Journal of Bioinformatics and Computational Biology
Vol. 5, No. 5 (2007) 1123–1138

ICP Imperial College Press
www.icpress.co.uk

A GENTLE INTRODUCTION TO SNP ANALYSIS: RESOURCES AND TOOLS

JAMES T. L. MAH

Data Mining Department, Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
tlmah@i2r.a-star.edu.sg

K. S. CHIA

Centre for Molecular Epidemiology
Department of Community, Occupational and Family Medicine
National University of Singapore
16 Medical Drive, Singapore 117597

Received 20 June 2007
Revised 15 July 2007
Accepted 16 July 2007

Bioinformatics is the use of informatics tools and techniques in the study of molecular biology, genetic, or clinical data. The field of bioinformatics has expanded tremendously to cope with the large expansion of information generated by the mouse and human genome projects, as newer generations of computers that are much more powerful have emerged in the commercial market. It is now possible to employ the computing hardware and software at hand to generate novel methodologies in order to link data across the different databanks generated by these international projects and derive clinical and biological relevance from all of the information gathered. The ultimate goal would be to develop a computer program that can provide information correlating genes, their single nucleotide polymorphisms (SNPs), and the possible structural and functional effects on the encoded proteins with relation to known information on complex diseases with great ease and speed. Here, the recent developments of available software methods to analyze SNPs in relation to complex diseases are reviewed with emphasis on the type of predictions on protein structure and functions that can be made. The need for further development of comprehensive bioinformatics tools that can cope with information generated by the genomics communities is emphasized.

Keywords: Single nucleotide polymorphism; complex disease; nonsynonymous.

1. Introduction

Bioinformatics is the use of informatics tools and techniques in the study of molecular biology, genetic, or clinical data. In recent years, there has been a large influx of international projects such as the human[1] and mouse[2] genome projects, necessitating the need for more powerful computing systems and tools to cope with

the large amounts of clinical and scientific data generated. The field of bioinformatics has expanded tremendously to cope with this need, as newer generations of computers that are much more powerful have emerged in the commercial market. These computer systems can handle and analyze enormous quantities of data easily and efficiently. It is now possible to employ the computing hardware and software at hand to generate novel methodologies in order to link data across the different databanks generated by these international projects and derive clinical and biological relevance from all of the information gathered.[3] An important goal would be to develop a computer program that can provide information correlating genes, their single nucleotide polymorphisms (SNPs), and the possible structural and functional effects on the encoded proteins with relation to known information on complex diseases with great ease and speed.

With the rapid advancement of the Human Genome Project (HGP), the major phases of which have finished, there is an increasing need to study naturally occurring sequence variations and understand human DNA polymorphisms. As SNPs comprise about 90% of the human DNA polymorphisms,[4] SNPs are instrumental in understanding the structure of the human genome and fathoming the causation of many disease and pathological processes.

Studies have elucidated that in the genome, SNPs occur about every 0.3–1 kilobases (kb).[5] An SNP is a mutation with a single DNA base substitution observed with a frequency of at least 1% in a given population.[6] Broadly, SNPs can be intronic or exonic in nature. Intronic SNPs are located in intronic or "silent" regions of the genome, and are thought not to have any effect on protein products if translated. Nonsynonymous SNPs (nsSNPs) are exonic SNPs which are situated in coding or exonic regions of the genome, and result in amino acid variants in the protein products or changes in protein length due to their effects on stop codons. In addition, SNPs are thought to be responsible for almost 90% of the interindividual variability, and about 100,000 amino acid differences.[7] However, in practice, as genomic DNA insertions and deletions of both multiple and single bases are frequently found during SNP discovery processes, these variants are also deposited in SNP databases.[7] Hence, the term SNP is loosely taken to include both single-base and multiple-base changes in the genomic DNA.[7]

SNPs are generally biallelic systems having two alleles for any particular marker or genetic locus in an individual. This means that the information content per SNP marker is relatively low when compared to microsatellite markers, which may have more than 10 alleles per marker. It has been estimated that it will take approximately five SNP markers to equal the information of each microsatellite marker, and approximately 2,000 SNPs will be required to equal a 10 centimorgan (cM) microsatellite map.

As SNPs are usually population-specific, this implies that a marker that is polymorphic in one population may not be very polymorphic in another. This means that polymorphic SNP markers will have to be generated to specifically target the population under study. By comparison, microsatellite markers have been shown to

be polymorphic across populations. This means that once they are generated, they can be used universally.

SNP markers are increasingly used in genetic analysis as they offer a number of benefits. SNPs, found approximately every kilobase, offer the potential for generating very high-density genetic maps, which are extremely useful for haplotyping genes or chromosomal regions of interest. They may also be the polymorphisms associated with the disease phenotypes under study. The low mutation rate of SNPs enables them to act as excellent markers for studying complex genetic traits.

Modern day genetics define genetic diseases to be caused by visible chromosomal defects and by mutant genes that disrupt specific function(s) of a single gene. These defects mostly cause Mendelian or single-gene diseases, of which more than 1,500 have been discovered. Most of these diseases are rare and can be identified by their specific patterns of transmission (dominant, recessive, X-linked).[8] However, there are many common chronic diseases with an adult-onset pattern that show a familial aggregation and do not follow a simple Mendelian family inheritance pattern. These diseases appear to be caused by multiple genes, usually interacting with environmental factors.[9]

The term "complex trait" refers to any phenotype that does not display classic, Mendelian recessive, dominant, or codominant inheritance patterns attributable to a single genetic locus. Often, many genes interact with environmental factors to produce the final complex phenotype displayed in the disease.[10] Genetic traits which are caused by a combination of environmental factors and mutations in multiple genes interacting with one another are known as complex disease traits or complex diseases. Such conditions include coronary heart disease, hypertension, diabetes, obesity, Alzheimer's disease, and Parkinson's disease.[8]

Due to the multifactorial nature of complex disease traits, it can be difficult to dissect out which are the single causal factors in disease causation. Therefore, there is a need for efficient bioinformatics tools to aid in the dissection of causal factors involved in complex diseases. This article serves to give the reader a general overview of the resources and bioinformatics tools available to achieve this. To this aim, we organize our review into the following three main sections. Section 2 presents SNP databases. Section 3 presents tools for manipulation of SNP data. Section 4 presents tools for exploring the function of SNPs.

2. SNP Databases for Research

At present, public SNP databases are estimated to contain more than five million raw SNPs spread throughout the genome and deposited in two major public repositories.[11] They are the National Center for Biotechnology Information (NCBI) databases of genetic variation named Single Nucleotide Polymorphism database (dbSNP)[12,13] (http://www.ncbi.nlm.nih.gov/SNP/) and the Human Genome Variation database (HGVbase)[14] (http://hgvbase.cgb.ki.se/). These databases are described in Secs. 2.1 and 2.2. Besides these two major public repositories, there are

SNP trace files which reside in the NCBI trace repository providing information on SNPs. If an SNP is HapMap-verified, one would not be interested in these trace files. However, if one was interested in an unverified SNP that was sequence-determined, the SNP itself as well as the quality of the surrounding data in the originating trace would be of interest to help one make one's own estimation of whether or not to believe that the SNP exists, that is, the SNP is not a sequence artifact.

2.1. *dbSNP database*

The most widely used public SNP database is dbSNP[13] (http://www.ncbi.nih.gov/SNP/), which is responsible for the collection and archival of SNPs discovered by computational means or direct observation submitted by the scientific community. It is an archival data resource of discovered genomic and cDNA sequence variations, mainly comprising single-base mutations (SNPs), microsatellite or short tandem repeats, and short deletions and insertions from any species,[13] with detailed information such as their assay and discovery methods, validation information, flanking sequence polymerase chain reaction (PCR) conditions,[10] and population statistics. The public can conveniently access its complete contents online together with other information sources like GenBank,[15] PubMed, LocusLink, and the Human Genome Project data at NCBI which are integrated with submissions to dbSNP.[8] Although the details on SNPs provided are very comprehensive, it lacks direct information on the types of disease the SNP could be associated with. The user has to click on multiple links to other sites to obtain this information. The average SNP density is found to be approximately 1.1 SNP per kilobase in dbSNP.[8] Altogether, there are about four million nonredundant human SNPs compiled from about 300 sources in the dbSNP database,[11] and the number is growing rapidly at a rate of about 90 SNPs per month.[13]

2.2. *HGVbase database*

The HGVbase (formerly known as the Human Genic Bi-Allelic Sequences or HGBASE) (http://hgvbase.cgb.ki.se/) is a gene-oriented database[16] cataloging known human intragenic sequence variations.[17,18] Information in the database is supplied by a European academic consortium involving Karolinska Institute (KI), European Bioinformatics Institute (EBI), and European Molecular Biology Laboratory (EMBL), with support from Interactiva GmbH (Germany).[14] This database consists of multiple types of variations, with SNPs predominating.[14] The information it provides online focuses on how SNPs are related to gene functions, giving details of human gene–related (promoter, exonic, and intronic) SNPs.[16] In contrast to dbSNP, HGVbase consists of about one million nonredundant human SNPs from nearly 800 unique sources.[11]

As the HGVbase[14] focuses on very precise manual curation and annotation of SNP information, it is smaller in comparison to dbSNP. However, it is very useful for queries on the genetic aspects of human phenotypic variation where detailed

information is required. The web browser at HGVbase includes information on the type of disease the SNP is associated with, which dbSNP lacks.

3. Manipulation of Information from SNP Databases

Due to the large amount of SNP information present in SNP databases, many new web-based applications have been developed by bioinformaticians to endeavor to facilitate the extraction of relevant SNPs from databases and to incorporate them in formats that are convenient to the user. Since none of them are exhaustive, the goal is to improve upon available systems and provide a unique and exhaustive platform for users to work on. The most notable of the web-based applications used by researchers are described below in Secs. 3.1–3.7.

3.1. *The UCSC Genome Browser Database*

Due to the exponential increase in the amount of gene sequence information in public databases and assemblies, accurate and complete annotation of information in these databases is important for data analysis and interpretation.[19] The University of California, Santa Cruz (UCSC) Genome Browser Database (http://genome.ucsc.edu) provides up-to-date access to current and archived genome assemblies as well as annotation data with each assembly.[19] Annotations include mRNA and expressed sequence tag (EST) alignments, gene predictions, cross-species homologies, and SNPs.[19] The web-based UCSC Genome Browser or UCSC Table Browser data retrieval tool is an interactive utility built on top of the database optimized for the rapid visualization and query of data in graphical and text-based formats.[19] It is based on the storage of sequence and annotation data in a MySQL relational database, which enables retrieval of data from indexed files.[19] Multifaceted functions of the browser make it a popular tool for data retrieval for scientists. The advance queries option allows the user to integrate and configure the results from separate tables to see common features in the genome. Flexibility in the filtering option allows for the setting of constraints for values and retrieving subregions of features in the query fields.[20]

3.2. *SNPper*

SNPper[6] (http://snpper.chip.org/) is a web-based application that automatically extracts SNP data from different databases. After analyzing them, it creates sets of candidate SNPs that fit certain user-defined criteria and exports the results in suitable formats like XML or HTML for further analysis. The user-defined criteria include SNP frequency, reliability of SNP information, location of the SNP (intergenic or intragenic), types of amino acid changes caused by the SNP, and their regulatory characteristics (promoter, intron, exon).[6] The results can also be viewed graphically in a web-based interface. Furthermore, the SNP filtering function allows

for additional refining of the dataset such as lowering the search distance and thus adding more SNP entries to be analyzed.

SNPper[6] depends on a local relational database and real-time connection to GoldenPath,[21] the draft human genome browser. The database is compiled from parsing dbSNP and GoldenPath, producing information on almost two million SNPs and over 15,000 genes.[6] Postprocessing will connect SNPs to related genes in the database and their location in the gene. If the SNP falls within a gene's coding region, it can determine the type of amino acid change.

The web-based interface, which allows user access, is written in the Common Lisp language.[22] Unique features in SNPper are mainly twofold: SNP sets management and facilities which are interoperable. The main aim of SNPper is to generate a collection of SNPs with some desired characteristics that lie in a specified region in the genome. This collection of SNPs is named an SNP set.[6] SNPper can uniquely manipulate, refine, save, and export SNP sets, and generate SNP sets through complex queries such as one using Gene Ontology classes.[23] It is the only SNP database known to the author enabling users to gain full access in the form of machine-readable XML files via a Remote Procedure Call Interface.[23] The user can upload, analyze, and display the SNP datasets through this interface. Compared to SNP search and analysis tools on dbSNP and GoldenPath websites, SNPper allows for good interoperability with other applications.[6] It reduces a lot of manual work on the part of researchers, since it produces SNP data results that are user-defined and easily exportable. However, at present, SNPper has no facility for disease or pathological process search. In this respect, SNPper is less useful for the clinical scientist interested in which SNPs are located in which genes that play a role in the causation of complex diseases. Database information is also limited to humans.[23]

3.3. *PARSESNP*

PARSESNP[24] (Project Aligned Related Sequences and Evaluate SNPs) (http://www.proweb.org/parsesnp/) is a web-based data analysis tool used in forward genetics[25] and reverse genetics.[26] This tool extracts polymorphism data from medical data in the Human Gene Mutation Database[27] (http://www.hgmd.org/), dbSNP, and the results of reverse genetic screens in plants[28] (http://tilling.fhcrc.org:9366/). It then integrates the information collected and displays the output in graphical or tabular forms. PARSESNP[24] flexibility is in its ability to curate polymorphism data in DNA sequence or translated protein sequence forms from different databases, and compares them with cDNA reference or genomic sequences.[24] It can determine the effects of nucleotide changes such as missense mutations on the gene and protein product and on restriction enzyme polymorphisms and restriction enzyme sites.[24] However, there is no facility for disease or pathological process search, hence making it less useful for the clinical scientist interested in SNPs that cause diseases.

3.4. *Functional Element SNPs Database*

The Functional Element SNPs Database[29] (FESD) tool is designed to organize functional elements into categories in human gene regions and to output their flanking sequences needed for genotyping experiments as well as provide a set of SNPs that lie within each area. In FESD,[29] human gene regions are divided into ten separate functional elements, namely promoter regions, CpG islands, 5′-untranslated regions (5′-UTRs), translation start sites, splice sites, coding exons, introns, translation stop sites, polyadenylation signals, and 3′-UTRs; following this, all known SNPs are allocated to each functional element at their specific location. By utilizing the FESD[29] web interface freely available at http://combio.kribb.re.kr/ksnp/resd/, researchers can manually choose a group of SNPs of special interest in certain functional elements along with their flanking sequences. The selected SNP datasets are used in gene-based haplotype or linkage studies and sequencing experiments, in the hope of discovering mutations that cause or contribute to complex diseases.[29] The advantages of FESD[29] over browsers in public databases like dbSNP,[13] UCSC Genome Browser (http://genome.ucsc.edu), or Ensembl browser (http://www.ensembl.org/index.html)[30] are that FESD provides user-friendly interfaces for working with a set of SNPs for biologists who are not familiar with public databases. This software application uses Perl scripts to parse data from the NCBI FTP site and import it into a MySQL relational database,[29] without the user having to handle Perl. For the clinical scientist, the disorder or clinical synopsis search facility in FESD is useful in identifying genes in diseases. Although the software attempts to link genes and disease processes, however, it does not link SNPs directly to any disease causation. The user has to manually link and infer which SNPs are located in specific genes and hence are involved in a particular disease of interest.

3.5. *The Genetic Association Database*

The Genetic Association Database[31] (GAD) (http://geneticassociationdb.nih.gov/) is a comprehensive public archival database of more than 5,000 public human genetic association studies. It allows systematic analysis of common complex human genetic diseases in the context of current high-throughput assay methods. As it is web-based and provides molecular and clinical search parameters, it is easily accessible to the scientific community in molecular biology and clinical disciplines.[31] Often in reporting genetic association studies, the scientific literature is filled with arbitrary gene names and symbols, making cross-comparisons and meta-analysis problematic. GAD overcomes this problem by using official Human Genome Organisation (HUGO) gene symbols to standardize gene nomenclature in its database.[31] By annotating each record and providing individual links to other databases such as HapMap and PubMed, cross-referencing and integration to other databases is efficient and systematic.[31]

3.6. *The SNP Consortium*

The SNP Consortium (TSC) was formed in 1999 as a collaboration effort from companies and institutions to develop an SNP public resource in the human genome.[32] The aim was to discover 300,000 SNPs in 2 years, but the end results surpassed this. A total of 1.4 million SNPs were released into the public domain at the end of 2001[33] by this Consortium.

The SNP Consortium website (http://snp.cshl.org) allows users to perform gene or SNP keyword searches and browse or deposit data into text files.[32] A graphical genome browsing interface displays SNPs mapped onto the genome assembly incorporating externally available gene predictions.[32] Other options include being able to view features on the genome assembly, zoom in and out, and customize which features are displayed. SNP allele frequency and genotype data can be downloaded via FTP-download and on separate SNP report web pages.[32] The user can download relevant SNP linkage maps and browse them in a comparative map viewer.[32]

Although the SNP Consortium website is relatively easy to use and navigate, it lacks the facility to search for SNPs involved in disease pathology.

3.7. *The Japanese JSNP database*

Started in 2000, JSNP (http://snp.ims.u-tokyo.ac.jp/) is a repository of Japanese Single Nucleotide Polymorphism (SNP) data. It was developed through the Prime Minister's Millennium Project, and was a collaboration between the Human Genome Center (HGC) of the Institute of Medical Science (IMS) at the University of Tokyo and the Japan Science and Technology Corporation (JST).[34] The goal was to identify and collect up to 150,000 SNPs from the Japanese population that were either located in genes or nearby regions which may influence the genes. JSNP acts as a facility for public use to allow researchers to obtain high-quality SNP data. The project essentially constructs a basic dataset to identify relationships between polymorphisms and common diseases or their reactions to drugs. Therefore, much emphasis is on the identification of SNPs that lie in candidate regions which may affect the phenotype, but not necessarily cause disease.[34] Although JSNP has a set of tools available for multiple queries including PCR profiles, drug reactions, basic local alignment search tool (BLAST) SNP, dbSNP, OMIM disease, and mapping, it does not have at present facilities for protein sequence information searches. Online Mendelian Inheritance in Man (OMIM), available at http://www.ncbi.nlm.nih.gov/omim/, is a database which catalogs human genes and genetic disorders.

4. Existing Software Tools to Explore the Function of SNPs

Besides the software and websites described above in previous sections, there are specific tools in the World Wide Web to explore the function of SNPs, such as to predict the effects of SNPs on proteins and final products. Sections 4.1–4.4 describe

the characteristic features of SNP effect prediction programs, and give two examples of these program as well as tools for functional analysis of SNPs such as SNP3D, Hap, and PupaSNPfinder.

4.1. *SNP effect prediction programs*

Although association and linkage studies as experimental techniques are commonly used to identify SNPs underlying complex disorders, they are very costly due to the large number of markers to be screened.[35] This is because linkage disequilibrium studies are based on whole-genome scanning.[36] Candidate gene studies[36,37] try to rectify this problem by reducing the number of SNPs to be studied to those that are located in genes mostly likely involved in the disease. However, if a large number of candidate genes are studied, multiple testing of large numbers of SNPs makes detection of association difficult.[35]

Clearly, to reduce the number of SNPs that need to be tested in candidate gene studies, one can prioritize SNPs according to their putative functional[37] and physiological significance and choose the important ones for the experiments. In addition, bioinformatics tools can help select significant SNPs by discriminating between neutral SNPs (which exert minimal or no effect) and functional SNPs (which may have some detrimental effect on protein functions and hence the associated disease).[38] Disease-associated nsSNPs can be distinguished from neutral nsSNPs by empirical rule-based and machine learning approaches.[38] Empirical rules differentiating disease-associated from neutral nsSNPs have been derived from structural information,[39] evolutionary information,[40] or both.[41] Classification models automatically learnt from training data were also derived by other studies.[42–44] With the exception of Wang and Moult,[39] all of the mentioned studies utilized position-specific evolutionary information in multiple sequence alignments for their prediction. Sunyaev *et al.*[41] estimated that ∼20% of common human nsSNPs damage proteins and that an average human carries about 2,000 harmful amino acid variants.

Two popular bioinformatics tools available in the World Wide Web, PolyPhen and SIFT, which are designed to predict amino acid changes that affect protein function, are summarized below.

4.2. *PolyPhen*

PolyPhen (Polymorphism Phenotyping) (http://genetics.bwh.harvard.edu/pph/) is a World Wide Web server performing the automatic functional annotation of coding nsSNPs.[35] The server annotates any SNP present in the HGVbase database.[35] All of the annotated datasets of SNPs can be accessed at http://www.bork.embl-heidelberg.de/PolyPhen/data.[35] PolyPhen, through a fully automated sequence of several programs constructed based on straightforward empirical rules, predicts the possible impact of an amino acid substitution on protein structure and function and then outputs the results in a user-friendly HTML interface. Input query is in

the form of amino acid or protein sequence in classical FASTA[45] format. The user can also input the SNP rs identifier number to access any available predictions stored in the PolyPhen server database for the human protein on that particular SNP. PolyPhen is useful and effective in analyzing large databases like dbSNP. For example, using PolyPhen to analyze dbSNP build 121, PolyPhen predicts that out of 50,919 entries in dbSNP that were mapped to human proteins, approximately 19%–20% are possibly damaging SNPs.

4.3. *SIFT*

Sorting Intolerant From Tolerant (SIFT) is a useful program available at http://blocks.fhcrc.org/sift/SIFT.html that can predict whether an amino acid substitution affects protein function and hence could potentially alter phenotype. It enables users to prioritize amino acid substitutions in protein sequences and their corresponding SNPs for further studies.[46] Ng and Henikoff[46] demonstrated that SIFT can differentiate between functionally neutral and deleterious amino acid substitutions in mutagenesis studies and on human polymorphisms. For example, SIFT had a false-positive error rate of $\sim 30\%$ in its prediction for neutral substitutions that did not alter LacI function.[47]

Presently, it is estimated that amino acid substitutions account for about half of the known gene lesions that cause human inherited disease.[48] SIFT uses sequence homology and relies completely on sequence in its prediction.[47] An advantage of SIFT not requiring any structural information is that SIFT can perform predictions on a greater number of substitutions.[46] In addition, as more genomes and protein sequences become available, the number of substitutions that SIFT can predict is expected to increase.[46]

The SIFT prediction algorithm presumes that important amino acids are well conserved in families of homologous proteins, and that changes at these well-conserved or consensus positions are deleterious to protein function. If, for example, a specific amino acid position in an alignment of a protein family only contains the amino acid isoleucine, it is assumed that substitution to another amino acid is selected against in evolution (negative selection) and that isoleucine is essential for protein function. Therefore, a change to an amino acid of different physiochemical properties at this position may be detrimental to protein function. Similarly, if a position in an alignment contains hydrophobic amino acids isoleucine, valine, and leucine, then SIFT assumes, in effect, that this position can only contain amino acids with hydrophobic character. Any changes to other hydrophobic amino acids at this position are predicted to be tolerated, but changes to other hydrophilic residues (such as charged or polar) will be predicted to adversely affect protein structure and function. Such predictions are invaluable in predicting the effects of amino acid substitutions in protein tertiary structures and in biological enzyme catalysis.

The algorithm of SIFT works in three simple steps. First, SIFT considers the position at which the change occurred and the type of amino acid change. Second,

given a protein sequence, SIFT chooses similarly related proteins and computes an alignment of these proteins with the query. Finally, SIFT calculates the probability that an amino acid at a position is tolerated conditional on the most frequent amino acid being tolerated. A normalized value is obtained based on the amino acids present at each position in the alignment. If this normalized value falls under a specific cutoff value, SIFT will predict that the amino acid substitution at that position is deleterious to protein function.[40]

4.4. *Other tools for functional analysis of SNPs*

There exist many other tools which serve to analyze SNP function. Below is a summary of three other tools, namely SNPs3D, Hap, and PupaSNPfinder. A web resource and database which can give information on disease/gene relationships at the molecular level is SNPs3D.[49] SNPs3D (http://www.SNPs3D)[49] is a tool that assigns molecular functional effects of nonsynonymous SNPs based on structure and sequence analysis methods. It has three primary modules. The first module identifies candidate genes involved in disease. The second module identifies relationships between different sets of candidate genes. Lastly, the impact of nsSNPs on protein function is analyzed by the third module.[49] SNP/protein function relationships are derived using principles of protein structure and stability as well as sequence conservation. Gene–gene interactions are displayed in an interactive graphical interface, which allows access to underlying information and navigation through the network.[49]

Recent studies show that haplotypes are structured blocks with limited diversity.[50] There are many methods of resolving haplotype data from genotypes of SNPs present in the population. An efficient algorithm for partitioning genotypes of SNPs into blocks and generating predictions for each block is Hap.[50] One inputs a population of genotypes, and Hap partitions the SNPs into blocks which show limited diversity.[50] For each block, Hap predicts the common haplotypes and haplotypes of individuals in the population sampled.[50] This resource for the prediction of haplotype structure is publicly available at http://www.calit2.net/compbio/hap.

PupaSNPfinder (PupaSNP)[51] is a web-based tool designed for high-throughput search of SNPs with potential phenotypic effect. The input for PupaSNP is a list of genes. It can also generate the required gene list from chromosomal coordinates.[51] PupaSNP retrieves SNPs that could affect conserved regions on intron/exon boundaries or exonic splicing enhancers, predicted transcription factor binding sites (TFBSs), and amino acid changes in proteins.[51] The program utilizes mapping information of SNPs provided by Ensembl. In the case of user-defined SNPs which are not yet mapped in the genome, PupaSNP can handle these with ease as well.[51] The advantage of PupaSNP as compared to other programs which analyze SNPs and their effects on proteins is that it includes SNPs with possible transcriptional effect.[51] PupaSNP is useful in studies of multifactorial disorders, where the analysis

of functional SNPs will enhance the sensitivity of identification of candidate genes in the disease.[51] PupaSNP is available at http://pupasnp.bioinfo.cnio.es.

5. Conclusion

As the number of SNP entries deposited in public databases continues to increase, there is an urgent need for more powerful and comprehensive computational tools to analyze and study them. These tools are essential to consolidate, integrate, and rationalize information from different databases and to provide an easy, up-to-date, complete, nonredundant, and unique database with user-friendly tools to query it. This end product should be freely available on the World Wide Web for the scientific community to easily use and access. In addition, international efforts to help derive all of the functional elements in the human genome sequence, such as the ENCyclopedia Of DNA Elements (ENCODE) Project,[52] will greatly facilitate *in silico* bioinformatics tools with the discovery of SNPs and their functions.

As the relationship between SNPs and risk factors of complex diseases becomes increasingly apparent, it is impractical and costly to solely depend on high-throughput laboratory methods to screen for candidate genes and causative SNPs. This is due to the increasingly large number of candidate genes and SNPs present in SNP and genome databases. It is too expensive and time-consuming to follow all of these potential candidates. There is therefore a need for cheaper and more

Table 1. Summary of SNP databases.

Name	URL	Comments
dbSNP	http://www.ncbi.nih.gov/SNP/	SNP collection/archive discovered by computational means or direct observation
HGV base	http://hgvbase.cgb.ki.se/	Gene-oriented database cataloging known human intragenic sequence variations

Table 2. Summary of tools for the manipulation of SNP data.

Name	URL	Comments
UCSC	http://genome.ucsc.edu/	Annotations of SNPs for data analysis
SNPper	http://snpper.chip.org/	Web application for SNP details extraction
PARSESNP	http://www.proweb.org/parsesnp/	Web tool for extraction of polymorphic data and data analysis
FESD	http://combio.kribb.re.kr/ksnp/resd/	Tool that organizes gene functional elements and SNPs, and performs searches on them
GAD	http://geneticassociationdb.nih.gov/	Public archive of human genetic association studies
SNP Consortium	http://snp.cshl.org/	Huge SNP public resource in the human genome
JSNP	http://snp.ims.u-tokyo.ac.jp/	Repository of Japanese SNP data

Table 3. Summary of tools which can be used to analyze functions of SNPs.

Name	URL	Comments
PolyPhen	http://genetics.bwh.harvard.edu/pph/	World Wide Web server performing the automatic functional annotation of coding nonsynonymous SNPs
SIFT	http://blocks.fhcrc.org/sift/SIFT.html/	Tool for prediction of amino acid substitution effect on protein function by SNP
SNPs3D	http://www.SNPs3D/	Tool which assigns molecular function effects to nonsynonymous SNPs
Hap	http://www.calit2.net/compbio/hap/	Tool for the resolution of haplotype data from genotypes of SNPs in the population
PupaSNP	http://pupasnp.bioinfo.cnio.es/	Web tool to search for SNPs with phenotypic effect

practical methods like *in silico* bioinformatics prediction tools to first select and prioritize a smaller subset of potential disease-associated SNPs and genes before high-throughput genomic screening methods are used to further study and analyze them. Tables 1–3 summarize the key resources and tools surveyed in this article.

Lastly, these methods described above for analyzing SNPs will be increasingly important in clinical medicine, especially in the field of pharmacogenomics. Pharmacogenomics is the study of how inherited variations in genes and SNPs dictate drug responses and predict how individuals respond to drugs by correlating gene expression data or SNPs with a drug's efficacy and toxicity. This field of science aims to develop methods to optimize drug treatments according to each patient's personal genotype in order to ensure maximum efficacy with minimal side effects. Such approaches herald the arrival of "personalized medicine" or "evidence-based medicine," where drug therapies are optimized for each individual's unique genetic makeup.

References

1. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, *Nature* **409**:860–921, 2001.
2. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K *et al.*, Initial sequencing and comparative analysis of the mouse genome, *Nature* **420**:520–562, 2002.
3. Baxevanis AD, Francis Ouellette BF (eds.), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, Wiley, New York, 1998.
4. Collins FS, Brooks LD, Chakravarti A, A DNA polymorphism discovery resource for research on human genetic variation, *Genome Res* **8**:1229–1231, 1998.
5. Schork NJ, Fallin JD, Lanchbury JS, Single nucleotide polymorphisms and the future of genetic epidemiology, *Clin Genet* **58**:250–264, 2000.
6. Riva A, Kohane IS, SNPper: Retrieval and analysis of human SNPs, *Bioinformatics* **18**:1681–1685, 2002.
7. Brookes A, The essence of SNPs, *Gene* **234**:177–186, 1999.
8. Motulsky AG, Genetics of complex diseases, *J Zhejiang Univ Sci B* **7**:167–168, 2006.

9. Davey Smith G, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR, Genetic epidemiology and public health: Hope, hype, and future prospects, *Lancet* **366**:1484–1498, 2005.
10. Marsh DG, Approaches toward the genetic analysis of complex traits: Asthma and atopy, *Allergy* **54**:198–205, 1999.
11. Jiang R, Duan J, Windemuth A, Stephens JC, Judson R, Xu C, Genome-wide evaluation of the public SNP databases, *Pharmacogenomics* **4**:779–789, 2003.
12. Sherry ST, Ward M, Sirotkin K, dbSNP — Database for single nucleotide polymorphisms and other classes of minor genetic variation, *Genome Res* **9**:677–679, 1999.
13. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K, dbSNP: The NCBI database of genetic variation, *Nucleic Acids Res* **29**:308–311, 2001.
14. Fredman D, Siegfried M, Yuan YP, Bork P, Lehväslaiho H, Brookes AJ, HGVbase: A human sequence variation database emphasizing data quality and a broad spectrum of data sources, *Nucleic Acids Res* **30**:387–391, 2002.
15. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL, GenBank, *Nucleic Acids Res* **28**:15–18, 2000.
16. Aerts J, Wetzels Y, Cohen N, Aerssens J, Data mining of public SNP databases for the selection of intragenic SNPs, *Hum Mutat* **20**:162–173, 2002.
17. Brookes AJ, Lehväslaiho H, Siegfried M, Boehm JG, Yuan YP, Sarkar CM, Bork P, Ortigao F, HGBASE: A database of SNPs and other variations in and around human genes, *Nucleic Acids Res* **28**:356–360, 2000.
18. Fredman D, Munns G, Rios D, Sjöholm F, Siegfried M, Lenhard B, Lehväslaiho H, Brookes AJ, HGVbase: A curated resource describing human DNA variation and phenotype relationships, *Nucleic Acids Res* **1**(Database issue):D516–D519, 2004.
19. Karolchik D, Baertsch R, Diekhans M *et al.*, The UCSC Genome Browser Database, *Nucleic Acids Res* **31**:51–54, 2003.
20. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ, The UCSC Table Browser data retrieval tool, *Nucleic Acids Res* **32**(Database issue):D493–D496, 2004.
21. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D, The human genome browser at UCSC, *Genome Res* **12**:996–1006, 2002.
22. Riva A, Ramoni M, LispWeb: A specialized HTTP server for distributed AI applications, *Comput Netw ISDN Syst* **28**:953–961, 1996.
23. Riva A, Kohane I, A SNP-centric database for the investigation of the human genome, *BMC Bioinformatics* **5**:33, 2004.
24. Taylor NE, Greene EA, PARSESNP: A tool for the analysis of nucleotide polymorphisms, *Nucleic Acids Res* **31**:3808–3811, 2003.
25. Ulrich CM, Bigler J, Sibert J, Greene EA, Sparks R, Carlson CS, Potter JD, Cyclooxygenase 1 (COX1) polymorphisms in African-American and Caucasian populations, *Hum Mutat* **20**:409–410, 2002.
26. Till BJ, Reynolds SH, Greene EA *et al.*, Large-scale discovery of induced point mutations with high-throughput TILLING, *Genome Res* **13**:524–530, 2003.
27. Krawczak M, Cooper DN, The human gene mutation database, *Trends Genet* **13**:121–122, 1997.
28. Colbert T, Till BJ, Tompa R, Reynolds S, Steine MN, Yeung AT, McCallum CM, Comai L, Henikoff S, High-throughput screening for induced point mutations, *Plant Physiol* **126**:480–484, 2001.
29. Kang HJ, Choi KO, Kim BD, Kim S, Kim YJ, FESD: A functional element SNPs database in human, *Nucleic Acids Res* **33**:D518–D522, 2005.
30. Hubbard TJP, Aken BL, Beal K *et al.*, Ensembl 2007, *Nucleic Acids Res* **35**(Database issue):D610–D617, 2007.

31. Becker KG, Barnes KC, Bright TJ, Wang SA, The genetic association database, *Nat Genet* **36**:431–432, 2004.
32. Thorisson GA, Stein LD, The SNP Consortium website: Past, present and future, *Nucleic Acids Res* **31**:124–127, 2003.
33. Sachidanandam R, Weissman D, Schmidt SC *et al.*, A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, *Nature* **409**:928–933, 2001.
34. Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y, JSNP: A database of common gene variations in the Japanese population, *Nucleic Acids Res* **30**:158–162, 2002.
35. Ramensky V, Bork P, Synyaev S, Human non-synonymous SNPs: Server and survey, *Nucleic Acids Res* **30**:3894–3900, 2002.
36. Risch NJ, Searching for genetic determinants in the new millennium, *Nature* **15**:847–856, 2000.
37. Emahazion T, Feuk L, Jobs M, Sawyer SL, Fredman D, St Clair D, Prince JA, Brookes AJ, SNP association studies in Alzheimer's disease highlight problem for complex disease analysis, *Trends Genet* **17**:407–413, 2001.
38. Bao L, Cui Y, Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information, *Bioinformatics* **21**:2185–2190, 2005.
39. Wang Z, Moult J, SNPs, protein structure, and disease, *Hum Mutat* **17**:263–270, 2001.
40. Ng PC, Henikoff S, Predicting deleterious amino acid substitutions, *Genome Res* **11**:863–874, 2001.
41. Sunyaev S, Ramensky V, Koch I, Lathe III W, Kondrashov A, Bork P, Prediction of deleterious human alleles, *Hum Mol Genet* **10**:591–597, 2001.
42. Chasman D, Adams RM, Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation, *J Mol Biol* **307**:683–706, 2001.
43. Saunders CT, Baker D, Evaluation of structural and evolutionary contributions to deleterious mutation prediction, *J Mol Biol* **322**:891–901, 2002.
44. Krishnan VG, Westhead DR, A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function, *Bioinformatics* **19**:2199–2209, 2003.
45. URL: http://dove.embl-heidelberg.de/Blast2/fasta.html.
46. Ng PC, Henikoff S, SIFT: Predicting amino acid changes that affect protein function, *Nucleic Acids Res* **31**:3812–3814, 2003.
47. Ng PC, Henikoff S, Accounting for human polymorphisms predicted to affect protein function, *Genome Res* **12**:436–446, 2002.
48. Krawczak M, Ball EV, Fenton I, Stenson PD, Abeysinghe S, Thomas N, Cooper DN, Human gene mutation database — A biomedical information and research resource, *Hum Mutat* **15**:45–51, 2000.
49. Yue P, Melamud E, Moult J, SNPs3D: Candidate gene and SNP selection for association studies, *BMC Bioinformatics* **22**:166, 2006.
50. E Halperin, Eskin E, Haplotype reconstruction from genotype data using imperfect phylogeny, *Bioinformatics* **20**:1842–1849, 2004.
51. Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, Robledo M, Dopazo J, PupaSNP Finder: A web tool for finding SNPs with putative effect at transcriptional level, *Nucleic Acids Res* **1**(Web Server issue):W242–W248, 2004.
52. ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements) Project, *Science* **306**:636–640, 2004.

James T. L. Mah is currently a Research Engineer in the Data Mining Department, Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. Concurrently, he is also a Principal Investigator of the Translational Interface Programme Projects at the Oncology Research Institute, National University of Singapore. His academic achievements include Bachelor of Medicine & Bachelor of Surgery, London, UK, and Bachelor of Science in Immunopathology and Infection, London, UK. He holds a National Science Scholarship (Ph.D.) from A*STAR. His area of interest is in clinical bioinformatics, particularly in the use of novel bioinformatics tools in the analysis of single nucleotide polymorphisms in complex diseases.

K. S. Chia is currently a Professor in the Department of Community, Occupational and Family Medicine, National University of Singapore (NUS). Concurrently, he is also the Director of the Centre for Molecular Epidemiology, NUS, and a Foreign Adjunct Professor for Karolinska Institute in Sweden. His academic achievements include Bachelor of Medicine & Bachelor of Surgery, Master of Science in Occupational Medicine, and Doctor of Medicine, all from NUS. At present, some of the projects in which Prof. Chia is the Principal Investigator are a record linkage facility for a cancer studies project within the Singapore Cancer Syndicate, establishment of the infrastructure for a population-based case-control studies project on the molecular epidemiology of cancer, and the Singapore Consortium of Cohort Studies project initiated by the Biomedical Research Council (BMRC).