

CHAPTER 3

Human Genetic Variation: Databases and Concepts

MICHAEL R. BARNES

GlaxoSmithKline Pharmaceuticals
Harlow, Essex, UK

- 3.1 Introduction
 - 3.1.1 Human genetic variation
 - 3.1.2 The genome as a framework for integration of genetic variation data
- 3.2 Forms and mechanisms of genetic variation
 - 3.2.1 Single nucleotide variation: SNPs and mutations
 - 3.2.1.1 The natural history of SNPs and mutations
 - 3.2.1.2 SNP and mutation databases united?
 - 3.2.2 Tandem repeat polymorphisms
 - 3.2.3 Insertion/deletion polymorphisms and chromosomal abnormalities
 - 3.2.4 Gross chromosomal aberrations
 - 3.2.5 Somatic mutations
 - 3.2.5.1 Somatic point mutations
 - 3.2.5.2 Genomic aberrations in cancer
- 3.3 Databases of human genetic variation
- 3.4 SNP databases
 - 3.4.1 The dbSNP database
 - 3.4.1.1 The Reference SNP dataset (RefSNPs)
 - 3.4.1.2 Searching dbSNP
 - 3.4.1.3 Submitting data to dbSNP
 - 3.4.1.4 Key SNP data issues
 - 3.4.1.5 Candidate SNPs — SNP to assay
 - 3.4.2 Human Genome Variation Database (HGVB)
- 3.5 Mutation databases
 - 3.5.1 The Human Gene Mutation Database (HGMD)
 - 3.5.2 Sequence Variation Database (SRS)
 - 3.5.3 The Protein Mutation Database (PMD)
 - 3.5.4 On-line Mendelian Inheritance in Man (OMIM)
- 3.6 Genetic marker and microsatellite databases
 - 3.6.1 dbSTS and UniSTS
 - 3.6.2 The Genome Database (GDB)

- 3.7 Non-nuclear and somatic mutation databases
 - 3.7.1 MITOMAP
 - 3.7.1.1 Searching MITOMAP
 - 3.7.2 The Mitelman Chromosome Abberations Map
 - 3.8 Tools for SNP and mutation visualization—the genomic context
 - 3.9 Tools for SNP and mutation visualization—the gene context
 - 3.9.1 LocusLink
 - 3.9.2 SNPper
 - 3.9.3 CGAP-GAI (<http://lpgws.nci.nih.gov/>)
 - 3.10 Conclusions
 - References
-

3.1 INTRODUCTION

Genetic variation is a key commodity for geneticists; not only as the much sought after basis of heritable phenotype, but also as a marker to aid in this search. For the wider biological research community, information on genetic variation can tell us many things about the functional parameters and critical regions of a gene, protein, regulatory element or genomic region. Study variation and a picture of the driving force of evolution begins to emerge. This knowledge can not only help us elucidate the function of genes and pathways by studying their function and dysfunction in normal and diseased states, it can also help us to understand the origins and diversity of mankind and other organisms. The availability of a complete human genome sequence finally puts this variation into context with all other biological data. In this chapter we will present an overview of the many forms of genetic variation, we will review current and past trends in the use of this data and highlight the key databases from which this data can be accessed and manipulated.

3.1.1 Human Genetic Variation

Human genetic variation and our environment are the two key factors that make each and every one of us different. Genetic variation takes many forms, although these variants arise from just two types of genetic mutation events. The simplest type of variant results from a single base mutation which substitutes one nucleotide for another. This mutation event accounts for the commonest form of variation, single nucleotide polymorphisms (SNPs). Many other types of variation result from the insertion or deletion of a section of DNA. At the simplest level this can result in the insertion or deletion of one or more nucleotides, so-called insertion/deletion (INDEL) polymorphisms. The most common insertion/deletion events occur in repetitive sequence elements, where repeated nucleotide patterns, so-called 'variable number tandem repeat polymorphisms' (VNTRs), expand or contract as a result of insertion or deletion events. VNTRs are further subdivided on the basis of the size of the repeating unit; minisatellites are composed of repeat units ranging from 10 to several hundred base pairs. Simple tandem repeats (STRs or microsatellites) are composed of 2–6-bp repeat units. The rarest insertion/deletion events involve deletions or duplications of regions ranging from a few kilobases to several megabases. These forms of variation were once thought to be restricted to rare genomic syndromes, however, sequencing of the human genome has presented a great deal of evidence to suggest that these events may be more common than previously expected.

The quantity of genetic variation in the human genome is something that until recently we have only been able to estimate by an educated guess. Empirical studies quite quickly identified that on average, comparison of chromosomes between any two individuals will generally reveal common SNPs (>20% minor allele frequency) at 0.3–1-kb average intervals, which scales up to 5–10 million SNPs across the genome (Altshuler *et al.*, 2000). The availability of a complete human genome has helped us considerably to estimate the number of potentially polymorphic STRs and minisatellites, as VNTRs over a certain number of repeats can be reliably predicted to be polymorphic. Viknaraja *et al.* (unpublished data) completed an *in silico* survey of potentially polymorphic VNTRs in the human genome and identified over 100,000 potentially polymorphic microsatellites. Other forms of variation such as small insertion/deletions are more difficult to quantify, although they are likely to fall somewhere between SNPs and VNTRs in numbers. Large deletions or duplications are the most unquantifiable form of variation in the genome. Quantification of these forms of variation is only possible by intensive cytogenetic methods (Gratacos *et al.*, 2001). They cannot be reliably identified from the genome sequence; in fact they are implicitly an obstacle to genome assembly, as large duplications are often incorrectly collapsed into a single assembly.

This huge quantity of genetic variation in the human genome led many to question the origin and maintenance of such a ‘genetic load’ in the human population. The traditional belief that most mutation was deleterious and subject to selection was quickly challenged by this data. In response to this observation Kimura (1983) and others formulated a ‘neutral theory of evolution’. This theory proposed that most sequence variation does not directly impact phenotypic variation and so is not directly subjected to the forces of selection. Thus, the overwhelming majority of genetic variants are likely to be phenotypically neutral, while many will define the diverse phenotypes that define individual humans. However a certain undefined number of these alleles will have deleterious effects, either directly causing or increasing susceptibility to disease. Some of this variation, so-called mutations, will be rare in populations whilst others will be common, so-called polymorphisms that increase susceptibility to common diseases. It will not usually be possible to identify these deleterious alleles directly, instead genetics has developed around the concept of using markers to detect nearby deleterious alleles. Fortunately for geneticists, the huge quantity of common polymorphism across the human genome makes it very likely that one or more of these polymorphisms will be in close enough vicinity to a rarer disease allele to detect it by common co-inheritance (linkage disequilibrium) between the two alleles.

Thus, one of the primary objectives of genetics is to utilize polymorphisms across the genome as markers which show co-inheritance with the phenotype under study. SNPs are the most obvious choice for these studies as they are the commonest form of human variation. However this choice has not always been so clear. Despite the abundance of SNPs in the genome, without knowledge of the genome sequence, SNP identification is a laborious process. This has made SNP availability very limited until very recently. Instead geneticists have used microsatellites as markers. These highly polymorphic markers can be isolated by relatively simple molecular methods and can detect disease-causing mutations in family-based studies over a larger distance than SNPs, often over 20 MB. The extent of this linkage enables whole genome linkage studies with as few as 200–500 microsatellite markers. Such linkage studies have been very successful in mapping mutations causing single gene disorders or Mendelian traits, but have been largely unsuccessful in detecting the multiple genes responsible for the commoner complex diseases (Risch, 2000).

The primary approach proposed for mapping complex disease genes is to use markers to detect population-based allelic association or linkage disequilibrium (LD) between

markers and disease alleles (see Chapter 8 for a detailed exploration of this area). These associations can be very strong even where the corresponding family linkage signal is weak or absent. This approach can localize disease alleles to very small regions, based on localized LD, which on average extends between 5–100 kilobases (kb) depending on a range of factors (Reich *et al.*, 2001). Detection of this association demands a massive increase in marker density with 200,000–500,000 markers estimated to be needed to cover the genome for an association scan compared to the 200–500 markers needed for a family-based linkage scan.

These population-based association studies call for ultra high-throughput genotyping methods. Technology developments to date suggest that SNPs are likely to be the most viable option for these studies for a number of reasons, but primarily because SNPs are more tractable to automated high-throughput analysis than microsatellite markers. Until very recently demand for SNPs completely outstripped SNP availability and so whole genome SNP association studies simply could not be attempted. This situation is now changing—completion of the genome has enabled several large-scale SNP discovery projects. Genetics is now entering a promising new era where marker resources and locus information are no longer the main factors limiting the success of complex disease gene hunting. The emphasis is now on good study design and carefully ascertained study populations. Effective informatics is critical to effectively exploit this data. More than ever, geneticists will need to be competent users of bioinformatics tools to construct sophisticated marker maps that can detect the full complexity of human genetic variation.

To find disease associations and ultimately disease alleles, it is necessary to study genetic variation at increasing levels of detail. At first, markers need to be identified at a sufficient density to build marker frameworks to detect linkage or association across the genome. Once this linkage or association is detected a denser framework of markers is needed to refine the signal. In the case of linkage analysis, marker density may not need to be increased beyond a few hundred kilobases as linkage is likely to remain intact over considerable distances in families. However in the case of association, marker density needs to be increased to a level at which all haplotype diversity in a population is captured (see Chapter 8). This may call for the construction of very dense marker maps down to a resolution of 5–10 kb. Ultimately, once LD is established between a marker and a phenotype it is necessary to identify all genetic variation across the narrowed locus, hopefully allowing the identification of the disease allele. This increasing resolution of analysis may involve a progression of bioinformatics tools and increasing ingenuity in the use of these tools as the requirements for detail increase. Variation can take many forms, any of which may have a bearing on the genetic mechanisms of disease. The very act of characterizing variation across a locus may help to cast light upon its genetic nature and the possible nature of the phenotype. For example, some genomic regions show hypermutability, while others show very low levels of mutation or polymorphism. The reasons for these differences are poorly understood, they may be based upon the physical properties of chromosomes, evolutionary selection or other unknown influences, all of which may have a bearing on disease.

3.1.2 The Genome as a Framework for Integration of Genetic Variation Data

Bioinformatics offers some powerful tools for detecting, organizing and analysing human genetic variation data. The value of these tools is totally dependent on the underlying quality and organization of the data. Ideally, variation data needs to be available in an organized and centralized form that will allow complex queries and integration with other

data sources. Without the benefit of a complete genome, such integration was little more than a pipe dream, but now we are presented with an opportunity to integrate data on the sequence framework. Generally it takes only two 20–30 base pairs of flanking sequence to unambiguously locate a sequence feature such as an SNP in the genome. This bioinformatics process is called electronic PCR (ePCR) and it is completely analogous to laboratory-based PCR. Two primers are used to map a sequence feature (e.g. a SNP). To validate the position both primers must map in the same vicinity spanning a defined distance, effectively producing an electronic PCR product. The possibilities for data integration are immense. For genetics, exact base pair localization of each variant allows the construction of absolutely precise physical maps, which can be accurately integrated with genetic maps. It is now possible to take a given region and place SNPs, mutations, microsatellites and insertion/deletions in exact order. Without a sequence map this simply would not have been possible as each marker may have been mapped by different laboratory methods—producing few directly comparable results (see Chapter 7 for a discussion of map integration issues).

3.2 FORMS AND MECHANISMS OF GENETIC VARIATION

In silico (bioinformatic) analysis of human sequence presents an opportunity to identify genetic variants by comparison of differences between two sequences. Most obviously *potential* SNPs can be identified by comparison of two sequences; these could be expressed sequence tags, cDNAs or genomic sequences. The same method can also be used to identify *potential* INDEL polymorphisms. *Potential* is a key word to apply to this *in silico* polymorphism discovery process which can be prone to false positives introduced by sequencing error and other issues (see Chapter 10).

Human genome sequence also gives us an opportunity to assess some of the less commonly studied forms of variation. Although under-represented in databases some potential forms of variation can be identified from a single DNA sequence, by sequence alone. Short tandem repeat sequences are the most obvious example of such variants, however, sequence analysis can also be used to identify minisatellites and segmental duplications which may also mediate large deletions or duplications. Our knowledge of these forms of variation is limited; this reflects studies to date which have focused on more technically tractable variants, such as SNPs, mutations and short tandem repeats. Databases have also as a matter of practicality tended to focus on these classes of variation, and in this chapter we will review these databases in detail. We will also attempt to draw the less studied forms of variation into context, reviewing the best tools to access this data. Where no database exists we will review the mechanisms which govern variation and which can assist detection by bioinformatics methods.

3.2.1 Single Nucleotide Variation: SNPs and Mutations

Terminology for variation at a single nucleotide position is defined by allele frequency. In the strictest sense, a single base change, occurring in a population at a frequency of >1% is termed a single nucleotide polymorphism (SNP). When a single base change occurs at <1% it is considered to be a mutation. However, this definition is often disregarded, instead ‘mutations’ occurring at <1% in general populations might more appropriately be termed low frequency variants. The term ‘mutation’ is often used to describe a variant identified in diseased individuals or tissues, with a proven role in the disease phenotype.

Mutation databases and polymorphism databases have generally been divided by this definition. Polymorphisms are generally considered widespread in populations and mutations are usually rare and are not generally thought to be spread widely in populations, but instead occur sporadically or are inherited in families in a Mendelian manner. A grey area exists, which argues against the rigidity of this division of data. Some autosomal recessive Mendelian mutations have been linked to complex disease susceptibility in a heterozygote form and indeed are relatively widely spread in populations. For example, homozygote mutations in the cystathione beta synthase gene cause homocystinuria, a rare disorder inducing multiple strokes at an early age. The heterozygotes do not share this severe disorder, but do have an increased lifetime risk of stroke (Kluijtmans *et al.*, 1996). In Caucasians the population frequency of homozygote homocystinuria mutations is only 1/126,000, but in the same population, heterozygote frequency is relatively high at 1/177. There are many other examples of ‘Mendelian mutations’ which actually exist at appreciable heterozygote levels in general populations, particularly isolated populations, e.g. mutations in the breast cancer susceptibility gene, BRCA1, have been found in 1–2% of Jewish populations (Bahar *et al.*, 2001) and mutations in the CFTR gene cause cystic fibrosis, the most common autosomal recessive disease in the Caucasian population, with a carrier frequency of around 2% (Roque *et al.*, 2001).

3.2.1.1 The Natural History of SNPs and Mutations

The presence of heterozygous ‘Mendelian mutations’ in general populations illustrates the point that it may not always be helpful to rigidly separate polymorphism and mutation data. Another factor which argues against division of these data is that both SNPs and mutations arise by the same mechanism, although selection may influence their spread in populations. Miller and Kwok (2001) presented a detailed review of the ‘life cycle’ of a single nucleotide variation, they defined SNP and mutation evolution in four phases (Figure 3.1):

- (1) Appearance of a new variant allele by mutation
- (2) Survival of the allele through early generations against the odds
- (3) Increase of the allele to a substantial population frequency
- (4) Fixation of the allele in populations

Each of these stages goes to the heart of the differences and similarities between SNPs and mutations. Both arise by the same mechanism; nucleotide substitution is DNA sequence context dependent—substitution rates are influenced by 5′ and 3′ nucleotides. This effect is most dramatic for CT and GA transitions; these CpG dinucleotides are methylated and tend to deaminate to either a TpG or CpA dinucleotide (Cooper and Youssoufian, 1988). This makes these dinucleotides the most likely locations for point mutation in the human genome, with G > A or C > T transitions accounting for 25% of all SNPs and mutations in the human genome (Miller and Kwok, 2001). In itself this molecular mechanism accounts for the deficiency of CG dinucleotides in the human genome. The creation of new CG dinucleotides is not an adequate counter balance against this effect, due to the lower frequency of transversions back to CpG. While SNPs and mutations both arise in the same way, their survival in populations is likely to be quite different. Most newly arisen SNPs and mutations are likely to be lost in early generations by random sampling of the gene pool alone. For example if a heterozygous individual for a selectively neutral mutation has two offspring, there is a 0.75 probability that the mutation will be found in at least one child. If each generation has two children, the probability of loss of the new mutation is $1-(0.75)^g$, where g = generations. To give a worked example, this relates to a 94% probability of loss of a mutation or SNP in 10 generations (approximately

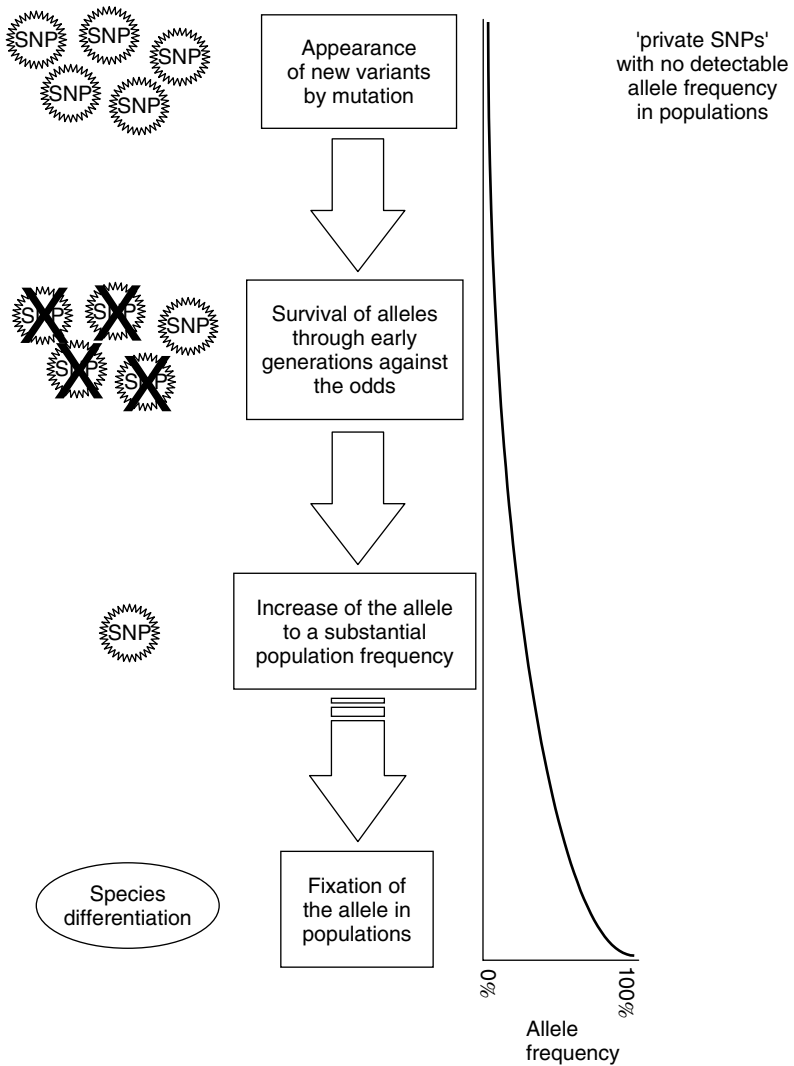


Figure 3.1 The life cycle of SNPs and mutations. SNP and mutation evolution occurs in four main phases: (1) appearance of a new variant allele by mutation; (2) survival of the allele through early generations against the odds; (3) increase of the allele to a substantial population frequency; (4) fixation of the allele in populations.

200 years). Where a heterozygous mutation has an early onset deleterious effect, natural selection is likely to further increase the rate of loss of the allele from populations. The same pressures do not apply to late onset diseases, perhaps explaining the proliferation of such diseases in humans.

If an SNP or mutation survives early generations and increases in frequency sufficiently to become homozygous in some individuals the risk of loss of the allele is reduced. At this stage the frequency of the allele in a population is likely to vary, with higher frequency

alleles being consistently favoured, especially when populations are subject to severe bottlenecks in size. Reich *et al.* (2001) presented convincing evidence for such a bottleneck in recent Northern European population history. In the face of these fluctuations of allele frequency, an SNP or mutation will cease to exist in populations, either by disappearing or by reaching a 100% allele frequency, in which case the variant becomes an allele that helps to define a species. Interestingly there is no evidence of shared SNPs between species, a study of variation between the human and orang-utan X chromosome found that 2.9% of nucleotide sites differ, but no SNPs were shared (Miller *et al.*, 2001). This suggests that the lifetime of an SNP is considerably shorter than the divergence of these two species. Based on this data, Miller *et al.* (2001) estimated that the average period from original mutation to species fixation of an allele was 284,000 years.

3.2.1.2 SNP and Mutation Databases United?

The high level of interest in SNP data has led to the development of an excellent central SNP database—dbSNP at the NCBI (Sherry *et al.*, 2001). Mutation databases are still lagging behind SNPs in terms of data integration and visualization on the human genome. However the many commonalities between these two forms of data may have inspired the SNP database HGBase to re-align and rename itself HGVBASE—a central database of human genetic variation including SNP and mutation data (Fredman *et al.*, 2002). This is a valuable step which will make mutation data much more accessible to geneticists in a well-integrated form. Other highly specialized mutation databases exist, including HGMD, GDB and a large range of locus-specific databases. It is not yet clear to what extent mutation and SNP data will be integrated, but the availability of a complete human genome presents an unbeatable opportunity to bring these two sources of data together in a genomic context, without compromising the necessary integrity of either form of data.

3.2.2 Tandem Repeat Polymorphisms

Tandem repeats or variable number repeat polymorphisms (VNTRs) are a very common class of polymorphism, consisting of variable length sequence motifs that are repeated in tandem in a variable copy number (Figure 3.2). VNTRs are only surpassed in quantity by SNPs in the human genome. They have been found in all organisms studied, although they

Repeat type	Example
Mononucleotide	AAAAAAAAAAAAAAAAAAAA
Dinucleotide	CACACACACACACACACA
Triplet/trinucleotide	CAGCAGCAGCAGCAGCAGCAGCAG
Tetranucleotide	TAAGTAAGTAAGTAAGTAAGTAAGTAAG
Pentanucleotide etc.	GAATTGAATTGAATTGAATTGAATTGAATT
Repeat terminology	Example
Perfect STR	CACACACACACACACACACACACACACA
Imperfect STR	CACATACACACACACACACGCACACACA
Interrupted STR	CACACACACACGGGCACACACACACACA
Compound STR	CACACACACACACATGTGTGTGTGTGTG

Figure 3.2 Tandem repeat types and terminology.

tend to occur at higher frequencies in organisms with large genomes. Viknaraja *et al.* (unpublished data) analysed the draft human genome sequence (December 2001 freeze) and identified several hundred thousand potentially polymorphic VNTRs. However there is little or no information on the heterozygosity and polymorphic nature of the vast majority of these polymorphisms. VNTRs have traditionally been subdivided into subgroups based on the size of the tandem repeat unit. Repeated sequences of one to six bases are termed microsatellites or short tandem repeats (STR), larger tandem repeats in units of 14–100 bp are termed minisatellites. Microsatellites and minisatellites are generally thought to show different mutational mechanisms which are influenced by sequence properties and lengths. In microsatellites the predominant mutational mechanism is thought to be DNA slippage during replication. In minisatellites the predominant mechanism appears to be gene conversion and unequal crossing over (Goldstein and Schlotterer, 1999). The distinction between microsatellites and minisatellites is somewhat arbitrary for repeat units between 7 and 13 bp and it has been suggested that highly repeated sequences or sequences which are more likely to form loops in these size categories should be called minisatellites. This somewhat vague definition may be academic, in effect microsatellites and minisatellites have quite different properties, dictated by their repeat size, copy number and the perfection of the repeat. For the specific needs of a genetic study it may be necessary to pick the tandem repeat which conforms most closely to the heterozygosity requirements for the marker (see Chapter 8). The polymorphic nature of a VNTR is thought to depend upon a range of factors: the number of repeats, their sequence content, their chromosomal location, the mismatch repair capability of the cell, the developmental stage of the cell (mitotic or meiotic) and/or the sex of the transmitting parent. (Debrauwere *et al.*, 1997).

Aside from their utility as highly polymorphic genetic markers, much evidence exists to demonstrate that tandem repeats exert a functional effect when located in or near gene coding or regulatory regions. Thus VNTRs in themselves can be candidates for disease-causing genetic variants. The best characterized of these are the triplet repeat expansion diseases. Triplet repeat expansion is an insertion process that occurs during meiosis. Insertion of new repeats is strongly favoured over loss of repeats—pathological triplet repeat expansions manifest through successive generations with worsening symptoms known as ‘anticipation’, as the repeat expands with increasingly pathological results. Most triplet repeat expansions have been identified in monogenic diseases and may occur in almost any genic region. Over five triplet repeat classes have been described so far, causing a range of diseases including, Fragile X, myotonic dystrophy, Friedreich’s ataxia, several spinocerebellar ataxias and Huntington’s disease (Usdin and Grabczyk, 2000). Spinocerebellar ataxia 10 (SCA10) is notably caused by the largest tandem repeat seen in the human genome (Matsuura *et al.*, 2000). In general populations the SCA10 locus is a 10–22mer ATTCT repeat in intron 9 of the SCA10 gene; in SCA10 patients, the repeat expands to >4500 repeat units, which makes the disease allele up to 22.5 kb larger than the normal allele.

Tandem repeats have also been associated with complex diseases, for example different alleles of a 14mer VNTR in the insulin gene promoter region, have been associated with different levels of insulin secretion. Different alleles of this VNTR have been robustly linked with type I diabetes (Lucassen *et al.*, 1993) and in obese individuals they have also been associated with the development of type II diabetes (Le Stunff *et al.*, 2000). Kubota (2001) took the concept of triplet repeat anticipation to an extreme by suggesting that every human chromosome suffers from a burden of accumulating trinucleotide repeats. Thus, he predicted the ‘mortality’ of human chromosomes with the passage of generations,

eventually leading to a deficiency of replication and to the mortality of *Homo sapiens* as a species! This is certainly a controversial theory, but the basic concept is interesting and illustrates that the burden of VNTR-mediated genetic disease is only likely to increase.

The value of tandem repeats as markers and functional elements is clear, although for practical reasons the focus of genetics is shifting to SNPs. However, VNTR markers will probably continue to be a fundamental tool and to overlook them could be unwise, as often a highly polymorphic VNTR may be more informative than several SNPs. In comparison to the relatively low heterozygosity of SNPs, much less dense VNTR maps are needed to match the equivalent detection power of a high density SNP map (see Chapter 7). A single polymorphic VNTR may even be as informative as a complex SNP haplotype. The drawback of tandem repeats are mainly technological—detection methods cannot currently match the highly automated microtitre plate-based or DNA chip-based assays that have characterized modern SNP assay development, although technology developments may eventually alter this situation (Krebs *et al.*, 2001).

In comparison to the hundreds of thousands of VNTR polymorphisms in the genome, only 18,000 VNTRs have been genetically characterized. Several highly characterized subsets of these markers have been arranged into well-defined linkage marker panels by the Marshfield Institute and Genethon (see Chapter 7 for details). These panels vary in marker spacing to allow different density genome scans. Almost all genetically characterized VNTRs are stored centrally in several sources, including GDB, CEPH and dbSTS (see below). Potentially polymorphic novel VNTRs can be identified from genomic sequence using the tandem repeat finder tool (Benson, 1999; <http://c3.biomath.mssm.edu/trf.html>). A complete analysis of the human genome sequence using tandem repeat finder is presented in the UCSC human genome browser in the ‘simple repeats’ track (see Chapter 9).

3.2.3 Insertion/Deletion Polymorphisms and Chromosomal Abnormalities

While tandem repeat polymorphisms are in themselves a major form of variation in genomes, they may also mediate other forms of variation by predisposing DNA to localized rearrangements between homologous repeats. Such rearrangements give rise to Insertion/Deletion (INDEL) polymorphisms. Indels appear to be quite common in most genomes studied so far, this probably reflects their association with common VNTRs. Indels have been associated with an increasing range of genetic diseases, for example, Cambien *et al.* (1992) found association between coronary heart disease and a 287-bp Indel polymorphism situated in intron 16 of the angiotensin converting enzyme (ACE). This Indel, known as the ACE/ID polymorphism, accounts for 50% of the inter-individual variability of plasma ACE concentration. The molecular mechanism of insertion/deletion polymorphism is still poorly understood, many different molecular mechanisms may account for an Indel event, although most are likely to be DNA sequence dependent. As discussed earlier, localized sequence repetitiveness in the form of direct tandem repeats or inverted repeats or ‘symmetric elements’, have been shown to predispose DNA to insertion/deletion events (Schmucker and Krawczak, 1997). Darvasi and Kerem (1995) found evidence to suggest that slipped-strand mispairing (SSM) was a common mechanism for insertion/deletion events. Analysis of sequences surrounding 134 disease-causing Indel mutations in the coding regions of three genes, the cystic fibrosis transmembrane conductance regulator, beta globin and factor IX, found that 47% of Indel mutations occurred within a unit repeated tandemly two- to seven-fold. The proportion of SSM mutations was significantly higher than expected by chance. The estimated net proportion of deletion and insertion mutations attributed to SSM was 27%. Further mechanisms have been

proposed; Deininger and Batzer (1999) suggested that many INDELs may be caused by the insertion of Alu elements, which number in excess of 500,000 copies in the human genome providing abundant opportunities for unequal homologous recombination events.

Although Indel polymorphisms are likely to be very widely distributed throughout the genome, relatively few have been characterized and there is no central database collating this form of polymorphism. The Marshfield website maintains the most comprehensive single source of short insertion/deletion polymorphisms (SIDPs), over 2000 are maintained in a form which can be searched by chromosome location. Other databases such as dbSNP and HGVBASE also capture SIDPs to some extent. Larger Indels are generally overlooked in databases unless associated with a specific gene or study, in which case they appear in GDB, OMIM and other similar sources.

3.2.4 Gross Chromosomal Aberrations

While minor Indel polymorphisms are thought to be relatively common in human populations, gross chromosomal abnormalities such as deletions, inversions or translocations were thought to be rare. Nevertheless as our knowledge of the genome develops an increasing number of clinically characterized genomic syndromes are being identified. Some of these affect multiple genes and cause pronounced phenotypes including velocardiofacial syndrome (VCFS) a deletion syndrome on 22q11.2 (Gong *et al.*, 1996) and Charcot-Marie-Tooth disease type 1A (CMT1A) a duplication syndrome on 17p11.2 (Thomas, 1999). Other much more subtle genomic syndromes are emerging which suggest that these syndromes may in fact be more common than previously believed. DUP25 is an interstitial duplication of 17Mb at 15q24–26, which is associated with joint laxity and panic disorder (Gratacos *et al.*, 2001). Changes in dosage of one or more of the 59+ genes in the DUP25 region are likely to contribute to the subtle clinical phenotype. Detection of DUP25 was not easy as it shows non-Mendelian transmission precluding straightforward linkage analysis. Instead researchers used laborious cytogenetic methods to detect the duplication. This analysis identified DUP25 in 90% of patients with one or more anxiety disorders, and in 80% of subjects with joint laxity and remarkably in 7% of French population-based controls.

These genomic disorders are generally thought to be caused by aberrant recombination at region- or chromosome-specific low-copy repeats, known as segmental duplications (Emanuel and Shaikh, 2001). This new class of repetitive DNA element has only been identified very recently, largely as a result of human genome sequencing. Segmental duplications result from the duplication of large segments of genomic DNA that range in size from 1 to 400 kb. These duplications can mediate interchromosomal or intrachromosomal recombination events. Knowledge that relatively common diseases can be caused by recurrent chromosomal duplications and deletions has demonstrated that potential for genomic instability could be directly related to the structure of the regions involved. The sequence of the human genome offers to add insight and understanding to the molecular basis of such recombination 'hot spots'. This insight is already being gained, in the case of VCFS on 22q11.2 complete genomic sequence across the region has revealed four segmental duplications flanking the VCFS deletion region (Shaikh *et al.*, 2001).

Availability of information on known deleted or duplicated regions varies greatly; some have been narrowed to fairly well-defined critical regions, others are very poorly defined. Details of some of the more extensively characterized deletion/rearrangement syndromes are captured in GDB and OMIM, although in most cases information is spread throughout the literature and basically needs to be hunted down on a case by case basis. The UCSC

human genome browser is a particularly useful ally in this hunt (see Chapter 5), as it annotates large duplicated regions in the human genome. The objective of this annotation is primarily to identify duplication errors in human genome contig assembly, but this also effectively identifies segmental duplications, such as the duplications flanking the VCFS region on 22q11.2.

3.2.5 Somatic Mutations

A completely distinct category of human mutations arises somatically during the process of tumourgenesis. These mutations may take many forms, the most commonly characterized are somatic point mutations identified during the screening of candidate genes in tumour tissues. Cytogenetic studies of human neoplasias have also identified a number of chromosomal aberrations involving large deletions and duplications (Shapira, 1998). As somatic mutations are not inherited it is obviously important to avoid mixing somatic point mutation data with human polymorphism and mutation data.

3.2.5.1 Somatic Point Mutations

Screening of candidate genes for point mutations in tumour material has identified a number of key genes with a role in cancer. There is no central database containing point mutation data identified during these screens, although some locus-specific databases do exist, it is not possible to list all these specialist resources. In some cases it may be possible to identify locus-specific databases by a gene-specific websearch (e.g. using SCIRUS, see Chapter 2). In most cases mutation data needs to be identified directly from the literature.

3.2.5.2 Genomic Aberrations in Cancer

Almost 100,000 neoplasia-associated chromosomal abnormalities have been characterized at the molecular level, revealing previously unknown genes that are closely associated with tumourgenesis. It is not clear if somatic chromosomal aberrations and genomic syndromes share any common mechanisms, such as mediation by segmental duplications, although this is a possibility. Prospects for informatic and laboratory study of chromosomal aberrations in cancer are assisted by the availability of a centralized database to capture this data, the Mitelman map of chromosomal aberrations in cancer. This resource has been integrated into the NCBI MapViewer tool and the Cancer Genome Anatomy Project (CGAP) (see Table 3.1).

3.3 DATABASES OF HUMAN GENETIC VARIATION

The vast range of human genetic variation is still largely uncharted and what information exists cannot be derived from a single database. At best the data needs to be gathered from several databases or worse still the data may not be readily available in a database at all, in which case detailed literature and internet searching or bioinformatic analysis approaches may be necessary. Having described the main forms of human variation, we will now introduce the key databases for mining this information. We will also examine how these genetic databases integrate with other databases and the human genome sequence to add a full genomic context to variation, to help in the characterization of a potential genetic lesion. Table 3.1 presents a selection of the best tools and databases for this purpose.

TABLE 3.1 Genetic Variation-Focused Databases and Tools on the Web**Mutation databases**

OMIM	http://www.ncbi.nlm.nih.gov/Omim/
HGMD	http://www.hgmd.org
GDB Mutation Waystation	http://www.centralmutations.org/
HUGO mutation database initiative	http://www.genomic.unimelb.edu.au/mdi/

Central databases (SNPs and mutations)

HGVbase	http://hgvbase.cgb.ki.se/
Sequence variation database (SRS)	http://srs.ebi.ac.uk/
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/
The SNP consortium (TSC)	http://snp.cshl.org/

Genetic marker maps (microsatellites, STSs other markers)

Marshfield maps	http://research.marshfieldclinic.org/genetics/
Genome Database (GDB)	http://www.gdb.org
dbSTS	http://www.ncbi.nlm.nih.gov/STS/
UniSTS	http://www.ncbi.nlm.nih.gov/genome/sts/

Somatic and non-nuclear mutation databases

MitoMap	http://www.gen.emory.edu/mitomap.html
Mitelman Map	http://cgap.nci.nih.gov/Chromosomes/Mitelman

Gene-orientated SNP and mutation visualization

LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/
PicSNP	http://picsnp.org
Protein Mutation Database	http://www.genome.ad.jp/htbin/www_bfind?pmd
Go!Poly	http://61.139.84.5/gopoly/
GeneLynx	http://www.genelynx.org
SNPper	http://bio.chip.org:8080/bio/snppeer-enter
GeneSNPs	http://www.genome.utah.edu/genesnps/
CGAP SNP database	http://lpgws.nci.nih.gov/

Genome-orientated for SNP and mutation visualization

Ensembl	http://www.ensembl.org
Human Genome Browser (UCSC-HGB)	http://genome.ucsc.edu/index.html
Map Viewer	http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch

3.4 SNP DATABASES

The deluge of SNP data generated over the past 2 years can primarily be traced to two major overlapping sources: The SNP Consortium (TSC) (Altshuler *et al.*, 2000) and members of the Human Genome Sequencing Consortium, particularly the Sanger Institute and Washington University. The predominance of SNP data from this small number of closely related sources has facilitated the development of a central SNP database — dbSNP at the NCBI (Sherry *et al.*, 2001). Other valuable databases have developed using dbSNP data as a reference, these tools and databases bring focus to specific subsets of SNP data, e.g. gene-orientated SNPs, while enabling further data integration around dbSNP.

3.4.1 The dbSNP Database

The National Center for Biotechnology Information (NCBI) established the dbSNP database in September 1998 as a central repository for both SNPs and short INDEL polymorphisms. In May 2002 (Build 104) dbSNP contained 4.2 million SNPs. These SNPs collapse into a non-redundant set of 2.7 million SNPs, known as Reference SNPs (RefSNPs). Approximately 10% of these RefSNPs do not currently map to the draft human genome, which leaves 2.43 million SNPs with potential utility for genetic mapping. These quantities of SNPs give a high level of coverage across the genome. One study estimated that 85% of all known exons are within 5 kb of an SNP in the dbSNP database (International SNP Map Working Group, 2001). These figures will have undoubtedly improved considerably by the time this book comes to press.

3.4.1.1 The Reference SNP Dataset (RefSNPs)

The non-redundant RefSNP dataset is produced by clustering SNPs at identical genomic positions and creating a single representative SNP (designated by an 'rs' ID). The sequence used in the RefSNP record is derived from the SNP cluster member with the longest flanking sequence; this sequence is derived from one individual and is not a composite sequence assembled from the cluster. The RefSNP record collates all information from each member of the cluster, e.g. frequency information. The availability of the RefSNP dataset considerably streamlines the process of integrating SNPs with other data sources. External resources generally use the RefSNP dataset which makes the RefSNP ID the universal SNP ID in the SNP research community. RefSNPs have also become an integral part of the NCBI data infrastructure, so that the user can effortlessly browse to dbSNP from diverse NCBI resources, including LocusLink, Map View and Genbank.

3.4.1.2 Searching dbSNP

There are a bewildering range of approaches for searching dbSNP. The database can be searched directly by SNP accession number, submitter, detection method, population studied, publication or a sequence-based BLAST search. The database also has a complex search form which allows more flexible freeform queries (<http://www.ncbi.nlm.nih.gov/SNP/easyform.html>). This allows the user to select SNPs which meet several criteria, for example it is possible to search for all validated non-synonymous SNPs in gene coding regions on chromosome 1 (Figures 3.3 and 3.4). The advanced form also includes a separate interface for retrieving all SNPs between two STS markers or two golden path locations.

There are many other tools which use the dbSNP dataset, e.g. LocusLink, SNPper and the human genome browsers (Table 3.1). These tools can offer powerful alternative interfaces for searching dbSNP, but be aware that third party tools and software may use filtering or repeat masking protocols, which can lead to the exclusion of SNPs with poor quality or short flanking sequence, or SNPs in repeat regions. If it is important to identify *all* SNPs in a given gene or locus then it is worth consulting several different tools and comparing the results. Some of the best SNP visualization tools are discussed later in this chapter.

3.4.1.3 Submitting Data to dbSNP

The dbSNP database accepts direct data submissions from researchers by e-mail or FTP. The submission process is generally intended for large batch submissions involving hundreds or thousands of SNPs, using a text flatfile submission format. Each SNP submission

NCBI Single Nucleotide Polymorphism

PubMed Entrez BLAST OMIM Taxonomy Structure

Search GenBank for [] Go

Search Form

Organism: ALL AND

Chromosome: 1 AND

Function class: Coding-Any AND

Genome mapping results: ALL AND

NCBI reference cluster ID (rs#): Between [] and [] AND

Success rate (integer 1-100): Between [] % and [] % AND

Heterozygosity (real 0.0 - 1.0): Between [] and [] AND

Validated: True AND

Gene symbol: [] AND

LocusID: [] AND

Accession: []

This search may take a few minutes to complete.

The maximum number of returned SNP id for each query is 30,000. Please download from the [ftp site](#) to obtain larger data set.

GENERAL: [Home Page](#) | [Overview](#) | [dbSNP Summary](#) | [How To Submit](#) | [Genome](#) | [FAQ](#) | [RefSNP Summary Info](#) | [FTP SERVER](#) | [Database Schema](#) | [Build History](#) | [Blast SNP](#)

Figure 3.3 The dbSNP freeform search interface.

contains many elements to describe the SNP, but primarily it should contain a report describing how to assay the SNP, the SNP sequence information and if available the SNP allele frequency. While the submission format is suitable for bulk submissions it may present the occasional submitter some problems. Preparation of any more than a handful of SNPs in this format really requires some grasp of a text manipulation language such as perl (Stein, 2001). In this case it may be a good idea to find a friendly perl programmer or contact dbSNP directly for guidance and assistance in the preparation of the submission. A web interface for form-based submission is currently in development, which should alleviate this problem.

Query:

Total number of SNPs found: 231

Click to download list of NCBI refSNP cluster ID (#RS)

Request result in other format :

The result will be sent to the email address you provide below.

Email:

Items 1-25 of 231 Page of 10

rs	Map	Gene	Het	Validation	Genotypes Avail	
					Linkout	Avail
rs242		LTC				
rs1085		LTC		☆		
rs1250		LTC		☆		
rs1344		LTC		☆		
rs1921		LTC		☆		
rs1956		LTC		☆		
rs3052		LTC		☆		
rs4230		LTC				
rs5257		LTC		☆		
rs5273		LTC				
rs5274		LTC				
rs5277		LTC				

Figure 3.4 Search results from a dbSNP freeform search.

3.4.1.4 Key SNP Data Issues

The sequencing of the human genome has provided a massive boost to human polymorphism discovery efforts. Table 3.2 presents a breakdown of dbSNP submission sources. From this table it is clear that 94% of SNPs in dbSNP originate from three main sources: the TSC, the Sanger Institute and the Kwok Laboratory (informatic analysis of data from the Whitehead Institute and Washington University). SNPs sourced from the TSC were identified by the major genome sequencing centres by detection of high-confidence base differences in aligned sequences primarily from reduced representation shotgun (RRS) sequencing (Altshuler *et al.*, 2000) and also by alignment of genomic clones (Mullikin *et al.*, 2000). RRS sequencing involves sequencing of random clones from the genomes of many individuals. This method has several advantages over other SNP identification methods, in that it does not require previous knowledge of genomic sequence or PCR, and it provides haploid genotypes, the alleles of which are easier to call (see Chapter 10 for an overview of these methods). The later two sources, SANGER and KWOK account for 64% of dbSNP SNPs. These represent SNPs generated by the major human genome sequencing centres. These SNPs were identified by overlapping genomic sequence reads.

TABLE 3.2 Main SNP Submission Sources in the dbSNP Database (BUILD 104)

Source	Total submissions	RefSNP clusters	Primary SNP ID method
TSC	1,279,099	1,275,272	Shotgun and Genomic
Kwok (WASHU)	1,182,884	493,536	Genomic overlap
Sanger	1,529,560	1,348,534	Genomic overlap
Lee	99,505	46,942	EST trace mining
Yusuke	73,720	73,720	SNP disc (Japanese)
Perlegen	25,326	25,315	Microarray (Chr21 only)
HGBASE	13,100	13,081	Various
CGAP	12,881	12,733	EST trace mining
Other	13,367	ND	Various
Total	4,229,442	2,673,925	

In the wake of the TSC and the genomic overlap SNP discovery projects, further SNP submissions to dbSNP will continue from the genome centres in the final stages of genome finishing, but further growth of dbSNP will depend on the next steps after completion of the human genome. The human genome is likely to be repeatedly re-sequenced in the next few years, either entirely or across defined regions. This will in turn generate further SNPs by comparison of genomic overlaps. The Sanger Institute has already announced a 5-year plan to re-sequence all known human exons in 96 individuals. This should detect 95% of SNPs with a frequency of >1%. Inevitably novel SNPs will become increasingly rare, based on a law of diminishing returns. Based on the observed SNP density in the genome, estimates suggest that the dbSNP dataset may currently represent 20–30% of common SNPs in the human genome. Different SNP discovery projects have sampled variation at very different levels. The TSC SNPs were discovered using a publicly available panel of 24 ethnically diverse individuals (Collins *et al.*, 1998). This panel would have a 95% chance of detecting SNPs down to a frequency of 5%. SNPs identified by genomic sequence overlap (which comprise 64% of dbSNP data), offer the shallowest sampling of human variation. Genomic overlap SNPs are candidate SNPs identified by comparison of two individuals, this approach has some major drawbacks, the SNP discovery method is more error prone (heterozygotic SNPs are often missed) and many SNPs discovered by this method are likely to be ‘private’ SNPs which are restricted to the individual and not generally represented in populations (see below for more details on candidate SNP issues).

Aside from the major SNP data submissions from the genome centres, dbSNP also accepts direct SNP submissions from researchers and most journals now require SNP submission to dbSNP before publication (a practice which needs to be encouraged). These have been estimated to add to dbSNP at a rate of about 100 primarily gene-orientated SNPs per month.

3.4.1.5 Candidate SNPs – SNP to Assay

As we have already demonstrated, the dbSNP dataset has one overwhelming caveat — most of the SNPs are ‘candidate’ SNPs of unknown frequency and are unconfirmed in a laboratory assay. This translates to the simple fact that many SNPs do not exist at a detectable frequency in any population. Over 60% of the SNPs in dbSNP were detected by statistical methods for identification of ‘candidate’ SNPs by comparison of DNA sequence traces from overlapping clones. Marth *et al.* (2001) investigated the reliability of these candidate

SNPs in some depth completing two pilot studies to determine how well candidate SNPs would progress to working assays in three common populations. In both studies, they found that between 52–54% of the characterized SNPs turn out to be common SNPs (above >10%) for each population. Significantly, between 30 and 34% of the characterized SNPs were not detected in each population. These results suggest that if a candidate SNP is selected for study in a common population, there is a 66–70% chance that the SNPs will have detectable minor allele frequency (1–5%) and a 50% chance that the SNPs are common in that population (>10%). Put another way, ~17% of candidate SNPs will have no detectable variation in common populations, these ‘monomorphic’ SNP candidates, are likely to represent ‘private SNPs’, which exist in the individual screened but not appreciably in populations. This probably reflects the massive increase in population size and admixture over the past 500 years (Miller and Kwok, 2001). Beyond validation of the SNP, the last hurdle is assay design—many SNPs are located in repetitive or AT rich regions, which makes assay design difficult, this can account for a further 10–30% fallout, depending on the assay technology.

Any genetic study needs to take these levels of attrition between SNP and working assay into account (Table 3.3). There is only one solution to this problem—to determine the frequency of the 2 million or so public SNPs in common ethnic groups. This is now widely recognized in the SNP research community and several public groups including the TSC are already undertaking or seeking to undertake large-scale SNP frequency determination projects.

3.4.2 Human Genome Variation Database (HGvbase)

The Human Genome Variation database, HGvbase, previously known as HGbase (Brookes *et al.* 2000; <http://hgbase.cgb.ki.se/>), was initially created in 1998 with a remit to capture all intra-genic (promoter to end of transcription) sequence polymorphism. One year later, the remit of the database expanded to a whole genome polymorphism (and nominally mutation) database, this ambitious expansion in remit was supported by the establishment of a European consortium comprising teams at the Karolinska Institute, Sweden, the European Bioinformatics Institute, UK and at the European Molecular Biology Laboratory, Germany. At this point, HGbase encompassed the same classes of variants as dbSNP. Both HGvbase and dbSNP make regular data exchanges to allow data synchronization. In November 2001, the HGbase project adopted the new name HGvbase (Human Genome Variation database; Fredman *et al.*, 2002). This change reflected another change in the scope of the database as it took on a HUGO endorsed role as a central repository for mutation collection efforts undertaken in collaboration with the Human Genome Variation Society (HGVS).

TABLE 3.3 Pitfalls from Candidate SNP to Assay (From Marth *et al.*, 2001)

SNP to assay conversion steps	Remaining RefSNPs (% attrition)
Reference SNP identified	2.4 M
Not mapped to human genome	2.16 M (10%)
Assay design not possible or assay fails	1.84 M (15%)
Not polymorphic in study population	1.52 M (17%)
Frequency <20% in chosen population	1.26 M (50%)
SNPs (>20% frequency) with assay available	0.63 M

There is no doubt that dbSNP has assumed the *de facto* position of the primary central SNP database. To accommodate this, HGVbase has assumed a complementary position, with a broader remit covering all single nucleotide variation—both SNPs and mutations. HGVBASE is also taking a distinct approach to dbSNP by seeking to summarize all known SNPs as a semi-validated, non-redundant set of records. HGVbase is seeking to address some of the problems associated with candidate SNPs and so in contrast to the automated approach of dbSNP, HGVbase is highly curated. The curators are aiming to provide a more-extensively validated SNP data set, by filtering out SNPs in repeat and low complexity regions and by identifying SNPs for which a genotyping assay can successfully be designed. The HGVbase curators have also identified SNPs and mutation data from the literature, particularly older publications before database submission was the norm. HGVbase currently contains 1.45 M non-redundant human polymorphisms and mutations (release 13–March 2002).

HGVbase is a highly applied database, which also provides some useful tools for experimental design, including a tool for defining haplotype tags—‘Tag ’n Tell’. This tool will find a minimum set of markers that uniquely characterize (or ‘tag’) chosen haplotypes. According to user preferences, not all entered haplotypes have to be considered in the tag-selection process, this is useful for determining optimal haplotype tag sets to capture common haplotypes (see Chapter 9 for an example of haplotype tagging using this tool).

The HGVbase search interface is relatively simple, tools are available to facilitate BLAST searching and keyword queries of the database. As these options are relatively limited, other tools which access HGVbase data, are a better option—most are from the EMBL and EBI organizations, including Ensembl and SRS (Table 3.3; described below). The *in silico* quality control approach adopted by HGVbase is valuable, particularly for the broader biological community of SNP data consumers. For the geneticist, HGVbase serves to identify SNPs with a higher chance of converting from ‘candidate SNPs’ to informative SNP assays. If you take the cost of failed assays into account, this is a valuable objective, although if all available SNPs need to be identified it may still be important to search dbSNP and other resources.

3.5 MUTATION DATABASES

The polymorphism data stored in dbSNP is valuable biological information that helps to define the natural range of variation in genes and the genome, however most of the polymorphisms can be assumed to be functionally neutral. By contrast human mutation data is functionally defined and has obvious implications for the nature and prevalence of disease and the pathways underlying disease. This makes the study of naturally-occurring mutations important for the understanding of human disease pathology, particularly the relationships between genotype and phenotype and between DNA and protein structure and function. A large number of Mendelian disease mutations have been identified over the past 20 years. These have helped to define many key biological mechanisms, including gene regulatory motifs and protein–protein interactions (see Chapter 13). Many highly specialized locus-specific databases (LSDBs) have been established to collate this data. This chapter could not hope to cover all these databases, but there are now several centralized resources which index and provide links to some of the larger resources. Other ‘boutique’ databases can sometimes be identified by general web searching (see Chapter 2).

3.5.1 The Human Gene Mutation Database (HGMD)

The HGMD was established in April 1996 to collate published germline mutations responsible for human inherited disease. In October 2001, HGMD contained 26,637 mutations

in 1153 genes. The scope of HGMD is limited to mutations leading to a defined inherited phenotype, including a broad range of mechanisms, such as point mutations, insertion/deletions, duplications and repeat expansions within the coding regions of genes. Somatic mutations and mutations in the mitochondrial genome are not included. HGMD invites submissions from researchers but most records are curated directly from mutation reports in more than 250 journals and directly from the LSDBs which are comprehensively linked. To be included, there must be a convincing association between the mutation and the phenotype. All mutations in HGMD are represented in a non-redundant form which unfortunately does not conserve all the redundant mutations constituting the cluster, so it is not possible to determine if mutations are identical by descent, also data is lost on the frequency of mutations. The HGMD search interface is primarily text based and targeted searching tends to rely on knowledge of the correct HUGO nomenclature for a gene.

3.5.2 Sequence Variation Database (SRS)

The sequence variation database forms part of the Sequence Retrieval Server (SRS) at the EBI, Hinxton UK. SRS is a flexible sequence query tool which allows the user to search a defined set of sequence databases by accession number, keyword or sequence similarity. Several categories of sequence variation are encompassed by SRS, including HGVbase and a large number of locus specific databases which are listed in Table 3.4.

3.5.3 The Protein Mutation Database (PMD)

The Protein Mutation Database (PMD) is unique among genetic variation databases as it contains both natural and artificial mutation data derived from human proteins (Kawabata *et al.*, 1999). The artificial mutation data is derived from the literature and mainly consists of site-directed and random mutagenesis data. It is important to clearly delineate artificial data and so each record is clearly defined as either natural or artificial. The database gives detailed description of the functional or structural effects of the mutations if known and provides links to the original publications. Relative differences in activity and/or stability, in comparison with the wild-type protein, are also indicated. PMD contains 119,190 natural and artificial mutations (January 2002) and these can be searched by keyword or sequence similarity (BLAST), a complete report on the mutated protein sequence is displayed which allows the user to see the position of altered amino acids. Where 3D structures have been experimentally determined, PMD displays mutated residues in a different colour on the 3D structure.

The Protein Mutation Database is very valuable for the functional analysis of proteins. The detailed functional characterization of mutations gives the user an opportunity to compare known mutations with variations in orthologous residues in related proteins. The data is also useful to aid in the delineation of the functional domains of proteins in the database and other homologous proteins (see Chapter 14 for further examination of such approaches for mutation analysis).

3.5.4 On-line Mendelian Inheritance in Man (OMIM)

OMIM is an on-line catalogue of human genes and their associated mutations, based on the long running catalogue Mendelian Inheritance in Man (MIM), started in 1967 by Victor McKusick at Johns Hopkins (Hamosh *et al.*, 2000). OMIM is an excellent resource for providing a brief background-biology on genes and diseases, it includes information on the most common and clinically significant mutations and polymorphisms in genes. Despite the name, OMIM also covers complex diseases in varying degrees of detail.

TABLE 3.4 Locus-Specific Databases Indexed by the Sequence Variation Database

Name	Description	Entries
General mutation databases		74,117
EMBLCHANGE	Sequence change features from EMBL	32,863
SWISSCHANGE	Sequence change features from SWISS-PROT	17,294
OMIMALLELE	Alleles from OMIM	9344
HUMUT	Protein Mutation Databank	14,616
Mitochondrial genome		9401
HUMAN_MITBASE	Human mitochondrial DNA variants	9401
Locus-specific mutation databases		240,73
P53LINK	p53 mutations database	14,834
APC	APC mutation database	825
BTKBASE	Bruton's tyrosine kinase mutations	454
VWF	von Willebrand factor gene variations	144
CFTR	Cystic fibrosis mutation database	809
PAH	Phenylalanine hydroxylase mutations	289
HAEMA	Haemophilia A, Factor VIII mutations	604
HAEMB	Haemophilia B	1722
LDLR	Low-density lipoprotein receptor	283
PAX6	PAX6 mutation database	118
EMD	Emery–Dreifuss muscular dystrophy	87
L1CAM	Neuronal cell adhesion molecule gene mutations	91
CD40LBASE	CD40 ligand defects	60
G6PD	Glucose-6-phosphate dehydrogenase variants	122
ANDROGENR	Androgen receptor mutations	514
RDS	Retinal degeneration slow gene mutations	33
RHODOPSIN	Rhodopsin gene mutations	133
FANCONI	Fanconi anaemia mutation database	32
HEXA	Hexosaminidase A mutations	89
XCGDBASE	X-linked chronic granulomatous disease	303
DMD	Duchenne/Becker muscular dystrophy	184

(continued overleaf)

TABLE 3.4 (*continued*)

Name	Description	Entries
FVII	Factor VII mutation database	176
ATM	Ataxia–telangiectasia mutation database	200
P16	CDKN2A/P16NK4A mutation database	146
GAA	Acid alpha-glucosidase mutation database	83
OTC	Ornithine transcarbamylase (OTCase) mutations	105
IL2RGBASE	Interleukin-2 receptor gamma mutations	161
BIOMDB	Database of tetrahydrobiopterin deficiency mutations	78
Central databases		984,093
HGVbase	Human Genome Variation database (SNPs and mutations)	984,093

In January 2002, the database contained over 13,285 entries (including entries on 9837 gene loci and 982 phenotypes). OMIM is curated by a dedicated but small group of curators, but the limits of a manual curation process mean that entries may not be current or comprehensive. With this caveat aside OMIM is a very valuable database, which usually presents a very accurate digest of the literature (it would be difficult to do this automatically). A major added bonus of OMIM is that it is very well integrated with the NCBI database family, this makes movement from a disease to a gene to a locus and vice versa fairly effortless.

3.6 GENETIC MARKER AND MICROSATELLITE DATABASES

3.6.1 dbSTS and UniSTS

dbSTS is an NCBI database containing sequence and mapping data for Sequence Tagged Sites (STSs) (Olson *et al.*, 1989). These STSs can include polymorphic sequences such as short tandem repeats (STRs), or non-polymorphic sequences. In fact any unique genomic landmark which can be amplified by PCR can be used as an STS marker. Both polymorphic and non-polymorphic STS markers have been used to construct extensive high resolution radiation hybrid maps of the human gene, while polymorphic markers have been used to construct genetic maps (see Chapter 7). The dbSTS database maintains complete records for over 133,202 STS markers, including ~18,000 STR markers and gives key information for each record such as primer sequences, map location and marker aliases. Searching dbSTS can be achieved in many ways. The UniSTS interface allows direct searches by keyword, the NCBI Map View application allows searching by genomic location or locus, while dbSTS is also available for BLAST searching by NCBI BLAST. This array of search options makes the dbSTS database a very reliable source for retrieval of both genetic and physical STS map markers.

3.6.2 The Genome Database (GDB)

The genome database (GDB) was established ahead of most other genetics databases in 1990 as a central repository for mapping information from the human genome project. Throughout the early 1990s GDB was the dominant genome database and served as the primary repository for genetic map-related information. In January 1998, after several years of uncertain US government funding, GDB funding was officially terminated. By December 1998 funding from another source was found, but at a significantly lower level. By this time other databases had inevitably overtaken GDB as 'central genome databases' (Cuticchia, 2000). Today GDB is still one of the most comprehensive sources for some forms of genetic data, including tandem repeat polymorphisms (it contains over 18,000), it also contains an eclectic range of information on fragile sites, deletions, disease genes and mutations, collected by a mixture of curation and direct submission. GDB development is ongoing and the historical focus of the database on genetic maps is broadening to a more integrated view of the genome ultimately down to the sequence level (which unfortunately is currently lacking). Plans to finally integrate a sequence map might well make GDB a prominent genetic resource again, although political issues still threaten to halt these aspirations (Bonetta, 2001).

The GDB graphical search interface was a truly pioneering tool of the field and was the first to introduce the kind of graphical map viewing applications that Ensembl and UCSC now excel at. Unfortunately the originals are not always the best and the graphical GDB interface is now starting to look very tired indeed. However, GDB also has a more productive text/table based search interface. This allows complex queries, for example it is possible to retrieve all known polymorphic or non-polymorphic markers between two markers. Advanced filters can also be used, for example markers above a defined level of heterozygosity can be retrieved. Results are retrieved and ordered based on the genetic distances of the markers, along with a very roughly estimated Mb location. As the markers are ordered by genetic distance, many markers cannot be resolved beyond a certain level, therefore markers with identical genetic distances are presented in an arbitrary order. However, high level order is quite reliable and supported by LOD scores. Clarification of genetic marker order and distance is a complex process, which involves integrating multiple maps ultimately down to the level of the human genome to build up a consensus order and distance. These issues of map and marker integration will be examined in detail in Chapter 7.

3.7 NON-NUCLEAR AND SOMATIC MUTATION DATABASES

3.7.1 MITOMAP

The sequencing of the human mitochondrial genome (mtDNA) was a landmark in genomics, being the first component of the human genome to be completely sequenced (Anderson *et al.*, 1981). The mitochondrial genome consists of a 16,569-bp closed circular molecule in the mitochondrion—each of the several thousand mtDNAs per cell encodes a control region encompassing a replication origin and the promoters, a large (16S) and small (12S) rRNA, 22 tRNAs, and 13 polypeptides. All of the mtDNA polypeptides are components of the mitochondrial energy generating pathway, oxidative phosphorylation, which is functionally essential and evolutionarily constrained (Wallace *et al.*, 1995). Despite this selection pressure, maternally inherited mtDNA has a very high mutation rate—mtDNA mutates 10–20 times faster than nuclear DNA as a result of inadequate proofreading by mitochondrial DNA polymerases and limited mtDNA repair capability. As

a result mtDNA mutations might be expected to be relatively common — this is supported by the relative abundance of mitochondrial disorders described to so far — although it is also important to note that such mutations, being comparatively easy to identify by sequencing, are likely to have been among the first to be characterized.

More than 100 mitochondrial diseases have now been described, including a broad spectrum of degenerative diseases involving the central nervous system, heart, muscle, endocrine system, kidney and liver. Information on the phenotypes and causative mutations of these diseases are covered briefly in OMIM and in detail in the mitochondrial mutation database, MITOMAP (Kogelnik *et al.*, 1998). The MITOMAP database (Table 3.1) integrates information on all known mtDNA mutations and polymorphisms with the broad spectrum of available molecular, genetic, functional and clinical data, into an integrated resource which can be queried from a variety of different perspectives.

MITOMAP places the clinical mutation dataset of over 150 disease-associated mutations into their genomic context. It also encompasses information on over 100 mtDNA rearrangements, including nucleotide positions of breakpoint junctions and sequences of associated repeat elements. Clinical characteristics are associated with the mutations and are accessible both through associated datasets in MITOMAP as well as through linkage to OMIM. MITOMAP also provides information on nuclear genes which impinge on mtDNA structure and function. Finally, a population variation dataset provides access to known mtDNA haplotypes and their continental distributions and population frequencies.

3.7.1.1 Searching MITOMAP

MITOMAP is searchable by gene, disease and enzyme — users can refine their search by function, polymorphism, or references (author, title, journal, year or keyword). MITOMAP data has been collated from published literature on the mitochondrial genome and regular searches are made to capture new publications. The database also accepts direct submissions, including over 199 unpublished polymorphisms and mutations.

3.7.2 The Mitelman Chromosome Abberations Map

Cytogenetic studies over the past few decades have revealed clonal chromosomal aberrations in over 100,000 human neoplasms. Many of these have been characterized at the molecular level, revealing previously unknown genes that may be closely associated with tumourigenesis. Information on chromosome changes in neoplasia has grown rapidly, making it difficult to identify all recurrent chromosomal aberrations. The Mitelman Map of Chromosome Aberrations in Cancer (Mitelman *et al.*, 1997) was first published over 15 years ago to compile this information; the database now contains over 7100 references encompassing some 100,000 aberrations in 97 different histological types of cancers. The catalogue has evolved from a book to a CD-ROM published by John Wiley and now it is also available as a web-based database (<http://cgap.nci.nih.gov/Chromosomes/Mitelman>; Mitelman *et al.*, 2002).

The Mitelman database actually consists of three databases. A generalized search form, allows one to search by abnormality, breakpoint, number of clones, number of chromosomes, sex, age, race, country, series, hereditary disorder, topography, immuno-phenotype, morphology, tissue, previous tumour, treatment, reference and/or cytogenetic characteristics to determine frequencies of balanced and unbalanced translocations. The results of a search provide a variety of information. For example, if you select a breakpoint and a gene, the search retrieves relevant PubMed references, diagnoses, the specific chromosome aberration and all genes involved. The Mitelman map is an extremely complex

and detailed database so it is well worth consulting the ‘Help’ section for specific instructions before commencing a search. A more immediately accessible breakdown of the recurrent neoplasia-associated aberrations described by Mitelman are presented by the NCBI MapView tool. This data is an updated version of the survey appearing in the April 1997 Special Issue of *Nature Genetics* (Mitelman *et al.*, 1997). To view the Mitelman aberrations across chromosome 22, for example, try the following URL: <http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/maps.cgi?ORG=hum&MAPS=ideogr,mit&CHR=22>

For cancer geneticists, the Mitelman database benefits greatly from inclusion in the Cancer Genome Anatomy Project (CGAP). CGAP and NCBI are also collaborating closely which has allowed information on chromosomal aberrations to be closely linked with the other CGAP and NCBI resources including mapped SNPs, FISH mapped BACs, and GeneMAP99. The CGAP catalogue is of particular value, serving as a comprehensive index to breakpoints, clones (BACs, cDNA), genes (expression, sequence, tissue), libraries and SNPs (primer pairs, linkage and physical maps). The Mitelman database is undoubtedly the most comprehensive listing of clinical cytogenetic studies in existence, integration of this data with MapViewer and soon hopefully with other viewers such as Ensembl, creates a great opportunity to study the genetics and the biological process of chromosomal aberration right down to the sequence level; this should in turn help to provide insight into the molecular mechanisms of tumourigenesis.

3.8 TOOLS FOR SNP AND MUTATION VISUALIZATION – THE GENOMIC CONTEXT

The human genome is the ultimate framework for organization of SNP and mutation data and so genome viewers are also one of the best tools for searching and visualizing polymorphisms. The three main human genome viewers, Ensembl, the UCSC Human Genome Browser (UCSC-HGB) and the NCBI Map Viewer (Table 3.1), all maintain variable levels of SNP annotation on the human genome, although none maintain annotation of mutation data. Most of the information in these viewers overlap, but each contains some different information and interpretation and so it usually pays to consult at least two viewers, if only for a second opinion. Consultation between viewers is easy as all three now use the same whole genome contig, known as ‘the golden path’ and so they link directly between viewers to the same golden path coordinates.

User defined queries with these tools can be based on many variables, STS, markers, DNA accessions, gene symbol, cytoband or golden path coordinate. This places SNPs and mutations into their full genomic context, giving very detailed information on nearby genes, transcripts and promoters. Ensembl and UCSC-HGB both show conservation between human and mouse genomes, UCSC-HGB also includes tetradon and fugu (fish) genome conservation. This may be particularly useful for identification of SNPs in potential functional regions, as genome conservation is generally restricted to genes (including undetected genes) and regulatory regions (Aparicio *et al.*, 1995). We examine the use of these tools in detail in Chapters 5, 9 and 12.

3.9 TOOLS FOR SNP AND MUTATION VISUALIZATION – THE GENE CONTEXT

For the biologist or candidate gene hunting geneticist, SNP information may be of most interest when located in genes or gene regions, where implicitly each SNP can be evaluated

for potential impact on gene function or regulation. Many tools are available to identify and analyse such SNPs and almost all are based on the dbSNP dataset, but most have somewhat different approaches to the presentation of data (see Table 3.1 for a list of these tools). Choice of tool may be a matter of personal preference so it is probably worth taking a look at a few. The drawback of using some of these tools is that some are maintained by very small groups so sometimes tools may not be comprehensive or current. New tools are constantly appearing in this area so it is often worth running a web search to look for new and novel contributions to this research area—for example ‘SNP AND gene AND database’ is all you need to enter as a search term in a general web search engine.

3.9.1 LocusLink

The NCBI LocusLink database is a reliable tool for gene-orientated searching of dbSNP. It can be queried by gene name or symbol, query results will show a purple ‘V’ link if SNP records have been mapped to a gene. Clicking on this link will take you to a report detailing all RefSNP records mapped across the gene. Almost all NCBI tools integrate directly with dbSNP; LocusLink is the central NCBI ‘gene view’ which links out to a wide range of resources, it also includes a RefSNP gene summary (a purple V or VAR link). This summary details all SNPs across the entire gene locus including upstream regions, exons, introns and downstream regions. Non-synonymous SNPs are identified and the amino acid change is recorded, analysis even accommodates splice variants. LocusLink has the advantage of the NCBI support so it is probably one of the most comprehensive and reliable data sources for gene-orientated SNP information.

Although LocusLink benefits from the reliability bestowed by the infrastructure and resources available at the NCBI, several other tools present gene-focused data with a subtly different approach. Some of these are worth trying, again the tool for you may be a matter of personal preference so try a few. There are many tools which fit into this category, some of these are listed in Table 3.1, but for the purposes of this chapter we will only review two of the more outstanding tools: SNPper and CGAP-GAI.

3.9.2 SNPper

SNPper is a web-based tool developed by the Children’s Hospital Informatics Program (CHIP), Boston (Riva and Kohane, 2001). The SNPper tool maps dbSNP RefSNPs to known genes, allowing SNP searching by name (e.g. using the dbSNP ‘rs’ name), or by the golden path position on the chromosome. Alternatively, you can first find one or more genes you are interested in and find all the SNPs that map across the gene locus, including flanking regions, exons and introns. SNPper produces a very effective gene report (Figure 3.5) which displays SNP positions, alleles and the genomic sequence surrounding the SNP. It also presents very useful text reports which mark up SNPs across the entire genomic sequence of the gene and another report which marks up all the amino acid-altering SNPs on the protein.

The great strength of SNPper lies in its data export and manipulation features. At the SNP report level, SNPs can be sent directly to automatic primer design through a Primer3 interface. At a whole gene level or even at a locus level, SNP sets can be defined and refined and e-mailed to the user in an excel spreadsheet with SNP names in the first column and flanking sequences in the second, ready for primer design.

SNPper currently contains information on around 1,900,000 SNPs and 12,479 genes (January 2002). These correspond to all the unambiguously mapped known SNPs and

SNPper

Gene: IFNAR2							
Name:	interferon (alpha, beta and omega) receptor 2			XmolXport			
Sequence:	Ensta - Annotated - Protein			Strand:	+		
Transcript Position:	chr21:31460142-31492817			Length:	32676		
Coding Sequence Position:	chr21:31471990-31492784			Length:	20795		
Look up this gene in:							
Genbank (mRNA):	NM_000874	Genbank (prot):	NP_000865	Entrez:	IFNAR2	LocusLink:	3455
PubMed:	IFNAR2	OMIM:	602376	Unigene:	IFNAR2	Ensembl:	IFNAR2

Exons:				
#	Start	End	Length	
1	31460142	(0)	31460284 (142)	143
2	31471907	(11765)	31472045 (11903)	139
3	31473740	(13598)	31473782 (13640)	43
4	31475018	(14876)	31475142 (15000)	125
5	31476785	(16643)	31476958 (16816)	174
6	31478776	(18634)	31478922 (18780)	147
7	31482729	(22587)	31482898 (22756)	170
8	31490664	(30522)	31490795 (30653)	132
9	31492628	(32486)	31492817 (32675)	190
XmolXport			Total:	1263

Known SNPs:

SNPset: SS397

Source: [IFNAR2](#)

Created on: **01/17/2002 08:38:53**

SNPs: **29** (avg dist: 1170)

Spacing: **0**

Commands: [Save this SNPset](#)
[Refine this SNPset](#)
[Email this SNPset to yourself](#)
[XmolXport](#)
[SNP graph](#)
[Get flanking sequences](#)

Name	Position	Genepos	Role
rs1476415	chr21:31456148	-15842	A/C Promoter
rs2843981	chr21:31458136	-13854	A/T Promoter
rs2248202	chr21:31461643	-10347	A/C Intron
rs2300370	chr21:31462320	-9670	A/G Intron
rs2248412	chr21:31463294	-8696	A/G Intron
rs2248420	chr21:31463541	-8449	C/T Intron
rs1051393	chr21:31472018	28	G/T Exon, Coding sequence
rs2834156	chr21:31473920	1930	C/T Intron
rs2834157	chr21:31474308	2318	A/G Intron
rs2236756	chr21:31474686	2696	A/C Intron
rs2834158	chr21:31474976	2986	C/T Intron, Exon/intron boundary

Refine SNPset

SNPset: **SS397** Total number of SNPs: 29

Size: 93932 Average distance: 1170

Resolution: 0 Visible SNPs: 29

Restrict to: TSC SNPs
 Validated SNPs
 Promoter 3' UTR
 Exons Coding sequences
 Introns Exon/intron boundary

New resolution:

Figure 3.5 The SNPper gene report. The report displays SNP positions, alleles and the genomic sequence surrounding the SNP. It also presents text reports which mark up SNPs across the entire genomic sequence of the gene and amino acid-altering SNPs on the protein.

genes in the human genome. By restricting the database to known genes, they have considerably simplified their task as all the gene annotation is well defined. SNPper uses this advantage to maximum effect by presenting the data very clearly and informatively. SNPper is a highly recommended tool for the laboratory-based geneticist.

3.9.3 CGAP-GAI (<http://lpgws.nci.nih.gov/>)

The Cancer Genome Annotation Project (CGAP)/Genetic Annotation Initiative (GAI) database is a valuable resource which identifies SNPs by *in silico* prediction from alignments of expressed sequence tags (ESTs) (Riggins and Strausberg, 2001). The database was established specifically to mine SNPs from ESTs generated by CGAP's Tumour Gene Index project (Strausberg *et al.*, 2000), which is generating more than 10,000 ESTs per week from over 200 tumour cDNA libraries. The analysis also encompasses other public EST sources.

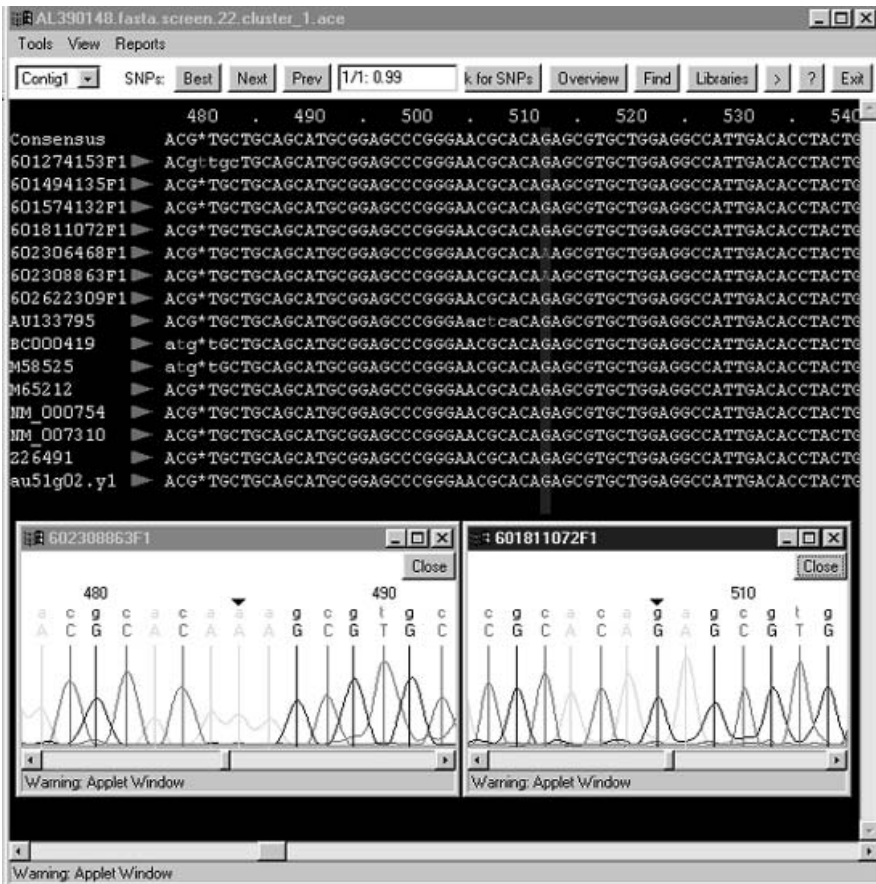


Figure 3.6 The CGAP-GAI web interface for identification of candidate SNPs in ESTs. The JAVA view of trace data helps to support the base call of a potential SNP in an EST, although laboratory investigation is the only reliable SNP confirmation.

Candidate SNPs in ESTs can easily be viewed with the CGAP-GAI web interface in a graphical JAVA assembly (Figure 3.6). SNPs in ESTs are identified by an automated SNP-calling algorithm, mining EST data with greater than 10 reads from the same transcribed region yielded predicted SNPs with an 82% confirmation rate (Riggins and Strausberg, 2001). All SNPs which meet the stringent calling criteria are submitted to dbSNP. It is also worthwhile searching CGAP directly if you are interested in a specific gene. The threshold for automated SNP detection is set very high, so many potential SNPs evade automatic detection, but these candidate SNPs can be identified quite easily by eye, simply by looking for single base conflicts where sequence is otherwise high quality. The JAVA view of trace data helps to support the base call of a potential SNP in an EST (Figure 3.6), although laboratory investigation is the only completely reliable SNP confirmation. Intriguingly this resource could potentially contain some somatic mutations from tumour ESTs which would probably be discarded by the automatic detection algorithm which requires some degree of redundancy to call the SNP.

3.10 CONCLUSIONS

The last few years have revolutionized our knowledge of polymorphism and mutation in the human genome. SNP discovery efforts and processing of genome sequencing data have yielded several million base positions and several hundred thousand VNTRs that might be polymorphic in the genome. This information is complemented by a more select collection of mutation data painstakingly accumulated over many years of disease-gene hunting and mutation analysis. The sheer scale of this data offers tremendous opportunities for genetics and biology. We are now entering a new phase in genetics where we can begin to design experiments to capture the full genetic diversity of populations. This may herald a revolution in genetics allowing rapid association of genes with diseases, alternatively it may simply identify further downstream bottlenecks in the progression to validated disease genes. The literature is already replete with reports of genetic associations and still more failures to replicate associations, but progressions from associated marker to validated disease gene are rare indeed. This may be the real challenge for genetics—to cast new insight into the structure and function of genes, proteins and regulatory regions. To achieve this we will need to integrate diverse sources of data to build up complete pictures of biological systems and their interactions with disease. Again an understanding of mutation and polymorphism may be an important aid in this process—with mutations representing the extreme boundaries beyond which genes begin to dysfunction and polymorphisms perhaps representing the functional range within which genes can operate. Our knowledge of the breadth and variety of human genetic variation can only increase our understanding of the mechanisms of disease and more importantly it may help us to define targets for intervention.

REFERENCES

- Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, *et al.* (2000). A SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**: 513–516.
- Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, *et al.* (1981). Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465.
- Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, *et al.* (1995). Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc Natl Acad Sci USA* **92**: 1684–1688.
- Bahar AY, Taylor PJ, Andrews L, Proos A, Burnett L, Tucker K, *et al.* (2001). The frequency of founder mutations in the BRCA1, BRCA2, and APC genes in Australian Ashkenazi Jews: implications for the generality of U.S. population data. *Cancer* **92**: 440–445.
- Benson G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.
- Bonetta L. (2001). Sackings leave gene database floundering. *Nature* **414**: 384.
- Brookes AJ, Lehvälaiho H, Siegfried M, Boehm JG, Yuan YP, Sarkar CM, *et al.* (2000). HGBASE: A database of SNPs and other variations in and around human genes. *Nucleic Acids Res* **28**: 356–360.
- Cambien F, Poirier O, Lecerf L, Evans A, Cambou J-P, Arveiler D, *et al.* (1992). Deletion polymorphism in the gene for angiotensin-converting enzyme is a potent risk factor for myocardial infarction. *Nature* **359**: 641–644.

- Collins FS, Brooks LD, Chakravarti A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* **8**: 1229–1231.
- Cooper DN, Youssoufian H. (1988). The CpG dinucleotide and human genetic disease. *Hum Genet* **78**: 151–155.
- Cuticchia AJ. (2000). Future vision of the GDB Human Genome Database. *Hum Mut* **15**: 62–67.
- Darvasi A, Kerem B. (1995). Deletion and insertion mutations in short tandem repeats in the coding regions of human genes. *Eur J Hum Genet* **3**: 14–20.
- Debrauwere H, Gendrel CG, Lechat S, Dutreix M. (1997). Differences and similarities between various tandem repeat sequences: minisatellites and microsatellites. *Biochimie* **79**: 577–586.
- Deininger PL, Batzer MA. (1999). Alu repeats and human disease. *Mol Genet Metab* **67**: 183–193.
- Emanuel BS, Shaikh TH. (2001). Segmental duplications: an ‘expanding’ role in genomic instability and disease. *Nature Rev Genet* **2**: 791–800.
- Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, Brookes AJ. (2002). HGVBbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res* **30**: 387–391.
- Goldstein DB, Schlotterer C. (Eds) (1999). *Microsatellites — Evolution and Applications*. Oxford University Press: Oxford, UK.
- Gong W, Emanuel BS, Collins J, Kim DH, Wang Z, Chen F, *et al.* (1996). A transcription map of the DiGeorge and velo-cardio-facial syndrome minimal critical region on 22q11. *Hum Mol Genet* **5**: 789–800.
- Gratacos M, Nadal M, Martin-Santos R, Pujana MA, Gago J, Peral B, *et al.* (2001). A polymorphic genomic duplication on human chromosome 15 is a susceptibility factor for panic and phobic disorders. *Cell* **106**: 367–379.
- Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. (2000). Online Mendelian Inheritance in Man (OMIM). *Hum Mut* **15**: 57–61.
- International SNP Map Working Group (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.
- Kawabata T, Ota M, Nishikawa K. (1999). The protein mutant database. *Nucleic Acids Res* **27**: 355–357.
- Kimura M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press: Cambridge, UK.
- Kluijtmans LA, van den Heuvel LP, Boers GH, Frosst P, Stevens EM, van Oost BA, *et al.* (1996). Molecular genetic analysis in mild hyperhomocysteinemia: a common mutation in the methylenetetrahydrofolate reductase gene is a genetic risk factor for cardiovascular disease. *Am J Hum Genet* **58**: 35–41.
- Kogelnik AM, Lott MT, Brown MD, Navathe SB, Wallace DC. (1998). MITOMAP: a human mitochondrial genome database—1998 update. *Nucleic Acids Res* **26**: 112–115.
- Krebs S, Seichter D, Forster M. (2001). Genotyping of dinucleotide tandem repeats by MALDI mass spectrometry of ribozyme-cleaved RNA transcripts. *Nature Biotechnol* **19**: 877–880.
- Kubota S. (2001). The extinction program for *Homo sapiens* and cloning humans: trinucleotide expansion as a one-way track to extinction. *Med Hypotheses* **56**: 296–301.
- Le Stunff C, Fallin D, Schork NJ, Bougneres P. (2000). The insulin gene VNTR is associated with fasting insulin levels and development of juvenile obesity. *Nature Genet* **26**: 444–446.

- Lucassen AM, Julier C, Beressi JP, Boitard C, Froguel P, Lathrop M, *et al.* (1993). Susceptibility to insulin dependent diabetes mellitus maps to a 4.1-kb segment of DNA spanning the insulin gene and associated VNTR. *Nature Genet* **4**: 305–310.
- Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, *et al.* (1999). A general approach to single-nucleotide polymorphism discovery. *Nature Genet* **23**: 452–456.
- Marth G, Yeh R, Minton M, Donaldson R, Li Q, Duan S, *et al.* (2001). Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genet* **27**: 371–372.
- Matsuura T, Yamagata T, Burgess DL, Rasmussen A, Grewal RP, Watase K, *et al.* (2000). Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nature Genet* **26**: 191–194.
- Miller RD, Kwok PY. (2001). The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine. *Hum Mol Genet* **10**: 2195–2198.
- Miller RD, Taillon-Miller P, Kwok PY. (2001). Regions of low single-nucleotide polymorphism incidence in human and orang-utan xq: deserts and recent coalescences. *Genomics* **71**: 78–88.
- Mitelman F, Mertens F, Johansson B. (1997). A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genet* **15**: 417–474.
- Mitelman F, Johansson B, Mertens F (Eds) (2002). Mitelman Database of Chromosome Aberrations in Cancer <http://cgap.nci.nih.gov/Chromosomes/Mitelman>.
- Mullikin JC, Hunt SE, Cole CG, Mortimore BJ, Rice CM, Burton J, *et al.* (2000). An SNP map of human chromosome 22. *Nature* **407**: 516–520.
- Olson M, Hood L, Cantor C, Botstein D. (1989). A common language for physical mapping of the human genome. *Science* **245**: 1434–1435.
- Reich D, Cargill M, Bolk S, Ireland J, Sabeti P, Richter D, *et al.* (2001). Linkage disequilibrium in the human genome. *Nature* **411**: 199–204.
- Riggins GJ, Strausberg RL. (2001). Genome and genetic resources from the Cancer Genome Anatomy Project. *Hum Mol Genet* **10**: 663–667.
- Risch N. (2000). Searching for genetic determinants in the new millennium. *Nature* **405**: 847–856.
- Riva AA, Kohane IS. (2001). A web-based tool to retrieve human genome polymorphisms from public databases. *Proc AMIA Symp* 558–562.
- Roque M, Godoy CP, Castellanos M, Pusiol E, Mayorga LS. (2001). Population screening of F508del (DeltaF508), the most frequent mutation in the CFTR gene associated with cystic fibrosis in Argentina. *Hum Mut* **18**: 167.
- Schmucker B, Krawczak M. (1997). Meiotic microdeletion breakpoints in the BRCA1 gene are significantly associated with symmetric DNA sequence elements. *Am J Hum Genet* **61**: 1454–1456.
- Shaikh TH, Kurahashi H, Emanuel BS. (2001). Evolutionarily conserved duplications in 22q11 mediate deletions, duplications, translocations and genomic instability. *Genet Med* **3**: 6–13.
- Shapira SK. (1998). An update on chromosome deletion and microdeletion syndromes. *Curr Opin Pediatr* **10**: 622–627.
- Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, *et al.* (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308–311.
- Stein LD. (2001). Using Perl to facilitate biological analysis. *Methods Biochem Anal* **43**: 413–449.
- Strausberg RL, Buetow KH, Emmert-Buck MR, Klausner RD. (2000). The cancer genome anatomy project: building an annotated gene index. *Trends Genet* **16**: 103–106.

- Thomas PK. (1999). Overview of Charcot-Marie-Tooth disease type 1A. *Ann NY Acad Sci* **883**: 1–5.
- Usdin K, Grabczyk E. (2000). DNA repeat expansions and human disease. *Cell Mol Life Sci* **57**: 914–931.
- Wallace DC, Shoffner JM, Trounce I, Brown MD, Ballinger SW, Corral-Debrinski M, *et al.* (1995). Mitochondrial DNA mutations in human degenerative diseases and aging. *Biochim Biophys Acta* **1272**: 141–151.