GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Mini Review

Identifying human disease genes: advances in molecular genetics and computational approaches

S.M. Bakhtiar[1,5], A. Ali[2], S.M. Baig[3], D. Barh[4], A. Miyoshi[5] and V. Azevedo[5]

[1]Department of Bioinformatics, Mohammad Ali Jinnah University, Islamabad Expressway, Islamabad, Pakistan
[2]Atta-ur-Rahman School of Applied Biosciences, National University of Science and Technology, Islamabad, Pakistan
[3]Health Biotechnology Division, National Institute for Biotechnology and Genetic Engineering, Faisalabad, Pakistan
[4]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Nonakuri, Purba Medinipur, India
[5]Laboratório de Genética Celular e Molecular, Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil

Corresponding author: V. Azevedo
E-mail: vasco@icb.ufmg.br

Genet. Mol. Res. 13 (3): 5073-5087 (2014)
Received May 20, 2014
Accepted May 28, 2014
Published July 4, 2014
DOI http://dx.doi.org/10.4238/2014.July.4.23

ABSTRACT. The human genome project is one of the significant achievements that have provided detailed insight into our genetic legacy. During the last two decades, biomedical investigations have gathered a considerable body of evidence by detecting more than 2000 disease genes. Despite the imperative advances in the genetic understanding

of various diseases, the pathogenesis of many others remains obscure. With recent advances, the laborious methodologies used to identify DNA variations are replaced by direct sequencing of genomic DNA to detect genetic changes. The ability to perform such studies depends equally on the development of high-throughput and economical genotyping methods. Currently, basically for every disease whose origen is still unknown, genetic approaches are available which could be pedigree-dependent or -independent with the capacity to elucidate fundamental disease mechanisms. Computer algorithms and programs for linkage analysis have formed the foundation for many disease gene detection projects, similarly databases of clinical findings have been widely used to support diagnostic decisions in dysmorphology and general human disease. For every disease type, genome sequence variations, particularly single nucleotide polymorphisms are mapped by comparing the genetic makeup of case and control groups. Methods that predict the effects of polymorphisms on protein stability are useful for the identification of possible disease associations, whereas structural effects can be assessed using methods to predict stability changes in proteins using sequence and/or structural information.

Key words: Human inherited disorders; Genetic diseases; Molecular diagnostics; Medical informatics

INTRODUCTION

The identification of susceptibility genes and underlying genetic variations is one of the major goals in medical genetics. Our genetic heritage was viewed in detail for the first time after completion of the Human Genome Project (HGP) at the cost of ~$3 billion. The main goal of the HGP was to sequence the entire human genome, which in turn opened new arenas for studying human diseases at the molecular level and encouraged genetics to bring the study of rare diseases of childhood to center stage in understanding the pathogenesis of all possible dysfunctions. The human genome is a complex structure of 3.2 billion nucleotides packed inside the nucleus, of which only 1% is translated into proteins and 0.5% codes for regulatory elements controlling the expression of genes. The function of the other 98.5% of the genome, referred to as the dark region, is yet to be discovered. In the last 20 years, there have been advances in DNA sequencing technologies starting from Sanger DNA sequencing to high-throughput next-generation sequencing (NGS) and ultimately whole exon sequencing where millions of DNA strands can be run in parallel. These high-throughput DNA sequencing approaches have not only increased the output but have also reduced the cost of DNA sequencing, making them the methods of choice to study both monogenic and common complex diseases. The emerging need of appropriate medical diagnosis and genetic counseling have given strong incentives to research in medical sciences (McClellan and King, 2010) and have gathered a considerable body of evidence with the identification of more than 2000 disease genes in just 20 years.

Inherited human diseases can be classified as rare/monogenic and common/complex or multi-factorial diseases. According to the United States Rare Diseases Act of 2002, a dis-

ease that affects less than 200,000 individuals in the country is called a rare disease [U.S. Rare Diseases Act of 2002 (H.R. 4013)]. There are about 7000 rare diseases (80% having a genetic origin) affecting millions of people worldwide (Yaneva-Deliverska, 2011). These diseases could be chronic, fatal and devastating; hence, a rare disease can affect individuals ranging from a few to hundreds, making it a challenge to develop proper diagnostic methods and therapeutic interventions for such diseases. Despite the massive progress and development in medical genetics, the heritability of both common and rare diseases, caused by rare variants, still needs to be properly explored. These genetic variations are classified on the basis of their effects on the function of genes. The most common type of variation is where a single base is altered. Owing to the redundancy of the genetic code, this variation within coding regions can be silent if it does not change the amino acid sequence of the protein encoded. But these substitutions can sometimes have a phenotypic effect due to varying levels of transfer RNAs decoding each amino acid and this can result in the modification of protein expression. The DNA sequence variants can be called either common or rare on the basis of their minor allele frequencies (MAFs) in the population. The variants with an MAF >5% (0.05) are called common variants and those with an MAF <1% are called rare variants, while those having an MAF between 1 and 5% are referred to as uncommon variants. Rare variants are considered individually rare, but each person will have thousands of such rare variants across the genome. Therefore, it is difficult to determine if this novel variant is a sequencing artifact or a true variant responsible for disease (Bailey-Wilson and Wilson, 2011).

Genetic markers, e.g., short tandem repeats (STRs), were documented to be fairly frequent in the human genome and were extensively used in the past to study heterozygosity in a population. Similarly, a single nucleotide polymorphism (SNP), a sequence variant that occurs at more than 99% frequency in a population, can make a certain population susceptible to a particular disease or even infection, which maybe a resistant allele in other population. As technology evolved, arduous identification and genotyping methods, such as restriction fragment length polymorphisms, for detecting SNPs associated with any medical condition were replaced by direct high-throughput sequencing of genomic DNA (Burgess, 2011). To pinpoint genetic markers on the human genetic map, it was necessary to have a collection of extended families with a large number of informative meiotic events. Similarly, the answer to the question of how many markers need to be genotyped to increase the power of a study and to detect linkage is reliant on the informativeness of the pedigree and the marker genotyped.

It is worth considering that with changing technology, the ability to identify disease-causing mutations in a single patient/family is gaining ground. To perform such studies, large-scale inexpensive SNP genotyping methods are required. Analytical methods for genome-wide association studies (GWAS) are still evolving, but a number of critical methods have emerged, which greatly strengthen the analysis. On the other hand, it has been recognized that with hundreds of thousands of SNPs genotyped in thousands of cases and controls, one can make very accurate assessments of the degree to which cases and controls are genetically well matched. Recently, there has been an explosion of GWAS for common diseases from a few deep re-sequencing studies done to date, including type II diabetes, coronary artery disease, and hypertension, but still a number of challenges remain (Schunkert et al., 2011). First, the ability to pinpoint functional mutations among the large set of rare and common variants in a gene is critical. Comparative genomic approaches seems promising in identifying loss of function alleles. Second, there is the need for very large cohorts; even for well-validated

candidates, thousands of subjects are required to detect signals from rare alleles. The complex multifactorial diseases are mainly investigated using either common-disease common variant hypotheses (a combined effect of a large number of common alleles resulting in a complex phenotype, where each variant has a modest effect) or common-disease rare variant hypotheses (a complex phenotype is caused by many rare variants, each with a large effect or contribution towards the heritability of a particular trait) (Schork et al., 2013).

Despite the important advances in understanding the molecular basis of numerous diseases, the pathogenesis of many others remains vague. Even with excellent descriptions of genes and mutations involved in disease, the underlying pathophysiology remains incomprehensible. In many cases, particular factors are associated with disease, but distinguishing whether these are casually related to the disease process or alternatively to a secondary consequence of disease is a really difficult task (Rubio et al., 2013). Complications in most frequent disease phenotypes, such as diabetes, have posed a vexing problem for genetic analysis. While single genes with a very large effect can often be ruled out on the basis of the patterns of transmission within families, the observed levels of familial reappearance could be ascribed to either the combined effects of three or four genes, for example, each with moderate effects in individual patients, or it could be the combined effects of 50-100 genes, each with very small effects. Genetic testing increasingly shows the potential to guide clinical practices and disease management in patients with these disorders.

Given the diversity and complex nature of problems in genetics and medicine, it is imperative to approach each problem with a comprehensive knowledge of available computational tools, so that the best tools can be selected for the problem at hand. Bioinformatic programming skills are becoming a necessity across many features of biology and medicine, owing in part to the continuing explosion of biological data aggregation and complexity as well as the scale of questions now being addressed through modern bioinformatics.

Genetic and computational approaches to seek disease genes

The identification of genes involved in diseases (gene mapping) can be traced back to the development of the first molecular method for constructing a linkage map of the human genome with RFLP. Large chromosomal aberrations, the earliest mutations that co-segregate with disease phenotype, used to be detected through cytogenetic analysis of chromosomes and karyotyping. These low-resolution methods of chromosomal analysis led to the identification of genetic causes of granulomatosis and X-linked Duchenne muscular dystrophy. In the 1980s, other approaches such as simple sequence repeats and genome-wide sequence-tagged site markers were also introduced. In 1997, copy-number variation (CNV) was used to detect genomic imbalances (Solinas-Toldo et al., 1997). Genome-wide CNV was found useful in typing genetic determinants in dominant diseases. Another method called homozygosity mapping was also successfully used for disease gene identification in recessive phenotypes, particularly those occurring in communities with consanguinity. In this method, selective individuals in a particular family are analyzed. These representative individuals in families show limited genetic heterogeneity because of their common ancestors, and thus, the method relies on the identification of homozygous mutation. Since 1987, homozygosity mapping has been successful in identifying multiple disease-causing genes. Currently, almost every disease whose origin is still unknown, genetic approaches may elucidate fundamental disease mechanisms

(Singleton et al., 2010).

To seek for a particular gene, either a pedigree/family-based approach or population-based case-controls are used. Monogenic disorders follow a Mendelian pattern of inheritance and show a varying degree of penetrance, whereas multifactorial common complex diseases involve environmental factors in addition to genetics. Genes, acting through proteins they encode, interact with numerous other proteins present within a cellular environment. These interactions are in turn influenced by the cell larger environments including organ, organisms, home, family, life standard, socioeconomic status, climate, etc. (Bollati and Baccarelli, 2010). Deletions and insertions of one or few bases or of larger DNA segment (more frequent in non coding regions) can cause both loss and gain of function. Monogenic diseases are relatively uncommon and rare and have strong familial inheritance, and thus, they can be well investigated through pedigree-based approaches, whereas a large unrelated case-control sample set is mainly used in GWAS. Once an approximate chromosomal location is recognized for the gene, investigators then have to inch along the chromosome to reach the actual gene using various molecular techniques, where the whole procedure is referred to as positional cloning.

One of the methods for gene mapping is through linkage studies, which rely on the co-segregation of causal variants with marker alleles within pedigrees. Since the frequency of recombination events per meiosis is relatively low, tagging a casual variant requires only few genetic markers per chromosome (Visscher et al., 2012). Linkage analysis is a powerful statistical approach used for mapping genes not only in rare monogenic disorders but in common complex traits as well. Parametric and nonparametric approaches can be used to detect linkage. The former method also called LOD score analysis requires specified parameters (e.g., mode of inheritance, marker frequency, allele frequency, penetrance, etc.), while the latter does not need the specification of parameters. Parametric methods are more powerful than nonparametric methods in detecting linkage, but nonparametric methods can be a better alternative in cases where the disease model cannot be assumed. In linkage analysis, genetic maps are effectively generated using either microsatellite markers or SNPs, both in candidate genes and in whole genome scan. Combination of the information from linkage analysis with advanced high-throughput genotyping and deep sequencing technology better guide researchers in identifying susceptibility genes in a cost-effective way. The limitation, on the other hand, is low mapping resolution, which means how close to the casual variant one can get through linked markers.

Another method is to search for a variety of deleterious genes in a particular group that exhibits characteristics such as originating from a small founding group, inbreeding, or a high incidence of several different disease genes. With the discovery that the human genome is riddled with small genetic differences called SNPs coupled with the publication of the human genome sequences, pedigree-independent strategies became more popular. For example, in the "candidate gene" method, the investigator hypothesizes a gene or variation in the gene that might have led to a particular genetic disability. This method requires prior knowledge of a particular gene in the pathogenesis of the disorder. The gene and surrounding DNA is analyzed through case-control association studies for comparison among people with and without disease to figure out alterations specific to people having the disease phenotype (Singh et al., 2014). Similarly, allele specific methods and direct sequencing of entire candidate region in both cases and controls are also exploited. To identify the novel variants, both common and rare, the entire gene with 3' and 5' flanking regions, entire exonic and intronic and regulatory

regions are deep sequenced with overlapping primer sets to generate a contig of whole candidate region. This is although a very expensive method, but it is a robust method for identification of novel common, rare and uncommon DNA sequence variants.

GWAS are completely independent and involve the comparison of thousands to millions of variants throughout the genome between cases and controls. GWAS have the potential for discovering differences related to genes that might not normally have been implicated in causing disease. This technique is mainly based on the principle of linkage disequilibrium (LD), the nonrandom association between alleles at different loci, at the population level (The International HapMap Consortium, 2005). Tagger SNPs (representative variants of a certain region) are selected from HapMap and dbSNP data sets of a given ethnic group to increase the power of the study and to reduce the overall cost of the investigation. Thus, common haplotypes can be identified that may tag many uncommon and disease-causing variants (Dickson et al., 2010). There are various factors that affect the overall power of GWAS to identify susceptibility genes, such as MAFs, LD between the markers, causal variants, phenotype, and the most important one, the size of the sample set. The cases should be properly diagnosed with standard clinical and investigation protocols and matching controls of the same ethnicity should be ascertained.

A discovery study, based on GWAS, is conducted by analyzing a complete panel of markers with a small sample set, and the most significant markers for the phenotype in question are selected and replicated in a separate sample set of cases and controls (Marian, 2012). Analysis of GWAS data requires extensive statistical expertise to increase the significance level and to reduce the chance of false positive associations. Usually, millions of SNPs are tested in a single GWAS against a large sample set, which generates a huge amount of data, and thus, a P value of less than 1×10^{-8} is considered to be a level of significant association between the marker and phenotype. The genomic sequence will also provide useful knowledge about intron/exon structure, promoters and other regulatory regions, clustering of related genes, syntenic relationships with model organism genomes, and overall chromosomal organization (Loman et al., 2012). Genetic maps are now very advanced and promise to become even more effective for candidate gene identification with the advent of high-density SNP collections for association studies.

All these methods are summarized in Figure 1,which shows that especially the GWAS are particularly well adapted for finding genetic factors underlying complex genetic diseases. About 1200 GWAS have identified the association of several loci in more than 200 complex diseases (Marian, 2012). Therefore, our understanding of the genetic basis of disease is beginning to improve with the help of large-scale GWAS and high-throughput sequencing technologies, although many molecular and physiological studies need to be conducted to confirm association with disease etiology. Geneticists have realized that they could exploit a statistical analysis dependent on population-based LD to map genes, which suggests that fine-mapping using population association could lead to closer linkage between a causative mutation and linked marker. Although GWAS are unbiased with respect to prior biological knowledge and genome location, they are not unbiased in terms of what is detectable. GWAS rely on LD between genotyped SNPs and ungenotyped casual variants. It is clear that for most complex traits that have been investigated by GWAS, various loci identified have genome-wide statistical significance (false positive) and there are (many) other loci that have not been identified because of a lack of statistical significance (false negative) (Anderson et al., 2011).

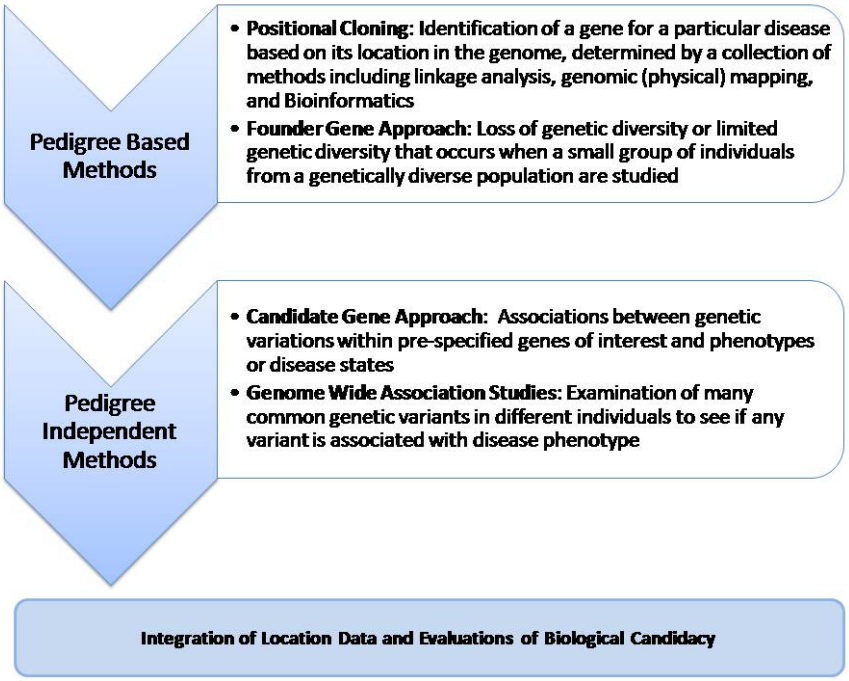

Figure 1. Molecular approaches to seek disease genes.

Computer algorithms and programs for linkage analysis have provided the foundation of most disease-gene discovery projects, and similarly, the databases of clinical findings are extensively used to support diagnostic decisions. One of the major challenges in biomedical informatics is to overcome the lack of standards for the collection, manipulation and sharing of research findings. Clinical medicine and research in human genetics is still far behind in incorporating ontologies and information technology as done by microbial geneticists, so the development and implementation of various standards have been a major focus in recent past. The first naming scheme for genetic disease developed by Victor McKusick and colleagues at Online Mendelian Inheritance in Man (OMIM) has long been used in the research community. Hamosh et al. (2013) provided an overview of the logic behind the OMIM nomenclature. The Orphanet portal, developed especially for clinical aspects of rare diseases (www.orpha.net), offers information for professionals and the layman regarding prognosis, available treatments, drugs, expert centers, clinical trials, patients' organizations, and other related concerns about rare diseases. For various disorders, it is important to record the phenotypic abnormalities along with diagnosis so that diseases can be clustered according to their phenotypic similarities. This clustering can be used to understand molecular relationships of the genes involved.

Proteins from disease genes usually exhibit distinct properties from the rest of genes; for example, they are longer, more conserved, phylogenetically extended and without close paralog. The interaction network of disease genes was recently the focus of research (Kann, 2007). Large scale experiments generate lists of several hundreds of disease gene candidates, and it is still a

challenge to identify the disease genes among them. Using distinguishable properties of a disease gene, computational tools could be developed to prioritize disease gene candidates.

Despite the success of linkage analysis in case of Mendelian disorders, involvement of different factors, both genetic and environmental, makes the identification of genetic traits or variants responsible for complex disease a troublesome task. Modern sequencing methods have tried to solve the complications by analyzing genetic variation between individuals in a fast and highly accurate manner. By particularly analyzing variations in single bases, supported by computational tools such as the ENCODE project and 1000 genome projects, we can understand disease mechanism at the molecular level (Martin et al., 2013).

A few years back, the bioinformatics data available were mainly derived from academic studies of individual human genes with just few extended regions. The advent of expressed sequence tags (ESTs) was a milestone, which also stirred research interest in bioinformatics (Mattick et al., 2010). Although the research was disconnected and highly error prone, bioinformatics data proved its worth even when simple BLAST (basic local alignment search tool) searches revealed exciting glimpses of novel genes in huge numbers. Similarly, gene-oriented UniGene resources consisting of EST sets were anchored on derived unique 3' sequences. ESTs proved helpful in studies involving splicing with few limitations such as artifactual ESTs containing intronic sequences and inappropriately grouped ESTs. Just two decades ago, only the beta-globin gene cluster of 73 kb on chromosome 11 was the most impressive stretch of contiguous human sequence available for bioinformatic analysis. To date, it is likely that a significant portion of all human genes are not represented in dbEST, with the reason being their low abundance or highly specific distribution in tissues or time expression.

Besides EST, there are many other available technologies to analyze patterns of gene expression, such as counting ESTs from various libraries that contribute to different transcripts or at least clusters and hybridization-based techniques, using “chips” or microarray gridding. They not only provide the opportunity to detect straightforward differences in the expression of individual genes, but also provide new opportunities to coordinate pattern detection (Norton et al., 2011). Going beyond expression data, efforts in proteomics can be expected to enhance the understanding of post-transcriptional events. Structural genomics is also drawing more attention, which is expected to increase the collection of protein structures conferring the effectiveness of approaches to functional genomics.

There are many examples of ad hoc database systems that are designed to analyze biological sequences, e.g., ACEDB. However, general purpose databases, with standardized and versatile query capability, are the backbone of most large-scale sequence databases. As discussed earlier, the advantages of the object-oriented databases are undermined because of the lack of standardization (Peterson et al., 2010). Therefore, the profusion of intelligent integration is required such as in the case of NCBI’s LocusLink or GeneCards. Restriction caused by the number of databases, degree of overlap, rapid changes in field and funding issues have devastating consequences, such as the end of some databases, e.g., Genome Database at Johns Hopkins University, and the commercialization of others, such as SWISS-PROT. On the other hand, a recent trend is an increase in the number of useful views on the data and comprehensive scope offered by institutions such as NCBI, which is also now subject to more active curation, as in the RefSeq collection.

Table 1 lists some databases used to study disease-associated genetic loci, such as HGMD, which includes a comprehensive core collection of data on published germ line muta-

tions in nuclear genes underlying human inherited diseases. Similarly, COSMIC, is designed to collect and display somatic mutation information and related details, including information concerning human cancers. TIDBase, on other hand, is a public website and database focusing only on type 1 diabetes (TID). Annotation of human variation data with protein structural information and other functionally relevant information are dealt with in MutDB (Jex et al., 2010). ModSNP is a portal to search for variants in Swiss-Prot entries of the UniProt Knowledgebase (UniProtKB) and gives direct access to the Swiss-Prot Variant page. SAAPdb is used for the integration of information on SNPs with analysis of the likely structural effects of these amino acid mutations (Xi et al., 2010).

Table 1. Database for disease-associated genetic loci.

Serial No.	Name	Web address	Reference
1	T1Dbase	http://www.t1dbase.org	(Hulbert et al., 2007)
2	COSMIC	http://www.sanger.ac.uk/genetics/CGP/cosmic/	(Forbes et al., 2008)
3	The European Genome-Phenome Archive	https://www.ebi.ac.uk/ega/	(Church et al., 2010)
4	ModSNP	http://swissvar.expasy.org	(Yip et al., 2004)
5	SwissVar	http://swissvar.expasy.org/	(Mottaz et al., 2010)
6	HGMD	http://www.hgmd.cf.ac.uk/ac/index.php	(Stenson et al., 2003)
7	Catalog of published Genome Wide Association Studies (NHGRI)	http://www.genome.gov/gwastudies/	(Gong et al., 2011)

Variant phenotype interaction

Understanding clinical phenotypes through their corresponding genotypes is essential to unveil inherited alterations that lead to pathological processes and syndromes. However, such comprehension can be very difficult with complex disorders, which frequently present different clinical phenotypes resulting from interactions between multiple and potentially unknown genetic loci. Moreover, different genetic alterations may cause very similar or even the same phenotype (Cantor et al., 2004). Thus, complex and multivariate analyses of the molecular processes underlying phenotypically similar disorders are required to obtain insights into the composite gene and protein interactions. To carry out these tasks computationally, structured information and controlled vocabularies are available and describe biological processes and molecular functions. Table 2 lists the databases and tools, which could be used to analyze the effects of genetic mutations on a particular phenotype.

Table 2. Databases to study effects of genetic mutations.

Serial No.	Program	Web address	Reference
1	SDM	www-cryst.bioc.cam.ac.uk/~sdm/sdm.php	(Worth et al., 2007)
2	FoldEF/FoldX	Foldx.crg.es/	(Guerois et al., 2002)
3	I-MUTANT	Gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant/I-Mutant.cgi	(Capriotti et al., 2004)
4	MU-Pro	Mupro.proteomics.ics.uci.edu/	(Cheng et al., 2006)
5	CUPSAT	Cupsat.tu-bs.de	(Parthiban et al., 2006)
6	ERIS	Troll.med.unc.edu/eris/login.php	(Yin et al., 2007)
7	Polyphen	Genetics.bwh.harvard.edu/	(Adzhubei et al., 2010)

Generally, the effects of mutation on protein stability or function are verified by comparing a mutated sequence with the wild-type protein sequence. The oligomeric state

of wild-type and mutant proteins are used to be analyzed by chemical cross-linking, size exclusion chromatography, analytical ultracentrifugation, dynamic light scattering, gel filtration chromatography and ion mobility-mass spectrometry analysis. The functional effects of mutations are investigated by measuring the effects of a mutation on binding to other proteins, such as co-immunoprecipitation followed by Western blotting, or to a ligand. Determining the experimental structure of wild-type and/or mutant proteins can provide further insight into the mechanism underlying observed biophysical or functional differences (Schnabel and Erlander, 2012). These types of experiments are certainly limited to investigate the functional effects of mutations *in vitro*, whereas to understand physiological effects of mutations, they need to be tested *in vivo*, for instance, using genetically engineered animal models. However, the sheer volume of SNP data generated in recent years from projects such as the Human Genome Project, HapMap Project and GWAS makes it nearly impossible to characterize all SNPs in this way. It has been estimated that up to 80% of disease-associated SNPs are a consequence of protein stabilization effects, and since this analysis was carried out on monogenic diseases, this pattern may not be valid for complex diseases (Kuhlenbäumer et al., 2011).

Using the collective results of various computational tools, such as SDM with functional site predictions made by CRESCENDO (Chelliah et al., 2004), and observed interaction sites stored in the databases PICCOLO, BIPA (Lee and Blundell, 2009) and CREDO (Schreyer and Blundell, 2009), the structural and functional effects of SNPs can be differentiated, thereby potentially aiding the identification of the causative mechanism of a disease. Bonds ON Graph (Bongo), on the other hand, is a graph-based method that analyzes the likelihood of point mutations causing disease by affecting its corresponding protein structures (Cheng et al., 2006). For a target mutation, Bongo identifies two sets of key residues from the residue interaction network of its corresponding wild-type and mutant protein structure. Another particular feature of Bongo is its pure prerequisite of structural information, since it considers a protein internal network alone, and thus, the prediction result is complementary to other contemporary approaches, such as SDM, SIFT and PolyPhen 2 (Adzhubei et al., 2010).

Another database called SAMUL provides comprehensive structural and functional annotations of amino acid residues and amino acid variations, which can be browsed and interpreted in reference with the structural and functional environments of wild-type amino acid residues. SAMUL stores amino acid sequence variants from the *Homo sapiens* genome annotation provided by the human variation database Ensemble, cancer somatic mutations from COSMIS, and UniProt human sequence variations (Yip et al., 2008). UniProt is a main store for protein sequences, which provides rich annotation on function and cross references. By having the knowledge of protein structure, investigators are able to gain a better understanding of its function, which allows the development of pharmacological agents to manipulate its activity. Recent advances have also been made in high-resolution *abinitio* protein structure prediction, where a model structure is built without the use of a template structure.

Cancer genetics

In case of cancer, knowledge of the human genome is used for anticancer drug design.

As genetic variation plays a key role in cancer risk, disease outcome is therefore based on the analysis of the genome and transcriptome to identify molecular changes in cancer tissue (Ozsolak and Milos, 2011). Genome analysis usually includes comparative genome hybridization and SNP analyses, whereas transcriptome analysis comprises genome wide gene expression profiling methods such as microarray, RNAi knockdown of gene expression and analysis of alternative splicing (Hammond et al., 2001). In cancer genetics, the basic requirement is to manage large datasets, and therefore, many molecular sequence data repositories are available, which are updated on a daily basis, as summarized in Table 3. Resources such as UniGene and RefSeq, present at the National Center for Biotechnology Information website, support new gene discovery and are supported by repositories including GenBank, EMBL, DDBJ, Genome Browser, Ensemble, and Golden Path server.

Table 3. Sequence data repositories for genetic changes involved in cancer.

Sr No.	Databases and resource	URL
1	GenBank	www.ncbi.nlm.nih.gov
2	EMBL	www.ebi.ac.uk/
3	DDBJ	www.ddbj.nig.ac.jp
4	Genome Browser	genome.ucsc.edu
5	Ensemble	www.ensembl.org
6	Golden Path Server	genome.ucsc.edu
7	Cancer Genome Anatomy Project	cgap.nci.nih.gov
8	Cancer Biomedical Informatics Grid	cabig.nci.nih.gov
9	National Cancer Institute Computational Biology Group	ncicb.nci.nih.gov
10	Biomolecular Interaction Network Database	www.bind.ca
11	Human Proteome Organization	www.hupo.org
12	Protein Structure Initiative	www.nigms.nih.gov/research/specificareas/PSI/pages/default.aspx
13	Gene drug disease association	www.pharmgkb.org/index.jsp

There are some special initiatives for cancer research, which include Cancer Genome Anatomy, Cancer Biomedical Informatics Grid, National Cancer Institute Computational Biology Group, Biomolecular Interaction Network Database, Human Proteome Organization, Protein structure Initiative, etc. The major role of bioinformatics in this field is to analyze the sequences and molecular data to define the differences between cancerous and normal tissue, as well as to collect information on all the human genes and proteins for genome-based therapies (Lind et al., 2010). Therefore, analysis is based on comparative genomics, including the analysis of the types of genes, gene families, and the location of genes on the chromosomes of various organisms. In the case of the human genome, there are tandem duplicated regions present that represent the regions of genetic instability and that are often associated with human disease.

Another important role is the analysis of sequence variations within gene promoters of primate species to search for conserved regions (Kingsley, 2011). As the sequence variation between haplotype blocks have shown genetic diversity in the human population, the major challenge in cancer bioinformatics is to design high-throughput data collection methods and bioinformatic tools to uncover the genomic variations that impair protein function and influence gene regulation and expression. SNPs provide a major contribution to the prediction of disease manifestations and drug side effects. They also provide information regarding loss of heterozygosity, which is used to identify genomic regions that harbor tumor suppressor genes and to characterize different tumor types. SNP analysis is a

common approach in cancer research to choose a candidate gene, screen for SNPs, and then to determine haplotypes, haplotype frequencies, and risk associated with each haplotype. However, this analysis requires consideration of sample size, for example, which must be large and should include a randomized scan of a large population in genetic equilibrium, and the interaction of SNPs with other genes and candidate gene must also be taken into account (Tang et al., 2004).

Cancer and noncancer haplotyping can be analyzed by using many web-based tools such as dbSNP (www.ncbi.nlm.nih.gov), SNP Consortium and linkage analysis. Databases are also available for splice variants (www.mdc-berlin.de) (Lee et al., 2003). Alternative splicing has been observed in 40% of the human genome, and splice variants in cancer cells and tissues are often analyzed by gene expression profiling using microarrays. For microarray experiments, the most difficult task faced by the biologists is to analyze the lost lists of genes from an expression microarray experiment. Therefore, an attempt is made to analyze genes by gene ontology, biochemical function, known biochemical and regulatory relationships, and known protein-protein and gene-gene interactions. To solve this, Pathway Miner is an important tool that produces spreadsheets of gene and pathway information and interactive graphs depicting known regulatory relationships among the upregulated and downregulated genes.

A totally different approach of microarray data analysis with understanding of genes as a prerequisite could be used, with prediction of complex gene relationships if more information on genes is incorporated in the analysis. The recent usage of information on the human genome and proteome has depended heavily on advances in bioinformatics, and similarly, drug discovery also relies heavily on well-integrated data management to identify drug targets and determine drug-gene interactions (Katsios et al., 2012). Integrated bioinformatic approaches for analyzing these target genes, their nucleotide polymorphisms, protein structures, protein-protein interactions and protein modification sites for the degradation, activation and sorting of the receptors are essential to understanding the response of individuals to such drugs.

Databases, tools and algorithms for SNP at DNA and protein level

Analyses of protein structure and function have revealed that SNPs are responsible for most (60%) inherited diseases (Wineinger et al., 2011). SNPs connect phenotype to genotype with the potential to be used in the study of disease prognosis. Although the pace of collecting SNP data is impressive, progress in annotating SNPs is relatively slow. The identification of disease-associated SNPs via informatic approaches is becoming a major challenge and requires urgent attention. There are just few thousand SNPs that have been found to be associated with a human genetic disorder (Hamosh et al., 2013), and as a consequence, understanding the contribution of SNPs to disease remains complicated and controversial. In terms of genetic effects, SNPs can cause various effects according to their location in the genome, where they could be regulatory if present in transcription initiation sites or may affect mRNA splicing sites (Musunuru et al., 2010). In the case of non synonymous SNPs (causing amino acid change), they can alter protein function, protein stability, protein aggregation or post-translational modifications. Table 4 summarizes the tools and databases that focus on SNP analysis and their effects on phenotypes.

Table 4. Tools and databases involved in SNP analysis.

Serial No.	Databases and resource	URL	Purpose
1	dbSNP	www.ncbi.nlm.nih.gov/SNP/	*De facto* central DNP database
2	HGMD	www.hgmd.cf.ac.uk/ac/index.php	Human Gene mutation Database
3	OMIM	www.ncbi.nlm.nih.gov/omim	Online mendelian inheritance of man
4	PharmGKB	www.pharmgkb.org	Pharmacogenetics knowledge base
5	dbGAP	www.ncbi.nlm.nih.gov/gap	Database of genotype and phenotype
6	PyMOL	Delanoscientific.com	Visualization of protein structure
7	Endeavouralgorithm	tomcat.esat.kuleuven.be/endeavour	Prioritization on the basis of machine learning
8	GeneSeeker	www.cmbi.ru.nl/geneseeker/	Produces list of candidate disease genes based on cytogenetic localization and expression
9	Gene2Disease	g2d2.ogic.ca	Identifies candidate disease gene by doing a homology search on Gene Ontology
10	SUSPECTS	www.genetics.med.ed.ac.uk/suspects/	Combines scores from PROSPECTER, InterPro, and expression liberaries
11	TOM	www-micrel.deis.unibo.it/~tom	Identifies candidate genes involved in inherited diseases
12	PRIORITIZER	genenetwork.nl/wordpress/prioritizer	Uses Bayesian approach to classify genes that are associated in disease
13	Gentrepid	www.gentrepid.org/	Gene prediction using structural bioinformatics and system biology
14	PhenoPred	www.phenopred.org	Uses available protein interactions, gene function, sequence features, and disease information to predict candidate gene
15	FitSNPs	Fitsnps.stanford.edu/	Human Gene Expression
16	LS-SNP	Modbase.compbio.ucsf.edu/LS-SNP	Large Scale human SNP annotation
17	SNPs3D	snps3d.org	Protein structure annotation
18	MutDB	www.mutdb.org	Annotate protein structure
19	PolyDoms	Polyview.cchmc.org/polyview3d.html	Protein structure annotation
20	Uniprot	www.pir.uniprot.org	Universal protein resource
21	SNPeffect	Snpeffect.vib.be/	Predicts functional site disruption on protein sequences
22	SIFT	http://sift.cchmc.org	Sorting intolerant from tolerant SNPs
23	PolyPhen	Genetics.bwh.harvard.edu/pph/	Polymorphism phenotyping
24	PMut	Mmb2.pcb.ub.es:8080/PMut/	Protein structure annotation
25	SAP	Sapred.cbi.pku.edu.cn/	Protein structure annotation
26	SNAP	http://snap.genomics.org.cn	Protein structure annotation
27	Parepro	www.mobioinfor.cn/prepro/	Predicting the amino acid replacement probability
28	PANTHER	www.pantherdb.org/	Protein analysis through evolutionary relationships
29	PolyMAP	www.owl-ontologies.com/Ontology_UT_Dallas.owl#polyMAP"	Polymorphisms minning and annotation program
30	SNPSeek	Snp.wustl.edu/cgi-bin/SNPseek/index.cgi	Protein structure annotation
31	PupaSuit	Pupasuite.bioinfo.cipf.es/	Protein structure annotation
32	SNP@Promoter	Variome.kobic.re.kr/SNPatpromoter/	Protein structure annotation
33	TRANSFAC	www.biobase-international.com/pages/index.php?id=transfec	Transcription factor database

CONCLUSION

The implication of the discovery of candidate genetic variation in human genetics can hardly be overstated. Methods to predict the effect of polymorphisms on protein stability are useful for the identification of possible disease associations, but with advances in computational tools together with rapid and economical sequencing, scientists can move ahead towards an era of personalized medicine. Personalized medicine will permit us to obtain profiles for a vast number of gene variants that increase the risk of an individual of having a genetic disease,

especially for complex disorders such as cancer.

ACKNOWLEDGMENTS

Research supported by CNPq, Brazil, and Higher Education Commission, Pakistan.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7: 248-249.

Anderson CA, Boucher G, Lees CW, Franke A, et al. (2011). Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* 43: 246-252.

Bailey-Wilson JE and Wilson AF (2011). Linkage analysis in the next-generation sequencing era. *Hum. Hered.* 72: 228-236.

Bollati V and Baccarelli A (2010). Environmental epigenetics. *Heredity* 105: 105-112.

Burgess DJ (2011). Human Disease: Next-generation sequencing of the next generation. *Nat. Rev. Genet.* 12: 78.

Cantor S, Drapkin R, Zhang F, Lin Y, et al. (2004). The BRCAl-associated protein BACH1 is a DNA helicase targeted by clinically relevant inactivating mutations. *Proc. Natl. Acad. Sci. U. S. A.* 101: 2357-2362.

Capriotti E, Fariselli P and Casadio R (2004). A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* 20 (Suppl 1): 163-168.

Chelliah V, Chen L, Blundell TL and Lovell SC (2004). Distinguishing structural and fünctional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.* 342: 1487-1504.

Cheng J, Randall A and Baldi P (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* 62: 1125-1132.

Church DM, Lappalainen I, Sneddon TP, Hinton J, et al. (2010). Public data archives for genomic structural variation. *Nat. Genet.* 42: 813-814.

Dickson SP, Wang K, Krantz I, Hakonarson H, et al. (2010). Rare variante create synthetic genome-wide associations. *PLoS Biol.* 8: el000294.

Forbes SA, Bhamra G, Bamford S, Dawson E, et al. (2008). The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr. Protoc. Hum. Genet.* Chapter 10: Unit 10.11.

Gong S, Worth CL, Cheng TM and Blundell TL (2011). Meet me halfway: when genomics meets structural bioinformatics. *J. Cardiovasc. Transl. Res.* 4: 281-303.

Guerois R, Nielsen JE and Serrano L (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* 320: 369-387.

Hammond SM, Caudy AA and Hannon GJ (2001). Post-transcriptional gene silencing by double-stranded RNA. *Nat. Rev. Genet*. 2: 110-119.

Hamosh A, Sobreira N, Hoover-Fong J, Sutton VR, et al. (2013). PhenoDB: a new web-based tool for the collection, storage, and analysis of phenotypic features. *Hum. Mutat.* 34: 566-571.

Hulbert EM, Smink LJ, Adiem EC, Allen JE, et al. (2007). TIDBase: integration and presentation of complex data for type 1 diabetes research. *Nucleic Acids Res.* 35: D742-D746.

Jex AR, Hall RS, Littlewood DT and Gasser RB (2010). An integrated pipeline for next-generation sequencing and annotation of mitochondrial genomes. *Nucleic Acids Res.* 38: 522-533.

Kann MG (2007). Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief. Bioinform.* 8: 333-346.

Katsios C, Ziogas DE, Liakakos T, Zoras O, et al. (2012). Translating câncer genomes sequencing revolution into surgical oncology practice. *J. Surg. Res.* 173: 365-369.

Kingsley CB (2011). Identification of causal sequence variante of disease in the next generation sequencing era. *Methods Mol. Biol*. 700: 37-46.

Kuhlenbäumer G, Hullmann J and Appenzeller S (2011). Novel genomic techniques open new avenues in the analysis of monogenic disorders. *Hum. Mutat.* 32:144-151.

Lee S and Blundell TL (2009). BIPA: a database for protein-nucleic acid interaction in 3D structures. *Bioinformatics* 25: 1559-1560.

Lee C, Atanelov L, Modrek B and Xing Y (2003). ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res.* 31:101-105.

Lind C, Ferriola D, Mackiewicz K, Heron S, et al. (2010). Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. *Hum. Immunol.* 71:1033-1042.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, et al. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30: 434-439.

Marian AJ (2012). The enigma of genetics etiology of atherosclerosis in the post-GWAS era. *Curr. Atheroscler. Rep.* 14: 295-299.

Martin TA, Lane J, Ozupek H and Jiang WG (2003). Claudin-20 promotes an aggressive phenotype in human breast cancer cells. *Tissue Barriers* 1: e26518.

Mattick JS, Taft RJ and Faulkner GJ (2010). A global view of genomic information - moving beyond the gene and the master regulator. *Trends Genet.* 26: 21-28.

McClellan J and King MC (2010). Genetic heterogeneity in human disease. *Cell* 141: 210-217.

Mottaz A, David FP, Veuthey AL and Yip YL (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics* 26: 851-852.

Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, et al. (2010). From noncoding variant to phenotype via SORT1 at the lpl3 cholesterol locus. *Nature* 466: 714-719.

Norton N, Li D, Rieder MJ, Siegfried JD, et al. (2011). Genome-wide studies of copy number variation and exome sequencing identify rare variants in BAG3 as a cause of dilated cardiomyopathy. *Am. J. Hum. Genet.* 88: 273-282.

Ozsolak F and Milos PM (2011). Transcriptome profiling using single-molecule direct RN A sequencing. *Methods Mol. Biol.* 733: 51-61.

Parthiban V, Gromiha MM and Schomburg D (2006). CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* 34: W239-W242.

Peterson TA, Adadey A, Santana-Cruz I, Sun Y, et al. (2010). DMDM: domain mapping of disease mutations. *Bioinformatics* 26: 2458-2459.

Rubio C, Bellver J, Rodrigo L, Bosch E, et al. (2013). Preimplantation genetic screening using fluorescence *in situ* hybridization in patients with repetitive implantation failure and advanced maternal age: two randomized trials. *Fertil. Steril.* 99: 1400-1407.

Schnabel CA and Erlander MG (2012). Gene expression-based diagnostics for molecular cancer classification of difficult to diagnose tumors. *Expert. Opin. Med. Diagn.* 6: 407-419.

Schork AJ, Thompson WK, Pham P, Torkamani A, et al. (2013). All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet.* 9: el003449.

Schreyer A and Blundell T (2009). CREDO: a protein-ligand interaction database for drug discovery. *Chem. Biol. Drug Des*. 73: 157-167.

Schunkert H, Kõnig IR, Kathiresan S, Reilly MP, et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* 43: 333-338.

Singh H, Arentson BW, Becker DF and Tanner JJ (2014). Structures of the PutA peripheral membrane flavoenzyme reveal a dynamic substrate-channeling tunnel and the quinone-binding site. *Proc. Natl. Acad. Sci. U.S. A.* 111: 3389-3394.

Singleton AB, Hardy J, Traynor BJ and Houlden H (2010). Towards a complete resolution of the genetic architecture of disease. *Trends Genet.* 26: 438-442.

Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, et al. (1997). Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* 20: 399-407.

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

Tang K, Oeth P, Kammerer S, Denissenko MF, et al. (2004). Mining disease susceptibility genes through SNP analyses and expression profiling using MALDI-TOF mass spectrometry. *J. Proteome Res.* 3: 218-227.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* 437*:* 1299-1320.

Visscher PM, Brown MA, McCarthy MI and Yang J (2012). Five years of GWAS discovery. *Am. J. Hum. Genet.* 90: 7-24.

Wineinger NE, Patki A, Meyers KJ, Broeckel U, et al. (2011). Genome-wide joint SNP and CNV analysis of aortic root diameter in African Americans: the HyperGEN study. *BMC Med. Genomics* 4: 4.

Worth CL, Bickerton GR, Schreyer A, Forman JR, et al. (2007). A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease. *J. Bioinform. Comput. Biol.* 5: 1297-1318.

Xi R, Kim TM and Park PJ (2010). Detecting structural variations in the human genome using next generation sequencing. *Brief. Funct. Genomics* 9: 405-415.

Yaneva-Deliverska M (2011). Rare diseases and genetic discrimination. *J. IMAB* 17:116-119.

Yin S, Ding F and Dokholyan NV (2007). Eris: an automated estimator of protein stability. *Nat. Methods* 4: 466-467.

Yip YL, Scheib H, Diemand AV, Gattiker A, et al. (2004). The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum. Mutat.* 23: 464-470.

Yip YL, Famiglietti M, Gos A, Duek PD, et al. (2008). Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Hum. Mutat.* 29: 361-366.