RESEARCH NOTE

Human Genetic Bi-allelic Sequences (HGBASE), a Database of Intra-genic Polymorphisms

Chandra Sarkar/+ , Flavio R Ortigão, Ulf Gyllensten*/, Anthony J Brookes****

Interactiva Biotechnologie GmbH, D-89077 Ulm, Germany *Swedish Genome Research Center **Department of Genetics and Pathology, Biomedical Center, Uppsala, Sweden

Key words: single nucleotide polymorphisms - polymorphisms - intra-genic polymorphisms - databases - bioinformatics

The Human Genome Project is providing a wealth of information about the human gene repertoire, and promises to furnish a complete genome sequence (and thereby a complete gene catalog) by the year 2005. This enormous output of data is beginning to be complemented by large scale studies designed to uncover normally occurring variations within human gene sequences. Much of this variability is very subtle, often comprises single nucleotide polymorphisms (SNPs) which are ideally compatible with a number of large scale detection procedures. SNPs will be the basis of future highly dense polymorphic marker maps, and those related to known genes can be exploited in genetic association studies aimed at defining the genetic basis of all manner of complex phenotypes, not least disorders such as mental illness, diabetes, cardiovascular disease and cancer. All indications are that 100,000-200,000 human genome SNPs will be identified within the next two years.

In light of the above developments, a database of gene based polymorphisms is obviously required. To fulfill this need we have constructed and recently released at http://hgbase.interactiva.de the HGBASE (human genic bi-allelic sequences) database of intra-genic sequence polymorphism. HGBASE is the result of a joint venture between Uppsala University Medical Genetics Department, the Swedish Genome Research Centre, and Interactiva Biotechnologie GmbH. Its primary purpose is to facilitate genotype-phenotype association studies based upon the rapidly growing number of known, gene related, single nucleotide polymorphisms (SNPs) and other intra-genic sequence variations. Furthermore, HGBASE will help towards the production of a dense SNP map of the human genome, which itself will be a valuable research tool.

HGBASE is not designed to include gene 'mutations', but instead is a catalog of intra-genic (promoter to transcription end point) sequence variants found in 'normal' individuals. Although the distinction between 'mutation' and 'variation' can be somewhat blurred, the general idea is that the content of HGBASE concerns frequently occurring 'normal polymorphisms', whether or not they are suspected to increase the risk of developing a particular phenotype. This is in contrast to 'mutant sequences' which are known to cause genetic disease. Despite its name, HGBASE contains all types of intra-genic variation and is not limited to bi-allelic polymorphisms (though these do represent most of the database content). Both functional polymorphisms (e.g. promoter and non-silent codon changes) and non-functional polymorphisms (e.g. intron sequence differences) are included. This is for two reasons. Firstly, it is often difficult to be certain about the functional consequence of a variation. Secondly, regardless of functional relevance, any intra-genic polymorphism can usually be employed as an effective surrogate marker for an unknown functional variant in an association study, due to close proximity and linkage disequilibrium.

Gene polymorphisms may be retrieved from HGBASE by using the database search facilities to query either by a text string or by a DNA sequence. Data submission to HGBASE is made simple by provision of a series of Web page data submission forms. All submitted data is made available to any other public database that wishes to download it, and continual efforts are made to access new relevant data from other databases and literature publications. The exponential growth in polymorphism discovery requires that scientists make every effort to submit their data to the HGBASE database to ensure it remains up to date. HGBASE does not claim any rights to publicly available or submitted data, instead this remains the property of the original submitter. Deposition of data into HGBASE requires only the allelic DNA sequences, the allele frequencies, the host gene name, and the intra-genic domain. Additional com-

+Corresponding author. Fax: +49-731-93579291. E-mail: sarkar@interactiva.de
Received 15 June 1998
Accepted 30 July 1998

ments, such as assay conditions, can be supplied though are not required. The submitted data is presented in HGBASE along with the submitters name and contact details to aid discussion and questioning. Database curators will subsequently enhance the submitted data by adding links to other databases, and by adding information concerning gene function, gene location, gene expression pattern, disease associations, and suggested assay formats. This 'added value' data is accessible to users following a simple registration procedure that is free to academia but for which a charge is made to industry to cover the costs of collecting and maintaining the additional data.