

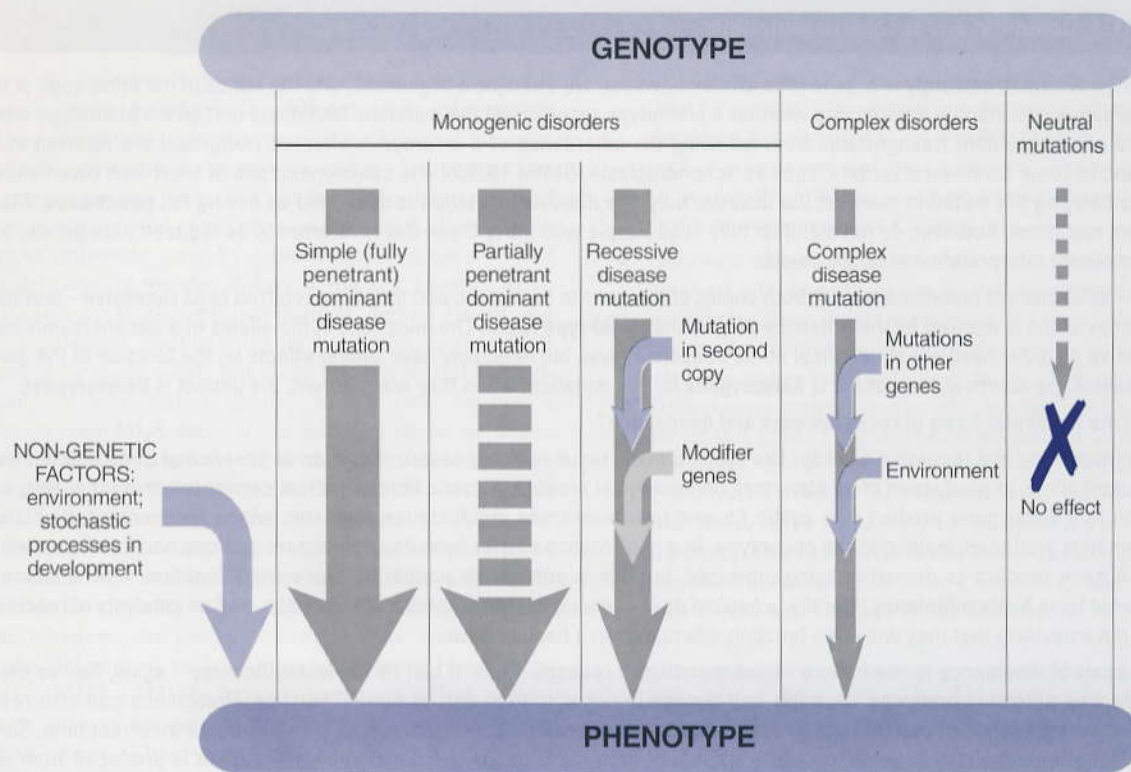


Figure 3.1: From genotype to phenotype.

The roles of genetic and nongenetic factors in generating the phenotype. Dominant mutations have a phenotypic effect even though a normal copy of the allele is present. This can be a full effect in all cases (full penetrance), or in only a subset of cases (reduced penetrance). Recessive mutations manifest a phenotypic effect only when no normal copy is present. Mutations involved in complex disorders require mutations in more than one gene and/or environmental factors to be manifested. Many mutations are neutral, having no known effect on phenotype. Much of human evolutionary genetics is concerned with analyzing and interpreting the patterns of diversity represented by these neutral mutations.

BOX 3.2 Polymorphism, variant, or mutation?

A simple base-substitutional difference between two human DNA sequences can be given a number of different labels, and unfortunately the rules for applying them are not always clear or consistent.

- ▶ A **polymorphism** has been rather arbitrarily defined as existing when at least two alleles are present in the population, and the minor allele is at a 1% or greater frequency (see e.g., Cummings, 2000).
- ▶ While all differences among sequences have mutation as their root cause, the word **mutation** is often reserved for use when pathogenic variation is being referred to, and is then used in contrast to 'polymorphism', which describes a sequence change in a gene which has no effect on function. This distinction in meaning is particularly prevalent among medical geneticists. However, there are clearly potential problems here, since it can be difficult or even impossible to know whether or not a sequence change is completely without a phenotypic effect (Cotton and Scriver, 1998). Also, in some populations disease-causing mutations are present at frequencies in excess of 1%, and therefore can under some definitions qualify as polymorphisms. Examples are a specific mutation in Europeans in the *CFTR* gene responsible for cystic fibrosis, or mutations in some African populations in the β -globin gene responsible for sickle cell anemia.
- ▶ An allele below the 1% frequency used to define a polymorphism is sometimes called a **variant**. Clearly, since allele frequencies often vary greatly between populations, one population's variant may be another's polymorphism. In addition, variant is sometimes used as a general term that includes both polymorphisms and mutations.

The term 'mutation' has negative connotations for the layperson (Condit *et al.*, 2002), and this fact has led to discussion about the choice of alternative words when scientists communicate about genetic diversity with the general public – 'variation', 'variant' or 'alteration' were favored.

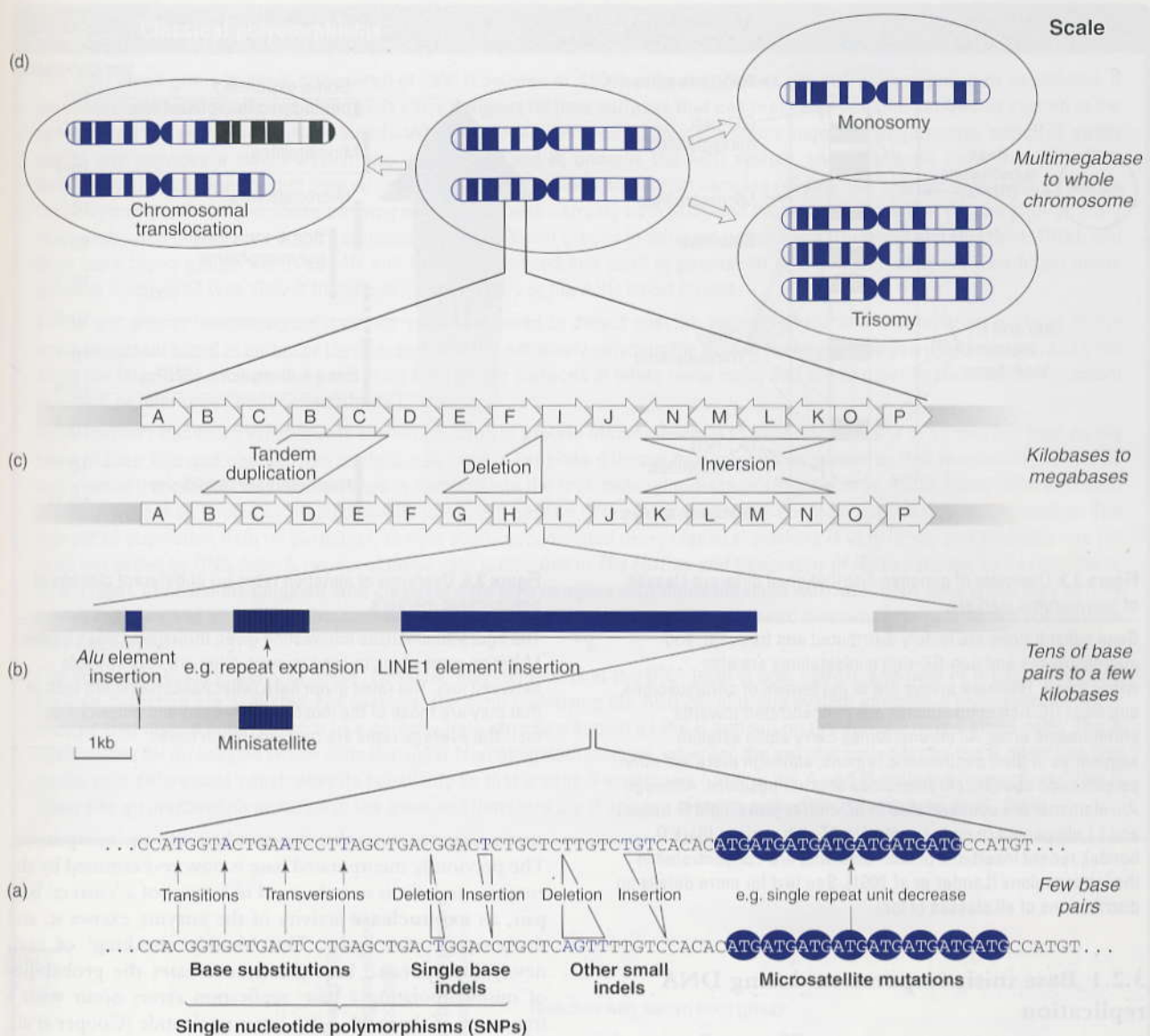


Figure 3.2: Overview of different classes and scales of mutation.

(a) Examples of mutations affecting a few base pairs, including examples of transition and transversion substitutions (see Figure 3.6), and a change in the number of ATG repeats at a trinucleotide microsatellite. (b) Examples of mutations affecting tens of base-pairs to a few kb, including insertion of a ~300-bp *Alu* element and a full length 6.1-kb *LINE1* element, and the expansion by several repeat units of a minisatellite. (c) Examples of segmental mutations on the kilobase to megabase scale. Arrows indicate the order and orientation of adjacent segments of DNA. (d) Mutations on the multimegabase or whole chromosome scale. Most translocations are deleterious, but are important in interspecies differences. Most aneuploidies (departure from the 46-chromosome complement, shown here as monosomy or trisomy) are early lethals; some, e.g., trisomy 21 and aneuploidies of the sex chromosomes, are tolerated. No aneuploidy can be passed on to offspring.

These kinds of differences are examples of **single nucleotide polymorphisms**, or **SNPs** (pronounced 'snips'). The insertion or deletion (**indel**) of a single base is also included in the category of SNPs – perhaps unfortunately, since the mechanisms which underlie these indels, and the analytical treatment of them, differ from those for base substitutions.

Two fundamental processes give rise to base substitution mutations: the misincorporation of nucleotides during replication, and mutagenesis caused by the chemical modification of bases, or physical damage due, for example, to ultraviolet or ionizing radiation. These two processes are summarized in Figure 3.7.

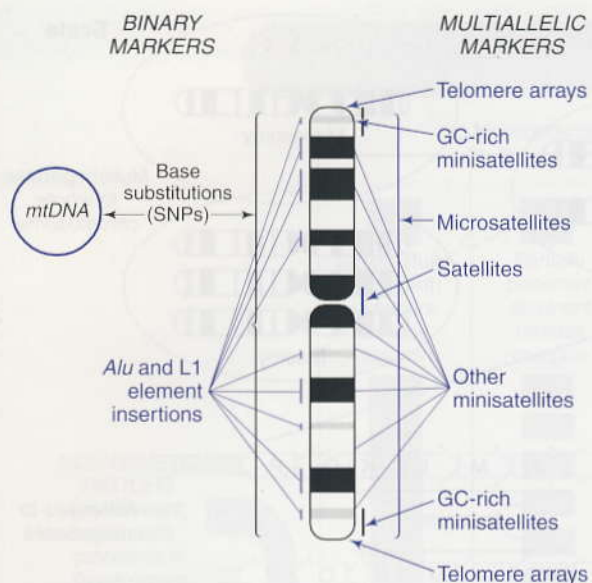


Figure 3.3: Overview of genomic distribution of different classes of polymorphic markers.

Base substitutions are widely distributed and frequent, and microsatellites and non-GC-rich minisatellites are also widespread. Telomere arrays are at the termini of chromosomes, and most GC-rich minisatellites are concentrated towards chromosome ends. All chromosomes carry alpha satellite sequences in their centromeric regions, although there are other satellites on specific chromosomes at other positions. Although *Alu* elements are concentrated in GC-rich regions (light G-bands) and L1 elements are concentrated in AT-rich regions (dark G-bands), recent insertions of both elements are concentrated in the latter regions (Lander *et al.* 2001). See text for more details on distributions of all classes of loci.

3.2.1 Base misincorporation during DNA replication

When a diploid cell divides, all 6400 Mb of its DNA must be replicated so that each daughter cell contains two copies of the haploid genome. DNA replication, the process that accomplishes this remarkable feat, proceeds with extremely high fidelity (reviewed by Kunkel and Bebek, 2000). A new base is incorporated if it pairs with the existing base in the single-stranded template. However, the existence of the correct number of hydrogen bonds between bases is in itself insufficient to ensure that A pairs only with T, and G with C. The **DNA polymerase**, the enzyme responsible for DNA synthesis, also requires the correct overall geometry of the base pair before the bond will be made with the growing daughter strand. Occasional incorporation of the incorrect base does occur, possibly because of rare, transient chemical forms of bases that have different base-pairing properties and geometries. However, as well as the high fidelity involved in this incorporation step, the primary DNA replication machinery has a 'proof-reading' function: after incorporation of a new base at the 3' end of the daughter strand, the

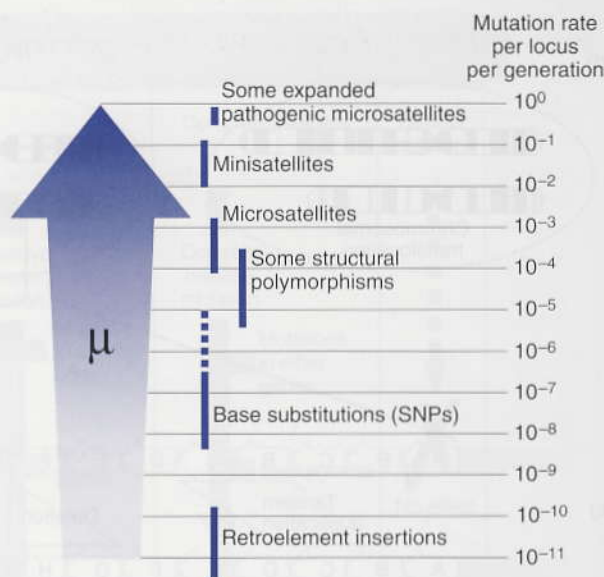


Figure 3.4: Overview of mutation rates (μ) of different classes of polymorphic markers.

The figure summarizes information given throughout this chapter. Mutation rates of micro- and minisatellites are very variable between loci. The rates given here reflect ascertainment bias, in that they are those of the mostly widely used and polymorphic loci. The average rates are probably much lower.

enzyme moves on so that the next base can be incorporated. The previously incorporated base is now re-examined by the enzyme, and if it is not observed to be part of a 'correct' base pair, an **exonuclease** activity of the enzyme excises it, and the enzyme tries again. This 'double checking' of each newly incorporated base greatly decreases the probability of misincorporating a base: replication errors occur with a frequency of about 10^{-9} – 10^{-11} per nucleotide (Cooper *et al.*, 1995).

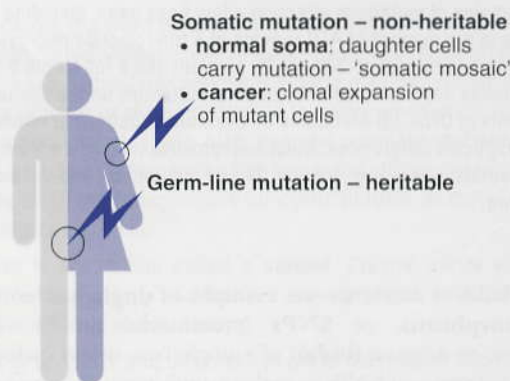


Figure 3.5: Somatic and germ-line mutations.

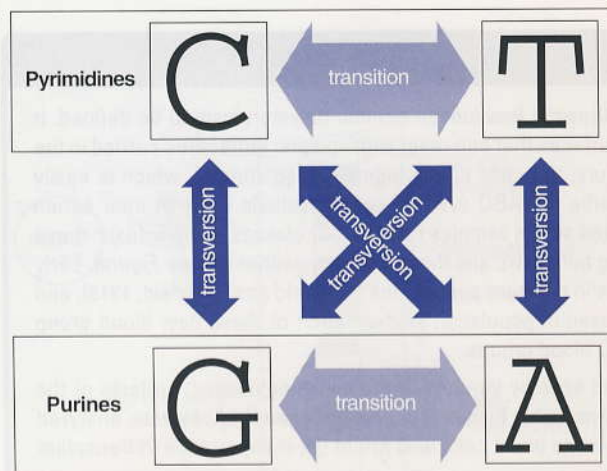


Figure 3.6: Transition and transversion mutations.

For each base there are two possible transversions, but only one transition. Nonetheless the mechanisms of mutation lead to transitions being more common than transversions (see Section 3.2.4).

3.2.2 Chemical and physical mutagenesis

The integrity of the genetic material, the sequence of DNA within a cell, is under constant assault by chemical and physical processes that alter bases or damage the physical structure of

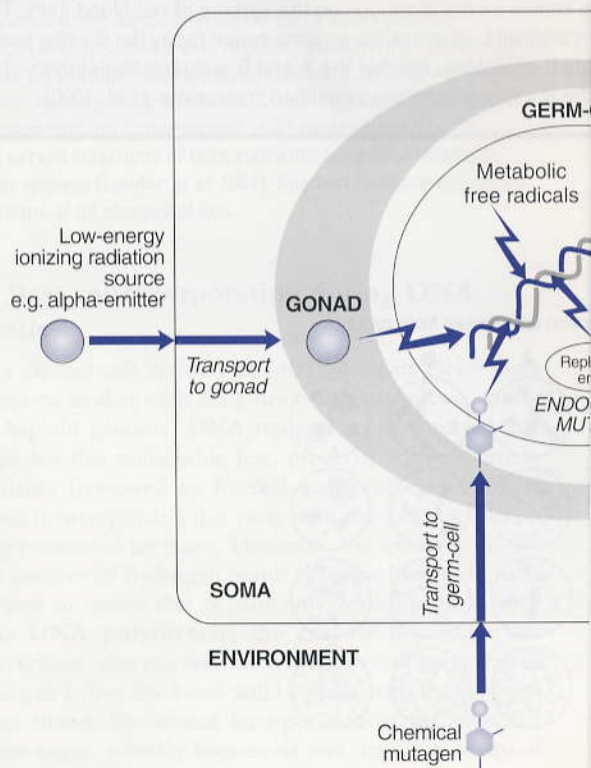


Figure 3.7: Chemical, physical and endogenous mutagenesis.

Overview of processes causing germ-line mutation. Ionizing radiation can produce damaging ions.

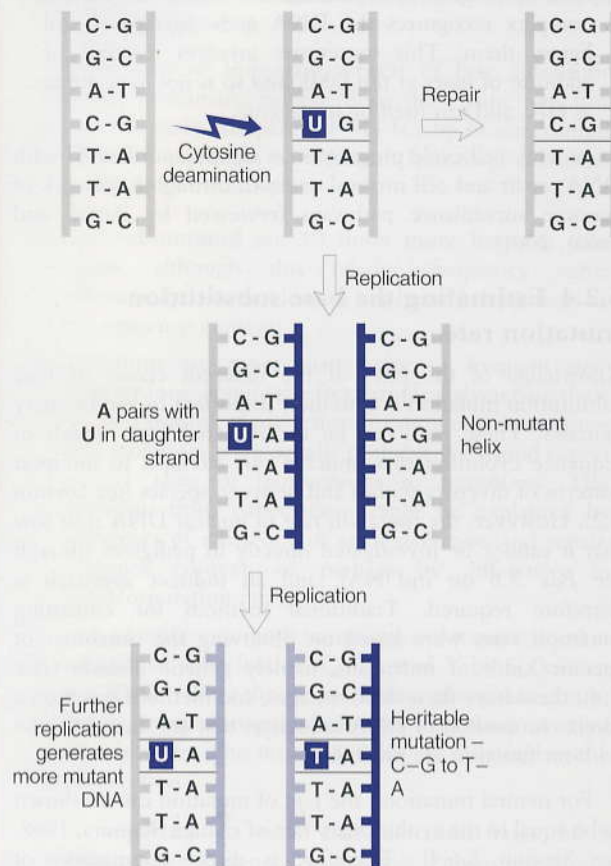


Figure 3.8: The fates of damaged DNA.

Damage to one strand of DNA, in this case cytosine deamination to yield uracil, can be repaired. Generation of a stable and heritable base-pair mutation requires two rounds of DNA replication.

synthesis have comparatively low fidelity. For example, most of the repair of damaged bases is carried out by the enzyme pol β , which excises and replaces only the damaged base. Pol β is the least accurate mammalian polymerase, with an average error rate (as determined *in vitro*) of about 10^{-3} (Osheroff *et al.*, 1999), so itself will lead to a large number of misincorporations during repair.

Double-stranded breaks (DSBs), which can be created by ionizing radiation, free radical action, and events during DNA replication, are particularly damaging lesions which can result in chromosomal aberrations, cell malfunction and cell death (reviewed by van Gent *et al.*, 2001). Complementary strand information cannot be used to repair such lesions, and instead a complex series of events can lead to repair of the DSB either through:

- homologous recombination (HR), in which the undamaged sister chromatid acts as template. This ensures accurate repair, but because it relies on the presence of the sister chromatid is dependent on the stage of the cell cycle at which damage occurs;

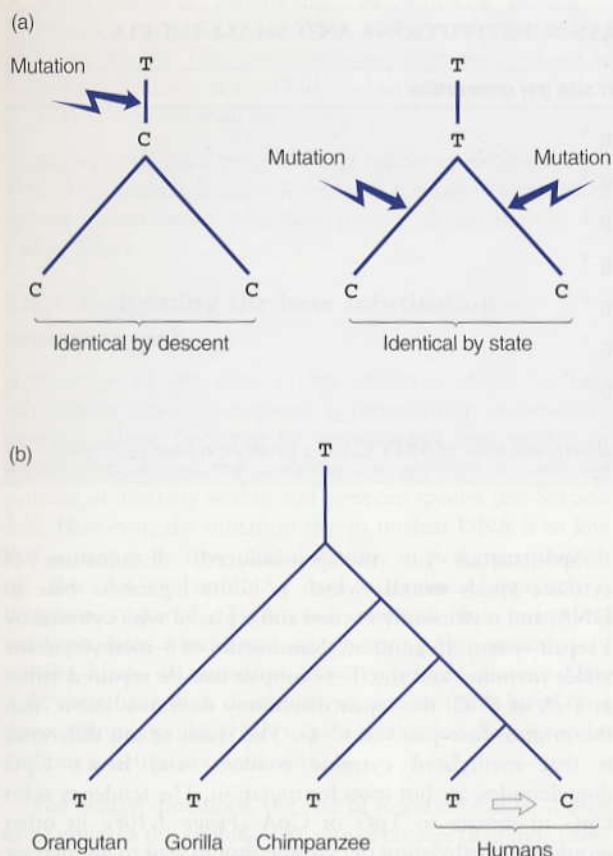


Figure 3.9: Identity by descent, and ancestral state determination of an SNP.

(a) When two alleles and their common ancestor share the same allelic state (here, a C-allele at SNP), they show identity by descent (left panel); when the common ancestor has a different allelic state (T-allele), they show identity by state (right panel). (b) Given the known phylogeny of our great ape relatives (see Chapter 7), the presence of a T at SNP site in great ape homologs of a human sequence showing a T/C polymorphism defines T as the ancestral state. The mutation underlying the polymorphism is a T to C transition. Note that polymorphism at the same base is on rare occasions expected to occur in the other primates. Analysis of more than one member of each species is useful.

in agreement with that for pseudogenes (Table 3.1), which may reflect differences in sequence composition].

3.2.6 Base substitutions and small indels within genes

Evolutionary genetic studies often focus on mutations that are assumed to be neutral – having no effect on phenotype. Many of these lie outside genes; however, because of the nature of the biological mechanisms which read information in DNA and express it as proteins, some sequence changes within genes are expected to have no or little effect on function. On the other hand, and as alluded to above, base substitutions within genes can be the cause of genetic disease,

and it is therefore important to understand the potential phenotypic effects of such changes in DNA sequences. Most mutations affecting phenotype do so adversely – there are more ways to impair the function of a gene than to enhance it. However, there is certainly an ascertainment bias here, since mutations causing diseases are noted and find their way into mutation databases. Mutations improving the performance of a protein or changing its function advantageously are not necessarily noted, but are nevertheless the stuff of positive selection and evolution. This discussion focuses on deleterious mutations.

Figure 3.11 shows some of the potential effects of base substitutions and small indels on a protein-coding gene; Table 3.2 gives the relative frequencies of the different types of mutation causing genetic disease, as listed in the Human Gene Mutation Database.

Outside the open reading frame, mutations can affect expression of the gene by altering **promoters**, **enhancers**, or polyadenylation signals. Mutation of the consensus splicing signals at the ends of introns can affect splicing, leading, for example, to the retention of an intron in the transcript or the skipping of an exon, and consequent loss of function. While the GT/AG dinucleotides at the ends of introns (Figure 3.11) are highly conserved and necessary for correct splicing, other adjacent and even distant nucleotides can influence its efficiency or accuracy (reviewed by Cartegni *et al.*, 2002). Alternatively, a new splice site can be created in an intron. To complicate matters even further, a mutation within an intron can affect the regulation of an adjacent gene; this appears to be the case for the human lactase gene, whose downregulation in adulthood is prevented by a mutation 14 kb upstream, in an intron within an unrelated gene (see Section 13.4.3).

In the open reading frame itself, base substitutions can have a wide range of possible effects, from complete neutrality to complete abolition of protein production. The redundancy of the genetic code (see Figure 2.6) provides a buffer against deleterious effects of base substitution: many amino acids are encoded by more than one codon (six different codons encode serine, for example), and in general the third position is particularly insensitive to mutation. Substitutions that do not alter an amino acid are known as **silent-site**, or **synonymous** substitutions, and are often assumed to be selectively neutral. However, such substitutions could in principle alter mRNA stability, or efficiency of translation, and hence affect function. Cases are also known where they have a deleterious effect through altered splicing (reviewed by Cartegni *et al.*, 2002) – for example, a silent-site substitution in an exon of the phenylalanine hydroxylase gene gives rise to a phenylketonuria disease allele (OMIM 261600) by inducing skipping of the next exon (Chao *et al.*, 2001).

Base substitutions which lead to a change of amino acid are termed **nonsynonymous**, or **missense** mutations, and are denoted, for example as 'Arg702Trp', a substitution of tryptophan for arginine at amino acid 702 of a protein. Beyond straightforward redundancy, the genetic code

Figure 3.10: The (a) Spontaneous glycosidase. De the 5-carbon of action of a spec affects the com

appears to be of base substitution assignment of means that mar amino acid wit properties, for change is terme

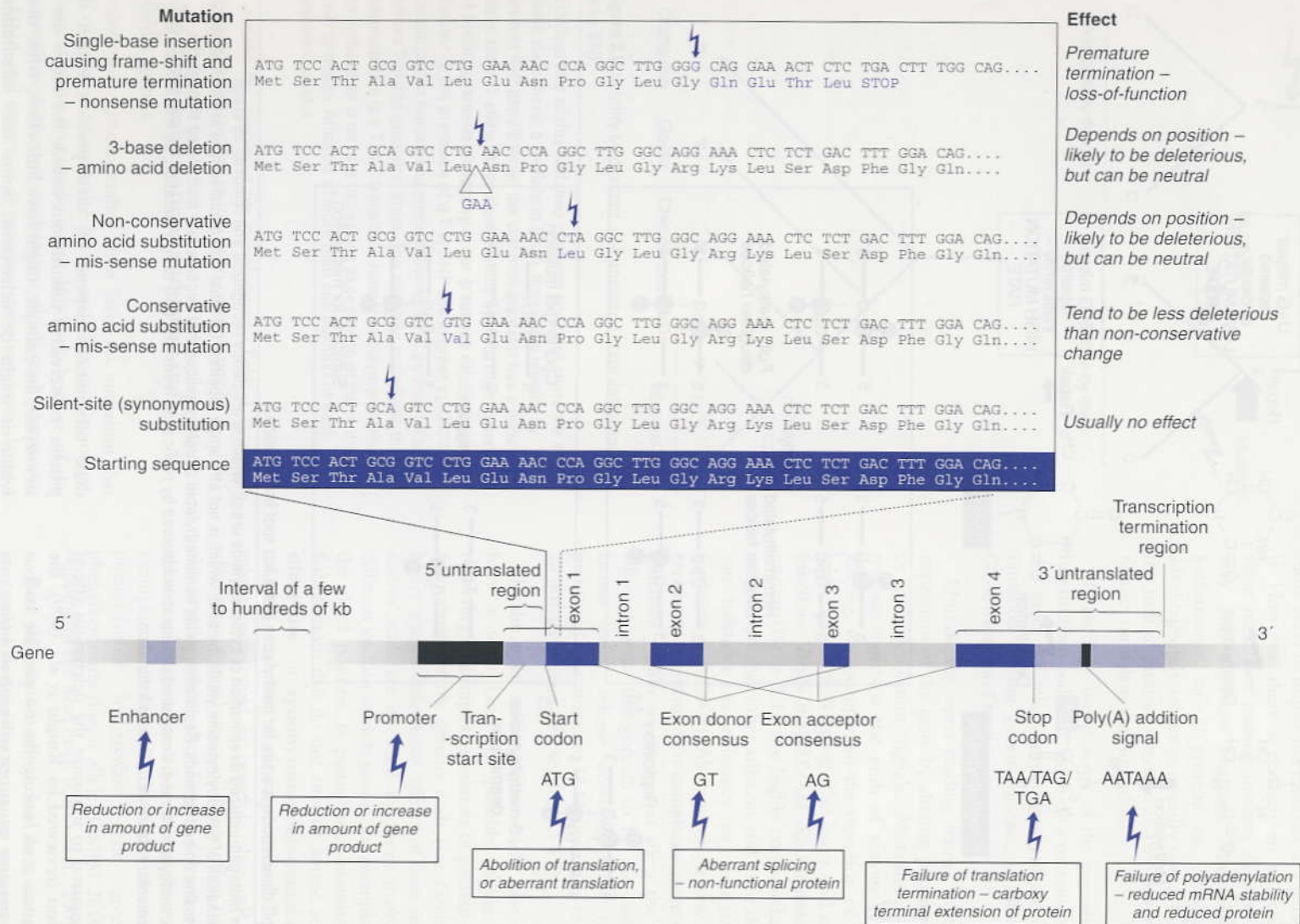


Figure 3.11: Small-scale mutations within a typical human protein-coding gene and their potential effects.

The different elements of the gene are not to scale, and in particular the introns illustrated here are small relative to the exons (see *Table 2.1* for typical sizes of these elements).

TABLE 3.2: SUMMARY OF STATISTICS FOR SMALL-SCALE PATHOGENIC MUTATIONS IN 1163 GENES.

Type of mutation	Number	Percentage
Base substitution	21 200	75% of total
Amino acid substitution (missense)	14 756	52% of base substitutions
Generation of stop codon (nonsense)	3198	11% of base substitutions
Abolition of stop codon	24	< 1% of base substitutions
Splice site mutation	2976 ^a	11% of base substitutions
Regulatory	246	1% of base substitutions
Small deletion	5049	18% of total
Small insertion	1922	7% of total
Total	28 171	

Data taken from the Human Gene Mutation Database (02/09/2002 release). The database lists a total of 30 641 mutations, including 2470 gross lesions and other small indels not given here.

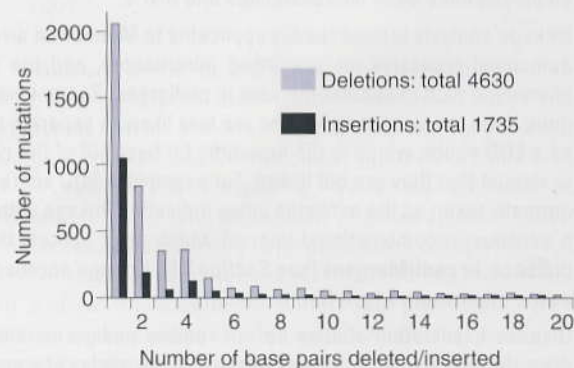
^aOf which 1680 occur at the GT/AG consensus dinucleotides.

An example is the Hb^S mutation, which changes a hydrophilic ('water-loving') glutamic acid on the molecule's surface for a hydrophobic ('water-hating') valine: the consequence is that the protein molecules associate together, rather than staying in solution, leading to changes in cellular behavior. A mutation changing an amino acid codon into a termination codon is termed a **nonsense** mutation, and is likely to be deleterious unless very close to the natural termination codon.

A small indel within an open reading frame is more likely to be deleterious than is a base substitution. Because the genetic code is based on triplets, any indel of a multiple of three bases will lead to an insertion or deletion of amino acid(s). Again, the effects of such changes will depend on their position in the protein. The best known example of a single amino acid deletion is the commonest cystic fibrosis mutation, $\Delta F508$ (see Section 14.2.1), which deletes a phenylalanine from the CFTR protein, leading to defective intracellular processing. If the indel is not a multiple of three bases, it changes the reading frame of translation. Such **frameshift** mutations are highly deleterious, since the amino acid sequence 3' of the mutation is completely different from that of the wild-type. Translation can proceed until a termination codon is reached, leading to prematurely terminated protein, or to abnormally extended protein, particularly if the mutation is close to the 3' end of the coding region. Alternatively, the change of reading frame may lead to unstable mRNA or aberrant splicing which abolishes protein production. Among disease-causing mutations, at least, small deletions predominate over small insertions, and the great majority of events involve 1–5 bases (Figure 3.12).

3.2.7 SNP diversity and SNP distribution

Interest in the identification of SNPs is high, largely because of their potential use as **molecular markers** in **disease association studies** (Box 3.5). A number of large-scale re-sequencing (sequencing the same locus in several individuals) studies of particular loci have now been done, and provide a picture of SNP diversity in these regions (reviewed by Przeworski *et al.*, 2000). As well as this, efforts to sequence the entire genome have used libraries of clones [mostly **bacterial artificial chromosome (BAC)** clones] constructed from the DNAs of more than one individual. This means that overlapping sequences between different clones readily reveal SNPs: the draft and finished sequence of the publicly funded sequencing project, for example, represented a 7.5-fold coverage of the genome, and identified 971 000

**Figure 3.12:** Incidence of small indels in human gene mutations.

Distribution of small indels in the Human Gene Mutation Database (21/06/2002 update) by size.

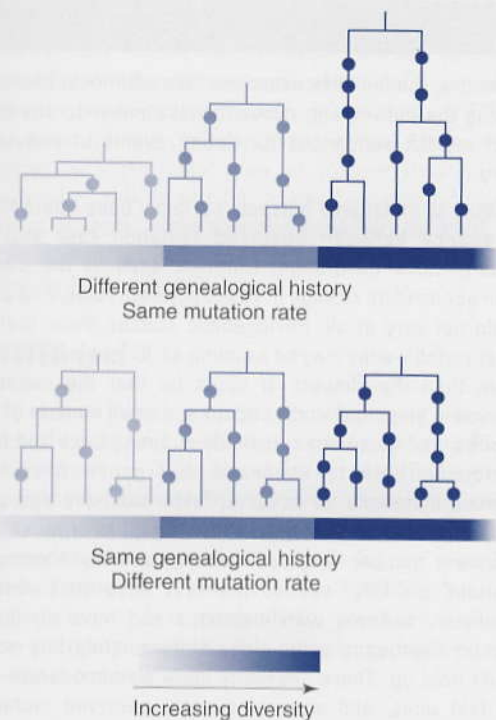


Figure 3.13: The effect of genealogical history and mutation rate on the sequence diversity of a locus.

substitutions appear to be neutral. Evolutionary genetic studies concentrate on these, but also consider other apparently neutral variations, such as change in length of a poly(C) tract in the control region, and the '9-bp deletion', a change in copy number of a 9-bp tandem repeat motif from two copies to one copy (rare mtDNAs even have three repeats). Early comparisons of human and other primate DNAs indicated that the base substitutional mutation rate in mtDNA is about 10 times higher than the average rate in nuclear DNA (Brown *et al.*, 1979). This high rate leads to a very large number of different sequences in human populations – one of the reasons for the popularity of mtDNA as an evolutionary tool. The observed mutation rate throughout the 16.5-kb molecule is nonuniform, with a relatively low rate in the coding regions, and a relatively high rate in the control region, which does not encode proteins or RNAs (see Figure 2.20). At some positions within the **hypervariable segments (HVS I and HVS II, sometimes also known as hypervariable regions, or HVRs)**, of the control region, mutation rate is so high that mutations can often be observed in pedigrees, and comparisons between the pedigree rate and the rate calculated by other means have caused controversy (see Box 3.6).

A good estimate of mtDNA mutation rate is important for human evolutionary genetic studies. The widely employed estimates are indirect, based on calibrating the amount of divergence or diversity accumulated since a well-dated event, such as a phylogenetic split between humans and a closely

related primate, or a particular human settlement, such as that of New Guinea. Most interest has focused on the mutation rate of the control region, since this is the part of mtDNA most widely used in population studies. mtDNA mutation rates are often expressed as base substitutions per site per million years, rather than the 'per base per generation' expression that has been used elsewhere in this chapter. A range of rates covering an order of magnitude has been provided by these studies: 0.025–0.26 per site per million years (these figures include confidence intervals, and were compiled from the literature by Parsons *et al.*, 1997). This range, assuming a 20-year generation time, equates to approximately 5×10^{-7} to 5×10^{-6} per base per generation. Some mutation rates focus on an even smaller region, the more variable of the two **hypervariable segments (HVS I)**. An average rate for a 275-bp section is given as one transition per 20 180 years (Richards *et al.*, 2000), which equates to 3.6×10^{-6} per base per generation, for a 20-year generation time. Outside the control region mutation rates have again been calculated by phylogenetic comparisons and assumptions about species divergence times: a rate from whole-mtDNA genome sequencing (Ingman *et al.*, 2000) is 3.4×10^{-7} .

There are a number of possible reasons for the high rate of mutation of mtDNA compared to nuclear DNA:

- ▶ because of its function in energy generation through oxidative phosphorylation, the mitochondrion contains a high concentration of mutagenic oxygen free radicals;
- ▶ mtDNA may have a higher turnover rate than nuclear DNA, requiring more replications per unit of time;
- ▶ because of the way it is replicated, mtDNA spends more of its time in the vulnerable single-stranded form than does nuclear DNA; indeed, the control region is sometimes called the 'D-loop' region to reflect its unusual structure;
- ▶ mtDNA is not packaged with histones, which may make it more vulnerable to mutation;
- ▶ repair systems in mitochondria are less effective than those in the nucleus (reviewed by Shadel and Clayton, 1997).

Mutation in the nuclear genome has been described in the text, and is easy to understand – mutations occur in the germ-line and are passed to diploid offspring via gametes in a single (haploid) dose. Mutations in mtDNA are different: as was explained in Chapter 2, each cell contains many mitochondria, each of which contains several mtDNA molecules; thus there is a large population of mtDNAs in any cell. A mtDNA mutation occurs in a *single* mtDNA molecule in a *single* mitochondrion within this population, yet can come to represent all the mtDNA in the soma of an individual in a subsequent generation. How does this occur?

Observations of mtDNA mutations, both neutral and disease-associated, indicate that the population of mitochondria passes through a bottleneck, at some point in oogenesis between the primordial germ cells and the primary oocyte