Chapter Seven - Discussion

7 Discussion

7.1 Summary of thesis results and discussion of major themes

The studies presented in this thesis were conducted over a period of rapid development in the field of genomics. Shortly after their beginning, the first complete sequence of a human chromosome was published (Dunham *et al.*, 1999). Shortly after their conclusion, the publication of the finished human genome sequence was announced. During that time other developments occurred, such as releases of sequences from large-scale cDNA sequencing projects, production of a physical map and draft sequence of the mouse genome, and the sequencing of genomes of organisms such as *Ciona intestinalis* (Dehal *et al.*, 2002), *Fugu rubripes* (Aparicio *et al.*, 2002) and *Arabidopsis thaliana* (Arabidopsis Genome initiative (2000)).

Chapter 3 described the production of a transcript map of the human Xq22-q23 region. This study provided an illustration of how effective the combination of the genomic sequence and results of large-scale cDNA sequencing projects is for gene-identification. Other aspects of genome structure only revealed through knowledge of the genomic sequence were also outlined. These included an inverted duplication with a very high level of sequence similarity, containing a gene from a family under intensive investigation, as well as an example of how the mitochondrial genome has integrated into the nuclear genome during evolution. Aspects of transcriptional control were briefly explored for genes with alternative polyadenylation sites.

Finally, and most importantly, the process of rigorous study of the genes in their genomic context revealed the presence of high numbers of duplicated genes, with some families contained within the region, some having paralogues on autosomes, and others sharing paralogy with genes on Xp. This finding stimulated the experimental and computational studies described in the subsequent chapters of the thesis, which were aimed at furthering our knowledge of the evolution of these gene families.

In Chapter 4, the generation of a sequence-ready BAC contig was described spanning *Mus musculus* X E3-F2, which is equivalent to human Xq22-q23. This comparative mapping was undertaken in order that the extent of sequence duplication within the corresponding region of the mouse genome could be assessed. It also contributed to efforts underway to sequence the whole mouse X chromosome. From analysis of the sequence generated from this region, it was found that many of the duplicated genes were also present in the mouse genome, indicating events which occurred before the divergence of the human and mouse lineages. Whilst many of the genes were conserved between human and mouse (as described in previous, lower-resolution studies), some differences were described including the KIR3DL1 gene, which is autosomal in human and rat. The mouse region studied also contained non-interspersed sequence repeats not detected in human.

In Chapter 5, the duplicated genes whose paralogues resided within Xq22 were studied in detail. Their degree of relatedness was assessed through sequence-based phylogenetic analysis, and their arrangement in comparison with the mouse X E3-F2 region was considered. In some cases, these analyses indicated orthologous relationships between human and mouse genes. However, in other cases, there were revealed potential examples of gene conversion between loci. This emphasised the benefit of being able to study the genes in their genomic context, revealing the details of paralogue proximity and orientation, and underscored problems with inferring orthology and paralogy based on sequence similarity alone. Studies of the expression patterns of the genes found patterns ranging from paralogues with similar, ubiquitous expression to those with more restricted patterns, with some paralogues showing differences within a family.

Overall, these studies described the striking degree of paralogy found within the Xq22 region and provide further avenues for targeted research into the evolution of the region and divergence of paralogue function and expression. As several of the genes across different families had been functionally characterised to some degree, this information can now also be used to focus studies for the various uncharacterised paralogues.

Chapter 6 presented evidence for a hypothesis suggesting a segmental duplication leading to the generation of paralogues with copies arrayed between the Xq22 region and Xp. This expanded further on previous observations and hypotheses, and the ability to examine the genes in their genomic context again proved its value in supporting the segmental duplication model. To provide further information on the evolutionary history of the Xp/Xq22 paralogous blocks, marsupial orthologues were identified and their genomic localisation determined. This confirmed previous studies which had shown that much of the region constituting human Xp is autosomal in marsupials (Glas et al., 1999), and demonstrated that many of the Xp paralogues were co-localised in the marsupial genome.

In order to try to estimate the minimum age of the duplication event leading to Xp/Xq22 paralogy, use was made of the recently completed draft sequence of *Fugu rubripes* and of the available literature for the paralogous genes. Phylogenetic and genomic organisation evidence suggested that the duplication occurred before the divergence of lineages leading to *Fugu* and humans approximately 430 Mya, and possibly before the divergence of cartilaginous and bony fish approximately 530 Mya. Comparative sequence analysis also supported this hypothesis, and demonstrated differences in sequence composition between the human Xp and Xq regions, and between marsupial, human and mouse.

An underlying theme throughout these studies has been illustration of the benefit of the availability of genomic sequence, particularly long-length finished sequence. Initially providing a basis for a comprehensive transcript map of a genomic region, the ability to see genes in their genomic context revealed aspects of the chromosome's evolution and biology that may not have been otherwise observed. The context information in itself provided evidence supporting a hypothesis for a model of evolution leading to generation of paralogy between Xp and Xq22.

The extent to which a gene identification strategy using mRNA and genomic sequence can be implemented is naturally limited by the availability of both types of data. The human genome sequence is now almost complete, and the efforts of several large-scale cDNA sequencing projects have provided large amounts of useful data, from which many genes have been identified and annotated. These have added to data on specific genes generated over the years by many investigators.

A limitation of any cDNA-based approach is the nature of generation of the cDNA, from RNA derived from tissue samples. The availability of different tissue types can be limiting, especially in humans. For example the cDNA libraries employed in gene identification in Chapter 3 omit different tissues that may express a putative gene. The sensitivity of RNA to degradation and the nature of steps involved in cloning of cDNAs can also result in incomplete cDNA representation of the mRNAs from a tissue. Finally, some mRNAs may have temporally-restricted expression which again may result in incomplete mRNA representation.

These limitations may be partially circumvented by the availability of genomic and mRNA sequence information from other organisms. As discussed in earlier chapters, this information can be used to identify conserved regions of genomic sequence, suggesting functional roles for these sequences (Gottgens *et al.*, 2002). For example, in Chapter 4, comparisons were based on gene annotation conducted independently in human and mouse. A direct comparison of the two regions may reveal further conserved regions indicative of genes. Tools such as TWINSCAN (Korf *et al.*, 2001) could be implemented in such an approach. Tissue availability is less of an issue with model organisms and a more comprehensive array of RNA samples can be accessed, thus leading to a better representation of the transcriptome.

A limiting factor in genomic sequence comparisons for different organisms is the degree of relatedness of the organisms being compared. If the organisms' genomes are not sufficiently divergent, the conserved regions may be hidden by "noise". This is illustrated in the comparative analysis of the CSTF2/NOX1/XK-L loci described in Chapter 6. If the genomes are too divergent, though, identification of conserved sequences will be problematic, especially when employing algorithms designed to analyse large amounts of sequence which may have to trade sensitivity for pragmatic reasons of computational intensity.

Gene annotation has been pioneered on organisms such as yeast, bacteria and worms, and has recently been applied on a genome-wide scale for a variety of higher organisms through initiatives such as the Ensembl project (EBI and Wellcome Trust Sanger Institute), Genome Browser (University of California at Santa Cruz) and genome resources at the NCBI (NIH). These initiatives have addressed the issues of tracking draft and finished genomic sequence records and their subsequent revisions, analysing the sequence using a variety of gene and repeat identification algorithms and combining results of similarity searches of mRNA and protein databases with the genomic sequences. Incorporation of other sources of data such as SNPs is also performed, and all of these data are made available via searchable graphical interfaces.

These initiatives have made great progress in facilitating use of the genomic sequence information by investigators worldwide. Indeed, anybody with access to the internet can view the genomic sequences and their annotation for organisms such as human, mouse, fruitfly and pufferfish, and use was made of this in the studies described in this thesis (for example the *Fugu* analysis in Chapter 6). A more complete annotation of genome sequences combines approaches such as these with manual annotation. For example, algorithms used to align mRNA sequences to genomic sequence can miss very small exons, and can fail to identify splice sites correctly. Assessment of how complete a gene structure appears to be is also best achieved through manual inspection. Finally, features such as unusual gene structures, as seen for the NXF2 gene (Chapter 3), and context-dependent features such as high levels of paralogy (Chapters 3-6) may be missed by purely computational approaches.

A combined approach to human genome annotation is now being employed at different centres, such as by the HAVANA (Human And Vertebrate ANalysis and Annotation) Group at the Wellcome Trust Sanger Institute. Such annotations are being collated in databases such as the VEGA (VErtebrate Genome Annotation) database. Gene annotation is also being complemented by targeted gene identification efforts for human, similarly to those described in Chapter 3, by the EGAG (Experimental Gene Annotation Group) group at the Wellcome Trust Sanger Institute, and by large-scale mRNA sequencing efforts as described in Chapter 1.

Combined with large-scale cDNA sequencing data and comparative analysis of genomes of different organisms, this provides a powerful approach for the comprehensive annotation of genes within a genome.

Knowledge of gene structure is of interest for several reasons. For one, it provides insight into how genes have evolved in different organisms under different evolutionary pressures. A key question in genetics remains as to how genes have evolved and the mechanisms of their transcription. Were introns always a feature of gene structure or have they evolved from precursors? The “intron-early” and “intron-late” hypotheses are still the subject of debate. Elucidation of the gene structures of all genes within the genome of an organism will provide comprehensive information regarding ranges of exon and intron sizes, and how they vary within different regions of the genome with different sequence compositions. Within the human genome for example, introns appear to be longer in Giemsa dark-band regions.

It has also been suggested that intron-length may be related to level of transcription of the gene. This would be logical, as it would take less time and energy to transcribe a shorter gene, secondary structure effects notwithstanding. Indeed, knowledge of gene structure is fundamental to studies aimed at understanding gene transcription, as *cis*-acting elements for transcription control such as promoters and enhancers are not fully represented in the transcriptome (although such elements may occur also in exonic and intronic regions).

It is also apparent that we probably still have much to learn regarding interpretation of genes by the process of transcription. The recently discovered process of trans-splicing in *Caenorhabditis elegans* (Blumenthal, 1995) and the growing research into the process of RNA-editing have served to remind us of this, in addition to active research into transcription control and mRNA localisation elements contained within untranslated portions of RNA transcripts. Gene structures provide a framework on which to base these studies, and the genomic sequence provides the information which contains within it the cis-acting elements involved in transcription.

Knowledge of gene structures is also valuable in studies attempting to identify genes involved in genetic disorders, and genetic differences between individuals giving rise to different traits. In order to focus methods for detection of genetic differences, information regarding exon/intron structure is invaluable in designing assay reagents (for example primers designed to amplify an exonic region for sequencing).

Finally, gene structure information is useful in identifying genes that have shared a common ancestor prior to gene duplication events, particularly in instances where the sequences have diverged sufficiently to be in the “grey-area” of homology. In such instances, aspects of gene structure such as exon sizes and intron phase can provide evidence of common origin. Exon size information provided compelling evidence for the paralogy discussed in Chapter 6.

The main aspect of this thesis has centred on the discovery and characterisation of extensive paralogy within Xq22 and between Xp and Xq. Studies of gene duplication have benefited greatly from the large amounts of genomic sequence now available. Although questions remain as to predominant mechanisms of duplication and functional importance of paralogues, it is clear that gene duplication is an important feature of all organisms whose genomes have been studied to date, from bacteria to humans.

One intriguing feature of gene duplication in the human genome is the apparent heterogeneity of the distribution of gene duplications. To date, the most intensively studied regions of gene duplications have been the Hox gene clusters and the MHC regions. Prior to elucidation of the genome sequence and annotation of genes, it would not be clear if gene duplications were indeed enriched in these regions, or if study of these regions had led to an ascertainment bias.

The gene duplications within Xq22 were striking in that in adjacent regions of the chromosome, multiple duplicated genes were not noted. However, the X chromosome does in fact contain a variety of duplicated loci (e.g. MAGE genes). It is possible that the X chromosome may be enriched for segmental duplications as it is largely unpaired at meiosis in males, and there may arise an opportunity for increased rearrangement. Complete annotation of the human X chromosome and comparative analysis with other organisms may shed light on this hypothesis.

Gene duplications provide challenges regarding assessment of function of gene products. In cases where sequence similarity is high, assay platforms such as hybridisation techniques may provide “composite” information regarding gene expression, for example. Mutation screening of genes may also be affected depending on the strategy used. Knowledge of the existence of paralogues provides useful information to the investigator in this regard. Examples where this could be an issue have been described in this thesis. In particular, the two thymosin-beta genes within Xq22 encode identical proteins, and any differences between the loci conferring selective advantages to the retention of both gene copies would more likely act at the level of transcription of the genes. Alternatively there may be selective advantages in keeping both copies similar, perhaps via gene conversion. Although the mechanisms of gene loss are incompletely understood and are also factors in retention of duplicated loci, these two loci appear to have been conserved over at least 90 million years of evolution.

In summary, sequence duplications are an indication of the dynamic nature of a genome, and how it may adapt to different selective pressures. Availability of genomic sequence allows careful and comprehensive study of these loci, which may otherwise be refractory to investigation, and provides important information regarding their context and distribution within the genome of an organism.

7.2 Future directions

Future directions for gene identification and annotation have been partly discussed in earlier chapters and sections. It is unlikely that there will be major advances in overcoming the inherent difficulties of working with human RNA samples, but large-scale sequencing efforts or SAGE analysis (Sun *et al.*, 2004) of cDNA libraries from a range of different tissues will continue to provide valuable resources. These will also aid in unravelling complexities of transcription, such as alternative splicing of pre-mRNAs. Whilst EST sequences can provide useful evidence of alternative splicing events, for full elucidation of the transcript sequence and hence gene structure they are often too short to allow definitive conclusions to be drawn. The cDNA clones themselves will also provide a resource for functional studies of the gene products, as in the *C. elegans* “ORFeome” effort (Reboul *et al.*, 2003).

One area of gene identification that would particularly benefit from development of methodology is in the representation of the 5’-ends of transcripts. It has proved difficult to ensure faithful representation of the beginning of an mRNA transcript for various technical reasons. However, representation of the 5’ end of a transcript is crucial in the annotation of core promoter regions. Truncated transcripts may result in erroneous annotation of promoters. Some advances have been made in this area, but it still remains a difficult topic and will benefit from continued research.

To fully characterise the genes of an organism, aspects such as alternative splicing, RNA-editing, alternative polyadenylation site usage, alternative promoters, enhancers, mRNA localisation signals and sequence elements affecting mRNA turnover and translation all need to be taken into account. This is in addition to any other protein-related aspects to be considered, such as post-translational modification. It is clear that there is still much to be investigated in these areas. As genes are studied further in different organisms, we will learn more regarding the use of different aspects of genetic organisation and control in different species, and those elements that are shared.

The study of gene duplications will benefit greatly from increasing genome mapping and sequencing of a variety of organisms from different evolutionary lineages and with different degrees of divergence. Gene duplication, via single gene tandem duplications and segmental duplication, and gene loss are balancing factors in shaping the genome of an organism. Extended knowledge of the genes within different organisms will aid in the understanding of the relative balances of these factors at different loci. Species-specific differences can also then be taken into account when attempting to elucidate the molecular evolution of paralogues and their functions, and in interpreting data derived from animal models.

Comprehensive annotation of highly related loci will also provide useful data for the experimental studies addressing functions of these loci. It may avoid confusion due to “composite” data being produced from highly related loci, and will highlight cases of possible functional redundancy. It will also provide a framework on which to base studies of differences in copy number of paralogous loci between individuals. As SNP data become available, this would be complemented by large-scale analysis of the gene complement of individuals, which may contribute to different traits, disorders, susceptibility to disease and response to drug treatment. It would be interesting to assess the composition of the Xq22 region in different individuals as an example.

As annotation and characterisation of paralogues within a genome progresses, studies of why some regions are rich in gene duplications will be aided. For example, why should so many duplicate genes be found within Xq22? What features of the genome have affected the evolutionary dynamics to cause this? How stable is Xq22 (and its mouse counterpart)? In this regard, it is of interest that the proteolipid protein (PLP) gene in Xq22, is often duplicated in Pelizaeus-Merzbacher disease. Whilst this may be unconnected to the presence of multiple repeat families (some of which are non-genic) within the region, the repeats may predispose to genetic rearrangement. In some cases however the relevant sequence features may have become unrecognisable if the gene duplications occurred a long time ago. The process of gene conversion, possible examples of which are also seen within Xq22, also clearly has to be taken into account in these studies. Analysis of the genes in other primates may shed light on whether this is the case (Rozen *et al.*, 2003).

The evidence provided for the segmental duplication generating the paralogues seen between Xp and Xq22 has raised further questions related to X chromosome biology and evolution. We have demonstrated that the segmental duplication appears to be a relatively ancient event, and that marsupial orthologues of the Xp paralogues are autosomal and co-localised (for those studied). Following the duplication, one copy (represented at Xq22) was localised on the autosome that in mammals became the ancestral X chromosome, whilst the second copy (represented at Xp) remained autosomal. It is unclear what the initial arrangement of the regions following the duplication was, and what rearrangements occurred subsequently.

It is also unclear whether the region of extensive paralogy *within* Xq22, which is flanked by genes with paralogues on Xp, was also part of the segmental duplication. A similar question remains as to the origin of the large non-paralogous region separating the two regions of Xp paralogues. Were these regions originally present in the duplication and have subsequently undergone different rates of gene loss or rearrangement, or have they been acquired more recently by the paralogous regions? The large number of non-paralogous genes interspersed throughout the blocks of paralogy are part of this same question. Given the apparent considerable age of the duplication, there have presumably been large numbers of both gene gains and losses which have degraded the original segments. Further study of these regions in the marsupial and more distantly related organisms may provide useful information to address these questions.

Finally, whilst initial translocation of the Xp paralogues to the X chromosome from an autosome would not result in problems of dosage if they were also transferred to the Y (i.e. into an existing pseudoautosomal region), subsequent degradation of their copies on the Y would result in dosage imbalance between sexes unless they became recruited into the X inactivation system. This is an issue relevant to any genes translocated to the sex chromosomes. It is unclear what the inactivation statuses of many of the genes are at present. If inactivated, what features did they gain (or lose?) in order to accomplish this?

As the X chromosome sequence and annotation approaches completion, the extent of paralogy within the chromosome can be fully assessed, and compared to other regions of the genome. It also paves the way for further studies of individual variation, gene functions and chromosome evolution. This thesis has illustrated how the genome sequence can be used to view not only the content of genomic regions, but also their contexts. Both perspectives are required to begin to understand fully the biology of chromosomes and the evolution that has shaped them. It is also clear from the studies presented here that there is much still to be explored in this area, and our understanding is far from complete. The unique biology and structure of the X chromosome ensures that the process of addressing these questions will be a fascinating one.