Sean Mooney
is an Assistant Professor of Medical and Molecular Genetics in the Center for Computational Biology and Bioinformatics at the Indiana University School of Medicine.

Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis

Sean Mooney

Date received (in revised form): 7th December 2004

Keywords: *single nucleotide polymorphism, protein structure/function, candidate gene, human variation, genome, mutation*

Abstract

Since the initial sequencing of the human genome, many projects are underway to understand the effects of genetic variation between individuals. Predicting and understanding the downstream effects of genetic variation using computational methods are becoming increasingly important for single nucleotide polymorphism (SNP) selection in genetics studies and understanding the molecular basis of disease. According to the NIH, there are now more than four million validated SNPs in the human genome. The volume of known genetic variations lends itself well to an informatics approach. Bioinformaticians have become very good at functional inference methods derived from functional and structural genomics. This review will present a broad overview of the tools and resources available to collect and understand functional variation from the perspective of structure, expression, evolution and phenotype. Additionally, public resources available for SNP identification and characterisation are summarised.

INTRODUCTION

Single nucleotide polymorphisms (SNPs; Table 1 provides a glossary of terms) are the most common form of genetic variation in humans comprising nearly 1/1,000th of the average human genome.[1] Traditionally, SNPs are assumed biallelic, ie only two of the four common nucleotides are found in that position, having their least common nucleotide found in greater than 1 per cent of the population. The distribution and function of SNPs are important areas of current research. Reviews are available for understanding how SNPs affect protein structure, the use of SNPs in genetics studies and identifying functional variants in candidate genes.[2–5] This review covers bioinformatics efforts to predict variation that is likely to have a functional effect (and possibly a phenotypic effect) and to classify the downstream molecular effects of those variants.

Research suggests that most SNPs fall in the 95 per cent non-coding region of the genome.[6] Of the SNPs that are near or in a gene, their effect on function is difficult to determine. SNPs can alter the function of DNA, RNA and proteins, and are generally classed by genomic location (Table 2). Non-synonymous SNPs alter the amino acid sequence of the protein product through either amino acid substitution or the introduction of a nonsense/truncation mutation. A variant may also affect the expression or translation of a gene product, either by interrupting a regulatory region or by interfering with normal splicing and mRNA function. This can include SNPs in regulatory regions, synonymous SNPs and intronic SNPs. The molecular effects of variation are now becoming better understood in many cases, and specific examples are discussed in detail later in this paper.

The two types of variation that are usually studied from a functional perspective are polymorphisms with no known phenotype, and phenotypically

Sean Mooney,
Center for Computational Biology and Bioinformatics,
Department of Medical and Molecular Genetics,
Indiana University School of Medicine,
714 N Senate Ave; EF 250,
Indianapolis, IN 46202, USA

E-mail: sdmooney@iupui.edu

Table 1: Glossary of terms

Allele	One of the forms of a variant that occurs at a given locus
Coding	In a region of the genome that is transcribed
Haplotype	The organisation of variation across a chromosome
Missense mutation	A variant that alters a codon to substitute one amino acid for another
Nonsense mutation	A mutation that introduces a stop codon
Rare variant	A variation where the least common allele occurs less than 1 per cent in the population
SNP (single nucleotide polymorphism)	An inherited single nucleotide substitution between individuals of a species. Commonly defined as having the least frequent allele occur at a rate greater than 1 per cent in a population. The most common form of human variation

Table 2: SNP functional classes

Coding SNPs	cSNP	Positions that fall within the coding regions of genes
Regulatory SNPs	rSNP	Positions that fall in regulatory regions of genes
Synonymous SNPs	sSNP	Positions in exons that do not change the codon to substitute an amino acid
Non-synonymous SNPs	nsSNP	Positions that incur an amino acid substitution
Intronic SNPs	iSNP	Positions that fall within introns

annotated or disease-associated variation. Polymorphisms without a known phenotype are usually discovered by SNP screening or genomic analysis, while the phenotypically annotated SNPs are often discovered from association studies. Because human mutations are often inferred to be disease-associated through genetics association studies, these mutations may not be causative; they may only be in linkage with the actual causative allele.

Today, the primary database of polymorphisms is dbSNP,[7] which currently contains more than 5,000,000 validated human SNPs. Disease-associated polymorphisms are available from databases such as OMIM,[8] Swiss-Prot,[9] the Human Gene Mutation Database (HGMD)[10] and HGVBase.[11] Together, these databases represent more than 40,000 non-synonymous, synonymous and non-coding polymorphisms.

The first efforts to understand the patterns of sequence variation in the coding regions of genes were studied on two different sets of genes by Cargill *et al.*[12] and Halushka *et al.*[13] Cargill *et al.* characterised variations in 106 genes that were hand selected for their potential relevance to human disease. The authors estimated between 36 and 54 per cent of the non-synonymous mutations were non-conservative, based on the BLOSUM62 matrix. Lau and Chasman have provided further analysis of these data sets.[14]

Functional bioinformatics approaches have been applied to the analysis of disease-associated mutations, as well. Several recent reports have focused very precisely on where diseased alleles are occurring on protein structures and what the properties of those mutations are. It has been shown that these mutation positions are conserved evolutionarily,[15–19] and that they are relevant to protein structure.[18–21] One of the difficulties in analysing disease-associated mutations is that it is very difficult to obtain a set of neutral alleles for comparison. Most approaches to collecting neutral mutations contain some amount of false positives. Some have compared against nsSNPs, while others have used accepted mutations in other species. Terp *et al.* identified structurally relevant features common in disease-associated mutations from the HGMD.[22] Further analysis of the biophysical and evolutionary distributions of disease-associated mutations was provided by

Ferrer-Costa *et al.*[23] Stitziel *et al.* further analysed the locations of non-synonymous disease-associated polymorphisms by classifying the mutational data into structural classes.[24] Evolutionary conservation and amino acid identities of mutations have been studied in detail. Mooney *et al.* showed that, in general, disease-associated mutations tend to occur in positions that are conserved.[25,26] Vitkup *et al.* examined the frequencies of mutations associated with disease in detail.[27] They found that the prevalence of observed disease-associated mutations correlates strongly with the mutability of the observed genetic code. Together, these reports and the papers discussed in the previous section give a good understanding of the properties of disease-associated mutations. Their positions are conserved evolutionarily and the nature of the mutations are far more likely to be non-conserved than unannotated polymorphisms.

Annotated genome databases are excellent resources for SNP annotations

Variation does not occur randomly across genetic sequences and often occurs in hotspots.[28] It is likely that selection has played a role in the evolution of human genetic variation.[29,30] With this in mind, the following sections summarise the methods and resources available to predict and characterise the function of non-synonymous, synonymous and non-coding polymorphisms using bioinformatics methods.

WEB RESOURCES AND SOFTWARE TOOLS FOR SNP CHARACTERISATION

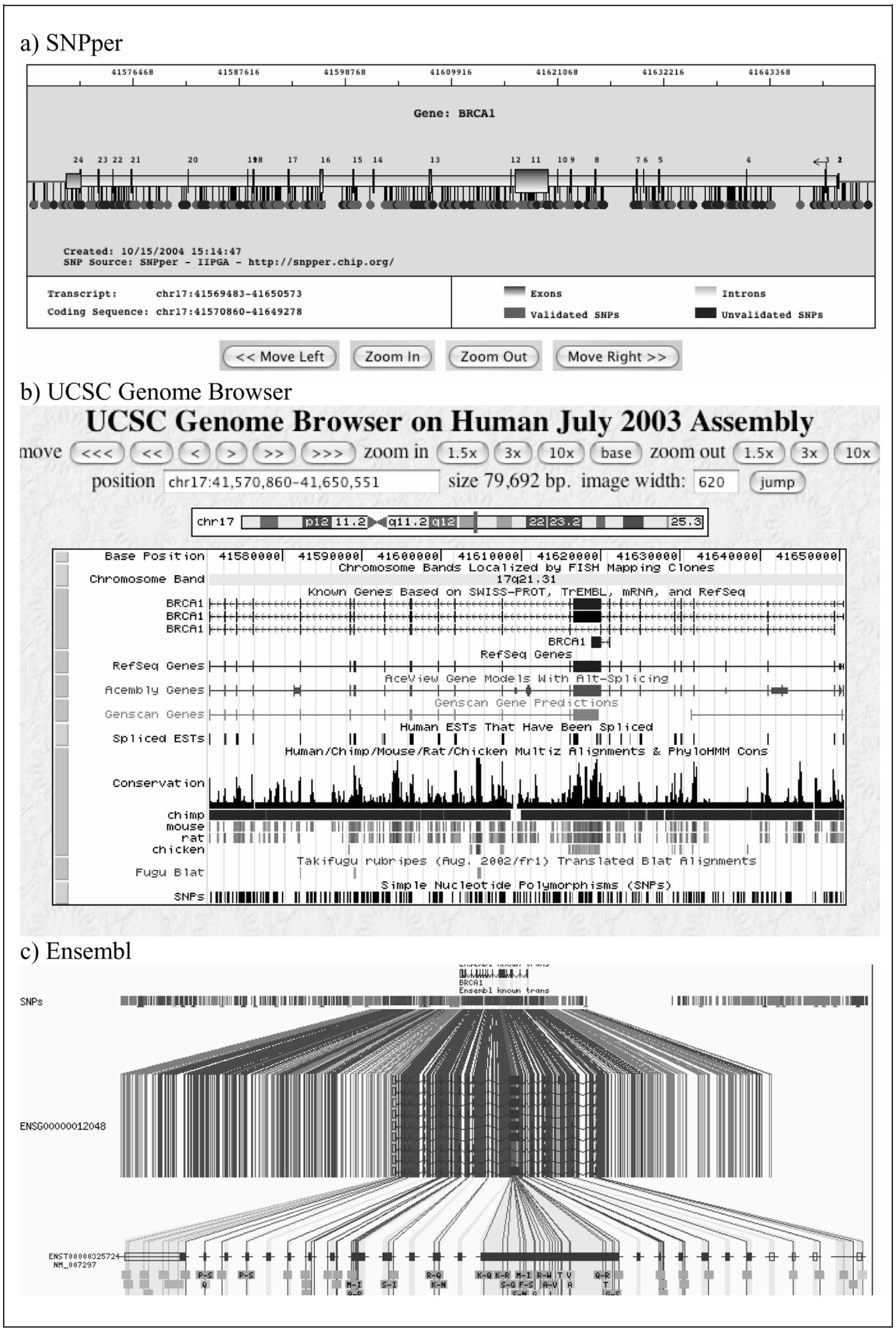
Many resources now annotate variation data with functional information. This section identifies interesting and novel resources for SNP annotation and analysis. The simplest approaches classify variants based on their relationship to genes. Information about whether variants occur near a gene, in a coding region, in an exon, in an intron or up- or downstream of the gene is relatively direct using several genome resources. The NCBI databases, such as dbSNP and OMIM,[31] and Ensembl[32] provide visualisation access and some annotations related to function, based on experiment. Furthermore, with the sequencing of several mammalian genomes, comparative sequence analysis is now possible, even for variation outside of exon boundaries.

Many locus or disease-specific databases are beginning to integrate functional information, such as protein structure, into their annotation sets. The Human Genome Variation Society maintains a comprehensive list of locus-specific databases, which often contain high-quality annotations on variation in specific genes or diseases.[33]

An excellent resource for visualisation of SNP locations and other genome annotations is GoldenPath, the UCSC Genome Browser and genome assembly.[34] Here, an incredible array of annotations has been assembled. Furthermore, the database is completely in the public domain, and trivially importable into a relational database, such as MySQL.[35] The other primary genome resource is Ensembl.[36] Within Ensembl, users can visualise variation in and around genes, and their data annotations are of high quality and embedded into an elegant interface. Another powerful resource for SNP analysis is SNPper.[37,38] SNPper was created in the Kohane Lab at Harvard University for the analysis of SNPs. SNPper focuses on SNP selection for genetic studies, and is a valuable tool for the geneticist. The National Cancer Institute has developed several SNP analysis tools as part of its Cancer Genome Anatomy Project's Genetic Annotation Initiative (CGAP-GAI).[39,40] Figure 1 displays SNPs in the *BRCA1* gene using SNPper, Ensembl and UCSC's GoldenPath Genome Browser. Table 3 summarises SNP databases and tools for SNP analysis.

BIOINFORMATICS APPROACHES TO PREDICTING FUNCTIONAL NON-SYNONYMOUS SNPS

Much effort has been invested in predicting the function of non-


a) SNPper

41576468
41587616
41598768
41609916
41621068
41632216
41643368
Gene: BRCA1
24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2

Created: 10/15/2004 15:14:47
SNP Source: SNPper - IIPGA - http://snpper.chip.org/

Transcript: chr17:41569483-41650573
Coding Sequence: chr17:41570860-41649278

Exons
Introns
Validated SNPs
Unvalidated SNPs
<< Move Left
Zoom In
Zoom Out
Move Right >>

b) UCSC Genome Browser

UCSC Genome Browser on Human July 2003 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position chr17:41,570,860-41,650,551 size 79,692 bp. image width: 620

jump
chr17 p12 11.2 q11.2 q12 22 23.2 25.3
Base Position 41580000 41590000 41600000 41610000 41620000 41630000 41640000 41650000
Chromosome Band
Chromosome Bands Localized by FISH Mapping Clones
17q21.31
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
BRCA1
BRCA1
BRCA1
BRCA1
RefSeq Genes
RefSeq Genes
AceView Gene Models With Alt-Splicing
Acembly Genes
Genscan Gene Predictions
Genscan Genes
Human ESTs That Have Been Spliced
Spliced ESTs
Human/Chimp/Mouse/Rat/Chicken Multiz Alignments & PhyloHMM Cons
Conservation
chimp
mouse
rat
chicken
Takifugu rubripes (Aug. 2002/fr1) Translated Blat Alignments
Fugu Blat
Simple Nucleotide Polymorphisms (SNPs)
SNPs

c) Ensembl

BRCA1
Ensembl known trans
SNPs
ENSG00000012048
ENST00000325724
NM_007297
P-S
Q
P-S
M-I
S-I
R-Q
K-N
K-Q
K-R
M-I
S-G
F-S
R-W
A-V
T
V
A
Q-R
T

Figure 1: Tools for identifying SNPs in relationship to genes and gene structure. The following are visualisations of SNP positions in the *BRCA1* gene. The resources are (a) SNPper,[32] (b) UCSC Genome Browser[41] and (c) Ensembl Genome Browser[36]

synonymous mutations, based on evidence that regulatory and coding SNPs are most likely to affect disease[42–44] and the wide availability of functional data on proteins. An important question is understanding whether a particular mutation will be tolerated.[45] There are several ways an nsSNP can affect gene product function. The most probable effect is a partial or complete loss of function of the mutated gene product. A less likely possibility is a gain of function

Table 3: Online SNP databases

	URL	Comments
Genome resources		
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/	The primary repository for SNP data
Ensembl	http://www.ensembl.org/	Genome database
GoldenPath	http://genome.ucsc.edu/	Genome database
HapMap Consortium	http://www.hapmap.org/	Haplotype block information
JSNP	http://snp.ims.u-tokyo.ac.jp/	Japanese SNP database
Mutation repositories		
HGVBase	http://hgvbase.cgb.ki.se/	Public genotype phenotype database
HGMD	http://www.hgmd.org/	Mutation database with many annotations
Swiss-Prot	http://us.expasy.org/	Protein database with extensive variant annotations
List of locus-specific databases	http://www.genomic.unimelb.edu.au/mdi/dblist/dblist.html	
CGAP-GAI	http://cgap.nci.nih.gov/	Cancer Gene Anatomy Project at the National Cancer Institute
Other databases and tools	http://ymbc.ym.edu.tw/dgd/hdgd.htm	Tools for SNP analysis and gene characterisation
Tools		
SNPper	http://snpper.chip.org/	Novel software for SNP analysis
BioPerl	http://www.bioperl.org/	A programming application program interface (API) for bioinformatics analysis
Genewindow	http://www.genewindow.nci.nih.gov/	Interactive tool for visualisation of variation

mutation, such as those that have been observed in somatic mutations of the androgen receptor ligand binding domain[46] or the activation (by loss of GTPase activity) of the RAS oncogene.[47]

Better characterized proteins often result in better classifications, because more functional and structural information is available

Researchers have taken several approaches to predict the function of nsSNPs. Almost all methods use categories, or discrete or continuous valued features to predict a deleterious mutation. These features range from sequence-based properties, physical properties of the wild-type and mutant amino acids, protein structural properties and evolutionary properties derived from a phylogeny or sequence alignment. To classify whether a mutation will be tolerated, a training set is usually constructed of mutations known to be deleterious. For example, these training sets can be derived from saturation mutagenesis experiments where mutation severity is determined in activity assays,[16,18,21,48,49] multiple sequence alignments where tolerance to mutation is derived from evolutionary analyses of sequence positions,[19] or known deleterious human mutations.[18]

The earliest studies analysed mutations using sequence properties based on how conservative a mutation was, using a BLOSUM62 matrix.[12] BLOSUM62 does not take into account the sequence or structural context of the mutation, so further efforts were employed to include position-specific conservation estimates and protein structural information. Ng and Henikoff continued this body of research by developing a position specific estimation of non-conservative mutations with their method, Sorting Intolerant From Tolerant (SIFT),[49,50] to find that 25 per cent of nsSNPs in dbSNP are likely to affect protein function.[17] SIFT is based on a position-specific scoring matrix (PSSM), and estimates positions that will be unfavourable to mutation, based on tolerated mutations in homologues. SIFT has recently been applied to SNPs in both DNA repair genes and separately to BRCA1.[51,52]

An early study to classify and survey non-synonymous SNPs that included protein structural features was that of Sunyaev *et al.*[20] The authors compared disease-associated mutations in orthologous genes and human cSNPs. To assess local functionality for a given position, both protein structural information and evolutionary information were taken into account. Protein structural parameters such as solvent

Saturation mutagenesis experiments are reasonable datasets for training machine learning methods

accessibility, location within beta strands or active sites, and participation in disulphide bridges were used. Sequence and sequence-based evolutionary conservation were also assessed. The authors found that approximately 70 per cent of disease-associated mutations were in protein structural sites described above and most likely to affect protein function. Additionally, their disease mutations were more likely (35 versus 9 per cent) to be solvent inaccessible than accepted mutations in orthologous genes. Chasman and Adams[16] continued this direction by training and annotating both continuous and categorically valued features on nsSNPs to build a probabilistic classifier of nsSNPs, utilising 16 different structure and evolutionary-based features. They estimated that between 26 and 32 per cent of nsSNPs affect protein function. Performing a similar rigorous analysis using heuristics instead of a probabilistic model, Sunyaev *et al.* published an estimate that approximately 20 per cent of the common nsSNPs would likely have a functional effect.[19] Using a strictly evolutionary approach, Fay *et al.* have estimated that 23 per cent of this deleterious variation is only slightly deleterious.[30] Wang and Moult[53] compared disease-associated variation within the HGMD with SNPs extracted from dbSNP. They classified mutations into five groups for altering protein stability, ligand binding, catalysis, allosteric regulation and post-translational modification. They found that 90 per cent of the disease-associated mutations affect molecular function and that 83 per cent of those that affect molecular function do so through disruption of protein stability. Conversely, they found that 30 per cent of the non-synonymous SNPs are classified as affecting protein stability. Ng and Henikoff followed this with an analysis of previous estimates on the number of functional nsSNPs in a typical genome.[17] They found that in an analysis of the Whitehead Institute nsSNPs the expected false positive error rates suggested that estimates of extrapolations to thousands of functional nsSNPs in a typical human genome are probably overstated.

In order to assess different features for prediction of intolerant mutations, Saunders and Baker[18] followed the analysis of Sunyaev *et al.* and Chasman and Adams with a machine learning perspective on the different proposed features. They applied decision trees and a linear logistic regression to find that a protein structure-derived solvent accessibility term (Cβ density) and an evolutionary term derived from a PSSM matrix (SIFT) were the most accurate terms for prediction. They carefully selected a human allele training set along with saturation mutagenesis data sets from *lac* repressor, HIV-1 protease and T4 lysozyme. The human allele training set was carefully selected to contain known deleterious and neutral alleles, as opposed to alleles that had simply been associated a phenotype. They found that decision trees had an overall classification error of 29.6 per cent on the human alleles and 22.9 per cent on the *in vitro* mutations. They also found that in both human alleles and *in vitro* cases, the SIFT and Cβ density terms classified the best, and that the normalised B-factor and Sunyaev-derived structural rules did not improve classification accuracy when incorporated with the former terms in a combined analysis.

Adding to the feature comparison performed by Saunders and Baker, Krishnan and Westhead[21] compared different approaches to classification. They rigorously compared decision trees to support vector machines (SVMs) and applied these methods to the same *in vitro* mutagenesis data sets as well as to SNPs in the nematode worm species *Caenorhabditis elegans*. One of their findings showed that introducing structural features reduced their error rate. The features that reduced the error rate include mass and hydrophobicity differences, buried charges, solvent accessibility and secondary structure. In another application of machine learning methods, Cai *et al.* applied a Bayesian method for predicting disease-associated SNPs and obtained relatively low false positive error rates, in exchange for a relatively high false negative rate.[48] In general, prediction methods must balance sensitivity and specificity, where low false positive error rates (high specificity) can only be achieved with low sensitivity.

More recently, at the 'Inferring SNP Function' session at the Pacific Symposium of Biocomputing (PSB), Karchin *et al.* continued these efforts to identify the most informative features for predicting deleterious mutations.[54] Using the saturation mutagenesis experiments discussed above, they ranked 32 features using mutual information and found that structural features, such as solvent accessibility of the wild type and mutant as well as an evolutionary term, derived from superfamily multiple alignments.

Currently the state-of-the-art classification tools are based on SVMs or decision trees and the best features for classification are based on structural and evolutionary properties. Structurally, solvent accessibility has consistently been shown to be important in determining whether a mutation will be tolerated.[16,18,19,55] Evolutionarily, non-tolerated mutations inferred using a PSSM matrix are generally better than using positional conservation approaches.[18]

SIFT and PolyPhen are commonly used resources for classifying uncharacterized nonsynonymous SNPs

Several web resources are available for functional annotation of variation, two excellent resources are PolyPhen[55] and SIFT.[50] For a summary, see Table 4. PolyPhen[55] uses a wide variety of features that are sequence-, evolutionary- and structurally based to predict whether a non-synonymous mutation is likely to affect protein function, and performs optimally if structural information is available. PolyPhen has been applied to all mutations in HGVBase, and a server is provided for the annotation of new mutations. Currently, more than 11,000 non-synonymous mutations are annotated. SIFT is available online for predicting intolerant mutations using position-specific information derived from sequence alignments, and requires only sequence and homologue information.[50] SNPeffect at EMBL annotates SNPs with three categories of functional and chemical properties, protein structure and dynamics, functional sites and cellular processing.[56] SNP3D is another resource for inferring the function of SNPs and incorporates structure, alternative splicing, systems biology and evolutionary information for annotation.[57] Additionally, data on the Swiss-Prot website link to homology models of mutations when available.

Generally, the quality of the method will depend on the amount of input data available to the researcher. If only a sequence is available without known homologues or known structure, the method will not perform as well as if these were available. If only sequence and homologue data are available, SIFT will likely give the best performance. If structure data are available, the PolyPhen method will add to an analysis by SIFT because of its use of that data. For any

Table 4: Tools for predicting the function of nsSNPs

Function prediction	URL	Comments
SIFT	http://blocks.fhcrc.org/sift/SIFT.html	Online tool for sequence-based annotation of mutations
PolyPhen	http://www.bork.embl-heidelberg.de/PolyPhen/	Server for functional analysis of mutations
SNP3D	http://www.snps3d.org/	Annotations of structure, systems biology, evolution and alternative splicing
SNPeffect	http://snpeffect.vib.be/index.php	Annotations based on structure, catalysis and cellular process
PicSNP	http://plaza.umin.ac.jp/∼hchang/picsnp/	Gene-centric mutation annotation
TopoSNP	http://gila.bioengr.uic.edu/snp/toposnp/	Protein structural annotations of SNPs
MutDB	http://www.mutdb.org/	Protein structural information of SNPs

The quality of a prediction depends on the amount of input data available

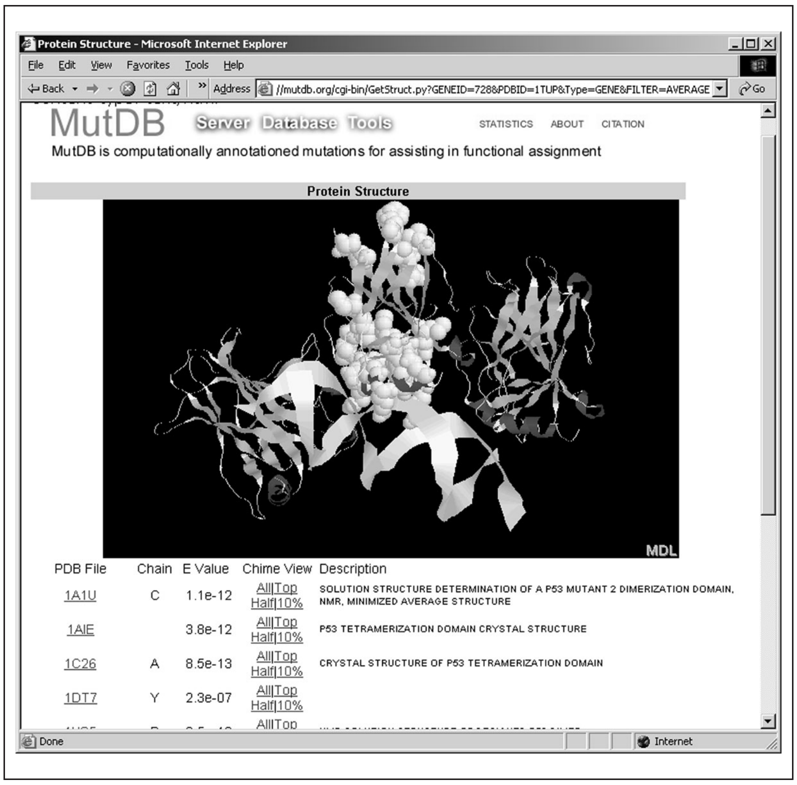
of these resources, better characterised genes will result in better quality predictions.

For protein structural annotations of variation in dbSNP and Swiss-Prot, MutDB[58] was developed to annotate known variation data with information relevant to identifying the molecular effects of a mutation or polymorphism. Mutations from Swiss-Prot and dbSNP are annotated with protein structure information, when available (Figure 2). Similarly, PicSNP[59] is a resource that has annotated more than 1.1 million SNPs and classified them into structural and functional groups. Additionally, they also associate mutations to Gene Ontology categories.

BIOINFORMATICS APPROACHES TO PREDICTING FUNCTIONAL SYNONYMOUS VARIATION

Synonymous variation has been shown to be functional as well. In particular, exonic splicing enhancers (ESEs) are short sequences that occur in exons, and they encourage exon recognition by the cell's splicing machine.[60] When mutated, ESEs can affect mRNA splicing and causing exon skipping.[60] Majewski and Ott[61] found that SNPs occur less often near intron–exon boundaries and the frequency of SNPs in both introns and exons increases as the distance from the boundary increases. Indeed, ESEs have relevance to human disease. Disease-associated variation that disrupts ESEs were found in the breast cancer-associated genes *BRCA1*[62,63] and *BRCA2*.[64] Furthermore, it has been shown that mutations that affect mRNA splicing are the most common type of mutations in neurofibromatosis type 1.[65]

An original approach to the analysis of variation that disrupts ESEs was recently reported by Fairbrother *et al*., where the authors aligned SNPs that are in predicted ESE sites and showed, by comparison to the chimpanzee genome, that these SNPs are under selective pressure. In fact nearly 20 per cent of the polymorphisms have been selected out, and that this is most notable near splicing sites.[66]


Protein Structure - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Address
//mutdb.org/cgi-bin/GetStruct.py?GENEID=728&PDBID=1TUP&Type=GENE&FILTER=AVERAGE
Go
MutDB
Server
Database
Tools
STATISTICS
ABOUT
CITATION
MDL
Done
Internet

MutDB is computationally annotationed mutations for assisting in functional assignment

Protein Structure

PDB File	Chain	E Value	Chime View	Description
1A1U	C	1.1e-12	All|Top Half|10%	SOLUTION STRUCTURE DETERMINATION OF A P53 MUTANT 2 DIMERIZATION DOMAIN, NMR, MINIMIZED AVERAGE STRUCTURE
1AIE		3.8e-12	All|Top Half|10%	P53 TETRAMERIZATION DOMAIN CRYSTAL STRUCTURE
1C26	A	8.5e-13	All|Top Half|10%	CRYSTAL STRUCTURE OF P53 TETRAMERIZATION DOMAIN
1DT7	Y	2.3e-07	All|Top Half|10%	
			All|Top	

Figure 2: Mutations in TP53 mapped to a protein structure in MutDB

BIOINFORMATICS APPROACHES TO PREDICTING FUNCTIONAL NON-CODING VARIATION

Non-coding variation has not received the attention that non-synonymous SNPs and disease-associated mutations have. This is due to difficulties in collecting functional variation information, not to lack of importance. Understanding how variation affects gene expression has been called one of the key challenges in human genetics.[67] The challenge arises from the difficulty in separating regulatory variation (*cis*-acting factors) from the cellular environment and variation on other chromosomes (*trans*-acting factors) and the environment.[68] A recent review succinctly summarises the efforts to understand this challenging problem.[69] A key problem for bioinformatics is developing methods to predict variation that is likely to affect expression levels.

A challenge for the future is prediction of *cis*–acting variation that affects gene expression

Mapping variation to known regulatory sites will not complete the entire story, however, because it is not sufficient to know whether a variant is present; it will be necessary to know whether a variant affects, or disrupts, a function, and therefore expression.

Although this area is difficult to study, there are a few studies that have attempted to examine the relationship between gene expression and variation. Most of the projects that aim to identify the prevalence of variation that alters gene expression at the genomic level have done so by coupling computational methods with experimental analysis of gene expression levels using microarrays.

Cowles *et al.* addressed the problem of removing *trans*-acting factors by focusing their studies on the expression levels in an F_1 hybrid mouse derived from two inbred mouse strains. This allowed them to remove *trans* regulation from the results.[68] They studied 69 genes in total and found that 6 per cent (with large error) of those genes had variants that affect detectable gene expression levels. Pastinen *et al.* examined 129 genes to identify 23 genes that had allele-dependent expression levels.[70] Additionally, Wittkopp *et al.* compared differences in gene expression between closely related *Drosophila* species and found that most of the genes with significant expression level differences had *cis*-regulatory differences.[71] They also found that *cis*-regulatory differences were more common than *trans*-regulatory differences.

Hoogendoorn *et al.* have screened different promoter variants to identify haplotypes that are likely to affect gene expression.[72,73] Their experiments found that a third of the variants can alter expression levels by more than 50 per cent. Later, Buckland *et al.* tested the ability of 20 variant promoters on chromosome 21 alter gene expression and found that approximately 18 per cent of the variants altered expression levels by 1.5-fold or more.[74]

Very little bioinformatics research has been performed to build predictors of variation that is likely to affect gene expression levels. Currently, identifying whether the position is conserved in model organisms and whether the polymorphism sits in a known regulatory motif remain the only computational way of roughly estimating whether a variant will affect expression levels. For example, Consite is a method that predicts transcription factor binding sites.[75,76] Surely when well-annotated databases begin to take form, regulatory relevant polymorphism classification will become possible. PupaSNP Finder[77,78] is a tool for identifying SNPs that could have an effect on transcription. Using Ensembl, the authors map SNPs in dbSNP to transcription factor binding sites, intron/exon border consensus sequences, ESE sequences and variations that are non-synonymous. Another resource is rSNP_Guide,[79,80] which contains annotations of SNPs based on potential effects to regulation.

PROGRAMMING TOOLS FOR SNP ANALYSIS

For bioinformatics researchers developing applications or web resources, the BioPerl project[81] has created an open source set of tools for the analysis of biological data. Using this powerful toolset, scientists can quickly analyse local and remote data from dbSNP, Ensembl and other resources. BioPerl is well supported and includes many tools for annotation, visualisation and analysis of genetic variation. Several tutorials are available for learning the intricate details of Perl and BioPerl.[82] Another tool is libsequence,[83] a C++ library containing tools for SNP analysis. Using similar tools, two groups have described the development of a SNP annotation and selection pipeline.[84,85]

THE FUTURE

The importance and quantity of SNP data now available provide many avenues for future research. First, prediction and classification of non-synonymous mutations using protein structure-based tools need to be improved. Reduction of false positive rates is required. Structure-based tools utilising *ab initio* and comparatively modelled structures will probably be required. For synonymous mutations and non-coding mutations, the spectrum of potential functional disruptions needs to be described. Additionally, for all prediction methods, better training sets are required. These sets need to be rich in both affected and neutral alleles.[17,18]

There is currently no centralised resource for phenotypically and functionally annotated variation

Finally, controlled vocabularies for describing the range of mutation affects could be valuable for building classification methods. Currently, molecular and physiological phenotypic information is often poorly annotated in either human-readable annotations and the subtleties of specific phenotypes are often unclear. There is no central database for phenotypically annotated mutations, and researchers must search many scattered resources to be thorough. There are also few, if any, resources that give information on complex or multifactorial disease where a condition may be caused by more than a single SNP.

The future for this area of research is bright. It is clear from the initial research efforts that bioinformatics methods that predict molecular effects of mutation will continue to improve. A word of caution must be added, however, that bioinformatic scientists building these methods will have the most success if they choose their learning tools carefully and their training sets to best represent the spectrum of predictions they will be making.

Acknowledgments

Sean Mooney is funded by INGEN, the Showalter Trust and a Pervasive Technology Laboratory Fellowship. The Indiana Genomics Initiative (INGEN) of Indiana University is supported in part by Lilly Endowment Inc. The author would like to thank Giselle Knudsen, Predrag Radivojac, Stuart Young and Tatiana Foroud for helpful comments and input.

References

⋆ Papers of particular interest published within the period of this review.

⋆⋆ Papers of extreme interest published within the period of this review.

1. Taillon-Miller, P., Gu, Z., Li, Q. *et al.* (1998), 'Overlapping genomic sequences: A treasure trove of single-nucleotide polymorphisms', *Genome Res.*, Vol. 8(7), pp. 748–754.

2. ⋆⋆Sunyaev, S., Lathe, W. 3rd and Bork, P.(2001), 'Integration of genome data and protein structures: prediction of protein folds, protein interactions and "molecular phenotypes" of single nucleotide polymorphisms', *Curr. Opin. Struct. Biol.*, Vol. 11(1), pp. 125–130.

3. ⋆⋆Steward, R. E., MacArthur, M. W., Laskowski, R. A. and Thornton, J. M. (2003), 'Molecular basis of inherited diseases: a structural perspective', *Trends in Genetics.*, Vol. 19(9), pp. 505–513.

An early review focusing on efforts to understand the protein structural effects of SNPs.

4. Marnellos, G. (2003), 'High-throughput SNP analysis for genetic association studies', *Curr. Opin. Drug Discov. Devel.*, Vol. 6(3), pp. 317–321.

5. ⋆⋆Rebbeck, T. R., Spitz, M. and Wu, X. (2004), 'Assessing the function of genetic variants in candidate gene association studies', *Nat. Rev. Genet.*, Vol. 5(8), pp. 589–597.

Recent review highlighting SIFT and PolyPhen, as well as other functional analysis of SNPs.

6. Hagmann, M. (1999), 'A good SNP may be hard to find', *Science*, Vol. 285(5424), pp. 21–22.

7. URL: http://www.ncbi.nih.nlm.gov/snp/

8. ⋆Hamosh, A.,Scott, A. F., Amberger, J. *et al.* (2000), 'Online Mendelian Inheritance in Man (OMIM)', *Human Mutat.*, Vol. 15(1), pp. 57–61.

9. Boeckmann, B., Bairoch, A., Apweiler, R. *et al.* (2003), 'The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003', *Nucleic Acids Res.*, Vol. 31(1), pp. 365–370.

10. ⋆Stenson, P. D.Ball, E. V., Mort, M. *et al.* (2003), 'Human Gene Mutation Database (HGMD): 2003 update', *Human Mutat.*, Vol. 21(6), pp. 577–581.

11. Fredman, D., Munns, D., Rios, F. *et al.* (2004), 'HGVbase: A curated resource describing human DNA variation and phenotype relationships', *Nucleic Acids Res.*, Vol. 32 (Database issue), pp. D516–519.

12. ⋆⋆Cargill, M., Altshuler, D., Ireland, J. *et al.* (1999), 'Characterization of single-nucleotide polymorphisms in coding regions of human genes', *Nat. Genet.*, Vol. 22(3), pp. 231–238.

Early analysis of many SNPs in important genes. Also the subject of many bioinformatics analysis publications.

13. ⋆⋆Halushka, M. K., Fan, J. B., Bentley, K.

et al. (1999), 'Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis', *Nat. Genet.*, Vol. 22(3), pp. 239–247.

Another early survey on the prevalence of variation in genes that are likely candidates for a specific phenotype. Like Cargill et al., the data are commonly used in other bioinformatics studies.

14. *Lau, A. Y. and Chasman, D. I. (2004), 'Functional classification of proteins and protein variants', *Proc. Natl Acad. Sci. USA*, Vol. 101(17), pp. 6576–6581.

15. Mooney, S. D. and Klein, T. E. (2002), 'Structural models of osteogenesis imperfecta-associated variants in the *COL1A1* gene', *Mol. Cell Proteomics*, Vol. 1(11), pp. 868–875.

16. **Chasman, D. and Adams, R. M. (2001), 'Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation', *J. Mol. Biol.*, Vol. 307(2), pp. 683–706.

Describes an effort to perform a statistical analysis using features trained from saturation mutagenesis experiments to estimate the number of function polymorphisms in a genome.

17. **Ng, P. C. and Henikoff, S. (2002), 'Accounting for human polymorphisms predicted to affect protein function', *Genome Res.*, Vol. 12(3), pp. 436–446.

The SIFT method is applied to SNPs. These results, along with an analysis of other estimates, suggests that earlier estimates of 20–30 per cent of nsSNPs affect protein function are probably overstated.

18. **Saunders, C. T. and Baker, D. (2002), 'Evaluation of structural and evolutionary contributions to deleterious mutation prediction', *J. Mol. Biol.*, Vol. 322(4), pp. 891–901.

Comparison of several features, including SIFT and some structural features, for applying to machine learning methods to predict intolerant mutations. A small human allele data set is collected that is unbiased and performance of features is tested.

19. **Sunyaev, S., Ramensky, V., Koch, I. *et al.* (2001), 'Prediction of deleterious human alleles', *Human Mol. Genet.*, Vol. 10(6), pp. 591–597.

Using a set of structural and evolutionary rules, the number of functional nsSNPs is estimated.

20. **Sunyaev, S., Ramensky, V. and Bork, P. (2000), 'Towards a structural basis of human non-synonymous single nucleotide polymorphisms', *Trends Genet.*, Vol. 16(5), pp. 198–200.

Analysis of SNPs using protein structural features.

21. **Krishnan, V. G. and Westhead, D. R. (2003), 'A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function', *Bioinformatics*, Vol. 19(17), pp. 2199–2209.

Application of different machine learning methods on the nsSNP classification problem.

22. Terp, B. N., Cooper, D. N., Christensen, I. T. *et al.* (2002), 'Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease', *Human Mutat.*, Vol. 20(2), pp. 98–109.

23. *Ferrer-Costa, C., Orozco, M. and de la Cruz, X. (2002), 'Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties', *J. Mol. Biol.*, Vol. 315(4), pp. 771–786.

24. *Stitziel, N. O., Tseng, J., Pervouchine, D. *et al.* (2003), 'Structural location of disease-associated single-nucleotide polymorphisms', *J. Mol. Biol.*, Vol. 327(5), pp. 1021–1030.

25. Mooney, S. D. and Klein, T. E. (2002), 'The functional importance of disease-associated mutation', *BMC Bioinformatics*, Vol. 3(1), p. 24.

26. Mooney, S. D., Klein, T. E., Altman, R. D. *et al.* (2003), 'A functional analysis of disease-associated mutations in the androgen receptor gene', *Nucleic Acids Res.*, Vol. 31(8), p. e42.

27. *Vitkup, D., Sander, C. and Church, G. M. (2003), 'The amino-acid mutational spectrum of human genetic disease', *Genome Biol.*, Vol. 4(11), p. R72.

28. Benzer, S. (1961), 'On the topography of the genetic fine structure', *Proc. Natl Acad. Sci. USA*, Vol. 47, pp. 403–426.

29. Akey, J. M., Zhang, G, Zhang, K. *et al.* (2002), 'Interrogating a high-density SNP map for signatures of natural selection', *Genome Res.*, Vol. 12(12), pp. 1805–1814.

30. Fay, J. C., Wyckoff, G. J. and Wu, C. I. (2001), 'Positive and negative selection on the human genome', *Genetics*, Vol. 158(3), pp. 1227–1234.

31. Wheeler, D. L., Church, D. M., Federhen, S. *et al.* (2001), 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Res.*, Vol. 29(1), pp. 11–16.

32. **Hammond, M. P. and Birney, E. (2004), 'Genome information resources – developments at Ensembl', *Trends Genet.*, Vol. 20(6), pp. 268–272.

Overview of the Ensembl genome database and browser, a fully functional, well-annotated genome information resource.

33. URL: http://www.genomic.unimelb.edu.au/mdi/dblist/dblist.html

34. **Kent, W. J., Sugnet, C. W., Furey, T. S. *et al.* (2002), 'The human genome browser at

UCSC', *Genome Res.*, Vol. 12(6), pp. 996–1006.

Overview of the UC Santa Cruz genome browser and its associated data.

35. URL: http://www.mysql.com/
36. URL: http://www.ensembl.org/
37. URL: http://snpper.chip.org/
38. ★★Riva, A. and Kohane, I. S. (2004), 'A SNP-centric database for the investigation of the human genome', *BMC Bioinformatics*, Vol. 5(1), p. 33.

A SNP-specific selection and analysis toolkit.

39. Clifford, R. J., Edmonson, M. N., Nguyen, C. U. *et al.* (2004), 'Bioinformatics tools for single nucleotide polymorphism discovery and analysis', *Ann. New York Acad. Sci.*, Vol. 1020, pp. 101–109.
40. ★★Staats, B., Qi, L., Beerman, M. *et al.* (2005), 'Genewindow: an international tool for visualization of genomic variation', *Nature Genetics*, Vol. 37(2), pp. 109–110.
41. URL: http://genome.ucsc.edu/
42. Chakravarti, A. (1998), 'It's raining SNPs, hallelujah?', *Nat. Genet.*, Vol. 19(3), pp. 216–217.
43. ★★Syvanen, A. C., Landegren, U., Isaksson, A. *et al.* (1999), 'First International SNP Meeting at Skokloster, Sweden, August 1998. Enthusiasm mixed with scepticism about single-nucleotide polymorphism markers for dissecting complex disorders', *Eur. J. Human Genet.*, Vol. 7(1), pp. 98–101.

SIFT is applied to variation in the BRCA1, illustrating its powerful utility.

44. Collins, F. S., Guyer, M. S. and Charkravarti, A. (1997), 'Variations on a theme: Cataloging human DNA sequence variation', *Science*, Vol. 278(5343), pp. 1580–1581.
45. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. and Sauer, R. T. (1990), 'Deciphering the message in protein sequences: Tolerance to amino acid substitutions', *Science*, Vol. 247(4948), pp. 1306–1310.
46. Zhao, X. Y., Malloy, P. J., Krishan, A. V. *et al.* (2000), 'Glucocorticoids can promote androgen-independent growth of prostate cancer cells through a mutated androgen receptor', *Nat. Med.*, Vol. 6(6), pp. 703–706.
47. Quilliam, L. A., Zhong, S., Rabun, K. M. *et al.* (1995), 'Biological and structural characterization of a Ras transforming mutation at the phenylalanine-156 residue, which is conserved in all members of the Ras superfamily', *Proc. Natl Acad. Sci. USA*, Vol. 92(5), pp. 1272–1276.
48. ★Cai, Z., Tsung, E. F., Marinescu, V. D. *et al.* (2004), 'Bayesian approach to discovering pathogenic SNPs in conserved protein domains', *Human Mutat.*, Vol. 24(2), pp. 178–184.
49. ★Ng, P. C. and Henikoff, S. (2001), 'Predicting deleterious amino acid substitutions', *Genome Res.*, Vol. 11(5), pp. 863–874.
50. ★★Ng, P. C. and Henikoff, S. (2003), 'SIFT: Predicting amino acid changes that affect protein function', *Nucleic Acids Res.*, Vol. 31(13), pp. 3812–3814.

The SIFT method, an evolutionary approach to classification of intolerant mutations, is described.

51. ★Zhu, Y., Spitz, M. R., Amos, C. I. *et al.* (2004), 'An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology', *Cancer Res.*, Vol. 64(6), pp. 2251–2257.
52. Fleming, M. A., Potter, J. D., Ramirez, C. J. *et al.* (2003), 'Understanding missense mutations in the *BRCA1* gene: An evolutionary approach', *Proc. Natl Acad. Sci. USA*, Vol. 100(3), pp. 1151–1156.
53. ★★Wang, Z. and Moult, J. (2001), 'SNPs, protein structure, and disease', *Human Mutat.*, Vol. 17(4), p. 263–270.

An early analysis of protein structural features for disease-associated mutations and nsSNPs.

54. Karchin, R., Kelly, L. and Sali, A. (2005), 'Improving functional annotation of non-synonymous SNPs with information theory', in Klein, T. E. *et al.*, Eds, 'Proceedings of the 10th Pacific Symposium in Biocomputing 2005', 4th–8th January, Hawaii (in press).

Selection and analysis of a subset of structural and evolutionary features for functional SNP classification.

55. ★★Ramensky, V., Bork, P. and Sunyaev, S. (2002), 'Human non-synonymous SNPs: Server and survey', *Nucleic Acids Res.*, Vol. 30(17), pp. 3894–3900.

Introduction and application of PolyPhen, a tool for predicting function nsSNPs.

56. URL: http://snpeffect.vib.be/
57. URL: http://www.snps3d.org/
58. ★★Mooney, S. D. and Altman, R. B. (2003), 'MutDB: Annotating human variation with functionally relevant data', *Bioinformatics*, Vol. 19(14), pp. 1858–1860.

A resource for identifying mutations and nsSNPs with known structure.

59. Chang, H. and Fujita, T. (2001), 'PicSNP: A browsable catalog of nonsynonymous single nucleotide polymorphisms in the human genome', *Biochem. Biophys. Res. Commun.*, Vol. 287(1), pp. 288–291.
60. ★★Blencowe, B. J. (2000), 'Exonic splicing enhancers: Mechanism of action, diversity and role in human genetic diseases', *Trends Biochem. Sci.*, Vol. 25(3), pp. 106–110.

Review of ESEs and their role in human disease.

61. Majewski, J. and Ott, J. (2002), 'Distribution and characterization of regulatory elements in the human genome', *Genome Res.*, Vol. 12(12), pp. 1827–1836.

62. Orban, T. I. and Olah, E. (2001), 'Purifying selection on silent sites – a constraint from splicing regulation?', *Trends Genet.*, Vol. 17(5), pp. 252–253.

63. **Liu, H. X., Cartegni, L., Zhang, M. Q. and Krainer, A. R. (2001), 'A mechanism for exon skipping caused by nonsense or missense mutations in *BRCA1* and other genes', *Nat. Genet.*, Vol. 27(1), pp. 55–58.

An analysis of ESE affecting SNPs in the BRCA1 gene.

64. Fackenthal, J. D., Cartegni, L., Krainer, A. R. and Olopade, O. I. (2002), 'BRCA2 T2722R is a deleterious allele that causes exon skipping', *Amer. J. Human Genet.*, Vol. 71(3), pp. 625–631.

65. Ars, E., Serra, E., García, J. *et al.* (2000), 'Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1', *Human Mol. Genet.*, Vol. 9(2), pp. 237–247.

66. Fairbrother, W. G., Holste, D., Burge, C. B. and Sharp, P. A. (2004), 'Single nucleotide polymorphism-based validation of exonic splicing enhancers', *PLoS Biol.*, Vol. 2(9), p. E268.

67. **Hudson, T. J. (2003), 'Wanted: Regulatory SNPs', *Nat. Genet.*, Vol. 33(4), pp. 439–440.

An article describing the need for SNPs that affect gene expression.

68. **Cowles, C. R., Joel, N. H., Altshuler, D. and Lander, E. S. (2002), 'Detection of regulatory variation in mouse genes', *Nat. Genet.*, Vol. 32(3), pp. 432–437.

An approach to discovering haplotypes that affect regulation of gene expression.

69. **Knight, J. C. (2004), 'Allele-specific gene expression uncovered', *Trends Genet.*, Vol. 20(3), pp. 113–116.

A review describing the challenge of identifying regulatory SNPs.

70. Pastinen, T., Sladek, R., Gurd, S. *et al.* (2004), 'A survey of genetic and epigenetic variation affecting human gene expression', *Physiol. Genomics*, Vol. 16(2), p. 184–193.

71. **Wittkopp, P. J., Haerum, B. K. and Clark, A. G. (2004), 'Evolutionary changes in *cis* and *trans* gene regulation', *Nature*, Vol. 430(6995), pp. 85–88.

An analysis of cis-regulatory variation in two closely related Drosophila species.

72. Hoogendoorn, B., Coleman, S. L., Guy, C. A. *et al.* (2003), 'Functional analysis of human promoter polymorphisms', *Human Mol. Genet.*, Vol. 12(18), pp. 2249–2254.

73. **Hoogendoorn, B., Coleman, S. L., Guy, C. A. *et al.* (2004), 'Functional analysis of polymorphisms in the promoter regions of genes on 22q11', *Human Mutat.*, Vol. 24(1), pp. 35–42.

An in vitro approach to testing SNPs that affect expression levels.

74. Buckland, P. R., Coleman, S. L., Hoogendoorn, B. *et al.* (2004), 'A high proportion of chromosome 21 promoter polymorphisms influence transcriptional activity', *Gene Expr.*, Vol. 11(5–6), pp. 233–239.

75. Sandelin, A., Wasserman, W. W. and Lenhard, B. (2004), 'ConSite: Web-based prediction of regulatory elements using cross-species comparison', *Nucleic Acids Res.*, Vol. 32 (web server issue), pp. W249–252.

76. URL: http://www.phylofoot.org/consite/

77. **Conde, L., Vaquerizas, J. M., Santoyo, J. *et al.* (2004), 'PupaSNP Finder: A web tool for finding SNPs with putative effect at transcriptional level', *Nucleic Acids Res.*, Vol. 32 (web server issue), pp. W242–248.

A resource for finding SNPs in known transcription factor binding sites.

78. URL: http://pupasnp.bioinfo.cnio.es/

79. **Ponomarenko, J. V., Merkulova, T. I., Orlova, G. V. *et al.* (2003), 'rSNP_Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation', *Nucleic Acids Res.*, Vol. 31(1), pp. 118–121.

A resource for characterizing regulatory SNPs.

80. URL: http://wwwmgs.bionet.nsc.ru/mgs/systems/rsnp/

81. **Stajich, J. E., Block, D., Boulez, K. *et al.* (2002), 'The Bioperl toolkit: Perl modules for the life sciences', *Genome Res.*, Vol. 12(10), pp. 1611–1618.

The BioPerl project is the leading open source API for biological data analysis.

82. URL: http://bio.perl.org/

83. Thornton, K. (2003), 'Libsequence: A C++ class library for evolutionary genetic analysis', *Bioinformatics*, Vol. 19(17), pp. 2325–2327.

84. Aerts, J., Welzels, Y., Cohen, N. and Aerssens, J. (2002), 'Data mining of public SNP databases for the selection of intragenic SNPs', *Human Mutat.*, Vol. 20(3), pp. 162–173.

85. Wjst, M. (2004), 'Target SNP selection in complex disease association studies', *BMC Bioinformatics*, Vol. 5(92).