to be similarly effective for EGFR mutant tumors [52].

Erlotinib and gefitinib, similar to other EGFR TKIs, bind to the ATP-binding site of the EGFR tyrosine kinase domain, thereby blocking catalytic activity of the kinase. The net result is inhibition of the downstream signaling of the pathways responsible for cellular proliferation. Recently, multiple mechanisms of primary and secondary acquired resistance of lung tumors to EGFR TKIs have been described. Secondary mutations in EGFR, namely, T790M and/or MET amplification, account for most cases of resistance. There are changes in epithelial to mesenchymal transition (EMT) that also account for EGFR TKI resistance [53]. Many new rationally directed strategies are being employed in clinical trials to overcome resistance, and these are discussed below.

5.11 Biology of EGFR Mutations

EGFR mutations are typically heterozygous, with the mutant allele demonstrating gene amplification [54]. Several genomic studies confirm that EGFR mutant NSCLC represents a distinct phenotype with unique expression, mutation, and copy number signature [55]. L858R and G719S TKI-sensitive EGFR mutants show that these substitutions activate the kinase, resulting in receptors with 50-fold more activity compared to their wild-type counterparts [56, 57]. The presence of a TKI-sensitive mutation results in preferred binding of gefitinib or erlotinib versus ATP. Inhibition of the EGFR survival pathway with targeted therapies such as gefitinib or erlotinib is mediated through the intrinsic apoptotic pathway. BIM, a BCL-2 pro-apoptotic family member regulated by ERK signaling, is essential for apoptosis triggered by EGFR kinase inhibitors [58, 59].

5.12 Resistance to EGFR TKI

Resistance to EGFR TKI drugs can be either de novo (primary) or acquired (secondary). In EGFR mutant tumors, a 75 % response rate with EGFR TKI is seen, indicating that 25 % of patients will have primary resistance. Small insertions or deletions in exon 20 are observed in approximately 5 % of NSCLC. In vitro studies suggest that such mutations are less sensitive to EGFR TKIs [60]. Other patients can develop de novo resistance via T790M, which is encoded by exon 20 [61]. This mutation is more commonly found in patients with acquired resistance.

Most tumors without EGFR kinase domain mutations are insensitive to tyrosine kinase inhibition. The initial observation that K-RAS mutant lung cancers are resistant to EGFR TKIs has been studied [62]. While K-RAS testing has been routinely adopted as a negative predictor of benefit from EGFR-directed therapy in colorectal cancer, K-RAS testing has not been widely adopted in lung cancer [63]. Other genomic alterations that may coexist with EGFR mutations, such as downstream mutations in PIK3CA, PI3K, and loss of PTEN, have been shown to be less responsive to treatment with EGFR TKI [64]. Cross talk between EGFR and other signaling pathways, such as insulin-like growth factor receptor 1 (IGFR1), may also mediate resistance. Recently, hepatocyte growth factor (HGF), the ligand for MET receptor tyrosine kinase, has been described as a mechanism of resistance. HGF binding increases MET-mediated activation of the PI3K-AKT pathway, decreasing the ability of EGFR TKI to abrogate this signaling cascade [65].

Despite initial response, after a median period of 10–14 months, EGFR mutant lung cancers acquire secondary resistance to EGFR TKI therapy. Jackman et al. proposed the following definition of acquired resistance to EGFR TKIs, which has largely become adopted as a routine standard: previous treatment with a single-agent EGFR TKI; a tumor that harbors an EGFR mutation known to be associated with drug sensitivity and/or objective clinical benefit from treatment with an EGFR TKI; systemic progression of disease (by RESIST or WHO criteria) while on continuous treatment with gefitinib or erlotinib within the past 30 days; and no intervening systemic therapy between cessation of TKI and initiation of new therapy [66].

A common mechanism of secondary resistance to EGFR TKI involves a secondary-site mutation in the threonine gatekeeper residue at position 790 of the EGFR gene (T790M). The T790M mutation in EGFR is found in 50 % of EGFR mutant tumors with acquired resistance to erlotinib or gefitinib. The above substitution of methionine for threonine leads to altered drug binding in the ATP pocket of EGFR. In addition, the secondary T790M mutation restores ATP affinity back to the level of wild-type EGFR, in effect abrogating the initial activating mutation [67]. Interestingly, in cases of acquired resistance, T790M may demonstrate a better prognosis than non-T790M mutations [68]. In addition, while T790M is more likely to show progression in lungs/pleura, non-T790M is more likely to progress distantly [68]. Other second-site mutations in EGFR have been associated with acquired resistance, including L747S (exon 19), D761Y (exon 19), and T854A (exon 21 in the activation loop) [69].

Table 5.2 Mechanisms of acquired resistance to EGFR tyrosine kinase inhibitor-directed therapies

T790M and rare second-site mutations	60 %
Unknown—includes epithelial to mesenchymal transformation	30 %
Small-cell transformation	6 %
MET amplification	4 %

Percentages are based on aggregate data from the two largest re-biopsy series to date (Arcila et al., $n=99$, and Sequist et al., $n=37$) [53, 132, 133]. Small-cell transformation includes cases with histologic change to neuroendocrine differentiation. MET amplification represents cases without coexisting EGFR T790M

MET oncogene amplification has been observed in 4–20 % of tumor samples harboring clinical resistance to EGFR TKI [70, 71]. Cells with MET amplification can signal through the ERBB3 pathway to maintain activation of AKT, despite presence of EGFR TKI. Recent studies have also demonstrated that increase in receptor tyrosine kinase (RTK) ligand, namely, hepatocyte growth factor—the ligand for MET—can act as a primary culprit of drug resistance [72]. These elegant studies have shown that increase in RTK ligand through autocrine tumor cell production, paracrine contribution from tumor stroma, or systemic production confers resistance to inhibitors of an oncogenic kinase. We also have evidence that MET can be amplified de novo, without any resistance mechanisms [73]. HGF, in particular, can reactivate both the MAPK and AKT signaling pathways [72, 74].

Large re-biopsy series have suggested other mechanisms of EGFR TKI acquired resistance, including an epithelial to mesenchymal transition (EMT) or transformation to a small-cell lung cancer phenotype [53] (Table 5.2). Such histologic changes may account for a substantial percentage of the unknown mechanisms of acquired resistance and remain an area of active investigation.

5.13 Overcoming Resistance to EGFR TKI

There is an ongoing rationale to develop trials to determine optimum upfront treatments for patients whose tumors harbor EGFR mutations. One strategy involves utilization of the requirement for BIM and adding a BCL-2 inhibitor to enhance TKI-induced apoptosis [58]. For drug-resistant EGFR mutations, such as exon 20 insertions or deletions, second generation (i.e., afatinib) and third generation EGFR TKIs with more potency may overcome such resistance [75]. For genomic alterations that co-occur with EGFR mutations, drug combinations are being pursued. For example, as IGF1R signaling can mediate disease resistance through sustained activation of PI3K-AKT, addition of a PI3K or AKT inhibitor to TKI treatment may prove to be beneficial. Recently, preclinical data suggests that combined EGFR/MET or EGFR/HSP90 inhibition is effective in the treatment of lung cancers codriven by mutant EGFR containing T790M and MET [76]. The above study provides preclinical proof of principle that combination targeting of EGFR and MET may benefit patients with NSCLC.

Second generation EGFR inhibitors were developed to overcome T790M-mediated resistance. In preclinical models, such agents were shown to be more potent against the second-site mutation than gefitinib or erlotinib. However, their clinical efficacy remains to be established. Perhaps the most promising of the second generation EGFR TKIs involves afatinib. In the LUX-1 study, afatinib was compared to placebo for patients with advanced, metastatic NSCLC after failure of erlotinib, gefitinib, or both and one or two lines of chemotherapy and demonstrated improvement in PFS. Recent LUX-Lung 3 trial data has shown that in patients with stage IIIB or IV EGFR mutation NSCLC taking afatinib as a first-line treatment, progression-free survival approached 1 year (PFS of 11.1 months) versus just over half a year (PFS of 6.9 months) for those treated with pemetrexed/cisplatin [77].

With respect to drug combinations, to simultaneously target EGFR and its downstream target AKT, irreversible EGFR inhibitors have been paired with mTOR inhibitors (such as rapamycin) [78]. Whether the above approach is effective in acquired resistance remains to be established. Dual inhibition of EGFR with afatinib and cetuximab showed promising results in a clinical trial and suggested the possibility of overcoming acquired resistance. Activity was not specific to the common T790M mutant. To date, trials of MET inhibition have been in the TKI naïve population, though combination strategies with EGFR TKI may be promising. Finally, the question of whether to continue treatment with an EGFR TKI in patients with acquired resistance is still debated. In standard practice, progression on TKI oftentimes leads to discontinuation of therapy, though some studies suggest that disease flares and re-responses to drug may be seen even after progression [79].

Recent concepts of tumor biology and resistance suggest that tumors are not a homogenous mass consisting of one population of cells. In fact, a heterogeneous mixture of resistant and sensitive cells likely comprise all tumors, raising questions about the role for re-biopsy and whether different populations become dominant under different stressors [80]. The above has implications for clinical trial development and design. Overall, with the discovery and advanced therapeutics for patients with EGFR mutation, patients with metastatic disease are achieving survival rates which are double that of patients with wild-type EGFR tumors.

5.14 ALK-Driven NSCLC

ALK, similar to EGFR, is a receptor tyrosine kinase. It is normally not expressed in lung tissue. The enzyme was originally identified in anaplastic large-cell lymphoma, with fusion of ALK (chromosome 2) to a nucleolar protein gene (chromosome 5), NPM, which allowed for a chimeric protein (NPM-ALK) [81]. In 2007, identification of a transforming echinoderm microtubule-associated protein-like 4 (EML4) gene and the anaplastic lymphoma kinase (ALK) fusion gene (both proximal to each other on chromosome 2) was first described in a surgically resected lung adenocarcinoma specimen [82]. A small inversion within chromosome 2p, in which the N terminus of the EML4 gene becomes fused to the intracellular kinase domain of ALK, results in the formation of the EML4-ALK fusion gene. In vitro, the fusions

yield gain of function. Transgenic mice expressing EML4-ALK develop hundreds of adenocarcinoma nodules in weeks, demonstrating that the above fusion has clear oncogenic activity. Over nine different fusion variants have been described [83]. This gene rearrangement occurs largely independent of EGFR and other mutations.

The overall incidence of the rearrangement has been reported to be 1–7 % with the use of reverse-transcriptase polymerase chain reaction (PCR), fluorescence in situ hybridization (FISH), and immunohistochemistry [84]. Patients with the EML4-ALK fusion oncogene tend to be younger, have adenocarcinoma histology, and be light or never smokers. Interestingly, in the largest series to date, NSCLC patients with at least two of the following characteristics were selected for genetic screening: female sex, Asian ethnicity, never/light smoking history, and adenocarcinoma histology. Thirteen percent of patients were found to be EML4-ALK mutant, while 22 % were EGFR mutant [84].

PF-02341066, now known as crizotinib, had been developed to target c-MET, but was known to also inhibit ALK. While patients with the fusion oncogene appear to be resistant to EGFR TKIs, small-molecule TKI crizotinib showed activity in cell lines containing the EML4-ALK fusion gene [85]. In a phase I trial, the maximum tolerated dose (MTD) of crizotinib was determined to be 250 mg twice a day, with an objective response rate of 61 % [86]. Preliminary results from the phase II study were reported at the World Lung Congress in July 2011. Among 133 patients with advanced, ALK-positive NSCLC, the objective response rate was 51 %, and the disease control rate at 12 weeks was 74 %. The follow-up of phase II patients was too short to evaluate PFS. In August 2011, the Food and Drug Administration (FDA) granted accelerated approval to crizotinib with an FDA-approved companion diagnostic test. The above example highlights the importance of defining a predictive biomarker assay early, or alongside new and targeted drug development.

Screening patients for EML4-ALK rearrangement can certainly offer ALK-positive patients the opportunity to benefit from a highly effective and well-tolerated therapy. Screening can be enriched by selecting patients who are younger, with adenocarcinoma histology, and whom are light or never smokers. The FDA-approved test is the FISH analysis; however, as we learn more about the molecular diagnostics, we will also incorporate sequencing methods, IHC, as well as other technologies. Not only are EML4-ALK translocations important, it will be important to determine the variants of EML4-ALK as well as other translocation partners for ALK. Like other targeted therapies, patients with ALK-positive NSCLC eventually relapse on crizotinib, typically within 1 year of therapy. An active area of research involves overcoming resistance, particularly to gatekeeper mutations within the ALK tyrosine kinase domain [87].

5.15 ROS-1

An original report identified a ROS-1 translocation in an Asian patient with NSCLC, but prevalence of this genetic alteration had been lacking until recently [88]. A recent study identified ROS-1 rearrangements in 1.7 % of patients with NSCLC

using fluorescence in situ hybridization [89]. Similar to EGFR- and ALK-positive patients, patients with ROS-1 translocation were more likely to have adenocarcinoma, be Asian, younger, and never smokers.

Preclinical studies identified TAE684 as a kinase inhibitor with activity against HCC78 lung cancer cell lines harboring a ROS-1 translocation [90]. In subsequent studies, crizotinib showed in vitro and early evidence of clinical activity in ROS-1 rearranged NSCLC [89]. Presently, there is little known about the signaling pathway from activated ROS-1 kinase. As FISH remains the gold standard in many contexts, future directions may involve alternative screening strategies using immunohistochemistry, due to its rapid and cost-effective manner to screen for such subsets.

5.16 K-RAS-Driven NSCLC

K-RAS mutations represent the most common molecular change in NSCLC, with the Lung Cancer Mutation Consortium identifying RAS mutations in 22 % of tumor samples [40]. The presence of RAS mutation has been shown to be associated with a poor prognosis with an OS hazard ratio of 1.4 (95 % CI 1.18–1.65, $p=0.01$) for K-RAS mutant compared to wild type for all studies and 1.5 (95 % CI 1.26–1.8, $p=0.02$) for studies specific for adenocarcinoma [91]. Given our limited number of effective pharmacologic drugs to target this pathway, the clinical utility of the mutation has been debated recently [63]. Previously, K-RAS status was used to predict and select patients who may benefit from EGFR TKI and anti-EGFR therapies. However, as the EGFR mutational status has demonstrated significant predictive value, it has become the preferred test over RAS testing.

5.17 RAS Biology

Oncogenes of the RAS family encode for proteins on the cytoplasmic surface of cell membranes. RAS proteins are guanosine diphosphate (GDP)/guanosine triphosphate (GTP)-regulated binary on-off switches. In quiescent cells, RAS is GDP bound and inactive until extracellular stimuli cause temporary activation of the GTP-bound form of RAS. Mutant RAS proteins render the proteins constitutively GTP bound and activated, leading to stimuli-independent, persistent activation of RAF, mitogen-activated protein kinase kinase (MEK), ERK, and phosphatidylinositol 3-kinase (PI3K), AKT, and mammalian target of rapamycin (mTOR), promoting cell proliferation, survival, and metastasis [92].

Although mutations in each of the three RAS genes—H-RAS, K-RAS, and N-RAS, have been linked with malignancy, K-RAS is most closely associated with NSCLC [93]. In lung cancer, K-RAS mutations occur primarily at codon 12 or 13 [94]. Point mutations have been described at codons 12, 13, and 61 which result in loss of intrinsic GTPase activity. It is unclear whether mutations at different residues result in unique mutant proteins with distinct clinical outcomes, though there

may be a suggestion as to the above. For example, NSCLC cell lines with mutant K-RAS (G12D) had activated PI3K and MEK signaling, whereas NSCLC cell lines with mutant K-RAS (G12C) or mutant K-RAS (G12V) had activated Ral signaling and decreased growth factor-dependent AKT activation. Such findings may have important ramifications for clinical trial design and the use of K-RAS mutational analysis as a biomarker for therapeutic approaches.

K-RAS mutations occur with 20–30 % prevalence in adenocarcinoma histology and are rarely found in squamous-cell cancers (less than 5 %) [95]. K-RAS mutations are associated with history of tobacco use, with a meta-analysis revealing higher rate of K-RAS mutation among former or current smokers compared with never smokers (25 % and 6 %, respectively) [96].

Retrospective pooled analyses investigated the prognostic and predictive effect of K-RAS mutational status. In these analyses, K-RAS mutational status did not seem to be prognostic or predictive of adjuvant chemotherapy benefit. Two meta-analyses have investigated the interaction between EGFR TKI benefit and K-RAS status. K-RAS mutants compared with wild-type tumors were significantly associated with a lack of response to EGFR TKI [97]. The predictive value and increasing use of EGFR mutation status for selection of patients for EGFR TKI have diminished the utility of K-RAS testing as a negative predictor of clinical benefit.

5.18 RAS Utility in Clinical Trials

Perhaps the primary utility of K-RAS testing is to prospectively select patients for clinical trials of targeted therapies that inhibit part of the pathway or demonstrate promise in K-RAS mutant lung cancer. MEK and PI3K inhibitors have the most evidence to support their clinical development as possible anti-RAS strategies. MEK protein kinases are downstream of RAS and RAF. In a phase II trial, selumetinib (AZD6244)—a potent and selective MEK inhibitor—revealed similar efficacy to pemetrexed in unselected advanced NSCLC in the second-line setting. In another randomized phase II trial, selumetinib in combination with docetaxel was compared to docetaxel alone, for the second-line treatment of 87 patients with K-RAS mutant, locally advanced or metastatic NSCLC. The combination arm demonstrated improved overall survival of 9.4 months, as compared to 5.2 months, although the above result did not reach statistical significance. However, selumetinib + docetaxel improved PFS, when compared to docetaxel + placebo (5.3 versus 2.1 months, respectively; HR 0.58; 80 % CI 0.42–0.79, $p=0.0138$). It has been suggested that selumetinib may provide benefit for K-RAS/p53 tumors but not K-RAS/LKB1 tumors, and further stratification by LKB1 status may prove to be helpful [98].

The other downstream effector of mutant K-RAS is the PI3K/AKT/mTOR signaling cascade. Several inhibitors against this pathway are in clinical development, and their role as monotherapy or in combination with standard therapies, remains to be elucidated. In fact, data from preclinical K-RAS mutant, genetically engineered mouse models of NSCLC suggest that dual inhibition of PI3K and MEK pathways

may be required. In one such study, when NVP-BEZ235 a dual pan-PI3K and mTOR inhibitor was combined with a MEK inhibitor, ARRY-142886, there was marked synergy in shrinking K-RAS mutant cancers [99]. Such in vivo studies suggest that inhibitors of the PI3K-mTOR pathway may be active in cancers with PIK3CA mutations and, when combined with MEK inhibitors, may effectively treat K-RAS mutated lung cancers.

K-RAS mutant non-small-cell lung cancers also demonstrate sensitivity to heat shock protein-90 (HSP-90) inhibitors, and mouse models have revealed marked response with such inhibitors [100]. Trials of HSP-90 inhibitors in combination with chemotherapy are ongoing and will include analysis of efficacy in K-RAS mutant tumors.

The concept of "synthetic lethality," in particular, has become a model for targeting K-RAS driven tumors. In one such study, a pooled shRNA-drug screen strategy was utilized to identify genes that, when inhibited, cooperate with MEK inhibitors to effectively treat K-RAS mutant cancer cells [101]. The antiapoptotic BH3 family gene BCL-XL emerged as a primary hit through this approach. ABT-263, a chemical inhibitor that blocks the ability of BCL-XL to bind and inhibit proapoptotic proteins, in combination with a MEK inhibitor led to dramatic apoptosis in K-RAS mutant cell lines and K-RAS mutant xenografts, suggesting synergistic lethality with BCL-XL/MEK inhibition as a potential therapeutic approach for K-RAS mutant cancers [101]. A synthetic lethal interaction between K-RAS oncogenes and Cdk4 as a therapeutic strategy for NSCLC in patients carrying K-RAS oncogenes has also been recently described [102]. A recent review outlines other pathways involved in synthetic lethality, including the WT1 pathway [103], TBK1 and the NF-KB pathway [104, 105], the *enhancer of zeste homolog 2* (EZH2) epigenetic gene silencing pathway [106], the *seven in absentia homolog 2* (Siah 2) ubiquitin pathway [107], *the GATA-binding factor 2* pathway [108], and the IL-8 (CXCL8) pathway [109]. While K-RAS was previously an elusive clinical target in NSCLC, new drug developments and downstream inhibitors of K-RAS seem promising.

5.19 MET as a Therapeutic Target in NSCLC

MET receptor tyrosine kinase is located on chromosome 7q21-q31 and can demonstrate activating mutations, aberrant overexpression, and amplification in certain subsets of lung cancers. Certain JM domain mutations (R988C, T1010I, alternative spliced JM-deleting variant) are oncogenic activating variants with enhanced oncogenic signaling, cell motility, and migration. There have also been reports of exon skipping with exon 14 of MET and gain of function [110]. Overall, in protein studies of human lung cancer tissue, 67 % of adenocarcinomas, 60 % of carcinoids, 57 % of large-cell carcinomas, 57 % of squamous-cell carcinomas, and 25 % of SCLC strongly express MET [111]. In large cohorts of patients with NSCLC not previously treated with EGFR-specific TKI, MET amplification is seen in 1.4–21 % of patients [112].

The sole ligand for MET is hepatocyte growth factor/scatter factor (HGF/SF). In normal cells, hepatocyte growth factor-induced MET activation is under tight regulation by paracrine ligand delivery. The ligand HGF can also be overexpressed in lung cancer or expressed in stroma, and both the MET receptor and the HGF ligand can be targets for therapeutics.

MET tyrosine kinase is activated when HGF ligand binds to the SEMA domain of MET at the plasma membrane [113]. Upon binding of the HGF ligand to MET, MET dimerization, autophosphorylation, and activation of tyrosine kinase catalytic activity occurs. Docking proteins, in turn, activate two major downstream pathways, including the RAS-RAF-MAPKK-ERK pathway as well as the PI3K-AKT-mTOR-NF-kB pathway [114]. MET has also been shown to cross talk with various signaling pathways, including VEGF and EGFR, among others [80, 115]. The protein product of MET gene, hepatocyte growth factor receptor (HGFR), has been implicated in various oncogenic processes including cell proliferation, survival, invasion, motility, and metastasis.

Clinically, the role of MET in the development of EGFR TKI resistance has been an active area of research. In an in vitro model, Engelman et al. showed that MET amplification causes gefitinib resistance through ERBB3 (Her-3)-dependent activation of PI3K [71]. In some preclinical models, resistance appears to be overcome by combined dual inhibition of EGFR and MET, as shown in lung, pancreatic, and breast tumor xenografts. These results provide a rationale for ongoing clinical studies of MET inhibitor, alone and in combination with EGFR TKI in NSCLC.

METMab, also known as onartuzumab, is a single-armed humanized monoclonal antibody that binds the SEMA domain of MET. By competing for the binding of HGF to MET, METMab acts like a classic receptor antagonist. METMab blocks ligand-induced MET dimerization and prevents activation of MET's kinase domain. Dual inhibition of c-MET and EGFR in NSCLC was recently studied in a randomized, double-blind, phase II study. This study compared METMab with erlotinib versus placebo with erlotinib as second/third line therapy in advanced NSCLC. METMab with erlotinib resulted in improved PFS and OS, with OS benefit noted in the arm with MET FISH ≥ 5 copies as well as FISH (−)/IHC (2+/3+) arm (HR = 0.37, median 12.6 months vs. 4.6 months, $p = 0.002$) [116].

The MET kinase inhibitor, ARQ-197 (tivantinib), has also been studied in locally advanced or metastatic NSCLC. ARQ 197–209, a randomized, placebo-controlled phase II clinical trial of erlotinib + ARQ 197 in previously treated EGFR inhibitor-naive patients, was found to be superior to erlotinib + placebo [117]. Although the phase III trial has been reported as negative, further biomarkers need to be assessed, and a potential biomarker-driven trial may provide additional insight.

Through detailed analysis of the lung cancer genome, other new and exciting biomarkers have been elucidated including RET, BRAF, and PI3K. Such biomarkers have come to fruition very quickly and represent clinically relevant subsets of lung carcinoma which galvanize excitement for additional specific targeted therapies and drug development.

5.20 Molecular Biomarker Testing

Practitioners have been faced with the challenge of efficient and effective molecular testing, as the underlying biology and molecular pathways involved in lung cancer have become better defined. Foundation Medicine, which represents one such cancer diagnostics company, provides comprehensive analysis of tumor tissue through genomic analysis. The company's first assay, FoundationOne, includes a genomic profile to identify a patient's individual molecular alterations and match them with relevant targeted therapies and clinical trials [118]. Caris Life Sciences uses techniques of both genomic and proteomic (i.e., immunohistochemistry) analysis. Utilizing strategies of IHC, FISH, PCR, and next generation sequencing, Caris allows practitioners to customize the level of tumor profiling, as necessary for each patient [119]. Interestingly, not only has therapy become personalized, but so has the ability to obtain relevant biomarkers. We have developed a customized algorithm for molecular biomarker testing in the thoracic oncology program. The goal of such testing is to provide complex genomic and proteomic results to oncologists, other health care providers, and their patients in order to choose the most effective therapy.

Biomarkers, and biomarker testing, are a rapidly moving target. One such example involves the excision repair cross-complementation group 1 (ERCC1) protein, which has been described as a potential prognostic biomarker of efficacy of cisplatin-based chemotherapy in NSCLC [120, 121]. Although ERCC1 as a biomarker of patient survival, treatment efficacy, or both has been studied at the genomic, transcriptional, and protein level, a recent study was unable to validate the predictive effect of immunostaining for ERCC1 protein [122]. This discordance suggested that immunohistochemical analysis with currently available ERCC1 antibodies has limited usefulness in guiding therapeutic decision making. The above example highlights the technical biases, limitations, and caution one must employ when conducting biomarker studies.

5.21 Genomic Characterization of Lung Cancer

The standard of care for patients with advanced NSCLC now involves selecting biomarker-driven treatment algorithms based on molecular profiling of a patient's tumor. The genomic characterization of lung cancer has become possible with recent advances in multiplex genotyping and high-throughput next generation sequencing.

Personalized cancer care has benefitted greatly from advances in DNA-based high-throughput genomic technologies. First generation Sanger sequencing has given way to next generation sequencing, which no longer requires a gel- or polymer-based matrix and no longer requires prior knowledge of the genome sequence [123]. Nucleic acid sequencing is now significantly faster, and with reduced error and cost, progressing from single biomarker tests to multiplex, hot

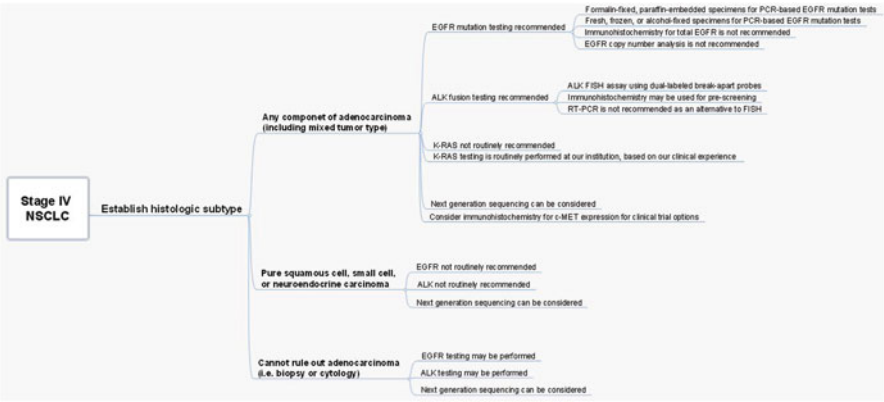


Fig. 5.2 Molecular genotyping as a mind map for differing histologic subtypes**.** Included are guidelines from the College of American Pathologists (CAP), International Association for the Study of Lung Cancer (IASLC), and Association for Molecular Pathology (AMP) for the most common "actionable" molecular subsets (i.e., EGFR, ALK, K-RAS) based on the histologic subtype. EGFR and ALK testing is recommended for adenocarcinomas and mixed lung cancers with an adenocarcinoma component, regardless of smoking status or histologic grade. EGFR and ALK testing is not routinely recommended in lung cancers that lack any adenocarcinoma component, such as pure squamous-cell carcinoma, pure small-cell carcinoma, or large-cell carcinoma. Next generation sequencing (NGS) is a tool being utilized with increasing frequency for molecular genotyping, though specific recommendations on testing with NGS have yet to be described. *CAP* College of American Pathologists, *IASLC* International Association for the Study of Lung Cancer, *AMP* Association for Molecular Pathology, *NGS* next generation sequencing, *NSCLC* non-small-cell lung cancer, *PCR* polymerase chain reaction, *FISH* fluorescence in situ hybridization

spot mutation tests to initial high-throughput technologies to, ultimately, next generation sequencing (NGS).

Next generation sequencing will likely have applications for genotype-based molecular biomarker development. Single-gene biomarkers have already proven successful in guiding selection of molecularly targeted agents in NSCLC. NCCN and ASCO, for example, both recommend EGFR mutation and ALK gene rearrangement testing on all NSCLC with an adenocarcinoma component, regardless of smoking status, histologic grade, or dominant histologic subtype. Such testing is not recommended for pure squamous-cell carcinoma, small-cell carcinoma, or neuroendocrine tumors. However, this too is not truly clear since there are some incidences of EGFR/ALK mutations in non-adenocarcinoma histology. With the knowledge of distinct genetic abnormalities in each histologic subtype of lung cancer, there will be opportunities to develop novel molecularly targeted and biomarker-driven strategies (Fig. 5.2).

An increasing need has been to develop methods to simultaneously query the mutational or expression status of many genes of interest. Multiplex polymerase chain reaction, with platforms such as Sequenom or SNaPShot, can identify potentially actionable molecular targets. Sequenom can use fresh, frozen, or formalin-fixed paraffin-embedded samples. It can detect and quantify mutation frequencies from at least 10 % of mutation-positive cells [124]. SNaPShot, on the other hand,

interrogates more than 50 hot spot mutation sites in up to 14 key cancer genes [125]. SNaPShot is fairly labor intensive, with a 2–3 week turnaround time, and requires more genomic DNA than Sequenom. Importantly, these multiplex genomic tests only detect expression of selected and known hot spot mutations and oncogenes and do not have the capability of discovering new targets [123]. As inactivation of a tumor suppressor can involve deletions or changes in a wide region of the loci, analysis becomes more challenging or at times, not feasible.

NGS technologies may ultimately offer rapid genome-wide characterization of DNA, mRNA, transcription factor regions, miRNA, chromatin structure, and DNA methylation which may prove to be most insightful. One primary advantage of NGS is that its coverage, or average number of sequencing reads that align to each base within the sample DNA, is highly adjustable [123]. During the ASCO 2012 annual meeting, it was noted that NGS is being used as a research tool in all types of malignancy to understand tumor molecular mechanism, discovery of novel drug targets, and screening candidate patients for clinical trials. Certainly informatics tools will need to be concurrently employed with the robust information generated with the above platform.

With the use of such platforms, other genetic alterations in non-small-cell carcinoma have been identified, including PIK3CA, BRAF, and HER-2. In one such study, 552 NSCLC tumors were analyzed for genetic abnormalities using a multiplexed PCR-based SNaPShot assay, plus FISH for ALK translocations as a part of routine clinical practice [126]. Prevalent mutations were in K-RAS (24 %), EGFR (13 %), PIK3CA (4 %), and ALK translocations (5 %). PIK3CA mutations were seen more commonly in squamous-cell carcinoma. Mutations in IDH and B-catenin were seen with lower frequency. The authors report that their molecular analysis steered patients toward a genotype-directed targeted therapy in 22 % of cases and also directed patients toward relevant clinical trials [126]. A whole-genome and transcriptome sequencing analysis of 17 patients with NSCLC was also reported, with a focus on the genomic landscape of smokers versus nonsmokers [127]. Novel alternations in genes involved in chromatin modification and DNA repair pathways were identified, along with DACH1, CFTR, RELN, ABCB5, and HGF. Analysis revealed 14 fusions, including ROS-1 and ALK, as well as novel metabolic enzymes. It was suggested that histone deacetylase (HDAC) inhibitors may play an increasing role given the number of events in chromatin modifier genes which were described. Interestingly, there were perturbations in 54 potentially targetable genes, with currently available drugs [127].

Similarly, parallel sequencing of non-small-cell lung cancer with adenocarcinoma histology was recently described through exome and genome sequences of 183 lung adenocarcinoma tumor samples [128]. Novel targets including recurrent somatic mutations in the splicing factor gene U2AF1 and truncating mutations affecting RBM10 and ARID1A were seen. Whole-genome sequence analysis revealed frequent structural rearrangements including in-frame exonic alternations in EGFR and SIK2 kinases [128].

Such integrative genome analyses in small-cell cancer identify inactivation of p53 and RB1 along with recurrent mutations in CREBBP, EP300, and MLL genes

Table 5.3 Alterations in targetable oncogenic pathways in squamous-cell lung cancer

PI3K/AKT pathway alterations	Frequency
PTEN	15 %
PIK3CA	16 %
AKT1	<1 %
AKT2	4 %
AKT3	16 %
STK11	2 %
TSC1	3 %
TSC2	3 %
RTK/RAS pathway alterations	**Frequency**
EGFR	9 %
ERBB2	4 %
ERBB3	2 %
FGFR1	7 %
FGFR2	3 %
FGFR3	2 %
K-RAS	3 %
H-RAS	3 %
N-RAS	<1 %
RASA1	4 %
NF1	11 %
BRAF	4 %

Percentage of samples ($n = 178$) with alterations in the PI3K/RTK/RAS pathways, as obtained by the Cancer Genome Atlas Research Network, through whole-exome sequencing and whole-transcriptome expression profiles. Alterations are defined by somatic mutation, homozygous deletion, high-level, focal amplification, and in some cases by significant up- or downregulation of gene expression (AKT3, FGFR1, PTEN) [131]

that encode histone modifiers [129]. In this study which sequenced 29 small-cell lung cancer exomes, 2 genomes, and 15 transcriptomes, mutations in PTEN, SLIT2, and EPHA7, as well as focal amplifications of the FGFR1 tyrosine kinase gene, were seen as well. Overall, the study implicated histone modification as a major feature of small-cell lung cancer while identifying possible novel therapeutic targets. A concurrent comprehensive genomic analysis identified SOX-2 as a frequently amplified gene in small-cell lung cancer, at a rate of approximately 27 % [130]. Suppression of SOX-2 using shRNAs blocked proliferation of SOX-2 amplified cell lines. RNA sequencing also identified multiple fusion transcripts and a recurrent RLF-MYCL1 fusion [130].

A comprehensive genomic characterization of squamous-cell lung cancer was undertaken as a part of The Cancer Genome Atlas [131] (Table 5.3). One hundred seventy-eight lung squamous-cell carcinomas were profiled. Statistically recurrent mutations were seen in 11 genes, including mutation of p53 in nearly all specimens.

Previously unreported loss of function mutations were seen in the HLA-A class I major histocompatibility gene. Altered pathways included NFE2L2 and KEAP1 in 34 %, squamous differentiation genes in 44 %, and PI3K pathway genes in 47 %, and CDKN2A and RB1 in 72 % of tumors [131]. Other frequently mutated targets included PTEN, cyclin-dependent kinase, PIK3CA, PI3K, fibroblast growth factor receptor, and FAK. Such genomic characterization of lung cancer will have marked implications for therapeutic intervention and drug development. Several pharmaceutical drugs are already available (or in development) to target the vast majority of the above pathways, and we will likely see these agents in clinical trials, alone or in combination, with standard chemotherapy.

NGS is not without its challenges, however. Challenges and opportunities surrounding the quantity and quality of tumor tissue, intra-patient tumor heterogeneity, dynamic changes within the cancer genome at various points in the disease course, and considerations of when to re-biopsy patients will need to be addressed as NGS technologies become more prevalent. In addition, even with enhanced informatics tools to synthesize the above genomic information, integrating the above information into practice will be a joint interdisciplinary endeavor.

5.22 Future Directions

Future directions in lung cancer care and research will involve incorporation of molecular characteristics and next generation sequencing into screening strategies to improve early detection, as well as into treatment decision making with patients and providers in the clinic. The incorporation of molecular biomarkers into clinical prediction models and the development of additional biomarkers from blood as opposed to tumor tissue may prove to be critical, as blood can be readily accessible and monitored over time. Candidate molecular biomarkers have been identified in the cellular components of blood, including circulating tumor cells, and will likely be further developed for screening and monitoring of residual disease. While targeted therapies have already revolutionized the field of lung cancer, next steps will involve strategies for managing acquired resistance to targeted therapies, including EGFR and ALK. Newer generation inhibitors and/or combination strategies to inhibit multiple pathways may ultimately prove fruitful in overcoming secondary resistance. Finally, new strategies against previously elusive targets, such as K-RAS, will likely be employed.

Personalization of therapy will involve close collaboration between the laboratory and the clinic as well as interdisciplinary collaboration between surgeons, pathologists, molecular biologists, translational researchers, and medical oncologists. Incorporating the above information in a systematic manner may not be simple, but will be a critical goal as we move forward. In our opinion, tumor board, where we discuss clinical care in a multidisciplinary fashion, should involve radiology findings, pathology findings, and now also biomarker findings. Integrating

Table 5.4 Molecular tumor board

Name	Age	PS	Stage	Histology	EGFR	ALK	K-RAS	ROS-1	MET	Other	Rx
JM	42	0	IIIB	NSCLC		+					crizotinib
RH	67	1	IIIA	SCCA						FGFR1	Phase I
GW	55	0	IV	NSCLC	+						erlotinib
EC	58	1	IV	NSCLC			+				Phase II
HS	63	2	E-S	SCLC							Pt/VP-16

A sample representation of molecular tumor board from the thoracic oncology program at the University of Chicago

PS performance status, *E-S* extensive stage, *EGFR* activating epidermal growth factor receptor mutation, *ALK* anaplastic lymphoma kinase fusion, *ROS-1* reactive oxygen species-1 translocation, *FGFR1* fibroblast growth factor receptor 1 amplification, *Rx* proposed treatment, *Pt/VP-16* platinum-etoposide chemotherapy

these discussions into a formal treatment plan for our patients and into a research platform will not only allow our patients but also allow future generations to benefit from such therapeutic and joint decision making. Tumor board in the era of personalized medicine will highlight and integrate molecular markers into the discussion of optimal patient care (Table 5.4).

5.23 Conclusions

Our historical binary division of lung cancer into non-small-cell lung cancer and small-cell lung cancer has been significantly expanded with the advent of genotyping and genomic profiling.

Advances in whole-genome sequencing and success of high-throughput functional genomics now allow us to supplement our conventional approaches with systems-level approaches that allow researchers to study human biology as systems of interacting genetic and epigenetic factors in relevant biological contexts. While non-small-cell lung cancer, in particular, has seen several of these advances, recent genomic characterizations have been under way in small-cell lung cancer [35, 129] and squamous-cell cancers. With our understanding of activating mutations in EGFR, we have seen a prototype and proof of principle for molecular targeting in lung cancer. At the same time, cancer therapeutics seems to be entering a new era, as traditional cytotoxic systemic chemotherapy is being supplemented with targeted drugs, which relies on specific pathways upregulated in malignancy. With the development of pharmacologic strategies against ALK fusion genes, ROS-1 rearrangements, and ongoing research in targeted therapies for K-RAS and MET, we are poised to change the landscape for cancer care and research in general. Indeed, understanding the genetic architecture underlying a complex biological system and heritable multigene disorder such as lung cancer will be one of the major goals of human genetics in the next decades.

References

1. Siegel R, Naishadham D, Jemal A (2012) Cancer statistics, 2012. CA Cancer J Clin 62(1):10–29
2. Malvezzi M, Bertuccio P, Levi F et al (2013) European cancer mortality predictions for the year 2013. Ann Oncol 24(3):792–800
3. American Cancer Society (2012) Cancer facts & figures. Atlanta, American Cancer Society
4. Altekruse SF, Kosary CL, Krapcho M, et al. SEER Cancer Statistics Review (1975–2007) National Cancer Institute, Bethesda, MD. http://seer.cancer.gov/csr/975_2007/based on Nov 2009 SEER data submission, posted to the SEER website, 2010. Accessed 24 Feb 2013
5. Health Service, Centers for Disease Control and Prevention, Washington, DC, 2010. Smoking and tobacco use – fact sheets. http://www.cdc.gov/tobacco/data_statistics/fact_sheets/. Accessed 24 Feb 2013
6. Schroeder SA (2013) New evidence that cigarette smoking remains the most important health hazard. N Engl J Med 368(4):389–390
7. Alberg AJ, Nonemaker J (2008) Who is at high risk for lung cancer? Population-level and individual-level perspectives. Semin Respir Crit Care Med 29(3):223–232
8. Godtfredsen NS, Prescott E, Osler M (2005) Effect of smoking reduction on lung cancer risk. JAMA 294(12):1505–1510
9. Bruske-Hohlfeld I (2009) Environmental and occupational risk factors for lung cancer. Methods Mol Biol 472:3–23
10. Sim MR (2013) A worldwide ban on asbestos production and use: some recent progress, but more still to be done. Occup Environ Med 70(1):1–2
11. Matakidou A, Eisen T, Houlston RS (2005) Systematic review of the relationship between family history and lung cancer risk. Br J Cancer 93(7):825–833
12. Ihsan R, Chauhan PS, Mishra AK et al (2011) Multiple analytical approaches reveal distinct gene-environment interactions in smokers and non smokers in lung cancer. PLoS One 6(12):e29431
13. Tockman MS (1986) Survival and mortality from lung cancer in a screened population. Chest 89(4 suppl):324S–325S
14. Tockman MS, Mulshine JL (1997) Sputum screening by quantitative microscopy: a new dawn for detection of lung cancer? Mayo Clin Proc 72(8):788–790
15. Prorok PC, Andriole GL, Bresalier RS et al (2000) Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. Control Clin Trials 21(6 Suppl):273S–309S
16. Bach PB, Mirkin JN, Oliver TK et al (2012) Benefits and harms of CT screening for lung cancer: a systematic review. JAMA 307(22):2418–2429
17. Aberle DR, Adams AM, Berg CD, et al, for the National Lung Screening Trial Research Team (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 365(5):395–409
18. Dancey JE, Dobbin KK, Groshen S et al (2010) Guidelines for the development and incorporation of biomarker studies in early clinical trials of novel agents. Clin Cancer Res 16(6):1745–1755
19. Johnson DH, Fehrenbacher L, Novotny WF et al (2004) Randomized phase II trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non-small-cell lung cancer. J Clin Oncol 22(11):2184–2191
20. Sandler A, Gray R, Perry MC et al (2006) Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. N Engl J Med 355(24):2542–2550
21. Scagliotti GV, Parikh P, von Pawel J et al (2008) Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. J Clin Oncol 26(21):3543–3551
22. Ciuleanu T, Brodowicz T, Zielinski C et al (2009) Maintenance pemetrexed plus best supportive care versus placebo plus best supportive care for non-small-cell lung cancer: a randomised, double-blind, phase 3 study. Lancet 374(9699):1432–1440

23. Giovannetti E, Mey V, Nannizzi S et al (2005) Cellular and pharmacogenetics foundation of synergistic interaction of pemetrexed and gemcitabine in human non-small-cell lung cancer cells. Mol Pharmacol 68(1):110–118
24. Maziak DE, Darling GE, Inculet RI et al (2009) Positron emission tomography in staging early lung cancer: a randomized trial. Ann Intern Med 151(4):221–228, W-48
25. Luke WP, Pearson FG, Todd TR et al (1986) Prospective evaluation of mediastinoscopy for assessment of carcinoma of the lung. J Thorac Cardiovasc Surg 91(1):53–56
26. Vilmann P, Krasnik M, Larsen SS et al (2005) Transesophageal endoscopic ultrasound-guided fine-needle aspiration (EUS-FNA) and endobronchial ultrasound-guided transbronchial needle aspiration (EBUS-TBNA) biopsy: a combined approach in the evaluation of mediastinal lesions. Endoscopy 37(9):833–839
27. Goldstraw P (2009) The 7th edition of TNM in lung cancer: what now? J Thorac Oncol 4(6):671–673
28. Groome PA, Bolejack V, Crowley JJ et al (2007) The IASLC Lung Cancer Staging Project: validation of the proposals for revision of the T, N, and M descriptors and consequent stage groupings in the forthcoming (seventh) edition of the TNM classification of malignant tumours. J Thorac Oncol 2(8):694–705
29. Doddoli C, D'Journo B, Le Pimpec-Barthes F et al (2005) Lung cancer invading the chest wall: a plea for en-bloc resection but the need for new treatment strategies. Ann Thorac Surg 80(6):2032–2040
30. Yan TD, Black D, Bannon PG et al (2009) Systematic review and meta-analysis of randomized and nonrandomized trials on safety and efficacy of video-assisted thoracic surgery lobectomy for early-stage non-small-cell lung cancer. J Clin Oncol 27(15):2553–2562
31. Petersen RP, Pham D, Burfeind WR et al (2007) Thoracoscopic lobectomy facilitates the delivery of chemotherapy after resection for lung cancer. Ann Thorac Surg 83(4):1245–1249, discussion 50
32. Ferguson MK, Lehman AG (2003) Sleeve lobectomy or pneumonectomy: optimal management strategy using decision analysis techniques. Ann Thorac Surg 76(6):1782–1788
33. Strauss GM, Herndon JE 2nd, Maddaus MA et al (2008) Adjuvant paclitaxel plus carboplatin compared with observation in stage IB non-small-cell lung cancer: CALGB 9633 with the Cancer and Leukemia Group B, Radiation Therapy Oncology Group, and North Central Cancer Treatment Group Study Groups. J Clin Oncol 26(31):5043–5051
34. Heon S, Johnson BE (2012) Adjuvant chemotherapy for surgically resected non-small cell lung cancer. J Thorac Cardiovasc Surg 144(3):S39–S42
35. Douillard JY, Rosell R, De Lena M et al (2006) Adjuvant vinorelbine plus cisplatin versus observation in patients with completely resected stage IB-IIIA non-small-cell lung cancer (Adjuvant Navelbine International Trialist Association [ANITA]): a randomised controlled trial. Lancet Oncol 7(9):719–727
36. Pisters KM, Evans WK, Azzoli CG et al (2007) Cancer Care Ontario and American Society of Clinical Oncology adjuvant chemotherapy and adjuvant radiation therapy for stages I-IIIA resectable non small-cell lung cancer guideline. J Clin Oncol 25(34):5506–5518
37. Douillard JY, Rosell R, De Lena M et al (2008) Impact of postoperative radiation therapy on survival in patients with complete resection and stage I, II, or IIIA non-small-cell lung cancer treated with adjuvant chemotherapy: the adjuvant Navelbine International Trialist Association (ANITA) Randomized Trial. Int J Radiat Oncol Biol Phys 72(3):695–701
38. Curran WJ Jr, Paulus R, Langer CJ et al (2011) Sequential vs. concurrent chemoradiation for stage III non-small cell lung cancer: randomized phase III trial RTOG 9410. J Natl Cancer Inst 103(19):1452–1460
39. Pao W, Hutchinson KE (2012) Chipping away at the lung cancer genome. Nat Med 18(3):349–351
40. Kris MG, Johnson BE, Kwiatkowski DJ et al (2011) Identification of driver mutations in tumor specimens from 1,000 patients with lung adenocarcinoma: The NCI's Lung Cancer Mutation Consortium (LCMC). J Clin Oncol 29:477s, suppl 15; abstr CRA 7506

41. Vignot S, Frampton GM, Soria JC et al (2013). Next-generation sequencing reveals high concordance of recurrent somatic alterations between primary tumor and metastases from patients with non-small-cell lung cancer. J Clin Oncol. Epub ahead of print.
42. Herbst RS, Heymach JV, Lippman SM (2008) Lung cancer. N Engl J Med 359(13):1367–1380
43. Herbst RS, Maddox AM, Rothenberg ML et al (2002) Selective oral epidermal growth factor receptor tyrosine kinase inhibitor ZD1839 is generally well-tolerated and has activity in non-small-cell lung cancer and other solid tumors: results of a phase I trial. J Clin Oncol 20(18):3815–3825
44. Fukuoka M, Yano S, Giaccone G et al (2003) Multi-institutional randomized phase II trial of gefitinib for previously treated patients with advanced non-small-cell lung cancer (The IDEAL 1 Trial) [corrected]. J Clin Oncol 21(12):2237–2246
45. Lynch TJ, Bell DW, Sordella R et al (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med 350(21):2129–2139
46. Paez JG, Janne PA, Lee JC et al (2004) EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 304(5676):1497–1500
47. Marks JL, Broderick S, Zhou Q et al (2008) Prognostic and therapeutic implications of EGFR and KRAS mutations in resected lung adenocarcinoma. J Thorac Oncol 3(2):111–116
48. Pao W, Chmielecki J (2010) Rational, biologically based treatment of EGFR-mutant non-small-cell lung cancer. Nat Rev Cancer 10(11):760–774
49. Mok TS, Wu YL, Thongprasert S et al (2009) Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N Engl J Med 361(10):947–957
50. Maemondo M, Inoue A, Kobayashi K et al (2010) Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. N Engl J Med 362(25):2380–2388
51. Mitsudomi T, Morita S, Yatabe Y et al (2010) Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial. Lancet Oncol 11(2):121–128
52. Rosell R, Moran T, Queralt C et al (2009) Screening for epidermal growth factor receptor mutations in lung cancer. N Engl J Med 361(10):958–967
53. Sequist LV, Waltman BA, Dias-Santagata D et al (2011) Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors. Sci Transl Med 3(75):75ra26
54. Soh J, Okumura N, Lockwood WW et al (2009) Oncogene mutations, copy number gains and mutant allele specific imbalance (MASI) frequently occur together in tumor cells. PLoS One 4(10):e7464
55. Ding L, Getz G, Wheeler DA et al (2008) Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455(7216):1069–1075
56. Yun CH, Boggon TJ, Li Y et al (2007) Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. Cancer Cell 11(3):217–227
57. Carey KD, Garton AJ, Romero MS et al (2006) Kinetic analysis of epidermal growth factor receptor somatic mutant proteins shows increased sensitivity to the epidermal growth factor receptor tyrosine kinase inhibitor, erlotinib. Cancer Res 66(16):8163–8171
58. Gong Y, Somwar R, Politi K et al (2007) Induction of BIM is essential for apoptosis triggered by EGFR kinase inhibitors in mutant EGFR-dependent lung adenocarcinomas. PLoS Med 4(10):e294
59. Deng J, Shimamura T, Perera S et al (2007) Proapoptotic BH3-only BCL-2 family protein BIM connects death signaling from epidermal growth factor receptor inhibition to the mitochondrion. Cancer Res 67(24):11867–11875
60. Greulich H, Chen TH, Feng W et al (2005) Oncogenic transformation by inhibitor-sensitive and -resistant EGFR mutants. PLoS Med 2(11):e313
61. Prudkin L, Tang X, Wistuba II (2009) Germ-line and somatic presentations of the EGFR T790M mutation in lung cancer. J Thorac Oncol 4(1):139–141

62. Pao W, Wang TY, Riely GJ et al (2005) KRAS mutations and primary resistance of lung adenocarcinomas to gefitinib or erlotinib. PLoS Med 2(1):e17
63. Roberts PJ, Stinchcombe TE (2013) KRAS mutation: should we test for it, and does it matter? J Clin Oncol 31(8):1112–1121
64. Sos ML, Koker M, Weir BA et al (2009) PTEN loss contributes to erlotinib resistance in EGFR-mutant lung cancer by activation of Akt and EGFR. Cancer Res 69(8):3256–3261
65. Yano S, Wang W, Li Q et al (2008) Hepatocyte growth factor induces gefitinib resistance of lung adenocarcinoma with epidermal growth factor receptor-activating mutations. Cancer Res 68(22):9479–9487
66. Jackman D, Pao W, Riely GJ et al (2010) Clinical definition of acquired resistance to epidermal growth factor receptor tyrosine kinase inhibitors in non-small-cell lung cancer. J Clin Oncol 28(2):357–360
67. Yun CH, Mengwasser KE, Toms AV et al (2008) The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. Proc Natl Acad Sci U S A 105(6):2070–2075
68. Oxnard GR, Arcila ME, Sima CS et al (2011) Acquired resistance to EGFR tyrosine kinase inhibitors in EGFR-mutant lung cancer: distinct natural history of patients with tumors harboring the T790M mutation. Clin Cancer Res 17(6):1616–1622
69. Bean J, Riely GJ, Balak M et al (2008) Acquired resistance to epidermal growth factor receptor kinase inhibitors associated with a novel T854A mutation in a patient with EGFR-mutant lung adenocarcinoma. Clin Cancer Res 14(22):7519–7525
70. Bean J, Brennan C, Shih JY et al (2007) MET amplification occurs with or without T790M mutations in EGFR mutant lung tumors with acquired resistance to gefitinib or erlotinib. Proc Natl Acad Sci U S A 104(52):20932–20937
71. Engelman JA, Zejnullahu K, Mitsudomi T et al (2007) MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science 316(5827):1039–1043
72. Wilson TR, Fridlyand J, Yan Y et al (2012) Widespread potential for growth-factor-driven resistance to anticancer kinase inhibitors. Nature 487(7408):505–509
73. Jagadeeswaran R, Surawska H, Krishnaswamy S et al (2008) Paxillin is a target for somatic mutations in lung cancer: implications for cell growth and invasion. Cancer Res 68(1):132–142
74. Straussman R, Morikawa T, Shee K et al (2012) Tumour micro-environment elicits innate resistance to RAF inhibitors through HGF secretion. Nature 487(7408):500–504
75. Li D, Ambrogio L, Shimamura T et al (2008) BIBW2992, an irreversible EGFR/HER2 inhibitor highly effective in preclinical lung cancer models. Oncogene 27(34):4702–4711
76. Xu L, Kikuchi E, Xu C et al (2012) Combined EGFR/MET or EGFR/HSP90 inhibition is effective in the treatment of lung cancers codriven by mutant EGFR containing T790M and MET. Cancer Res 72(13):3302–3311
77. Yang JC-H, Schuler MH, Yamamoto N et al LUX-Lung 3: a randomized, open-label, phase III study of afatinib versus pemetrexed and cisplatin as first-line treatment for patients with advanced adenocarcinoma of the lung harboring EGFR-activating mutations. 2012 ASCO annual meeting. Abstract LBA7500. Presented 4 June 2012.
78. Li D, Shimamura T, Ji H et al (2007) Bronchial and peripheral murine lung carcinomas induced by T790M-L858R mutant EGFR respond to HKI-272 and rapamycin combination therapy. Cancer Cell 12(1):81–93
79. Chaft JE, Oxnard GR, Sima CS et al (2011) Disease flare after tyrosine kinase inhibitor discontinuation in patients with EGFR-mutant lung cancer and acquired resistance to erlotinib or gefitinib: implications for clinical trial design. Clin Cancer Res 17(19):6298–6303
80. Turke AB, Zejnullahu K, Wu YL et al (2010) Preexistence and clonal selection of MET amplification in EGFR mutant NSCLC. Cancer Cell 17(1):77–88
81. Morris SW, Kirstein MN, Valentine MB et al (1995) Fusion of a kinase gene, ALK, to a nucleolar protein gene, NPM, in non-Hodgkin's lymphoma. Science 267(5196):316–317
82. Soda M, Choi YL, Enomoto M et al (2007) Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 448(7153):561–566

83. Horn L, Pao W (2009) EML4-ALK: honing in on a new target in non-small-cell lung cancer. J Clin Oncol 27(26):4232–4235
84. Shaw AT, Yeap BY, Mino-Kenudson M et al (2009) Clinical features and outcome of patients with non-small-cell lung cancer who harbor EML4-ALK. J Clin Oncol 27(26):4247–4253
85. Koivunen JP, Mermel C, Zejnullahu K et al (2008) EML4-ALK fusion gene and efficacy of an ALK kinase inhibitor in lung cancer. Clin Cancer Res 14(13):4275–4283
86. Camidge DR, Bang Y, Kwak EL et al (2011) Progression-free survival (PFS) from a phase 1 study of crizotinib (PF-02341066) in patients with ALK-positive non-small cell lung cancer (NSCLC). J Clin Oncol 29(15 suppl):2501
87. Choi YL, Soda M, Yamashita Y et al (2010) EML4-ALK mutations in lung cancer that confer resistance to ALK inhibitors. N Engl J Med 363(18):1734–1739
88. Rikova K, Guo A, Zeng Q et al (2007) Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell 131(6):1190–1203
89. Bergethon K, Shaw AT, Ou SH et al (2012) ROS1 rearrangements define a unique molecular class of lung cancers. J Clin Oncol 30(8):863–870
90. McDermott U, Iafrate AJ, Gray NS et al (2008) Genomic alterations of anaplastic lymphoma kinase may sensitize tumors to anaplastic lymphoma kinase inhibitors. Cancer Res 68(9):3389–3395
91. Mascaux C, Iannino N, Martin B et al (2005) The role of RAS oncogene in survival of patients with lung cancer: a systematic review of the literature with meta-analysis. Br J Cancer 92(1):131–139
92. Roberts PJ, Der CJ (2007) Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. Oncogene 26(22):3291–3310
93. Slebos RJ, Kibbelaar RE, Dalesio O et al (1990) K-ras oncogene activation as a prognostic marker in adenocarcinoma of the lung. N Engl J Med 323(9):561–565
94. Riely GJ, Kris MG, Rosenbaum D et al (2008) Frequency and distinctive spectrum of KRAS mutations in never smokers with lung adenocarcinoma. Clin Cancer Res 14(18):5731–5734
95. Graziano SL, Gamble GP, Newman NB et al (1999) Prognostic significance of K-ras codon 12 mutations in patients with resected stage I and II non-small-cell lung cancer. J Clin Oncol 17(2):668–675
96. Mao C, Qiu LX, Liao RY et al (2010) KRAS mutations and resistance to EGFR-TKIs treatment in patients with non-small cell lung cancer: a meta-analysis of 22 studies. Lung Cancer 69(3):272–278
97. Linardou H, Dahabreh IJ, Kanaloupiti D et al (2008) Assessment of somatic k-RAS mutations as a mechanism associated with resistance to EGFR-targeted agents: a systematic review and meta-analysis of studies in advanced non-small-cell lung cancer and metastatic colorectal cancer. Lancet Oncol 9(10):962–972
98. Chen Z, Cheng K, Walton Z et al (2012) A murine lung cancer co-clinical trial identifies genetic modifiers of therapeutic response. Nature 483(7391):613–617
99. Engelman JA, Chen L, Tan X et al (2008) Effective use of PI3K and MEK inhibitors to treat mutant Kras G12D and PIK3CA H1047R murine lung cancers. Nat Med 14(12):1351–1356
100. Sos ML, Michel K, Zander T et al (2009) Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions. J Clin Invest 119(6):1727–1740
101. Corcoran RB, Cheng KA, Hata AN et al (2013) Synthetic lethal interaction of combined BCL-XL and MEK inhibition promotes tumor regressions in KRAS mutant cancer models. Cancer Cell 23(1):121–128
102. Puyol M, Martin A, Dubus P et al (2010) A synthetic lethal interaction between K-Ras oncogenes and Cdk4 unveils a therapeutic strategy for non-small cell lung carcinoma. Cancer Cell 18(1):63–73
103. Vicent S, Chen R, Sayles LC et al (2010) Wilms tumor 1 (WT1) regulates KRAS-driven oncogenesis and senescence in mouse and human models. J Clin Invest 120(11):3940–3952
104. Meylan E, Dooley AL, Feldser DM et al (2009) Requirement for NF-kappaB signalling in a mouse model of lung adenocarcinoma. Nature 462(7269):104–107

105. Barbie DA, Tamayo P, Boehm JS et al (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 462(7269):108–112
106. Huqun, Ishikawa R, Zhang J et al (2012) Enhancer of zeste homolog 2 is a novel prognostic biomarker in nonsmall cell lung cancer. Cancer 118(6):1599–1606
107. Ahmed AU, Schmidt RL, Park CH et al (2008) Effect of disrupting seven-in-absentia homolog 2 function on lung cancer cell growth. J Natl Cancer Inst 100(22):1606–1629
108. Kumar MS, Hancock DC, Molina-Arcas M et al (2012) The GATA2 transcriptional network is requisite for RAS oncogene-driven non-small cell lung cancer. Cell 149(3):642–655
109. Sunaga N, Imai H, Shimizu K et al (2012) Oncogenic KRAS-induced interleukin-8 overexpression promotes cell growth and migration and contributes to aggressive phenotypes of non-small cell lung cancer. Int J Cancer 130(8):1733–1744
110. Seo JS, Ju YS, Lee WC et al (2012) The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res 22(11):2109–2119
111. Ma PC, Jagadeeswaran R, Jagadeesh S et al (2005) Functional expression and mutations of c-Met and its therapeutic inhibition with SU11274 and small interfering RNA in non-small cell lung cancer. Cancer Res 65(4):1479–1488
112. Cappuzzo F, Marchetti A, Skokan M et al (2009) Increased MET gene copy number negatively affects survival of surgically resected non-small-cell lung cancer patients. J Clin Oncol 27(10):1667–1674
113. Peruzzi B, Bottaro DP (2006) Targeting the c-Met signaling pathway in cancer. Clin Cancer Res 12(12):3657–3660
114. Birchmeier C, Birchmeier W, Gherardi E et al (2003) Met, metastasis, motility and more. Nat Rev Mol Cell Biol 4(12):915–925
115. Sulpice E, Ding S, Muscatelli-Groux B et al (2009) Cross-talk between the VEGF-A and HGF signalling pathways in endothelial cells. Biol Cell (Under the auspices of the European Cell Biology Organization) 101(9):525–539
116. Spigel DR, Ervin TJ, Ramlau R et al. Final efficacy results from OAM4558g, a randomized phase II study evaluating MetMAb or placebo in combination with erlotinib in advanced NSCLC. J Clin Oncol 29: 2011 (suppl; abstr 7505)
117. Schiller JH, Akerley WL, Brugger W et al (2010) Results from ARQ 197–209: A global randomized placebo-controlled phase II clinical trial of erlotinib plus ARQ 197 versus erlotinib plus placebo in previously treated EGFR inhibitor-naive patients with locally advanced or metastatic non-small cell lung cancer (NSCLC). J Clin Oncol 28:18s (suppl; abstr LBA7502)
118. Burke A. Foundation medicine: personalizing cancer drugs. MIT Technol Rev. http://www.technologyreview.com/. Accessed 21 Feb 2012
119. Heger M. Caris Adds Next-Gen Sequencing to portfolio of molecular tumor profiling technologies. Clinical sequencing news. http://www.genomeweb.com/. Accessed 06 Feb 2013
120. Roth JA, Carlson JJ (2011) Prognostic role of ERCC1 in advanced non-small-cell lung cancer: a systematic review and meta-analysis. Clin Lung Cancer 12(6):393–401
121. Reynolds C, Obasaju C, Schell MJ et al (2009) Randomized phase III trial of gemcitabine-based chemotherapy with in situ RRM1 and ERCC1 protein levels for response prediction in non-small-cell lung cancer. J Clin Oncol 27(34):5808–5815
122. Friboulet L, Olaussen KA, Pignon JP et al (2013) ERCC1 isoform expression and DNA repair in non-small-cell lung cancer. N Engl J Med 368(12):1101–1110
123. Li T, Kung HJ, Mack PC et al (2013) Genotyping and genomic profiling of non-small-cell lung cancer: implications for current and future therapies. J Clin Oncol 31(8):1039–1049
124. Thomas RK, Baker AC, Debiasi RM et al (2007) High-throughput oncogene mutation profiling in human cancer. Nat Genet 39(3):347–351
125. Su Z, Dias-Santagata D, Duke M et al (2011) A platform for rapid detection of multiple oncogenic mutations with relevance to targeted therapy in non-small-cell lung cancer. J Mol Diagn 13(1):74–84
126. Sequist LV, Heist RS, Shaw AT et al (2011) Implementing multiplexed genotyping of non-small-cell lung cancers into routine clinical practice. Ann Oncol 22(12):2616–2624

127. Govindan R, Ding L, Griffith M et al (2012) Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell 150(6):1121–1134
128. Imielinski M, Berger AH, Hammerman PS et al (2012) Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell 150(6):1107–1120
129. Peifer M, Fernandez-Cuesta L, Sos ML et al (2012) Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nat Genet 44(10):1104–1110
130. Rudin CM, Durinck S, Stawiski EW et al (2012) Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat Genet 44(10): 1111–1116
131. Cancer Genome Atlas Research Network (2012) Comprehensive genomic characterization of squamous cell lung cancers. Nature 489(7417):519–525
132. Arcila ME, Oxnard GR, Nafa K et al (2011) Rebiopsy of lung cancer patients with acquired resistance to EGFR inhibitors and enhanced detection of the T790M mutation using a locked nucleic acid-based assay. Clin Cancer Res 17(5):1169–1180
133. Oxnard GR, Arcila ME, Chmielecki J et al (2011) New strategies in overcoming acquired resistance to epidermal growth factor receptor tyrosine kinase inhibitors in lung cancer. Clin Cancer Res 17(17):5530–5537

Index

N. Maltsev et al. (eds.), *Systems Analysis of Human Multigene Disorders*, Advances in Experimental Medicine and Biology 799, DOI 10.1007/978-1-4614-8778-4,