Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

SECTION 1

AN INTRODUCTION TO BIOINFORMATICS FOR THE GENETICIST

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 1

Introduction: The Role of Genetic Bioinformatics

MICHAEL R. BARNES[1] and IAN C. GRAY[2]

[1]*Genetic Bioinformatics and* [2]*Discovery Genetics*
Genetics Research Division
GlaxoSmithKline Pharmaceuticals, Harlow, Essex, UK

1.1 INTRODUCTION

In February 2000, scientists announced the draft completion of the human genome. If media reports were accepted at face value, then it might be reasonable to predict that most geneticists would be unemployed within a decade of this announcement and human disease would become a distant memory. As we all know this is very far from the truth, the human genome is many things but it is not in itself a panacea for all human ailments, nor is it a revelation akin to the elucidation of the DNA double helix or the theory of evolution. The human genome is simply a resource borne out of technical prowess, perhaps with a little human inspiration. One thing that is certain is that we do not yet understand the functional significance of the majority of our genome, but what we do know is finally put into context. Over the past 200 years mankind has developed an

ever increasing understanding of genetics; Darwin and Mendel provided the 19th century theories of evolution and inheritance, while Bateson, Morgan and others established a framework for the mechanisms of genetics at the beginning of the 20th century. The tentative identification of DNA as the genetic material by Avery and colleagues in the 1940s preceded the elucidation of the structure of the DNA molecule in 1953 by Watson and Crick, which in turn provided a mechanism for DNA replication and ushered in the era of modern molecular genetics. In 2003, precisely 50 years after this landmark discovery it is anticipated that the entire human genome sequence will be completed in a final, polished form; a fully indexed but currently only semi-intelligible 'book of life'. Here lies the most overlooked property of the genome — its value as a framework for data integration, a central index for biology and genetics. Almost any form of biological data can be mapped to a genomic region based on the genes or regulatory elements that mediate it. So the sequencing of the human genome means new order for biology. This order is perhaps comparable to the order the periodic table brought to chemistry in the 19th century. Where elements were placed in an ordered chemical landscape, biological elements will be grouped and ordered on the new landscape of the human genome. This presents excellent opportunities to draw together very diverse biological data; only then will the 'book of life' begin to make sense.

The human genome and peripheral data associated with and generated as a result of it require increasingly sophisticated data storage, retrieval and handling systems. With the promises and challenges that lie ahead, bioinformatics can no longer be the exclusive realm of the Unix guru or the Perl hacker and in recent years web browsers have made tools accessible and user friendly to the average biologist or geneticist. Bioinformatics is now both custodian and gatekeeper of the new genome data and with it most other biological data. This makes bioinformatics expertise a prerequisite for the effective geneticist. This expertise is no mystery; modern bioinformatics tools coupled with an inquiring mind and a willingness to experiment (key requirements for any scientist, bioinformatician or not) can yield confidence and competence in bioinformatic data handling in a very short space of time. The objective of this book is not to act as an exhaustive guide to bioinformatics, other texts are available to fulfil this role, but instead is intended as a specialist guide to help the typical geneticist navigate the internet jungle to some of the best tools and databases for the job, that is, associating genes, polymorphisms and mutations with diseases and genetic traits. In this chapter we give a flavour of the many processes in modern genetics where bioinformatics has a major impact and refer to subsequent chapters for greater detail.

At the risk of over simplifying a very complex issue, the process of understanding genetic disease typically proceeds through three stages. First, recognition of the disease state or syndrome including an assessment of its hereditary character; second, discovery and mapping of the related polymorphism(s) or mutation(s) and third, elucidation of the biochemical/biophysical mechanism leading to the disease phenotype. Each of these stages proceeds with a variable degree of laboratory investigation and bioinformatics. Both activities are complementary, bioinformatics without laboratory work is a sterile activity as much as laboratory work without bioinformatics can be a futile and inefficient one. In fact these two sciences are really one, genetics and genomics generate data and computational systems allow efficient storage, access and analysis of the data — together, they constitute bioinformatics. Almost every laboratory process has a complementary bioinformatics process, Table 1.1 lists a few of these — building on these basic applications will maximize the effect of bioinformatics on workflow efficiency.

TABLE 1.1 Examples of Bioinformatics Applications in Genetics Research

Data	Related Laboratory Techniques	Associated Bioinformatics Applications
Human genome sequence	DNA sequencing PCR Novel gene identification by expression analysis *In vitro* characterization of regulatory elements	Gene and regulatory region prediction BLAST homology searching Electronic PCR PCR primer design Electronic translation and protein secondary structure prediction *In silico* design of expression constructs
Genetic markers	Genotyping	Identification of optimal marker sets Genotyping assay design QC checking and statistical analysis of genotype data
Model organism genome sequence	Comparative genetics (e.g. linkage) and genomics (e.g. transgenics and gene knock-outs)	Linkage analysis of models of human diseases Comparative genetic and physical maps for cross-species analysis of linkage regions Functional assessment of gene regulatory regions by cross-species comparison *In silico* drafting of gene knock-out and transgenic constructs
Expression RNA and protein	Microarrays Serial analysis of gene expression (SAGE) Proteomics	Gene regulatory analysis Tumour and other disease tissue classification Elucidation of gene–gene interactions and disease pathway expansion
Three-dimensional protein structure	Crystallography/NMR	Prediction and visualization of molecular structures related to disease and mutation

1.2 GENETICS IN THE POST-GENOME ERA — THE ROLE OF BIOINFORMATICS

In the role of genome data custodian and gatekeeper, bioinformatics is an integral part of almost every field of biology, including of course, genetics. In the broadest sense it covers the following main aspects of biological research:

- Knowledge management and expansion
- Data management and mining
- Study design and support
- Data analysis
- Determination of function

These categories are quite generic and could apply to any field of biology, but are clearly applicable to genetics. Both genetics and bioinformatics are essentially concerned with asking the right questions, generating and testing hypotheses and organizing and interpreting large amounts of data to detect biological patterns.

1.3 KNOWLEDGE MANAGEMENT AND EXPANSION

Few areas of biological research call for a broader background in biology than the modern approach to genetics. This background is tested to the extreme in the selection of candidate genes to test for involvement with a disease process, where genes need to be chosen and prioritized based on many criteria. Often biological links may be very subtle, for example a candidate gene may regulate a gene which regulates a gene that in turn may act upon the target disease pathway. Faced with the complexity of relationships between genes, geneticists need to be able to expand pathways and identify complex cross talk between pathways. As this process can extend almost interminably to a point where virtually every gene is a candidate for every disease, knowledge management is important to help to weigh up evidence to prioritize genes. The geneticist may not be an authority in the disease area under study, and in today's climate of reductionist biology an expert with a global picture of the disease process at the molecular level may be hard to find. Therefore effective tools are needed to quickly evaluate the role of each candidate and its related pathways with respect to the target phenotype.

Literature is the most powerful resource to support this process, but it is also the most complex and confounding data source to search. To expedite this process, some databases have been constructed which attempt to encapsulate the available literature, e.g. On-line Mendelian Inheritance in Man (OMIM). These centralized data resources can often be very helpful for gaining a quick overview of an unfamiliar pathway or gene, but inevitably one needs to re-enter the literature to build up a fuller picture and to answer the questions that are most relevant to the target phenotype or gene. The internet is also an excellent resource to help in this process. In Chapter 2, we offer some pointers to help the reader with effective literature searching strategies and give suggestions as to some of the best disease databases and related resources on the internet.

1.4 DATA MANAGEMENT AND MINING

Efficient application of knowledge relies on well organized data and genetics is highly dependent upon good data, often in very large volumes. Accessing available data, particularly in large volumes is often the biggest informatic frustration for geneticists. Here

we focus on aspects of accessing data from public databases; solutions for in-house data collection, either in the form of 'off the shelf' or custom-built laboratory information management systems (LIMS) belong to a specialist area that lies beyond the scope of this book.

Genetic data have grown exponentially over the last few years, fuelled by the expressed sequence tag (EST) cDNA sequence resources generated largely during the 1990s and more recently the increasing genomic sequence data from the human genome and other genome sequencing projects. Genetic database evolution has matched this growth in some areas, with some resources leading the efforts towards whole genome integration of genetic data, particularly the combined human genome sequence, genetic map, EST and SNP databases exemplified by the Golden Path. Curiously, development in many of the older more established genetic resources (for example, GDB and HGMD) has been somewhat stagnant. This may be partly due to the difficulties involved in data integration with the draft genome sequence, which is effectively a moving target as the data are updated on a regular basis. Many of the traditional genetic databases have not seized the opportunity to integrate genetic data with the human genome sequence. The future survival of these databases will certainly depend on this taking place and there is no question that the role of these databases will change. One might question the value of some of the older genetic datasets, for example, why would we need radiation hybrid maps of the human genome, when we have the ultimate physical map — the human genome sequence? These painstakingly collected datasets have already played a critical role in the process of generating the maps that allowed the sequencing of the human genome and they may still have some value as an aid for QC of new data and perhaps more importantly as a point of reference for all the studies that have previously taken place.

A key problem that frequently hinders effective genetic data mining is the localization of data in many independent databases rather than a few centralized repositories. A clear exception to this is SNP data which has now coalesced around a single central database — dbSNP at NCBI (Sherry *et al*., 2001). By contrast human mutation data, which has been collected over many years, is still stored in disparate sources, although moves are afoot to move to a similar central database — Hobbies (Fredman *et al*., 2002). These developments are timely; human mutation and polymorphism data both hold complementary keys to a better understanding of how genes function and malfunction in disease. The availability of a complete human genome presents us with an ideal framework to integrate both sets of data, as our understanding of the mechanisms of complex disease increase, the full genomic context of variation will become increasingly significant.

With the exception of dbSNP most recent database development has not been implicitly designed for geneticists, instead genomic databases and genome viewers have developed to aid the annotation of the human genome. Of course this data is vital for genetics, but this explains why the available tools often appear to lack important functionality. One has to make use of what functionality is available, although sometimes this means using tools in ways that were not originally intended (for example many geneticists use BLAST to identify sequence primer homology in the human genome, but few realize that the default parameters of this tool are entirely unsuited for this task). We will attempt to address these issues throughout this book and offer practical solutions for obtaining the most value from existing tools wherever possible. In Chapter 5 we examine the use of human genome browsers for genetic research. Tools such as Ensembl and the UCSC human genome browser annotate important genetic information on the human genome, including SNPs, some microsatellites and of course, genes and regulatory regions. User-defined queries place genes and genetic variants in their full genomic context, giving very

detailed information on nearby genes, promoters or regions conserved between species, including mouse and fish. It is difficult to overstate the value of this information for genetics. For example, cross-species genome comparison is invaluable for the analysis of function, as inter-species sequence conservation is generally thought to be restricted to a functionally important gene or regulatory regions and so this is one of the most powerful tools for identifying potential regulatory elements or undetected genes (Aparicio *et al*., 1995). Several chapters in this book cover tools and databases to support these approaches (see Chapters 12 and 13).

As technology developments have scaled up the throughput of genotyping to enable studies of tens (and possibly hundreds) of thousands of polymorphisms and provided the capability to generate equally impressive amounts of microarray transcript data to name just two examples, the need for more effective data management has intensified. This reveals the major drawback of the ultra user-friendly 'point and click' interfaces to most genetics and genomics tools — they often do not allow retrieval of bulk datasets; instead data often has to be retrieved on a point by point basis. For many applications this is highly inefficient at best or simply non-viable at worst. One solution to this problem is to query the database directly at a UNIX or SQL level, but this may not be a trivial process for the occasional user with no or limited knowledge of command lines and in many cases it will not be possible to access the data directly in this manner. If the raw data are available, it may be possible to build custom databases, using database tools such as Microsoft ACCESS. However, the authors accept that this is not a straightforward option nor the method of choice of most users and instead this book will focus on web-based methods for data access. Where there is no web-based method to achieve a data mining goal, geneticists should consider contacting the developers of databases to request new functionality, such requests are generally welcomed by database developers, many of whom would be very pleased to know that their tools are being used! Several developers have already improved their methods for bulk data retrieval (probably as a result of requests from users), but interfaces are still lacking in some critical areas for genetics. For example, several tools allow the user to generate a list of SNPs across a locus (e.g. dbSNP, Ensembl and UCSC), but only one allows the user to retrieve the flanking sequence of each SNP in one batch to allow primer design (SNPper — see Chapter 3). We will try to tackle these problems as they arise throughout the book.

1.5 GENETIC STUDY DESIGNS

There are a number of approaches to disease gene hunting and many arguments to support the merits of one approach over another. Whatever the method, comprehensive informatics input at the study design stage can contribute greatly to the quality, efficiency and speed of the study. It can help to define a locus clearly in terms of the genes and markers that it contains and supports a logical and systematic approach to marker and gene selection and subsequent genetic analysis, simultaneously reducing the cost of a project and improving the chances of successfully discovering a phenotype–genotype correlation.

Despite the recent improvements in the throughput of genetic and genomic techniques and the increased availability of gene and marker data, genes which contribute to the most common human diseases are still very elusive. By contrast, the identification of genes mutated in relatively rare single gene disorders (so-called Mendelian or monogenic disorders) is now straightforward if suitable kindreds are available. The identification of the genes responsible for a plethora of monogenic disorders is one of the genetics

success stories of the late 1980s and the 1990s; genes identified include, to name but a few — *CFTR* (cystic fibrosis; Riordan *et al*., 1989), Huntingtin (Huntington's disease; Huntington's Disease Collaborative Research Group, 1993), Frataxin (Friedreich's ataxia; Campuzano *et al*., 1996) and *BRCA1* in breast and ovarian cancer (Miki *et al*., 1994).

Unfortunately, success in the identification of genes with a role in complex (i.e. multigenic) disease has been far less successful. Notable examples are the involvement of *APOE* in late-onset Alzheimer's disease and cardiovascular disease and the role of *NOD2* in Crohn's disease (Hugot *et al*., 2001; Saunders *et al*., 1993). However, genes for most of the common complex diseases remain elusive. Our ability to detect disease genes is often dependent on the analysis method applied. Methods for the identification of disease genes can be divided neatly into two broad categories, linkage and association. Although many common principles apply to both of these study types, each approach has distinct informatics demands.

1.5.1 The Linkage Approach

The vast majority of Mendelian disease genes have been identified by linkage analysis. This involves identifying a correlation between the inheritance pattern of the phenotypic trait (usually a disease state) with that of a genetic marker, or a series of adjacent markers. Because of the relatively low number of recombination events observed in the 2–5 generation families typically used for linkage analyses (around one per Morgan, which is roughly equivalent to 100 megabases, per meiosis), these marker/disease correlations extend over many megabases (Mb), allowing adequate coverage of the entire human genome with a linkage scan of only 300–600 simple tandem repeat (STR) markers giving an average spacing of 10 or 5 cM respectively. STRs are the markers of choice for linkage analysis, due to the fact that they show a high degree of heterozygosity. Markers with a heterozygosity level of $>70\%$ are typically selected for linkage panels (i.e. from 100 individuals selected at random, at least 70 would have two different alleles for a given marker; clearly the higher the heterozygosity the greater the chance of following the inheritance pattern from parent to offspring). Such marker panels are well characterized and can be accessed from several public sources at various densities (see Chapter 7). Just over 16,000 STR markers have been characterized in humans, which represents a small fraction of the estimated total numbers of polymorphic STRs. Analysis of the December 2001 human genome draft sequence suggests that there may be somewhere in the order of 200,000 potentially polymorphic STRs in the human genome (Viknaraja *et al*., unpublished data). Software tools are now available to assist in the sequence-based identification of these potentially polymorphic STR markers across a given locus, should additional markers be required to narrow a linkage region (see Chapter 9 for details).

Clearly the limited degree of recombination that facilitates linkage analysis with sparse marker panels is a double-edged sword; the investigator may be left with several megabases of DNA containing a large number of potential candidate genes. However, combining data from several different families often results in reduction of the genetic interval under study, and the high-throughput sequencing capabilities available in many modern genetics laboratories coupled with complete genome sequence render the systematic screening of a large number of candidate genes a far less daunting task than it was 10 years ago.

Unlike single gene Mendelian diseases, complex genetic diseases are caused by the combined effect of multiple polymorphisms in a number of genes, often coupled with environmental factors. The successes of linkage analysis in the rapid identification of

Mendelian disease genes has spawned large-scale efforts to track down genes involved in the more common complex disease phenotypes. This approach is not restricted to academic research groups; many pharmaceutical and biotechnology companies have joined what many would perceive to be a 'genetic gold-rush', in an attempt to identify new drug targets for common diseases such as asthma, diabetes and schizophrenia, in a manner reminiscent of the rush to mine drug targets from expressed sequence tags (ESTs) in the late 1990s (Debouck and Metcalf, 2000). The application of a linkage approach to complex disease typically involves combining data from a large number of affected sib-pairs. Publicly available software for linkage analysis of sib-pairs is described in detail in Chapter 11.

Unfortunately the identification of genes involved in common diseases using a linkage strategy has been largely unsuccessful to date, mainly because each gene with phenotypic relevance is thought to make a relatively small contribution to disease susceptibility. These small effects are likely to be below the threshold of detection by linkage analysis in the absence of unfeasibly large sample sizes (Risch, 2000). In an attempt to circumvent this problem researchers using linkage approaches to identify genes involved in complex disease typically relax the threshold of acceptable 'log of the odds' (LOD) score (see Chapter 11) from 3, the traditionally accepted threshold of evidence for linkage in monogenic disease to 2, or sometimes even lower (Pericak-Vance *et al.*, 1998). However we would expect to see a number of hits due to chance alone with a comprehensive genome scan at this threshold. The rationale for lowering the threshold for detection of linkage, i.e. the effect of each contributing gene in a complex disease is smaller than would be expected for a monogenic disease, can result in a situation where a true signal is indistinguishable from background noise. In order to distinguish true linkage from false positives, many investigators are now using a combination of both linkage and association, relying on linkage analysis to reveal tentative, broad map positions which are subsequently confirmed and narrowed with an association study (see Chapter 8).

1.5.2 The Association Approach

In its simplest form, the aim of a genetic association study is to compare an allele frequency in a disease population with that in a matched control population. A significant difference may be indicative that the locus under test is in some way related to the disease phenotype. This association could be direct, i.e. the polymorphism being tested may have functional consequences that have a direct bearing on the disease state. Alternatively, the relationship between a genetic marker and phenotype may be indirect, reflecting proximity of the marker under test to a polymorphism predisposing to disease. The phenomenon of co-occurrence of alleles (in this case a disease-conferring allele and a surrogate marker allele) more often than would be expected by chance is termed linkage disequilibrium (LD). Suitable population structures for genetic association studies and statistical methods and software tools for the analysis of data resulting from such studies are discussed in detail in Chapters 8 and 11. Our aim here is to give the reader the briefest of introductions.

Association studies have three main advantages over linkage studies for the analysis of complex disease: (i) case–control cohorts are generally easier to collect than extended pedigrees; (ii) association studies have greater power to detect small genetic effects than linkage studies; a clear example is the insulin gene, which shows extremely strong association with type 2 diabetes, but very weak linkage (Speilman *et al.*, 1993); (iii) LD typically stretches over tens of kilobases rather than several megabases (Reich *et al.*, 2001), allowing focus on much smaller and more manageable loci. Among other reasons (discussed in

Chapter 8), this is because an association-based approach exploits recombination in the context of the entire population, rather than within the local confines of a family structure.

Of course, this last point is the other side of the double-edged sword of marker density and resolution mentioned in the context of linkage analysis above. The trade-off is reduced range over which each marker can detect an effect, resulting in a need for increased marker density. The required marker density for an association-based genome scan is unknown at present as we do not have enough information regarding human genome diversity in terms of polymorphic variability and genome-wide patterns of LD. However, typical guesses are in the range of 30,000–300,000 markers (Collins *et al*., 1999; Kruglyak, 1999); orders of magnitude higher than the numbers required for linkage analysis. The high cost of generating the several million genotypes for such an experiment has prevented any such undertaking at the time of writing, although several proof of concept studies have demonstrated that high-density SNP maps can be efficiently generated using existing technologies and should be achievable in a reasonable time-frame (Antonellis *et al*., 2002; Lai *et al*., 1998). In the meantime, it is likely that research groups will continue to test individual genes for association with disease (the 'candidate gene' approach — see Section 1.7 below).

Once the genomic landscape, in terms of polymorphism and LD, is known with some degree of accuracy, it is highly likely that the number of markers required for a whole genome association study can be reduced by an intelligent study design with heavy reliance on bioinformatics input. Testing all available markers in a given region for association with a disease is expensive, laborious and frequently unnecessary; a simple example to illustrate this would be two adjacent markers which always demonstrate co-segregation; in other words, the genotypic status of one can always be predicted by genotyping the other — there is no point in genotyping both. Although this example is simple in the extreme, as adjacent markers typically show varying degrees of (rather than absolute) co-segregation, there is a trade-off between minimizing the amount of required genotyping whilst minimizing loss of information. Selection of optimal non-redundant marker sets, coupled with an initial focus on gene-rich regions, is the key to providing lower overall genotyping costs whilst retaining high power to detect association. This will require extensive knowledge of the blocks of preserved marker patterns (haplotypes) in the population under study; bioinformatics tools for constructing and analysing haplotypes and selecting optimal marker sets based on haplotypic information are discussed in detail in Chapters 8 and 11.

1.5.3 Markers for Association Studies

STRs were (and still are) the vanguard of linkage analysis, mainly because of their high levels of heterozygosity and hence increased informativeness when compared to an earlier marker system, the restriction fragment length polymorphism (RFLP); the majority of RFLPs are the result of a single nucleotide polymorphism (SNP) which creates or destroys a restriction site. SNPs have made a comeback worthy of Lazarus in recent years and are now the marker of choice for genetic association studies. The main reasons for the return to favour of SNPs are their abundance (an estimated 7 million with a minor allele frequency of greater than 5% in the human genome; Kruglyak and Nickerson, 2001) and binary nature which renders them well suited to automated, high-throughput genotyping. As mentioned above, tens or hundreds of thousands of SNPs will be required for whole genome association scans (even with optimized marker sets). Until very recently, studies on this scale were unfeasible, not only as a result of unacceptably high genotyping costs,

but also due to the lack of available markers. Large-scale SNP discovery projects such as the SNP consortium (TSC; Altshuler *et al*., 2000a) have increased the number of known SNPs dramatically. We now have a great deal of SNP data (3.4 million non-redundant SNPs deposited in dbSNP at the time of going to press), however it is becoming apparent that even this number of markers will be insufficient for comprehensive association studies (note that the figure of 3.4 million includes a considerable number of SNPs with a minor allele frequency of less than 5%, which may be of limited use in association studies; this is discussed in Chapter 8).

We have already touched on the importance and potential impact of defining haplotypes as the basis for identifying optimal marker sets. This method has already been applied in small-scale studies with striking results. For example, in a study of nine genes spanning a total of 135 kb, Johnson *et al*. (2001) found that just 34 SNPs from a total of 122 could be used to define all common haplotypes (those with a frequency of greater than 5%) across the nine genes, an impressive validation of the approach of defining maximally informative minimal marker sets based on haplotypic data. However this study also highlighted the inadequacy of the current public SNP resource; only 10% of the SNPs identified by Johnson *et al*. were found to be present in dbSNP. Using dbSNP data alone, it was impossible to capture comprehensive haplotype data; in fact for four of the nine genes, no SNPs whatsoever were registered in dbSNP. Unfortunately it appears that our current public SNP resource represents the tip of the iceberg in terms of requisite information for the proper implementation of modest candidate gene association studies, let alone whole genome scans. However, given the burgeoning nature of dbSNP, we are optimistic that this situation is transient.

As a footnote to this section, it should be noted that although STRs have been largely swept aside by the wave of SNP euphoria, STRs may still be useful for association studies; indeed, it is possible that LD can be detected over far greater distances with STRs than SNPs under some circumstances, as discussed in Chapter 8.

1.6 PHYSICAL LOCUS ANALYSIS

In recent years, as the human genome sequence has neared completion, practical approaches to physical characterization of a genetic locus have changed quite dramatically. The laborious laboratory-based process of contig construction using yeast and bacterial artificial chromosome (YAC and BAC) clones or cosmids, involving consecutive rounds of library screening, clone characterization and identifying overlaps between clones, has become largely redundant, as has clone screening for the identification of novel polymorphic markers and genes. Today this process, which took many months or even years, can be completed in an afternoon using web-based human genome browsers. This shifts the initial focus of a study from contig construction and characterization to very detailed locus characterization using a range of bioinformatics tools; it is now possible to characterize a locus *in silico* to a very high level of detail before any further laboratory work commences. When the wet work does start, good prior use of bioinformatics will have rendered many procedures superfluous and the study is far more efficient and focused as a result. Figure 1.1 illustrates some of the key stages in the genetic analysis of candidate genes and loci — the role of informatics at each stage of this process is explored in detail in this book and the relevant chapters addressing each issue are indicated.

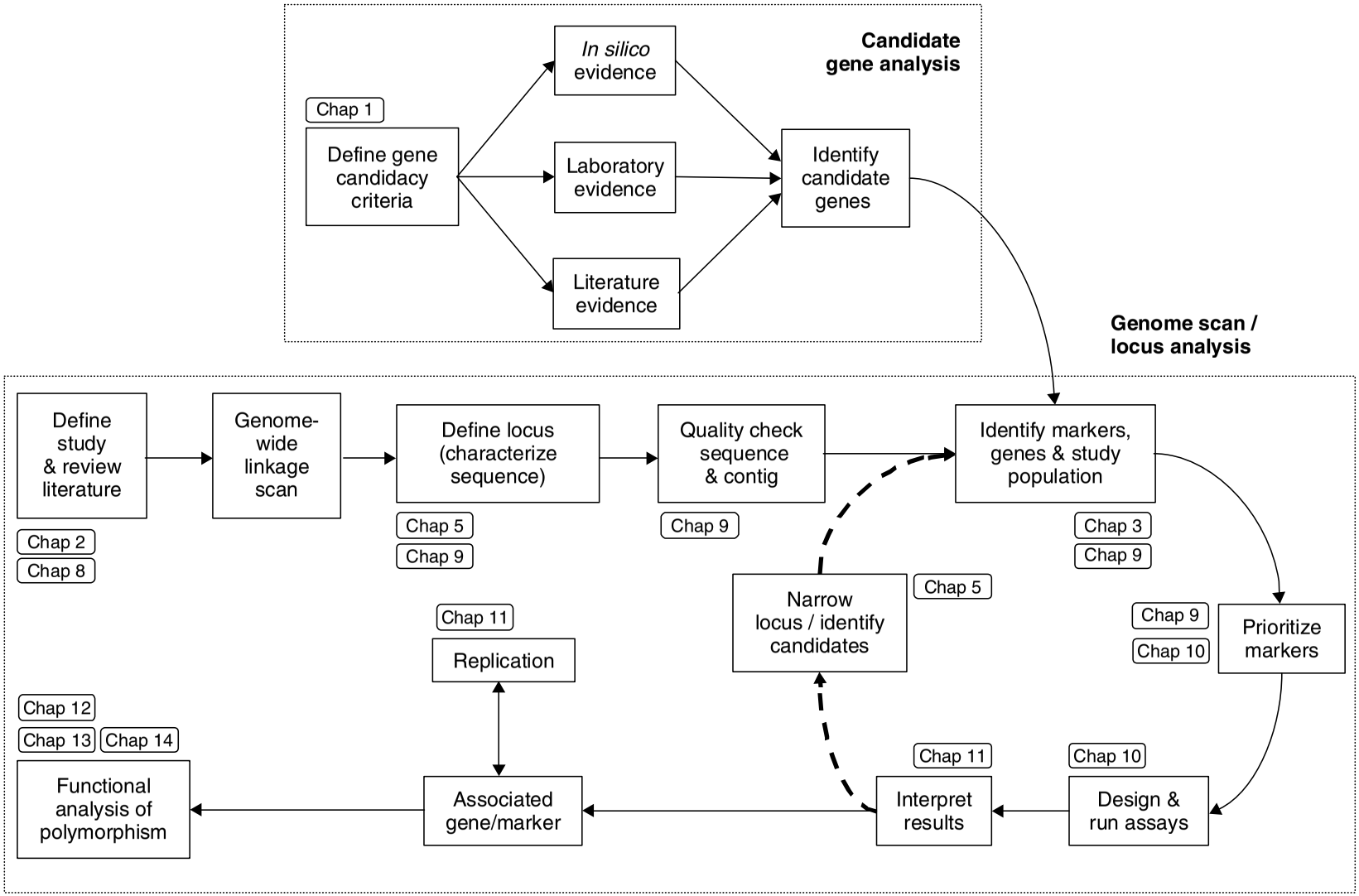

Figure 1.1 The genetic study process for complex disease, both candidate gene approaches and the follow-up of genome-wide linkage scans, highlighting chapters covering informatics aspects of each key step.

1.7 SELECTING CANDIDATE GENES FOR ANALYSIS

Candidate genes are typically selected for testing for association with a disease state on the basis of either (i) biological rationale; the gene encodes a product which the investigator has good reason to believe is involved in the disease process, (ii) the fact that the gene in question is located under a linkage peak, or (iii) both. The biggest problem with candidate gene analysis is that apparently excellent candidates are usually highly abundant and this surfeit of 'good' candidates is often difficult to rationalize.

Bioinformatics can be one of the most effective ways to help shorten, or more correctly prioritize, a candidate list without immediate and intensive laboratory follow-up. Firstly candidate criteria need to be determined based upon the phenotype in question. Detailed searches of the literature may help to flesh out knowledge of the disease and related pathways. Once a set of criteria is defined (for example which tissues are likely to be affected, which pathways are likely to be involved, and what types of genes are likely to mediate the observed phenotype), further literature review will help to 'round up the usual suspects', genes in known pathways with an established role in the phenotype under study. This is probably the most time-consuming step, but some tools can help to expedite this process, for example tools like OMIM can provide concise summaries of a disease area or gene family. Other databases encapsulate knowledge of pathways and regulatory networks, e.g. the *Kyoto Encyclopedia of Genes and Genomes* (KEGG; Kanehisa *et al*., 2002). An alternative or parallel approach at this stage is to use a broader net to identify all genes which *could* be involved in the disease based on relaxed criteria such as tissue expression. Many *in silico* gene expression resources are available, including data derived from EST libraries, serial analysis of gene expression (SAGE; Velculescu *et al*., 1995) data, microarray and RT-PCR data (see Chapter 15). For example, if the disease manifests in the lung, it is possible to distinguish genes that show lung expression from those that do not. This gives an opportunity to reduce emphasis on genes that show expression patterns which conflict with the disease hypothesis. However, it should be noted that electronic expression data is typically not comprehensive and care must be taken in using it to exclude the expression of a gene in a specific tissue. Low-level expression may not be detected by the method used; furthermore, gene expression may show temporal and spatial regulation — a gene may only be expressed during a specific phase of development or under particular conditions, e.g. cellular stress or differentiation.

1.8 PROGRESSING FROM CANDIDATE GENE TO DISEASE-SUSCEPTIBILITY GENE

In recent years, countless associations between genes and disease have been published, however many of these are likely to be spurious. Many reported associations show marginal p-values and subsequent studies often fail to replicate initial findings. Clearly p-values of around 0.05, generally accepted as the cut-off for a significant finding, will occur by chance for every 20 tests performed; this largely explains the general failure to reproduce promising primary results. However, real but very small effects giving marginal p-values are also difficult to replicate, leaving the investigator unsure as to the meaning of a failure to replicate. One approach for resolving the issue is to perform a rigorous meta-analysis using all available data, including both positive and negative associations. This type of analysis was recently used to demonstrate an association between the nuclear hormone receptor PPARγ and diabetes, using data (previously regarded as equivocal)

drawn from a range of publications (Altshuler *et al*., 2000b). Nonetheless, this approach relies on a lack of publication bias, i.e. the improbable assumption of an equal chance of publication for both positive and negative results.

Ultimately the biologist requires functional data to support an hypothetical genetic association; bioinformatics has a role to play here too. For example, DNA variants that alter subsequent amino-acid sequences can be checked for potential functional consequences using software tools (Chapters 12 and 14). Similarly, a thorough bioinformatic characterization of putative regulatory elements can give an indication of the possible impact of polymorphisms on *cis*-acting transcriptional motifs and the consequence on expression levels (Chapter 13). Bioinformatics can also assist in laboratory-based functional assessment of genes and polymorphisms; simple sequence manipulation tools coupled with genome sequence data can be used to design constructs for the *in vitro* and *in vivo* analysis of genes and polymorphisms using expression assays, transgenic mice and a host of other systems. However, perhaps the largest impact from bioinformatics on the field of functional characterization of genes will come from the development of powerful pattern recognition software for the identification of relationships between multitudes of transcripts analysed using microarrays. This approach has already proved useful in tumour classification by relating patterns of gene expression to response to chemotherapeutic agents (Butte *et al*., 2000). An extension of this method should allow the elucidation of gene–gene interactions and the identification of common or converging biochemical pathways. Coupled with a knowledge of putative disease-related polymorphisms and comparable expression profiles in disease tissue, microarrays (together with the nascent field of proteomics; see Chapter 16) promise to be an extremely powerful future tool for the dissection of complex disease processes. Figure 1.2 illustrates approaches for gene characterization which are useful for both prioritizing candidate genes for analysis and establishing causality in a disease process. The chapter detailing each aspect is indicated.

1.9 COMPARATIVE GENETICS AND GENOMICS

We have already touched on the role of bioinformatics in relation to the identification of functionally important DNA motifs by cross-species comparison. This area is covered more fully in Chapters 9 and 12. Recently the sequencing of a number of genomes has been completed, including the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, the fruit fly *Drosophila melanogaster* and the nematode worm *Caenorhabditis elegans;* soon these will be joined by the puffer fish species *Fugu rubripes* and *Tetraodon nigroviridis*, the zebra fish *Danio rerio* and of course the mouse and rat. This has provided an unprecedented opportunity for large-scale genome comparisons, allowing researchers to make inferences not only with regard to the identification of conserved regulatory elements, but also about genome evolutionary dynamics. Whole genome availability also provides a complete platform for the design of *in vivo* paradigms of human disease, for example transgenic and gene knock-out animal models and more sophisticated spatially and temporally regulated conditional mutants.

Large-scale approaches to biochemical pathway dissection using expression microarrays in relatively simple organisms, particularly yeast, are also proving extremely promising. Whole genome expression profiles can be generated and correlated transcription profiles identified for related groups of genes. Coincident expression patterns are frequently indicative of subsequent protein–protein interactions and co-localization in protein complexes (Jansen *et al*., 2002). Similar tissue-specific experiments can be performed for

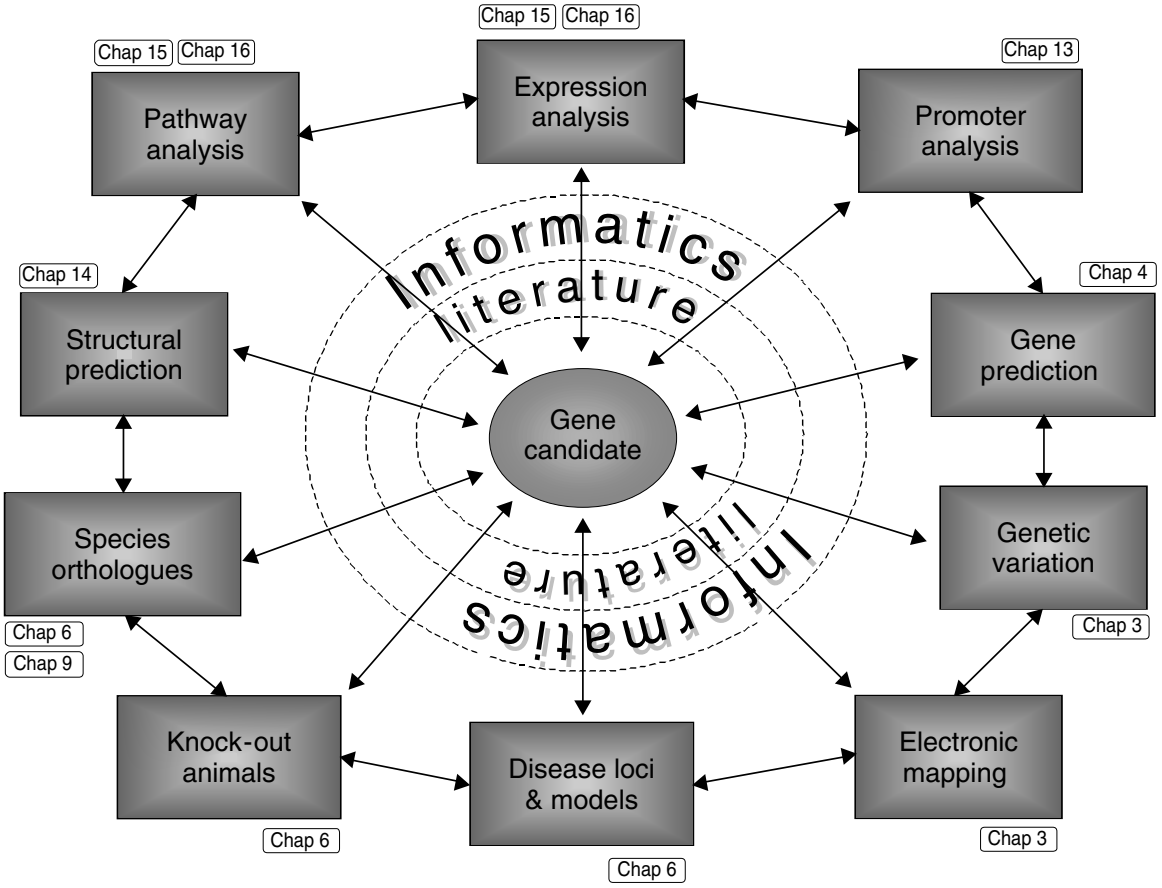

Figure 1.2 Approaches for gene characterization, indicating chapters detailing each aspect.

higher organisms, both for the purposes of identifying coincident transcription profiles for unravelling biochemical pathways and for comparison of diseased and normal tissues (see, for example Mody *et al*., 2001; Saban *et al*., 2001). Tissue derived from animal models such as mice can have advantages over using diseased human tissue: the disease model can be generated under a controlled environment, typically on an identical genetic background to the control tissue, and procurement of a significant number of high-quality tissue samples (essential for the extraction of good quality RNA) is more straightforward (see Chapter 15).

Thus far we have given a few examples of the impact of combining model organisms with high-throughput genomics technologies for improving our understanding of gene function and interaction, biochemical pathways and human disease (comparative genomics). Similar strides are being made in the field of comparative genetics (here we define genetics as phenotype-driven gene identification using genetic mapping procedures), particularly in the areas of mouse and rat genetics. The ability to perform controlled crosses such as inter-crosses and backcrosses (see Silver, 1995; Chapter 11) coupled with the development of fairly high density genetic maps over the last few years has rendered the mapping of monogenic traits in both mouse and rat a reasonably straightforward exercise. The impact of the completion of the mouse and rat genome sequences in the near future will be similar to the impact of the availability of the human genome on human genetics; indeed, the partially completed mouse and rat genomes are already giving significant improvements in speed of mapping and candidate gene

identification. These developments together with recently implemented large-scale mutagenesis programmes for the generation of monogenic mutants (see Chapter 6) promise to provide a significant increase in the mutant mouse resource in terms of simple disease models.

Significant progress has also been made in mapping complex traits in both the mouse and rat in recent years, including the development of software packages for the identification of quantitative trait loci (QTL; see Chapter 11). However, although experimental crosses can be designed to maximize the chances of success (unlike human studies), complex trait analysis in model organisms is still plagued by the difficulties in identifying and precisely localizing genes of relatively small effect. QTL linkage peaks are typically broad due to lack of absolute correspondence between genotype and phenotype and a consequent inability to identify unequivocal recombinant animals. In an attempt to overcome this limitation, mapping methods using 'heterogenous stocks' have recently been developed (Mott *et al*., 2000). The heterogenous stock comprises a mouse line resulting from inter-crossing several different inbred strains and maintaining the resulting mixed stock through several generations (typically 30–60). Each chromosome from a mouse derived from a heterogenous stock consists of a mosaic of DNA from the different founding strains, allowing a fine mapping approach based on a knowledge of the ancestral alleles in the original inbred lines. Mott *et al*. have developed publicly available software for the analysis of heterogenous stocks (see Chapter 11).

Perhaps one of the most exciting developments in model organism genetics is the fusion of classical genetics with high-throughput genomics techniques. Microarrays provide a means of checking all genes within a QTL linkage peak for subtle differences in expression levels, potentially pinpointing the culprit gene. This tactic was used successfully to reveal the role of *Cd36* in metabolic defects, following linkage analysis in the rat (Aitman *et al*., 1999). As an extension of this method, a gene expression profile may be treated as a quantitative trait and used as a phenotypic measure in linkage analysis for the identification of genes influencing the expression level, as a route to biochemical pathway expansion. Jansen and Nap (2001) recently coined the phrase 'genetical genomics' for this type of approach.

1.10 CONCLUSIONS

We hope this book will help the geneticist to design and complete more effective genetic analyses. Bioinformatics can have far-reaching effects on the way that a laboratory scientist works but obviously it will never entirely replace the laboratory process and is simply another set of tools to expedite the research of the practising biologist. Misconceptions regarding the power of bioinformatics as a stand-alone science are perhaps among the biggest mistakes that computer-based bioinformatics specialists can make and may even explain a degree of prejudice against bioinformatics — perceived by some as an '*in silico* science' with little basis in reality. Taken to an extreme and without a balanced understanding of both the application of software tools and a good appreciation of basic biological principles, this is exactly what bioinformatics can be, but where bioinformatics proceeds as part of 'wet' and 'dry' cycles of investigation, both processes are stronger as a result. In this introduction we have briefly examined some of the experimental genetics processes which can be assisted by informatics; we now invite the reader to read on for more detailed coverage of each of these processes in the remaining chapters of this book.

REFERENCES

Aitman TJ, Glazier AM, Wallace CA, Cooper LD, Norsworthy PJ, Wahid FN, *et al*. (1999). Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. *Nature Genet* **21**: 76–83.

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.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, *et al*. (2000a). An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**: 513–516.

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, *et al*. (2000b). The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genet* **26**: 76–80.

Antonellis A, Rogus JJ, Canani LH, Makita Y, Pezzolesi MG, Nam M, *et al*. (2002). A method for developing high-density SNP maps and its application at the Type 1 Angiotensin II Receptor (AGTR1) Locus. *Genomics* **79**: 326–332.

Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci USA* **97**: 12182–12186.

Campuzano V, Montermini L, Molto MD, Pianese L, Cossee M, Cavalcanti F, *et al*. (1996). Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271**: 1423–1427.

Collins A, Lonjou C, Morton NE. (1999). Genetic epidemiology of single-nucleotide polymorphisms. *Proc Natl Acad Sci USA* **96**: 15173–15177.

Debouck C, Metcalf B. (2000). The impact of genomics on drug discovery. *Ann Rev Pharmacol Toxicol* **40**: 193–207.

Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, Brookes AJ. (2002). Hobbies: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res* **30**: 387–391.

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, *et al*. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**: 599–603.

Huntington's Disease Collaborative Research Group. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**: 971–983.

Jansen RC, Nap JP. (2001). Genetical genomics: the added value from segregation. *Trends Genet* **17**: 388–391.

Jansen R, Greenbaum D, Gerstein M. (2002). Relating whole-genome expression data with protein–protein interactions. *Genome Res* **12**: 37–46.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, *et al*. (2001). Haplotype tagging for the identification of common disease genes. *Nature Genet* **29**: 233–237.

Kanehisa M, Goto S, Kawashima S, Nakaya A. (2002). The KEGG databases at GenomeNet. *Nucleic Acids Res* **30**: 42–46.

Kruglyak L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet* **22**: 139–144.

Kruglyak L, Nickerson DA. (2001). Variation is the spice of life. *Nature Genet* **27**: 234–236.

Lai E, Riley J, Purvis I, Roses A. (1998). A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. *Genomics* **54**: 31–38.

Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, *et al.* (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**: 66–71.

Mody M, Cao Y, Cui Z, Tay KY, Shyong A, Shimizu E, *et al.* (2001). Genome-wide gene expression profiles of the developing mouse hippocampus. *Proc Natl Acad Sci USA* **98**: 8862–8867.

Mott R, Talbot CJ, Turri MG, Collins AC, Flint J. (2000). A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc Natl Acad Sci USA* **97**: 12649–12654.

Pericak-Vance MA, Bass ML, Yamaoka LH, Gaskell PC, Scott WK, Terwedow HA, *et al.* (1998). Complete genomic screen in late-onset familial Alzheimer's disease. *Neurobiol Aging* **19** (1 Suppl): S39–S42.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, *et al.* (2001). Linkage disequilibrium in the human genome. *Nature* **411**: 199–204.

Risch NJ. (2000). Searching for genetic determinants in the new millennium. *Nature* **405**: 847–856.

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, *et al.* (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**: 1066–1073.

Saban MR, Hellmich H, Nguyen NB, Winston J, Hammond TG, Saban R. (2001). Time course of LPS-induced gene expression in a mouse model of genitourinary inflammation. *Physiol Genomics* **5**: 147–160.

Saunders AM, Schmechel D, Pericak-Vance M, Enghild J, Salvesen GS, *et al.* (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci USA* **90**: 1977–1981.

Saunders AM, Strittmatter WJ, Schmechel D, St. George-Hyslop PH, Pericak-Vance MA, Joo SH, *et al.* (1993a). Association of apolipoprotein E allele E4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43**: 1467–1472.

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.

Silver LM. (1995). *Mouse Genetics: Concepts and Applications*. Oxford University Press: Oxford, UK.

Spielman RS, McGinnis RE, Ewen WJ (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**: 506–516.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. (1995). Serial analysis of gene expression. *Science* **270**: 484–487.

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 2

Internet Resources for the Geneticist

MICHAEL R. BARNES[1] and CHRISTOPHER SOUTHAN[2]

[1]*GlaxoSmithKline Pharmaceuticals*
Harlow, Essex, UK

[2]*Oxford GlycoSciences UK Ltd*
The Forum, 86 Milton Science Park
Abingdon OX14 4RY, UK

2.1 INTRODUCTION

The World Wide Web ('the web') and our knowledge of human genetics and genomics are both expanding rapidly. By allowing swift, universal and largely free access to data, particularly the human genome sequence, the web has already played an important role in the study of human genetics and genomics. Increased data accessibility is dramatically changing the way the scientific community is communicating and carrying out research. The internet biology community is expanding daily with an organic development of websites, tools and databases, which could eventually replace the conventional scientific paper as the predominant form of communication. Already we are starting to see successful website/journal hybrids such as *Genome Biology* (http://genomebiology.com/) and biomednet (biomednet.com) which offer high quality peer-reviewed scientific articles and reviews alongside bioinformatics databases and tools. Many more established journals like *Nature* and *Science* are rapidly following suit with user-friendly websites, which offer much more than the full text of the journal.

The web is offering more than just information. Virtual research communities have been organized around databases and specialist research groups. These communities are even influencing the way bioinformatics tools are being developed, a good example being Ensembl the human genome browser developed at the EBI and Sanger Institute in Hinxton, Cambridgeshire (Hubbard *et al*., 2002). In the spirit of open source community projects such as the free UNIX operating system Linux, the Ensembl development team has developed Ensembl on an 'open source' basis. This means all code is freely available to anyone who wishes to download it. But further still, Ensembl is developed by a 'virtual community' of developers from institutes, industry and academia around the world who are free to modify and add to the central software code (subject to a peer review). So the tools and interfaces, though primarily developed in Hinxton, may include contributions from developers in Singapore, North Carolina and New York or elsewhere.

2.1.1 Hypothesis Construction and Data Mining — essentials for Genetics

Genetics is a science which calls for analysis and interpretation across a wide range of biological research. Many chapters in this book deal with focused tools. Beyond these specialist applications however, geneticists need access to a wide range of databases and literature, both to update particular research areas and formulate new hypotheses. This requires expertise across the gamut of biological data on the internet. This ranges from the review literature to highly specific databases. This can illuminate biology from gene function to biological pathways. Effective data mining needs an understanding of the general principles by which it is organized, particularly the sequence-based data resources. This needs to be backed up by good scientific judgment concerning quality and significance.

An exhaustive description of biological data and databases on the internet would be beyond the scope of this book. Confucius might not have been thinking of internet searching when he said 'give a man a fish and he will live for a day, teach a man to fish and he will live forever', but the principle still applies. So, instead of reviewing the data

resources themselves the most useful thing we can do here is to review search methods to help identify both current and future resources.

2.2 SUB-DIVISION OF BIOLOGICAL DATA ON THE INTERNET

Biological information on the internet can be roughly subdivided into two broad categories, which we will term 'the biological internet' and 'biological information on the internet'. This distinction may not be immediately apparent — we define 'the biological internet' as purpose-built biological tools and databases which index and contain detailed biological information, such as the human genome sequence, nucleotide and protein sequences, genetic markers, polymorphisms and the full range of biological literature. The majority of these tools and databases are maintained in a highly integrated form by major biological organizations such as NCBI and EMBL. We define 'biological information on the internet' as biological data which is less formally maintained on the web, this could include information on research laboratory homepages, conference abstracts, tools, boutique databases and any other data that scientists have seen fit to present on the web.

These distinctions are more clearly defined by the tools that are available to search the data. Firstly there are general purpose web search engines, such as Google, Lycos and Excite (see Table 2.1 for a full list), these tools index and search the full range of the internet and have the capability to identify webpages, tools and databases by simple keyword searching. A second category of tools are the specialist biological search tools, such as Entrez-PubMed and BLAST (see Chapter 4). The former uses keyword searching or accession number queries, the latter uses similarity searching to find related sequences.

The choice of search tool depends on the kind of information that needs to be retrieved. The scope of biological and genetic information on the internet is so broad that no single tool is available to index all data. The key point to understand is which tool is most suitable to identify a specific form of data. For example literature is most comprehensively indexed by PubMed or Scirus (see below), whereas nucleotide records can only be identified with any specificity by Entrez or BLAST. This is in contrast to a laboratory homepage or a boutique web resource. Unless a description is published in PubMed these resources may only be identified by a web search tool. If it is not clear what information needs to be retrieved then clearly both specific and general search tools should be used.

TABLE 2.1 Key Internet Search Engines with Reported Index Size (Equivalent to the Number of Documents Indexed)

Search Engine	URL	Reported Index Size
Google	http://www.google.com/	560 M
AltaVista	http://www.altavista.com/	350 M
FAST	http://www.alltheweb.com/	340 M
Northern Light	http://www.northernlight.com/	265 M
Excite	http://www.excite.com/	250 M
HotBot	http://www.hotbot.com/	110 M
Lycos	http://www.lycos.com/	110 M
MetaCrawler	http://www.metacrawler.com/	ND
Scirus	http://www.scirus.com/	69 M (science only)

2.3 SEARCHING THE INTERNET FOR GENETIC INFORMATION

The World Wide Web began as an information-sharing and retrieval project at the European particle physics laboratory CERN (Berners-Lee *et al*., 1999). It has only recently evolved into the mass media beast that we all know. But just as the internet began, so it continues as an information-sharing resource for scientists in all fields. One cannot deny that commercial proliferation has not been an unmitigated success for the growth of the web but this has led many scientists to perceive the internet as a rising tide of irrelevant noise that has largely washed away any intrinsic value. This is a misconception. We will demonstrate that some web resources for biological sciences are both outstanding and indispensable. Internet biology suffers as much as any other field of scholarship from: data of dubious provenance, broken links, outdated sites and newsgroup spam. But it also contains valuable and novel data which can be crucial for scientific discovery. The skill is to recognize chaff and know how to sift the wheat from it. To do this we need tools that are capable of highlighting relevant information in an organized manner.

In the process of linking genotypes to phenotypes it is important to know about the function of a gene or gene family, for example to prioritize candidate disease-association genes. In such cases biological search tools and internet search tools may provide complementary results. To give an hypothetical example let us assume that a genetic locus associated with a familial form of basal cell carcinoma includes a novel gene with homology to WNT genes. With no knowledge of WNT genes it would be difficult to include or exclude this gene as a candidate. A search of PubMed would reveal a daunting range of over 1000 publications mentioning members of the WNT gene family. Some might contain specific information to link WNT genes to carcinoma but it would take a long time to read and digest all the available information. Using Google to search for 'WNT gene' would identify a range of conference abstracts and laboratory homepages. Towards the top of the hit-list this would include the 'World Wide WNT Window' (www.stanford.edu/□rnusse/wntwindow.html). This is an excellent summary of the whole WNT signalling pathway maintained by prominent researchers in the WNT signalling field. The page includes a detailed and regularly maintained summary of all genes in this highly complex pathway, which is currently unpublished. Examination of this pathway would identify the Patched receptor upstream, which has been shown to cause 80% of sporadic basal cell carcinomas. This is just one of many examples of how a thriving unpublished and unpublicized on-line research community can be identified by opportunistic internet searching.

2.4 WHICH WEB SEARCH ENGINE?

In a nutshell the availability of full-text search engines allows the web to be used as a searchable 15-billion-word encyclopedia. However, because the web is a distributed, dynamic, and rapidly growing information resource, it presents many difficulties for traditional information retrieval technologies. This why the choice of the search methodology used for searching can lead to very different results.

An important point to make is that all search engines are not the same. A common misconception is that most internet search engines index the same documents for a large proportion of the web. In fact the coverage of search engines may vary by an order of magnitude. An estimated lower boundary on the size of the indexable web is 0.8 billion pages

(http://www.neci.nec.com/□lawrence/websize.html). Many engines index only a fraction of the total number of documents on the web and so the coverage of any one engine may be significantly limited. Combining the results of multiple engines has been shown to significantly increase coverage. This is done automatically with metasearch engines such as MetaCrawler (www.metacrawler.com), which search and combine the results of several search engines. Table 2.1 presents a selection of web search engines with direct applicability to biological searching. We also recommend the website, SearchEngineWatch.com, for reviews and reports on new search engines.

2.4.1 Google

It is apparent from Table 2.1 that Google offers the widest indexing capacity. This is an innovative search engine based on scientific literature citation indexes (Butler, 2000). Conventional search engines use algorithms and simple rules to rank pages based on the frequency of the keywords specified in a query. Google exploits the links between webpages to rank hits. Thus the highly cited pages of the web world with many links pointing to them are ranked highest in the results. This is an efficient searching mechanism which effectively captures the internet community 'word of mouth' on the best and most frequently used webpages.

2.4.2 Scirus

The greatest limitation for web search engines is unindexed databases. These include many of the databases that make up the biological internet, such as sequence databases and some subscription-based resources such as full-text journals, and commercial databases. Although limited material from these sites, such as front pages, documentation and abstracts are indexed by search engines, the underlying data is not available because of database firewalls and/or blocks on external indexing.

In an attempt to solve this problem, the publisher Elsevier has developed Scirus (http://www.scirus.com/). This is a joint venture with FAST, a Norwegian search engine company who have produced an excellent specialist scientific search engine. Scirus enhances its specificity and scope by only indexing resources with scientific content. These include webpages, full-text journals and Medline abstracts. This makes Scirus an effective tool for both web and literature searching tool. Both full text and PDF format journal content is indexed by performing a MetaSearch of the other major providers of full text — Elsevier's ScienceDirect and Academic Press's IDEAL collection. Scirus also searches the web for the same key words, including Medline, patents from the databases of the US Patent Office, science-related conferences and abstracts. The Medline database is provided on the BioMedNet platform, which requires a free BioMedNet login and password for access. Scirus offers the user several options to customize their searches to search only free sites, only membership sites or only specific sites. The advanced interface also allows boolean queries (see below). By March 2002 Scirus had indexed 69 million science-related pages, including PDF files and peer-reviewed articles, thereby covering the majority of the biologically relevant internet.

Although Scirus expands the scope of biological data searching beyond other search engines it falls short in some areas. For example the full-text journals are restricted to Elsevier and Academic Press. Coverage is also restricted by index pre-filtering that might miss some websites. Another disadvantage is that search results tend to be redundant. Although for literature searching there are alternative full text searching tools such as

HighWire (see below) Scirus is tantalizingly close to what a universal biological search engine should be.

2.5 SEARCH SYNTAX: THE MATHEMATICS OF SEARCH ENGINE USE

The best search engine in the world will not retrieve relevant results unless the query is correctly defined. This is easy to master and a few basic commands can turn a poor specificity keyword search into a highly targeted query. The key to successful sifting of the web is to select for the minimum number of irrelevant hits (maximize specificity) but avoiding the exclusion of relevant hits (minimize false negatives).

2.5.1 Using the '+ and −' Symbols to Filter Results

Sometimes it is necessary to ensure that a search engine finds pages that have all the words you enter, not just some of them. This can be achieved by using the '+' symbol. Similarly you may wish to exclude a specific word from your search by using the '−' symbol. These commands work with nearly all the major web search engines and are similar in function to the boolean operators 'AND' and 'NOT' respectively.

As an example let's say you wish to find information about human promoter prediction tools. You could search using [+ promoter + prediction + tool]. This will only retrieve pages that contain all three words. If the search returns excessive information by including tools for plant and bacterial promoter prediction, one could further refine the search by using the following search query [+ promoter + prediction + tool − plant − prokaryote]. This will subtract pages which mention plants and prokaryotes. Be aware though that this might filter out valid hits to tools which analyse *both* prokaryote and eukaryote sequences.

2.5.2 Using Quotation Marks to Find Specific Phrases

The most complex filtering syntax on our promoter prediction query still manages to retrieve over 1000 results, so we need to consider other methods of reducing the number of hits. One approach is to use a phrase search that will find only those pages where the terms appear in exactly the order specified. This is achieved by putting quotation marks around the phrase, so we might search with ['promoter prediction tool']. This retrieves six relevant hits but clearly many sources have been filtered out, so it is important to beware of over-specifying search terms.

2.5.3 Restricting the Searching Domain of a Query

A final measure that can be taken to fine tune your query is to restrict the internet domain. For example you can restrict your search to only identify hits in the .edu (educational) domain or to ignore hits from the .com (company) domain. This is achieved in Google and most other sites by using the [+ site:.edu] to include a domain or [− site:.com] to exclude a domain. This command can be extended further to search only a specific site, e.g. to search the NCBI website for SNP information try [+ SNP + site:ncbi.nlm.nih.gov].

Table 2.2 includes the search results obtained from the different variations on the search for promoter prediction tools, using both Google and Scirus. This shows the improvements

TABLE 2.2 Different Results Obtained from Different Query Targeting Methods. Results Compare the Number of Hits Returned by the General Search Engine Google and Specialist Science Search Engine Scirus

Query	Google Hits	Scirus Hits[]
+ promoter + prediction + tool	4050	2379
'promoter prediction tool'	6	2
'promoter prediction tools'	14	8
+ promoter + prediction + tool − plant	2630	1312
+ promoter + prediction + tool − plant − bacterial	2080	936
+ promoter + prediction + tool − plant − bacterial − site:.com	1750	NA

[]Queries to Scirus were designed using the equivalent boolean syntax in the advanced search form.

from filtering on the query. The final word on fine tuning web search queries is to be as flexible as possible. Try to use keywords which are likely to be specific to the kind of website or tool you are looking for. Sometimes it is useful to go to a page or tool similar to the one you are looking at to check for very specific words that might be shared by similar sites. For example, in the case of promoter prediction tools, a commonly occurring word was 'server'; exchanging this for 'tool' significantly improves the relevance of the hits.

2.6 BOOLEAN SEARCHING

Although the familiar boolean search commands (AND, OR, NOT) are widely used for many forms of database searching, including PubMed, they are not universally supported by all web search engines. Table 2.3 lists those supported by the most popular search engines. The functionality offered by AND and NOT mirrors the functionality of [+ and −]. Other commands have a distinct function, for example [SNP OR Analysis] will retrieve all webpages that contain the words SNP or analysis. The NEAR command is not

TABLE 2.3 Boolean Commands Supported by Popular Web Search Engines

Command	How	Supported by
Or	OR	AltaVista, Excite, Google, Lycos, Northern Light
	None	FAST, LookSmart,
And	AND	AltaVista, Excite, Lycos, Northern Light
	None	FAST, Google, LookSmart
Not	NOT	Excite, Lycos, Northern Light
	AND NOT	AltaVista
	None	FAST, Google, LookSmart,
Near	NEAR	AltaVista (10 words), Lycos (25 words)
	None	FAST, Google, LookSmart

widely supported but can be useful to help to identify two keywords in close proximity to each other.

2.7 SEARCHING SCIENTIFIC LITERATURE — GETTING TO 'STATE OF THE ART'

Effective mining of the literature is important at the stages of conception, design and construction of genetic studies. At the most basic level it is important to be aware of the 'state of the art' in a research area before embarking on new efforts. At the very least this avoids duplication of effort, but it can also provide previously unrecognized clues which need to be followed up. Unfortunately this important informatic process is still lacking truly innovative tools and databases. We are still struggling with tools that cover the fundamentals of literature searching, such as making the full text of *all* journals available for searching. Even with unlimited access to full text, the problems with effective literature mining are profound. Some of these problems stem from the limitation of language as a precise query tool — there is simply too much vocabulary to describe or specify the same target information. Some databases attempt to minimize the impact of this problem by the use of controlled vocabulary and gene nomenclature. But in the absence of such measures, flexible composition of queries becomes quite critical to obtain comprehensive coverage of a research area.

There are many commercial and publicly available tools and databases for mining scientific literature which vary in their data content. Some offer access to proprietary curated databases but they all employ essentially similar keyword-based interfaces with a facility for boolean operators to combine and subtract keywords.

2.7.1 PubMed

PubMed is the most widely used free literature searching tool for biologists. It forms part of the Entrez-integrated database retrieval system at the NCBI and is essentially a web interface to the Medline database which indexes >11 million journal abstracts. It also provides links to the full text of more than 1100 journals available on the web, although search queries are restricted to the text in abstracts. The interface allows the user to specify a search term (any alpha numeric string) and a search field (e.g. title, text word, journal or author). Queries retrieve abstracts from most of the major journals, although not all journals are indexed, particularly newer journals or journals with lower impact factors. There is a surprisingly stringent threshold applied before a journal will be considered for Medline indexing.

Many of the same guiding principles applied to searching the web also apply to PubMed, but there are some differences between this tool and other more general web search engines. Firstly the boolean operators are limited to the three main operators AND, OR and NOT. One major improvement over most web search engines is the availability of a wildcard function (□) to designate any character or combination of characters. The creative use of wildcards and boolean terms is important to widen the search without retrieving excessive and irrelevant results. For example, to find publications which present evidence of schizophrenia association on chromosome 8q21, an appropriate PubMed query might be [schizo□ AND 8q□] searching the *text word* field. Using a wildcard search with 'schizo□' instead of 'schizophrenia' retrieves articles which mention schizoaffective, schizophrenia or schizophrenic, all of which may be relevant. By using a wildcard with

'8q' the search will retrieve nearby loci or larger loci which may encompass 8q21, e.g. 8q13–8q22. These are simple points but they are integral to a successful search strategy. Those using these facilities extensively will find additional searching guidelines on the NBCI website.

2.8 SEARCHING FULL-TEXT JOURNALS

Prospects for literature searching have improved recently with the greater availability of full-text articles. We have already described the advances offered by Scirus in searching full-text journals and the web simultaneously. Other highly recommended websites are HighWire which is approaching comprehensive coverage of available full-text journals and Medline (see below). However, searching scientific publications is still somewhat decentralized and there is still no completely comprehensive central tool to search all full-text journals, although it is possible to search the full text of most of the major genetics journals by visiting the top three or four major publishers. Table 2.4 lists the major sites which index the full text of a large range of science journals. As a benchmarking test we queried each tool, with a standard full-text query for the keyword [WNT], where searching Medline was also an option we identified the combined number of full text and Medline hits in parentheses. The highest number of results was retrieved with Scirus, however these results were very redundant. The HighWire tool seemed most effective in the benchmarking test, identifying a high number of hits with no redundancy.

2.8.1 HighWire

HighWire was set up as a non-profit making organization in 1995 by Stanford University to help universities and societies to publish on the web at low cost (Butler, 2000). Since its launch HighWire has expanded to become the world's second-largest scientific repository, after the US space agency NASA's Astrophysics Data System (which contains no biological information). Many journals available on the HighWire site make their content free immediately, or 1 or 2 years after print publication often coupled with an early view service for papers in press. In March 2002, HighWire had indexed 410,821 free full-text articles, derived from a list of 324 full-text journals. These are listed on the website along with Medline records from January 1948 through to April 2002. In our benchmark test

TABLE 2.4 Major Websites Providing Full-text Journal Access and Searching

Site/Publisher	Test Query Hits (with Medline)	URL
PubMed	(1615)	http://www.ncbi.nlm.nih.gov/entrez
Scirus	5061 (7015)[]	http://www.scirus.com/
HighWire	2651 (3738)	http://highwire.stanford.edu/
Biomednet	1192 (2749)	http://www.bmn.com
ScienceDirect (Elsevier)	1264	http://www.sciencedirect.com
IDEAL	565	http://www.idealibrary.com/
Nature Publishing Group	255	http://www.nature.com/nature/
Wiley InterScience	196	http://www.interscience.wiley.com/

[]Results from Scirus were redundant.

against other full-text search tools a comparative search of PubMed and HighWire with the keywords [Wnt16 OR Wnt-16] identified two papers with PubMed and eight papers with HighWire.

2.8.2 Literature Digests and Locus-specific Databases

The literature searching process can be simplified by searching locus-specific databases. The most widely used is On-line Mendelian Inheritance in Man (OMIM). As the name suggests, this focuses on Mendelian monogenic disorders, although it also offers some coverage of complex diseases. As a manually curated digest of the literature extracted from the full text of publications it can contain more information than PubMed. Although this has the disadvantage that not all entries are fully comprehensive or current, the database usually captures the most salient information and is therefore a good place to start. In addition OMIM is fully integrated with the NCBI database family. This facilitates rapid and direct linking between disease, gene sequence and chromosomal locus.

Other databases are available which provide curated information about genes and diseases which can also help to speed up the literature searching process. One of these is GeneReviews (www.geneclinics.org). This complements the molecular genetics emphasis of OMIM by offering a distinctly different focus. GeneReviews is a medical genetics information resource aimed at physicians and other healthcare providers. The site provides current, expert-authored, peer-reviewed, full-text articles describing the application of genetic testing to the diagnosis, management and genetic counselling of patients with specific inherited conditions. It also contains an international genetic testing Laboratory Directory and an international genetic and prenatal diagnosis Clinic Directory.

2.9 SEARCHING THE HEART OF THE BIOLOGICAL INTERNET — SEQUENCES AND GENOMIC DATA

So far we have reviewed a range of tools and approaches for searching the wider internet and the specialist scientific literature for biological information which may be useful for genetics. All of the tools reviewed so far may provide links, but will stop short of direct retrieval of actual biological database records, such as DNA or protein sequence records. This biological information is the heart of the biological internet. However, the flood of sequence data from genome sequencing has rapidly pushed biological sequence data beyond the reach of general internet searching tools. Instead sequence data can be searched and retrieved by using specialist bioinformatics tools based on sequence homology, map location, keyword, accession number and other features in the records. At a basic level this can be done by keyword searching using search tools such as, Entrez at the NCBI (Schuler *et al*., 1996) or SRS at the EBI (Zdobnov *et al*., 2002). Moving beyond simple searching methods the biological databases are constantly being updated and re-engineered to allow more powerful data query methods. These methods are covered in many other chapters throughout this book.

2.10 NUCLEOTIDE AND PROTEIN SEQUENCE DATABASES

There are three major organizations that collaborate to collect publicly available nucleotide and protein sequences. These organizations share data on a daily basis but they are distinguished by different international catchment areas for submissions, different formats and

sometimes differences in the nature of their submitter annotations. Genbank is maintained by the NCBI in the United States (http://www.ncbi.nlm.nih.gov/Genbank/index.html). EMBL is maintained by the European Bioinformatics Institute in the United Kingdom (http://www.ebi.ac.uk/). The third member is the DNA Database of Japan (DDBJ) in Mishima, Japan (http://www.ddbj.nig.ac.jp/). All three organizations offer a wide range of tools for sequence searching and analysis but two integrated database query tools have become pre-eminent. These are Entrez from the NCBI and SRS from the EBI.

2.10.1 Entrez

Entrez (http://www.ncbi.nlm.nih.gov/Entrez/) is the backbone of the NCBI database infrastructure. It is an integrated database retrieval system that allows the user to search and browse all the NCBI databases through a single gateway. Entrez provides access to DNA and protein sequences derived from many sources, including genome maps, population sets and, as already described, the biomedical literature via PubMed and Online Mendelian Inheritance in Man (OMIM). New search features are being added to Entrez on a regular basis. Most recently facilities have been added to allow searches for DNA by 'ProbeSet' data from gene-expression experiments and for proteins by molecular weight range, by protein domain or by structure in the Molecular Modelling Database of 3D structures (MMDB).

2.10.2 Sequence Retrieval Server (SRS)

The sequence retrieval server (SRS) serves a similar role to Entrez, for the major European sequence databases. SRS is a flexible sequence query tool which allows the user to search a defined set of sequence databases and knowledge-bases by accession number, keyword or sequence similarity. SRS encompasses a very wide range of data, including all the major EMBL sequence divisions (Table 2.5). SRS goes one step further than Entrez by enabling the user to create analysis pipelines by selecting retrieved data for processing by a range of analysis tools, including ClustalW, BLAST and InterProScan.

2.11 BIOLOGICAL SEQUENCE DATABASES — PRIMARY AND SECONDARY

Anyone entering the heart of the biological internet encounters a bewildering number of accession numbers, identifiers and gene names. To get to grips with this flood of terminology it is important to understand the difference between primary and secondary databases and their associated accession numbers. This is not proposed as a rigorous definition but it does have a utility for understanding the information flow between sequence databases.

2.11.1 Primary Databases

Primary accession numbers have a number of key attributes; they refer to nucleic acid sequences derived directly from a sequencing experiment, the results are submitted by authors in a standardized format to GenBank, EMBL or DDBJ, the accession numbers are both unique and stable (if they are updated or amended by the submitting authors the accession number will signify a version change as .1, .2 etc.), the data records from every accession number can be retrieved, a contactable submitter is included in every record,

TABLE 2.5 Databases Indexed by the Sequence Retrieval Server at the EBI

Data Type	Database
Scientific literature	Medline, GO, GOA
Protein sequence libraries	European, Japanese and US protein patents, SWISS-PROT, SpTrEMBL
DNA sequence libraries	EMBL, Ensembl HUMAN, global DNA patents
Protein motifs	INTERPRO, PROSITE, PRINTS, PFAM, PRODOM, NICEDOM
DNA sequence related	UTR, UTRSITE, BLOCKS, TAXONOMY, GENETICCODE, REBASE, EPD, CPGISLAND, ENSEMBLCPG, UNIGENE
Transfac (Transcription factor analysis)	TFSITE, TFFACTOR, TFCELL, TFCLASS, TFMATRIX, TFGENE
Protein3DStruct	PDB, DSSP, HSSP, FSSP, RESID
Mutations	SWISSCHANGE, EMBLCHANGE, OMIM, HUMUT, HUMAN_MITBASE, P53LINK, Locus Specific Mutations (see Chapter 3)
SNPs	HGBASE, HGBASE_SUBMITTER
RH mapping	RHDB, RHEXP, RHMAP, RHPANEL
Metabolic pathways	LENZYME, LCOMPOUND, PATHWAY, ENZYME, EMP, MPW, UPATHWAY, UREACTION, UENZYME, UCOMPOUND
SRS pipelineapplications	FASTA, FASTX, FASTY, NFASTA, BLASTP, BLASTN, CLUSTALW, NCLUSTALW, PPSEARCH, RESTRICTIONMAP, HMMPfam, InterProScan, FingerPRINTScan, PfScan, BlastPRODOM, ScanRegExp

they are explicitly redundant in that all submissions are accepted regardless of partial or complete overlap with existing entries and lastly the growth rate remains close to exponential and now exceeds 16 million sequence records. The concept of authors' needs stretches to encompass consortia that run high-throughput sequencing projects. One of the most valuable and perhaps overlooked principals of these unique public repositories is that there is always (with the exception of patent data, see below) an identified individual or laboratory representative listed with the sequence record who can be contacted for any queries regarding experimental details, data quality and availability of source material. There is a large amount of information associated with primary sequence records. These include primary accession numbers, version numbers, protein ID numbers, gene identifier (GI) numbers, header records and feature identifiers. These cannot be covered in detail here but full descriptions are given in database guides (http://www.ebi.ac.uk/embl/index.html) and release notes (ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt).

Geneticists should be encouraged to contact submitting authors in cases where anything seems non-obvious about primary data records for an mRNA or a finished genomic clone. They may have extra information that has a crucial bearing on the interpretation of genetic experiments. Authors may be difficult to track down if they have moved institutions but they are usually pleased to assist in the utilization of their data, because as with scientific publishing, this is the principle behind public sequence databases. Technical errors,

anomalies, miss-annotation in submissions or artefacts are entirely the responsibility of submitting authors not the database administrators. Although we should be sanguine concerning anomalies in the high-throughput data divisions (EST, GSS, STS, HTG, HTC and SNP) if problems are pointed out authors can certainly amend or update their entries or in some cases may withdraw them. The primary data is deposited in good faith so authors should certainly not be harshly judged if an error has occurred in the rough and tumble of cloning, sequencing and submitter annotation. The exception to author responsibility for GenBank records is the patent division (gbPAT) where inventors are not equivalent to academic authors. These sequence records are processed by the US, European and Japanese patent offices and forwarded on to the databases. Although author contact may not be practical database users should be aware that patent applications are public documents and for an increasing number of gbPAT records the documentation can be accessed via the patent number on-line and free of charge (http://ec.espacenet.com/espacenet/ and http://www.uspto.gov/patft/). It is also possible to get to these patent full-text links directly from sequence entries via SRS.

2.11.2 Secondary Databases — Nucleic Acids and Proteins

By definition secondary databases are derived from the primary data. The word secondary should not be taken to imply lower value; indeed they include sources of the highest utility for genetic research. However they are defined, it is important to understand how they are linked back to the experimental data. The good news for geneticists is that there is now a comprehensive selection of high quality secondary databases that extract and collate subsets of mRNA, genomic or protein sequences from primary GenBank entries. The bad news is that the proliferation of features that make secondary databases so powerful also presents a bewildering range of options to the user. Testimony to both the good and bad news is given by the 2002 update of the Molecular Biology Database Collection (http://nar.oupjournals.org/cgi/content/full/30/1/1/DC1). This covers no less than 355 databases, up from 281 in 2001, of which the primary databases, GenBank, EMBL and DDJB, constitute only three entries. Although this compendium includes many non-human data sources almost all of these secondary databases contain information that could be pertinent to mammalian genetics. These review issues appear every January in *Nucleic Acids Research* and are definitely worth browsing. Are the genome portals secondary databases? This is where the definitions become blurred. Because NCBI generate their own genomic contig accessions (NT numbers) and Ensembl generate their own exon and gene identifiers they could be considered secondary databases. In so far as the UCSC genome portal marks up only external sequence record identifiers (primary and secondary) they are not strictly a secondary database. However, because they usefully give every type of gene prediction in the display a retrievable identity number, they could be considered as a secondary database.

The value of secondary databases includes the following:

- Distilling down a massive number of overlapping and/or redundant primary GenBank entries to a manageable range of genomic sections, unique transcripts and translated protein sequences
- Maintaining a running total of gene products, they partition human gene products and other vertebrates with extensive genomic data such as mouse, rat and zebra fish
- The inclusion of informative graphic displays for sequence features
- Providing access to a vast amount of pre-processed bioinformatic data

- Extensive interconnectivity through web hot-links
- Many of them are backed up by extensive institutional resources and expertise

However, users of these secondary databases also need to be aware of their shortcomings:

- They all suffer from the snapshot problem i.e. the time to re-build or update massive data sets means they are always out of date with respect to the new data cascading into the primary databases (given the complexity of the processes this is entirely expected but they often do not display the dates when the primary records were extracted)
- They all have different look-and-feel interfaces thereby necessitating regular practice to get the best out of them
- The web-based interoperativity can leave a lot to be desired; e.g. broken links, link-outs to databases that are not maintained to the same standards and overkill by linking out to too many similar sources
- Their automated annotation schema can be confounded by sequence artefacts (Southan *et al*., 2002)
- The overlap between utility and content between major databases is extensive but is never enough for any of them to be the mythical 'one-stop-shop'
- Non-redundant transcript and protein collections may seem conceptually similar but because they diverge in schema details and update frequency they all give different statistics
- Some secondary databases such as SwissProt keep sequence identifiers both unique and stable but for technical reasons others, such as UniGene EST clusters or Ensembl genes, may change identifiers between builds
- Many specialized 'boutique' databases are never updated when their originators move on or run out of resources
- Last but not least some secondary databases that initially had free access can become commercial and require a subscription fee

2.11.3 Nucleic Acid Secondary Databases

For the analysis of their results the geneticist must become acquainted with these feature-rich sources of gene product information. A key example, based around nucleic acid sequence but including protein of secondary databases is LocusLink/RefSeq (LLRS) for mRNAs. The LLRS system is built round a reference sequence (RefSeq) which is usually the longest available mRNA of those coding for the same protein. RefSeq includes splice variants and if only genomic sequence is available, such as for many of the 7TM receptors, the system defaults to the predicted coding sequence annotated as a 'CDS' in the database entry. For example there is no experimentally determined human rhodopsin mRNA in GenBank, only a model mRNA predicted from the genomic sequence U49742. This presents an immediate problem for the geneticist, as the untranslated region (UTR) of the rhodopsin locus, which defines the boundaries and functional regions of the gene may be extensive. Chapter 4 takes a detailed look at approaches to help define the true extent of gene loci.

The end-product of the RefSeq pipeline is a unique mRNA, coding sequence (CDS), or set of splice variants for those gene products where data or predictions are available.

The LocusLink side of things, as suggested by the title, is directed towards mapping the RefSeq gene products onto the genomic sequence and checking the consistency between the two. LocusLink has linked sections of key importance to the geneticist. These are: variation which assigns SNP data, OMIM which includes verified monogenic disease links, homologene which indicates close homologues in other species, UniGene which specifies ESTs clusters associated with the gene product, and PubMed that links to all publications that can be specifically linked to the primary GenBank accession numbers. There are also links to all three genome portals, NCBI, UCSC and Ensembl. There has been some confusion in the past where the portals could not synchronize their builds and track displays with GP version updates but this problem has been addressed and they should all be on version 28 (from December 2001) at the time of writing.

The RefSeq identifier is secondary in the sense that it is a supplementary identifier assigned to one particular mRNA chosen as the reference sequence. These accession numbers have the prefix NM_ for mRNA entries and NP_ for protein entries. The LocusLink/RefSeq system goes one step further in assigning a third identifier, XM_ for nucleic acid and XP_ for proteins, which are the genomic counterparts of the NM and NP numbers. A BLAST search against the NCBI protein database will show all three entries, the primary accession number, the NM_ and the XM_ entries. There is the added complication that the XP_ sequences have a variable evidence support level and include *ab-initio* genomic predictions both with and without EST support. Secondary accession numbers are also important for ESTs. ESTs can be considered as mRNA fragments that, with sufficient sampling (now just exceeding 4 million human entries in dbEST) can be clustered or assembled to form a contiguous extended transcription product and in some cases, the splice variants from the tissue types sampled for EST preparation. The main post-genomic utility of EST collections is as exon detectors. In addition to splice variants these can reveal possible gene transcription activity where no extended mRNA has been experimentally verified. The primary data source for ESTs is the dbEST division of GenBank.

The geneticist should be aware of two major secondary EST databases, UniGene (Wheeler *et al*., 2002) and the TIGR human gene index (Liang *et al*., 2000). The principles by which these different databases are constructed, are explained in the appropriate source references but in fact they both converge to a similar set of 'virtual' surrogate transcripts. In the TIGR case, the virtual transcripts assembled from overlapping ESTs can be retrieved; in the Unigene case, the individual EST reads can be batch downloaded. As with most secondary databases, built from the same source data, the two databases have both overlap and complementarity. The TIGR assemblies are particularly useful for extending the 3′ UTR of known mRNAs but the assemblies are re-compiled at long time intervals. UniGene is updated more frequently and is fully interlinked to the LocusLink/RefSeq system but the clusters are built on mRNAs from the preceding version of GeneBank.

2.11.4 STSs and SNPs

These are two of the most important data sources for the geneticist involved in disease mapping. The dbSTS database contains sequence and mapping data on short genomic landmark sequences. Although they have a primary sequence record and GB accession number they also have a number of alternative marker names. These have been cross-referenced into a secondary database called UniSTS that integrates all available marker and mapping data (http://www.ncbi.nlm.nih.gov/genome/sts/). The dbSNP database is an interesting exception in that it is not a division of GenBank so it is not strictly a primary database. The

submissions (SS numbers) are equivalent to a primary record but overlapping sequences with the same polymorphism are collapsed into the Reference SNP Cluster Report with an RS number. This can be considered a secondary database where the RS numbers are non-redundant and stable. These RS numbers, currently at 2,640,509 for human, are integrated with other NCBI genomic data and primary GenBank records containing overlapping sequences deduced or stated to be from the same location. The HGVbase has a smaller set of 984,093 highly curated records (http://hgvbase.cgb.ki.se/). They have their own secondary accession/ID number and these can be queried and retrieved from the Ensembl genome annotation. Chapter 3 presents detailed examination of the major databases of genetic variation.

2.11.5 Protein Databases and Websites

A website of central importance in protein analysis is the Expert Protein Analysis System (ExPAS; http://www.expasy.ch/). In addition to protein analysis tools, such as PROSITE (http://www.expasy.ch/prosite/) and Swiss-3Image (http://www.expasy.ch/sw3d/) Swiss-Prot protein database contains high-quality annotation and web-linked cross-references to 60 other databases. It is accompanied by TrEMBL, a computer-annotated supplement that contains the translations of all coding sequences present in primary nucleotide sequence databases not yet in SwissProt. Sequence records are merged where possible to minimize the redundancy. Sequence conflicts and splice variants are indicated in the feature table of the corresponding entry. The combined database is referred to as SwissProt/TrEMBL (SPTR). Amongst the links in SPTR it is worth mentioning the InterPro system which is of very high utility for finding protein family-specific domain matches (Apweilwer *et al*., 2000). Acquiring this information is one of the main goals of the bioinformatic analysis of proteins so it is useful to find that this piece of the work is already done and updated with new releases of InterPro. Other major sites provide PFAM, PROSITE, and other tools for protein sequence analysis. The Sanger Institute (http://www.sanger.ac.uk/) provides access and maintains PFAM and multiple other useful links and genomic tools, including three-dimensional protein structure prediction (http://genomic.sanger.ac.uk/123D/123D.shtml).

Any division between the universe of DNA and protein sequences is clearly artificial. Protein information can be accessed from within the LLRS system, just as it is also possible to link out to primary nucleic acid sequence record accession numbers from SPTR. However, the complementarity between LocusLink/RefSeq and SPTR is clear. The focus is on nucleic acid sequences in the former and protein sequences in the latter. The message for the user is that both sources will be essential for interpreting the results of genetic experiments.

2.12 CONCLUSIONS

In this chapter we have introduced the major data sources available on the internet that geneticists increasingly need to access for their research. The choice was based on our direct working experience of their utility. Rather than restrict ourselves to just cataloguing these, we have also included some discussion of the principles behind the organization of biological data, such as the concept of primary and secondary sequence databases. We have also demonstrated the power of web search engines, both of the specialist and common variety. Mastering these is essential for interrogating biological resources on the

internet. They also allow the user to search for new developments, tools and databases. This is something we strongly recommend to future-proof your own research, even if we cannot future-proof this book!

REFERENCES

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, *et al*. (2000). InterPro — an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16**: 1145–1150.

Berners-Lee T, Fischetti M, Dertouzos M. (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. Harper: San Francisco.

Butler D. (2000). Biology back issues free as publishers walk HighWire. *Nature* **404**: 117.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, *et al*. (2002). The Ensembl genome database project. *Nucleic Acids Res* **30**: 38–41.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet* **25**: 239–240.

Southan C, Cutler P, Birrell H, Connell J, Fantom KG, Sims M, *et al*. (2002). The characterization of novel secreted Ly-6 proteins from rat urine by the combined use of two-dimensional gel electrophoresis, microbore high performance liquid chromatography and expressed sequence tag data. *Proteomics* **2**: 187–196.

Schuler GD, Epstein JA, Ohkawa H, Kans JA. (1996). Entrez: molecular biology database and retrieval system. *Methods Enzymol* **266**: 141–162.

Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, *et al*. (2002). Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res* **30**: 13–16.

Zdobnov EM, Lopez R, Apweiler R, Etzold T. (2002). The EBI SRS server — recent developments. *Bioinformatics* **18**: 368–373.

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 3

Human Genetic Variation: Databases and Concepts

MICHAEL R. BARNES

GlaxoSmithKline Pharmaceuticals
Harlow, Essex, UK

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 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 tranversions 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

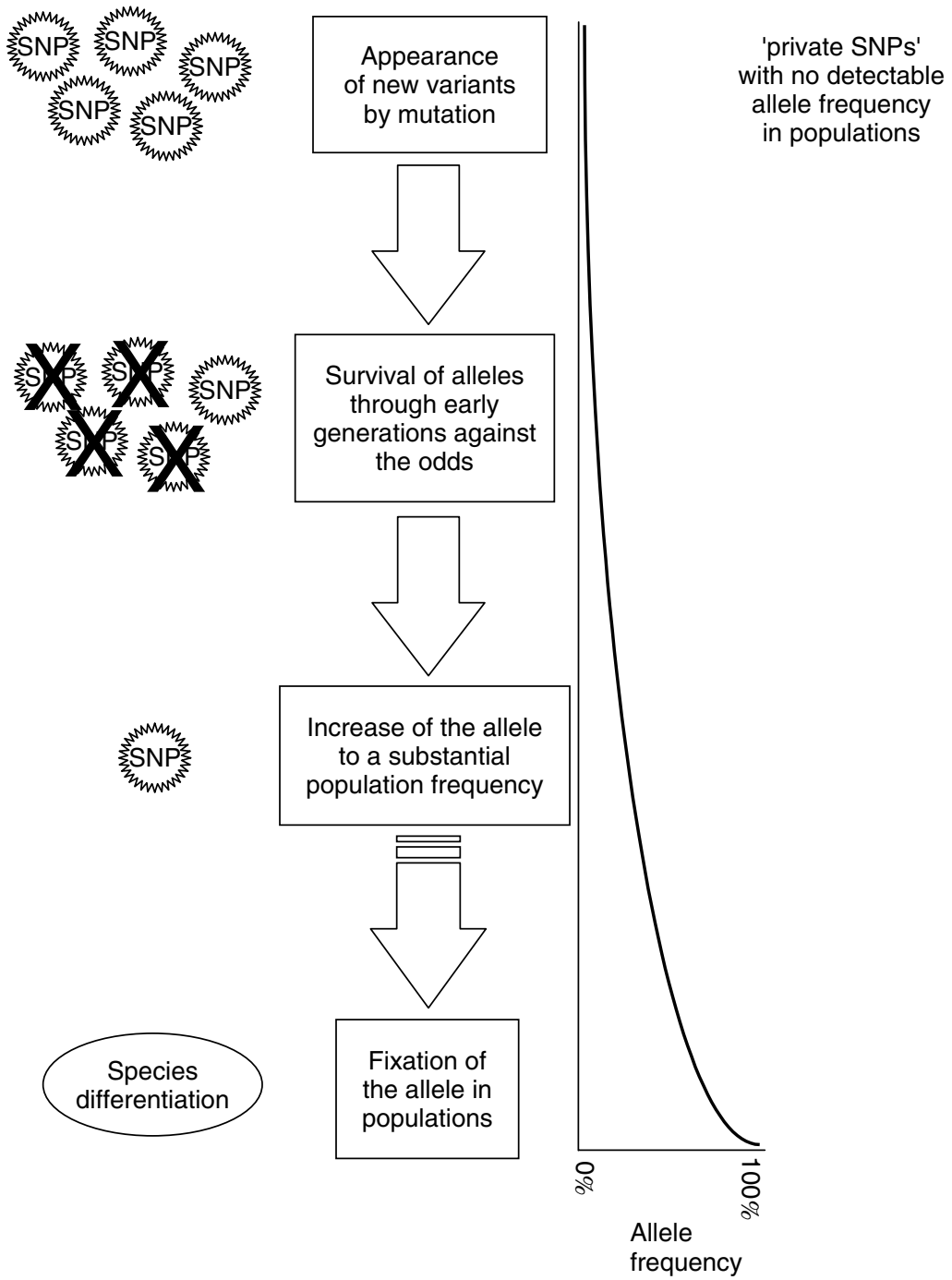

Figure 3.1 The life cycle of SNPs and mutations. SNP and mutation evolution occurs in four main phases: (1) appearance of a new variant allele by mutation; (2) survival of the allele through early generations against the odds; (3) increase of the allele to a substantial population frequency; (4) fixation 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	**A**AA**A**AAAA**A**AAA**A**AAA**A**
Dinucleotide	**CA**CA**CA**CA**CA**CA**CA**
Triplet/trinucleotide	**CAG**CAG**CAG**CAG**CAG**CAG**CAG**CAG**CAG**
Tetranucleotide	**TAAG**TAAG**TAAG**TAAG**TAAG**TAAG**TAAG**
Pentanucleotide etc.	**GAATT**GAATT**GAATT**GAATT**GAATT**GAATT
Repeat terminology	**Example**
Perfect STR	CACACACACACACACACACACACACA
Imperfect STR	CACA**T**ACACACACACACAC**G**CACACACA
Interrupted STR	CACACACACAC**GGG**CACACACACACACA
Compound STR	CACACACACACACA**TGTGTGTGTGTGTG**

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 17 Mb 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 is 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 tumourigenesis. 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 chromosome 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/snpper-enter
GeneSNPs	http://www.genome.utah.edu/genesnps/
CGAP SNP database	htpp://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

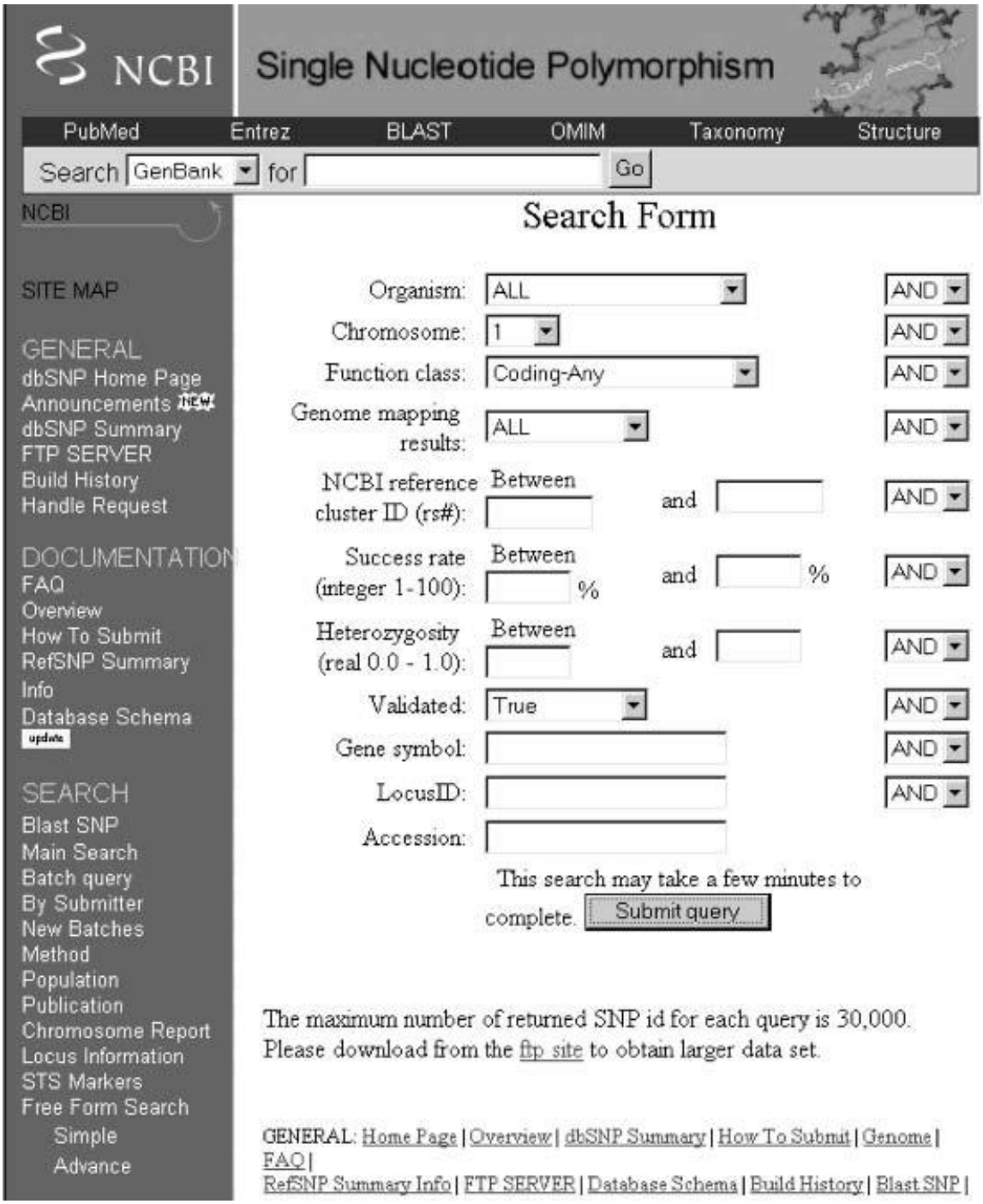

Figure 3.3 The dbSNP freeform search interface.

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.

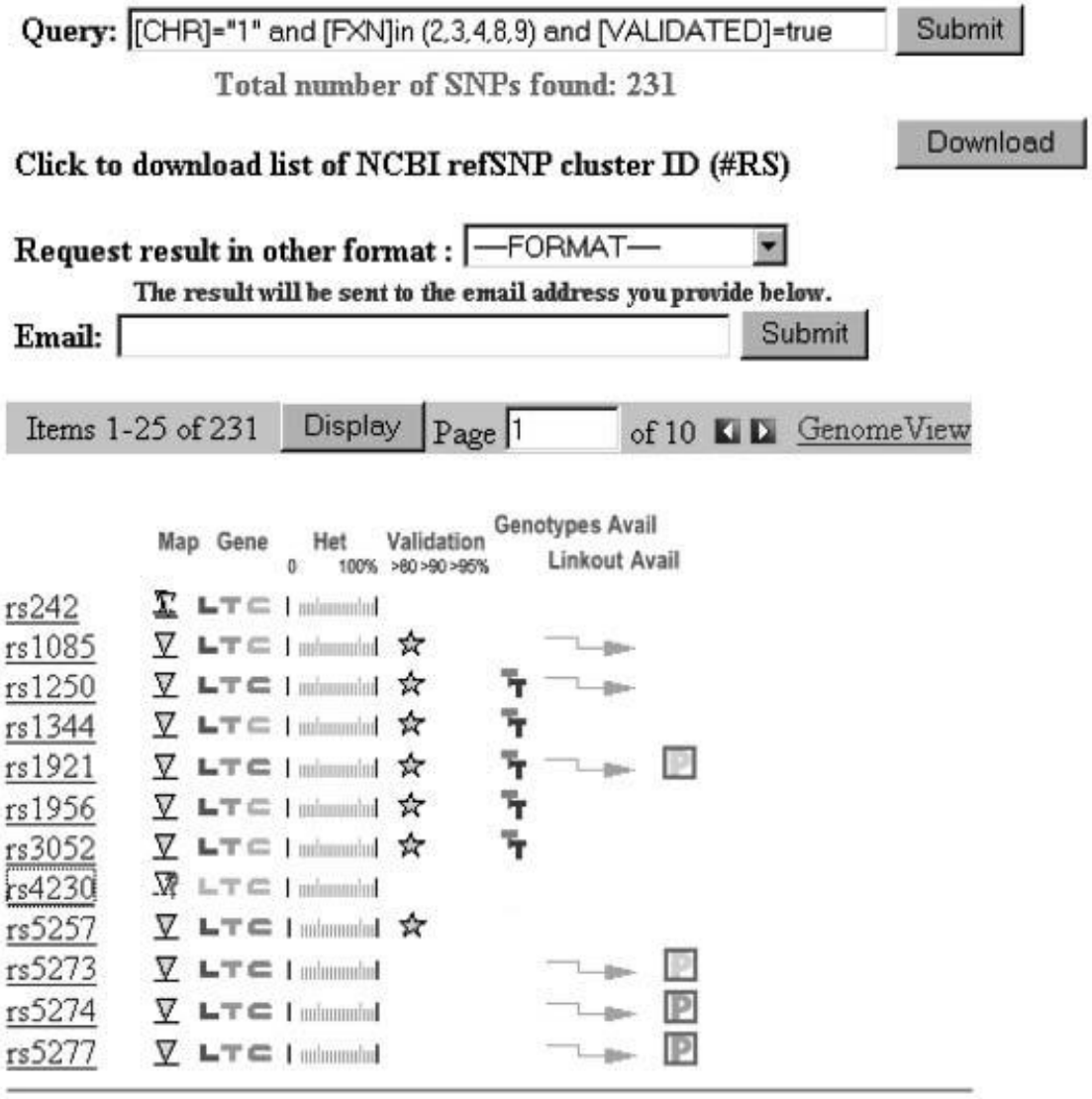

Figure 3.4 Search results from a dbSNP freeform search.

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 on 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://hgvbase.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 convertion 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 (16 S) and small (12 S) 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 abberations 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

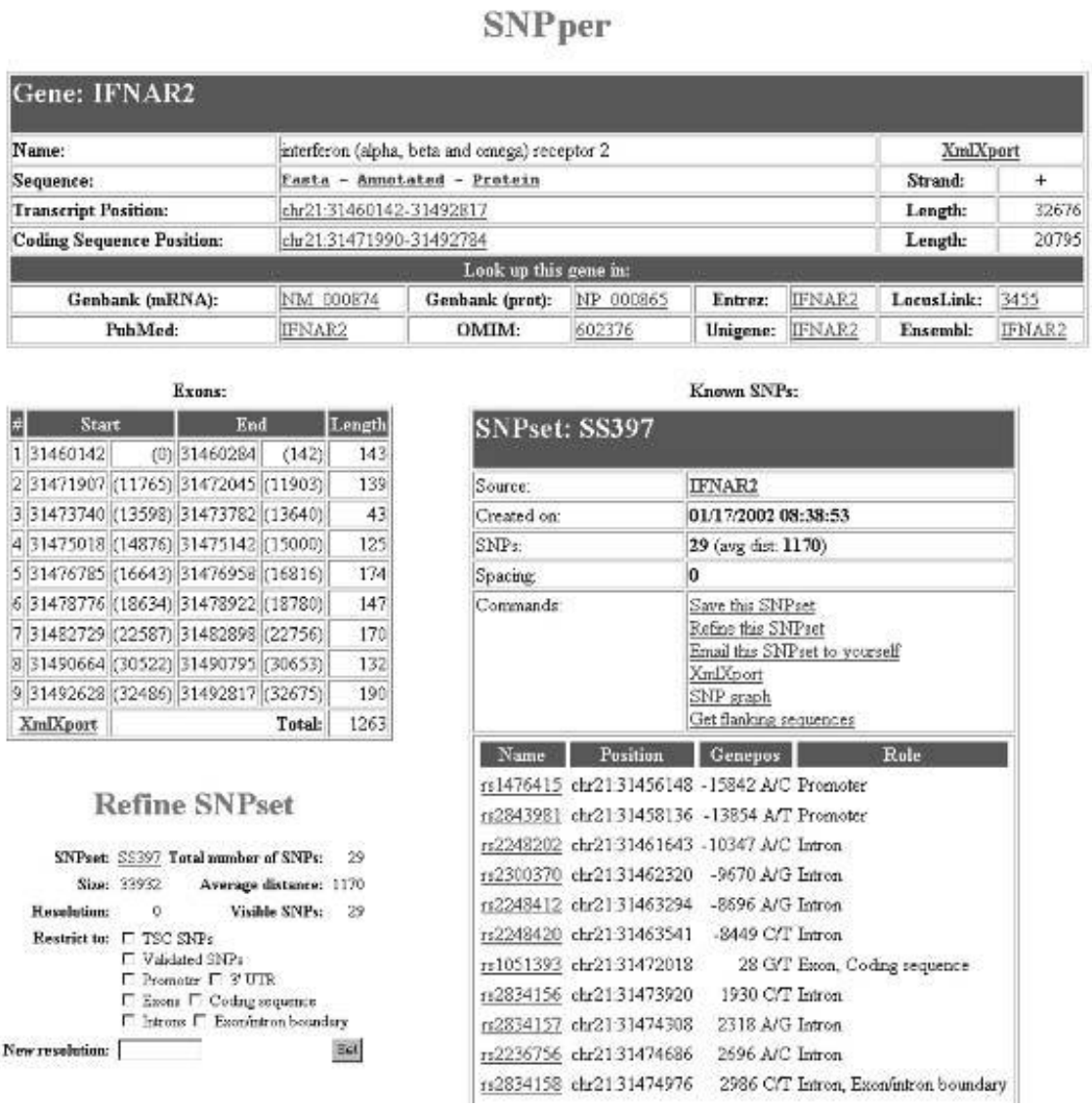

Figure 3.5 The SNPper gene report. The report displays SNP positions, alleles and the genomic sequence surrounding the SNP. It also presents text reports which mark up SNPs across the entire genomic sequence of the gene and amino acid-altering SNPs on the protein.

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 (htpp://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.

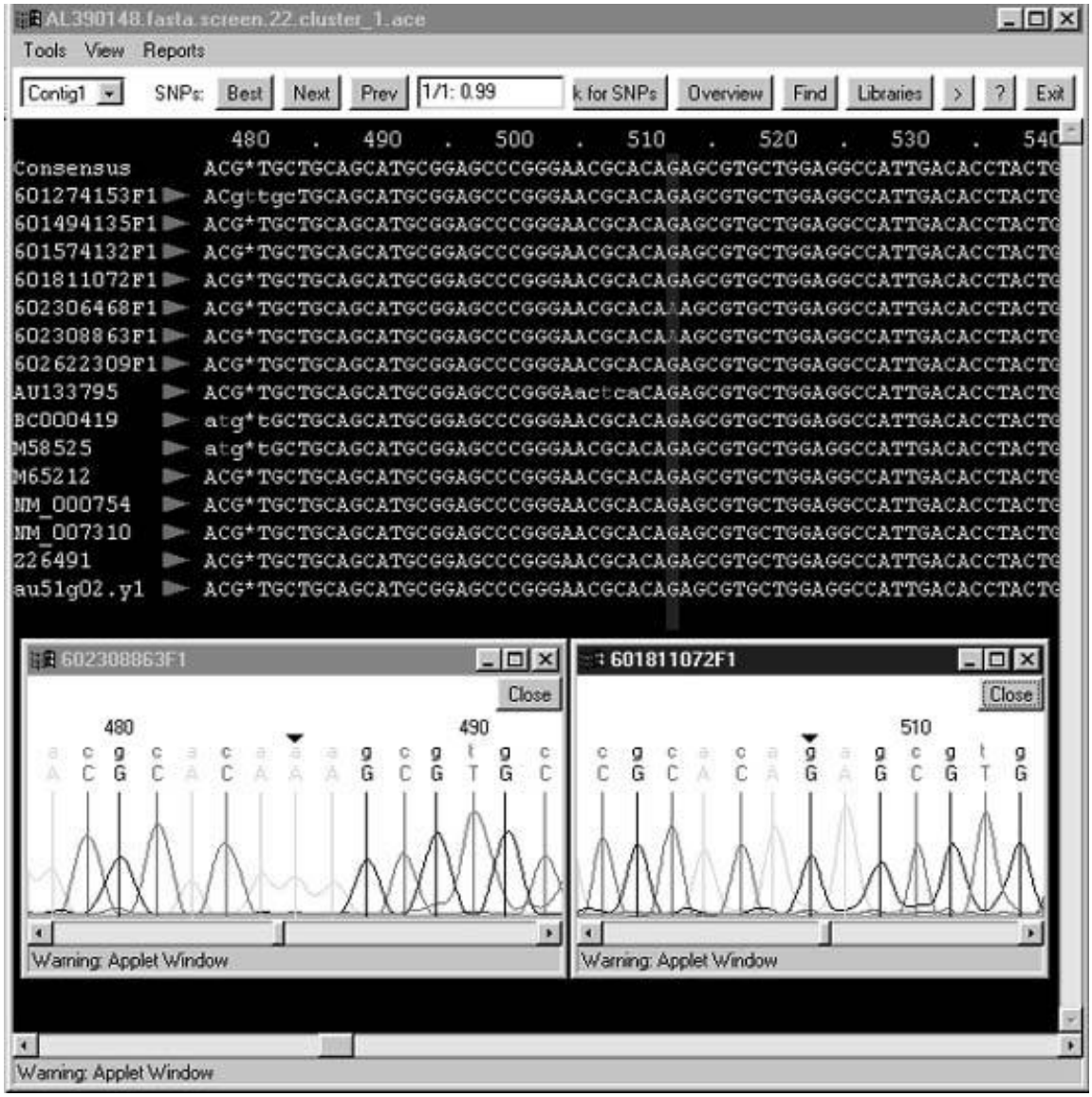

Figure 3.6 The CGAP-GAI web interface for identification of candidate SNPs in ESTs. The JAVA view of trace data helps to support the base call of a potential SNP in an EST, although laboratory investigation is the only reliable SNP confirmation.

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, Lehväslaiho 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). HGVbase: 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.

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 4

Finding, Delineating and Analysing Genes

CHRISTOPHER SOUTHAN

Oxford GlycoSciences UK Ltd
The Forum, 86 Milton Science Park
Abingdon OX14 4RY, UK

4.1 INTRODUCTION

This chapter will describe ways to interrogate human genome (HG) data with the results of genetic experiments in order to locate known genes on the current Golden Path (GP) chromosomal assemblies. It will also describe the assessment of evidence for genes that do not yet have experimental support and some analytical choices that may reveal more about them. In addition to some general aspects of gene detection some specific examples will be worked through in some detail. This illustrates technical subtleties that are not easy to capture at the overview level. As an introduction to the HG, GP and gene annotation the following chapter by Semple is recommended. Chapter 2 also provides some useful background on the organization of sequence databases. A caveat needs to be added here that many roads lead to Rome. Some particular ways of hacking through the genome jungle

are implicitly recommended by being used as the examples in this chapter. They will also be restricted to public databases and web tools. These are the personal choices of the author based on an assessment of their availability and utility. Other experts may propose alternative routes to the same information, either using different public resources, locally downloaded datasets, Unix-based tools, commercial software or subscription databases.

Genetic investigations are concerned with discerning the complex relationships between genotype and phenotype. The statement that phenotype is determined by the biochemical consequences of gene expression is equally obvious. However, the reason for making this explicit is to recommend that those performing and interpreting genetic experiments may find it more useful to conceptualize the gene as a cascade of evidence that connects DNA to a protein product rather than abstract ideas about what might constitute a gene locus. The idea of focusing on gene products also makes it easier to design experiments to verify predicted transcripts and proteins. It must also be remembered that many gene products are non-message RNA molecules but they will not be covered in this chapter. Before describing the evidence used to classify gene products it is necessary to define some of the terminology encountered in the literature and database descriptions. These are variously classified as known, unknown, hypothetical, model, predicted, virtual or novel. There are no widely accepted definitions of these terms but their usage in this chapter will be as follows. A known gene product is experimentally supported and would be expected to give close to a 100% identity match to a unique GP location. The term ‘unknown’ is typically applied to gene products that are supported experimentally but that lack any detectable homology or experimentally determined function. The term ‘predicted’, also referred to as ‘model’ or ‘hypothetical’ by the NCBI, will be reserved for an mRNA or protein ORF predicted from genomic DNA. Virtual mRNAs will refer to constructs assembled from overlapping ESTs that exceed the length of any single component. The term ‘novel’ has diminishing utility and will simply refer to a protein with no extended identity hits in the major protein databases.

4.2 THE EVIDENCE CASCADE FOR GENE PRODUCTS

So what kinds of evidence need to be considered before we assess the likelihood of a stretch of genomic DNA giving rise to a gene product and what kind of numbers can be assigned to these evidence levels? The most solid evidence of a gene is the experimental verification of the protein product by mass spectrometry and/or Edman sequencing. Although these techniques are commonly used to analyse proteins produced by heterologous expression *in-vitro* surprisingly few genes from *in-vivo* or cell line sources have been verified at this level. From the entire SP/TR collection of human proteins only 311 are cross-referenced as having at least a fragment of their primary structure identified directly from a 2D-PAGE experiment (http://ca.expasy.org/ch2d/) (Hoogland *et al*., 2000). Numerous mass spectrometry-based identifications and peptide sequences from human proteins are reported in the literature but little of this data has been formally submitted to the public databases and therefore has not been captured by SwissProt or other secondary databases. However, even this most direct of gene product verifications is rarely sufficient to confirm the entire open reading frame (ORF). For example secreted proteins are characterized by the removal of signal peptides and frequent C-terminal processing. This precludes defining the N and C translation termini by protein chemical means.

The next level down in the evidence cascade is of course an extended mRNA. There are currently 48,681 human mRNAs in GenBank. However transcript coverage is by no means

complete as they collapse down by shared identity to a set of 13,429 human transcripts (excluding splice variants) in the NCBI RefSeq collection (http://www.ncbi.nlm.nih.gov/LocusLink/RSstatistics.html) (Pruitt and Maglott, 2001). Although this collection attempts to provide a non-redundant snapshot of gene transcription it must be remembered that they are not all full-length transcripts. If the databases do not contain an extended mRNA the assembly of overlapping and/or clone-end clustered ESTs can be considered as a virtual mRNA (Schuler, 1997). The ESTs have the additional utility that many of them can be ordered as clones. Alternatively, the virtual consensus sequence, backed up by comparisons to the genomic DNA, can be used for PCR cloning. The fact that 94% of known mRNAs are covered by at least one EST makes them strong supporting evidence for a transcript, especially if they include a plausible splice junction and are derived from multiple clones from different tissue cDNA libraries (http://www.ncbi.nlm.nih.gov/UniGene/). The TIGR gene indexes are a useful source of pre-assembled virtual sequences that they term tentative human consensus sequences or THCs (Quackenbush *et al*., 2001). These can also be selected in the UCSC genome display. The use of unspliced ESTs as evidence for a transcribed gene is unreliable as they can arise from genomic contamination. However human EST-to-genome matches for exon detection can be further supported where orthologous ESTs from other vertebrates, such as mouse or rat, match uniquely in the same section of GP. If an assembly of mouse ESTs is consistent with a human gene model then the existence of an orthologous human transcript is strongly implicated.

The protein databases occupy the centre of the evidence cascade for gene products. Those mRNAs that translate to an open reading frame (ORF) are experimentally supported even if they are not full-length and/or there can be ambiguity about the choice of potential initiating methionines. However, the fact that the protein databases have now expanded to include human ORFs derived solely from genomic predictions (described in the next section) means that the evidence supporting them as gene products becomes circular. The highest curation level is provided by SwissProt sequences from the Human Proteomics Initiative set (HPI) (http://ca.expasy.org/sprot/hpi/hpi_stat.html). The March 2002 number comprised 7895 unique gene products and 2039 splice variants (O'Donovan *et al*., 2001). The SwissProt/TrEMBL (SP/TR) total for human proteins in February 2002 was 24,147, including splice variants (http://www.ebi.ac.uk/proteome/HUMAN/interpro/stat.html). The current Ensembl release, 4.28.1, contains 21,619 proteins classified as 'knowns' by an identity above 95% to a human SP/TR entry (Hubbard *et al*., 2002). The International Protein Index (IPI) maintains a database of cross references between the data sources SwissProt, TrEMBL, RefSeq and Ensembl. This provides a minimally redundant yet maximally complete set of human proteins with one sequence per transcript (http://www.ebi.ac.uk/IPI/IPIhelp.html). The March 2002 release contains 65,082 protein sequences but this includes 28,350 XP RefSeq ORFs predicted by the NCBI which are not supported by mRNAs.

The next level of evidence can be classified as genomic prediction i.e. where a cDNA, a translated ORF and a plausible gene splice pattern can be predicted from a stretch of genomic DNA (Burge and Karlin, 1997). This proceeds more effectively on finished sequence or at least where unfinished sequence contains the exons in the correct order. This is done after filtration of repeats which can be considered as another link in the evidence chain. A very high local repeat density certainly suggests where exons are unlikely but the converse is not true i.e. the absence of repeats does not prove the presence of genes. The shortcomings of *ab initio* gene prediction have been pointed out but the geneticist should at least be aware of possible false positives and false negatives (Guigo *et al*., 2000). The Ensembl statistics of the ratio of genes

predicted by Genscan over genes with a high evidence-supported threshold is currently 7.5 : 1 (http://www.ensembl.org/Homo_sapiens/stats/). Although this clearly represents over-prediction some may be 'genes-in-waiting' which more accumulated evidence may verify, for example by the cloning of an extended mRNA. Looking for a consensus or at least common exons from a number of gene prediction programs with different underlying gene model assumptions can strengthen this type of evidence but this can become a circular argument where the programs are both trained and benchmarked with known genes. For unfinished genomic sequence the presence of gaps and local miss-ordering of contigs within the clone degrades the performance of *ab initio* methods. The most effective way of filtering down genomic predictions without experimental evidence is homology support i.e. the predicted protein shows extended similarity with other proteins. This is described in detail in the Ensembl documentation but in essence all possible protein similarity sections from translated DNA are identified and used to build homology-supported gene predictions using GeneWise (Birney and Durbin, 2000). The advantage of gene detection by homology is that the entirety of protein sequence space can be used. The caveat is that predicted gene products with low similarity to extant proteins would be discarded in this filter, although the entire set of Genscan predictions are preserved for searching in Ensembl and can also be displayed at UCSC.

The next link in the evidence chain is a special case of the similarity principle but in this case utilizing comparisons between the genomes of other vertebrates such as mouse and fish for which extended data are now available (Wiehe *et al.*, 2001). Mouse genome assemblies have recently appeared on the Ensembl and UCSC sites. Although the initial assembly is only □20% the total depth in the trace archives and HTGS divisions is approaching complete coverage. Cross-species data can be assessed at three levels. The first is a simple DNA similarity on pieces of mouse DNA known to be syntenic from the location of known mouse genes and/or the extended similarity score which, with appropriate masking, locates it uniquely to a human locus. This approach is termed phylogenetic footprinting (Susens and Borgmeyer, 2001). The problem for gene product detection is that this is too sensitive i.e. mouse/human syntenic regions have many conserved similarity 'patches' outside the boundaries of known exons. They are likely to be important for functions not yet understood but are difficult to discriminate from potential coding regions. The second level is mouse BLAT as used on the UCSC site. This goes a step back by doing a translation similarity comparison rather than direct DNA-to-DNA. This makes it more likely to pick up reading frame similarities across exons. The third level is the so-called exofish. By the detection of translation similarities at the amino acid level this is capable of detecting those exons that are conserved between human and fish. This will be more useful when exofish updates to a complete fish genome rather than a partially assembled one.

The last link in the evidence chain, the *in silico* recognition of transcriptional control regions, is circumstantial but is likely to increase in utility (Kel-Margoulis *et al.*, 2002). These could include potential start sites in proximity to CpG islands, promoter elements, transcription factor binding sites, and potential polyadenylation acceptor sites in 3□ UTR. When considered in isolation these signals have poor specificity but taken in combination with a consensus gene prediction and conservation of these putative control regions between human and mouse, they can become a useful part of the evidence chain.

In summary there is currently direct experimental evidence for □15,000 genes and strong evidence to support a lower gene limit of around 25,000. The confirmation rates for the types of evidence listed above has not been calibrated experimentally so we cannot come up with any kind of scoring function to rank gene likelihood. Going to the

extremities of the evidence cascade, for example with the 65,082 ORFs from the IPI or the 62,271 UniGene clusters containing at least two ESTs, would result in a higher upper limit. This uncertainty becomes a key issue for genetic experiments. Let us suppose, for example, that a linkage study has defined a trait within the genomic region bounded by two microsatellite markers. If the lower limit gene number is true then the investigator merely needs to check the annotations from any of the three gene portals to produce a list of gene products between the positioned markers from which to choose candidates for further work. If the upper limit is true this approach has a major limitation because many of the genes between the markers will not be annotated. However, the different levels of gene evidence described above can be visualized in the display tracks of the genome viewers. Consideration of the evidence will enable the geneticist to decide what experiments need to be designed to confirm potential novel gene products. An example of working through this evidence is given in the examples below.

4.3 SHORTCOMINGS OF THE STANDARD GENE MODEL

One of the conclusions that could be drawn from the draft human genome sequence was that the standard gene model of a defined gene locus → a single mRNA species → a single protein, is no longer adequate to describe the increasingly complex relationship between the genome and its products. Attempts to fit transcript data into the standard gene model highlight a number of 'grey' areas. The first of these is delineating the extreme 5′ and 3′ ends of the mRNA transcripts (Pesole *et al*., 2002; Suzuki *et al*., 2002). The fact that many mRNAs are labelled as partial is testimony to the difficulty of finding library inserts that are complete at the 5′ end. In many cases the mRNAs are considered finished when a plausible ORF has been delineated. However, very few cDNAs are full-length in that they have been 'walked out' to determine the true 5′-most initiation of transcription in the 5′ UTR. The same problem applies to the UTR at the 3′ end. There may be substantial stretches of 3′ UTR extending downstream of the first polyadenylation position at which further cloning attempts have ceased. The problem is compounded by the poor performance of gene prediction programs for 5′ and 3′ ends. The first step towards resolving uncertainties about transcript extremities, is to survey the coverage of all available cDNA sequences, whether nominally full-length or partial, ESTs and patent sequences. These can often extend the UTR sections. The second grey area concerns pseudogenes. In some cases genomic sequence is so severely degraded that transcription is unlikely. However, from the current pseudogene listing in RefSeq of 1598 loci, at least 30 are recorded as having detectable transcripts (http://www.ncbi.nlm.nih.gov/LocusLink/statistics.html). The third grey area is gene product heterogeneity. In some cases there may be alternative upstream initiation methionines or alternatively spliced exons in the 5′ UTR. The causes for 3′ heterogeneity include variations in the pattern of intron splicing from a pre-mRNA, as well as alternative poladenylation positions inside the 3 UTR. The fourth grey area concerns overlapping genes. As genomic annotation proceeds we can find more examples of this both from gene products reading from opposite strands and same-strand genes in close proximity.

Considering these grey areas as a whole, they can all be seen as deviations from the simple gene model. Many individual examples of such complexities had been documented before the genome draft of May 2001. However, it is only since then that assessments of their overall incidence could be made, most recently for completed chromosomes such as 20 (Deloukas *et al*., 2001). It is therefore essential for the geneticist to keep an open mind about the extremities and plurality of gene products.

4.4 LOCATING KNOWN GENES ON THE GOLDEN PATH

Genes can be located by one of the following: a section of raw sequence data, a primary accession number, a secondary accession number, a similarity search, a gene product name, or a set of Golden Path (GP) coordinates. Each of these has advantages and disadvantages and, although the three gene portals are generally consistent, they may not give the same answers in every case. Bearing in mind that only the first two of these are stable and (almost) free of potential ambiguity it is better to use at least two ways to define and store the results, for example a section of raw sequence and a gene name, or a primary accession number and a set of GP coordinates. The BACE gene will be used as an example of a known gene to locate. The potential complexity of this task is illustrated by the example of the Ensembl gene report for BACE that includes no less than 46 separate terms (Figure 4.1).

4.4.1 Raw Sequence Data

The availability of GP means that most features can now be unambiguously located in the genome with as little as 100 bp. This means that storing a sequence string, preferably with a longer sequence context of 200–1000 bp, is a useful method of locking-on to a genomic location. It is also immune to the vagaries of shifting secondary accession numbers, naming ambiguities or GP sequence finishing that can change the genomic coordinates. Performing nucleotide searches against GP using tools such as BLAT (UCSC) or SAHA (Ensembl) or BLAST (NCBI), means that sequence matches can be quickly located. The disadvantage for raw sequence is that it has to be stored in its entirety, it may contain errors, it needs the operation of a similarity search to be located and similarity matches across repeat containing sections or duplicated regions of the genome need close inspection to sort out. This can be a particular problem for STSs and SNPs

Figure 4.1 The Ensembl gene report page for BACE (release 4.28.1).

if the GP match is in the region of 98 to 95% identity. Within this range it is difficult to discriminate technical sequencing errors from multiple genomic locations or assembly duplication errors. It can also be useful to search the primary genomic data, especially if GP is not complete in that section. For example although BACE is linked by Ensembl to AP001822 as the finished GP sequence, a database search reveals another four matching primary genomic accession numbers from chromosome 11, AP000892 (finished at version 4) with AC020997, AP000685 and AP000761 still unfinished. One less obvious advantage of these five overlapping genomic contigs is that if they proceed to finishing more SNP positions may be revealed. As described below the genome portals capture mRNA entries for most gene products unless they are very recent. However, because of the thin annotation they do not capture sequences from the patent divisions. A BLAST search of gbPAT with any BACE mRNA gives 18 high-identity DNA matches. These are clearly mRNAs that could be usefully compared with all other mRNA sequences for polymorphisms, splice variants or UTR differences. However users should be aware that not only are some of these 18 entries identical versions of the same sequence derived from multiple claims in the patent documents but they may also be identical to a public accession number if the authors and inventors are from the same institution. Another reason for using raw sequence data for gene product checking is because all secondary databases suffer from the snapshot effect where updates lag behind the content of the primary databases. For example the SNP or EST assignments made for BACE in the secondary databases (see below) could be checked by BLAST searches against the updates of dbSNP or dbEST (remember the latest EST data needs to be searched in 'month' as well as dbEST).

4.4.2 Primary Accession Numbers

Because these uniquely define stretches of sequence they are stable except where genomic and occasionally mRNAs, undergo version changes. They can be used in any of the major genome query portals to go directly to a genomic location. The disadvantage is redundancy for mRNAs, short sequence context for some STSs, both redundancy and large multi-gene sequence tracts for genomic mRNA, and very recent accessions may not be indexed in genome builds. If the query fails to connect to a genome feature the sequences can be searched as raw sequence. Taking the BACE example there are eight mRNA accession numbers listed in Figure 4.1 that can be used as a genome portal query. Interrogating UCSC with BACE retrieves nine mRNA entries, LocusLink connects directly to only three but the UniGene cluster Hs.49349 connects to 12. Users need to be aware that although an mRNA accession number can provide a specific route into GP the variable number of links to the genome portals is related to their update frequency.

4.4.3 Secondary Accession Numbers

From Figure 4.1 we can read eight secondary accession numbers that designate protein translations for each of the BACE mRNAs. It also has three RefSeq numbers NM_012104 for the mRNA, NP_036236 for the protein and NT_009151 for the genomic contig. There is one SwissProt accession BACE_HUMAN (P56817) and one TREMBL splice variant Q9BYB9. The LocusID, 23621, in turn links out to many other accession numbers which point to the BACE genome sequence. These include the Hs.49349 UniGene cluster that includes 336 ESTs with primary accession numbers. Via the LocusLink Variation link

the RefSNP numbers can be located. In this case they consist of 43 intronic SNPs, three within the mRNA, including one (rs539765) which causes an Arg > Cys exchange, and seven SNPs in the 3′ UTR. It is possible to use a RefSNP (rs) number to go directly to the SNP location in Ensembl or UCSC. However because of multiple GP matches in Ensembl it is necessary to know the genomic location beforehand.

4.4.4 Gene Names

Including abbreviations Figure 4.1 there are nine synonyms or aliases for this enzyme. This illustrates the problem where gene products are given different names by different authors. The best way to cross-check names, spelling variations and frequency of use, is to search PubMed. Checking title lines only is more specific but does not capture all occurrences. In this case a title search found a new name extension, BACE1, with five citations compared with 22 for BACE. This seems logical since the discovery of the BACE2 paralogue on chromosome 21. However, the Human Gene Nomenclature Committee have not been consistent because they have only listed BACE and BACE2 as official symbols even though they have listed ACE1 as an alias for ACE since the recent discovery of ACE2 (http://www.gene.ucl.ac.uk/nomenclature/). The most frequent specific term was 'beta-secretase precursor' at 30 citations. The alternative 'membrane-associated aspartic protease 2′ gave eight citations and 'beta-site app cleaving enzyme' was the least frequent with only two. Paradoxically this has been chosen for the LocusLink name. The least specific name was aspartylprotease 2 with two false positives and ASP 2 with 143 title matches, also mostly false positives. The imprecision of name searching was reinforced by checking ASP-2 with three matches and ASP2 with five. Only one was a true positive and two of the citations referred to ASP2 as an odorant-binding protein from the honeybee. The complexity of the aliases for just one gene product makes it clear that any gene name lists, for example as candidate genes to be screened for mutations, must be backed-up by accession numbers and/or raw sequence. It also illustrates the need to cross-check aliases and their spellings when attempting a comprehensive literature search on a particular gene product. The formal sequence-literature links that can be followed in Entrez, LocusLink or SwissProt are not comprehensive because they are dependent on the journal–author–database system that usually only makes these links explicit for a new accession number. Much important literature remains outside this system. Review articles, for example, do not typically include primary accession numbers when describing genes so the specificity of literature searches remains dependent on the name links. Information trawling with gene names can also be done with the standard internet search portal. Putting the term 'beta-site app cleaving enzyme' into the Google search engine gave 249 hits (http://www.google.com/). The listing included duplicates but very few false positives.

4.4.5 Genome Coordinates

Since the adoption of a unified GP assembly this method of genomic location has become more reliable but users are advised to check the synchronization of new GP versions between the three portals. Users should refer to the individual portals for the details of using these coordinates but for the BACE example the NCBI showed a region described as 120,533K–120,594K, the Ensembl viewer specified the coordinates as 120549397–120575715 bp (with a zoom setting 120.5 Mb) on chromosome 11 and the UCSC viewer designated the position in the form chr11 : 120545299–120599798.

4.5 GENE PORTAL INSPECTION

From the descriptions above it should be possible to locate any known gene or genetic marker such as an STS or a SNP. Descriptions of the genome viewer features for Ensembl, UCSC and NCBI are included in the chapter by Semple. However two examples are included below (Figures 4.2 and 4.3) because they illustrate technical differences and highlight the deviations from the standard gene model. The UCSC display (Figure 4.2) includes 12 mRNA sequences for BACE where Ensembl (Figure 4.1) has included accession number links for only eight. The display in Figure 4.2 also shows there are significant differences in the lengths of the 5′ and 3′ ends. Clearly AF201468 (5878 bp) and AB032975 (5814 bp) are the longest reads but in fact AB032975 is labelled as a partial CDS because of what may be a sequencing error at the 5′ end. The matches to the spliced ESTs together with the rat and mouse mRNAs suggest the 5′ UTR may be full-length for these entries i.e. they extend to the start of transcription. This is in contrast to the shorter 5′ ends for the majority of mRNAs. A detailed analysis of the 3′ ends by EST distribution profiles indicates that the different UTR lengths in this case arise not from incomplete cloning but from three alternative polyadenylation positions (Southan, 2001). Further heterogeneity is illustrated by three splice variants affecting exons 3 and 4. The representative mRNAs are AB050436, AB050437 and AB050438. There is also an alternative protein reading frame from AF161367, a partial mRNA cloned from CD34+ stem cells. Opening up the spliced EST tracks in the viewer shows individual ESTs corresponding to these splice forms. Approximately midway between exons 1 and 2 (from the 5′ end) is a spliced EST, AL544727, derived from spleen. This suggests the possibility of another splice form but this would need analysis for canonical splice sites and experimental verification. Similarly an EST from spinal cord AL589586 suggests an alternative exon just on the 5′ side of exon 3. Although the rat and mouse mRNAs displayed in Figure 4.2 show the same exon positions as most human sequences there are suggestions of splice variants in non-human ESTs but these tracks were not expandable in the version tested.

The NCBI display for BACE mRNAs and ESTs (Figure 4.3) shows concordance and discrepancies with the UCSC display (Figure 4.2). The exon positions are identical. They include the same RefSeq mRNA and genomic secondary accession numbers. The EST matches are in broad agreement towards the 3′ end but two additional potential exon matches are indicated at the 5′ end. Although these may be unspliced matches that would need further investigation, one of these coincides with the XM_084660 reference sequence

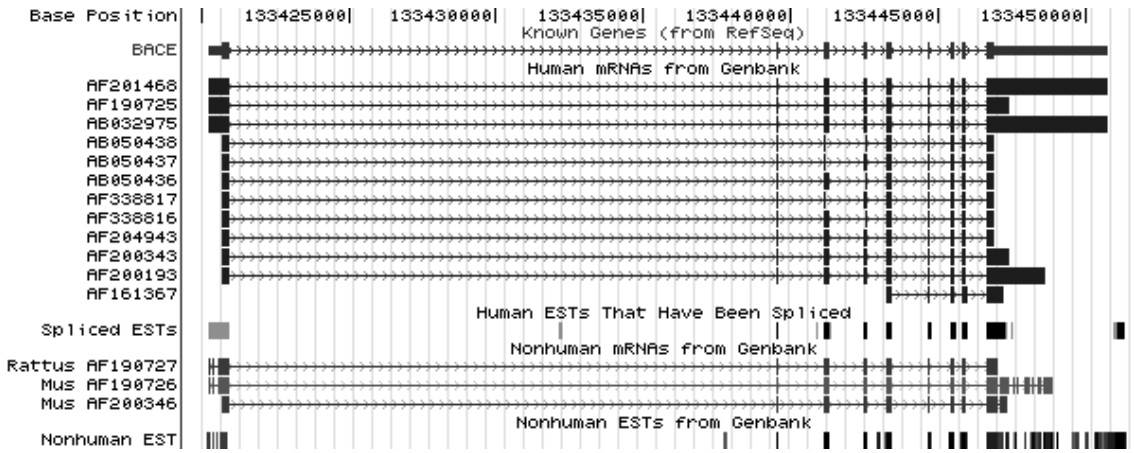

Figure 4.2 The UCSC display for BACE mRNAs and ESTs.

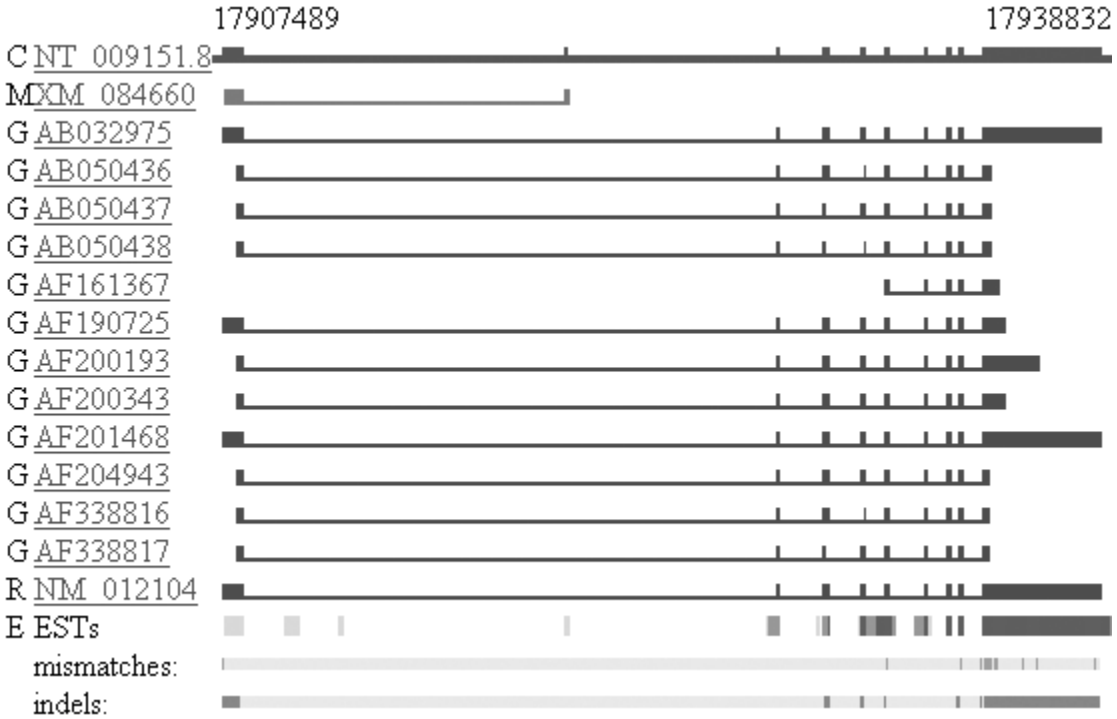

Figure 4.3 The NCBI display for BACE mRNAs and ESTs.

predicted by NCBI from the contig NT_009151. There is no mRNA verification for this prediction so it will be of interest to see if additional EST data will appear and, if not, how long this prediction will be maintained as genome annotation. The mismatches and INDEL tracks are a useful feature unique to NCBI. The mismatches within the set of 12 mRNA sequences could represent SNPs or technical sequence errors. The INDELs also show major length discrepancies. In Figure 4.2 these highlight the three splice positions in agreement with UCSC but the INDEL in exon 8 could not be interpreted from the link provided.

4.6 LOCATING GENES WHICH ARE NOT PRESENT IN THE GOLDEN PATH

Estimates suggest the GP is still missing □2.5% of the genome, there are still small gaps in the unfinished sections and the latest Ensembl release locates only 92% of known proteins (http://www.ensembl.org/Dev/Lists/announce/msg00070.html). This means that some genetic markers in close proximity to genes are either not covered by GP or are not fully annotated in unfinished sequence. Two human proteins that have no matches on the current GP version 28 from December 2001 illustrate this problem. The first of these, spP83110 serine protease HTRA3, has an mRNA entry AY040094. The second protein spP83105 serine protease HTRA4 has an mRNA accession but the entire ORF is covered by two long EST reads AL545759 and AL576444. Because it has a full length mRNA HTRA3 has a LocusLink ID of 94031 but no mapping links. Searching HTRA3 by BLASTN against the NCBI nr nucleotide database, containing 1,184,532 sequences, hits only the probable mouse orthologous mRNA, AY037300, at 86% identity within the reading frame. However checking monthly updates at 811,100 sequences reveals a 99% identity to a new genomic entry AC113611 of 190,038 bp from chromosome 4. This sequence was also in the unfinished High Throughput Genomic Sequences (HTGS)

division, with 47,855 sequences, along with the probable rat orthologous genomic section, AC110369, at 87% identity. There were no mouse genome matches from this search. A check on the nucleotide patent databases, with 582,838 sequences, showed a new mRNA match, AX338509 from patent WO0183775. The HTRA3 mRNA has EST matches to UniGene cluster Hs.60440 with four STSs from chromosome 4. Presumably these STSs will be located on GP when the AC113611 genomic sequence is assembled into chromosome 4. Checking the chromosome 4 SNPs at 105,568 sequences by BLAST search, recorded no hits within the 2552 bp mRNA of AY040094 but found over 100 matches within the repeat-masked sections of AC113611. Using the same sequence to BLAST against the 115,608 sequences in the STS division gives eight hits above 95% identity, although only three looked like unique matches. Interestingly the HTRA3 mRNA AY040094 has no STS matches although four chromosome 4 STSs were picked up in the UniGene entry. A possible explanation is that the cluster included clone links to ESTs that extend past the 3′ end of the mRNA.

Performing the same database checks for the HTRA4 ESTs, AL545759 and AL576444, produces a different pattern of findings. There were no hits in nr or gbPAT. However, the HTGS search located extended identity hits to no less than four genomic entries. These comprised of three recently sequenced sections of chromosome 8 AC108863, AC105089, AC105088, and a short match to an entry without a chromosomal assignment, AC107926. Checking for HTRA4 in LocusLink could find no IDs because of the absence of a full-length mRNA. It was picked up as the UniGene cluster Hs.322452 with nine ESTs but no mapping information was included even though our search update had located it to chromosome 8. No reading frame SNPs could be detected from the 92,110 chromosome 8 entries. By using the genomic contig, AC108863, (198,743 bp) as a BLAST query only three SNP identity matches were detected, rs1467190, rs2010445 and rs2056170, but three STS markers G60989, G23343, and G04735, were located.

In summary; although these two gene products cannot be located on the latest GP a series of manual database checks have established a mixture of patent mRNAs, unfinished genomic matches, ESTs, STSs and SNPs. It will be interesting to track how soon these features find their way into the GP annotation pipelines. If genetic studies should need this location data in the interim, the searches have established that HTRA3 probably has enough SNPs in the genomic vicinity for association studies, but that there is a very low SNP density in proximity to HTRA4. If the overlapping genomic coverage for HTRA4 could detect all the exons it might be possible to assemble a ‘mini golden path’ across this particular section. However if it became necessary to re-order and re-assemble the contigs within the unfinished entries this would be a challenging task to perform with web-based tools.

4.7 ANALYSING A NOVEL GENE

Sooner or later experimental results will locate a piece of GP where there are no fully annotated known genes. Figures 4.4, 4.5 and 4.6 show selected tracks from the Ensembl, UCSC and NCBI displays between the 3′ side of the BACE gene and the 5′ end of the next known gene PCSK7. The known genes are marked in brown in Ensembl and blue in UCSC. The latter are mRNA mappings and therefore include the UTR sections. Let us assume a genetic linkage study had found significant associations in this area, either from the two STS markers or the 50 or so SNPs that lie in this interval but are outside the boundaries of the two neighbouring genes. The question immediately arises as to what

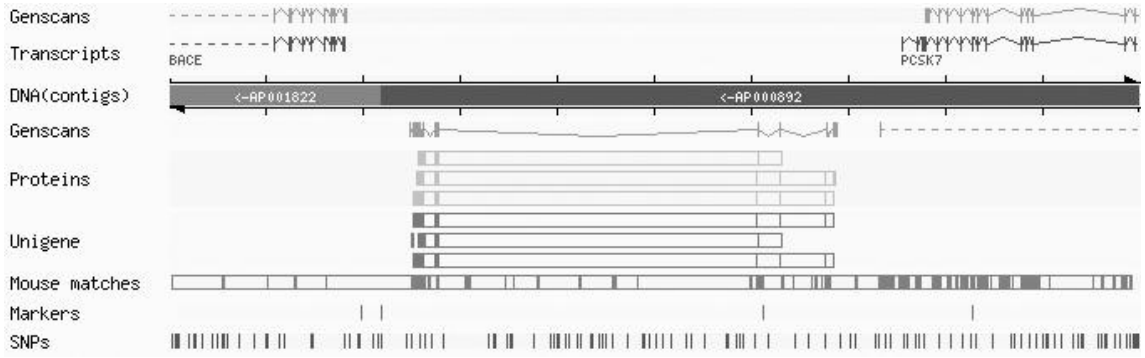

Figure 4.4 The Ensembl display for the unknown gene between BACE (left) and PCSK7 (right).

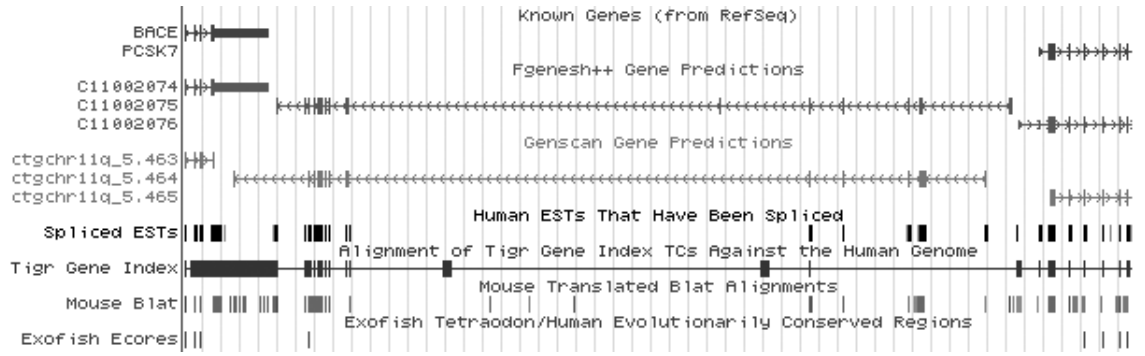

Figure 4.5 The UCSC display for the unknown gene between BACE (left) and PCSK7 (right).

other gene product(s) might be located between the two knowns. The first step is to check the continuity of this section of GP. This can be done in any of the viewers and in this case there is complete clone overlap across this section.

Inspection of all three displays indicates a possible novel gene product with a variety of supporting evidence. They include gene predictions which include both common and different exon positions. The UCSC Genscan prediction number 464 overlaps with the 3′ UTR of BACE making this a less plausible (but still possible) exon. Reading vertically down the Ensembl tracks first we see evidence for three protein homologies (yellow) as judged by the matches in register with the Genscan exon predictions. These are Q96RS9, a novel DZIP3, Q02455 a myosin-like peptide from yeast and P53804 a tetratricopeptide repeat protein. There is the same pattern of exon matches to three UniGene cluster entries (red) Mm.3679 *Mus musculus* for the tetratricopeptide repeat domain protein, Hs.165662 for *Homo sapiens* KIAA0675 unknown protein and Hs.118174 for *Homo sapiens* TTC3 tetratricopeptide repeat domain 3. There is a denser pattern of matches to mouse DNA (pink) that includes many sections outside the Genscan predicted exons.

Moving down the UCSC tracks in Figure 4.5 we see the spliced ESTs (black) in register with Genscan exons. However these identity EST matches are not equivalent to the homology-based UniGene matches in Ensembl. Interestingly the internal exon predicted only by Fgenes has no spliced EST support. Exploring the EST coverage further we see that the (brown) THC tracks include an assembly that matches the predicted exons at the BACE end of the Fgenesh++ prediction. The NCBI tracks go into more detail by not only mapping UniGene cluster components directly back to putative genomic exons by

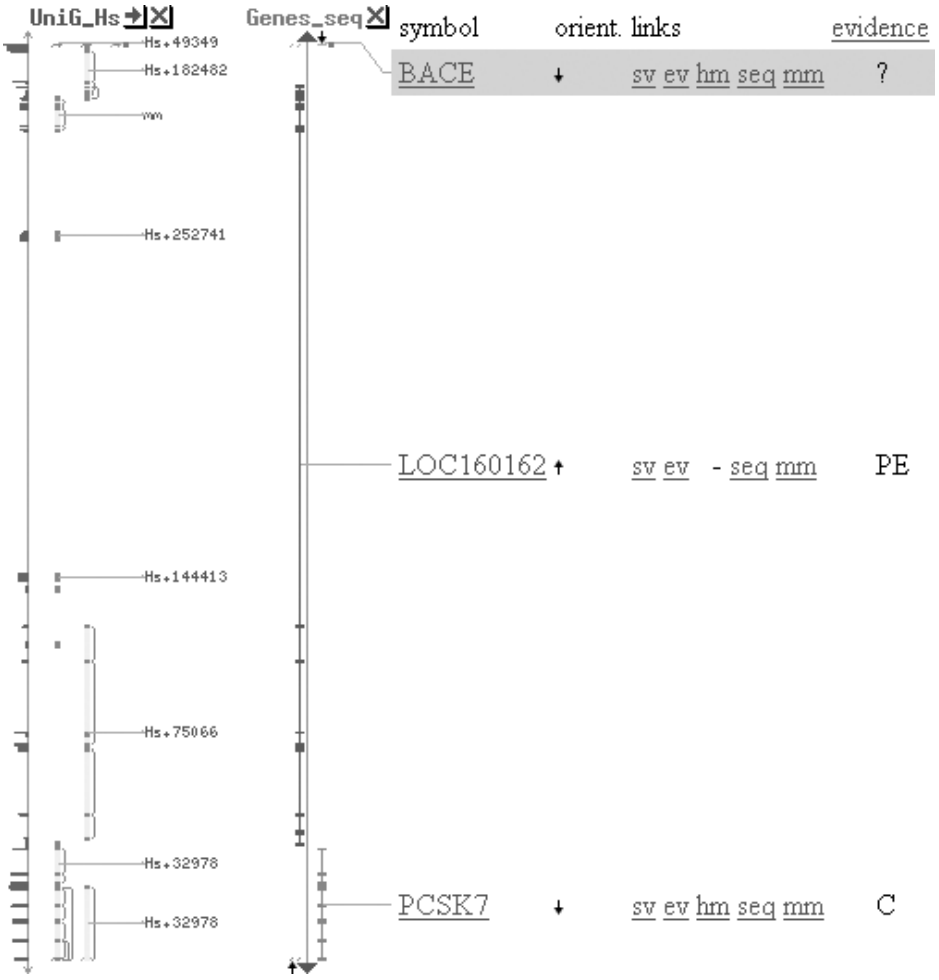

Figure 4.6 The NCBI display for the unknown gene between BACE (top) and PCSK7 (bottom). The leftmost track shows the EST distribution. The next track to the right marks the UniGene clusters. The central track is the gene prediction for LOC160162 and the gene structure for the N-terminal section of PCSK7 (bottom).

identity matches but also, on the left hand edge, showing an identity block proportional to the number of EST matches. Surprisingly there are five EST clusters which raises the possibility of more than one gene. The mouse BLAT track (brown) is equivalent to the Ensembl (pink) mouse track but the translation mode filters down to fewer features. The exofish track in UCSC (blue) supports just one single exon at the 5′ end of the putative novel gene compared with many conserved exons in both gene neighbours. In isolation this would be considered as weak evidence for the gene product. However it could simply mean that this predicted protein is not conserved between fish and human or the puffer fish ORFs are not complete across this section.

Up to this point our analysis of the genomic region between the 3′ end of BACE and the 5′ end of PKSC7 points strongly to the presence of a gene product on the basis of gene prediction and EST coverage. So where do we go from here? One option is to do some searches with the available mRNA and protein sequence from the Fgenesh++ prediction (numbered C11002075 in Figure 4.5) that can be downloaded from the UCSC site. The result brings us a long way forward in the evidence cascade because we record an 81% protein identity to what is likely to be the recently deposited mouse orthologue mRNA, BC023073. Interestingly this level of similarity should result in this gene passing

the Genwise threshold for marking a novel gene position (black) in the next release of Ensembl. At these similarity levels we can back-check this mouse sequence against human GP by the very fast BLAT search (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start). The result (Figure 4.7) clearly supports both the orientation (3′-to-3′ relative to BACE) and seven of the exons from C11002075. However the mouse sequence is clearly missing the 5′ end.

The next step involved searching the entire genomic DNA section of 54 kb from which C11002075 was predicted against human ESTs. This was performed using MEGABLAST with a 90% match stringency and masking of the repeat sections in the genomic query section. The result (Figure 4.8) is equivalent in principal to the UniGene clusters in the NCBI viewer but it is easier to pick out the ESTs that bridge several exons. Another reason for doing this analysis is that over 1 million human ESTs have been added to dbEST since the UniGene clusters were built. We can identify three ESTs that cover 35 kb of genomic sequence across three exons and performing the analogous search against mouse ESTs, with an 80% identity cut-off, finds a long EST spanning the four central exons. This gives us more confidence of a single rather than multiple gene products. The next step was to search ESTs against the TIGR THCS to establish if any virtual mRNAs could be found. In fact two of these, THC856832 and THC796698, represented the 5′ and 3′ ends respectively and to join these assemblies a bridging EST was found, BM055167. By using a web version of the CAP3 assembler (http://bio.ifom-firc.it/ASSEMBLY/assemble.html) it was possible to construct an extended virtual mRNA of 2720 bp. This was translated into a protein of 474 amino acids using the translation tool (http://ca.expasy.org/tools/dna.html) (Figure 4.9).

So far so good, but what else can we do to verify this putative novel protein *in silico*? The first step is a cross-check for reading frame consistency and species orthologues by performing TBLASTN against all ESTs (Figure 4.10). The results show the complete coverage of the entire ORF by human ESTs but also suggests potential splice variants

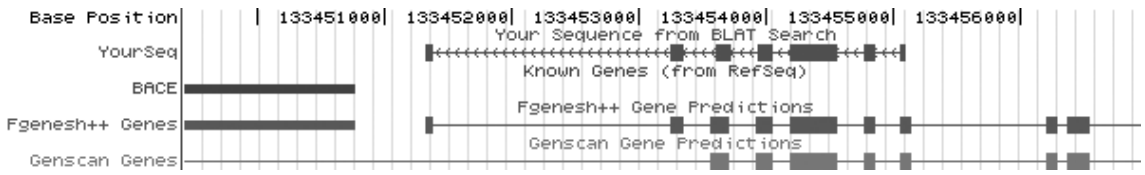

Figure 4.7 The alignment of the mouse protein from BC023073 after a BLAT search against the UCSC GP. The BACE gene is on the left hand side.

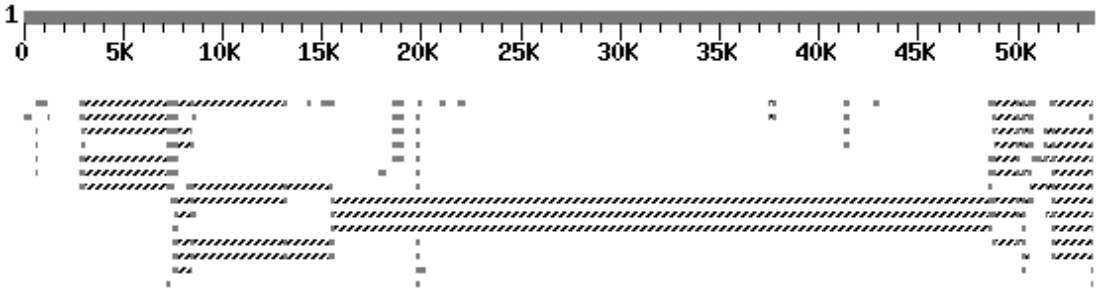

Figure 4.8 Result of a MEGABLAST search of the genomic sequence between BACE and PCSK7 against human ESTs. The solid lines indicate gaps in the same ESTs. The solid sections are putative exon matches.

```
gcgggtcctgtccctcccccactttcctcccgggggcgcggcgcgggagagcataatggc
agcgtctgaggttgctggtgttgtggccaatgcccccagtcctccggaatcttctagttt
atgtgcttccaaatcagacgaaggtctcccagatggtctaagcaccaaagactctgcaca
gaagcagaagaactcgcctctgttgagtgtaagtagccaaacaataaccaaggagaataa
cagaaatgtccatttggagcactcagagcagaatcctggttcatcagcaggtgacacctc
agcagcgcaccaggtggttttaggagaaaacttgatagccacagcccttgtctttctggc
agtgggtctcagtctgatttgaaggatgtggccagcacagcaggagaggaggggacaca
agccttcgggagagcctccatccagtcactcggtctcttaaggcagggtgccatactaag
cagcttgcctccaggaattgctctgaagagaaatccccacaaacctccatcctaaaggaa
ggtaacagggacacaagcttggatttccgacctgtagtgtctccagcaaatggggttgaa
ggagtccgagtggatcaggatgatgatcaagatagctcttccctgaagctttctcagaac
attgctgtacagactgactttaagacagctgattcagaggtaaacacagatcaagatatt
gaaaagaatttggataaaataatgacagagagaaccctgttgaaagagcgttaccaggag
                      M  T  E  R  T  L  L  K  E  R  Y  Q  E
gtcctggacaaacagaggcaagtggagaatcagctccaagtgcaattaaagcagcttcag
V  L  D  K  Q  R  Q  V  E  N  Q  L  Q  V  Q  L  K  Q  L  Q
caaaggagagaagaggaaatgaagaatcaccaggagatattaaaggctattcaggatgtg
Q  R  R  E  E  E  M  K  N  H  Q  E  I  L  K  A  I  Q  D  V
acaataaagcgggaagaaacaaagaagaagatagagaaagagaagaaggagtttttgcag
T  I  K  R  E  E  T  K  K  K  I  E  K  E  K  K  E  F  L  Q
aaggagcaggatctgaaagctgaaattgagaagctttgtgagaagggcagaagagaggtg
K  E  Q  D  L  K  A  E  I  E  K  L  C  E  K  G  R  R  E  V
tgggaaatggaactggatagactcaagaatcaggatggcgaaataaataggaacattatg
W  E  M  E  L  D  R  L  K  N  Q  D  G  E  I  N  R  N  I  M
gaagagactgaacgggcctggaaggcagagatcttatcactagagagccggaaagagtta
E  E  T  E  R  A  W  K  A  E  I  L  S  L  E  S  R  K  E  L
ctggtactgaaactagaagaagcagaaaaagaggcagaattgcaccttacttacctcaag
L  V  L  K  L  E  E  A  E  K  E  A  E  L  H  L  T  Y  L  K
tcaactcccccaacactggagacagttcgttccaaacaggagtgggagacgagactgaat
S  T  P  P  T  L  E  T  V  R  S  K  Q  E  W  E  T  R  L  N
ggagttcggataatgaaaaagaatgttcgtgaccaatttaatagtcatatccagttagtg
G  V  R  I  M  K  K  N  V  R  D  Q  F  N  S  H  I  Q  L  V
aggaacggagccaagctgagcagccttcctcaaatccctactcccactttacctccaccc
R  N  G  A  K  L  S  S  L  P  Q  I  P  T  P  T  L  P  P  P
ccatcagagacagacttcatgcttcaggtgtttcaacccagtccctctctggctcctcgg
P  S  E  T  D  F  M  L  Q  V  F  Q  P  S  P  S  L  A  P  R
atgcccttctccattgggcaggtcacaatgcccatggttatgcccagtgcagatccccgc
M  P  F  S  I  G  Q  V  T  M  P  M  V  M  P  S  A  D  P  R
tccttgtctttcccaatcctgaaccctgccctttcccagcccagccagccttcctcaccc
S  L  S  F  P  I  L  N  P  A  L  S  Q  P  S  Q  P  S  S  P
cttcctggctcccatggcagaaatagccctggcttgggttcccttgtcagcccccacggt
L  P  G  S  H  G  R  N  S  P  G  L  G  S  L  V  S  P  H  G
ccacacatgccccctgccgcctccatcccacctcccccaggcttgggcggtgttaaggct
P  H  M  P  P  A  A  S  I  P  P  P  P  G  L  G  G  V  K  A
tctgctgaaactccccggccccaaccagtagacaaactggagaagatcctggagaagctg
S  A  E  T  P  R  P  Q  P  V  D  K  L  E  K  I  L  E  K  L
ctgacccggttcccacagtgcaataaggcccagatgaccaacattcttcagcagatcaag
L  T  R  F  P  Q  C  N  K  A  Q  M  T  N  I  L  Q  Q  I  K
acagcacgtaccaccatggcaggcctgaccatggaggaacttatccagttggttgctgca
T  A  R  T  T  M  A  G  L  T  M  E  E  L  I  Q  L  V  A  A
cgactggcagaacatgagcgggtggcagcaagtactcagccacttggtcgcatccgggcc
R  L  A  E  H  E  R  V  A  A  S  T  Q  P  L  G  R  I  R  A
ttgttccctgctccactggcccaaatcagtaccccaatgttcttgccttctgcccaagtt
L  F  P  A  P  L  A  Q  I  S  T  P  M  F  L  P  S  A  Q  V
tcatatcctggaaggtcttcacatgctccagccacctgtaagctatgtctaatgtgccag
S  Y  P  G  R  S  S  H  A  P  A  T  C  K  L  C  L  M  C  Q
aaactcgtccagcccagtgagctgcatccaatggcgtgtacccatgtattgcacaaggag
K  L  V  Q  P  S  E  L  H  P  M  A  C  T  H  V  L  H  K  E
tgtatcaaattctgggcccagaccaacacaaatgacacttgtccctttgtccaactctt
C  I  K  F  W  A  Q  T  N  T  N  D  T  C  P  F  C  P  T  L
aaatgacggacctgactggggaggaagaagagaaactgatgtgaacaggaagcgcgg
K
gttcaagatttctaaaactctatatttatacagtgacatatactcatgccatgtacattt
ttattatataggtaatgtgtgtatagaaagtctgtattccaatgttcgtaaatgaaacta
tgtatattatgcagaaacagtctgttccccctcatcttgcaattcctttgggggatgcag
attgtagggaagatgatgtttagtttggccttgaaattatgatatccctgcccagggctg
ttttcaaatacaatataaaaaccacctaggaacctgctgttgctctaaggccattctgct
ttggtttggctcagcctctagtccatttccttaaggctcatgtatgcagatttaaagcct
ggtgctcacccactgtccaaccagatgccttgcttaccgaaagcctccagaagcctcagt
attgttttagccactctactccaaatggataaaatgagactctgattgaggaaaaaaaag
taaccctagtagtttgaaa
```

Figure 4.9 Predicted ORF for a novel protein. This was produced by assembling the appropriate assemblies and ESTs into a virtual mRNA. This was then translated to give the putative full-length protein sequence.

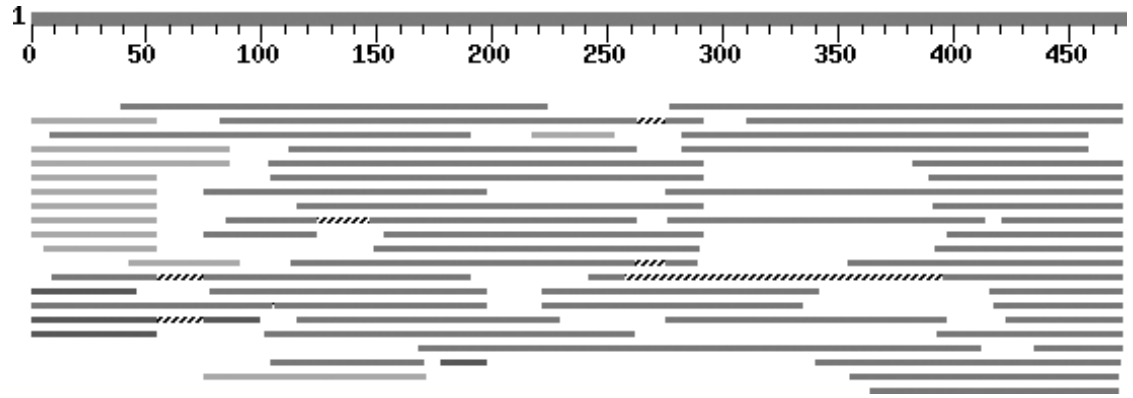

Figure 4.10 Checking for continuity of reading frame by translation searching (TBLASN) of the unknown ORF against all ESTs. The hatched lines represent deletions in ESTs that could represent splice variants.

in these matches, for example AI351632, represented as hatched lines in Figure 4.10. In addition to a bovine sequence BE75593 we also see a likely orthologous match to AL640079 from a toad. The support for the ORF now seems unassailable. The next step using BLAT again, is to map it back to GP (Figure 4.11). This reveals the matching of 15 exons from putative 5′ UTR to 3′ UTR. This is consistent with the Fgenes++ prediction at the 5′ end but this included two extra exons at the 3′ end. The fact that the virtual mRNA butts up very close to both neighbouring genes suggests that this could be a full-length transcript.

Clearly the analysis of what, for example, might be a candidate disease-associated gene, has to move on from the identification of an ORF to the assignment of function that is both mechanistically plausible and experimentally testable. The subject of assigning functions to new proteins is outside the scope of this chapter. However the two basic steps are a protein database search and motif analysis. The protein search (Figure 4.12) only shows significant similarity scores over the C-terminal section of the protein but the hits include the same proteins assigned as UniGene homologies by Ensembl. A comprehensive domain analysis using InterPro recognizes two domains (Kriventseva *et al*., 2001; Southan, 2000). One of the domains identified, IPR000694, is a proline-rich domain that may be involved in protein–protein interactions (Figure 4.13). However, the motif recognition specificity is low and therefore this could be a spurious match arising from a general high proline composition. An SRS query shows 1152 of these domains have been recorded in Ensembl (Zdobnov *et al*., 2002). The second domain, IPR001841, is more specific because it only occurs 187 times in the Ensembl gene set. The RING-finger is a specialized type of Zn-finger of 40 to 60 residues that binds two atoms of zinc, and is probably also involved

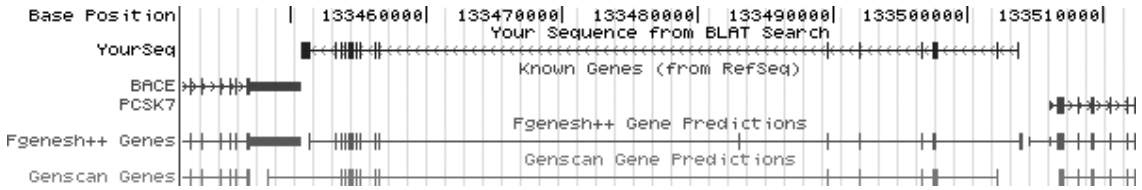

Figure 4.11 Matching the virtual mRNA back against GP using the BLAT search at UCSC. This delineates 15 exons with the gene reading in the opposite orientation to its neighbouring genes, i.e. 3′ end to the left, on the same strand.

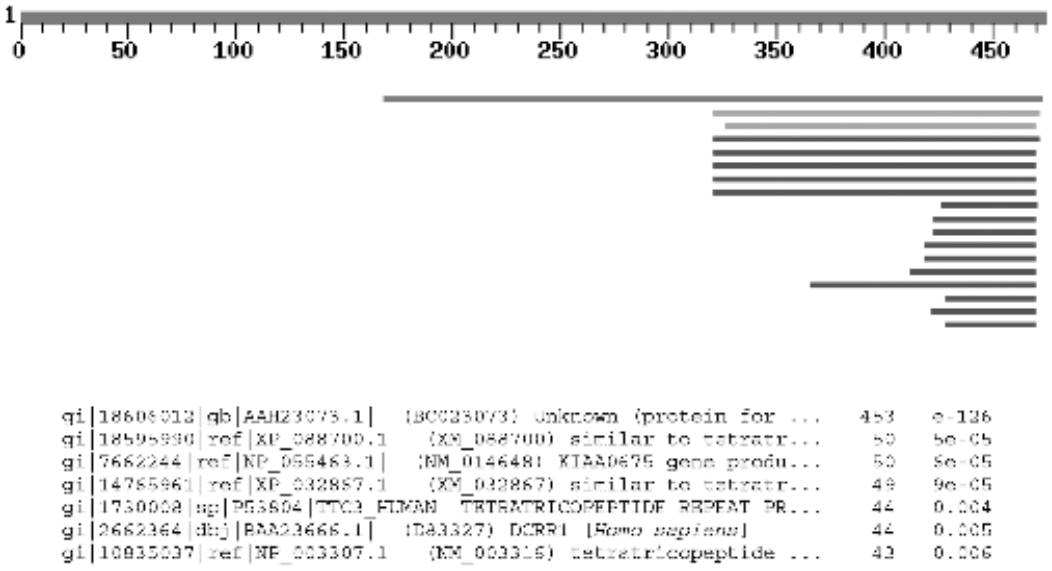

Figure 4.12 The sequence similarity scores of the novel ORF against the NCBI non-redundant protein database.

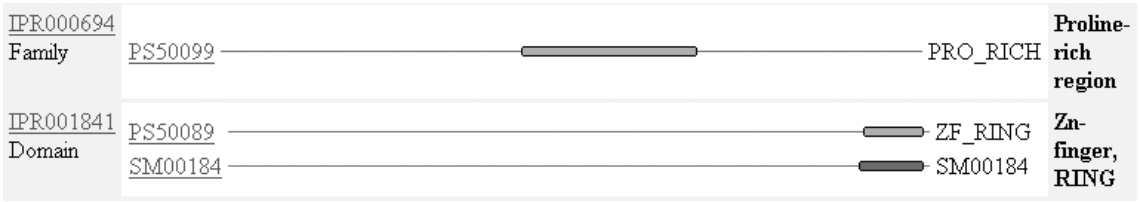

Figure 4.13 The InterPro domain/protein family analysis result for the novel ORF. The praline-rich domain is defined from a Prosite profile. The zinc finger is defined by both a Prosite profile and a SMART domain.

in mediating protein–protein interactions. They can also bind DNA however, since they contain many Lys, Ser and Thr residues. In fact combining the two domain searches finds intersecting hits (i.e. containing both domains) for only 17 Ensembl proteins. Inspecting the graphical displays shows one of these gene products, ESP0000020915, to be similar in domain orientation and spacing to the novel protein. Unfortunately the trail went cold here because this identifier has been changed in the latest Ensembl release and the SRS link to the protein sequence was dead.

So how did the three major gene portals do? Quite well considering they all included the potential novel gene product as a gene prediction although they disagreed on exon number. They also displayed key supporting evidence in different forms of track annotation. Only a small subset of the display options has been presented here. Was the use of all three portals essential? Strictly speaking we could have accessed sufficient supporting evidence from each one. However to collect all the available data it was necessary to use all three. The other aspect is that each portal has particular facilities that even if not unique at the technical level is easier to use at one portal compared to the other three. Consequently this kind of detailed analysis becomes a *de facto* three-stop-shop. For example the UniGene homology assignments, available from Ensembl, were all correct as judged by the agreement with the protein similarities (red tracks in Figure 4.4). Having said that, one of the direct protein homology assignments (yellow tracks in Figure 4.4),

the myosin-like peptide from yeast, was probably erroneous because of the low complexity of the query protein amino acid composition. In terms of markers, SNPs and genes Ensembl does particularly well for combined export options. The UniGene identity matches on the NCBI display together with the graphical stacks proportional to the number of EST matches are useful but in this case what is likely to be a single transcript was split into four clusters. One of these was illegible on the graphic and two others are dubious because of being unspliced. The UCSC displays were useful to see the two alternative gene models as well as being the only source of the TIGR EST assemblies. Another useful facility on this site is the ability of BLAT to display the hits of any externally constructed model or new database sequence. This can then be compared directly with the other display options (e.g. in Figures 4.6 and 4.11). The NCBI have recently introduced a gene model builder that can reproduce some of the steps above (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/ModelMakerHelp.html).

4.8 COMPREHENSIVE DATABASE SEARCHING

The protein matches and the InterPro analysis have already given functional clues about our novel protein. However if this particular gene product was located in close proximity to an SNP with a disease association we would need to find out as much as possible, not only to provide more supporting evidence for the gene product but also testable predictions about function that can be followed up. Performing a comprehensive search is not a trivial exercise since it involves 17 divisions of GenBank and sources of trace data that have not yet been submitted to GenBank. So where do we start? The two large repositories labelled nr protein or nucleotide on the NCBI BLAST server are a useful first choice (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi). We have already checked nr protein at 891,607 sequences but we need to compliment this with month, which in this case yields another 61,254 protein sequences but no additional high-scoring hits. The search against nr nucleotide with 1,192,858 sequences records three extended matches. This includes the mouse sequence already described, BC023073, and the primary accession number of the finished genomic section AP000892. The third match, XM_100696, is a secondary accession number for a reference mRNA sequence predicted by the NCBI Annotation Project from a genomic contig NT_009151. This is the same prediction labelled LOC160162 in Figure 4.5. There is an accompanying 56-residue predicted ORF that is in the NCBI protein database but has no supporting evidence. Inspection of the genomic location suggests it may be a spurious prediction.

Checking public patented proteins at 88,019 sequences gave no hits. However the patent nucleotide division, gbPAT, at 581,001 sequences, gives three solid hits, AX321627, AX192589 and AX072029. The first of these is a 2114-bp DNA from patent WO0172295. The document indicates this protein was isolated from a lung cancer sample (http://ep.espacenet.com/). These hits constitute a partial mRNA level of confirmation for the novel protein but a reciprocal check (i.e. a BLASTN of AX321627 against the nr nucleotide database) indicates this clone may be a chimera from two separate gene products. A search against a commercial patent database, containing 673,453 protein sequences, reveals identity matches for the N-terminal section from patent WO200060077 and a C-terminal identity match from WO200055350, both of which are reported as cancer-associated transcripts (http://www.derwent.com/geneseq/index.html). Checking the GSS division by TBLASTN gives four genomic hits; AZ847251 from mouse, AG114530 from chimpanzee, BH306228 from rat and BH406519 from chicken. Using BLAST against the

TABLE 4.1 Useful Resources for Gene Finding and Analysis

Site description	URL
Ensembl at EBI/Sanger Centre	http://www.ensembl.org/
Human Genome Browser at UCSC	http://genome.ucsc.edu/
Map Viewer at NCBI	http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_ search
Protein Atlas of the genome	http://www.confirmant.com/
SWISS-2DPAGE database	http://ca.expasy.org/ch2d/
Ensembl 4.28.1 announcement	http://www.ensembl.org/Dev/Lists/announce/msg00070.html
NCBI gene model builder	http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/ ModelMakerHelp.html
UniGene EST clusters	http://www.ncbi.nlm.nih.gov/UniGene/
InterPro at EBI	http://www.ebi.ac.uk/interpro/
Proteome analysis at EBI	http://www.ebi.ac.uk/proteome/
Google general search portal	http://www.google.com/
RefSeq at NCBI	http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html
International Protein Index	http://www.ebi.ac.uk/IPI/IPIhelp.html
Derwent sequence patent databases	http://www.derwent.com/geneseq/index.html
BLAST at NCBI	http://www.ncbi.nlm.nih.gov/BLAST/
BLAT at UCSC	http://genome.ucsc.edu/cgi-bin/hgBlat?command = start)
DAS — distributed annotation	http://biodas.org/
Exofish at Genoscope	http://www.genoscope.cns.fr/externe/tetraodon/
Fgenesh at Sanger Institute	http://genomic.sanger.ac.uk/gf/Help/fgenesh.html
Expasy translation tool	http://ca.expasy.org/tools/dna.html
CAP3 nucleotide assembly tool	http://bio.ifom-firc.it/ASSEMBLY/assemble.html
GeneWise at Sanger Institute	http://www.sanger.ac.uk/Software/Wise2/
Genscan at MIT	http://genes.mit.edu/GENSCAN.html
SSAHA at Sanger Institute	http://www.sanger.ac.uk/Software/analysis/SSAHA/

Ensembl mouse peptides detected a C-terminal similarity that is a zinc finger domain match. However both the human and mouse mRNA have unique and solid hits against mouse chromosome 9.40 Mb. This suggests the gene product is derived from this locus although it has not been annotated yet by Ensembl. Interestingly the gene lies between two odour receptors, unlike the human positioning between BACE and PCSK7, showing the position is non-syntenic.

Drawing detailed conclusions from these results is outside the scope of this chapter but the example makes clear how much extra information a comprehensive database search can yield. Was the protein unknown and/or novel? The difficulty of answering this question illustrates the diminishing utility of these terms. The protein has at least one function-related motif that can be recognized at high specificity so it can no longer be classified as an unknown. It remains novel only in the strict sense of not being represented in the current protein databases. It is not novel in the wider sense because both the mRNA and ORF were substantially covered as predicted by sequence data entries in the public and patent databases respectively.

4.9 CONCLUSIONS AND PROSPECTS

The geneticist is in the fortunate position of having access to secondary databases and GP genomic viewers of increasing quality, content and utility. This is making the process of finding and analysing gene products easier. However the examples used in this chapter also show that there are many subtle details in genomic annotation and the implications of these will take some time to unravel. This requires comprehensive inspection and may ultimately need experimental verification. The expansion of web-linked interoperativity and interrogation tools means that new options will already be available by the time this is in print. One consequence of these advances could be the perception of a diminished necessity to perform bioinformatic analysis. Although this is true in the sense that secondary databases include an increasing amount of 'pre cooked' bioinformatic data, there is a paradox in that the more sophisticated the public annotation becomes the more important it is to understand the underlying principles. For example, it is important to be able to discriminate between gene products defined by *in-vitro* data or only by *in-silico* prediction.

So what of the future? There are four developments worth highlighting. The first is that the combination of increasing transcript coverage, finished golden path and extensive mouse synteny data will diminish the uncertainty limits of gene numbers. The ability to pick out SNP haplotype blocks in relationship to gene products, already available as tracks on the UCSC display options for chromosome 21 will be a big step forward for association studies (Patil *et al*., 2001). The proliferation of DAS servers will enable more groups to share their own specialized annotation tracks with the wider community (http://biodas.org/). Last but not least defining gene products at the protein level is likely to have a major impact on annotation quality, and efforts are already underway to do this on a genome-wide scale (http://www.confirmant.com/).

REFERENCES

Birney E, Durbin R. (2000). Using GeneWise in the Drosophila annotation experiment. *Genome Res* **10**: 547–548.

Burge C, Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**: 78–94.

Deloukas P, Matthews LH, Ashurst J, Burton J, Gilbert JG, Jones M, *et al*. (2001). The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**: 865–871.

Guigo R, Agarwal P, Abril JF, Burset M, Fickett JW. (2000). An assessment of gene prediction accuracy in large DNA sequences. *Genome Res* **10**: 1631–1642.

Hoogland C, Sanchez JC, Tonella L, Binz PA, Bairoch A, Hochstrasser DF, *et al*. (2000). The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res* **28**: 286–288.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, *et al*. (2002). The Ensembl genome database project. *Nucleic Acids Res* **30**: 38–41.

Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E. (2002). TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res* **30**: 332–334.

Kriventseva EV, Biswas M, Apweiler, R. (2001). Clustering and analysis of protein families. *Curr Opin Struct Biol* **11**: 334–339.

O'Donovan C, Apweiler R, Bairoch A. (2001). The human proteomics initiative (HPI). *Trends Biotechnol* **19**: 178–181.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, *et al*. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.

Pesole G, Liuni S, Grillo G, Licciulli F, Mignone F, Gissi C, *et al*. (2002). UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res* **30**: 335–340.

Pruitt KD, Maglott DR. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* **29**: 137–140.

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, *et al*. (2001). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res* **29**: 159–164.

Schuler GD. (1997). Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med* **75**: 694–698.

Southan C. (2000). Website review: Interpro (the integrated resource of protein domains and functional sites). *Yeast* **17**: 327–334.

Southan C. (2001). A genomic perspective on human proteases as drug targets. *Drug Discov Today* **6**: 681–688.

Susens U, Borgmeyer U. (2001). Genomic structure of the gene for mouse germ-cell nuclear factor (GCNF). II. Comparison with the genomic structure of the human GCNF gene. *Genome Biol* **2**: research No. 0017.

Suzuki Y, Yamashita R, Nakai K, Sugano S. (2002). DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res* **30**: 328–331.

Wiehe T, Gebauer-Jung S, Mitchell-Olds T, Guigo R. (2001). SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Res* **11**: 1574–1583.

Zdobnov EM, Lopez R, Apweiler R, Etzold T. (2002). The EBI SRS server — recent developments. *Bioinformatics* **18**: 368–373.