Published in final edited form as:

Methods Mol Biol. 2010 ; 628: 307–319. doi:10.1007/978-1-60327-367-1_17.

Bioinformatic Tools for Identifying Disease Gene and SNP Candidates

Sean D. Mooney, Vidhya G. Krishnan, and Uday S. Evani

Abstract

As databases of genome data continue to grow, our understanding of the functional elements of the genome grows as well. Many genetic changes in the genome have now been discovered and characterized, including both disease-causing mutations and neutral polymorphisms. In addition to experimental approaches to characterize specific variants, over the past decade, there has been intense bioinformatic research to understand the molecular effects of these genetic changes. In addition to genomic experimental assays, the bioinformatic efforts have focused on two general areas. First, researchers have annotated genetic variation data with molecular features that are likely to affect function. Second, statistical methods have been developed to predict mutations that are likely to have a molecular effect. In this protocol manuscript, methods for understanding the molecular functions of single nucleotide polymorphisms (SNPs) and mutations are reviewed and described. The intent of this chapter is to provide an introduction to the online tools that are both easy to use and useful.

Keywords

Single nucleotide polymorphism; SNP; Genetic disease; Candidate gene; Genome; Bioinformatics; Machine learning

1. Introduction

Over the past decade, considerable effort has been placed on understanding how genetic changes give rise to the molecular effects that cause diseases and phenotypes (1–3). These efforts have given rise to many databases, web resources, and tools for prioritizing candidate single nucleotide polymorphisms (SNPs) or hypothesizing the molecular causes of genetic disease. In this review, these resources and online tools are described within the genomic context of the annotations they provide. Most of the focus is on human annotations, although some resources provide insight into SNP data from model organisms such as mouse, fruit fly, or chimpanzee.

There are now many databases that provide access to SNP or disease mutation data. Most SNP data is eventually deposited in the de facto central SNP database, The Single Nucleotide Polymorphism database (dbSNP, <http://www.ncbi.nlm.nih.gov/SNP/>). There are also now many genotype-phenotype databases available as well including the Human Gene Mutation Database (HGMD, <http://www.hgmd.cf.ac.uk/ac/index.php>) (4), Online Mendelian Inheritance in Man (OMIM, <http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim>) (5), the Pharmacogenetics Knowledge Base (PharmGKB, <http://www.pharmgkb.org/>) (6), and database of Genotype and Phenotype (dbGAP, <http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap>) (7). There are also a growing number of databases of resequencing polymorphism data including the SeattleSNPs project (<http://pga.mbt.washington.edu/>) and sequencing of somatic mutations in cancer (8, 9). This has led to a wealth of genetic variation data.

Typically, SNP data is used as a marker in the context of a linkage or population-based association study. Here, we are focusing on SNPs as the elements that cause disease and alter phenotypes through alteration of some molecular function. There are a number of challenges to identifying these so-called functional variants. First, the marker variants themselves are likely in linkage disequilibrium (or linkage, depending on the study) with the causal variant. Second, identification of candidate disease genes may be the first challenge to narrowing a region for SNP prioritization. Finally, our understanding of how SNPs disrupt molecular function is poorly understood. Here, we focus on two important areas, identification of candidate genes that may have causal variants and identification of candidate causal SNPs.

2. Materials

In general, most of the tools here are deployed as a website or a web resource, requiring only a computer with an internet connection. Occasionally, other software may be required. For visualization of protein structure, UCSF Chimera (10) or Delano Scientific PyMOL (<http://delanoscientific.com/>) maybe useful. Some tools require Flash or Scalable Vector Graphics (SVG).

3. Methods

3.1. Prediction of Genes Likely to Cause or be Associated with Disease

A recent disease gene prioritization tool is FitSNPs (Functionally interpolated SNPs; <http://fitsnps.stanford.edu/>) (11). The tool is claimed to provide a new way to distinguish disease-associated genes from false positives in genome-wide association studies. The feature is based on human microarray data, and it reveals the association between gene expression and disease-associated variants.

Another relatively recent addition to the library of tools that use biochemical information to aid in genetic studies are algorithms for identification of candidate disease genes or genes likely to have associated disease causing genetic changes. These approaches are generally supervised, that is they require knowledge of genes that cause a disease. Most tools use several points of data to infer candidates, and here we discuss web-based tools available for prioritization. One comprehensive example is the Endeavour algorithm (<http://homes.esat.kuleuven.be/~bioiuser/endeavour/>), first published in 2006 in *Nature Biotechnology* (12). Endeavour uses a variety of publicly “-omic” features to predict candidate genes including protein interaction, gene expression, function, sequence, and literature. The tool consists of either Java or web-based clients and is easy to use. It requires a list of training set genes and a list of test set genes.

GeneSeeker (<http://www.cmbi.ru.nl/geneseeker/>) (13) produces a list of candidate disease genes based on cytogenetic localization and expression/phenotypic data from various human and mouse databases. GeneSeeker connects to these databases directly online to guarantee the user to be able to access the most recent data instead of having to download the updated repositories periodically. Although this tool is best for Mendelian diseases that show difference in gene expression patterns in affected tissues, it can also be used to predict candidate genes in other complex diseases.

Gene2Disease (G2D, http://www.ogic.ca/projects/g2d_2/) (14), a system that identifies the candidate disease gene by doing a homology search on Gene Ontology (GO, <http://www.geneontology.org/>) (15) annotated disease-associated genes. G2D uses biomedical literature searches and associated disease conditions with GO terms.

The automated server, SUSPECTS (<http://www.genetics.med.ed.ac.uk/suspects/>) (16) combines the scores from PROSPECTR (<http://www.genetics.med.ed.ac.uk/prospectr/>, based on sequence features) (17), Gene Ontology (GO), InterPro (18) and expression libraries to rank candidate genes in large regions of interest. This tool assumes that the candidate genes in general, share similar domains, annotation, and expression pattern. It provides a 3-D graphical output of the region of interest and hyperlinks to enhance the depth of information about the gene.

Transcriptomics of OMIM (TOM, <http://www-micrel.deis.unibo.it/~tom/>) (19) identifies candidate genes involved in inherited diseases. The algorithm uses mapping, expression, and functional online data repositories. This tool, in general, can be used to predict gene-locus and locus-locus query. It offers flexibility to the user to be able to make a choice between expression data alone or functional analysis using GO terms or both to filter candidate disease genes.

PRIORITIZER (<http://pcdoeglas.med.rug.nl/prioritizer/>) (20) uses a Bayesian approach to classify genes that are associated in diseases. This tool uses data from GO, the Kyoto Encyclopedia of Genes and Genomes (KEGG, <http://www.genome.ad.jp/kegg/>) (21), Biomolecular Interaction Network Database (BIND, <http://binddb.org/>) (22), the Human Protein Reference Database (HPRD, <http://www.hprd.org/>) (23, 24), and Reactome, predicted PPI and expression data. Disease genes are identified through common interactions of proteins in multiple disease intervals that have common phenotypes. This method is based on the assumption that candidate genes are functionally closely related.

Gentrepid (<https://www.gentrepid.org/>) (25) aims to improve some of the existing methods for candidate gene prediction by using structural bioinformatics and system biology approaches such as domain comparison, pathways, and protein–protein interaction data. This tool is based on two assumptions. First, newly identified disease genes and the known disease genes participate in the same complex or pathway. Second, candidate genes that have same phenotype as known disease genes have similar functions. Gentrepid is reported to have better performance than the updated version of the G2D tool which outperformed earlier tools.

PhenoPred (<http://www.phenopred.org/>) (26) utilizes publically available protein interaction, gene function, sequence features, and disease information to prioritize genes associated with disease. The authors have automatically mapped protein-disease annotations to the Disease Ontology (SVM) hierarchy. Then, for each disease, a support vector machine is trained using random genes as negative examples. Then, each of the SVMs is applied back to genes not used in training, and the prediction scores are ranked. A web service for all of the annotations is available on the website, and either genes or diseases can be queried. Several of these tools have been compared and used in concert to identify genes in complex diseases including type 2 diabetes and obesity (27). It should be noted that how each of these methods compared to each other is unclear. Each method is listed in Table 1.

It is worth being aware of the drawbacks of using the various features in the described tools. The disadvantage of the tools that rely mainly on GO terms is that GO annotation is not complete due to the ongoing process of annotation and also includes a bias to well characterized or studied diseases. Earlier tools such as SUSPECTS, POCUS (28), and G2D are based on descriptive keyword search to identify candidate disease gene. In the case of prediction tools based on structural characteristics of gene products, one can leave out the specificity of the gene-by-gene insight that is available in the case of ontology based tools. TOM tries to merge both the methods.

3.2. Prioritization of Functional SNPs and Mutations

The first useful approach a researcher should undertake for identification of functional sites near genetic variation data is to identify functional features that reside on or near the site of variability. This will enable hypothesis generation and guide the researcher toward the first experiments to assay a potential functional effect. The first approach is almost always visualization upon a genome browser, such as UCSC Genome Database (<http://genome.ucsc.edu/>) (29) or Ensembl (<http://ensembl.org/>) (30). However, in addition to these resources, several SNP or mutation specific databases have been developed that provide a variety of genomic annotations. Below they are described, separated by the types of genomic features they can provide, such as at the protein level, at the mRNA/transcript level, and at the genomic level. Each of the following resources is generally freely available and can be accessed on the Internet.

3.2.1. Protein Level

3.2.1.1. Protein Structure Annotation: One of the most common annotations of a SNP is identification of its location on a known or predicted protein structure (see review (31) for understanding the importance of protein structure in genetics). Several web-based databases annotate protein structure and provide a variety of services to query, and these include Large Scale human SNP annotation (LS-SNP, <http://modbase.compbio.ucsf.edu/LS-SNP/>) (32), SNPs3D (<http://snps3d.org/>) (33), MutDB (<http://www.mutdb.org/>) (34), and PolyDoms (<http://polydoms.cchmc.org/polydoms/>) (35). LS-SNP stands out as a useful and unique resource because it provides annotations of nsSNPs that have been mapped to homology models from the MODBASE (<http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi>) (36) dataset.

While visualizing protein structure is useful to an expert in the biochemistry of that protein, it may or may not be useful for hypothesizing the effects that an amino acid substitution will have on that site. This is because effects on protein structure can be very subtle and may be visually nonobvious.

3.2.1.2. Annotation of Known Functional Sites: Many bioinformatic tools are available to predict functional sites upon protein sequences and structures. These tools generally are developed in laboratories of individual researchers and are widely distributed. Examples include prediction of catalytic residues in enzymes (37), protein and DNA binding residues (38), and post-translational modifications (39). Several papers have discussed the importance of stability (40), protein interaction (41), and other functions, such as posttranslational modifications, on disease proteins (42). Reviewing all of them is beyond the scope of this chapter. However, there are some resources that integrate several annotations together for a more comprehensive analysis. First, the Universal Protein Resource (Uniprot, <http://www.pir.uniprot.org/>) database (43) contains annotations of both variation (VARIANT features) and sites of interest, such as posttranslational modification sites. Second, several datasets directly integrate genetic variation data and known protein functional sites such as the SNP Function Portal (44) and SNPeffect (<http://snpeffect.vib.be/>) (45). The SNPeffect and PupaSuite (<http://pupasuite.bioinfo.cipf.es/>) (46) tools have been updated to combine annotations and provide predictions of functional site disruption on protein sequences and structures (47).

If any of these predictive tools are used, however, the accuracies of the methods should be scrutinized by referring back to the paper that originally described the method. Again, these methods should be used to hypothesize the effect, and should not be considered definitive or causative, because they generally have high false discovery rates and sensitivity may be low

(1, 2). Further biochemical experiments are almost always required for confirmation. These methods are summarized in Table 2.

3.2.1.3. Prediction of Whether an Amino Acid Substitution Will Affect Protein

Function or Phenotype: Many tools have been developed to prioritize a given amino acid substitution and many analyses have been applied to understanding the effects of nsSNPs and mutations that are not included in the tools below (48–52). These tools are all supervised, that is, they use a training set of positive and negative examples to “learn” sites. They usually use features based on sequence, structure, or known function. Some of these tools use experimental amino acid substitutions as training (49, 50, 53), others use substitutions based on disease-associated human alleles (32, 33, 54, 55). Two of the first published methods were Sorting Intolerant from Tolerant (SIFT, <http://blocks.fhcrc.org/sift/SIFT.html>) (56) and Polymorphism Phenotyping (PolyPhen, <http://genetics.bwh.harvard.edu/pph/>) (55), and both are widely accepted and easy to use. SIFT uses conservation in a multiple sequence alignment as its sole feature, and experimental mutations as its training data. PolyPhen includes protein structure data and other features, while its training is based on human allele data. More recently, other methods have been developed and deployed online, including SNPs3D (33), LS-SNP (32), PMut (<http://mmb2.pcb.ub.es:8080/PMut/>) (54), the SAP prediction method (<http://sapred.cbi.pku.edu.cn/>) (57), Screening for Nonacceptable Polymorphisms (SNAP, <http://cubic.bioc.columbia.edu/services/SNAP/>) (58), Predicting the Amino Acid Replacement Probability (Parepro, <http://www.mobioinform.cn/parepro/>) (59) and Protein Analysis Through Evolutionary Relationships (PANTHER, <http://www.pantherdb.org/>) (60). For a recent comparison of most of these methods, see the review of Ng and Henikoff (2). The SVM utilized by LS-SNP (32) and the method SNAP (58) are two more recent additions to this library of tools that have web sites available for prediction.

Two considerations should be made when choosing a tool to use. First, training sets used for prediction are an important issue to consider when choosing a method; recently, an overview of this issue was published (53). Second, the approach for classification should also be considered, although in general, more recent machine learning approaches appear to be more accurate. Overall, characterizing protein amino acid substitutions remains the most well studied area of predicting the effects of genetic variation. Current research efforts are focusing on improving accuracy through better features, training sets, and classification approaches. The methods described here are summarized in Table 3.

3.2.2. Transcript Level

3.2.2.1. Annotation of Sites that May Affect Splicing: Several resources annotate SNPs with transcript level features. It is well understood that pathogenic mutations can occur in splicing factor binding sites such as intron–exon splice sites, exonic splicing enhancers (ESE) and silencers (ESS). A recent review highlights the importance of splicing function on genetic disease (61). There are now several tools available for annotation of splicing effects including Polymorphism Mining and Annotation Programs (PolyMAPr, <http://pharmacogenomics.wustl.edu/>) (62), SNPseek (<http://snp.wustl.edu/cgi-bin/SNPseek/index.cgi>), PupaSuite (<http://pupasuite.bioinfo.cipf.es/>) (46) and the SNP Function Portal (44). These resources generally use motif or position specific scoring matrix (PSSM) based prediction of splicing signals or known sites in humans or comparative sites in model organisms such as ESEFinder (63).

3.2.3. Genome Level

3.2.3.1. Identification of Genomic Features near a Candidate SNP: It is now clear that genetic variation affects gene expression and can affect phenotype (see introduction of (64)

for brief review). The molecular mechanisms underlying changes in gene expression continue to be unclear, although there are now insights. One challenge in identification of human functional SNPs is that many SNPs may be in linkage disequilibrium (LD) with each other. That is, pairs, or groups, of SNPs may be highly correlated within a population, preventing accurate statistical identification of the causal element. This challenge has kept experimentally validated functional SNPs for use as bioinformatic training data for predicting expression altering SNPs elusive (65).

There are several SNP browsing tools that can identify features in the promoter region and relate that information to SNPs that are present upon them. These include the Ensembl, NCBI, and UCSC genome databases (29, 30, 66), SNPper (67), SNPSeek, SNP@Promoter (<http://variome.kobic.re.kr/SNPatPromoter/>) (68), the SNP Function Portal (44), PupaSuite (46), and PolyMAPr (62). Generally, these tools can provide annotations of sequence conservation from genome alignments, transcription factor binding sites using databases such as the Transcription Factor Database (TRANSFAC, <http://www.biobase-international.com/pages/index.php?id=transfac>) (69), CpG islands, and other genomic features. Other features shown to be of interest, such as microRNA binding sites, are currently not available outside of the genome browsers (70).

3.2.3.2. Identification of SNPs that May Affect Expression of Genes: Although this is still an ongoing area of research, there are now insights into the mechanisms of cis-acting alleles. A recent survey of features for prediction of regulatory SNPs found that distance to the transcription start site, local repetitive content, sequence conservation, allele frequency, and CpG islands were the most important features for discrimination of regulatory SNPs (71). Transacting regulation appears to be more complicated (64).

Accurate prediction of genetic regulatory networks appears to be in its infancy. Recently, sequence based prediction of expression was shown to be feasible in *Drosophila* using the sequences of transcription factor binding sites (72). However, this approach has not been shown to work for changes as small as a SNP.

4. Conclusions

In summary, there are now many resources for prediction of candidate genes (Table 1) and functional SNPs (Tables 2 and 3). Much research has been performed in predicting the effects of protein amino acid substitutions. Many functional SNPs are synonymous or fall outside of coding regions. This has led to more research focus on predicting the effects of these variants, and we are now beginning to understand the features that are important for determining molecular disruption.

Acknowledgments

We are graciously supported by K22LM009135 (PI: Mooney), R01LM009722 (PI: Mooney), P01AG018397 (PI: Econs), U01GM061373 (PI: Flockhart), and the Indiana Genomics Initiative. The Indiana Genomics Initiative (INGEN) is supported in part by the Lilly Endowment.

References

1. Mooney S. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief Bioinform.* 2005; 6:44–56. [PubMed: 15826356]
2. Ng PC, Henikoff S. Predicting the effects of amino Acid substitutions on protein function. *Annu Rev Genomics Hum Genet.* 2006; 7:61–80. [PubMed: 16824020]
3. Steward RE, MacArthur MW, Laskowski RA, Thornton JM. Molecular basis of inherited diseases: a structural perspective. *Trends Genet.* 2003; 19:505–513. [PubMed: 12957544]

4. Cooper DN, Stenson PD, Chuzhanova NA. The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms. *Curr Protoc Bioinformatics*. 2006; Chapter 1(Unit 1.13)
5. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). *Hum Mutat*. 2000; 15:57–61. [PubMed: 10612823]
6. Altman RB. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat Genet*. 2007; 39:426. [PubMed: 17392795]
7. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007; 39:1181–1186. [PubMed: 17898773]
8. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. *Science*. 2006; 314:268–274. [PubMed: 16959974]
9. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. *Nature*. 2007; 446:153–158. [PubMed: 17344846]
10. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem*. 2004; 25:1605–1612. [PubMed: 15264254]
11. Chen R, Morgan AA, Dudley J, Deshpande T, Li L, Kodama K, Chiang AP, Butte AJ. FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. *Genome Biol*. 2008; 9:R170. [PubMed: 19061490]
12. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, et al. Gene prioritization through genomic data fusion. *Nat Biotechnol*. 2006; 24:537–544. [PubMed: 16680138]
13. van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG, Vriend G. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res*. 2005; 33:W758–W761. [PubMed: 15980578]
14. Perez-Iratxeta C, Wjst M, Bork P, Andrade MA. G2D: a tool for mining genes associated with disease. *BMC Genet*. 2005; 6:45. [PubMed: 16115313]
15. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000; 25:25–29. [PubMed: 10802651]
16. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*. 2006; 22:773–774. [PubMed: 16423925]
17. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*. 2005; 6:55.
18. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. New developments in the InterPro database. *Nucleic Acids Res*. 2007; 35:D224–D228. [PubMed: 17202162]
19. Rossi S, Masotti D, Nardini C, Bonora E, Romeo G, Macii E, et al. TOM: a web-based integrated approach for identification of candidate disease genes. *Nucleic Acids Res*. 2006; 34:W285–W292. [PubMed: 16845011]
20. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*. 2006; 78:1011–1025. [PubMed: 16685651]
21. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 1999; 27:29–34. [PubMed: 9847135]
22. Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*. 2003; 31:248–250. [PubMed: 12519993]
23. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, et al. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*. 2004; 32:D497–D501. [PubMed: 14681466]
24. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, et al. Human protein reference database – 2006 update. *Nucleic Acids Res*. 2006; 34:D411–D414. [PubMed: 16381900]
25. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res*. 2006; 34:e130. [PubMed: 17020920]

26. Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, Boyle SM, Mooney SD. An integrated approach to inferring gene-disease associations in humans. *Proteins*. 2008; 72:1030–1037. [PubMed: 18300252]
27. Tiffin N, Adie E, Turner F, Brunner HG, van Driel MA, Oti M, et al. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res*. 2006; 34:3067–3081. [PubMed: 16757574]
28. Turner FS, Clutterbuck DR, Semple CA. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol*. 2003; 4:R75. [PubMed: 14611661]
29. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, et al. The UCSC Genome Browser Database. *Nucleic Acids Res*. 2003; 31:51–54. [PubMed: 12519945]
30. Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, et al. Ensembl 2004. *Nucleic Acids Res*. 2004; 32(Database issue):D468–D470. [PubMed: 14681459]
31. Laskowski RA, Thornton JM. Understanding the molecular machinery of genetics through 3D structures. *Nat Rev Genet*. 2008; 9:141–151. [PubMed: 18160966]
32. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, et al. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*. 2005; 21:2814–2820. [PubMed: 15827081]
33. Yue P, Melamud E, Moul J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*. 2006; 7:166. [PubMed: 16551372]
34. Singh A, Olowoyeye A, Baenziger PH, Dantzer J, Kann MG, Radivojac P, et al. MutDB: update on development of tools for the biochemical analysis of genetic variation. *Nucleic Acids Res*. 2007; 36 (Database issue):D815–D819. [PubMed: 17827212]
35. Jegga AG, Gowrisankar S, Chen J, Aronow BJ. PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease. *Nucleic Acids Res*. 2007; 35:D700–D706. [PubMed: 17142238]
36. Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC, et al. MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res*. 2004; 32(Database issue):D217–D222. [PubMed: 14681398]
37. Youn E, Peters B, Radivojac P, Mooney SD. Evaluation of features for catalytic residue prediction in novel folds. *Protein Sci*. 2006; 16:216–226. [PubMed: 17189479]
38. Ofran Y, Rost B. Predicted protein–protein interaction sites from local sequence information. *FEBS Lett*. 2003; 544:236–239. [PubMed: 12782323]
39. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*. 2004; 32:1037–1049. [PubMed: 14960716]
40. Wang Z, Moul J. SNPs, protein structure, and disease. *Hum Mutat*. 2001; 17:263–270. [PubMed: 11295823]
41. Ye Y, Li Z, Godzik A. Modeling and analyzing three-dimensional structures of human disease proteins. *Pac Symp Biocomput*. 2006; 11:439–446. [PubMed: 17094259]
42. Radivojac P, Baenziger PH, Kann MG, Mort ME, Hahn MW, Mooney SD. Gain and loss of phosphorylation sites in human cancer. *Bioinformatics*. 2008; 24:i241–i247. [PubMed: 18689832]
43. UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res*. 2008; 36:D190–D195. [PubMed: 18045787]
44. Wang P, Dai M, Xuan W, McEachin RC, Jackson AU, Scott LJ, et al. SNP Function Portal: a web database for exploring the function implication of SNP alleles. *Bioinformatics*. 2006; 22:e523–e529. [PubMed: 16873516]
45. Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F. SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics*. 2006; 22:2183–2185. [PubMed: 16809394]
46. Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, Robledo M, Dopazo J. PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res*. 2004; 32:W242–W248. [PubMed: 15215388]

47. Reumers J, Conde L, Medina I, Maurer-Stroh S, Van Durme J, Dopazo J, et al. Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeff and Pupa Suite databases. *Nucleic Acids Res.* 2008; 36:D825–D829. [PubMed: 18086700]
48. Cai Z, Tsung EF, Marinescu VD, Ramoni MF, Riva A, Kohane IS. Bayesian approach to discovering pathogenic SNPs in conserved protein domains. *Hum Mutat.* 2004; 24:178–184. [PubMed: 15241800]
49. Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol.* 2001; 307:683–706. [PubMed: 11254390]
50. Krishnan VG, Westhead DR. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics.* 2003; 19:2199–2209. [PubMed: 14630648]
51. Saunders CT, Baker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol.* 2002; 322:891–901. [PubMed: 12270722]
52. Vitkup D, Sander C, Church GM. The amino-acid mutational spectrum of human genetic disease. *Genome Biol.* 2003; 4:R72. [PubMed: 14611658]
53. Care MA, Needham CJ, Bulpitt AJ, Westhead DR. Deleterious SNP prediction: be mindful of your training data! *Bioinformatics.* 2007; 23:664–672. [PubMed: 17234639]
54. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics.* 2005; 21:3176–3178. [PubMed: 15879453]
55. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 2002; 30:3894–3900. [PubMed: 12202775]
56. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003; 31:3812–3814. [PubMed: 12824425]
57. Ye ZQ, Zhao SQ, Gao G, Liu XQ, Langlois RE, Lu H, Wei L. Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). *Bioinformatics.* 2007; 23:1444–1450. [PubMed: 17384424]
58. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* 2007; 35:3823–3835. [PubMed: 17526529]
59. Tian J, Wu N, Guo X, Guo J, Zhang J, Fan Y. Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. *BMC Bioinformatics.* 2007; 8:450. [PubMed: 18005451]
60. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, et al. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* 2005; 33:D284–D288. [PubMed: 15608197]
61. Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet.* 2007; 8:749–761. [PubMed: 17726481]
62. Freimuth RR, Stormo GD, McLeod HL. PolyMAPr: programs for polymorphism database mining, annotation, and functional analysis. *Hum Mutat.* 2005; 25:110–117. [PubMed: 15643605]
63. Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ, Krainer AR. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum Mol Genet.* 2006; 15:2490–2508. [PubMed: 16825284]
64. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, et al. Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat Genet.* 2003; 35:57–64. [PubMed: 12897782]
65. Hudson TJ. Wanted: regulatory SNPs. *Nat Genet.* 2003; 33:439–440. [PubMed: 12665861]
66. Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* 2001; 29:137–140. [PubMed: 11125071]
67. Riva A, Kohane IS. SNPper: retrieval and analysis of human SNPs. *Bioinformatics.* 2002; 18:1681–1685. [PubMed: 12490454]
68. Kim BC, Kim WY, Park D, Chung WH, Shin KS, Bhak J. SNP@ Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions. *BMC Bioinformatics.* 2008; 9(Suppl 1):S2. [PubMed: 18315851]

69. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 2003; 31:374–378. [PubMed: 12520026]
70. Chen K, Rajewsky N. Natural selection on human microRNA binding sites inferred from SNP data. *Nat Genet.* 2006; 38:1452–1456. [PubMed: 17072316]
71. Montgomery SB, Griffith OL, Schuetz JM, Brooks-Wilson A, Jones SJ. A survey of genomic properties for the detection of regulatory polymorphisms. *PLoS Comput Biol.* 2007; 3:e106. [PubMed: 17559298]
72. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature.* 2008; 451:535–540. [PubMed: 18172436]
73. Kawabata T, Ota M, Nishikawa K. The Protein Mutant Database. *Nucleic Acids Res.* 1999; 27:355–357. [PubMed: 9847227]

Table 1

Bioinformatic tools for prioritization of genetic disease candidate genes

Name	URL	Features
FitSNPs	http://fitsnps.stanford.edu/	Human Gene expression
Endeavour	http://homes.esat.kuleuven.be/~bioiuser/endeavour/	Gene expression, protein interaction, protein sequence and domain, Kyoto Encyclopedia of Genes and Genomes (KEGG), literature, others
Gene2Disease (G2D)	http://www.ogic.ca/projects/g2d_2/	Gene Ontology (GO) and biomedical literature searches using MEDLINE
GeneSeeker	http://www.cmbi.ru.nl/geneseeker/	Cytogenetic localization and gene expression patterns from mouse and human databases
Gentrepid	https://www.gentrepid.org/	Domain comparison, pathways and protein–protein interaction.
PhenoPred	http://www.phenopred.org/	Protein interaction, gene function, sequence features and disease information
PRIORITIZER	http://pcdoeglas.med.rug.nl/prioritizer/	GO, KEGG, Biomolecular Interaction Network Database (BIND), Human Protein Reference Database (HPRD), Reactome, predicted PPI and expression data
SUSPECTS	http://www.genetics.med.ed.ac.uk/suspects/	Sequence, GO, InterPro and expression libraries
Transcriptomics of OMIM (TOM)	http://www-micrel.deis.unibo.it/~tom/	GO, genomic location and expression data

Table 2

Useful annotation resources for characterization and hypothesizing of SNP function. The following resources aggregate many annotations from other resources

		Genome	Transcript	Protein
LS-SNP	http://modbase.compbio.ucsf.edu/LS-SNP/			X
MutDB	http://mutdb.org/			X
PolyDoms	http://polydoms.cchmc.org/			X
PolyMAPr	http://pharmacogenomics.wustl.edu/	X	X	X
PromoLign	http://polly.wustl.edu/promolign/main.html	X		
PupaSuite	http://pupasuite.bioinfo.cipf.es/	X	X	X
SNP function portal	http://brainarray.mbni.med.umich.edu/Brainarray/Database/SearchSNP/snpfunc.aspx	X	X	X
SNP@Promoter	http://variome.kobic.re.kr/SNPatPromoter/	X		
SNPPer	http://snpper.chip.org/	X		
SNPs3D	http://www.snps3d.org/			X
SNPSeek	http://snp.wustl.edu/cgi-bin/SNPseek/index.cgi	X	X	X

Table 3

Tools for predicting functional nonsynonymous single nucleotide polymorphisms (nsSNP)

Name	URL	Training set
LS-SNP	http://modbase.compbio.ucsf.edu/LS-SNP/	Human allele
PANTHER	http://www.pantherdb.org/tools/csnpscoreForm.jsp	Evolution/human allele
Parepro	http://www.mobioinform.cn/parepro/	Human allele
PMut	http://mmb2.pcb.ub.es:8080/PMut/	Human allele
PolyPhen	http://genetics.bwh.harvard.edu/pph/	Human allele
SAPRED	http://sapred.cbi.pku.edu.cn/	Human allele
SIFT	http://blocks.fhcrc.org/sift/SIFT.html	Experimental ^a
SNAP	http://cubic.bioc.columbia.edu/servers/SNAP/	Experimental ^b
SNPs3D	http://www.snps3d.org/	Human allele

^aTraining set consists of saturation mutagenesis experimental data in *LacI*, HIV-1 protease, T4 lysozyme

^bTraining set consists of amino acid substitutions in the Protein Mutant Database (73) and Swiss-Prot