

resistance markers or surrogate markers for resistance is available, it may impact on the selection of the drug for therapy. Specific examples would be the point mutations in key topoisomerase loci associated with quinolone resistance or the presence of the *mec* region or an *ermB*-bearing transposon. Another example would be type and expression levels of  $\beta$ -lactamases or drug efflux pumps. A 'profile' of resistant or resistance-prone strains may be ascertained, and this profile could emerge as predictive of the chance of resistance emergence in a particular pathogen population. A pharmacogenomics analysis of gene expression of a pathogen may provide a *predictive* outcome on therapy. At present, the methods are too costly and time consuming to contemplate this for routine analysis. However, it is not too early to begin to devise methods and gather data on these questions.

Pharmacogenomics offers the opportunity to exploit knowledge of the change in mRNA expression and the change in protein expression in response to a novel antimicrobial agent. These changes in expression can be in the host, in the pathogen, or even in an intermediary commensal organism (Figure 1). Knowing what genes are turned on, or which pro-

teins' expression are changed in the pathogen upon infection provides additional targets for antimicrobial chemotherapy. Knowing that same information for the host offers additional targets that can be assessed to help the host fight off an infection, eg by regulating the innate immune response or blocking inappropriate inflammation. This knowledge will occur first in the laboratory, and the hunt for correlations will take time. Once correlations are established, though, there are two immediate paths available: screening to identify agents to stop the infection and development of rapid, highly specific diagnostics to assess the pathogen population.

Genomics, pharmacogenomics, and proteomics are technologies that offer tremendous promise in anti-infective drug discovery. The key to productively using these technologies is to think beyond the current research paradigm ie, identify a compound that kills or prevents the infective agent from growing. Instead, these technologies give us the potential for tens, if not hundreds, of novel anti-infective targets—from the pathogen's gene regulation and expression in response to therapy, to the host's gene regulation and expression in response to

infection. We now have the opportunity to look deeper into the infective process and find other points at which to stop the pathogen. It is a tremendously exciting time to be working in anti-infectives: the promise is considerable, and the goal the best possible: saving millions of lives.

#### DUALITY OF INTEREST

None declared.

#### Correspondence should be sent to

DB Davison, Bioinformatics, HW3-0.07, Bristol-Myers Squibb Pharmaceutical Research Institute, 311 Pennington-Rocky Hill Road, Pennington, NJ 08534, USA.  
Tel: +1 609 818 4224  
Fax: +1 609 818 3100  
E-mail: Daniel.Davison@bms.com

- 1 Peet NP, Bey P. *Drug Discov Today* 2001; **6**: 495–498.
- 2 Wieczorek SJ, Tsongalis GJ. *Clin Chim Acta* 2001; **308**: 1–8.
- 3 Schmitz G *et al.* *Clin Chim Acta* 2001; **308**: 43–53.
- 4 Stephenson J. *JAMA* 2001; **286**: 1441–1442.
- 5 Dietrich WF. *Curr Biol* 2001; **11**: 1503–1511.
- 6 Martinez JL, Baquero F. *Antimicrob Agents Chemother* 2000; **44**: 1771–1777.
- 7 Oliver A *et al.* *Science* 2000; **289**: 391–392.

## The HUGO Mutation Database Initiative

RGH Cotton<sup>1,2</sup> and O Horaitis<sup>1</sup> on behalf of the HUGO Mutation Database Initiative

<sup>1</sup>Genomic Disorders Research Centre, St Vincent's Hospital Melbourne, Fitzroy, Australia; <sup>2</sup>The University of Melbourne, Department of Medicine, Melbourne, Australia

*The Pharmacogenomics Journal* (2002) **2**, 16–19. DOI: 10.1038/sj/tpj/6500070

The human genome has somewhere around 30000 genes.<sup>1</sup> If we consider that some genes such as cystic fibrosis have nearly 1000 mutations causing this rare inherited disorder, it is possible that there may be up to  $30 \times 10^6$

mutations causing single gene disorders if mutations in all genes cause disease. A more conservative figure is  $3 \times 10^6$ . If we consider also non-disease causing polymorphisms that are thought to occur every 200–1000 bases in the  $3 \times 10^9$  genome, we arrive at 3–15 million possible polymorphisms. In the case of polymorphisms these are

important in common disease, in variation in drug metabolism and as markers in linkage studies. When one considers single base changes in the  $3 \times 10^9$  bases and that each of these can change to one of three others, there are potentially  $9 \times 10^9$  base changes possible (without insertions or deletions). Thus it is clear that there are likely to be at least tens of millions of base changes that are important to human health. In the case of single gene disorders, each mutational event needs to be characterized by at least 10 extra pieces of data, ideally more like 50,<sup>2</sup> whereas polymorphisms perhaps need less. This means that there are at least hundreds of millions of pieces of data that are needed to fully record variation in the human genome. This is only one order of magnitude less

than the task of recording the human genome sequence of  $3 \times 10^9$  units. Thus it is in the interest of medical science that a system be put in place to systematically collect accurate variation data, safely store it, and make it available to those who need the data. It was this impending scenario and a need for a system to cope with it that induced the formation of the HUGO Mutation Database Initiative in 1994. Its history and progress can be seen on its website.<sup>3</sup> It has been supported by the Human Genome Organization (HUGO) and the March of Dimes and has around 600 members in 34 countries.

Phenotype variation has been known and used for thousands of years, however molecular variation was only revealed in the 1950s once protein sequencing of mutant proteins was established. The rate of discovery of molecular variation accelerated considerably in the 1960s and 1970s when gene cloning and DNA sequencing were applied to disease genes in rare inherited disorders. The globin gene was perhaps the first human disease gene in which mutations were described. These were collected by those interested in such variation and printed as a book<sup>4</sup> and, with the advent of computer databases, were made with relevant software and placed on the WWW. In the case of the globin genes the data from the book have been transferred to a website.<sup>5</sup> Also in the 1960s Victor McKusick began collecting inherited syndromes<sup>6</sup> and later began listing mutations in the genes that were found mutant and ultimately this listing, online Mendelian Inheritance in Man, has been placed on the Web.<sup>7</sup>

Since these early developments there has been an expansion of numbers of databases. Those databases collecting mutations in single genes are called locus specific mutation databases (LSDBs), whereas those collecting mutation in all or many genes are referred to as central or general mutation databases.

Central or general mutation databases collect mutation in all genes but those that exist differ because of their reasons for being initiated. These have

recently been reviewed.<sup>8</sup> OMIM<sup>7</sup> began as a systematic record of inherited syndromes in print form. As genes causing the syndromes were identified, the records in this compilation began to include mutations identified in such genes. Because it cannot keep up with all mutations it only collects the first mutation and then the most interesting after that. For example in cystic fibrosis and phenylketonuria (12 September 2001), OMIM contains 127 of 989 and 65 of 443 mutations in these diseases respectively, compared with the mutations in the Locus Specific Database for these genes. HGMD<sup>9</sup> began as a research tool to document the different types of mutations occurring in humans and ultimately led to the finding that mutations in CpG doublets were the most frequent and then to exploration of why this was so. This collection from the published literature has become a useful compilation so that users could find if a particular mutation had been described and, if so, who by and where. Currently this compilation is some months behind due to a commercial agreement necessary for its funding. dbSNP<sup>10</sup> was initiated by the NCBI as the US government's public compilation of variation and specifically for polymorphisms (or SNPs—single nucleotide polymorphisms as they were known from that time), which were discovered from a major funding initiative to do so. However, despite the name, this database collects all variation that is submitted including that causing single gene disorders. HGBASE<sup>11</sup> collects any and all variation types regardless of clinical correlations or not, as well as carrying frequency data. Strict criteria are applied to variants hosted (single copy, mapped uniquely, fully consistent allied data features). Variations from other databases and the literature are actively sought out, giving a significantly broader scope than dbSNP. Other databases reside at the Whitehead Institutes and there is another, initiated by a group of pharmaceutical companies.

As a general rule, those central databases recording variation causing single gene disorders, only record

published variation. The SNP database tends to record published SNPs as well as large numbers submitted from major funding efforts. Collection from the literature may not be sustainable in the long term when we have a need for documentation of millions of mutations.

The listing of the mutations in the globin gene(s) was in fact the first locus specific mutation database (LSDB), where the main author was interested in collecting the details of the mutation and the phenotype. Today there are around 260 LSDBs mounted on nearly 100 websites.<sup>2</sup> These databases vary in almost every aspect (except those on the same website where their characteristics are similar), because not only do they use 10 or so different software types but also the initiators have had different interests and different objectives in mind. Also, some are better funded than others so appear more professional. There are three main types of LSDBs, those focusing on the mutation only and describing only the first example of each, eg the PAH database,<sup>12</sup> those cataloguing patients with specific diseases and noting the mutations, eg MUTBASE<sup>13</sup> and those cataloguing somatic mutations, eg TP53.<sup>14</sup>

Mutation View<sup>15</sup> was originally described as an integrated system of central and LSDBs. It does catalogue each published case of a particular mutation.

There are two major differences between LSDBs and Central Databases that have important consequences regarding utility for specific purposes. First is that LSDBs are run by experts in the gene involved and secondly most of them collect unpublished mutations. The consequence of the first point is that many of the LSDBs are more knowledge bases of the genes, eg PAH<sup>12</sup> with enormous amounts of information ranging from that for biochemists to that for patients. On the second point, the consequences are that a recent survey<sup>16</sup> showed that LSDBs contained around 100% more mutations than HGMD that only collects published mutations.

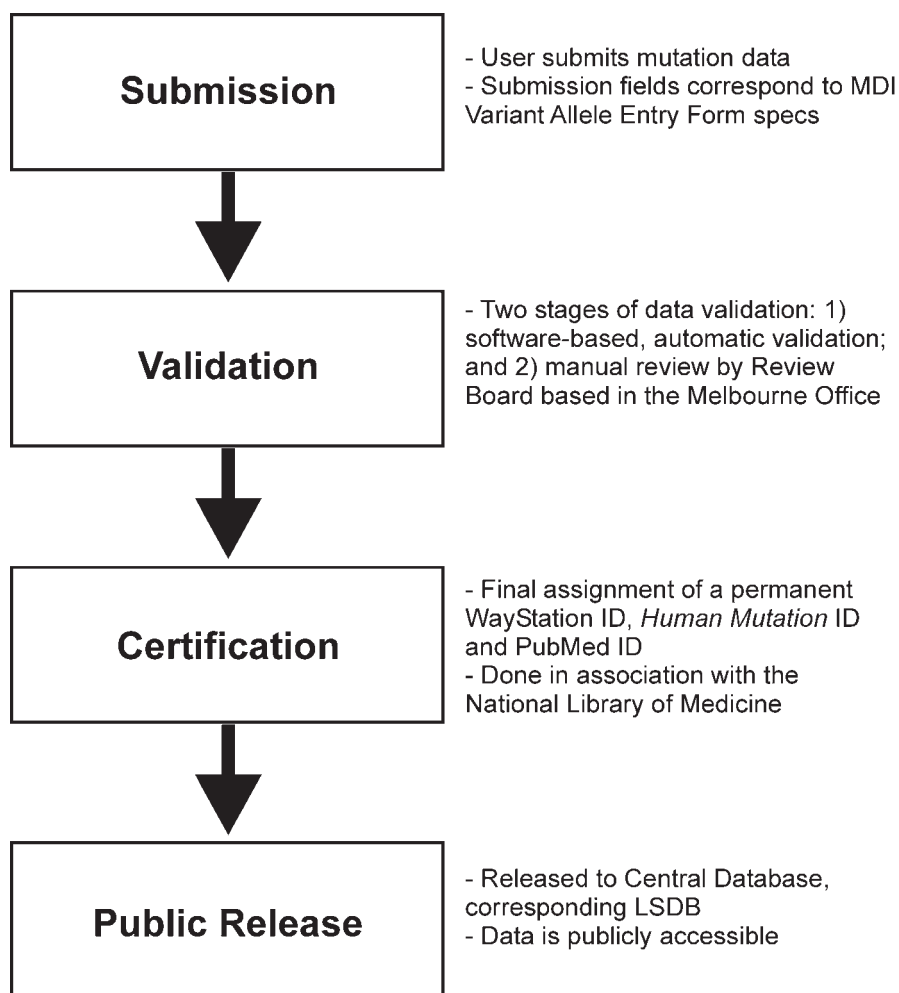


Figure 1 Proposed data submission to release scheme.

The main overarching objective in achieving the aims of the HUGO MDI has been to combine the strengths of the central database and the LSDBs. Thus in broad terms the Initiative set out to establish a federation of LSDB curators to ensure capture and work with central databases to ensure storage and distribution on a proper bioinformatics basis.

It is an enormous challenge to aim towards the day when tens of millions of sequence variations will have been accurately collected and stored and made available to the public. However, because of the contribution such a compilation will have to medical care, medical research and biological science generally, there has been considerable interest, enthusiasm and drive at the twice yearly Mutation

Database Initiative meetings. Members of the Initiative have published extensively on the topic particularly focusing on topics of concern (see HUGO-MDI website<sup>3</sup>).

Mutation *nomenclature* was an early concern, as whilst there were several systems in use, proper discussion with consequent recommendations had never occurred. The outcome of such a process has resulted in a HUGO-MDI recommended nomenclature for the simple changes<sup>17</sup> with a further discussion for more complex mutations.<sup>18</sup>

Recommendations for *content* have been published<sup>19</sup> as well as a recommended entry form<sup>20</sup> produced by many members of the Initiative. Some members of the Initiative have produced off-the-shelf software for

the initiation of web-based databases.<sup>21,22</sup>

*Quality control and peer review* has rightly received considerable attention. One of the outcomes of this has been automated mutation by mutation checkers<sup>23</sup> that simply check if the correct amino acid change, base numbers, etc have been deduced, also a set of rules have been developed for the acceptance of a mutation as causing single gene disorders.<sup>24</sup>

The most daunting problem is how to ensure complete collection of all variation that is being uncovered. This problem is being compounded by the fact that journals are generally not accepting reports of single mutations after the initial wave once a disease gene is discovered. This is especially so for the 452nd mutation causing PKU or even a group of them. Initially the journal *Human Mutation* accepted such publications electronically and published them electronically but this has ceased. The Initiative members have thus been moved to plan an integrated system of receipt, review, publication, PubMed ID registration, and public storage. This has resulted in a pilot receipt point, the 'WayStation'<sup>25</sup> and agreement for publication of data by Wiley-Liss in *Human Mutation* and agreement by HGBASE<sup>11</sup> to be the storage database for the data.

Another approach to ensuring mutation capture has been to encourage *National Databases* who are likely to be able to contact all diagnostic and research laboratories in their country to induce collection of mutations. One such database is the Turkish database.<sup>26</sup> Besides ensuring mutation collection, such national (or ethnic) databases are a vital aid to delivery of national genetic health care. Because of the past and current huge transnational migration such national/ethnic databases are of international importance.

Because of concerns to ensure genetic privacy, mutation databases need to consider *ethical aspects* whereby patients may not wish to have their perhaps identifiable mutation on the WWW. Another concern has been *Copyright and intellectual property aspects*. There have been cases of data-

bases being taken from a site without permission and placed on another without attribution. There needs to be some mechanism to avoid this problem.

Of concern to readers of this journal is the collection of SNPs now that wholesale collection by concerted public and industry funding has ceased. We expect SNPs in single gene disorder genes to be collected through the same mechanisms particularly from diagnostic laboratories.

The eternal problem of such projects is *funding*. In the case of MDI we have had generous support from HUGO and the March of Dimes and we are currently looking at government support (see below) and possibly commercial funding.

To ensure that we achieve our most difficult remaining aims there are several initiatives underway. Coupled with the agreement with Wiley/*Human Mutation*, the HUGO MDI is about to create a Society, with *Human Mutation* as its journal. Regarding funding we have been invited to and have submitted a P41 grant application to NIH to fund the collection (WayStation), the storage (an update of HGBASE<sup>11</sup>) and the Administrative office.

The current plan is to receive mutation reports at the WayStation; once they have been automatically checked (mutation checker<sup>23</sup>) they will be sent for expert review. LSDB curators will be asked to do this, where one exists or an expert in the gene in question where one doesn't. Once the submission has been approved it will be sent first to NCBI for a PubMed ID, then to HGBASE and the LSDB if one exists. The current scheme is illustrated in Figure 1.

To ensure systematic and complete capture of all variation described,

there will be a need not only for voluntary work, as is now occurring in the HUGO MDI, but certainly (besides software) a need to pay key individuals to ensure the grass roots are searched for mutations and SNPs. Thus the success of our enterprise depends on funding and this will have to be either from government, from sales of updates of our data or straight commercial funding. The latter of course will need to allow the database content to be fully and immediately public.

Much has been achieved since the initiative began in November 1994. However, it is natural that what remains is the most difficult task so we will have to be patient in carrying it out. The indications are that all is in place to make it achievable. This has been and will be a community activity and we invite all who are able to assist to do so by contacting the authors.

#### DEFINITIONS

*Uses of the words mutation, polymorphism and SNP have been problematic. Throughout biology, mutation is any base change but in clinical genetic usage mutations refer to a deleterious change causing single gene disorder. In the same clinics polymorphism is used for harmless base changes. Further confusion has been added by dbSNP at NCBI, which receives not only single nucleotide polymorphisms as 'advertised' but also other types of polymorphisms and base changes of any type causing single gene disorders. The simplest nomenclature is to call them base changes as such or refer to sequence variation without making a judgment on its effect as is made in the clinic.*

#### DUALITY OF INTEREST

None declared.

#### Correspondence should be sent to

RGH Cotton, Genomic Disorders Research Centre, St Vincent's Hospital Melbourne, PO Box 2900, Fitzroy VIC 3065, Australia.  
Tel: +61 3 9288 2980

Fax: +61 3 9288 2989

E-mail: cotton@ariel.ucs.unimelb.edu.au

- 1 Das M *et al.* *Genomics* 2001; **77**: 71–78.
- 2 Claustres M *et al.* *Genome Res* (under review).
- 3 HUGO MD URL: <http://www.genomic.unimelb.edu.au/mdl/>
- 4 Huisman THJ *et al.* *A Syllabus of Human Haemoglobin Variants*, 1st edn. The Sickle Cell Anaemia Foundation: Augusta, GA, USA.
- 5 The Globin Gene Server URL: <http://globin.cse.psu.edu/>
- 6 McKusick VA. *Mendelian Inheritance in Man: Catalogs of Autosomal Dominant, Autosomal Recessive, and X-Linked Phenotypes*. The Johns Hopkins Press: Baltimore, MD, USA, 1966.
- 7 Online Mendelian Inheritance in Man: <http://www3.ncbi.nlm.nih.gov/omim/>
- 8 Porter CJ *et al.* *Hum Mut* 2000; **15**: 1236–1244.
- 9 Krawczak M, Cooper DN. *Trends Genet* 1997; **13**: 121–122. URL: <http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html>
- 10 NCBI dbSNP URL: <http://www.ncbi.nlm.nih.gov/SNP/>
- 11 Human Genic Bi-Allelic Sequences HGBASE URL: <http://hgbase.cgr.ki.se/>
- 12 Phenylalanine Hydroxylase Locus Database URL: <http://ww2.mcgill.ca/pahdb/>
- 13 MUTBASE: <http://www.uta.fi/imt/bioinfo/mutdatbas.html#idmdb>
- 14 TP53 database: <http://www.iarc.fr/P53/index.html>
- 15 Minoshima S *et al.* *NAR* 2001; **1**: 327–328. <http://131.113.190.126/mutview3/mutview/>
- 16 Cotton RGH. *Hum Mut* 2000; **15**: 4–6.
- 17 Antonarakis SE and the Nomenclature Working Group. *Hum Mut* 1998; **11**: 1–3.
- 18 den Dunnen JT, Antonarakis SE. *Hum Mut* 2000; **15**: 7–12.
- 19 Scriver CR *et al.* *Hum Mut* 1999; **13**: 344–350.
- 20 Entry Form: <http://www.genomic.unimelb.edu.au/mdl/entry.html>
- 21 Universal Mutation Database Software: <http://www.umd.necker.fr/>
- 22 MuStar<sup>®</sup>: <http://www.hgu.mrc.ac.uk/Softdata/Mustar/>
- 23 DNA Mutation checker: <http://www2.ebi.ac.uk/cgi-bin/mutations/check.cgi>
- 24 Cotton RGH, Horaitis O. *Hum Mut* 2000; **15**: 16–21.
- 25 WayStation Pilot: <http://www.centralmutations.org>
- 26 Turkish mutation database: <http://bioserver.bio.boun.edu.tr>