

# Genetic Variation: Polymorphisms and Mutations

Alan F Wright, *MRC Human Genetics Unit, Edinburgh, UK*

The amount of sequence variation in different regions of the human genome varies by an order of magnitude. Mutations give rise to all variation, but their survival in the genome is influenced by many factors including effects on reproductive fitness, human population history, chromosomal location and recombination rates.

## Introduction

Genetic variation is generated continuously by the mutational process, but its persistence in the genome is determined by different historical and genomic factors. Some of these factors leave an imprint on sequence variation across the whole genome, others only influence local patterns of variation. A new variant with a beneficial effect on reproductive success (fitness) can leave a detectable imprint on the local pattern of genetic variability, although this may extend for some distance from the selected variant. By contrast, the pattern of variability across the whole genome may be influenced by demographic factors such as a population bottleneck, in which the current population is descended from a few antecedents. A new mutation will, on average, persist for a longer period of time if it has a beneficial effect on reproductive fitness. Similarly, it will persist more readily in a large population than in a small population, because it is less likely to be lost as a result of sampling effects. In addition, the overall extent of genetic variability is strongly influenced both by mutation rates, which vary across the genome, and by population size. Large populations tend to show more variability than do small populations, because they generate more mutations.

Most new mutations that affect gene function have deleterious effects on reproductive fitness. But because genes represent only a small fraction of the human genome, most mutations are thought to have no effect on reproductive fitness and are effectively invisible to natural selection – a category referred to as ‘selectively neutral’. Most deoxyribonucleic acid (DNA) variants in the human genome are thought to be selectively neutral for three main reasons. First, the main portion of the genome, estimated as about 97%, neither codes for a functional product, such as protein or ribonucleic acid (RNA), nor indirectly affects gene function, by regulating expression or replication. Second, if a new variant does occur in the 1.5% of the genome that encodes a functional product (coding regions), it may not result in a change of amino acid (i.e. it may be a ‘synonymous’ substitution). Third, variants that do affect regulatory regions or coding regions and do

## Introductory article

### Article contents

- Introduction
- Mutation and Genetic Variation
- Types of Genetic Variation
- Distribution of Variants
- Genetic Variation and Ethnicity

doi: 10.1038/npg.els.0005005

change an amino acid (nonsynonymous substitutions) may have no effect on reproductive fitness.

By contrast, a very few mutations have large functional effects, most of which are deleterious to the organism. The extreme case is that of lethal mutations, which are destined to disappear from the population but may still persist for many generations (in heterozygous form) if they are partially or completely recessive. It is estimated that every individual carries one or two recessive lethal mutations in heterozygous form. Mutations that are advantageous persist for longer periods of time and their frequency in the population will tend to be higher (e.g. >1%) than that of deleterious variants, which tend to be rare (<<1%).

Between the extremes of selectively neutral and lethal mutations lie those variants that influence physiological, morphological and pathological variation in the human population. These variants provide the genetic substrate for evolutionary change in response to infection, starvation, stressful or extreme environments, and sexual or other forms of behavioral selection. Most such functional variants are likely to be selectively deleterious, but their effects on reproductive success may be so small that they can still persist in the human population for long periods of time, especially if they are recessive. Most naturally occurring variants in experimental organisms are partially recessive, and there is good reason to believe that this also applies to humans. Other variants are deleterious in some contexts (e.g. stages of development or environments) but advantageous in others – a situation that can lead to their maintenance in the genome at a relatively high population frequency.

## Mutation and Genetic Variation

The source of all genetic variation lies in the mutational process, which occurs at different rates in

different parts of the genome and in the male and female germ lines (sperm and eggs). Mutation rates in humans vary across the genome from about  $10^{-7}$  to  $10^{-9}$  per nucleotide base per generation. These rates are roughly equivalent to  $10^{-4}$ – $10^{-6}$  per gene per generation. There are  $6 \times 10^9$  nucleotides in the human genome so that a continuous input of new mutations occurs at an estimated average rate of about 200 per diploid genome (zygote) per generation. Although the coding regions of genes make up only about 1.5% of the genome, as much as 3% of the genome may have small effects on genome function or expression. This is shown by the fact that many DNA sequences in noncoding regions show unexpectedly high sequence identity (conservation) across different species, suggesting that they do have a function, such as regulating chromatin structure or gene expression.

Most human traits, such as height or disease susceptibility, are genetically complex and influenced by many genetic and environmental factors (i.e. they are multifactorial). The distribution of genetic effects in such traits is generally thought to be L-shaped (leptokurtotic), with many variants of small effect and few variants of larger effect. There are many more genomic sites at which a variant can exert a small or functionally remote effect on a trait than there are sites at which it can exert a large or direct one. Also increasingly recognized are complex alleles, whereby specific effects on a trait result from combinations of variants at several sites in the gene (sometimes even spanning adjacent genes). Finally, variants often influence several different traits or biological processes, and the magnitude of these effects can vary.

A small but medically important group of traits is strongly influenced by single genes. Whereas 1–2% of the human population is affected by Mendelian or monogenic traits, more than 60% will develop a multifactorial disorder at some time in life. A list of genetic variants that result in monogenic disorders is compiled in the Human Gene Mutation Database (see Web Links), which documents 28 000 variants in more than 1100 genes.

Genes that show high mutation rates contribute disproportionately to human disease. The highest mutation rates ( $\mu$ ) for genes causing human disease are estimated to be in the region of 5–10 per 100 000 gametes per generation. Examples include the polycystic kidney disease 1 (*PKDI*) gene, the most common cause of adult polycystic kidney disease, which affects one per 1000 of the general population ( $\mu = 7$ – $12 \times 10^{-5}$  per generation); the neurofibromin 1 (*NFI*) gene, which causes type 1 neurofibromatosis and affects one in 3000 ( $\mu = (4$ – $10) \times 10^{-5}$  per generation); and the dystrophin (*DMD*) gene, the cause of Duchenne muscular dystrophy, which affects one in 3500 males ( $\mu = (4$ – $11) \times 10^{-5}$  per generation). Several

of these genes are large and so provide large ‘targets’ for the mutational process.

Most genes have substantially lower mutation rates, such that variants are rarely identified unless they have severe effects. Factors influencing the variability of mutation rates across the genome are poorly understood, but high rates are found in certain identifiable sequences, such as CG dinucleotides, some purine-rich and repetitive sequences. For example, a single CG-containing codon in the fibroblast growth factor receptor 3 (*FGFR3*) gene is associated with a glycine to arginine mutation at residue 380 (G380R) in all individuals with achondroplasia, the most common cause of short-limbed dwarfism. The mutation rate at this codon is extremely high (at  $10^{-5}$  per generation). Different mutations in the *FGFR3* gene also give rise to a diverse range of phenotypes, including seven differently named disorders that affect either skeletal or skin development – a situation that is not uncommon in human genetics.

## Types of Genetic Variation

### Single nucleotide variants

By far the most common type of mutational change is the single nucleotide (nt) substitution, in which a nucleotide changes from one purine base to another (e.g. adenine to guanine), which is called a transition, or from a purine to a pyrimidine base (e.g. guanine to cytosine, or vice versa), which is called a transversion. A single nucleotide polymorphism (SNP) is a specific class of single nucleotide substitution that has the additional property of being common in the population. A ‘polymorphism’ is defined as a sequence variant that has a population frequency of at least 1%. There are an estimated 3–10 million SNP variants in the human genome with a frequency higher than 1%.

A widely used measure of genetic variability is ‘nucleotide diversity’ ( $\pi$ ), which is defined as the frequency with which any two sequences from a random sample of the population differ at a particular nucleotide site. It is not a good measure of diversity resulting from rare mutations, but it is a useful measure of the diversity owing to common variants such as SNPs. The average value of  $\pi$  in the human genome is 0.0008, which means that 1 nt substitution is expected every 1250 nt base pairs (bp). Most of these variants are SNPs that lie outside gene coding regions, so they are expected to be selectively neutral. This measure of diversity should be qualified further, because the value of  $\pi$  also depends on the population frequency of the variant. A selectively neutral SNP is predicted to occur every 500 bp for those with allele

frequencies in the 1–10% range every 1500 bp for those in the 10–20% range, and every 3000 bp for those in the 40–50% range. Most detected SNP loci show two common alleles with frequencies of at least 10–20%. Many of these are predicted to be old, predating the emergence of anatomically modern humans from Africa about 100 000 years ago. Other measures of single nucleotide diversity are more sensitive to the presence of rare variants and take into account the fact that large samples may be required to detect them (e.g. the ‘number of segregating sites’).

Most noncoding SNPs are likely to be nonfunctional and selectively neutral, but there is an important subgroup that can influence multifactorial diseases or traits. **Table 1** gives examples of SNPs influencing human disease-related traits. A prototype example is the apolipoprotein E (*APOE*) gene, which has three SNP alleles (*E2*, *E3* and *E4*) resulting in variant proteins (apoε2, apoε3 and apoε4) with different amino acids at two sites (arginine or cysteine at residues 112 or 158). The *E4* and *E2* variants are each implicated in human disease. More than 90% of individuals who have two copies of the *E2* allele (homozygotes) are clinically normal, but the rest develop type III hyperlipidemia, with abnormal blood lipid levels that predispose them to premature arterial disease. Other factors seem to determine whether an *E2* homozygote develops arterial disease, such as obesity, age, gender and hormonal influences, so this is an example of a variant influencing a multifactorial trait.

The *E4* allele is associated with another genetically complex disease, Alzheimer disease, in which it seems to advance the onset of dementia in susceptible individuals by between 5 years (in carriers of one *E4* copy) and 15 years (in carriers of two *E4* copies). The relative risk of developing Alzheimer disease in *E4* carriers compared with noncarriers is about 3 for the roughly 25% of Western populations who carry one *E4* copy, and 6–15 for the roughly 2% who carry two *E4* copies. The protein product of the *E4* allele, apoε4 (with arginine at residues 112 and 158), is present in all global populations at frequencies that vary from 5% to 35%. Almost all animals, including our closest ancestors the great apes, have arginine at residue 112, and only humans are known to have variant forms of the protein containing a cysteine at this site (in apoε2 and apoε3) and either cysteine (apoε2) or arginine (apoε3) at residue 158. This suggests that an ancestral *E4* allele mutated first to *E3* and then to *E2* in the past 5 million years. A single base substitution, with a C to T transition at codon 112 would change *E4* to *E3* (now the most common allele), and a further C to T transition at codon 158 in *E3* would result in *E2* (the least common allele). It has been proposed that *E4*

became disadvantageous, perhaps during the growth of urbanization in Middle Eastern centers of agriculture, and the previously neutral *E3* and *E2* alleles (otherwise destined to be lost) became favorable. Other explanations are possible, such as the chance survival of otherwise rare variants when human population numbers were small.

An example of an SNP that seems to have risen to high frequency in a single population because it conferred a selective advantage is the Duffy blood group allele, *FY\*O*, which protects against malaria resulting from *Plasmodium vivax* infection. This allele causes loss of the red blood cell FY antigen as a result of mutation in an ‘enhancer’ sequence regulating gene expression. Its frequency reaches 100% in sub-Saharan African populations, but it is essentially absent elsewhere. Similarly, sickle cell (*HBB\*S*) and other β-globin gene variants of hemoglobin seem to protect against *Plasmodium falciparum* malaria. About 400 million people are estimated to carry at least one copy of such malaria-resistance variants in the hemoglobin, alpha (*HBA1* and *HBA2*) or hemoglobin, beta (*HBB*) globin genes – about 1–15 worldwide – resulting in an enormous burden of childhood anemia. Many of these variants are thought to have risen to such high frequencies in the past 10 000 years because of their strong selective advantage.

### Rare variants

Most SNPs occur in noncoding regions of the genome, but an estimated 10 000–50 000 SNPs occur in coding regions and give rise to a change in amino acid in the protein product. These coding SNPs (cSNPs) are therefore capable of directly influencing gene function and represent an important class of variant, with potential relevance to morphological, physiological and pathological traits. Most coding variants are likely to be selectively deleterious but the more frequent subgroup of cSNPs may represent the least deleterious or neutral end of the spectrum. Current estimates suggest that 1–2 deleterious mutations arise per zygote per generation as a result of such nonsynonymous mutations. A subset of these increases in frequency (>1%) to become cSNPs, but most remain at low frequency. It is estimated that about 20% of nonsynonymous coding variants (many of which are cSNPs) are selectively neutral; of those that are slightly deleterious, more than 80% are expected to be present at frequencies below 1%.

Most new mutations are lost with a probability of about  $e^{-1}$  (0.37) per generation as a result of stochastic sampling, but their persistence depends on the ‘effective population size’ ( $N_e$ ). This is defined as the number of individuals in a population that

**Table 1** Examples of allelic diversity in human disease<sup>a</sup>

Locus	Allele	Trait	Frequency	Effect	Comments
<b>Common variants influencing human disease</b>					
<i>Cardiovascular</i>					
<i>APOE</i>	* <i>E4</i>	Alzheimer disease	0.10–0.15 (Caucasian)	Early onset	Allele present in primates and all world populations; may account for 20% of Alzheimer disease
		Age-related macular degeneration	0.10–0.15	Decreased risk	Well-established protective effect on age-related macular degeneration
		Cardiovascular disease	0.10–0.15	Increased risk	Accounts for 10–16% of cholesterol variance in Western populations; increases risk of cardiovascular disease (odds ratio ≈1.5)
<i>F5</i>	R506Q	Venous thrombosis	0.02–0.08	Increased risk	Carriers have around 10% lifetime risk for significant venous thrombosis
<i>Metabolic/nutritional</i>					
<i>PPARG</i>	P12A	Type 2 diabetes mellitus	0.85 (Caucasian)	Increased risk	Relative risk 1.25
<i>HFE</i>	C282Y	Hemochromatosis	0.05 (Caucasian)	About 40% risk for homozygotes	High frequency in Caucasians, low in Asiatics, suggesting that this may be a recent mutation (< 50 000 years)
<i>CFTR</i>	ΔF508	Cystic fibrosis	0.04 (Caucasian)	Homozygotes are affected	Accounts for 66% of CF chromosomes in Caucasians, >1000 rare alleles also found.
<i>Cancer</i>					
<i>ELAC2</i>	S217L and A541T	Prostate cancer	0.30 and 0.04 (Caucasian)	Increased risk	Odds ratio 2.4–3.1
<i>BRCA2</i>	N372H	Breast cancer	0.22–0.29 (Caucasian)	Increased risk	Relative risk 1.31 for HH compared to NN genotypes
<i>Infectious/inflammatory</i>					
MHC class I	<i>HLA-B*2702, 04, 05</i>	Ankylosing spondylitis	0.09 (Caucasian)	Increased risk	Odds ratio ≈170; mechanism unclear; also associated with reactive arthritis and uveitis; about 2% of B27-positive carriers develop ankylosing spondylitis
MHC class II	<i>DQB1*0302–DRB1*0401/DQB1*0201–DRB1*03</i>	Type 1 diabetes mellitus	0.05 (European)	Increased risk	About 10% of heterozygotes for these high-risk haplotypes develop type 1 diabetes mellitus; relative risk ≈20
<i>IL12B</i>	3' UTR allele 1	Type 1 diabetes mellitus	0.79 (Caucasian)	Increased risk	Interaction with <i>HLA</i> ; increased expression of <i>IL12B</i> <i>in vitro</i>
<i>NOD2</i>		Inflammatory bowel disease	0.04 (3020insC, G881R); 0.01 (R675W)	Increased risk (threefold one copy, 40-fold two copies for 3020insC)	Western populations, but not Japanese, show three common variants predisposing to inflammatory bowel disease, and many rare variants
<i>G6PD</i>	A <sup>-</sup> (V68M/N126D)	G6PD deficiency	~0.20 (West African)	Decreased risk of severe malaria	High allele frequency proposed to be due to balancing selection
<i>HBB</i>	<i>HGS</i> (E6V)	Anemia (homozygotes)	0.12 (West African)	Decreased risk of severe malaria	High allele frequency proposed to be due to balancing selection
<i>CCR5</i>	Δ32- <i>CCR5</i>	Human immunodeficiency virus 1 transmission	0.09 (Caucasian)	Decreased HIV-1 transmission	Recent origin, estimated about 700 years ago
<i>Developmental</i>					
<i>PDGFRA</i>	Promoter <i>H1/H2α</i> haplotypes	Neural tube defect	0.23 (Caucasian)	Increased risk for sporadic neural tube defect	At least six polymorphic sites in each haplotype
<i>MTHFR</i>	A222V	Neural tube defect	0.38 (Caucasian)	Increased risk for homozygotes (val/val) in presence of low red blood cell folate	Interaction between genotype and environment (dietary folate)

Table 1 Continued

Locus	Allele	Trait	Frequency	Effect	Comments
<i>Metabolic/nutritional</i>					
<i>CFTR</i>	>960 alleles	Cystic fibrosis	Most rare, but $\Delta F508$ common in Caucasians, accounting for ~70% of CF alleles	High risk	$\Delta F508$ allele estimated as ~3000 years old
<b>Rare variants influencing human disease</b>					
<i>Cardiovascular disease</i>					
<i>LDLR</i>	>735 alleles	Coronary artery disease (CAD)	All rare, except in isolate or founder populations	High risk of CAD	
<i>APOB</i>	>32 alleles	Coronary artery disease	R3500Q with frequency 0.002, remainder rare	High risk of CAD	
<i>Cancer</i>					
<i>BRCA1</i>	>1200 alleles	Familial breast/ovarian cancer	All rare, except in isolate or founder populations	High risk	
<i>BRCA2</i>	>1400 alleles	Familial breast cancer	All rare, except in isolate or founder populations	High risk	Only one common allele (N372H) with a small increase in relative risk (1.31)
<i>MLH1</i>	>168 alleles	Hereditary nonpolyposis colorectal cancer (HNPCC)	All rare	High risk	
<i>MSH2</i>	>130 alleles	HNPCC	All rare	High risk	
<i>TP53</i>	>144 alleles	Multiple cancers	All rare	High risk	
<i>Neurosensory</i>					
<i>ABCA4</i>	>350 alleles	Stargardt disease, retinitis pigmentosa	Most rare, G863A allele ~0.014 (Europeans)	High risk	
<i>RHO</i>	>100 alleles	Retinitis pigmentosa, congenital stationary night blindness	All rare	High risk	
<i>GJB2</i>	>59 alleles	Nonsyndromic deafness	Most rare, 30delG allele around 0.015 (Europeans)	High risk	30delG absent from non-European populations

<sup>a</sup>Data are from Online Mendelian Inheritance in Man and the Human Gene Mutation Database (see Web Links).

contribute genes to succeeding generations and that account for the observed random fluctuations in gene frequency resulting from sampling effects (genetic drift). Human  $N_e$  has been estimated consistently to be in the region of 10 000 on the basis of measurements of human sequence variability, despite the fact that the true population size is six billion. (The reason for this seems to be the recent exponential growth of human populations, the impact of which on genome diversity is not readily detectable in small sequence surveys.) Most deleterious mutations are destined to be lost in a timescale in the region of the natural logarithm of  $N_e$  – that is, in about 10 generations – but with a very wide distribution around this value. Such mutations are therefore held at low frequency in the population, close to that expected if there is an equilibrium

between input of new mutations and loss by natural selection (mutation–selection balance), assuming an infinitely large population.

These results suggest that each of us probably carries several hundred slightly deleterious mutations, most of them at low population frequency and predominantly in heterozygous form. Analyses of SNP frequency distributions in human populations show that nonsynonymous variants in coding regions have significantly lower frequencies than synonymous variants, suggesting that most are at least mildly deleterious from the point of view of reproductive fitness.

For neutral variants, there are likely to be as many variants with a frequency below 1% as above it (on the basis of estimates of human  $N_e$ ). With deleterious

variants, the proportion present at low population frequencies (<1%) is considerably greater. There is another factor that contributes to the predominance of rare variants in the human population, although it is hard to quantify experimentally. The expansion of the human population from a few tens of thousands to six billion in the past 100 000 years has generated an enormous diversity of mutations. Few of these have had time to spread and equilibrate in the population so that, in the absence of strong selection, most are currently at low frequency. In short, most variant sites in the human genome are predicted to be rare, with many of them occurring in only a few individuals and unlikely to be detected in small sequence surveys. These rare and deleterious variants, especially those with large effects on a disease-related trait (e.g. blood pressure or extreme obesity), often come to medical attention and so tend to be overrepresented in disease populations.

Most mutations implicated in severe monogenic disorders that influence reproductive fitness are rare, with a frequency close to that predicted by a mutation–selection balance. These disorders typically result from rare alleles arising at many different sites in the gene. **Table 1** gives examples of this allelic diversity in a range of monogenic and genetically complex diseases. In the extreme case of a dominant disorder that is lethal, each causal mutation is unique, occurring *de novo* in one of the parental gametes. What is more surprising is that many adult-onset disorders, with no evident effects on reproductive success, are also associated with a diverse set of rare alleles. For example, mutations in the breast cancer, early onset 1 (*BRCA1*) and breast cancer, early onset 2 (*BRCA2*) genes, which account for a high proportion of familial and monogenic forms of breast cancer (although less than 3% of all breast cancer), are extremely diverse: there are 1200–1400 known disease-causing alleles, but only 6–7 common polymorphic (SNP) alleles, only one of which shows a very small effect on cancer susceptibility.

The contribution of rare variants to the heritable component of genetically complex, multifactorial traits is likely to be significant, although it is currently unknown whether or not they predominate over those that are more common (e.g. SNPs). The hypothesis that SNPs rather than rare variants provide the main genetic substrate for common clinical disorders (the ‘common disease common variant’ hypothesis) is currently a main focus of research.

## Insertion and duplications

Other common forms of genetic variation include insertions, deletions and inversions of one or more bases. Estimates suggest that 5–10% of the human

genome is duplicated one or more times. These regions include both relatively large DNA segments of 10–400 kilobases (kb) that are duplicated a few times throughout the genome, and a large class of short (1–5 bp), dispersed and tandemly repeated segments. The latter show a relatively high rate of insertion or deletion mutation and so commonly show polymorphic variation, with different individuals showing different numbers of short tandem repeat (STR) sequences, such as  $(A)_n$ ,  $(CA)_n$  or  $(AAG)_n$ , where  $n$ , the number of repeats, is variable. This class of repetitive sequence is important in human genetics.

Microsatellite repeats or STRs are an important source of genetic markers that have helped to revolutionize the mapping of human disease genes. They are, in general, selectively and functionally neutral (except for certain trinucleotide repeats discussed below) and are widely dispersed in noncoding regions of the genome. The most widely occurring dinucleotide repeat,  $(CA)_n$ , is present at an estimated 50 000–100 000 sites per genome. The  $(CA)_n$  repeats occur on average every 30–60 kb, of which about 8000 have been identified and are listed in human genome databases. Tri- and tetranucleotide repeats are also useful genetic markers, and the collective frequency of all microsatellites is one in every 10–30 kb. They provide highly informative markers for ‘tagging’ specific segments of the genome, which then can be tracked through families. This provides the necessary information to measure the genetic distance between the marker and its associated trait, as is required for genetic mapping by linkage analysis.

A few STRs are neither functionally nor selectively neutral because they occur in functionally important regions of the human genome and give rise to human disease. These trinucleotide repeat diseases generate abnormally large alleles that interfere with the expression or function of associated genes. They encompass four main types. The first is represented by a group of 14 progressive neurological disorders resulting from expansion of  $(CAG)_n$  repeats located in the exonic (coding) region of the genes concerned. The CAG trinucleotide in an exon encodes the amino acid glutamine, so that pathologically extended tracts of polyglutamine occur in the protein. A second type of trinucleotide expansion is associated with a CTG motif in the downstream (or 3′) untranslated region of the dystrophin myotonic-protein kinase (*DMPK*) gene, which interferes with gene expression in this and possibly neighboring genes. A third type also occurs in a noncoding region, but in this case it is in the upstream (or 5′) untranslated ‘promoter’ region, which initiates and regulates transcription. This and other trinucleotide expansions are unstable and can increase in size over one or a few generations during meiosis, from a mildly abnormal expansion (premutation) to a

severely abnormal and highly unstable state. In this situation, the disease mechanism involves abnormal modification of CGG sequences by methylation, which switches off the gene promoter. The fourth form of trinucleotide expansion is associated with a GAA repeat expansion in an intron – the noncoding region of a gene that separates exons. It seems to cause disease by forming abnormal DNA structures that interfere with transcription of the associated gene by RNA polymerase.

Another class of polymorphic repeat is the mini-satellite or ‘variable number tandem repeat’ sequence. These are tandemly repeated sequences of 5–64 bp that often extend over several kilobases of DNA and occur at more than 10 000 sites throughout the genome. They are less evenly distributed than are microsatellites, but are highly informative markers owing to the very large number of alleles in the population. They are, therefore, particularly important in forensic work.

Large genomic duplications (e.g. 10–400 kb) can also give rise to variation between individuals, as a result of deletion or further duplication. These can result in genomic instability, owing to unequal meiotic pairing of homologous segments. Most occur in genetically ‘silent’ regions and so have no untoward effect; however, some lead to duplication or deletion of important genes and give rise to disease. The progressive muscle wasting disorder type 1A Charcot–Marie–Tooth disease is commonly associated with a large genomic duplication of a region of 1.5 megabases (Mb), which includes the myelin-associated peripheral myelin protein 22 (*PMP22*) gene on chromosome 17.

## Deletions

Short repetitive regions such as microsatellites are associated with both insertion and deletion of repeats. Deletion of repetitive sequences also occurs in and between genes. The *DMD* gene on the X chromosome is the largest known human gene, with 79 exons spanning 2400 kb. In Duchenne muscular dystrophy (DMD), 65% of affected males have *DMD* deletions and 6–7% have duplications involving contiguous exons. A third of all mutations in DMD are new or *de novo* because the disease is not compatible with survival beyond 25 years of age. The large size of the *DMD* gene contributes to its high mutation rate. This is also the case with the *NF1* gene, although it also contains a duplicated sequence of 85 kb that is associated with large (1.5 Mb) deletions in 5–20% of affected individuals. Another fatal muscle wasting disease of infancy, type 1 spinal muscular atrophy, results from mispairing and subsequent deletion of the region lying between two adjacent and recently duplicated genes (survival of motor neuron 1, telomeric;

*SMN1* and survival of motor neuron 2, centromeric; *SMN2*) that are in inverted orientation and 99% identical. About 1 in 80 people are asymptomatic carriers of causal *SMN1* deletions, whereas affected homozygotes occur at a frequency of one in 10 000.

A polymorphic deletion variant ( $\Delta F508$ ), which is responsible for a major Mendelian disorder, is found in the cystic fibrosis transmembrane regulator (*CFTR*) gene. This variant arose by deletion of 3 bp at codon 508 in a functionally important domain of the protein, resulting in loss of the amino acid phenylalanine (symbol F) at this site. The  $\Delta F508$  mutation is present (in one copy) in asymptomatic carriers in 4% of Caucasians, but causes a severe obstructive lung and pancreatic disorder, cystic fibrosis, when present in homozygous form (two copies). Cystic fibrosis affects one in 2000 Caucasians but is rare in Asian and African populations. The high frequency of this deleterious variant is puzzling. One possibility is that it conferred a selective advantage to carriers during infectious epidemics such as typhoid fever in Europe in the past 10 000 years, thereby rising to its present high frequency. Alternatively, the high frequency might have arisen through genetic drift (the change in frequency of a variant as a result of sampling processes or ‘chance’) when the population size was small.

Deletion of large genomic segments containing many genes is rarely compatible with survival. Occasionally this occurs in relatively gene-poor regions of the genome and gives rise to ‘contiguous gene syndromes’. An example is the WAGR syndrome, which is associated with a type of early-onset kidney tumor (Wilm tumor), aniridia (absent iris) and genito-urinary and renal abnormalities, owing to deletion of several genes in chromosomal region 11q13.

## Inversions

Small sequence inversions are not uncommon but most have no functional consequences unless they occur in a gene. Inversion of large chromosomal segment is less common but occasionally gives rise to disease. The X-linked disease hemophilia A is associated with deficiency of factor VIIIc, which is essential for normal blood clotting. A common cause of severe factor VIIIc deficiency results from a ‘flip inversion’ in the coagulation factor VIII-associated (*F8C*) gene. This is due to the presence of several copies of a small gene (gene A) near the chromosomal end or telomere of the X-chromosome long arm, one of which occurs in intron 22 of the *F8C* gene. Recombination between the intronic and distal copies of gene A results in inversion of the distal part of *F8C* with consequent loss of function.

## Distribution of Variants

The availability of the human genome sequence has made it possible to investigate the distribution of SNP variants and their relationships in different individuals and populations. Assuming an idealized random mating population, the alleles at different SNP loci are expected to be associated with each other at random in the population, in accordance with their allele frequencies (linkage equilibrium). However, this assumes that, first, the loci are far apart on the same chromosome (unlinked) or that they are on different chromosomes (unlinked); or second, they are close together on the same chromosome (linked) but there has been sufficient time for exchange and randomization of alleles between gametes as a result of recombination. Samples from human populations tend to show an excess of nonrandomly associated SNP alleles (in linkage disequilibrium) when compared with the expectation based on human population history. A possible explanation is that a recent population ‘bottleneck’ substantially reduced the number of generations that separates us from the common ancestor, so that there has been insufficient time for randomization of adjacent SNP alleles to occur. Alternatively, ‘stratification’ of human populations into nonrandom mating groups could prevent the exchange of alleles between pairs of loci and slow the process of allelic exchange between chromosomes.

Evidence suggests that human population bottlenecks did indeed occur, both in the past 30 000–50 000 years in northern Europeans, and around 140 000 years ago, when anatomically modern humans emerged from Africa. This last event is consistent with the greater extent of linkage equilibrium and greater nucleotide diversity in sub-Saharan Africans (who had no such bottleneck) than in other human populations (see below). The number of common SNP haplotypes – combinations of SNP alleles on a single chromosome – found in human populations is consistently less than expected, again suggesting that population bottlenecks have occurred, thereby reducing the diversity of SNP haplotypes.

SNP haplotypes tend to show a block-like structure, with combinations of SNP alleles invariantly associated on the same chromosome (i.e. in linkage disequilibrium). These blocks are generally small (10–100 kb) and partly reflect the recent common ancestry of most humans. For example, there is a 40–50% chance that any two copies of a 10-kb sequence taken from members of the population have been inherited (without recombination) from a single common ancestor.

There is, however, another explanation for the existence of such haplotype blocks of common

variants. Growing evidence indicates that there are preferential sites of recombination (‘hot spots’) on human chromosomes, whereas previously it was assumed that recombination could occur at any site more or less at random. For example, about 95% of recombination events in a 200-kb region in the major histocompatibility complex (MHC) on chromosome 6 is restricted to six hot spots, each less than 2 kb. Sequence variants lying between hot spots of recombination will therefore tend to form haplotype blocks that remain intact over thousands of generations. Variants lying on either side of such hot spots are more likely to be randomly associated in the population (i.e. in linkage equilibrium).

Different chromosomal sites show different extents of nucleotide diversity. For example, X chromosomes are present in a single copy (hemizygous) in males and so their  $N_e$  is three-quarters of that of an autosome. The overall diversity reflects many different mechanisms, only one of which is  $N_e$ , but average nucleotide diversity ( $\pi$ ) is lower for the X chromosome (0.00047) than for autosomes (0.0008). Similarly, the nonrecombining region of the Y chromosome ( $\sim 80\%$  of its total length) has an  $N_e$  that is one-fifth ( $0.8 \times 1/4$ ) of that of the autosomes and has an even lower nucleotide diversity (0.00015). The diversity of sex chromosomes also reflects the 2- to 4-fold higher mutation rate in sperm as compared with eggs, owing to the larger number of cell divisions (and thus DNA replication errors) during spermatogenesis. Since, the X chromosome undergoes only a third of all meioses in males, this is another reason for its lower diversity.

Sequence diversity is influenced by natural selection at neighboring loci. If a genetic variant is under strong selection in a population, this tends to reduce the diversity of neutral loci, such as most SNPs, in the surrounding region. The extent of this reduction depends on the local recombination rate, because recombination will separate the selected region from its flanking sequences. Regions of low recombination therefore tend to show lower nucleotide diversity. This seems to be a significant factor in explaining the observed 10-fold differences in sequence variability across the genome. For example, it can be seen that the sequences surrounding a region of high or low variability tend to show similar patterns of variability. Possible reasons include local differences in mutation rate, local differences in the extent to which natural selection has acted on nearby genes and local differences in recombination rates. The available evidence suggests that local differences in mutation rate provide only a partial explanation, but a combination of historical factors such as the extent of natural selection plus the patchy distribution of recombination sites (hot spots) explain most of these differences.

## Genetic Variation and Ethnicity

The nature of the genetic variation underlying commonly observed ethnic differences in facial or other morphological features is mostly unknown, but such differences are likely to be quantitative rather than qualitative. In other words, differences in the population frequency of variants such as cSNPs, rather than their presence or absence, seem most likely to be involved.

Sequence surveys in different ethnic groups consistently show that most genetic diversity occurs within rather than between ethnic groups. About 85% of all sequence variants are found in members of the same population, whereas only 5–10% of variants account for differences between populations in the same continent. Genetic differences between continental populations account for 5–20% of all differences. Continent- and population-specific variants are certainly present, but most of them are uncommon. An example would be the disease-causing alleles responsible for the disorder Tay–Sachs disease, a fatal neurodegenerative disorder of childhood that is 100 times more common in Ashkenazi Jews than in other populations. The carrier (heterozygote) frequency of this autosomal recessive disease is about one in 30 in the Ashkenazi population, possibly as a result of a founder effect sometime between AD 70 and AD 1100 before the major migrations into Poland and Russia, implying that there was a small population size at the time. Similarly, certain  $\beta$ -globin gene variants responsible for thalassemia in the Mediterranean and sickle cell disorders in Africans are specific to those regions, where they show relatively high frequencies. In general, such population- or region-specific variants are uncommon. One that is common, however, is the Duffy blood group allele, *FY\*O*, which is present only in sub-Saharan Africans for the reasons discussed above.

These findings have been replicated with different classes of genetic variant, including SNPs, microsatellites and mitochondrial DNA. A caveat is required however, because variants that are present at high population frequency, such as SNP alleles (with a frequency  $> 20$ ), are more likely to be shared by all populations, as they both predate the divergence of ethnic groups and are the ones most likely to be sampled. In contrast, rare variants are the ones most likely to be population-specific, but are more difficult and expensive to detect and to analyze.

Human populations show less genetic diversity than do other species such as large mammals, despite occupying a larger geographic range. For example, the frequency of nuclear DNA sequence (SNP) variants was found to be 2- to 4-fold higher in great apes than in humans. Similar results have been

obtained with the mitochondrial genome, in which diversity is 3- to 4-fold higher in nonhuman primates such as chimpanzees. The census size of chimpanzees is around 100 000–2 000 000 and yet their  $N_e$  is 35 000, which is consistent with the finding of greater diversity. Possible explanations for the reduced diversity of humans include our recent evolutionary origin and demographic factors contributing to a recent common ancestry, leaving little time for sequence divergence. Alternatively, the occurrence of large-scale population movements and migrations may have tended to homogenize ethnic groups, further reducing differences. In support of these possibilities, there is evidence for population expansions between 70 000 and 160 000 years ago, and for at least one significant population bottleneck (as discussed above) resulting in reduced sequence diversity. The comparative homogeneity of our species therefore argues strongly against concepts of race that have fueled eugenic movements and cast a shadow over some aspects of human genetics in the past 50–100 years.

### See also

Human Genetic Diversity  
Mutational Change in Evolution  
Single Nucleotide Polymorphism (SNP)

### Further Reading

- Botstein D and Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics* **33**(supplement): 228–237.
- Reich DE, Schaffner SF, Daly MJ, *et al.* (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genetics* **32**: 135–142.
- Risch NJ (2000) Searching for genetic determinants in the new millennium. *Nature* **405**: 847–856.
- The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.
- Wright A, Charlesworth B, Rudan I, Carothers A and Campbell H (2003) A polygenic basis for late-onset disease. *Trends in Genetics* **19**: 97–106.

### Web Links

- Online Mendelian Inheritance in Man  
<http://www.ncbi.nlm.nih.gov/omim>
- Human Gene Mutation Database  
<http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html>
- Apolipoprotein E (*APOE*); LocusID: 348. LocusLink:  
<http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=348>
- Dystrophin (*DMD*); LocusID: 1756. LocusLink:  
<http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=1756>
- Coagulation factor VIII-associated (*F8A*); LocusID: 8263. LocusLink:  
<http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=8263>
- Neurofibromin 1 (*NF1*); LocusID: 4763. LocusLink:  
<http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=4763>
- Polycystic kidney disease 1 (*PKDI*); LocusID: 5310. LocusLink:  
<http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=5310>

Apolipoprotein E (*APOE*); MIM number: 107741. OMIM:  
<http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispim?107741>

Dystrophin (*DMD*); MIM number: 300377. OMIM:  
<http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispim?300377>

Coagulation factor VIII-associated (*F8A*); MIM number: 305423.  
OMIM:

<http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispim?305423>

Neurofibromin 1 (*NF1*); MIM number: 162200. OMIM:  
<http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispim?162200>

Polycystic kidney disease 1 (*PKD1*); MIM number: 601313. OMIM:  
<http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispim?601313>