MINI REVIEW Genomics & Informatics Vol. 6(4) 166-172, December 2008

In Silico Functional Assessment of Sequence Variations: Predicting Phenotypic Functions of Novel Variations

Hong-Hee Won[1,2] and Jong-Won Kim[3]*

[1]Samsung Biomedical Research Institute, Samsung Medical Center, Seoul 135-710, Korea, [2]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea, [3]Department of Laboratory Medicine and Genetics, Sungkyunkwan University School of Medicine, Samsung Medical Center, Seoul 135-710, Korea

Abstract

A multitude of protein-coding sequence variations (CVs) in the human genome have been revealed as a result of major initiatives, including the Human Variome Project, the 1000 Genomes Project, and the International Cancer Genome Consortium. This naturally has led to debate over how to accurately assess the functional consequences of CVs, because predicting the functional effects of CVs and their relevance to disease phenotypes is becoming increasingly important. This article surveys and compares variation databases and *in silico* prediction programs that assess the effects of CVs on protein function. We also introduce a combinatorial approach that uses machine learning algorithms to improve prediction performance.

Keywords: sequence variation, amino acid substitution, nonsynonymous single nucleotide polymorphism, missense mutation, prediction, protein function

Introduction

Single amino acid substitutions in protein-coding sequences are common in the human genome. These protein-coding sequence variations (CVs) are important diagnostic markers and therapeutic targets in genetic disease studies. Although most CVs are functionally neutral, some CVs affect phenotype, including nonsynonymous single nucleotide polymorphisms (nsSNPs) that contribute to normal phenotypic differences in hair color, skin color (Sulem *et al*., 2007; Han *et al*., 2008), and disease susceptibility (WTCCC, 2007; Amos *et al*., 2008; Harley *et al*., 2008; Tenesa *et al*., 2008). Other CVs result in deleterious missense mutations that cause highly penetrant Mendelian-inherited diseases (Kim *et al*., 2007). These deleterious mutations have been of great interest in biomedical research and clinical practice for decades and account for approximately half of the genetic variations that are known to cause disease.

Using recent advancements in sequencing technologies, several studies have reported a number of sequence variations in certain cancers (Sjoblom *et al*., 2006; Greenman *et al*., 2007; Campbell *et al*., 2008; Jones *et al*., 2008), in which mutational patterns have differed greatly between patients with the same disease. Furthermore, major initiatives, such as the Human Variome Project, the 1000 Genomes Project, and the International Cancer Genome Consortium, will generate a vast amount of variation data. Consequently, it is important to assess variations in conjunction with protein function and disease phenotype. Several databases, such as the Online Mendelian Inheritance in Man (OMIM) and the Human Gene Mutation Database (HGMD), have documented CVs that correspond to Mendelian-inherited human diseases. In addition, many computational programs have been created to predict the functional effects of unknown CVs (Ng *et al*., 2006; Care *et al*., 2007). Database searches and bioinformatic predictions can be useful in prioritizing novel CVs for further analysis.

In this review, we summarize the databases that are most helpful in interpreting the functional effects of CVs. We perform an extensive survey of existing *in silico* prediction methods and compare their performance. Finally, we introduce a combination method as a promising approach to improve prediction performance.

Polymorphism and Mutation Databases

Several databases that are helpful in assessing the functional effects of CVs or their relevance to disease phenotype are listed in Table 1. Each of two broad-category mutation databases, general mutation databases (GMDBs) and locus-specific mutation databases (LSDBs), has unique strengths and weaknesses (Porter *et al*., 2000). Because polymorphism and mutation databases have been developed for different uses, they complement each other.

*Corresponding author: E-mail kimjw@skku.edu
Tel +82-2-3410-2705, Fax +82-2-3410-2719
Accepted 25 November 2008

Table 1. Polymorphism and mutation databases

Database	Recent release date*	Data type	Features	Website
OMIM (Hamosh *et al*., 2005)	Updated daily	Deleterious mutations	Full-text descriptions of published disease-causing variations	http://www.ncbi.nlm.nih.gov/omim
HGMD (Stenson *et al*., 2008)	Sept 2008	Deleterious mutations	Comprehensive collection of published disease-causing variations	http://www.hgmd.cf.ac.uk
LSDB in HGVS	Nov 2008	Deleterious mutations	Specialized collection of a particular gene or locus	http://www.hgvs.org/dblist/glsdb.html
Swiss-Prot (Yip *et al*., 2004)	Nov 2008	Deleterious mutations and neutral polymorphisms	Well-summarized list of variations and corresponding proteins	http://www.expasy.org/cgi-bin/lists?humsavar.txt
dbSNP (Sherry *et al*., 2001)	Apr 2008	Neutral and (few) deleterious SNPs	Broad collections of SNPs regardless of clinical associations (clinically associated SNPs linked to source sites)	http://www.ncbi.nlm.nih.gov/projects/SNP
dbGaP (Mailman *et al*., 2007)	Nov 2008	Deleterious or phenotype-affecting SNPs	Collections of SNPs affecting clinical phenotypes or nonclinical traits	http://www.ncbi.nlm.nih.gov/gap
HapMap (Frazer *et al*., 2007)	Oct 2008	Neutral SNPs and (very few) deleterious SNPs	Collections of SNPs of 270 individuals randomly selected from African, Asian, and European populations	http://www.hapmap.org

*Accessed Nov 2008. Because LSDBs are individually updated, the presented release date is for HGVS.

OMIM

OMIM is among the most representative and well-documented GMDBs, and it contains a full-text overview of human genes and genetic disorders. The contents of OMIM are considered comprehensive, authoritative, and timely, because they are written at Johns Hopkins University School of Medicine and edited daily with input from scientists and physicians from around the world (Hamosh *et al*., 2005). Many CVs that are registered in OMIM may have deleterious effects on protein function and cause Mendelian-inherited diseases. OMIM also includes some disease susceptibility variations that are found in association analyses. OMIM is therefore a valuable resource to study the characteristics of variations that severely affect a certain phenotype.

HGMD

HGMD is a comprehensive and publicly available GMDB of gene lesions that underlie human inherited diseases (Krawczak *et al*., 2000; Stenson *et al*., 2003; Stenson *et al*., 2008). Two different versions of the database exist: an up-to-date commercial version and an older public version-both of them provide comprehensive mutation information. The total number of public entries that are available to users from academic institutions or nonprofit organizations is 61,447, comprising 35,168 missense or nonsense mutations, 10,035 small deletions, and 5805 splicing mutations. Because data are collected by a combination of manual and computerized searches, the contents are considered highly reliable. GMDBs, such as OMIM and HGMD, have several limitations, however, most of which are attributable to less-specialized knowledge of particular genetic loci (Porter *et al*., 2000).

LSDB and Swiss-Prot

LSDBs are usually maintained by experts on a particular gene or locus, provide a greater depth of information about each variation, and often present unpublished data that are submitted directly by researchers in related fields. In contrast, LSDBs can often become stagnant or disappear because they are likely to depend on limited funding resources and part-time scientists (Porter *et al*., 2000). Hundreds of LSDBs for 718 genes are currently listed on the website of the Human Genome Variation Society (HGVS). The Swiss-Prot database (release version 56.4), with 46,799 CVs for 9085 proteins, complements LSDBs. These CVs are particularly useful for developing prediction algorithms of functionality (Care *et al*., 2007), because Swiss-Prot provides a well-summarized list of CVs with corresponding protein identifications, sequence positions, amino acid changes, and disease associations (disease *vs.* polymorphism).

dbSNP

NCBI's single nucleotide polymorphisms database (dbSNP, build 129) contains 14.7 million human reference SNPs. Their broad collection of simple genetic polymorphisms includes SNPs, small-scale multibase deletions or insertions, retrotransposable element insertions, and microsatellite repeat variations (Sherry *et al.*, 2001). The database provides the frequency of the polymorphism by population or individual, allowing for estimates of prevalence in a specific population. The database archives variations regardless of their clinical associations and contains some clinically associated SNPs that are linked to OMIM, LSDB, or the clinical laboratory. It should be noted that SNPs that lack clinical associations and have not been functionally validated may still be relevant.

dbGaP

The database of genotype and phenotype (dbGaP) archives the results of studies that investigate the interaction between genotype and phenotype (Mailman *et al.*, 2007). The database includes SNPs that affect both clinical and nonclinical phenotypes that are found in genome-wide association studies, medical sequencing, and molecular diagnostic assays. The results are categorized by study and by disease. More than 30 studies are listed, each comprising thousands of case-control sets or parent-offspring trios. Authorized users can download individual-level data for their own research.

HapMap

The haplotype map (HapMap) (release #24) contains the genotypes and frequencies of over 3.8 million SNPs. The SNPs were obtained by analyzing DNA samples from 270 individuals, comprising 30 trios of two parents and an adult child of African ancestry, 30 trios of European ancestry, and 90 unrelated individuals of Asian ancestry (Frazer *et al.*, 2007). Because the individuals were randomly selected, one would expect that very few variations in HapMap are functionally deleterious; therefore, HapMap data could be used as a reference set of neutral CVs.

In conclusion, the OMIM, HGMD, and LSDB databases catalog known deleterious mutations that result in severe disease, while the Swiss-Prot and dbGaP databases record those that have modest effects on the resulting protein. The Swiss-Prot, dbSNP, and HapMap databases provide fundamental information on neutral polymorphisms.

Prediction Programs for Functional Assessment of Sequence Variations

Because it was shown that protein structure and sequence-based attributes could provide information to distinguish deleterious mutations from neutral single-base changes (Sunyaev *et al.*, 2000; Chasman *et al.*, 2001), many prediction programs have been developed and implemented on a web server to provide *in silico* prediction of CV functionality (Ng *et al.*, 2006; Care *et al.*, 2007). These programs employ a variety of rule-based models and machine learning algorithms, using information on protein structure, sequence, physicochemical properties, phylogenetics, and evolutionary features. The widely used programs are listed below and summarized in Table 2.

SIFT and PolyPhen

The program Sorting Intolerant From Tolerant (SIFT) uses sequence homology to calculate a scaled probability for the substitution that is observed (Ng *et al.*, 2001; Ng *et al.*, 2002; Ng *et al.*, 2003). Substitutions that have a low scaled probability are predicted to affect protein function. The Swiss-Prot/TrEMBL databases and PSI-BLAST were used for sequence alignment. As the first program that was implemented on a web server, SIFT is one of the most frequently used, along with Polymorphism Phenotyping (PolyPhen) (Sunyaev *et al.*, 2001; Ramensky *et al.*, 2002). PolyPhen uses empirically derived rules to predict the effect of CVs on protein function. The rules are based on known protein structures, sequence conservation between homologous proteins, and sequence-based characterization of the substitution site (*e.g.*, binds lipid, metal). These two programs are among the earliest developed programs and have been recently updated.

MSRV

A method that was published by Jiang *et al.* adopts Multiple Selection Rule Voting (MSRV), which includes three physicochemical properties (molecular weight, pI value, and hydrophobicity scale) of amino acids, three relative frequencies for the presence of amino acids in secondary structures (helices, strands, and turns), and two evolutionary conservation scores (Jiang *et al.*, 2007). These authors compared the areas under the receiver operating curves (AUCs) of MSRV, SIFT, and PolyPhen and showed that MSRV employs optimal feature sets, outperforming SIFT and PolyPhen in prioritizing disease mutations that are responsible for monogenic and polygenic diseases.

Table 2. Programs for predicting functional effects of coding sequence variations

Method	Recent release date*	Algorithm	Performance†	Source code	Website
SIFT (Ng *et al*., 2003)	March 2008	Calculates a scaled probability for the substitution using sequence homology	FN error: 31% FP error: 20%	Available	http://blocks.fhcrc.org/sift
PolyPhen (Ramensky *et al*., 2002)	March 2008	Empirical rules based on characterization of the substitution site, conservation between homologous proteins, and protein structures	FN error: 31% FP error: 9%	On request	http://coot.embl.de/PolyPhen
MSRV (Jiang *et al*., 2007)	Aug 2007	Multiple Selection Rule Voting (MSRV) using physicochemical properties, relative frequencies in secondary structures, and evolutionary conservation	AUC: 82～87% SIFT AUC: 75% PolyPhen AUC: 70～75%	Not specified	http://msms.usc.edu/msrv
PANTHER (Thomas *et al*., 2003)	Aug 2007	Calculates the functional likelihood using a hidden Markov model with a protein family library	FN error: 59% FP error: N/A	Available	http://www.pantherdb.org/tools/csnpScoreForm.jsp
SNAP (Bromberg *et al*., 2007)	Sept 2008	Combines sequence analysis tools and uses protein annotation, solvent accessibility, secondary structure, flexibility, SIFT results, and conservation	FN error: 20% FP error: 24%	On request	http://rostlab.org/services/SNAP
PMUT (Ferrer-Costa *et al*., 2005)	May 2005	Uses two neural networks trained with human mutation data using structural and evolutionary information	FN error: 12～21% FP error: 10～17%	Not specified	http://mmb2.pcb.ub.es:8080/PMut
nsSNPAnalyzer (Bao *et al*., 2005)	Feb 2005	Random forest trained with structural information, sequence conservation, and SIFT prediction using Swiss-Prot data	FN error: 21% FP error: 38%	Available	http://snpanalyzer.utmem.edu

*Accessed Nov 2008. If the release date was not available on the website, the publication date was presented.
†Performance was summarized based on the literature in which the method was introduced. False negative (FN) error rate is the percentage of the deleterious variations predicted to be neutral. False positive (FP) error rate is the percentage of the neutral variations predicted to be deleterious. Area under the receiver operating characteristic curve (AUC) was calculated in the Jiang *et al*. study (Jiang *et al*., 2007).

PANTHER

The program PANTHER can predict the effect of CVs on protein function by relating sequence to function (Thomas *et al*., 2003; Thomas *et al*., 2004). The program uses a hidden Markov model and a library of protein families to score the functional likelihood of different amino acid substitutions. The phenotypic effect is determined by the position-specific evolutionary conservation (PSEC) scores that are obtained from the model. Because the source codes for the PANTHER predictor and the PANTHER library are available online, this method is useful for analyzing a large number of CVs.

SNAP and PMUT

Screening for Non-Acceptable Polymorphisms (SNAP) predicts non-neutral substitutions by using annotations from the protein mutant database (Kawabata *et al*., 1999) and by combining many sequence analysis tools in neural networks (NNs) (Bromberg *et al*., 2007; Bromberg *et al*., 2008a; Bromberg *et al*., 2008b). It also uses solvent accessibility, secondary structure, flexibility, SIFT results, and conservation information. SNAP gives a reliability index of the prediction, ranging from 0 (low) to 9 (high reliability). PMUT is similar to SNAP with regard to its methods and output. PMUT is also based on the use of two NNs that are trained with human mutation data (Ferrer-Costa *et al*., 2002; Ferrer-Costa *et al*., 2004; Ferrer-Costa *et al*., 2005). It displays a pathogenicity index that ranges from 0 to 1 (>0.5 signals pathological mutations), a confidence index that ranges from 0 (low) to 9 (high confidence), and the mutation site on the protein structure to trace its pathogenicity.

nsSNPAnalyzer

The program nsSNPAnalyzer uses a machine learning algorithm called random forest to combine heterogeneous sources of predictors (Bao *et al.*, 2005; Bao *et al.*, 2005). Random forest was trained with various features, such as structural information, sequence conservation, and SIFT prediction using a dataset from Swiss-Prot. The source codes of nsSNPAnalyzer are also available on the website.

As reviewed in this section, many programs use common sources and methods while exploiting different algorithms, features, and databases. Combined use of these *in silico* programs is encouraged to mitigate their limitations in prediction performance. Source codes are necessary to analyze large-scale data with several programs; code availability is summarized in Table 2.

Combination Approach to Predicting Function

Given that predictions of the functional consequences of amino acid substitutions can be more accurate by combining different *in silico* methods (Ng *et al.*, 2006), a combination approach has been proposed to improve prediction accuracy (Won *et al.*, 2008). To assess the effectiveness of this approach, the prediction performance of individual programs must be evaluated. A support vector machine (SVM) was used to combine three representative *in silico* prediction programs (SIFT, PolyPhen, and PANTHER) to predict the phenotypic effects of CVs.

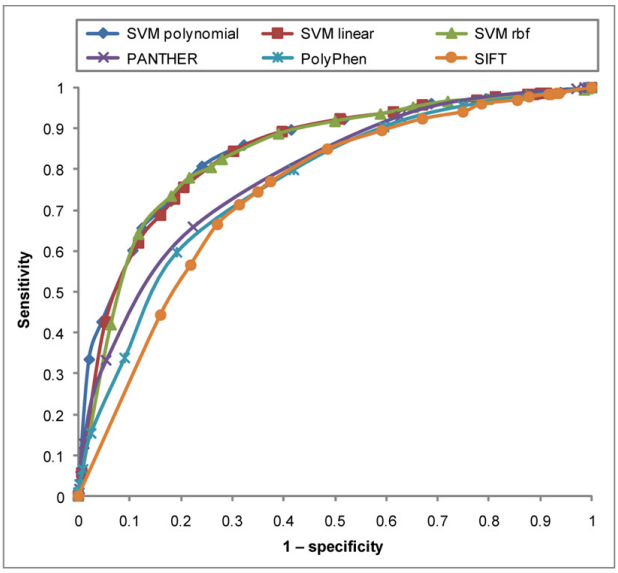

Fig. 1. Receiver operating characteristic (ROC) curves of individual predictors and their combinations. The ROC curves of SVM combinations tend toward the upper left corner of the plot more than the three individual prediction programs, indicating superior performance. This shows that the appropriate combination can noticeably improve prediction accuracy. Figure from Won *et al.*, 2008.

Assuming that the HapMap dataset comprises mainly nonpathogenic variations (negative samples) while the HGMD dataset comprises pathogenic variations (positive samples), we compared the prediction performances of SVM combinations and individual predictors, including SIFT, PolyPhen, and PANTHER (see Won *et al.*, 2008 for details). The three different kernel functions-a linear kernel, a polynomial kernel, and a radial basis function kernel-were used to train SVMs. The experimental results show that the SVM combinations outperform the individual prediction programs (Fig. 1). In particular, $SVM_{polynomial}$ has a slightly superior predictive value than the other two SVM combinations. PANTHER outperforms PolyPhen and SIFT in terms of sensitivity over all specificity regions. The superior performance of $SVM_{polynomial}$ indicates that the appropriate combination can effectively improve prediction accuracy.

Conclusion

Interpreting the functionality of newly found variations in gene coding regions is of much importance to both biomedical research and clinical practice. The first step to understanding these variations is to examine them using valuable resources, such as variation databases and functional prediction programs. Furthermore, automated prediction methods are essential for analyzing CVs on a genome-wide scale. This review summarizes representative examples of useful resources and emphasizes the ongoing need for improvement in the performance of individual prediction methods. We suggest that comprehensive analyses that use a combination of complementary databases and *in silico* programs are necessary to overcome the relative weakness of each program. In the case of SVM combinations, we showed that prediction can be improved effectively if the results of the individual programs are appropriately combined.

Acknowledgments

This study was supported by the Korean HapMap Project, funded by the Korean Ministry of Education, Science and Technology.

References

Amos, C.I., Wu, X., Broderick, P., *et al.* (2008). Genome-

wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.* 40, 616-622.

Bao, L., and Cui, Y. (2005). Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics* 21, 2185-2190.

Bao, L., Zhou, M., and Cui, Y. (2005). nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.* 33(Web Server issue), W480-482.

Bromberg, Y., and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* 35, 3823-3835.

Bromberg, Y., and Rost, B. (2008a). Comprehensive *in silico* mutagenesis highlights functionally important residues in proteins. *Bioinformatics* 24, i207-212.

Bromberg, Y., Yachdav, G., and Rost, B. (2008b). SNAP predicts effect of mutations on protein function. *Bioinformatics* 24, 2397-2398.

Campbell, P.J., Pleasance, E.D., Stephens, P.J., *et al.* (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 105, 13081-13086.

Care, M.A., Needham, C.J., Bulpitt, A.J., and Westhead, D.R. (2007). Deleterious SNP prediction: be mindful of your training data! *Bioinformatics* 23, 664-672.

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.* 307, 683-706.

Ferrer-Costa, C., Gelpi, J.L., Zamakola, L., *et al.* (2005). PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* 21, 3176-3178.

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.* 315, 771-786.

Ferrer-Costa, C., Orozco, M., and de la Cruz, X. (2004). Sequence-based prediction of pathological mutations. *Proteins* 57, 811-819.

Frazer, K.A., Ballinger, D.G., Cox, D.R., *et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851-861.

Greenman, C., Stephens, P., Smith, R., *et al.* (2007). Patterns of somatic mutation in human cancer genomes. *Nature* 446, 153-158.

Hamosh, A., Scott, A.F., Amberger, J.S., *et al.* (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33(Database issue), D514-517.

Han, J., Kraft, P., Nan, H., *et al.* (2008). A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet.* 4, e1000074.

Harley, J.B., Alarcon-Riquelme, M.E., Criswell, L.A., *et al.* (2008). Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nat. Genet.* 40, 204-210.

Jiang, R., Yang, H., Zhou, L., *et al.* (2007). Sequence-based prioritization of nonsynonymous single-nucleotide polymorphisms for the study of disease mutations. *Am. J. Hum. Genet.* 81, 346-360.

Jones, S., Zhang, X., Parsons, D.W., *et al.* (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 321, 1801-1806.

Kawabata, T., Ota, M., and Nishikawa, K. (1999). The protein mutant database. *Nucleic Acids Res.* 27, 355-357.

Kim, H.J., Sohn, K.M., Shy, M.E., *et al.* (2007). Mutations in PRPS1, which encodes the phosphoribosyl pyrophosphate synthetase enzyme critical for nucleotide biosynthesis, cause hereditary peripheral neuropathy with hearing loss and optic neuropathy (cmtx5). *Am. J. Hum. Genet.* 81, 552-558.

Krawczak, M., Ball, E.V., Fenton, I., *et al.* (2000). Human gene mutation database-a biomedical information and research resource. *Hum. Mutat.* 15, 45-51.

Mailman, M.D., Feolo, M., Jin, Y., *et al.* (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39, 1181-1186.

Ng, P.C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.* 11, 863-874.

Ng, P.C., and Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* 12, 436-446.

Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812-3814.

Ng, P.C., and Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* 7, 61-80.

Porter, C.J., Talbot, C.C., and Cuticchia, A.J. (2000). Central mutation databases--a review. *Hum. Mutat.* 15, 36-44.

Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30, 3894-3900.

Sherry, S.T., Ward, M.H., Kholodov, M., *et al.* (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308-311.

Sjoblom, T., Jones, S., Wood, L.D., *et al.* (2006). The consensus coding sequences of human breast and colorectal cancers. *Science* 314, 268-274.

Stenson, P.D., Ball, E., Howells, K., *et al.* (2008). Human Gene Mutation Database: towards a comprehensive central mutation database. *J. Med. Genet.* 45, 124-126.

Stenson, P.D., Ball, E.V., Mort, M., *et al.* (2003). Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* 21, 577-581.

Sulem, P., Gudbjartsson, D.F., Stacey, S.N., *et al.* (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* 39, 1443-1452.

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

Sunyaev, S., Ramensky, V., Koch, I., *et al.* (2001).

Prediction of deleterious human alleles. *Hum. Mol. Genet.* 10, 591-597.

Tenesa, A., Farrington, S.M., Prendergast, J.G., *et al.* (2008). Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat. Genet.* 40, 631-637.

Thomas, P.D., Campbell, M.J., Kejariwal, A., *et al.* (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 13, 2129-2141.

Thomas, P.D., and Kejariwal, A. (2004). Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc. Natl. Acad. Sci. U. S. A.* 101, 15398-15403.

Won, H.H., Kim, H.J., Lee, K.A., and Kim, J.W. (2008). Cataloging coding sequence variations in human genome databases. *PLoS ONE* 3, e3575.

WTCCC. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661-678.

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