Chapter 4
Analysis of the wild-type *C. elegans* transcriptome

4.1. Introduction

The *C. elegans* genome was the first of any metazoan to be completely sequenced, this feat having been achieved in 1998 (C. elegans Sequencing Consortium, 1998). Furthermore it was only the second eukaryotic genome to be completed, after *S. cerevisiae*. Annotation of the ~100Mb genome of *C. elegans* is excellent and arguably more advanced than that of other animals. Regardless of this a completely stable set of gene annotations has not yet been achieved, with new releases (albeit with only minor changes) every month or so. My intention was to determine how well gene annotations corresponded to the transcribed regions of the *C. elegans* genome using whole-genome Affymetrix GeneChip® *C. elegans* Tiling 1.0R Arrays. Similar studies done in *Arabidopsis thaliana*, *Drosophila melanogaster* and humans had revealed that vastly more of each genome is transcribed than could be accounted for by then current annotations (Bertone *et al*., 2004; Hanada *et al*., 2007; Manak *et al*., 2006; The FANTOM Consortium and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group), 2006). The genome of *C. elegans* is already considered to be transcriptionally dense, with ~62% of the genome thought to be genic and ~33% exonic (WS150 release of Wormbase). The Affymetrix tiling arrays used can survey the transcriptome to a resolution of 35bp. When this project was conceived these microarrays were not yet commercially available. This project was therefore a collaboration with the laboratory of T.R. Gingeras, Affymetrix Inc., Santa Clara, CA., USA where the microarray hybridizations were performed. The informatics was performed in association with Arun Ramani, a postdoctoral researcher in the Fraser lab.

In order to achieve adequate cover of the transcriptome for this study total RNA from six different developmental stages in the *C. elegans* life-cycle, specifically embryos, L2, L3, L4, young adults and gravid adults was hybridized in at least duplicate. This RNA was derived from the wild-type reference strain Bristol N2. Not only was this done to give us maximum coverage of the transcriptome, but also to give an adequate data set for comparison with the NMD-deficient transcriptome, as will be seen in chapter 5. The output of this platform is a set of probe intensities for the ~3 million probes arrayed on each chip, analysis of which reveals the regions of the genome for which transcript is present in the sample.

The use of tiled microarrays allows us to survey all transcribed regions of the genome and therefore examine how transcript structures change as well as transcript levels. Historically, however, single colour tiled microarrays have not been used to generate gene intensities and determine differential expression between conditions. With no established methodology and pipeline by which to do this it was required that we develop our own analysis strategy. Also, as with any other technology platform, validation of the output was required before the data could be considered reliable. One possible method of validation would be exhaustive RT-PCR and sequencing to confirm the existence and identity of novel transcribed regions and structural changes indicated by the tiling data. A superior alternative now available to us, however, is ultra-high density sequencing of cDNAs. This automatically allows validation of novel features and gives information on connectivity of structures by identification of reads that span exon-exon boundaries.

Furthermore the number of reads that map to a given structure act as an expression value with which we can compare gene intensity values derived from tiling data. This therefore allows validation of transcript prediction and intensity values simultaneously. Consequently, we produced ultra-high density sequence data using the Illumina platform for two developmental stages individually (L4 and young adult), as well as a mixed stage sample containing RNA derived from all developmental stages in the worm lifecycle in order to give us maximum coverage of the transcriptome at the depth available. The Illumina sequence data have the advantage of being of greater resolution than the tiling array data but could not adequately replace the tiling array data, being of insufficient depth (i.e. insufficient number of unique reads) and providing stage-specific information at fewer stages. The purification of RNA for sequencing and tiling array analysis excludes RNAs <200nts. Consequently such RNAs are not represented in the data.

In this chapter I will demonstrate the quality of the tiling array data by comparison with the sequence data. I will then present the protocols established using the two forms of data produced and how they inform us on the current state of gene annotations. I will discuss how our data relate to the density and accuracy of gene predictions as well as how they can be used to predict changes in splice forms and connectivity between annotated and predicted structures.

4.2. Tiling array data normalization

All Affymetrix GeneChip® *C. elegans* Tiling 1.0R Array data presented in this thesis was quantile normalized prior to use. Quantile normalization is a standard approach

applied to one-colour microarray data (Bolstad *et al*., 2003). As discussed in chapter 3, there are two forms of variation that occur between individual microarray experiments – biological variation and technical variation. The goal of normalization is to reduce technical variation. Differences in labeling efficiency of samples, quantity of material hybridized and the gain of lasers used to scan the arrays are all examples of what introduces technical variation. The key assumption made by quantile normalization is that the true biology-driven distribution of probe intensities on a one-colour microarray is the same between all arrays. Quantile normalization takes all probes on an array and sorts them in order of intensity. This is done for all arrays that are to be compared. The mean of the probes for each array at each sorted position then becomes the normalized probe intensity at that position (e.g. the tenth highest probe intensity on all arrays is now the same – the mean of the non-normalized intensities). Each array now has the exact same probe intensities but the intensities are not assigned to the same probe, rather the ranking of intensities for each probe within an array is the same as before but the distribution of probe intensities is now the same for all arrays. Consequently the mean probe signal for all arrays is also the same. All microarrays are now comparable.

4.3. Defining regions of tiling array signal along genomic coordinates

In order to call regions of the genome as expressed using tiling array data it is first necessary to define the methodology and criteria by which this is to be done. There are two distinct ways in which this has been done in previous studies, each with its advantages and disadvantages. A method originally implemented by Wolfgang Huber at the EBI, involves aligning the signal acquired along genomic coordinates and then

dividing the signal into runs of probes showing similar intensities, thus defining transcript and intron-exon boundaries (David *et al*., 2006). This methodology has the advantage of not using gene annotations as a reference and is therefore completely unbiased. A disadvantage is that it requires the user to pre-define the number of partitions that should be drawn in the signal, which is distinctly problematic without reference to a defined set of controls, such as annotated gene structures. Knowledge of the annotated gene structures would permit optimization in order to ensure that expressed exons are not partitioned or fused during the analysis, ensuring that an accurate number of partitions are drawn in the data. Ultimately, however, the number of transcribed regions called by this method is defined by the user rather than the data, which may not be the best method for the purposes of transcript discovery where the user cannot know in advance how many regions of expression to expect.

An alternative way of defining regions of signal is by identifying runs of probes above a calculated background. Again, theoretically this requires no prior knowledge of or reference to annotated gene structures but the complexities of the methodology eventually demand optimization of the technique relative to a set of controls, of which annotated genes are likely to be best. An assumption when optimizing this technique therefore, is that the gene annotations used for comparison are close to correct. This is appropriate for the purposes of transcript discovery, as it makes no assumptions as to the number of genomic regions that correspond to a retained RNA but does ensure that the number of regions discovered is represented as accurately as possible relative to the known characteristics of the transcriptome. The output of such an analysis is discreet

regions of the genome for which transcript exists at a detectable level. Such detected regions are referred to as transcribed fragments or "transfrags" (figure 4.1). Satisfied that this was the most appropriate methodology to identify transcribed regions of the genome, this was the approach we used.

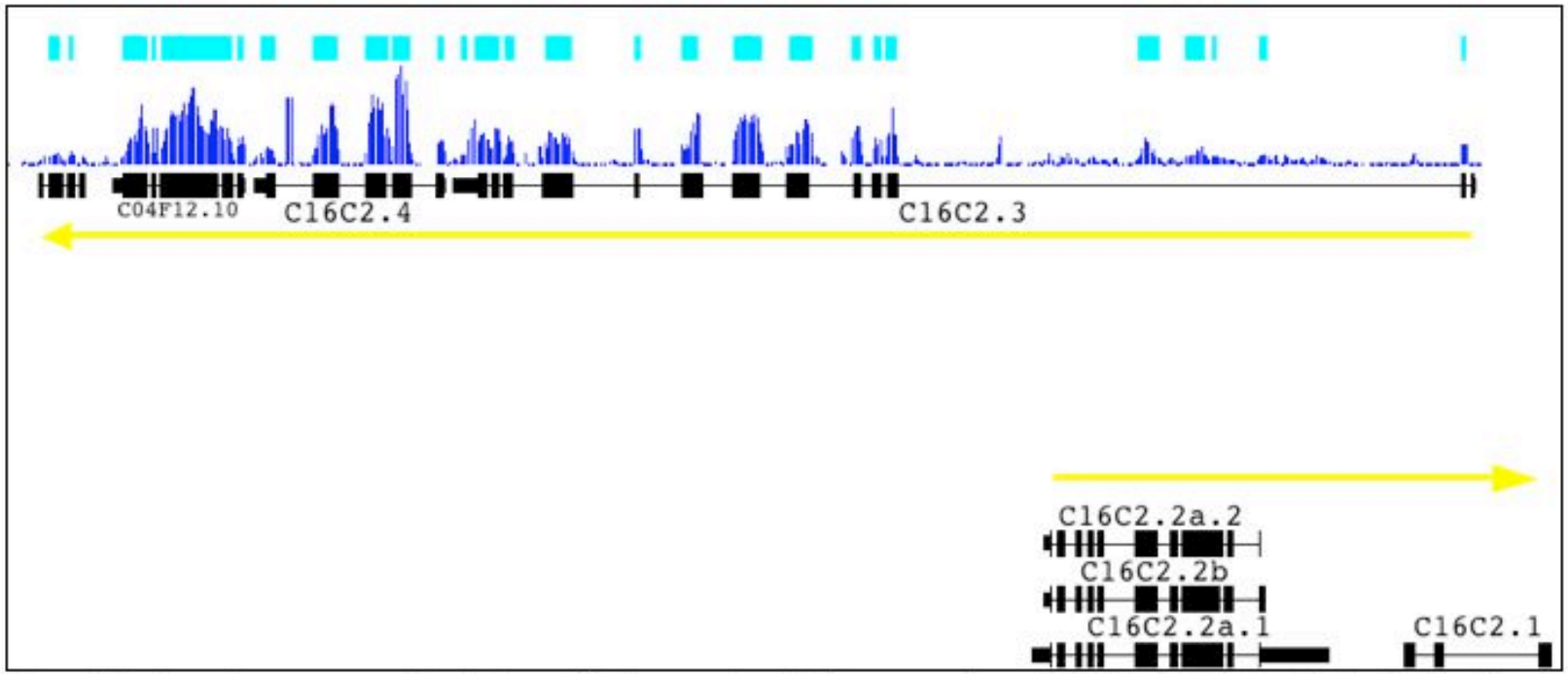

Figure 4.1. Transfrags corresponding to transcribed genes. Annotated genes are shown in black and are transcribed in the direction of their neighbouring yellow arrow. Normalized probe signal is shown in dark blue and the transfrags generated from that signal in light blue. As can be seen, transfrags broadly represent individual exons. There is not necessarily a transfrag for every exon for lowly expressed genes and transfrags may not represent full-length exons. Where short introns exist such that few or no probes map to that structure then exons may be merged into a single transfrag. Broadly, however, one transfrag = one exon; one exon = one transfrag.

4.4. Idealizing parameters for building transfrags

The interval analysis that defines transfrags for any given data set was performed using Affymetrix Tiling Analysis Software (TAS) version 1.1. Prior to interval analysis the data from each replicate are quantile normalized together in R (http://www.r-project.org). The three key parameters that then need to be defined for the interval analysis are the background, the maximum gap (maxgap) and the minimum run (minrun). The background is the threshold above which a probe intensity is considered. The minrun represents the number of consecutive probes that must be above background before a transfrag can be identified spanning that region, in terms of the number of bases of genome represented by those probes. The maxgap is the maximum amount of genome for which there is no signal above background that can be tolerated before a transfrag is terminated. In optimizing the interval analysis relative to gene annotations there are three assumptions that are made. The first is that for each expressed exon (i.e. exon to which a transfrag maps) there should be only one transfrag. If exons are being artificially split into numerous corresponding transfrags this is an indication that the maxgap is too low. Alternatively it could be that the minrun is too low and therefore low-level random noise is being called as transfrags. The second key assumption is that for each transfrag that maps to a gene, it should only span one exon. If a transfrag spans multiple exons then maxgap is likely to be too high, leading to the artificial fusion of transfrags. All of this assumes that the background threshold has been set such that noise is maximally reduced without loss of real signal. Background threshold was calculated to include the top 5% of non-genic probes on the array. This is summarized in figure 4.2. In order to satisfy the "one exon, one transfrag; one transfrag, one exon" optimization strategy a range of

maxgap and minrun combinations were tested and the combination most closely matching the criteria was selected. This was maxgap = 35bp, minrun = 70bp. As the tiling array is made up of 25mer probes tiled at an average genomic distance of 10bp, this is effectively a minrun of two probes and a maxgap of one probe.

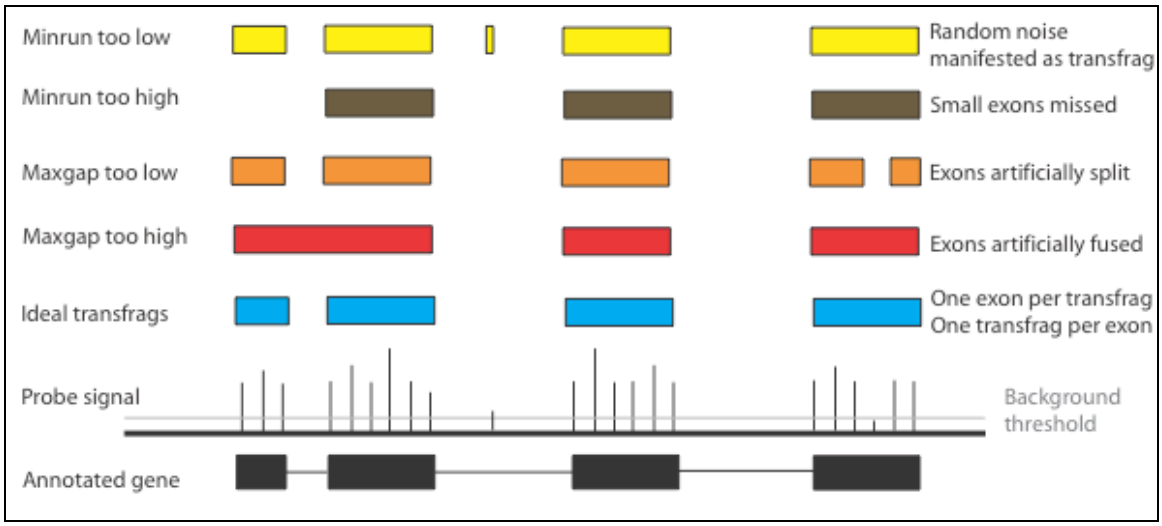

Figure 4.2. Selection of transfrag building parameters schematic. The parameters for building transfrags to represent transcribed regions of the genome were optimized such that one transfrag corresponded to one exon and one exon corresponded to one transfrag. This required that exons were not artificially fused or split by the use of inappropriate maxgap and minrun values.

4.5. Comparison of transfrags with the genome

Each transfrag was classified as either overlapping an annotated gene (genic) or not (extra-genic). The genic transfrags were then further classified as exonic if overlapping an exon. The number and percentage of transfrags within each category detected at each stage is shown in table 4.1.

Stage	Total transfrags	Genic	Percent	Exonic	Percent	Extra-genic	Percent
Embryo	36205	34886	96.36	33610	92.83	1319	3.64
L2	57564	53778	93.42	49499	85.99	3786	6.58
L3	49968	47717	95.50	45219	90.50	2251	4.50
L4	45770	43804	95.70	42050	91.87	1966	4.30
Young adult	46126	44139	95.69	42644	92.45	1987	4.31
Gravid adult	43507	41439	95.25	40045	92.04	2068	4.75

Table 4.1. Transfrag distribution at each developmental stage.

As is clear from table 1, the vast majority of transfrags detected are genic suggesting that the *C. elegans* genome is well annotated and there is not much novel transcription. This will be discussed further at the end of the chapter.

4.6. Measuring gene expression using tiling arrays

By the specification of the microarray design there is a probe every ~35bp, thus there are many probes per gene. Owing to the constraints of the array design, however, probes are not idealised and all behave differently. Furthermore for any given condition the probes that cover a gene which are above background may be different as a consequence of both biological and technical variability. Probes on a microarray are considered to behave differently as a consequence of their different binding capabilities owing to their different nucleotide constituents. The problem of how to derive a gene intensity from a set of probe intensities is therefore not as simple as taking the mean or median intensity across all probes above background as different probes will be used for each calculation. There are two possible methods of reducing technical variability introduced by using different probes for such a calculation. One approach is to correct for probe behaviour and the

other is to consider only exons and genes for which there is a sufficiently high number of probes for which there is signal above background. In the latter case the variability in individual probe intensity should be neutralized by the use of many probes.

The method of correcting for probe behaviour that has previously been used is to correct probe intensities from cDNA hybridizations relative to probe intensities derived from hybridization of genomic DNA (David *et al*., 2006). Hybridized genomic DNA is theoretically present at a ratio of 1:1 between probes and so the consequent probe intensities are representative of the binding characteristics of each probe. By this method all probe intensities should become more consistent relative to each other within a transcribed structure. Fewer probes should therefore be required to give a representative gene or exon intensity. Ultimately, however, this approach requires the optimization and performance of genomic hybridizations for potentially minimal gain as before an exon or gene could confidently called as expressed it is desirable that the majority of probes within any structure to be considered are above background. Furthermore for structures with relatively few probes above background it is possible that they are expressed at a low level but low intensity probes are more susceptible to errors in detection regardless of correction for probe behaviour. Calculating a gene intensity based on a small number of such probes is therefore inadvisable. Consequently we opted to stringently filter structures for which the majority of probes were above background and calculate intensities accordingly. Our criteria for doing this were to consider only exons for which more than 50% of probes were above background and only genes for which more than 50% of unique exons matched this criterion. We consider this to be reasonable as genes

not matching these criteria are generally too lowly expressed and probe intensities too close to the background cut-off to be considered accurate. The gene intensity is then taken as the median intensity of the probes filtered by the above criteria. Median rather than mean intensity was used, as this method is less susceptible to skewing by outlying probe intensities. The background threshold is calculated to include the top 5% of non-genic probes. A schematic of how gene intensities is calculated from both tiling array and Illumina sequence data is shown in figure 4.3.

4.7. Measuring expression using ultra-high density sequence data

The output of Illumina sequencing technology is ~3 million 35bp reads per sample sequenced. Alignment of reads uniquely mappable to the genome or annotated transcriptome leads to a certain number of reads overlapping each nucleotide. Each nucleotide can therefore be given an intensity score, which is the number of times it occurs in mapped reads. Gene intensities from sequence data are therefore calculated as the median number of reads that map to a nucleotide for which there is at least one read.

Table 4.2 shows the number of genes at each stage for which an intensity score can be derived by the criteria discussed for each of the technologies. The generation of gene intensities allows comparisons to be drawn between conditions to infer changes in overall gene expression or change in major splice form of genes. It is necessary, however, to demonstrate that these gene intensities are truly representative before such analyses can be undertaken. To this end gene intensities derived from tiling data were compared to gene intensities derived from sequence data. If these intensities look similar this is solid

evidence that the derived gene intensities are reliable and representative of true transcript abundance.

Stage	Tiling	Sequence	Overlap
Embryo	4471	NA	NA
L2	7323	NA	NA
L3	7208	NA	NA
L4	6355	7043	5164
Young adult	7220	6716	5681
Gravid adult	6577	NA	NA

Table 4.2. Number of genes called as expressed by each technology and the overlap between these lists.

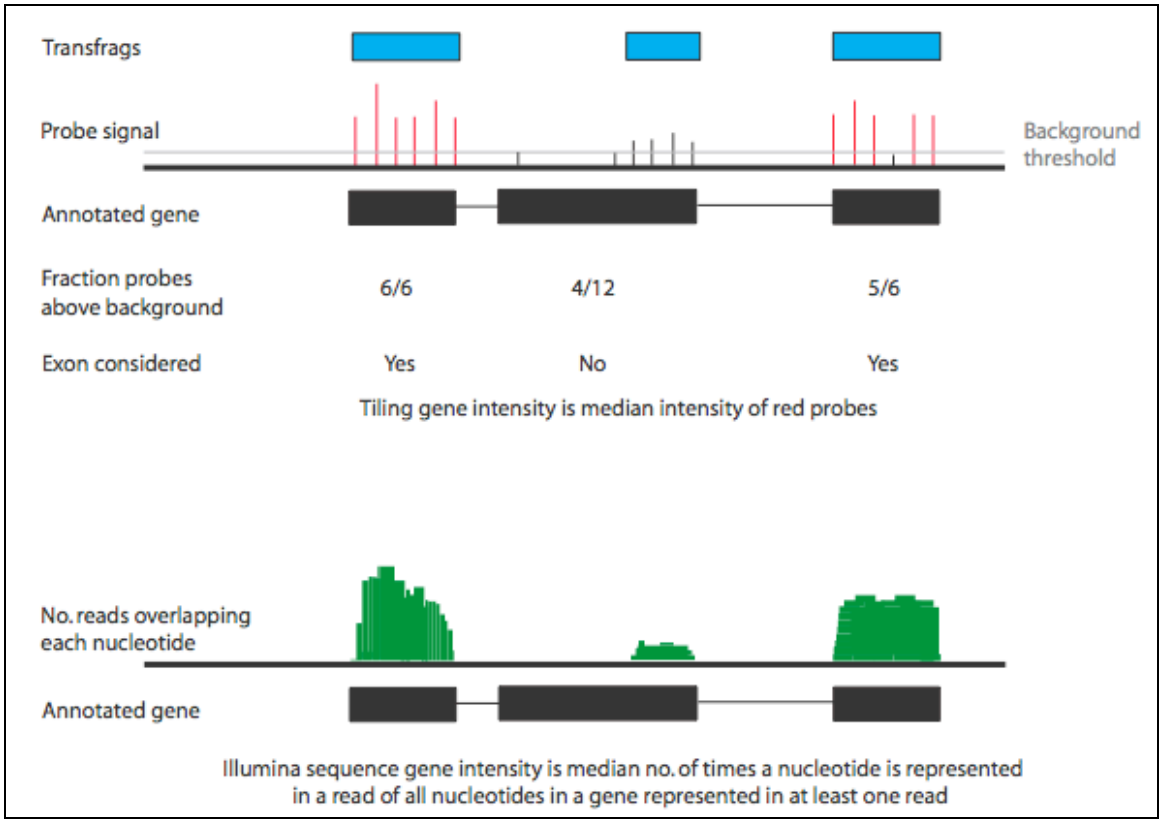

Figure 4.3. Calculating gene intensity values from tiling array and Illumina sequence data. For the tiling array data the gene intensity is the median probe intensity of all probes above background in exons for which ≥50% are above background (red probes). The background threshold is calculated to include the top 5% of non-genic probes on the array. The gene intensity derived sequence data is based on the number of times a base within a gene is represented within reads uniquely alignable to the genome. The gene intensity is the median number of times a single base is represented of all bases represented at least once within the gene.

Figure 4.4 shows the plot of gene intensities derived from the two different technologies. Gene intensities from the tiling data were binned at 0.1 increments of gene intensity (log_2 scale) and the mean gene intensity calculated. This was then plotted against the mean of gene intensities for the same genes in the sequence data. The plot indicates that there is good agreement (R = 0.82). Consequently we consider the intensities derived from our tiling data to be representative and usable. This correlation is greater than that previously reported for analogous comparisons made in *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* (Nagalakshmi *et al*., 2008; Wilhelm *et al*., 2008) (R = 0.68 and R = 0.48). The manner in which tiling and sequence expression scores were calculated between these studies and that presented here are different. Critically both of these studies compensate for the inevitable 5' drop-off observed in the sequence data caused by oligo(dT) priming, by calculating the sequence expression scores based on n (30 and 300) 3' coding nucleotides. This is feasible given sequence data of sufficient depth such that the 3' end for genes for which there are reads are always represented. Our data are not of this depth and consequently expression scores are the median count of detected nucleotides.

Despite the clear correlation between gene intensities generated by the two technologies as exhibited in figure 4.4, there are clear discrepancies, especially in the top bins. There are a number of potential causes of this. Firstly, only polyadenylated transcripts are considered by the sequencing technologies whereas total RNA is hybridized to the microarrays. Secondly, the technical difference between the two technologies, such as

amplification of the (ds)cDNA for sequencing are likely to lead to discrepancies. The former difference may be the more likely cause as the discrepancies are most marked for the most abundant transcripts. The correlation observed, however, is most striking leading us to believe that the derived gene intensities are representative and usable.

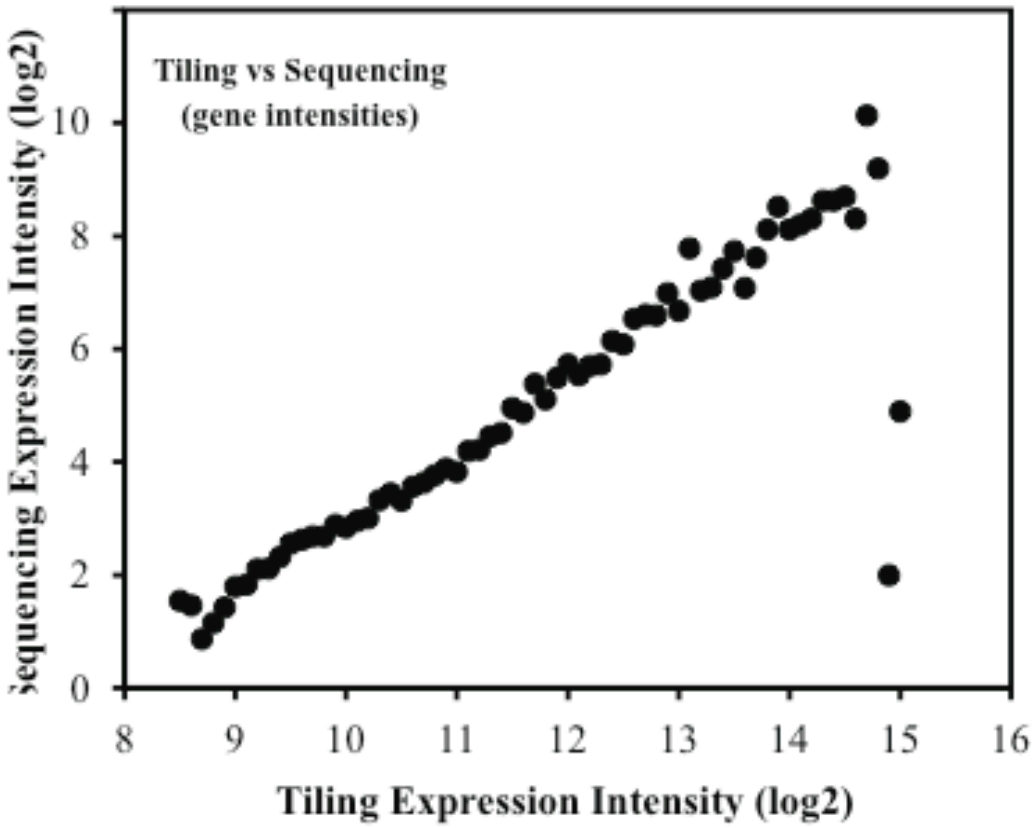

Figure 4.4. Correlation of gene intensities derived from tiling array and sequence data. Gene intensities from the tiling data were binned at 0.1 increments of gene intensity ($\log_2$ scale) and the mean gene intensity calculated. This was then plotted against the mean of gene intensities for the same genes in the sequence data. R = 0.82. This demonstrates good agreement between gene intensities derived from both technologies, thus validating our approach.

4.8. Validation of tiling data by sequence data

One method of validating the novel transfrags identified from the tiling array data would be exhaustive RT-PCR. This, however, would be time consuming and complicated by the fact that validation of small structures requires prior knowledge of their connectivity to surrounding structures. A more favourable alternative therefore is comparison of tiling data with ultra-high density sequence data. Not only does this allow the validation of novel transfrags, but should also allow them to be connected to other structures by identifying sequence reads which overlap transfrags. The number of transfrags identified by tiling arrays and validated by sequencing for stages at which we have stage-specific sequence data is shown in table 4.3. The ability of the sequence data to validate the tiling data is inevitably dependent on the depth of sequencing. It is clear then that the greater the intensity of the transfrag the more likely it is to be validated by the sequence data. The stringent background threshold set prior to the identification of transfrags, however, leads us to believe that were the sequence data of greater depth the rate of transfrag validation would have been consistently high across a greater range of transfrag intensities. We therefore consider our tiling data to be adequately validated and of a very high quality. That said, the marked difference in the fraction of genic and non-genic transfrags validated suggests that there may be a high rate of false discovery of novel transfrags.

The precise overlap between genes detected by the two technologies at all stages is illustrated in figure 4.5. Importantly, this is for genes called as expressed by the 50% criteria, rather than genes that have overlapping transfrags. It is these genes that will be

considered from this point on. The discrepancies between the two technologies are inevitably due to the differences in depth as well as stringency of the two technologies. The tiling data represents signal for more individual transcripts and is therefore of a greater depth than the sequence data. The presence of only one uniquely mappable read corresponding to a gene in the sequence data, however, is enough for that gene to be considered expressed whereas a transcript detected at a low level on the tiling array is more likely to be discarded as noise. Further to this, total RNA was hybridized to the tiling arrays whereas polyA+ RNA was used for sequencing in order to eliminate reads derived from rRNA. It is therefore inevitable that there will be differences in coverage by the two technologies.

Stage	Total transfrags	Genic	Exonic	Extra-genic	Total validated by seq	Percent
L4	45770	43804	42050	1966	42502	92.86
Young adult	46126	44139	42644	1987	42074	91.22
Stage	**Genic validated by seq**	**Percent**	**Exonic validated by seq**	**Percent**	**Extra-genic validated by seq**	**Percent**
L4	41521	94.79	40529	96.38	981	49.90
Young adult	40974	92.83	40152	94.16	1100	55.36

Table 4.3. Tiling array transfrags confirmed by sequencing. This table represents the proportions of genic, exonic and extra-genic transfrags validated by sequencing for the stages at which we have stage-specific sequence information. We note that more genic than extra-genic (novel) transfrags are validated by the sequence data. This may be due both to noise in our data and novel transcripts being expressed beneath the level of detection by the sequence data. ~91-93% of all transfrags are validated at stages for which we have stage-specific sequence data. We therefore consider our tiling data to be of high quality.

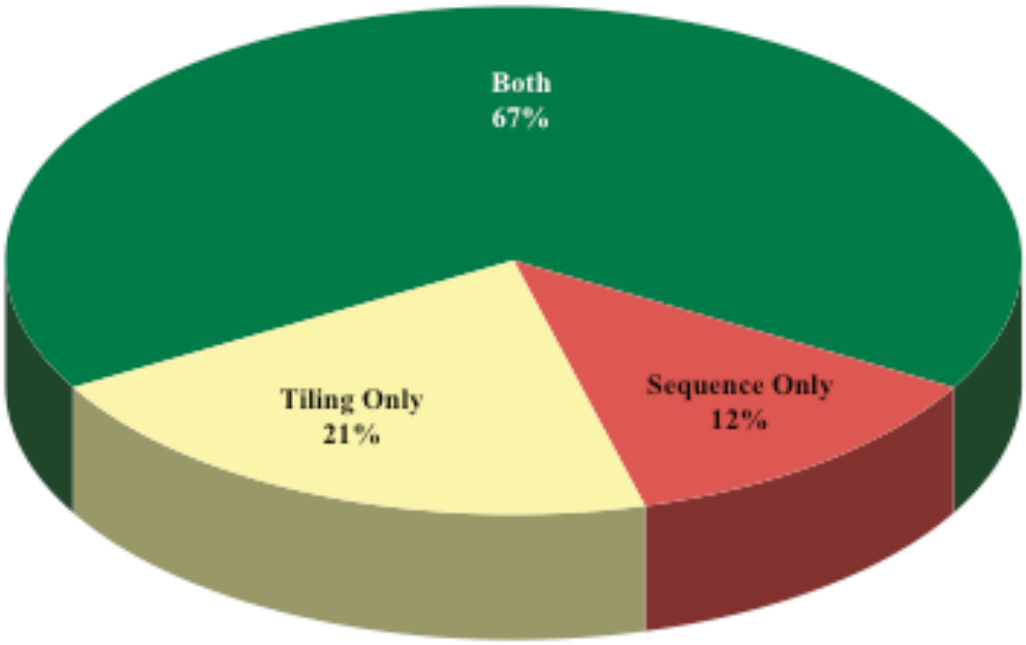

Figure 4.5. Overlap between genes detected by tiling arrays and by sequencing. Genes are defined as expressed in the tiling array data if ≥50% of probes per exon are above background and ≥50% of unique exons match that criterion. For a gene to be detected in the sequence data at least one uniquely mappable read must map to the gene.

4.9. Addressing alternative splicing using tiling data

High density tiling data theoretically allows the comparison of each exon of a gene in terms of expression level and in so doing, the identification of changes in major spliceform between conditions. Tiling arrays, however, can only provide data that allow the user to comment on differential inclusion of a given exon within the repertoire of splice forms of a gene. It gives no information in terms of connectivity of an exon to the other exons within a gene. Here we use our tiling data to generate a "splice index" (SI) for the change in expression of an exon relative to the expressed gene between conditions. More specifically, SI = $(E_i/G_i)t_1/(E_i/G_i)t_2$ where E_i is the median probe intensity above background of the exon, G_i of the gene and t_1 and t_2 are the different timepoints. The SI is used to infer a major change in splice form. It is essentially a

measure of how the intensity of a given exon changes relative to the whole gene between developmental stages. This then allows us to compare the genes for which a change in spliceform is detected to those for which different splice forms are known.

It is essential that intensity values can be assigned to exons with high confidence. For this reason exons with fewer than three probes were omitted from the analysis. At least one exon changes at least 2-fold in 5% of detected genes, which is to say that 5% of detected genes clearly exhibit a change in major isoform across development. While 18% of annotated genes have at least two annotated isoforms, this is the first systematic analysis of how these isoforms change across development. Of the 870 genes that show a change in spliceform between any two stages (>2-fold change for any given exon), 459 have multiple annotated isoforms in WS150. The remaining 47% of the genes we detect by this method are therefore not predicted to be alternatively spliced in WS150. These 411 genes, however, correspond to only 2% of annotated coding genes. This therefore does not conclusively demonstrate that alternative splicing is grossly underrepresented in current gene annotations. Since we detect less than 5% of genes to be alternatively spliced at this fold-change in SI, however, it may be that a relaxation of this threshold would show that the trend continues as the gene list expands. This would inevitably lead to an increase in false discovery, however, and it seems that our sequence data may offer a better alternative in addressing alternative splicing.

4.10. Addressing alternative splicing using sequence data

We have demonstrated that tiling array data can be used to indicate changes in major splice form for expressed genes between conditions. Connectivity of exons, however, cannot be inferred from tiling data. High-density sequence data can be used to this end by looking for reads that span exon-exon boundaries. This is the single biggest advantage of sequence data over tiling data – the information it provides on connectivity within expressed structures. The proportion of reads that span any set of exon boundaries relative to another may give an indication of the relative combinations of exons used in a given condition. This would be extremely useful in that it gives information on exon connectivity within transcripts. The methodology involves identifying sequence reads that do not map to the genome. These reads are then aligned with all combinations of adjacent and non-adjacent exons for all annotated isoforms of all genes using Maq. The output of this is reads that map to annotated exon-exon junctions and reads which span previously unidentified exon-exon junctions for annotated exons. The technique is therefore limited by the accuracy and completeness of exon boundary annotations. A schematic of the approach is shown in figure 4.6 and an example of the output in figure 4.7. A summary of the number of reads mapping to the genome and spanning annotated and non-annotated exon-exon boundaries across all samples is shown in table 4.4.

Ultimately sequence data at the required depth may render tiling data completely redundant in addressing alternative splicing. A constant issue which we face in our tiling analysis is what fold-changes are reasonable cut-offs, allowing us to call events. This is not an issue with sequence data for this analysis. If we identify a uniquely mappable read

spanning an exon-exon junction it is reasonable to assume that those exons are connected, allowing us to determine changes in spliceform. This also allows us to identify exons that are connected where no such connectivity exists in current gene annotations. Though the depth of our sequence data is inadequate to comprehensively map all splice events and spliceform changes at this stage, our current data show unannotated splice events for ~1% of detected genes. Critically, ~80% of genes identified to have alternative splicing in this manner also were found to have at least one exon with a splice index ≥1.5 by tiling analysis, confirming that the changes in transcript structure that we monitor by tiling analysis are likely to be real. Thus the high resolution mapping of the transcriptome using tiling arrays and the gene expression levels and transcript structural features that we derive from these data appear to be accurate.

	NO. OF NUCLEOTIDES	PERCENTAGE OF TOTAL
Nucleotides aligned to genome at ≥Q30	621610325	73.02
Nucleotides aligned to annotated transcriptome at ≥Q30	36085206	4.24
Nucleotides aligned to non-adjacent exons at ≥Q30	47205	0.01
Non-aligned nucleotides	193584559	22.74
Total Nucleotides	851327295	100.00

Table 4.4. Reads mapping to the genome and spanning exon-exon boundaries. Shown are the number of nucleotides across all samples mapped to the genome, and transcriptome as described above using Maq version 0.6.6 at a mapping quality of ≥Q30.

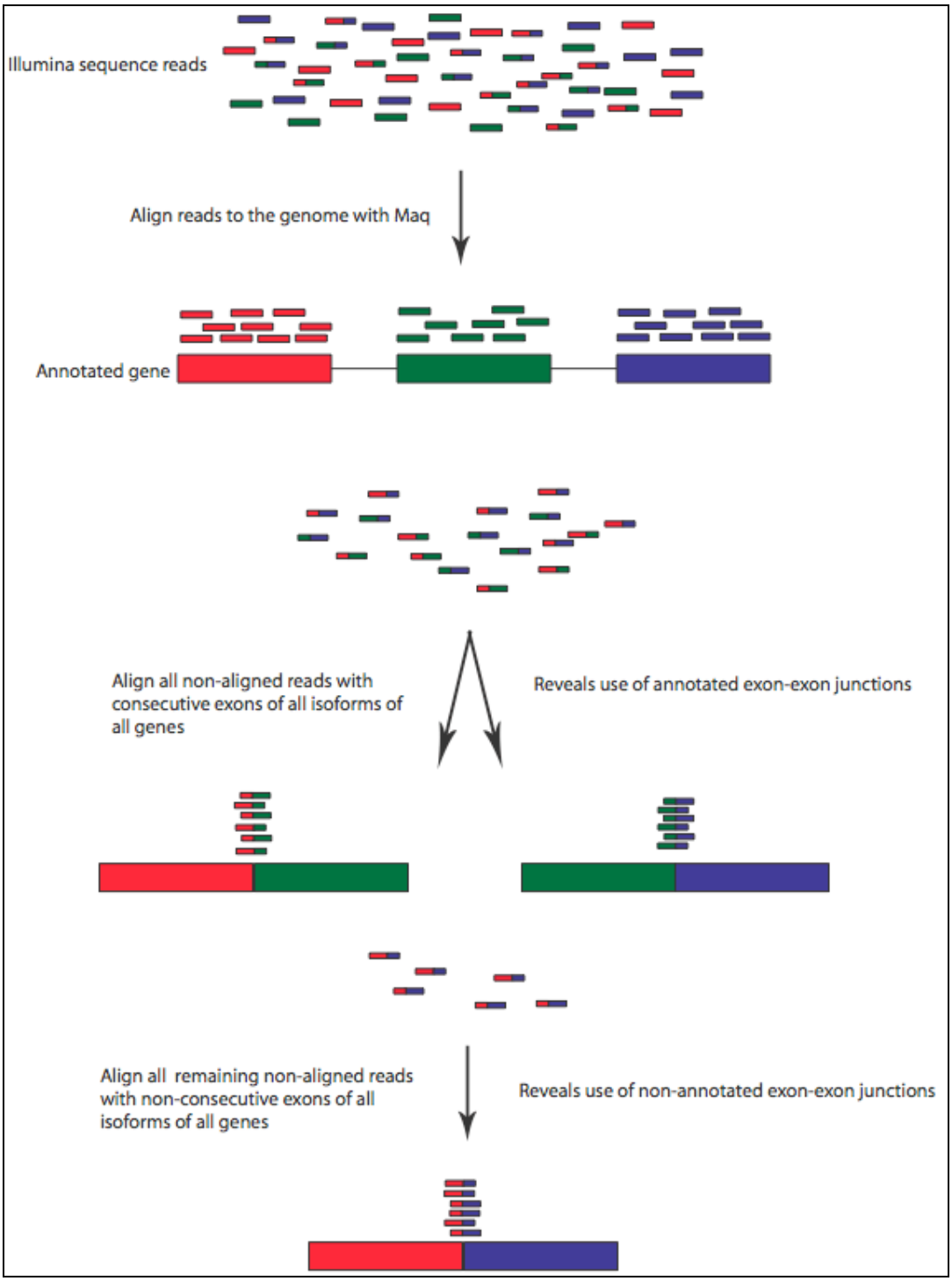

Figure 4.6. Use of Illumina sequence reads to identify utilized exon-exon junctions. Alignment of uniquely mappable reads to the genome using Maq removes reads not spanning exon boundaries and can be used to generate gene intensities. The remaining reads are aligned with consecutive exons for all isoforms of all genes. This reveals annotated exon-exon junctions. The remaining reads are then aligned to all combinations of non-consecutive exons for all isoforms of all genes. This reveals non-annotated exon-exon junctions.

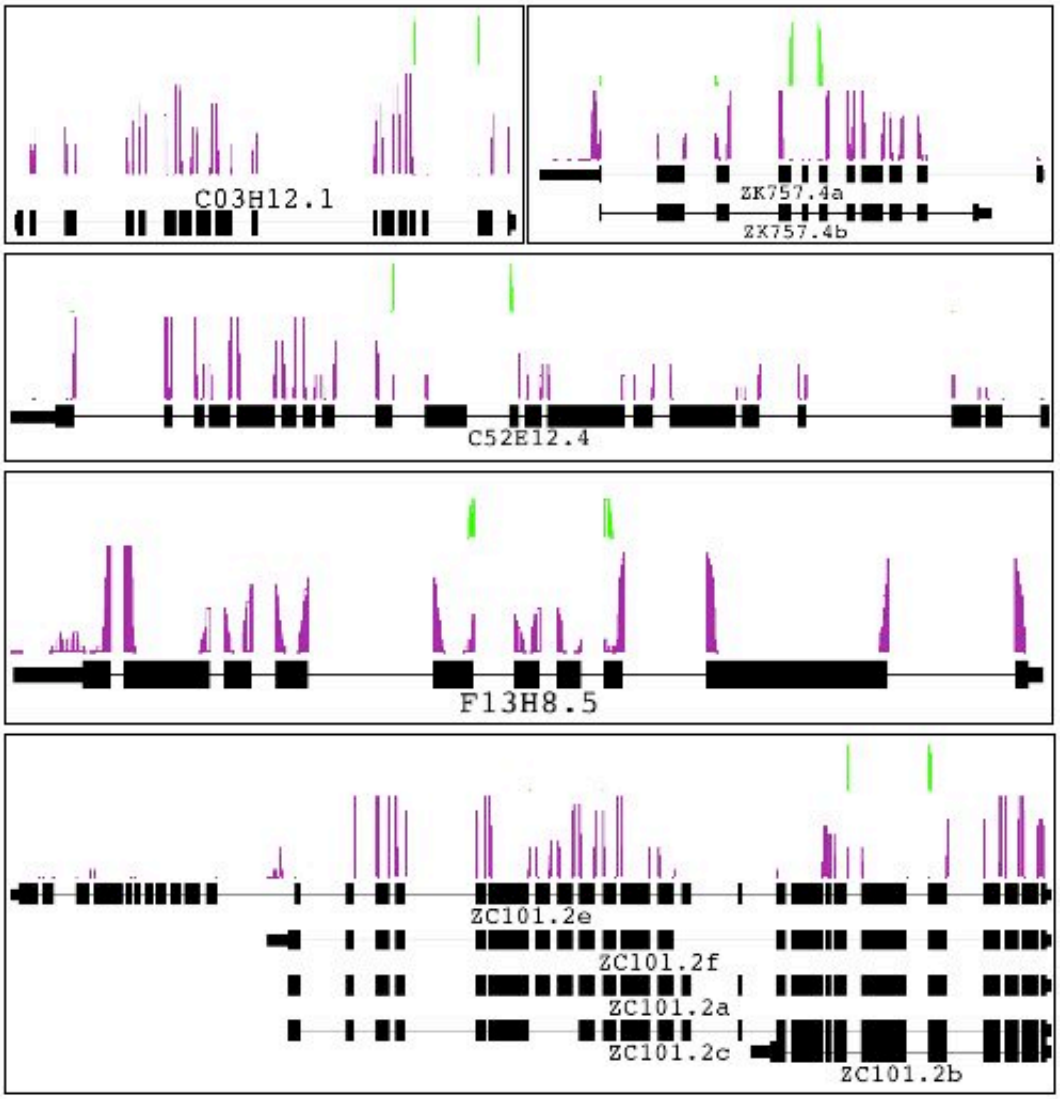

Figure 4.7. Ultra-high density sequence reads reveal novel splice junctions. Illumina sequence reads which cannot be aligned to the genome are aligned to adjacent annotated exons and all combinations of non-adjacent exons for all isoforms of all genes with Maq. Reads spanning annotated exon boundaries are shown in purple. Novel exon boundaries are shown in green. Relative numbers of reads spanning each exon-exon junction may reveal relative usage. At the current depth of sequencing 1% of genes appear to undergo at least one novel splice event.

4.11. Discussion

The work presented here clearly demonstrates the utility and complementarity of these technologies in forwarding our knowledge and understanding of gene annotations. It represents the first splicing analysis of its kind in *C. elegans* and the potential to become the most comprehensive analysis of its kind in any organism. The utility of the approaches developed in this work have been clearly demonstrated, as has the redundancy of tiling array data given the resolution and connectivity information of ultra-high density sequence data. Tiling array data, however, represents a more cost-effective approach in addressing the same questions. Sequence data to a greater depth will be required in order to more completely identify the complete repertoire of exon-exon junctions. At the time of printing this thesis sequence data had been produced at 15x the depth of the data utilized here. The tools are now in place to utilize these data to great effect. The splicing analysis performed using the tiling data does imply that alternative splicing is far more prevalent than can be accounted for by current annotations. It would be most interesting to see if this is further borne out by the newly acquired sequence data.

The dearth of novel transfrags detected in the tiling data and the low frequency of their validation by the sequence data suggest that the genome of *C. elegans* is well annotated. We do, however, provide clear evidence for novel transcription. Furthermore we do not discount the possibility that many of the novel transfrags which were not validated are expressed at a low level are beneath the level of detection of our sequence data, the scarcity of these transcripts being in part causative of their prior anonymity. Also, certain developmental stages are inevitably under-represented in the sequence data, leading to a

reduced possibility of validating novel transfrags detected at these stages. It will be interesting to see how many more transfrags are validated by the newly acquired sequence data.

The identity of these novel transfrags as additional exons of annotated genes, entire novel coding genes, or non-coding transcripts is yet to be tackled. This can be addressed using the sequence data but represents a more complex problem than the study of connectivity between annotated exons. Our splicing analysis thus far had involved looking for reads that span annotated exon boundaries. No such boundaries are defined by the transfrags or novel sequence reads mapped to the genome. A shotgun approach to assembling sequence reads into transcripts may represent the best possibility of connecting transcribed units. Whatever the approach taken and whomever implements it, it is likely to be extremely complex and computationally intensive.

Regarding the scarcity of novel transfrags, validated or otherwise, relative to analogous studies in other organisms – perhaps this is not surprising. The density of gene annotation in *C. elegans* surpasses that of human and *Drosophila*. Furthermore our study considered whole animals, i.e. all cells and tissues at once. If there are low levels of tissue-specific expression of novel genes we were unlikely to detect them in this study. Regardless, the output of this study has proven it a worthwhile undertaking and an ideal dataset for comparison with the nonsense-mediated mRNA decay transcriptome as we shall see. In terms of data quality, it is noted that markedly fewer genes are detected for any condition than were seen using two-colour microarrays in chapter 3. Importantly, the

expression microarrays used in chapter 3 have only one probe per gene, and are 70mers rather than the 25mer probes on Affymetrix arrays. The increased specificity per probed, coupled with the greater probe number per gene give us greater confidence in the output of the Affymetrix microarrays, even if the depth of detection is lesser. Our confidence in these data is further strengthened by the correlation of gene intensities between our tiling array and Illumina sequence data. We therefore consider the datasets presented in this chapter and the methodologies applied to them to be an ideal framework for comparison with the nonsense-mediated mRNA decay deficient transcriptome.