'A variant of uncertain significance' and the proliferation of human disease gene databases

*David R. Nelson**

Department of Molecular Sciences and The UT Center of Excellence in Genomics and Bioinformatics, University of Tennessee, Memphis, TN 38163, USA
**Correspondence to*: Tel: +1 901 448 8303; Fax: +1 901 448 7360; E-mail: dnelson@utmem.edu

Date received (in revised form): 9th October 2004

Abstract

The rapid accumulation of mutation data has led to the creation of nearly 300 locus-specific mutation databases. These sites may contain a few dozen to almost 20,000 mutations for a given gene. Many of the mutations are uncharacterised and have no known effects on the gene product, the 'variant of uncertain significance'. Here, the statistics of mutation distribution are examined for six different gene databases: *BRCA1* and *BRCA2*, haemoglobin-beta (*HBB*), *HPRT1*, *CFTR* and *TP53*. The percentage of all possible point mutations for a protein (the mutation space) is calculated for each gene and the question 'How much mutation data is enough?' is raised.

Keywords: *variant of uncertain significance, human mutation databases,* BRCA1, BRCA2, *haemoglobin-beta,* HBB, HPRT1, CFTR, TP53, *mutations*

Mutations and human diseases

The number of people on earth is now more than 6 billion. It is generally understood that every individual has a few mutations in their genome that did not exist in their parents' genomes. An estimate of the spontaneous mutation rate in humans is 1.8×10^{-8} per nucleotide per generation.[1] With 3.2 billion bases per genome, this would translate into 115 mutations per diploid genome per generation. Multiplying that by 6 billion people gives 690 billion mutations per human generation, or about 216 mutations for each nucleotide in the human genome.

Our species is undergoing a massive worldwide mutational experiment, with every nucleotide in our genomes being tested by mutation. What this means is that, if we look hard enough, we should find all non-lethal sequence variants in any given human gene. The genetics research community has begun this search by identifying genes that, when defective, lead to a human disease. The Online Mendelian Inheritance in Man (OMIM) database[2] is a repository for information on human genes. OMIM suggests searching LocusLink,[3] limiting to human, the search with the term 'disease_known' to find a human disease gene count (2,440 entries). Refining this search with 'disease_known AND has_seq' will limit the result to human diseases with a known gene sequence. As of 23rd September, 2004, this number is 1,692. The current gene count in humans is 42,716 according to the Genome Alignment and Annotation (GALA) database.[4] Therefore, about 4 per cent of human genes today have a disease associated with them.

To show that a gene is responsible for a disease, a mutation search is usually carried out in patients and unaffected relatives. The location of stop codons, splice-site mutations, in-frame deletions, frameshifts or missense mutations that affect function in patients but not in unaffected individuals proves that the gene is responsible for the phenotype of the disease; however, this is only the beginning. If a disease is common or of longstanding interest to the research community, there is often a database started for that gene to compile the known mutations and phenotypes. Such databases start small, but they can become quite large. The Human Gene Mutation database[5] has a link to 292 locus-specific mutation databases.[6] Because there is some duplication (such as six different TP53 databases) and some databases cover more than one gene, these 292 links cover a minimum of 257 different human genes. Many are free to the public, but some require registration, a password or subscription — as with the 'hypoxanthine guanine phosphoribosyltransferase-1' (HPRT) database.[7,8] These 257 genes currently represent 15 per cent of the 1,692 human disease genes with known sequences. This list is not comprehensive, since it does not contain the HbVar database of haemoglobin variants.[9,10]

What can be expected from intense scrutiny of a gene? As mentioned above, if we look hard enough, mutations in every codon and even every nucleotide — that do not cause a lethal

phenotype — should be discovered. In the real world, how close are we to achieving this hypothetical coverage? Herein, are presented six examples of well studied genes: *BRCA1* and *BRCA2*, the breast cancer susceptibility genes; haemoglobin-beta (*HBB*); *HPRT1*; the cystic fibrosis transmembrane conductance regulator (*CFTR*); and the *TP53* tumour suppressor.

BRCA1 and *BRCA2*

The *BRCA1* and *BRCA2* genes are special cases of great human interest. More than 20,000 complete coding sequences of *BRCA1* from patients with breast cancer or from their relatives have been determined.[11] This includes some flanking sequence around each of the 24 exons. The protein is 1,863 amino acids long, which is equivalent to 5,589 base pairs (bp) just for the coding sequence. Usually, the test that is ordered involves sequencing both *BRCA1* and *BRCA2* (27 exons, 3,418 amino acids, 10,254 bp for the coding region) at a cost of $2,975 for the first member of a family and $350 for subsequent members. The cost may be partially covered as part of a health insurance plan. Because Myriad Genetic Laboratories has a patent on the test, they perform all of these tests with uniform quality control, and all of the mutation data goes into one database, the Breast Cancer Information Core (BIC).[12] Access to the BIC is password-protected for members only. Membership may be obtained by online application[13] and it is clearly intended for researchers, not patients. Results from the BIC have been summarised.[14–16]

As of 22nd September, 2004, there were 9,556 entries for *BRCA1* and 9,217 entries for *BRCA2*. Each entry represents a mutation found in one person. Since 20,000 individual sequence tests have been performed,[11] one can infer that more than 10,000 of them are without a single mutation, else they would have been counted as an entry in the BIC. For *BRCA1*, 1,539 of the 9,556 entries are distinct mutations, polymorphisms or variants. The remainder are duplicates, and some are very common founder mutations. Of the 1,539 unique mutations, 878 were observed only once (57 per cent); these are the rare variants mentioned above and are seen once in 20,000 sequences (or even at a lower frequency, but discovered by accident in 20,000 sequences). By examining the BIC database for mutations in the coding region, it was shown that 850 of the 1,863 codons contain at least one mutation (46 per cent). By looking at the cDNA mutation map, the largest non-mutated region is 43 bp, from nucleotides 4,530 to 4,572. The mutations on this histogram plot are so thick that the whole gene looks like a dense bar code.

BRCA2 is a longer gene. It has 1,893 distinct mutations, polymorphisms or variants, with 1,146 reported only once (60 per cent); this is very similar to *BRCA1*. In *BRCA2*, 1,323 of the 3,418 codons have at least one mutation (39 per cent). From the cDNA mutation map, the largest non-mutated region is 48 bp, from nucleotides 1,282 to 1,329. This brings us back to the earlier question of how saturated is our mutation catalogue. To discuss this, I will introduce the concept of *mutation space*. Considering only single-base changes, and leaving out insertions and deletions, if each codon can theoretically be mutated by nine different one-base changes, then the mutation space for a protein-coding gene is nine times the number of codons. For illustrative purposes, the number of distinct mutations in *BRCA1* is 1,539. The mutation space is 9 × 1,863 codons = 16,767. Thus, 1,539 distinct mutations is 9 per cent of the mutation space — although that number does include frameshift mutants, so the 9 per cent value is slightly inflated. For *BRCA2*, the percentage of mutation space is only 6 per cent. In other words, more than 90 per cent of the possible mutations are yet to be discovered. Because the sampling of this space is so low, a high percentage of the mutants found in *BRCA1* and *BRCA2* screening will never have been seen before. One of a kind missense mutations that do not truncate the protein will be the so-called 'variants of uncertain significance'.

Patients who have such a mutation are given a pamphlet entitled *Testing for Hereditary Cancer Risk: WHAT DOES A "VARIANT OF UNCERTAIN SIGNIFICANCE" MEAN?*[17] This is a looming question for geneticists faced with mutation data but no experimental data. The paper of Abkevich *et al.*[11] tries to address this by bioinformatics methods. Of 314 missense mutations in *BRCA1* from the 20,000 sequence tests, they state that only 21 are classified as deleterious, or suspected deleterious, mutants. By comparing human *BRCA1* with orthologues from other species, however, and by taking into account the properties of the side chains in the normal and mutated amino acids, they predicted that 50 more of these missense mutations would be suspected to be deleterious. Of the 243 other missense mutants, 14 had previously been deemed neutral or harmless. Their analysis added 92 more mutants into this category — for a total of 177/314 mutants with predictions either of deleterious, suspected deleterious or of little clinical significance. The remaining 137 mutants are still unclassified. The authors caution that this method only gives predictions, and that care must be used in interpretation in a clinical setting. It should be remembered that patients are making decisions about mastectomies, oophorectomies, chemotherapy options and radiation treatments based, in part, on this mutation information. Another study, by Goldgar *et al.*,[18] also uses a bioinformatics approach and integrates the results from several analyses into a weighted probability of significance. Their model was applied to three mutants each from *BRCA1* and *BRCA2*, allowing a risk to be assigned to five of the six mutants. It is not clear how the two methods will compare against each other.

Haemoglobin-beta

The second gene example is haemoglobin-beta (*HBB*). This was one of the earliest human sequences to be studied; in fact,

many of the mutations were found by protein sequencing methods, rather than DNA sequencing. HBB is among the most intensively studied human proteins; it is responsible for sickle-cell anaemia and the beta-thalassaemias. The OMIM entry for *HBB* has 1,157 references and entries for 522 allelic variants.[19] In 1996, Huisman *et al.* stated that 138 of the 146 codons of the *HBB* gene have been mutated.[20] Inspection of the HbVar database[9,10] reveals that this number is now 141/146 codons, with only Thr-4, Thr-12, Thr-50, Pro-125 and Val-137 not having a known mutation in humans. Some codons have five or six different mutations. Note that the Val after the start Met has been assigned as amino acid 1. *HBB* has 461 distinct point mutants in 146 codons; this equals 35 per cent of the mutation space ($9 \times 146 = 1{,}314$ possible point mutants). This value is much higher than that seen in the larger *BRCA1* and *BRCA2* genes. One observation that affects this value is the near-complete lack of synonymous substitutions in this database, which reflects the protein rather than DNA sequencing analyses. There are only two synonymous substitutions reported: both are Gly to Gly mutations. Certainly synonymous substitutions have been observed, but they do not seem to be reported. Including synonymous mutations would make the 35 per cent figure significantly higher.

Examination of the Pfam database of 7,503 protein family sequence alignments[21,22] for the globin family reveals that Thr-12 was seen as a Ser in the mouse-eared bat (*Myotis velifer*) and Thr-50 was also often seen as a Ser — even in primates such as *Macaca mulatta*. Pro-125 was almost always Gln, except in a few primates such as the gorilla, human and chimpanzee. Val-137 was invariant in all mammalian *HBB* sequences. Val-137 was an Ile in the South American lungfish (*Lepidosiren paradoxus*) and Ile or Leu in some frog *HBB* sequences. Thr-4 was outside the alignment edge in Pfam, but rabbits and black lemur (*Eulemur macaco*) both had a Ser at this position. These observations show that at least four of these five amino acids do vary in other mammals and probably could vary in humans without serious consequences. Val-137 seems to be the most invariant of the five, and it may have a significant effect if mutated to another residue.

HPRT1

Mutations in the *HPRT1* gene lead to the Lesch–Nyhan syndrome, a defect in the purine salvage pathway. The protein is fairly small, with only 218 amino acids (217 in the mature protein), yet there are 2,500 mutations reported in this gene.[7,8] The database for *HPRT1* is available only by subscription, so, without access, one is limited to published reports that give a detailed breakdown of the types of mutations and the number of unique mutations. In public sources, a total of 218 different mutations have been found in 271 patient cases studied.[23] The Human Gene Mutation Database now lists 223 mutants,[5] with 115 being missense or nonsense point mutants in 80 codons (about 6 per cent of the mutation space). In people, very few mutants have been found independently more than once. Somatic functional mutations in *HPRT1* can be selected *in vitro* in 6-thioguanine-resistant T lymphocytes from normal people. These are not naturally occurring mutants, they arise by inactivation of the *HPRT1* gene. Inactivation prevents this purine salvage pathway enzyme converting 6-thioguanine to a toxic nucleotide analogue. These *in vitro* mutants make up the majority of the *HPRT1* database that is not publicly available. Since they are inactivating mutations, they have to be at critical sites in the gene. The study by Duan *et al.* examined the known mutations (including patient mutations) and related them to the crystal structures of the HPRT protein.[24] Duan *et al.* report that 155 of the 217 amino acids (72 per cent) have known missense mutations that cause an amino acid change: these arise from a subset of 963 single-base-substitution mutants that cause missense mutations. The number of synonymous mutations in this set is not given. In addition to the 963 missense mutations, there are 51 (of a possible 66) nonsense mutations in the database; this gives a total of 1,014 single base mutations in 217 codons for 52 per cent of the sequence space ($9 \times 217 = 1{,}953$). After accounting for missense, nonsense and synonymous mutations, there remain 46/217 codons with no known mutations.

CFTR

There are currently 1,338 distinct mutations listed in the CFTR Mutation Database.[25] By examining the list of mutations and not including the mutations in non-coding regions or insertions and deletions, there were 702 unique point mutations in 501 of the 1,481 codons (34 per cent). This covers about 5 per cent of the *CFTR* gene's mutation space, which is very similar to that of the *BRCA2* gene.

TP53

The International Agency for Research on Cancer (IARC) TP53 Mutation Database[26,27] version R9 of June 2004, contains 19,809 somatic mutations (74 per cent missense, 7 per cent nonsense) and 264 germline mutations. More than 1,700 different point mutations at more than 310 distinct codons (out of a total of 393 codons) have been found as described in the slideshow on the database site. The exact numbers were not easily obtainable from the database. 1,700 mutations represent 48 per cent of the mutation space for 393 codons ($9 \times 393 = 3{,}537$); this is an even higher percentage than that seen in the *HBB* gene. This number includes synonymous mutations that do not change the amino acid. *TP53* is unusual in the distribution of its mutations. The vast majority of the mutations are in the middle third of the gene, coding for the DNA-binding portion of the protein.

Interpreting the data

The data on these six genes present a trend that will be followed with other disease genes. More and more sequence data will be collected until a high percentage of all the possible mutations are found. We could simply perform a 'thought experiment' and ask: 'What does one do when all possible (non-lethal) point mutants in a gene are documented in real people and the missing mutants are made and tested for function?' Nearly all of these mutants will be 'variants of uncertain significance'.

We are heading in this direction — as indicated by the 12th February, 2004 Request for Application (RFA) from the National Institutes of Health, entitled 'Revolutionary Genome Sequencing Technologies: The $1,000 Genome'.[28] Once the low-cost genome is available, perhaps in 10 years, millions of human genomes will have been sequenced and every gene will have sequence data exceeding what is now available for the *TP53* gene. All cancer patients will be reading the same pamphlet.

There are several approaches to addressing the massive information overload. First, genotype must be linked to phenotype by examining family histories; this will already have been done for the best-studied disease genes. Of course, this is not possible on a mutation that is seen only once. Secondly, the bioinformatics approach[11,18] will predict a mutation's probable significance by sequence comparison to orthologues and by rating the severity of an amino acid change. An enhancement of this method would be to scan for any sequence motifs for structure or targeting that may be affected by the mutation. Thirdly, examine the crystal structure of the protein, or of a close relative of that protein if known, to see if the mutation would cause a significant structural effect. Fourthly, assay the mutant protein in a variety of function tests or apply spectroscopic methods to detect changes in chromophores. The assay method may not be possible if the function of the protein is unknown. In many cases, the primary biology of the protein will need to be discovered — that is, the pathway in which it participates, any interacting proteins, expression levels, subcellular localisation, post-translational modifications, rates of turnover, etc.

This is really an outline for the biochemistry, cell biology and molecular biology of human beings at the level of every gene. Such understanding only comes slowly, whereas sequencing is far faster. In the interim, we will have more data than we can possibly use. Considering the cost of analysing each mutation, there must be some point that crosses a practical return on investment. Do we know everything we need to know about mutants of haemoglobin-beta, or should we make the remaining 843 point mutants just to be sure we have not missed anything? Nicholas Murray Butler (1862–1947), president of Columbia University, wrote: 'An expert is one who knows more and more about less and less'. The question ultimately becomes, how 'expert' do we need to be?

Acknowledgements

I would like to thank the owners of the databases mentioned in this paper, without which there could be no paper: The Breast Cancer Information Core (BIC), Online Mendelian Inheritance in Man (OMIM Database), LocusLink, The GALA Genome Alignment and Annotation Database, The Human Gene Mutation Database, HPRT Database, HbVar Database, Pfam Database, CFTR Mutation Database and IARC TP53 Mutation Database.

References

1. Kondrashov, A.S. (2003), 'Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases', *Hum. Mutat.* Vol. 21, pp. 12–27.
2. Online Mendelian Inheritance in Man (OMIM) Database. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = OMIM.
3. LocusLink. http://www.ncbi.nlm.nih.gov/LocusLink/.
4. Genome Alignment and Annotation (GALA) Database, Human; July 2003 data release. http://gala.cse.psu.edu/stats/galahg16_cnts.html.
5. Human Gene Mutation Database. http://www.hgmd.org/.
6. Locus-Specific Mutation Databases at HGMD. http://archive.uwcm.ac.uk/uwcm/mg/docs/oth_mut.html.
7. HPRT Database. http://www.ibiblio.org/dnam/mainpage.html.
8. Cariello, N.F., Douglas, G.R., Gorelick, N.J. *et al.* (1998), 'Databases and software for the analysis of mutations in the human *TP53* gene, human *HPRT1* gene and both the *lacI* and *lacZ* gene in transgenic rodents', *Nucleic Acids Res.* Vol. 26, pp. 198–199.
9. HbVar Database of haemoglobin variants. http://globin.cse.psu.edu/hbvar/menu.html.
10. Hardison, R.C., Chui, D.H.K., Giardine, B. *et al.* (2002), 'HbVar: A relational database of human hemoglobin variants and thalassemia mutations at the globin gene server', *Hum. Mutat.* Vol. 19, pp. 225–233.
11. Abkevich, V., Zharkikh, A., Deffenbaugh, A.M. *et al.* (2004), 'Analysis of missense variation in human *BRCA1* in the context of interspecific-sequence variation', *J. Med. Genet.* Vol. 41, pp. 492–507.
12. Breast Cancer Information Core, BIC (password required). http://research.nhgri.nih.gov/projects/bic/Member/index.shtml.
13. BIC membership application link. http://research.nhgri.nih.gov/projects/bic/application.cgi.
14. Couch, F.J. and Weber, B.L. (1996), 'Mutations and polymorphisms in the familial early-onset breast cancer (*BRCA1*) gene. Breast Cancer Information Core', *Hum. Mutat.* Vol. 8, pp. 8–18.
15. Shen, D. and Vadgama, J.V. (1999), '*BRCA1* and *BRCA2* gene mutation analysis: Visit to the Breast Cancer Information Core (BIC)', *Oncol. Res.* Vol. 11, pp. 63–69.
16. Hohenstein, P. and Fodde, R. (2003), 'Of mice and (wo)men: Genotype–phenotype correlations in *BRCA1*', *Hum. Mol. Genet.* Vol. 12(2), pp. R271–R277.
17. Anon (2000), Pamphlet: 'Testing for Hereditary Cancer Risk: WHAT DOES "A VARIANT OF UNCERTAIN SIGNIFICANCE" MEAN?', Myriad Genetic Laboratories, Salt Lake City, UT.
18. Goldgar, D.E., Easton, D.F., Deffenbaugh, A.M. *et al.* and the Breast Cancer Information Core (BIC) Steering Committee (2004), 'Integrated evaluation of DNA sequence variants of unknown clinical significance: Application to *BRCA1* and *BRCA2*', *Am. J. Hum. Genet.* Vol. 75, pp. 535–544.
19. + 141900 HEMOGLOBIN-BETA LOCUS; HBB. http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id = 141900
20. Huisman, T.H.J., Carver, M.-F.H. and Efremov, G.P. (1996), 'A syllabus of human hemoglobin variants', The Sickle Cell Anemia Foundation, Augusta, GA.
21. Pfam Database. http://www.sanger.ac.uk/Software/Pfam/.
22. Bateman, A., Coin, L., Durbin, R. *et al.* (2004), 'The Pfam protein families database', *Nucleic Acids Res.* Vol. 32, pp. D138–D141.
23. Jinnah, H.A., De Gregorio, L., Harris, J.C. *et al.* (2000), 'The spectrum of inherited mutations causing HPRT deficiency: 75 new cases and a review of 196 previously reported cases', *Mutat. Res.* Vol. 463, pp. 309–326.

24. Duan, J., Nilsson, L. and Lambert, B. (2004), 'Structural and functional analysis of mutations at the human hypoxanthine phosphoribosyl transferase (*HPRT1*) locus', *Hum. Mutat.* Vol. 23, pp. 599–611.
25. CFTR Mutation Database. http://www.genet.sickkids.on.ca/cftr/.
26. Olivier, M., Eeles, R., Hollstein, M. *et al.* (2002), 'The IARC TP53 Database: New online mutation analysis and recommendations to users', *Hum. Mutat.* Vol. 19, pp. 607–614.
27. IARC TP53 mutation database version R9, July 2004. http://www.iarc.fr/P53/.
28. Anon (2004), 'Revolutionary genome sequencing technologies: The $1,000 genome', National Institutes of Health Request for Application, 12th February, 2004 (available at http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-04-003.html).