D810–D815 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 25 October 2008*
doi:10.1093/nar/gkn798

VarySysDB: a human genetic polymorphism database based on all H-InvDB transcripts

Makoto K. Shimada[1,2], Ryuzou Matsumoto[3], Yosuke Hayakawa[2,3], Ryoko Sanbonmatsu[1,2], Craig Gough[1,2], Yumi Yamaguchi-Kabata[1], Chisato Yamasaki[1,2], Tadashi Imanishi[1,*] and Takashi Gojobori[1,4]

[1]Integrated Database and Systems Biology Team, Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, [2]Japan Biological Informatics Consortium (JBIC), [3]Hitachi Software Engineering Co., Ltd., Tokyo and [4]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Shizuoka, Japan

Received August 14, 2008; Revised October 8, 2008; Accepted October 10, 2008

ABSTRACT

Creation of a vast variety of proteins is accomplished by genetic variation and a variety of alternative splicing transcripts. Currently, however, the abundant available data on genetic variation and the transcriptome are stored independently and in a dispersed fashion. In order to provide a research resource regarding the effects of human genetic polymorphism on various transcripts, we developed VarySysDB, a genetic polymorphism database based on 187 156 extensively annotated matured mRNA transcripts from 36 073 loci provided by H-InvDB. VarySysDB offers information encompassing published human genetic polymorphisms for each of these transcripts separately. This allows comparisons of effects derived from a polymorphism on different transcripts. The published information we analyzed includes single nucleotide polymorphisms and deletion–insertion polymorphisms from dbSNP, copy number variations from Database of Genomic Variants, short tandem repeats and single amino acid repeats from H-InvDB and linkage disequilibrium regions from D-HaploDB. The information can be searched and retrieved by features, functions and effects of polymorphisms, as well as by keywords. VarySysDB combines two kinds of viewers, GBrowse and Sequence View, to facilitate understanding of the positional relationship among polymorphisms, genome, transcripts, loci and functional domains. We expect that VarySysDB will yield useful information on polymorphisms affecting gene expression and phenotypes. VarySysDB is available at http://h-invitational.jp/varygene/.

INTRODUCTION

Accumulated information on human genetic polymorphisms has encouraged genome-wide association studies that use polymorphisms as markers. This approach is now commonly used in various studies (1,2), and has led to a greater understanding of the diversity in phenotypes as well as pathogenic biological processes.

Currently, several kinds of human genetic polymorphism databases aid researchers in exploring genetic information for various applications. Examples of such databases include the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/index.html) (3) containing information on single nucleotide polymorphisms (SNPs) and short deletion and insertion polymorphisms (DIPs) as submitted by the corresponding authors of the published data. The Database of Genomic Variants (http://projects.tcag.ca/variation/) (4) provides genomic regions involved in structural variations as defined by alternations in DNA segments larger than 1 kb. Short Tandem Repeats (STRs), also known as simple sequence repeats or microsatellites are a different type of major source of genomic diversity. Information on human STRs is available from public domains such as UgMicroSat*db* (http://www.veenuash.info/web1/index.htm) (5), UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway) (6) and GDBS/H-GOLD (http://hinv.jp/gdbs/) (7). Accordingly, these polymorphism data are described by

*To whom correspondence should be addressed. Tel: +81 3 3599 8800; Fax: +81 3 3599 8801; Email: t.imanishi@aist.go.jp
Correspondence may also be addressed to Takashi Gojobori. Tel: +81 55 981 6847; Fax: +81 55 981 6848; Email: tgojobor@genes.nig.ac.jp
Present address:
Makoto K. Shimada, Institute for Comprehensive Medical Science, Fujita Health University, Aichi 470-1192, Japan

position in the human reference genome (i.e. genome coordinate).

Recently, the accumulated knowledge on alternative splicing and the regulation of gene expression has reinforced the importance of transcript multiplicity as another source of diversity of protein function. H-InvDB catalogs a comprehensive annotation of the human transcriptome including transcript diversities and gene expression profiles (8,9). H-InvDB is concentrating on full-length cDNA annotation to overcome the limitation of conventional databases based on high-throughput EST data, such as scarce distributions around the 5′-ends of mRNAs and absence of some combination of the alternative splicing (AS) exons (10). Thus, H-DBAS, a satellite database of H-InvDB which is a specialized AS database, is the comprehensive database containing AS accurately annotated manually and automatically based on highly reliable cDNA sequences (11).

The currently available human genetic polymorphism databases described above have not yet been integrated with well-annotated AS isoforms conforming to a uniform standard. Therefore, we developed VarySysDB, a database of human genetic polymorphisms based on all of the 187 156 matured mRNA transcripts from 36 073 loci provided by H-InvDB. [Hereinafter, these matured mRNA transcripts annotated by H-InvDB and the loci defined by transcript clusters will be called H-inv transcripts (HITs) and H-Inv clusters (HIXs), respectively]. VarySysDB provides separately annotated genetic polymorphisms for each HIT, even from multiple transcripts forming a HIX. It provides information regarding SNPs, DIPs, STRs, single amino acid repeats (SARs), structural variation (or copy number variations; CNVs), linkage disequilibrium (LD) regions and their relationship with the genome, HITs, and functional domains. Moreover, we designed VarySysDB to include annotations we made, which covers intronic SNPs located on conserved dinucleotide splice sites, nonsynonymous SNPs that affect functional (InterPro) and protein structural (SCOP) domains, and polymorphic tandem repeat sequences, as well as other publicly available information. Since VarySysDB is a satellite database of H-InvDB, it is well designed to provide appropriate links for each HIT to H-InvDB, as well as to other related public databases. All of the annotation data in VarSysDB is available to all users, with no restriction to academic users only. We hope that VarySysDB will deliver an even greater understanding of the various biological processes, permit a detailed evaluation of how polymorphisms affect different phenotypes, and foster a rich research environment focused on exploring the causes of genetic variation through genome-wide association studies.

CONSTRUCTION OF THE DATABASE

Source of data

Table 1 lists the data used to construct VarySysDB. This database includes the transcript data from H-InvDB, as well as published genetic polymorphism data.

Mapping genetic polymorphism on H-Inv transcripts

We mapped all the genetic polymorphism data onto the exact transcript position using our in-house program to convert their location from genome coordinates to those of the HIT.

VarySysDB contains these polymorphism data with the following conditions as well as our own annotations. (i) SNPs and DIPs: SNP and DIP data were downloaded from dbSNP (Table 1). We eliminated SNP and DIP

Table 1. Data used in VarySysDB

	Number of data available in VarySysDB	Database: name and version (or date of download)	Provider	URL	References
H-Inv Transcripts (HITs)	187 156	H-InvDB 5.0[a]	BIRC[b]	http://h-invitational.jp/hinv/	(8,9)
SNPs & DIPs	11 817 893[c]	dbSNP build 128	NCBI[d]	http://www.ncbi.nlm.nih.gov/projects/SNP/	(3,14)
STRs	18 637	H-InvDB 5.0	BIRC	http://h-invitational.jp/hinv/	(8,9)
SARs	33 007	H-InvDB 5.0	BIRC	http://h-invitational.jp/hinv/	(8,9)
CNVs	11 966	DGV (hg18.v3)[e]	TCAG[f]	http://projects.tcag.ca/variation/	(4)
LD-bins	99 921	D-HaploDB[g]	Kyushu University	http://orca.gen.kyushu-u.ac.jp/	(13)
OMIM allelic variants	950	OMIM (1.28. 2008)[h]	NCBI	http://www.ncbi.nlm.nih.gov/omim/	(14,15)
Functional domain	–	InterPro 15.1	EBI[i]	http://www.ebi.ac.uk/interpro/	(16,17)
Structural domain (SCOP)	–	GTOP (2.4. 2008)[j]	NIG[k]	http://sybock.genes.nig.ac.jp/~hinv4/gtop.html	(18)

[a]H-Invitational Database Release 5.0 (used human genome sequence version: hg18, NCBI Build36.2).
[b]Biomedicinal Information Research Center.
[c]The number of SNP & DIP data downloaded from dbSNP and analyzed for VarySysDB. The numbers of annotated SNP and HIT pairs are as follows: 568 982 for non-synonymous, 431 433 for synonymous, 747 for synonymous at stop-codon, 7227 for termination, 1510 for stop codon to amino acid, 8945 for NMD.
[d]National Center for Biotechnology Information.
[e]Database of Genomic Variants.
[f]The Centre for Applied Genomics.
[g]Database of Definitive Haplotypes.
[h]Online Mendelian Inheritance in Man™.
[i]European Bioinformatics Institute.
[j]PDB 2007-Apr-6, Swissprot 52.1, SCOP 1.69, Pfam 21.0, ProSite 20.0, Wormpep 174, HUGE 2003-11-6(kiaa2038).
[k]National Institute of Genetics.

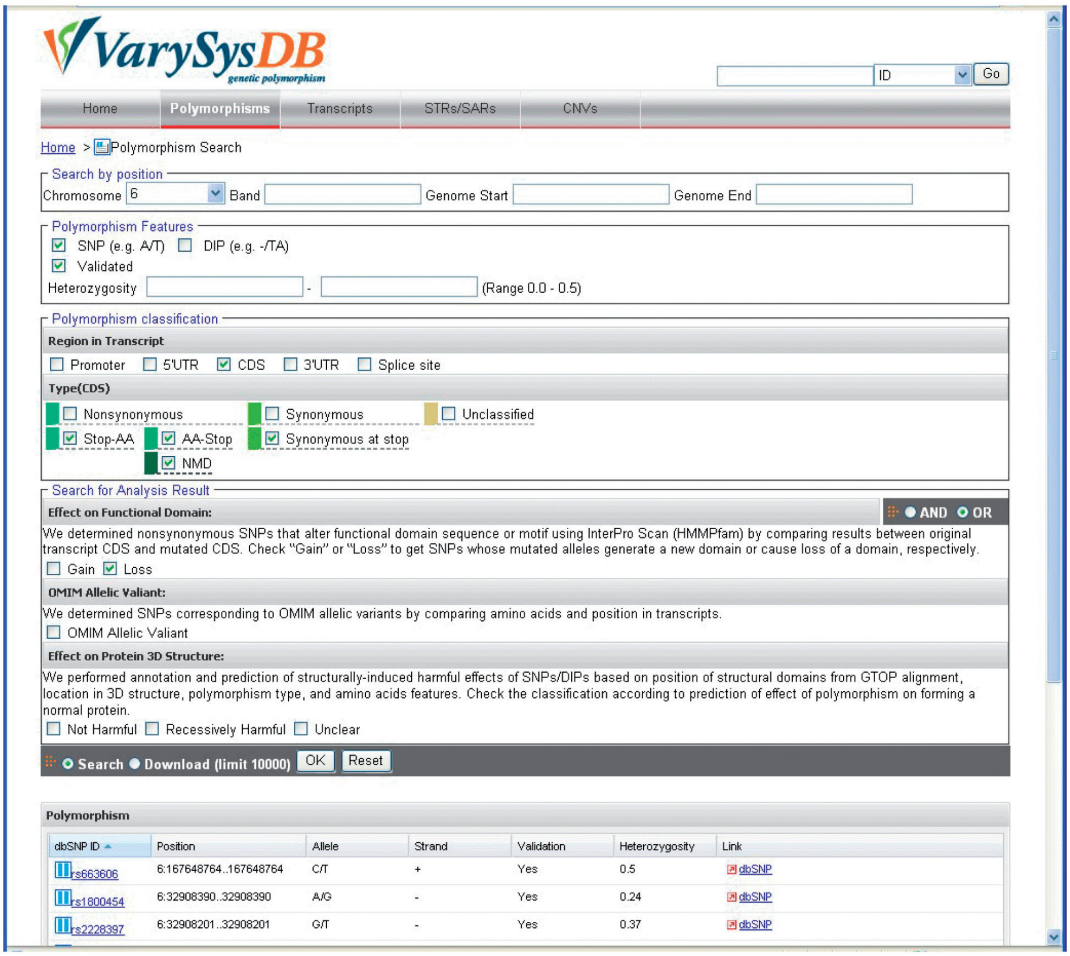

Figure 1. View of polymorphism search page, which is one of the search pages contained in VarySysDB. In the polymorphism search page, users can search the polymorphism data by features, classification and our analysis results such as effects on functional domains and protein 3D structures. Four boxes ('Search by postion', 'Polymorphism features', 'Polymorphism classification', 'Search for analysis result') organize the search criteria by subject. When multiple search criteria are specified 'over' these boxes, an 'and' search is conducted, offering polymorphisms matching all the specified criteria.

data if their alleles contradicted the transcript sequences (12). (ii) STRs and SARs: We searched HIT sequences for STRs, with an STR defined as a repeat of ten or more dinucleotides and a repeat of five or more tri-, tetra- and penta-nucleotide sequences. For SARs, we searched the amino acid sequences translated from HIT sequences for single amino acid repeats of five or more. (iii) OMIM allelic variant (OMIM AV): we downloaded OMIM AVs with MIM Number Prefixes of 'gene with known sequence' using the 'limit' GUI of the OMIM web page (Table 1). We filtered the OMIM AVs in the exonic region included in the dbSNP by checking each location in the HIT using our in-house programs to annotate separately.

Annotation

We classified SNPs according to their effect on translation based on each HIT sequence (Table 1). This highlighted our unique annotation regarding the effect of each SNP on different HITs within a HIX. We also classified SNPs and DIPs according to their locations in HITs, which includes the promoter (defined as the region within 2 kb upstream of first exon), the splice dinucleotide site or the exonic regions. Furthermore, we annotated SNPs and DIPs in the coding region into the following categories: (i) those that alter functional (InterPro) domain sequences so drastically that InterProcScan results change (effect on functional domain); (ii) those that are located in protein structural (SCOP) domains and change amino acid characters so as to result in harmful effects on the protein 3D structure; (iii) those that match their location in HITs and alleles to descriptions of OMIM AVs (OMIM Allelic Variants) (Figure 1). Cases in category (ii), those that have an effect on protein structure, are subdivided into three subcategories chosen according to the effect of the polymorphism: (a) 'Not Harmful'; (b) 'Recessively Harmful' due to loss or reduction of function; (c) 'Possible to be Harmful (Unclear)' because of a drastic change of a structural domain which may induce toxic aggregation.

Within STRs and SARs, we distinguished the polymorphic cases by transcript sequence alignments.

For CNVs, we downloaded from DGV (Table 1), and classified them according to the detection methods described in the downloaded data into six divisions for the convenience of users.

ACCESSING THE DATABASE

Database contents and organization

Table 2 lists the web-interfaces or GUIs in VarySysDB containing six search pages. The results of searches can be downloaded as well as easily displayed on the computer screen. VarySysDB is composed of three subsystems, including Varygene2, LD Search System and GBrowse. A menu bar of Varygene2 is designed to select search pages from 'Polymorphisms', 'Transcripts', 'STRs/SARs' and 'CNVs.'

By clicking 'Polymorphisms', users can search by feature and our aforementioned annotation regarding SNPs and DIPs (Figure 1).

STRs/SARs with length polymorphisms proven by our sequence alignment can be extracted from an STR/SAR Search page. VarySysDB can retrieve STRs and SARs according to features such as the repeat unit sequence (e.g. 'at' nucleotides for STR, 'P' amino acid for SARs) and number of repeats.

By clicking 'CNVs', users can search by features of CNVs, such as CNV class (i.e. copy number variation or inversion) and detection method.

Table 2. System and web-interface design of VarySysDB

Subsystem	Web-interface	Function
Varygene 2	Polymorphism search	Retrieving and displaying genetic polymorphisms.
	Polymorphism table	Displaying detailed information on polymorphisms.
	Transcript search	Retrieving and displaying transcript information.
	Transcript table	Displaying detailed information on transcripts.
	Sequence view	Displaying cDNA sequence with information on polymorphisms and functional domains.
	STR/SAR search	Retrieving and displaying STRs and SARs.
	CNV search	Retrieving and displaying CNVs.
	CNV table	Displaying detailed information on CNVs.
	Keyword search	Retrieving and displaying by ID, gene name or definition.
	System information	Displaying summary table showing total numbers of transcripts and polymorphisms in VaryGene2.
LD-Search	–	Retrieving and displaying LD-bins within the specified region.
GBrowse	–	Displaying genomic region specified with HITs, HIXs and polymorphisms.

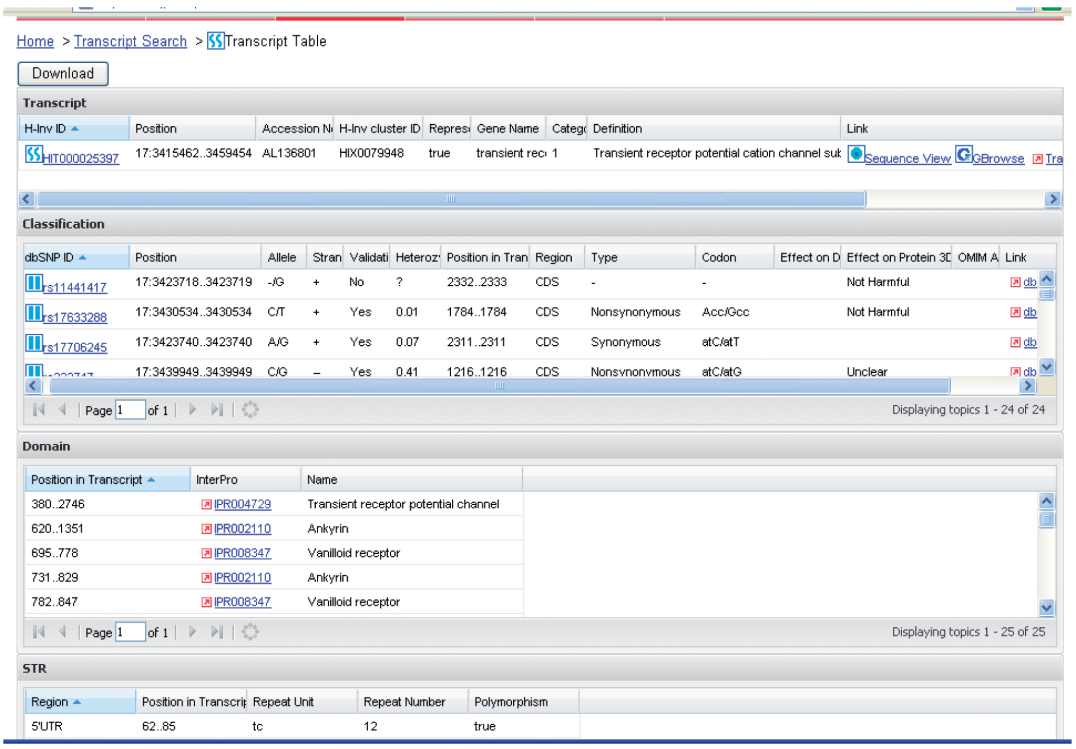

Figure 2. Transcript table containing information on HIT, polymorphism mapped on the HIT (SNP classification, STRs and SARs), and functional domain included in the HIT.

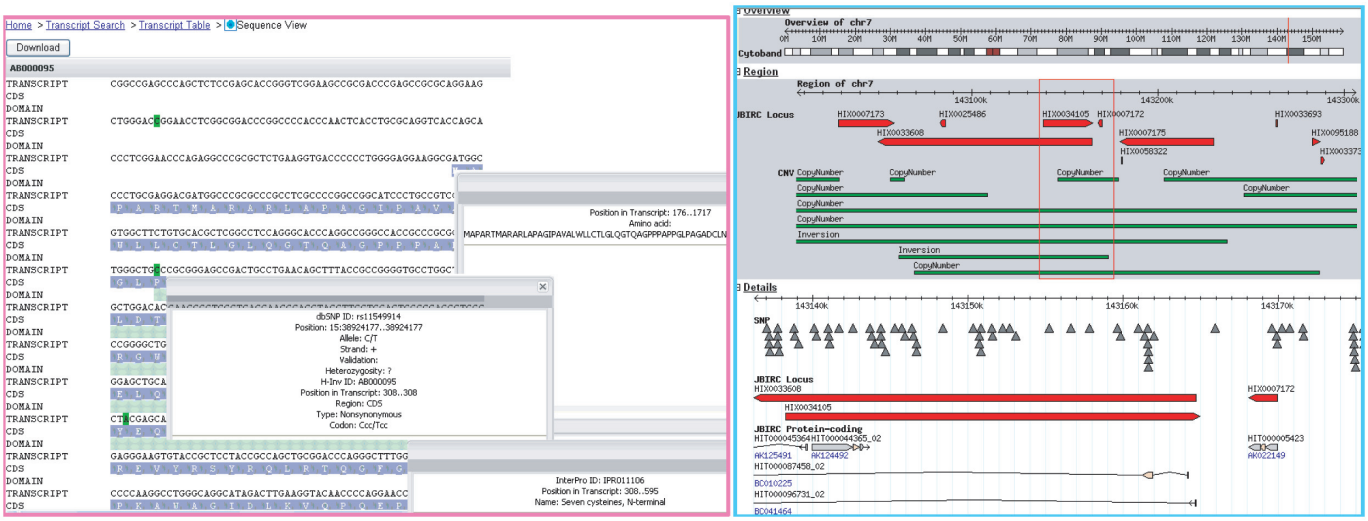

Figure 3. Two graphical viewers in VarySysDB. Left: Sequence View containing polymorphism, domain, sequence of HIT and amino acid sequence. Right: GBrowse showing position of SNP, CNV, HIT and HIX.

The genetic polymorphism data in VarySysDB are also searchable by the features of the HIT on which the polymorphism is located (Transcript Search in Table 2). The HIT features defined in H-InvDB include 'representative transcript', that is, the best HIT to represent a HIX, and 'similarity category' as determined by the level of similarity to known human proteins or InterPro domains (Figure 2).

Sequence View shows the sequence of a HIT and the corresponding amino acid sequence with positional relationship among SNPs, DIPs and functional domains (Figure 3).

VarySysDB also has an LD Search System. This is a subsystem to retrieve LD-bin data distributed within a specified region of the chromosome. The LD-bins offered here are definitive haplotypes that originate from a single sperm, indicating that they are free from errors, which are typically caused by the inference from diploid genotypes (13). This enables users to detect associations among polymorphisms.

GBrowse in VarySysDB can be used to navigate positional relationships among HITs, HIXs and polymorphisms. Since GBrowse is an open-source architecture with various functions, users can conveniently download information from the retrieved region and upload their own data to make comparisons with the information in VarySysDB (Figure 3).

These various web interfaces enable users to extract human genetic polymorphism annotations with user-friendly search systems.

Availability

VarySysDB can be downloaded and freely accessed, with no restriction to academic users only, from http://h-invitational.jp/varygene/. A help document is also available from http://www.h-invitational.jp/hinv/help/Documents/VarySysDB_help.pdf.

ACKNOWLEDGEMENTS

We thank members of the Integrated Database and Systems Biology Team from the Biomedicinal Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST) for their helpful suggestions and cooperation. Especially we thank Akihiro Matsuya, Takuya Habara and Tomohiro Endo for their technical support for constructing and publishing the database. We are also grateful to Drs Shinsei Minoshima (Hamamatsu Univ. School of Medicine), Satoshi Fukuchi (NIG) and Kenshi Hayashi and Koichiro Higasa (Kyusyu Univ.) for effective discussion on this work.

FUNDING

The Ministry of Economy, Trade and Industry of Japan; Japan Biological Informatics Consortium. Funding for open access publication charge: Japan Biological Informatics Consortium.

Conflict of interest statement. None declared.

REFERENCES

1. Maresso,K. and Broeckel,U. (2008) Genotyping platforms for mass-throughput genotyping with SNPs, including human genome-wide scans. In Rao,D.C. and Gu,C.C. (eds), *Advance in Genetics*. Vol. 60, Elsevier, Amsterdam, pp. 107–139.
2. Seng,K.C. and Seng,C.K. (2008) The success of the genome-wide association approach: a brief story of a long struggle. *Eur. J. Hum. Genet.*, **16**, 554–564.
3. 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.
4. Iafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W. and Lee,C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
5. Aishwarya,V. and Sharma,P.C. (2008) UgMicroSatdb: database for mining microsatellites from unigenes. *Nucleic Acids Res.*, **36**, D53–D56.

6. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
7. Tamiya,G., Shinya,M., Imanishi,T., Ikuta,T., Makino,S., Okamoto,K., Furugaki,K., Matsumoto,T., Mano,S., Ando,S. *et al.* (2005) Whole genome association study of rheumatoid arthritis using 27 039 microsatellites. *Hum. Mol. Genet.*, **14**, 2305–2321.
8. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, e162.
9. Yamasaki,C., Murakami,K., Fujii,Y., Sato,Y., Harada,E., Takeda,J., Taniya,T., Sakate,R., Kikugawa,S., Shimada,M. *et al.* (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res.*, **36**, D793–D799.
10. Takeda,J.-i., Suzuki,Y., Nakao,M., Barrero,R.A., Koyanagi,K.O., Jin,L., Motono,C., Hata,H., Isogai,T., Nagai,K. *et al.* (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56 419 completely sequenced and manually annotated full-length cDNAs. *Nucleic Acids Res.*, **34**, 3917–3928.
11. Takeda,J.-i., Suzuki,Y., Nakao,M., Kuroda,T., Sugano,S., Gojobori,T. and Imanishi,T. (2007) H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. *Nucleic Acids Res.*, **35**, D104–D109.
12. Yamaguchi-Kabata,Y., Shimada,M.K., Hayakawa,Y., Minoshima,S., Chakraborty,R., Gojobori,T. and Imanishi,T. (2008) Distribution and effects of nonsense polymorphisms in human genes. *PLoS ONE*, **3**, e3393.
13. Higasa,K., Miyatake,K., Kukita,Y., Tahira,T. and Hayashi,K. (2007) D-HaploDB: a database of definitive haplotypes determined by genotyping complete hydatidiform mole samples. *Nucleic Acids Res.*, **35**, D685–D689.
14. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
15. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
16. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
17. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P., Bucher,P. *et al.* (2002) InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform.*, **3**, 225–235.
18. Kawabata,T., Fukuchi,S., Homma,K., Ota,M., Araki,J., Ito,T., Ichiyoshi,N. and Nishikawa,K. (2002) GTOP: a database of protein structures predicted from genome sequences. *Nucleic Acids Res.*, **30**, 294–298.