Guest Editorial

Collection of variation causing disease – The Human Variome Project

Collection of mutations (defined as variation causing disease in this paper) causing disease began soon after the cause of thalassaemia as a mutation in the β-globin gene was established.[1] As other disease genes were defined, collection of individual mutations and their effects increased dramatically and it is still increasing today. The reason for this collection was/is for research, clinical guidance and diagnostic strategies in specific diseases. These collections have been made by expert and dedicated workers world- and disease-wide for their own purposes. Some collection has been occurring centrally, with mutations in all genes involved; for example, Online Mendelian Inheritance in Man (OMIM; http://www.ncbi.nlm.nih.gov/omim/[2]), Human Gene Mutation Database (HGMD; http://www.hgmd.cf.ac.uk/[3]) and MutDB (http://mutdb.org/[4]).

Collection of mutation data in individual genes into databases (locus-specific databases [LSDBs]) is difficult to fund and is rarely funded.[5] Thus, the fact that so many exist[6] indicates they are needed. These 'curators' consequently usually work in isolation in their spare time, using a variety of software packages, with data consequently being out of date, or, sometimes, LSDBs are removed from the internet. Other problems besides these are lack of incentives to submit data, lack of standard methods, lack of coordination, large amounts of data in clinics not available to those who need it and lack of secure repositories, among others.

Attempts to raise the profile of this activity and its funding began in the mid-1990s with the formation of the HUGO-Mutation Database Initiative,[7] which developed into the Human Genome Variation Society (HGVS; http://www.hgvs.org/). This organisation developed standards and provided advice and some incentives, and generated more LSDBs. To raise the profile further, the concept of the Human Variome Project (HVP; http://www.humanvariomeproject.org/[8]) was developed. A meeting, co-sponsored by the World Health Organization and the American College of Medical Genetics (ACMG), was held in Melbourne, involving the United Nations Educational, Scientific and Cultural Organisation (UNESCO), Organisation for Economic Co-operation and Development (OECD), top genetics organisations and representatives from over 30 countries.[9,10] The 96 recommendations published[10] indicated the poor state of the field. A further meeting was held in May 2008 and the plans developed from this were recently published.[11] Both these meetings and fora organised by HVP, have stimulated much useful activity (see below).

In general terms, it is clear that the genomics/genetics community is working towards not only annotating regulatory regimes and other features in the Encyclopedia of DNA Elements (ENCODE) project[12] (http://www.genome.gov/10005107), but also, ultimately, in five to ten years, towards having a standard reference sequence numbered from 1 to 3×10^9. Each base may have a significance, and this significance will be accessed by clicking on the base. The single nucleotide polymorphism, HapMap, 1000 Genomes Project and copy number variations (CNV) consortia will provide many variations and some will be linked to common disease or pharmacological utility. It is important to realise that the vast majority have no effect on humans or

will have no recorded phenotype. On the other hand, data derived from inherited disease *will* have phenotype and will be useful not only in the clinical sphere, but also in the 90 per cent in general of human disorder outside the 10 per cent (average) of all diseases inherited in a Mendelian manner. The number of mutations causing inherited disease, at least within or near the coding sequence, is estimated to be between 200,000 to 20 million (20,000 genes × 10–1,000 mutations per gene). The variations collected by the major consortia are assured of being in databases, as such deposition is part of the funding conditions, including for the now ubiquitous Genome-Wide Association Studies (GWAS) sphere in the general area of translational medicine. In fact, there is a huge chasm in understanding and funding between the common disease consortia and the necessarily fragmented inherited disease world. The HVP hopes to enhance the funding and the recognition of this latter field.

One target in the field of high-throughput sequencing activity is personal genomics. Currently, this is focused on potential drug dosing and the elucidation of any small risk increases if a particular variation is present. Those receiving their own sequence, however, will need to know the significance of, say, changes in the breast cancer antigen or colon cancer genes. This is where a list of pathogenic mutations pursued by the HVP will be needed. Further, two individuals planning to have children will be able to match their sequence to see if they each have damaging mutation in any one of their genes which might cause recessive disease. If so, pre-implantation/prenatal counselling will be indicated in an analogous way to that for thalassaemia in Greece.

So how is the project 'to collect all mutations in all genes from all countries' progressing? This seems to be a massive task but because the data are needed, it needs to be, and will be, done. It is simply a case of working out how, and spreading the load to perhaps 10,000 people in, say, 100 countries.

There are two critical pilot studies under way which will provide guidance for the whole project.

The HVP/InSiGHT inherited colon cancer pilot

Those dealing with patients with inherited colon cancer, like any inherited disease, need lists of mutations and their effects when considering what to tell a patient and how to treat them. InSiGHT (http://www.insight-group.org) is a group that has been working together for many years, but the databases they created were incomplete and the software for each was different. Four databases have now been placed on the Leiden Open (source) Variation Database (LOVD) software,[13] developed over many years by HGVS members and others — particularly Drs den Dunnen and Fokkema. This database now includes the databases of published mutations, unpublished mutations from laboratories, *in vitro* experiments and pathology. Data from countries around the world now include those from Canada, China, Germany and the US Cancer Family Registry. The next step in the pathway is mounting on general databases and browsers;[14] this has been performed by the tuberous sclerosis 2 (TSC2) database (http://www.LOVD.nl/TSC2), and many databases are on the University of California Santa Cruz (UCSC) Genome Browser.[15] Other databases are also pursuing global data (Fanconi Anaemia [http://www.rockefeller.edu/fanconi/mutate/] and Cystic Fibrosis [http://www.genet.sickkids.on.ca/cftr/]). Access to clinical data as a routine is now being examined.

Country-specific data

LSDBs typically collect *novel* mutations, but for decisions on healthcare in a country, all instances of a mutation are now used not only for diagnostic strategies and connecting families, but also for discovery of modifier genes and in therapeutic trials of mutation specific drugs. There have been patchy attempts to make such a collection but they have not been systematic, have not led to routine practice or not developed methods readily transported to other countries. Recently, a consortium was formed in Australia to set up such a system. Other countries, such as Korea, Argentina, Saudi Arabia, Kuwait, China and Japan, are enthusiastically moving in this direction.

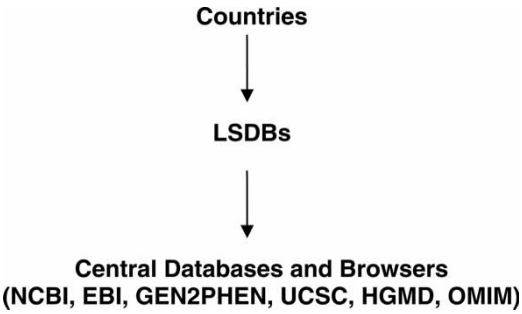

Figure 1. Mutation flow needed.

After the collection of legacy data in countries and deposition in databases, systems for effortless transfer of data to databases will need to be developed.

The proposed final pathway is shown in Figure 1. Much needs to be done, in addition to the above, but this will be done around these core activities. Besides clinical and diagnostic activities, curation, software, funding and so on will be needed. As funding is difficult to obtain for this project, and because so many people are needed to undertake it, we need all those interested to come forward to assist and join the upcoming formal HVP Consortium, which is based around those who have contributed to the project, directly or indirectly, over many years. Otherwise, individuals can contact the Author.

References

1. Chang, J.C. and Kan, Y.W. (1979), 'Beta 0 thalassemia, a nonsense mutation in man', *Proc. Natl. Acad. Sci. USA* Vol. 76, pp. 2886–2889.
2. Hamosh, A., Scott, A.F., Amberger, J., Valle, D. and McKusick, V.A. (2005), 'Online Mendelian Inheritance in Man (OMIM)', *Hum. Mutat.* Vol. 15, pp. 57–61.
3. 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.
4. Dantzer, J., Moad, C., Heiland, R. and Mooney, S. (2005), 'MutDB services: Interactive structural analysis of mutation data', *Nucleic Acids Res.* Vol. 33 (Web Server issue): W311–W314.
5. Cotton, R.G., Phillips, K. and Horaitis, O. (2007), 'A survey of locus-specific database curation', *J. Med. Genet.* Vol. 44, p. e72.
6. Horaitis, O., Talbot, C.C. Jr, Phommarinh, M., Phillips, K.M. *et al.* (2007), 'A database of locus-specific databases', *Nat. Genet.* Vol. 39, p. 425.
7. Cotton, R.G., McKusick, V. and Scriver, C.R. (1998), 'The HUGO Mutation Database Initiative', *Science* Vol. 279, pp. 10–11.
8. Ring, H.Z., Kwok, P.Y. and Cotton, R.G. (2006), 'Human Variome Project: An international collaboration to catalogue human genetic variation', *Pharmacogenomics* Vol. 7, pp. 969–972.
9. Axton, M. (2007), 'What is the Human Variome Project?', *Nat. Genet.* Vol. 39, p. 423.
10. Cotton, R.G., Appelbe, W., Auerbach, A.D., Becker, K. *et al.* (2007), 'Recommendations of the 2006 Human Variome Project meeting', *Nat. Genet.* Vol. 39, pp. 433–436.
11. 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.
12. ENCODE Project Consortium (2004), 'The ENCODE (ENCyclopedia Of DNA Elements) Project', *Science* Vol. 306, pp. 636–640.
13. Fokkema, I.F., den Dunnen, J.T. and Taschner, P.E. (2005), 'LOVD: Easy creation of a locus-specific sequence variation database using an 'LSDB-in-a-box' approach. *Hum. Mutat.* Vol. 26, pp. 63–68.
14. den Dunnen, J.T., Sijmons, R.H., Andersen, P.S., Vihinen, M. *et al.* (2009), 'Sharing data between LSDBs and central repositories', *Hum. Mutat.* Vol. 30, pp. 493–495.
15. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M. *et al.* (2002), 'The human genome browser at UCSC', *Genome Res.* Vol. 12, pp. 996–1006.

Richard G.H. Cotton
Head Genomic Disorders Research Centre and Convenor, Human Variome Project, Howard Florey Institute, Carlton South; and Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Parkville, Vic 3010, Australia
Tel: +61 3 8344 1893; Fax: +61 3 9347 6842;
E-mail: cotton@unimelb.edu.au