BRIEFINGS IN BIOINFORMATICS. VOL 11. NO 1. 96–110
Advance Access published on 10 December 2009
doi:10.1093/bib/bbp048

Advances in translational bioinformatics: computational approaches for the hunting of disease genes

Maricel G. Kann

Submitted: 11th August 2009; Received (in revised form): 15th September 2009

Abstract

Over a 100 years ago, William Bateson provided, through his observations of the transmission of alkaptonuria in first cousin offspring, evidence of the application of Mendelian genetics to certain human traits and diseases. His work was corroborated by Archibald Garrod (Archibald AE. The incidence of alkaptonuria: a study in chemical individuality. *Lancert* 1902;**ii**:1616–20) and William Farabee (Farabee WC. Inheritance of digital malformations in man. In: *Papers of the Peabody Museum of American Archaeology and Ethnology.* Cambridge, Mass: Harvard University, 1905; 65–78), who recorded the familial tendencies of inheritance of malformations of human hands and feet. These were the pioneers of the hunt for disease genes that would continue through the century and result in the discovery of hundreds of genes that can be associated with different diseases. Despite many ground-breaking discoveries during the last century, we are far from having a complete understanding of the intricate network of molecular processes involved in diseases, and we are still searching for the cures for most complex diseases. In the last few years, new genome sequencing and other high-throughput experimental techniques have generated vast amounts of molecular and clinical data that contain crucial information with the potential of leading to the next major biomedical discoveries. The need to mine, visualize and integrate these data has motivated the development of several informatics approaches that can broadly be grouped in the research area of 'translational bioinformatics'. This review highlights the latest advances in the field of translational bioinformatics, focusing on the advances of computational techniques to search for and classify disease genes.

Keywords: *translational bioinformatics; disease genes; computational biology*

INTRODUCTION

More than 100 years ago, Archibald Garrod confirmed, with his study of the incidence of alkaptonuria in men, the Mendelian laws of inheritance of this disorder. Dr William Bateson, a keen follower of Mendel, had previously hypothesized that alkaptonuria in offspring resulting from mating of first cousins might be the due to the fact that 'first cousins will frequently be the bearer of similar gametes' dispelling the previous notion that mating of first cousins in general might lead to the diseases, and hypothesizing that the disease follows similar inheritance laws observed by Mendel in plants. Just after the terms genotype and phenotype were coined [1], in 1905, William Farabee [2], a recognized anthropologist, recorded the familial tendencies of inheritance for malformations of human hands and feet and also recognized the Mendelian patterns of inheritance for those anomalies.

It would take over 90 more years of genetic research to identify mutations in the BRCA1 gene with clear relationships to familial breast cancer [3]. This breakthrough knowledge has had important implications for the diagnosis and prognosis of

Corresponding author. Maricel G. Kann, University of Maryland, Baltimore County 1000 Hilltop Circle, Baltimore, MD 21250, USA. Tel: +1-410-455-2258; Fax: +1-410-455-3875; E-mail: mkann@umbc.edu

Maricel G. Kann is an assistant professor at the University of Maryland, Baltimore County. Her research interests include methods for alignment of protein sequences, predictors of protein–protein interactions and the study of protein domains and their associations with disease. She has co-chaired several sessions at international bioinformatics conferences related to the field of translational bioinformatics.

cancer and familial forms of other complex diseases. However, we are still far from resolving the subtleties involved in the intricate pathways and molecular relationships responsible for these disorders, in particular for complex diseases, and, most importantly, we are still unable to deliver a cure for most diseases. The shift to large-scale sequencing of individual human genomes and the availability of new techniques for probing thousands of genes provide new sources of meaningful medical insights. The informatics issues related to the accession, integration, visualization and representation of this knowledge in a systematic manner are quite challenging. On the other hand, the stakes are high; for instance, by identifying molecular patterns that characterize each individual genome and discerning which of these individual variations is related to a particular disease or response to treatment, bioinformaticians could provide the foundations for the development of tools for the diagnosis, prognosis and personalized treatment of diseases.

Translational bioinformatics is an emerging field addressing the computational challenges in biomedical research and the analysis of the vast amount of clinical data generated from it [4]. It is difficult to define such a broad field and, due to the inherently interdisciplinary nature of the research, impossible to detach translational bioinformatics from other related fields. The American Medical Informatics Association (AMIA) has defined the field of Translational Bioinformatics as:

> the development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of translational bioinformatics is newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders, including biomedical scientists, clinicians, and patients [5].

The combination of novel experimental techniques with the emergence of translational bioinformatics has changed how the search for disease genes is performed. In the past, searching for disease genes was done mainly using positional cloning. In modern approaches, bioinformatics is an integral part of the search for disease-associated genes.

Besides the potential impact on personalized medicine, the field of translational bioinformatics provides a wide range of tools and resources that are invaluable in biomedical research. Due to the space limitations, however, this review will focus on the latest accomplishments in the hunt for disease genes.

SEARCHING FOR DISEASE GENES

This review provides a summary of the computational approaches related to the search for disease genes and it is divided into three parts. The first section is focused on the study of the properties and characteristics of disease genes. The second section provides a description of the methodologies and the available resources for the identification of disease genes. Finally, the last section highlights the advances of the study of specific gene disruptions associated with diseases, i.e. the analysis of human single nucleotide polymorphisms (SNPs) and structural variations.

Characterization of disease genes

The two main intrinsic properties of the genes that hamper the study of their functions and their associations with diseases are:

(1) Diseases are caused by the effect of several genes: for instance, comprehensive studies on mutations in complex diseases, such as breast cancer [6] or other types of cancer [7, 8], reported hundreds of mutated genes. While this is certainly the case for complex diseases, even simple Mendelian diseases can lead to complex genotype–phenotype associations.
(2) Genes can often perform several functions (gene pleiotropy): Mutational analyses of a particular gene, like the *BRAF* gene, reveal dozens of mutational sites that lead to different phenotype associations to cancer [7].

In addition, environmental factors also make disease traits difficult to detect and complicate the search for the genes responsible for such traits. The use of medication or xenobiotic substances is an example of an environmental variant. For example, it is difficult to detect whether alcohol-induced toxicity is normal

to the phenotype, or whether it is a result of the individual variation in human liver alcohol dehydrogenase and other enzymes. Other environmental factors of great relevance to human disease are viruses and other infectious agents [9, 10]. Environmental factors, in conjunction with epigenetic regulation of the genes, might also be responsible for the low penetrance of certain alleles.

Disease genes are those genes involved in the causation of, or associated with [e.g. in genome-wide association studies (GWAS)], the disease. For example, for cystic fibrosis (CF), the gene *CTFR* was mapped to chromosome 7q31-q32 by linkage analysis in 1985 and later cloned by Francis Collins and co-workers [11]. The deletion of three base pairs in CFTR's nucleotide sequence results in the absence of a phenylalanine residue at position 508 of the protein. In the endoplasmic reticulum, CFTR proteins with this deletion are targeted for degradation. As a result, there is an imbalance of the sodium and chloride ion concentrations that creates a thick, sticky mucus layer that leads to chronic infections. Environmental and genetic factors influence this disease, and as a result, individuals with the same mutation might have different disease outcomes. Despite advances in understanding CF, there is still much to learn and understand about this disease. In particular, the mechanism for lung disease in CF patients, which is lethal, is still unknown.

The number of disease genes discovered has been steadily increasing throughout the years. Figure 1 depicts the growth of disease gene data from 1981 to 2009 (D.Magglot and J.Amberger, personal communication). The analyses of the characteristics that differentiate disease from non-disease genes have been used to develop disease classifiers, a research area of major importance due to the medical relevance of disease genes.

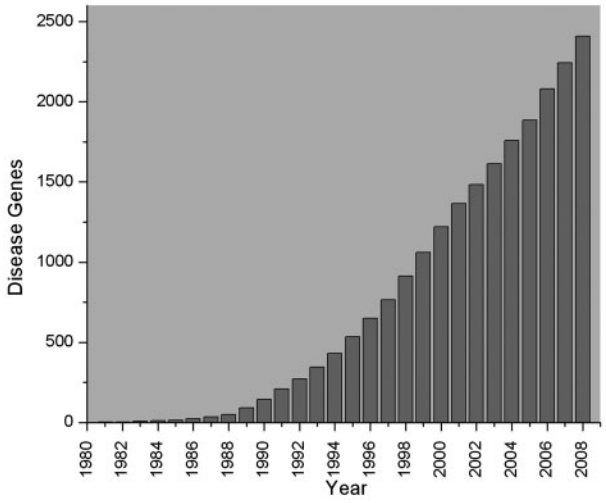

Figure 1: Histogram of cumulative growth of disease gene discovery. Counts from 1981 to 2005 correspond to the number of diseases for which the underlying genetic defect is known. Values for the last 3 years also include some selected diseases for which a genetic association has been reported, but no causation has been shown.

Disease gene properties

Proteins derived from disease genes have been found to have properties that distinguish them from all other genes: they are longer [12], more conserved, phylogenetically extended and without close paralog [13]. In addition, when compared against housekeeping genes, they present different patterns of conservation, function and DNA coding lengths [14]. Since inherited disease genes are more likely to be non-essential, one could hypothesize that they arrive later in the evolution of the human species. Surprisingly, Domazet-Loso and Tautz [15] studies showed that non-essential disease genes are of ancient origin. In agreement with previous findings about the disease gene length, their analysis also showed that ancient genes tend to be longer. The authors, confirming what others had reported, found no significant differences in the rates of evolution of disease versus non-disease genes [14]. The question of evolutionary divergence of disease genes, however, remains open, with some findings indicating that there is a higher rate of non-synonymous substitutions than synonymous ones [16] and others the opposite [14, 17].

If disease genes were to evolve at higher speed, could this be just an effect of the weak dominance of these disease genes? To answer this question, Osada *et al.* [18] compared evolutionary rates and the degree of polymorphism of the dominant and recessive disease genes. They found a higher rate of non-synonymous polymorphisms in recessive genes. In their analysis, the differences in selection intensity are still significant even after taking into account the dominance, suggesting that there are significant differences in the deleterious effect of the dominant and recessive genes.

The interaction network of disease genes has been the subject of many studies [19, 20]. The relevance of protein interactions in diseases and the development of computational tools applied to disease gene

identification has been previously discussed by Kann [21]. Protein interaction networks have been studied for Alzheimer's disease [22], ataxias and disorders of Purkinje cell degeneration [23] and for cancer genes [19]. Genes involved in the same disease have been found to form subnetworks [24]. Finding functional modules associated with each disease could reveal important aspects of the disease mechanisms and aid in disease classification [25]. Feldman *et al*. [26] compared disease genes against essential genes (from mouse orthologs of human genes) in terms of their connectivity and found the two set of genes to be clearly distinct. In their analysis, the authors used a network of interactions derived from the analysis of hundreds of articles obtained with the Gene Ways [27] natural language system. The resulting network includes almost 13 000 physical interactions and 4458 genes. This work exemplifies the impact of text mining in the field of translational bioinformatics for the gathering of data from millions of existing manuscripts.

In addition to the extensive study of the structured regions of the protein and their effect on disease, the study of intrinsically disordered regions of the proteins from disease genes has recently lead to 'unfoldomics', or mapping of disordered proteins to human diseases [28, 29]. A number of intrinsically disordered proteins have been shown to be associated with cancer [30], cardiovascular disease [31], diabetes, neurodegenerative diseases [32] and other human diseases [33, 34].

To summarize, not all the properties that characterize disease genes have been probed or can be easily explained. Previous studies require functional, evolutionary and statistical hypotheses to explain the observations about disease genes. Disease genes might need to interact with each other and might also need to be co-expressed as they participate in the same functional pathways. Fewer paralogs within the human genome might explain the inability of the system to compensate for disruptions created when these genes are modified. Longer genes can be explained statistically as they will have more possible sites for mutations. Figure 2 depicts the distribution of lengths of disease and non-disease genes (from OMIM [35] and RefSeq [35], respectively). Based on the two-sample Kolmogorov–Smirnov test statistic, the distribution of lengths for the disease and non-disease genes are significantly different with a *P*-value of 3.0e-21. In addition, we estimated the number of protein domains (from CDD [36]) of disease genes (from OMIM) to be higher, on average, than non-disease genes (Kolmogorov–Smirnov test statistic with a *P*-value 1e-6).

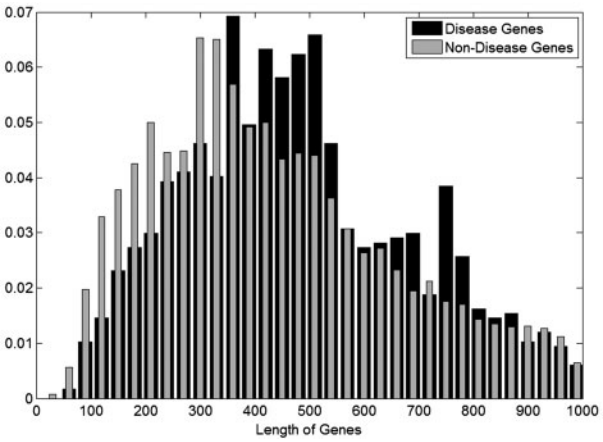

Figure 2: Distribution of length of proteins from disease and non-disease genes (black and gray, respectively). Disease genes (from OMIM [99]) are significantly longer than non-disease genes (RefSeq [35]).

Disease gene prioritization

Large-scale experiments generate lists of several hundreds of disease gene candidates, and it is still a challenge to identify the disease genes among them. Certain gene properties, as described above, differentiate disease genes and have been used as the bases for computational tools to prioritize disease gene candidates derived from these experiments. Table 1 provides a sample of the most recent publicly available sites that offer tools to rank disease gene candidates. All of these approaches are based on the integration of different sources. A summary of the data sources used by these methods and a brief description of the results are provided subsequently.

Data sources

Protein interaction (PPINT): It has been observed that disease genes are highly connected with other genes from the same disease. Differences in the network properties, such as higher connectivity, have been used to generate several gene-prioritization tools [12, 37–45]. PPINT is a feature that has been integrated with other gene properties into most of the tools highlighted in Table 1.

Gene function (gene ontology): Disease genes are expected to share common functional properties, as annotated in the gene ontology (GO) [46].

Table 1: Resources for gene prioritization

Name	Citation	Website	Data type
ToppGene	Chen *et al*. [65]	toppgene.cchmc.org	PPINT, GO, PATH, PDOM,GEXP, TXT, TFBS, MOUSE
PhenoPred	Radivojac *et al.* [70]	www.phenopred.org	PPINT, GO, STR, SEQ
CGPRIO	Furney *et al*. [74]	bg.upf.edu/cgprio	PCONS, GSTRU,PDOM, PPINT,REG
CAESAR	Gaulton *et al.* [80]	visionlab.bio.unc.edu/Caesar	TXT + GO, MOUSE, GEXP,eVOC, MP, PDOM
GenTrepid	George *et al.* [79]	www.gentrepid.org	PPINT, PATH, PDOM
ENDEAVOUR	Aerts *et al*. [78]	www.esat.kuleuven.be/endeavour	PATH, PPINT, PDOM, GEXP, GO, TFBS, SEQ, TXT
SUSPECTS	Adie *et al.* [77]	www.genetics.med.ed.ac.uk/suspects	GO,PDOM,GEXP + Prospectr
PROSPECTR	Adie *et al*. [75]	www.genetics.med.ed.ac.uk/prospectr	LEN,PHYL
Prioritizer	Franke *et al*. [74]	pcdoeglas.med.rug.nl/prioritizer	GO,PATH,PPINT,GEXP

The table summarizes the data types used to develop the different approaches for gene prioritization highlighted here and their abbreviations as found in the main text: PPINT, PATH, PDOM, GO, GEXP, gene–disease associations extracted from biomedical literature (TXT), TFBS, information about orthologous sequences from mouse (MOUSE), structure (STR) and sequence (SEQ) similarity, DOs, conservation at the protein level (PCONS), gene regulation (REG), gene structure (GSTRU), eVOC anatomical ontology (eVOC), MP, gene length (LEN) and phylogenetic information (PHYL).

This hypothesis was tested and validated in a set of disease genes from OMIM for each of the three branches of GO, namely biological process, cellular components and molecular function [47]. Goh *et al*. [47] showed that the GO homogeneity in each branch of GO is significantly higher for each disorder compared with random. Therefore, most methods for gene prioritization will increase the score of the candidate gene that share GO annotation with other genes from the same disease [48].

Pathway (PATH): As with functional annotation, disease genes are, as it is with functional annotation, most likely to share common pathways as annotated in KEGG [49], Reactome [50], BioCarta [51], BioCyc [52], GenMAPP [53], MSigDB [54] and others.

Gene expression (GEXP): Disease genes are expected to be co-expressed; thus gene expression data can be used in combination with GO, PPINT and other features described in this section to increase the performance of gene-prioritization methods [45]. In addition, the availability of gene-expression data with clinical phenotypes has generated many approaches for the integration of these data, all with enormous potential for the diagnosis and prognosis of cancer and other complex diseases [55–63]. A full description of the advances in microarray analysis is beyond the scope of this manuscript.

Protein domain (PDOM): Candidate disease genes might have functions that are more similar to those of known disease genes [64]. The function affected might be due to the protein domains, which represent the functional units of the proteins. The presence of a certain domain when genes with that particular domain are enriched in the disease, has been used as an indication of the association of that gene with the disease.

Gene regulation (REG and TFBS): Genes within the same gene-regulation network are expected to affect similar diseases. Thus, similarities in transcription factor binding sites (TFBSs) have been also incorporated into several of the approaches highlighted in Table 1.

Sequence properties (SEQ: LEN, GSTRU): Sequence properties such as gene and protein length or structure could distinguish disease genes from non-disease genes (see previous section). Sequence similarity has also been incorporated into the ENDEAVOUR gene-prioritization tool.

Expression and phenotypic data from orthologs (ORTH, MOUSE): Functional information about genes in other species is the only source for functional information when the human data is not available or impossible to produce. Thus, studies on model organisms are key to biomedical progress. ToppGene [65] incorporates mouse data, and van Driel *et al*. [66] includes several other species into the GeneSeeker method.

Other ontologies used: The other ontologies used are eVOC anatomical ontology [67] and mammalian phenotype ontology (MP) [68] are used in CAESAR. The disease ontology (DO) information (http://diseaseontology.sourceforge.net) provides hierarchical organization for disease types based on

the Unified Medical Literature system (UMLS) [69]. The DO was incorporated into PhenoPred to cluster similar diseases into higher levels of aggregation, improving the confidence on PhenoPred predictions [70].

Text mining (TXT): There are over 19 million biomedical records in PubMed today [71], and this repository constitutes one of the best sources of information about disease genes. In addition to CAESAR (see below), several other approaches have been used to integrate text-mining tools with disease ontologies to derive gene–disease associations [43, 72].

Overview of the methods

Methods for gene prioritization rely on the information provided by one or more of the experimental techniques described above. Therefore, the amount and quality of the available experimental data generated by these techniques is a major limitation of the gene-prioritization techniques. For instance, protein–protein interaction-based methods suffer from the incompleteness and low quality of the data currently available for interaction networks in mammals. Another source of uncertainty is the disease mapping information used to train and evaluate the computational methods, for it is of variable resolution and expected to contain large numbers of false positives. Furthermore, gene-prioritization methods have been hampered by the complexity and difficulty in creating functional and disease ontologies. Methods that rely on text mining, also face the difficulties inherited from natural language processing, such as issues related with extracting gene names from the biomedical literature [73].

Prioritizer, developed by Franke *et al*. [74] is available for download in their site (Table1). Franke *et al*. [74] studied the effect of using three different gene networks—GO, PPINT and GEXP—to correctly rank the disease genes for a set of 96 disorders with 409 known disease genes. Combining PPINT and GEXP, a better ranking was achieved than what could have been obtained randomly. The method showed considerable improvement (represented by an increase in the area under the ROC curve) when GO was added. The best ranking of disease genes was reached when the three types of data were combined. The authors used the combination of all sources to prioritize genes in artificial susceptibility loci and found a 2.8-fold increase in the chance of detecting disease genes with respect to random selection.

Another tool, PROSPECTR, uses an alternating decision tree which has been trained to differentiate between genes 'likely to be involved in disease' and 'genes unlikely to be involved' in disease [75]. The method uses gene properties that are characteristic of disease genes (see previous section) to provide each gene with a score, which is a measure of confidence in the classification. In a test set of 675 genes from the Human Gene Mutation Database [76], and 675 picked at random from Ensembl (had no association with disease), PROSPECTR performed with a sensitivity of 0.71 and a specificity of 0.58. SUSPECTS adds an extra layer to PROSPECTR, using GO annotation, protein domain and gene expression data to rank and score each gene [77]. The program scores each gene of the test set based on its relationship to the networks in the training set. Similarly, ENDEAVOUR uses a larger set of over 12 data sets (listed in Table 1) and order statistics to score and rank genes from the test set [78]. The test set of disease genes used to benchmark SUSPECTS was, on average, within the top 13% of the candidates, while results using ENDEAVOUR indicate that disease genes from a test set of 200 genes ranked 13 on average, representing a 7- and 9-fold enrichment over random classifiers for each method, respectively.

GenTrepid uses two methods: common module profiling (CMP), based on similarity of protein-domain composition, and common pathway scanning (CPS), based on common protein interactions and metabolic pathways among disease genes [79]. George *et al*. [79] found the two methods to be complementary, with the combination of the two approaches yielding the best performance. However, a meta-analysis combining both methods into a consensus would have decreased the performance compared to using both methods independently. When used side by side, the two methods were reported to have a sensitivity of 0.52 and a specificity of 0.97 in a benchmark of 170 genes (29 diseases) representing a 13-fold enrichment in disease genes. In other words, a list of 100 gene candidates could be reduced to 8 with significant cost and time reduction in the posterior experimental analysis of these candidate genes.

CAESAR, developed by Vision and co-workers [80], is a tool primarily based on text mining of disease information mainly from review articles and the

integration of other gene data (Table 1). The authors have addressed the challenge of analyzing complex traits, In their study of 18 genes complex human trait susceptibility genes, CAESAR selected 7 of the genes within the top 2% of the ranked genes. From almost 15 000 genes, 16 of the 18 genes were ranked with a median rank of 549.5. This represents a 67-fold average enrichment.

PhenoPred, developed by Radivojac *et al.* [81] studies the network of interactions and functional relationships of the target protein and detects its local neighborhoods in the network. The authors devised a supervised approach to find local signatures of the disease and find new candidate genes that are not necessarily in close proximity to the known disease genes. PhenoPred can be queried using a gene or a disease name. If the input is a gene, the program will return a list of diseases that the gene could be associated with (based on the network properties of the genes known for that disease). If starting with a disease name, the program retrieves all the genes predicted to be related to the disease. In both cases, PhenoPred provides a similarity score that represents the chance of the gene–disease association to be true. The authors showed that this approach works best when combined with the molecular function of the query gene and physicochemical properties of its protein product.

ToppGene Suite [65] integrates a vast number of genomic data from humans and mice (Table 1). This state-of-the-art resource includes ToppFun and ToppGene methods that can be used for the analysis of gene functional enrichment and for the prioritization of disease gene candidates, respectively. It uses a fuzzy-based similarity measure between the genes in the training and test set based on their semantic annotation. It also derives the probability (*P*-value) that each annotation is related to the gene in question, using random sampling of the whole genome. The authors analyzed 20 gene–disease associations from five disorders (from recently reported GWAS) and found that ToppGene ranked 19 of 20 candidate genes within the top 20%. The mean rank for ToppGene was 6.8 (excluding diseases that lacked interaction data [65]).

Lastly, CGPRIO, a tool recently developed by Furney *et al.* [82] is based on gene properties such as length and structure for identifying those features that characterize cancer genes. Based on distinguishing features, a naïve Bayes model is used to classify genes as proto-oncogene or tumor suppressor genes (Table 1).

From the user's perspective, the most desirable features for these methods are:

(i) Online availability: all the methods in Table 1 are freely available and can be interactively used within their websites, with the exception of CAESAR and Prioritizer, which are downloadable from their sites as stand-alone applications.
(ii) Advanced interface for input of training and test sets: the option of a custom-made list of genes for training is highly desirable for biologists (input of ENDEAVOUR and ToppGene). Additional input options such as those found in SUSPECTS and GeneTrepid (including disease name or keywords within the gene description) facilitate the analysis when no customized training set is available.
(iii) Clear summary of the results with an overall ranking of the genes: some methods like ToppGene and SUSPECTS provide both a measure of the likelihood that the gene is responsible for the disease or associated with the training set and also a clear visualization of the results of ranking the genes with each feature used in the process.

In summary, the methods for disease gene prioritization have led to an improvement in the detection of disease genes and to an increase in our knowledge about the integration of the several data sources for gene function and disease association. However, these methods can only be as accurate as the data they are based upon, which is an important issue, given the low quality of some of the experimental data on which they rely (e.g. protein–protein interaction data is incomplete and unreliable). Producing good ontologies for complex processes and improving the methods for mining and integrating the multisource data are difficult tasks that, unless addressed, will continue to severely limit the progress of gene-prioritization techniques.

Computational approaches for the analysis of disease mutations

Recent advances in sequencing techniques are generating data about individual human genomes at a relatively low cost. The identification of disease-related SNPs derived from large-scale techniques has

the potential to create personalized tools for the diagnosis, prognosis and treatment of diseases. Mutations in the genomic code often produce changes in the protein sequence, leading to diseases. The key to approaches that identify disease mutations lies in distinguishing between SNPs that are functionally relevant from those that are not. For the non-synonymous SNPs within coding regions (coding nsSNPs), methods rely on the study of the functional disruptions produced in the protein. An in-depth discussion of the online resources available for the analysis of SNPs can be found in Karchin's recent review article [83]. Here, I discuss the recent advances in computational methodologies developed for the analysis of coding nsSNPs and, briefly, for the analysis of structural variants.

Analysis of SNPs

Large-scale GWAS and human-sequencing projects are producing hundreds of SNPs with putative relevance to cancer [84] and other diseases (see review [85]). Some of these sequence disruptions in the protein produce changes in the stability, regulation, ability to interact or to be modified, and are ultimately associated with the disease. Computational approaches developed to prioritize SNPs can reduce the number of experimental trials by focusing on sites that are functionally relevant. Ideally, one would also like to deduce from the analyses of SNPs the mechanistic changes produced by the mutation and the cause of the disease. Methods used to predict whether a mutation is deleterious combine structural, conservation and/or other sequence properties that identify the mutational site as a potential site. The properties used in these approaches are highlighted below.

Protein structure

Disease mutations have been found to affect the stability of the proteins [86] or to cause protein aggregation [87]. It has been shown by several authors that the impact of the coding nsSNPs can be investigated by studying the 3D structure of the protein [88–94]. Polyphen, a method developed by Sunayev and colleagues [95], relies on functional annotation and structure predictors for evaluating the deleteriousness of the SNPs.

Conservation

Early studies showed that disease mutations are located in conserved sites [94, 96]. Conservation across species is often an indication of functional relevance. One of the earlier approaches, SIFT, combined conservation with physicochemical properties of the amino acids to produce a list of mutations that are not tolerated at a particular protein site [89].

Protein domain

Location of the mutation within a particular protein domain is also critical to predicting deleterious effects. Clifford *et al.* [97] incorporated a score based on the protein domain's position specific scoring matrix (PSSM). The score, or logR.*E*-value, is calculated as the log10 (*E*-value_variant/*E*-value_canonical), where the *E*-values are generated from the domain's alignment of the variant and canonical proteins using HMMer [98]. The logR.*E*-value is a measure of how a particular mutation affects the total score of the alignment to the domain's PSSM. The authors found that this measure is a good predictor of whether or not the SNP is deleterious. Recently, Kann and co-workers have mapped all human SNPs and disease mutations (from OMIM [99] and Swiss-Prot [100]) to their corresponding protein domain sites. We have created a freely available resource for the domain mapping of domain mutations, the DMDM site. A screenshot of the DMDM protein domain webpage for the DNA-binding homeodomain is depicted in Figure 3. DMDM aggregates all the information about human mutations and provides coordinates of all mutations within the human domains. DMDM is available at http://bioinf.umbc.edu/DMDM and can be used to identify domain sites with high incidence of disease mutations.

Posttranslational modifications

Mutations that affect post-translational modifications might produce a gain or loss of function causative of disease. In a recent study of cancer mutations, Radivojac *et al.* [81] found that mutations predicted to have an effect on phosphorylation function are enriched in somatic cancer data. These results suggest that both gain and loss of phosphorylation might be important features for identifying cancer mutations, especially drivers. This approach was generalized to incorporate other post-translational modifications (methylation, glycosylation, ubiquitination) together with functional site predictors (e.g. catalytic residues, DNA-binding residues) towards probabilistically identifying molecular mechanisms of disease [101].

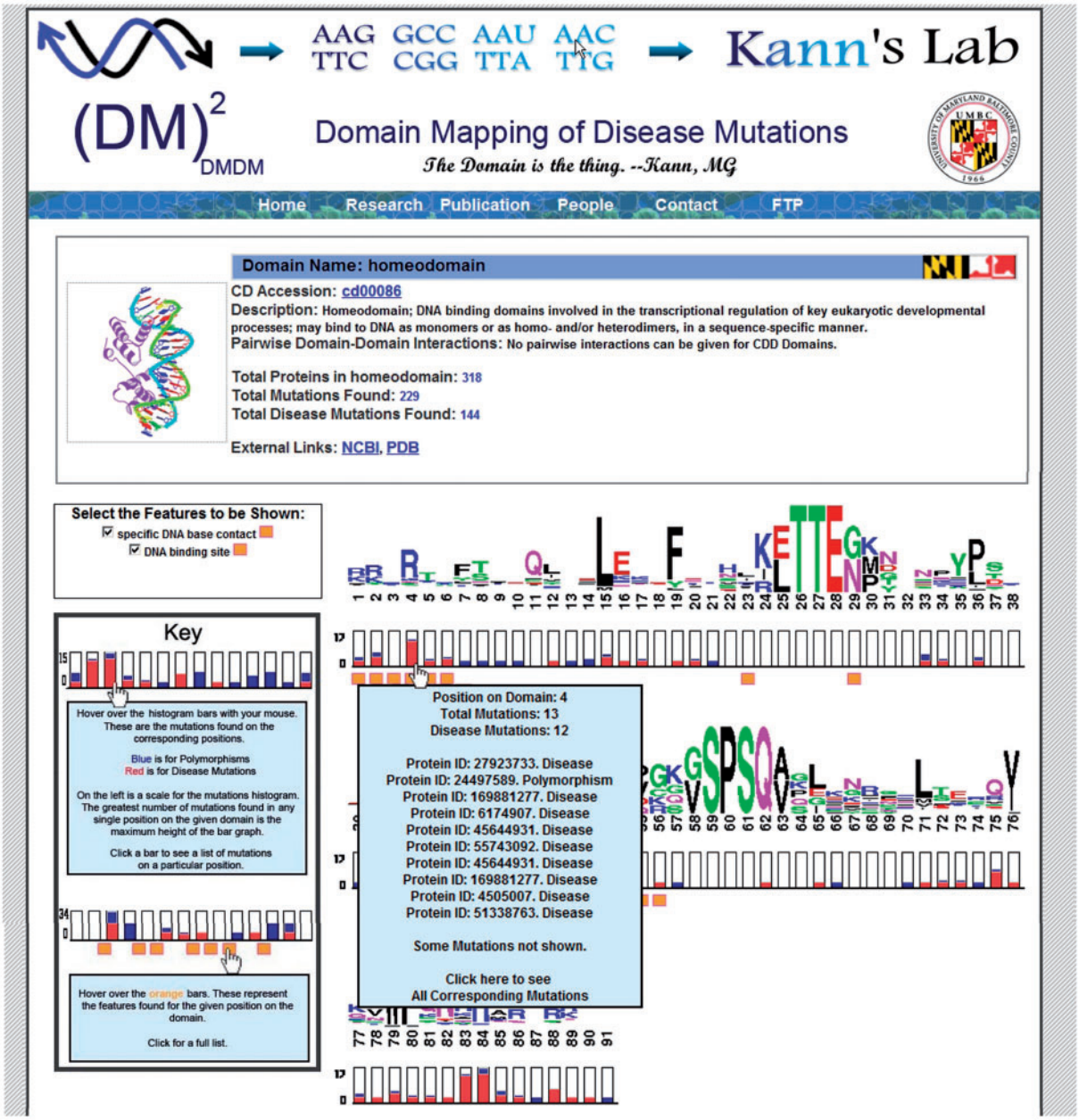

Figure 3: Screenshot of a protein domain page from the DMDM website. The query result for the homeodomain, a DNA-binding protein domain, is depicted. For each position of the protein domain, the weblogo is shown aligned to a histogram indicating the number of SNPs and disease mutations known in all the human proteins aligned to the domain (from 318 human proteins). In addition, bars underneath each position indicate the functional sites of the domain (e.g. DNA-binding site). The cursor shows data for domain position number 4 for which 12 mutations associated with disease (from OMIM [99] and Swiss-Prot [100]) and one SNP (from dbSNP [138]) are known. The data show only a subset of the results.

Modern approaches integrate multiple molecular features and have been applied to genes from several diseases [102–107]. However, there are still discrepancies among the predictions from the different approaches. In addition, with the exception of the phosphorylation function, these approaches are unable to provide hypotheses for the actual cause of the disease.

Analysis of non-coding SNPs

In addition to the study of coding SNPs, other computational approaches (only briefly mentioned here) focus on SNPs within the non-coding regions of the genome. Non-coding SNPs could be located within TFBSs, microRNA-binding motifs, regulatory-potential sequences or splice sites, and account for most of the human variation found in GWAS [108]. Because SNPs located within the TFBS of a gene may affect the level or timing of the gene expression, computational methods that identify the TFBS-SNPs are valuable resources for selecting candidate regulatory polymorphisms of biomedical significance. An example of such an approach is RAVEN [109], which combines phylogenetic footprinting and TFBS prediction to identify variations in candidate *cis*-regulatory elements. Other methods like UTRScan [110] and FASTSNP [111] also focus their analysis on SNPs within the non-coding regions. UTRScan can be used for the analysis of 5′- and 3′-UTR of eukaryotic mRNAs. UTRScan relies on user-submitted experimental data about the biological activity of functional patterns of UTR sequences to predict whether a particular UTR SNP has functional relevance [110]. FASTSNP has been used to identify intronic SNPs that may lead to defects in RNA and mRNA processing. FASTSNP is based on a decision tree principle to predict whether the SNP has an effect on the TFBS of the gene [111]. SNPs located within two base pairs of an intron–exon junction, or at exonic splicing enhancer (ESE) or exonic splicing silencer (ESS)-binding sites may disrupt mRNA splicing and severely affect proein function [111]. Methods such as ESEfinder or RESCUE ESE can predict ESE motifs [112, 113]. ESS sites can also be predicted using the FAS-ESS method [114]. An excellent review of the different approaches used to predict and identify functional polymorphisms within microRNA-binding sites was provided by Chen *et al*. [115].

Analysis of structural variants

The study of variations of the human genome is not limited to the analysis of SNPs. Other structural variants can also be linked to diseases (see articles [116, 117] and reviews [118, 119]). These structural variants include duplications, inversions and deletions that can currently be identified by array comparative genomic hybridization (aCGH) and paired-end mapping [120–127]. Addressing the need for a common framework of reference for structural variant comparison, Raphael and co-workers proposed a computational approach for the localization of the breakpoints of these modifications [128]. They introduced an algorithm for the identification of data from aCGH and paired-end mapping and provided a framework for comparing structural variants across the different techniques. Advances in the analysis of structural variants will have great implications for the analysis of the human-genome and cancer-genomesequencing projects in the near future.

CONCLUSIONS

The study of disease genes has evolved from basic assumptions that genes follow Mendelian laws to modern computational techniques that are capable of providing insight on hundreds of genes and discriminate particular mutations associated with diseases. The major breakthroughs in the field have lead to general knowledge of the functional, networking and evolutionary properties of disease genes as well as to the identification of genes for specific diseases. Our understanding of the molecular interactions within systems and the phenotypes they are capable of causing could still change dramatically, e.g. by devising the role of microRNAs in normal regulation and disease regulation. Bioinformaticians are addressing the challenges created by the availability of molecular and clinical data produced by new techniques. Integration of data from regulation, interaction and other functional activity of the genes has become essential in medical research.

Ideally, one would like to create the framework for the integration of the experimental, biological and clinical data with existing molecular data and to provide experimental validation of the computational findings. An example of such an approach is the work by Leach *et al*. [129] that introduced a knowledge-based system that combines reading, reasoning and reporting methods to facilitate analysis of experimental data, which was then applied to the analysis of a large-scale gene expression array data sets relevant to craniofacial development. Their tool, Hanalyzer, provided functional hypotheses regarding the role of four genes (Apobec2, E430002G05Rik, Hoxa2, Zim1) in the development of the murine

tongue. Experimental validation of these results indicated that all four were expressed in the tongue. Further analysis will be required to determine if these genes have specific roles in tongue development and function, and if they act as specific markers for individual components of the intrinsic and extrinsic tongue musculature.

Recently the relations between diseases, phenotype and mechanisms have been exploited in an attempt to identify potential new applications for already approved drugs, which could accelerate drug development and reduce overall costs [130]. Data from several pharmaceutical knowledge sources (Drug Bank, Anatomical Therapeutic Chemical Classification) and molecular networks (BIND, BioGRID, KEGG, HPRD) were aggregated using the Resource Description Framework (RDF) standard. The linked ensemble was analyzed for all associations between genes, phenotypes, diseases, clinical symptoms, drug mechanisms and indications; relations showing substantially greater disease to drug associations via phenotypes and mechanisms were ranked and investigated in more detail. One strong association found was between systemic lupus erythematosus and the breast cancer drug Tamoxifen.

Studies in yeast and other model organisms have led to the development of techniques for the integration of functional data in humans. Troyanskaya and colleagues [131] have recently introduced a Bayesian integration system to provide functional maps for human data. The functional maps are available at http://function.princeton.edu/hefalmp and allow for interactive visualization of large-scale experimental data. Another example of integration of experimental data is the work of Califano and co-workers [132] on the identification of post-translational modulators of transcription factor activity and the integration of networks from multiple sources. For this purpose, the authors created the Modulator Inference by Network Dynamics (MINDy) algorithm [132] and the interactome dysregulation enrichment analysis (IDEA) algorithm. MINDy was recently applied to analyze the interface between signaling pathways and transcriptional networks in human B cells [133]. The IDEA algorithm is focused on the search for interactions (instead of genes) that might affect the disruption causing the diseases, and integrates data from different sources, including protein interactions. It has been successfully used to predict oncogenes and molecular perturbation targets in B-cell lymphomas [134].

REMARKS ABOUT FUTURE DIRECTIONS

The completion of the human genome has changed the way the search for disease genes is performed. In the past, the approach was to focus on one or a few genes at a time. Now, projects like the cancer genome atlas exemplify the efforts to systematically analyze all the gene alterations involved in different cancer types [84]. The next step is to produce a complete picture of the mechanistic aspects of the diseases and the design of drugs against them. For that, a combination of two approaches will be needed: a systematic search and in-depth study of each gene.

The future of the field will be defined by new techniques to integrate large bodies of data from different sources and to incorporate functional information into the analysis of large-scale data. The response of bioinformatics to new experimental techniques brings a new perspective into the analysis of the experimental data, as demonstrated by the advances in the analysis of data from microarray and other technologies. It is expected that this trend will continue with novel approaches to respond to new techniques, such as next-generation sequencing technologies. For instance, the availability of large numbers of individual human genomes will promote the development of computational analyses of rare variants, including the statistical mining of their relations to lifestyles, drug interactions and other factors.

Biomedical research will also be driven by our ability to efficiently mine the large body of existing and continuously generated biomedical data. Text-mining techniques, in particular, when combined with other molecular data, can provide information about gene mutations and interactions and will become crucial to stay ahead of the exponential growth of data generated in biomedical research. Another field that is benefiting from the advances in mining and integration of molecular, clinical and drug analysis is pharmacogenomics [135–137]. *In silico* studies of the relationships between human variations and their effect on diseases will be key to the development of personalized medicine.

In summary, translational bioinformatics has already transformed the search for disease genes and has the potential to become a crucial component of other areas of medical research.

Key Points

- Some properties of the disease genes can distinguish them from other genes: they are longer, with fewer paralogs and more homologs in other species.
- Disease genes tend to interact with each other and to be co-expressed. Also, their network of interaction is significantly different than that of the housekeeping genes.
- Mutations in disease genes can be identified by their ability to disrupt the gene function or structure. The study of sequence conservation, structure and other gene properties provides the basis for *in silico* methodologies to predict whether a mutation is deleterious.
- Integration of expression, interaction, evolutionary and sequence data constitutes one of the most powerful tools in translational bioinformatics.

Acknowledgements

Thanks to all the scientists (included or not in this review) that contributed with their excellent work to the field of research reviewed here. Many thanks to Eric Neumann, Mileidy Gonzalez, Simone Gupta and Predrag Radivojac for their comments on the manuscript; to Richard Blissett for his editorial work; and to Attila Kertesz-Farkas and Tom Peterson for their help with the figures. Thanks to Donna Magglot (NCBI) and Joanna Amberger (OMIM) for kindly providing the data for Figure 1 and to anonymous reviewers for their helpful comments.

FUNDING

National Institutes of Health, National Cancer Institute [grant number: 1K22CA143148 - 01].

References

1. Johannsen W. Om arvelighed i samfund og i rene linier. Oversigt over det Kongelige Danske Videnskabernes Selskabs Forhandlinger. In: Fischer G (ed). *Erblichkeit in Populationen und in reinen Linien*, Vol. 3. Jena. 1903:247–70.
2. Farabee WC. Inheritance of digital malformations in man. In: *Papers of the Peabody Museum of American Archaeology and Ethnology*. Cambridge, Mass: Harvard University, 1905; 65–78.
3. Miki Y, Swensen J, Shattuck-Eidens D, *et al*. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 1994;**266**:66–71.
4. Butte AJ. Translational bioinformatics: coming of age. *J Am Med Inform Assoc* 2008;**15**:709–14.
5. American Medical Informatics Association. AMIA Strategic Plan, 2006. http://www.amia.org/inside/stratplan/ (3 August 2009, date last accessed).
6. Sjoblom T, Jones S, Wood LD, *et al*. The consensus coding sequences of human breast and colorectal cancers. *Science* 2006;**314**:268–74.
7. Parsons DW, Jones S, Zhang X, *et al*. An integrated genomic analysis of human glioblastoma multiforme. *Science* 2008; **321**:1807–12.
8. Jones S, Zhang X, Parsons DW, *et al*. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 2008;**321**:1801–6.
9. Pagano JS, Blaser M, Buendia MA, *et al*. Infectious agents and cancer: criteria for a causal relation. *Semin Cancer Biol* 2004;**14**:453–71.
10. Fan H. A new human retrovirus associated with prostate cancer. *Proc Natl Acad Sci USA* 2007;**104**:1449–50.
11. Collins FS, Drumm ML, Cole JL, *et al*. Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science* 1987;**235**:1046–9.
12. Mushegian AR, Bassett DE Jr, Boguski MS, *et al*. Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proc Natl Acad Sci USA* 1997;**94**:5831–6.
13. Lopez-Bigas N, Ouzounis CA. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res* 2004;**32**:3108–14.
14. Tu Z, Wang L, Xu M, *et al*. Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics* 2006;**7**:31.
15. Domazet-Loso T, Tautz D. An ancient evolutionary origin of genes associated with human genetic diseases. *Mol Biol Evol* 2008;**25**:2699–707.
16. Smith NG, Eyre-Walker A. Human disease genes: patterns and predictions. *Gene* 2003;**318**:169–75.
17. Kondrashov FA, Ogurtsov AY, Kondrashov AS. Bioinformatical assay of human gene morbidity. *Nucleic Acids Res* 2004;**32**:1731–37.
18. Osada N, Mano S, Gojobori J. Quantifying dominance and deleterious effect on human disease genes. *Proc Natl Acad Sci USA* 2009;**106**:841–6.
19. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. *Bioinformatics* 2006;**22**: 2291–7.
20. Iossifov I, Zheng T, Baron M, *et al*. Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network. *Genome Res* 2008;**18**: 1150–62.
21. Kann MG. Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform* 2007;**8**:333–46.
22. Chen JY, Shen C, Sivachenko AY. Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pac Symp Biocomput* 2006;**11**:367–78.
23. Lim J, Hao T, Shaw C, *et al*. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* 2006;**125**:801–14.
24. Gandhi TK, Zhong J, Mathivanan S, *et al*. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet* 2006;**38**: 285–93.
25. Oti M, Brunner H. The modular nature of genetic diseases. *Clin Genet* 2007;**71**:1–11.
26. Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci USA* 2008;**105**:4323–8.
27. Rzhetsky A, Iossifov I, Koike T, *et al*. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform* 2004;**37**: 43–53.

28. Midic U, Oldfield CJ, Dunker AK, *et al*. Protein disorder in the human diseasome: unfoldomics of human genetic diseases. *BMC Genomics* 2009;**10**(Suppl 1):S12.
29. Uversky VN, Oldfield CJ, Midic U, *et al*. Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics* 2009;**10**(Suppl 1):S7.
30. Iakoucheva LM, Brown CJ, Lawson JD, *et al*. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 2002;**323**:573–84.
31. Cheng Y, LeGall T, Oldfield CJ, *et al*. Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry* 2006;**45**:10448–60.
32. Uversky VN. Intrinsic disorder in proteins associated with neurodegenerative diseases. In: Ovádi J, Orosz F (eds). *Protein Folding and Misfolding: Neurodegenerative Diseases*. New York: Springer, 2008;21–75.
33. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* 2008;**37**:215–46.
34. Uversky VN. Amyloidogenesis of natively unfolded proteins. *Curr Alzheimer Res* 2008;**5**:260–87.
35. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2007;**35**:D61–5.
36. Marchler-Bauer A, Anderson JB, Derbyshire MK, *et al*. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Research* 2007;**35**:D237–40.
37. Sam L, Liu Y, Li J, *et al*. Discovery of protein interaction networks shared by diseases. *Pac Symp Biocomput* 2007;**12**: 76–87.
38. Chen J, Aronow BJ, Jegga AG. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics* 2009;**10**:73.
39. Kohler S, Bauer S, Horn D, *et al*. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008;**82**:949–58.
40. Gonzalez G, Uribe JC, Tari L, *et al*. Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. *Pac Symp Biocomput* 2007;**12**:28–39.
41. Freudenberg J, Propping P. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* 2002;**18**(Suppl 2):S110–5.
42. Oti M, Snel B, Huynen MA, *et al*. Predicting disease genes using protein-protein interactions. *J Med Genet* 2006;**43**: 691–8.
43. Lage K, Karlberg EO, Stãrling ZM, *et al*. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007;**25**: 309–16.
44. Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics* 2006;**22**:2800–5.
45. Ma X, Lee H, Wang L, *et al*. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics* 2007;**23**: 215–21.
46. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
47. Goh KI, Cusick ME, Valle D, *et al*. The human disease network. *Proc Natl Acad Sci USA* 2007;**104**:8685–90.
48. Turner FS, Clutterbuck DR, Semple CAM. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol* 2003;**4**:R75.
49. Kanehisa M, Araki M, Goto S, *et al*. KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 2008;**36**:D480–4.
50. Matthews L, Gopinath G, Gillespie M, *et al*. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 2009;**37**:D619–22.
51. Biocarta, http://www.biocarta.com/ (1 December 2009, date last accessed).
52. Karp PD, Ouzounis CA, Moore-Kochlacs C, *et al*. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 2005; **33**:6083–9.
53. Dahlquist KD, Salomonis N, Vranizan K, *et al*. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* 2002;**31**:19–20.
54. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**:15545–50.
55. van't Veer LJ, Dai H, van de Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;**415**:530–6.
56. van de Vijver MJ, He YD, van't Veer LJ, *et al*. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002;**347**:1999–2009.
57. Sotiriou C, Wirapati P, Loi S, *et al*. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 2006;**98**:262–72.
58. Lapointe J, Li C, Higgins JP, *et al*. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci USA* 2004;**101**:811–6.
59. Dhanasekaran SM, Barrette TR, Ghosh D, *et al*. Delineation of prognostic biomarkers in prostate cancer. *Nature* 2001; **412**:822–6.
60. Ma XJ, Salunga R, Tuggle JT, *et al*. Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci USA* 2003;**100**:5974–9.
61. Acharya CR, Hsu DS, Anders CK, *et al*. Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer. *J Am Med Assoc* 2008;**299**:1574–87.
62. Golub TR, Slonim DK, Tamayo P, *et al*. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;**286**:531–7.
63. Troyanskaya O, Cantor M, Sherlock G, *et al*. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001;**17**:520–5.
64. Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. *Nature* 2001;**409**:853–5.
65. Chen J, Bardes EE, Aronow BJ, *et al*. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 2009;**37**:W305–11.
66. van Driel MA, Cuelenaere K, Kemmeren PP, *et al*. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res* 2005;**33**:W758–61.

67. Kelso J, Visagie J, Theiler G, *et al*. eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res* 2003;**13**:1222–30.
68. Smith CL, Goldsmith CA, Eppig JT. The Mammalian Phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol* 2005; **6**:R7.
69. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267–70.
70. Radivojac P, Peng K, Clark WT, *et al*. An integrated approach to inferring gene-disease associations in humans. *Proteins* 2008;**72**:1030–7.
71. Wheeler DL, Barrett T, Benson DA, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2008;**36**:D13–21.
72. Gudivada RC, Qu XA, Chen J, *et al*. Identifying disease-causal genes using semantic web-based representation of integrated genomic and phenomic knowledge. *J Biomed Inform* 2008;**41**:717–29.
73. Hirschman L, Colosimo M, Morgan A, *et al*. Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics* 2005;**6**(Suppl 1):S11.
74. Franke L, Bakel H, Fokkens L, *et al*. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 2006;**78**:1011–25.
75. Adie EA, Adams RR, Evans KL, *et al*. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* 2005;**6**:55.
76. Cooper DN, Stenson PD, Chuzhanova NA. The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms. *Curr Protoc Bioinformatics* 2006; Chapter 1:Unit 1.13.
77. Adie EA, Adams RR, Evans KL, *et al*. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics* 2006;**22**:773–4.
78. Aerts S, Lambrechts D, Maity S, *et al*. Gene prioritization through genomic data fusion. *Nature Biotechnol* 2006;**24**: 537–44.
79. George RA, Liu JY, Feng LL, *et al*. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res* 2006;**34**:e130.
80. Gaulton KJ, Mohlke KL, Vision TJ. A computational system to select candidate genes for complex human traits. *Bioinformatics* 2007;**23**:1132–40.
81. Radivojac P, Baenziger PH, Kann MG, *et al*. Gain and loss of phosphorylation sites in human cancer. *Bioinformatics* 2008;**24**:i241–7.
82. Furney SJ, Calvo B, Larranaga P, *et al*. Prioritization of candidate cancer genes—an aid to oncogenomic studies. *Nucleic Acids Res* 2008;**36**:e115.
83. Karchin R. Next generation tools for the annotation of human SNPs. *Brief Bioinform* 2009;**10**:35–52.
84. Collins FS, Barker AD. Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci Am* 2007;**296**:50–7.
85. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science* 2008;**322**:881–8.
86. Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* 2005;**353**:459–73.
87. Chiti F, Dobson CM. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem* 2006;**75**: 333–66.
88. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet* 2000;**16**:198–200.
89. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res* 2002;**12**: 436–46.
90. Saunders CT, Baker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol* 2002;**322**:891–901.
91. Sunyaev S, Ramensky V, Koch I, *et al*. Prediction of deleterious human alleles. *Hum Mol Genet* 2001;**10**: 591–7.
92. Wang Z, Moult J. Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain. *Proteins* 2003;**53**: 748–57.
93. Wang Z, Moult J. SNPs, protein structure, and disease. *Hum Mutat* 2001;**17**:263–70.
94. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Ann Rev Genomics Hum Genet* 2006;**7**:61–80.
95. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 2002;**30**: 3894–900.
96. Miller MP, Kumar S. Understanding human disease mutations through the use of interspecific genetic variation. *Hum Mol Genet* 2001;**10**:2319–28.
97. Clifford RJ, Edmonson MN, Nguyen C, *et al*. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics* 2004;**20**:1006–14.
98. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;**14**:755–63.
99. Hamosh A, Scott AF, Amberger JS, *et al*. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;**33**: D514–7.
100. Boeckmann B, Bairoch A, Apweiler R, *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003;**31**:365–70.
101. Li B, Krishnan VG, Mort ME, *et al*. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 2009;**25**:2744–50.
102. Kaminker JS, Zhang Y, Watanabe C, *et al*. CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res* 2007;**35**:W595–8.
103. Kaminker JS, Zhang Y, Waugh A, *et al*. Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Res* 2007;**67**:465–73.
104. Mathe E, Olivier M, Kato S, *et al*. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res* 2006;**34**:1317–25.
105. Hon LS, Zhang Y, Kaminker JS, *et al*. Computational prediction of the functional effects of amino acid substitutions

in signal peptides using a model-based approach. *Hum Mutat* 2009;**30**:99–106.

106. George Priya Doss C, Sudandiradoss C, Rajasekaran R, *et al*. Applications of computational algorithm tools to identify functional SNPs. *Funct Integr Genomics* 2008;**8**: 309–16.
107. Gong S, Worth CL, Bickerton GR, *et al*. Structural and functional restraints in the evolution of protein families and superfamilies. *Biochem Soc Trans* 2009;**37**:727–33.
108. Hindorff LA, Sethupathy P, Junkins HA, *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009;**106**:9362–7.
109. Andersen MC, Engstrom PG, Lithwick S, *et al*. In silico detection of sequence variations modifying transcriptional regulation. *PLoS Comput Biol* 2008;**4**:e5.
110. Pesole G, Liuni S. Internet resources for the functional analysis of 5′ and 3′ untranslated regions of eukaryotic mRNAs. *Trends Genet* 1999;**15**:378.
111. Yuan HY, Chiou JJ, Tseng WH, *et al*. FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Res* 2006;**34**: W635–41.
112. Fairbrother WG, Yeo GW, Yeh R, *et al*. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res* 2004;**32**:W187–90.
113. Cartegni L, Wang J, Zhu Z, *et al*. ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res* 2003; **31**:3568–71.
114. Wang Z, Rolish ME, Yeo G, *et al*. Systematic identification and analysis of exonic splicing silencers. *Cell* 2004;**119**: 831–45.
115. Chen K, Song F, Calin GA, *et al*. Polymorphisms in microRNA targets: a gold mine for molecular epidemiology. *Carcinogenesis* 2008;**29**:1306–11.
116. Rodriguez-Revenga L, Mila M, Rosenberg C, *et al*. Structural variation in the human genome: the impact of copy number variants on clinical diagnosis. *Genet Med* 2007; **9**:600–6.
117. Marshall CR, Noor A, Vincent JB, *et al*. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* 2008;**82**:477–88.
118. Sharp AJ, Cheng Z, Eichler EE. Structural variation of the human genome. *Annu Rev Genomics Hum Genet* 2006;**7**: 407–42.
119. Buckley PG, Mantripragada KK, Piotrowski A, *et al*. Copy-number polymorphisms: mining the tip of an iceberg. *Trends Genet* 2005;**21**:315–7.
120. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. *Nature Genet* 2005;**37**(Suppl):S11–17.
121. Pinkel D, Segraves R, Sudar D, *et al*. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet* 1998; **20**:207–11.
122. Volik S, Raphael BJ, Huang G, *et al*. Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res* 2006;**16**:394–404.
123. Iafrate AJ, Feuk L, Rivera MN, *et al*. Detection of large-scale variation in the human genome. *Nature Genet* 2004;**36**: 949–51.
124. Kidd JM, Cooper GM, Donahue WF, *et al*. Mapping and sequencing of structural variation from eight human genomes. *Nature* 2008;**453**:56–64.
125. Korbel JO, Urban AE, Affourtit JP, *et al*. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 2007;**318**:420–6.
126. Sebat J, Lakshmi B, Troge J, *et al*. Large-scale copy number polymorphism in the human genome. *Science* 2004;**305**: 525–8.
127. Tuzun E, Sharp AJ, Bailey JA, *et al*. Fine-scale structural variation of the human genome. *Nature Genet* 2005;**37**: 727–32.
128. Sindi S, Helman E, Bashir A, *et al*. A geometric approach for classification and comparison of structural variants. *Bioinformatics* 2009;**25**:i222–30.
129. Leach SM, Tipney H, Feng W, *et al*. Biomedical discovery acceleration, with applications to craniofacial development. *PLoS Comput Biol* 2009;**5**:e1000215.
130. Qu XA, Gudivada RC, Jegga AG, *et al*. Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships. *BMC Bioinformatics* 2009;**10**(Suppl 5):S4.
131. Huttenhower C, Haley EM, Hibbs MA, *et al*. Exploring the human genome with functional maps. *Genome Res* 2009;**19**: 1093–106.
132. Wang K, Banerjee N, Margolin AA, *et al*. Genome-wide discovery of modulators of transcriptional interactions in human B lymphocytes. *Lec Notes Comput Sci* 2006;**3909**: 348–62.
133. Wang K, Alvarez MJ, Bisikirska BC, *et al*. Dissecting the interface between signaling and transcriptional regulation in human B cells. *Pac Symp Biocomput* 2009;**14**:264–75.
134. Mani KM, Lefebvre C, Wang K, *et al*. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol Syst Biol* 2008;**4**:169.
135. Jiang Z, Zhou Y. Using bioinformatics for drug target identification from the genome. *Am J Pharmacogenomics* 2005;**5**:387–96.
136. Watters JW, McLeod HL. Cancer pharmacogenomics: current and future applications. *Biochim Biophys Acta* 2003; **1603**:99–111.
137. Pahl A, Benediktus E, Chialda L. Pharmacogenomics of asthma. *Current Pharmaceutical Design* 2006;**12**:3195–206.
138. Sherry ST, Ward MH, Kholodov M, *et al*. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001; **29**:308–11.
139. Garrod AE. The Incidence of Alkaptonuria: A Study in Chemical Individuality. *Lancet* 1902;**2**:1616–1620.