Annu. Rev. Genomics Hum. Genet. 2005. 6:381–406
doi: 10.1146/annurev.genom.6.080604.162234

DISEASE GENE DISCOVERY THROUGH INTEGRATIVE GENOMICS

Cosmas Giallourakis,[1,2] Charlotte Henson,[1] Michael Reich,[1] Xiaohui Xie,[1] Vamsi K. Mootha[1,3,4]
[1]*Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02139;*
[2]*Gastrointestinal Unit, Massachusetts General Hospital, Boston, Massachusetts 02114;*
[3]*Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02446;*
[4]*Center for Human Genetics Research, Massachusetts General Hospital, Boston, Massachusetts 02114; email: vamsi@hms.harvard.edu*

Key Words human genetics, positional cloning, functional genomics, machine learning

■ **Abstract** The availability of complete genome sequences and the wealth of large-scale biological data sets now provide an unprecedented opportunity to elucidate the genetic basis of rare and common human diseases. Here we review some of the emerging genomics technologies and data resources that can be used to infer gene function to prioritize candidate genes. We then describe some computational strategies for integrating these large-scale data sets to provide more faithful descriptions of gene function, and how such approaches have recently been applied to discover genes underlying Mendelian disorders. Finally, we discuss future prospects and challenges for using integrative genomics to systematically discover not only single genes but also entire gene networks that underlie and modify human disease.

INTRODUCTION

Elucidating the inherited basis of human disease fundamentally involves linking genomic variation to clinical phenotype. Establishing this relationship, however, can be challenging for several reasons. First, many disease phenotypes are difficult to ascertain, may be heterogeneous, and can be influenced by environmental factors. Second, current genotyping technologies do not permit routine, comprehensive characterization of genomic variation in a large cohort of cases and controls; hence, it is still necessary to focus on variation within high-priority regions of the genome, such as protein-encoding genes. Finally, even when phenotype and a genotype are ascertained in a comprehensive and reliable manner, establishing reliable linkage or association may be statistically challenging, due to the limited number of cases, limited recombination resolution, or admixture.

Despite these challenges, human genetics has been extremely successful, especially for Mendelian diseases, during the past 15 years. The Online Mendelian

Inheritance in Man (OMIM) website lists a total of 1655 inherited human diseases for which genes have been identified, as well as an additional 1436 inherited diseases for which an underlying genetic basis has not yet been discovered (OMIM statistics, November 30, 2004). Much of this success can be attributed to the availability of genetic tools, initially genetic maps and more recently the sequence of the entire human genome (59, 111). Botstein & Risch (14) suggest that the disease genes discovered to date likely represent the easy ones, and that discovering the genetic basis of the remaining Mendelian and complex disorders will be more challenging, perhaps due to the rarity of the phenotypes, due to genetic heterogeneity, or because of complex genetics, i.e., multiple genes and modifiers contributing to a phenotype.

Fortunately, genomics has sparked the creation of vast new functional clues about genes and genomic elements that can aid in our search for human disease genes. New technologies, such as microarrays and tandem mass spectrometry, now enable genome-scale monitoring of RNA, protein, and metabolite abundance, under baseline and perturbed states. Complete genome sequences are available for a variety of organisms, facilitating the annotation of gene structures and regulatory elements. Embedded within these vast databases of information are correlations that weave together genes and genomic elements into functional networks. These networks include well-characterized genes (including the ~1500 genes previously linked to human disease) as well as the vast majority of the genes and genomic elements about which very little is known (Figure 1).

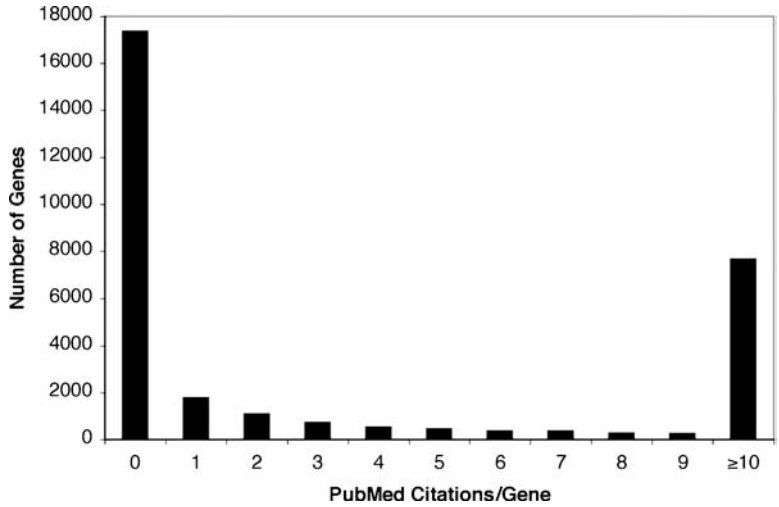

Figure 1 Distribution of literature citations per protein-encoding gene in the human genome. Shown on the x-axis is the number of PubMed citations/gene, and on the y-axis is the number of genes with that number of citations. Data were generated October 19, 2004.

These data sets can be mined to systematically prioritize genes that can be tested individually or collectively for variation in human diseases. Moreover, analyzing these large-scale data sets may help shed insight into disease mechanisms for genes implicated by association studies.

In this article, we review recent progress in utilizing and integrating functional genomic data sources (i.e., integrative genomics) to expedite human disease gene discovery. We begin with a brief overview of the traditional approaches for disease gene discovery. Next we review the wide array of genomics technologies and data sets now becoming available and how they are being used individually to aid in our search for candidate disease genes. Then we discuss practical approaches for integrating these data sets to boost sensitivity and specificity to construct more faithful functional relationships among genes. Finally we discuss future opportunities and challenges for disease gene discovery through integrated analysis of genome-scale information.

TRADITIONAL APPROACHES FOR DISEASE GENE DISCOVERY

Two approaches have traditionally been used to discover genes underlying human diseases: the candidate gene approach and positional cloning via linkage analysis.

The candidate approach relies on prior biochemical knowledge about the disease genes, such as tissues in which they are expressed or putative functional protein domains. Genes are prioritized using these clues and sequentially tested in association studies for segregating mutations or polymorphisms. Genes underlying retinitis pigmentosa (27, 30), familial hypertrophic cardiomyopathy (35), and Li-Fraumeni syndrome (65) were all discovered in this manner.

Positional cloning, on the other hand, does not formally require prior knowledge about gene function. Traditionally these studies are performed in large families with multiple affected members using microsatellite markers and other DNA polymorphisms. Alleles of markers that segregate with the disease help delineate a critical region within which the disease gene lies. This method has been quite effective for mapping the genetic variation underlying Mendelian disease, even though the nature of positional cloning limits its resolution to relatively large regions of the genome. Given the spacing of markers and the observed number of meioses, the resolution limit is on the order of 1–10 centiMorgans (cM). In most favorable cases the critical interval consists of a few dozen genes within 1 cM, but in other cases the interval may include several hundred genes. Researchers must then sift through the candidate genes within this critical region to identify mutations in genes that segregate with the disease.

It is useful to consider the search for the cystic fibrosis (CF) gene—a positional cloning expedition that occupied several labs for many years—and how the search might be performed differently today. In 1985, Lap-Chee Tsui and colleagues (106) used linkage analysis to map the disease to chromosome 7. By testing

additional markers, they mapped the disorder to a 1.5-Mb interval flanked by the protooncogene *MET* and the marker *D7S8* (107). Researchers used chromosome jumping and cloning in yeast artificial chromosomes to further delineate the interval and relied on other clues—evolutionary conservation, presence of an mRNA transcript, hypomethylated CpG islands—to ascertain gene structures. Simultaneously, clinical researchers discovered that CF patients exhibit defects in chloride transport. By 1990 mutations in the *CFTR* gene were identified, and researchers demonstrated that when the fully cloned gene was reintroduced into CF cells, defective ion transport could be rescued (26, 82). Together, these studies established *CFTR* as the gene underlying CF.

How might this search be different if it occurred today? First, the availability of the genome sequence offers a virtually unlimited source of markers for positional cloning (of course, many of these may be linked), thus assigning the disease to a narrower genomic locus. After mapping CF to the interval flanked by *MET* and *D7S8*, we could (in a single afternoon) examine the human genome browser (11, 56) and rapidly identify 14 known and predicted gene structures within the 1.6-Mb interval. We could then ask which genes are functionally associated with "clues" provided by the disease. For example, it was known that CF was likely due to a defect in ion channel activity and that the pancreas, lungs, and glands are particularly affected. Today, we can immediately determine that three (*MET*, *ST7*, *CFTR*) of the 14 genes within this interval encode transmembrane proteins (92). Of these 14 genes, *ST7* shows nearly ubiquitous expression whereas *MET* exhibits limited expression in bronchoepithelial cells. *CFTR*, on the other hand, shows enriched expression in fetal lung, pancreas, and salivary glands (http://symatlas.gnf.org/SymAtlas/), precisely the tissues most affected in CF. Hence, *CFTR* emerges as an attractive candidate by joining these publicly available data sets. Of course, we would still have to sequence the gene in patients and controls and demonstrate segregating mutations as well as additional functional support, but this example illustrates how rapidly we can prioritize candidate genes with freely available functional genomics data.

In the next few sections, we review new genomic resources that are now becoming available and how they are being used in clever ways to discover genes underlying human diseases.

HUMAN GENOME SEQUENCE AND ITS FUNCTIONAL ELEMENTS

A draft sequence of the human genome was published in 2001 (59, 111) and more recently in completed form (48a), representing the most valuable resource for disease gene discovery. An international effort is currently underway to systematically catalog common variation across selected populations of humans, which promises to expedite the mapping of human phenotypic traits by providing a virtually unlimited collection of markers (83).

Analysis of the human genome has revealed that of the 2.85 billion bases in the genome, only 1.2% of the sequence encodes the estimated 22,500 proteins. However, comparative sequence analysis suggests that about 5% of the genome is under evolutionary selection based on human-mouse comparisons, and thus is likely to be functionally important (113). Hence, in addition to protein-coding exons, there are a vast number of "features" present in the genome's landscape. In principle, these additional, conserved regions represent functional elements that may represent high-quality candidate disease genes.

One subset of the conserved elements encodes an estimated 200–400 microRNAs (10) that help regulate the expression of thousands of human genes (54, 61, 115). MicroRNAs are evolutionarily conserved genes whose transcripts are processed to form short, single-stranded 21–23 nucleotide RNA species that typically bind to the untranslated regions (UTRs) of genes to cooperate with a set of proteins to either halt translation or promote RNA cleavage/degradation (8). Another subset of conserved elements encodes thousands of antisense transcripts, which are developmentally regulated and expressed in a tissue-specific manner to regulate target genes (13, 57).

While the above elements are transcribed, another large fraction of conserved features represent putative regulatory elements. Such features include *cis*-elements that control expression of individual or small groups of transcripts, such as promoters, enhancers, and insulators, or structural elements such as locus control regions and matrix attachment sites that may control the architecture of large chromosomal territories (75, 85, 101). Comparative sequence analysis has helped in the discovery and annotation of hundreds of such regulatory elements that are enriched upstream or downstream of functionally related genes (9, 115), and elegant experimental approaches are being developed to elucidate their roles (64).

Although theoretically any nucleotide in the genome can contribute to human diseases, in the near future we will still have to prioritize segments of the genome. In addition to protein-encoding genes, these additional classes of functional elements naturally expand the inventory of candidate genomic elements that ought to be prioritized in disease gene expeditions. In the next few sections, we discuss some experimental and computational approaches for collectively understanding the function of these genomic elements.

INFERRING FUNCTION THROUGH GENOME-SCALE DATA SETS

Having a handle on the function of a gene enables researchers to assess its candidacy in an inherited disease. Currently, only a small fraction (~25%) of all protein encoding genes are well characterized using traditional approaches (Figure 1). Historically, candidate gene approaches for rare and common diseases have focused on this small fraction of well-characterized genes, and, as stated above, these protein-encoding genes represent only a fraction of all evolutionarily conserved elements.

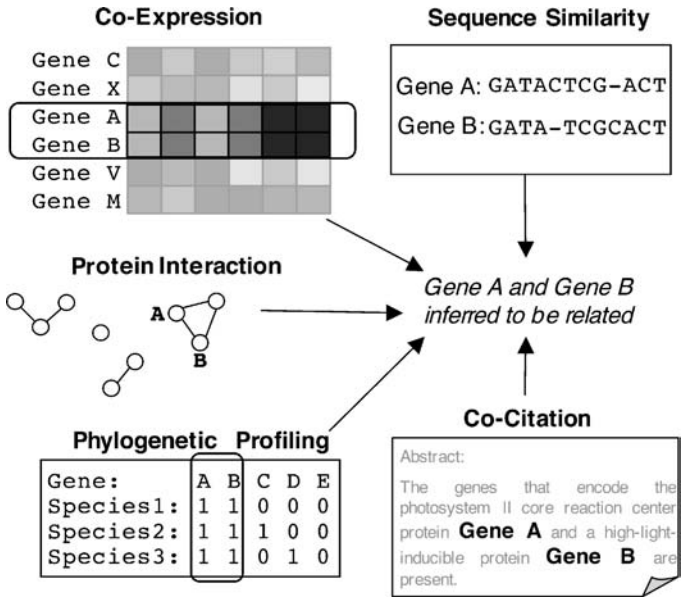

Figure 2 Some of the systematic strategies for inferring gene function.

Genome-scale experiments afford an opportunity to rapidly annotate the other genes in the genome so that they can be considered in such studies. These technologies include DNA microarrays, mass spectrometry-based proteomics, and genome-wide RNAi screens. Thanks to sharing policies enforced by funding agencies and journals, these data sets are being deposited into the public domain. Gene expression profiles of cells in response to radiation (108), proteomic surveys of malaria during developmental stages (31), and genome-wide RNAi screens in worms (32) represent just a few examples of the types of data that shed insight into many of the genes. Such experiments do not provide an in-depth understanding of an individual gene, and they tend to be noisy relative to traditional experiments. But they do provide simultaneous snapshots of all the genes in the genome that collectively can be useful. Simple "guilt by association" strategies can be used to mine these large-scale data sets to infer the function of poorly characterized genes (Figure 2). For example, two genes that share RNA expression profiles, or whose protein products physically interact, may be functionally related. Hence, these large-scale data sets enable us to search for relationships between genes that may not be apparent at the level of sequence. These different functional genomic experiments provide complementary views of gene function and facilitate more reliable grouping of genes based on shared roles in the cell. In the next few sections, we consider some of the functional genomic data sets and analytical strategies already being used to discover human disease genes.

Fully Sequenced Genomes

Sequencing technologies and computational algorithms have matured so that sequencing and assembling entire genomes in a matter of weeks to months is relatively straightforward. Draft genome sequences are currently available or will soon become available for a number of vertebrates, including mouse, rat, dog, chimpanzee, and chicken (37, 45, 113). In addition, hundreds of genome sequences are available from other animals, plants, and fungi.

Phylogenetic profiling is a powerful computational strategy that leverages these completed genome sequences to infer gene function (78, 100). The strategy is based on the assumption that functionally related genes will likely evolve in a correlated fashion, and therefore they are likely to share homologs among organisms. A phylogenetic profile for each gene can be created in the form of a binary vector representing whether the homolog of the gene is present in a set of sequenced organisms. Phylogenetic profiles can then be organized based on similarity.

Recently, two groups published elegant studies (21, 62) in which they integrated genetic linkage intervals with phylogenetic profiles to discover genes underlying Bardet-Biedel syndrome (BBS). BBS is a multisystem disorder characterized by retinal degeneration, obesity, polydactyly, renal and genital malformations, and learning disabilities. Defective basal body function has been implicated in the pathogenesis of this pleiotropic disease. Six genes associated with BBS have been identified (*BBS1*, *BBS2*, *BBS4*, *BBS6/MKKS*, *BBS7*, and *BBS8*), and all encode protein components of the flagellar and basal body (FABB). The two studies discovered additional genes underlying other forms of BBS by beginning with the clinical clue that previously characterized BBS forms are due to defects in FABB.

Li et al. (62) considered *BBS5*, which resides within one of the intervals associated with this syndrome. They compared three genomes—human, *Chlamydomonas*, and *Arabidopsis*—to compile a list of putative FABB components. Specifically, they reasoned that because this apparatus is found in humans and in *Chlamydomonas* but not in plants, proteins contained in the first two genomes but not in the third would serve as a high-quality list enriched in the FABB list. They noted that proteins mutated in five of the previously described six BBS proteins were in this list of 688 proteins, and that 52 of the previously known 58 FABB proteins were in this list, demonstrating the sensitivity of their approach. Finally, they crossed this list with the list of genes residing within the *2q31* genetic interval linked to *BBS5*. This is a large region, with ~230 protein-encoding genes. Only two of these genes intersected with the FABB proteome, and one had a splice site mutation that segregates with the disease in one family, and additional genetic data supported its involvement in *BBS5*.

Chiang et al. (21) investigated the molecular etiology of *BBS3*, which had been linked to a 10-cM interval (containing an estimated 64 genes) in a single Bedouin family. Chiang et al. reasoned that organisms containing orthologues to the known BBS genes likely contain orthologues to as-yet-unidentified BBS genes. They compared human genes with those from 11 fully sequenced metazoan genomes

and identified a total of 1588 genes that shared phylogenetic profiles with the known BBS genes. Four of these genes landed within the 10-cM critical region, one of which harbored a truncation mutation that segregated with the disease phenotype. These studies demonstrate the power of phylogenetic profiling for homing in on candidate disease genes, especially for syndromes and for other Mendelian disorders.

Global Profiles of RNA Expression

Systematic RNA expression profiling represents one of the earliest functional genomics technologies (1, 19, 86, 110). Some of these technologies, such as expressed sequence tags (ESTs) and serial analysis of gene expression (SAGE), enable the discovery and quantitation of expressed genes in a particular tissue or cell type. Other technologies, such as oligonucleotide and cDNA microarrays, enable facile profiling of a predefined set of genes. These technologies have been widely used and have already yielded vast collections of freely available data (see Appendix).

How can these RNA expression resources be used for disease gene discovery? First, some disorders are due to defects in genes that are expressed in a limited number of tissues. Hence, catalogs of tissue-specific expression provide excellent candidate genes. Second, these large-scale data sets can be mined, using coexpression analysis, to infer the function of poorly characterized genes or modules. We consider each of these applications.

TISSUE EXPRESSION ATLASES Some human disease genes are expressed only in tissues exhibiting a pathologic phenotype. Several rich sources of information about tissue-specific expression are currently available.

The dbEST database at NCBI has >4 million ESTs derived from >7000 cDNA libraries representing more than 600 cell types/states. GeneAtlas (www.symatlas.org) is a tissue expression compendium of human, mouse, and rat samples that allows users at the most basic level to view a gene's expression profile across multiple tissue types (96, 97). The Gene Expression of the Nervous System Atlas (GENSAT) is intended to provide a spatiotemporal expression map of 5000 genes in the developing and adult mouse brain. This resource may help spotlight genes expressed in specific neurons that are altered in specific human brain diseases (42). Both of these valuable resources can be exploited in searching for diseases believed to exhibit restricted patterns of tissue expression.

Recently, Katsanis and colleagues computationally mined dbEST to identify clusters that exhibit preferential expression in the retina and integrated the results with retinal disease gene loci (55). This approach identified 88% (22/25) of known retinal disease genes exclusively expressed in the retina. It also yielded positional candidates for 42 mapped but unidentified disease genes. In a complementary approach, Blackshaw and colleagues coupled SAGE data and large-scale in situ hybridization of 1085 transcripts that showed dynamic changes and preferential expression in the retina to provide a valuable resource for mapping retinopathies

(12). These strategies can be applied to numerous other disorders, such as cardiomyopathies, muscular dystrophies, deafness, and others, where diseases are likely to be due to mutations in genes expressed in those tissues.

COEXPRESSION ANALYSIS Atlases of gene expression in combination with coexpression analysis provide valuable insight into the function of poorly characterized genes. Perhaps the richest sources of RNA expression data have come from cDNA and oligonucleotide microarrays (19, 86), which have had numerous applications in classifying cancers (41), elucidating pathogenesis of complex diseases (70), deciphering mechanism of drug action (47), characterizing genomic activity during various cellular processes, such as the cell cycle (22, 93) and response to serum (51), and profiling expression across different tissues (96). Data submission standards have been established and enforced by a variety of journals and funding agencies; hence, a wealth of expression data is now available from the Stanford University Microarray Database (40), Gene Expression Omnibus (GEO) (28), and Array Express (EBI) (15). As a public resource, these expression databases are valuable substrates for coexpression analysis, which can detect gene properties that are subtler than simple tissue-specific expression patterns.

Coexpression analysis attempts to group genes together on the basis of shared expression similarity across a battery of "conditions." Genes that exhibit coexpression likely share the same function (24, 52, 101). A variety of similarity metrics (e.g., Euclidean distance, Pearson correlation coefficient, or Spearman rank correlation coefficient) can be coupled with different clustering algorithms [e.g., hierarchical clustering (29), k-means (101), and self-organizing maps (SOMs) (99)]. These algorithms often have varying strengths and applicability. A commonly shared disadvantage of these algorithms is that they rely on similarity metrics defined over all experimental conditions. Often one would like to organize genes into different modules in which genes share similar expression profiles only among a subset of experimental conditions. Recently, several "bi-clustering" algorithms were developed that attempt to group genes together within the context of a subset of experimental conditions (20, 36, 73).

Another computational strategy seeks to score genes on the basis of their expression similarity not to a single gene, but rather to a set of genes. Our group introduced the "neighborhood analysis" algorithm (71), and a related methodology was developed and applied to *C. elegans* expression studies (77). With these simple strategies, the user defines a gene set corresponding to a pathway or process of interest—it could even represent previously discovered disease genes for a related set of disorders. The algorithms then identify all other genes in the microarray data set that share expression similarity to the gene set.

Tiranti and colleagues (103) applied neighborhood analysis to prioritize the 130 candidate genes located within the locus for ethylmalonic encephalopathy (EE). Given the clinical features of the disease, they hypothesized that the defect was due to mutations in genes related to mitochondrial functioning. Using neighborhood analysis (71), they identified other genes within the interval coexpressed

with well-characterized nuclear-encoded mitochondrial genes. One of these co-expressed genes, *HSCO*, harbors homozygous mutations in all probands from the four consanguineous families that were originally used for the mapping. Nearly all of these mutations were loss-of-function mutations, producing premature stop, frameshift, or aberrant splicing defects, providing definitive genetic proof of *HSCO* involvement in EE.

Proteomics

Proteomics refers to the systematic analysis of proteins, protein complexes, and their interactions (23). The technologies underlying proteomics are less mature than microarray technologies for RNA expression, but they are already providing complementary information that can be useful in studying disease processes. Two types of proteomic data sets that are emerging are catalogs of organelle proteins and genome-wide interaction maps.

ORGANELLE PROTEOMICS To date, proteomic catalogs of proteins residing in the nucleolus, centrosome, nuclear speckles, golgi, splicesome, midbody, lysosome, mitochondria, and nuclear envelope have been generated (2, 3, 7, 72, 81, 84, 87, 91, 114). Comprehensive analysis of subcellular localization in yeast was recently achieved using large-scale epitope tagging (48).

The analysis of cellular substructures provides powerful functional clues about genes as certain protein complexes and cellular organelles can be associated with human diseases. For example, human respiratory chain disorders are often due to defects in the mitochondrion. Cardiomyopathies are often due to mutations in the cardiac myocyte's contractile machinery (88). Recent proteomic surveys of organelles can help expand candidate genes for diseases while also helping us to understand the function of known disease genes.

Autosomal recessive malignant infantile osteopetrosis (ARO) is a genetically heterogeneous disease characterized by a spectrum of phenotypes including severe osteosclerosis, pathologic fractures, hepatosplenomegaly, pancytopenia, and retinal degeneration (102). In many cases, mutations in *TCIRG1* or in *CLCN7*, which encode lysosomal proteins, underlie this disorder (33, 58, 98). Recently, scientists discovered that mutations in a third gene, *OSTM1*, can also result in ARO, and that the mouse orthologue of *OSTM1* is mutated in another osteosclerotic mouse mutant (17). Interestingly, a recent lysosomal proteomic survey (7) assigns OSTM1 to this organelle, demonstrating how dysfunction of three proteins, all located in the same organelle, can conspire in the pathogenesis of ARO. Here we see a striking example of the emerging link between this heterogeneous Mendelian disease and lysosomal biology.

PROTEIN INTERACTION MAPS Several methodologies now exist for high-throughput construction of protein interaction networks based on yeast two-hybrid (Y2H) screening, affinity tag purification coupled with mass spectrometry, directed

peptide libraries, and protein arrays (23). Two large-scale, mass spectrometry-based studies of protein interactions in yeast have been performed to date, each focusing broadly on gene sets involved in signal transduction or genes involved in the DNA damage response (34, 46). Both studies yielded interactions for about 25% of the yeast proteome. In contrast to mass spectrometry-based proteomics, which interrogates protein complexes, Y2H detects pairwise interactions, and has been applied on genome-wide scales to create interaction maps in yeast, *C. elegans*, and *Drosophila* (39, 50, 63, 109).

Several recent studies demonstrate the value of protein interaction maps, even from model organisms, in our search for human disease genes. Syndromes such as xeroderma pigmentosum (XP), Cockayne syndrome (CS), and trichothiodystrophy (TTD)—all of which have overlapping clinical and cellular phenotypes associated with UV DNA damage repair—are associated with mutations in genes encoding components of the TFIIH complex, which is involved in DNA transcription and repair. A rare form of TTD, termed TTD-A, had been identified in three families in whom the TFIIH complex exhibited biochemical instability; however, none of the previously known components were mutated (112). Recently, Ranish et al. (80) applied yeast proteomics to identify proteins in the polymerase II initiation complex, identifying a previously unrecognized tenth member, TFB5. TFB5 is evolutionarily conserved and its orthologue in *Chlamydomonas reinhardtii* is a suppressor of an UV-sensitive mutant (38). Ranish and colleagues discovered mutations in *TFB5* in patients with TTD-A and performed additional functional studies to provide definitive evidence that *TFB5* underlies this disorder (80).

Physiology and Phenomics

An organism's DNA sequence, via RNA, proteins, and metabolites, is ultimately expressed as a context-dependent phenotype—a phenotype could correspond to yeast fitness on selected media or the outcome of a host-pathogen interaction in humans. Several recent high-throughput approaches illustrate the utility of phenotypic screens in prioritizing disease genes. One class of such experiments utilizes deletion strains, whereas another is based on systematic perturbations.

SYSTEMATIC DELETION PROJECTS Several efforts are currently in progress to systematically knock out each gene in a genome. If the resulting phenotype resembles a disease phenotype, the underlying gene may represent a candidate gene.

Ron Davis's group (89) developed systematic deletion strains of yeast that have been used for various functional genomics projects. Steinmetz et al. (94) used 4706 viable deletion strains in a high-throughput assay for mitochondrial respiratory function to link novel genes to mitochondrial biology, yielding high-quality, candidate genes for heritable respiratory chain disorders. They performed a retrospective analysis of known Mendelian mitochondrial disease genes and reported that many of their yeast orthologues, when deleted, exhibit a respiratory petite phenotype. Prospectively, their screen promises to accelerate positional cloning

by providing 11 new disease candidates for mutational screens for 7 putative mitochondrial disease loci.

Perhaps the most interesting use of the deletion strains has been the systematic survey of synthetic lethal interactions (104). Synthetic lethality results when two mutations in two different genes are each viable as single mutations, but lethal when combined in the same haploid genome. This study began with a subset of 132 query genes and generated all pairwise crosses with the ~4700 mutants carrying viable gene deletions. They discovered approximately 1000 synthetic lethal interactions in their sampling. When extrapolated, their investigation implies a tremendous genomic load of epistatic interactions in humans. Such synthetic lethal screens help group genes together based on functional redundancy and provide a complementary approach for annotating gene networks, as well as pairs that could be jointly considered in human disease studies.

A repository of mouse knockout strains, analogous to deletion strains available for yeast, was proposed by the Knockout Mouse Project (5). Currently, approximately 10% (2600) of the mouse genes have been knocked out, although only 415 are readily available in the public domain via Jackson Laboratory. In addition, there are several gene-trap consortia through which embryonic stem (ES) cells can be obtained for genes of interest. However, knockouts to date have not been subjected to a standard set of phenotyping protocols and are often characterized using the expertise of the lab that generated the mice. Hence, many of these mice are grossly normal, resulting in "no-phenotype" publications. In such cases, it is likely that more subtle physiological parameters are not being appreciated. Efforts are needed to systematically phenotype these mice, as was proposed by several recent conferences and consortia (5).

A noteworthy study in the shift to examine physiological phenotypes at the genome-wide level is Howard Jacob's group's (95) Herculean task of completing measurements of a constellation of 239 parameters related to cardiovascular or renal physiology. Genotyping intercrossed F2 rats allowed his group to map 81 quantitative loci in rat and, with comparative genomics, localize these loci to the human genome.

HIGH-THROUGHPUT GAIN AND LOSS OF FUNCTION A new class of genome-wide experiments attempts to systematically perform genetic or chemical perturbation followed by cellular phenotyping (e.g., reporter gene activity or high-content microscopy). These approaches are possible in part due to large, high-quality collections of cDNAs, RNAi libraries, and growing collections of chemical compounds. Some of these studies have also been facilitated by the development of new methods for introducing DNA and chemicals into cells in a high throughput fashion (117).

For instance, Labow and colleagues tested the ability of 20,704 cDNAs to activate transcription from an IL-8 promoter reporter construct, identifying an unrecognized cAMP-like response element and a novel coactivator (*TORC1*) (49). In a similar fashion, others have reported large-scale screens of minimal synthetic

reporters seeking to identify genes that regulate AP1 or NF-κB activation (18, 67). RNAi screens have been used to investigate deubiquitinating enzymes in cancer-related pathways (16) as well as to study modulators of TRAIL-induced apoptosis and NF-κB activation (6, 116).

Several genome-wide RNAi screens in model organisms have helped discover entire catalogs of candidate genes for human diseases. For example, a genome-wide RNAi screen in *C. elegans* identified 417 genes that modulate fat metabolism (4). This screen employed a clever visual screen for fat-staining. The catalog of fat-related genes identified in this study includes the human orthologue of a human gene mutated in maturity onset diabetes of the young. The remaining genes identified in this study represent excellent candidate genes for human obesity or lipodystrophies.

Other Emerging Technologies

A number of other emerging technologies hold promise for facilitating disease gene discovery. Genome-scale location analysis is a technology for systematically detecting nucleic acid-protein interactions and was recently applied to discover target genes of diabetes-related transcription factors (76). Similarly, genome-wide profiling of RNA-binding proteins with their cognate transcripts (44) promises to shed insight into human diseases related to RNA processing. Systematic profiling of metabolites using NMR or tandem mass spectrometry, often dubbed "metabolomics" or "metabonomics," represents additional high-content readouts of cellular function that will complement RNA and protein profiling (74).

INTEGRATING GENOME-SCALE DATA SETS

In the examples described thus far, human disease genes were prioritized using information from a single type of functional genomics data set. In our search for human disease genes, we would ideally rely on multiple tiers of support that an individual gene is involved in a process before pursuing a costly association study. There are biological and statistical rationales for integrating diverse genomic data sources.

First, each technology interrogates different aspects of gene function. For example, affinity tag-based protein-interaction methods tend to discover membership in the same physical complex, whereas the Y2H technique discovers direct interactions (stable or transient). Synthetic lethality screens tend to discover genes that can compensate for each other, whereas coexpression analysis identifies genes that are likely under similar regulatory control. Combining these complementary viewpoints could be useful, providing a more comprehensive description of functional gene networks.

Second, each technology tends to produce noisy data and can be associated with its own inherent experimental limitations. For example, mass spectrometry-based proteomics systematically misses low-abundance proteins (43). Oligonucleotide

microarrays are sensitive to even low-abundance transcripts, but they can only quantify transcripts predefined on the chip. Metabolite profiling can be powerful, but the class of metabolites surveyed in a single experiment may depend on properties of the chromatography column. In addition, many of these large-scale experimental data sets are extremely noisy, so making genome-wide predictions using information from a single large-scale data set can lead to high numbers of false positives.

Several recent studies have shown that integration of different types of functional genomics data sets can produce more reliable predictions of yeast protein function and interaction (53, 60, 105). Specifically, these studies demonstrated that data integration can improve the sensitivity and specificity for detecting true functional relationships among genes. The benefits of integration are particularly valuable in prioritizing candidate human disease genes, where genomic intervals may be extremely large, and the cost of mutation screening or follow-up can be tremendous. In the CF example presented earlier, both transmembrane domain predictions and patterns of tissue distribution supported a role for *CFTR* in cystic fibrosis.

Here we briefly review ad hoc and formal approaches for integrating functional genomics data sets and discuss how such an integrated approach has been applied successfully to the identification of a human disease gene.

Approaches for Integrating Data Sets

A simple but intuitive approach for integrating data from diverse data sets uses simple logical operators such as AND and OR. The AND rule predicts a functional relationship only when all data sets agree, e.g., gene product A and gene product B share similar functions if A and B interact in a protein interaction network AND A and B exhibit coexpression in microarray experiments. The OR rule predicts an interaction when any of the experimental data sets supports the functional interaction. The AND rule is more stringent (and is expected to yield a higher specificity), whereas the more permissive OR rule provides greater sensitivity to detect functional interactions at the cost of specificity. Another way to combine different sources is to use majority voting. In this case, a functional relationship is predicted only when the majority of data sets agree. All of these methods suffer from one major disadvantage: They are all based on the assumption that each prediction from a data set has equal weight of confidence. This is not true because some methods can be more reliable than others.

Machine-learning methods provide more sophisticated data integration procedures that consider data reliability and redundancy as well as missing data, often leading to better results. An effective method is Bayesian inference, which was previously applied successfully in computational biology research, ranging from the prediction of subcellular localization of proteins (25) to the prediction of protein interactions in yeast (53). Bayesian inference combines information from heterogeneous data sets in a probabilistic manner, assigning a probability to the prediction result rather than just a binary classification (105). Each individual data set is essentially weighted by its accuracy and redundancy, which are determined

using gold standard "true positives" and "true negatives." Here we briefly review the principles of Bayesian inference through a simple example.

Imagine that we are interested in identifying candidate genes for aging, and that our hypothesis is that genes associated with reactive oxygen species (ROS) underlie this process. Our goal is to enumerate all genes in the genome that might be associated with ROS, as these will be reasonable candidate genes for aging. We must begin with a prior estimate of the number of ROS-related genes in the genome. With such an estimate we can compute the "prior odds" of finding an ROS gene, given by $O_{prior} = \frac{P(ROS)}{P(\sim ROS)}$. We can also consider the posterior odds of finding an ROS gene given N genome-scale data sets with values $g_1 \cdots g_N$:

$$O_{posterior} = \frac{P\,(ROS|g_1 \cdots g_N)}{P\,(\sim ROS|g_1 \cdots g_N)}$$

Posterior refers to the fact that the odds have changed after we have additional information from the large-scale data set. According to Bayes' theorem, the posterior odds can be calculated as $O_{posterior} = L(g_1 \cdots g_N)O_{prior}$, where $L(g_1 \cdots g_N)$ is the likelihood ratio defined as

$$L(g_1 \cdots g_N) = \frac{P\,(g_1 \cdots g_N|ROS)}{P\,(g_1 \cdots g_N|\sim ROS)}.$$

The two probabilities are estimated separately using a positive control set of ROS genes as well as a collection of genes known not to participate in this process. When the data $g_1 \ldots g_N$ are discrete, the probabilities are often constructed using contingency tables. Estimating the two probabilities can be rather challenging when N is large. However, if the N genome-scale data sets are independent of each other (i.e., they provide uncorrelated data), in which case the scenario is often termed naïve, then the L can be simplified to

$$L(g_1 \cdots g_N) = \prod_{i=1}^{N} \frac{P(g_i|ROS)}{P(g_i|\sim ROS)}.$$

In this case, different sources of data are decoupled. The likelihood ratio for each data source can be calculated separately and multiplied together to form L. The naïve Bayesian network is more easily computed and yields optimal results when the different data sets contain uncorrelated evidence; but even when this condition is not met, the results are often useful.

Such Bayesian approaches have been valuable in predicting yeast protein subcellular localization (25), protein interactions (53), and functional gene networks (105) using publicly available data.

Discovery of a Human Disease Gene via Integrative Genomics

Can such integrated approaches be applied to human diseases? We recently combined evidence from publicly available atlases of gene expression with organelle proteomics data (using a simple AND rule) to home in on the gene underlying Leigh Syndrome French Canadian variant (LSFC) (71).

LSFC is an autosomal recessive disorder characterized by a subacute degeneration of the brainstem as well as by a cytochrome *c* oxidase (*COX*) deficiency. The genes underlying four other inherited forms of *COX* deficiency were previously identified, and all encode mitochondrial proteins involved in assembling this multisubunit complex (90). Based on the clinical features of the disease (lactic acidosis and Leigh syndrome) and biochemical features of the disease (COX deficiency), we hypothesized that the gene underlying this disease encodes a protein involved in mitochondrial biology.

Beginning with this clinical clue, we integrated three sources of data: genome sequence, RNA abundance, and protein expression, with the goal of identifying genes in the genome that encode proteins related to mitochondrial function (Figure 3). First, we used genome browsers and ab initio gene predictions to compile a

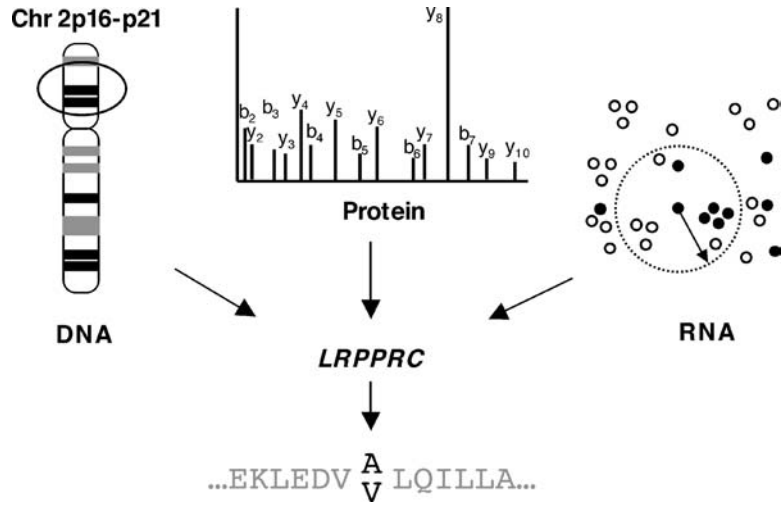

Figure 3 Discovery of a human disease gene through the integrated analysis of large-scale biological data sets (71). Leigh Syndrome French Canadian variant (LSFC) is an autosomal recessive, fatal metabolic disease that was previously mapped to a 2-Mb interval on chromosome 2. Its clinical and biochemical features suggested a disorder secondary to mitochondrial dysfunction. To prioritize these candidates, the authors used neighborhood analysis of publicly available microarray data sets to discover genes in the genome (of unknown function) that are coexpressed with the known mitochondrial genes. The authors also mapped tandem mass spectra (corresponding to peptides) from a mitochondrial proteomics project to this interval. When the two large-scale data sets and the genomic interval were integrated with a simple AND rule, one gene, *LRPPRC*, emerged as a candidate that is coregulated with known mitochondrial genes and gives rise to mitochondrial peptides. Based on this analysis, this gene was prioritized as the top candidate and resequenced in patients and controls. Mutations in *LRPPRC* provided strong genetic proof that *LRPPRC* underlies LSFC.

comprehensive list of candidate genes within the genetic linkage interval, thus identifying 30 genes total. Second, we explored four large-scale, publicly available atlases of RNA expression (69, 79, 96) and applied neighborhood analysis (described earlier) to score a query gene's expression correlation with the known, nuclear-encoded mitochondrial genes. Using this metric, we scored all the genes in the genome for their correlation in expression to the previously known mitochondrial genes. Third, we took tandem mass spectra (each corresponding to a single peptide) from a mitochondrial proteomics project and mapped them directly onto the genome.

We then integrated these three data sets to discover that exactly one gene, *LRPPRC*, had a high neighborhood analysis score and peptide support from the proteomics project (Figure 4). Hence, it emerged as a high-quality candidate gene for a disease characterized by mitochondrial dysfunction.

Prior to screening *LRPPRC* for mutations in the patients, we needed to ascertain its proper gene structure. By mapping the proteomic data directly onto the genomic interval, we determined that *LRPPRC* actually had a 38-exon structure, contrary to previous reports. We reasoned that the gene was misannotated and performed rapid amplification of cDNA ends (RACE) to validate a 38-exon structure of *LRPPRC*. With the complete gene structure in hand, we resequenced *LRPPRC* in patients, parents, and unrelated controls to discover two mutations in this gene that underlie LSFC. Hence, the combined analysis of genome, RNA, and protein enabled us to

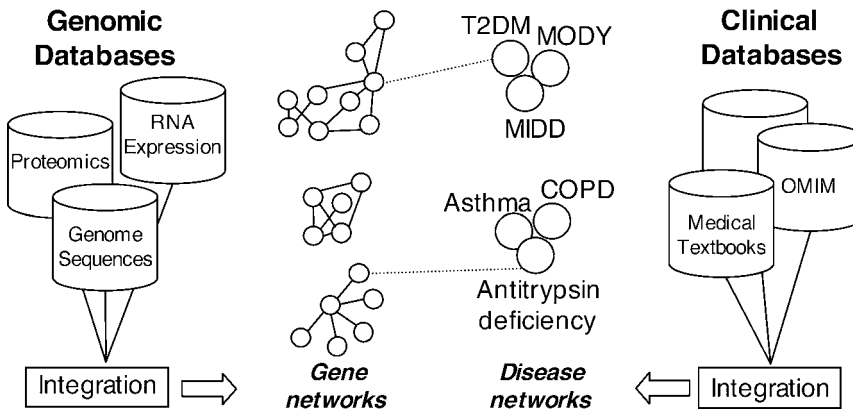

Figure 4 Genomics-based disease gene discovery in the future. Genome-scale data sets, like those described in the text, can be used to decipher the functional network relationships among all the genes in the human genome. Via accessible databases, it will soon be possible to cluster human diseases on the basis of clinical signs, symptoms, etiology, and pathogenesis to help construct disease networks. Gene networks can be mapped to disease networks via the ~1500 known disease-gene relationships. Such mappings will enable researchers to start with clinical features of a disease to discover the gene networks that underlie and modify them.

quickly move from clinical clues to a candidate gene, annotate that gene structure, and then subject it to systematic resequencing. This example illustrates how one can link clinical features of a disease to genes through the integrated analysis of genomic data.

FUTURE PROSPECTS

Genome-scale experiments are generating a wealth of data that provide systematic and complementary views of gene function. These resources, combined with new computational methodologies, are already accelerating disease gene discovery. Studies in yeast and in other model organisms have been extremely valuable: Not only have they generated valuable data that has directly assisted in the discovery of human disease genes, but they have provided important lessons on how best to integrate diverse data sets to infer gene function. Botstein & Risch (14) suggested that the disease genes discovered to date likely represent the easy ones, but hopefully the integrated analysis of genome-scale information will facilitate the discovery of those that remain.

We anticipate that in the very near future, strategies will be developed that systematically link clinical features to genomic elements (Figure 4). We can consider two separate networks: networks of genes and networks of disease. Diseases can be related to each other on the basis of shared clinical signs and symptoms, pathophysiology, etiology, or cellular endophenotypes. Genes and genomic elements can be related to each other using the growing wealth of functional genomics data with the approaches described in this review. The established ~1500 disease-gene relationships provide links between these two spaces. For a new disease (possibly complex) of unknown etiology, we can identify other disorders sharing similar clinical features, a subset of which may be previously associated with human genes. Gene networks containing these genes naturally represent excellent candidates or modifiers for the query disease. The mapping between clinical features (phenotype) and genes may become so robust that genes underlying a sporadic disease may be identified on the basis of the presenting symptoms in a single individual. Achieving this goal fundamentally requires integrating clinical informatics databases with genomics databases and carries with it key challenges.

First, we need improved nosology, i.e., methods for disease classification. Traditionally, diseases have been categorized on the basis of pathophysiology or on etiology, but often these characterizations break down and more ad hoc approaches are used, resulting in the celebrated debate between splitters and lumpers (68). An ontology-based approach to disease classification, in which a fixed vocabulary is used to annotate diseases, can improve this process. The Unified Medical Language System (UMLS) represents a set of knowledge sources developed at the U.S. National Library of Medicine (http://umlsinfo.nlm.nih.gov) and is a promising resource for improving disease classification. Medical textbooks and other clinical data sources need to adopt such a standard so that information can be freely exchanged.

Second, to construct disease networks, it is essential that we can access the tremendous wealth of knowledge stored in medical textbooks, scientific literature, and clinical journals. The recent collaboration between the Internet search engine company Google and leading research libraries at Harvard University, Oxford University, Stanford University, University of Michigan, and New York Public Library promises to provide searchable access to millions of texts in the public domain or excerpted from copyrighted materials. In addition to making a great number of current texts available online, this initiative will add nineteenth and early twentieth century texts to the body of electronically searchable knowledge, transcending limitations of predigital publishing technologies (66). When combined with improved data-mining tools, such data sets promise to help us construct informative and structured mappings among human diseases.

Third, we need more freely accessible genome-scale data sets. Most of the currently available large-scale data sets focus on a subset of protein-encoding genes, making it difficult to extend functional predictions to other genomic elements. If we are to implicate noncoding conserved elements in human disease, it's essential that we generate large-scale data sets that annotate their function. We also need improved data standards and tools for accessing and visualizing data. At present, genome sequence information and microarray data sets are beginning to become freely available and accessible to all users via standard formats. Similar standards and resources will be required for other large-scale data sets.

Finally, we need improved methods for integrating large-scale data sets that can properly manage the nuances of these genomic data sets. Ideal integrative strategies would handle categorical as well as continuous measures, would take into account positive and negative controls, and would make reasonable predictions without overfitting. Such strategies would also have to handle missing data or sparse data as well as highly correlated data. Bayesian approaches represent a reasonable approach to this challenge, but other techniques will certainly be needed.

Genomics is yielding a tremendous amount of information on the nature and function of all features of the human genome. In the coming years, as comprehensive genotyping and sequencing technologies mature, we will see a rapid shift from candidate gene studies to genome-wide association studies for rare and for common human diseases. The challenge then will lie in determining which statistical associations are true and relevant to disease biology. As we embark on these exciting new studies, the integration of genome-wide association studies with functional genomics data sets will enable us to spotlight not only single genes but also entire networks of genes that underlie and modify human disease.

ACKNOWLEDGMENTS

We thank Tracey Petryshen, Emily Walsh, Debora Marks, and Joel Hirschhorn for valuable comments on the manuscript. We thank Yanhui Hu for generating the data in Figure 1. VKM is funded by a Career Award in the Biomedical Sciences from the Burroughs Wellcome Fund.

The *Annual Review of Genomics and Human Genetics* is online at http://genom.annualreviews.org

LITERATURE CITED

1. Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, et al. 1992. Sequence identification of 2,375 human brain genes. *Nature* 355:632–34
2. Andersen JS, Lyon CE, Fox AH, Leung AK, Lam YW, et al. 2002. Directed proteomic analysis of the human nucleolus. *Curr. Biol.* 12:1–11
3. Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, Mann M. 2003. Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* 426:570–74
4. Ashrafi K, Chang FY, Watts JL, Fraser AG, Kamath RS, et al. 2003. Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. *Nature* 421: 268–72
5. Austin CP, Battey JF, Bradley A, Bucan M, Capecchi M, et al. 2004. The knockout mouse project. *Nat. Genet.* 36:921–24
6. Aza-Blanc P, Cooper CL, Wagner K, Batalov S, Deveraux QL, Cooke MP. 2003. Identification of modulators of TRAIL-induced apoptosis via RNAi-based phenotypic screening. *Mol. Cell.* 12:627–37
7. Bagshaw RD, Mahuran DJ, Callahan JW. 2004. A proteomics analysis of lysosomal integral-membrane proteins reveals the diverse composition of the organelle. *Mol. Cell. Proteomics*
8. Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116:281–97
9. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–25
10. Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RH, Cuppen E. 2005. Phylogenetic shadowing and computational identification of human microRNA Genes. *Cell* 120:21–24
11. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, et al. 2004. An overview of Ensembl. *Genome Res.* 14:925–28
12. Blackshaw S, Harpavat S, Trimarchi J, Cai L, Huang H, et al. 2004. Genomic analysis of mouse retinal development. *PLoS Biol.* 2:E247
13. Bolland DJ, Wood AL, Johnston CM, Bunting SF, Morgan G, et al. 2004. Antisense intergenic transcription in V(D)J recombination. *Nat. Immunol.* 5:630–37
14. Botstein D, Risch N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* 33(Suppl.):228–37
15. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, et al. 2003. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 31:68–71
16. Brummelkamp TR, Bernards R, Agami R. 2002. A system for stable expression of short interfering RNAs in mammalian cells. *Science* 296:550–53
17. Chalhoub N, Benachenhou N, Rajapurohitam V, Pata M, Ferron M, et al. 2003. Grey-lethal mutation induces severe malignant autosomal recessive osteopetrosis in mouse and human. *Nat. Med.* 9:399–406
18. Chanda SK, White S, Orth AP, Reisdorph R, Miraglia L, et al. 2003. Genome-scale functional profiling of the mammalian AP-1 signaling pathway. *Proc. Natl. Acad. Sci. USA* 100:12153–58
19. Chee M, Yang R, Hubbell E, Berno A, Huang XC, et al. 1996. Accessing genetic information with high-density DNA arrays. *Science* 274:610–14

20. Cheng Y, Church GM. 2000. *Biclustering of expression data*. Presented at 8th Intl. Conf. Intell. Syst. Mol. Biol. (ISMB)
21. Chiang AP, Nishimura D, Searby C, Elbedour K, Carmi R, et al. 2004. Comparative genomic analysis identifies an ADP-ribosylation factor-like gene as the cause of Bardet-Biedl syndrome (BBS3). *Am. J. Hum. Genet.* 75:475–84
22. Cho RJ, Huang M, Campbell MJ, Dong H, Steinmetz L, et al. 2001. Transcriptional regulation and function during the human cell cycle. *Nat. Genet.* 27:48–54
23. de Hoog CL, Mann M. 2004. Proteomics. *Annu. Rev. Genomics Hum. Genet.* 5:267–93
24. DeRisi JL, Iyer VR, Brown PO. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680–86
25. Drawid A, Jansen R, Gerstein M. 2000. Genome-wide analysis relating expression level with protein subcellular localization. *Trends Genet.* 16:426–30
26. Drumm ML, Pope HA, Cliff WH, Rommens JM, Marvin SA, et al. 1990. Correction of the cystic fibrosis defect in vitro by retrovirus-mediated gene transfer. *Cell* 62:1227–33
27. Dryja TP, McGee TL, Reichel E, Hahn LB, Cowley GS, et al. 1990. A point mutation of the rhodopsin gene in one form of retinitis pigmentosa. *Nature* 343:364–66
28. Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30:207–10
29. Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863–68
30. Farrar GJ, Kenna P, Jordan SA, Kumar-Singh R, Humphries MM, et al. 1991. A three-base-pair deletion in the peripherin-RDS gene in one form of retinitis pigmentosa. *Nature* 354:478–80
31. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, et al. 2002. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* 419:520–26
32. Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J. 2000. Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. *Nature* 408: 325–30
33. Frattini A, Orchard PJ, Sobacchi C, Giliani S, Abinun M, et al. 2000. Defects in TCIRG1 subunit of the vacuolar proton pump are responsible for a subset of human autosomal recessive osteopetrosis. *Nat. Genet.* 25:343–46
34. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415:141–47
35. Geisterfer-Lowrance AA, Kass S, Tanigawa G, Vosberg HP, McKenna W, et al. 1990. A molecular basis for familial hypertrophic cardiomyopathy: a beta cardiac myosin heavy chain gene missense mutation. *Cell* 62:999–1006
36. Getz G, Levine E, Domany E. 2000. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA* 97:12079–84
37. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428:493–521
38. Giglia-Mari G, Coin F, Ranish JA, Hoogstraten D, Theil A, et al. 2004. A new, tenth subunit of TFIIH is responsible for the DNA repair syndrome trichothiodystrophy group A. *Nat. Genet.* 36:714–19
39. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al. 2003. A protein interaction map of *Drosophila melanogaster*. *Science* 302:1727–36
40. Gollub J, Ball CA, Binkley G, Demeter J, Finkelstein DB, et al. 2003. The Stanford Microarray Database: data access and

quality assessment tools. *Nucleic Acids Res.* 31:94–96
41. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531–37
42. Gong S, Zheng C, Doughty ML, Losos K, Didkovsky N, et al. 2003. A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* 425:917–25
43. Gygi SP, Rochon Y, Franza BR, Aebersold R. 1999. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* 19:1720–30
44. Hieronymus H, Silver PA. 2003. Genome-wide analysis of RNA-protein interactions illustrates specificity of the mRNA export machinery. *Nat. Genet.* 33:155–61
45. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695–716
46. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415:180–83
47. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* 102:109–26
48. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, et al. 2003. Global analysis of protein localization in budding yeast. *Nature* 425:686–91
48a. International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–45
49. Iourgenko V, Zhang W, Mickanin C, Daly I, Jiang C, et al. 2003. Identification of a family of cAMP response element-binding protein coactivators by genome-scale functional analysis in mammalian cells. *Proc. Natl. Acad. Sci. USA* 100:12147–52
50. 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. USA* 98:4569–74
51. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, et al. 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* 283:83–87
52. Jansen R, Greenbaum D, Gerstein M. 2002. Relating whole-genome expression data with protein-protein interactions. *Genome Res.* 12:37–46
53. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, et al. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302:449–53
54. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. 2004. Human microRNA targets. *PLoS Biol.* 2:e363
55. Katsanis N, Worley KC, Gonzalez G, Ansley SJ, Lupski JR. 2002. A computational/functional genomics approach for the enrichment of the retinal transcriptome and the identification of positional candidate retinopathy genes. *Proc. Natl. Acad. Sci. USA* 99:14326–31
56. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. 2002. The human genome browser at UCSC. *Genome Res.* 12:996–1006
57. Kiyosawa H, Yamanaka I, Osato N, Kondo S, Hayashizaki Y. 2003. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* 13:1324–34
58. Kornak U, Kasper D, Bosl MR, Kaiser E, Schweizer M, et al. 2001. Loss of the ClC-7 chloride channel leads to osteopetrosis in mice and man. *Cell* 104:205–15
59. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
60. Lee I, Date SV, Adai AT, Marcotte EM.

2004. A probabilistic functional network of yeast genes. *Science* 306:1555–58
61. Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120:15–20
62. Li JB, Gerdes JM, Haycraft CJ, Fan Y, Teslovich TM, et al. 2004. Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. *Cell* 117:541–52
63. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, et al. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science* 303:540–43
64. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, et al. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288:136–40
65. Malkin D, Li FP, Strong LC, Fraumeni JF Jr, Nelson CE, et al. 1990. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* 250:1233–38
66. Markoff J, Wyatt E. 2004. Google is adding major libraries to its database. *N.Y. Times*, p. 1
67. Matsuda A, Suzuki Y, Honda G, Muramatsu S, Matsuzaki O, et al. 2003. Large-scale identification and characterization of human genes that activate NF-kappaB and MAPK signaling pathways. *Oncogene* 22:3307–18
68. McKusick VA. 1969. On lumpers and splitters, or the nosology of genetic disease. *Perspect. Biol. Med.* 12:298–312
69. Miki R, Kadota K, Bono H, Mizuno Y, Tomaru Y, et al. 2001. Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc. Natl. Acad. Sci. USA* 98:2199–204
70. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34:267–73
71. Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, et al. 2003. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc. Natl. Acad. Sci. USA* 100:605–10
72. Mootha VK, Bunkenborg J, Olsen JV, Hjerrild M, Wisniewski JR, et al. 2003. Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* 115:629–40
73. Murali TM, Kasif S. 2003. Extracting conserved gene expression motifs from gene expression data. *Pac. Symp. Biocomput.* 8:77–88
74. Nicholson JK, Connelly J, Lindon JC, Holmes E. 2002. Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discov.* 1:153–61
75. Nobrega M, Pennacchio LA. 2004. Comparative genomic analysis as a tool for biological discovery. *J. Physiol.* 554:31–39
76. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, et al. 2004. Control of pancreas and liver gene expression by HNF transcription factors. *Science* 303:1378–81
77. Owen AB, Stuart J, Mach K, Villeneuve AM, Kim S. 2003. A gene recommender algorithm to identify coexpressed genes in *C. elegans*. *Genome Res.* 13:1828–37
78. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* 96:4285–88
79. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, et al. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA* 98:15149–54
80. Ranish JA, Hahn S, Lu Y, Yi EC, Li XJ, et al. 2004. Identification of TFB5,

a new component of general transcription and DNA repair factor IIH. *Nat. Genet.* 36:707–13
81. Rappsilber J, Ryder U, Lamond AI, Mann M. 2002. Large-scale proteomic analysis of the human spliceosome. *Genome Res.* 12:1231–45
82. Rich DP, Anderson MP, Gregory RJ, Cheng SH, Paul S, et al. 1990. Expression of cystic fibrosis transmembrane conductance regulator corrects defective chloride channel regulation in cystic fibrosis airway epithelial cells. *Nature* 347:358–63
83. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–33
84. Saitoh N, Spahr CS, Patterson SD, Bubulya P, Neuwald AF, Spector DL. 2004. Proteomic analysis of interchromatin granule clusters. *Mol. Biol. Cell* 15: 3876–90
85. Sakai E, Bottaro A, Davidson L, Sleckman BP, Alt FW. 1999. Recombination and transcription of the endogenous Ig heavy chain locus is effected by the Ig heavy chain intronic enhancer core region in the absence of the matrix attachment regions. *Proc. Natl. Acad. Sci. USA* 96:1526–31
86. Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467–70
87. Schirmer EC, Florens L, Guan T, Yates JR III, Gerace L. 2003. Nuclear membrane proteins with potential disease links found by subtractive proteomics. *Science* 301:1380–82
88. Schonberger J, Seidman CE. 2001. Many roads lead to a broken heart: the genetics of dilated cardiomyopathy. *Am. J. Hum. Genet.* 69:249–60
89. Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW. 1996. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet.* 14:450–56
90. Shoubridge EA. 2001. Cytochrome c oxidase deficiency. *Am. J. Med. Genet.* 106: 46–52
91. Skop AR, Liu H, Yates J III, Meyer BJ, Heald R. 2004. Dissection of the mammalian midbody proteome reveals conserved cytokinesis mechanisms. *Science* 305:61–66
92. Sonnhammer EL, von Heijne G, Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6:175–82
93. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, et al. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* 9:3273–97
94. Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, et al. 2002. Systematic screen for human disease genes in yeast. *Nat. Genet.* 31:400–4
95. Stoll M, Cowley AW Jr, Tonellato PJ, Greene AS, Kaldunski ML, et al. 2001. A genomic-systems biology map for cardiovascular function. *Science* 294:1723–26
96. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. USA* 99: 4465–70
97. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* 101:6062–67
98. Sun-Wada GH, Wada Y, Futai M. 2003. Lysosome and lysosome-related organelles responsible for specialized functions in higher organisms, with special emphasis on vacuolar-type proton ATPase. *Cell Struct. Funct.* 28:455–63
99. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, et al. 1999. Interpreting

patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96:2907–12

100. Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* 278:631–37
101. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. 1999. Systematic determination of genetic network architecture. *Nat. Genet.* 22:281–85
102. Teitelbaum SL, Ross FP. 2003. Genetic regulation of osteoclast development and function. *Nat. Rev. Genet.* 4:638–49
103. Tiranti V, D'Adamo P, Briem E, Ferrari G, Mineri R, et al. 2004. Ethylmalonic encephalopathy is caused by mutations in ETHE1, a gene encoding a mitochondrial matrix protein. *Am. J. Hum. Genet.* 74:239–52
104. Tong AH, Lesage G, Bader GD, Ding H, Xu H, et al. 2004. Global mapping of the yeast genetic interaction network. *Science* 303:808–13
105. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. USA* 100:8348–53
106. Tsui LC, Buchwald M, Barker D, Braman JC, Knowlton R, et al. 1985. Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* 230:1054–57
107. Tsui LC, Rommens JM, Burns J, Zengerling S, Riordan JR, et al. 1988. Progress towards cloning the cystic fibrosis gene. *Philos. Trans. R. Soc. London B. Biol. Sci.* 319:263–73
108. Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98:5116–21
109. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403:623–27
110. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. *Science* 270:484–87
111. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51
112. Vermeulen W, Rademakers S, Jaspers NG, Appeldoorn E, Raams A, et al. 2001. A temperature-sensitive disorder in basal transcription and DNA repair in humans. *Nat. Genet.* 27:299–303
113. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–62
114. Wu CC, MacCoss MJ, Mardones G, Finnigan C, Mogelsvang S, et al. 2004. Organellar proteomics reveals Golgi arginine dimethylation. *Mol. Biol. Cell* 15: 2907–19
115. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434:338–45
116. Zheng L, Liu J, Batalov S, Zhou D, Orth A, et al. 2004. An approach to genomewide screens of expressed small interfering RNAs in mammalian cells. *Proc. Natl. Acad. Sci. USA* 101:135–40
117. Ziauddin J, Sabatini DM. 2001. Microarrays of cells expressing defined cDNAs. *Nature* 411:107–10

APPENDIX Online resources for accessing large-scale biological data sets

Genome Browsers	
UCSC	http://genome.ucsc.edu/
Ensembl	http://www.ensembl.org/
NCBI	http://www.ncbi.nlm.nih.gov/Genomes/
Gene Expression Repositories	
ArrayExpress EBI	http://www.ebi.ac.uk/arrayexpress/
Stanford Microarray Database (SMD)	http://genome-www.stanford.edu/microarray
Gene Expression Omnibus (GEO)	http://www.ncbi.nlm.nih.gov/geo/
Protein Databases	
Database of Interacting Proteins	http://dip.doe-mbi.ucla.edu/
Biomolecular Interaction Network Database	http://bind.ca/
Human Reference Protein Database	http://www.hprd.org
Pathway Databases and Resources	
Reactome	http://www.reactome.org/
KEGG	http://www.genome.ad.jp/kegg/
BioCarta	http://www.biocarta.com/
GenMAPP	http://www.genmapp.org/
Disease Databases	
Online Mendelian Inheritance in Man	http://www.ncbi.nlm.nih.gov/omim/
Model Organism Databases	
Mouse Genome Database	http://www.informatics.jax.org/
Rat Genome Database	http://rgd.mcw.edu/
FlyBase	http://flybase.org
WormBase	http://www.wormbase.org/
Saccharomyces Genome Database	http://www.yeastgenome.org
Zebrafish Information Network	http://zfin.org/
Other Resources	
miRNA Registry	http://www.sanger.ac.uk/Software/Rfam/mirna/
ChemBank	http://chembank.broad.harvard. edu/