

The Genetics of Variation in Gene Expression

Chris J. Cotsapas

A thesis submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy

University of New South Wales

2005

Abstract

The majority of genetic differences between species and individuals have been hypothesised to impact on the regulation, rather than the structure, of genes. As the details of genetic variation are uncovered by the various genome sequencing projects, understanding the functional effects on gene regulation will be key to uncovering the molecular mechanisms underlying the genesis and inheritance of common phenotypes, such as complex human disease and commercially important traits in plants and animals. Unlike coding sequence polymorphisms, genetic variants affecting gene expression will reside in the transcriptional machinery and its regulatory inputs. As these are largely specific to cell- or tissue-types, we would expect that regulatory variants will also affect final mRNA levels in a tissue specific manner. Genetic variation between individuals may therefore be more complex than the sum total of sequence differences between them.

Demonstrating this hypothesis is the main focus of this thesis. We use microarrays to measure mRNA levels of approximately 22,000 transcripts in inbred and recombinant inbred strains of mice, and present compelling evidence that the genetic influences on these levels are tissue-specific in at least 85% of cases. We uncover two loci which apparently influence transcript levels of multiple genes in a tissue-specific manner. We also present evidence that failure of microarray data normalisation may cause spurious linkage of expression phenotypes leading to erroneous biological conclusions, and detail a novel, extensible mathematical framework for performing tailored normalisation which can remove such systematic bias. The wider context of these results is then discussed.

Contents

1	Introduction	11
1.1	Summary	12
1.2	Detection strategies	13
1.2.1	Allelic discrimination	13
1.2.2	Genetical genomics	15
1.3	Genetics of regulatory variation	18
1.3.1	Extent of genetic effects	18
1.3.2	<i>cis</i> , <i>trans</i> , and master regulators	20
1.3.3	Heritability, epistasis, and the number of determinants	21
1.3.4	Tissue specificity	22
1.4	Biological implications	23
1.5	Outline	25
2	Microarray normalisation for genetical genomics	26
2.1	Introduction	27
2.1.1	A note on microarray data visualisation	28
2.2	Normalisation – mathematical bias removal	29
2.2.1	Scaling	30
2.2.2	Analysis of Variance	31
2.2.3	Principal Components Analysis	31
2.2.4	Intensity–dependent smoothing	33
2.3	Correcting multiple non–linear biases in microarray data	34
2.3.1	Non–linear artefacts	35
2.3.2	Additive model normalisation	36

2.4	Failure of normalisation in genetical genomics experiments . . .	41
2.4.1	Experimental design	42
2.4.2	Lack of agreement between normalisation results . . .	43
2.5	Conclusions	46
2.6	Materials and Methods	47
2.6.1	Sample handling	47
2.6.2	Expression profiling	47
2.6.3	Normalisation	48
2.6.4	Linkage analysis	48
3	Genetic influence on mRNA levels is tissue specific	49
3.1	Introduction	50
3.2	Experimental design	51
3.3	Tissue specificity of influences on gene expression	53
3.3.1	The majority of genetic influences are tissue specific .	54
3.3.2	Expression levels do not reflect complexity of influences	56
3.4	Functional bias in influenced transcripts	57
3.4.1	Over-representation of functional themes	59
3.5	Discussion	61
3.6	Materials and Methods	63
3.6.1	RNA preparation	63
3.6.2	Microarray hybridisation and washing	63
3.6.3	Data processing	64
3.6.4	Overlap analysis	64
3.6.5	Detecting changes in expression between strains	64
3.6.6	Gene Ontology analysis	66
4	Dissection of genetic influences on mRNA levels in a Re-	67
	combinant Inbred panel	
4.1	Introduction	68
4.2	Experimental design	69
4.2.1	Moderated linkage statistics	70
4.3	Independent tissue analysis	72

4.3.1	Linkage complexity	73
4.4	Expression level correlation analysis detects biological themes under genetic influence	74
4.4.1	Correlation analysis of genetically variant expression levels identifies biological pathways	75
4.4.2	Correlated clusters have common genetic determinants	76
4.5	Discussion	77
4.6	Materials and Methods	79
4.6.1	RNA preparation	79
4.6.2	Microarray hybridisation and washing	80
4.6.3	Data processing	80
4.6.4	Linkage analysis	81
5	Resolvable genetic determinants of mRNA levels are tissue specific	82
5.1	Introduction	83
5.2	Experimental design	84
5.3	Tissue specificity of parentally influenced genes	84
5.4	Transgressive segregation of mRNA levels	87
5.5	Some loci influence multiple transcript mRNA levels	89
5.5.1	A region on chromosome 1 influences multiple tran- scripts in brain	90
5.5.2	A region on chromosome 8 influences transcripts in all tissues	92
5.6	Discussion	94
5.7	Materials and Methods	97
5.7.1	RNA preparation	97
5.7.2	Microarray hybridisation and washing	97
5.7.3	Data processing	98
5.7.4	Linkage analysis	98
5.7.5	Overlap analysis	99

6 Discussion	100
6.1 Results summary	101
6.2 Defining regulatory circuits	102
6.2.1 Regulatory interactions as networks	103
6.2.2 Tissue specific regulatory interactions	104
6.2.3 Mapping regulatory metaphenotypes	105
6.3 The implications of tissue specificity	106
6.3.1 Understanding relationships between individuals . . .	107
Literature cited	109
A Significant linkages in the BxD panel	128
B Over-represented GO terms in correlation clusters	204

List of Tables

1.1	Summary of genetic influences on gene expression levels . . .	19
2.1	Effect of normalisation on gene identification	44
2.2	Effect of normalisation on locus identification	45
3.1	Genetic influences on mRNA levels in each tissue	54
3.2	Expression of genetically influenced genes across tissues . . .	57
3.3	Extrapolating genetic influence between tissues	58
3.4	Enriched Gene Ontology terms for genetically influenced genes.	60
4.1	Linkage results in three tissues	73
4.2	Complexity of expression variation	74
4.3	Linkage aggregation in RI brain	76
4.4	Linkage aggregation in RI kidney	77
4.5	Linkage aggregation in RI liver	78
5.1	Linkage results for genetically influenced genes	84
5.2	<i>cis</i> and <i>trans</i> effects on gene expression	86
5.3	Genetic dissection of transgressive mRNA levels	88
5.4	<i>cis</i> and <i>trans</i> effects on transgressively segregating genes . . .	89
5.5	A locus affecting multiple transcripts in brain	90
5.6	Loci affecting transcript levels in multiple tissues	92
5.7	Linkage significances for genes affected by chromosome 8 . . .	93
A.1	Gene linkages in brain	129

LIST OF TABLES

A.2	Gene linkages in kidney	186
A.3	Gene linkages in liver	195
B.1	Over-represented GO terms in BxD brain	205
B.2	Over-represented GO terms in BxD kidney	206
B.3	Over-represented GO terms in BxD liver	207

List of Figures

2.1	Basic microarray visualisation	29
2.2	Non-linear biases in microarray data	37
2.3	Systematic biases in microarrays	39
2.4	Systematic biases in microarrays	39
2.5	Deposition order biases in each channel	40
3.1	Tissue-specific genetic influence on mRNA levels	55
4.1	Significance comparisons for linkage analysis	71
5.1	Overlap of linkage results for genetically influenced genes	85
5.2	Overlap for transgressive genes	88
5.3	Effects mapping to chromosome 1	91
5.4	Linkage scan showing effects in all tissues	93

Acknowledgments

This thesis is the culmination of four years of work with Professor Peter Little, without whom it would have been impossible. I owe him a great debt of gratitude for years of support and friendship, but most of all for the patience he exhibited whilst I transformed myself from dilettante to scientist. The trust and friendship he has extended to me have been an honour and a privilege, and I can only hope they have not been misplaced.

Dr. Rohan Williams has played a central role in this work, as collaborator, office-mate, food buddy, aesthetic sounding-board and purveyor of exotic quantitative methods. His broad knowledge and burning curiosity made our wide-ranging scientific conversations fascinating, and he has taught me much. The Gene Ontology analyses presented in Chapter 3 are his work, and many points in the Discussion have arisen as a consequence of brainstorming sessions.

The statistical aspects of this work are the fruit of a collaboration with the School of Mathematics, UNSW, with Professors William Dunsmuir and Matt Wand, and especially Dr. David Nott. David has gently tutored me into at least passing competence with elementary statistics; the B-statistic modifications in Chapter 3 are his work, and he has had a major influence on the normalisation methods presented in Chapter 2. He has kindly proof-read and corrected some parts of this work. The additive model normalisation procedure in Chapter 2 was devised by Professor Matt Wand, who also wrote an early implementation.

The material presented here rests, directly or indirectly, on the work of other members of our laboratory. The expression data in Chapter 3 was generated by Jeremy Pulvers as part of his Honours project; Eva Chan kindly provided genetic map data; and Mark Cowley has contributed program code for several analyses. They have provided a stimulating environment to work in, and I am fortunate to count them as friends and colleagues.

Bronwyn Robertson and Geoff Kornfeld of the Ramaciotti Centre for Gene Function Analysis, UNSW, provided microarrays and facilities for the experiments described here. A/Prof. Russell Standish of the High Perfor-

mance Computing Support Unit at UNSW wrote an optimised implementation of the bootstrapped Student's t -test used in Chapter 4, and kindly arranged for compute time on the Barossa cluster. John Schimenti (Jackson Laboratories) and Maja Bucan (University of Pennsylvania) and their laboratories kindly assisted in obtaining BxD mouse strains. They all have my sincere thanks.

On a personal note, I would never have embarked on so ambitious a move without the support of my family. They had always taught me that everything was possible, and although I suspect this is not what they had in mind, were directly responsible for me moving hemispheres on a whimsical decision to follow science. Friends I have made in Australia made me welcome and provided respite and refreshment over the last four years: Neil Saunders, Stephen Harrop, and Greg Tyrelle as Team Linux; Kai and Kerry Schindlmayr for endless friendship and hospitality; Julie Lim, Emma Collinson, Yael Azriel were my female consciences; Jo Gibson forebore to judge, and Laura Coleman taught me I was not alone. All have my thanks and love.