EDITORIAL

nature genetics

WayStation to HUGOBase

Within a century, the language was established: a Samoyedic Lithuanian dialect of Guarani, with classical Arabian inflections. The content was also deciphered: some notions of combinative analysis, illustrated with examples of variations with unlimited repetition. These examples made it possible for a librarian of genius to discover the fundamental law of the Library.
—Jorge Luis Borges, "The Library of Babel"

The hope of database projects is that data gathered for one researcher's purpose can be used by others, thereby avoiding duplication of effort. In the case of human genetic variation and its associated phenotypes, data are being gathered in a profusion of different contexts and assembled into a proliferating cacophony of databases. The databases themselves need to be organized, with tools, incentives and standards that will induce researchers to submit, credit and use genomic data in a more efficient way. We suggest that researchers and publishers could work together under the auspices of the Human Genome Organization (HUGO) to develop a database of genome variants. Such a project should be sustainable and reflect the different needs of its users. As a catalog, such a project seems dauntingly large, but it would not necessarily entail the documentation of the nearly infinite information space imagined by Borges in the quote above, and the longer we wait, the more data will be collected in a Babel of formats.

Human genetic variation forms a spectrum, from rare mutations with large effects, through rare variants, chromosomal rearrangements and structural polymorphisms, to common SNPs and haplotypes. Genetic causes of disease and phenotypic variation have been found in every part of the spectrum, and discoveries in mendelian genetics and population epidemiology are increasingly mutually illuminating. Recognizing this, the Human Genome Variation Society (http://www.hgvs.org/) was formed under the auspices of HUGO and is beginning to pull together, examine and discuss the disparate genomic repositories with a view to overall coordination. With core databases, locus-specific databases and viewing tools, some might conclude that the last thing we need is another database. In fact, amid the proliferation of resources, HGVS database proponents George Patrinos and Tony Brookes described the current provision as "disastrously deficient" in a recent commentary (*Trends. Genet.* **21**, 333–338; 2005).

The needs of researchers are various, from the clinician who wants to know whether a particular gene variant has been seen in a pedigree with a certain disease, to a testing company wanting a catalog of all the disease-causing mutations at a particular locus, to researchers studying common and complex diseases where allele frequency and haplotype background become factors used to calculate genetic risk. The minimum transferable information gained from the databases may be only that a variant occurs somewhere in the human population, so that it can be added to flexible high-throughput genotyping assays. Still, adding phenotypic information and population-specific allele frequencies will certainly help with prioritizing this process.

Three stumbling blocks are the size of the project, attributing credit and sustaining funding. One part of the solution to a big data set is the provision of tools to access database information, such as Genewindow (*Nat. Genet.* **37**, 109–110; 2005).

New reports of mutations are published in some journals (*e.g.*, *Human Mutation*), but we lack an appropriate way to credit allele frequency information or inconclusive but properly executed genetic association studies. Creative solutions are likely to arise from dialog between researchers and journal publishers. One successful model for community-publisher collaboration in cell biology is the Signaling Gateway (http://www.signaling-gateway.org/), which provides peer-reviewed, citable publication units (molecule pages), each with its own digital object identifier. Because different interest groups will be providing different kinds of data, with varying levels of motivation to publish, a sustainable database will probably need to run on a mixed economy of funding sources and incentives.

The challenge of documenting mutations, polymorphisms and even structural rearrangements seems minor and technical in comparison with that of constructing the catalog of phenotype information needed to make the variants meaningful, let alone linking phenotypes in a usable way to the genetic data. Options range from a distinctive phenome project (*Nat. Genet.* **34**, 15–21; 2003), through ontologies, to a standardized mark-up language for annotating polymorphisms, like the one advocated by Patrinos and Brookes in their commentary. One suggestion is that phenotyping should borrow from the national gene bank projects. Because these enormous undertakings stand to contribute a large proportion of the next decade's data, our solutions will need to have their origins in the standards being established for them. These biobanks have also run headlong into the ethical issues of consent and genetic privacy (*Nat. Biotechnol.* **23**, 539–545; 2005), and so some solutions can be borrowed for the database of databases.

There are many options and many stakeholders, but the essentials are certain: a tool to make data submission easy, a credit and citation system to make it attractive, review to validate it and a powerful cross-platform browser to access data. We can consolidate our knowledge and ignorance into one library now or evolve a solution eventually, at a greater cost. ■

A SPECTRUM OF DATABASES

OMIM	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
HGMD	http://www.hgmd.cf.ac.uk/
WayStation	http://www.centralmutations.org/
ALFRED	http://alfred.med.yale.edu/alfred/index.asp
CDC GDP	http://apps.nccd.cdc.gov/genomics/GDPQueryTool/
HapMap	http://www.hapmap.org/
GAD	http://geneticassociationdb.nih.gov/