Chapter One - Introduction

1 Introduction

1.1 Genome mapping and sequencing

1.1.1 Genomes

The genome of a cell contains all of the information necessary for the cell's function, although that function is shaped and facilitated by its environment. The genome remains the ultimate determinant of a cell or organism, because no matter how the information encoded within it is interpreted, only the genomic sequence contains all of the instructive components. Following the discovery of DNA as the inherited material within a cell, the discovery of the structure of DNA in 1953 by Watson and Crick, and the subsequent efforts of many scientists, led to rapid advances in our understanding of the molecular basis of genetic inheritance. Whilst retroviruses, for example, utilise RNA rather than DNA as their genomic material, the central concept of the genome containing all of the information required for a cell's function remains intact.

Although the concept of a genome is well established, the diversity of genome structure and organisation displayed by different organisms is remarkable. Genomes can show dramatic variation in sizes between organisms, as well as in their gene complement. For example, the genome of the SV40 virus is ~ 5 kb, and contains ~ 6 genes, but the human genome is ~3000 Mb and contains ~30-40,000 genes. *Mycoplasma genitalium*, a prokaryote with one of the smallest genomes known, is thought to approach a "minimal genome" of ~0.6 Mb, containing ~503 genes. A striking 85-90% of its genome is coding. In contrast, the genomes of some plants and amphibians can be of the order of 10^5 Mb in size.

Genome organisation also shows large variations between organisms. Some viruses for example, utilise overlapping genes to increase the coding capacity of their genomes, and the use of introns in eukaryotic but not prokaryotic genes is a clear distinction. Prokaryotic genomes comprise circular double-stranded DNA, compacted into a nucleoid structure; in comparison eukaryotic genomes are linear double-stranded DNA molecules, and are tightly packed into chromatin within a nucleus. The ciliated protozoa, an example of which is *Tetrahymena thermophila*, are remarkable in dividing their genome between two distinct nuclei to form a macronuclear and micronuclear genome. The micronuclear genome undergoes extensive rearrangements in producing

the macronuclear genome, from which transcription occurs. These examples of complex genome dynamics exemplify our need to further explore the variety of genome dynamics in different organisms.

Within phyla however, there can be significant variation in genome size. Gene number does not seem to increase dramatically with organismal complexity. For example, the human genome is expected to contain only approximately twice the number of genes found in the nematode *Caenorhabditis elegans*. The lack of understanding of relationships between genome size and content and organismal complexity reflects our limited knowledge of gene organisation and regulation, genome dynamics and an often anthropocentric view of evolutionary relationships. The availability of genome sequences for many different organisms will undoubtedly shed light on these aspects of biology.

1.1.2 Genome mapping and sequencing

The term "genome" reached iconic status during the latter part of the 20th Century. From the elucidation of the first genome sequence, that of the bacteriophage *phiX*174 virus in 1977 (Sanger *et al.,* 1978), to the completion of the finished human genome sequence by a consortium of research centres in 2003, the field of genome mapping and sequencing has advanced substantially.

The availability of genomic sequence for an organism is useful for many studies, such as gene identification, genetic trait mapping and the study of genome evolution. This has been illustrated by research involving organisms for which genome sequence has been available for some time, such as *Saccharomyces cerevisae*, *Drosophila melanogaster* and *Caenorhabditis elegans*, as well as many microbial organisms and viruses.

Before genome sequencing was considered practical, mapping and sequencing of the genome was on a smaller scale. These approaches involved isolating clones containing segments of genomic sequence, usually through hybridisation-based approaches. These studies were often motivated by evidence suggesting the location of a gene of interest in a particular region of the genome, based on linkage of traits or diseases to genetic markers placed on genetic maps.

Genetic maps are more readily generated in model organisms such as yeast, flies, worms and even mice, due to the ability to generate genetic crosses and follow

genes. In comparison, in humans genetic marker availability at a reasonably high resolution was not available until relatively recently. Also, in model organisms selective breeding can be used to produce informative genetic crosses. Early markers included protein isozymes and restriction fragment length polymorphism (RFLP) markers. These provided limited resolution but were nevertheless useful, and led to the production in 1987 of the first genetic map of the whole human genome (Donis-Keller *et al.,* 1987).

The subsequent availability of highly polymorphic microsatellite markers greatly increased genetic mapping power in humans, and in 1992 a second-generation genetic map of the human genome was produced using this type of marker (Weissenbach *et al.,* 1992). Further improvements in screening techniques have facilitated further refinement of genetic maps (Kong *et al.,* 2002).

These maps were developed in concert with refinements in physical mapping approaches. Initial restriction mapping, RH-mapping and cytogenetic approaches were improved by the advent of the polymerase chain reaction (PCR) which enabled clone contig maps to be produced using PCR-based STS-content approaches. In addition, restriction fingerprinting approaches aided the production of mapped clone substrates for genomic sequencing.

Early whole-genome mapping approaches used cosmid and phage genomic clones but their small size rendered them impractical for the mapping of larger genomes, driving improvements in cloning technologies, initially resulting in the advent of yeast artificial chromosomes (YACs). The first human genome YAC map was produced in 1993 (Cohen *et al.,* 1993). YAC clones are however limited by their propensity for chimaerism and unsuitability for sequencing. Although not suitable for providing sequencing substrates, the STS-content of YAC maps nevertheless provided ordered sets of markers for the production of bacterial clone maps (e.g. the YAC map of chromosome 22 (Collins *et al.,* 1995)).

Subsequent clone-based approaches used the smaller, more stable, cosmid (Collins and Bruning, 1978) and P1 artificial chromosome (PAC) genomic clones propagated in bacteria (Sternberg, 1992). Later, larger insert bacterial artificial chromosome (BAC) clones increased mapping efficiency further and were shown to be remarkably stable, presumably on account of their low copy number (Shizuya *et al.,* 1992). BACs were ultimately adopted as the substrate of choice for genome

sequencing. Developments in restriction fingerprinting techniques and increasing availability of STS markers further aided progress.

As sequencing methodologies improved and the density of markers placed on genetic and physical maps increased, the potential to sequence large genomes was explored. The first microbial genome to be sequenced was that of *Haemophilus influenzae* in 1995. The genome of this organism was sequenced using a whole-genome random sequencing strategy (Fleischmann *et al.,* 1995), also called a whole-genome shotgun (WGS) strategy. The first eukaryotic genome to be sequenced was that of *Saccharomyces cerevisae*, using a hierarchical shotgun strategy (Bussey *et al.,* 1997). The same strategy was also applied to the sequencing of the genome of *Caenorhabditis elegans,* the first multicellular organism to have its genome sequence determined (The *C. elegans* Sequencing Consortium, 1998). Publication of the genome sequence of *Drosophila melanogaster* in 2000 demonstrated the use of the whole-genome shotgun approach in the sequencing of complex genomes (Adams *et al.,* 2000).

1.1.3 The human and mouse genome projects

In 2001, two draft versions of the human genome were published, as well as BAC maps covering the different chromosomes (Lander *et al.,* 2001; Venter *et al.,* 2001). The physical maps covered more than 96% of the euchromatic genome and the draft sequence covered approximately 94% of the genome. At the time of writing in 2003, the finished sequence of the human genome was announced.

Debate over the strategies used to sequence the human genome, map-based clone sequencing versus whole-genome shotgun has been much reported, with the publicly-funded effort adopting the map -based clone sequencing approach and a private company, Celera Genomics formed some time later to adopt a whole-genome shotgun approach. Whilst offering the potential to provide information on the value of each approach in sequencing large, complex and repetitive genomes, the ability to make these comparisons was somewhat hampered by the commercial nature of the privately-funded approach.

A clone-based (hierarchical shotgun) sequencing strategy provides information regarding the positioning of the individual clones, information which can prove valuable in studies of the genome. In addition, in instances where short but very highly homologous repeats occur, clone-based mapping can resolve these regions where the

clones also contain some unique sequence, unlike a WGS approach in which the repeats would collapse. Sequencing of large genomic clones also allows the process of finishing to be performed, where ambiguities can be resolved. This is impractical with WGS-only approaches.

Whole-genome shotgun approaches can however provide data where there may be gaps in coverage in a clone library, and for regions that prove difficult to clone in larger segments. They also do not require investment in preliminary mapping studies and so can provide rapid sequence coverage of an organism's genome. They are also generally quicker to generate large amounts of sequence coverage. The WGS approach does need careful thought however as to how finishing of the sequence will progress, in the absence of a physical map. As with many areas, a combination of clone-based mapping and WGS strategies (hybrid approach) may prove to be the optimal approach.

A combination of clone-based mapping and WGS sequencing is currently being employed in the sequencing of the mouse genome. Early assemblies using WGS sequence have been released, constituting a "draft" version of the *Mus musculus* genome. Currently, approximately 6x genome coverage of WGS sequence has been generated, and the latest assembly (NCBI 30 assembly) comprises ~ 2,500 Mb of sequence in 136483 contigs (statistics taken from Ensembl Jan 2004). This approach provided an opportunity to test two independently produced assembly programs, Phusion (Mullikin and Ning, 2003) and Arachne (Batzoglou *et al.,* 2002), which were developed simultaneously. This WGS approach has provided a useful sequence resource at an early stage, which will be combined with physically-mapped, clone-based sequence to provide a "finished" genome. This illustrates the utility of a combinatorial approach.

The effort to assemble a BAC map of the mouse genome (Gregory *et al.,* 2002) was greatly facilitated by the availability of the human genome sequence, which provided a framework to allow the mouse map to be generated remarkably quickly, as will be illustrated in Chapter 4. BLAST analyses were used to match mouse BAC-end sequences to human chromosomes. This allowed mouse BAC contigs, assembled by restriction fingerprinting, to be ordered and oriented in regions of conserved synteny between the two genomes. Neighbouring mouse contigs could then be analysed for possible fingerprint overlaps at lower stringency. The availability of comprehensive

data from large-scale BAC–end sequencing and whole BAC library restriction-fingerprinting efforts was crucial in this regard.

This combined approach provides an efficient model for the mapping and sequencing of genomes for which related genome data are available, and provides early sequence data as well as a high-quality finished genome sequence and mapped clone resources. This hybrid approach is also currently being applied to the zebrafish genome.

1.1.4 Future directions and related studies

A WGS approach has also been utilised in the sequencing of the genomes of *Fugu rubripes* (Aparicio *et al.,* 2002) and *Ciona intestinalis* (Dehal *et al.,* 2002). In the case of *Ciona intestinalis*, high levels of haplotype diversity have complicated early assemblies of the sequence data. This illustrates the point that WGS approaches alone may not be optimal, and that assembly approaches will need to be continually refined.

For organisms with less well-developed genomic resources, techniques such as HAPPY mapping have shown utility in generating physical mapping data, as was displayed in the *Dictostylium* genome project (Williams and Firtel, 2000). This relatively simple technique assays pools of sheared genomic DNA for shared markers (Dear and Cook, 1989). This will be particularly important for those organisms for which research funding is limited, such as where examination of the genomes of organisms at key evolutionary positions may shed light on genome evolution.

Generation of reference genomic sequence data is complemented by efforts to understand the differences between individuals within a species, by studies of nucleotide variation such as single-nucleotide polymorphisms (SNPs) where different alleles of single base-pairs are found, small insertions and deletions (indels) and larger DNA segment copy number differences. This information provides markers for linkage and association studies, and information regarding the differences between individuals that may lead to differences in gene expression or protein function. In parallel with the release of the human genome draft sequence came efforts to generate large amounts of SNP data. These efforts are underway in many centres, both publicly and privately funded. At the time of writing, ~2.17 million SNPs had been mapped to the human genome assembly (statistics from dbSNP at NCBI).

The utility of SNP data for association studies aimed at discovering genes involved in complex disease has been debated. The success of such approaches will depend on the degree to which the polymorphism observed is causative or confers susceptibility (the "common-variant, common-disease" argument), and on the underlying haplotype structures (conservation of allelic variations over a genomic region) reflecting population stratification.

Efforts are underway to generate a haplotype map of the human genome to determine the extent of linkage disequilibrium (LD) within the genome (Gibbs *et al.*, 2003), and early studies have suggested that relatively large segments of the genome appear to exist as "haplotype blocks" (Wall and Pritchard, 2003). This may reduce the number of SNPs needed to perform genome-wide screens, but at the expense of resolution. Nevertheless, these studies will doubtless shed light on the evolution of the human genome and will further our knowledge of the differences that contribute to our individual traits, susceptibility to disease and response to therapies.

1.2 Gene identification

1.2.1 Genes

Studies of genes, greatly facilitated by the availability of large amounts of genomic sequence, have shown us that gene structures can display great diversity. At the most fundamental level, the difference between eukaryotic and prokaryotic genes is in the use of introns. In prokaryotes, transcription of genes occurs and the mRNA is translated directly. In eukaryotes, the mRNA is subjected to processing including the addition of a 5' CAP structure, addition of a 3' polyA tract and the splicing of introns. Only then is the mature mRNA translated. However, the discovery of trans-splicing in *C. elegans* (Blumenthal, 1995) hints at further undiscovered complexities in the expression and composition of genes.

The word "gene" has also proved difficult to define. Alternative splicing of transcribed mRNA, the use of alternative promoters and different polyadenylation sites (as explored in Chapter 3) can all result in further complexity, leading to the production of mature mRNA species encoding different proteins, in different tissues and with different physiological half-lives. Alternative splicing is seen for many genes (see also Chapter 3) and is currently an area of intensive research.

One of the most extreme examples of alternative splicing seen to date is the *Drosophila melanogaster* Dscam gene, reported potentially to encode greater than 38,000 different mRNA species (Schmucker *et al.*, 2000). Verifying which variants are actually produced *in vivo*, and have physiological relevance rather than being just products of inefficient or aberrant splicing, is technically difficult due to incomplete representations of different tissues and different developmental stages, especially in human. In this regard, model organisms often provide a more comprehensive array of resources. The development of microarrays is also maturing, and in the future may allow such high-resolution studies on a larger scale. A striking example of the physiological relevance of alternative splicing is the sex-lethal (sxl) gene in *Drosophila melanogaster*. Skipping of an exon containing a premature termination codon in females allows production of functional sxl protein, leading to further regulation of splicing in other genes to result in sexual differentiation effects.

The use of alternative promoters and of other regulatory elements is another poorly understood facet of genes. Utilisation of these elements allows a gene's expression to be controlled, both temporally and spatially within the tissues throughout the organism. For example, the dystrophin gene uses at least eight different promoters to generate cell-type specific transcripts. One of the clear benefits of the availability of genomic sequence, as opposed to studying only mRNA, is that promoters and other regulatory elements will be contained within the sequence. One of the challenges in this area is the identification of such regions. Experimental techniques for studying these elements are low-throughput and labour-intensive, and whilst various computational approaches have been used to try and identify promoters, they are limited and suffer from high over-prediction (and potentially under-prediction) rates. However, an approach demonstrating utility in this regard is comparative analysis of genomic sequence from multiple species, as will be discussed further in Section 1.4.

Utilisation of different polyadenylation sites has been noted for many genes, and the study of these has benefited from the production of large numbers of 3' region Expressed Sequence Tags (ESTs). By noting when 3' ESTs cluster together and share a common polyadenylation site, and by aligning these against genomic sequence, it can be determined if different EST clusters are present at different regions of a gene's 3' UTR. As regions in the UTRs of genes have been shown to regulate their rate of decay, and so indirectly the level of protein produced from them, utilisation of alternative 3'

UTR sites could allow a gene to produce mRNA species with different physiological properties. Signals present within the 3' UTR controlling the localisation of mRNA within the cell, and hence localisation of resultant protein, have also been described (Veyrune *et al.,* 1996) and could be controlled by alternative polyadenylation site usage.

Other complexities of gene structure include the complex VDJ segment joining of immunoglobin genes, in which gene structures become rearranged with the potential for a great variety of structures to be created. Genetic aberrations can also occur, resulting in fusions of genes in some instances. Fused genes can result from chromosomal translocations bringing different regions of different genes within proximity of one another, resulting in production of hybrid mRNA transcripts. An example of this is the translocation that can occur between chromosome 22 and 9, generating "Philadelphia" chromosomes in chronic myelogenous leukaemia (Nowell and Hungerford, 1960). This translocation generates a fusion between the BCR and ABL genes, with the resulting BCR-ABL gene causing activation of transforming pathways.

The availability of large amounts of genomic and mRNA sequence, and both computational analysis and manual curation (as seen in Chapters 3 and 4) have provided data on the physical dimensions and sequence compositions of a large number of gene structures in complex organisms. Great heterogeneity is apparent, with genes ranging in size from approximately 0.1 kb to 2.4 Mb in human (Zhang, 1998). The large variation in size is largely attributed to differing intron sizes, as the average exon size is relatively uniform at approximately 200 bp. For the largest genes, a very small proportion of the transcribed RNA becomes mature mRNA. This would appear wasteful in terms of energy requirements of the cell, but it has been proposed that this may constitute a form of control due to the length of time taken to transcribe these genes with respect to other events occurring within the cell (e.g. ~ 16 hrs for the transcription of the 2.4 Mb dystrophin gene (Tennyson *et al.,* 1995)).

Depending on the region of the genome in which a gene resides, the repeat and GC content of different loci is seen to vary. In different organisms, orthologous genes can show marked differences in their size and repeat content. For example, the genome of the pufferfish *Fugu rubripes* is more compact than that of *Homo sapiens*. It contains fewer repeat sequences, and its genes generally have shorter introns (Aparicio *et al.,* 2002).

As gene structures and their interpretation by the transcriptional apparatus of the cell are so complex, defining a gene is non-trivial. It is often in practice taken to mean the boundaries of the genome containing the promoter and exons (although not necessarily all regulatory regions) used to produce a range of related transcripts. That is the definition that will be used in this thesis, and the word "locus" will be used to describe defined regions of the genome, which may often be synonymous with a gene.

1.2.2 cDNA-based gene identification methods (direct selection) and exon trapping

Before large amounts of genomic sequence were available, gene discovery techniques focussed on the cloning and sequencing of cDNA sequences generated from mRNA, and on methods such as direct selection.

Direct selection (Lovett *et al.*, 1991), uses the ability of mRNA to be converted to cDNA, and the ability of genomic DNA and cDNA strands to form duplexes via hybridisation. Genomic DNA clones are digested by restriction enzymes, and linkers are attached to the free ends facilitating attachment to magnetic beads. The clone fragments are incubated with cDNA clones amplified from a cDNA library by PCR, and hybridisation can occur if exons are present in the genomic fragment and the corresponding mRNA is represented in the cDNA source library.

Hybridised cDNAs are captured using a magnetic column, the cDNA fragments eluted and re-amplified by PCR, and the process repeated using this refined pool of cDNAs. Successive rounds of this procedure result in an enrichment of cDNA species, which are eventually sequenced. This approach is limited by the representation of genes in the cDNA library chosen, however, and by the low hybridisation efficiency of small exons.

Some of the benefits of this method are that information is generated regarding the expression pattern of the transcript, based on the sources of the mRNAs, and also on the splicing of the transcript as intron splicing has occurred by the stage when processed (polyA+) mRNAs are isolated. In this way, different splice variants can be identified, and their expression patterns determined in a variety of tissues.

Exon-trapping (Duyk *et al.*, 1990), involves the insertion of a genomic DNA fragment into a "mini-gene" vector, containing two exons separated by an intron, which contains a multiple cloning site. If the genomic DNA fragment contains an exon, it can be spliced in-between the existing vector exons following transient expression in

mammalian cells, and its presence or absence is revealed by PCR analysis of the mRNA produced. This approach is however limited to small-insert clones, and can also be limited by slow or cryptic splicing.

The cDNA sequence alone, however, does not allow the structure of the gene to be elucidated and does not contain information on promoters or regulatory sequences, other than those which may be present in exons. In addition, mRNA is technically difficult to handle, as it is very susceptible to degradation, and tissue availability as a source of mRNA can be an issue in some species.

1.2.3 Sequence-based gene and regulatory-sequence identification methods

The genomic sequence contains all of the information needed to transcribe functional genes, including exons, promoters and other regulatory sequences such as enhancers. As such, it provides a powerful resource for gene discovery. The problem then becomes one of determining where all the different features of the genes lie. To address this, two main approaches are adopted – *ab initio* gene prediction and annotation using mRNA sequence.

Various programs have been developed that predict exons or gene structures in genomic DNA of different organisms. Such prediction is more straightforward in bacteria and yeast as open reading frames are more easily discernible and the genes lack introns. More sophisticated approaches are required in higher organisms. The prediction algorithms generally all make use of the different composition of coding regions compared to the rest of the genome, as they are constrained by codon usage, to predict coding regions.

Exon prediction programs such as GRAIL (Xu *et al.*, 1994) are used to predict single exons. Gene prediction programs such as FGENES (Solovyev and Salamov, 1997) and GENSCAN (Burge and Karlin, 1997) make further use of conserved splicing signals in introns to attempt to predict exon/intron structure. Whilst all programs suffer to varying degrees from lack of specificity and sensitivity (Guigo *et al.*, 2000), they have nevertheless proved invaluable in the annotation of genomic sequence and can attain high levels of accuracy in some instances (>90% for Genscan (Guigo *et al.*, 2000)). Use of multiple programs can increase sensitivity and confidence in prediction to some extent.

The initiation of large scale cDNA sequencing projects and release of the sequence information into the public domain has had a dramatic effect on genome-scale gene identification in a variety of species, including mouse and human. Some of these projects are listed in Table 1-1.

The cDNA sequences produced can be mapped and aligned onto the genomic sequence by programs such as EST2GENOME, allowing gene structures to be annotated. These programs are not perfect though, and sometimes find difficulty defining splice sites correctly and can miss very small exons.

Project / Centre	Comments	Statistics to date	Link
NEDO – University of Tokyo, Helix Research Institute, Kazusa DNA Research Institute; Japan.	Three-centre human cDNA library generation and sequencing project. Some centres utilise the oligo-capping methodology. Non-Kasuza clones prefixed "FLJ".	Total of 29,314 clones registered with DDBJ.	http://www.nedo.go.jp/bio-e/index.html
Kazusa DNA Research Institute, Japan.	Kazusa cDNA sequencing project. Part of the NEDO project. Utilise size-fractionated human cDNA clones to select larger clones. Clones prefixed "KIAA" (or "FLJ" for NEDO clones).	2,031 KIAA clones in HUGE database. 362 FLJ NEDO database clones (adult spleen, not oligo-capped).	http://www.kazusa.or.jp/en/
RIKEN – Genomics Sciences Centre, Japan	Mouse cDNA library production, sequencing and functional annotation. Developed many methodologies for high-throughput cDNA library preparation and sequencing, including the CAP-trapper method.	60,770 mouse clones in FANTOM2 dataset.	http://www.gsc.riken.go.jp/
MGC – NIH Mammalian Gene Collection. Multi-Institute trans-NIH initiative.	Human and mouse cDNA library production and sequencing of clones containing full-length ORFs.	Human 14,878 full ORF clones (11,061 nr genes). Mouse 10,947 full ORF clones (9,019 nr genes).	http://mgc.nci.nih.gov/
DKFZ – Heidelberg, Germany.	Consortium of eight centres to produce (at the DKFZ) and sequence human cDNA clones.	>3200 FL cDNA clones (tissue/development specific)	http://www.dkfz-heidelberg.de/mga/GCC/

Table 1-1 Major human and mouse cDNA sequencing projects in progress at time of writing.

Key references:

HUGE – (Kikuno *et al.*, 2002)

RIKEN – (Shibata *et al*., 2000), (Carninci *et al.*, 2000)

NEDO – (Yudate *et al.*, 2001)

MGC – (Strausberg *et al*., 1999)

There are also technical limitations to be considered in large-scale cDNA sequencing projects. One major problem is avoiding clone size bias due to the relative inefficiencies of ligations for larger insert clones. This is often overcome by conducting a size fractionation step with the source mRNA, with the different size fractions then sub-cloned separately to avoid smaller inserts out-competing larger ones (e.g. KIAA cDNA clones – Kazusa DNA Research Institute, Chiba, Japan). Another problem is ensuring that the clones represent the full 5' UTR of the mRNA transcript. Here, various different approaches have been applied such as oligo-capping, which targets the 5' CAP structure of mature mRNA, and hence selects for full-length mRNA species (e.g. RIKEN cap-trapper methodology (Carninci *et al.*, 2000)).

Although improvements have been made in this area, it is still an issue that impacts on the study of promoters and transcription, as knowledge of the true 5' end of a mature mRNA species can determine where in the genome the transcription start site and core promoter lie. Conversely, mRNA that is incomplete at the 5' end can lead to erroneous conclusions, and other methods for determining the true 5' end such as primer extension are laborious. Additional technologies, such as subtractive hybridisation are also used to enrich the diversity of transcripts represented in the cDNA libraries chosen, in order to reduce redundant sequencing (Konietzko and Kuhl, 1998).

cDNA sequences provide direct information about transcription and RNA processing. However, even the application of the measures described does not provide access to all mRNAs (for example those expressed in tissues that cannot be obtained). Conversely, the genome sequence contains information on all of the exons, but these are small signals hidden in a mass of other sequences. An additional, powerful method of genomic sequence-based gene identification utilises genomic sequence from different species, and the observation that functionally important regions of the genome, such as exons, are usually more conserved between species than non-coding DNA. This approach will be discussed later in Section 1.4.

Genomic sequence alone cannot provide information on which splice variants are produced or where they are expressed. As is often the case, the optimal strategy combines two approaches – use of genomic sequence and gene prediction algorithms combined with mRNA sequence information. The approach to gene identification described in Chapter 3 and 4 utilises mRNA sequence aligned against the genome sequence to facilitate manual annotation of gene structures. Additionally in Chapter 3, results from gene prediction algorithms are used to target the design of primers to putative exons, to interrogate cDNA sources and generate sequence which can be used to elucidate gene structure.

A difficult problem in genomic sequence analysis is the issue of promoter prediction. Unlike gene prediction, promoter prediction must contend with the problems of short transcription factor binding site consensus sequences, which occur by chance many times within the genome. As a result, many *ab initio* promoter prediction approaches suffer from very high false positive rates. In addition, promoter function is poorly understood due to the technical difficulties involved, and there are relatively few verified promoters that can be used for training the computational methods. Promoter function can also involve elements that are at a considerable distance upstream or downstream of a gene (presumably involving looping of DNA to bring these elements into proximity with the gene). These effects present a formidable or insurmountable challenge from a genomic analysis perspective.

Progress has been made however, and *ab initio* promoter prediction packages (e.g. PROMOTERINSPECTOR (Scherf *et al.,* 2000), FirstEF (Davuluri *et al.,* 2001)) and programs that predict transcription start sites (e.g. EPONINE (Down and Hubbard, 2002)) are available, alongside a database of experimentally determined transcription factor binding sites (TRANSFAC (Wingender *et al.,* 2001)). Increasing representation of the true 5' ends of mRNA transcripts in libraries used for large-scale sequencing projects (e.g. RIKEN cDNA sequencing project) and from Rapid Amplification of cDNA Ends (RACE (Frohman *et al.,* 1988)) technologies also lead to information regarding where the core promoter lies for more refined analysis. Further effort is needed in this area though, as neither of these methodologies is completely effective.

1.2.4 Genomic sequence analysis

The availability of the sequence of an organism's genome also allows the study of the composition of genomes as a whole. It allows comparisons between the genomes

of different species (see later) and studies of how the composition of a genome varies between species and between different chromosomal regions within a species. Earlier studies of mammalian genome composition utilised approaches based on differential staining of genomic regions by different dyes, such as Giemsa. The banding pattern produced by this stain reflects differences in GC composition and degree of condensation, with dark "G" bands having a relatively low % GC (and also lower gene content) and the lighter "R" bands a higher % GC (and raised gene content). The availability of large amounts of genomic sequence allows questions of genome composition to be addressed directly.

The study of isochores is one such application. Bernardi and others first demonstrated that genomes appear to contain regions with differing GC compositions, termed isochores, from gradient ultacentrifugation of nuclear DNA producing fractions with differing GC content (Bernardi *et al.*, 1985). These isochores comprise regions between 100 kb and many megabases in size. The human genome is thought to contain five classes of isochore – two light (AT-rich) classes; L1 and L2, and three heavy classes (GC-rich); H1, H2 and H3. The differing GC composition of isochores appears to correlate with features such as repeat density, gene content, replication timing and recombination frequency. Building understanding of these correlations and possible mechanisms of genome composition evolution is an area of current research, and benefits greatly from the framework provided by the genomic sequence.

GC content also varies on a finer scale. Within the vertebrate genome, so-called CpG islands are found, and in humans are associated with approximately 56% of genes (Antequera and Bird, 1993). CpG di-nucleotides are depleted within the genome, due to methylation of cytosine pre-disposing it to undergo deamination to form thymine. Due to active promoter regions being methylation-free, CpG di-nucleotides within these regions are preserved, forming CpG "islands". CpG islands are found within the promoter regions or first exons of housekeeping genes, and a proportion of genes with restricted expression. This fact was used in early transcript mapping approaches and estimation of human gene number using restriction digestion of DNA using methylation-sensitive restriction enzymes such as *HpaII* (which cut unmethylated CpG regions). Programs are available to find CpG regions within genomic sequence (e.g. as part of the GRAIL program, and Gos Micklem personal communication), and these can be correlated with gene structures as annotation continues.

The overall GC content of genomes has also been shown to vary substantially. Between orders large variations in overall GC content can be seen. For example, the genome of the Archael extremophile *Methanococcus jannaschii* which occupies an environment with very high pressures (>200 atmospheres) and temperatures (~85^0C) (TIGR website) is ~ 31% GC, compared to an average of ~ 41% GC for the human genome (Bult *et al.,* 1996).

Another feature of genomes that can be studied using sequence analysis approaches is the repeat landscape. The genomes of different organisms contain differing numbers and types of repeats. The genomes of the pufferfish *Fugu rubripes* and *Tetraodon nigroviridis* appear to be relatively repeat-poor, and so are considered compact genomes. In contrast, mammalian genomes contain a variety of different interspersed repeats. In human, interspersed repeats constitute approximately 45% of the genome. The predominant repeats are LINEs (Long Interspersed Nuclear Elements) and SINEs (Short Interspersed Nuclear Elements), with smaller contributions from LTR elements and DNA transposons. Their genomic distribution appears to correlate with factors such as GC content and gene density.

Availability of genomic sequence allows repeat composition to be studied in context with other features of the genomic landscape, and these analyses are facilitated by the availability of repeat element databases (e.g. REPBASE (Jurka, 2000)) and search tools (e.g. REPEATMASKER, A. F. A. Smit and P. Green unpublished). Some aspects of genome structure and function are very likely to remain intractable using a pure sequence analysis approach. However, in many areas a combination of sequence analysis and experimental approaches is proving highly productive in building understanding about the composition of genomes, how they function and how their structure and dynamics are related to these functions.

1.3 Gene duplications and evolution of genomes

1.3.1 General evolution framework

The age of the Earth is debatable, but one estimate is that it is approximately 4.55 billion years old and life is thought to have become established on Earth approximately 3.5 billion years ago. Since then, evolution has resulted in an extraordinarily diverse range of species. Studies of fossils and of the morphology and behaviour of extant species have developed our understanding of the relationships

between different orders and species. The more recent application of increasingly sophisticated molecular techniques, to compare protein and nucleic acid sequences among species, has improved our understanding further.

A much-studied aspect of life's evolutionary history is the so-called "Cambrian explosion". During the early Cambrian period approximately 545 million years ago (Mya), there appears to have been a massive diversification of life forms, the reasons for which are the subject of study and debate. The availability of genome sequences of different species will allow us to gain a deeper understanding of the relationships between species, the evolution of molecular functions and the evolution of genomes.

Our current understanding of broad organismal relationships is still expanding (for a recent overview, see (Pennisi, 2003)). Twenty-five metazoan organisms for which representative genome sequences are available, are in the process of being sequenced, or are selected for sequencing (Ureta-Vidal *et al.*, 2003). Further to this, many organisms from earlier evolutionary branches are having their genomes sequenced and smaller selected genomic regions from other organisms can be sequenced with modern technologies and improving genomic clone resources.

Of particular interest with regards to studies presented in this thesis is the evolution of vertebrates. An overview of our current understanding of vertebrate evolution is shown in Figure 1-1. Also shown are key geological features of the Earth during these periods, which help to put this evolutionary period in perspective. These include developments such as the establishment of life on land, continental fragmentation and formation, formation of mountain ranges and changes in atmospheric conditions. Several species at key points in this evolutionary tree have been the subject of studies of particular molecules which have provided insight into species relationships and molecular evolution. Mitochondrial sequences have proved particularly useful in this regard.

A central theme of biological research in the last two centuries is the theory that selective pressures act upon the organism to shape the evolution of functions at the molecular level. This is ultimately reflected in the genome of an organism, the genes it contains and its composition. The following sections discuss key theories under active investigation in this area.

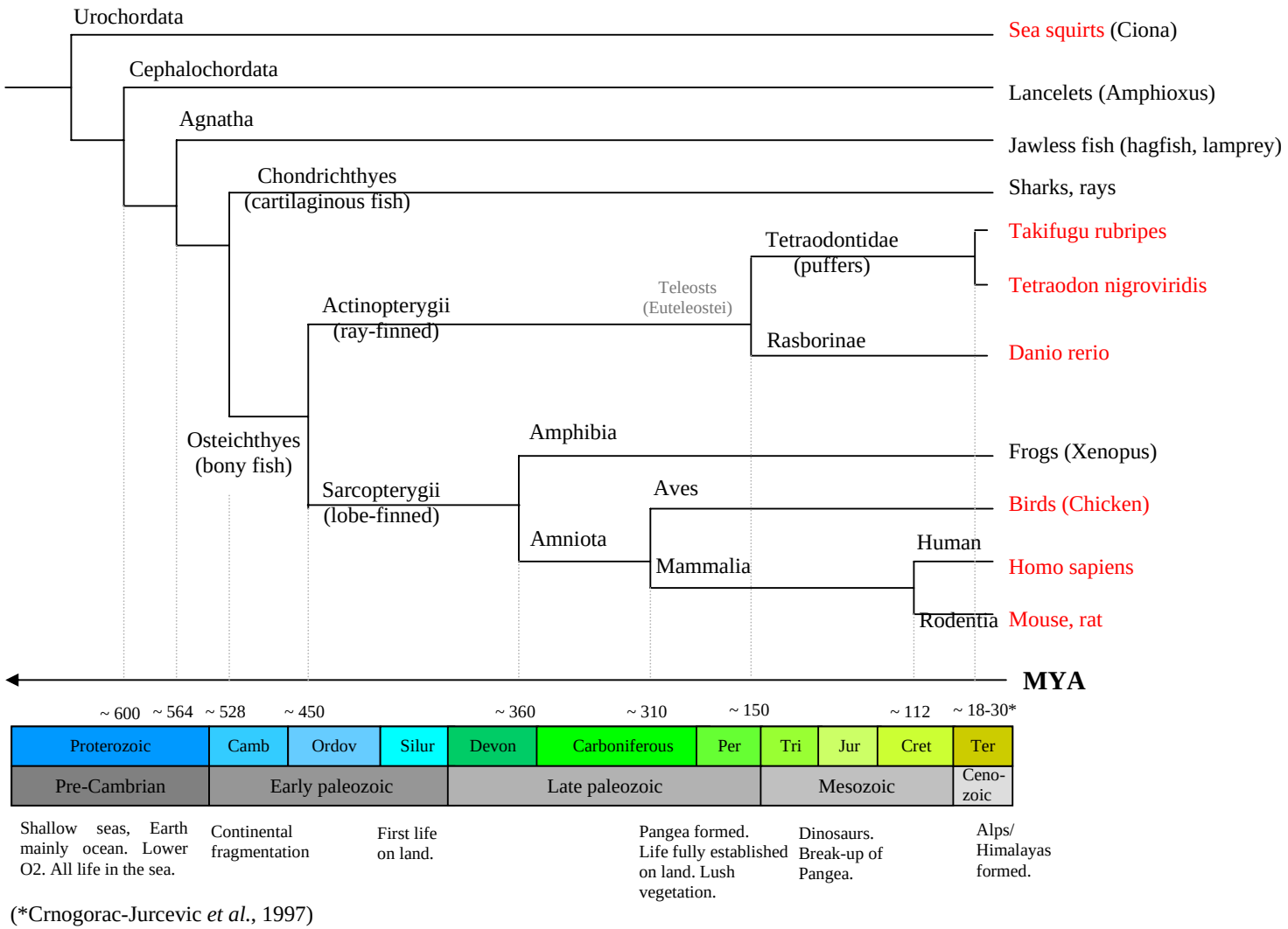

Figure 1-1 Schematic cladogram of selected branches of chordate/vertebrate evolutionary relationships. Estimates of divergence times are from Kumar and Hedges (Kumar and Hedges, 1998), unless otherwise indicated (Camb – Cambrian, Ordov – Ordovician, Silur – Silurian, Devon – Devonian, Per – Permian, Tri – Triassic, Jur – Jurassic, Cret – Cretacious, Ter – Tertiary). Organisms for which large amounts of genomic sequence are being generated are in red font.

1.3.2 *Whole genome duplication hypothesis*

It had been noted previously that for several genes, vertebrates have several copies of genes which are represented by a single copy in invertebrates. The most notable examples are the *Hox* genes. These genes, present in paralogous clusters, have key roles in developmental regulation and have been the target of intensive study (Averof, 2002).

In 1970, Susumu Ohno proposed that, during the course of evolution, the genomes of organisms giving rise to the vertebrate lineage underwent many gene duplication events, and that this contributed to an increase in organism complexity (Ohno, 1999). This has been proposed to have incorporated rounds of genome duplication. This theory could explain why some organisms contain multiple copies of a particular gene, depending on when a genome-duplication occurred. Differences in the exact ratios seen could be explained by gene loss or tandem duplications in different species. Figure 1-2 summarises currently proposed genome-duplication events.

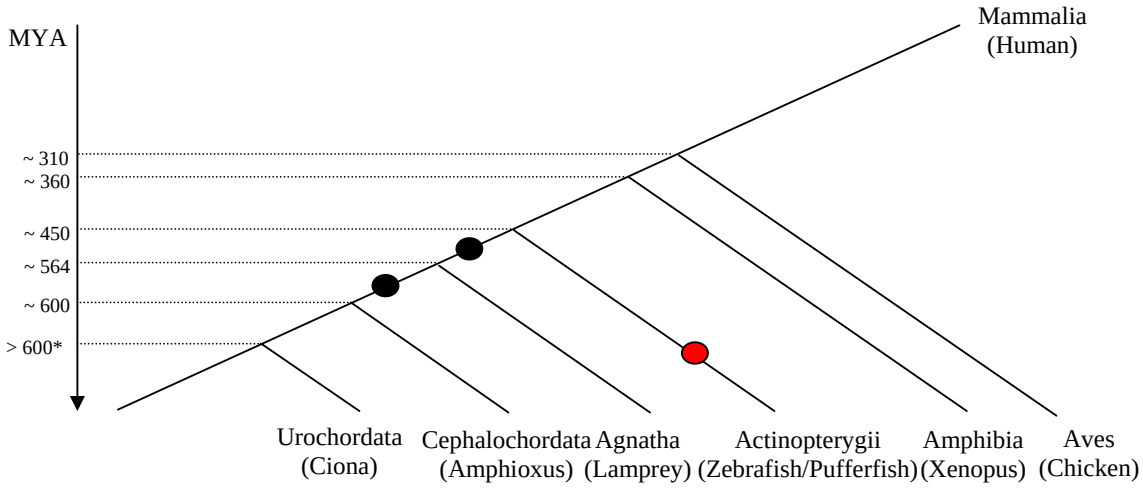

Figure 1-2 Schematic representation of proposed genome duplication events (black circles) during vertebrate evolution (including the genome duplication occurring in the teleost lineage – red circle). Divergence timings are from Kumar and Hedges (1998) (except *, Makalowski, 2001)

Additional evidence supporting this theory comes from the observation of tetraploidy in an amphibian, *Xenopus laevis*, which could represent an intermediate stage in whole-genome duplication. It has also been shown that the yeast genome has undergone whole-genome duplication (Wolfe and Shields, 1997), as well as the

Arabadopsis genome. Proponents of the theory suggest that two rounds of genome duplication occurred early in the vertebrate lineage ("2R"), with a further round of genome duplication occurring in the lineage producing the teleost fish, for example zebrafish (Figure 1-2).

This theory is still the subject of intensive investigation and vigorous debate, with no clear decision yet between the whole-genome duplication model or an alternative in which genomes were shaped by extensive segmental and tandem duplications, as discussed in the following section. It should be noted that the theories are not necessarily mutually exclusive, and that both mechanisms may have played a role in shaping the vertebrate genomes.

Evidence for the theory of whole-genome duplications has come from studies of DNA content and gene/cluster number in a wide variety of organisms, and has been reviewed extensively. Evidence against it has come particularly from the viewpoint that genes duplicated in this manner should maintain a symmetrical ((A,B)(C,D)) phylogeny where genes generated by a genome duplication are more similar to one another than counterparts prior to the duplication. However, it has been noted that this assumption may be violated if an allotetraploidy scenario is involved, whereby mating of closely related species results in an increase in ploidy. In addition, some plants and animals (e.g. salmonid fish) contain a mixture of tetraploid and diploid loci, indicating that genome duplication may not necessarily occur in its entirety, but may instead only involve a number of chromosomes.

The central theme of these scenarios, whichever proves to be correct, is that by increasing its gene complement through duplication of genetic material, an organism increases the repertoire of molecules on which selective pressures can act, and potentially increases its ability to evolve to meet changing evolutionary pressures.

1.3.3 Segmental duplications and tandem duplications

Apart from whole-genome duplication, an organism can increase its gene complement through the processes of segmental duplications and tandem gene duplications, segmental duplication being the duplication of complete region of genomic DNA and tandem duplication being the duplication of a gene within the same region of the genome.

Segmental duplications can lead to duplication of multiple genes in a single event. They are characterised by common order and transcriptional direction of the paralogous genes. These may be tolerated by the organism to different degrees, depending on the copy-number effects of the genes involved. This is non-trivial, as there are examples where an increase in copy number of a gene has deleterious effects, such as in Pelizaeus-Merzbacher disease, which can be caused by duplication of the PLP gene region (Inoue *et al.*, 1996). These effects will likely depend on the complexity of the organism and the functions of the genes involved. Large numbers of segmental duplication would also lead to greater homogeneity of the genome's composition, depending on the size and numbers of duplications, as before the sequences diverge through mutation, they will be equivalent.

In humans, segmental duplications have been implicated in disease, such as DiGeorge and velocardiofacial syndromes, due to their promotion of genomic rearrangements by a variety of mechanisms including unequal crossover (Emanuel and Shaikh, 2001). These rearrangements can lead to duplication or deletion of genes. Such rearrangements have also been attributed to other low-copy repeat (LCR) regions which are not necessarily part of segmental duplications (Mazzarella and Schlessinger, 1998; Stankiewicz and Lupski, 2002).

Tandem duplications of genes also increase the gene complement of an organism and are again constrained by copy-number effects, but as fewer genes are involved in the duplication event, the functional impact is expected to be lessened. This is however still dependent on the gene in question and the context in which it operates. There are many well studied examples of tandemly duplicated genes, including the hox and globin genes.

The generation of genome-scale sequence data for many organisms such as *Saccharomyces cerevisae*, *Drosophila melanogaster*, *Caenorhabditis elegans* and others has allowed studies on the processes of segmental and tandem duplications in these organisms, allowing statistical techniques to be refined and shedding light on the contribution of these processes to genome evolution (Wolfe and Shields, 1997), Friedman and Hughes, 2001), Achaz *et al.*, 2001).

The relevant contributions of these processes to the shaping of the human genome have recently been the subject of two gene-centric genome-scale studies (Gu, *et*

al., 2002); McLysaght *et al.,* 2002), and were made possible by the completion of the draft version of the human genome (Lander, Linton *et al.,* 2001). These studies utilised different methodologies, based on molecular-clock analysis of duplicated genes. Both concluded that extensive gene duplication, or at least one round of polyploidy, occurred early in chordate evolution, consistent with widespread segmental or chromosomal duplications. Gu *et al.,* (2000) also noted a wave of duplications subsequent to the mammalian radiation, which were attributed to tandem or segmental duplications.

Further studies of segmental duplications have been performed using the human genome sequence. Bailey *et al.,* (Bailey *et al.,* 2002) made use of the whole-genome shotgun data from Celera to identify regions of the genome that had high-levels of sequence identity to one another, thus representing relatively recent duplication events. In this approach, over-representation of a region with WGS reads was used as evidence of a duplication. The authors concluded that approximately 5% of the human genome consists of highly related duplications, and their study illustrates another application of combined genome-sequencing strategies.

All duplication event studies are faced with two main problems: the first is the loss of genes subsequent to duplication; and the second is the inability to detect similarities over deep evolutionary time, due to rearrangement events disrupting synteny and gene order, and accumulation of separate mutations causing sequences to diverge. The study of whole genomes from different species provides important assistance here. These studies can sometimes help to establish whether a gene has been lost in one species or gained in another. Here a consensus approach is used, as duplication of a gene independently in more than one closely-related species (or gene loss) is less likely than an ancestral event having given rise to the situation seen in both species.

In cases where sequences have diverged substantially and paralogy is less certain, genomic sequence can be extremely powerful in supporting the case for paralogy. Conservation of gene order, transcriptional orientation with respect to neighbouring genes and conserved exon structure (including intron phases) are indicative of paralogy in these circumstances. An example of this is seen in Chapter 6.

Much work has been conducted on paralogous genes, to study how they have diverged in function. Immediately following the duplication event, the paralogues would presumably share a common function and expression pattern, depending on

whether all relevant regulatory sequences were also duplicated. Over time, the paralogues will diverge and one might acquire a novel function or alternatively might become a non-functional pseudogene. To some extent, it might be possible to infer the function of one paralogue from information on the functional role of the other. The extent to which this is appropriate will depend on the extent and nature of the divergence at the protein-encoding level and of their expression patterns, both temporally and spatially.

An obvious question relating to paralogues is that immediately after two paralogues are created by a duplication event, why should an organism keep both copies? One possibility is that after the duplication the organism may not have any mechanism for removing the second copy. If such a mechanism exists, the answer will depend on what selective advantages are conferred by divergence of expression pattern and the proteins encoded, balanced against the rate of gene loss and the additional DNA synthesis and transcriptional burden placed on the organism.

Many studies have focussed on the divergent functions of proteins encoded by paralogous genes, such as the *Hox* genes. Recently attention has focussed on the hypothesis that mutations that affect protein function substantially are likely to occur at a lower rate than those in regulatory sequences bound by transcriptional factors, which are generally short motifs. This is thought to be because mutations within the short motif may be more likely to affect transcription factor binding than a mutation occurring within a larger protein coding region, much of which may not be functionally critical, causing a change in function.

In this way, divergence of transcription patterns could occur much earlier with loss of tissue-specific transcription factor binding sites (or reduced efficiency) in different paralogues. If the transcription pattern of the ancestral gene is now encompassed by transcription from the two separate genes, there will be selective pressure to keep both paralogues. This sub-functionalization hypothesis has recently been formalised by Force *et. al.* (Force *et al.,* 1999) as the DDC (Duplication, Degeneration and Complementation) model.

In order to test this hypothesis, expression patterns of paralogous genes need to be determined. Studies of expression patterns in large complex organisms are made more difficult by the problems of generating expression data at the sub-tissue or

temporal level. In this way, information regarding subtle differences of expression, which may be crucial functionally, is missed. Studies to date have still been informative though, and are also explored in Chapter 5 of this thesis. Recent advances in microarray technology, imaging techniques and improved tissue collections and probes promise further understanding in this area.

Studies of paralogues are also aided by the availability of data from model organisms such as the mouse and zebrafish. Both organisms are amenable to genetic analysis and to high-resolution gene expression studies. Genes studied in zebrafish in support of the DDC hypothesis include the *engrailed1* genes (Force *et al.*, 1999), and Na+/K+ ATPase α genes (Serluca *et al.*, 2001) and illustrate the utility of studies in model organisms. Other techniques such as gene knock-out/knock-in approaches and RNAi methodologies may also shed light on paralogous gene sub-functionalization.

In summary, the increase in availability of genomic sequence for organisms representing different evolutionary lineages and improvements in experimental approaches to studying gene expression and protein function offers an unprecedented opportunity to gain understanding of the evolutionary processes which have shaped genomes, and led to the diversity of life seen today.

1.4 Comparative genomic analysis

1.4.1 Identification of functionally important sequences

The limitations in using experimental and computational approaches in one species to identify functional elements in its genome have been described above. Comparing the genome sequences of different species provides an enormously powerful tool for identifying these functional elements, as these regions of the genome should diverge more slowly than other, non-functional regions.

With the recent availability of genomic sequence from a variety of organisms has come the development of tools for aligning these much larger segments of sequence and visualising sequence conservation. The choice of method is important for analysing orthologous genomic regions. Some methods, such as PIPMAKER (Schwartz *et al.*, 2000), use a local alignment approach to align different genomic sequences. This is particularly useful for detecting homology between regions that may have been rearranged with respect to one another. Other methods, such as Vista (Mayor *et al.*,

2000), employ a global alignment approach which retains information regarding overall arrangements of the regions, but will miss rearranged segments.

Combinations of these approaches yield optimal analyses of similarities between genomic sequences. In common with both of these broad approaches is the need for databases of repeat elements contained within the genomes of different organisms, to allow masking of these regions in the sequences under study. The RepBase database provided by Arian Smit is an excellent example of such a database, which has proved indispensable in the studies of genomic sequences (Jurka, 2000).

An example of the utility of these comparative sequence analysis approaches is shown by studies of the SCL locus, in which sequences of the loci in five different vertebrates were compared, revealing conserved elements important for function (Gottgens *et al.,* 2002). Promoter motifs were found that were conserved across human, mouse, chicken, pufferfish and zebrafish.

A key element in comparative genomic analyses for gene and other functional element identification is the choice of organisms to compare. Ideally, a balance must be reached whereby the chosen organisms are sufficiently closely related to retain significant sequence similarity in functionally conserved regions, and yet be sufficiently divergent that non-functional sequences do not appear as "noise" in alignments. In practice no single species will satisfy these requirements for all types of functional element.

Pilot studies using regions of genomic sequence from a variety of different organisms have been instructive in this regard, such as the comparative genomic sequencing being conducted for a variety of mammals and chicken at the NIH intramural sequencing centre (http:\\www.nisc.nih.gov) (Thomas *et al.,* 2003). BAC libraries are also being generated for a wide variety of species (http://bacpac.chori.org/) which will facilitate sequence-ready contig generation for defined regions in the absence of any larger scale sequencing initiative.

For gene identification purposes, comparison of mouse and human sequences displays considerable sequence conservation in non-coding sequences, and so whilst useful for detecting exons, potential false-positives are also seen. This non-coding conservation was used in the mapping of the mouse genome. Some groups have attempted to circumvent the problem of noise in the human-mouse alignments by using

a gene prediction algorithm which also takes sequence alignment data into consideration (e.g. TWINSCAN (Korf *et al.*, 2001) and SGP2 (Parra *et al.*, 2003)). Decreased false-positives can be achieved by using more distantly related organisms. The marsupial genome sequence represents a potentially useful resource in this regard, as it is more distantly related to humans than the mouse, and yet is perhaps not so distant that functionally important regions would not be recognised. An example is seen in Chapter 6 and in the study by Chapman *et al.* (2003). Chicken genomic sequence is also useful for identifying coding regions (Thomas *et al.*, 2003).

The true power of these approaches will perhaps be seen in the study of regulatory sequences. Difficulties in their computational prediction and experimental study may be partially overcome by the use of comparative sequence analyses to highlight regions for further study and reduce downstream effort. The complexities of gene expression regulation are formidable, as illustrated by the studies of regulation of the globin loci. Understanding of the regulation of expression of these genes has been gained over many years of careful study. Improvements in our ability to predict these regions would aid our understanding of this process.

1.4.2 Evolutionary studies

Comparative genome analyses on the whole-genome scale can also provide information on evolutionary relationships between species, how different chromosomes have evolved, and which rearrangements have occurred in different evolutionary lineages. Knowledge of the relationships between genomic regions of different organisms, shared synteny, allows inferences of likely orthology to be drawn from studies in different organisms. For example, knowing the syntenic relationships between human and mouse allows knowledge from genetic studies in mice to be applied in the search for genes implicated in a particular trait.

Many approaches have been used to generate data on shared synteny, such as the use of FISH, radiation-hybrid or HAPPY mapping techniques. The generation of genomic sequence data provides the highest possible level of resolution for comparative mapping. This is important when small-scale rearrangements have occurred within regions of shared synteny, which may not be detected by other comparative mapping approaches. Examples of this are seen in Chapter 4.

Comparative mapping studies have been particularly useful in attempts to elucidate the events involved in the evolution of the mammalian genome. With karyotypes ranging from three pairs of chromosomes in the Indian muntjac deer, to 67 pairs in the black rhinoceros (O'Brien, *et al.,* 1999), it is clear that many genomic rearrangements have occurred within the mammalian lineage. Techniques such as Zoo-FISH have been used to elucidate patterns of large-scale conserved synteny in some mammals (Chowdhary *et al.,* 1998). As mapping data are produced for mammalian species, from sequencing, cytogenetic, genetic and physical mapping approaches, they can now be related back to a high resolution human, mouse and rat genomic framework (Murphy *et al.,* 2001).

1.4.3 Value of comparative genomic analysis for functional studies

Comparative genomic analyses also aids in the establishment of orthology for genes in different species. Establishment of orthology between genes is helpful in studies aiming to understand the function of a gene product in one species, if functional information is available from studies in a related species. In the absence of comparative mapping data, establishing orthology is difficult when the representation of mRNA sequences is much lower for one of the organisms, and the gene in question may show high levels of similarity to multiple genes.

The ability to see genes in their genomic context allows the investigator to collate additional evidence to support orthology, such as the presence of neighbouring genes that are also shared between the different species, and conservation of transcription direction. When taken over a large region and supported by lower-resolution studies of chromosome conservation (such as by cross-species chromosome painting), this is persuasive evidence that the regions represent either orthologous segments or, less likely, a segmental duplication, and that the sequence similarities seen do not refer to paralogous sequences. As will be seen in Chapter 5, this can be beneficial when sequence similarity data alone may be misleading, due to processes such as gene conversion causing sequence homogeneity between genes within a species. In these cases, phylogenetic analyses suggest different relationships to those seen with the benefit of gene context information.

Genome sequence resources are now well developed for many important model organisms as discussed earlier. These will aid functional studies of genes in these

organisms directly, as well as shed light on the relationships between orthologues and paralogues. In these organisms, knowledge of the genomic sequence allows the researcher to search actively for different splice variants suggested by gene identification approaches (as discussed earlier) and can provide additional genetic markers (including SNPs) to refine linkage analysis and association approaches. It also provides further information for "knock-out" or "knock-in" approaches and facilitates identification of genes in gene-trap experimental approaches.

In all these cases, knowledge of the relationships between genes in different species is required to fully understand how the functions of those genes differ between species in different physiological settings.

1.4.4 Current and future prospects/projects

Future prospects in the field of comparative genomics include the expansion of number of species for which high-quality, comprehensive genomic sequence data are provided. Currently, organisms have been chosen depending on their prominence as important model organisms, or on their positions in the evolutionary tree. The production of genomic sequences from organisms with key, intermediary evolutionary relationships will further our knowledge of genome evolution. An example is *amphioxus*, which represents an early chordate lineage.

Another key future area for comparative genomics is that of within-species comparisons. It has already been noted that people differ in the copy number of certain genes, some of which may have important functional or medical consequences. Examples include the X-linked cone pigment genes responsible for red-green colour vision (Neitz and Neitz, 1995), and the cytochrome P450 CYP2D6 genes involved in the metabolism of drugs and other compounds (Agundez *et al.*, 2001). A project to sequence the ~4.6 Mb MHC regions of at least 8 different haplotypes is also already underway (Allcock *et al.*, 2002). As well as gross differences between individuals or closely related species, knowledge of the SNP differences between individuals is expected to facilitate refined association studies for complex diseases, and move further towards the goal of personalised medicine, through pharmacogenomic approaches. The development of a haplotype map has been proposed to aid these approaches. In each of these cases, studies are being made of how the genomes of individuals differ at a high resolution.

Epigenetic studies will also complement these approaches, by providing information about how individual genomes are "programmed". These forms of comparative genomic analyses between individuals will further our understanding on how differences in how the genome is "interpreted", affects phenotype.

Common to all these approaches is that by comparing genomic information between species or individuals, we can gain further understanding of how genomes and the genes within them have evolved, and on the functions of those genes. This is a more powerful approach than the study of a single genome, and allows useful transfer of information between studies. A more complete understanding will result.

1.5 The X chromosome

1.5.1 Human X chromosome overview

Comprising approximately 153 Mb of euchromatic DNA (Ensembl v16.33.1), the human X chromosome represents ~ 5% of the human genome. A sub-metacentric chromosome, almost its entire genomic sequence is now known. The GC content of the X chromosome, at 39%, is slightly lower than the genome average of 41%. It is also notably richer in interspersed repeats, containing approximately 57% repeats compared to a genome average of 45%. Much of this increase is attributed to an abundance of LINE1 elements, constituting 30% of the chromosome compared to a genome average of 17%. It appears to be relatively gene-poor compared to the genome average, containing an estimated 800-1000 genes (~ 3% of the human gene complement). Initially estimated from CpG island and EST RH-mapping approaches, this has been supported by transcript mapping approaches, as described in Chapter 3.

The X chromosome has been mapped and sequenced by a consortium of centres, led by the Wellcome Trust Sanger Institute (Figure 1-3), which has generated three-quarters of the finished sequence. The other centres involved include the Baylor College of Medicine (Texas, USA), the Washington University Genome Sequencing Center (St. Louis, USA), the Max Planck Institute for Molecular Genetics (Berlin, Germany), and the Institute of Molecular Biotechnology (Jena, Germany). The chromosome was mapped in bacterial clones using a combination of STS-content mapping and restriction-fingerprinting approaches, to identify BAC and PAC clones for sequencing. In the final phases of the project, YAC clones have also been utilised to close remaining gaps.

The human X chromosome has long been the subject of intensive study due to its unique biology as one of the sex chromosomes. Many disease genes have been mapped to the X chromosome, partly facilitated by the manifestation of many recessive conditions in males giving rise to a characteristic inheritance pattern.

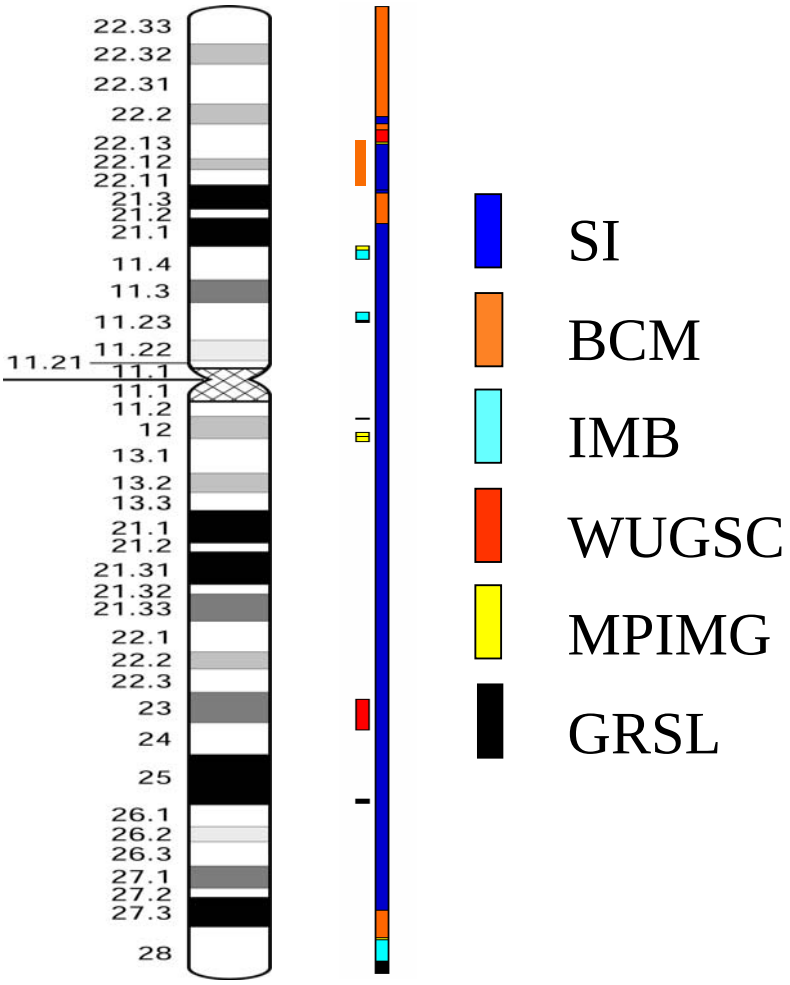

Figure 1-3 Overview of the regions being mapped and sequenced by each of the major sequencing centres. Illustration kindly provided by Dr. Mark Ross, Wellcome Trust Sanger Institute.

At the time of writing, 208 disorders showing mendelian inheritance have been mapped to the X chromosome (Ensembl v16.33.1 – data derived from OMIM), including well known disorders such as haemophilia A and colour-blindness. In particular, there appears to be an abundance of genes implicated in mental retardation located on the X chromosome, many of which have now been characterised (Frints *et al*., 2002).

One distinctive facet of X chromosome biology is X inactivation (Lyon, 1998). As female cells possess two copies of the X chromosome, and males only one copy, a dosage-compensation mechanism has evolved whereby one of the female's X chromosomes is inactivated (Lyon, 1999). Our current understanding of the mechanism is that early in female embryonic development, the X chromosomes are somehow "counted", and all but one of the X chromosomes are subjected to inactivation (Avner and Heard, 2001). The utilisation of a counting mechanism is illustrated by the observation that in XXX cells, two of the X chromosomes are inactivated.

The inactivation mechanism involves coating of the inactive X by the non-coding RNA transcript of the *XIST* locus, which is expressed from the inactive X chromosome (Brown *et al*., 1991). The *XIST* gene is located at the X inactivation centre (Xic) at Xq13. As the choice of X chromosomes for inactivation is usually a random process, either the paternal or maternal chromosome may become inactivated. All descendants of the cell in which the decision was made inactivate the same X chromosome. As the embryo matures, this leads to mosaicism with patches of cells containing either an active paternally or maternally inherited X chromosome. This in turn can lead to mosaic phenotypes, the most striking example of which is the coat patterning of piebald cats. Male spermatogonia also inactivate their single X chromosome, but this becomes reactivated during the production of sperm.

Whilst X inactivation appears to have evolved as a way of maintaining X chromosome gene dosage between sexes, there are a number of genes that escape inactivation (Brown and Greally, 2003). In some instances these are genes with functional homologues on the Y chromosome; for others it may simply be the case that dosage of the gene involved is not important or that the gene has not yet been drawn into the process of inactivation.

1.5.2 Sex chromosome evolution

A key question regarding the X and Y sex chromosomes is - how did they evolve? The X and Y chromosomes are morphologically distinct, and are very different in their gene and repeat composition. However, the human X and Y chromosomes share various regions of homology. Two of these - the major and minor pseudoautosomal regions (PARs)(Rappold, 1993) - enable the sex chromosomes to pair during male meiosis (Figure 1-4). There is an obligatory recombination during meiosis in the ~ 2.5

Mb major pseudoautosomal region (PAR1). This equates to an exceedingly high recombination rate of 20 centiMorgans (cM) per Mb. Recombination within the smaller (~320 kb) minor pseudoautosomal region (PAR2) is not obligatory but is still elevated above the genome average. Between PAR1 and PAR2, which reside at the ends of the chromosome arms, there is normally no recombination in male meiosis.

In addition to these two regions, there are other regions of homology shared between the sex chromosomes. For example there is an XY homology block located at Xq21 and Yp (Sargent *et al.,* 2001). Many X chromosome genes also have homologous counterparts on the Y chromosome.

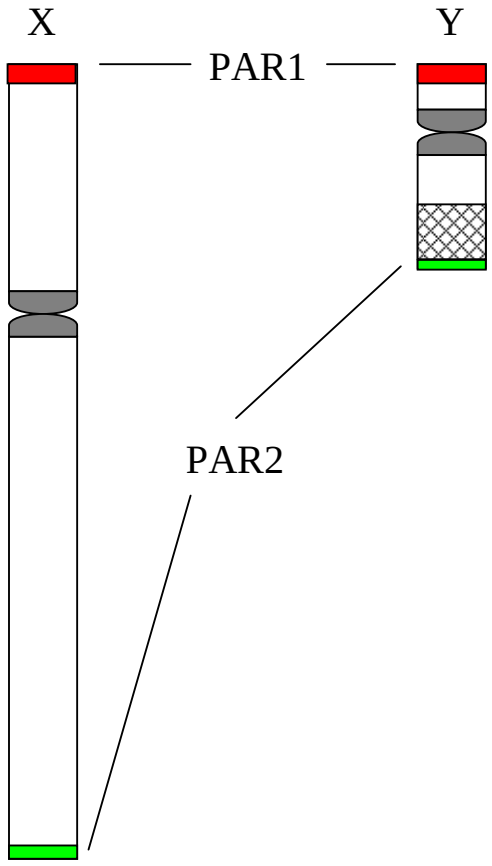

Figure 1-4 Schematic diagram of the human X and Y chromosomes. PAR1 and PAR2 are illustrated. Heterochromatic region of the Y is displayed as a hatched box.

The current theory of mammalian sex chromosome evolution suggests that the X and Y chromosomes were initially a pair of homologous autosomes. When one chromosome, the Y, gained a major sex-determining region (thought to be SRY), it would lead to a need for reduction in recombination between the two chromosomes in order to maintain sex differences.

As the pair of chromosomes evolved, lack of recombination between them would lead to the genetic isolation of most of the Y chromosome, whilst the X chromosome was still able to pair and recombine in females. This isolation of the Y led to its gradual degradation, a process termed "Muller's ratchet", as genes accumulated mutations and became inactive and genetic material was lost. In concert, the evolution of the X inactivation mechanism described earlier would ensure that dosage was largely conserved for X-linked genes between the sexes. Indeed, it has been proposed that the Y is heading for extinction, although this could overlook unknown or poorly understood aspects of Y chromosome biology (including the recent suggestion that the Y chromosome recombines with itself in palindromic regions instead of with a homologous chromosome).

Studies of the sex chromosomes in monotremes (prototheria) and marsupials (metatheria) have shed further light on the evolution of the mammalian sex chromosomes. Mammalian evolutionary relationships are summarised in Figure 1-5.

The X chromosome in marsupials is considerably smaller than that of eutherian mammals, accounting for approximately 3% of the genome, and the Y chromosome is tiny (~ 10 Mb) (Toder *et al.*, 2000). The marsupial X and Y chromosomes do not appear to pair at meiosis and there is no evidence that they possess pseudoautosomal regions. Monotremes differ, in that both the X and Y chromosomes are large, and pairing takes place between the entire short arm of the X and the long arm of the Y. In addition, monotremes appear unique amongst animals studied to date, due to the presence of a number of small chromosomes which form a paired end-to-end chain, pairing to the Y chromosome short arm, at meiosis.

X chromosome inactivation is known to occur in marsupials, but inactivation is imprinted, always affecting the paternal X chromosome. This imprinted inactivation has been also been observed in the extra-embryonic tissues of the mouse. It remains to be established whether or not X chromosome inactivation occurs in monotremes, owing to the lack of a chromatin body in female cells.

Comparative mapping studies between monotreme, marsupial and eutherian X chromosomes has revealed that many genes show conserved synteny, constituting an X Conserved Region (XCR). However, many of the genes located on human Xp are found to be autosomal in monotremes and marsupials (Spencer *et al.*, 1991). This was

further demonstrated by a striking experiment in which a marsupial X chromosome 'paint' probe was hybridised to human metaphase chromosomes (Glas *et al.*, 1999). This clearly demonstrated that homology was restricted to Xq and a small proximal region of Xp. This is depicted in Figure 1-6. PAR1 is also likely to have been added to the X during eutherian evolution.

In contrast, the synteny for the X chromosome (if not gene order) is remarkably conserved within the eutherian lineage, possibly due to the strong evolutionary pressures to maintain synteny once genes were recruited into the X inactivation system (Ohno's Law).

The most parsimonious explanation for these observations is that during the eutherian radiation, there was at least one addition to the X and Y chromosomes from autosomes. These studies demonstrate the utility of comparative mapping approaches in the elucidation of events during evolution of the mammalian sex chromosomes.

1.5.3 Human Xq22-q23

When studies for this thesis began, Xq22-q23 was the region with the most comprehensive coverage of finished genomic sequence and presented an ideal region for sequence-based gene identification and annotation approaches. Comprising approximately 15 Mb of DNA, the region spans light, dark and intermediate Giemsa-staining bands. As such, it could be expected to display a heterogenous gene density, isochore content and repeat content. These features combined to make the region ideal for preliminary studies as human X chromosome sequencing progressed.

The Xq22-q23 region has also been shown to contain genes responsible for a variety of Mendelian, inherited disorders, some of which have been cloned. At least 13 mendelian-inherited conditions are associated with Xq22-q23 (OMIM). Gene identification and annotation studies of the region would therefore be of potential benefit for studies aiming to identify the remaining, uncloned disease genes.

A first generation transcript map of the region was published by Srivastava *et al*., (1999). In addition, other genes had been localised to the region in other studies including the Gene Map '99 RH mapping effort (Deloukas *et al.*, 1998). It was reasoned that a genomic-sequence based approach would be complementary to these studies.

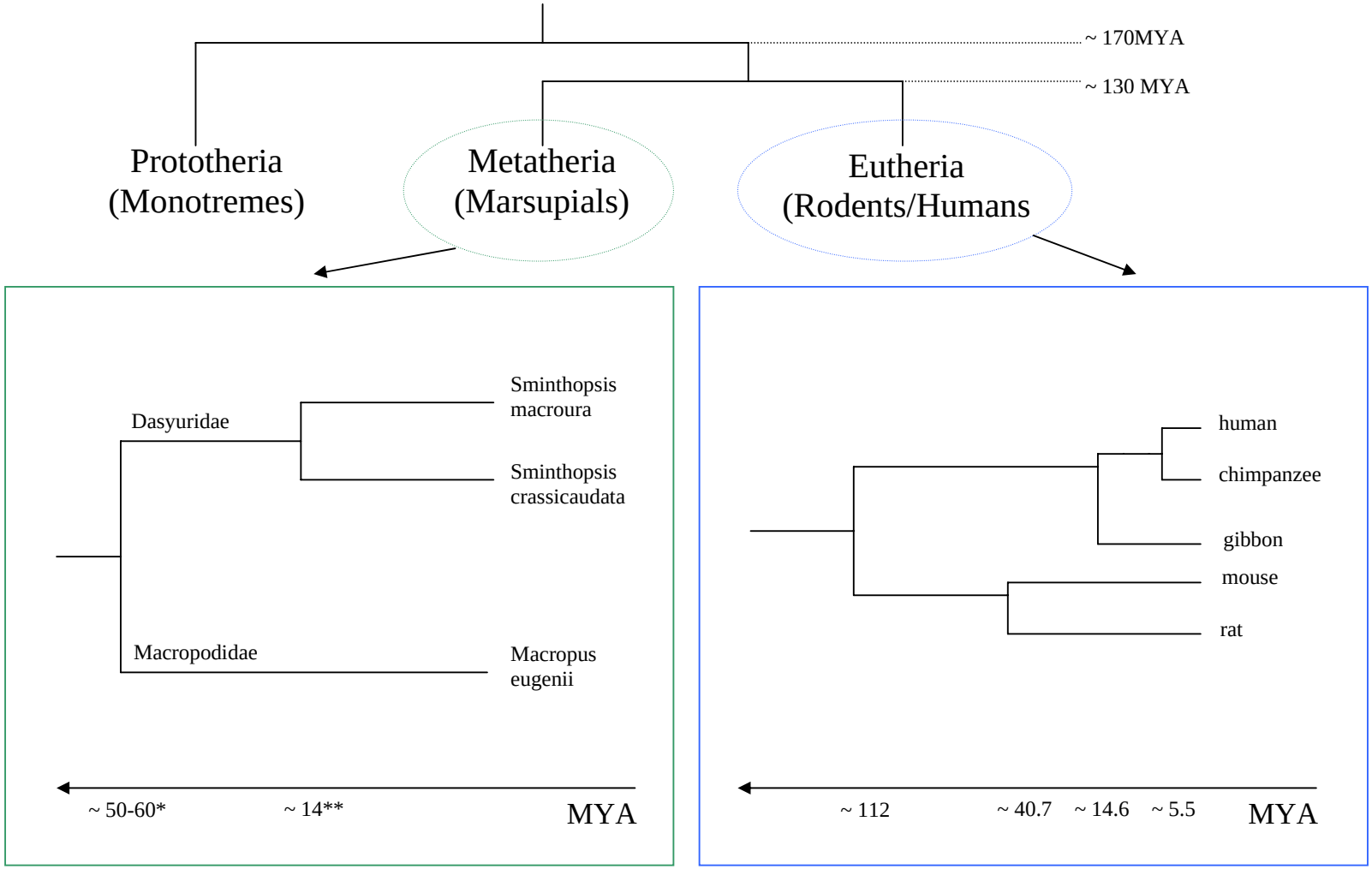

Figure 1-5 Figure depicting selected aspects of mammalian phylogenies. Timings of divergence (except therian divergences) are according to Kumar and Hedges, (1998) except *(Graves and Westerman, 2002) and ** (Blacket *et al.*, 1999).

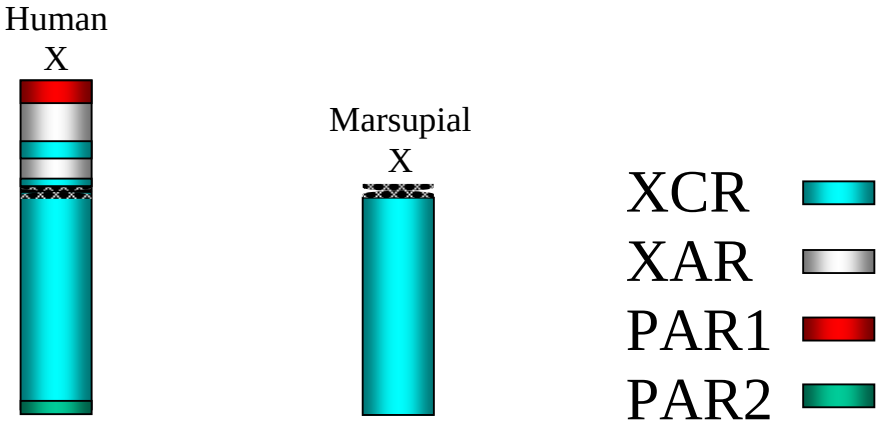

Figure 1-6 Diagram illustrating conservation between human and marsupial X chromosomes. X-conserved regions (XCR), regions added to the X during eutherian evolution (XAR) and pseudo-autosomal regions (PAR1 and 2) are indicated.

The X chromosome is well conserved in gene content between human and mouse, and indeed in all eutherian mammals studied to date, for reasons discussed in the previous section. Gene order is not always conserved on a large scale however, particularly between human and mouse, due to multiple inversion events in the different lineages. Human Xq22-q23 has shared synteny with the E3-F2 region in mouse, and this region has been shown to contain genes leading to mutant phenotypes. An example is the "jimpy" mutant, caused by defects in the Plp gene (Sidman *et al.*, 1964). Investigation of the shared synteny between human and mouse in this region would reveal detail about the fine-structure of conservation present. It would also prove useful in the transfer of information from observed phenotypes in mouse in searches for possible causative genes for disorders mapped to human Xq22-q23, particularly from imminent or future mouse saturation-mutagenesis studies.

In summary, Xq22-q23 provided an ideal region for pilot sequence-based studies of gene identification and annotation on the human X chromosome, which could then be applied on a chromosome-wide basis as sequencing of the chromosome progressed. It would also provide important information for searches for genes involved in disorders mapped to the region, as well as gene structure information for those genes already shown to be causative in disorders, to facilitate mutation screening and studies of the biology of the genes.