

Analysis of the  
transcriptomes of wild-type  
and mutant *C. elegans*

Andrew Christopher Nelson

This dissertation is submitted for the  
degree of Doctor of Philosophy  
September 2008

Corpus Christi College  
University of Cambridge

The Wellcome Trust Sanger Institute  
Hinxton  
Cambridge, UK

## **Declaration**

I hereby declare that my dissertation contains material that has not been submitted for a degree or diploma or any other qualification at any other university. This thesis describes my own work and does not include work that has been done in collaboration, except when specifically indicated in the text.

Andrew C. Nelson

29/09/2008

## Abstract

A key question in biology is how genotype can inform us of phenotype. For model organisms, most phenotypes reported have been at the level of the morphology and behaviour of the whole organism. However, recent advances in technology allow gene expression to be assessed on a genome-wide scale and pioneering work in yeast has shown that such expression profiles can be used as high density, quantitative phenotypes. I wanted to test whether expression profiles can also serve as useful phenotypes of whole animals rather than single cells. More specifically I sought to test whether the expression profiles resulting from perturbations of genes in one pathway looked more like those of other perturbations of the same pathway than another pathway. To do this I used two-colour DNA expression microarrays to survey gene expression in the nematode *Caenorhabditis elegans*. Expression profiles were produced for a number of different worm strains with mono-genic perturbations in different pathways involved in germline development. Clustering of the resulting expression profiles rediscovered the known pathways. This then allowed me to query perturbations of candidate modulators of EGF signalling against the compendium of expression profiles. I conclude that, as in yeast, expression profiles serve as reliable high-density phenotypes that allow meaningful biological comparisons to be drawn.

The quality of an expression microarray can only be as high as the gene annotations on which it is based. I therefore sought to evaluate how well characterised the transcribed genome of *C. elegans* is. To do this I used a combination of whole genome tiled microarrays and ultra-high density sequencing to assess the transcript complement of whole animals throughout development. We found that the vast majority (~95%) of expression is genic but the combinations and numbers of splice sites used are greater than previously predicted, suggesting that current annotations are largely complete, but that our knowledge of splice variation across development is still far from finished.

Whilst surveying transcripts in wild-type animals yields valuable data, it is known that there are many transcripts that are produced and subsequently degraded by the nonsense-mediated mRNA decay pathway (NMD). To identify these transcripts we compared the transcripts of wild-type animals to those of mutants of the NMD pathway. We find that ~13% of endogenous genes are NMD targets. The majority of these transcripts have upstream start codons in the 5' UTR or are alternatively spliced leading to a premature in-frame stop codon. Finally, we find that ~10% of all gene expression changes throughout development require NMD and thus that NMD is a bona fide regulator of gene expression.

<b>DECLARATION .....</b>	<b>II</b>
<b>ABSTRACT.....</b>	<b>III</b>
<b>LIST OF FIGURES .....</b>	<b>VI</b>
<b>LIST OF TABLES.....</b>	<b>VI</b>
<b>CHAPTER 1 - INTRODUCTION .....</b>	<b>1</b>
1.1. OUTLINE .....	2
1.2. <i>CAENORHABDITIS ELEGANS</i> AS A MODEL SYSTEM.....	3
1.2.1. <i>The germline</i> .....	5
1.2.2. <i>The vulva</i> .....	18
1.3. RNA INTERFERENCE IN <i>CAENORHABDITIS ELEGANS</i> .....	24
1.3.1. <i>The mechanism of dsRNA-induced gene silencing in C.elegans</i> .....	24
1.3.2. <i>RNAi by feeding</i> .....	28
1.4. MICROARRAY TECHNOLOGIES .....	30
1.5. MICROARRAYS AS A PHENOTYPING TOOL .....	32
1.6. AIMS OF CHAPTER 3 .....	34
1.7. TRANSCRIPTOME INTERROGATION.....	36
1.8. NONSENSE-MEDIATED MRNA DECAY.....	37
1.9. METHODS OF SURVEYING THE TRANSCRIPTOME .....	44
<b>CHAPTER 2 - MATERIALS AND METHODS .....</b>	<b>48</b>
2.1. REAGENTS .....	49
2.1.1. <i>C. elegans</i> .....	49
2.1.2. <i>Bacteria</i> .....	51
2.1.3. <i>Buffers used for Affymetrix tiling microarray hybridization and processing</i> .....	52
2.1.4. <i>10x PCR reaction buffer</i> .....	54
2.2. PROTOCOLS.....	55
2.2.1. <i>Maintenance of C. elegans stocks</i> .....	55
2.2.2. <i>Bleach sterilization of C. elegans strains and synchronization</i> .....	55
2.2.3. <i>Freezing and recovery of C. elegans stocks</i> .....	55
2.2.4. <i>RNAi by feeding on plates, RNA extraction and visual phenotyping</i> .....	56
2.2.5. <i>DAPI staining</i> .....	57
2.2.6. <i>Generation of mixed-stage RNA reference sample</i> .....	58
2.2.7. <i>RNA labelling and two-colour microarray hybridization</i> .....	57
2.2.8. <i>Affymetrix tiling microarray hybridization</i> .....	59
2.2.9. <i>(ds)cDNA production for Illumina sequencing</i> .....	63
2.2.10. <i>Reverse transcription and PCR</i> .....	63
2.2.11. <i>Two-colour expression microarray data analysis</i> .....	64
2.2.12. <i>Identifying transcribed regions and visualization of tiling microarray data</i> .....	65
2.2.13. <i>Affymetrix tiling microarray expression data analysis</i> .....	65
2.2.14. <i>Illumina sequence data analysis</i> .....	66
<b>CHAPTER 3 - MICROARRAY ANALYSIS OF GERMLINE PERTURBATIONS .....</b>	<b>67</b>
3.1. INTRODUCTION .....	68
3.2. OUTLINE OF APPROACH .....	71
3.3. INITIAL MICROARRAY DATA PROCESSING, NORMALISATION AND ASSESSMENT OF DATA QUALITY .....	76
3.4. PROOF-OF-PRINCIPLE EXPERIMENTS .....	80
3.5. LOW-RESOLUTION PHENOTYPIC ANALYSIS OF PATHWAY PERTURBATIONS.....	85
3.6. IDENTIFICATION OF NOVEL MODULATORS OF RAS/MAPK SIGNALLING IN THE GERMLINE.....	90
3.7. THE DIFFERENTIALLY EXPRESSED GENES.....	98
3.8. DISCUSSION .....	100

<b>CHAPTER 4 - ANALYSIS OF THE WILD-TYPE <i>C. ELEGANS</i> TRANSCRIPTOME .....</b>	<b>103</b>
4.1. INTRODUCTION .....	104
4.2. TILING ARRAY DATA NORMALIZATION .....	106
4.3. DEFINING REGIONS OF TILING ARRAY SIGNAL ALONG GENOMIC COORDINATES .....	107
4.4. IDEALIZING PARAMETERS FOR BUILDING TRANSFRAGS .....	111
4.5. COMPARISON OF TRANSFRAGS WITH THE GENOME .....	112
4.6. MEASURING GENE EXPRESSION USING TILING ARRAYS.....	113
4.7. MEASURING EXPRESSION USING ULTRA-HIGH DENSITY SEQUENCE DATA.....	115
4.8. VALIDATION OF TILING DATA BY SEQUENCE DATA.....	119
4.9. ADDRESSING ALTERNATIVE SPLICING USING TILING DATA .....	121
4.10. ADDRESSING ALTERNATIVE SPLICING USING SEQUENCE DATA .....	123
4.11. DISCUSSION .....	127
<b>CHAPTER 5 - NONSENSE-MEDIATED MRNA DECAY IS A REGULATOR OF DEVELOPMENTAL GENE EXPRESSION.....</b>	<b>130</b>
5.1. INTRODUCTION .....	131
5.2. THE TARGETS OF NMD.....	133
5.3. STRUCTURAL FEATURES WHICH DEFINE NMD TARGETS .....	134
5.4. TRANSLATION INITIATION AND NMD.....	145
5.5. NMD REGULATES THE EXPRESSION OF GENES IN OPERONS .....	149
5.6. NMD REGULATES DEVELOPMENTAL GENE EXPRESSION.....	153
5.7. GLD-1 AS A PROTECTOR OF TRANSCRIPTS FROM NMD.....	159
5.8. DISCUSSION .....	163
<b>CHAPTER 6 - GENERAL DISCUSSION AND FUTURE WORK.....</b>	<b>168</b>
<b>REFERENCES .....</b>	<b>178</b>

## List of Figures

Figure 1.1.	Cartoon representation of gonadogenesis	7
Figure 1.2.	Regulation of the mitosis/meiosis decision by the interplay of pro- and anti-meiotic factors	13
Figure 1.3.	The canonical EGF/ras/MAPK and Notch signalling pathways as they are known to act in the vulva and germline	17
Figure 1.4.	Vulval specification and lineage	22
Figure 1.5.	Mechanism of RNAi gene silencing	28
Figure 1.6.	L4440 RNA interference feeding vector	30
Figure 1.7.	The recognized post-transcriptional causes of NMD targeting	43
Figure 1.8.	Technical flow-through of Affymetrix tiling array and Illumina sequencing technologies	48
Figure 3.1.	Clustering of differentially expressed genes between N2 and each genic perturbation	85
Figure 3.2.	Relative fecundity of germline perturbations	88
Figure 3.3.	DAPI staining of whole animals to assess quantity of germline	90
Figure 3.4.	Screening for modulators of EGF/ras/MAPK signalling in the vulva	92
Figure 3.5.	<i>pkc-1</i> clusters with the EGF/ras/MAPK signalling pathway	96
Figure 3.6.	The activity of PKC is modulated by the activities of PLC and DGK	98
Figure 4.1.	Transfrags corresponding to transcribed genes	112
Figure 4.2.	Selection of transfrag building parameters schematic	114
Figure 4.3.	Calculating gene intensity values from tiling array and Illumina sequence data	118
Figure 4.4.	Correlation of gene intensities derived from tiling array and sequence data	120
Figure 4.5.	Overlap between genes detected by tiling arrays and by sequencing	123
Figure 4.6.	Use of Illumina sequence reads to identify utilized exon-exon junctions	127
Figure 4.7.	Ultra-high density sequence reads reveal novel splice junctions	128
Figure 5.1.	Structural changes in SR gene transcripts leading to NMD	140-142
Figure 5.2.	Increasing 5' UTR length correlates with increased magnitude of NMD	146
Figure 5.3.	An A nucleotide -3 of the annotated start codon correlates with NMD regulation	150
Figure 5.4.	Examples of operonic gene regulation by NMD	153
Figure 5.5.	NMD regulation via a shift in promoter usage	154
Figure 5.6.	Structural changes leading to NMD targeting	159
Figure 5.7.	Model of gene regulation by NMD	168

## List of Tables

Table 1.1.	Components of the NMD machinery known to exist in model organisms	39
Table 3.1.	Genes involved in germline development perturbed in this study	74
Table 3.2.	Relative Pearson correlations using different normalization methods	79
Table 3.3.	Genes upregulated and downregulated relative to N2 for each condition	82
Table 3.4.	Selected genes suppressing the Muv phenotype in RNAi screens in 100% Muv mutants.	94
Table 4.1.	Transfrag distribution at each developmental stage.	115
Table 4.2.	Number of genes detected by each technology and overlap.	118
Table 4.3.	Tiling array transfrags confirmed by sequencing.	122
Table 4.4.	Reads mapping to the genome and spanning exon-exon boundaries.	126
Table 5.1.	Novel NMD regulated genes detected on <i>gld-1(RNAi)</i> .	164