

Gene Expression Biomarkers for Colorectal Neoplasia

by

L. C. LaPointe

B.Sc. (Florida State University) 1991

Department of Medicine

FLINDERS UNIVERSITY OF SOUTH AUSTRALIA

Committee in charge:

Prof. Graeme P. Young, Chair

Dr. Robert A. Dunne

Dr. Peter L. Molloy 2008

Contents

1	Introduction	1
1.0.1	Colorectal neoplasia	3
1.0.2	Adenomas as a target for cancer prevention	4
2	Review of Colorectal Gene Expression	8
2.1	Gene expression in the large intestine	9
2.1.1	Gene expression patterned during organogenesis	9
2.1.2	Expression along the proximal-distal axis	10
2.2	Gene expression along the crypt-axis	11
2.2.1	Wnt signalling	12
	Canonical Wnt pathway	13
	Non-canonical Wnt	18
2.2.2	TGF- β Superfamily	18
	Mechanisms of TGF β superfamily signalling	19
2.2.3	Notch control of lineage differentiation	20
2.2.4	Hedgehog Pathway	21
2.3	Molecular biology of Adenoma Formation	22
2.3.1	Cell cycle balance and oncogenesis	22
2.3.2	The adenoma-carcinoma sequence	27
2.3.3	Disruptive Wnt signalling and neoplasia	27
2.3.4	Chromosomal instability pathway	30

2.3.5	The microsatellite instability pathway	31
2.3.6	The methylator phenotype	32
2.3.7	Serrated polyp pathway	33
2.3.8	Other Pathways	34
2.3.9	Acceleration of cancer progression by TGF- β and the Epithelial-Mesenchymal Transition	34
2.4	Colorectal neoplasia biomarker research	35
2.4.1	Microarray data for discovery	35
2.4.2	The need for validation	37
2.5	Conclusions	37
2.5.1	Hypothesis in the context of the literature	38
3	Discriminant Analysis	40
3.1	Background	40
3.1.1	Discrimination between two classes	43
3.2	Statistical decision theory	43
3.2.1	The base case: Disease incidence known, no training data	44
3.2.2	General case: Disease incidence known, data available . .	45
3.2.3	Cost and risk Functionals	47
3.3	Discriminant functions	48
3.3.1	Distance metrics for class separation	49
3.3.2	Linear discriminant analysis	52
3.3.3	Least squares (regression) solution	54
3.3.4	Quadratic discriminant analysis	57
3.3.5	Overfitting and the bias vs. variance trade-off	57
3.4	Conclusions	61

4	High dimensional analysis	63
4.1	Aims	63
4.2	Analysing data with more features than observations	63
4.3	Feature Selection and Subset Methods	66
4.3.1	Best subset regression	66
4.4	Feature Extraction	66
4.4.1	Principal Component Regression	67
4.5	Regularization and Penalization Methods	68
4.5.1	Ridge regression	68
4.5.2	The Lasso	71
4.6	Shortest Least Squares	72
4.7	Conclusions	73
5	Materials and Methods	74
5.1	Aims	74
5.2	Discovery data	75
5.2.1	Differential display discovery	75
5.2.2	GeneLogic data	76
5.3	Validation data:	78
5.3.1	Custom microarray	78
5.3.2	Microarray geometry and design considerations	78
5.3.3	Perfect match (PM) vs. mismatch (MM) probes	79
5.3.4	Labelled cRNA vs. cDNA	80
5.4	Laboratory methods	81
5.4.1	Human tissue samples	81
5.4.2	RNA extraction	82
	Method I	82
	Method II	82
5.4.3	Microarray processing	83

	HG U133 Plus 2.0 GeneChips	83
	CG_AGP custom microarray	84
5.4.4	RT-PCR	85
5.5	Statistical methods	86
5.5.1	Statistical software and data processing	86
5.5.2	Affymetrix GeneChip data reduction	86
5.5.3	Annotation of discovery data	87
	BLAST-based annotation of differential display sequences	87
	HG U133 (A/B/Plus2) annotation	88
	Custom microarray annotation	89
5.5.4	Hypothesis testing of differentially expressed biomarkers	89
5.5.5	Inter-segment modeling of the large intestine	90
5.5.6	Logistic regression modeling	91
5.5.7	Estimates of performance characteristics	91
5.5.8	Receiver operator characteristic curves and D-Value . . .	93
5.5.9	Tissue specific expression patterns	94
5.5.10	Gene set enrichment analysis	97
5.5.11	K-nearest neighbor clustering	97
5.5.12	Genetic algorithm for KNN	98
5.5.13	Principal components analysis	98
5.5.14	Supervised principal components analysis	100
5.6	Conclusions	101
6	Normal Gene Expression	102
6.1	Aim	102
6.2	Introduction	102
6.3	Results	105
6.3.1	Gene expression data	105
	Discovery data	105

Test data	106
6.3.2 Gene variation along the colon: univariate analyses . . .	106
6.3.3 Patterns of gene expression along the colon	110
PCA and supervised PCA	110
6.4 Discussion	113
6.4.1 A map of differential gene expression along the colon . .	113
6.4.2 Expression patterns of selected genes	116
6.4.3 The nature of gene expression change along the colon . .	119
6.5 Conclusions	121
7 Discovery of Neoplasia Markers	122
7.1 Aim	122
7.2 Differential display discovery	123
7.2.1 Nucleotide sequences to genes	123
7.2.2 Preliminary validation: RT-PCR experiments	123
7.2.3 Univariate analysis	124
7.2.4 Multivariate analysis	126
Logistic regression modeling	126
K-Nearest Neighbor analysis	127
Principal component analysis	129
7.2.5 A closer look at mis-classified specimens	130
7.3 Discovery using full genome microarrays.	130
7.3.1 Quality control	131
7.3.2 Principal components analysis	131
7.3.3 Genes differentially expressed in neoplastic tissues	132
7.3.4 Discovery of neoplasia-specific genes	135
7.3.5 Comparing expression between adenomatous and cancerous tissues	140
7.3.6 Multivariate models built from univariate candidates . .	140

7.4	Pathway analysis by gene set enrichment analysis	142
7.4.1	Wnt pathway analysis	144
7.4.2	Supervised PCA using pathway probesets	146
7.5	Literature based discovery	148
7.6	Intersection of discovery results	148
7.7	Conclusions	149
8	Validation	154
8.1	Aims	154
8.2	Custom chip design results	155
8.2.1	Composition of the custom microarray	155
8.3	Clinical specimens	156
8.4	Quality control analysis of the custom microarray data	158
8.5	Hypothesis testing of differential display candidates	161
8.5.1	Custom probes against sequence IDs	161
8.5.2	Commercial probes for presumed gene symbols	163
8.5.3	Multivariate analysis: logistic regression	163
8.6	Hypothesis testing of microarray-derived candidates	165
8.6.1	Testing proximal vs. distal expression patterns	165
8.6.2	Hypothesis testing of probesets for neoplasia discrimination	168
8.6.3	Neoplasia specific probesets	170
8.6.4	Probesets differentially expressed in adenoma versus cancer	172
8.7	Hypothesis testing of literature-based candidates	173
8.8	Candidate biomarkers in common	173
8.8.1	Validated genes discovered in this research	173
8.8.2	Biomarkers common to all discovery sources	175
8.9	Discussion and conclusions	177
8.9.1	Thesis aim achieved	177
	Comparison to the colorectal biomarker discovery literature	179

Neoplasia biomarker panel	182
8.9.2 Conclusion	188
9 Conclusions	189
9.1 Overview	189
9.2 Analysis of gene expression microarrays	190
Univariate vs. multivariate results	190
Identification of phenotype-specific RNA transcripts	192
The utility of gene set enrichment analysis	194
The utility of PCA to visualize high dimensional data	195
Critical impact of quality control	196
9.3 Gene expression along the normal colon	197
Value of understanding normal gene expression patterns	197
Influence of colorectal location on gene expression	198
How do genes change longitudinally?	199
Intrinsic vs. extrinsic expression patterns	199
9.4 Neoplastic gene expression in the colorectum	200
Design and validation of the custom microarray	200
Transcript expression trends	201
Neoplasia phenotype and gene expression	201
Wnt expression pattern	202
9.5 Biomarkers for colorectal neoplasia	202
9.5.1 A list of biomarker candidates	203
9.6 Future work	204
9.6.1 Biomarker assay development	204
9.6.2 Further research directions	206
Improved biological understanding	206
Improved phenotype-specific gene detection	207
9.7 In closing	208

A	Gene expression literature	209
A.0.1	Differential display literature	209
A.0.2	Microarray-based discovery	210
A.1	Conclusion	227
B	Quality control methods	229
B.1	Aim	229
B.2	Description of Gene Logic data	229
B.3	Quality control of Affymetrix Gene Chips	230
B.3.1	Scaling factors	231
B.3.2	Background values	232
B.3.3	Percent present	232
B.3.4	Spike-in probesets	233
B.3.5	Control probe degradation	235
B.4	RNA degradation analysis	236
B.4.1	28S:18S ratio	236
B.4.2	Within-probeset degradation	238
B.5	Principal component analysis	242
B.6	Conclusion	243
C	Machine learning algorithms	244
C.1	Support Vector Machines	244
C.1.1	Wolfe dual	246
C.1.2	Soft margin optimisation	249
C.1.3	Importance of regularisation	250
C.1.4	KKT conditions	251
C.1.5	The SVM solution	253
C.1.6	Nonlinear learning boundaries	253
C.1.7	Implementation	255
C.2	Conclusions	256

D	Extended Tables and Figures	257
D.1	Materials & methods	257
D.1.1	Covariates provided with GeneLogic data	257
D.1.2	KEGG gene pathways	258
D.1.3	Gene sets used for GSEA analysis	260
D.2	Normal tissue analysis	262
D.2.1	Genes elevated in proximal tissues	262
D.2.2	Genes elevated in distal tissues	263
D.2.3	RT-PCR validation of proximal-distal genes	264
D.3	Discovery - differential display	265
D.3.1	Annotation of differential display sequences	265
D.4	Discovery - GeneLogic microarray data	271
D.4.1	QC: Principal component plots	271
D.4.2	Probesets upregulated in neoplastic tissues	274
D.4.3	Probesets downregulated in neoplastic tissues	276
D.4.4	Probesets upregulated in adenomas vs. cancer tissues	282
D.4.5	Probesets upregulated in cancer vs. adenoma tissues	283
D.5	Hypothesis testing and validation	287
D.5.1	Validated differential display candidates	287
D.5.2	Adenoma specific biomarkers from differential display	290
D.5.3	Common genes validated by custom and commercial probesets	293
D.5.4	Validated microarray discovered genes	295
D.5.5	Validated biomarkers discriminating adenoma vs. cancer	295
D.5.6	Validated biomarkers elevated in cancers relative to adenomas	296
D.5.7	Validation of turned-off biomarkers	297
D.5.8	ROC curves for novel genes	298
D.5.9	List of validated genes	301

E Appendix: Publications and Patents Arising	305
E.1 Peer reviewed articles	305
E.2 Invited talks	305
E.3 Conference posters	306
E.4 Patents submitted	307

Gene Expression Biomarkers for Colorectal Neoplasia

L. C. LaPointe

Flinders University of South Australia

Department of Medicine

Prof. Graeme P. Young

The aim of this research was to assemble sufficient experimental evidence about candidate gene transcript expression changes between non-neoplastic and neoplastic colorectal tissues to justify future assay development involving promising leads. To achieve this aim, this thesis explores the hypothesis that gene expression-based biomarkers can be used to accurately discriminate colorectal neoplastic tissues from non-neoplastic controls.

This hypothesis was tested by first analysing multiple, large, quality controlled data sets comprising gene expression measurements across colorectal phenotypes to discover potential biomarkers. Candidate biomarkers were then subjected to validation testing using a custom-design oligonucleotide microarray applied to independently derived clinical specimens. A number of novel conclusions are reached based on these data. The most important conclusion is that a defined subset of genes expressed in the colorectal mucosa are reliably differentially expressed in neoplastic tissues. In particular, the apparently high prediction accuracy achieved for single gene transcripts to discriminate hundreds of neoplastic and non-neoplastic tissues provides compelling evidence that the resulting candidate genes are worthy of further biomarker research.

In addition to addressing the central hypothesis, additional contributions are made to the field of colorectal neoplasia gene expression profiling. These contributions include:

The first systematic analysis of gene expression in non-diseased tissues along the colorectum To better understand the range of gene expression in non-diseased tissues, RNA extracts taken from along the longitudinal axis of the large intestine were studied.

The development of quality control methodologies for high dimensional gene expression data Complex data collection platforms such as oligonucleotide microarrays introduce the potential for unrecognized confounding variables. The exploration of quality control parameters across five hundred microarray experiments provided insights about quality control techniques.

The design of a custom microarray comprised of oligonucleotide probe-sets hybridising to RNA transcripts differentially expressed in neoplastic colorectal specimens A custom design oligonucleotide microarray was designed and tested combining the results of multiple biomarker discovery projects.

Introduction of a method to filter differentially expressed genes during discovery that may improve validation efficiencies of biomarker discovery based on gene expression measurements Differential expression discovery research is typically focused only on quantitative changes in transcript concentration between phenotype contrasts. This work introduces a method for generating hypotheses related to transcripts which may be qualitatively “switched-on” between phenotypes.

Identification of mRNA transcripts which are differentially expressed between colorectal adenomas and colorectal cancer tissues Transcripts differentially expressed between adenomatous and cancerous RNA extracts were discovered and then tested in independent tissues.

In conclusion, these results confirm the hypothesis that gene expression profiling can discriminate colorectal neoplasia (including adenomas) from non-neoplastic controls. These results also establish a foundation for an ongoing biomarker development program.

Declaration

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief does not contain any material previously or written by another person except where due reference is made in the text.

.....

Lawrence Charles LaPointe

Acknowledgements

Firstly, I would like to thank my supervisors Prof. Graeme Young of Flinders University, Dr. Rob Dunne of CSIRO Mathematical and Information Sciences and Dr. Peter Molloy of CSIRO Molecular Health Technologies. I am indebted to Peter Molloy for reminding me that good science requires precision and careful consideration and that patience is often rewarded. I am grateful to Rob Dunne for teaching me skills that I will use for the rest of my career and for his excellent instruction of complex subject matter. I express my greatest thanks to Graeme Young, without whose guidance I would not have been able to start, conduct, or complete this research.

Collectively, my supervisors' guidance, scientific instruction, and ability to provide insightful criticism made this work possible.

I would like also to thank Clinical Genomics Pty Ltd and Enterix Inc for support of this research, including providing me ample time to dedicate to this study. In particular, I thank Howard Chandler, Max Mawhinney, and Peter Horrobin who have shared my vision that good science makes good business. With their support, I have been able to invest considerable time and energy into this research.

I thank my wife and family for love and support. I especially thank Karen for enduring my absence, inattention, and stress through these years without a single word of objection. Thank you for helping me to make this investment.

Finally, I express my deepest gratitude to the nameless patients and volunteers whose generous gift of clinical specimens forms the cornerstone of this research. To these individuals: your decision to contribute to the benefit of others even while you are confronted by the tragedy of colorectal cancer is inspirational. This thesis is aimed at discovering biomarkers which I hope will help others avoid your pain and I dedicate this work to you.

