© 2002 Oxford University Press Nucleic Acids Research, 2002, Vol. 30, No. 1 387–391

HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources

D. Fredman, M. Siegfried, Y. P. Yuan[1], P. Bork[1], H. Lehväslaiho[2] and A. J. Brookes*

Center for Genomics and Bioinformatics, Karolinska Institute, Berzelius väg, S171 77 Stockholm, Sweden, [1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany and [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

Received September 20, 2001; Accepted September 26, 2001

ABSTRACT

HGVbase (Human Genome Variation database; http://hgvbase.cgb.ki.se, formerly known as HGBASE) is an academic effort to provide a high quality and non-redundant database of available genomic variation data of all types, mostly comprising single nucleotide polymorphisms (SNPs). Records include neutral polymorphisms as well as disease-related mutations. Online search tools facilitate data interrogation by sequence similarity and keyword queries, and searching by genome coordinates is now being implemented. Downloads are freely available in XML, Fasta, SRS, SQL and tagged-text file formats. Each entry is presented in the context of its surrounding sequence and many records are related to neighboring human genes and affected features therein. Population allele frequencies are included wherever available. Thorough semi-automated data checking ensures internal consistency and addresses common errors in the source information. To keep pace with recent growth in the field, we have developed tools for fully automated annotation. All variants have been uniquely mapped to the draft genome sequence and are referenced to positions in EMBL/GenBank files. Data utility is enhanced by provision of genotyping assays and functional predictions. Recent data structure extensions allow the capture of haplotype and genotype information, and a new initiative (along with BiSC and HUGO-MDI) aims to create a central repository for the broad collection of clinical mutations and associated disease phenotypes of interest.

INTRODUCTION AND OVERVIEW

HGVbase (Human Genome Variation database; http://hgvbase.cgb.ki.se) is the new name adopted by the HGBASE project in order to better reflect the scope of the database and its additional new role (see below) as a central depository for mutation collection efforts undertaken in allegiance with the newly inaugurated Human Genome Variation Society (HGVS; http://www.hgvs.org). HGVbase is motivated by the intuitive premise that gene and genome sequence variations underpin a large fraction of human phenotypic variation. Thus, HGVbase was constructed to provide a non-redundant collection of all known and suspected human DNA variants of any type, emphasizing high data quality and active data collection from a broad range of sources. Records are assigned unique and permanent HGVbase identification numbers to facilitate cross-database referencing, publication and similar use by the community. The inclusion of a variation does not depend upon whether or not information exists pertaining to allele frequencies, functional/phenotypic consequences or validation status. Records include polymorphisms (sequence variations in which the most abundant allele has a frequency of <99%) as well as variations with rare or single occurrence alleles, plus disease-related and disease-causing clinical mutations. In this article, polymorphisms and mutations are referred to equally as 'variations' or 'variants'.

From its original inception as a joint academic–industry initiative, HGVbase is now run as a fully academic project. Overall scientific planning, data curation work and hosting/development of the main web pages are the responsibility of Anthony Brookes and colleagues at the Center for Genomics and Bioinformatics at the Karolinska Institute (KI), in Stockholm, Sweden. Database design and implementation, as well as data ascertainment and provision to the community (downloads and search functions) are tasks shared between the KI team and the group of Heikki Lehväslaiho at the European Bioinformatics Institute (EBI) in Hinxton, UK. Programming developments that improve and automate the tasks of data processing/curation are contributed by both the KI and EBI teams, with more ambitious and powerful systems being created by Yan P. Yuan and Peer Bork at the European Molecular Biology Laboratories (EMBL) in Heidelberg, Germany. Exploitation of the Ensembl project resources developed at EBI is also central to the HGVbase curation and annotation pipeline. Project activities are funded by non-conditional research support and financial donations provided by the KI, EBI and EMBL institutions, plus industry groups Pharmacia, Celera and Doubletwist. In the spirit of a public-domain community project, we would welcome expressions of interest from additional groups that might wish to become involved in

*To whom correspondence should be addressed. Tel: +46 8 7286630; Fax: +46 8 331547; Email: anthony.brookes@cgb.ki.se

funding or helping to execute any aspects of the HGVbase venture.

DATA PRESENTATION AND AVAILABILITY

A full database description is provided online and accompanying web pages summarize single nucleotide polymorphism (SNP)-related meetings, genotyping methods, SNP-databases and analysis software.

Interrogating HGVbase returns flat-file descriptions that use feature tags to organize the data. Records are placed in the context of surrounding genome or cDNA sequence, showing 25 bases on either side. In many cases (work underway) entries are also related to some human genes in which they reside, detailing the intragenic region (intron, exon, untranslated region, promoter) and summarizing consequences to coding regions and splice sites. Wherever possible, exon-located variants are described at the DNA, RNA and protein sequence levels. Variation positions are indicated within a reference EMBL or GenBank sequence, and all entries have been uniquely mapped to the human draft genome. Allele frequency data is also provided. Hypertext links are provided to the given EMBL or GenBank file, as well as to representations of source information in other databases, to PubMed for literature sources, to host gene summaries, to submitter contact details, to restriction enzymes available for RFLP–PCR genotyping and to HGVbase feature definitions. The full HGVbase content is freely available according to copy-left principles (http://www.gnu.org/copyleft) through our web pages in a variety of formats (XML, Fasta, SRS, SQL dumps and tagged-text flat-files).

Online search tools facilitate data interrogation of all information fields in HGVbase, comprising five major categories: (i) sequence location, (ii) predicted functional importance, (iii) validation status, (iv) allele frequencies in populations and (v) data sources. By sequence similarity, use of keyword queries or setting thresholds for inclusion, users can extract subsets of records. Continuous enhancements to the existing interfaces have been made, such as the possibility to retrieve flanking sequences of an arbitrary length for a given set of HGVbase polymorphisms, and enabling searches by equivalent IDs in other major polymorphism/mutation databases.

A key ongoing development is the implementation of new search tools that will allow extraction of all variations residing within genomic domains, as defined by flanking elements that can be other variants or genetic markers, cytogenetic map coordinates, genes and absolute chromosomal positions.

DATABASE CONTENT

Sequence variation data is acquired from multiple sources (see below). However, not all accessible variants become included in HGVbase due to our high demands for data quality and our desire to emphasize those loci most likely to be immediately useful in the laboratory setting. Specifically, acceptance criteria and extensive curation efforts are applied to all potentially new records. This entails an assessment of (i) internal data consistency, (ii) accuracy of correlation to a unique position in a reference DNA sequence in a public database and (iii) accordance with our list of required data features (see below). By strict application of these rules, we find that <60% of presently reported sequence variants are acceptable for direct incorporation, and many records submitted directly to us by the community need to be passed through several cycles of review, return to submitter, correction and resubmission. By repeatedly (every few months) harvesting and processing larger public-domain datasets as they evolve, and as pertinent allied resources improve, we expect to be able to retain this uniquely high data quality and progress ultimately towards a comprehensive representation of all valid DNA variations.

Having processed the vast majority of publicly available sequence variations, HGVbase content in September 2001 comprised 983 480 non-redundant entries, representing 1 008 457 individual source records. Amongst these were 982 584 single nucleotide polymorphisms (SNPs), 814 insertion–deletion variants (indels), 46 simple tandem repeats (STRs) and 36 'generic' (or complex) changes involving alterations not covered by any of the preceding three definitions. Relating these to known gene structures is partially complete (~50% of all gene relationships identified) and remains ongoing, representing a challenging annotation task will be driven forward continually as the known human gene complement evolves and gene structures become more precisely defined.

DATA SOURCES

HGVbase input data is acquired from multiple sources, including public databases, the literature, our own and collaborative discovery efforts, and direct submissions provided by the community. In addition to the variations themselves, allele frequency data in different populations (even for previously known and well-studied variations) is highly valuable information and its supply is very much appreciated. No claim of ownership of any underlying HGVbase data is made by us, though our specific compilation and representation of it are subject to our copyleft and ownership claims (ftp://ftp.ebi.ac.uk/pub/databases/variantdbs/hgvbase/LICENSE). These are designed to ensure this resource remains freely available to everyone for research purposes.

A great deal of HGVbase data originates from various public polymorphism and mutation databases. Approval is always sought before data is harvested from any other depository. To date, without exception, every public resource we have approached has been willing and proactive in helping us to access and process their information. No attempt has yet been made to acquire any private-domain lists of sequence variations. Bidirectional data exchange with dbSNP (1) was established at the end of the year 2000, though we only incorporate those records that succeed in passing all of our quality requirements summarized above.

Beyond harvesting of large public datasets, individual researchers make frequent submissions to HGVbase. Smaller lists of variants (up to tens of records) may be submitted as Microsoft Excel or Microsoft Word submission sheets that are filled in locally and then emailed to us. For larger submissions (up to a few thousand records) we will work with the submitter to create purpose-built software tailored towards extracting data from whatever format is convenient for the submitter. In both these above cases, we strive to fully manually curate (with the aid of established purpose-built visualization and data processing software) all aspects of the supplied data to ensure absolute data consistency and full coverage according to our list of record features summarized below. For even larger datasets, and where regular submissions might be anticipated (e.g. dbSNP downloads), we would establish fully automated curation procedures.

From 1998 to early 2001 we worked hard to manually extract new variations and related information from the published literature every week. Recent dramatic expansion of research in the field has put this task beyond the scope of our available manpower, and so we are now working to establish more automated procedures whereby authors of such papers will automatically be contacted and asked to make pre-formatted data submissions of the information they have published. These submissions will then be automatically validated as much as possible, and entered into HGVbase.

Details of data sources (and submitter contact details where appropriate) are provided within each HGVbase record. Existing HGVbase records as of September 2001 are a composite of information from 791 different sources that provided data as follows: 714 publications, 142 batch submissions, plus information from 30 web databases (AD Study Results Database, Albinism Database, ALFRED Database, Androgen Receptor Mutation Database, Ataxia-Telangiectasia Mutation Database, Breast Cancer Mutation Database, Canvas Database, CGAP-GAI Database, Cystic Fibrosis Mutation Database, Cytokine Gene Polymorphism Database, dbSNP Database, EGP Database, Factor VIII Database, Fanconi Anemia Mutation Database, GM2 Gangliosidoses Database, GSD II Database, Human Gene Mutation Database, Human Ornithine Transcarbamylase Database, Human Type I and Type III Collagen Mutation Database, Hypertension Candidate Gene SNP Database, ICG-HNPCC Database, IMS-JST Database, Leiden Muscular Dystrophy Database, Neuronal Ceroid Lipofuscinoses NCL Mutations Database, Online Mendelian Inheritance in Man Database, p53 Database, Phenylalanine Hydroxylase Locus Database, von Willebrand Factor Database, Whitehead SNP Database, WRN Mutations Database).

DATABASE STRUCTURE AND ORGANIZATION

HGVbase categorizes variations as either SNPs, indels, STRs or 'generic' alterations, for which the three letter codes SNP, IND, STR and GEN, respectively, are employed. For each newly included variant, a progressive integer (padded to nine digits with zeros) is added to the appropriate three letter code to create a unique HGVbase identifier. These HGVbase IDs are immediately allocated to newly processed submissions, and passed directly back to the data submitter to help with manuscript preparation or similar record referencing tasks. Redundancies between newly processed variants and existing HGVbase entries are identified on the basis of both (i) equivalent unique map location and corresponding offset in a public reference DNA sequence (deduced or claimed position based upon high-stringency sequence similarity searches) and (ii) identical variation type. Distinct allele details between two variants do not prevent record merging, but instead the allele alternatives are simply combined into a composite list.

The total set of information potentially specified for each record is as follows:

1. DNA sequence (required): comprising 25 bp 5′ of the polymorphism, the allelic bases themselves and 25 bp 3′ of the polymorphism—specified as either genomic or cDNA level data and in the same orientation as the direction of transcription if within the span of an identified gene. The purpose of this is to provide a convenient string for first-pass record comparison and localization within a short stretch of (e.g. gene) sequence.
2. A DDBJ/EMBL/GenBank reference DNA sequence (required): this comprises accession number(s) and affected component base(s) for at least one reference file, with the goal of providing both an ideal genomic and cDNA reference (if appropriate) for all. The purpose of this is to enable unambiguous whole genome positional referencing. Since most gene structures and coding domains are still far from completely defined, no attempt is presently made to use any formalized or standardized naming system for variations.
3. Approved gene name and symbol (optional): wherever possible, we strive to relate variations to the genes in which they reside by means of HUGO nomenclature committee approved definitions. These assignments are created by HGVbase curation work, submitted claims, Ensembl project coordinate information and direct high-stringency BLAST analyses. Some variations are presently associated only with anonymous EST sequences.
4. Source details (required): comprising all known sources of information pertaining to the variation, including, but not restricted to, source database (with a hypertext link to the database and source entry if possible), literature details (with a hypertext link to PubMed plus a list of MESH terms) and submitted information (including submitter name and contact details).
5. Intragenic location (optional): the functional site of a variation in its host gene is detailed wherever possible, e.g. exonic, intronic, coding sequence, 5′- or 3′-untranslated region, splice site and codon plus amino acid consequences (continually processing all HGVbase records to extend and update this level of annotation).
6. Validation status (required): as supplied by the data source, an indication of whether experimentally proven or merely suspected to exist. If at least one source claims the variant to be 'proven', it is so marked in HGVbase.
7. Allele frequency (optional): according to source definitions for any type of human 'population', this measure is presented in terms of percentage observation in a stated number of individuals.
8. Functional predictions (optional): all amino acid variants are currently being analyzed to predict functional significance based upon various aspects of protein structure and amino acid conservation.
9. Repeat element content (required): repeat element features of the variation region and its immediate flanking DNA are given to forewarn users of potential assay difficulties and experimental artifacts that may occur due to the non-uniqueness of the locus sequence.
10. Assay suggestions (optional): to aid laboratory study of the variation, assay suggestions are given comprising all possible enzymes for RFLP discrimination of the alleles (implemented), an assay design for genotyping by DASH (2) (due for complete incorporation by the end of 2001) and recommendations for any other assay platform that may be submitted to us for any or all HGVbase entries.

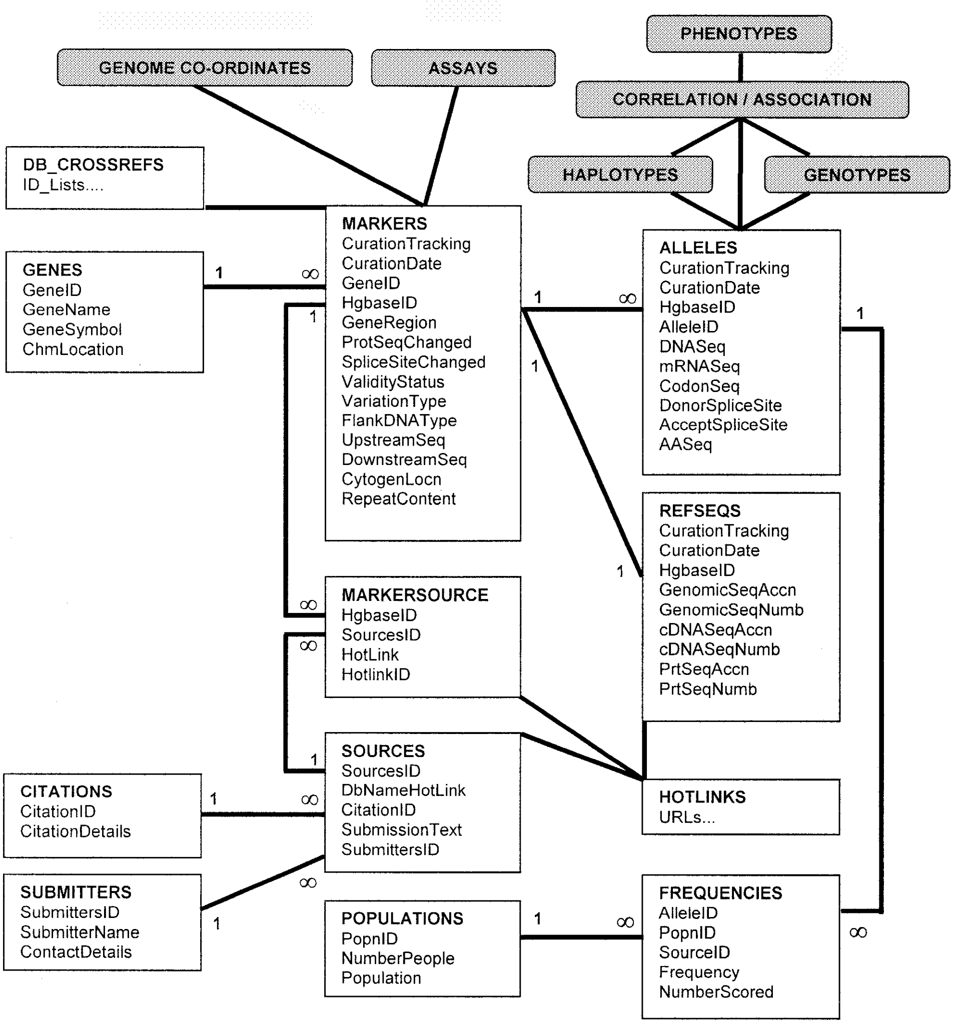

Figure 1. A summary view of the tables and fields that comprise HGVbase. An abbreviated set of key tables and component fields is shown as white boxes, with principal table joins represented. Complex joins are shown between box corners. In development extensions to this table structure are outlined above in shaded boxes, with join details not shown as they are yet to be optimized.

11. Marker coordinates (required): as from the end of 2001 chromosome and base pair coordinates shall be presented for each variation, these data having been extracted from the Golden Path human genome sequence map hosted by the Ensembl project where all HGVbase record are graphically represented. This new coordinate feature will enhance our ability to extract variation–gene relationships, and enable us to offer new positional and regional search functionality (see below).
12. Curation status (required): an indication of the extent of HGBASE staff curation and data checking that has been applied is given by the code letters M, R and A, indicating full Marker (locus), Reference sequence or Allele level curation respectively.

HGVbase development efforts recently established systems for incorporating unlimited haplotype classifications (groups of marker alleles and/or haplotypes) and genotype definitions (diploid genome representations of marker alleles or haplotypes), as shown in Figure 1. However, genotype archiving for lists of specific individuals will not be implemented in HGVbase. These developments reflect our expectation that genotype and haplotype data elements will soon become the principal functional units for the study of correlations between sequence variations and phenotypes on an increasingly large scale. It is anticipated that the first examples of haplotype data will appear in HGVbase in early 2002.

A particularly exciting aspect to the continued evolution of HGVbase is an agreement we have with researchers at the Bioinformatics Supercomputing Centre (BiSC) and members of the HUGO Mutation Database Initiative (MDI) (3) to try to capture extensive numbers of proven and likely phenotype altering mutations. Part of this will entail federating efforts with extant Locus Specific Mutation Databases (LSDBs). Additionally, our partners will establish a data processing 'WayStation' and undertake promotional campaigns that together encourage and simplify the submission of clinical mutations from diagnostic and research laboratories. HGVbase will then be the final 'Warehouse' where the information becomes integrated into the bigger picture of the genome variation via the systems we already have in place. To enable this, we and the BiSC team are now working together to create an effective and yet streamlined means for representing most clinical phenotypes of interest. We will also endeavor to create a flexible standard for mutation/polymorphism and phenotype description based on UML/XML.

SOFTWARE DETAILS

HGVbase is stored locally in a MySQL relational database running on a Digital UNIX server. Manual and semi-automated curation is performed through a series of Microsoft Access/ Visual Basic interfaces which are linked to the MySQL database. Tools for internal consistency checking, export, import and automated annotation procedures are implemented in Visual Basic and Perl. The online text search interface uses an SRS 6 (4) server, whilst sequence searches utilize Fasta3 (5) enabled at EBI.

Increasingly large volumes of data have prompted us to develop tools for fully automated annotation and curation, some of which are publicly available online through our web site. 'HNP Blast' performs sequence similarity searches of various GenBank sub-divisions, and extracts information on genes and gene structure from matching entries. The 'DNA Mutation Checker' has been created to verify the transcription and translation effects of a DNA level sequence variation. In line with our policy of running an open community project, these powerful tools are available online, and our general record curation software is available upon request.

ACKNOWLEDGEMENTS

We thank Thermo-Hybaid (Interactiva Division, Germany) for support during early development of the database and for transferring the project to the public domain. We also gratefully acknowledge the KI, EBI, EMBL, Pharmacia, Celera and Doubletwist for their unconditional financial and practical support of the HGVbase project.

REFERENCES

1. 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.
2. Prince,J.A., Feuk,L., Howell,W.M., Jobs,M., Emahazion,T., Blennow,K. and Brookes,A.J. (2001) Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): design criteria and assay validation. *Genome Res.*, **11**, 152–162.
3. Cotton,R.G., McKusick,V. and Scriver,C.R. (1998) The HUGO Mutation Database Initiative. *Science*, **279**, 10–11.
4. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
5. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.