Chapter 6
General Discussion and Future Work

The work detailed in this thesis represents clear progress in the fields to which it belongs. The approaches applied are either novel or are significant improvements on previous studies and have already yielded valuable results. The data and approaches taken also strongly indicate that future pursuit of the ultimate aims of these projects is worthwhile and are very likely to prove fruitful. I will now discuss the projects detailed in the previous chapters individually, focusing on the outcomes thus far and the future potential of the projects.

Expression profiles have been used with tremendous success in yeast as a means of describing the phenotype of different mutant strains. Comparing the expression profiles of two different mutants provides a high-resolution way to ask whether the two genes are likely to act in the same pathway – genes that act in the same pathway or complex have very similar profiles. This is a powerful approach to identify how novel genes act – if they share profiles with well-characterised genes, one can infer that they act in similar processes. The key question that I sought to address is whether this kind of approach can be used in a far more complex system than yeast, in a whole animal. To investigate this, I used standard two-colour array technology to generate expression profiles for a number of worm populations, either carrying loss-of-function mutations in genes known to play key roles in different signalling pathways affecting germline development, or having had these genes targeted through RNAi. I then used standard clustering algorithms to compare these expression profiles and used this to ask several questions: do two genes in the same pathway tend to cluster together? Do genes in different pathways cluster in different branches? Does the expression profile of one perturbation of a gene cluster very near another perturbed expression profile of the same gene? Finally, since all the profiles were of perturbations of genes affecting

brood size, I asked whether the clustering was directly correlated with strength of phenotype i.e. do genes of similar brood size cluster together?

I found first that genes in similar pathways do tend to cluster together far more strongly than associations between genes of different pathways. The implication of this is clear: if a novel gene is known to affect germline function then we can discover how it acts by comparing its profile with that of all those examined; if it clusters with several genes of a known pathway, it is highly likely to act in that pathway. Second, I found that in general the RNAi phenotype of a gene, as monitored by its perturbed expression profile, looks very similar to its genetic loss-of-function mutant – this is reassuring. Finally, I found that genes cluster independent of RNAi phenotype strength – clustering is thus driven by the underlying pathway affected and not by the extent to which it is affected. I thus concluded that expression profiles are indeed effective tools for identifying the mechanism of action of a novel gene.

To test whether we can really use the expression profile compendium to confirm how a novel gene acts, I turned to a dataset generated by Catriona Crombie, a postdoc in the Fraser lab. She had isolated a number of mutants that are candidate modulators of the EGF/ras/MAPK signalling pathway in the *C. elegans* vulva. Since EGF/ras/MAPK signalling is also required for germline development, I reasoned that I could confirm the role of these novel modulators by testing whether their expression profiles clustered with those of known EGF/ras/MAPK pathway genes. I selected one of these genes, *pkc-1*, to test this and find that it does indeed cluster tightly with other genes in this pathway, thus confirming both the approach and the pathway in which *pkc-1* appears to act. This result evidently requires further follow-up, for example by

detailed staining and microscopy, but it is very encouraging for such an approach. There is a wealth of genes identified as potential modulators of EGF/ras/MAPK and Notch pathway signalling revealed by screens in our lab and others. These are now candidates for testing against our compendium of expression profiles to provide further evidence of their roles in these pathways.

This approach appears to have much potential in adding evidence for the roles of genes in germline development. Importantly, in justifying the nature of the approach, it appears to be sensitive to even small weak gene perturbations resulting in only slight brood-size defects and considers populations rather than individuals. This may suggest that other approaches may be less sensitive to such changes in phenotype. But is there much value in increasing our knowledge of *C. elegans* germline development? Obviously the primary motivation behind any such biological study should be the downstream development of our understanding of human biology. That the signalling pathways being considered in the study are conserved from worms to humans and that the identified involvement of PKC signalling with EGF/ras/MAPK signalling had already been demonstrated in mammals suggests that there is relevant potential in this methodology. Identification of modulators of these pathways in *C. elegans* may be to identify candidate genes in human disease where these pathways are dysregulated, such as in cancer. The next step is clearly to increase the size of the compendium with other known regulators of germline development and gametogenesis as well as novel genes giving brood-size defects and novel genes, which are candidate regulators of the pathways of interest.

Our interrogation of the *C. elegans* transcriptome appears to have been similarly fruitful. We used tiled microarrays and ultra-high density sequencing to assess genome-wide transcription at multiple stages during worm development. We initially set out to address two key issues – whether there is a substantial amount of transcription beyond current annotations, and how complete current splicing annotations are. Widespread novel transcription has recently been shown to exist in a number of other organisms. Using whole genome tiled microarrays we have demonstrated that throughout development only ~5% of expressed regions of the genome lie outside annotated structures. This is reassuring in that it increases confidence in current gene annotations. It does, however, demonstrate that there are regions of novel transcription, which require further characterization. Ultra-high density sequencing technologies appear to be an ideal tool to do this, offering greater resolution than tiling array data and providing connectivity data in the form of reads that span exon-exon boundaries. We have used these data to identify reads spanning exon-exon boundaries of exons annotated as connected as well as novel exon-exon boundaries. Thus far our data indicate that novel splice events occur for ~1% of annotated genes. Critically, however, this approach is limited by the depth of coverage of the transcriptome provided by our sequence data. We have recently acquired sequence data to a greater depth, which will allow more thorough identification of novel splice events. Whilst ultra-high density sequence data offers a better option in studying splicing than tiled microarray data, our microarray data have nevertheless given us an interesting insight into the extent of unannotated splicing. Using our tiling array data to look at changes in relative exon intensities throughout development we have identified genes that exhibit major changes in exon use, indicating alternative spliceforms. Of the genes exhibiting novel splicing events in

our sequence data ~80% were also identified in our tiling analysis leading us to believe that the genes we discover using the tiling data are alternatively spliced. ~50% of the genes identified as alternatively spliced at high confidence using our tiling data have only one annotated isoform, suggesting that annotation of spliceforms is far less complete than that of transcription as a whole. It will be very interesting to see if this trend continues when our sequence data analysis is extended to our newly acquired data set. The approach discussed to study alternative splicing using sequence data could also be expanded to catalogue trans-splicing events by identifying reads that span independently transcribed structures. A further application of our sequence data may be to uncover the identity of novel transcribed regions in terms of their connectivity to already annotated genes, or each other as previously unannotated spliced or unspliced transcripts. This is a far more complex problem than studying connectivity of annotated exons. It will require an approach that does not rely on gene annotations. A shotgun approach to assemble sequence reads may offer a possible method of connecting novel structures and may also allow better annotation of exon boundaries in already annotated genes and novel splice sites within annotated introns and exons.

The quality of our data and the approaches applied represent a major step forward in transcriptome analysis towards the ultimate set of gene annotations. The value of this is difficult to overestimate. Identification of all genes may lead to the discovery of transcripts and proteins of novel function. Knowledge of all isoforms of all genes will allow a more complete study of protein structure and consequent biological properties and how these change between different conditions. It will also lead to improvement in approaches to quantify transcript levels by allowing more comprehensive

transcriptome coverage by expression microarrays. Any benefit to microarray design, however, may be short-lived. It seems that the key advantage of microarrays over ultra-high density sequencing is the cost-differential for the same depth of coverage of the transcriptome. Were funds unlimited it is difficult to identify many applications for which microarrays would be the preferred platform. Should ultra-high density sequencing become more affordable and of higher throughput then, the use of microarrays may become a thing of the past.

Our interrogation of the wild-type *C. elegans* transcriptome and the approaches applied to it provided the ideal framework for comparison with the NMD-deficient transcriptome. Our motivation in studying the NMD was to determine whether the identity of NMD targets, their structures and how those structures change could provide an insight into the role and mechanism of NMD. *C. elegans* appeared to be an ideal system in which to do this as it allowed us to study NMD in a dynamic biological environment i.e. throughout development. Whole genome tiling array data was produced for comparison with our wild-type dataset. Comparison of the resulting gene intensities between wild-type and NMD-deficient animals revealed genes regulated by NMD. Analysis of the properties of these transcripts confirmed features that have previously been reported as being NMD causative, such as identification of alternative spliceforms and transcripts containing uORFs. Interestingly the strength of the annotated translation initiation sequence appears to be critical to the predisposition of transcripts to NMD. NMD may therefore act to regulate steady-state transcription in accordance with the strength of the translation initiation sequence. Whilst translation initiation events occurring after the annotated translation initiation site leading to an in-frame premature termination codon have been recognized to lead

to NMD, this direct relationship of the strength of the translation initiation sequence at the annotated start site and NMD was previously unrecognized and is likely to be of great importance. It is known that many disease-associated mutations and variants result in mRNAs harbouring PTCs. The clinical outcome of harbouring such alleles is NMD dependent (Khajavi *et al*., 2006). Sequence variation at the translation initiation site may occur leading to effective under- or over-expression of a transcript due to a shift in its susceptibility to NMD. It may therefore have a significant link to human disease.

Amongst the repertoire of NMD targets are operonic genes. This is most interesting as it is known that whilst genes in operons are transcribed at equal levels, the measured abundance of transcripts for genes in the same operon are often different. Whilst not a complete explanation for this inequity of effective expression, NMD does appear to one mechanism by which this occurs.

Perhaps the most interesting finding of the work detailed in this thesis is the requirement of NMD for ~10% of developmentally regulated gene expression changes via regulated changes in transcript structure. Such structural changes may be a switch in spliceform or a shift in the position of transcription initiation to include or omit a uAUG. This demonstrates that the timing of gene expression is not dictated by the rate of transcription alone, rather in some cases the position of transcription initiation and also splicing events may act via NMD to dictate the effective level of gene expression in a temporally controlled manner. This represents a hitherto unrecognized mechanism of gene expression regulation and will inevitably alter perception of how such regulation is achieved. Whilst it seems likely that this method

of gene expression regulation occurs in wild-type animals, we cannot discount the possibility that the changes in transcript structure that lead to NMD targeting are through the action of splicing and transcription factors which are also NMD regulated. If this is so then much of the signal may be artifactual. This does appear extremely unlikely but we cannot discount the possibility. Testing this through identifying targets of NMD regulated splicing and transcription factors would be a huge undertaking and assumes that we have already comprehensively identified all such factors, which is unlikely. As previously stated then, possibly the best method of validating this finding and adding value to it would be repeating the study in another biological system where NMD is not essential such as yeast or fly. The same possibility of NMD regulation of transcripts due to regulation of upstream factors would still stand however.

Taken together our findings regarding NMD will have a significant impact on current perception of NMD and have real potential to influence perception of human disease biology and gene regulation. Though NMD is not essential in yeast, fly and worm it appears to required for mouse embryonic development (Medghalchi *et al*., 2001). NMD being required for correct developmental gene expression rather than just acting as a surveillance mechanism may serve as a partial or complete explanation of this. The possibility of variation at the translation initiation site leading to variation in effective gene expression via NMD and consequent modulation of a disease phenotype seems very real and worthy of investigation.

The output of all of these individual studies is indicative of the continuing value of *C. elegans* as a tool for large-scale biological studies. Whilst the requirement of

performing expression analyses at the level of the whole animal may often be cited as a disadvantage, it was the ease of studying expression throughout development and in the germline at the level of a whole animal that led to the manner of all of these studies. The utility and ease of RNAi in *C. elegans* and the wealth of genetic mutants were also key advantages that were essential to these studies. That the use of *C. elegans* continues to be just as valid as technology and biological research moves on demonstrated that *C. elegans* remains at the cusp of cutting-edge research and is likely to for the foreseeable future.