Bioinformatics for personal genome interpretation

Emidio Capriotti*, Nathan L. Nehrt*, Maricel G. Kann* and Yana Bromberg*

Submitted: 19th September 2011; Received (in revised form): 8th November 2011

Abstract
An international consortium released the first draft sequence of the human genome 10 years ago. Although the analysis of this data has suggested the genetic underpinnings of many diseases, we have not yet been able to fully quantify the relationship between genotype and phenotype. Thus, a major current effort of the scientific community focuses on evaluating individual predispositions to specific phenotypic traits given their genetic backgrounds. Many resources aim to identify and annotate the specific genes responsible for the observed phenotypes. Some of these use intra-species genetic variability as a means for better understanding this relationship. In addition, several online resources are now dedicated to collecting single nucleotide variants and other types of variants, and annotating their functional effects and associations with phenotypic traits. This information has enabled researchers to develop bioinformatics tools to analyze the rapidly increasing amount of newly extracted variation data and to predict the effect of uncharacterized variants. In this work, we review the most important developments in the field—the databases and bioinformatics tools that will be of utmost importance in our concerted effort to interpret the human variome.

Keywords: genomic variation; genome interpretation; genomic variant databases; gene prioritization; deleterious variants

INTRODUCTION
In 1990, the Human Genome Project was launched and, almost 14 years later, the complete sequence of the human genome (over 3 billion bp) was made available [1] at an estimated cost of $2.7 billion. Since then, genomic data has been collected at a continually increasing rate (Figure 1). The strategy for relating a genotype to a phenotype experimentally depends on the type of trait or disease being studied. Re-sequencing the associated gene in affected individuals versus a control population can elucidate variants causing Mendelian pathologies. For analyzing complex, multigenic diseases, sequencing all possible disease-associated regions is necessary.

Although many types of genetic variations exist, the Single Nucleotide Variants (SNVs; mutations...
affecting exactly 1 nt in the genomic sequence) are most prevalent [4]. Many common Single Nucleotide Polymorphisms (SNPs; SNVs that substitute a nucleotide by another one and occur in at least 1% of the population) have been extensively evaluated for disease associations. The International HapMap Consortium was established in 2001 to detect the patterns in human DNA sequence variation and to determine their frequencies in different populations. One major goal of the effort was to enable the discovery of disease-related SNPs [5]. In 2007, using linkage data from HapMap, a genome-wide association study (GWAS) on 17,000 samples/500 K SNVs was performed to detect variants associated with seven common diseases [6]. To date, nearly 1,000 GWAS, genotyping at least 100 K SNPs per individual, have been published [7]. Although these studies are important for identifying disease-associated variants, only a few thousand common SNPs have been significantly associated to specific phenotypic traits [8]. Moreover, a systematic analysis of lower frequency, rare SNVs was not possible with GWAS [9].

With the advent of modern high-throughput technology, the cost of sequencing whole genomes has continued to decrease to reach ~$3,000 today. These technological advances have also enabled the sequencing of individual genomes [10, 11] and the establishment of the 1000 Genomes Project Consortium. In the publication of its pilot results, the Consortium reported detecting over 16 million SNVs [4] — data that now needs to be analyzed for its association with various phenotypes. In addition, direct to consumer (DTC) companies are offering limited genotyping to provide insight into personal traits and disease risks [12]. It is expected that in the next few years, we will witness a second phase of the personal genomics wave, where SNP genotyping chips will be replaced by whole-genome sequencing. Despite more comprehensive databases and better methods for the analysis of genetic variants, the problem of genome interpretation is still far from being solved. Thus, the idea of a ‘$1000 genome, $100,000 genome interpretation’ was expressed by the president of the American College of Medical Genetics, Bruce Korf. In this review, we summarize the newly available genetic variation resources, methods for gene prioritization and algorithms for the prediction of variant effects for use in interpreting personal genomes.

DATABASES AND RESOURCES

As next-generation sequencing technologies continue to decrease in cost and increase in throughput, SNP chip-based genotyping will rapidly be replaced...
by whole-exome and whole-genome sequencing. Once an individual’s genotype has been accurately determined, the first step in genotype interpretation is to identify each variant as known or novel, rare or common, and to determine if it has been previously associated with a disease. Identifying the structural and functional context of each variant is also critical for variant prioritization. A number of databases are freely available to aid the variant annotation process.

**Databases for short length and structural variations**

Several databases aid in the classification of variants as either known or novel, and rare or common (Table 1). The National Center for Biotechnology Information (NCBI) dbSNP database [2] is the largest source of short genetic variation data. dbSNP currently contains over 40 million, both common and rare human SNVs, short indels and microsatellites (Build 134, August 2011). Where available, the database also reports SNV clinical significance. The 1000 Genomes Project Consortium is a major contributor of novel variants to dbSNP, aiming to catalog 95% of human variants with an allele frequency of at least 1% in each of five major human population groups. The Consortium is expanding on the work of the International HapMap Project [5] to catalog genetic variation shared within and between members of various populations. So far, over 38 million variant sites have been identified within the framework of this effort (Phase 1 Low Coverage Data, May 2011). In addition, the Consortium data includes inferred genotypes for individual samples, useful for future association studies utilizing genotype imputation. The recently initiated UK10K Project (http://www.uk10k.org) will have even greater power to discover rare variants, identifying those with as little as 0.1% allele frequency. The project will conduct low coverage, whole-genome sequencing for 4000 healthy individuals, and whole-exome sequencing for 6000 individuals with a variety of extreme disease phenotypes to facilitate the discovery of rare variants associated with these diseases. Although more exhaustive in scope, the UK10K data is less accessible—access is managed by a consortium and requires acceptance of ‘terms and conditions’ to protect the privacy and interests of the study participants.

While databases like dbSNP and HapMap and projects like 1000 Genomes and UK10K focus primarily on short-length variants like missense, nonsense and short insertion and deletion mutations (indels), larger-scale structural rearrangements, copy number variants (CNVs) and large indels can also dramatically affect human phenotypes. NCBI’s database of genomic structural variation (dbVar) [13] and the collaborative effort Database of Genomic Variants (DGV) [14] are two of the largest repositories for large-scale (typically >1 kb in length) structural variations. DGV only contains entries from healthy human controls, while dbVar contains entries from all species and includes variants with associated phenotypes. The DGV archive (DGVa) [13] is a new database maintained by the European Bioinformatics Institute that also contains structural variants from all species with associated phenotypes when available.

**Genotype/phenotype annotation databases**

Many specialized databases contain variant–disease associations that are commonly used to identify known deleterious mutations (Table 1). The Online Mendelian Inheritance in Man (OMIM) database [16] is a catalog of human genes and diseases. OMIM is manually curated and contains descriptions of over 13,000 genes and almost 7000 phenotypes (September 2011). Over 2600 genes in OMIM contain listings of specific allelic variants associated with disease. The SwissVar database [3] is another manually curated source of variant–phenotype association data. The database also includes a number of variant features, e.g. physico-chemical properties, affected functional features and conservation profiles for amino-acid changing variants in SwissProt proteins. SwissVar currently contains information on over 24,000 deleterious variants linked to over 3300 diseases (September 2011). The Human Gene Mutation Database (HGMD) [15] is a large collection of variants associated with human inherited diseases. HGMD is available in two versions: a free version for academic/nonprofit users, and a more regularly updated, paid professional version. The free version of HGMD contains associations of approximately 82,000 variants of all kinds to approximately 3000 diseases (September 2011). NCBI’s ClinVar database, currently in development, aims to provide a freely available, comprehensive listing of variants associated with phenotypes along with links to regularly updated evidence for the associations. In addition to OMIM, SwissProt and HGMD, which list variants in all disease-associated genes, locus specific...
Table 1: Databases and resources for personal genome interpretation

<table>
<thead>
<tr>
<th>Database</th>
<th>URL</th>
<th>Description</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short variations—SNVs, short indels</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structural variations—structural rearrangements, CNVs, large indels</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGV</td>
<td><a href="http://projects.tcag.ca/variation">http://projects.tcag.ca/variation</a></td>
<td>Structural variants from healthy human controls</td>
<td>[14]</td>
</tr>
<tr>
<td>DGVa</td>
<td><a href="http://www.ebi.ac.uk/dgva">http://www.ebi.ac.uk/dgva</a></td>
<td>Structural variants from all species</td>
<td>[13]</td>
</tr>
<tr>
<td>General variants associated with phenotypes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HGMD</td>
<td><a href="http://www.hgmd.org">http://www.hgmd.org</a></td>
<td>Human variant–disease associations (inherited diseases)</td>
<td>[15]</td>
</tr>
<tr>
<td>OMIM</td>
<td><a href="http://www.omim.org">http://www.omim.org</a></td>
<td>Human variant–disease associations (includes extensive gene and phenotype descriptions)</td>
<td>[16]</td>
</tr>
<tr>
<td>SwissVar</td>
<td><a href="http://swissvar.expasy.org">http://swissvar.expasy.org</a></td>
<td>Human variant–disease associations (non-synonymous SNVs only)</td>
<td>[3]</td>
</tr>
<tr>
<td>GWAS and other association studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dbGaP</td>
<td><a href="http://www.ncbi.nlm.nih.gov/gap">http://www.ncbi.nlm.nih.gov/gap</a></td>
<td>Controlled access to individual genotype/phenotype data from association studies</td>
<td>[17]</td>
</tr>
<tr>
<td>EGA</td>
<td><a href="http://www.ebi.ac.uk/ega">http://www.ebi.ac.uk/ega</a></td>
<td>Controlled access to individual genotype/phenotype data from association studies</td>
<td></td>
</tr>
<tr>
<td>GAD</td>
<td><a href="http://geneticassociationdb.nih.gov">http://geneticassociationdb.nih.gov</a></td>
<td>Mainly complex disease SNVs from association studies</td>
<td>[18]</td>
</tr>
<tr>
<td>NHGRI GWAS Catalog</td>
<td><a href="http://www.genome.gov/gwasstudies">http://www.genome.gov/gwasstudies</a></td>
<td>Significant SNVs from GWAS</td>
<td>[7]</td>
</tr>
<tr>
<td>Cancer genes and variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICGC</td>
<td><a href="http://www.icgc.org">http://www.icgc.org</a></td>
<td>Somatic variants from tumor sequencing projects</td>
<td>[9]</td>
</tr>
<tr>
<td>COSMIC</td>
<td><a href="http://sanger.ac.uk/genetics/CGP/cosmic">http://sanger.ac.uk/genetics/CGP/cosmic</a></td>
<td>Somatic variants from tumor sequencing and literature</td>
<td>[20]</td>
</tr>
<tr>
<td>Cancer Gene Census</td>
<td><a href="http://sanger.ac.uk/genetics/CGP/Census">http://sanger.ac.uk/genetics/CGP/Census</a></td>
<td>Comprehensive list of cancer-related genes</td>
<td>[21]</td>
</tr>
<tr>
<td>TCGA</td>
<td><a href="http://cancergenome.nih.gov">http://cancergenome.nih.gov</a></td>
<td>Somatic variants from tumor sequencing projects</td>
<td>[22]</td>
</tr>
<tr>
<td>Pharmacogenomic genes and variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DrugBank</td>
<td><a href="http://drugbank.ca">http://drugbank.ca</a></td>
<td>Drug properties and protein amino acid target sequences</td>
<td>[23]</td>
</tr>
<tr>
<td>Crowdsourced genes and variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNPedia</td>
<td><a href="http://www.snpedia.com">http://www.snpedia.com</a></td>
<td>Human SNV–disease associations</td>
<td></td>
</tr>
<tr>
<td>WikiGenes</td>
<td><a href="https://www.wikigenes.org">https://www.wikigenes.org</a></td>
<td>Gene annotations from all species</td>
<td>[26]</td>
</tr>
<tr>
<td>Viewers of the structural and functional impact of variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMDM</td>
<td><a href="http://bioinf.umbc.edu/dmdm">http://bioinf.umbc.edu/dmdm</a></td>
<td>Aggregates human protein mutations at individual domain positions</td>
<td>[27]</td>
</tr>
<tr>
<td>LS-SNP/PDB</td>
<td><a href="http://ls-snp.icm.jhu.edu/ls-snp-pdb">http://ls-snp.icm.jhu.edu/ls-snp-pdb</a></td>
<td>Variant/PDB structure viewer (includes multiple filters for selection of variants)</td>
<td>[28]</td>
</tr>
<tr>
<td>MutDB</td>
<td><a href="http://mutdb.org">http://mutdb.org</a></td>
<td>Variant/PDB structure viewer (includes SIFT predictions for nonsynonymous mutations)</td>
<td>[29]</td>
</tr>
<tr>
<td>SAAFdb</td>
<td><a href="http://bioinf.org.uk/saafpdb">http://bioinf.org.uk/saafpdb</a></td>
<td>Variant/PDB structure viewer (includes impact on physico-chemical and functional features)</td>
<td>[30]</td>
</tr>
<tr>
<td>StSNP</td>
<td><a href="http://llyninlab.org/StSNP">http://llyninlab.org/StSNP</a></td>
<td>Variant/PDB structure viewer (includes physico-chemical impact for nonsynonymous mutations)</td>
<td>[31]</td>
</tr>
<tr>
<td>SNPeff</td>
<td><a href="http://snpeffct.vib.be">http://snpeffct.vib.be</a></td>
<td>Variant/PDB structure viewer (includes predictions for variants to cause protein aggregation)</td>
<td>[32]</td>
</tr>
<tr>
<td>TopoSNP</td>
<td><a href="http://gila-fw.bioengr.uic.edu/snp/toposnp">http://gila-fw.bioengr.uic.edu/snp/toposnp</a></td>
<td>Variant/PDB structure viewer (includes location of variant on surface, in pocket or in core)</td>
<td>[33]</td>
</tr>
</tbody>
</table>
databases (LSDBs) report variants in a single gene, often related to a single disease. LSDBs are used as a source of variant data for both SwissProt and HGMD. A comprehensive list of links to LSDBs is provided on the Human Genome Variation Society website (http://www.hgvs.org/dblist/glsdb.html). The number of LSDBs and the quality of data contained therein has grown over recent years. However, at least one comprehensive study shows [34] that LSDBs would be better equipped to serve the research (and treatment) community with standardization of reported results and improvement of variant effect (and disease) annotation.

GWAS and other association study databases

As previously noted, large-scale GWAS have identified thousands of variants associated with disease. The National Human Genome Research Institute Catalog of Published Genome-Wide Association Studies (NHGRI GWAS catalog) [35] conveniently lists significantly associated marker SNPs from these studies in a manually curated, online database. The catalog includes data on study designs, individual SNV P-values, odds ratios and links to the published studies. The Genetic Association Database (GAD) [18] predates the NHGRI GWAS catalog, and contains curated information on both positive and negative variant associations from GWAS and candidate gene association studies primarily from studies of common, complex diseases. In addition to the summary data in the GWAS catalog and GAD, the database of Genotype and Phenotype (dbGaP) [17] and the European Genome-phenome Archive (EGA) provide controlled access to individual-level genotype and phenotype data from many large-scale association studies.

Cancer gene and variant databases

Given the significance of somatic mutations in oncogenesis, several large-scale projects sequencing multiple cancer types have emerged including the Cancer Genome Atlas (TCGA) [22] and the Cancer Genome Project (CGP) (http://www.sanger.ac.uk/genetics/CGP). The International Cancer Genome Consortium (ICGC) [19] was developed to coordinate cancer sequencing projects around the world, including TCGA and the CGP, for over 50 different cancer types and subtypes. Data portals for ICGC and TCGA are available to retrieve open access variant data, and individual level, controlled access genotype data by application. Data from the CGP and curated mutations from the literature for a list of genes previously associated with cancer (the Cancer Gene Census list [21]) are available from COSMIC, the Catalog of Somatic Mutations in Cancer [20]. COSMIC (Release 54) currently contains data on over 177,000 mutations from almost 620,000 tumors. The NCI Cancer Gene Index is another comprehensive source of genes related to cancer, containing gene–disease and gene–drug relationships text-mined and manually validated from over 20 million MEDLINE abstracts.

Pharmacogenomic genes and databases

Specialized databases also now exist to link genes and genotypes with drug targets and drug response. The Pharmacogenomics Knowledgebase (PharmGKB) [36] contains both manually curated and automatically text-mined associations of human variations to drug response. The database includes information for variants in over 1500 genes related to approximately 375 drugs and almost 300 diseases (September 2011). DrugBank [23] is a more drug-focused resource containing structural, chemical and pharmacologic properties for over 6800 drugs (September 2011). DrugBank also contains the amino acid target sequences for individual drugs, enabling the identification of variants falling in drug binding sites.

Other resources and tools for personal genome interpretation

Crowdsourcing

Many annotation databases use automated searches followed by expert human curation to identify and validate variant–disease associations from literature. As the pace of association studies continues to increase, this process will become increasingly unsustainable. To address this problem, several databases have been developed to harness a crowdsourcing model for gene and variant annotation including Gene Wiki [25], WikiGenes [26] and SNPedia (http://www.SNPedia.com). While all of these resources include some information automatically extracted from public sources like PubMed, OMIM and dbSNP, the community contribution and curation could potentially provide more comprehensive and update-to-date information as new studies are published.

Viewers

A variant’s structural/functional context is critical to determining its likelihood of disease involvement. Several web tools are available for viewing SNVs superimposed onto the corresponding Protein Data Bank (PDB) [37] protein structures. These tools
highlight the changes in physico-chemical properties and display proximity to structural features like domain interfaces, binding sites, etc. LS-SNP/PDB [38], MutDB [29], SAAPdb [30], StSNP [31], SNPeff [32] and TopoSNP [33] are just a few resources that provide this type of structural and functional annotation. By aggregating all SNPs and disease mutations from dbSNP, OMIM and SwissVar at the protein and domain levels, the Domain Mapping of Disease Mutations (DMDM) database [27] also adds the critical functional context provided by protein domains for variant characterization.

**Epigenetics and transcriptome Databases**

Changes in gene regulation due to epigenetic mechanisms, other than variation in the DNA sequence, can also be disease associated. Local patterns of DNA methylation, chromatin structure and histone modification states, and nonprotein-coding RNAs (ncRNA), e.g. microRNAs, affect gene expression levels. Thus, genome-wide studies to comprehensively catalog the various structural and functional elements of the genome, as well as studies to map the epigenetic elements affecting gene expression levels, are now being undertaken. These will lead to a better understanding of genome complexity and gene regulation. The ENCyclopedia of DNA Elements Project (ENCODE) [39] includes studies to catalog the full human transcriptome including protein coding, noncoding and pseudogene transcripts, in addition to local chromatin states and methylation patterns. The National Institutes of Health Roadmap Epigenomics Mapping Consortium [40] was recently organized to map DNA methylation, histone modifications, chromatin accessibility and ncRNA transcripts in each human tissue and cell type. The ENCODE Project and NIH Roadmap Epigenome data are both available as annotation tracks viewable from the UCSC Genome Browser at http://genome.cse.ucsc.edu/encode and http://www.genomebrowser.org, respectively. Interesting reviews about large-scale epigenomics projects, association studies of epigenetic variation and computational epigenetics analysis have been recently published [41–43].

**GENE PRIORITIZATION**

Aberrations in normal gene function that result in the development of a disease define a ‘disease gene’. Proving a gene–disease relationship experimentally is expensive and time-consuming. Ranking candidate genes prior to experimental testing reduces the associated costs. Computational gene prioritization uses heterogeneous pieces of evidence to associate each gene with a given disease. Whereas experimental studies provide a lot of information, incorporation of other sources of evidence is necessary to narrow down the candidate search space. Gene prioritization techniques effectively translate heterogeneous experimental data into legible disease–gene associations.

**Making sense of available data**

A functional module, or molecular pathway, is generally defined as a series of interactions between molecules in the cell leading to a specific end point in cellular functionality. For the body to remain disease-free, all normally occurring processes, molecular interactions and pathways should function without major alterations. Moreover, since it is an oversimplification to view a single pathway as a discrete and independent entity, it is increasingly evident that different diseases, resulting from aberrations in different pathways, are also interdependent.

Identifying the pathways affected in the observed disease is a major challenge. A given pathway can be altered by gene expression changes, gene-product malformation, introduction of new pathway members, and/or environmental disruptions. Identification of gene–disease associations is complicated by gene pleiotropy, multigenic nature of most diseases, varied influence of environmental factors and overlying genome variation; i.e. any one source of information about a disease may or may not be sufficient to identify its specifics. Moreover, the available experimental data describing each of the biological concepts involved is itself very heterogeneous. Thus, using a combination of resources requires knowing how to meaningfully combine the extracted information (Table 2).

The lines of evidence most robust in identifying genes as prime suspects for disease involvement are: GWAS or linkage analysis studies, similarity or linkage to and co-regulation/-expression/-localization with known disease genes, and participation in disease-associated pathways/compartments. Five notions commonly define these associations: (i) functional association: participation in a common pathway with other disease genes, (ii) cross-species association: orthologues generating similar phenotypes in
other organisms, (iii) disease association: gene–product presence in disease pathways or associated cellular compartments and/or localization to affected tissues, (iv) mutant implication: candidate genes harboring functionally deleterious mutations in diseased individuals and (v) text implication: co-occurrence of gene and disease terms within scientific texts.

**Functional association**

**Common pathway**

Many gene prioritization tools use gene–gene (protein–protein) interaction and/or pathway information to prioritize candidates. Since diseases are results of pathway malfunction, disabling any of the pathway members may lead to similar phenotypes. Cancer-associated proteins, for example, are very strongly interconnected [69]. In general, genes responsible for generating similar diseases tend to participate in the same protein–protein interaction networks [70].

**Regulatory information and genetic linkage**

Co-regulated genes are often thought to be involved in the same molecular pathways [71]. However, co-regulated genes may reside in distinct pathways [72]. Moreover, co-expressed nonparalogous genes often demonstrate conservation of clustering across species [73], suggesting that co-expression clusters are evolutionarily advantageous and naturally selected [74]. Some researchers [75] argue that these clusters may represent groups of genes involved in high-level cooperation beyond the canonical description of cellular pathways. As such, deregulation of these clusters may manifest in disease. Note that whereas genetic linkage/co-regulation are valuable markers of disease association, they are not specific; i.e. a given disease-associated gene may be co-regulated with

---

**Table 2: Available data sources and gene prioritization tools**

<table>
<thead>
<tr>
<th>Method</th>
<th>URL</th>
<th>Description</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAESAR</td>
<td><a href="http://polaris.med.unc.edu/projects/caesar/">http://polaris.med.unc.edu/projects/caesar/</a></td>
<td>ESPNOML</td>
<td>[44]</td>
</tr>
<tr>
<td>CANDID</td>
<td><a href="https://dsweb.wustl.edu/hutz/candid.html">https://dsweb.wustl.edu/hutz/candid.html</a></td>
<td>ESPNL</td>
<td>[45]</td>
</tr>
<tr>
<td>DADA</td>
<td><a href="http://compbio.case.edu/dada/">http://compbio.case.edu/dada/</a></td>
<td>PL</td>
<td>[46]</td>
</tr>
<tr>
<td>DomainRBF</td>
<td><a href="http://bioinfo.au.tsinghua.edu.cn/domainRBF/gene/">http://bioinfo.au.tsinghua.edu.cn/domainRBF/gene/</a></td>
<td>SOML</td>
<td>[47]</td>
</tr>
<tr>
<td>ENDEAVOR</td>
<td><a href="http://www.esat.kuleuven.be/endeavour">http://www.esat.kuleuven.be/endeavour</a></td>
<td>ESPNOL</td>
<td>[48]</td>
</tr>
<tr>
<td>G2D</td>
<td><a href="http://www.wogic.ca/projects/g2d2/">http://www.wogic.ca/projects/g2d2/</a></td>
<td>ESPOL</td>
<td>[49, 50]</td>
</tr>
<tr>
<td>GeneDistiller</td>
<td><a href="http://www.genedistiller.org/">http://www.genedistiller.org/</a></td>
<td>ESPNOL</td>
<td>[51]</td>
</tr>
<tr>
<td>GeneWanderer</td>
<td><a href="http://compbio.charite.de/genewanderer/GeneWanderer">http://compbio.charite.de/genewanderer/GeneWanderer</a></td>
<td>SNML</td>
<td>[52]</td>
</tr>
<tr>
<td>Genie</td>
<td><a href="http://cbsd.mdcc-berlin.de/tools/genie/">http://cbsd.mdcc-berlin.de/tools/genie/</a></td>
<td>PNML</td>
<td>[53]</td>
</tr>
<tr>
<td>Gentrivid</td>
<td><a href="https://www.gentrivid.org/">https://www.gentrivid.org/</a></td>
<td>ESPNL</td>
<td>[54]</td>
</tr>
<tr>
<td>MedSim</td>
<td><a href="http://www.fusimmat.de/">http://www.fusimmat.de/</a></td>
<td>ESPL</td>
<td>[55]</td>
</tr>
<tr>
<td>MimMiner</td>
<td><a href="http://www.cmbi.ru.nl/MimMiner/">http://www.cmbi.ru.nl/MimMiner/</a></td>
<td>SPNOL</td>
<td>[56]</td>
</tr>
<tr>
<td>PGMapper</td>
<td><a href="http://www.genediscovery.org/pgmapper/">http://www.genediscovery.org/pgmapper/</a></td>
<td>SL</td>
<td>[57]</td>
</tr>
<tr>
<td>PhenoPred</td>
<td><a href="http://www.phenopred.org/">http://www.phenopred.org/</a></td>
<td>ESPL</td>
<td>[58]</td>
</tr>
<tr>
<td>PINTA</td>
<td><a href="http://www.esat.kuleuven.be/pinta/">http://www.esat.kuleuven.be/pinta/</a></td>
<td>SPO</td>
<td>[59]</td>
</tr>
<tr>
<td>PROSPECTR</td>
<td><a href="http://www.genetics.med.ed.ac.uk/prospectr/">http://www.genetics.med.ed.ac.uk/prospectr/</a></td>
<td>L</td>
<td>[61]</td>
</tr>
<tr>
<td>SNP3D</td>
<td><a href="http://www.snps3d.org/">http://www.snps3d.org/</a></td>
<td>SNML</td>
<td>[62]</td>
</tr>
<tr>
<td>SUSPECTS</td>
<td><a href="http://www.genetics.med.ed.ac.uk/suspects/">http://www.genetics.med.ed.ac.uk/suspects/</a></td>
<td>SPNOL</td>
<td>[63]</td>
</tr>
<tr>
<td>ToppGene</td>
<td><a href="http://toppgene.cchmc.org/">http://toppgene.cchmc.org/</a></td>
<td>ESPNOL</td>
<td>[64]</td>
</tr>
<tr>
<td>VAAST</td>
<td><a href="http://www.yandell-lab.org/software/vaast.html">http://www.yandell-lab.org/software/vaast.html</a></td>
<td>EM</td>
<td>[65, 66]</td>
</tr>
</tbody>
</table>

A wide range of data sources can be used to formulate the associations and implications described herein. Existing tools try to take advantage of many of them. This table summarizes current web-available, state of the art gene prioritization methods. Note, some methods (e.g. PRINCE [68]) are not available online, but are downloadable for local installation. Letters in the Description column denote the data sources used: (E) Experimental observation: Linkage, association, pedigree, relevant texts and other data. (S) 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. (P) Pathway, protein–protein interaction, genetic linkage, expression: Disease–gene associations, pathways and gene–gene/protein–protein interactions/interaction predictions and gene expression data. (N) Non-human data: Information about related genes and phenotypes in other species. (O) Ontologies: Gene, disease, phenotype and anatomic ontologies. (M) 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. (L) Literature: mixed information of all types extracted from literature references (e.g. disease–gene correlation and nonontology based gene-function assignment).
or linked to another disease-associated gene, where the two diseases are not necessarily identical.

**Similar function**
It is common to observe reduced/absent phenotypic effect in response to gene knockout/inactivation [76, 77]. This phenomenon is largely explained by functional compensation via partial interchangeability of paralogous genes. When functional compensation is insufficient, inactivation of any of the paralogs leads to same/similar disease. Thus, many tools use functional similarity to establish disease–gene association. Defining functional similarity is nontrivial. When utilizing ontologies, like GO [78], the question becomes a problem of assigning a ‘score’ to the similarity of two ontology nodes/subtrees. ‘Functional distance’ [79–82] calculations for any two genes within the context of the ontology suggest a well-defined way of annotating functional similarity.

**Sequence/structure association**
Sequence/structure homology is often used for transferring functional annotations from characterized genes/proteins to new ones [83]. Since functionally similar genes are likely to produce similar disease phenotypes, homology is also important in disease gene prioritization. Additionally, disease genes are distinctly associated with specific canonical features such as higher exon number, gene, protein and 3’-UTR length and distance to a neighboring gene, as well as lower sequence divergence from their orthologs [64, 84]. Proteins with signal peptides are also more commonly disease associated [64], whereas disordered proteins are often implicated in cancer [85].

**Cross-species association**
A high number of orthologs suggests essential genes prone to disease involvement. Orthologs also generally participate in similar molecular pathways across species, although different levels of function may be necessary for different organisms. Cross-species, tissue-specific phenotypic differences due to slightly varied sequences are thus useful for gene prioritization. Note that phenotype ontologies are necessary to facilitate [86] this comparison of organism phenotypes.

A correlation of co-expression of genes in different species is also a predictive measure for annotating disease genes [87, 88]. As discussed above, there is evidence for co-expression of genes that are not related in any known functional manner [89–91]. These co-expressed clusters may be evolutionarily advantageous [74, 75, 90], but are only evident as such if conserved throughout other species. Cross-species comparisons of protein co-expression are therefore useful for validating disease–gene co-expression associations (e.g. [87]).

**Disease compartment association**
Altered gene expression is expected in association with many common complex diseases [92]. Genes that are preferentially expressed in disease-affected tissues are likely candidates for disease association. Some proteins interact only in some tissues [93], so tissue specificity is important for finding the right protein–protein interaction networks. Similarly, disease-association with cellular pathways (e.g. ion channels) and compartments (e.g. plasma membrane) may indicate that pathway/compartment-specific gene-products are also disease associated.

**Mutant implication**
Every genetic disease is associated with some sort of mutation that alters normal functionality. Selection of candidates for further analysis is often based on observations of variants in diseased individuals, which are absent in healthy controls. Not all observed variants are deleterious. Most observed variations do not manifest phenotypically, and some are weakly deleterious or even beneficial. Many gene prioritization methods use mutation effect predictions to make their own inferences. Tools used to make these predictions are described in the ‘Genetic variant interpretation’ section of this article.

**Text implication**
Experimental findings of gene–disease associations are often reported in the literature. Some of the data is also machine accessible via various databases described above such as dbSNP [11], GAD [18] and COSMIC [20], or by depositing manually curated knowledge into databases like GeneRIF [94] and UniProt. However, a vast sea of data remains ‘hidden’ in the natural language text of scientific publications. Text mining tools have recently come of age [94–97], allowing for the identification of possible gene–gene and disease–gene correlations [98–100]. For example, the Information Hyperlinked Over Proteins, IHOP method [101] extracts gene/protein...
names in scientific texts and links these via pathology, phenotype, physiology and interaction information. Gene prioritization techniques also often rely on term co-occurrence statistics (e.g. PosMed [102], GeneDistiller [51]) and gene-function annotations (e.g. ENDEAVOR [103], PolySearch [61]).

The inputs, outputs and processing
Gene prioritization methods vary based on the inputs they require and the types of outputs they produce. For an excellent summary of methods and their inputs, see [48] and Table 3. Method inputs rely on two distinct notions: previously known information about the disease and the candidate search space. The disease information may include a list of genes known to be associated with the disease, the tissues and pathways it afflicts and any relevant keywords. The candidate search space may include the entire genome, or may be defined by the suspect genomic region, overexpression in the affected tissue, or other experimental results. Not surprisingly, the accuracy of the prioritization method often depends on the accuracy and specificity of the input data. Outputs of prioritization methods are generally limited to ranked lists of genes, often associated with test-performance values (e.g. P-values). Some methods only rank/order the top genes they select, while others manage the entire submitted list. Selected input and output requirements are important for a tool’s acceptance by the biological community. A given method’s ease of use often defines its utility as strictly as do its speed and accuracy.

Finally, gene prioritization methods also differ in the algorithms they use to make sense of the data. Tools used include mathematical/statistical models and scoring methods (e.g. SUSPECTS [64], GeneProspector [52]), fuzzy logic (e.g. ToppGene [65, 66]), artificial learning devices (e.g. decision trees in PROSPECTR [62], neural networks in PosMed [102]), network/topology analysis.

Table 3: Tools for the interpretation of single nucleotide variants

<table>
<thead>
<tr>
<th>Method</th>
<th>URL</th>
<th>Description</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Methods for the prediction of stability change upon mutation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AutoMUTE</td>
<td><a href="http://proteins.gmu.edu/automute/">http://proteins.gmu.edu/automute/</a></td>
<td>Delaunay tessellation and combined machine learning methods</td>
<td>[104]</td>
</tr>
<tr>
<td>CUPSAT</td>
<td><a href="http://cupsat.tu-bs.de/">http://cupsat.tu-bs.de/</a></td>
<td>Physics-based energy function</td>
<td>[105]</td>
</tr>
<tr>
<td>D-Mutant</td>
<td><a href="http://sparks.informatics.iupui.edu/hzhou/mutation.html">http://sparks.informatics.iupui.edu/hzhou/mutation.html</a></td>
<td>Statistical-based energy function</td>
<td>[106]</td>
</tr>
<tr>
<td>Fold-X</td>
<td><a href="http://foldx.org/">http://foldx.org/</a></td>
<td>Physics-based energy function</td>
<td>[107]</td>
</tr>
<tr>
<td>PoPMuSiC</td>
<td><a href="http://babylone.ulb.ac.be/popmusic">http://babylone.ulb.ac.be/popmusic</a></td>
<td>Statistical-based energy function optimized by ANN</td>
<td>[109]</td>
</tr>
<tr>
<td><strong>Methods for the prediction of deleterious non-synonymous SNVs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PANTHER</td>
<td><a href="http://www.pantherdb.org/">http://www.pantherdb.org/</a></td>
<td>Protein family HMM-based method</td>
<td>[110]</td>
</tr>
<tr>
<td>PhD-SNP</td>
<td><a href="http://gpcr.biocomp.unibo.it/PhD-SNP.htm">http://gpcr.biocomp.unibo.it/PhD-SNP.htm</a></td>
<td>Sequence and profile-based SVM method</td>
<td>[111]</td>
</tr>
<tr>
<td>PolyPhen</td>
<td><a href="http://genetics.bwh.harvard.edu/pph">http://genetics.bwh.harvard.edu/pph</a></td>
<td>Decision Tree-based method</td>
<td>[112]</td>
</tr>
<tr>
<td>MuxPred</td>
<td><a href="http://mutdb.org/mutpred">http://mutdb.org/mutpred</a></td>
<td>Random forest approach including multiple scores</td>
<td>[113]</td>
</tr>
<tr>
<td>SIFT</td>
<td><a href="http://sift-dna.org">http://sift-dna.org</a></td>
<td>Protein block alignment-based method</td>
<td>[114]</td>
</tr>
<tr>
<td>SNAP</td>
<td><a href="http://rostalk.org/services/snap">http://rostalk.org/services/snap</a></td>
<td>Sequence profile-based neural network</td>
<td>[115]</td>
</tr>
<tr>
<td>SNPs3D</td>
<td><a href="http://www.snp3d.org">http://www.snp3d.org</a></td>
<td>Structure-based SVM predictor</td>
<td>[63]</td>
</tr>
<tr>
<td>SNPs&amp;GO</td>
<td><a href="http://snps-and-go.biocomp.unibo.it">http://snps-and-go.biocomp.unibo.it</a></td>
<td>Sequence profile and functional-based SVM</td>
<td>[116, 117]</td>
</tr>
<tr>
<td><strong>Predictors of the impact of SNVs at DNA level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANNOVAR</td>
<td><a href="http://www.openbioinformatics.org/annovar">http://www.openbioinformatics.org/annovar</a></td>
<td>Scoring functional and evolutionary information</td>
<td>[118]</td>
</tr>
<tr>
<td>binCons</td>
<td><a href="http://zoo.nhgri.nih.gov/binCons/index.cgi">http://zoo.nhgri.nih.gov/binCons/index.cgi</a></td>
<td>Evolutionary analysis with window-based approach</td>
<td>[119]</td>
</tr>
<tr>
<td>GERP</td>
<td><a href="http://mendel.stanford.edu/SidowLab/downloads/gerp/">http://mendel.stanford.edu/SidowLab/downloads/gerp/</a></td>
<td>Site-specific mutative analysis</td>
<td>[120]</td>
</tr>
<tr>
<td>Gunby</td>
<td><a href="http://pga.jgi-psf.org/gunby/">http://pga.jgi-psf.org/gunby/</a></td>
<td>Statistical-based mutative analysis</td>
<td>[121]</td>
</tr>
<tr>
<td>Is-rSNP</td>
<td><a href="http://www.genomics.csse.unimelb.edu.au/is-rSNP/">http://www.genomics.csse.unimelb.edu.au/is-rSNP/</a></td>
<td>Effect of variants in regulatory regions</td>
<td>[122]</td>
</tr>
<tr>
<td>MutationTaster</td>
<td><a href="http://www.mutationtaster.org/">http://www.mutationtaster.org/</a></td>
<td>Evolutionary conservation, splicing site changes and loss of protein features</td>
<td>[123]</td>
</tr>
<tr>
<td>PhastCONS</td>
<td><a href="http://compgen.bscb.cornell.edu/phast">http://compgen.bscb.cornell.edu/phast</a></td>
<td>Phylogenetic HMM-based method</td>
<td>[124]</td>
</tr>
<tr>
<td>SCONES</td>
<td><a href="http://genetics.bwh.harvard.edu/scone">http://genetics.bwh.harvard.edu/scone</a></td>
<td>Site-specific mutative analysis</td>
<td>[125]</td>
</tr>
<tr>
<td>Skippy</td>
<td><a href="http://research.nhgri.nih.gov/skippy">http://research.nhgri.nih.gov/skippy</a></td>
<td>Predicts variants affecting splicing sites</td>
<td>[126]</td>
</tr>
<tr>
<td>VISTA</td>
<td><a href="http://genome.lbl.gov/vista/index.shtml">http://genome.lbl.gov/vista/index.shtml</a></td>
<td>Integrated approach including scores from different methods</td>
<td>[127]</td>
</tr>
</tbody>
</table>
approaches (e.g. DiseaseNet [127]), and vector/profile comparisons (e.g. CAESAR [44], MedSim [56]) among others. Most often, some combination of the above methods is used, but there is no single methodology that is objectively better than the others for the compilation of data from all sources.

GENETIC VARIANT INTERPRETATION

The interpretation of the functional impact of genetic variation is one of the most important tasks in personal genomics and personalized medicine [129]. Genomic variants have different effects depending on whether they occur in coding or noncoding regions. In coding regions, variants can change the amino acid sequence of the coded protein. In noncoding regions, they can affect transcription, splicing and binding. The recent 1000 Genomes Project Consortium work confirms that single nucleotide variants (SNVs) are the most common type of genetic variation [2]. Thus, understanding the functional effect of SNVs is one of the main goals of modern genetics/genomics studies [130]. Over the past 10 years, several methods have been developed to predict the impact of SNVs [131–133]. Here, we describe the information used by these algorithms, and present a selection of the most popular web-available tools for genome interpretation.

Computational approaches for genome interpretation

As noted, experimental studies to characterize the impact of SNVs are still expensive and time consuming. To partially overcome this limitation, several algorithms have been implemented to predict the effect of genetic variants (Table 3). All such methods take input information derived from sequence conservation, because it has been observed that functionally important regions of the genome tend to be more conserved through evolution than nonfunctional ones [134, 135]. The detection of functional and conserved sites depends on their locations in the genome. We currently have a better understanding of the relationship between the DNA sequence and function for coding regions than for noncoding ones [136]. Hence, the majority of methods have been designed to predict the effect of nonsynonymous SNVs (nsSNVs) and, only recently, a few supervised methods have been developed to evaluate the impact of genetic variants in noncoding regions.

Predicting the effect of nsSNVs

Methods for predicting the effect of nsSNVs estimate their probability of being disease-associated or functionally deleterious. The catalog of the relationships between molecular phenotypes and disease is far from complete. However, it is believed that the pathologic state results from amino acid substitutions affecting functionally critical residues and/or causing alterations in the structure of the folded protein, structural instability or protein aggregation [137]. Several methods have been developed for predicting the effects of amino acid substitutions. In particular, we describe two (not fully separable) classes: (i) predictors of nsSNV functional effects; i.e. modifying the catalytic site of an enzyme or affecting a residue involved in crucial interactions with partner molecules and (ii) those predicting the effect of nsSNVs on protein stability (Table 3).

Methods for the prediction of protein stability changes

Incorrect protein folding mechanisms and decreased stability are the major consequences of pathogenic nsSNVs [138], as they can cause a reduction in hydrophobic area, overpacking, backbone strain and/or loss of electrostatic interactions [139]. Although different thermodynamics measures can be used to assess the variation of stability upon mutation, one of the most common is the difference of the folding free energy change between the wild-type and mutated proteins (ΔΔG). Several methods have been developed to predict if a given nsSNV changes the stability of the protein structure. Some algorithms implement an energy function to compute the ΔΔG [106, 140–145], whereas others are based on machine-learning approaches [108, 146–148].

The methods relying on energy functions can be subdivided into (i) physics-based approaches that use a force field to describe the atomic interactions involved in the folding process [142–144] and (ii) statistical potential approaches that use an empirical energy function derived from the statistical analysis of the structural environment around the mutated site [106, 140, 141, 145]. More recently, an increasing amount of thermodynamic data, collected in web-available databases such as ProTherm [149], has allowed training machine learning methods to predict both the value and sign of the difference
between the folding free energy of the wild-type and mutated proteins (ΔΔG). In 2010, the accuracy of the most popular web-available algorithms was assessed by reporting method performances on a set of thermodynamic data not included in the training set [150]. Although tested on data sets of different sizes, Dmutant [106], FoldX [107] and I-Mutant3.0 [147] scored the highest for predicting protein stability changes. This assessment showed that current methods for the prediction of stability changes due to nsSNVs reach a moderate level of accuracy (~60%). Further improvements are therefore necessary to provide more reliable predictions.

Methods for the prediction of functional effects of nsSNVs

Efforts to design accurate algorithms for the prediction of functionally deleterious nsSNVs have resulted in a slew of available methods [130]. Considering evolution as the ultimate mutagenesis experiment, comparative sequence analysis is a powerful source of information taken into account by all the algorithms. A simple study performed on a dataset of nsSNVs extracted from SwissVar database [3] showed significant differences between the distribution of the frequencies of wild-type and mutant residues for the subsets of disease-related and neutral variants (Figure 2A and B). The median values for the frequencies of the wild-type residues in disease-related and neutral variants (0.66 and 0.34, respectively) confirm the idea that wild-type residues are more conserved in disease-associated nsSNVs. Analyzing the distributions of the frequencies of the mutant residues, it was shown that in ~60% of the deleterious mutations, the mutant residue does not appear in any sequence of the multiple sequence alignment, whereas in ~71% of the neutral mutations, the mutant residue appears at least once. In addition, the distribution of the difference between the frequencies of the wild-type and mutant residues, in Figure 2C, confirms the previous observations. Similar results were obtained when calculating the distributions of the conservation index as defined in Ref. [151] (Figure 2D).

The discriminative power of evolutionary information is used in all prediction methods, although in different ways. For example, SIFT [114] uses blocks of conserved regions in proteins, PhD-SNP [111] and SNPs&GO [116] calculate the sequence profile by running the BLAST algorithm [152] over a database of sequences, PolyPhen [112] and SNAP [154] also include position-specific independent count (PSIC) scores, PANTHER [134] compares the sequence against a library of hidden Markov models from protein families and other methods perform the analysis of the DNA sequence by evaluating the selective pressure acting at the codon level [155, 156]. Predictors also use features from

![Figure 2](https://academic.oup.com/bib/article-abstract/13/4/495/181806/1344695-181806)
predicted or experimentally determined protein structures and available functional information [28, 112, 113, 117, 138, 154].

The algorithms for the detection of deleterious nsSNVs also differ in their training sets and underlying classification methodologies. Most are trained on an annotated set of nonsynonymous variants from OMIM, HGMD and/or SwissVar. Others, like SNAP, use mutations from the PMD database [157], collecting functional (as opposed to disease associated) data from mutagenesis experiments. In a recent work, the performance of a pool of methods was assessed on a curated set of nsSNVs [158]. This study showed that more sophisticated methods, such as MutPred [113] and SNAP [154] which include features from predicted structures and SNP&GO [116] which uses functional information encoded in a Gene Ontology-based score, had the best performance.

With the increasing number of annotated nsSNVs, new algorithms can also be trained on a set of mutations associated to a specific class of diseases and/or proteins. Methods for the prediction of cancer-causing nsSNVs [159–161] are an interesting example, as is a new tool for predicting the effect of genetic variants in voltage-gated potassium channels [162].

**Predicting the impact of genomic variants in noncoding regions**

Until recently, the analysis of the effect of genetic variations strongly focused on those altering the protein sequence. The interpretation of genetic variants occurring in noncoding regions is also a challenging task. Although variants in noncoding regions may exhibit weaker effects than nsSNVs, it is evident that they constitute the majority of human genetic variations [4, 136], and are also likely to be disease-associated; i.e. ~88% of weakly trait-associated variants from GWAS studies are noncoding [5]. Noncoding variants under purifying selection are five times more common than those in coding regions [163], and the detection of numerous regulatory variants with significant effect [15] has recently spurred interest in their computational annotation. Thus, a considerable number of methods are currently available to perform an evolutionary analysis of the nucleotide sequence to determine conserved regions across species. This approach, also applicable to protein sequences, is more complex for noncoding regions where there is no detectable conservation outside vertebrates [164]. The available algorithms for the detection of deleterious noncoding SNVs estimate the rate of evolution at the mutated position or consider a sliding window around the mutation site. Methods like binCons [119] and phastCons [124] implement a context dependent approach or a Hidden Markov model, in contrast to other algorithms such as GERP [120], SCONE [125] and Gumby [121] which calculate a position-specific score. This class of methods was also reviewed in a recent publication [165].

New approaches to predict the effect of mutations in noncoding regions focus specifically on genetic variations in regulatory regions and splicing sites. For example, Is-rSNP [122] uses a transcription factor position weight matrix and novel convolution methods to evaluate the statistical significance of the score. The RAVEN algorithm combines phylogenetic information and transcription factor binding site prediction to detect variations in candidate cis-regulatory elements [166]. Recently, a new method including features associated with the mutated site and its surrounding region and gene-based features has been used for the identification of functional, regulatory SNVs involved in monogenic and complex diseases [167]. SNVs affecting splicing sites and their surrounding regions can be evaluated using Skippy [126]. In Table 3, we listed a selection of methods for the prediction of the effect of SNVs.

**Integrated tools for variant annotation**

The steps for interpreting the net effects of variants from an individual genome or from a disease association study have previously been performed one at a time: filtering out common polymorphisms, identifying known deleterious mutations, functionally annotating and predicting the effects of novel variants and prioritizing variants for experimental follow-up. A number of integrated tools are now emerging to automate various portions of this pipeline including ANNOVAR [118], the Ensembl Variant Effect predictor [168], GAMES [169], SeqAnt [170], Sequence Variant Analyzer (SVA) [171] and MutationTaster [123]. Frameworks for storing patient data along with associated analysis tools like i2b2 [172] and caBIG [173], and workflow management systems like Galaxy [174] and Taverna [175] that can be installed and run ‘on the cloud’, are also now available to automate and dramatically speed up variant annotation pipelines.
FUTURE OUTLOOK

Advances in high-throughput sequencing technology are generating a large amount of genetic variation data, thereby creating more complete models relating genotype to phenotype. The release of this information to publicly available databases has stimulated the development of several tools for genome interpretation. Although these methods have reached a promising level of accuracy, there are still many challenges to overcome before they will be directly applicable in a clinical setting. A number of recent studies address this concern [176, 177]. To make personal genome analysis a routine practice in the diagnosis and treatment of genetically determined phenotypes, the following challenges must be met: (i) defining standard and unified protocols for testing functional variation, (ii) designing integrated and publicly available resources of annotated genetic variants, (iii) developing holistic approaches to score the effect of multiple genetic variants, (iv) implementing user-friendly methods for the application of personal genomics in the health care context. The first challenge will require outlining easily reproducible experimental procedures necessary for data consistency. Curated data sets with standardized nomenclatures for the functional effects of genetic variants (easily parseable from the literature) will also be necessary. These resources will be useful for the development and benchmarking of new and more accurate methods for genome interpretation.

One of the most challenging aims for personal genomics will be the development of models able to capture the full complexity of the human genome. These models should take into account the linkage disequilibrium between different genomic regions and the possible effects of compensatory mutations. Bioinformatics will be particularly important for this challenge, enabling the design of heuristics to reduce the computational complexity of the problem. Since one of the primary goals of personal genomics is the development of computational methods for use in clinical diagnostics, an important issue is the usability of these tools. New clinical applications should be easily accessible, return useful and comprehensible results and perform their analyses in a reasonable run time. It will thus be crucial to adopt open access policies that, avoiding privacy/copyright issues, will allow the sharing of large sets of data and developed analysis tools. In particular these algorithms can be used to define a set of markers important for genetic counseling.

In the near future we expect to have accurate disease-specific protocols for estimating the disease development and transmission risks inherent to a personal variome. These will be useful in the diagnosis of inherited disease, in preventative management and/or in family planning.

Recently, the Critical Assessment for Genome Interpretation (CAGI) experiments, assessing the accuracy of computational methods for genome interpretation over a blind set of data, and international meetings, e.g. ISMB SNP-SIG, AIMM and PSB, have drawn attention in the bioinformatics community to the challenges of the analysis of a personal genome. In the near future, these types of initiatives will be essential for organizing the necessarily interdisciplinary scientific environment for cracking the code of the human genome.

Key Points

- Vast amounts of variation data from genome sequencing studies need to be analyzed to understand its association with various phenotypes.
- Well-curated databases, reliable tools for gene prioritization and accurate methods for predicting the impact of variants will be essential for the interpretation of personal genomes.
- Standard and unified protocols for testing the functional impact of genetic variations are critical for their accurate annotation.
- Experimental studies and computational models describing the gene/protein interaction networks and aiming at capturing the full complexity of the human genome will be key to leveraging personal genomic data for the precise diagnosis and effective treatment of disease.

Acknowledgements

We acknowledge all the colleagues who contributed in the organization and panel discussions at the last PSB and ISMB SNP-SIG meetings. In particular, we would like to thank Can Alkan, Christopher Baker, Steven Brenner, Sean Mooney, John Moult, Pauline Ng, Burkhard Rost, Janita Thusberg and Mauno Vihinen. The mention of commercial products herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services. This is a contribution of the Food and Drug Administration and is not subject to copyright.

FUNDING

The European Community through the Marie Curie International Outgoing Fellowship program (PIOF-GA-2009-237225 to E.C.); this work was supported by the National Institutes of Health (NIH) (1K22CA143148 to M.G.K.); Rutgers University, New Brunswick, School of Environmental and Biological Science (SEBS)
start-up funds (to Y.B.); the Research Participation Program administered by Oak Ridge Institute for Science and Education (ORISE) through an inter-agency agreement between Department of Energy (DOE) and Food and Drug Administration (FDA) (to N.N.).

References


