Published online 7 September 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D815–D819***
doi:10.1093/nar/gkm659

MutDB: update on development of tools for the biochemical analysis of genetic variation

Arti Singh[1], Adebayo Olowoyeye[1], Peter H. Baenziger[1], Jessica Dantzer[1], Maricel G. Kann[2], Predrag Radivojac[3], Randy Heiland[4] and Sean D. Mooney[1,*]

[1]Center for Computational Biology and Bioinfomatics, Department of Medical and Molecular Genetics, Indiana University School of Medicine; 410 W. 10th Street, Suite 5000, Indianapolis, IN 46202, [2]Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, MD 21250, [3]School of Informatics, Indiana University; 1900 East 10th Street, Bloomington, IN 47406 and [4]Pervasive Technology Labs, Indiana University; Informatics and Communications Technology Complex, 535 West Michigan Street, Indianapolis, IN 46202, USA

Received February 6, 2007; Revised and Accepted August 9, 2007

ABSTRACT

Understanding how genetic variation affects the molecular function of gene products is an emergent area of bioinformatic research. Here, we present updates to MutDB (http://www.mutdb.org), a tool aiming to aid bioinformatic studies by integrating publicly available databases of human genetic variation with molecular features and clinical phenotype data. MutDB, first developed in 2002, integrates annotated SNPs in dbSNP and amino acid substitutions in Swiss-Prot with protein structural information, links to scores that predict functional disruption and other useful annotations. Though these functional annotations are mainly focused on nonsynonymous SNPs, some information on other SNP types included in dbSNP is also provided. Additionally, we have developed a new functionality that facilitates KEGG pathway visualization of genes containing SNPs and a SNP query tool for visualizing and exporting sets of SNPs that share selected features based on certain filters.

INTRODUCTION

Understanding how coding single nucleotide polymorphisms (cSNPs) and disease-associated mutations cause molecular alterations and expression changes in gene products is important to many fields of biological and medical research (1,2). We believe that linking disease with basic research data will enable hypothesis generation that can be experimentally tested in the laboratory with functional assays.

Recently, several servers and databases aiming to understand the biochemical effects of nonsynonymous SNPs and disease-associated mutation have been developed. These include SIFT (3), PolyPhen (4), SNPs3D (5), PANTHER (6), PMUT (7), LS-SNP (8), PolyDoms (9) and SNPEffect (10). These methods and their resulting datasets generally apply DNA and protein sequence, protein structure and/or evolutionary features to classify a query amino acid substitution using a training set of putative neutral and causative amino acid substitutions (4,5,8,11–17).

Similarly, MutDB (18,19) is an online resource that serves as a step toward better understanding the potential molecular effects of a mutation. MutDB integrates genetic variation from two public databases, Swiss-Prot (20) and dbSNP (21), and annotates the variants with biochemically relevant information. These two databases are chosen because they are freely available and represent a significant breadth of available amino acid substitutions. However, neither of these databases annotates disease causing amino acid substitutions particularly well. dbSNP contains few links to OMIM (22), and Swiss-Prot does not identify disease causing amino acid substitutions from other amino acid substitutions. Therefore, a researcher studying a specific disease should have some prior knowledge of the proteins and mutations of interest, and MutDB provides some helpful links to useful databases with disease and phenotype annotations such as OMIM and dbGAP (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap), (22).

In addition to updating to the latest mutation and SNP datasets, here we present several additions to the MutDB resource. First, we have developed a pathway visualization add-on to MutDB that leads the biologist from mutations in a gene to KEGG (23) biological pathways involving the gene. This enables the researcher to view the systems context of both a mutation and its associated phenotype. Second, we have constructed an AJAX (Asynchronous

*To whom correspondence should be addressed. Tel: +1 317 278 9221; Fax: +1 317 278 9217; Email: mooney@compbio.iupui.edu

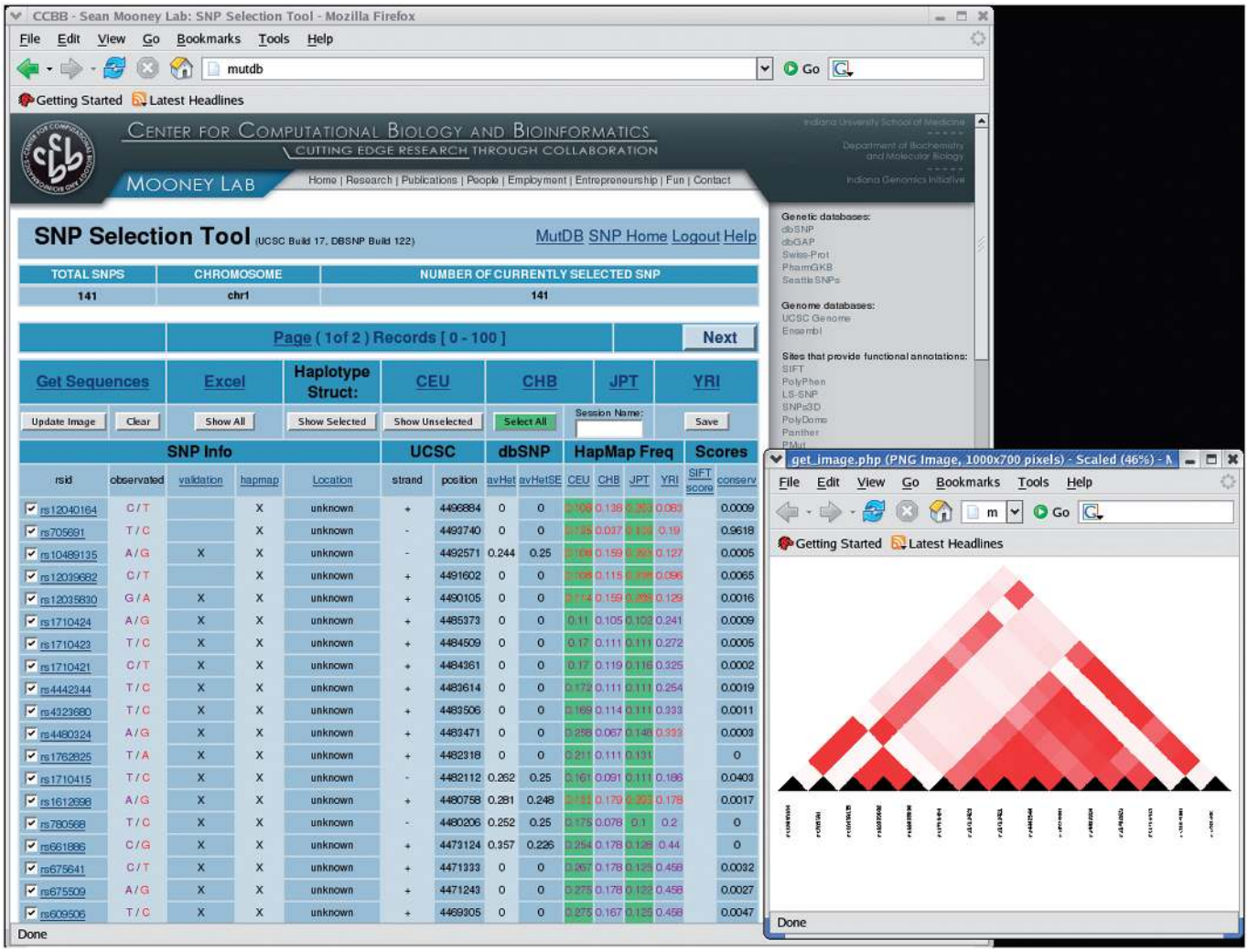

Figure 1. SNP query tool snapshot highlighting SNP filtering. Multiple filtering options include: validation (T/F), HapMap (31) (T/F), Location, avHet, avHetSE, HapMap Frequency (CEU, CHB, JPT, YRI), SIFT score and UCSC conservation. Users can preview current filtering criteria by scrolling over pop-up window link. Once SNPs are selected, Haploview like images can be rendered showing HapMap LD structure (lower right).

JavaScript and XML) based SNP query tool that allows users to save searches, view Haploview-like haplotype structure (24), and select subsets of SNPs based on frequencies and SNP scores. Together these tools represent a useful addition to our existing library of genetic research tools (Figure 1).

METHODS AND USAGE

Web interface and data organization

The SNP and mutation data is parsed directly from Swiss-Prot (currently build 51.0) and dbSNP (currently build 126) without curation, other than to remove any annotations that do not map to the wild-type amino acid in the referenced sequence. The gene model provided by MutDB is organized, using gene information extracted from a local copy of the UCSC Human Genome Annotation Database (ver. Hg18, http://genome.ucsc.edu/) (25). We also use Ensembl (ver. 41_36c, http://www.ensembl.org/) (26) for some gene name cross-references. We attempt to keep pages organized by Entrez Gene ID with the most representative transcript as the primary gene page. Other known mRNA transcripts annotated in the UCSC Genome Annotation Database are listed at the bottom of the page with their annotations. This data may be browsed alphabetically by gene symbol or by employing one of several search methods, including keyword, gene symbol, protein or Refseq ID, and individual variant identifier. Each gene is given its own page for display, providing a list of related SNPs and mutations classified by their effects on the protein product, as well as a pictorial representation of the sequence including points of conservation, location of exons and location of variants. Links to corresponding Swiss-Prot and dbSNP pages, a short description of the gene, and the related chromosome name are supplied.

Each variant is annotated within its own page providing further details, which includes the protein sequence, if known, and any related Protein Data Bank (PDB)

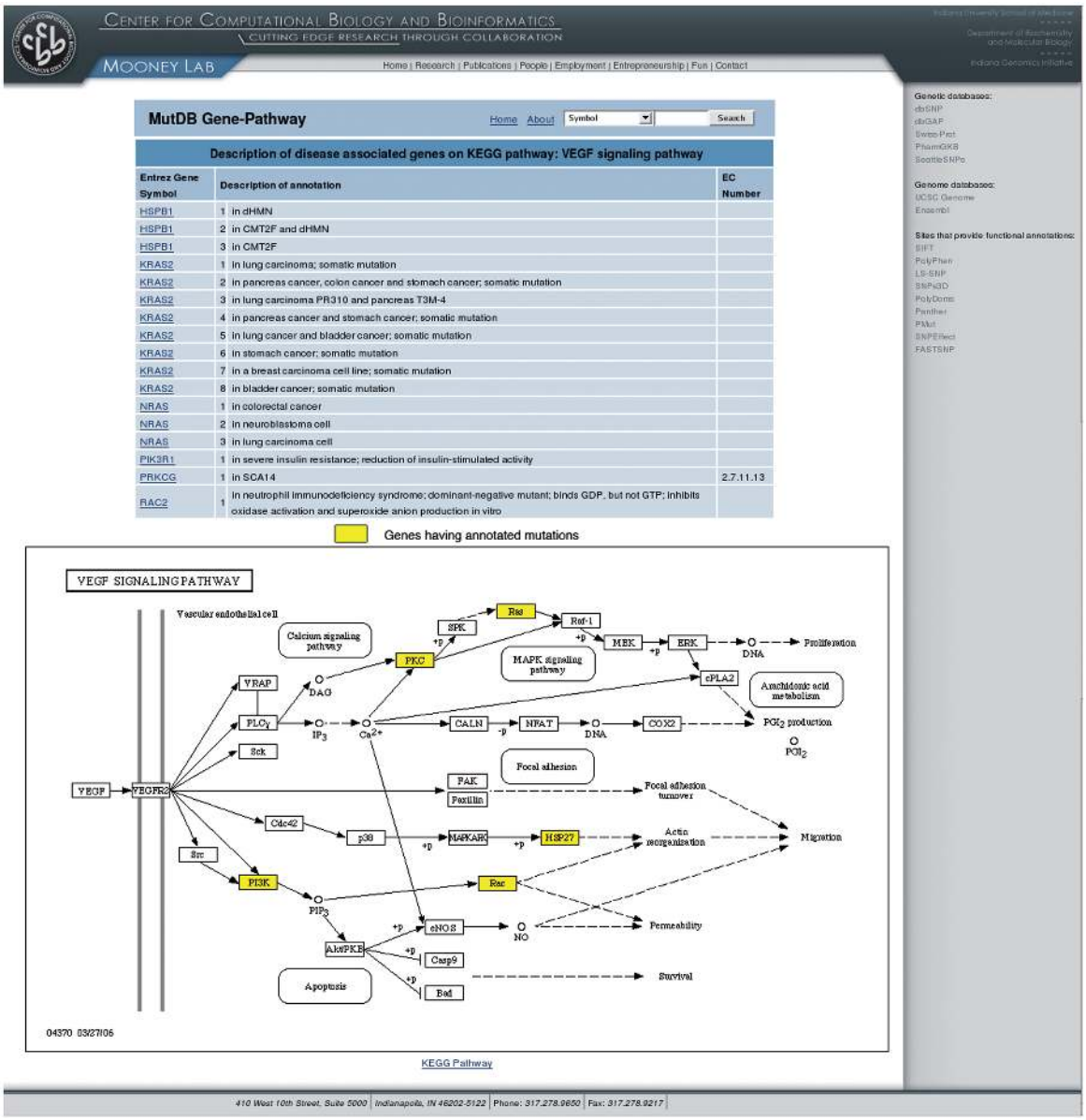

Figure 2. MutDB-KEGG integration example of the VEGF Pathway. This pathway shows all proteins with SNPs or Swiss-Prot mutations and all unique diseases and comments provided by Swiss-Prot (top). The VEGF signaling pathway showing proteins with mutations in yellow (bottom).

(http://www.rcsb.org/pdb/) (27) structures, KEGG Pathways, HapMap data and Entrez Gene information. We describe important aspects of our annotation pipeline below.

Protein structure annotations

Protein structural mapping for each amino acid substitution is performed by aligning the query sequence to each high scoring segment pair (HSP) from BLAST (28) search results using BioPerl scripts (29). BLAST results used for alignment come only from PDB (using a sequence data file downloaded in January of 2007) and are limited to those with 100% identity to the original sequence. These pairwise alignments are then used to map wild-type and mutation sequence to structure sequence. The annotated mutations that are mapped to a structure can be displayed using the integrated Jmol visualization tool (http://jmol.sourceforge.net/) or in extensions developed for UCSF Chimera (30) and Delano Scientific PyMOL (http://pymol.sourceforge.net/). To download the extensions visit http://lifescienceweb.org/.

Function annotations

We provide links to other tools that provide predictions of functional or molecular disruptions caused by an amino acid substitution. These include SNPs3D (5), PolyPhen (4), SIFT (3), PolyDoms (9), PMUT (7) and PANTHER (6) and are deep linked directly to the gene or SNP page, if available. Sorting Intolerant from Tolerant (SIFT) scores (3) and their associated predictions are supplied for

each variant causing an amino acid substitution. Variants with low confidence scores are marked with an asterisk. Here, again, the source Swiss-Prot and dbSNP pages are linked.

Visualization on KEGG pathways

We have augmented MutDB annotations with KEGG pathways using KEGG web services (23). This enables visualizing proteins, mutations and pathways on approximately 188 human pathways found in KEGG. The addition of a link, 'Visualize Pathways', on the MutDB gene page takes the user to a page listing the names of all KEGG pathways involving the gene. When a pathway is chosen, the user is taken to a new page displaying the pathway and a list of involved genes and their associated phenotypes.

All genes containing a SNP denoted as having a disease annotation or comment (per Swiss-Prot) are colored yellow in the pathway. This page is also hyper-linked to KEGG and MutDB. This functionality makes use of KEGG SOAP-based web services with supplementary data saved locally (Figure 2).

SNP query tool

A recent addition to our toolset is a SNP query tool that enables querying and exporting sets of SNPs that share selected features. The SNP query tool requires two sequence-tagged site (STS) markers or dbSNP reference cluster IDs (rs#) as input and returns all SNPs between the markers. The tool uses AJAX and a paging scheme to increase responsiveness upon large data sets. AJAX enhances speed by exchanging small amounts of data with the server, so the entire web page need not be reloaded each time the user makes a change. This technique along with the broad filtering options provide for an interactive tool.

Users may filter SNPs by manual selection or one of the filtering criteria. There are currently eleven filter options: validation status in dbSNP, hapmap status, location (functional class), avHet (average heterozygosity in dbSNP), avHetSE (SE for the average heterozygosity in dbSNP), CEU (CEPH—Utah residents with ancestry from northern and western Europe frequencies in HapMap), CHB (Han Chinese in Beijing, China), JPT (Japanese in Tokyo, Japan), YRI (Yoruba in Ibadan, Nigeria), SIFT score (3) and conservation score [based on the UCSC Genome Annotation Database conservation (25)]. The conservation score is the averaged 10-mer window of conservation values around each SNP derived from alignments of the 16 vertebrate species in the UCSC Annotated Genome Database.

A user can authenticate to enter the tool or visit as a guest, and may save each session and return later. Retrieval of sequence surrounding the SNP and exportation of SNP data to Microsoft Excel are easily performed via provided links. Excel output includes the dbSNP rsID, primer sequences, and the polymorphic alleles. The tool displays a PNG image containing RefSeq transcript information and location information for all selected SNPs indexed by function type using the UCSC Genome Annotation Database. A user may also visualize linkage disequilibrium for up to 200 selected SNPs in a Haploview (24) like structure. The SNP query tool is located at http://www.mutdb.org/snp and is linked from each page (Figure 1).

Table 1. Top 15 accessed genes on MutDB from October 2005 to January 2007

Symbol	Name
1. BRCA1	Breast cancer 1, early onset
2. CFTR	Cystic fibrosis transmembrane conductance
3. AR	Androgen receptor
4. APOE	Apolipoprotein E precursor
5. ATP7B	ATPase, Cu++transporting, beta polypeptide
6. TP53	Tumor protein p53
7. CD53	CD53 antigen
8. BRCA2	Breast cancer 2, early onset
9. FBN1	Fibrillin 1
10. APC	Adenomatosis polyposis coli
11. NOTCH3	Notch homolog 3
12. KALRN	Kalirin, RhoGEF kinase
13. CYP2D6	Cytochrome P450, family 2, subfamily D
14. RET	Ret proto-oncogene
15. HBB	Beta globin

BRCA1, CFTR, AR and APOE are the most requested pages within MutDB.

Continued web services support

MutDB continues to support its SOAP-based web services. The web services can be accessed via http://www.lifescienceweb.org. This interface is used to communicate to the structural visualization extensions for UCSF Chimera and Delano Scientific PyMOL.

Most accessed gene pages

In MutDB, the most accessed genes may give insight into the current interests of researchers. The most accessed genes from October 2005 to January 2007 are listed in Table 1. Not surprisingly, the most accessed genes also have many mutations associated with them and are what we would consider to be well-studied disease-associated genes.

Future

Understanding the underlying molecular causes of disease remains an important area for research. We continue to investigate annotations that are useful for hypothesis generation and directing experimental validation. While we continue to update the database as new annotations become available, we are also adding useful annotations outside of protein amino acid changes such as noncoding, synonymous and intronic variation.

ACKNOWLEDGEMENTS

We would like to thank Shoji Ichikawa and Somying Promso for helpful comments on the SNP query tool. We are supported by NLM K22LM009135 (PI: Mooney), P01AG018397 (PI: Econs), a grant from IU Biomedical

Research Council, an RSFG grant from IUPUI, the Showalter Trust and the Indiana Genomics Initiative. The Indiana Genomics Initiative (INGEN) is supported in part by the Lilly Endowment. RH is supported by Indiana Pervasive Computing Research (IPCRES) Initiative. Funding to pay the Open Access publication charges for this article was provided by NLM K22LM009135 (PI: Mooney).

Conflict of interest statement. None declared.

REFERENCES

1. Mooney,S. (2005) Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief Bioinform.*, **6**, 44–56.
2. 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.
3. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
4. Sunyaev,S., Ramensky,V., Koch,I., Lathe,W., III, Kondrashov,A.S. and Bork,P. (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.
5. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.
6. Mi,H., Lazareva-Ulitsky,B., Loo,R., Kejariwal,A., Vandergriff,J., Rabkin,S., Guo,N., Muruganujan,A., Doremieux,O. *et al.* (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.*, **33**, D284–D288.
7. Ferrer-Costa,C., Gelpi,J.L., Zamakola,L., Parraga,I., de la Cruz,X. and Orozco,M. (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics*, **21**, 3176–3178.
8. Karchin,R., Diekhans,M., Kelly,L., Thomas,D.J., Pieper,U., Eswar,N., Haussler,D. and Sali,A. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*, **21**, 2814–2820.
9. Jegga,A.G., Gowrisankar,S., Chen,J. and Aronow,B.J. (2007) PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease. *Nucleic Acids Res.*, **35**, D700–D706.
10. Reumers,J., Schymkowitz,J., Ferkinghoff-Borg,J., Stricher,F., Serrano,L. and Rousseau,F. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res.*, **33**, D527–D532.
11. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
12. Saunders,C.T. and Baker,D. (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.*, **322**, 891–901.
13. Cavallo,A. and Martin,A.C. (2005) Mapping SNPs to protein sequence and structure data. *Bioinformatics* (Oxford, England), **21**, 1443–1450.
14. 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.
15. Dobson,R.J., Munroe,P.B., Caulfield,M.J. and Saqi,M.A. (2006) Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. *BMC bioinformatics*, **7**, 217.
16. 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* (Oxford, England), **19**, 2199–02209.
17. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
18. Mooney,S.D. and Altman,R.B. (2003) MutDB: annotating human variation with functionally relevant data. *Bioinformatics*, **19**, 1858–1860.
19. Dantzer,J., Moad,C., Heiland,R. and Mooney,S. (2005) MutDB services: interactive structural analysis of mutation data. *Nucleic Acids Res.*, **33**, W311–W314.
20. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
21. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
22. Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M. *et al.* (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32**(Database issue), D35–D40.
23. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
24. Barrett,J.C., Fry,B., Maller,J. and Daly,M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
25. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
26. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
27. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
28. Altschul,S., Madden,T., Schaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped BLAST and PSI-BLAST: A New Generation Of Protein Database Search Tools. *Nucleic Acids Res.*, **25**, 3389–3402.
29. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
30. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
31. The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.