PLOS | COMPUTATIONAL BIOLOGY

Chapter 15: Disease Gene Prioritization

Yana Bromberg*

Department of Biochemistry and Microbiology, School of Environmental and Biological Sciences, Rutgers University, New Brunswick, New Jersey, United States of America

Abstract: Disease-causing aberrations in the normal function of a gene define that gene as a disease gene. Proving a causal link between a gene and a disease experimentally is expensive and time-consuming. Comprehensive prioritization of candidate genes prior to experimental testing drastically reduces the associated costs. Computational gene prioritization is based on various pieces of correlative evidence that associate each gene with the given disease and suggest possible causal links. A fair amount of this evidence comes from high-throughput experimentation. Thus, well-developed methods are necessary to reliably deal with the quantity of information at hand. Existing gene prioritization techniques already significantly improve the outcomes of targeted experimental studies. Faster and more reliable techniques that account for novel data types are necessary for the development of new diagnostics, treatments, and cure for many diseases.

This article is part of the "Translational Bioinformatics" collection for *PLOS Computational Biology*.

1. Introduction

In 1904 Dr. James Herrick reported [1] the findings of "peculiar elongated and sickle shaped" red blood cells discovered by Dr. Ernest Irons in a hospital patient afflicted with shortness of breath, heart palpitations, and various other aches and pains. This was the first documented case of sickle cell disease in the United States. Forty years later, in 1949, sickle cell anemia became the first disease to be characterized on a molecular level [2,3]. Thus, implicitly, the first disease-associated gene, coding for beta-globin chain of hemoglobin A, was discovered.

It took another thirty years before in 1983 a study of the DNA of families afflicted with Huntington's disease has revealed its association with a gene on chromosome 4 called huntigtin (HTT) [4]. Huntington's became the first genetic disease mapped using polymorphism information (G8 DNA probe/genetic marker), closely followed by the same year discovery of phenylketonuria association with polymorphisms in a hepatic enzyme phenylalanine hydroxylase [5]. These advances provided a route for predicting the likelihood of disease development and even stirred some worries regarding the possibility of the rise of "medical eugenics" [6]. Interestingly, it took another ten years for HTT's sequence to be identified and for the precise nature of the Huntigton's-associated mutation to be determined [7].

The recent explosion in high-throughput experimental techniques has contributed significantly to the identification of disease-associated genes and mutations. For instance, the latest release of SwissVar [8], a variation centered view of the Swiss-Prot database of genes and proteins [9,10], reports nearly 20 thousand mutations in 35 hundred genes associated with over three thousand broad disease classes. Unfortunately, the improved efficiency in production of association data (*e.g.* genome-wide association studies, GWAS) has not been matched by its similarly improving accuracy. Thus, the sheer quantity of existing but yet unvalidated data resulted in information overflow. While association and linkage studies provide a lot of information, incorporation of other sources of evidence is necessary to narrow down the candidate search space. Computational methods - gene prioritization techniques, are therefore necessary to effectively translate the experimental data into legible disease-gene associations [11].

2. Background

The Merriam-Webster dictionary defines the word "disease" as a "a condition of the living animal or plant body or of one of its parts that impairs normal functioning and is typically manifested by distinguishing signs and symptoms." Thus, disease is defined *with respect to normal function* of said body or body part. Note, that this definition also describes the malfunction of individual cells or cell groups. In fact, many diseases can and should be defined on a cellular level. Understanding a disease, and potentially finding curative or preventive measures, requires answering three questions: (1) What is the affected function? (2) What functional activity levels are considered normal given the environmental contexts? (3) What is the direction and amount of change in this activity necessary to cause the observed phenotype?

Contrary to the view that historically prevailed in classical genetics it is rarely the case that one gene is responsible for one function. Rather, an assembly of genes constitutes a functional module or a molecular pathway. By definition, a molecular pathway leads to some specific end point in cellular functionality via a series of interactions between molecules in the cell. Alterations in any of the normally occurring processes, molecular interactions, and pathways lead to disease. For example, folate metabolism is an important molecular pathway, the disruptions in which have been associated with many disorders including colorectal cancer [12] and coronary heart disease [13]. Because this pathway involves 19 proteins interacting via numerous cycles and feedback loops [14], it is not surprising that there are a

Citation: Bromberg Y (2013) Chapter 15: Disease Gene Prioritization. PLoS Comput Biol 9(4): e1002902. doi:10.1371/journal.pcbi.1002902

Editors: Fran Lewitter, Whitehead Institute, United States of America and Maricel Kann, University of Maryland, Baltimore County, United States of America

Published April 25, 2013

Funding: YB is funded by Rutgers, New Brunswick start-up funding, Gordon and Betty Moore Foundation grant, and USDA-NIFA and NJAES grants Project No. 10150-0228906. The funders had no role in the preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: YanaB@rci.rutgers.edu

What to Learn in This Chapter

- Identification of specific disease genes is complicated by gene pleiotropy, polygenic nature of many diseases, varied influence of environmental factors, and overlying genome variation.
- Gene prioritization is the process of assigning likelihood of gene involvement in generating a disease phenotype. This approach narrows down, and arranges in the order of likelihood in disease involvement, the set of genes to be tested experimentally.
- The gene "priority" in disease is assigned by considering a set of relevant features such as gene expression and function, pathway involvement, and mutation effects.
- In general, disease genes tend to 1) interact with other disease genes, 2) harbor functionally deleterious mutations, 3) code for proteins localizing to the affected biological compartment (pathway, cellular space, or tissue), 4) have distinct sequence properties such as longer length and a higher number of exons, 5) have more orthologues and fewer paralogues.
- Data sources (directly experimental, extracted from knowledge-bases, or text-mining based) and mathematical/computational models used for gene prioritization vary widely.

number of different ways in which it can be broken. The changes in concentrations and/or activity levels of any of the pathway members directly affect the pathway end-products (*e.g.* pyrimidine and/or methylated DNA). The specifics of a given change define the severity and the type of the resulting disease; see Box 1 for discussion on disease types. Moreover, since the view of a single pathway as a discrete and independent entity (with no overlap with other pathways) is an oversimplification, it is increasingly evident that different diseases are also interdependent.

Box 1. Genetic similarities of different disease types.

Diseases can be very generally classified by their associated causes: *pathogenic* (caused by an infection), *environmentally determined* (caused by "inanimate" environmental stressors and deficiencies, such as physical trauma, nutrient deficiency, radiation exposure and sleep deprivation), and *genetically hereditary* or *spontaneous* (defined by germline mutations and spontaneous errors in DNA transcription, respectively). Moreover, certain genotypes are more susceptible to the effects of pathogens and environmental stress, contributing to a deadly interplay between disease causes. Regardless of the cause of disease, its manifestations are defined by the changes in the affected function. For example, cancer is the result of DNA damage occurring in a normal cell and leading toward a growth and survival advantage. The initial damage is generally limited to a fairly small number of mutations in key genes, such as proto-oncogenes and tumor suppressor genes [135]. The method of accumulation of these mutants is not very important. A viral infection may cause cancer by enhancing proto-oncogene function [136] or by inserting viral oncogenes into host cell genome. An inherited genetic variant may disrupt or silence a single allele of a mismatch-repair gene as in Lynch syndrome [137]. Spontaneous transcription errors and influence of environmental factors, *e.g.* continued exposure to high levels of ionizing radiation, may result in oncogene and tumor suppressor-gene mutations leading to the development of cancer [138]. Thus, the same broad types of disease can be caused by the disruption of the same mechanisms or pathways resulting from any of the three types of causes.

3. Interpreting What We Know

Identifying the genetic underpinnings of the observed disease is a major challenge in human genetics. Since disease results from the alteration of normal function, identifying disease genes requires defining molecular pathways whose disrupted functionality is necessary and sufficient to cause the observed disease. The pathway function changes due to the (1) changes in gene expression (*i.e.* quantity and concentration of product), (2) changes in structure of the gene-product (*e.g.* conformational change, binding site obstruction, loss of ligand affinity, etc.), (3) introduction of new pathway members (*e.g.* activation of previously silent genes), and (4) environmental disruptions (*e.g.* increased temperatures due to inflammation or decreased ligand concentrations due to malnutrition). While all members of the affected pathways can be construed as disease genes, the identification of a subset of the true causative culprits is difficult. Obscuring such identification are individual genome variation (*i.e.* the reference definition of "normal" is person-specific), multigenic nature and complex phenotypes of most diseases, varied influence of environmental factors, as well as experimental data heterogeneity and constraints.

Disease genes are most often identified using: (1) genome wide association or linkage analysis studies, (2) similarity or linkage to and co-regulation/co-expression/co-localization with known disease genes, and (3) participation in known disease-associated pathways or compartments. In bioinformatics, these are represented by multiple sources of evidence, both direct, *i.e.* evidence coming from own experimental work and from literature, and indirect, *i.e.* "guilt-by-association" data. The latter means that genes that are in any way related to already established disease-associated genes are promoted in the suspect list. Additionally, implied gene-disease links, such as functional deleteriousness of mutations affecting candidate genes, contributes to establishing associations. The manner in which each guilty association is derived varies from tool to tool and all of them deserve consideration. Very broadly, gene-disease associations are inferred from (Figure 1):

1. *Functional Evidence* – the suspect gene is a member of the same molecular pathways as other disease-genes; inferred from: direct molecular interactions, transcriptional co-(regulation/expression/localization), genetic linkage, sequence/structure similarity, and paralogy (in-species homology resulting from a gene duplication event)
2. *Cross-species Evidence* – the suspect gene has homologues implicated in generating similar phenotypes in other organisms
3. *Same-compartment Evidence* – the suspect gene is active in disease-associated pathways (*e.g.* ion channels), cellular compartments (*e.g.* cell membrane), and tissues (*e.g.* liver).
4. *Mutation Evidence* – suspect genes are affected by functionally deleterious

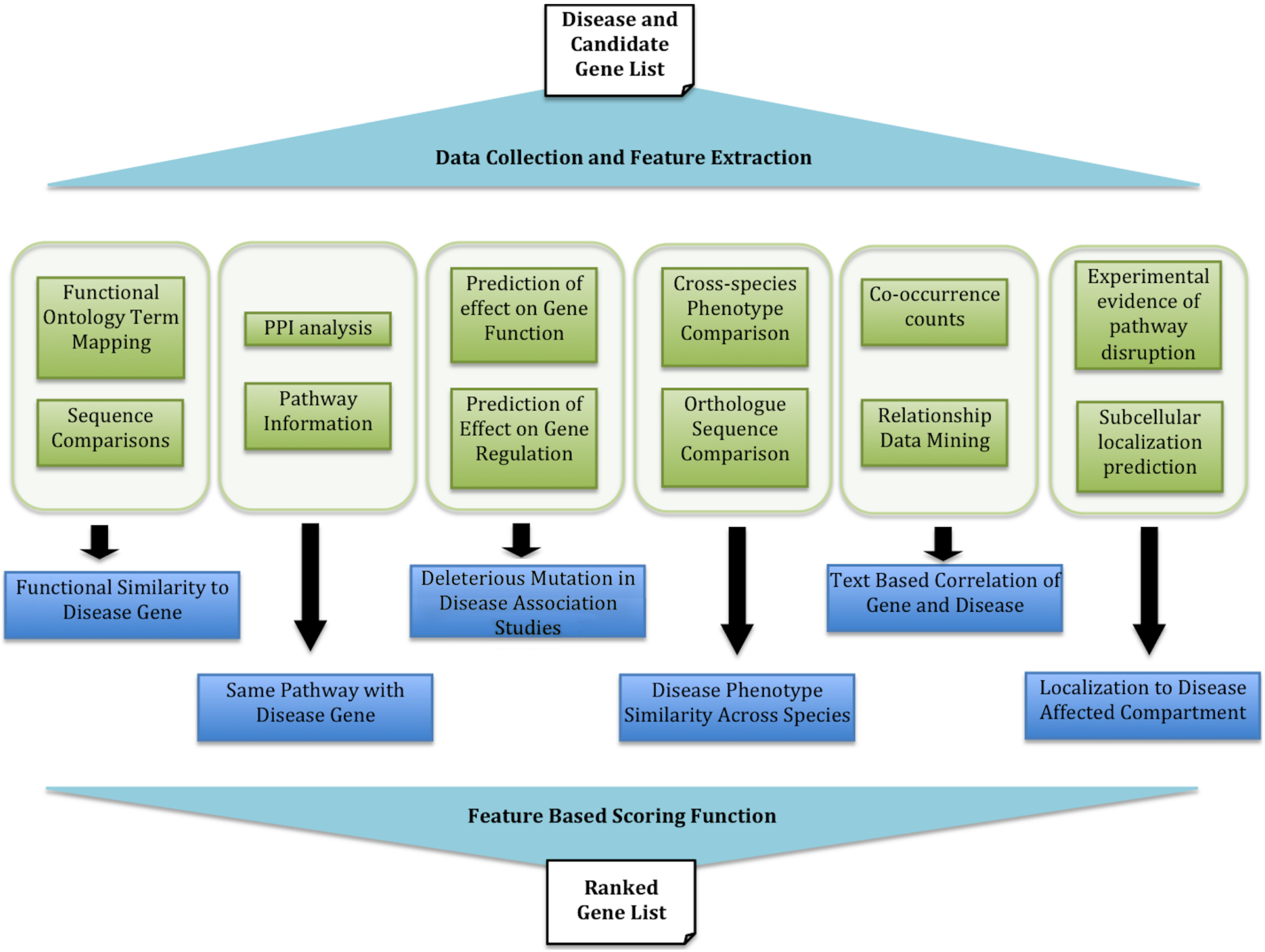

Disease and Candidate Gene List
Data Collection and Feature Extraction
Functional Ontology Term Mapping
Sequence Comparisons
PPI analysis
Pathway Information
Prediction of effect on Gene Function
Prediction of Effect on Gene Regulation
Cross-species Phenotype Comparison
Orthologue Sequence Comparison
Co-occurrence counts
Relationship Data Mining
Experimental evidence of pathway disruption
Subcellular localization prediction
Functional Similarity to Disease Gene
Same Pathway with Disease Gene
Deleterious Mutation in Disease Association Studies
Disease Phenotype Similarity Across Species
Text Based Correlation of Gene and Disease
Localization to Disease Affected Compartment
Feature Based Scoring Function
Ranked Gene List

Figure 1. Overview of gene prioritization data flow. In order to prioritize disease-gene candidates various pieces of information about the disease and the candidate genetic interval are collected (green layer). These describe the biological relationships and concepts (blue layer) relating the disease to the possible causal genes. Note, the blue layer (representing the biological meaning) should ideally be blind to the content green layer (information collection); *i.e.* any resource that describes the needed concepts may be used by a gene prioritization method.
doi:10.1371/journal.pcbi.1002902.g001

mutations in genomes of diseased individuals

5. *Text Evidence* – there is ample co-occurrence of gene and disease terms in scientific texts. Note that textual co-occurrence represents some form of biological evidence, which does not yet lend itself to explicit documentation.

3.1 Functional Evidence

3.1.1 Molecular interactions. Gene prioritization tools, from the earliest field pioneers like G2D [15,16,17] to the more recent ENDEAVOUR [18,19] and GeneWanderer [20,21], among many others, have used gene-gene (protein-protein) interaction and/or pathway information to prioritize candidate genes. Biologically this makes sense, because if diseases result from pathway breakdown then disabling any of the pathway components can produce similar phenotypes; *i.e.* genes responsible for similar diseases often participate in the same interaction networks [22,23]. To illustrate this point, consider the interaction partners of the melanocortin 4 receptor (MC4R) in STRING [24,25] server generated Figure 2. Note, not all known interactions are shown – the inclusion parameter is STRING server likelihood >0.9.

MC4R is a hypothalamic receptor with a primary function of energy homeostasis and food intake regulation. Functionally deleterious polymorphisms in this receptor are known to be associated with severe obesity [26,27,28]. Here, MC1R, MC3R, and MC5R are membrane bound melanocortin (1,3,5) receptors that interact with MC4R via shared binding partners. Syndecan-3 (SDC3), agouti signaling protein precursor (ASIP), agouti related protein precursor (AgRP), pro-opiomelanocortin (POMC) and/or their processed derivatives directly bind MC4R for varied purposes of the MC4R signaling pathway. Finally, the reported interactions with Neuropeptide Y-precursor (NPY) and the growth hormone releasing protein (GHRL) are literature derived and may reflect indirect, but tight connectivity. By the token of "same pathway" evidence, *MC4R* interactors, whether agonists or antagonists, may be predicted to be linked to obesity. In fact, mutations that negatively affect normal POMC production or processing have been shown to be obesity-associated [29,30] and gene association studies have linked AgRP with anorexia and bulimia nervosa behavioral traits [31], representative of food intake abnormalities. Other pathway participants have also been marked and extensively studied for obesity association.

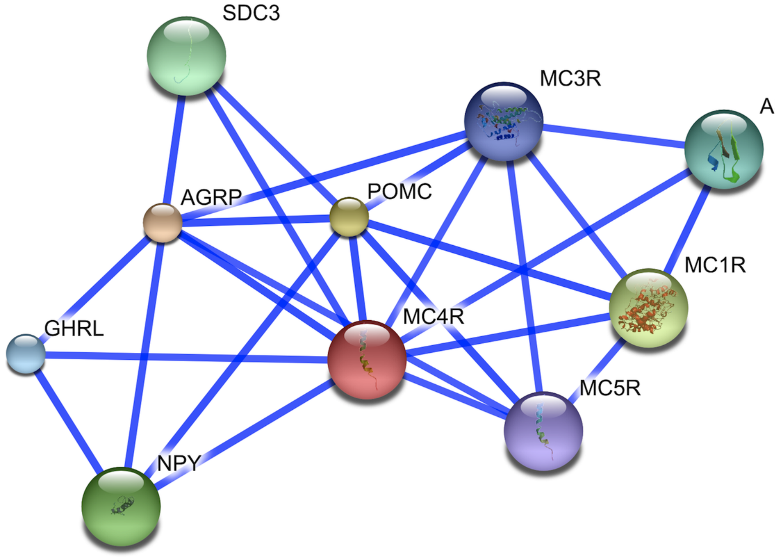


Figure 2. MC4R-centered protein-protein interaction network. The figure illustrates protein-protein interaction neighborhood of the human melanocortin 4 receptor (MC4R) as illustrated by the confidence view of the STRING 8.3 server. The nodes of the graph represent human proteins and the connections illustrate their known or predicted, direct and indirect interactions. The connection between any two protein-nodes is based on the available information mined from relevant databases and literature. The network includes all protein interactions that have >0.9 estimated probability.
doi:10.1371/journal.pcbi.1002902.g002

3.1.2 Regulatory and genetic linkage. Co-regulation of genes has traditionally been thought to point to their involvement in same molecular pathways [32] and, by that token, to similar disease phenotypes; *e.g.* [33,34]. For example, GPR30 a novel G-protein coupled estrogen receptor is co-expressed with the classical estrogen receptor ERβ [33]. The former (GPR30) has been linked to endometrial carcinoma [35] so it is no surprise that the latter (ERβ) is also associated with this type of cancer [33].

However, co-regulation doesn't *always* have to mean the same pathway – studies have shown that consistently co-expressed genes, while possibly genetically linked [36,37], may also reside in distinct pathways [38]. Additionally, co-expressed non-paralogous genes, independent of common pathway involvement, often cluster together in different species and fall into chromosomal regions with low recombination rates [39,40], suggesting genetic linkage [39,40]. These finding suggests that clusters of co-expressed genes are selectively advantageous [36]. Possibly, these clusters are groups of genes that despite the apparent functional heterogeneity may be jointly involved in orchestrating complicated cellular functionality [41]. Evolutionary pressure works on maintaining co-expression of these genes and on keeping recombination rates within the clusters low. Thus, the fine-tuned cooperation of alleles is not broken by recombination, but rather transmitted as one entity to the next generation. Deregulation of these clusters is therefore likely to be deleterious to the organism and develop into disease.

Genes co-expressed with or genetically linked to other disease genes are also likely to be disease-associated. However, while genetic linkage and co-regulation are valuable markers of disease association, they also pose a specificity problem; *i.e.* a given disease-associated gene may be co-regulated with or linked to another disease-associated gene, where the two diseases are not identical. Genetic linkage similarly poses a problem for GWAS where it is difficult to distinguish between "driver" mutations, the actual causes of disease, and "passenger" mutations, co-occurring with the disease-mutations due to genetic linkage.

3.1.3 Similar sequence/structure/function. Reduced or absent phenotypic effect in response to gene knockout/inactivation is a common occurrence [42,43], largely explained by functional compensation, *i.e.* partial interchangeability of paralogous genes. In humans, genes with at least one paralogue, approximated by 90% sequence identity, are about three times less likely to be associated with disease as compared to genes with more remote homologs [44]. However, in the cases where paralogous functional compensation is insufficient to restore normal function, inactivation of any of the paralogues leads to same or similar disease. Prioritization tools thus often use functional similarity as an input feature. For example, one GeneOntology (GO, [45]) defined MC4R function, is "melanocyte-stimulating hormone receptor activity" (GO:0004980). There are two other human gene products sharing this function: MSHR (MC1R, 52% sequence identity) and MC3R (61%). Predictors relying on functional similarity to annotate disease association would inevitably link both of these with obesity. These findings are confirmed by the recent studies for MC3R [46], but the jury still remains out for MC1R involvement.

Quantifying functional similarity is of utmost importance for the above approach. Using ontology-defined functions (*e.g.* GeneOntology) this problem reduces to finding a distance between two ontology nodes/subtrees (*e.g.* [47,48,49,50]). For un-annotated genes, however, sequence and structure homology is often used to transfer functional annotations from studied genes and proteins [51,52]. Since functionally similar genes are likely to produce similar disease phenotypes, sequence/structure similarities are good indicators of similar disease involvement. Additionally, disease genes are often associated with specific gene and protein features such as higher exon number and longer gene length, protein length, presence of signal peptides, higher distance to a neighboring gene and 3′ UTR length, and lower sequence divergence from their orthologues [53,54]. Moreover, disordered proteins are often implicated in cancer [55].

3.2 Cross-species Evidence

Animal models exist for a broad range of human diseases in a number of well-studied laboratory organisms, *i.e.* mouse, zebrafish, fruit fly, etc. However, straightforward cross-species comparisons of orthologues and their associated phenotypic traits are also very useful. A high number of orthologues (consistent presence in multiple species) generally highlights essential genes that are prone to disease involvement. Orthologues generally participate in similar molecular pathways although different levels of function are necessary for different organisms (*e.g.* human MC4R is more functional then its polar bear orthologue [56]). Thus, cross-species tissue-specific phenotypic differentiation due to slightly varied sequences may be useful for gene prioritization. For example, the human MC4R and almost all of its close orthologues (*e.g.* in mouse, rat, pig, and cow) contain a conserved valine residue in the 95th position of the amino acid sequence. In the polar bear orthologue, however, this position is frequently occupied by an isoleucine residue [56]. When considering MC4R involvement in generating an obesity phenotype, it is useful to note that polar bears have a need for increased body fat content for thermal insulation, water buoyancy, and energy storage requirements [56] as compared to humans and to other organisms that share a conserved V95. Thus, one can imagine that the V95I mutation, while deleterious to the function of the receptor, is a polar bear specific adaptation to its environment, and may have a similar (increased body fat) effect in humans. In fact, V95I does inactivate the human receptor [57,58] and associates with obesity.

Comparing human and animal phenotypes is not always straightforward. Washington *et al* [59] have shown that phenotype ontologies facilitate genotype-phenotype comparisons across species. Disease phenotypes recorded in their ontology (OBD, ontology based database) can be compared to the similarly built cross-species phenotype ontologies using a set of proposed similarity metrics. Finding related phenotypes across species suggests orthologous human candidate genes. For instance, phenotypic similarities of eye abnormalities recorded in human and fly suggest that *PAX6*, a human orthologue of the phenotype-associated fly gene *ey*, is a possible disease-gene candidate. Further investigation shows that mutations in *PAX6* may result in aniridia (absence of iris), corneal opacity (aniridia-related keratopathy), cataract (lens clouding), glaucoma, and long-term retinal degeneration (Figure 3) [59].

A correlation of gene co-expression across species is also useful for annotating disease genes [60,61]. Genes that are part of the same functional module are generally co-expressed. Also, there is evidence for co-expression of visibly functionally unrelated genes [37,62,63]. The explanation of these co-expression clusters having an evolutionary advantage only holds true for otherwise unjustified conservation of these clusters throughout different species; *i.e.* cross-species comparison of protein co-expression may be used for validation of disease-gene co-expression inference. Using this assumption, Ala *et al* [61] had narrowed down the initial list of 1,762 genes in the loci mapped via genetic linkage to 850 OMIM (Online Mendelian Inheritance in Man) [64] phenotypes to twenty times fewer (81) possible disease-causing genes. For example, in their analysis a cluster of functionally unrelated genes co-expressed in human and mouse contained a *bona fide* disease-gene KCNIP4 (partial epilepsy with pericentral spikes).

3.3 Compartment Evidence

Changes in gene expression in disease-affected tissues are associated with many complex diseases [65]. Tissue specificity is also important for choosing correct protein-protein interaction networks, as some proteins interact in some tissues, but rarely in others [66]. Disease-associated cellular pathways (*e.g.* ion channels or endocytic membrane transport) and compartments (*e.g.* membrane or nucleus) implicate pathway/compartment-specific gene-products in disease as well. For example, autosomal recessive generalized myotonia (Becker's disease) (GM) and autosomal dominant myotoniacongenita (Thomsen's disease, MC) are characterized by skeletal muscle stiffness [67]. This phenotype is the result of muscle membrane hyperexcitability and, in conjunction with observed alterations in muscle chloride and sodium currents, points to possible involvement of deficiencies of the muscle chloride channel. In fact, studies point to the mutations in the transmembrane region of CLC-1, the muscle chloride channel coding gene, as the culprit [67]. Another example is that of the multiple storage diseases, such as Tay-Sachs, Gaucher, Niemann-Pick and Pompe disease, which are caused by the impairment of the degradation pathways of the intracellular vesicular transport. In fact, many of the genes implicated in these diseases encode for proteins localized to endosomes (*e.g. NPC1* in Neimann-Pick [68]) or lysosomes (*e.g. GBA* [69] in Gaucher, *GAA* in Pompe [70] and *HEXA* in Tay Sachs [71]).

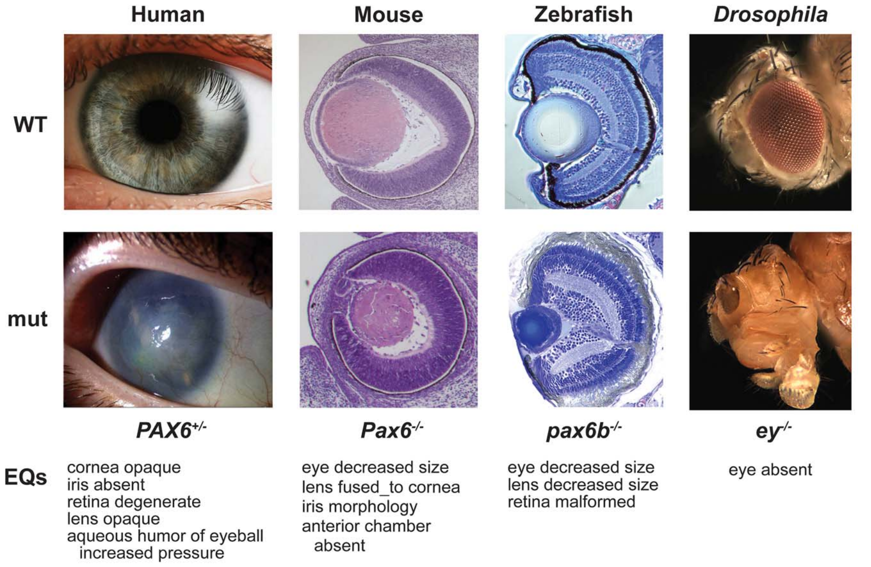


Figure 3. Correlating cross-species phenotypes. Phenotypes of wild-type (top) and PAX6 ortholog mutations (bottom) in human, mouse, zebrafish, and fly can be described with the EQ method suggested by Washington et al [59]. Once phenotypic descriptions are standardized across species, genotypic variations can be assessed as well.
doi:10.1371/journal.pcbi.1002902.g003

3.4 Mutant Evidence

By definition, every genetic disease is associated with some sort of mutation that alters normal functionality. In fact, primary selection of candidates for further analysis is often largely based on observations of polymorphisms in diseased individuals, which are absent in healthy controls (*e.g.* GWAS). However, not all observed polymorphisms are associated with deleterious effects. Note, that on average gain and loss of function mutations are considered to alter normal functionality equally deleteriously. Most of the observed variation does not at all manifest phenotypically, some is weakly deleterious with respect to normal function, and less still is weakly beneficial. In nature strongly beneficial mutations are very rare; they spread rapidly in the population and cannot be considered disease-associated. On the other hand, strongly deleterious or inactivating mutations are often incompatible with life. A small percentage of mutations of this type, affecting genes whose function is not life-essential, are often associated with monogenic Mendelian disorders. Strongly deleterious mutations in the genes whose function may somehow be compensated (*e.g.* via paralogue activity) are associated with complex disorders, where the level of compensation affects the observed phenotype. Complex disorders may also accumulate weakly deleterious mutations to generate a strongly negative phenotype. Intuitively it is clear that a selected candidate gene, carrying a deleterious mutation in an affected individual is more likely to be disease-associated than one which contains functionally neutral mutants or no variation at all.

3.4.1 Structural variation. Structural variation (SV) is the least studied of all types of mutations. It has long been assumed that less than 10% of human genetic variation is in the form of genome structural variants (insertions and deletions, inversions, translocations, aneuploidy, and copy number variations - CNVs). However, because each of the structural variants is large (kb-Mb scale), the total number of base pairs affected by SVs may actually be comparable to the number of base pairs affected by the much more common SNPs (single nucleotide polymorphisms). Moreover, high throughput detection of structural variants is notoriously difficult and is only now becoming possible with better sequencing techniques and CNV arrays. Thus, more SVs may be discovered in the near future. We do not currently know what proportion of genetic disease is caused by SVs, but we suspect that it is high.

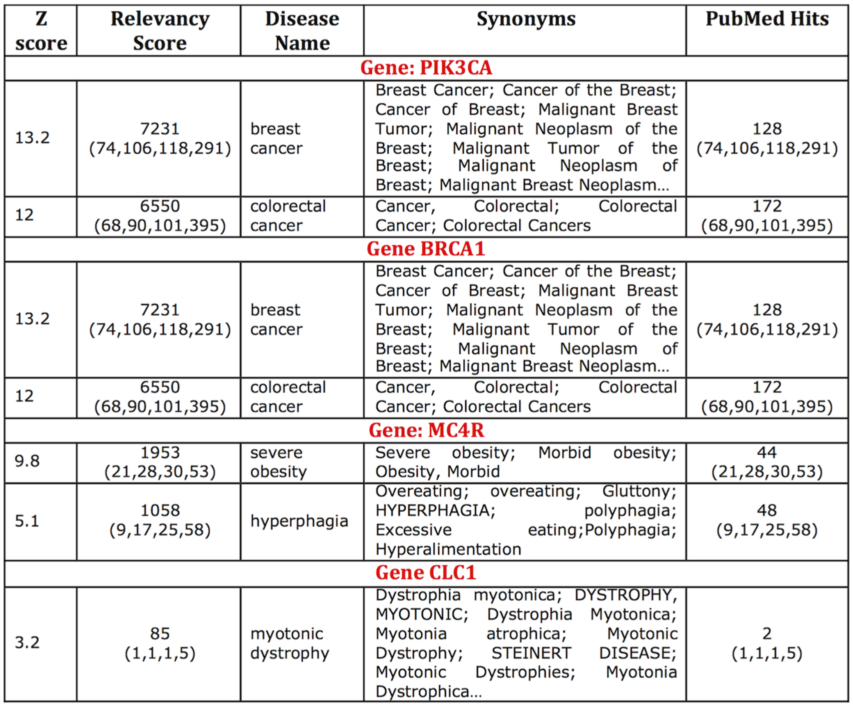
Due to the above mentioned constraints on SV identification, there are only ~180 thousand structural variants reported in one of the most complete mutation collections – the Database of Genomic Variants, DGV [72]. Gross changes to genome sequence are very likely to be disease associated, but also frequently gene non-specific. For instance, Down's syndrome, trisomy 21, is an example of a whole extra chromosome gain and *cri du chat* syndrome results from the deletion of the short arm of chromosome 5 [73]. All of the genes found in these regions of the genome are, by default, associated with the observed disease but neither can be considered primarily causal. When the damage is less extensive the genes involved may be further evaluated for causation. For instance, several epilepsy-associated genes are known, but functionally-significant mutations in these account for only a small fraction of observed disease cases. One study [74] reports that CNV mutants found in epileptic individuals but not in the general population account for nearly nine percent of all cases. Among these are CNVs resulting from deletions in AUTS2 and CNTNAP2 genes. Both of these genes have been implicated in other neurological disorders [75,76] reaffirming the possible disease link. Inversions, translocations and large deletions and insertions have all been implicated in different forms of disease. Even very small indels, resulting in an open-reading frame shift (frameshift mutations), are often sufficient to cause disease. For instance, one of the causes of Tay-Sachs is a deletion of a single cytosine nucleotide in the coding sequence of a lysosomal enzyme beta-hexosaminidase [71].

In most cases of diseases that are associated with SVs the prioritization of disease-causing genes is reduced to finding those that are directly affected by the mutation. Lots of work has been done in this direction, including development of the CNVinetta package [77] for mining and visualizing CNVs, GASV approach for identifying structural variation boundaries more precisely [78], and software created by Ritz *et al* for searching for structural variants in strobe sequencing data [79]. SV identification is still a new field, but the advances in methodologies will have a great impact on our understanding and study of many of the known diseases.

3.4.2 Nucleotide polymorphisms. The other ~90% of human variation exists in the form of SNPs (single nucleotide polymorphisms) and MNPs (multi-nucleotide polymorphisms; consecutive nucleotide substitutions, usually of length two or three). A single human genome is expected to contain roughly 10–15 million SNPs per person [80]. As many as 93% of all human genes contain at least one SNP and 98% of all genes are in the vicinity (~5 kb) of a SNP [81]. The latest release of NCBI dbSNP database [82] (build 137) contains nearly 43 million validated human SNPs, 17.5 million of which have been experimentally mapped to functionally distinct regions of the genome (*i.e.* mRNA UTR, intron, or coding regions). Non-coding region SNPs (~17.2 million) are trivially more prevalent than coding SNPs (~432 thousand) as non-coding DNA makes up the vast majority of the genome. Coding SNPs, however, are over-represented in disease associations; *e.g.* OMIM contains 2430 non-coding SNPs (0.0001% of all) and 5327 coding ones (0.01% of all – 100-fold enrichment). Due to the redundancy of the genetic code, coding SNPs can be further subdivided into synonymous (no effect on protein sequence) and non-synonymous (single amino acid substitution) SNPs. Simple statistics of the genetic code suggest that synonymous SNPs should account for 24% of all coding-region SNPs. dbSNP data suggests an even larger percentage of synonymity – ~188 thousand (44%), which is possibly due to evolutionary pressure eliminating functionally deleterious non-synonymous SNPs. MNPs are rare as compared to SNPs, but are over-represented amongst the protein altering variants, almost always changing the affected amino acid, or two neighboring ones, or introducing a nonsense mutation (stop-codon) [83].

Identifying and annotating functional effects of SNPs and MNPs is important in the context of gene prioritization because genes selected for further disease-association studies are more likely to contain a deleterious mutation or be under the control of one (*e.g.* mutations affecting transcription factor or microRNA binding sites). In recent years a number of methods were created for identifying mutations as functionally deleterious. PromoLign [84], PupaSNP finder [85], and RAVEN [86] look for SNPs affecting transcription, SNPper [87] finds and annotates SNP locations, conservation, and possible functionalities so that they can be visually assessed, and SNPselector [88] and FASTSNP [89] assess various SNP features such as whether it alters the binding site of a transcription factor, affects the promoter/regulatory region, damages the 3′ UTR sequence that may affect post-transcriptional regulation, or eliminates a necessary splice site. Coding synonymous SNPs have recently been shown to have the same chance of being involved in a disease mechanism as non-coding SNPs [90]. This effect may be due to codon usage bias or to changes in splicing or miRNA binding sites [91]. However, few (if any) computational methods are able make predictions with regard to their functional effects.

Non-synonymous SNPs are somewhat more studied. Early termination of the protein is very often associated with disease so genes with nonsense mutants are automatically moved up in the list of possible suspects. Missense SNPs and MNPs, which alter the protein sequence without destroying it, may or may not be disease associated. In fact, most methods estimate that only 25–30% of the nsSNPs negatively affect protein function [92]. Databases like OMIM [93], and more explicitly, SNPdbe [94], SNPeffect [95], PolyDoms [96], Mutation@A Glance [97] and DMDM [98] map SNPs to known structural/functional effects and diseases. Computational tools that make predictions about functional and disease-associated effects of SNPs include SNAP [99,100], SIFT [101,102], PolyPhen [103,104],



Z score	Relevancy Score	Disease Name	Synonyms	PubMed Hits
Gene: PIK3CA				
13.2	7231 (74,106,118,291)	breast cancer	Breast Cancer; Cancer of the Breast; Cancer of Breast; Malignant Breast Tumor; Malignant Neoplasm of the Breast; Malignant Tumor of the Breast; Malignant Neoplasm of Breast; Malignant Breast Neoplasm...	128 (74,106,118,291)
12	6550 (68,90,101,395)	colorectal cancer	Cancer, Colorectal; Colorectal Cancer; Colorectal Cancers	172 (68,90,101,395)
Gene BRCA1				
13.2	7231 (74,106,118,291)	breast cancer	Breast Cancer; Cancer of the Breast; Cancer of Breast; Malignant Breast Tumor; Malignant Neoplasm of the Breast; Malignant Tumor of the Breast; Malignant Neoplasm of Breast; Malignant Breast Neoplasm...	128 (74,106,118,291)
12	6550 (68,90,101,395)	colorectal cancer	Cancer, Colorectal; Colorectal Cancer; Colorectal Cancers	172 (68,90,101,395)
Gene: MC4R				
9.8	1953 (21,28,30,53)	severe obesity	Severe obesity; Morbid obesity; Obesity, Morbid	44 (21,28,30,53)
5.1	1058 (9,17,25,58)	hyperphagia	Overeating; overeating; Gluttony; HYPERPHAGIA; polyphagia; Excessive eating;Polyphagia; Hyperalimentation	48 (9,17,25,58)
Gene CLC1				
3.2	85 (1,1,1,5)	myotonic dystrophy	Dystrophia myotonica; DYSTROPHY, MYOTONIC; Dystrophia Myotonica; Myotonia atrophica; Myotonic Dystrophy; STEINERT DISEASE; Myotonic Dystrophies; Myotonia Dystrophica...	2 (1,1,1,5)

Figure 4. PolySearch gene-disease associations. PolySearch uses PubMed lookup results to prioritize diseases associated with a given gene. Here, screen shots of the top two results (where available; sorted by relevancy score metric) from PolySearch are shown. According to these, BRCA1 and PIK3CA are associated with breast cancer, while MC4R and CLC1 are not. These results quantitatively confirm intuitive inferences made from simple PubMed searches.
doi:10.1371/journal.pcbi.1002902.g004

PHD-SNP [105], SNPs3D [106], and many others. Most of these methods are binary in essence – that is they point to a deficiency without suggesting specifics of the disease or molecular mechanisms of functional failure. Nevertheless, they are very useful in conjunction with other data described above. The recent trend in mutation analysis has seen the development of tools, like SNPNexus [107] and SNPEffectPredictor [108] that are no longer limited by DNA type and predict effects for both non-coding and coding region SNPs.

3.5 Text Evidence

The body of science that addresses gene-disease associations has been growing in leaps and bounds since the mapping of a hemoglobin mutation to sickle cell anemia. Some researchers have been proactive in making their data computationally available from databases like dbSNP, GAD [109], COSMIC [110], *etc.* Others have contributed by depositing knowledge obtained through reading and manual curation into the likes of PMD [111], GeneRIF [112] and UniProt [9]. However, huge amounts of data, which could potentially improve the performance of any gene prioritization method, remains hidden in plain site in natural language text of scientific publications. Consider, for example, a scientist who is interested in prioritizing breast cancer genes. A casual search in PubMed for the term combination *breast cancer* generates over two hundred thousand matches. Limiting the field to *genetics of breast cancer* reduces the count to slightly fewer than fifty thousand. The past thirty days have brought about 46 new papers. Thus, someone interested in getting all the genetic information out of the PubMed collection would need to dedicate his or her life to reading. Fortunately, scientific text mining tools have recently come of age [113,114,115]. The new tools will allow for intelligent identification of possible gene-gene and disease-gene correlations [116,117,118]. For example, the Information Hyperlinked Over Proteins, IHOP method [119] links gene/protein names in scientific texts via associated phenotypes and interaction information. For automated link extraction, however, the existing gene prioritization techniques rely mostly on term co-occurrence statistics (*e.g.* PosMed [120] and GeneDistiller [121]) and gene-function annotations (*e.g.* ENDEAVOR [122] and PolySearch [123]), which can then be related to diseases as described above.

For a significantly oversimplified example of this type of processing consider searching PubMed for the terms *breast cancer* and *BRCA1*. The initial search returns 50 articles, as compared to 21 for *breast cancer* with *BRCA2*, 6 with *PIK3CA*, 1 with *TOX3*, and 0 for *MC4R* or *CLC1* associations. While the number of publications reflects many extraneous factors such as the popularity and "research age" of the protein, it is also very much reflective of the possibility of gene-disease association. Thus, BRCA1 and BRCA2 would be the most likely candidates for cancer association, followed by PIK3CA and TOX3. MC4R and CLC1 would not make the cut. Note that PubMed now defaults to a smart search engine, which identifies all aliases of the gene and the disease while cutting out more promiscuous matches; *i.e.* turning off the translation of terms would result in significantly more less accurate matches. Using specialized tools like PolySearch (or IHOP) to perform the same queries produces more refined and quantifiable results (Figure 4).

4. The Inputs and Outputs

Existing disease-gene prioritization methods vary based on the types of inputs that they use to produce their varied outputs. Functionality of prioritization methods is defined by previously known information about the disease and by candidate search space [124], which may be either submitted by the user or automatically selected by the tool. Disease information is generally limited to lists of known disease-associated genes, affected tissues and pathways and relevant keywords. The candidate search space does not have to be input at all (*i.e.* the entire genome) or be defined by the suspect (for varied experimental reasons) genomic region. The prioritization accuracy, in large part, depends on the accuracy and specificity of the inputs. Thus, providing a list of very broad keywords may reduce the performance specificity, while incorrect candidate search space automatically decreases sensitivity. Prioritization methods generally output ranked/ordered lists of genes, oftentimes associated with p-values, classifier scores, etc.

Overall, input and output requirements and formats are a very important part of establishing a tool's relevance for its users. As with other bioinformatics methods, the ease use and the steepness of learning curve for a given gene prioritization method often define the user base at least as strictly as does its performance.

Box 2. Illustrating basic functionality of a standard (on-line fully-interconnected feed-forward sigmoid-function back-propagating) neural network.

In Figure 5A example network there are three fully interconnected layers of neurons (input, hidden, and output layers); *i.e.* each neuron in one layer is connected to every neuron in the next layer. The three input neurons encode biologically relevant pieces of data relating a given gene G to a given disease D. For each G and D, *i_neuron1* is the fraction of articles (out of 1000) containing in-text co-occurrences of G and D and *i_neuron2* represents the presence/absence of a sequence-similar gene G′ associated with D (*i_neuron3* = G/G′ sequence identity). The hidden (inference) layer consists of two neurons *h_neuron1* and *h_neuron2* with activation thresholds θ_1 and θ_2, respectively. The single output, *o_neuron* (threshold θ_O) represents the involvement of G in causing D: 0 = no involvement, 1 = direct causation. The starting weights of the network ($w_{i1\text{-}h1}$, $w_{i1\text{-}h2}\ldots w_{h2\text{-}o}$) are arbitrarily assigned random values between 0 and 1. Intuitively, the function of the network is to convert input neuron values into output neuron values via a network of weights and hidden neurons. Mathematically, the network is described as follows:

The value (d_x) of neuron x is the sum of inputs into x from the previous layer of neurons ($Y_{i=1\to n}$ in general; in our example: $I_{1\to 3}$, $H_{1\to 2}$). Each of the n inputs is a product of value of neuron Y_i and weight of connection between Y_i and x ($w_{Yi\to x}$).

$$d_x = \sum_{i=1}^{n} Y_i w_{Y_{i\to x}}$$

The value of the output (z_x) of a neuron x based on its d_x and its threshold θ_x is:

$$z_x = f(d_x + \theta_x)$$

In our case, the function (f) is a sigmoid, where a is a real number constant (optimized for any given network, but generally initially chosen to be between 0.5 and 2).

$$f(x) = \frac{1}{1 + e^{-ax}}$$

Thus, to compute the output of every neuron in the network we need to use the formula:

$$z_x = \frac{1}{1 + e^{-a(d_x + \theta_x)}}$$

Note, that to compute the output of the *o_neuron* (z_O; the prediction made by the network) we first have to compute the outputs of all *h_neurons* ($z_{Hi=1\to n}$).

In a supervised learning paradigm, experimentally established pairs of inputs and outputs are given to the network during training (Figure 5C). After each input, the network output (z_O) is compared to the observed result (R). If the network makes a classification error its weights are adjusted to reflect that error. Establishing the best way to update weights and thresholds in response to error is of the major challenges of neural networks. Many techniques use some form of the delta rule – a gradient descent-based optimization algorithm that makes changes to function variables proportionate to the negative of the approximate gradient of the function at the given point. [It's OK if you didn't understand that sentence – the basic idea is to change the weights and thresholds in the direction opposite of the direction of the error]. In our example, we use the delta rule with back-propagation. This means that to compute the error of the hidden layer, the threshold of the output layer (θ_O) and the weights connecting the hidden layer to the output layer ($w_{h1\to O}$, $w_{h2\to O}$) need to be changed first.

The steps are as follows:

1. Compute the error (e_O) of z_O as compared to result R. Note, that the difference between the expected and the observed values defines the gradient (g) at the output neuron.

$$e_O = z_O(1 - z_O)(R - z_O)$$

2. Compute the change in the threshold of the output layer ($\Delta\theta_O$), using a variable λ, the learning rate constant - a real number, often initialized to 0.1–0.2 and optimized for each network)

$$\Delta\theta_O = \lambda e_O$$

3. Compute the change in the weights connecting the hidden layer to the output, $w_{Hi \to O}$.

$$\Delta W_{H_i \to O} = \Delta \theta_O H_i$$

4. Compute the gradient (g_i) at hidden neurons

$$g_i = e_O w_{H_i \to O}$$

Note, from here all steps are the same as above

5. Compute the error at z_{Hi}

$$e_{H_i} = z_{H_i}(1 - z_{H_i}) g_i$$

6. Compute the change in θ_{Hi}

$$\Delta \theta_{H_i} = \lambda e_{H_i}$$

7. Compute the change in $w_{Ij \to Hi}$

$$\Delta W_{I_j \to H_i} = \Delta \theta_{H_i} I_j$$

In on-line updating mode of our example, weights and thresholds are altered after each set of input transmissions. Once the network has "seen" the full set of input/output pairs (one epoch/iteration), training continues re-using the same set until the performance is satisfactory. Note that neural networks are sensitive to dataset imbalance. *I.e.* it is preferable to "balance" the training data, such that the number of instances of each class is presented a roughly equal number of times.

In testing, updating of the weights no longer takes place; *i.e.* the z_O for any given set of inputs is constant over time. See Exercise 8 for an experience with testing. Note, there are many variations on the type and parameters of network learning (propagation mode and direction, weight update rules, thresholds for stopping, etc.) Please consult the necessary literature for more information, *e.g.* [134].

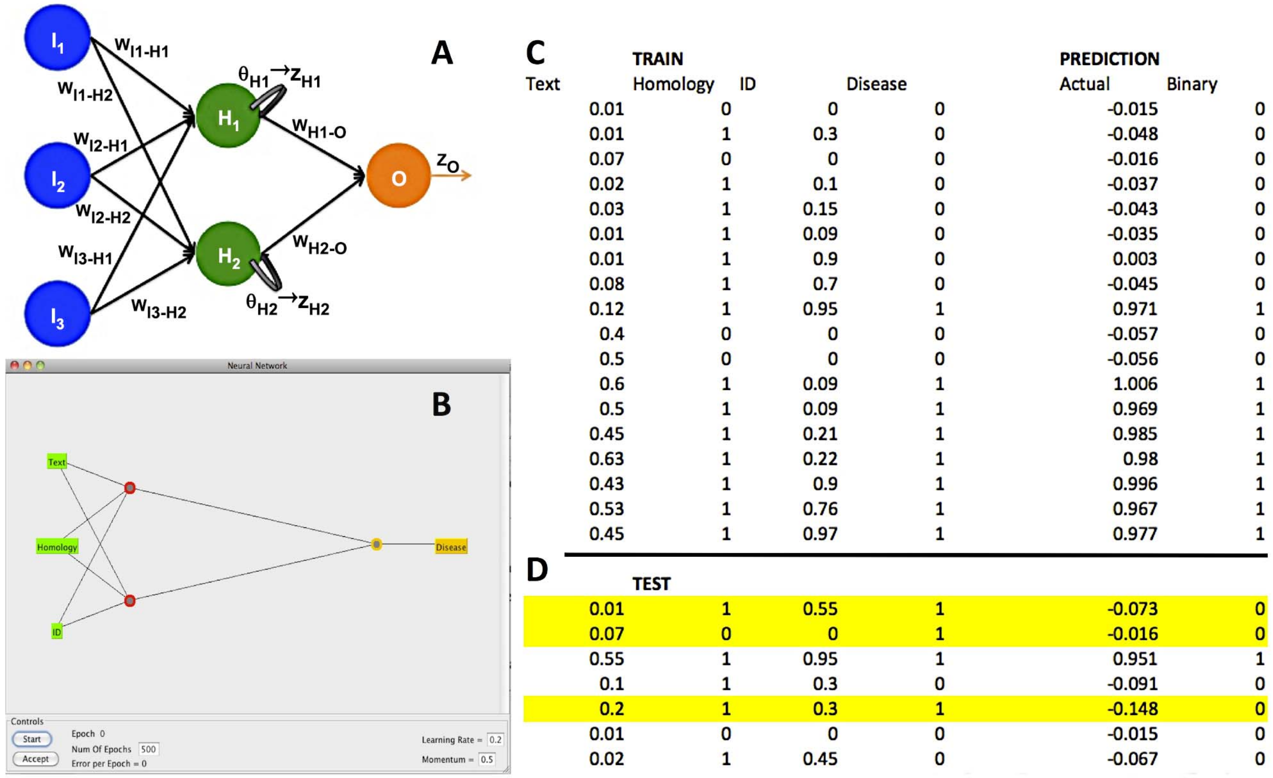
5. The Processing

Gene prioritization methods use different algorithms to make sense of all the data they extract, including mathematical/statistical models/methods (*e.g.* GeneProspector [125]), fuzzy logic (*e.g.* ToppGene [126,127]), and artificial learning devices (*e.g.* PROSPECTR [54]), among others. Some methods use combinations of the above. Objectively, there is no one methodology that is better than the others for all data inputs. For more details on computational methods used in the various approaches please refer to relevant tool publications and method-specific computer science/mathematics literature, *e.g.* [128,129,130,131,132,133,134].

To illustrate the general concepts of relying on the various computational techniques for gene prioritization we will consider the use of an artificial neural network (ANN). Keep in mind that while methods and their requirements differ, the notion of identifying patterns in the data that may be indicative disease-gene involvement remains the same throughout. In simplest terms, a neural network is essentially a mathematical model that defines a function *f*: *X*→*Y*, where a distribution over X (the inputs to the network) is mapped to a distribution over *Y* (the outputs/classifications). The word "network" in the name "artificial neural network" refers to the set of connections between the "neurons" (Figure 5). The functionality of the network is defined by the transmission of signal from activated neurons in one layer to the neurons in another layer via established (and weighed) connections. Besides the choice and number of inputs and outputs, the parameters defining a given ANN are (1) interconnection patterns, (2) the process by which the weights of connections are selected/updated (learning function), and (3) the activation thresholds (functions) of any one given neuron. "Training" a network means optimizing these parameters using an existing set of inputs (and, possibly, outputs). Ultimately, a trained network could then relatively accurately recognize learned patterns in previously unseen data. For more details regarding the possible types and parameters of neural networks see [132,134]. For an illustration of network application see Box 2 and Figure 5.

6. Summary

The development of high throughput technologies has augmented our abilities to identify genetic deficiencies and inconsistencies that lead to the development of diseases. However, a large portion of information in the heaps of data that these methods produce is incomprehensible to the naked eye. Moreover, inferences that could potentially be made from combining



Controls

Epoch 0

Num Of Epochs 500

Error per Epoch = 0

Learning Rate = 0.2

Momentum = 0.5

C

TRAIN				PREDICTION	
Text	Homology	ID	Disease	Actual	Binary
0.01	0	0	0	-0.015	0
0.01	1	0.3	0	-0.048	0
0.07	0	0	0	-0.016	0
0.02	1	0.1	0	-0.037	0
0.03	1	0.15	0	-0.043	0
0.01	1	0.09	0	-0.035	0
0.01	1	0.9	0	0.003	0
0.08	1	0.7	0	-0.045	0
0.12	1	0.95	1	0.971	1
0.4	0	0	0	-0.057	0
0.5	0	0	0	-0.056	0
0.6	1	0.09	1	1.006	1
0.5	1	0.09	1	0.969	1
0.45	1	0.21	1	0.985	1
0.63	1	0.22	1	0.98	1
0.43	1	0.9	1	0.996	1
0.53	1	0.76	1	0.967	1
0.45	1	0.97	1	0.977	1

D

TEST					
0.01	1	0.55	1	-0.073	0
0.07	0	0	1	-0.016	0
0.55	1	0.95	1	0.951	1
0.1	1	0.3	0	-0.091	0
0.2	1	0.3	1	-0.148	0
0.01	0	0	0	-0.015	0
0.02	1	0.45	0	-0.067	0

Figure 5. Predicting gene-disease involvement using artificial neural networks (ANNs). In a supervised learning paradigm, the neural networks are trained using experimental data correlating inputs (descriptive features relating genes to diseases) to outputs (likelihood of gene-disease involvement). The training and testing procedures for the generalized network (Panel A) are described in text. In our example, the WEKA [129,130,131,139] ANN (Panel B; a = 0.5, λ = 0.2) is trained using the training set (Panel C) repeated 500 times (epochs). The network "memorizes" (Predictions in Panel C) the patterns in the training set and is capable of making accurate predictions for four out of seven instances it has not seen before (test set, Panel D). It is important to note here that the erroneously assigned instances (yellow highlight) in the test set are, for the most part, unlike the training. The first one has very little literature correlation (0.01), while sequence similarity to another disease-involved gene is fairly high 0.55). The second maps an unlikely candidate gene (very low literature, no homology) to disease, and the third has barely enough literature mapping and borderline homology. Representation of neither of these instances was consistently present in the training set. This example highlights the importance of training using a representative training set, while testing on a set that is not equivalent to training.

doi:10.1371/journal.pcbi.1002902.g005

different studies and existing research results are beyond reach for anyone of human (not cyborg) descent. Gene prioritization methods (Table 1) have been developed to make sense of this data by extracting and combining the various pieces necessary to link genes to diseases. These methods rely on experimental work such as disease gene linkage analysis and genome wide studies to establish the search space of candidate genes that may possibly be involved in generating the observed phenotype. Further, they utilize mathematical and computational models of disease to filter the original set of genes based on gene and protein sequence, structure, function, interaction, expression, and tissue and cellular localization information. Data repositories that contain the necessary information are diverse in both content and format and require deep knowledge of the stored information to be properly interpreted. Moreover, the models utilizing the various sources assign different weights to the information they extract based on perceived quality and importance of each piece of data available in the context of the entire set of descriptors – a function unlikely to be reproduced in manual data interpretation. Thus, computational gene prioritization techniques serve as interpreters of both of newly retrieved data and of information contained in previous studies. They also are the bridge that connects seemingly unrelated inferences creating an easily comprehensible outlook on an important problem of disease gene annotation.

7. Exercises

1. Search the GAD (http://geneticassociationdb.nih.gov/) database for all genes reported to be associated with diabetes. Refine this set to find only the positively associated genes. How many are there? Why was the total data set reduced? Count the number of unique diabetes associated genes or explain why this is not feasible. How many SNPs associate these genes with diabetes? Is it realistically possible to experimentally evaluate individual effects of each SNP in this set?
2. Using STRING (http://string-db.org/), find ***all*** genes (hint: use limit of 50) interacting with insulin (confidence >0.99). *Note, this confidence limit is extremely high – computational techniques would normally deal with lower limits and thus larger data sets.* What is the insulin gene name used by STRING? How many interaction partners does your query return? Switch to STRING evidence view. Pick three genes connected to insulin via text mining, but without "insulin" in their full name, and find one reference for each in PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) suggesting that these genes are involved with diabetes. Report Gene IDs (*e.g.* MC4R), PubMed IDs and publication citations. Use PolySearch

Table 1. The available data sources and gene prioritization tools.

Data Type	Data Content	Possible Sources	Tools
Experiment, observation	Linkage, association, pedigree, relevant texts and other data	User provided	CAESAR [140], CANDID [141], ENDEAVOR [122], G2D [15,16,17], Gentrepid [142], GeneDistiller [121], PGMapper [143], PRINCE [144], Prioritizer [145], SUSPECTS [146], ToppGene [126,127]
Sequence, structure, meta-data	Sequence conservation, exon number, coding region length, known structural domains and sequence motifs, chromosomal location, protein localization, and other gene-centered information and predictions	SCOP [147], PFam [148,149], ProSite [150], UniProt, Entrez Gene [151], ENSEMBL [152], InterPro [153], LocDB [154], GeneCards [155], PredictProtein [156]	CAESAR, CANDID, ENDEAVOR, G2D, Gentrepid, GeneDistiller, GeneProspector [125], MedSim [157], MimMiner [158], PGMapper, PhenoPred [159], Prioritizer, PROSPECTR [54], SNPs3D [106], SUSPECTS, ToppGene
Pathway, protein-protein interaction, genetic linkage, expression	Disease-gene associations, pathways and gene-gene/protein-protein interactions/ interaction predictions, and gene expression data	KEGG [160,161], STRING, Reactome [162,163], DIP [164], BioGRID [165], GEO [166,167], ArrayExpress [168], ReLiance [169]	CAESAR, CANDID, DiseaseNet [170], ENDEAVOR, G2D, Gentrepid, GeneDistiller, GeneWanderer [20], MaxLink [171], MedSim, PGMapper, PhenoPred, PRINCE, Prioritizer, SNPs3D, SUSPECTS, ToppGene
Non-human data	Information about related genes and phenotypes in other species	OrthoDisease [172], OrthoMCL [173], MGD [174], Pathbase [175]	CAESAR, CANDID, ENDEAVOR, GeneDistiller, GeneProspector, GeneWanderer, MedSim, Prioritizer, PROSPECTR, SNPs3D, SUSPECTS, ToppGene
Ontologies	Gene, disease, phenotype, and anatomic ontologies	GO, DO [176], MPO [177,178], HPO [179], eVOC [180]	CAESAR, ENDEAVOR, G2D, GeneDistiller, MedSim, PhenoPred, Prioritizer, SNPs3D, ToppGene
Mutation associations and effects	Information about existing mutations, their functional and structural effects and their association with diseases, predictions of functional or structural effects for the mutations in the gene in question	dbSNP, PMD [111], GAD, DMDM, SNAP, PolyDoms, SNPdbe, SNPselector, RAVEN, SNPeffect, PHD-SNP, Mutation@A Glance, PromoLign, SIFT, PolyPhen, PupaSNP finder, FASTSNP	CAESAR, CANDID, GeneProspector, GeneWanderer, PROSPECTR, SNPs3D, SUSPECTS
Literature	Mixed information of all types extracted from literature references (*e.g.* disease-gene correlation and non-ontology based gene-function assignment)	PubMed, PubMed Central, HGMD [181], GeneRIF, OMIM	CAESAR, CANDID, DiseaseNet, ENDEAVOR, G2D, Gentrepid, GeneDistiller, GeneProspector, GeneWanderer, MedSim, MimMiner, PGMapper, PolySearch [123], PRINCE, Prioritizer, PROSPECTR, SNPs3D, SUSPECTS, ToppGene

There is a wide range of data sources that can be used to infer the above-described pieces of evidence. The existing tools try to take advantage of many (if not all) of them. This table summarizes the collections and methodologies that make current state of the art in gene prioritization possible. Note, not all resources mentioned here are utilized by all gene prioritization tools nor are all data sources available listed. Moreover, some resources may be classified as more than one data-type. Many of the resources reported here are available electronically through the gene prioritization portal [124].
doi:10.1371/journal.pcbi.1002902.t001

(http://wishart.biology.ualberta.ca/polysearch) **gene** to **disease** mapping with your gene IDs to do the same. Does your experience confirm that the functional "molecular interaction" evidence works? Why?

3. In AmiGO (GO term browser, http://www.geneontology.org), find the human insulin record (hint: use the insulin ID obtained above). What is the Swiss-Prot ID for insulin? Go to the term view. How many GO term associations does insulin have? Reduce the view to "molecular function" terms. How many terms are left? Create a tree view of these terms (hint: use the "Perform an action" dropdown). Which of the terms is the most exact in defining the likely molecular function of insulin (lowest term in a tree hierarchy)? Display gene products in "GO:0005158: insulin receptor binding", reduce the set to human proteins, and look at the inferred tree. How many gene products are in this term? Pick a set of three gene products (report IDs) and use them to search PolySearch for diabetes associations. In question 3 we used the "common pathway" evidence to show the relationship of genes to diabetes. What type of predictive evidence is used here?
4. Search the Mammalian Phenotype Ontology for keyword "diabetes" and select increased susceptibility (MPO, http://www.informatics.jax.org/searches/MP_form.shtml). How many genotypes are returned? Display the genotypes and click on the Aire^{tm1Mand}/Aire^{+} genotype for further exploration. What is the affected gene? Click on gene title (Gene link in Nomenclature section) to display further information. What is an orthologue? What is the human orthologue of your mouse gene? Look up this gene in OMIM (http://www.ncbi.nlm.nih.gov/omim) for association with diabetes. Copy/paste the *citation* from OMIM, describing the gene relationship to diabetes in humans. Do your

results confirm the "cross-species" evidence?

5. Search GeneCards (http://www.genecards.org, utilize advanced search) for genes expressing in the pancreas (hint: pancreatic tissue is often affected in diabetes). How many are there? Explore the GeneCard for CCKBR for diabetes association. Do you find that this gene confirms the "disease compartment" evidence? What database, referenced in GeneCards, contains the CCKBR-diabetes association? Now look at the GeneCard of PLEKHG4. Is there evidence for this gene being associated with diabetes (whether in the GeneCards record or otherwise)? Explain your ideas in detail, paying special attention to the "disease compartment" line of evidence.
6. Search UniProt (http://www.uniprot.org) for all reviewed [reviewed:yes] human [organism:"Homo sapiens [9606]"] protein entries that contain natural variants with reference to diabetes [annotation:(type:natural_variations diabetes)]. Use advanced search with specific limits (*i.e.* sequence annotation, natural_variations, term diabetes). How many proteins fit this description? Locate the entry for insulin (identifier from question 3) and find the total number of known coding variants of this sequence. How many are annotated as associated with any form of diabetes? (hint: read the general annotation section for correspondence of abbreviations to types of diabetes). Run SNAP (http://www.rostlab.org/services/snap/) to predict functional effects of all variants. (hint: use comma separated batch submit). How many are predicted to be functionally non-neutral? Do SNAP predictions of functional effect correlate with annotated disease associations? Does this result confirm the "mutant implication" for nsSNPs?
7. Search PolySearch for all genes associated with diabetes. How many results are returned? Look at the PubMed articles that associate "hemoglobin" with diabetes (follow the link from PolySearch). How many are there? Do you find this number large enough to convince you of hemoglobin-diabetes association and why? From reading article titles/extracted sentences, can you identify a biological reason for connecting hemoglobin to diabetes? If one looks especially convincing, cite that article (hint: its OK to not find one). For the first three articles, can you identify a biological reason for connecting hemoglobin to diabetes? Go back to the list of diabetes related genes and look at TCF7L2 articles. Are the biological reasons for matching TCF7L2 to diabetes clearly defined? Cite the most convincing article. Why do you think TCF7L2 is ranked lower in association than hemoglobin? Is there significant evidence for calcium channel (CACNA1E) involvement in diabetes? Consider the PubMed citations. Do you agree with PolySearch classification of this gene-disease association? Does your experience with PolySearch confirm the "text evidence" function of gene prioritization methods?
8. WEKA exercises (choose one).

   Download and install WEKA (http://www.cs.waikato.ac.nz/~ml/weka/). Using a text-editor (or Microsoft Excel) create comma delimited values (CSV) files identical to the ones described in Figure 5C–D (*i.e.* copy over the training and testing files and replace spaces with commas). Save the files and open the training file in WEKA's Explorer GUI. Open the training file in WEKA's Explorer GUI. You should have four columns of data (Text, Homology, ID, Disease) corresponding to four attributes of each data instance.

   8.1. Defined Questions: Run the MultiLayer Perceptron with parameters (momentum = 0.5, learning = 0.2, trained using the training set, Figure 5C, repeated 500 times/epochs). Test with the test set (Figure 5D) and output predictions for each test entry (make a screenshot). Assuming that everything predicted below 0 is 0, and everything above is 1. What is your performance (number of true/false positives/negatives, positive/negative accuracy/coverage, overall accuracy)? Try using the Decision Stump classifier with default parameters (take screenshot of output). If everything below 0.5 is 0, and everything above is 1, what is your performance? Is it better or worse than the neural net?

   8.2. Open ended: Experiment with different tools available from WEKA's Classify section setting the testing set to your test-file's location. First, run the MultiLayer Perceptron with parameters as described in Figure 5, then try to alter the parameters (momentum term, learning rate, and number of epochs). Try using Linear Regression, Decision Table, or Decision Stump classifiers with default parameters. Is your performance on the test set better or worse? Close the WEKA Explorer, reformat your train/test files in the text editor to replace Disease column values by Booleans (True/False) values, and re-open the training file. Use BayesianNet and RandomForest classifiers to test on the testing file. Does you performance improve? Note, that without further understanding of each of the tools, it is nearly impossible to determine which method is applicable to your data.

Answers to the Exercises can be found in Text S1.

Supporting Information

Text S1 Answers to Exercises.
(DOCX)

Acknowledgments

The author would like to thank Chengsheng Zhu (Rutgers University), Chani Weinreb (Columbia University) and Nikolay Samusik (Max Planck Institute, Dresden) for critical reading and comments to the manuscript. She also acknowledges the help of Gregory Behringer (Rutgers University) and of all of the students of the Spring 2012 Bioinformatics course at Rutgers in testing the exercises.

Glossary

- **Annotation** – any additional information about a genetic sequence. Annotation types are extremely varied, including functional, structural, regulatory, location-related, organism-specific, experimentally derived, predicted, etc.
- **CNV**, copy number variation – an alteration of the genome, which results in an individual having a non-standard number of copies of one or more DNA sections.
- **Gene prioritization** – the process of arranging possible disease causing genes in order of their likelihood in disease involvement.
- **GWAS**, genome wide association studies – the examination of all genes in the genome to correlate their variation to phenotypic trait variation across individuals in a given population.
- **Genetic linkage** – tendency of certain genetic regions on the same chromosome to be inherited together more often than expected due to limited recombination between them.
- **Genetic marker** – a DNA sequence variant with a known location that can be used to identify specific subsets of individuals (cells, species, individual organisms, etc.).
- **Homologue** – a gene derived from a common ancestor with the reference gene. Generally, gene A is a homologue of gene B if both are derived from a common ancestor.
- **Linkage disequilibrium** – tendency of certain genetic regions (not necessarily on the same chromosome) to be inherited together more often that expected from considering their population frequencies. In reference to gene prioritization, this phenomenon may complicate establishment of causal genes due to their consistent inheritance in complex with non-causal genetic regions.
- **Orthologues** – homologous genes separated by a speciation event. Generally, gene A is an orthologue of gene B if A and B are homologous, but reside in different species. Orthologues often perform the same general function in different organisms.
- **Paralogues** – homologous genes separated by a duplication event (often followed by copy differentiation). Generally, gene A is a paralogue of gene B if A and B are homologous and reside in the same species. A and B can be functionally identical or, on contraire, very different, but are often only slightly dissimilar.
- **Pleiotropy** – the influence of a single gene on a number of phenotypic traits.

Further Reading

- Alterovitz G, Ramoni M, eds. (2010) Knowledge-based bioinformatics: from analysis to interpretation. Padstow, Cornwall: John Wiley and Sons Ltd.
- Bromberg Y, Capriotti E, eds. (2012) SNP-SIG 2011: identification and annotation of SNPs in the context of structure, function and disease. Proceedings from SNP-SIG 2011 conference, Vienna, Austria. BMC Genomics 13 Supp 4.
- Chen JY, Youn E, Mooney SD (2009) Connecting protein interaction data, mutations, and disease using bioinformatics. Methods Mol Biol 541: 449–461.
- Dalkilic MM, Costello JC, Clark WT, Radivojac P (2008) From protein-disease associations to disease informatics. Front Biosci 13: 3391–3407.
- Evans JA, Rzhetsky A (2011) Advancing science through mining libraries, ontologies, and communities. J Biol Chem 286: 23659–23666.
- Kann MG (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief Bioinform 8: 333–346.
- Krallinger M, Leitner F, Valencia A (2010) Analysis of biological processes and diseases using text mining approaches. Methods Mol Biol 593: 341–382.
- Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, et al. (2012) The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci 21: 769–785.
- Maulik U, Bandyopadhyay S, Wang JTL, eds. (2010) Computational intelligence and pattern analysis in biological informatics. Hoboken, NJ: John Wiley and Sons, Inc.
- Mooney SD, Krishnan VG, Evani US (2010) Bioinformatic tools for identifying disease gene and SNP candidates. Methods Mol Biol 628: 307–319.
- Moreau Y, Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet 13: 523–536.
- Oti M, Brunner HG (2007) The modular nature of genetic diseases. Clin Genet 71: 1–11.
- Piro RM, Di Cunto F (2007) Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS J 279: 678–696.

References

1. Herrick JB (2001) Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. 1910. Yale J Biol Med 74: 179–184.
2. Pauling L, Itano HA, Singer SJ, Wells IC (1949) Sickle cell anemia, a molecular disease. Science 109: 443.
3. Ingram VM (1956) A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. Nature 178: 792–794.
4. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306: 234–238.
5. Woo SL, Lidsky AS, Guttler F, Chandra T, Robson KJ (1983) Cloned human phenylalanine hydroxylase gene allows prenatal diagnosis and carrier detection of classical phenylketonuria. Nature 306: 151–155.
6. Robertson M (1984) Towards a medical eugenics? Br Med J (Clin Res Ed) 288: 429–430.
7. (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell 72: 971–983.
8. Yip YL, Famiglietti M, Gos A, Duek PD, David FP, et al. (2008) Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Hum Mutat 29: 361–366.
9. UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: D142–148.
10. Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28: 45–48.
11. Moreau Y, Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature reviews Genetics 13: 523–536.
12. Potter JD (1999) Colorectal cancer: molecules and populations. J Natl Cancer Inst 91: 916–932.
13. Frosst P, Blom HJ, Milos R, Goyette P, Sheppard CA, et al. (1995) A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. Nat Genet 10: 111–113.
14. Thomas DC, Conti DV, Baurley J, Nijhout F, Reed M, et al. (2009) Use of pathway information in molecular epidemiology. Hum Genomics 4: 21–42.
15. Perez-Iratxeta C, Bork P, Andrade MA (2002) Association of genes to genetically inherited diseases using data mining. Nat Genet 31: 316–319.
16. Perez-Iratxeta C, Bork P, Andrade-Navarro MA (2007) Update of the G2D tool for prioritization of gene candidates to inherited diseases. Nucleic Acids Res 35: W212–216.
17. Perez-Iratxeta C, Bork P, Andrade MA (2010) G2D: Candidate Genes to Inherited Diseases.
18. Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, et al. (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res 36: W377–384.
19. Tranchevent LC, Moreau Y (2009) ENDEAVOUR.
20. Kohler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82: 949–958.
21. Kohler S (2008) GeneWanderer.
22. Sun J, Zhao Z (2010) A comparative study of cancer proteins in the human protein-protein interaction network. BMC Genomics 11 Suppl 3: S5.
23. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, et al. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38: 285–293.
24. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37: D412–416.
25. Snel B, Lehmann G, Bork P, Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 28: 3442–3444.
26. Huszar D, Lynch CA, Fairchild-Huntress V, Dunmore JH, Fang Q, et al. (1997) Targeted disruption of the melanocortin-4 receptor results in obesity in mice. Cell 88: 131–141.
27. Lubrano-Berthelier C, Le Stunff C, Bougneres P, Vaisse C (2004) A homozygous null mutation delineates the role of the melanocortin-4 receptor in humans. J Clin Endocrinol Metab 89: 2028–2032.
28. Farooqi IS, Keogh JM, Yeo GS, Lank EJ, Cheetham T, et al. (2003) Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. N Engl J Med 348: 1085–1095.
29. Challis BG, Coll AP, Yeo GS, Pinnock SB, Dickson SL, et al. (2004) Mice lacking pro-opiomelanocortin are sensitive to high-fat feeding but respond normally to the acute anorectic effects of peptide-YY(3-36). Proc Natl Acad Sci U S A 101: 4695–4700.
30. Yaswen L, Diehl N, Brennan MB, Hochgeschwender U (1999) Obesity in the mouse model of pro-opiomelanocortin deficiency responds to peripheral melanocortin. Nat Med 5: 1066–1070.
31. Helder SG, Collier DA (2011) The genetics of eating disorders. Curr Top Behav Neurosci 6: 157–175.
32. van Noort V, Snel B, Huynen MA (2003) Predicting gene function by conserved co-expression. Trends Genet 19: 238–242.
33. Huang GS, Gunter MJ, Arend RC, Li M, Arias-Pulido H, et al. (2010) Co-expression of GPR30 and ERbeta and their association with disease progression in uterine carcinosarcoma. Am J Obstet Gynecol 203: 242 e241–245.
34. Jesmin J, Rashid MS, Jamil H, Hontecillas R, Bassaganya-Riera J (2010) Gene regulatory network reveals oxidative stress as the underlying molecular mechanism of type 2 diabetes and hypertension. BMC Med Genomics 3: 45.
35. Smith HO, Leslie KK, Singh M, Qualls CR, Revankar CM, et al. (2007) GPR30: a novel indicator of poor survival for endometrial carcinoma. Am J Obstet Gynecol 196: 386 e381–389; discussion 386 e389–311.
36. Elizondo LI, Jafar-Nejad P, Clewing JM, Boerkoel CF (2009) Gene clusters, molecular evolution and disease: a speculation. Curr Genomics 10: 64–75.
37. Spellman PT, Rubin GM (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol 1: 5.
38. Yu CL, Louie TM, Summers R, Kale Y, Gopishetty S, et al. (2009) Two distinct pathways for metabolism of theophylline and caffeine are coexpressed in Pseudomonas putida CBB5. J Bacteriol 191: 4624–4632.
39. Singer GA, Lloyd AT, Huminiecki LB, Wolfe KH (2005) Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol Biol Evol 22: 767–775.
40. Hurst LD, Williams EJ, Pal C (2002) Natural selection promotes the conservation of linkage of co-expressed genes. Trends Genet 18: 604–606.
41. Dawkins R (1976) The Selfish Gene. New York City: Oxford University Press.
42. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63–66.
43. Conant GC, Wagner A (2004) Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans. Proc Biol Sci 271: 89–96.
44. Hsiao TL, Vitkup D (2008) Role of duplicate genes in robustness against deleterious human mutations. PLoS Genet 4: e1000014. doi:10.1371/journal.pgen.1000014
45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
46. Mencarelli M, Walker GE, Maestrini S, Alberti L, Verti B, et al. (2008) Sporadic mutations in melanocortin receptor 3 in morbid obese individuals. Eur J Hum Genet 16: 581–586.
47. Lord PW, Stevens RD, Brass A, Goble CA (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19: 1275–1283.
48. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics 23: 1274–1281.
49. del Pozo A, Pazos F, Valencia A (2008) Defining functional distances over gene ontology. BMC Bioinformatics 9: 50.
50. Schlicker A, Albrecht M (2010) FunSimMat update: new features for exploring functional similarity. Nucleic Acids Res 38: D244–D248.
51. Punta M, Ofran Y (2008) The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS Comput Biol 4: e1000160. doi:10.1371/journal.pcbi.1000160
52. Rentzsch R, Orengo CA (2009) Protein function prediction–the power of multiplicity. Trends Biotechnol 27: 210–219.
53. Lopez-Bigas N, Ouzounis CA (2004) Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res 32: 3108–3114.
54. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS (2005) Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics 6: 55.
55. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 323: 573–584.
56. Staubert C, Tarnow P, Brumm H, Pitra C, Gudermann T, et al. (2007) Evolutionary aspects in evaluating mutations in the melanocortin 4 receptor. Endocrinology 148: 4642–4648.
57. Xiang Z, Litherland SA, Sorensen NB, Proneth B, Wood MS, et al. (2006) Pharmacological characterization of 40 human melanocortin-4 receptor polymorphisms with the endogenous proopiomelanocortin-derived agonists and the agouti-related protein (AGRP) antagonist. Biochemistry 45: 7277–7288.
58. Hinney A, Hohmann S, Geller F, Vogel C, Hess C, et al. (2003) Melanocortin-4 receptor gene: case-control study and transmission disequilibrium test confirm that functionally relevant mutations are compatible with a major gene effect for extreme obesity. J Clin Endocrinol Metab 88: 4258–4267.
59. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, et al. (2009) Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biol 7: e1000247. doi:10.1371/journal.pbio.1000247
60. 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 U S A 100: 605–610.

61. Ala U, Piro RM, Grassi E, Damasco C, Silengo L, et al. (2008) Prediction of human disease genes by human-mouse conserved coexpression analysis. PLoS Comput Biol 4: e1000043. doi:10.1371/journal.pcbi.1000043
62. Michalak P (2008) Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics 91: 243–248.
63. Fukuoka Y, Inaoka H, Kohane IS (2004) Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC Genomics 5: 4.
64. McKusick-Nathans Institute of Genetic Medicine (JHUB, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD) (2010) Online Mendelian Inheritance in Man, OMIM (TM).
65. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217–1224.
66. Jiang B-B, Wang J-G, Wang Y, Xiao J-F (2009) Gene Prioritization for Type 2 Diabetes in Tissue-specific Protein Interaction Networks. Systems Biology 10801131: 319–328.
67. Koch MC, Steinmeyer K, Lorenz C, Ricker K, Wolf F, et al. (1992) The skeletal muscle chloride channel in dominant and recessive human myotonia. Science 257: 797–800.
68. Greer WL, Riddell DC, Gillan TL, Girouard GS, Sparrow SM, et al. (1998) The Nova Scotia (type D) form of Niemann-Pick disease is caused by a G3097→T transversion in NPC1. American journal of human genetics 63: 52–54.
69. Liou B, Kazimierczuk A, Zhang M, Scott CR, Hegde RS, et al. (2006) Analyses of variant acid beta-glucosidases: effects of Gaucher disease mutations. The Journal of biological chemistry 281: 4242–4253.
70. Shieh JJ, Wang LY, Lin CY (1994) Point mutation in Pompe disease in Chinese. Journal of inherited metabolic disease 17: 145–148.
71. Lau MM, Neufeld EF (1989) A frameshift mutation in a patient with Tay-Sachs disease causes premature termination and defective intracellular transport of the alpha-subunit of beta-hexosaminidase. J Biol Chem 264: 21376–21380.
72. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al. (2004) Detection of large-scale variation in the human genome. Nat Genet 36: 949–951.
73. Chen H (2007) Cri du chat syndrome. Medscape Reference. Available: http://emedicine.medscape.com/article/942897-overview. Accessed 16 January 2013.
74. Mefford HC, Muhle H, Ostertag P, von Spiczak S, Buysse K, et al. (2010) Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. PLoS Genet 6: e1000962. doi:10.1371/journal.pgen.1000962
75. Kalscheuer VM, FitzPatrick D, Tommerup N, Bugge M, Niebuhr E, et al. (2007) Mutations in autism susceptibility candidate 2 (AUTS2) in patients with mental retardation. Hum Genet 121: 501–509.
76. Alarcon M, Abrahams BS, Stone JL, Duvall JA, Perederiy JV, et al. (2008) Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. Am J Hum Genet 82: 150–159.
77. Wittig M, Helbig I, Schreiber S, Franke A (2010) CNVineta: a data mining tool for large case-control copy number variation datasets. Bioinformatics 26: 2208–2209.
78. Sindi S, Helman E, Bashir A, Raphael BJ (2009) A geometric approach for classification and comparison of structural variants. Bioinformatics 25: i222–230.
79. Ritz A, Bashir A, Raphael BJ (2010) Structural variation analysis with strobe reads. Bioinformatics 26: 1291–1298.
80. Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33 Suppl: 228–237.
81. Chakravarti A (2001) To a future of genetic medicine. Nature 409: 822–823.
82. Sherry ST, Ward M, Sirotkin K (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9: 677–679.
83. Rosenfeld JA, Malhotra AK, Lencz T (2010) Novel multi-nucleotide polymorphisms in the human genome characterized by whole genome and exome sequencing. Nucleic Acids Res 38: 6102–6111.
84. Zhao T, Chang LW, McLeod HL, Stormo GD (2004) PromoLign: a database for upstream region analysis and SNPs. Hum Mutat 23: 534–539.
85. Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, et al. (2004) PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. Nucleic Acids Res 32: W242–248.
86. Andersen MC, Engstrom PG, Lithwick S, Arenillas D, Eriksson P, et al. (2008) In silico detection of sequence variations modifying transcriptional regulation. PLoS Comput Biol 4: e5. doi:10.1371/journal.pcbi.0040005
87. Riva A, Kohane IS (2002) SNPper: retrieval and analysis of human SNPs. Bioinformatics 18: 1681–1685.
88. Xu H, Gregory SG, Hauser ER, Stenger JE, Pericak-Vance MA, et al. (2005) SNPselector: a web tool for selecting SNPs for genetic association studies. Bioinformatics 21: 4181–4186.
89. Yuan HY, Chiou JJ, Tseng WH, Liu CH, Liu CK, et al. (2006) FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. Nucleic Acids Res 34: W635–641.
90. Chen R, Davydov EV, Sirota M, Butte AJ (2010) Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. PLoS ONE 5: e13574. doi:10.1371/journal.pone.0013574
91. Parmley JL, Hurst LD (2007) How do synonymous mutations affect fitness? Bioessays 29: 515–519.
92. Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7: 61–80.
93. Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 37: D793–796.
94. Schaefer C, Meier A, Rost B, Bromberg Y (2012) SNPdbe: constructing an nsSNP functional impacts database. Bioinformatics 28: 601–602.
95. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, et al. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. Nucleic Acids Res 33: D527–532.
96. Jegga AG, Gowrisankar S, Chen J, Aronow BJ (2007) PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease. Nucleic Acids Res 35: D700–706.
97. Hijikata A, Raju R, Keerthikumar S, Ramabadran S, Balakrishnan L, et al. (2010) Mutation@A Glance: an integrative web application for analysing mutations from human genetic diseases. DNA Res 17: 197–208.
98. Peterson TA, Adadey A, Santana-Cruz I, Sun Y, Winder A, et al. (2010) DMDM: domain mapping of disease mutations. Bioinformatics 26: 2458–2459.
99. Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35: 3823–3835.
100. Bromberg Y, Yachdav G, Rost B (2008) SNAP predicts effect of mutations on protein function. Bioinformatics 24: 2397–2398.
101. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4: 1073–1081.
102. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812–3814.
103. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30: 3894–3900.
104. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248–249.
105. Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22: 2729–2734.
106. Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7: 166.
107. Chelala C, Khan A, Lemoine NR (2009) SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics 25: 655–661.
108. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, et al. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069–2070.
109. Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. Nat Genet 36: 431–432.
110. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, et al. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 39: D945–950.
111. Kawabata T, Ota M, Nishikawa K (1999) The Protein Mutant Database. Nucleic Acids Res 27: 355–357.
112. Mitchell JA, Aronson AR, Mork JG, Folk LC, Humphrey SM, et al. (2003) Gene indexing: characterization and analysis of NLM's GeneRIFs. AMIA Annu Symp Proc: 460–464.
113. Hirschman L, Yeh A, Blaschke C, Valencia A (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 6 Suppl 1: S1.
114. Altman RB, Bergman CM, Blake J, Blaschke C, Cohen A, et al. (2008) Text mining for biology–the way forward: opinions from leading scientists. Genome Biol 9 Suppl 2: S7.
115. Blaschke C, Andrade MA, Ouzounis C, Valencia A (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. Proc Int Conf Intell Syst Mol Biol: 60–67.
116. Laurila JB, Naderi N, Witte R, Riazanov A, Kouznetsov A, et al. (2010) Algorithms and semantic infrastructure for mutation impact extraction and grounding. BMC Genomics 11 Suppl 4: S24.
117. Caporaso JG, Baumgartner WA, Jr., Randolph DA, Cohen KB, Hunter L (2007) MutationFinder: a high-performance system for extracting point mutation mentions from text. Bioinformatics 23: 1862–1865.
118. Mika S, Rost B (2004) NLProt: extracting protein names and sequences from papers. Nucleic Acids Res 32: W634–637.
119. Hoffmann R, Valencia A (2004) A gene network for navigating the literature. Nat Genet 36: 664.
120. Thornblad TA, Elliott KS, Jowett J, Visscher PM (2007) Prioritization of positional candidate

genes using multiple web-based software tools. Twin Res Hum Genet 10: 861–870.

121. Seelow D, Schwarz JM, Schuelke M (2008) GeneDistiller–distilling candidate genes from linkage intervals. PLoS ONE 3: e3874. doi:10.1371/journal.pone.0003874

122. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, et al. (2006) Gene prioritization through genomic data fusion. Nat Biotechnol 24: 537–544.

123. Cheng D, Knox C, Young N, Stothard P, Damaraju S, et al. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Res 36: W399–405.

124. Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, et al. (2011) A guide to web tools to prioritize candidate genes. Briefings in bioinformatics 12: 22–32.

125. Yu W, Wulf A, Liu T, Khoury MJ, Gwinn M (2008) Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. BMC Bioinformatics 9: 528.

126. Chen J, Bardes EE, Aronow BJ, Jegga AG (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res 37: W305–311.

127. Chen J, Xu H, Aronow BJ, Jegga AG (2007) Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinformatics 8: 392.

128. Nilsson N (1997) Artificial Intelligence: A New Synthesis. San Francisco: Morgan Kaufmann Publishers. 513 p.

129. Bouckaert R, Frank E, Hall M, Holmes G, Pfahringer B, et al. (2010) WEKA-experiences with a java open-source project. . Journal of Machine Learning Research 11: 2533–2541.

130. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20: 2479–2481.

131. Gewehr JE, Szugat M, Zimmer R (2007) BioWeka–extending the Weka framework for bioinformatics. Bioinformatics 23: 651–653.

132. Steeb W-H (2008) The nonlinear workbook: chaos, fractals, cellular automata, neural networks, genetic algorithms, gene expression programming, support vector machine, wavelets, hidden Markov models, fuzzy logic with C++, Java and symbolic C++ programs. 4th edition. Singapore: World Scientific Publishing. 628 p.

133. Ben-Gal I (2007) Bayesian networks. In: Ruggeri F, Kennett R, Faltin F, editors. Encyclopedia of statistics in quality and reliability. Chichester, England: John Wiley and Sons.

134. Habra A (2005) neural networks - an introduction. Available: http://www.tek271.com/documents/others/into-to-neural-networks. Accessed 16 January 2013.

135. Sarasin A (2003) An overview of the mechanisms of mutagenesis and carcinogenesis. Mutat Res 544: 99–106.

136. Parsonnet J (1999) Microbes and malignancy : infection as a cause of human cancers. New York: Oxford University Press. xii, 465 p.

137. Hitchins MP (2010) Inheritance of epigenetic aberrations (constitutional epimutations) in cancer susceptibility. Adv Genet 70: 201–243.

138. Williams D (2008) Radiation carcinogenesis: lessons from Chernobyl. Oncogene 27 Suppl 2: S9–18.

139. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, et al. (2009) The WEKA Data Mining Software: an update. SIGKDD Explorations 11: 10–18.

140. Gaulton KJ, Mohlke KL, Vision TJ (2007) A computational system to select candidate genes for complex human traits. Bioinformatics 23: 1132–1140.

141. Hutz JE, Kraja AT, McLeod HL, Province MA (2008) CANDID: a flexible method for prioritizing candidate genes for complex human traits. Genet Epidemiol 32: 779–790.

142. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, et al. (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. Nucleic Acids Res 34: e130.

143. Xiong Q, Qiu Y, Gu W (2008) PGMapper: a web-based tool linking phenotype to genes. Bioinformatics 24: 1011–1013.

144. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R (2010) Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol 6: e1000641. doi:10.1371/journal.pcbi.1000641

145. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, et al. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78: 1011–1025.

146. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics 22: 773–774.

147. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.

148. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211–222.

149. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, et al. (2000) The Pfam protein families database. Nucleic Acids Res 28: 263–266.

150. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, et al. (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38: D161–166.

151. Maglott D, Ostell J, Pruitt KD, Tatusova T (2011) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 39: D52–57.

152. Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2011) Ensembl 2011. Nucleic Acids Res 39: D800–806.

153. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37: D211–215.

154. Rastogi S, Rost B (2011) LocDB: experimental annotations of localization for Homo sapiens and Arabidopsis thaliana. Nucleic Acids Res 39: D230–234.

155. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D (1997) GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13: 163.

156. Rost B, Yachdav G, Liu J (2004) The PredictProtein server. Nucleic Acids Res 32: W321–326.

157. Schlicker A, Lengauer T, Albrecht M (2010) Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. Bioinformatics 26: i561–567.

158. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA (2006) A text-mining analysis of the human phenome. Eur J Hum Genet 14: 535–542.

159. Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, et al. (2008) An integrated approach to inferring gene-disease associations in humans. Proteins 72: 1030–1037.

160. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.

161. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355–360.

162. D'Eustachio P (2011) Reactome knowledgebase of human biological pathways and processes. Methods Mol Biol 694: 49–61.

163. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37: D619–622.

164. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, et al. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30: 303–305.

165. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) The BioGRID Interaction Database: 2011 update. Nucleic Acids Res 39: D698–704.

166. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, et al. (2011) NCBI GEO: archive for functional genomics data sets–10 years on. Nucleic Acids Res 39: D1005–1010.

167. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207–210.

168. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, et al. (2011) ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Res 39: D1002–1004.

169. Iacucci E, Tranchevent LC, Popovic D, Pavlopoulos GA, De Moor B, et al. (2012) ReLiance: a machine learning and literature-based prioritization of receptor–ligand pairings. Bioinformatics 28: i569–i574.

170. Navlakha S, Kingsford C (2010) The power of protein interaction networks for associating genes with diseases. Bioinformatics 26: 1057–1063.

171. Ostlund G, Lindskog M, Sonnhammer EL (2010) Network-based Identification of novel cancer genes. Mol Cell Proteomics 9: 648–655.

172. O'Brien KP, Westerlund I, Sonnhammer EL (2004) OrthoDisease: a database of human disease orthologs. Hum Mutat 24: 112–119.

173. Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189.

174. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA (2008) The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res 36: D724–728.

175. Schofield PN, Gruenberger M, Sundberg JP (2010) Pathbase and the MPATH ontology: community resources for mouse histopathology. Vet Pathol 47: 1016–1020.

176. Osborne JD, Lin S, Zhu L, Kibbe WA (2007) Mining biomedical data using MetaMap Transfer (MMtx) and the Unified Medical Language System (UMLS). Methods Mol Biol 408: 153–169.

177. Smith CL, Eppig JT (2009) The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdiscip Rev Syst Biol Med 1: 390–399.

178. Smith CL, Goldsmith CA, Eppig JT (2005) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol 6: R7.

179. Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, et al. (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 83: 610–615.

180. Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, et al. (2003) eVOC: a controlled vocabulary for unifying gene expression data. Genome Res 13: 1222–1230.

181. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, et al. (2003) Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21: 577–581.