HGBASE: a database of SNPs and other variations in and around human genes

Anthony J. Brookes[1,*], Heikki Lehväslaiho[2], Marianne Siegfried[1,3], Jana G. Boehm[3], Yan P. Yuan[4], Chandra M. Sarkar[3], Peer Bork[4] and Flavio Ortigao[3]

[1]Center for Genomics Research, Karolinska Institute, Theorells väg 3, S-171 77 Stockholm, Sweden, [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK, [3]Interactiva Biotechnologie GmbH, Sedanstrasse 10, D-89077 Ulm, Germany and [4]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Received August 24, 1999; Revised and Accepted October 18, 1999

ABSTRACT

Human genome polymorphism is expected to play a key role in defining the etiologic basis of phenotypic differences between individuals in aspects such as drug responses and common disease predisposition. Relevant functional DNA changes will probably be located in or near to transcribed sequences, and include many single nucleotide polymorphisms. To aid the future analysis of such genome variation, HGBASE (Human Genic Bi-Allelic SEquences) was constructed as a means to gather human gene-linked polymorphisms from all possible public sources, and show these as a non-redundant set of records in a standardized and user-friendly database endowed with text and sequence based search facilities. After 1 year of presence on the WWW, the HGBASE project has compiled data for over 22 000 records, and this number continues to triple every 6–12 months with data harvested or submitted from all major public genome databases and published literature from the previous decade. Extensive annotation enhancement, internal consistency checking and manual review of every record is undertaken to address potential errors and deficiencies sometimes present in the original source data. The fully polished and comprehensive database is made freely available to all at http://hgbase.cgr.ki.se

INTRODUCTION AND OVERVIEW

The HGBASE (Human Genic Bi-Allelic SEquences) database is an attempt to catalogue all known sequence variations [particularly single nucleotide polymorphisms (SNPs)] as a non-redundant set of records, and present each variant in the context of its physical relationship to the nearest human gene. The home page for HGBASE is reached via the address http://hgbase.cgr.ki.se . The represented data within HGBASE is a composite of all pertinent information available within the public domain, which we then enhance, standardise and check thoroughly for internal consistency, completeness and non-redundancy. Together, the comprehensive nature and high accuracy of the records in HGBASE are two of the principle features that will hopefully make the database a significant resource to the scientific community. It is hoped that the ready accessibility of such information will assist researchers that are trying to use experimental or computational methods to analyse the functional and phenotypic consequences of genome variation. Furthermore, the data in HGBASE may assist investigations into the global patterns and distributions of these gene located variations, impacting on basic questions of human history and natural selection on the individual allele, single gene and regional levels.

A primary motivation behind HGBASE is the widely held belief that gene polymorphisms will be central to our understanding of phenotypic variation in aspects such as common disease risk, drug responses and perhaps even some psychological traits (1). Against a back-drop of many different strategies and opinions about how best to advance this field, HGBASE was designed to support primarily the ‘candidate gene association study’ principle. Thus, HGBASE does not place a major emphasis upon highly penetrant disease ‘causing’ mutations (though these are not explicitly excluded), but instead catalogues human genome sequence variations of all common types (see below), and presents each along with its physical relationship to the human gene that it is either encompassed by or most closely positioned to. HGBASE records describe polymorphisms (sequence variations in which the least abundant allele has a frequency of ≥1%) and variations with rarer alleles; however, since allele frequencies can vary dramatically between populations, no formal distinction is made in HGBASE between polymorphisms and variations (nor subsequently in this text). The gross location of each variation relative to its ‘host’ gene is described by a simple code that categorizes those residing in exons, introns, and 5′ and 3′ gene flanking regions. These polymorphisms will include some alleles that in isolation may contribute towards a certain phenotype, as well as others that are neutral variations in close proximity to, and therefore often in high linkage disequilibrium with, unknown functionally pathogenic alleles of human genes. All of these markers would thus be appropriate starting points for gene based association studies.

HGBASE is a joint academic and industry initiative, founded on the principle that all gathered data will be provided free and without restriction to all. In this context, we would

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

encourage the scientific community to maintain the positive support for HGBASE that it has so far shown, by continuing with frequent submission of newly discovered/reported polymorphisms. Overall design and scientific responsibility for HGBASE, plus data collection and data processing, are undertaken by Anthony Brookes and colleagues at the Center for Genomics Research in the Karolinska Institute (Stockholm, Sweden). Programming developments that improve and automate the tasks of data processing and data enhancement are contributed by Yan P. Yuan and Peer Bork at EMBL (Heidelberg, Germany). The database design and data structure are the result of joint efforts by Anthony Brookes, Chandra Sarkar at Interactiva GmbH (Ulm, Germany), and Heikki Lehväslaiho at EBI (Hinxton, UK), and these aspects are constantly being evolved and updated. Primary database coding and implementation is provided by Chandra Sarkar (formerly), and Heikki Lehväslaiho (currently). Maintenance of the Web interfaces, database distribution, links to other sites of interest, and additional curation activities are taken care of by Flavio Ortigao and staff at Interactiva GmbH (Ulm, Germany).

Reflecting the general nature of human genome polymorphism, the majority of HGBASE records concern single base variants that for the most part comprise only two alleles, but insertion–deletion differences (indels), simple tandem repeats and other short-range rearrangements are also represented. Thus, the scope of the database is in reality wider than suggested by the term ‘bi-allelic’ in its name. Upon first release (August 1998) there were 2400 distinct polymorphisms in HGBASE. This total has now tripled, and will soon become almost tripled again to over 22 000 with the inclusion of records currently being processed. Based upon the scale of established SNP discovery efforts currently initiated worldwide, we would expect this exponential rate of increase in database size to continue and probably speed up for several years to come. Eventually, the number of coding sequence variants alone will be expected to plateau at ~200 000 and other variants should sum to several fold more than this.

DATABASE STRUCTURE AND ORGANIZATION

The core of HGBASE is a list of non-redundant polymorphism records, implemented as follows. Four categories of variation are defined: (i) single base differences, (ii) insertion–deletion variants, (iii) simple tandem repeat polymorphisms, and (iv) ‘generic’ (or complex) changes involving alterations not described by the preceding three alternatives. Polymorphisms for inclusion in HGBASE are considered to be equivalent, and are, therefore, combined into a single record if they are of the same category and affect the same base(s) in the identical gene, regardless of the precise allele details. For example, a newly submitted SNP involving a T–C change at base ‘N’ of gene ‘XYZ’ would be merged with an existing record of an SNP involving a T–G change at base ‘N’ of the ‘XYZ’ gene, so producing one HGBASE record with three alleles (T, C and G). Details of the two distinct information sources and any other submitted or newly acquired data would be jointly presented within this record, along with a unique and permanent HGBASE accession number by which the underlying polymorphism can always be referenced. This accession number is structured as a progressively increasing numeric with a three letter prefix (SNP, IND, MIC, GEN) that indicates the category to which the polymorphism belongs (a record property that is also indicated in a separate information field). In this way, a concise and non-redundant catalog is maintained, simplifying tasks of data extraction and subsequent experimental planning by users of the database. The HGBASE accession number is therefore a suitable reference number for use in research communications to identify specific polymorphisms represented in HGBASE. As such, HGBASE accession numbers are immediately allocated to all newly received submissions and this information is passed directly back to the data submitter for possible use in manuscript preparation.

The additional information specified for each record is as follows:

1) DNA sequences comprising 25 bp 5′ of the polymorphism, the allelic bases themselves, and 25 bp 3′ of the polymorphism—all sequences shown in the same orientation as the direction of transcription. Currently this is declared as either genomic or cDNA sequence, but both will be included once the full human genome sequence becomes publicly available.
2) The HUGO nomenclature committee approved name and symbol for the host gene. When the host gene is presently known only as an anonymous Expressed Sequence Tag (EST) or computationally predicted gene, then this is stated and no name or symbol is given.
3) A DDBJ/EMBL/GenBank accession number for at least one nucleotide reference sequence, with specification of the residues therein that are altered by the polymorphism. Where possible, cDNA level and genomic DNA level references are provided, and in many cases an accession number and residue position for a reference protein sequence file (SWISS-PROT) may also be given. Since most gene structures and coding domains are still far from completely defined, no attempt is made to use any formalised or standardised naming system for polymorphisms represented in HGBASE. Instead, unambiguous and doubly foolproof polymorphic base specification is achieved by providing (i) the gene name/symbol plus 25 bp 5′ and 25 bp 3′ flanking sequences (effective since within any one stated gene the given ≥50 base string surrounding each variation is highly likely to be unique), and (ii) the numbered polymorphic base(s) in a reference DNA sequence file (effective since each given DDBJ/EMBL/GenBank accession plus version number combination will indicate an unequivocal position in a definitive sequence).
4) Indication of all known sources of the polymorphism, comprising a standardised comment (e.g., database, literature, *in silico*) as well as a detailed citation/pointer to each information source.
5) The intra-genic location of the polymorphism (e.g., exon, intron, coding sequence, 3′ untranslated region), and details of any predictable or known consequences thereof (e.g., codon and deduced amino-acid changes, splice site modifications, altered transcription factor binding sites).
6) Indication as to whether the variant is experimentally proven or merely suspected to exist, and why—in each case this judgement is required to be made by the data submitter (or copied from the data source) based upon their own criteria, and further details in specific cases must be sought from the original data source. A free text comments box is provided for the submitter to expand on this if they wish to do so.
7) Allele frequency for any number of ‘populations’ (as defined by the researchers who made the frequency determinations), plus

for each determination, the number of individuals studied (so that the reliability of the given frequency can be estimated).

8) A free text comments box for any additional information not covered by the above.

For database organisation we have established a two-level system. Local storage and handling of individual records is performed using the MS Access relational database tool. Purpose built scripts are then used to transfer data to either a different relational database platform (Oracle Server 8) or a simple flat-file format, which together provide inputs to the various Web Page interfaces. This arrangement is designed to allow convenient implementation of custom interface programs for advanced tasks like a Java bulk submission program. Final data presentation when viewed by the user is always given as simple text flat-files for easy download.

DATA COLLECTION AND SUBMISSION

The data content of HGBASE is designed to be a comprehensive summary of all known human gene related polymorphisms that are in the public domain. Clearly, with the state of such knowledge advancing so rapidly at the present time, no single database can ever be completely up to date in this undertaking. However, our experience in 2 years of HGBASE data collection is that, without exception, every public resource we have approached for permission to include their data has been both willing and, more often than not, proactive in helping us efficiently harvest and process their publicly released and sometimes unpublished polymorphism records. This impressive generosity on the part of the scientific community ensures that the time delay between new data appearing elsewhere and a representation of it being included in HGBASE is kept to a minimum. We are, of course, most grateful to all the individuals that have helped us in this regard, and believe it reflects a healthy recognition that genome polymorphism is both a research tool and part of our shared human heritage that should not be treated as a commercial entity in its own right.

Submission of data to HGBASE may be done in one of several ways. For small-scale submissions (one or a few polymorphisms) Web Page forms are provided that can be completed and dispatched directly. Medium-scale submissions (from one to several tens of polymorphisms) are more easily handled by entering the complete dataset into a MS Excel sheet which may be downloaded with instructions from the HGBASE Web Pages. The completed Excel sheet is then simply emailed to the database curators. Finally, for larger data-sets, any convenient data format may be supplied, and the HGBASE curators will then work to re-structure this, perhaps with guidance from the original submitter. The minimum amount of information that must be supplied when submitting a polymorphism includes (i) the allelic sequences plus 25 bp 5′ and 25 bp 3′—all in the direction of gene transcription, and declared as genomic or cDNA, (ii) reference DDBJ/EMBL/GenBank cDNA and/or genomic sequence files, plus the base numbers in those files that are altered by the polymorphism, and (iii) a submitter name plus contact details. All HGBASE information fields in addition to these are optional but are preferably supplied. When the additional information fields are not completed for any submission, these data are entered by database curators. Information updates to any and all records may be made by anyone by means of the same submission mechanisms described above.

HGBASE is operated under a policy that no claim whatsoever is made to the ownership of any data submitted by others or harvested by us from independent sources. In line with this, all records and subsequent updates carry details of the various data ‘sources’. These sections of each record include details of the individual submitter who provided the information (and whom should be contacted for further details of its mode of initial discovery) or the public sources from where data was acquired (typically literature references and/or Web Site addresses plus reference IDs relevant to those databases).

When polymorphisms are submitted, or when data are copied from literature or other public resources by HGBASE curators, our general finding is that they carry a variable, but sometimes disturbingly high, level of inaccuracy. For example, exonic bases may be stated to be intronic, codon-deduced and claimed amino-acid changes may be in conflict, or the wrong bases could be indicated in a referenced nucleic acid sequence file. Therefore all details that can be checked manually or computationally via Web resources (starting from just the submitted sequences) are re-determined. This is undertaken by a team of database curators using purpose-built software tools and Web-based resources in a structured data processing regime which involves a considerable degree of manual review of every record. Gaps and inaccuracies in the submitted information are thereby replaced by higher accuracy data. Finally, semi-automated internal consistency checks are performed upon each record before the information is eventually released to the public in 2–3 monthly batches as highly polished and richly annotated records.

DATABASE CONTENT

In the autumn of 1999 HGBASE held ~7000 distinct records (with a further 15 000 being processed at that time and due for inclusion in the database before January 2000), made up of the following:

- 98.6% SNPs, 1.2% indels, 0.1% simple tandem repeats, 0.1% other variants
- 87% in exons, 9% in introns, 4% in 5′ or 3′ flanking domains
- 27.5% in known coding regions (50% of these are non-synonymous)

To date, the sources from which the HGBASE content has been derived include the following (note: totals exceed the number of records in HGBASE due to multiple ascertainment or supply of some polymorphisms).

1) 247 unpublished polymorphisms from in-house discovery efforts of Drs Brookes and Ortigao, based upon 910 sequence-unique human genes for which complete cDNAs were available in 1998.

2) 1226 polymorphisms extracted from the total scientific literature of the past decade. This exercise, undertaken by Drs Brookes and Ortigao over a 2-year period, is estimated to have identified ~80% of all published intra-genic polymorphisms reported with sufficient information to allow definition of the allele sequences. However, this activity is unsustainable in the future due to the ever-increasing rate at which polymorphisms are being described. To ensure that all this emerging information does not become lost to the bio-informatics era, we implore all genetics journals to

adopt a policy similar to that typically applied to DNA sequences, whereby all polymorphisms have to be submitted to a database and assigned an accession number (HGBASE or other) before literature publication will be accepted.

3) 221 polymorphisms discovered and submitted in sets of one or a few at a time by independent researchers and directly contributed to HGBASE without solicitation.

4) 1062 polymorphisms extracted from various Web resources such as sequence catalogues and gene specific mutation databases. These include the PAH gene database (http://www.mcgill.ca/pahdb/) maintained by Charles Scriver (Québec, Canada) *et al.*; the HNPCC gene database (http://www.nfdht.nl/) maintained by Hans Vasen (Leiden, The Netherlands) *et al.*; the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/) maintained by Steve Sherry (Bethesda, USA) *et al.*; the AR gene database (http://www.mcgill.ca/androgendb/data.htm) maintained by Bruce Gottlieb (Quebec, Canada) *et al.*; the ATM database (http://www.vmresearch.org/atm.htm) maintained by Pat Concannon (Seattle, USA) *et al.*; the Bic database (http://www.nhgri.nih.gov/Intramural_research/Lab_transfer/Bic) maintained by Barb Weber (Philadelphia, USA) *et al.*; the CF database (http://www.genet.sickkids.on.ca/cftr/) maintained by Lap-Chee Tsui (Ontario, Canada) *et al.*; the NCL database (http://www.ucl.ac.uk/ncl/) maintained by Sara Mole (London, UK) *et al.*; the COL3A1 database (http://www.le.ac.uk/genetics/collagen/) maintained by Raymond Dalgleish (Leicester, UK) *et al.*; the Muscular Dystrophy database (http://www.dmd.nl/) maintained by Johan den Dunnen (Leiden, The Netherlands) *et al.*; the Fanconi Anemia database (http://www.rockefeller. edu/fanconi/mutate/) maintained by Peter Verlander (New York, USA) *et al.*; the GSDII database (http://www.eur.nl/FGG/CH1/pompe/index.htm) maintained by A. J. Reuser (Rotterdam, The Netherlands) *et al.*; the OTC gene database (http://www.peds.umn.edu/otc/) maintained by Norma Allewell (Minnesota, USA) *et al.*; the p53 gene database (http://perso.curie.fr/Thierry.Soussi/) maintained by Thierry Soussi (Paris, France) *et al.*; the VWF gene database (http://mmg2.im.med.umich.edu/vWF/) maintained by D. Ginsburg (Ann Arbor, USA) *et al.*; the WRN database (http://www.pathology.washington.edu/werner/ws_wrn.html) maintained by Michael Moser (Seattle, USA) *et al.*; the Albinism database (http://www.cbc.umn.edu/tad/) maintained by Richard King and Gail Summers (Minnesota, USA) *et al.*; the CANVAS database (http://ifr69.vjf.inserm.fr/~canvas/) maintained by François Cambien (Paris, France) *et al.*

5) 4423 polymorphisms (with 15 000 more currently being processed) provided after solicitation by researchers involved in large-scale SNP discovery and/or database projects, including Eric Lander (2) (Human SNP Database: Whitehead Institute for Biomedical Research, USA), Michael Krawczak (3) (Human Gene Mutation Database: Institute of Medical Genetics, UK), Leslie Picoult-Newberg (4) (Orchid Biocomputer, Inc.: Baltimore, USA), Kenneth Buetow (5) (Cancer Genome Anatomy Project, Genetic Annotation Initiative: National Cancer Institute, USA), Michele Cargill (6) (Whitehead Institute for Biomedical Research, USA), Aravinda Chakravarti (7) (Case Western Reserve University, USA).

HGBASE USAGE AND ACCESS

Usage of the HGBASE database has increased steadily in its first year of life, from an initial 12 000 hits per month to over 20 000 hits per month. Throughout this period, access to HGBASE data was free to all users with no restrictions, and this open access policy will be maintained in the future. A new feature that is due to be installed in late 1999 is the potential to download the complete database, with or without all update alterations included, as a simple text flat-file.

Three search modes have been implemented to allow users to interrogate HGBASE records. Two are based upon text. One of these is a ‘simple search’ tool, the use of which requires one to enter a single keyword to query all data except the sequences in HGBASE. The other is the Sequence Retrieval System (SRS), which offers users the possibility of making more complex text based searches via an interface that is becoming familiar to users of many life sciences databases (8). However, text searches are intrinsically limited by the lack of complete standardisation of gene naming and the fact that most transcribed elements have yet to be named (presently represented simply as ESTs in HGBASE, and yet comprising a wealth of polymorphism data). Therefore, the third search option allows one to begin with a test DNA sequences for the gene of interest (up to 10 kb of genomic or cDNA sequence) and use this to interrogate the allele and flanking DNA sequences represented in HGBASE. This is achieved via an interface to NCBI’s BLAST DNA sequence alignment algorithm (9). Both the SRS and BLAST search methods get their input automatically from data exports out of the Oracle database and/or flat-file copies of the database.

ONGOING DEVELOPMENTS

Collaborations are being pursued to maximally exploit HGBASE content for the scientific community. Heikki Lehväslaiho has worked to make HGBASE content available for searching via the SRS based multi-database interrogation system they have established at the European Bioinformatics Institute (8), and in this context many other database links are automatically provided. Similarly, it has been agreed with Richard Mural (Oak Ridge National Laboratory, USA) to include HGBASE data in the Genome Channel database (10), a deeply annotated graphical representation of the emerging human genome sequence map.

Enhancements to HGBASE content and structure are also under development. New data fields are planned for inclusion within the next 12 months. These will give information on (i) the chromosome map position of each polymorphism, along with an appropriate search tool, (ii) links to expression pattern and disease relationship data in other databases for each of the host genes, and (iii) suggested assay conditions for RFLP-PCR and DASH assay (11) scoring of the polymorphisms in genomic DNA.

Continued effort is being made to establish more rapid, automated and precise data processing procedures, in order to retain the high level of data accuracy in HGBASE whilst keeping pace with the accelerating rate of global SNP discovery. We shall endeavour to maintain as much manual review of records as is reasonably possible, since we find this is key to pruning out the myriad of inconsistencies and nonstandard record details and that are endemic in many data sources. Our principle objective is to build an extremely accurate, high utility and ultimately fully comprehensive catalogue of normal human gene variation, that will be useful as a principle tool to aid the study of the bewildering range of human phenotypic variation.

REFERENCES

1. Brookes,A.J. (1999) *Gene*, **234**, 177–186.
2. Wang,D.G., Fan,J.B., Siao,C.J., Berno,A., Young,P., Sapolsky,R., Ghandour,G., Perkins,N., Winchester,E., Spencer,J., Kruglyak,L., Stein,L., Hsie,L., Topaloglou,T., Hubbell,E., Robinson,E., Mittmann,M., Morris,M.S., Shen,N., Kilburn,D., Rioux,J., Nusbaum,C., Rozen,S., Hudson,T.J., Lander,E.S., *et al.* (1998) *Science*, **280**, 1077–1082.
3. Cooper,D.N., Ball,E.V. and Krawczak,M. (1998) *Nucleic Acids Res.*, **26**, 285–287.
4. Picoult-Newberg,L., Ideker,T.E., Pohl,M.G., Taylor,S.L., Donaldson,M.A., Nickerson,D.A. and Boyce-Jacino,M (1999) *Genome Res.*, **9**, 167–174.
5. Buetow,K.H., Edmonson,M.N. and Cassidy,A.B. (1999) *Nature Genet.*, **21**, 323–325.
6. Cargill,M., Altshuler,D., Ireland,J., Sklar,P., Ardlie,K., Patil,N., Lane,C.R., Lim,E.P., Kalayanaraman,N., Nemesh,J., Ziaugra,L., Friedland,L., Rolfe,A., Warrington,J., Lipshutz,R., Daley,G.Q. and Lander,E.S. (1999) *Nature Genet.*, **22**, 231–238.
7. Halushka,M.K., Fan,J.B., Bentley,K., Hsie,L., Shen,N., Weder,A., Cooper,R., Lipshutz,R. and Chakravarti,A (1999) *Nature Genet.*, **22**, 239–247.
8. Etzold,T., Ulyanov,A. and Argos,P. (1996) *Methods Enzymol.*, **266**, 114–128.
9. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
10. Mural,R.J., Parang,M., Shah,M., Snoddy,J. and Uberbacher,E.C. (1999) *Trends Genet.*, **15**, 38–39.
11. Howell,W.M., Jobs,M., Gyllensten,U. and Brookes,A.J. (1999) *Nature Biotechnol.*, **17**, 87–88.