Chapter 7

Discussion

7.1 Advances in mapping genomes using bacterial clones

7.2 Mining the human genome sequence

7.3 Comparing different genomes to aid human genome sequence analysis

7.4 Functional analysis of gene products

7.5 Conclusion

7.1 Advances in mapping technology and strategy

The strategy for mapping the human genome was developed using the methods and experience gained from mapping the genomes of model organisms. Bacterial clone maps covering the genomes of both *C. elegans* and *S. cerevisiae* were constructed by restriction digest fingerprinting whole genomic libraries of clones, which were assembled into contigs. For *C. elegans*, once the bacterial clone resources had been exhausted, the remaining gaps were bridged using YACs. The increased complexity of the human genome (30 times larger than the *C. elegans* genome) meant that the strategy implemented for mapping the *C. elegans* genome could not be directly applied to the human genome. For instance, whole genome fingerprinting of a human cosmid library was not feasible given the increased size of the human genome. This increase in genome complexity would have meant a large increase in the number of cosmids required (approaching half a million cosmids for a six fold coverage of the human genome), which in turn would have made the interactive contig assembly in FPC too laborious.

By contrast, strategies for generating clone maps covering small regions of the human genome relied on the use of YACs from the outset. Mapping regions of the human genome involved the use of landmarks taken from previously published genetic maps or RH maps which were used to identify and order YAC clones for contig construction. This approach was scaled up to construct maps on some chromosomes but attempts to adapt the whole genome fingerprinting strategy employed for the model genomes failed because of the instability and chimaerisms that was present in the YACs. YACs were also considered to be inappropriate substrates for sequencing,

compared to bacterial clones. As a result, work on the human genome was concentrated on finding ways to construct bacterial clone maps. Early on, cosmid contigs were constructed using data from the available YAC contigs and associated landmarks. For example, whole YACs were hybridised to gridded cosmid libraries to identify underlying cosmids. The cosmids were then fingerprinted to assemble contigs using the same method of restriction digest fingerprinting that was developed for the *C. elegans* mapping project. In one case, this approach resulted in the assembly of contigs covering 80% of the 450 kb region, and the bacterial clone contigs were bridged by the starting YAC (Holland, J.*, et al.*, 1993). This integrated YAC-cosmid map therefore closely resembled the product of the *C. elegans* mapping effort, although it was generated in a different way. This approach was initially employed as the basis for constructing the Xq22 contig described in chapter 3, in which cosmids were identified using probes derived both from YACs and from the available landmarks ordered in the YAC contig. Cosmids were then assembled into contigs using restriction digest fingerprinting. The progress was slow given the small size of cosmids (40 kb) and resulted in only 50% coverage of the 6.5 Mb region in cosmid contigs.

A major advance in human genome mapping was the development of two larger insert bacterial-based cloning systems (PACs and BACs) with the combined advantages of a much larger insert size (up to 300 kb), and little or no chimaerism or instability in contrast to that seen in YACs. This development was reflected in phase 2 of the Xq22 project, in which the focus of the mapping switched to PAC identification. Cosmids at the ends of contigs were labelled and used as hybridisation probes to isolate PACs. The PACs were fingerprinted and incorporated into the existing contigs. Difficulties

arose in using a combination of small and large clones together in terms of both fingerprint comparisons and in the selection of a minimum set of clones for sequencing. For fingerprinting, false overlaps between PACs were observed and true overlaps between PACs and cosmids were missed; for minimum set selection, the large insert clones made the cosmids largely redundant. Increasingly, sequence-ready contigs were constructed in new regions of the genome using only the larger insert PACs and BACs, including the region discussed in chapter 4, where contigs were constructed as part of the whole X chromosome mapping project.

At this stage in the human genome mapping project, bacterial clone contig construction was still reliant on landmarks ordered on available YAC contigs. On some chromosomes, including chromosome 22 and the X chromosome, the continued development of YAC contigs resulted in the provision of a sufficiently high density of marker (greater than 1 per 70 kb for chromosome 22 – Collins, J. E. *et al*, 1995) to enable construction of PAC and BAC contigs directly. This was illustrated in phase 3 of the Xq22 project, where new landmarks which became available from detailed YAC maps were used to identify new PACs. Sufficient coverage of the region was obtained to allow closure by walking in the final phase (phase 4) of the Xq22 map.

YAC maps, or later RH maps, of similar quality provided a high density of ordered landmarks along each chromosome and provided the means to apply the same strategy to map the rest of the human genome. However, the increased insert size of BACs and PACs compared to cosmids led to a further development late in the project. Whole genome fingerprinting, which had been used successfully to map the *C. elegans* genome in cosmids, was introduced using BACs to obtain very rapid coverage of the

rest of the human genome independently of landmarks, thus reducing the number of landmarks required to anchor and orient the bacterial clone map.

Walking to close gaps in the human genome was aided by the generation of sequence at the ends of BAC clones. These sequences provided a large resource of new landmarks both within contigs to confirm fingerprint assemblies but also at the ends of contigs to screen for additional BACs for gap closure. Bacterial clone maps now cover most of the euchromatic portion of the human genome, and the remaining gaps are being closed. An interesting development is the return to YACs to bridge gaps for which there is no bacterial clone coverage, mimicking the final gap closure carried out in the *C. elegans* mapping project.

The advances in bacterial clone mapping and sequencing that evolved for the construction of the sequence-ready maps for the human genome are now being applied to other organisms, particularly mouse and zebrafish. For mouse, whole genome fingerprinting and assembly of mouse-derived bacterial clones generated 7,500 contigs which were extended and joined to form less than 400 contigs covering approximately 90% of the mouse genome. Mouse BAC end sequences were used to align the mouse contigs to the human genome sequence and accelerated the manual joining process (Gregory, S., unpublished, http://mouse.ensembl.org/). A similar project is now well underway to generate bacterial clone contig maps of the zebrafish genome (see http://www.sanger.ac.uk/Projects/D_rerio). The speed and accuracy with which these large genomes are being mapped is a reflection of the technologies applied to map the human genome sequence. The ability to use the human genome sequence as a framework to anchor the mouse bacterial clone maps takes advantage of

the sequences conserved between the two genomes. The mapping of both the mouse and zebrafish genomes does not rely solely on using the human genome as a framework. Independent analysis of data, using whole genome fingerprinting is ensuring that the species-specific organisation of the mouse and zebrafish genomes will be maintained. However, the ability to compare the maps and sequences with the human map and sequence allows the syntenic relationships to be established and further characterised.

The developments in mapping, including large insert bacterial cloning systems and high density landmark generation/ordering, have made large-scale map construction a very efficient process and may lead to the construction of maps covering other genomes. These maps could be used as reagents to generate high quality sequence as was seen for the human genome. Alternatively the maps could be used as frameworks to carry out either clone-based or whole genome shotgun sequencing without finishing, using the clone maps to anchor and orient the shotgun sequence. The quality of this draft product produced would be much greater than from whole genome shotgun alone. The amount of large-scale sequencing of other genomes will depend in part on the contribution the maps and sequence currently being generated in mouse and zebrafish play in interpreting the human genome, the costs involved and the value to the research of the other organisms.

7.2 Mining the human genome sequence

The complexity of an organism was always thought to be a reflection of the complexity of the gene content. Analysis of the sequence of *S. cerevisiae*, a single-celled organism predicted 6,000 genes. The sequence of the genome of *C. elegans*, a multi-cellular organism, was predicted to contain19,000 genes. However, when the sequence of *D. melanogaster* was produced, less than 14,000 genes were predicted. It became clear that gene number alone was not a direct indication of complexity. Complexity may arise from not only the number of genes, but also from the use of promoters and regulatory elements, post-transcriptional modification where alternative splicing produces multiple transcripts per gene, and protein diversity such as differences in structure and interaction. For example, early analysis of the sequence of *D. melanogaster* predicted 13,601 genes would produce at least 14,113 transcripts but that this was likely to be an underestimate (Adams, M.D., *et al*, 2000). Analysis of part of the draft sequence of the human genome, using chromosome 22 and chromosome 19 sequence and annotation, predicted an average of 2.6 (for chr22) and 3.2 (for chr19) transcripts per gene (Lander, E.S., *et al*, 2001). Mining the human genome, arguably the most complex organism of all is significantly more difficult. At the same time, virtually all the information which gives rise to the greater complexity of protein function by the mechanisms described above is nevertheless encoded within the DNA sequence of the human genome. The challenge is to find those features and understand their role.

Prior to the availability of genomic sequence, gene identification relied upon the isolation of cDNA clones using YACs or bacterial clones from the physical maps.

Physical maps were generated across regions thought to contain genes involved in particular diseases. Candidate genes in the region were identified using techniques such as cDNA direct selection and exon trapping. However, these methods were laborious and subject to artefacts. For example, cDNA direct selection suffered from false positives due to repeats, pseudogenes and gene families. The techniques also yielded significant levels of false positives, often only providing limited levels of enrichment for the cDNAs of interest, resulting in the analysis of a larger number of cDNA clones than was desirable. A significant advantage of exon trapping was that it did not rely on expressed sequences to identify genes, but used the signals encoded in the genomic sequence to identify regions that splice. Exon trapping is also prone to false positives and because it relies on the presence of flanking intron sequence, was not able to clone first or last exons, or single exon genes. Some development of the initial procedure does allow for the cloning of 5' and 3' exons but this was also prone to false positives. However, exon trapping has been applied on a large scale to identify 6,400 potential exons on chromosome 22, and estimates suggested over half represented true exons (Trofatter, J. A., *et al.*, 1995).

The availability of the genomic sequence provides the foundation for a full investigation into the features contained within it, including the genes and regulatory elements. Moreover, this analysis can be carried out in part computationally, making the process less labour-intensive. Genes can be predicted by two complementary methods, similarity searches and *de novo* gene prediction, and confirmed where necessary by generating novel cDNA sequence. The ability to predict the genes in a given region which can then be investigated by experimental methods is a much more efficient system for gene identification than was previously seen prior to the

availability of genomic sequence. This strategy for gene identification was described in chapter 4, but revealed several limitations to identifying genes in this manner.

The first limitation is the incompleteness of the genomic sequence. A complete transcript map cannot be constructed if gaps remain in both the map and sequence. This was the case in the Xq23-q24 region studied where eleven gaps were present between twelve sequence contigs, and for a small number of clones only draft sequence was available. In order to carry out a detailed analysis of the genes contained within a given region, it is important to generate high quality finished sequence covering as much of the region of interest as possible. This was highlighted in the analysis of the critical region for MRX23 for which the bacterial clone map contained a gap of approximately 500 kb. A complete analysis of the genes in the region was not possible until this sequence becomes available.

A second limitation is the lack of available supporting evidence for predicted gene structures in the DNA and protein sequence database (e.g EMBL and Swissprot). In order to confirm a predicted gene, human cDNA sequence covering at least the predicted ORF is required. Also, supporting evidence may be present for part of the gene structure and may be incomplete. For instance, a gene having a large mRNA may be incompletely reverse transcribed to form a partial cDNA clone or the cDNA clone is incompletely sequenced as is the case in EST sequencing. These problems are being addressed in part by the full length cDNA sequencing programs such as the Mammalian Gene Collection (MGC - http://mgc.nci.nih.gov/). However, in many cases it is still necessary to carry out targeted cDNA sequencing to confirm predicted gene structures. Recent estimates from the genes identified on chromosome 20

showed that of the 727 genes identified, 350 need additional confirmatory human cDNA sequence (Graeme Bethel, personal communication). A third limitation in gene identification is the inability to confirm predicted genes. Some genes remain unconfirmed as they are not represented in the cDNA libraries available. They may be expressed in tissues not represented in the library collection, or temporally or transiently expressed. Testing a wider variety of cDNA libraries will increase the likelihood of confirming the gene. Predicting the mouse orthologue and testing for the presence of the cDNA in mouse cDNA libraries provides more flexibility in terms of the range of tissues that can be utilised and this approach is already providing valuable information for human sequence annotation. For example, the RIKEN Institute are sequencing cDNA clones, derived from over 200 different tissues and cell types, with the aim of collecting data on as many full-length cDNAs as possible (http://genome.gsc.riken.go.jp). In conjunction with the human cDNA sequencing effort, the Mammalian Gene Collection (MGC) is also sequencing cDNA clones derived from mouse tissues (http://mgc.nci.nih.gov/). As an alternative to cDNA analysis, the screening of RNA by RT-PCR may identify transcripts not previously detected as the RNA has undergone fewer manipulations. Very low level transcripts have also been detected by nested RT-PCR in tissues not expected to express the gene and may represent illegitimate transcription which nevertheless confirms the biological activity of the predicted gene (Roberts, R. G.*, et al.*, 1991; Kaplan, J. C.*, et al.*, 1992).

It is unlikely that all genes will be confirmed by identification and sequencing of cDNA. Therefore, the predicted genes for which no cDNA sequence is available require the level of confidence associated with each prediction to be assessed. In

chapter 4, predicted genes were only followed up if there was a certain level of supporting evidence, such as regions predicted by two separate gene prediction programmes, or an EST or cDNA sequence that spliced exactly on to the genomic DNA. Given the need to set a series of criteria by which predicted genes were analysed in order to reduce over prediction of genes to a minimum, it is possible that some real genes are missed. Therefore it is important to generate as much information as possible to be confident that a high percentage of the real genes have been identified, whilst limiting the number of false predictions. In the absence of supporting cDNA or protein sequence, the best indication that predicted genes or exons from partially confirmed genes are real is the observation that the predicted gene is conserved in other species as was carried out in chapters 5 and 6 (see Section 7.4).

As discussed in chapter 4, the first exon of a gene, containing the 5' untranslated region is often the most difficult to identify. The lack of coding sequence limits the use of sequence similarity searches. In addition, the gene prediction programs are not designed to predict UTR, because the prediction process involves codon usage analysis. Given the ever- increasing amount of gene structure information being generated, it may be possible to design computational tools to specifically predict first exons. In a similar fashion to other software development, a high quality training set of known first exons could be scanned and compared to identify signals specific to first exons. The prediction program could be tested and improved using a second calibration set.

The first exon of a gene is often adjacent to the promoter elements and transcription start site and analysis of these features may aid in the identification of the first exon. Three methods for the analysis of 5' sequence elements are described in chapter 4: CpG island detection, PromotorInspector and Eponine, which begin to address this. However, the analysis is currently limited to genes which are associated with CpG rich sequences at their 5' end. These tools are continuing to improve and the development could be widened to include genes which are not associated with a CpG island.

The manually curated annotation of the genomic sequence described in chapter 4 neither confirms all of the predictions, or identifies all of the functional elements contained within the sequence but still goes further than most of the automatic human sequence analysis currently being carried out (e.g genome browsers such as ENSEMBL and UCSC). The majority of the genes identified by the methods described in chapters 4, 5, and 6 are protein-coding genes, because these are the ones that can be more readily identified by primary sequence analysis. However, other genes produce non-coding RNAs (ncRNA) that function directly as structural, regulatory or even catalytic RNAs (Eddy '99). Unlike protein-coding genes these ncRNA genes have no obvious primary sequence patterns that can be used to identify them. Separate tools need to be developed to identify these sequences. One strategy that may be applicable to human sequence has been used successfully to identify potential novel ncRNAs in *E. coli*. The method took regions conserved in multiple species of bacteria and identified candidate ncRNA based on secondary structure analysis, such as identification of hairpin loops, rather than primary sequence analysis.

7.3 Comparing different genomes to aid human genome sequence analysis

Comparing genome sequences from different species is a powerful method for increasing the confidence in predicted genes, or identifying novel functional units. When two species diverge from a common ancestor those sequences that maintain their original function are likely to remain conserved in both species throughout their subsequent divergent evolution. In order to identify the functionally important units in the human genome it may be necessary to compare genome sequences from a variety of organisms, although any human-specific features will not be detected by this strategy. The more distantly related organisms such as yeast and worm are likely to show sequence conservation in coding regions alone. This may also be the case for distantly related vertebrates such as fish. The more closely related organisms, such as mouse, are likely to be conserved in coding regions, but also in other functional elements such as regulatory sequences. However, the closer the evolutionary relationship with human, the more 'sequence noise' is likely to arise where non-functional sequence appears similar because insufficient time has elapsed for the two sequences to diverge. For instance, comparing sequence between human and chimpanzee is only really useful to determine the differences between the two species rather than the similarities. The extreme case of this is to compare genomes within the same species to identify variation within a population and to determine the functional significance of the variation.

Identification of conserved sequences outside regions predicted to be coding may indicate the presence of a novel gene or a region involved in gene regulation. The challenge when comparing genome sequences is to identify the sequences conserved

between different organisms given that many are short and interspersed with large regions of non-conserved, non-functional DNA, and to distinguish between the conserved sequences that are functional and those that have no function. The coding regions of many vertebrate genes, unlike genes of other lower organisms such as bacteria and most of the genes in yeast, are disrupted by introns and exons can be small. Other functionally important segments of DNA such as regulatory sequences can be very small and placed far away from the gene they influence. In chapters 5 and 6, a genomic region in Xq24 was analysed in mouse and zebrafish, to identify potential novel functional elements using comparative sequence analysis.

There is a range of comparative sequence analysis tools available for identifying sequences conserved between species, some of which are described and used in chapters 5 and 6. Generating either local alignments or global alignments between pairs of sequences identifies conserved sequences between different species. Three tools, PIPMAKER, VISTA and BLAST, were used in chapters 5 and 6 to identify conserved sequences between human and mouse, human and zebrafish, and sequences conserved in all three species, but further work needs to be carried out to determine whether or not these novel conserved sequences represent functional units, for example involved in gene regulation.

Experiments have been carried out to identify those conserved elements that have regulatory function but these have been carried out on individual regions of interest and not on a large scale. In the case of the SCL locus, a region conserved in human and mouse was shown using a transgenic xenopus assay to be a novel neural enhancer (Gottgens, B*., et al.*, 2000). Touchman, J. W., *et al*. (2001) identified a novel

regulatory element of the human and mouse α-Synuclein genes using comparative genome sequence analysis and subsequently confirmed its regulatory function using a reporter gene assay (Touchman, J. W*., et al.*, 2001). Other methods of functional testing include mobility shift assays and DNA footprinting, two methods that examine the binding of proteins to the DNA fragments of interest, and site-directed mutagenesis, followed by functional analysis can identify individual bases that affect regulation (Sambrook, J*., et al.*, 1989).

7.4 Functional analysis of gene products

The identification of the genes encoded within the human genome is only the first step to a complete understanding of the genes encoded in the human genome. An important second step is to understand the role these genes play in the correct functioning of a cell. One method for determining function is to see the phenotypic effect if the function of a gene is altered. In humans, this analysis has been provided by naturally occuring mutations that cause diseases. Cross-species comparison is also a powerful tool for gaining an insight into the function of a gene where the function of one gene can be inferred based on the function of the homologue. The advantage of functional analysis in other organisms is that these mutations can be engineered. This was first pioneered with mutation analysis in bacteria and drosophila as large numbers of mutant phenotypes can be analysed.

More recently chemical mutagenesis programmes have been undertaken in yeast, worm, zebrafish and mouse (reviewed in Justice, M. J., 2000) where alterations in

phenotype are being used to categorise the function of genes that have been mutated. Genetic screens in the zebrafish have proven particularly useful in identifying genes involved in development. Zebrafish eggs develop externally and they can be easily visualised. The embryos are relatively transparent which aids the detection of phenotypic abnormalities. Two studies which screened for both defects in embryogenesis and essential functions and identified more than 2000 mutant phenotypes (Driever, W*., et al.*; 1996, Haffter, P*., et al.*, 1996). The fact that embryo development in the mouse takes place in the uterus means that in mouse mutagenesis programmes these types of mutations are difficult to analyse. Two groups (UK ENU Mutagenesis programme and German ENU-mouse mutagenesis screening project) are using a breeding strategy to identify dominant and haplo-insufficient mutations. The UK programme is screening for visible phenotypes including sensory, neurological, neuromuscular alterations and behavioural assays, and the German project is screening for haematological, clinical chemistry, immunological and allergy defects. An alternative to the *in vivo* mutagenesis programs is *in vitro* mutagenesis by gene trapping, or gene targeting in mouse ES cells. These methods allow mutations to be characterised in cell culture before translation to mouse (Evans, M. J*., et al.*, 1997). Gene trapping involves random integration in to the genome of a promoter-less reporter gene construct. Incorporation of the reporter gene within a gene may abolish the function of that gene as well as enable the selection of the mutant on the basis of the transcribed reporter gene. As an alternative to random mutagenesis, it is possible to target specific genes for disruption. One of the best methods for incorporating mutations into the mouse genome has been gene targeting by homologous recombination in ES cells and modifications to this procedure enable virtually any designer modification to be introduced to known cloned mouse genes (Koller, B. H*.,*

et al., 1992). However, gene targeting requires some knowledge of the sequence of the gene of interest, although recent developments require only a small amount of homologous sequence (approximately 80 bp) (Zhang, P*., et al.*, 2002).

Obtaining functional information for genes in other organisms allows inferences to be drawn about the function of the orthologous genes in humans. However, inferring the function of a gene in one organism based on evidence from another organism does have limitations. As soon as speciation from a common ancestor occurs, each new species is able to evolve independently. Even though initially orthologous genes may have the same function, over time novel function for one or both genes may evolve (see Figure 7.1). For instance, duplication of DNA sequence that includes a single gene will create two copies of the orthologue in one species. The duplicated copies are now termed paralogues. Although the one paralogue may maintain its original function, the other paralogue may evolve a novel function, as there may be little or no selective pressure for it to maintain its original function. Inferring the function of the orthologous gene, based on the function of the second paralogue would be incorrect. An example of this is the mouse p53 tumour suppressor, a transcription factor that regulates the progression of the cell through its cycle and cell death in response to environmental stimuli. In contrast, p63, presumed to be a paralogue based on its sequence similarity to p53, has a completely separate function and is essential for several aspects of ectodermal differentiation during embryogenesis (Mills, A. A*., et al.*, 1999).

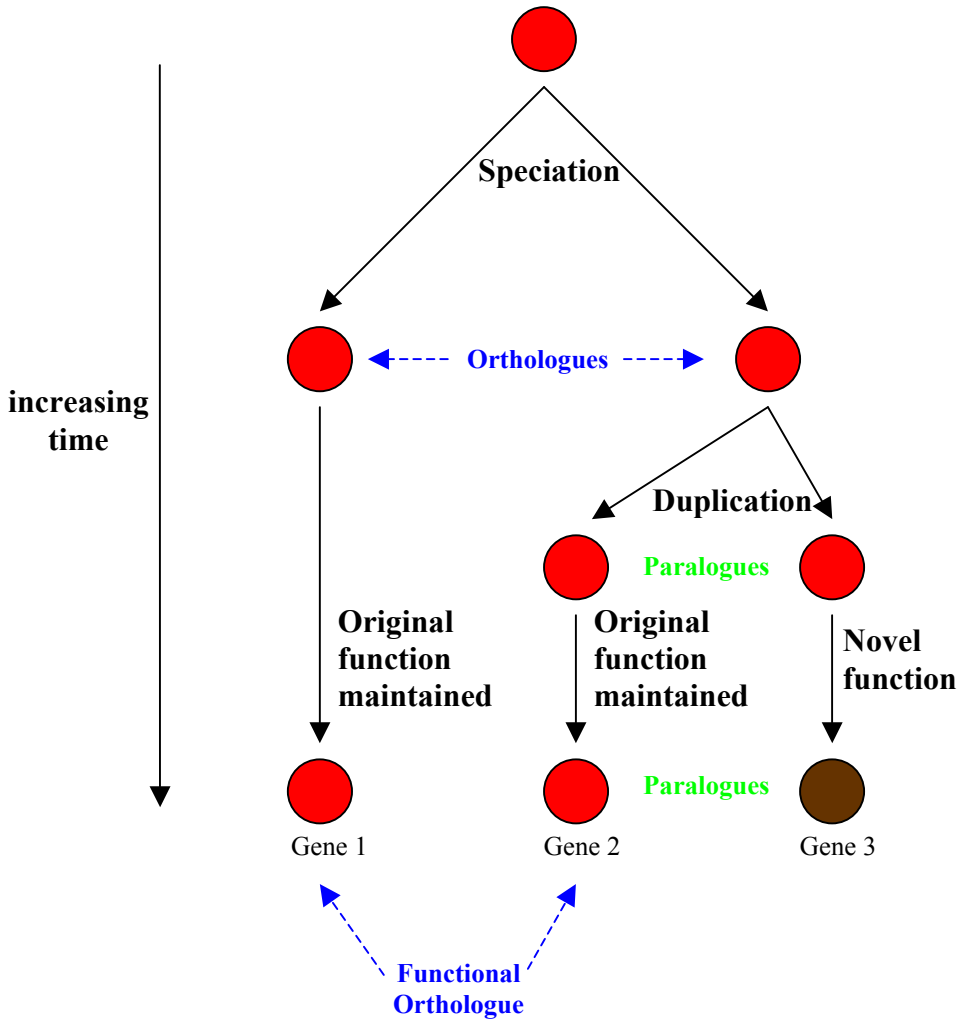

Figure 7.1: *A representation of how speciation and gene duplication can influence comparative genome analysis. The example shows a gene (circle) with a specific function (colour of circle). Two species diverge from a common ancestor (through speciation) which creates two orthologoues. Gene duplication in one species, creates two paralogues, one of which is free to evolve a novel function (represented by a brown circle). Although originally descended from the same gene, inferring function of gene 1 from the function of gene 3 would be incorrect. The true functional orthologue of Gene 1 is Gene 2, whose presence may not be detected by localised genome comparison.*

Analysis of a small region in each organism may not detect other copies of the same gene. Establishing the extent of paralogy within each organism is essential before the function of a particular gene is inferred by cross-species analysis. Detailed synteny maps such as those generated in chapters 5 and 6, and complete analysis of the potential orthologous genes are required to increase the confidence that true orthologues have been identified.

Further insight into the function of a protein can also be obtained by analysing the protein sequence for functional domains, using sequence-motif searches such as those available at INTERPRO (see chapter 4). Although these methods may be used to gain information regarding the potential function of a protein, experimental verification is always required to determine the exact function. The improvements in the speed and accuracy of gene identification have meant that an increasing number of genes have been and are being identified in a semi-automated fashion. However, little experimental data is being generated to confirm the predicted protein sequences.

Ultimately, confirmation of the structure and function of proteins is required by species-specific investigation where possible. For instance, the structure of a protein can be determined using X-ray crystallography, nuclear magnetic resonance and mass spectrometry. Determining the cellular localisation of a protein by generating and expressing a fusion construct containing the gene of interest linked to a reporter molecule such as green fluorescent protein (GFP) will provide insight into where in the cell the protein is functional. The hybrid protein is very different to the original protein of interest, therefore this type of analyses is often carried out using different reporter molecules, or antibodies to universal tags, and attaching the reporter molecule

to both the 5' and the 3' end of the protein of interest. Alternatively, an antibody specific to the protein of interest can be generated but this is laborious and costly so would be difficult to do on all human proteins.

Protein binding assays can be carried out using the yeast two hybrid and mammalian two hybrid system. In the yeast two-hybrid (Y2H) system, the gene of interest is commonly linked to the Gal4 DNA binding domain and is co-transfected into yeast cells with a library of genes linked to the Gal4 activation domain. If the protein of interest binds to a target protein, the two Gal4 domains will be brought together, and the expression of the downstream *LacZ* reporter gene can be measured using the β-galactosidase assay. One advantage of the Y2H system is that when a positive match is detected, the ORF is identified by simply sequencing the relevant clone(s). Therefore Y2H system is amenable to high-throughput screening of protein-protein interactions, such as has been reported for *S. cerevisiae* and *C. elegans* (reviewed in Walhout, A. J.*, et al.*, 2000).

7.5 Conclusion

Prior to the generation of whole genome sequences, individual research focussed on understanding individual genes or gene families, their protein products and their function. The availability of large amounts of genome sequence from human and other organisms is enabling large-scale interpretation of the sequences, including all the functional elements encoded within. This has seen the growth of large-scale mapping and sequence production, along with the development of computational hardware and software to both store and analyse this data. However, the complete interpretation of the genome sequence of these organisms includes the complete characterisation the genes encoded within and the role each gene plays within a cell and the contribution to the whole organism. This characterisation will need to be carried out for each gene and each functional element in each genome. Individual research will focus on understanding individual genes, gene families, their protein products and their function.