Guest Editorial

The Human Gene Mutation Database: Providing a comprehensive central mutation database for molecular diagnostics and personalised genomics

Reading the recent Editorial in the Journal by Richard Cotton,[1] which publicised the proposed new Human Variome Project, one could be forgiven for thinking that no central repository for inherited human gene mutations of pathological significance currently exists. In practice, however, the Human Gene Mutation Database (HGMD®; http://www.hgmd.org)[2] already constitutes a comprehensive collection of single base-pair substitutions in coding (missense and nonsense), regulatory and splicing-relevant regions of human nuclear genes, micro-deletions and micro-insertions, indels, repeat expansions, as well as gross gene lesions (deletions, insertions and duplications) and complex gene rearrangements. This unique resource currently contains in excess of 96,000 different germline mutations and disease-associated/functional polymorphisms—in a total of over 3,600 nuclear genes (December 2009 release)—causing or associated with human inherited disease.

HGMD currently provides free access to the bulk of its mutation data to over 30,000 registered academic/non-profit users worldwide. In the absence of any public funding, HGMD is maintained courtesy of a subscription-based version (HGMD Professional), distributed through BIOBASE GmbH (http://www.biobase-international.com). HGMD Professional not only provides access to the very latest mutation data, but also contains valuable extra features, including an expanded search engine, genomic coordinates, additional literature references, Human Genome Variation Society (HGVS) nomenclature (http://www.hgvs.org/mutnomen) and a suite of advanced search tools that greatly enhance the utility of the database. Together with BIOBASE, we are working toward being able to make all HGMD data and search tools available to the academic community free of charge and in a timely fashion, with the costs of upkeep being borne primarily by industry and commerce. We believe that this funding model should not only guarantee the financial viability of HGMD, but also allow this unique resource to be sustainable into the long term, to the benefit of the scientific community.

Although HGMD has now become the *de facto* central disease-associated mutation database, there are several other valuable sources of human mutation data available. Online Mendelian Inheritance in Man (OMIM;[3] http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim) is a formidable genotype–phenotype knowledge base, but it is not comprehensive with respect to mutation data on account of its policy of providing only specimen examples of allelic variants (in a total of 2,328 human genes). In addition to inherited pathological mutations, it should be noted that OMIM also contains some somatic lesions, neutral

polymorphisms, disease-associated single nucleotide polymorphism (SNP) haplotypes and data from genome-wide association studies (GWAS), none of which are included in HGMD. The HGVS website (http://www.hgvs.org/dblist/glsdb.html) provides a list of URLs for the ~670 internet-accessible locus-specific mutation databases (LSDBs). Although curated by experts in specific genes/proteins, these databases cannot be screened in combination and, even if they could, they would only represent a fraction of the mutational lesions listed in HGMD. Moreover, as Cotton[1] himself readily admits, many of these LSDBs are incomplete, out of date, inaccessible, moribund or otherwise non-functional. The National Center for Biotechnology Information (NCBI)'s Single Nucleotide Polymorphism Database (dbSNP; http://www.ncbi.nlm.nih.gov/projects/SNP/) has recently begun to include variants of clinical significance from OMIM, from some LSDBs and by direct submission, thereby extending its original remit of listing only human polymorphisms. These variants are currently listed under 'Clinical/LSDB variation'. Although dbSNP currently only contains ~8,669 such entries, it is anticipated that this number will steadily increase in the future. For now, however, utility is somewhat limited owing to the difficulty in accessing phenotypic information. Finally, the Pharmacogenomics Knowledge Base (PharmGKB;[4] http://www.pharmgkb.org) displays some overlap with HGMD's own ~4,600 functional and disease-associated polymorphism entries (http://www.functional-polymorphism.org).

PharmGKB contains only 1,575 annotated SNPs, however, and much of the data in this knowledgebase would appear to relate to GWAS data and human genetic variants that confer differential responsiveness to drugs, neither of which are covered by HGMD. Complementing the above germline mutation databases are those that focus exclusively on somatic mutations, usually in association with tumorigenesis. The most comprehensive of these is the Catalogue of Somatic Mutations in Cancer (COSMIC;[5] http://www.sanger.ac.uk/genetics/CGP/cosmic/), which currently (v. 43) contains over 88,000 mutations in 2,927 different genes (including 290 known cancer genes) and which represents the somatic equivalent of HGMD.

The above resources notwithstanding, HGMD Professional remains the only comprehensive database of germline mutations in nuclear genes underlying or associated with human inherited disease. It can be used to search for newly identified gene lesions to determine whether or not they are novel. It can be searched on a gene-wise basis to obtain an overview of the known mutational spectrum for a given gene (via a dynamic mutation viewer which depicts coding region mutations superimposed on the cDNA sequence of a gene). It can also be searched for other examples of a specific type of mutation in a specific location (eg at position +5 to a donor splice site) in order to garner evidence for the pathological authenticity of a given lesion.

The 'advanced search' facility, available as part of HGMD Professional, is a suite of software tools which has been designed to enhance mutation searching, viewing and retrieval. Currently, two of the main types of mutation listed in HGMD (single nucleotide substitutions and small micro-lesions, accounting for >90 per cent of all entries) may be interrogated with this toolset. The datasets for each mutation type may be readily combined (eg micro-deletions, micro-insertions and indels) in order to allow more powerful searching across comparable types of mutation. When utilising the advanced search, users may tailor their queries with more specific criteria, including amino acid exchange, nucleotide substitution, micro-deletion/insertion/indel size and composition, motif searching (both created and abolished), dbSNP number and article title/abstract keywords. The Mutation Interpretation Software offered by Alamut (http://www.interactive-biosoftware.com/alamut.html) is similar to some of the tools provided by HGMD Professional; the utility of the Alamut software will soon be greatly enhanced by the provision of links to the mutation data present in HGMD Professional.

HGMD/HGMD Professional will be continually improved by the inclusion of novel mutation data and new informational or software features.

Indeed, the provision of further supplementary information—including additional clinical phenotypes observed with a given mutation, fully annotated genomic sequences for all HGMD genes, genomic coordinates for as many mutations as possible, multiple additional references for each mutation, gene and disease ontologies and *in vitro* characterisation data—is already well underway. Further developments in the pipeline include tools for the *in silico* annotation of mutational information and the curation of *in vitro*/*in vivo* functional data to allow the identification of specific types of gene lesion from a functional standpoint. This should allow the rapid identification of, for example: all characterised/predicted mutations in microRNA binding sites, exon splice enhancers or polyadenylation sites; or all characterised/predicted mutations giving rise to the gain or loss of a glycosylation or phosphorylation site in the protein product; or (courtesy of BIOBASE's transcription factor database, TRANSFAC®; http://www.biobase-international.com/index.php?id=transfac[6]) the gain or loss of a given transcription factor binding site within gene regulatory regions.

So, the question therefore arises as to whether HGMD can be dovetailed into the Human Variome Project. The answer, unfortunately, is: 'Not easily', given the current thinking of the project organisers. Many of the objectives of the Human Variome Project[7] are certainly laudable. Plans floated for a consortium of up to 100,000 members plus at least 2,500 centrally coordinated experts, presumably responsible for at least ten genes each (http://www.humanvariomeproject.org), would, however, if put into practice, be cumbersome to administer and could doom the entire venture to failure through sheer inertia. Such an initiative would require control and coordination on a scale that has never before been attempted in the biomedical sciences. Further, even if such a project could be initiated and maintained on such a scale, it is most unlikely that it would be cost-effective. At a projected (and fairly conservative) $2,000 per expert per annum,[8] it would still cost upwards of $5 million per annum on a continuing basis, before even considering the additional costs inherent in organising, coordinating, monitoring and remunerating this number of experts. By contrast, HGMD has already managed to build a central mutation database for a small fraction of this proposed budget and is financially self-sustaining.

The first of the ten stated 'key objectives' of the Human Variome Project is to 'capture and archive all human gene variation associated with human disease in a central location'.[7] We urge the organisers of the Human Variome Project to abandon any plans they might have to coordinate centrally the various (largely independent and dissimilarly funded) existing entities such as dbSNP, European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk), GEN2PHEN (http://www.gen2phen.org; one of whose partners is BIOBASE), University of California Santa Cruz (UCSC) (http://genome.ucsc.edu), HGMD and OMIM[1] in favour of a 'light touch' umbrella role. Too much emphasis on central control is likely to inflate administration, stifle initiative and hinder progress. Integration would, in our view, be best facilitated if this process were allowed to evolve naturally, with individual components of the potential whole finding common cause through the establishment of links and the formation of ever-closer working relationships and partnerships. We believe that what is needed is a little less emphasis on regulation and regimentation, and rather more on consultation, inclusion, facilitation and 'incentivisation'. By presiding over a confederation of richly linked yet operationally independent information resources, a decentralised Human Variome Project could free itself up to fulfil the umbrella role through the promotion of links between publicly funded initiatives such as dbSNP, EBI, GEN2PHEN and OMIM, and by supporting established initiatives such as HGMD and the LSDBs, which have never been in receipt of any public funding. Thus, simply by adopting a more federated model, the Human Variome Project would ensure that all funds raised could be used effectively and efficiently to promote the integration of existing resources, thereby assisting what already works well.

Peter D. Stenson, Edward V. Ball, Katy Howells, Andrew D. Phillips, Matthew Mort and David N. Cooper
Institute of Medical Genetics
Cardiff University, Heath Park
Cardiff CF14 4XN, UK
Tel: +44 2920 744062;
E-mail: CooperDN@cardiff.ac.uk

References

1. Cotton, R.G.H. (2009), 'Collection of variation causing disease – The Human Variome Project', *Hum. Genomics* Vol. 3, pp. 301–303.
2. Stenson, P.D., Mort, M., Ball, E.V., Howells, K. *et al.* (2009), 'The Human Gene Mutation Database: 2008 update', *Genome Med.* Vol. 1, p.13.
3. Amberger, J., Bocchini, C.A., Scott, A.F. and Hamosh, A. (2009), 'McKusick's Online Mendelian Inheritance in Man (OMIM)', *Nucleic Acids Res.* Vol. 37 (Database issue), pp. D793–D796.
4. Sangkuhl, K., Berlin, D.S., Altman, R.B. and Klein, T.E. (2008), 'PharmGKB: Understanding the effects of individual genetic variants', *Drug. Metab. Rev.* Vol. 40, pp. 539–551.
5. Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E. *et al.* (2008), 'The Catalogue of Somatic Mutations in Cancer (COSMIC)', *Curr. Protoc. Hum. Genet.* Chapter 10, Unit 10.11.
6. Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I. *et al.* (2006), 'TRANSFAC and its module TRANSCompel: Transcriptional gene regulation in eukaryotes', *Nucleic Acids Res.* Vol. 34 (Database issue), pp. D108–D110.
7. Kaput, J., Cotton, R.G., Hardman, L., Watson, M. *et al.* (2009), 'Planning the Human Variome Project: The Spain Report', *Hum. Mutat.* Vol. 30, pp. 496–510.
8. Cotton, R.G.H., Phillips, K. and Horaitis, O. (2007), 'A survey of locus-specific database curation. Human Genome Variation Society', *J. Med. Genet.* Vol. 44, p. e72.