

# **CHAPTER 1**

## **AN INTRODUCTION TO GENETIC POLYMORPHISM**

**David Hopkinson & David Whitehouse**

### **1.1 Introduction**

Polymorphism is a term which literally conjures up an image of variability of form, shape, size, structure and composition and it has a currency in a wide variety of disciplines in science and art. Genetic polymorphism is a much more specific term and describes frequent variation at a specific locus in a genome. A useful practical definition says that a locus is polymorphic when there are two or more allelic forms in the same population and the commonest allele has a frequency of 0.99 or less.<sup>1</sup> This implies that at least 2% of the population will be heterozygous at a polymorphic locus, defined in these terms. Alleles with a frequency less than 1% are arbitrarily designated “rare” variants and those with a higher frequency are termed “polymorphic”.

Genetic polymorphisms arise from mutation at a locus followed by the action of evolutionary forces, such as natural selection or drift, which spread the mutant allele through the population in which it arose. E.B. Ford<sup>2</sup> in early studies of numerous species defined genetic polymorphism as a product of population dynamics, well before the molecular basis of such variation was understood. Genetic polymorphism, he proposed, is variation which occurs at a frequency too high to be accounted for by recurrent mutation alone.

Most mutations are harmless and most are probably lost for ever from the population within which they arise after relatively few generations.<sup>3</sup> Some survive and drift up to “polymorphic” frequency by chance fluctuations in population dynamics and a few persist and flourish through selection pressure as balanced or transient polymorphisms. Only the tiny minority lead directly to genetic disorders. Also, we now know that less than 5% of the human genome is coding DNA and so the vast majority of genetic polymorphisms lie in non-coding relatively less significant regions of our DNA. Thus genetic polymorphisms mostly represent normal background variation of uncertain functional significance.

In practice, genetic polymorphisms are most valuable as marker loci for genetic mapping. They are extremely valuable as tools to locate and thus identify human genes responsible for single locus disorders. Gradually the emphasis is changing towards applications in complex polygenic disorders and traits. Studies of human evolution and genetic diversity are also very much enhanced by the use of polymorphic loci. Another very important practical application is the use of genetic polymorphisms as tools in forensic science to determine identity and genetic relationships.

Genetic polymorphisms have been studied at various levels, from simple non-invasive phenotypic assessment (such as tests of colour vision), through cellular and serological tests (such as the determination of blood groups and HLA types), metabolic analysis (such as the acetylator status), gene product analysis (such as serum protein and red cell isozyme polymorphisms) up to direct examination of the nuclear and mitochondrial DNA itself. The genomic analysis allows the scrutiny of coding and non-coding DNA sequences at a gene locus as well as the investigation of the vast acres of intervening DNA sequence for person-to-person variation. Thus the level of resolution of genetic polymorphisms has gradually sharpened through technological change and has emphasised the universal nature of this phenomenon.

Throughout this review, we shall quote Online Mendelian Inheritance in Man (OMIM) accession numbers wherever a specific

human gene or disorder is cited. Other references which contain details that are germane to this article are listed in the References, and cited as superscript numbers in the text.

## 1.2 Molecular Basis of Genetic Polymorphisms

### 1.2.1 Classical Protein Polymorphisms

Recombinant DNA technology has allowed the structural basis of many of the classical human polymorphisms to be defined at the molecular level. For example, the functional differences between the products of the *A* and *B* alleles at the *ABO* blood group locus [OMIM 110300] are due to a few single base substitutions and the null allele, *O*, is characterised by a single base deletion leading to frame shift (see Chap. 2 — Henry & Samuelsson for details). The RH *D/d* polymorphism [OMIM 111680] is characterised by a much more extensive gene duplication/deletion event, which was predicted in part from protein and serological studies. The basis of the RH *C/c* and *E/e* polymorphisms [OMIM 111700] in contrast was less easily predicted from gene product analysis, since molecular studies have shown that they do not involve two separate genes as originally believed but differential splicing at the same locus, plus single nucleotide polymorphisms (see Chap. 3 — Avent for details). In general however, the pattern of mutations underlying the classical polymorphisms is exactly the same as was predicted from protein analysis; viz. the majority are single nucleotide polymorphisms (SNPs) leading to missense changes in coding sequence, some are the result of deletion/duplication events [classical examples here are the several human mucin polymorphisms, such as *MUC1*, OMIM 158340] and a tiny minority involve major rearrangement of coding sequence [such as the serum haptoglobin polymorphism, *HP*, OMIM 140100]. This having been said however, it is important to recognise the role of recombination and gene conversion in the generation and maintenance of protein polymorphism; for example, in the human globin gene

families [4, OMIM 141800 *et seq*], in the *HLA* superfamily [OMIM 142800 *et seq* and see Chap. 8 — Ross & Harvey] and at the human *PGM1* locus [5, OMIM 171900].

### 1.2.2 *DNA Polymorphisms*

The first DNA variants were detected in Southern blots of genomic DNA digested with restriction enzymes. These restriction fragment length polymorphisms (RFLPs) are due to cleavage or non-cleavage at enzyme recognition sites in DNA, and the variability in the length of the DNA fragment is mainly due to the occurrence of SNPs which create or abolish the short sequences recognised by the specific enzymes used to digest the DNA. The RFLPs show a wide range of allelic frequencies, similar to the classical protein polymorphisms in terms of worldwide distributions and ethnic differences, though as strictly biallelic polymorphisms the maximum heterozygosity at any one site cannot exceed 50%. In the early days, every instance of DNA variation detected by restriction enzyme digestion was referred to as a polymorphism, irrespective of frequency. This led to some confusion about nomenclature.

Restriction enzyme (RE) analysis detects only a proportion of DNA polymorphisms, in the same way that electrophoresis of gene products detects only a subset, about a third of protein polymorphisms, that are due to charge-change amino acid substitutions. On the whole, RE analysis is more effective in detecting polymorphism than gene product analysis since it provides a direct view of non-coding as well as coding DNA sequence. The incidence of DNA variation is much greater in non-coding than coding sequence but it has been difficult to derive precise estimates of the frequency of sites from indirect studies. Early estimates<sup>6,7</sup> suggest there are very roughly one or two simple polymorphic sites in every 1000 human base pairs but more refined estimates are beginning to emerge from deliberate large scale DNA sequence comparisons [Refs. 8 and 9 and see below].

### 1.2.3 Minisatellite Polymorphisms

RE analysis led to the discovery of a second class of polymorphisms in which the fragment length variability is due to variable number tandem repeats (VNTRs). These particular DNA sequences showed some resemblance to the enormously large repetitive DNA satellite regions and thus were christened minisatellites. The most amazing polymorphisms were discovered by Alec Jeffreys in 1985<sup>10</sup> and the electrophoretic patterns of these hypervariable tandem repeats gained enormous importance, and worldwide renown, in the multilocus DNA "fingerprint" technique. The bands were detected on Southern blots after digestion of genomic DNA, often with *HinF1*, followed by hybridisation to a radiolabelled "core" probe. However, these DNA "fingerprints" provide a phenotype not a genotype since although the individual bands show simple Mendelian segregation, there is no information relating alleles to loci or numbers of loci. The information is from multiple unlinked loci and no two bands are likely to be seen co-segregating in a family pedigree.

In contrast, probes which recognise single VNTR minisatellites can produce locus-specific DNA hybridisation patterns, from which genotypes can be deduced in family studies. They occur on all autosomes but are less abundant on the X- and Y-chromosomes being restricted to the X/Y pairing region. These are highly polymorphic loci, most of them display at least three alleles and repeat lengths of between 10 and 100 bp. The average heterozygosity is about 70%, though it can exceed 95%. Loci with heterozygosities greater than 95% tend to be rather unstable markers due to recurrent mutation. Hundreds of human minisatellite loci have been cloned and mapped,<sup>11</sup> and it is estimated that there are between 15,000 to 20,000 such loci in the human genome. However, they have a limited usefulness in gene mapping since the VNTR minisatellite loci tend to cluster in pro-terminal regions of chromosomes or in association with other dispersed repeats. Nevertheless, these polymorphisms have a very significant role in forensic science for the unequivocal identification of an individual and his/her relatives. In this context, it is interesting

to note that there is a further source of DNA variation within these VNTR loci due to small sequence differences among the repeating segments. These SNPs are distributed at intervals along the minisatellite and are known as minisatellite variant repeats (MVRs). Their presence and individual distribution can be revealed in a brilliantly graphic format as a sort of unique personal DNA “bar-code” in the PCR-MVR system devised by Jeffreys and his colleagues.<sup>12</sup>

#### *1.2.4 Microsatellite Polymorphisms*

Another extremely rich source of genetic polymorphism is the microsatellite tandem repeat polymorphisms which display allelic variation in the numbers of copies of short nucleotide repeat sequences. These short tandem repeats (STRs), consisting of di-, tri- and tetranucleotide repeats are currently the most informative and useful marker loci for gene mapping and other forms of genetic analysis.<sup>13,14</sup> This is partly due to the fact that minisatellite polymorphisms are easily scored by the polymerase chain reaction (PCR) technique and are amenable to semiautomation (see below).

Early studies demonstrated the abundance of  $(CA)_n$  dinucleotide polymorphisms, and hybridisation analysis suggested that there are 50,000 to 100,000  $(CA)_n$  repeat blocks between five and 30 repeats long in the human genome. The length of the repeat tends to be correlated with heterozygosity up to about  $n = 21$  and repeats less than  $n = 12$  tend to be monomorphic. The major alleles, that is the most frequent variant alleles, tend to be within  $3n$  of the most predominant allele and the difference in repeat units is usually no greater than  $10n$ . The information content of the  $(CA)_n$  polymorphisms is enormous and their use alone in human analysis provides a map resolution well below 5 cM.

The tri- and tetranucleotide polymorphisms are also extremely valuable STR markers since they can be scored more easily than the  $(CA)_n$  and other dinucleotide repeats. Yet again this has led to practical applications in gene mapping and also in the forensic field, although in the latter case there has been some controversy about

the ethics of establishing DNA databases. Notwithstanding, a standard panel of six STR loci, selected for their stability and high information content and ease of analysis has been introduced by the UK Forensic Science Service within the past few years for routine testing in all serious criminal investigations to aid the identification and apprehension of criminals. A database of DNA genotypes from persons with criminal records is thus being assembled and used to combat crime. There are approximately 400,000 individuals on the current database and the aim is for five million. Current evidence suggest the system has led to several "hits" in identifying criminals involved in multiple crimes such as rape and burglary.

Some trinucleotide DNA sequence repeats have been found to play a significant role in the aetiology of medical genetic conditions, such as the neurological disorders, Huntington disease, HD [OMIM 143100], Spinal bulbar muscular atrophy, SBMA [OMIM 313200], Myotonic Dystrophy, DMPK [OMIM 160900] and the Fragile X syndrome, FMR1 [OMIM 309550], which is associated with mental retardation and a very obvious change in chromosome morphology. These conditions, which may show the unusual genetic phenomenon of "anticipation" in which the age of onset tends to decrease and the severity of a disorder tends to increase from one generation to the next, can be attributed to trinucleotide repeat expansion within or close to their specific gene locus. For example, in myotonic dystrophy the mutational event is the expansion of a CTG repeat within the transcribed region of a gene which encodes a product that is active at the neuromuscular junctions. Healthy people are polymorphic and carry alleles with between five and 30 copies of this repeat, but in individuals with myotonic dystrophy the number of copies is as high as 100 or even 1000 or more in severe cases. The larger the number of copies, the earlier in life is the onset and the more severe the symptoms. This special type of polymorphism, which is probably entirely driven by recurrent mutation due to intrinsic features of the CNG trinucleotide repeat motif, is very different in character from single nucleotide polymorphisms. Indeed, all of the highly repetitive DNA sequence polymorphisms tend towards

mutational instability and thus lie outside or at the very edge of the classical definition of genetic polymorphism.

### **1.3 Functional Significance of Genetic Polymorphisms**

The vast majority of genetic polymorphisms probably have no functional significance whatsoever, when they occur in non-coding sequences far away from functional genes, unless by chance they were to affect the configuration of a significant motif, e.g. in a remote locus control region. Polymorphic variation close to and actually within genes is less frequent and is likely to be more significant. However, more often than not, such polymorphisms appear to have minor functional significance when considered locus by locus. These generalisations are predicted by the data from species comparisons of homologues, which show that all genes can tolerate a significant degree of point mutation, outside the relatively small regions of highly conserved sequence, seemingly without major change in gene function. Major changes in conserved sequences tend to be associated with frank pathology. Other changes, and these include our polymorphisms, probably contribute in varying degrees to the development of normal traits in appearance and behaviour and also to disease susceptibilities in a complex multifactorial pattern.

Direct analysis of enzyme activity in a series of classical isozyme polymorphisms revealed that differences in electrophoretic mobility are often associated with marked differences in catalytic activity but in no case was it possible to identify a phenotype which reflected enzyme overactivity or enzyme deficiency.<sup>1</sup> Studies of this kind reinforced the assumption that most polymorphisms are established in a population by the random processes of genetic drift as opposed to natural selection. There is only a handful of human loci where there is good evidence for a selective agent, malaria, maintaining balanced polymorphisms through advantage to heterozygous carriers of sickle cell haemoglobin [OMIM 141900], thalassaemia [OMIM 141800] and Glucose-6-phosphate dehydrogenase deficiency [OMIM

305900 and see Chap. 7 — Mason & Vulliamy for details]. Each of these polymorphisms occurs at high frequencies in populations with a high risk of exposure to malaria, though there is evidence to suggest that other infections might play a significant role.<sup>15</sup> The Duffy (Fy) blood group [OMIM 110700] is another polymorphism with alleles which show interaction with malaria.

It is very rare to obtain clear cut evidence for natural selection and to identify the agent. For example, the Cystic fibrosis, CF [OMIM 219700] allele,  $\Delta F508$  occurs with polymorphic frequencies in European populations but it is not certain whether this is a transient state, reflecting the action of a formerly prevalent agent such as an infection, or simply chance fluctuation. The haemochromatosis gene [OMIM 235200] is another puzzling example of a “pathological polymorphism” which is maintained at a high frequency in European populations.<sup>16,17</sup> In this case, there are at least two polymorphic alleles *C282Y* and *H63D*, which may lead to iron overload, and it is conceivable that they interact to affect the functional properties of the gene product. A number of examples of allelic interactions have been recorded where a polymorphism can have a marked effect on clinical outcome.<sup>18</sup> For example, a symptom-free polymorphic variant of a human spectrin gene, *SPTA1* [OMIM 182860] occurring *in trans* with a pathological variant produces severe elliptocytosis, but the same substitutions *in cis* are associated with a mild phenotype [Ref. 19, and Chap. 6 — Dhermy for details]. A similar compound effect is observed in the prion protein gene, *PRNP* [OMIM 176640] where an intragenic polymorphism situated upstream of a pathological mutation determines whether the clinical outcome is Creutzfeldt-Jacob disease or fatal insomnia.<sup>20</sup>

Population-based association studies have been used to investigate the functional significance of genetic polymorphism. For example, there are many well-established reports of positive associations of clinical disorders with classical polymorphisms, such as the blood groups and tissue typing antigens. A particularly well-known association is between the *HLA-B27* allele and ankylosing spondylitis [OMIM 106300]. The data are very convincing, 90% of patients have the

allele compared with a 10% frequency in the general population, but it is not yet certain whether the allele directly predisposes to the pathology or whether there is a closely adjacent causative mutation. Questions of this kind are likely to be significant targets in the post-genomic era, when the entire human DNA sequence has been defined in the Human Genome project. In principle, it will then be possible to identify all the common functional genetic polymorphisms and thus provide the vital tools for direct assessment of genetic risk factors in multifactorial disorders.<sup>21</sup>

## 1.4 Detection Techniques

Since the early work on protein and enzyme polymorphism, gel electrophoresis has been the mainstay for detecting and scoring genetic variation at both the protein and DNA levels. Current large-scale gene mapping studies frequently involve the analysis of large panels of STR loci by gel electrophoresis<sup>22,23</sup> and these can yield several hundred genotypes from a single gel. Here, the STR loci are amplified using fluorophore-labelled PCR primers and the polymorphic bands are detected by laser-stimulated fluorescence as they pass by a detector that records all the lanes in the gel. A newer generation genotyping system uses capillary tube electrophoresis instead of gels.<sup>24</sup> Capillary electrophoresis offers the possibility of greater automation since the highly skilled gel-loading step has been replaced with automated capillary-loading.

Although the completion of the first reference sequence of the human genome is still some way off, there is an increasing demand for new polymorphic markers on an unprecedented scale, in anticipation of mapping and establishing the function of the whole genome. Much of the effort is being placed on generating large panels of SNPs. Data emerging from recent detailed studies of human DNA suggest that SNPs are extremely abundant; about one in 600 bp of non-coding and one in 2000 bp of coding DNA may be heterozygous for single nucleotide substitutions.<sup>8,9</sup> Furthermore, the structural simplicity of SNPs in comparison to deletions, insertions

and STRs is ideally suited to full-scale automation for scoring genotypes. However, it should not be forgotten that short insertion/deletion polymorphisms do provide a source of biallelic markers, albeit less abundant than SNPs, that can be readily scored by simple fragment size analysis of PCR products.

A wide variety of established laboratory approaches is used to screen for new polymorphisms including single strand conformation polymorphism (SSCP) analysis, denaturing gradient gel electrophoresis (DGGE), gel-based heteroduplex detection, chemical cleavage mismatch analysis and direct DNA sequencing.<sup>21,25-27</sup> However, new methodologies are steadily flowing into the arena of SNP discovery. Examples of some very promising gel-independent developments include denaturing high performance liquid chromatography (DHPLC), which rapidly and efficiently identifies heteroduplexes, and "MALDITOF" mass spectrometry that compares DNA fragments according to their molecular mass.<sup>21,28</sup>

Bioinformatics gives an alternative opportunity for the *in silico* discovery of SNPs. The colossal and rapidly expanding public sequence database of genomic clones and sequenced tagged sites (STS), including expressed sequence tags, increasingly represent overlapping DNA sequence from several different individuals and tissue sources. These sequence overlaps are proving to be rich resources for SNP discovery<sup>9</sup> and short insertion/deletion polymorphisms can also readily be identified in this way.

The development of methods for high throughput scoring of individual human genetic variation on a scale to match that of the whole genome is a major challenge for the future. In principle, all of the laboratory mutation detection methods may be used to type polymorphisms although dedicated "read out" technologies offer the advantage of speed and automation. Current high throughput methods include "DNA mini-sequencing" which involves allele-specific nucleotide incorporation by primer extension, allele-specific hybridisation/5' nuclease digestion that utilises emission or quenching of a fluorescent signal and allele-specific oligonucleotide ligation.<sup>21</sup> These methods are all used in solid-phase array formats, often based

around 96- or 384-well microtitre plates, and are capable of providing highly efficient assay systems for polymorphism in PCR products.

The medium term future is likely to witness the widespread introduction of the genotyping microchip or biochip.<sup>29</sup> Microchips represent the extreme miniaturisation of molecular array technology. Chips contain arrays of thousands of detecting oligonucleotides attached to the surface of the microchip. A genotyping chip with allele-specific arrays for over 500 SNP loci has been produced<sup>9</sup> and the simultaneous scoring of 2000 SNP genotypes on a single chip is regarded by some as a realistic target for the millennium. To use a genotyping chip, multiplex PCR of SNP loci from an individual is carried out, the products are labelled with fluorescence and introduced to the chip. The result is a differential hybridisation pattern that is read by confocal microscopy and analysed by complex algorithms. Enzymic reactions using polymerases or ligase may be useful alternatives to differential hybridisation for analysing DNA variation on microarrays. Biochips may offer the solution to the massive case control studies necessary in the hunt for susceptibility genes for common diseases although one significant drawback may be their high cost.

## **1.5 Databases and Other Genetic Resources**

The Online Mendelian Inheritance in Man (OMIM) database is a timeless, regularly updated electronic resource of common and rare genetic conditions closely integrated with information on every human genetic polymorphism which has a phenotype. Population data on allele frequencies of blood groups, serum protein and isozyme polymorphisms, together with some information on RFLPs are tabulated in Roychoudhury and Nei.<sup>30</sup> Data of this kind are becoming more and more accessible via the internet. For example, the UK Human Genome Mapping Project Resource Centre, <http://www.hgmp.mrc.ac.uk/> contains hyperlinks to a vast number of genome project resources that are in the public domain including the OMIM database referred to above. There are also numerous

databases of genetic markers including locus specific information, mutations found in genetic diseases, genetic maps constructed from linkage analysis of STRs and other markers, as well as growing collections of SNP data of which four examples are shown below:

Human Genic Biallelic Sequences (HGBASE):

<http://hgbase.interactiva.de/>

SNP Polymorphism Repository (NIH):

<http://www.ncbi.nlm.nih.gov/SNP>

Human SNP Database at the Whitehead Institute:

<http://waldo.wi.mit.edu/SNP/human>

SNPs in the Human Genome (Washington University):

<http://www.ibr.wustl.edu/snp>

## References

1. Harris H., *The Principles of Human Biochemical Genetics* (Elsevier/North Holland, Amsterdam, 1980), pp. 329–379.
2. Ford E.B., Polymorphism and Taxonomy, in *The New Systematics* (ed.) Huxley J. (Clarendon Press, Oxford, 1940), pp. 493–513.
3. Fisher R.A., *The Genetical Theory of Natural Selection* (Clarendon Press, Oxford, 1930).
4. Smith R.A., Ho P.J., Clegg J.B., Kidd J.R. and Thein S.L., *Blood* **92** (1998), 4415–4421.
5. Yip S.P., Lovegrove J.U., Hopkinson D.A. and Whitehouse D.B., *Hum. Mol. Genet.* **8** (1999), 1699–1706.
6. Jeffreys A.J., *Cell* **18** (1979), 1–10.
7. Cooper D.W. *et al.*, *Hum. Genet.* **69** (1985), 201–205.
8. Nickerson D.A. *et al.*, *Nature Genetics* **19** (1998), 233–240.
9. Wang *et al.*, *Science* **280** (1998), 1077–1082.
10. Jeffreys A.J., Wilson V. and Thein S.L., *Nature* **314** (1985), 67–73.
11. Armour J.A.L., Povey S., Jeremiah S. and Jeffreys A.J., *Genomics* **8** (1990), 501–512.
12. Tamaki K. *et al.*, *Hum. Mol. Genet.* **1** (1992), 401–406.
13. Weber J.L. and May P., *Amer. J. Hum. Genet.* **44** (1989), 388–396.
14. Weissenbach J. *et al.*, *Nature* **359** (1992), 794–801.

15. Allen S.J. *et al.*, *Proc. Natl. Acad. Sci.* **94** (1997), 14736–14741.
16. Merryweather-Clarke A.T., Pointon J.J., Shearman J.D. and Robson K.J., *J. Med. Genet.* **34** (1997), 275–278.
17. Worwood M., *Clin. Lab. Haematol.* **20** (1998), 65–75.
18. Edwards Y.H. and Swallow D.M. Mutation and Protein Dysfunction, in *Protein Dysfunction in Human Genetic Disease* (eds.) Swallow D.M. and Edwards Y.H. (Bios, Oxford, 1997), pp. 1–14.
19. Lecomte M.C., Delaunay J. and Dhermy D., Spectrin and Other Red Cell Membrane Proteins in Hereditary Elliptocytosis and Spherocytosis, in *Protein Dysfunction in Human Genetic Disease* (eds.) Swallow D.M. and Edwards Y.H. (Bios, Oxford, 1997), pp. 203–218.
20. Petersen R.B., Parchi P., Capellari S. and Gambetti P., Fatal Fetal Insomnia, Creutzfeldt-Jacob Disease and the Prion Protein, in *Protein Dysfunction in Human Genetic Disease* (eds.) Swallow D.M. and Edwards Y. H. (Bios, Oxford, 1997), pp 243–254.
21. Schafer A.J. and Hawkins J.R., *Nature Biotech.* **16** (1998), 33–39.
22. Reed P.W. *et al.*, *Nature Genetics* **7** (1994), 309–395.
23. Ziegler J.S. *et al.*, *Genomics* **14** (1992), 1026–1031.
24. Wenz H. *et al.*, *Genome Res.* **8** (1998), 69–80.
25. Hayashi K. and Yandell D.W., *Human Mutation* **2** (1993), 338–346.
26. White M.B. *et al.*, *Genomics* **12** (1992), 301–306.
27. Cotton R.G., *Am. J. Hum. Genet.* **59** (1996), 289–291.
28. Underhill P.A. *et al.*, *Genome Res.* **7** (1997), 996–1005.
29. Castellino A.M., *Genome Res.* **7** (1997), 943–946.
30. Roychoudhury A.K. and Nei M., *Human Polymorphic Genes — World Distribution* (Oxford University Press, Oxford, 1988).