ADVANCING THE SERIAL ANALYSIS OF GENE EXPRESSION TECHNIQUE AND ITS APPLICATION TO THE STUDY OF THE DEVELOPMENT OF SQUAMOUS CELL LUNG CANCER

by

Scott Dorjan Zuyderduyn

B.Sc.(Hon.), The University of British Columbia, 1999

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Biochemistry and Molecular Biology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2009

© Scott Dorjan Zuyderduyn, 2009
ABSTRACT

Lung cancer is one of the most common and deadliest forms of cancer. Squamous cell lung carcinomas (SCC), a common lung cancer subtype, feature a series of identifiable pre-malignant and early malignant forms that progress sequentially into full-blown tumours. This thesis describes a sophisticated and statistically rigorous analysis of global gene expression profiles taken from samples of several key stages of progression. This dataset was generated using serial analysis of gene expression (SAGE), a powerful transcriptome profiling technique that captures small sequence tags from each transcript in an mRNA population. These tags can then be counted and mapped back to a matching transcript sequence to quantitatively determine the expression of a given gene. The analysis identified several genes which show changes in expression that are highly correlated with the progressive steps of SCC. In addition, gene expression changes were identified in samples of bronchial epithelium that correspond to an acute response to tobacco smoke exposure, a major contributor to SCC development.

The use of multiple sample types, the presence of extensive cellular heterogeneity, and the rarity of biological material for the purpose of validation introduced an additional layer of complexity that are not well-suited to conventional methods of SAGE analysis. To address these challenges, this thesis describes the development of two methodological improvements to SAGE data analysis. The first describes a computational strategy to identify additional sequence information that effectively increases the length of SAGE tag sequences, greatly enhancing the fidelity of tag to gene mapping. The second describes a new statistical method that shows improved performance in modelling SAGE data. The Poisson mixture model used in this work provides better estimates of statistical significance, is highly effective when using multiple sample types, and is a flexible framework for more complex meta-analyses.
## TABLE OF CONTENTS

ABSTRACT ................................................................................................................... ii  
TABLE OF CONTENTS ................................................................................................... iii  
LIST OF TABLES ........................................................................................................ vi  
LIST OF FIGURES ....................................................................................................... vii  
LIST OF ABBREVIATIONS ........................................................................................... x  
ACKNOWLEDGEMENTS .............................................................................................. xiv  
CO-AUTHORSHIP STATEMENT ................................................................................ xiv  

## CHAPTER I  INTRODUCTION ............................................................................. 1  
1.1 THE CHALLENGE OF CANCER ................................................................. 1  
1.2 TYPES OF GENETIC CHANGE IN CANCER .............................................. 5  
1.3 LUNG CANCER .............................................................................................. 9  
  1.3.1 Histological subtypes ........................................................................ 10  
    1.3.1.1 Small cell lung cancer (SCLC) ........................................ 10  
    1.3.1.2 Adenocarcinoma (AC) .................................................. 13  
    1.3.1.3 Atypical subtypes .......................................................... 13  
    1.3.1.4 Squamous cell lung cancer (SCC) ............................. 14  
      1.3.1.4.1 Stages of progression ................................ 14  
      1.3.1.4.2 Common genetic alterations ........................ 17  
1.4 GENOMICS AND GENE EXPRESSION PROFILING ............................. 19  
1.5 SERIAL ANALYSIS OF GENE EXPRESSION (SAGE) ............................ 23  
  1.5.1 Constructing a library ................................................................. 23  
1.6 DATA ANALYSIS AND STATISTICS ......................................................... 27  
  1.6.1 Statistical error and bias .............................................................. 28  
  1.6.2 Choice of test statistic .............................................................. 29  
  1.6.3 Multiple testing ........................................................................... 30  
  1.6.4 Let the data challenge assumptions ........................................ 31  
  1.6.5 Information visualization, data clustering, and  
      global comparisons ..................................................................... 32  
  1.6.6 Cross-validation ........................................................................... 36  
1.7 THESIS OBJECTIVES ............................................................................. 37  
BIBLIOGRAPHY ............................................................................................... 39  

## CHAPTER II  DETERMINING ADDITIONAL SEQUENCE DATA TO IMPROVE  
SAGE TAG TO GENE MAPPING .............................................................................. 46  
2.1 INTRODUCTION .......................................................................................... 46  
2.2 MATERIALS AND METHODS .................................................................... 48  
  2.2.1 SAGE data ............................................................................. 48
CHAPTER III  STATISTICAL INFERENC E FROM SAGE USING A POISSON MIXTURE MODEL ............................................................................ 78

3.1 INTRODUCTION .................................................................................. 78
3.2 MATERIALS AND METHODS ................................................................. 82
  3.2.1 Test datasets ................................................................................ 82
  3.2.2 Model fitting ................................................................................. 82
    3.2.2.1 Log-linear (Poisson) regression model ................................. 85
    3.2.2.2 Overdispersed log-linear regression model ......................... 85
    3.2.2.3 Poisson mixture model ......................................................... 86
3.3 RESULTS ............................................................................................. 87
  3.3.1 Goodness of fit ............................................................................. 87
  3.3.2 Tags with ambiguous mappings are represented by a greater number of components ......................................................... 89
  3.3.3 Component assignment of libraries is non-random ...................... 91
  3.3.4 Determining differentially expressed genes ................................. 91
3.4 DISCUSSION ....................................................................................... 99
BIBLIOGRAPHY ..................................................................................... 101

CHAPTER IV  TRANSCRIPTOME EVOLUTION IN THE DEVELOPMENTAL STAGES OF SQUAMOUS CELL LUNG CARCINOMA .......... 103

4.1 INTRODUCTION .................................................................................. 103
4.2 MATERIALS AND METHODS ................................................................. 106
  4.2.1 Sample collection and preparation ............................................. 106
  4.2.2 Data processing .......................................................................... 106
4.2.3 Multidimensional scaling ............................................................ 106
4.2.4 k-means clustering ..................................................................... 107
4.2.5 Hierarchical clustering .............................................................. 108
4.2.6 Gene Ontology and KEGG pathway enrichment ....................... 108
4.2.7 Statistical analysis ...................................................................... 109
  4.2.7.1 Preprocessing .......................................................................... 109
  4.2.7.2 Feature selection ...................................................................... 109
  4.2.7.3 Estimating the selectivity of a candidate tag list and
the generation of optimal signatures ........................................... 112
4.2.8 Tag to gene mapping ................................................................. 112
4.2.9 Microarray validation ................................................................. 113
4.2.10 Tissue samples .......................................................................... 115
4.2.11 MMP11 antibody ........................................................................ 115
4.2.12 Immunohistochemistry ............................................................... 115

4.3 RESULTS ............................................................................................... 117
  4.3.1 A global view of the transcriptome during the
development of SCC .................................................................. 117
    4.3.1.1 Multidimensional scaling analysis ..................................... 117
    4.3.1.2 Common gene expression patterns identified by
k-means clustering ...................................................................... 124
  4.3.2 Transcriptional signatures of developmental stages .............. 128
    4.3.2.1 Changes associated with tobacco smoke exposure ... 128
      4.3.2.1.1 Acute response to tobacco smoke exposure ............. 130
      4.3.2.1.2 Persistent response to tobacco smoke exposure ...... 135
    4.3.2.2 Changes associated with SCC development ........... 138
      4.3.2.2.1 Changes associated with pre-malignant
transformation ................................................................. 141
      4.3.2.2.2 Changes associated with malignancy ........ 148
      4.3.2.2.3 Changes associated with an invasive
phenotype .......................................................................... 159
  4.3.3 Comparison to existing squamous cell lung cancer profiles ..... 167

4.4 DISCUSSION ......................................................................................... 170
BIBLIOGRAPHY ............................................................................................. 173

CHAPTER V CONCLUSION AND FUTURE PROSPECTS ................................. 179

6.1 OVERALL DISCUSSION AND CONCLUSION ....................................... 179
6.2 FUTURE PROSPECTS .......................................................................... 184
BIBLIOGRAPHY ............................................................................................. 187

APPENDIX I LIKELIHOOD OF OBSERVING A DITAG GIVEN BOTH
CONTRIBUTING TAG SEQUENCES ................................................................ 186
APPENDIX II MODEL FITTING R SOURCE CODE ...................................... 188
APPENDIX III DERIVATION OF POISSON MIXTURE MODEL
DIFFERENTIAL EXPRESSION CONFIDENCE SCORE .................................. 192
**LIST OF TABLES**

**CHAPTER I**

| Table 1.1 | Estimated new cases and deaths for the top 5 cancer types | 1 |
| Table 1.2 | The histological subtypes of lung cancer | 11 |
| Table 1.3 | Comparison of SAGE and DNA microarray technologies | 24 |

**CHAPTER II**

| Table 2.1 | Estimated solutions for the distribution of tag lengths for 71 publicly available SAGE libraries | 54 |
| Table 2.2 | Improvement of tag to gene mapping with addition of extra nucleotides | 70 |
| Table 2.3 | Putative source of 14bp tags with ambiguous mappings that no longer map when lengthened to 16bp | 71 |
| Table 2.4 | Putative source of 14bp tags with unambiguous mappings that no longer map when lengthened to 16bp | 72 |
| Table 2.5 | Putative source of 14bp tags that fail to map | 74 |

**CHAPTER III**

| Table 3.1 | Datasets used to evaluate models | 83 |
| Table 3.2 | Comparison of model fits to a single group of biological replicates | 88 |
| Table 3.3 | Mean number of mixture model components | 90 |
| Table 3.4 | Top component memberships | 92 |

**CHAPTER IV**

| Table 4.1 | Summary of SAGE libraries | 118 |
| Table 4.2 | Source patient information for bronchial epithelium brushings samples | 121 |
| Table 4.3 | Description of clusters identified by k-means analysis | 127 |
| Table 4.4 | Microarray validation of smoke-exposure signatures | 134 |
LIST OF FIGURES

CHAPTER I

Figure 1.1 The hallmarks of cancer .............................................................. 4
Figure 1.2 Histology of common subtypes of lung cancer ...................... 12
Figure 1.3 Histological progression of squamous cell lung cancer .......... 15
Figure 1.4 The role of gene expression in the cell ...................................... 21
Figure 1.5 The serial analysis of gene expression library construction protocol ................................................................. 25
Figure 1.6 Examples of common data visualization methods .................. 33

CHAPTER II

Figure 2.1 Summary of genetic algorithm to find solutions for proportion of tag lengths (P(l_t)) ................................................................. 51
Figure 2.2 The beta distribution ................................................................. 53
Figure 2.3 Comparison of the observed and expected count of simulated tags with two additional nucleotides ................................. 68

CHAPTER III

Figure 3.1 Probability density of several models applied to data generated from two Poisson components ............................................... 80
Figure 3.2 Comparison to significance scores for a test of differential expression calculated using a negative binomial model ............ 94
Figure 3.3 Counts for two tags assessed using a negative binomial model and the Poisson mixture model where one model shows significance and the other does not ........................................ 95
Figure 3.4 Comparison to Bayes error rate for a test of differential expression calculated using a beta binomial model .................... 97
Figure 3.5 Counts for two tags assessed using a Bayes error rate and the Poisson mixture model where one model shows significance and the other does not ........................................ 98

CHAPTER IV

Figure 4.1 A confidence score calculation for a tag .............................. 111
Figure 4.2 Multidimensional scaling (MDS) analysis of 40 SAGE libraries from samples reflecting different stages of SCC development ......................................................... 119
Figure 4.3 Multidimensional scaling (MDS) analysis of 24 SAGE libraries from brushings of bronchial epithelium with different levels of tobacco smoke exposure ........................................ 122
Figure 4.4 Multidimensional scaling (MDS) analysis of 16 SAGE libraries from bulk samples reflecting the different stages of SCC development ......................................................... 123
Figure 4.5  Determination of cluster number for $k$-means analysis and the four major clusters ................................................................. 125
Figure 4.6  The five minor $k$-means clusters ................................................................. 129
Figure 4.7  Candidate smoke exposure tag selection plots ................................................................. 131
Figure 4.8  Heatmap with sample-wise hierarchical clustering of 70 tags upregulated in bronchial epithelium from current smokers ...... 132
Figure 4.9  Heatmap with sample-wise hierarchical clustering of 9 tags downregulated in bronchial epithelium from current smokers .. 133
Figure 4.10  Heatmaps with sample-wise hierarchical clustering of 4 tags upregulated and 1 tag downregulated in former, but not current, smokers ................................................................. 137
Figure 4.11  Candidate squamous cell lung cancer progression tag selection plots ................................................................. 139
Figure 4.12  Venn diagram of the number of candidate tags identified for different combinations of SCC progression sample types ...... 140
Figure 4.13.1  Heatmap with sample-wise hierarchical clustering of the first 69 of 138 tags upregulated in metaplasia and later stages of SCC development ................................................................. 142
Figure 4.13.2  Heatmap with sample-wise hierarchical clustering of the final 69 of 138 tags upregulated in metaplasia and later stages of SCC development ................................................................. 143
Figure 4.14.1  Heatmap with sample-wise hierarchical clustering of the first 79 of 316 tags downregulated in metaplasia and later stages of SCC development ................................................................. 144
Figure 4.14.2  Heatmap with sample-wise hierarchical clustering of the second 79 of 316 tags downregulated in metaplasia and later stages of SCC development ................................................................. 145
Figure 4.14.3  Heatmap with sample-wise hierarchical clustering of the third 79 of 316 tags downregulated in metaplasia and later stages of SCC development ................................................................. 146
Figure 4.14.4  Heatmap with sample-wise hierarchical clustering of the final 79 of 316 tags downregulated in metaplasia and later stages of SCC development ................................................................. 147
Figure 4.15  Microarray validation of optimal metaplasia progression-associated upregulated gene signature ................................................................. 149
Figure 4.16.1  Heatmap with sample-wise hierarchical clustering of the first 72 of 143 tags upregulated in malignant stages of SCC development ................................................................. 150
Figure 4.16.2  Heatmap with sample-wise hierarchical clustering of the final 71 of 143 tags upregulated in malignant stages of SCC development ................................................................. 151
Figure 4.17  Heatmap with sample-wise hierarchical clustering of the 26 tags downregulated in the malignant stages of SCC development ................................................................. 152
Figure 4.18.1  Microarray validation of optimal malignant progression-associated upregulated gene signature ................................................................. 154
Figure 4.18.2  Microarray validation of optimal malignant progression-associated downregulated gene signature ................................................................. 155
Figure 4.19  A model of MCM7 and CKS1B function based on current knowledge ................................................................. 156
Figure 4.20  Expression of select members of the MCM family and related genes ................................................................. 158
Figure 4.21  Heatmap with sample-wise hierarchical clustering of 29 tags upregulated in the invasive stage of SCC development .......... 160
Figure 4.22  Microarray validation of optimal invasive progression-associated upregulated gene signature .................................. 162
Figure 4.23  Western blot confirming correct activity of MMP11 antibody ............................................................... 163
Figure 4.24  MMP11 detection by immunohistochemistry ................................................................. 164
Figure 4.25  Role of serine hydroxymethyltransferase in thymidine biosynthesis ............................................................... 166
Figure 4.26.1  Heatmap of SAGE tag expression of candidate SCC genes identified from the Nacht and Bhattacharjee datasets ............. 168
Figure 4.26.2  Heatmap of SAGE tag expression of candidate SCC genes identified from the Erez/Dehan and Wachi datasets ............ 169
Throughout this thesis, individual genes are referred to according to the symbols and guidelines specified by the HUGO Gene Nomenclature Committee (www.genenames.org) (White, 1997). Specifically: a) the standard HUGO gene symbol is used in all cases, although alias(es) will be mentioned with the first appearance of the symbol if they are still in common use (e.g. CDKN2A (a.k.a. p16INK4a)); b) when referring to a gene locus or mRNA transcript arising from a locus, the gene symbol is italicized (e.g. GAPDH); c) when referring to a protein product, the gene symbol is not italicized (e.g. MDM1). The full name of the gene is not typically mentioned, but will always be present in this list of abbreviations.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABL</td>
<td>c-abl oncogene 1, receptor tyrosine kinase</td>
</tr>
<tr>
<td>AC</td>
<td>adenocarcinoma</td>
</tr>
<tr>
<td>ADH7</td>
<td>alcohol dehydrogenase 7 (class IV), mu or sigma polypeptide</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike information criterion</td>
</tr>
<tr>
<td>AE</td>
<td>anchoring enzyme</td>
</tr>
<tr>
<td>AKR1B10</td>
<td>aldo-keto reductase family 1, member B10 (aldo reductase)</td>
</tr>
<tr>
<td>AKR1C2</td>
<td>aldo-keto reductase family 1, member C2 (dihydrodiol dehydrogenase 2; bile acid binding protein; 3-alpha hydroxysteroid dehydrogenase, type III)</td>
</tr>
<tr>
<td>AKR1C3</td>
<td>aldo-keto reductase family 1, member C3 (3-alpha hydroxysteroid dehydrogenase, type II)</td>
</tr>
<tr>
<td>ALDH3A1</td>
<td>aldehyde dehydrogenase 3 family, member A1</td>
</tr>
<tr>
<td>AML</td>
<td>acute myeloid leukemia</td>
</tr>
<tr>
<td>ATP1A1</td>
<td>ATPase, Na+/K+ transporting, alpha 1 polypeptide</td>
</tr>
<tr>
<td>ATP1A2</td>
<td>ATPase, Na+/K+ transporting, alpha 2 (+) polypeptide</td>
</tr>
<tr>
<td>ATP1A3</td>
<td>ATPase, Na+/K+ transporting, alpha 3 polypeptide</td>
</tr>
<tr>
<td>ATP1B3</td>
<td>ATPase, Na+/K+ transporting, beta 3 polypeptide</td>
</tr>
<tr>
<td>BCR</td>
<td>breakpoint cluster region</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian information criterion</td>
</tr>
<tr>
<td>BIRC5</td>
<td>baculoviral IAP repeat-containing 5</td>
</tr>
<tr>
<td>BRCA1</td>
<td>breast cancer 1, early onset</td>
</tr>
<tr>
<td>BRCA2</td>
<td>breast cancer 2, early onset</td>
</tr>
<tr>
<td>C16orf89</td>
<td>chromosome 16 open reading frame 89</td>
</tr>
<tr>
<td>C3</td>
<td>complement component 3</td>
</tr>
<tr>
<td>C5orf32</td>
<td>chromosome 5 open reading frame 32</td>
</tr>
<tr>
<td>CAV1</td>
<td>caveolin 1, caveolae protein, 22kDa</td>
</tr>
<tr>
<td>CBFB</td>
<td>core-binding factor, beta subunit</td>
</tr>
<tr>
<td>CBR1</td>
<td>carbonyl reductase 1</td>
</tr>
<tr>
<td>CCND1</td>
<td>cyclin D1</td>
</tr>
<tr>
<td>CCNE1</td>
<td>cyclin E1</td>
</tr>
<tr>
<td>CDC45L</td>
<td>CDC45 cell division cycle 45-like (S. cerevisiae)</td>
</tr>
<tr>
<td>CDC6</td>
<td>cell division cycle 6 homolog (S. cerevisiae)</td>
</tr>
<tr>
<td>CDC7</td>
<td>cell division cycle 7 homolog (S. cerevisiae)</td>
</tr>
<tr>
<td>CDH1</td>
<td>cadherin 1, type 1, E-cadherin (epithelial)</td>
</tr>
<tr>
<td>CDK2</td>
<td>cyclin-dependent kinase 2</td>
</tr>
<tr>
<td>CDK4</td>
<td>cyclin-dependent kinase 4</td>
</tr>
<tr>
<td>CDKN1A</td>
<td>cyclin-dependent kinase inhibitor 1A (p21, Cip1)</td>
</tr>
<tr>
<td>CDKN1B</td>
<td>cyclin-dependent kinase inhibitor 1B (p27, Kip1)</td>
</tr>
</tbody>
</table>
CDKN2A  cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
cDNA  complementary deoxyribonucleic acid
CDT1  chromatin licensing and DNA replication factor 1
CGH  comparative genomic hybridization
CI  confidence interval
CIS  carcinoma in situ
CKS1B  CDC28 protein kinase regulatory subunit 1B
CML  chronic myelogenous leukemia
CNV  copy number variation
COL1A2  collagen, type I, alpha 2
COL3A1  collagen, type III, alpha 1
COL6A3  collagen, type VI, alpha 3
CRIP2  cysteine rich protein 2
CSTA  cystatin A (stefin A)
CV  cross-validation
CYP1A1  cytochrome P450, family 1, subfamily A, polypeptide 1
CYP1B1  cytochrome P450, family 1, subfamily B, polypeptide 1
DAB  3,3'-diaminobenzidine
DBF4  DBF4 homolog (S. cerevisiae)
DHFR  dihydrofolate reductase
ECM  extracellular matrix
EGFR  epidermal growth factor receptor (a.k.a. ERBB1)
EPHX1  epoxide hydrolase 1, microsomal (xenobiotic)
ERBB2  v-erb-b2 erythroblastic leukemia viral oncogene homolog 2 (a.k.a. HER2/neu)
ERG  v-ets erythroblastosis virus E26 oncogene homolog (avian)
EST  expressed sequence tag
ETV1  ets variant 1
ETV4  ets variant 4
FDR  false discovery rate
GAPDH  glyceraldehyde-3-phosphate dehydrogenase
GEO  Gene Expression Omnibus
GMNN  geminin, DNA replication inhibitor
GO  Gene Ontology
GPX2  glutathione peroxidase 2 (gastrointestinal)
GPX4  glutathione peroxidase 4 (phospholipid hydroperoxidase)
GSR  glutathione reductase
GSTA2  glutathione S-transferase alpha 2
GSTM1  glutathione S-transferase pi 1
HIF1A  hypoxia-inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)
IRLS  iteratively reweighted least-squares
JUP  junction plakoglobin (a.k.a. γ-catenin)
KLK3  kallikrein-related peptidase 3 (a.k.a. PSA)
KRAS  v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
KRT6A  keratin 6A
KRT6B  keratin 6B
KRT14  keratin 14 (epidermolysis bullosa simplex, Dowling-Meara, Koebner)
LAMA1  laminin, alpha 1
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRF</td>
<td>serum response factor (c-fos serum response element-binding transcription factor) (a.k.a. MCM1)</td>
</tr>
<tr>
<td>SVM</td>
<td>support vector machine</td>
</tr>
<tr>
<td>TALDO1</td>
<td>transaldolase 1</td>
</tr>
<tr>
<td>TE</td>
<td>tagging enzyme</td>
</tr>
<tr>
<td>TFF3</td>
<td>trefoil factor 3 (intestinal)</td>
</tr>
<tr>
<td>TMPRSS2</td>
<td>transmembrane protease, serine 2</td>
</tr>
<tr>
<td>TP63</td>
<td>tumor protein p63</td>
</tr>
<tr>
<td>TPM</td>
<td>tags per million</td>
</tr>
<tr>
<td>TSG</td>
<td>tumour suppressor gene</td>
</tr>
<tr>
<td>TYMS</td>
<td>thymidylate synthetase</td>
</tr>
</tbody>
</table>
I would like to thank my research supervisor Dr. Victor Ling for his unwavering support, enthusiasm, and commitment to good science. His ability to ask penetrating questions about subjects he should know nothing about was always a source of amazement. He is truly a scholar who any aspiring scientist could look to as a role model.

I appreciate the support and companionship of past and present Ling Lab members, the most diverse mix of individuals I have ever had the pleasure to work with. I tip my hat to my student predecessors Dr. Dennis Leveson-Gower, Dr. Maria Ho, and Dr. Maisie Lo who each showed the way through graduate school is never a straight line but, eventually, it does end. Long-distance running hurts, but it feels so good when you stop. I thank Dr. Tania Kastelic for her wisdom, experience, and superhuman attention to detail. I thank Dr. Jonathan Sheps for his knowledge, candour, and for his enthusiasm in engaging in the prolonged discussions that computer geeks like me are compelled to start not only about science and technology, but wild tangents about the state of the world and the issues of the day. I thank Dr. Renxue Wang for sharing his knowledge and his willingness to help with any problem at a moment’s notice. I thank Dr. Jaclyn Hung for her insights and expertise on the problem of lung cancer, and her willingness to share her material and equipment. I would like to acknowledge the often thankless and unnoticed tasks performed by Barb Schmidt, Lin Liu, and Sue Smith that keep the lab together, running smoothly, and able to carry out good science.

I am particularly indebted to the support, candour, ideas and continued friendship of Dr. Greg Vatcher.

I appreciate the expertise and clinical material provided by lung oncologist Dr. Stephen Lam. I also acknowledge the technical support provided by fellow students Alvin Ng (Hung Lab) and Gerald Li (MacAulay Lab).

I gratefully acknowledge the generous funding support provided by the Canadian Institutes for Health Research (CIHR) and the Michael Smith Foundation for Health Research (MSFHR).

I am grateful for the sponsorship of Dr. Raymond Ng in obtaining access to the Westgrid computing cluster, as well the efforts of the Westgrid support staff.

Finally, I acknowledge the guidance and assistance of my thesis committee: Dr. Wan Lam, Dr. Calum MacAulay, Dr. Ross McGillivray, and Dr. Raymond Ng.

“Science... never solves a problem without creating ten more.”

- George Bernard Shaw
Several chapters describe work done in collaboration with others. Chapter II was done in collaboration with Dr. Greg Vatcher. I developed and performed all data analyses as presented, and wrote the manuscript. I am the sole author of the work in Chapter III. Chapter IV was done in collaboration with Drs. Greg Vatcher, Stephen Lam, Wan Lam, Calum MacAulay, Raymond Ng, and Victor Ling. I developed and performed all data analyses as presented, and wrote the manuscript.
CHAPTER I

INTRODUCTION

Cancer is one of the most feared diseases and has afflicted humans for millennia. The first account of cancer was written on papyrus in 1500 BCE, and described a series of breast tumours that were treated with crude cauterization (Diamondopoulos, 1996). In the intervening three and a half thousand years, great progress has been made in understanding and treating the disease. However, the exceedingly complex nature of cancer and its ability to evolve and adapt, turning the intricate systems of the body against itself, continues to challenge medical science.

This introduction will provide an overview of the characteristics that make cancer so complex, before focusing on lung cancer, a multifaceted disease in its own right. Particular attention will be given to squamous cell carcinoma, a common lung cancer subtype. This will be followed by a discussion of genomics, and gene expression profiling in particular, as a means to unravel cancer’s complexity. Serial analysis of gene expression (SAGE), a particularly powerful technique for obtaining a global portrait of gene expression and the methodology that informs the work in this thesis, will be described. Finally, the methods and challenges of analyzing large-scale datasets, like those obtained from gene expression profiling, are explored.

1.1 THE CHALLENGE OF CANCER

Cancer is currently the second leading cause of death in Canada (behind cardiovascular disease), and is similarly placed throughout the Western world (Canadian Cancer Society, 2008; Jemal, 2007; Pisani, 2002). In 2007, the disease affected about 160,000 Canadians and contributed to over 72,000 deaths (Table 1.1) (Canadian Cancer Society, 2008). An additional 69,000 are affected by highly treatable forms of non-melanoma skin cancer (Canadian Cancer Society, 2008). Although a greater awareness of the health risks associated with cancer and
Table 1.1: Estimated new cases and deaths for the top 5 cancer types†

<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>Males</th>
<th></th>
<th>Females</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Deaths</td>
<td>Cases</td>
<td>Deaths</td>
</tr>
<tr>
<td>lung</td>
<td>12,400</td>
<td>11,000</td>
<td>10,900</td>
<td>8,900</td>
</tr>
<tr>
<td>breast</td>
<td>170</td>
<td>50</td>
<td>22,300</td>
<td>5,300</td>
</tr>
<tr>
<td>prostate</td>
<td>22,300</td>
<td>4,300</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>colorectal</td>
<td>11,400</td>
<td>9,000</td>
<td>4,700</td>
<td>4,000</td>
</tr>
<tr>
<td>lymphoma</td>
<td>3,700</td>
<td>1,700</td>
<td>3,100</td>
<td>1,400</td>
</tr>
</tbody>
</table>

Statistics from the National Cancer Institute of Canada (NCIC), 2007. † 69,000 cases of non-melanoma skin cancers are excluded.
continuing advances in diagnosis and treatment have led to a decline in cancer over the last fifteen years, a concordant decline in deaths from heart disease has led to cancer becoming the number one killer except in those over 85, and is expected to include even this group by 2018 (Jemal, 2005). For this reason, substantial progress must be made in cancer research and related fields in order to continue improving the long-term health prospects of this and future generations.

The pathology of cancer always involves the formation of malignant cells that destroy tissue or interfere with health-maintaining systems of the body. In most cases, this is associated with tumour formation, although haematological neoplasms (i.e. leukemia, lymphoma) are exceptions. Current scientific consensus postulates that cancer is, at its heart, a disease of the genes. The initiation and progression of carcinogenesis are driven by successive genetic changes that interfere with the integrity, function, or regulation of particular genes. Unfortunately, the identity of these loci and the mechanisms that disrupt them are only partially known, and can vary widely from tumour to tumour, even among cancers that arise from the same tissue and display identical histological characteristics. This is in stark contrast to major killers like heart attack, stroke or infection where the aetiology is largely uniform and new treatments can be widely applied.

Despite the uncertainty surrounding the genetic basis of most cancers, a number of so-called “hallmarks” have been suggested that are considered requirements for cancer to thrive and kill successfully: a) malfunctions in the molecular signals that promote or inhibit cellular growth, b) independence from programmed cell death, c) unlimited proliferative capacity (or immortalization), d) sustained neovascularisation, and e) acquiring the ability to escape the barriers of organs and tissues that gave rise to the cancer and further spread to others (Figure 1.1) (Hanahan, 2000). As a general principle, a causal genetic change will affect one or more of these processes. However, the order in which these requirements are met is not consistent, and the
**Figure 1.1: The hallmarks of cancer.** Depicted are the capabilities a cell must acquire in order to become malignant. The figure here is remastered from the original published in Hanahan and Weinberg (2000).
genetic changes that directly cause or attenuate them are highly variable.

Of greatest concern from a clinical perspective is how these hallmarks conspire to make a tumour progress, to resist treatment outright, or to provide the foundations for further mutations that allow the disease to evolve and recur in a more resistant form even after successful treatment. Although it may take many years of research to decipher the full complexity of cancer, a partial understanding of some of the molecular players has positively impacted patient outcomes. Familial mutations that increase the risk of disease have helped to identify segments of the population who will benefit from aggressive surveillance or preventative medicine. For example, certain inherited variations in the \textit{BRCA1} and \textit{BRCA2} genes are an indicator of increased risk for breast cancer (reviewed in Palacios, 2008). Early detection screens have been developed based on the identification of molecular markers that manifest in premalignant or early stage neoplasms. For example, the serum level of KLK3 (a.k.a. PSA) is a commonly used, although controversial, test for the presence of prostate cancer (reviewed in Lilja, 2008). More recently, tailored treatments have been developed based on markers associated with drug response. For example, the monoclonal antibody trastuzumab (Herceptin) is an effective therapy for breast cancers that overexpress ERBB2 (reviewed in Nanda, 2007).

The reality is that there is neither a “magic bullet” to attack cancer, nor a simple molecular mechanism to explain its occurrence or clinical course. The large number of cancer types, each associated with highly variable pathology and response to treatment, are a result of a complex genetic basis; it is this complexity that makes cancer such a difficult challenge.

\section*{1.2 \textbf{TYPES OF GENETIC CHANGE IN CANCER}}

Evolution is the means by which life adapts to a changing environment or enables survival in a new environment; this process is carried out through heritable modifications to the genes and random changes to the overall genetic makeup of populations. However, not all
changes are beneficial and cancer development is a downside of life’s genetic plasticity. Much of the complexity of cancer is due to the fact that all known natural, and normally beneficial, mechanisms of genetic change can be exploited to promote tumourigenesis.

DNA changes most often occur on the scale of a single or very small group of nucleotides. This can result from errors made during replication or by exposure to radiation or certain chemicals. Point mutations refer to the change of a single nucleotide from one base to another. When these occur in the coding region of a gene it is possible to introduce either: a) missense mutations that change the amino acid sequence of the encoded protein, altering or abolishing its function or b) nonsense mutations that introduce a premature stop codon, truncating the resulting protein. A classic example occurs in colon cancer, where point mutations in specific exons of the \textit{KRAS} gene occur in about 50\% of cases (Forrester, 1987; Bos, 1987). Mutations can also occur in non-coding regions such as regulatory elements, altering gene expression. A “C” to “G” transversion in the promoter region of \textit{BIRC5}, a known antiapoptotic factor, is found in a variety of cancer cell lines and correlates with the increased expression of this gene (Xu, 2004). Small deletions or insertions (indels) in a coding region can change the structure of a protein by altering its amino acid sequence. Despite the small number of nucleotides affected, an indel can introduce a frameshift that will change the entire amino acid sequence downstream of the mutation site. As with point mutations, indels can occur in regulatory elements as well. These alterations are variable in size; most often a single base is involved, but they usually do not exceed 15-20bp (Stenson, 2008). Such events occur in germ-line mutations of \textit{CDH1} and are strongly associated with a type of hereditary gastric cancer (Humar, 2002; Brooks-Wilson, 2004).

Other genetic changes occur on a larger scale. These copy number variations (CNVs) involve segments of a chromosome that can vary in size from a few hundred nucleotides to several megabases in size. These segments of DNA can be duplicated, amplifying the expression
of the genes contained within. For example, amplification of the *MYCN* oncogene is frequently observed in neuroblastomas, and is strongly correlated with the stage and aggressiveness of the disease (Seeger, 1985). Correspondingly, segments of DNA can be lost, reducing the function of the affected genes. Pediatric meningioma, a rare childhood cancer, is usually accompanied by a deletion of the tumour suppressor *NF2* (Begnami, 2007). Gains and losses of whole chromosomes (and even aneuploidy and polyploidy, where the entire set of chromosomes is affected) is often observed in cancer cells, although it remains unclear to what extent this is a contributing cause of tumourigenesis or a symptom of genome instability (Storchova, 2004).

Perhaps the most fascinating alterations are complex chromosomal rearrangements. When a section of one chromosome exchanges places with a section of a different chromosome, a translocation has occurred and can result in the disruption of a gene or its regulatory elements or put these elements into functional proximity of another gene. The most well-known example is the Philadelphia Chromosome, where the end of chromosome 9 is exchanged with that of chromosome 22 (t(9;22)(q34.1;q11.2)), resulting in a fusion gene *BCR-ABL*. This rearrangement occurs in 95% of cases of CML (Rowley, 2001). Inversions are similar, occurring within one chromosome when a segment of DNA is reversed. One subtype of AML is almost always associated with a specific inversion of chromosome 16 (inv(16)(p13;q22)), resulting in the fusion gene *MYH11-CBFB* (Liu, 1993). Interestingly, this inversion actually confers sensitivity to chemotherapy and a higher percentage of these patients achieve complete remission (Liu, 1995).

A growing appreciation has emerged for the role of epigenetic modifications in cancer (Feinberg, 2006). These modifications result in heritable changes to DNA that effect gene expression without modifying the sequence itself. Epigenetic modification generally refers to changes arising from two mechanisms. The first involves the post-translation modification of histones which facilitate chromatin remodelling, packing DNA into a form inaccessible to transcriptional machinery (Jenuwein, 2001). The second involves the methylation of the
cytosine in CpG dinucleotides. The majority of CpG sites in mammalian genomes are methylated, except for clusters of these dinucleotides called CpG islands that are present in the 5’ regulatory region of many genes (Bird, 1986). These CpG islands provide a mechanism to stably repress the expression of certain genes, and aberrant hyper- or hypomethylation, can result in unscheduled transcriptional repression or activation (Miranda, 2007). For example, hypermethylation has been shown to repress the transcription of GSTPI, an enzyme that plays an important role in cellular detoxification, in the majority of prostate cancers (Lin, 2001; Maruyama, 2002).

Cancer-promoting changes can be the result of a combination of the above mechanisms, often through loss of heterozygosity (LOH). When one allele has become inactivated, the remaining functioning allele can usually compensate. This heterozygous state is then lost when the remaining gene is disrupted. This two-hit model was first proposed in a study of childhood retinoblastoma (Knudson, 1971). The age of diagnosis in patients with bilateral tumours tended to be younger than those with unilateral tumours. Moreover, the distribution of the age of diagnosis for bilateral tumours was consistent with a single hit, whereas the distribution for unilateral tumours was consistent with two hits. This suggested that both copies of a single gene were inactivated and in bilateral cases, a disrupted copy was inherited. This hypothesis was later confirmed with the discovery of the RB1 gene, where one copy carries an inherited mutation and the second is subsequently lost through deletion in most cases of inherited retinoblastoma (Friend, 1986). LOH has since been implicated in the loss of function of a large number of tumour suppressor genes, and is now one of the core concepts in molecular oncology (Mendelsohn, 2001).

The recent discovery of microRNA (miRNA) has introduced an additional molecular player in cancer biology. These short, single-stranded RNAs act to regulate the expression of genes by targeting mRNA with close complementary sequence and bringing about their
destruction (reviewed in Ambros, 2004). The specific activities of miRNAs in cancer is an area of intense investigation, but one study has demonstrated that the expression levels of several hundred miRNAs are able to distinguish between normal and tumour samples across a wide panel of tissue types (Lu, 2005).

It is clear that there are a large number of genes that show alterations in cancer, a problem compounded by the variety of mechanisms that cause these changes. Furthermore, different malignancies arise from diverse, tissue-specific environments which can drastically influence the effect a given alteration will have. For example, \textit{CAV1} appears to act as a tumour-suppressor gene in one subtype of lung cancer, while being required for survival in another (Sunaga, 2004). Finally, genome instability is a characteristic of the vast majority of cancers, and is capable of inducing genetic changes that have little or no causal or supporting role in tumourigenesis, but nonetheless complicate the search for key alterations.

1.3 LUNG CANCER

Lung cancer is the third most commonly diagnosed cancer in both men and women (behind skin and breast/prostate cancers), but is the leading cause of cancer death in both genders (Canadian Cancer Society, 2008; Jemal, 2007). These tumours typically progress to an advanced and inoperable stage before symptoms compel a patient to seek medical attention. Indeed, a mass of considerable size would be expected before a patient would manifest common lung cancer indicators such as dyspnoea, haemoptysis and chest pain. The result is dismal outcomes – in non-small cell lung cancers (NSCLC), the most common subtype (see Section 1.3.1), advanced disease (stage III and IV) has a 20-30% treatment response rate and an 8-10 month median survival time (Scagliotti, 2007). However, when these tumours are diagnosed early, the treatment success rate is high. Individuals with stage I or stage II NSCLC have average 5-year survival rates of 65% and 41%, respectively (Nesbitt, 1995).
Like all cancers, lung tumourigenesis has a complex genetic basis. As discussed in the previous section, this complexity arises from the diversity of mechanisms that can alter DNA and the large number of potentially affected genes. The abundance of mutagenic agents in tobacco smoke, which is a clear causative factor in lung cancer initiation, adds a particularly insidious dynamic. Polycyclic aromatic hydrocarbons (PAH) such as benzopyrene and nicotine-derived nitrosamines such as nitrosamine 4-methylnitrosamino-1-(3-pyrdiy1)-1-butanone (NNK), are a few of the more potent examples of the hundreds of carcinogens present in tobacco smoke that have non-specific, damaging effects on DNA (Phillips, 1983; Hecht, 1998; Besaratinia, 2002).

1.3.1 Histological subtypes

Histology dictates the classification of lung cancer into several subtypes; each are associated with different clinical parameters such as prognosis, recurrence, and response to treatment. Malignant subtypes are separated into two categories: 1) small cell lung cancer forms its own group distinct from 2) non-small cell lung cancer (NSCLC), which comprises the adenocarcinoma, squamous cell carcinoma, large-cell carcinoma and other atypical subtypes. These two categories are largely historical and reflect the major differences in the primary choice of treatment at the time they were classified; the former being typically treated with chemotherapy, and the latter with surgery and radiotherapy (Holland, 2003). Each subtype and its approximate prevalence are summarized in Table 1.2 and photos providing a visual reference are found in Figure 1.2. A brief overview of the major subtypes follows, with a focus on the squamous cell subtype which is the focus of the work in Chapter IV of this thesis.

1.3.1.1 Small cell lung cancer (SCLC)

SCLC almost always occurs in smokers and generally arises in the central airways. The cells are flat and small, containing exiguous cytoplasm (Figure 1.2A). SCLC is thought to arise from the Kulchitsky cell, which has important neuroendocrine functions (Hattori, 1972). This
<table>
<thead>
<tr>
<th>malignancy</th>
<th>category</th>
<th>subtype†</th>
<th>prevalence (%)‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>benign</td>
<td>papillomas</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>adenomas</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>malignant</td>
<td>small cell lung cancer (SCLC)</td>
<td>small-cell carcinoma</td>
<td>14.6</td>
</tr>
<tr>
<td></td>
<td>non-small cell lung cancer (NSCLC)</td>
<td>squamous cell carcinoma</td>
<td>21.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>adenocarcinoma</td>
<td>35.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>large-cell carcinoma</td>
<td>7.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>adenosquamous carcinoma</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>carcinoid tumour</td>
<td>22.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bronchial gland carcinoma</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>others</td>
<td></td>
</tr>
</tbody>
</table>

‡ Incidence rates are U.S. statistics (Alberg, 2003).
Figure 1.2: **Histology of common subtypes of lung cancer.** All slides are stained with haematoxylin and eosin. Images courtesy of PathoPic (http://alf3.urz.unibas.ch/pathopic).
subtype is highly aggressive and is notorious for early and extensive metastases. Moreover, these tumours are rarely operable and are refractory to chemotherapy; in general, there is a very good initial response, but SCLC nearly always recurs in a highly resistant form (Rosti, 2006). Median survival is only about 7-20 months, depending on the extent of disease (El Maalouf, 2007).

1.3.1.2 Adenocarcinoma (AC)

Lung AC has recently become the most common subtype and occurs mainly in the peripheral airways. Although predominantly associated with smoking history, it is the most likely of all the subtypes to occur when the patient is young or has never smoked (Kreuzer, 1999; Subramanian, 2007). ACs are thought to arise from both type II pneumocytes (which reside in the alveoli) and Clara cells (which reside in the bronchioles), which secrete substances such as structure-maintaining surfactant proteins (Wistuba, 2006) (Figure 1.2B). These tumours are histologically heterogeneous, and can be classified into even further subtypes. Treatment options are numerous (e.g. surgery, chemotherapy, radiotherapy) and depend heavily on stage and histological characteristics, among other factors (Collins, 2007).

1.3.1.3 Atypical subtypes

Rarer subtypes comprise tumours that either: a) have unusual histology, making them difficult to definitively assign to other groups, or b) affect pulmonary structures that are uncommon sites of tumour formation (e.g. bronchial gland carcinoma). The most common of these atypical types are large cell carcinomas which get their name from the presence of large, undifferentiated cells (Figure 1.2C). Since they show no evidence of squamous or glandular differentiation, which would place them in the squamous cell carcinoma and adenocarcinoma categories, respectively, they are assigned a separate group by default. On the other end of the spectrum, adenosquamous carcinomas display characteristics of both the adenocarcinoma and
squamous cell carcinoma subtypes.

1.3.1.4 Squamous cell lung cancer (SCC)

SCC occurs mainly in the central airways and is strongly associated with smoking history. At one time, this subtype occurred with a much greater prevalence (similar to or greater than adenocarcinoma), but it is likely that the introduction of filtered, low-tar cigarettes and a decrease in smoking rates have contributed to their decline (Kreuzer, 1999). SCCs are thought to arise from the phenotypically similar squamous epithelium, a common cell type that plays important functional roles throughout the body. In the lungs, the flat squamous cells line the alveoli and promote gas exchange; in the central airways, where SCCs occur, squamous epithelium results from a transformation of the normal layer of pseudostratified columnar epithelium as a result of genetic mutation (e.g. from carcinogens in tobacco smoke) (Puchelle, 2006) (Figure 1.2D). Unlike other lung cancer subtypes, SCCs have a clearly defined and observable progression from normal epithelium to malignancy. Squamous cell carcinomas at other sites of the body have similar pre-malignant lesions (e.g. skin, oesophagus, and cervix) (Neville, 2002; Greer, 2006; Shimizu, 2007).

1.3.1.4.1 Stages of progression

Despite the relative difficulty in obtaining samples of pre-malignant lesions, lung neoplasms of the squamous subtype are histologically well-studied, and a detailed series of early progressive steps have been defined (Colby, 1998; Franklin, 2000; Kerr, 2001). The majority of cases are thought to progress sequentially from normal epithelium through, in order, hyperplasia, squamous metaplasia, dysplasia (mild, moderate, severe), carcinoma in situ (CIS), and finally an invasive carcinoma (Figure 1.3). However, this series can only be considered generally correct since each individual lesion may regress, fluctuate between steps, and even skip some steps entirely (Breuer, 2005). Furthermore, histologists may differ on their classification scheme.
Figure 1.3: Histological progression of squamous cell lung cancer. Adapted from Wistuba and Gazdar (2006).
depending on a variety of subjective factors. For example, squamous metaplasia may be considered mild dysplasia by some health facilities; dysplasia itself, which is a continuum of changes with overlapping features, is also subject to inter-observer variability with respect to severity (Nicholson, 2001).

Bearing this in mind, a brief description of the stages is as follows: hyperplasia is characterized by a thickening of the basal cell layer with the cells appearing otherwise normal; squamous metaplasia shows characteristic cytologic atypia - the cells flatten and may show keratinisation; mild dysplasia show a further thickening of the cell layer, a pleomorphic increase in the overall size of the cell and the relative size of the nucleus, and crowding of the cells at the basal layer; moderate dysplasia are more marked in these characteristics, along with evidence of rapid division apparent from cells in visible mitosis; severe dysplasia show very apparent abnormalities of the previous types involving most, but not all, cells; in carcinoma in situ all cells are completely atypical, showing extensive defects in cellular architecture and mitotic activity; finally, an invasive carcinoma is one in which the cells have broken through the basement membrane, infiltrating the surrounding tissue (Nicholson, 2001).

The step in this progression that represents a shift from an at-risk lesion to a committed carcinoma is unclear. Indeed, the increased risk of progressing to a carcinoma associated with each subsequent step is strong evidence for their assigned order. However, greater than 90% of the lesions identified as CIS eventually become a full-blown tumour, suggesting this step represents a de facto committed carcinoma (Venmans, 2000; Bota, 2001). Detection of SCC at this early stage of development has a drastic effect on survival, as treatment of CIS is essentially curative (>90% 5-year survival) (Lam, 2000).
1.3.1.4.2 Common genetic alterations

Some characterization of the genetic basis of SCC has been performed. However, no single causal alteration has been identified. Moreover, many of the identified markers consist of chromosomal aberrations where the specific gene(s) affected are as yet unknown. Nevertheless, there are a handful of genetic factors that currently stand out in SCC.

Deletions of 3p

Allelic losses of the short arm of chromosome 3 (3p) are an early and consistent genetic change associated with all lung cancers, including SCC (reviewed in Zabarovsky, 2002). In SCC, the frequency and extent of these losses is strongly correlated with the stage of progression; moreover, changes are often observed in histologically normal tissue in smokers, but never in non-smokers (Wistuba, 1999; Wistuba, 2000). Although initially thought to harbour a critical lung tumour suppressor gene (TSG), the inconsistency in the precise chromosomal locations of these changes and the failure to identify a single candidate gene has led to speculation that 3p harbours several such genes that have a complex relationship with cancer initiation and progression (Zabarovsky, 2002).

Loss of CDKN2A (p16\(^{INK4a}\))

Cyclin-dependent kinase inhibitor 2A, a classic TSG, plays a key role in cell cycle regulation (Kamb, 1994). The CDKN2A locus encodes several transcripts which inhibit CDK4, a kinase necessary for G1 phase progression; in addition, an alternate open reading frame (ORF) codes for a protein that sequesters MDM1, which is responsible for p53 degradation. Restricted CDKN2A function is observed in a wide variety of human cancers (Herman, 1995). In SCC, loss of CDKN2A expression is nearly universal and is often observed in preneoplastic lesions (Belinsky, 1998; Dessy, 2008). As with other cancers, this loss occurs through a variety of
mechanisms; in lung cancer this commonly involves homozygous deletion or promoter hypermethylation (Gazzeri, 1998).

**Increase in EGFR**

The epidermal growth factor receptor, as its name suggests, is a transmembrane tyrosine kinase which is activated by binding members of the epidermal growth factor (EGF) family, which initiates a signalling cascade that promotes cell proliferation (reviewed in Carpenter, 2000). Overexpression of *EGFR*, usually through an increase in gene copy number, is common in NSCLC and SCC in particular (Hirsch, 2003; Nakamura, 2006; Jeon, 2006). Recently, EGFR has garnered intense interest for lung cancer treatment when it was discovered that particular somatic mutations predict sensitivity to the newly developed EGFR inhibitor gefitinib (Paez, 2004). Unfortunately, the results implicated only the adenocarcinoma subtype.

**Loss of TP53**

*TP53* is the most frequently mutated gene known in human cancer, with modifications observed in more than half of cases (Benard, 2003). *TP53* is often referred to as the “guardian of the genome” because of its role in initiating growth arrest in the event of DNA damage, and inducing apoptosis if the damage cannot be repaired (Lane, 1992). In lung cancer, *TP53* mutations, most often accompanied with LOH, are observed in about 50% of cases of NSCLC (both AC and SCC subtypes) and over 70% of cases of SCLC (Weston, 1989; Miller, 1992).

**Increased telomerase**

Telomeres consist of TTAGGG repeats that reside at the ends of each chromosome. After mitosis, the ends of chromosomes are shortened and the telomeres provide a buffer that delays the degradation of important genetic material. In normal somatic cells, this progressive telomere shortening confers an upper bound (called the Hayflick limit) to the number of possible cell divisions. Once the telomeres are lost, DNA damage responses are triggered causing the cell to
lose the ability to divide in a process called senescence. Telomerase is able to replace the DNA lost at the telomeres, and is vital in stem cells that are required to divide throughout the lifetime of the organism. In many cancers, telomerase activity is increased and is likely an essential mechanism for allowing the uncontrolled cell division of a malignant tumour. This appears to be a common feature of lung cancer as well, with about 80% of SCCs showing increased telomerase activity; other lung cancer subtypes show similar incidence (Lantuéjoul, 2007).

1.4 GENOMICS AND GENE EXPRESSION PROFILING

Genomics, the study of the genes at the scale of the whole genome, is a field that has existed for less than thirty years. Advances in gene sequencing and computer technology have made genomic studies practical and cost effective, and the field continues to develop rapidly. There are two overarching rationales for using genomics as an approach in cancer research: a) hitherto unknown genetic determinants of tumourigenesis and other important factors, such as susceptibility or clinical response, can be more easily identified by a wide-ranging examination of the genome; and b) as a complex genetic disease, resulting from an intricate interplay of genetic aberrations, a global view of molecular change can provide a fuller understanding not possible with more selective study.

Success in sequencing whole genomes, notably the human genome, was the result of significant advances in high-throughput, automated techniques (Lander, 2001; Venter, 2001). As DNA sequencing became more inexpensive and technically undemanding, the application of these technologies to mRNA in order to characterize the transcriptome was a natural progression. Early gene expression profiles were generated by sequencing ESTs (random, single pass reads from the ends of a collection of cDNAs) (Adams, 1991). As the sequence of expressed genes became better defined and prediction models of gene loci were developed, hybridization-based, low-cost and highly parallelized spotted cDNA and oligonucleotide microarrays became possible
(Pease, 1994; Schena, 1995). Finally, improvements to the original EST method were developed, including SAGE and MPSS; these techniques decreased the amount of sequence required to identify a gene, and increased the number of identified transcripts per sequence read (Velculescu, 1995; Brenner, 2000; Saha, 2002; Matsumura, 2005).

The central dogma of molecular biology states that information flow in biological systems travels in one direction – from DNA to RNA to protein (Crick, 1970). Thus, changes to DNA affect RNA, which in turn affect the proteins that carry out the functions of the cell. In this sequential process, RNA is the first intrinsically dynamic layer, resulting from the interplay between the genome and the regulatory factors it interacts with (Figure 1.4). Thus, a transcriptome profile represents one of the most revealing looks at the current status of a cell, capable of indicating both genome-level changes and environmental influences. DNA deletions, amplifications, changes in the activity of transcription factors and changes to the methylation status of individual loci are all expected to influence gene expression levels. However, there are inherent limitations. Small mutations in the coding region of a gene will be missed. Complex chromosomal rearrangements may result in situations that are not reflected in a gene expression profile. Post-transcriptional splicing of mRNA can be detected in some, but not all, circumstances. The myriad number of post-translational modifications to proteins (e.g. alkylation, glycosylation, proteolytic cleavage, and so on) can have drastic affects on function, and cannot be detected.

However, despite the inability of transcriptome profiling to directly identify these events, one can adopt an optimistic point of view. Most, if not all, biochemical modifications will be expected to alter the transcriptome to some extent because the change a) has direct regulatory effects on other genes, or b) alters the steady-state of the cell requiring a response in order to maintain homeostasis. An example of the former is TP53, which is most often inactivated through missense point mutations not reflected at the transcriptional level (Soussi, 2001).
Figure 1.4: The role of gene expression in the cell. The three asterisks signify processes that will be reflected in transcriptome profiles.
However, TP53 is a potent transcriptional regulator, modulating the expression of dozens of genes such as \textit{CDKN1A}, \textit{MDM2}, \textit{PTEN}, and \textit{SFN} (Kanehisa, 2008). Thus, an alteration to TP53 may not result in a change to its own expression, but will be reflected in changes to the expression of downstream target genes. An example of the latter is the Warburg effect, which describes the increased glycolytic rates of tumour cells (Warburg, 1956). Rapid tumour growth results in a hypoxic environment that prevents oxidative phosphorylation, the normal source of energy, from functioning. HIF1A, a transcription factor that is activated under anaerobic conditions, is activated in the majority of cancers and induces the expression of several genes, including enzymes that drive the glycolysis pathway (Maxwell, 2001).

A common objective of gene expression profiling is to generate “signatures” that differentiate between two or more phenotypes. In the best case scenario, these signatures will include genes that have some direct, causal relationship with a phenotypic change; failing that, these signatures can identify genes or processes that can be used as a starting point to hunt down causal events through alternate methods. Moreover, improvements in clinical care can be achieved without identifying the underlying genetic aetiology, as long as the resulting signature has sufficient sensitivity and specificity.

An early demonstration of the power of gene expression profiling used microarrays to identify a signature that could distinguish between acute myeloid leukaemia (AML) and acute lymphocytic leukaemia (ALL) (Golub, 1999). Later, it was shown that gene expression profiles could identify specific signatures that could distinguish between many different tumour types (Ramaswamy, 2001). Such studies were successful in demonstrating that transcriptome profiles could yield sufficient information to make medically relevant phenotypic distinctions. Many of these distinctions are already possible with existing medical technology (for example, imaging techniques such as X-ray, CT, and MRI or with histological observation of tumour biopsies), but as genome profiling technologies becomes more advanced and economical they may become a
superior alternative. Moreover, gene expression profiling studies have proven successful in
discovering novel tumour subtypes that improve existing classifications, or have identified
signatures that correlate with important clinical variables such as prognosis and treatment
response (Lin, 2008; van’t Veer, 2008).

1.5 SERIAL ANALYSIS OF GENE EXPRESSION (SAGE)

Serial analysis of gene expression (SAGE) and DNA microarrays are the two methods
which have emerged to obtain high-throughput gene expression profiles (Schena, 1995;
Velculescu, 1995). It is instructive to compare SAGE to the more commonly used DNA
microarray, since the two have similar purposes (Table 1.3). SAGE is based on the capture of
small sequence tags that are extracted from a defined position of each mRNA molecule in an
input sample. The resulting SAGE data consists of quantitative tag counts; this is different from
microarrays which use DNA hybridization to obtain an analog signal which is much more
qualitative in nature. Moreover, microarray hybridization occurs on a pre-fabricated chip where
the target sequences must be known a priori. In contrast, SAGE samples the entire mRNA
population, generating a comprehensive profile that permits novel gene discovery. Although
originally developed to study the cancer transcriptome, SAGE has found wide use in profiling
diverse cell types from a variety of organisms.

1.5.1 Constructing a library

There is a lengthy protocol required to construct a SAGE library (Figure 1.5). First, the
mRNA from a sample of biological material is converted to cDNA using biotinylated
oligonucleotide deoxythymidine primers. The incorporated biotin allows the cDNA to associate
with streptavidin-coated magnetic beads so the molecules can be isolated by pulling down the 3’
end. The cDNA are then subjected to restriction digest by an anchoring enzyme (AE) (most
frequently NlaIII, which cuts at the recognition site CATG). The 3’ ends of the cDNA, which are
Table 1.3: Comparison of SAGE and DNA microarray technologies

<table>
<thead>
<tr>
<th>consideration</th>
<th>DNA microarray</th>
<th>SAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>data output</td>
<td>hybridization-based (analog)</td>
<td>count-based (digital)</td>
</tr>
<tr>
<td>detection</td>
<td>limited to probes present on the array</td>
<td>limited by sequencing depth of library</td>
</tr>
<tr>
<td>source of noise</td>
<td>background hybridization</td>
<td>sequencing and PCR amplification errors</td>
</tr>
<tr>
<td>re-use</td>
<td>depends on type of array</td>
<td>libraries from different sources can be compared</td>
</tr>
<tr>
<td>cost</td>
<td>low (thousands of $)</td>
<td>high (tens of thousands of $)</td>
</tr>
</tbody>
</table>
Figure 1.5: The serial analysis of gene expression library construction protocol. Image by Jiang Long for the UBC Science Creative Quarterly (sqc.ubc.ca). Minor modifications were made to correct the size and sequence of the TE recognition site and two cases where the AE site was incorrect. Reproduced under the Creative Commons license.
now shortened to the 3’-most AE recognition site, are again isolated using the magnetic beads. The sample is divided into two pools and a ligation step is performed to attach one of two different linkers to the 5’ end of the cDNA pool. Both linkers introduce a recognition site for a type IIS restriction enzyme, referred to as the tagging enzyme (TE) (for the original SAGE protocol this is \textit{BsmFI}, which cuts 14bp downstream from the recognition site GGGAC). When the sample is subjected to restriction digest by the TE, short tags are released while the remainder of the cDNA remains attached to the magnetic beads. The tags are then blunt-ended and ligated to form ditags. The population of ditags undergoes PCR amplification using primers for each of the two linker sequences. A further digestion with the AE removes the linkers, and the final ditags can be isolated. Finally, the ditags are ligated into long concatemers which are cloned into an appropriate vector for sequencing. Several variants of SAGE have been introduced, each differing in the length of the captured tag (14, 21, and 26bp) (Velculescu, 1999; Saha, 2002; Matsumura, 2003).

Computational methods are then required to analyze the resulting sequence. Initially, this involves an \textit{in silico} digest with the AE to reveal the ditags. These undergo further processing to remove erroneously captured linker sequences and duplicate ditags. The latter is part of an internal control for preferential PCR amplification that may introduce bias into the final tag counts. Indeed, the use of “ditags” is not strictly necessary to capture sequence tags. Since the number of different transcripts in a typical sample is large, it is unlikely that any two tags will associate more than once and any instances of this event can be ascribed to preferential amplification. Once this processing has occurred, the sequence and number of each individual tag can then be determined. Finally, tags must be assigned to a source sequence in order to identify the expressed gene.

Invariably, SAGE libraries are constructed from samples with some phenotypic difference, and the purpose is to identify tags (and ultimately genes) that show a significant
change in number. The application of statistical methods can be relatively straightforward for comparisons of two libraries, but as sequencing costs and the technical challenges of SAGE library construction have diminished, studies featuring many biological replicates and deeply sequenced libraries have become more common. More sophisticated methods are required to analyze datasets with multiple samples, but also to fully leverage the information potential gained from deeply sequenced libraries.

1.6 DATA ANALYSIS AND STATISTICS†

“Data mining” – the extraction of relevant information from large amounts of data – has wide applicability in many scientific fields such as epidemiology and in non-scientific fields such as business. Naturally, the large-scale nature of genomics has required the field to drawn upon accepted data mining strategies. The use of these methods is required to isolate informative segments of large datasets and help determine which observations are most relevant to the subject of study. In the case of SAGE, this involves the identification of tag sequences that have some change in count consistent with a given hypothesis. However, data mining has known issues that must be carefully considered when analyzing and interpreting large datasets, and SAGE data is no exception.

These issues can be more clearly demonstrated by example. A recent study featured an analysis of SAGE profiles obtained from the lung epithelium of individuals who have never smoked, had quit smoking, or are current smokers (Chari, 2007). The study highlighted the differences in gene expression between these three groups, with a particular focus on a substantial number of changes that appear to persist after an individual has stopped smoking. These results were intriguing because they suggest that some aetiology is permanently

† The examples used in this section of the introduction are taken from Zuyderduyn, S.D. (2009) Correspondence regarding 'Effect of active smoking on the human bronchial epithelium transcriptome'. *BMC Genomics*, 10:82. A fuller discussion of certain statistical pitfalls and other more esoteric flaws in the analysis of SAGE data can be found in the complete manuscript.
maintained after smoking cessation, that this manifests itself at the level of gene expression, and that such changes may contribute to an increased risk of developing lung cancer. Under scrutiny, however, the statistical methods employed by the authors seriously affected the validity of these conclusions. However, the Chari et al. study is a useful tool with which to discuss some of the common pitfalls of analyzing large datasets and their practical consequences. For simplicity, the language of frequentist statistics is used in the following overview, although the issues are equally applicable to Bayesian approaches.

1.6.1 Statistical error and bias

Statistical errors are commonly divided into two categories and operate under the assumption of a null hypothesis - the default premise accepted as true until statistical evidence indicates otherwise. Type I error, or a false positive, refers to the rejection of a null hypothesis that is actually true. Type II error, or a false negative, refers to the failure to reject a null hypothesis that is actually false. Other types of error have been proposed, many of which are somewhat lighthearted descriptions of statistical pitfalls. For example, Type III error has been described as correctly rejecting the null hypothesis for the wrong reason and Type IV error as the incorrect interpretation of a correctly rejected null hypothesis (Mosteller, 1948; Marascuilo, 1970).

Statistical bias is easy to define, but is often difficult to identify in specific cases without some vigilance. When some set of data is collected or manipulated in a manner that makes it more likely that an undesired or unaccounted for property is represented, there is a risk of bias. If the overrepresented property is associated in some way with a variable being tested for a significant change or effect, then the experiment is biased.

An example of this occurs in Chari et al. during the filtering of their dataset. When comparing a set of SAGE libraries, it is not unusual to remove tags with uniformly low
expression since there is not enough signal to provide any statistical significance. However, Chari et al. restrict their testing to tags with a “mean tag count of ≥20 tags per million (TPM) in at least one of never, former or current smoker SAGE libraries”. This introduces a bias because the reduction criterion includes smoking status, the variable being tested. This becomes problematic when estimating the amount of Type I error (discussed in the Section 1.6.3), since the dataset has now been enriched for data that will show changes between the three groups. This bias could be reasonably addressed by filtering using a criterion of a mean expression of 20TPM across all libraries.

1.6.2 Choice of test statistic

A test statistic is a value calculated from the observed data, which can be used to determine significance by calculating the chance of observing an equally extreme value if the null hypothesis is true (i.e. the $p$-value). The choice of test statistic relies on certain assumptions about the data. Test statistics can be classified as parametric, where the frequency distribution of the data is known or can be reasonably assumed, and non-parametric, when the frequency distribution is not known. For example, Student’s $t$-test is a parametric test used to compare the means of two groups of observations where the scale of the variance is unknown, but is assumed to follow a normal distribution (also referred to as a Gaussian distribution). The Mann-Whitney U test is equivalent, but discards the assumption of normality by comparing the relative ranks of the observed values. Although non-parametric tests are free of difficulties arising from incorrectly assuming an underlying distribution, they always have less power than a parametric test.

In Chari et al., the authors choose the Mann-Whitney U test to determine if a tag is differentially expressed between groups. Although the test itself is perfectly valid, some difficulties arise from its use with SAGE data. First, since the relative rank of the tag count is
used to calculate the test statistic, the following two comparisons are equivalent:

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 3, 5</td>
<td>4, 6, 7, 9</td>
<td></td>
</tr>
<tr>
<td>1, 3, 5, 10</td>
<td>9, 30, 50, 60</td>
<td></td>
</tr>
</tbody>
</table>

Common sense would suggest that the values in the second comparison are more statistically significant that the first, but the Mann-Whitney U test will determine that both are equally significant. Second, Chari et al.’s dataset consists of groups containing 4, 12, and 8 libraries. When comparing such a small number of samples, the power of the Mann-Whitney U test is limited. For example, assume for some tag that expression is lower in all libraries of the group of 4 than the group of 8. The lowest possible two-tailed $p$-value is 0.004. Although this value is more than acceptable for a single hypothesis test it can, as discussed in the next section, become problematic when testing many tags.

1.6.3 Multiple testing

Statistical hypothesis testing aims to determine if there is a true difference between two (or more) individual of groups of observations. This was introduced in the previous section, where the use of a test statistic to determine a $p$-value is described. It naturally follows that when one performs such a test on enough observations, cases will appear where the null hypothesis is true but the observed value suggests otherwise. For example, the commonly accepted threshold for significance ($\alpha$) is 0.05. If a hypothesis test is applied to 10,000 items (e.g. genes) where the null hypothesis is true, then one expects 500 will be identified as extreme, or significant, by chance alone. If 1,500 items are subsequently identified from real data, one can estimate that a third of these will be false positives.

Despite the relative simplicity of the multiple testing problem, it continues to be inadequately addressed in many studies using high-throughput data (Dupuy and Simon, 2007).
There are numerous methods to control the false discovery rate (FDR); a simple Bonferroni correction or the Benjamini-Hochberg method are widely accepted and commonly used (E, 1936; Benjamini and Hochberg, 1995). Moreover, modern computers have made powerful resampling approaches such as bootstrapping and Monte Carlo simulation easy to perform.

In Chari et al., over 8,000 tags are tested for a change in expression between smoking status groups, and comparisons resulting in a $p$-value $<0.05$ were considered significant. However, since there is no correction for multiple testing, the actual significance of any given result will be much lower. To show this directly, the same procedure can be applied to a randomization of the data. The class labels are randomly re-shuffled to produce a “null” dataset where one knows that any finding of significance will be spurious. When the $p<0.05$ criterion is used, 418 tags are identified; on the actual dataset, 885 tags are identified. Thus, the false discovery rate can be estimated at $418/885=47.2\%$; thus, about half of the reported findings are false positives. Moreover, the bias introduced by pre-filtering based on group membership (see Section 1.6.2.1) means this value is an underestimate.

### 1.6.4 Let the data challenge assumptions

When a corollary assumption or supposition is made at the outset of an experiment, it is never a mistake to let the data challenge it. This serves several functions: a) provides quality control, as a failure for strong assumptions to be supported may suggest a mistake in implementing or analyzing an experiment; b) allows rejection or refinement of the assumption, and appropriate measures can be taken to account for the uncertainty or pursue an alternative hypothesis.

Chari et al. provides a good example of how a core assumption underlying the experiment was accepted without letting the data challenge it. The authors report only the findings of a test of one null hypothesis: the expression of a tag in the never smoker group is the
same as the current smoker group. The selection of tags showing reversible or irreversible expression is later selected from this subset based on the expression in the former smoker group. This approach makes the tacit assumption that the limits of expression values are defined by the never smoker and current smoker groups, and that the expression of the former smoker group must be somewhere in between. A test of the other potentially informative null hypothesis, specifically never smokers versus former smokers and former smokers versus current smokers are not reported by the authors. When these are tested on the real and null datasets, an interesting result emerges; specifically, the estimated FDR is almost identical in all three of the possible comparisons. This disagrees with the assumption that former smokers will have gene expression profiles lying within a continuum of changes between never and current smokers, which would certainly result in a higher FDR for one or both of the comparisons involving former smokers. Although this doesn’t disprove the assumption, it does suggest that the statistical methods used should be re-examined. One likely explanation is that the criterion for statistical significance is simply not stringent enough to remove random variation, and may be obscuring the expected dynamics of the data.

1.6.5 Information visualization, data clustering, and global comparisons

When analyzing large-scale datasets, it is often desirable to identify broad similarities and differences. Sometimes this serves to efficiently summarize a dataset, but can also be used in exploratory data analysis where the formulation of a hypothesis is desired. Common approaches are principal component analysis (PCA), multidimensional scaling (MDS), $k$-means clustering, and hierarchical (or agglomerative) clustering (Figure 1.6).

PCA serves to reduce a multidimensional dataset to a smaller number of dimensions by identifying a principal component that accounts for as much of the variance as possible (Pearson, 1901). Further principal components can be identified by repeating the procedure on the
Figure 1.6: Examples of common data visualization methods. Clock-wise from top-left are principal component analysis (PCA), multidimensional scaling (MDS), $k$-means clustering, and two-way hierarchical clustering. The PCA is from Yeung (2001), and the MDS and hierarchical clustering from Koinuma (2006).
remaining variance. In gene expression analysis, PCA is typically applied to a large list of genes in order to identify large trends that define the structure of the data (Yeung, 2001).

MDS reduces a matrix of pair-wise similarities into a lower-dimensional space (Gower, 1966). For example, a set of ten objects would require nine dimensions to fully visualize the similarity of each object to all others. Several types of MDS algorithm are available, but in general the procedure re-positions the objects in the lower dimensional space to minimize a “loss function”, which describes the amount of similarity information in the original matrix not accounted for. In gene expression analysis, MDS can be applied to a set of samples to identify the main dynamics that account for the observed similarity.

$k$-means clustering partitions a list of objects into $k$ clusters based on their similarity (Steinhaus, 1956). The $k$-means algorithm begins by placing the objects into $k$ clusters randomly or using some simple heuristic, and then calculating the average vector (or centroid) for each cluster. Each object is then moved to the cluster with the nearest centroid, which are recalculated after each round of re-arrangement. The procedure is repeated until the clusters are stable. In gene expression analysis, $k$-means clustering is useful for identifying families of genes with similar expression patterns across a set of samples (Tavazoie, 1999).

Hierarchical clustering partitions data into subsets (or clusters) based on some distance metric, which can be represented as a tree structure, or hierarchy, called a dendrogram (Eisen, 1998). There are a large number of variations in the hierarchical clustering algorithm, primarily concerned with the calculations of the distance between a) individual observations, and b) the clusters themselves. Hierarchical clustering is useful for determining similarities between samples or genes, and can be presented simultaneously (e.g. two-way clustering where both the rows and columns of a data matrix are clustered).

A challenge of all of these approaches is determining how many components, dimensions or, in the case of hierarchical clustering, what level of a dendrogram adequately captures the
significant features of the data. In the case of PCA, MDS, and \( k \)-means clustering a common strategy is to plot the amount of variance accounted for as a function of the number of components or dimensions, and selecting a point where the amount of additional variance removed is noticeably smaller. This is sometimes referred to as an “elbow” plot, because the datapoints appear as a bent arm with the elbow corresponding to the optimal number of components or dimensions. In the case of hierarchical clustering, one can choose a cutoff similarity where partitions within a cluster are no longer meaningful.

A potential pitfall of these methods arises from their use during class discovery. Identifying subsets of observations that show similar properties can identify meaningful relationships. For example, an analysis of expression profiles from different cancer samples may reveal similarities that correspond to the stage or aggressiveness of the tumour, or to previously unknown molecular subtypes. However, it is important to ensure that claims of correlations identified by clustering are not made on data previously selected based on the same correlation. Like the multiple testing problem, many studies using high-throughput data commit this error (Dupuy and Simon, 2007).

In Chari et al., the authors refer to the use of “supervised clustering”, which is a neologism, but captures the essential nature of this error. In the study, several hundred SAGE tags are selected based on their differential expression between never and current smokers, and it is implied that the observed clustering of these two groups is meaningful. Furthermore, for these specific tags, the expression values from the former smokers are included and then it is claimed that the separation of all three groups into distinct clusters is meaningful. However, the additional samples form a distinct cluster not because of any biological significance, but because they have not undergone the same \( a \text{ priori} \) selection procedure applied to the other two groups. Ironically, if the former smokers had actually merged into the clusters of either the never or current smokers one could make a strong argument that gene expression changes arising from
smoke exposure revert or remain stable upon smoking cessation. However, the argument as presented is only valid if the study used all of the tags (or a subset selected on an unrelated variable such as average expression) and clusters corresponding to smoke exposure emerged.

1.6.6 Cross-validation

Cross-validation (CV) is a powerful technique to confirm the validity of a statistical analysis, to estimate how well a set of putative descriptive features will perform on future data (e.g. predicting the disease status of a sample), and to guard against overfitting. When performed correctly, CV can be an effective means of confirming the findings of a statistical analysis when additional samples are not available or are difficult to obtain. Even when the availability of additional samples is not an issue, CV can help in the development of an effective model of the existing data before expending further resources on validation efforts.

The overall strategy involves partitioning the data into a training set, where a statistical model is applied, and a testing set, which is used to determine the performance of the model. The partitioning procedure is performed a number of times so that a given sample will sometimes be part of the training set and sometimes the testing set. $K$-fold cross-validation involves partitioning the data into $K$ subsets, selecting one of the subsets as the testing set, and combining the remaining subsets to form the training set. A single estimate of performance is then obtained by averaging the results of the $K$ runs of CV. Leave-one-out cross-validation (LOOCV) is a special case of $K$-fold CV where $K$ is equal to the number of samples. In other words, each sample acts as an individual testing set and the procedure involves the maximum number of possible rounds of validation.

It is vital that the construction of the predictive model, or classifier, from the training set not involve the testing set in any way. For example, one may wish to identify a set of genes that can predict whether a biopsy is a malignant tumour. A set of gene expression profiles from a set
of normal or benign tumour and malignant tumours is collected, and a statistical test is used to rank the genes according to how well they differentiate between the sample types (using a metric such as the \( p \)-value or correlation coefficient). CV can then be used to determine how many genes should be used to obtain an optimal classifier, and to estimate the performance of this classifier. A common error is to utilize the ranking metric calculated using all of the samples during the construction of the classifier for the training set in each round of CV. The correct procedure is to re-calculate the ranking metric for each of the training sets. Failure to do so will result in an overestimate of the classifier performance, and may incorrectly calculate the number of genes that should be included in the optimal classifier. Along with the failure to correct for multiple testing and incorrectly assigning significance to clusters identified during class discovery, this is one of the most common errors in studies of cancer using gene expression profiles (Dupuy and Simon, 2007).

1.7 THESIS OBJECTIVES

The overall objective of this thesis was to identify changes in gene expression associated with the development of squamous cell lung cancer. This was accomplished using a large number of SAGE profiles generated from tissue samples of several pre-malignant and malignant stages. However, the rarity of this material made validation studies of any appreciable size extremely difficult. Furthermore, the nature of the samples was such that a straightforward analysis would have had limited effectiveness. As a result, much of the work in this thesis describes novel techniques that were developed to maximize the value of the SAGE profiles, and mitigate the requirement for large follow up studies. The three chapters describe:

1. the development of a technique to extract additional sequence information from SAGE data in order to provide increased confidence in the assignment of a gene to a given tag
2. the development of a statistical model to analyze multiple groups of SAGE libraries that outperforms current methods

3. the characterization of the transcriptome of the developmental stages of squamous cell lung cancer and the identification of specific biomolecules that are highly correlated with each transition using bioinformatic techniques, including those described in Chapters II and III
BIBLIOGRAPHY


CHAPTER II

DETERMINING ADDITIONAL SEQUENCE DATA TO IMPROVE SAGE TAG TO GENE MAPPING†

2.1 INTRODUCTION

One of the key components of SAGE analysis is matching a tag to an mRNA sequence in order to identify the parent transcript or gene. Several modifications of the SAGE protocol have been developed to produce longer tag sequences to make this identification easier and more accurate (Saha, 2002; Matsumura, 2003; Gowda, 2004). However, generating longer tags directly increases the cost per unit of information. Moreover, although recent studies using SAGE typically use these improved protocols, there are over 700 profiles in the public domain that were generated using the original method (which produces 10bp of identifying sequence) (Lash, 2000; Barrett, 2006). The ability to use previously generated profiles in new studies is one of the distinct advantages of SAGE, so efforts to increase the information content of this existing knowledge base can be of great benefit.

Difficulties with tag to gene mapping fall into two categories: 1) inability to assign a parent gene because two or more transcripts share the same tag (ambiguous tag), and 2) identifying the source of a tag when no transcript is matched (novel tag). The original SAGE protocol uses the restriction enzyme BsmFI to cleave 14bp from the start of the recognition site (GGGAC) to release the tag (Figure 1.5). The 3' cytosine of the sequence is also the 5' cytosine of the anchoring enzyme recognition site (CATG). Thus, the final tag should be 15bp long. However, it is apparent from the downstream ditags that the length of the cleaved tag is variable. This effect is likely influenced by sequence, temperature, and the conditions of the reaction mixture (salt content, pH, etc.). The standard data processing for SAGE ignores these additional

† A version of this chapter will be submitted for publication. Zuyderduyn, S.D. and Vatcher, G. Additional sequence information from SAGE can reduce mapping ambiguity and aid novel gene discovery.
nucleotides, ostensibly as a trade-off between extra information and a uniform, high-confidence processed set of tags. In fact, the original description of the SAGE technique considered only the first 13 nucleotides (Velculescu, 1995).

Although SAGE tag mapping is conceptually simple, there are several hazards that are seldom appreciated. First, several PCR amplification steps are required during library construction. Even with the use of a high-fidelity DNA polymerase, errors in replication do occur. When introduced during an early PCR cycle, these errors will be carried through to later cycles and can appear in significant numbers in the final dataset. Second, the anchoring enzyme \textit{NlaIII} (or the less used alternative, \textit{Sau3A}) may not always cut at the 3’-most recognition site of an mRNA molecule. This results in the release of a product where the 2\textsuperscript{nd} or even 3\textsuperscript{rd} tag is captured. Third, tags corresponding to transcript sequences read in the 3’→5’ direction are often observed. Although there is speculation that some “antisense” tags may arise from regulatory RNA molecules, most are conspicuously positioned, sharing the same AE site as the expected tag and expressed at a lower level, suggesting they are a procedural artefact. In any event, the mechanism by which they arise is unclear. Finally, significant numbers of SAGE tags arising from regions of the mitochondrial genome that do not correspond to polyadenylated transcripts and are ostensibly the result of mitochondrial DNA contamination.

The addition of even a small number of nucleotides can potentially reduce the number of ambiguous tag to gene mappings, provide sufficient sequence information to allow the use of a whole-genome sequence search to identify the source of novel tags, and better identify artefacts arising from the SAGE protocol. The following chapter describes a method to determine these additional nucleotides.
2.2 MATERIALS AND METHODS

2.2.1 SAGE data

Sequence data from several hundred SAGE libraries was obtained from the SAGEmap project (ftp://ftp.ncbi.nlm.nih.gov/pub/sage.fasta) at the National Center for Biotechnology Information (NCBI) (Lash, 2000). Corresponding processed data and library information was obtained from the Gene Expression Omnibus (GEO), also at the NCBI (Barrett, 2007). The standard procedure for extracting SAGE tags from sequence data was performed using the Bio::SAGE::DataProcessing Perl module (Zuyderduyn, 2004).

2.2.2 Software development

The tag length distribution estimator (Section 2.3.1) was programmed using the Python language (version 2.4.4) (van Rossum, 2007). The XBP-SAGE algorithm (Section 2.3.5) was programmed using the C++ language (Stroustrup, 2000) and compiled using the GNU C Compiler (GCC) (version 4.1.2). Debugging was assisted using valgrind (version 3.2.1) (Nethercote, 2007). CPU profiling to assist in optimizing algorithm implementation was performed using the Google Performance Tools (google-perftools) (version 0.8) (http://code.google.com/p/google-perftools). Post-processing of the XBP-SAGE output was facilitated through the use of scripts written in Perl (Wall, 2000). Statistical analysis was performed using the R software package (version 2.6.1) (R Core Development Team, 2007). Sequence bending/curvature predictions were performed using the FORTRAN program BEND (Goodsell, 1994). Minor source code modifications were required for a successful build. Wrapper scripts written in Perl were created to allow BEND to be run on many sequences, as the original program was written to provide estimates for a single input sequence.
2.2.3 Tag to gene mapping

Unless otherwise stated, tags were assigned a gene using SAGE Genie (March 03, 2008 release) (Boon, 2002). When a more exhaustive tag to gene mapping was performed the following resources were used:

1) SAGE Genie (March 03, 2008 release)

2) Exact-match BLASTN (Altschul, 1997) to the human genome (NCBI 36 assembly) using the web-based Ensembl BLAST resource (Flicek, 2008)

3) Exact-match BLASTN to the NCBI human EST collection (8,137,888 sequences) (Benson, 2008) using the web-based NCBI BLAST resource (Johnson, 2008)

4) Custom Perl scripts that match a specified tag to other tags in the library that differ by a single nucleotide or a single insertion or deletion. The tag was considered an artefact only if the closely related tag was expressed at least 10-fold greater.

5) Match to a complete database of antisense tags and tags occurring at anchoring enzyme sites other than the 3’-most position from a meta-catalogue of sequences from Refseq (Pruitt, 2007) and Unigene (Wheeler, 2008) (developed by G. Vatcher)
## 2.3 RESULTS

### 2.3.1 Additional sequence information is available for the majority of SAGE tags

During library construction, SAGE tags are ligated tail-to-tail to form ditags. These ditags then undergo an additional ligation step to form long concatemers, which are cloned and sequenced. The anchoring enzyme (i.e. *Nla*III) recognition site (i.e. CATG) then serves to identify the boundaries of each ditag in the resulting sequence data (Figure 1.5). Using the lengths of these ditags, one can determine the distribution of the tag sizes used to construct the library. Simple inspection reveals that the vast majority (>95%) of ditags range in size from 28-32bp. Since the library construction protocol includes a ditag size selection step, sequences that are considerably larger or smaller likely represent cases where sequencing errors have introduced an anchoring enzyme recognition site where none exists or have destroyed a site that does exist. Based on this range, the associated tags must range in size from 14-16bp. Therefore, the following relationships must hold (where P(\(l_d\)) is the proportion of ditags of size \(l_d\), and P(\(l_t\)) is the proportion of tags of size \(l_t\)):

\[
\begin{align*}
P(l_d=28) &= P(l_t=14)^2 \\
P(l_d=29) &= 2P(l_t=14)P(l_t=15) \\
P(l_d=30) &= 2P(l_t=14)P(l_t=16) + P(l_t=15)^2 \\
P(l_d=31) &= 2P(l_t=15)P(l_t=16) \\
P(l_d=32) &= P(l_t=16)^2
\end{align*}
\]

Throughout the text, \(P(l_d)\) denotes a vector containing the values from \(l_d=28\) to \(l_d=32\), and \(P(l_t)\) denotes a vector containing the values from \(l_t=14\) to \(l_t=16\). Using the relationships above, the proportion of each size of tag \(P(l_t)\) must satisfy the observed values for the proportion of ditag lengths \(P(l_d)\). A simple genetic algorithm was developed to estimate the values of \(P(l_t)\) (Figure 2.1). The starting population consisted of 100 random solutions for \(P(l_t)\). Each of these
100 random solutions for $P(l_t)$

Child with highest fitness – the sum of squared errors for the $P(l_d)$ calculated from candidate $P(l_t)$ vs. actual $P(l_d)$

Create 100 child solutions:
- 90% undergo change to random $P(l_t)$ by amount drawn from Beta distribution ($\alpha=1$, $\beta=20$)
- 10% undergo “translocation” where two random $P(l_t)$ are swapped

Stop algorithm when best solution is stable for 10 generations

Figure 2.1: Summary of genetic algorithm to find solutions for proportion of tag lengths $P(l_t)$. 
solutions produces 100 children, where 90% will undergo a change to the value of a random element of $P(l_i)$ (i.e. mutation) and the remaining 10% will swap the solutions of two random elements of $P(l_i)$ (i.e. translocation). The amount a value of $P(l_i)$ changes due to “mutation” is governed by a Beta distribution, with shape parameters $\alpha=1$ and $\beta=20$ (Figure 2.2). This produces very small changes in the vast majority of offspring, but larger changes still have the opportunity to occur from time to time. The fitness of an individual solution is the sum of square errors between the $P(l_d)$ observed from the data, and the expected $P(l_d)$ calculated from the proposed solution; the smaller this value, the better the solution. The single fittest offspring was carried over to the successive round of reproduction, and the process repeated until the best solution remained stable for 10 generations. The problem is relatively straightforward, and little optimization of the algorithm was required. A stable solution is usually found in <30 generations. On a Pentium 4 2.4GHz computer, execution took less than 30 seconds.

The algorithm was applied to 71 publicly available SAGE libraries (see Section 2.2.1). It is observed that the vast majority of SAGE tags have additional sequence beyond the canonical 14bp (>95%) (Table 2.1). Average values for $P(l_i)$ were $P(l_i=14)=0.036\pm0.039$, $P(l_i=15)=0.69\pm0.074$, and $P(l_i=16)=0.27\pm0.069$ (95% confidence intervals shown). Thus, it is clear that additional nucleotides beyond the commonly used 14bp are available in the majority of cases.

### 2.3.2 Theoretical improvement in mapping accuracy from extra nucleotides

In order to quantify the potential benefit of obtaining additional nucleotides, the SAGE Genie tag to gene mapping resource (March 03, 2008 release) (Boon, 2002) was examined. This resource contains several components; relevant here are a) a “master” list of tags and associated counts for 112 LongSAGE libraries, and b) a list of “best gene” mappings, where tags that have been previously observed in (a) are matched to a Unigene entry (Wheeler, 2008).
Figure 2.2: The beta distribution. Plotted is the value of the beta function with parameters $\alpha=1$ and $\beta=20$ (y-axis) as a function of $x$ (x-axis).
Table 2.1: Estimated solutions for the distribution of tag lengths for 71 publicly available SAGE libraries

<table>
<thead>
<tr>
<th>Library name</th>
<th>P(lt=14)</th>
<th>P(lt=15)</th>
<th>P(lt=16)</th>
<th>Sum square error</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAGE_293-CTRL</td>
<td>0.03</td>
<td>0.70</td>
<td>0.27</td>
<td>5.61E-05</td>
</tr>
<tr>
<td>SAGE_293-IND</td>
<td>0.03</td>
<td>0.70</td>
<td>0.27</td>
<td>1.26E-04</td>
</tr>
<tr>
<td>SAGE_95-259</td>
<td>0.02</td>
<td>0.70</td>
<td>0.28</td>
<td>2.42E-05</td>
</tr>
<tr>
<td>SAGE_95-260</td>
<td>0.03</td>
<td>0.69</td>
<td>0.28</td>
<td>8.35E-05</td>
</tr>
<tr>
<td>SAGE_95-347</td>
<td>0.02</td>
<td>0.70</td>
<td>0.28</td>
<td>1.89E-04</td>
</tr>
<tr>
<td>SAGE_95-348</td>
<td>0.03</td>
<td>0.71</td>
<td>0.26</td>
<td>2.94E-05</td>
</tr>
<tr>
<td>SAGE_A+</td>
<td>0.03</td>
<td>0.63</td>
<td>0.33</td>
<td>3.32E-05</td>
</tr>
<tr>
<td>SAGE_A2780-9</td>
<td>0.05</td>
<td>0.71</td>
<td>0.25</td>
<td>6.79E-06</td>
</tr>
<tr>
<td>SAGE_BBS42_whitematter</td>
<td>0.03</td>
<td>0.72</td>
<td>0.24</td>
<td>3.04E-04</td>
</tr>
<tr>
<td>SAGE_Br_N</td>
<td>0.03</td>
<td>0.68</td>
<td>0.29</td>
<td>4.80E-06</td>
</tr>
<tr>
<td>SAGE_CAPN1</td>
<td>0.04</td>
<td>0.72</td>
<td>0.24</td>
<td>6.28E-05</td>
</tr>
<tr>
<td>SAGE_CPDPR_LNCP-P-C</td>
<td>0.05</td>
<td>0.68</td>
<td>0.27</td>
<td>2.26E-06</td>
</tr>
<tr>
<td>SAGE_CPDPR_LNCP-T</td>
<td>0.03</td>
<td>0.69</td>
<td>0.28</td>
<td>9.92E-06</td>
</tr>
<tr>
<td>SAGE_Caco_2</td>
<td>0.04</td>
<td>0.74</td>
<td>0.22</td>
<td>4.80E-04</td>
</tr>
<tr>
<td>SAGE_Chen_LNCP</td>
<td>0.06</td>
<td>0.70</td>
<td>0.24</td>
<td>1.88E-05</td>
</tr>
<tr>
<td>SAGE_Chen_Normal_Pr</td>
<td>0.06</td>
<td>0.70</td>
<td>0.24</td>
<td>9.64E-06</td>
</tr>
<tr>
<td>SAGE_Chen_Dcbx</td>
<td>0.08</td>
<td>0.67</td>
<td>0.25</td>
<td>1.48E-04</td>
</tr>
<tr>
<td>SAGE_DCIS-3</td>
<td>0.03</td>
<td>0.64</td>
<td>0.33</td>
<td>1.94E-06</td>
</tr>
<tr>
<td>SAGE_DCIS-4</td>
<td>0.03</td>
<td>0.71</td>
<td>0.25</td>
<td>4.09E-06</td>
</tr>
<tr>
<td>SAGE_DCIS-5</td>
<td>0.04</td>
<td>0.68</td>
<td>0.26</td>
<td>1.29E-06</td>
</tr>
<tr>
<td>SAGE_DCIS_2</td>
<td>0.03</td>
<td>0.69</td>
<td>0.28</td>
<td>9.56E-05</td>
</tr>
<tr>
<td>SAGE_Duke-H988</td>
<td>0.02</td>
<td>0.73</td>
<td>0.25</td>
<td>4.07E-05</td>
</tr>
<tr>
<td>SAGE_Duke_1273</td>
<td>0.03</td>
<td>0.72</td>
<td>0.26</td>
<td>6.04E-05</td>
</tr>
<tr>
<td>SAGE_Duke_40N</td>
<td>0.02</td>
<td>0.67</td>
<td>0.30</td>
<td>7.49E-06</td>
</tr>
<tr>
<td>SAGE_Duke_48N</td>
<td>0.03</td>
<td>0.70</td>
<td>0.27</td>
<td>2.20E-05</td>
</tr>
<tr>
<td>SAGE_Duke_757</td>
<td>0.03</td>
<td>0.71</td>
<td>0.26</td>
<td>1.00E-05</td>
</tr>
<tr>
<td>SAGE_Duke_96-349</td>
<td>0.03</td>
<td>0.59</td>
<td>0.38</td>
<td>7.49E-04</td>
</tr>
<tr>
<td>SAGE_Duke_BBS42_normal_cerebellum</td>
<td>0.12</td>
<td>0.62</td>
<td>0.26</td>
<td>5.43E-05</td>
</tr>
<tr>
<td>SAGE_Duke_GBM_H1110</td>
<td>0.03</td>
<td>0.73</td>
<td>0.24</td>
<td>1.30E-04</td>
</tr>
<tr>
<td>SAGE_Duke_H1020</td>
<td>0.02</td>
<td>0.71</td>
<td>0.27</td>
<td>2.83E-05</td>
</tr>
<tr>
<td>SAGE_Duke_H1043</td>
<td>0.01</td>
<td>0.68</td>
<td>0.31</td>
<td>4.06E-05</td>
</tr>
<tr>
<td>SAGE_Duke_H1126</td>
<td>0.02</td>
<td>0.69</td>
<td>0.29</td>
<td>2.40E-04</td>
</tr>
<tr>
<td>SAGE_Duke_H247_Hypoxia</td>
<td>0.03</td>
<td>0.74</td>
<td>0.23</td>
<td>5.17E-05</td>
</tr>
<tr>
<td>SAGE_Duke_H247_normal</td>
<td>0.03</td>
<td>0.75</td>
<td>0.22</td>
<td>9.17E-06</td>
</tr>
<tr>
<td>SAGE_Duke_H341</td>
<td>0.05</td>
<td>0.71</td>
<td>0.25</td>
<td>2.83E-05</td>
</tr>
<tr>
<td>SAGE_Duke_H392</td>
<td>0.02</td>
<td>0.70</td>
<td>0.28</td>
<td>1.12E-03</td>
</tr>
<tr>
<td>SAGE_Duke_H54_EGFRvIII</td>
<td>0.02</td>
<td>0.72</td>
<td>0.25</td>
<td>1.79E-04</td>
</tr>
<tr>
<td>SAGE_Duke_H54_lacz</td>
<td>0.02</td>
<td>0.71</td>
<td>0.27</td>
<td>1.82E-04</td>
</tr>
<tr>
<td>SAGE_Duke_HMVCEL-VGF</td>
<td>0.03</td>
<td>0.72</td>
<td>0.25</td>
<td>2.81E-05</td>
</tr>
<tr>
<td>SAGE_Duke_HMVPEC</td>
<td>0.04</td>
<td>0.69</td>
<td>0.27</td>
<td>2.61E-06</td>
</tr>
<tr>
<td>SAGE_Duke_Kidney</td>
<td>0.05</td>
<td>0.70</td>
<td>0.25</td>
<td>1.57E-04</td>
</tr>
<tr>
<td>SAGE_Duke_leukocyte</td>
<td>0.04</td>
<td>0.73</td>
<td>0.23</td>
<td>2.83E-06</td>
</tr>
<tr>
<td>SAGE_Duke_mh1-1</td>
<td>0.04</td>
<td>0.71</td>
<td>0.25</td>
<td>2.64E-05</td>
</tr>
<tr>
<td>SAGE_Duke_post_crisis_fibroblasts</td>
<td>0.05</td>
<td>0.59</td>
<td>0.36</td>
<td>2.70E-05</td>
</tr>
<tr>
<td>SAGE_Duke_premiss_fibroblasts</td>
<td>0.06</td>
<td>0.66</td>
<td>0.29</td>
<td>7.20E-05</td>
</tr>
<tr>
<td>SAGE_Duke_thalamus</td>
<td>0.02</td>
<td>0.66</td>
<td>0.32</td>
<td>2.43E-04</td>
</tr>
<tr>
<td>SAGE_ES2-1</td>
<td>0.02</td>
<td>0.79</td>
<td>0.19</td>
<td>4.36E-05</td>
</tr>
<tr>
<td>SAGE_H1126</td>
<td>0.04</td>
<td>0.70</td>
<td>0.27</td>
<td>5.43E-06</td>
</tr>
<tr>
<td>SAGE_H1126</td>
<td>0.02</td>
<td>0.70</td>
<td>0.29</td>
<td>1.09E-04</td>
</tr>
<tr>
<td>SAGE_H1127</td>
<td>0.04</td>
<td>0.69</td>
<td>0.28</td>
<td>1.02E-06</td>
</tr>
<tr>
<td>SAGE_H408</td>
<td>0.03</td>
<td>0.72</td>
<td>0.25</td>
<td>6.32E-06</td>
</tr>
<tr>
<td>SAGE_HCT116</td>
<td>0.03</td>
<td>0.70</td>
<td>0.27</td>
<td>5.04E-04</td>
</tr>
<tr>
<td>SAGE_HMEC-B41</td>
<td>0.13</td>
<td>0.59</td>
<td>0.27</td>
<td>4.56E-04</td>
</tr>
<tr>
<td>SAGE_HOSE_4</td>
<td>0.02</td>
<td>0.68</td>
<td>0.29</td>
<td>6.91E-05</td>
</tr>
<tr>
<td>SAGE_LN1</td>
<td>0.04</td>
<td>0.69</td>
<td>0.28</td>
<td>1.16E-04</td>
</tr>
<tr>
<td>SAGE_Hemangioma_146</td>
<td>0.04</td>
<td>0.70</td>
<td>0.26</td>
<td>3.56E-05</td>
</tr>
<tr>
<td>SAGE_IDC-3</td>
<td>0.02</td>
<td>0.61</td>
<td>0.37</td>
<td>3.86E-06</td>
</tr>
<tr>
<td>SAGE_IDC-4</td>
<td>0.02</td>
<td>0.61</td>
<td>0.37</td>
<td>1.92E-05</td>
</tr>
<tr>
<td>SAGE_IDC-5</td>
<td>0.04</td>
<td>0.70</td>
<td>0.26</td>
<td>1.52E-06</td>
</tr>
<tr>
<td>SAGE_IDC3-2</td>
<td>0.02</td>
<td>0.61</td>
<td>0.37</td>
<td>3.86E-06</td>
</tr>
<tr>
<td>SAGE_IDC-4</td>
<td>0.02</td>
<td>0.61</td>
<td>0.37</td>
<td>1.92E-05</td>
</tr>
<tr>
<td>SAGE_IDC-5</td>
<td>0.04</td>
<td>0.70</td>
<td>0.26</td>
<td>1.52E-06</td>
</tr>
</tbody>
</table>

Average: 0.04 0.69 0.27 1.18E-04
The “master” list contained 4,953,569 distinct tag types for a total of 158,291,068 tag counts and can be assumed to contain almost all possible SAGE tags present in human tissues. However, many of these tags will represent artefacts such as erroneously sequenced tags. During the processing of a single library, it is not uncommon to remove tags that only appear once in order to ameliorate the contribution of such artefacts. In this case, a slightly expanded criteria was utilized; a tag was considered an artefact if it appeared at a frequency less than or equal to 10 tags per million (TPM) and was observed in only one library. This removed the vast majority of tag types, but most of the total tag count was retained. The resulting list contained 188,525 distinct tag types for a total of 130,393,800 tag counts.

For each LongSAGE tag type, both the first 14bp and 16bp (an additional two nucleotides longer than a canonical short SAGE tag) were used to search the “best gene” mappings for a match. Of the 188,525 tag types: 140,524 were unmatched, 44,462 mapped unambiguously to a single gene, 2,564 were ambiguous for 14bp but were completely resolved for 16bp, 313 were ambiguous for 14bp but were partially resolved for 16bp, and 746 were equally ambiguous for both 14bp and 16bp. Two conclusions can be made: 1) the majority of SAGE tags have no clear match to a gene (74.5%) and 2) of the 14bp tags that do match a gene, 7.4% are ambiguous, but the majority of these (77.0%) can be wholly or partially resolved with the addition of two extra nucleotides.

For SAGE tags that do not match a known mRNA sequence, an attempt to match the genome sequence can identify the source of the tag. However, for canonical 14bp tags this search will likely result in many spurious matches. The estimated size of the human genome is 3.2Gb (3.2x10^9 bp). The average number of locations in the genome that will match a given 14bp tag sequence is 3.2x10^9 / 4^{14} = 11.2. However, for longer sequences, this number drops to 3.2x10^9 / 4^{15} = 2.8 and 3.2x10^9 / 4^{16} = 0.70 matches for 15bp and 16bp tags, respectively. Thus, a small amount of additional sequence information can drastically increase the sensitivity of a
whole-genome search; in the case of 16bp, a single hit is expected to occur in the majority of attempts.

This demonstrates that the addition of even a small number of extra nucleotides has the potential to substantially improve the tag to gene mapping accuracy for a short SAGE library by a) reducing ambiguity and b) providing additional sequence information for a whole-genome search when a tag fails to map a known transcript.

2.3.3 Influence of tag nucleotide composition on length is small, but significant

To investigate the influence of sequence on the length of a tag, a regression analysis was used to identify any relationship between ditag length (the response variable) and nucleotide content (the explanatory variable). This procedure was applied to each of the 71 publicly available SAGE libraries (see Section 2.2.1). Ditag length was modelled as a Gaussian (normal) distributed variable. It should be noted that although ditag length is a discrete value, the binomial distribution was a poor fit to the observed data. As a result, the estimated magnitude of the effect of nucleotide content should be considered a suboptimal, albeit close, approximation.

Let \( L_i \) be the length of some ditag \( i \), \( \mu \) be the mean of all ditag lengths, \( \sigma^2 \) be the variance of all ditag lengths, \( \beta \) be the vector of linear coefficients to be estimated, and \( x \) be the vector of explanatory variables (i.e. \( x_{\text{nuc}} \) is the percentage contribution of some arbitrary nucleotide(s)). The linear regression model can then be expressed as:

\[
L_i \sim \text{Normal}(\mu, \sigma^2)
\]

\[
L_i = \beta_0 + \beta_{\text{nuc}}x_{\text{nuc}}
\]

The model fit was performed using the standard regression functions of the R statistical software (R Core Development Team, 2007). Significance was determined by testing the null hypothesis that \( \beta_{\text{nuc}}=0 \) (nucleotide composition has no effect on ditag length) using the \( t \)-test. Let \( \text{s.e.}(\beta_{\text{nuc}}) \) be the standard error of the estimated coefficient \( \beta_{\text{nuc}} \) and let \( \text{d.f.} \) be the degrees of
freedom. The $t$ statistic is:

$$t = \frac{\beta_{\text{nuc}}}{\text{s.e.}(\beta_{\text{nuc}})} , \text{d.f.} = n_{\text{obs}} - 2$$

An effect was sought for each individual nucleotide (A, C, G, T) and all groups of two nucleotides (AC, AG, AT, CG, CT, GT). In most of the 71 libraries tested, there was a clear association between ditag length and the number of AT or CG nucleotides. In 43/71 libraries, the effect was highly significant ($p<0.001$) and in 6/71 libraries, the effect was moderately significant ($p<0.01$). Of the remaining libraries, 8/22 showed AT or CG content as the strongest association although not to the chosen levels of significance ($p<0.01$). In no case did another possible nucleotide composition show a significant effect. However, the magnitude of this effect was not large nor consistently positive or negative. $\beta_{\text{AT}}$ was positive in 23/71 libraries (0.359±0.162; 95% CI) and negative for 48/71 libraries (-0.191±0.038; 95% CI).

This suggests that a more descriptive explanatory variable exists that is somehow associated with AT and/or CG content. One hypothesis is that certain motifs substantially alter the structure of the target DNA, and those that shorten the tag product tend to be AT rich. Another possibility, which is not mutually exclusive, is that random effects (e.g. local fluctuations in temperature, reaction conditions) simply outweigh the effect of sequence.

2.3.4 Predicted curvature of tag sequence is correlated with ditag length

Based on the results of Section 2.3.3, it was hypothesized that tag sequence influences the shape of the DNA molecule in a manner sufficient to affect the point at which the tagging enzyme ($Bsm$FI) cuts its target sequence. Research into the influence of sequence on DNA shape has been an ongoing concern in investigations into areas such as gene regulation (e.g. transcription factor binding) and DNA packaging (e.g. nucleosome binding). The rod model of DNA has been broadly adopted, and is appealing in the context of the problem here. This model describes short DNA sequences as resembling a “rod” which will bend and twist, depending on
the constituent sequence (reviewed in Munteanu, 1998). If the tagging enzyme cleaves DNA at a fixed distance from the recognition site, then significant bending or other distortion could effectively “shorten” the target DNA, resulting in a larger cleavage product. Indeed, a well known example is that of interspersed adenine tracts that are in-phase with the turns of the helix. Such a configuration will introduce rigidity on one face of the helix that will cause the DNA to bend (Reich, 1992). This example is consistent with the results in Section 2.3.3 where, in many libraries, the presence of AT nucleotides tends to result in shorter tags.

The BEND program predicts local bending and global curvature from arbitrary sequence using a choice of several models (Goodsell, 1994). The BEND developers found that a trinucleotide model based on nucleosome positioning data performs best and so it is used here. For each ditag, the two constituent 14bp tags were analyzed by the program and bend angles for each trinucleotide were calculated. If each trinucleotide X-Y-Z represents a distance of 2 units, then one would have to “walk” 13 units to navigate the helix along the backbone of a SAGE tag. Upon the introduction of a bend, the distance between nucleotide X and nucleotide Z will be shortened. Let \( \theta_i \) be the bend angle of trinucleotide \( i \) from the set of all length-1 trinucleotides. The modified length can be calculated using the expression:

\[
L_{\text{tag}} = \sum_i \cos(\theta_i)
\]

For each ditag, the \( L_{\text{tag}} \) value for the two constituent tags is combined to create a predicted length \( L_{\text{ditag}} \). As in Section 2.3.3, a linear regression analysis was performed to determine if there is a relationship between \( L_{\text{ditag}} \) and the observed ditag length. Let \( \mu \) and \( \sigma^2 \) be the mean and variance of the ditag length, respectively; and let \( \beta_0 \) and \( \beta_L \) be the coefficients of the linear fit.

\[
L_i \sim \text{Normal}(\mu, \sigma^2)
\]

\[
L_i = \beta_0 + \beta_L L_{\text{ditag}}
\]

The model fit was performed using the standard regression functions of the R statistical software (R Core Development Team, 2007). Significance was determined by testing the null hypothesis.
that $\beta_L=0$ (the value for $L_{\text{ditag}}$ has no relationship with ditag length) using the $t$-test, where:

$$t = \frac{\beta_L}{\text{s.e.}(\beta_L)}, \text{ d.f.} = n_{\text{obs}} - 2$$

In this case, the relationship was highly significant in 67/71 libraries ($p<0.001$) and moderately significant in the remaining libraries 4/71 ($p<0.05$). This analysis provides compelling support for the notion that the local shape of the DNA at the endonuclease target region plays a significant role in the size of the cleavage product obtained. However, the relative magnitude of the relationship is not strong. Although a significant proportion of the total variance in ditag length is accounted for by predicted shape effects, a majority of the variance remains. Therefore, calculating the bending of an arbitrary sequence can provide a prediction of ditag length that will outperform a random guess but not to a degree of accuracy that would be of substantial practical use. Thus, it is not possible to determine which tag the additional nucleotides of a ditag arose from based on a particular sequence.

2.3.5 XBP-SAGE: an algorithm to estimate additional nucleotides

One can imagine a library of ditags as being the realization of a population of tags, each having a length of 16bp prior to being subjected to restriction digest by the tagging enzyme (in reality, a tag can be arbitrarily long prior to digestion, but since it is already established that the tagging enzyme does not release a product longer than 16bp, any additional sequence can be ignored and we can suppose a length of 16bp as the effective maximum). The tagging enzyme will release a product that is either this length, or one or two nucleotide(s) shorter. This population is then randomly ligated to form ditags. Thus, we are interested in determining what tag sequences in the original population would have resulted in the observed ditag sequences. This is a classic problem where one can find a solution using the method of maximum likelihood (ML). Whereas a probability (more specifically, a posterior probability) represents the chance than an observed event (A) would have occurred given a known set of parameter values (B) (i.e.
P(A|B)), likelihood represents the chance that some set of parameter values would have resulted in the observed event (i.e. P(B|A)). It is worth noting this distinction since, although the underlying mathematics and resulting values are similar, the terms refer to two separate concepts.

2.3.5.1 The likelihood function

Let \( x \) be an observed variable and \( \theta \) be some parameter of interest. A likelihood function is defined as:

\[
\theta \mapsto P(x|\theta)
\]

\[
L(\theta|x) = P(x|\theta)
\]

In this case, let \( x \) be a possible ditag sequence and \( \theta \) be two candidate tag sequences. Obviously, if the first 14 bp of the tag sequences do not match the 14 bp ends of the ditag sequence, then \( L(\theta|x) = 0 \). Otherwise, \( L(\theta|x) \) will depend on the sequence of the ditag, the probability that a ditag of that length will be observed, and in what manner the extra nucleotides match those of the 16 bp of the proposed tag sequences. The likelihood function can be expressed in more specific terms. Let \( \lambda \) be a ditag ligation site, which results in some putative assignment of the extra nucleotides to the 3' or 5' tag. Then each \( L(\theta,\lambda|x) \) can be used to determine \( L(\theta|x) \):

\[
L(\theta|x) = P(x|\theta) = P(l_d) \sum_{\lambda} [P(x|\theta,\lambda)P(\lambda)]
\]

Consider the following example:

\[
x = 5'\text{-CATGTTTTTTTTTTCCCCCCCATG}-3' \\
3'\text{-GTACAAAAAAAAAAAGGGGGGGGTAC}-5' \quad (29 \text{ bp})
\]

\[
\theta = (5'\text{-CATGTTTTTTTTTTGG}-3', 5'\text{-CATGGGGGGGGGGAA}-3')
\]

We know that the extra nucleotide [T] has an equal probability of arising from either the 5' or 3' tag, thus \( P(\lambda=\text{CATGT}_{10}\text{T}\|C_{10}\text{CATG})=P(\lambda=\text{CATGT}_{10}\|TC_{10}\text{CATG})=0.5 \) (where the character \( \| \) denotes the ligation site). If it arose from the 5' tag, then observing \( \theta \) is impossible.
Thus $P(x|\theta,\lambda=\text{CATGT}_{10}\parallel\text{C}_{10}\text{CATG})=0$; if it arose from the 3' tag, then $\theta$ is plausible ($[\text{T}] = [\text{T}]$), thus $P(x|\theta,\lambda=\text{CATGT}_{10}\parallel\text{C}_{10}\text{CATG})=1$. In addition, we must also consider $P(l_d)$, the probability that a ditag of the observed length results from the ligation of two tags (see Section 2.3.1). Therefore:

$$
L(\theta|x) = P(l_d=29) \times \left[ P(x|\theta,\lambda=\text{CATGT}_{10}\parallel\text{C}_{10}\text{CATG})P(\lambda=\text{CATGT}_{10}\parallel\text{C}_{10}\text{CATG}) + \\
P(x|\theta,\lambda=\text{TGC}_{10}\parallel\text{CATG})P(\lambda=\text{TGC}_{10}\parallel\text{CATG}) \right] 
= P(l_d=29) \times \left[ 0 \times 0.5 + 1 \times 0.5 \right] 
= 0.5 \times P(l_d=29)
$$

In other words, if a 29bp ditag arose from a ligation of the two tags $\theta$, then the observed ditag $x$ would be expected to occur at a frequency equal to one half the probability of observing a ditag of length 29bp. Consider a more complicated example:

$$
x = \begin{array}{c}
\begin{array}{c}
5' - \text{CATGTTTTTTTTTTTG} \\
3' - \text{GTACAAAAAAGGC} \\
(30 \text{ bp})
\end{array}
\end{array}
\begin{array}{c}
\begin{array}{c}
\text{CGGGGGGTTTAC} \\
5' - \text{CATGTTTTTTTTTTA} \\
3' - \text{CATGCCCCCCTTCA} \end{array}
\end{array}

\theta = (5' - \text{CATGTTTTTTTTTTA} - 3', 5' - \text{CATGCCCCCCTTCA} - 3')
$$

We know that the two additional nucleotides $[\text{TG}]$ could be a) both from the 5' tag, b) both from the 3' tag, or c) one from each tag. The probability of each possibility (e.g. $P(l_t=16|l_d=30)$) varies from library to library and can be estimated as previously described (see Section 2.3.1). The unconditional probability of each of these three possibilities is:

$$
a) P(\lambda=\text{CATGT}_{10}\text{TG} \parallel \text{C}_{10}\text{CATG})=P(l_t=16)P(l_d=30)/P(l_d=30) \\
b) P(\lambda=\text{CATGT}_{10}\parallel\text{TGC}_{10}\text{CATG})=P(l_t=14)P(l_d=30)/P(l_d=30) \\
c) P(\lambda=\text{CATGT}_{10}T \parallel \text{GC}_{10}\text{CATG})=P(l_t=15)P(l_d=30)/P(l_d=30)
$$

Now, consider each possibility in turn: a) if both nucleotides came from the 5' tag, then $\theta$ is impossible ($[\text{TG}] \neq [\text{TA}]$), thus $P(x|\theta,\lambda=\text{CATGT}_{10}\text{TG} \parallel \text{G}_{10}\text{CATG})=0$; b) if both nucleotides came from the 3' tag, then $\theta$ is plausible ($[\text{CA}] = [\text{CA}]$), thus $P(x|\theta,\lambda=\text{CATGT}_{10}\parallel\text{TGC}_{10}\text{CATG})=1$; and c) if one nucleotide came from each tag, then $\theta$ is plausible ($[\text{T}] = [\text{T}]$ and $[\text{C}] = [\text{C}]$), thus $P(x|\theta,\lambda=\text{CATGT}_{10}T \parallel \text{GC}_{10}\text{CATG})=1$. Therefore:
\[ L(\theta|x) = P(l_d=30) \times \left[ P(x|\theta, \lambda = \text{CATGT}_{10} \text{TG} \| \text{C}_{10}\text{CATG}) P(\lambda = \text{CATGA}_{10} \text{TG} \| \text{C}_{10}\text{CATG}) + P(x|\theta, \lambda = \text{CATGT}_{10} \text{TG} \| \text{GC}_{10}\text{CATG}) P(\lambda = \text{CATGA}_{10} \text{TG} \| \text{GC}_{10}\text{CATG}) + P(x|\theta, \lambda = \text{CATGT}_{10} \text{TG} \| \text{TGC}_{10}\text{CATG}) P(\lambda = \text{CATGA}_{10} \text{TG} \| \text{TGC}_{10}\text{CATG}) \right] \]

\[ = P(l_d=30) \times \left[ 0 \times P(l_t=16)/P(l_d=30) + 1 \times P(l_t=14)/P(l_d=30) + 1 \times P(l_t=15)/P(l_d=30) \right] \]

\[ = P(l_t=14)P(l_t=16) + P(l_t=15)^2 \]

Tables of likelihood values covering all possible solutions for ditags of length 28-32bp are provided for reference in Appendix I.

However, these equations assume it is known which particular tag arose from any one ditag. In practice, one can only reduce the possible source to the set of ditags that match the first 14bp of sequence. Thus, one must determine the likelihood of observing some ditag given two sets of possible child tags (e.g. the counts of the pair of 14bp tags). Let us define \( \delta \) as a ditag having a sequence 5’-aXb-3’, where \( a \) and \( b \) is the first 14bp of the 5’ and 3’ tag, respectively, and \( X \) are the additional nucleotide(s). Furthermore, let \( A \) and \( B \) be the set of tags where the first 14bp is \( a \) or \( b \), respectively. The size of these sets is denoted by \( n_A \) and \( n_B \). Since one cannot impose an assumption of which specific tag arose from a ditag, the likelihood must be expressed in terms of observing the set of tags \( A \) and \( B \), which is the average of all possible joinings of a tag in \( A \) to a tag in \( B \):

\[ L(A,B|\delta) = P(\delta|A,B) = \frac{\sum_{a} \sum_{b} \sum_{\lambda} P(\delta|\lambda) P(\lambda|a,b)}{n_A n_B} \]

Finally, the global likelihood of all of the observed ditags in a library given some set of tags is simply the joint likelihood over all ditags. Let \( \Delta \) be the entire library of ditags and let \( \Gamma \) be the entire set of 16bp tags (where \( A \) and \( B \) denote subsets of \( \Gamma \) having the same first 14bp of sequence defined by the two flanking portions of some ditag \( \delta \)).

\[ L(\Gamma|\Delta) = \prod_\delta L(A,B|\delta) \]
\[ \log L(T|\Delta) = \sum_i \log L(A,B|\delta) \]

And so, the objective is to determine a set of 16bp tags T that maximizes the likelihood of observing the ditag sequences \( \Delta \).

### 2.3.5.2 The problem of global optimization

Although it is theoretically possible to find the most likely solution by testing all possible T, this approach is not practical. Consider a very small \( \Delta \) of 50 ditags. For each of the 100 tags, the two additional nucleotides can have one of 16 \( (4^2) \) possible sequences. Therefore, the solution space would consist of 100\(^{16} \) \( (1\times10^{12}) \) possible solutions. Raising the number of ditags one order of magnitude (500 ditags) results in a 16 order of magnitude increase in the size of the solution space \( (1\times10^{48}) \). Problems that grow rapidly in this fashion are said to undergo a “combinatorial explosion”, making a complete search of the solution space for any reasonably sized realization of the problem intractable. In the parlance of computational complexity theory, a brute force algorithm would run in \( O(c^n) \) time. In this example, if one assumes that calculating a single likelihood is exceedingly fast (1 calculation/picosecond), the search would take about \( 3\times10^{12} \) years.

### 2.3.5.3 Simulated annealing

One approach to solving T without requiring an exhaustive search would be to modify each tag one at a time and accept the solution if the result was an increase in \( L(T|\Delta) \). This would effectively make the running time linear, or \( O(n) \). However, this strategy runs the danger of getting trapped in local maxima due to the interdependency of different tags in calculating the likelihood. For example, a change in the additional nucleotides of a tag in the set of tags \( A \) having the same starting 14bp will affect \( L(A,B|\delta_{ab}) \), \( L(A,C|\delta_{ac}) \), \( L(A,D|\delta_{ad}) \), and so on. A

---

1 Big O notation is used to describe an algorithm’s resource usage (time in this case) in relation to the size of the input data. For example: \( O(1) \) is constant time, \( O(n) \) is linear time, and \( O(c^n) \) is exponential time.

2 The current age of the universe is estimated to be about \( 1.4\times10^{10} \) years (Hinshaw, 2009).
subsequent change to the additional nucleotides of a tag in, say, the set of tags $B$ may alter the likelihood $L(A, B|\delta_{ab})$ to such an extent that the previous solution for $A$ no longer provides optimal $L(T|\Delta)$.

This problem can be addressed by a simulated annealing (SA) strategy. This is a computational approach for finding a solution that is a good approximation of the global optimum given a large search space (Kirkpatrick, 1983). The rationale of SA is inspired by the metallurgical annealing process, where a material is subjected to high heat and then cooled slowly to remove defects. SA relies on an acceptance function, which utilizes a temperate parameter $t$. SA starts with a random solution to the problem. At the beginning of the algorithm, when $t$ is highest, the selection of a nearby solution is nearly random, while at lowest $t$, when the algorithm terminates, the chosen nearby solution will have the best fitness. By slowly decreasing $t$, the risk of getting trapped in local extrema is reduced since there is ample opportunity for escape at higher temperatures. SA is not a deterministic algorithm, and cannot guarantee an optimal solution; however, it can provide an excellent approximation in a reasonable amount of time.

2.3.5.4 Simple reductions to the solution space and the use of an “unknown” nucleotide

Two simple assumptions can be imposed to reduce the size of the solution space that must be evaluated. First, longer ditags can reduce the number of possible solutions since the extra nucleotides can be assigned to two contributing tags with certainty. 32bp ditags must arise from two 16bp tags, so both extra nucleotides are known; and 31bp ditags must arise from two tags that are at least 15bp, so one extra nucleotide is known. Second, we can ignore any solution for a given tag where there is no supporting evidence from other tags with the same 14bp starting sequence. For example, if there is a set of three tags where we have proposed that two of these have extra nucleotides AT and AG based on the ditag evidence, then the extra nucleotides of the
Finally, we can define an additional nucleotide which represents an “unknown” (denoted by an X in the program’s output). An “unknown” nucleotide is assigned when supporting evidence is unavailable and any of the four possible nucleotides are equally likely. Specifically, an “unknown” 1st extra nucleotide is assigned if, and only if, no evidence of any kind is available. Either all of the tags of a sequence must be unknown at the first position, or none at all. For example, a solution of {XX, XX, XX} is acceptable, but {AA, XX, XX} is not. An unknown 2nd extra nucleotide is similar, except this restriction only applies with respect to the subset of tags with the same 1st extra nucleotide. For example, a solution of {AA, AA, CX, CX} is acceptable, but {AA, AA, CT, CX} is not.

2.3.5.5 Algorithm summary

The XBP-SAGE algorithm proceeds as follows:

Step 1

An initial solution is created by randomly choosing the ligation point for each ditag. Any missing nucleotides are filled in, if possible, according to the rules described in Section 2.3.5.4. Two global likelihood values are generated: 1) the preliminary global likelihood, $L_p$, based only on the sequence arising from cutting the ditag at a proposed ligation point; and the full global likelihood, $L_f$, based on the additional filled-in sequence. The initial annealing temperature $t$ is set to 100.

Step 2

Proceed through the set of ditags in a random order, and for each ditag calculate $L_p$ for each possible ligation point (this is the set of nearby solutions). A choice of which ligation point to proceed with is made using $t$. Let $\Lambda$ be the set of $L_p$ for each possible ligation point $\lambda_1$, $\lambda_2$, ...
\( \lambda_n \). Calculate a modifier value:

\[
\mu = \left\lfloor \frac{\text{abs}(t-100)}{10} \right\rfloor + 1
\]

Alter the likelihoods such that:

\[
\lambda_i' = \exp\left\{ \lambda_i - \min(\Lambda) \right\}^\mu
\]

This procedure ensures that the ligation point that results in the highest likelihood will always be chosen when \( t=0 \). At the start of the algorithm, when \( t=100 \), the choice is a stochastic process influenced by the values of \( \Lambda \). For example, if \( \lambda_1 \) is twice the size of \( \lambda_2 \), then the ligation point corresponding to \( \lambda_1 \) will be chosen two-thirds of the time.

**Step 3**

In a randomly assigned order, generate additional nucleotides for tags where required (i.e. to avoid violating the assumptions described in Section 2.3.5.4) or change existing additional nucleotides if possible, and recalculate \( L_f \). If \( L_f \) does not improve and the previous extra nucleotide assignments are still valid, then discard the changes.

**Step 4**

Decrease the annealing temperature by 10. If \( t < 0 \) then accept the current solution and terminate the algorithm, otherwise return to Step 2.

**2.3.5.6 Performance on a test dataset**

To assess the performance of the XBP-SAGE algorithm, a publicly available normal human cervical epithelium LongSAGE library was used (GEO accession: GSM144012) (Barrett, 2007). LongSAGE tags are 21bp in length (including the anchoring enzyme site) and 100,000 were randomly selected and shortened to 16bp to create a SAGE library with an extra two nucleotides for each tag. This creates a test set with a realistic distribution of counts and sequence differences and known extra nucleotides. Each tag was then trimmed based on the following probabilities: \( P(l_i=14)=0.05, P(l_i=15)=0.65 \), and \( P(l_i=16)=0.30 \). Finally, the tags were...
randomly joined to create a simulated set of 50,000 ditags. The genetic algorithm (Section 2.3.1) to estimate values for $P(l_t)$ from the ditag lengths yielded excellent agreement with values of: $P(l_t=14)=0.0499$, $P(l_t=15)=0.650$, $P(l_t=16)=0.300$.

The data was subjected to the XBP-SAGE algorithm which completed in 2.5 hours on a single Pentium 4 2.4GHz CPU. 10,911 tag types (12,713 tag counts) were assigned no additional nucleotides, 12,921 tag types (26,584 tag counts) were assigned one additional nucleotide, and 5,777 tag types (60,703 tag counts) were assigned two additional nucleotides. Thus, it can be generally stated for this dataset that a tag with a count $>2$ will reveal one additional nucleotide, and a tag with a count $>9$ will reveal two additional nucleotides. The extra nucleotides were compared to the actual data and 98,472/100,000 (98.5%) were assigned correctly (Figure 2.3). When an incorrect assignment did occur, the tag was expressed at low levels ($<10$ counts) and was usually a low-expressed member of a highly-expressed group of tags with the same starting 14bp. For example:

<table>
<thead>
<tr>
<th>sequence</th>
<th>actual count</th>
<th>estimated count</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCAAGGTGGCC</td>
<td>44</td>
<td>38</td>
</tr>
<tr>
<td>CCAAGGTGGCCT</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

However, these situations were rare and, in general, the algorithm identified the breakdown of counts closely. For example:

<table>
<thead>
<tr>
<th>sequence</th>
<th>actual count</th>
<th>estimated count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTTTTTCAAAA</td>
<td>414</td>
<td>412</td>
</tr>
<tr>
<td>ACTTTTTCAAG</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>ACTTTTTCAACA</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ACTTTTTCAAG</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

### 2.3.5.7 Improvements to tag to gene mapping on real data

The XBP-SAGE algorithm was run an unmodified SAGE library generated from colonic epithelium (GEO accession: GSM728) (Velculescu, 1995; Barrett, 2007). This library consisted of 25,570 ditags. The genetic algorithm (Section 2.3.1) to estimate values for $P(l_t)$ from the ditag
Figure 2.3: Comparison of the observed and expected count of simulated tags with two additional nucleotides. Tags assigned zero, one, and two additional nucleotides are separated into three plots. The top three plots are scatterplots with the expected (known) count on the x-axis and the observed (estimated) count on the y-axis. The bottom three plots are density plots corresponding to the three scatterplots. These are provided to convey the number of datapoints at each x-y coordinates, as more than one tag can have the same expected and observed count. A dot in the density plot represents a case where a single observation is present at a given coordinate.
lengths yielded values of: \( P(l_t=14)=0.05 \), \( P(l_t=15)=0.67 \), \( P(l_t=16)=0.28 \). The consensus tag list contained 788 14bp tag types (815 tag counts), 9,646 15bp tag types (14,275 tag counts), and 7865 16bp tag types (36,048 tag counts). The success of tag to gene mapping was estimated using SAGE Genie (Boon, 2002) and was compared between the tags for which additional nucleotides were obtained and their canonical 14bp version (Table 2.2). To ameliorate the effect of spurious artefacts (e.g. PCR amplification or sequencing errors) tags which were observed only once were discarded. Of those tags that mapped ambiguously to a gene, 50% and 64% were no longer ambiguous with the addition of one or two extra nucleotides, respectively.

Interestingly, some tags that were ambiguous or unambiguous at 14bp became unmapped with the addition of two nucleotides (40/459 ambiguous tags and 47/409 unambiguous tags). Ostensibly, these tags represent rare cases where XBP-SAGE assigned an additional nucleotide incorrectly. Tags with counts \( \geq 5 \) were investigated in detail. Using a meta-list of all public LongSAGE libraries where each tag was shortened to 16bp, there was strong evidence (i.e. the 16bp tag was the highest expressed member of the group of tags with same starting 14bp) for correctness in 5/12 previously ambiguous tags and 13/21 previously unambiguous tags. However, when a comprehensive mapping was performed (see Section 2.2.3) an XBP-SAGE miscall was supported in only 3/12 previously ambiguous tags (Table 2.3) and 5/21 previously unambiguous tags (Table 2.4). The discrepancy is predominantly the result of sequencing artefacts. Of the previously ambiguous tags, assignments were made to the mitochondrial genome (4), PCR amplification artefacts (3), an antisense artefact (1), and an uncharacterized splice variant supported by EST evidence (1) (Table 2.3). Of the previously unambiguous tags, assignments were made to the mitochondrial genome (5), PCR amplification artefacts (3), antisense artefacts (2), probable antisense transcripts (2), an artefact of resulting from tag capture near an internal polyadenylation sequence (1), and uncharacterized splice variants supported by EST evidence (3) (Table 2.4).
### Table 2.2: Improvement of tag to gene mapping with addition of extra nucleotides

<table>
<thead>
<tr>
<th>Tag Length = 16bp</th>
<th>14bp</th>
<th>16bp</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td>unmapped</td>
<td>80</td>
<td>167</td>
<td>+108.7%</td>
</tr>
<tr>
<td>ambiguous</td>
<td>459</td>
<td>165</td>
<td>-64.0%</td>
</tr>
<tr>
<td>unambiguous</td>
<td>409</td>
<td>616</td>
<td>+50.6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tag Length = 15bp</th>
<th>14bp</th>
<th>15bp</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td>unmapped</td>
<td>299</td>
<td>433</td>
<td>+44.8%</td>
</tr>
<tr>
<td>ambiguous</td>
<td>1069</td>
<td>534</td>
<td>-50.0%</td>
</tr>
<tr>
<td>unambiguous</td>
<td>1042</td>
<td>1443</td>
<td>+38.5%</td>
</tr>
</tbody>
</table>

Data was compiled from the results of the XBP-SAGE algorithm run on a public SAGE library from colonic epithelium (GEO accession: GSM728). The table includes tags where an additional one or two nucleotides were assigned and where the tag count exceeds one.
Table 2.3: Putative source of 14bp tags with ambiguous mappings that no longer map when lengthened to 16bp

<table>
<thead>
<tr>
<th>Tag</th>
<th>Count</th>
<th>LongSAGE</th>
<th>Mapping Before</th>
<th>Mapping After</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATTTGAGAAGCC</td>
<td>333</td>
<td>yes</td>
<td>CT: RAD23B</td>
<td>mitochondrial genome</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TG: Hs.656343</td>
<td></td>
</tr>
<tr>
<td>GGGGCAGGGCCC</td>
<td>46</td>
<td>yes</td>
<td>AG: LOC729591</td>
<td>EIF5A antisense artefact</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AT: GFAP</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CA: CRELD1</td>
<td></td>
</tr>
<tr>
<td>GGGAAGCAGATT</td>
<td>32</td>
<td>yes</td>
<td>GG: F11R</td>
<td>mitochondrial genome</td>
</tr>
<tr>
<td>GGGGTTCAGGGGT</td>
<td>22</td>
<td>yes</td>
<td>AC: Hs.605638</td>
<td>mitochondrial genome</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TG: PYGB</td>
<td></td>
</tr>
<tr>
<td>CGCTGCTCCAC</td>
<td>21</td>
<td>no</td>
<td>AG: RPL11</td>
<td>unknown (probable miscall)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GG: MYO1G</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TT: LTBP3</td>
<td></td>
</tr>
<tr>
<td>GCCAACCTCTCTA</td>
<td>20</td>
<td>yes</td>
<td>AC: LOC286058</td>
<td>mitochondrial genome</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AG: EP400NL</td>
<td></td>
</tr>
<tr>
<td>AGCCACCGCGTA</td>
<td>11</td>
<td>no</td>
<td>31 different genes</td>
<td>unknown (probable miscall)</td>
</tr>
<tr>
<td>GCCATCTCCAG</td>
<td>9</td>
<td>no</td>
<td>TG: SLC39A13</td>
<td>maps to ESTs from Hs.591502, possible splice variant</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TT: RNF26</td>
<td></td>
</tr>
<tr>
<td>CTACTGACTCCG</td>
<td>9</td>
<td>no</td>
<td>7 different genes</td>
<td>artefact of CCACTGC (353 tags)</td>
</tr>
<tr>
<td>GTGAAACTCTGC</td>
<td>8</td>
<td>no</td>
<td>9 different genes</td>
<td>artefact of GTGAAACCCT (190 tags)</td>
</tr>
<tr>
<td>ATGGAGACTTCG</td>
<td>6</td>
<td>no</td>
<td>CA: CS</td>
<td>unknown (probable miscall)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GG: Hs.626951</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TG: ATP2B1</td>
<td></td>
</tr>
<tr>
<td>CCTGTAATACCG</td>
<td>6</td>
<td>no</td>
<td>16 different genes</td>
<td>artefact of CCTGTAAT^CC (472 tags)</td>
</tr>
</tbody>
</table>

Columns correspond to the elongated tag sequence, the observed count, whether the elongated sequence is supported by LongSAGE data, and the putative mapping(s) before and after the addition of extra nucleotides.
Table 2.4: Putative source of 14bp tags with unambiguous mappings that no longer map when lengthened to 16bp

<table>
<thead>
<tr>
<th>Tag</th>
<th>Count</th>
<th>LongSAGE</th>
<th>Mapping Before</th>
<th>Mapping After</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAACATTTCTCC</td>
<td>170</td>
<td>yes AC: Hs.707482 mitochondrial genome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTCAGGATACT</td>
<td>114</td>
<td>yes GA: TGFA mitochondrial genome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGGTGGCAAGAA</td>
<td>56</td>
<td>yes GG: LOC644075 mitochondrial genome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTCCACCCGA AA</td>
<td>48</td>
<td>yes GG: TFF3 possible variant of TFF3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATTCTGAATAA</td>
<td>48</td>
<td>yes AC: Hs.689535 mitochondrial genome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACAGGGTGACCT</td>
<td>26</td>
<td>no CC: EDF1 unknown (probable miscall)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCCATCCCTTA</td>
<td>25</td>
<td>yes CC: Hs.622702 mitochondrial genome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGAACCTTCAG</td>
<td>21</td>
<td>yes AA: HLA-A possible variant of HLA-A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACAAAAACTAGG</td>
<td>20</td>
<td>no GC: Hs.703814 unknown (probable miscall)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AATCACAATAAA</td>
<td>17</td>
<td>yes AC: CTAGE5 mispredicted 3’-UTR of CEACAM7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACAAAACTCCAC</td>
<td>15</td>
<td>yes TC: Hs.704518 ATP1B1 antisense transcript</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCCCTGCAGGG</td>
<td>14</td>
<td>no GA: SIRT6 unknown (probable miscall)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATGATGGCACCT</td>
<td>12</td>
<td>yes TA: LOC387647 TSPAN8 antisense transcript</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AATGAGAAGGTA</td>
<td>11</td>
<td>yes AA: PAF1 B4GALT1 internal poly-A tract</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCCCTATTAAGC</td>
<td>8</td>
<td>no TT: NOM1 artefact linker TCCCTATTAAGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTACCGCAGCA</td>
<td>7</td>
<td>no CC: GLTP unknown (probable miscall)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCTGGTTTGTCT</td>
<td>7</td>
<td>yes GG: WDR82 TMSB10 antisense transcript</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TGGCTAATGATG</td>
<td>5</td>
<td>yes CG: KPNR2 MALL antisense artefact</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCGTGGCTGGCT</td>
<td>5</td>
<td>no TC: RDH5 unknown (probable miscall)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATTTGAGAACCCT</td>
<td>5</td>
<td>no AT: REV1 artefact of ATTTGAGAACCCT (333 tags)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCACAGGTCAAC</td>
<td>5</td>
<td>no CA: EFCAB1 artefact of GCCCAAGTCA (547 tags)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Columns correspond to the elongated tag sequence, the observed count, whether the elongated sequence is supported by LongSAGE data, and the putative mapping(s) before and after the addition of extra nucleotides.
There were 80 tags that were initially unmapped. A comprehensive mapping was attempted for those with counts $\geq 5$. The additional nucleotides were supported by existing LongSAGE data in 17/25 cases. In the majority of cases (20/25), a source of the tag could be determined. Assignments were made to the mitochondrial genome (9), PCR amplification artefacts (5), an antisense artefact (1), an antisense transcript (1), a second position tag likely due to an incomplete AE restriction digest (1), and several uncharacterized splice variants supported by EST evidence (3) (Table 2.5).
## Table 2.5: Putative source of 14bp tags that fail to map

<table>
<thead>
<tr>
<th>Tag</th>
<th>Count</th>
<th>LongSAGE</th>
<th>Mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTCATAAGGAAA</td>
<td>107</td>
<td>yes</td>
<td>mitochondrial genome</td>
</tr>
<tr>
<td>TTTAACGGCCGC</td>
<td>86</td>
<td>yes</td>
<td>mitochondrial genome</td>
</tr>
<tr>
<td>AGACCCCAACACA</td>
<td>52</td>
<td>yes</td>
<td>mitochondrial genome</td>
</tr>
<tr>
<td>GTAAGTGACTCG</td>
<td>47</td>
<td>yes</td>
<td>mitochondrial genome</td>
</tr>
<tr>
<td>TCCGTACATCGA</td>
<td>38</td>
<td>no</td>
<td>artefact linker TCCCCGTACA</td>
</tr>
<tr>
<td>ACTTTCCAAAAA</td>
<td>32</td>
<td>yes</td>
<td>mitochondrial genome</td>
</tr>
<tr>
<td>GCTAGGTTTATA</td>
<td>28</td>
<td>yes</td>
<td>mitochondrial genome</td>
</tr>
<tr>
<td>CCTACAAGCAAG</td>
<td>21</td>
<td>yes</td>
<td>mitochondrial genome</td>
</tr>
<tr>
<td>CCGTGCTGCCAG</td>
<td>14</td>
<td>no</td>
<td>EP400NL (probable expressed pseudogene)</td>
</tr>
<tr>
<td>TCCCTATAAGCC</td>
<td>13</td>
<td>no</td>
<td>artefact linker TCCCTATTA</td>
</tr>
<tr>
<td>TCACCCACACCA</td>
<td>12</td>
<td>yes</td>
<td>RPL23 antisense transcript</td>
</tr>
<tr>
<td>ACCCCCTACAGAG</td>
<td>8</td>
<td>yes</td>
<td>mitochondrial genome</td>
</tr>
<tr>
<td>TTCTTGTGCGCC</td>
<td>8</td>
<td>yes</td>
<td>RPS11 antisense artefact</td>
</tr>
<tr>
<td>AACATCCCTATC</td>
<td>7</td>
<td>yes</td>
<td>mitochondrial genome</td>
</tr>
<tr>
<td>GCCGGAGGGCCC</td>
<td>6</td>
<td>no</td>
<td>unknown</td>
</tr>
<tr>
<td>TCACCTGACTAC</td>
<td>6</td>
<td>no</td>
<td>artefact linker TCCCCGTACA</td>
</tr>
<tr>
<td>GCCCCTTTTCC</td>
<td>6</td>
<td>no</td>
<td>unknown</td>
</tr>
<tr>
<td>GGGGTCCCAATTC</td>
<td>5</td>
<td>no</td>
<td>unknown</td>
</tr>
<tr>
<td>CGGAACACCGTG</td>
<td>5</td>
<td>yes</td>
<td>EZR splice variant</td>
</tr>
<tr>
<td>GAGGCCCCCTCAC</td>
<td>5</td>
<td>yes</td>
<td>unknown</td>
</tr>
<tr>
<td>CTAACTAGTTAC</td>
<td>5</td>
<td>yes</td>
<td>single genome match, predicted ncRNA (EnsEMBL:AL163011.3)</td>
</tr>
<tr>
<td>AGCTGTCGCCAC</td>
<td>5</td>
<td>yes</td>
<td>COX2 second position tag</td>
</tr>
<tr>
<td>CTGAAAAAAAA</td>
<td>5</td>
<td>yes</td>
<td>unknown</td>
</tr>
<tr>
<td>TCCCTATTGGGCGC</td>
<td>5</td>
<td>no</td>
<td>artefact linker TCCCTATTA</td>
</tr>
<tr>
<td>CCCCCCTGCATCA</td>
<td>5</td>
<td>yes</td>
<td>artefact of CCCCCCTGGAT (140 tags)</td>
</tr>
</tbody>
</table>

Columns correspond to the elongated tag sequence, the observed count, whether the elongated sequence is supported by LongSAGE data, and the putative mapping after the addition of extra nucleotides.
2.4 DISCUSSION

Identifying the correct source of a tag is the most crucial element of SAGE analysis. The power of the technique is diminished if a tag is assigned incorrectly or cannot be confidently determined. This chapter demonstrates that the tagging enzyme usually releases a tag one or two nucleotides longer than the 14bp assumed during standard data processing. Although sequence has an effect on the size of the tag released, this accounts for a very small amount of the variation observed. The determination of extra nucleotides is complicated by the ambiguity introduced by the ligation of two tags head-to-head to form ditags, a procedure critical to control for PCR amplification bias. The XBP-SAGE algorithm is designed to model these effects and determine a high-confidence list of extended tags. As a result, cases where a tag matches to more than one transcript or gene are reduced, artefacts are easier to identify, and the source of novel tags can be made with increased sensitivity (e.g. using a whole-genome search).

Although the algorithm is an estimate of the maximum likelihood solution for the additional nucleotides, the performance of XBP-SAGE on a test dataset suggests the typical solution is of excellent quality (error rate <3%) and the small number of errors that do occur usually correspond to tags with very low counts. These would typically be of little interest in a SAGE study. Moreover, the enhanced mapping on a real dataset demonstrates that these errors are far outweighed by the ambiguities that are resolved and the correction of erroneous mappings.

The use of the XBP-SAGE algorithm will increase the overall accuracy of a list of genes identified by SAGE, reducing the error introduced in downstream, large-scale bioinformatic analyses (e.g. GO enrichment, pathway analysis, etc.). The algorithm is also helpful in confirming the identity of high-confidence targets identified by statistical analyses before expending resources on further experiments.
BIBLIOGRAPHY


CHAPTER III

STATISTICAL INFERENCE FROM SAGE USING A POISSON MIXTURE MODEL†

3.1 INTRODUCTION

As a counting technology, SAGE produces profiles consisting of a digital output that is quantitative in nature. For example, a statement can be made with reasonable certainty that a SAGE tag observed 30 times in a library of 100,000 tags corresponds to a transcript that comprises 0.03% of the total transcriptome; the same statement cannot be made reliably with analog values, like those obtained from a microarray. Accordingly, a reliable statistical model should account for the discrete, count-based nature of SAGE observations. Statistical tests that incorporate a continuous probability distribution (e.g. the normal distribution assumed by Student’s \( t \)-test) are not appropriate. These tests require tag counts be normalized by division with the total library size to convert the data to the same continuous scale, discarding a statistically informative facet of the data.

The sampling of SAGE tags can be modeled by the binomial distribution that describes the probability of observing a number of successes in a series of Bernoulli trials. Here, the library size corresponds to the number of trials and the count of a particular tag is the number of successful trial outcomes. As the number of trials increases, the binomial distribution approaches the Poisson distribution. This is the case for SAGE (since the tag counts are small relative to a large library size), so the form of the Poisson and binomial distribution is essentially the same. A fortunate characteristic of both of these distributions is they are a function of a single parameter only, since the variance in the observed data is directly calculable from the mean.

However, in practice, the variance of SAGE data is often larger than can be explained by sampling alone. Several authors have attributed this effect, termed “overdispersion”, to a latent biological variability (Baggerly, 2003; Baggerly, 2004; Lu, 2005). Baggerly et al. referred to this as “between”-library variability, as opposed to “within”-library variability caused by sampling (Baggerly, 2003). Factors that could contribute to this variability are numerous. For example: sample preparation or quality, artefacts intrinsic to the library construction protocol, differences in gene transcription due to environment, and the intrinsic stability or regulatory complexity of transcription at a particular locus. This will adversely affect statistical analysis because additional variance results in an overstated significance. Procedures for using hierarchical models which incorporate a continuous prior distribution to model the excess variance have been presented for both the binomial (viz. beta-binomial using logistic regression (Baggerly, 2004) or \(t_w\)-test (Baggerly, 2003)) and Poisson (viz. negative binomial a.k.a. hierarchical gamma-Poisson using log-linear regression (Lu, 2005)) distributions. Attempts to use the log-normal and inverse-Gaussian as prior distributions (both of these have longer tails) did not show an appreciable improvement and are computationally difficult to fit (data not shown).

Here, it is argued that the excess variation is due to a mixing of two or more distinct Poisson (or binomial) components, and this mixing is the predominant source of total variation. This assumption corresponds to a finite mixture model, which has found wide applicability in several fields (for a general introduction, see McLachlan and Peel, 2000). To illustrate, consider a tag from ten SAGE libraries of equal size (e.g. 100,000 tags) that has observed counts where half are realizations of an expression of 0.0003 and the other half of 0.0004. As a result, the probability distribution of observing a particular tag count will be a combination of these two components (Figure 3.1). Note the similarity between the shapes of the probability distributions estimated from a fitted negative binomial (which assumes sampling variability drawn from a latent biological variability) and a Poisson mixture model (which assumes a set of independent
Figure 3.1: Probability density of several models applied to data generated from two Poisson components. 10 observations were randomly drawn from each of two Poisson distributions, one with a mean of 30, the other 40. The values drawn from the first component were [ 40, 34, 37, 28, 31, 21, 41, 27, 34, 27 ] and the values drawn from the second component were [ 36, 42, 26, 57, 43, 37, 38, 39, 35, 35 ]. The probability densities are shown for a single Poisson distribution, the negative binomial distribution, and a two-component Poisson mixture distribution using maximum likelihood estimates (see Methods). The probability densities of the individual Poisson components from which the data were actually drawn are also shown. The individual observations are represented by triangles at the bottom of the plot.
components, each having sampling variability only).

If the Poisson mixture model is an accurate foundation to explain SAGE observations, it is attractive for several reasons. First, this approach does not rely on a vague and continuous prior distribution to explain additional variance. Rather, the model asserts that a gene’s expression level will take on one of a number of distinct states. Second, overdispersed models applied to SAGE data tend to show a wide range of excess variation; in many cases, the excess is far greater than can be attributed to counting. This is a troubling prospect for studies that utilize a limited number of libraries (e.g. pair-wise comparisons), since the observed count may differ wildly from the underlying expression. If a mixture model provides an improved fit to SAGE data, this concern would be assuaged. Finally, mixture models, by nature, allow for the concept of subsets (or latent classes) in the expression values of each tag. Dysregulation of genes in disease processes such as cancer are often observed in only a proportion of profiled samples, and these will be naturally identified during model fitting. This property can also be utilized to identify sets of co-expressed genes.
3.2 MATERIALS AND METHODS

3.2.1 Test datasets

Test datasets were obtained from the Gene Expression Omnibus (GEO) (Barrett, 2007) and reflect a range of cancer studies, including malignancies of the skin (Cornelissen, 2003; van Ruissen, 2002; Weeraratna, 2004), breast (Porter, 2003; Porter 2001), blood (Lee, 2006) and brain (Boon, 2004) (Table 3.1). In the case of breast and skin data, libraries from a combination of studies were used. Datasets were filtered to remove tags expressed at a mean less than 100 tags per million.

3.2.2 Model fitting

The open-source statistical software package R was used to perform all calculations (R Core Development Team, 2007). For each of the models, let $Y_i$ be the observed tag count in library $i$, $n_i$ be the total number of tags in library $i$, and $N$ be the total number of libraries. Also, let $x_i$ be the vector of explanatory variables (e.g. normal=0 and cancer=1) associated with library $i$, and $\beta$ be the vector of coefficients. Example R code used to fit each model is included in Appendix II.
Table 3.1: Datasets used to evaluate models

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accession</th>
<th>Library Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>acute myeloid leukemia (AML) purified patient samples</td>
<td>GSM73364</td>
<td>100003</td>
<td>inv(16)</td>
</tr>
<tr>
<td></td>
<td>GSM73365</td>
<td>46329</td>
<td>inv(16)</td>
</tr>
<tr>
<td></td>
<td>GSM73366</td>
<td>83577</td>
<td>inv(16)</td>
</tr>
<tr>
<td></td>
<td>GSM73367</td>
<td>77604</td>
<td>inv(16)</td>
</tr>
<tr>
<td></td>
<td>GSM73368</td>
<td>50604</td>
<td>inv(16)</td>
</tr>
<tr>
<td></td>
<td>GSM73369</td>
<td>80131</td>
<td>t(8;21)</td>
</tr>
<tr>
<td></td>
<td>GSM73370</td>
<td>48397</td>
<td>t(8;21)</td>
</tr>
<tr>
<td></td>
<td>GSM73371</td>
<td>30067</td>
<td>t(8;21)</td>
</tr>
<tr>
<td></td>
<td>GSM73372</td>
<td>68267</td>
<td>t(8;21)</td>
</tr>
<tr>
<td></td>
<td>GSM73373</td>
<td>48278</td>
<td>t(8;21)</td>
</tr>
<tr>
<td></td>
<td>GSM73374</td>
<td>54669</td>
<td>t(9;11)</td>
</tr>
<tr>
<td></td>
<td>GSM73375</td>
<td>67803</td>
<td>t(9;11)</td>
</tr>
<tr>
<td></td>
<td>GSM73376</td>
<td>55209</td>
<td>t(9;11)</td>
</tr>
<tr>
<td></td>
<td>GSM73377</td>
<td>51254</td>
<td>t(9;11)</td>
</tr>
<tr>
<td></td>
<td>GSM73378</td>
<td>83102</td>
<td>t(9;11)</td>
</tr>
<tr>
<td></td>
<td>GSM73379</td>
<td>20712</td>
<td>t(9;11)</td>
</tr>
<tr>
<td></td>
<td>GSM73380</td>
<td>84229</td>
<td>t(9;11)</td>
</tr>
<tr>
<td></td>
<td>GSM73381</td>
<td>72483</td>
<td>t(15;17)</td>
</tr>
<tr>
<td></td>
<td>GSM73382</td>
<td>48733</td>
<td>t(15;17)</td>
</tr>
<tr>
<td></td>
<td>GSM73383</td>
<td>48394</td>
<td>t(15;17)</td>
</tr>
<tr>
<td></td>
<td>GSM73384</td>
<td>66647</td>
<td>t(15;17)</td>
</tr>
<tr>
<td></td>
<td>GSM73385</td>
<td>57563</td>
<td>t(15;17)</td>
</tr>
<tr>
<td>developmental stages of breast cancer bulk tissue samples</td>
<td>GSM691</td>
<td>7165</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM692</td>
<td>12142</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM14756</td>
<td>58181</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM14801</td>
<td>59327</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM1731</td>
<td>43902</td>
<td>DCIS</td>
</tr>
<tr>
<td></td>
<td>GSM2389</td>
<td>58801</td>
<td>DCIS</td>
</tr>
<tr>
<td></td>
<td>GSM14800</td>
<td>50875</td>
<td>DCIS</td>
</tr>
<tr>
<td></td>
<td>GSM1733</td>
<td>70099</td>
<td>invasive¹</td>
</tr>
<tr>
<td></td>
<td>GSM670</td>
<td>40223</td>
<td>invasive</td>
</tr>
<tr>
<td></td>
<td>GSM672</td>
<td>67386</td>
<td>invasive</td>
</tr>
<tr>
<td></td>
<td>GSM2382</td>
<td>65045</td>
<td>invasive</td>
</tr>
<tr>
<td></td>
<td>GSM2383</td>
<td>61480</td>
<td>invasive</td>
</tr>
<tr>
<td></td>
<td>GSM14797</td>
<td>21951</td>
<td>invasive</td>
</tr>
<tr>
<td>melanoma bulk tissue samples</td>
<td>GSM11123</td>
<td>15167</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM11124</td>
<td>11563</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM11125</td>
<td>10080</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM3242</td>
<td>37292</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM14751</td>
<td>26032</td>
<td>cancer</td>
</tr>
<tr>
<td></td>
<td>GSM14775</td>
<td>41338</td>
<td>cancer</td>
</tr>
<tr>
<td></td>
<td>GSM14778</td>
<td>11399</td>
<td>cancer</td>
</tr>
<tr>
<td>brain tumour bulk tissue samples</td>
<td>GSM676</td>
<td>94876</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM695</td>
<td>58826</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM761</td>
<td>51280</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM763</td>
<td>63208</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM786</td>
<td>77968</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM7498</td>
<td>31538</td>
<td>normal</td>
</tr>
<tr>
<td></td>
<td>GSM14799</td>
<td>308589</td>
<td>normal²</td>
</tr>
<tr>
<td></td>
<td>GSM31931</td>
<td>41773</td>
<td>normal³</td>
</tr>
<tr>
<td></td>
<td>GSM31935</td>
<td>305546</td>
<td>normal⁴</td>
</tr>
<tr>
<td></td>
<td>GSM697</td>
<td>52479</td>
<td>astrocytoma</td>
</tr>
<tr>
<td></td>
<td>GSM698</td>
<td>77004</td>
<td>astrocytoma</td>
</tr>
<tr>
<td></td>
<td>GSM699</td>
<td>28159</td>
<td>astrocytoma</td>
</tr>
<tr>
<td></td>
<td>GSM715</td>
<td>17576</td>
<td>astrocytoma</td>
</tr>
<tr>
<td></td>
<td>GSM1732</td>
<td>81495</td>
<td>astrocytoma</td>
</tr>
<tr>
<td></td>
<td>GSM2443</td>
<td>80265</td>
<td>astrocytoma</td>
</tr>
<tr>
<td>Dataset</td>
<td>Accession</td>
<td>Library Size</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>brain tumour bulk tissue samples (continued)</td>
<td>GSM2451</td>
<td>38634</td>
<td>astrocytoma</td>
</tr>
<tr>
<td></td>
<td>GSM2578</td>
<td>69513</td>
<td>astrocytoma</td>
</tr>
<tr>
<td></td>
<td>GSM14737</td>
<td>105764</td>
<td>astrocytoma</td>
</tr>
<tr>
<td></td>
<td>GSM14739</td>
<td>88568</td>
<td>astrocytoma</td>
</tr>
<tr>
<td></td>
<td>GSM14763</td>
<td>106982</td>
<td>astrocytoma</td>
</tr>
<tr>
<td></td>
<td>GSM14765</td>
<td>102439</td>
<td>astrocytoma</td>
</tr>
<tr>
<td></td>
<td>GSM14766</td>
<td>107344</td>
<td>astrocytoma</td>
</tr>
<tr>
<td></td>
<td>GSM14773</td>
<td>118733</td>
<td>astrocytoma</td>
</tr>
<tr>
<td></td>
<td>GSM1497</td>
<td>46928</td>
<td>ependymoma</td>
</tr>
<tr>
<td></td>
<td>GSM1735</td>
<td>74499</td>
<td>ependymoma</td>
</tr>
<tr>
<td></td>
<td>GSM2384</td>
<td>52934</td>
<td>ependymoma</td>
</tr>
<tr>
<td></td>
<td>GSM2408</td>
<td>52659</td>
<td>ependymoma</td>
</tr>
<tr>
<td></td>
<td>GSM14740</td>
<td>122690</td>
<td>ependymoma</td>
</tr>
<tr>
<td></td>
<td>GSM14741</td>
<td>120431</td>
<td>ependymoma</td>
</tr>
<tr>
<td></td>
<td>GSM14762</td>
<td>68614</td>
<td>ependymoma</td>
</tr>
<tr>
<td></td>
<td>GSM14776</td>
<td>75379</td>
<td>ependymoma</td>
</tr>
<tr>
<td></td>
<td>GSM14786</td>
<td>84073</td>
<td>ependymoma</td>
</tr>
<tr>
<td></td>
<td>GSM793</td>
<td>56871</td>
<td>ependymoma</td>
</tr>
<tr>
<td></td>
<td>GSM14767</td>
<td>100600</td>
<td>glioblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14768</td>
<td>102322</td>
<td>glioblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14769</td>
<td>99099</td>
<td>glioblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM696</td>
<td>70087</td>
<td>glioblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM745</td>
<td>60069</td>
<td>glioblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM765</td>
<td>61886</td>
<td>glioblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM1498</td>
<td>62675</td>
<td>glioblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM690</td>
<td>38933</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM693</td>
<td>19572</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14731</td>
<td>52645</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14732</td>
<td>48451</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14733</td>
<td>43068</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14734</td>
<td>69971</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14761</td>
<td>85376</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14772</td>
<td>60454</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14774</td>
<td>85984</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14779</td>
<td>72318</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14781</td>
<td>83671</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14782</td>
<td>68392</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14783</td>
<td>61853</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14787</td>
<td>57469</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14788</td>
<td>74295</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14791</td>
<td>32570</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14794</td>
<td>74612</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM14795</td>
<td>67404</td>
<td>medulloblastoma</td>
</tr>
<tr>
<td></td>
<td>GSM689</td>
<td>28133</td>
<td>oligodendroglioma</td>
</tr>
<tr>
<td></td>
<td>GSM14742</td>
<td>32442</td>
<td>oligodendroglioma</td>
</tr>
</tbody>
</table>

Accessions, sizes, and descriptions of the libraries included in the study of the Poisson mixture model. Superscripts denote: 1. de novo translocation, 2. treatment induced translocation, 3. small DCIS component, 4. GSM31935 is a shortened version of LongSAGE library GSM14799 (not included in analysis), 5. shortened LongSAGE library.
3.2.2.1 Log-linear (Poisson) regression model

The log-linear model assumes that the observed tag counts are distributed as

$$Y_i \sim \text{Poisson}(\mu_i)$$

$$\mu_i = np_i$$

where $p_i$ is the actual expression in terms of the proportion of all expressed tags. Here, the unconditional mean and variance are $E(Y_i) = \text{Var}(Y_i) = \mu_i$. Using the log link function, which linearizes the relationship between the dependent variables and the predictors(s), we obtain the linear equation

$$\log(Y_i) = \log(n_i) + x_i \beta$$

$$p_i = \exp(x_i \beta)$$

Using iteratively reweighted least-squares (IRLS), the value(s) for the coefficient(s) $\beta$ are estimated. The *stats* library included with R is used to fit the log-linear model.

3.2.2.2 Overdispersed log-linear regression model

In contrast to a canonical log-linear model, we assume the actual expression is distributed as

$$\theta_i \sim \text{Gamma}(\mu_i, 1/\phi)$$

$$\mu_i = np_i$$

where, as in Section 3.2.2.1, $p_i$ is the actual expression in terms of the proportion of all expressed tags. Here, the unconditional mean and variance are $E(\theta_i) = \mu_i$ and $\text{Var}(\theta_i) = \mu_i^2 \phi$. Since we are now sampling from this latent Gamma distribution, the observed tag counts are conditional on this underlying expression and are distributed as

$$Y_i \mid p_i, \phi \sim \text{Poisson}(\theta_i)$$

Now, the unconditional mean and variance are $E(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \mu_i(1+\mu_i \phi)$. As above, using the log link function we obtain the linear equation
\[
\log(Y_i) = \log(n_i) + x_i \beta \\
p_i = \exp(x_i \beta)
\]

Here, a maximum likelihood estimate of the values for the coefficient(s) \( \beta \) and the overdispersion parameter (\( \phi \)) can be performed. The MASS library for R is used to fit the overdispersed log-linear model (Venables, 2002). A more complete discussion of this model and its application to SAGE, including significance testing, is described by Lu et al. (Lu, 2005).

### 3.2.2.3 Poisson mixture model

Like the canonical log-linear regression model, we assume the observed tag counts are Poisson distributed. However, the counts are conditional on the choice of a Poisson-distributed component, such that

\[
Y_i \mid k \sim \text{Poisson}(\mu_{ik})
\]

\[
\mu_{ik} = n_i p_{ik}
\]

where the component \( k = 1, 2, \ldots, K \) and \( p_{ik} \) is the actual expression for component \( k \) in terms of the proportion of all expressed tags. The posterior probability that an observed tag count belongs to a component \( k \) is given by

\[
P(k \mid Y_i, \psi) = \frac{\pi_k f(Y_i \mid \mu_{ik})}{\sum_k \pi_k f(Y_i \mid \mu_k)}
\]

where \( \psi \) is the parameter vector containing the component means (\( \theta_1, \ldots, \theta_K \)) and mixing coefficients (\( \pi_1, \ldots, \pi_{K-1} \)). \( f(.) \) is the Poisson probability density function. To fit the model, one must estimate the values in \( \psi \). This can be done by maximum likelihood estimation (MLE) using the EM algorithm (Dempster, 1977). The flexmix library for R uses the EM algorithm to fit a variety of finite mixture models (Leisch, 2004).
3.3 RESULTS

3.3.1 Goodness of fit

In order to evaluate the efficacy of a mixture model approach, a comparison of the goodness of fit of this and previously described models on 15 sets of biological replicates from publicly available SAGE data was performed (see Section 3.2.1). Goodness of fit was assessed for: 1) the canonical log-linear (Poisson) model, 2) negative binomial (i.e. hierarchical gamma-Poisson or overdispersed log-linear) model, and 3) $k$-component Poisson mixture model (see Section 3.2.2). Since maximum likelihood estimation (MLE) is used to fit each of these models, the log-likelihood was the basis for assessing relative goodness of fit. A comparison of the Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978) (both of which use the log-likelihood and a term to penalize a model for estimating a larger number of parameters) was performed on each of the datasets (Table 3.2).

As expected, the canonical Poisson model, which does not account for excess variance, performs poorly in all cases. The Poisson mixture model consistently outperforms the negative binomial model regardless of the metric used. The competitiveness of the negative binomial model is perhaps not surprising since a comparison of the fit of these two models to simulated data indicates that the negative binomial can often fit better to data generated from a two-component Poisson mixture. This becomes more problematic as the component means draw closer (data not shown, Figure 3.1 is a good example). However, several hypotheses can be tested to further strengthen the case for the mixture model approach. These are considered in turn.
Table 3.2: Comparison of model fits to a single group of biological replicates

<table>
<thead>
<tr>
<th></th>
<th>mean AIC</th>
<th>mean BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>tags</td>
</tr>
<tr>
<td><strong>BRAIN</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>astrocytoma</td>
<td>14</td>
<td>1141</td>
</tr>
<tr>
<td>ependymoma</td>
<td>10</td>
<td>1205</td>
</tr>
<tr>
<td>glioblastoma</td>
<td>7</td>
<td>1197</td>
</tr>
<tr>
<td>medulloblastoma</td>
<td>18</td>
<td>1045</td>
</tr>
<tr>
<td>normal</td>
<td>8</td>
<td>1099</td>
</tr>
<tr>
<td><strong>AML</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>inv(16)</td>
<td>5</td>
<td>900</td>
</tr>
<tr>
<td>t(8;21)</td>
<td>5</td>
<td>1037</td>
</tr>
<tr>
<td>t(15;17)</td>
<td>5</td>
<td>709</td>
</tr>
<tr>
<td>t(9;11) de novo</td>
<td>4</td>
<td>954</td>
</tr>
<tr>
<td>t(9;11) treatment</td>
<td>3</td>
<td>1061</td>
</tr>
<tr>
<td><strong>BREAST</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>normal</td>
<td>6</td>
<td>1259</td>
</tr>
<tr>
<td>DCIS</td>
<td>4</td>
<td>598</td>
</tr>
<tr>
<td>invasive</td>
<td>3</td>
<td>1069</td>
</tr>
<tr>
<td><strong>SKIN</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>normal</td>
<td>4</td>
<td>1015</td>
</tr>
<tr>
<td>melanoma</td>
<td>3</td>
<td>992</td>
</tr>
</tbody>
</table>

SAGE tag counts from fifteen sets of biological replicates were fit to log-linear (Poisson), negative binomial (overdispersed log-linear), and Poisson mixture models. The table contains the number of replicates (N), tags tested, and mean number of mixture components (k). For each model, mean goodness of fit scores calculated using Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) are shown. For both scores, a lower value indicates a better fit.
3.3.2 Tags with ambiguous mappings are represented by a greater number of components

Consider an idealized situation where the expression of a gene can take on one of two states (and can therefore be modelled by a two-component Poisson mixture). A significant proportion of SAGE tags are ambiguous (correspond to more than one gene) and, under the idealized model, would result in tag counts that are modelled by $2^g$ components (where $g$ is the number of expressed genes the tag corresponds to). Therefore, the number of components in the mixture should be higher for ambiguous tags.

Simply partitioning the data into ambiguous and unambiguous tags and comparing the number of components is unlikely to be informative since, for any given ambiguous tag, it is not known how many of the possible genes are actually expressed. However, two normal brain libraries used in this study were generated using LongSAGE (GSM31931 and GSM31935), which provides 17 base pairs of information rather than 10. The tag sequences in these libraries were shortened before inclusion in the normal brain dataset used in the previous section. However, by comparing the shortened tag list to the original library, tags that actually correspond to two or more LongSAGE tag sequences (and presumably represent different transcripts) were identified. Tags counts of one or two were considered artefacts of PCR amplification or sequencing and were not used in this determination.

The number of ambiguous and unambiguous tags was tallied for each estimated number of components (Table 3.3). Ambiguous tags are represented more highly in the set of model fits that consist of a larger number of components. This effect, which is statistically significant, is consistent with the mixture model hypothesis.
Table 3.3: Mean number of mixture model components

<table>
<thead>
<tr>
<th>Library</th>
<th>$k$</th>
<th>Unambiguous</th>
<th>Ambiguous</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM31931</td>
<td>1</td>
<td>93</td>
<td>0</td>
<td>$p&lt;2.2E-16$ ($\chi^2=134.1; df=4$)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>405</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>210</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>27</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>GSM31935</td>
<td>1</td>
<td>74</td>
<td>34</td>
<td>$p=1.8E-6$ ($\chi^2=32.1; df=4$)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>317</td>
<td>246</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>149</td>
<td>171</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>17</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

Expressed ambiguous and unambiguous 10 base pair tags for two LongSAGE libraries were distinguished based on the number of 17 base pair sequences that give rise to the same tag. The tags in each of these two groups were binned according to the number of estimated mixture components. The $\chi^2$ statistic was used to test the null hypothesis that these two groups are equivalent.
3.3.3 Component assignment of libraries is non-random

If the mixture model approach holds, then the Poisson components should cluster the libraries into recurring groups. Such an enrichment of certain component assignments would be expected for a number of reasons. Two possibilities are: a) if one or more libraries are mislabelled, the tag expression in those libraries should show a preferential assignment to a separate component; and b) if the genes corresponding to a set of tags are co-expressed, the component assignment should be similar amongst these genes. Conversely, if the negative binomial model is more appropriate then component assignments should essentially be random, since the distribution assumed to give rise to biological variability is continuous and unconditional.

For each of the datasets, the component assignments for tags where the estimated number of components is two were tallied. The individual assignment was based on the component with the highest posterior probability, given a tag count and library size. In all cases, there were a significant number of tags where the parent libraries were partitioned into the same two components (Table 3.4). For example, in the AML libraries containing the cytogenetic abnormality t(8;21), of the 225 tags that had expression that could be fit to two Poisson components, 110 were partitioned in the form -++-- \( (p=4.5E-67; \text{binomial test}) \). In other words, almost half of the tags that fit to two components were assigned to a single component configuration (for 5 libraries, \((2^5/2)-1=15\) such configurations are possible).

3.3.4 Determining differentially expressed genes

In previously described overdispersed models, the identity of a library is included \textit{a priori} as a model covariate. Significance is then determined by testing the null hypothesis that the fitted \( \beta \) coefficient for this covariate is zero (Baggerly, 2004; Lu, 2005). A Bayesian significance score has also been described, although this was developed using a beta-binomial model (Vencio,
## Table 3.4: Top component memberships

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Component assignment</th>
<th>Freq.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BRAIN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>astrocytoma (N=14 n_k=2=454)</td>
<td>-+-----------------</td>
<td>14</td>
<td>4.3E-28</td>
</tr>
<tr>
<td></td>
<td>++----------------++-</td>
<td>12</td>
<td>2.7E-23</td>
</tr>
<tr>
<td>ependymoma (N=10 n_k=2=544)</td>
<td>-+--------------</td>
<td>16</td>
<td>6.0E-14</td>
</tr>
<tr>
<td></td>
<td>++---------------</td>
<td>15</td>
<td>9.3E-13</td>
</tr>
<tr>
<td>glioblastoma (N=7 n_k=2=607)</td>
<td>-+-</td>
<td>48</td>
<td>3.2E-19</td>
</tr>
<tr>
<td></td>
<td>+--------------++</td>
<td>42</td>
<td>5.2E-15</td>
</tr>
<tr>
<td>medulloblastoma (N=18 n_k=2=438)</td>
<td>-+-+-+-+-+-++</td>
<td>4</td>
<td>1.5E-9</td>
</tr>
<tr>
<td></td>
<td>-+-------------++</td>
<td>4</td>
<td>1.5E-9</td>
</tr>
<tr>
<td>normal (N=9 n_k=2=588)</td>
<td>-++-</td>
<td>41</td>
<td>7.4E-37</td>
</tr>
<tr>
<td></td>
<td>++++-</td>
<td>21</td>
<td>7.6E-14</td>
</tr>
<tr>
<td><strong>AML</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>inv(16) (N=5 n_k=2=387)</td>
<td>-++++</td>
<td>110</td>
<td>2.5E-39</td>
</tr>
<tr>
<td></td>
<td>+++++</td>
<td>36</td>
<td>0.028</td>
</tr>
<tr>
<td>t(8;21) (N=5 n_k=2=387)</td>
<td>-++-</td>
<td>110</td>
<td>2.5E-39</td>
</tr>
<tr>
<td></td>
<td>+++-</td>
<td>36</td>
<td>0.028</td>
</tr>
<tr>
<td>t(15;17) (N=5 n_k=2=225)</td>
<td>-+-</td>
<td>110</td>
<td>4.5E-67</td>
</tr>
<tr>
<td></td>
<td>++-</td>
<td>36</td>
<td>1.0E-6</td>
</tr>
<tr>
<td>t(9;11) de novo (N=4 n_k=2=502)</td>
<td>-++-</td>
<td>143</td>
<td>1.6E-16</td>
</tr>
<tr>
<td></td>
<td>++-</td>
<td>132</td>
<td>1.4E-12</td>
</tr>
<tr>
<td>t(9;11) treated (N=3 n_k=2=405)</td>
<td>++-</td>
<td>216</td>
<td>1.1E-16</td>
</tr>
<tr>
<td><strong>BREAST</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>normal (N=6 n_k=2=571)</td>
<td>-+-+-</td>
<td>82</td>
<td>3.6E-29</td>
</tr>
<tr>
<td></td>
<td>---++</td>
<td>65</td>
<td>4.0E-18</td>
</tr>
<tr>
<td>DCIS (N=4 n_k=2=154)</td>
<td>-++</td>
<td>97</td>
<td>1.4E-43</td>
</tr>
<tr>
<td></td>
<td>++-</td>
<td>41</td>
<td>4.5E-5</td>
</tr>
<tr>
<td>invasive (N=3 n_k=2=765)</td>
<td>-+-</td>
<td>337</td>
<td>4.6E-10</td>
</tr>
<tr>
<td><strong>SKIN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>normal (N=4 n_k=2=500)</td>
<td>-++</td>
<td>215</td>
<td>1.8E-54</td>
</tr>
<tr>
<td></td>
<td>-+++</td>
<td>405</td>
<td>2.1E-51</td>
</tr>
</tbody>
</table>

For each set of biological replicates, the top one or two component states were selected from tags where the estimated number of components is 2. One component was represented with −, the other with + (i.e. −− is equivalent to +−). The significance was calculated using a zero-truncated binomial test. The number of possible ways for the libraries to be assigned to the two components is \((2^N/2)-1\), where N is the number of libraries.
In contrast, the Poisson mixture model does not require the identity of the libraries be included (although the addition of such covariates is possible). Rather, once a mixture model has been fit, the posterior probabilities of membership in a particular component given the observed tag counts can be used to determine how well the components can differentiate between two or more sample types (e.g. normal versus cancer). Here, a score is presented based on the confidence that a sample is of type $\omega$ given that it arises from component(s) $k$. Using Bayes Theorem, one can derive the following expression (the complete proof is included in Appendix III):

$$P(\omega | k) = \frac{\sum_{j \in \omega} \tau_{jk}}{N_\omega \pi_k}$$

where $\omega$ is the set of libraries corresponding to some label of interest (e.g. normal or cancer), $\tau_{jk}$ is the posterior probability of the tag count from library $j$ arising from component(s) $k$, and $\pi_k$ is the mixing coefficient for component(s) $k$. Using this expression, one can determine which tags have a set of mixture components that are closely linked with the sample type(s) of interest.

To illustrate, SAGE libraries from normal brain ($n=8$) and ependymoma ($n=10$) (a type of brain tumour) were analyzed using both the overdispersed log-linear and Poisson mixture models. In the former case, significance was calculated using the method described by Baggerly et al. (Baggerly, 2004) (see also example R code in Appendix II). In the latter case, the method described above was used. A plot of the two sets of scores shows a moderate correlation and tags that are found highly significant in one test tend to be so in the other (Figure 3.2).

However, a number of observations are found to be significant using the overdispersed log-linear model and not the Poisson mixture model, and vice versa. A closer look at the most extreme examples illustrates the superior performance of the mixture approach (Figure 3.3). In the first example, tag ACAACAAAGA seems clearly expressed in normal libraries, but is completely abolished in the ependymoma libraries. However, according to the overdispersed
Figure 3.2: Comparison to significance scores for a test of differential expression calculated using a negative binomial model. Using the tag counts from 8 normal brain libraries versus 10 ependymoma libraries, differential expression between these two sample types was assessed using two methods. Plotted are the significance scores calculated for a negative binomial model versus a Poisson mixture model. The negative binomial (x-axis) is a p-value, so smaller values are more significant. The Poisson mixture (y-axis) is a confidence score, so larger values are more significant. Circled are two examples of SAGE tags where one model shows significance while the other does not.
Figure 3.3: Counts for two tags assessed using a negative binomial model and the Poisson mixture model where one models shows significance and the other does not. The figure is divided to show separate plots of the expression level of two tags observed in 8 normal brain libraries and 10 ependymoma libraries. The x-axis is the normalized expression (count/library size*100,000) and the y-axis is divided into the two sample types. In the top plot, the negative binomial model is not significant and the Poisson mixture is significant; in the bottom plot, the situation is reversed. Light gray guide lines denote the expected expression level of the Poisson components.
model, the observation is not at all significant ($p=0.9998$). The mixture model, however, produces a confidence score of 99.42%, which suggests this tag is highly informative with respect to sample type. This example demonstrates the difficulty that the log-linear model has with fitting groups where tag counts are zero, a problem that is even more pronounced when using a logistic regression model (for a more thorough discussion of this problem see Lu et al. (Lu, 2005)).

In the second example, tag CAGTTGTGGT clearly has increased expression in some libraries from both the normal and ependymoma groups. However, according to the overdispersed model, the observation is highly significant ($p=8.8E-7$). The mixture model, however, produces a confidence score of 59.8% which is only nominally better than chance. This example demonstrates how the log-linear model seems to downweight the occasional extreme observation in one group, even if it is in agreement with observations in the other group. This can result in candidate lists based on the log-linear significance containing tags that have extreme observations that occur at a higher rate in one group over another, which are typically of little interest.

Similar results were obtained when comparing to the Bayes error rate described by Vencio et al. (Vencio, 2004). Again, a moderate correlation is seen and tags found highly significant in one test tend to be so in the other (Figure 3.4). Overall, the Bayes error rate is in better agreement with the mixture model confidence score and appears to be more robust in assessing tags with zero counts in one group. However, the assumption of a hierarchical model (in this case, a beta-binomial) used to calculate the Bayes error rate versus a Poisson mixture model results in differences between the two methods. Two examples, analogous to those described above, are highlighted (Figure 3.5). In both cases, the Poisson mixture model appears to give confidence values that are in better agreement with the observations.
Figure 3.4: Comparison to Bayes error rate for a test of differential expression calculated using a beta binomial model. Using the tag counts from 8 normal brain libraries versus 10 ependymoma libraries, differential expression between these two sample types was assessed using two methods. Plotted are the Bayes error rate described in Vencio et al. (2004) versus a Poisson mixture model confidence score. For the Bayes error rate (x-axis) smaller values are more significant. The Poisson mixture (y-axis) is a confidence score, so larger values are more significant. Circed are two examples of SAGE tags where one model shows significance while the other does not.
Figure 3.5: Counts for two tags assessed using a Bayes error rate and the Poisson mixture model where one models shows significance and the other does not. The figure is divided to show separate plots of the expression level of two tags observed in 8 normal brain libraries and 10 ependymoma libraries. The x-axis is the normalized expression (count/library size*100,000) and the y-axis is divided into the two sample types. In the top plot, the Bayes error rate is not significant and the Poisson mixture is significant; in the bottom plot, the situation is reversed. Light gray guide lines denote the expected expression level of the Poisson components.
3.4 DISCUSSION

The exploration of statistical approaches to SAGE analysis is important since the number of studies using the technology has resulted in a continuing rise in the amount of available data. The notion of sampling variability being the predominant source of “within”-library variability and distinct components being the predominant source of “between”-library variability is reassuring for investigators who choose the SAGE technique to obtain a comprehensive profile of gene expression in a limited number of samples. Nevertheless, there is certainly a contribution by a latent biological variability as evidenced by the increased performance of the negative binomial as the number of libraries increases. However, this work demonstrates that a simple overdispersed model may overstate this effect, and that certainly there is a clustering of expression into distinct components, which are then sampled. This is consistent with the view of gene transcription for any one locus consisting of (possibly several) inactivated or activated state(s). The same idea holds for some known mechanisms of genetic disease, such as loss of heterozygosity (LOH) or amplification of a particular locus (e.g. cancer).

For this reason, it is recommended that investigators try the mixture model approach in comparisons of groups of biological replicates. Failing this, some of the difficulties that can be encountered with the negative binomial model can be lessened by: a) setting a tolerance for how much overdispersion ($\varphi$) is acceptable in a final list of candidate tags, although such a cutoff would be somewhat arbitrary; and b) add a small value to the tag count to avoid the problems the model has with groups consisting of many zero counts. One strategy is to assume equal odds that the next tag drawn is the one of interest by adding 1 to the count, and 2 to the library size (i.e. (count+1)/(size+2)) (K. Baggerly, personal communication).

In the future, it may be worthwhile to combine both approaches by defining a negative binomial mixture model. However, at this point, such an approach is unlikely to provide
significant improvement given the small number of libraries in a typical set of available biological replicates. In addition, applying the concept of “information sharing” between tags may provide estimates of statistically informative variables that apply library-wide, and could be utilized to improve the power of the method described in this chapter (Kuznetsov, 2002; Thygesen, 2006).

The Poisson mixture model appears to be a rational means to represent SAGE data that are biological replicates and as a basis to assign significance when comparing multiple groups of such replicates. The use of a mixture model can improve the process of selecting differentially expressed genes, and provide a foundation for ab initio identification of co-expressed genes and/or biologically-relevant sample subsets.
BIBLIOGRAPHY


Leisch, F. (2004) "FlexMix: A general framework for finite mixture models and latent class regression in R." Journal of Statistical Software Volume, DOI:


CHAPTER IV
TRANSCRIPTOME EVOLUTION IN THE DEVELOPMENTAL STAGES OF
SQUAMOUS CELL LUNG CARCINOMA†

4.1 INTRODUCTION

Over 80% of cancers are carcinomas, which are malignancies that arise from epithelial cells. Carcinoma development is driven by a series of genetic alterations that are accompanied by histological changes that occur in a specific chronology and can often be observed in pre-malignant stages of tumourigenesis. A well-known example is the Fearon-Vogelstein model of colorectal cancer formation which correlates certain genetic alterations with the histopathological stages of progression (Fearon, 1990). The identification and characterization of the genetic changes that drive development is obviously critical to an overall understanding of tumourigenesis, and those that occur in pre-malignant and early malignant stages have particular importance as early detection markers. These genetic changes can potentially predict the risk for disease advancement and allow for more informed treatment decisions. However, studying these lesions is difficult since biological material is limited and clinical presentation is relatively rare.

This chapter describes the analysis of transcriptome profiles of the early stages of squamous cell carcinoma (SCC) of the lung. This dataset represents, to date, the most complete catalogue of gene expression during the development of a solid tumour. For several reasons, this cancer type is attractive as a target for this effort. Other than highly treatable forms of skin cancer, lung cancers are the single most commonly diagnosed and are the leading cause of cancer mortality, killing more people than breast, prostate and colon cancer combined (Jemal, 2007). The SCC subtype of lung cancer comprises 25-30% of these cancers, representing a large fraction of total cancer cases. Furthermore, this tumour subtype has a more defined and readily

† A version of this chapter will be submitted for publication. Zuyderduyn, S., Vatcher, G., Lam, S., Lam, W., MacAulay, C., Ng, R., and Ling, V. Transcriptome evolution in the developmental stages of squamous cell lung cancer.
identifiable progression of pre-invasive lesions, a feature common to squamous cell carcinomas in other tissues.

In lung, SCC progression starts with normal epithelium that (usually through exposure to tobacco smoke) advances to hyperplasia, squamous metaplasia (referring to the hallmark change in morphology of the normal columnar epithelium to a squamous cell type), varying degrees of dysplasia, carcinoma in situ (CIS), and finally a full blown invasive carcinoma (Hirsch, 2001). This nomenclature is similar in other tissues (e.g. skin, oesophagus, cervix) (Neville, 2002; Greer, 2006; Shimizu, 2007). Some plasticity is possible in this sequence, although it is thought to occur in the vast majority of cases. For example, lesions may spend very little time at a given stage of progression, or perhaps skip some stages altogether. Lesions in the precursor stages of malignant development often fail to progress and can even regress, particularly following smoking cessation (Colby, 1999). Observations of these events strongly suggest that an irreversible commitment to carcinogenesis does not occur until later stages of dysplasia, although it has almost certainly occurred at the CIS stage (Breuer, 2005).

By associating these histological stages of progression with the occurrence of certain changes in gene expression, a refined picture of the molecular basis of malignant potential and commitment in lung SCC can be developed. Previous gene expression profiles of SCC have compared normal samples (either bulk lung tissue or cell lines) to primary, invasive tumours (Bhattacharjee, 2001; Nacht, 2001). Since the intervening stages are omitted, it is impossible to assign a chronology to the observed expression changes and distinguish between benign, pre-malignant, and malignant events. By identifying the contribution to each stage of development, a refinement of the phenotype (e.g. pre-malignant vs. malignant) associated with certain gene expression changes can be determined.

This study profiles gene expression in several stages of SCC progression using serial analysis of gene expression (SAGE), a powerful technique for obtaining a comprehensive
snapshot of the transcriptome (Velculescu, 1995). This technique generates small sequence tags extracted from a defined position of each mRNA sequence. These tags are serially ligated and cloned and can then be sequenced and counted, generating a nearly unbiased profile of the mRNA population. SAGE has the advantage of being: 1) based on counts and therefore provides observations that are quantitative (compared to the more qualitative analog hybridization signals arising from microarray data); 2) capable of increased statistical power simply by increasing the sampling depth and/or including additional libraries generated at a later time; and 3) unbiased with respect to genes detected since no a priori knowledge of mRNA sequence is required.

The analysis presented here focuses on two objectives: 1) investigate the relationship between the global gene expression profile and the histological stages of SCC development, and 2) identify a set of gene expression signatures that can differentiate the different neoplastic stages of SCC.
4.2 MATERIALS AND METHODS

4.2.1 Sample collection and preparation

Details on the sampling and preparation of normal bronchial epithelium have been previously described (Lonergan, 2006). Pre-malignant lesions and tumour samples were obtained from biopsied material. Samples I11 and I12 are both pools of material from four patients and were obtained from Dr. Ming-Sound Tsao (Ontario Cancer Institute).

4.2.2 Data processing

Automated sequencer traces were base-called and assigned quality scores using the Phred software package (Ewing, 1998a; Ewing, 1998b). SAGE tags were extracted from sequence data using Bio::SAGE::DataProcessing (ver. 1.20), a module written in Perl, and default parameters (Zuyderduyn, 2004).

4.2.3 Multidimensional scaling

Multidimensional scaling (MDS), or principal coordinates analysis, was used to characterize the overall relationship between sets of libraries. This method projects a multidimensional set of values into a lower-dimensional space to aid interpretation. The distance between all possible library pairs was calculated using the Pearson correlation coefficient (Gower, 1966). Let $x_{ik}$ be the normalized expression (tag count/library size) of the $k$-th tag in library $i$. Let $N$ be the total number of tags compared. The distance between two libraries $i$ and $j$ is then:

$$r_{ij} = \frac{\sum_{k}^{N} x_{ik} x_{jk} - \frac{\sum_{k}^{N} x_{ik} \sum_{k}^{N} x_{jk}}{N}}{\sqrt{\left(\sum_{k}^{N} x_{ik}^{2} - \left(\frac{\sum_{k}^{N} x_{ik}}{N}\right)^{2}\right)\left(\sum_{k}^{N} x_{jk}^{2} - \left(\frac{\sum_{k}^{N} x_{jk}}{N}\right)^{2}\right)}}$$
A dissimilarity matrix was created where each cell contained the value $1 - r_{ij}$, for each possible pair of libraries $i$ and $j$. The cmdscale function of the R statistical software was used to determine the optimal coordinates to represent the libraries in 3 dimensions (Gentleman, 2004).

### 4.2.4 k-means clustering

The objective of $k$-means clustering is to separate a set of $n$ objects into $k$ partitions by maximizing the similarity of objects within each partition (Steinhaus, 1956). A Poisson-based clustering algorithm specifically designed for use with SAGE data was utilized (Cai, 2004). The algorithm is unable to handle tag counts of zero, so a value of 0.5 was used for these tags. In addition, the algorithm determines cluster centres in terms of raw tag count only and this can complicate interpretation. Therefore, the equations for the algorithm were modified slightly to produce cluster centres corrected for library size. Finally, there is no guarantee that a $k$-means analysis will identify the optimal set of clusters. However, this shortcoming is greatly assuaged when running the analysis multiple times with different starting (or seed) clusters and choosing the outcome with the least residual variance. In this case, the procedure is performed 20 times. In addition, a strategy called $k$-means++ can greatly improve the selection of the starting clusters and enhance the quality of the final clusters (Arthur, 2007). The $k$-means algorithm, supplemented with these improvements, was implemented using the C++ programming language (Stroustrop, 2000).

Since the number of partitions is not known a priori, the “elbow criterion” was used to determine a reasonable value for $k$ (Jain, 1999). This involves performing the clustering over a range of $k$ and plotting the amount of residual variance. In this case, the total within-cluster dispersion ($S$) was used as a metric for variance (Cai, 2004). As $k$ increases, the amount of variance not explained by the clusters will decrease (by definition, the variance will be zero when $k$ is equal to $n$). Under most circumstances, the plot will reveal an “elbow” where the
addition of more clusters shows a marked decline in the amount of variance removed. This point was chosen as the optimal $k$.

### 4.2.5 Hierarchical clustering

Sample-wise hierarchical clustering was performed using the same Poisson-based distance metric described in Section 4.2.4 averaged across all tags included in the clustering. Cluster-to-cluster distances were calculated using the average linkage method, where the mean distance between objects in the two clusters is used.

### 4.2.6 Gene Ontology and KEGG pathway enrichment

The Gene Ontology (GO) project is a collaborative effort to assign biological processes, cellular components, and molecular functions to descriptions of genes using a controlled vocabulary (Ashburner, 2000). In addition, GO terms are assigned to genes along with an evidence code, which describes the criteria used to make the assignment. In cases where a large list of genes makes a detailed examination prohibitive, a GO enrichment analysis can quickly uncover highly represented biological themes.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) features a collection of about 120 manually curated pathways grouped in broad categories of metabolism, genetic information processing, environmental information processing, cellular processes, and human diseases (Kanehisa, 2008). As with GO, a KEGG enrichment analysis can indicate units of the cell “circuitry” that show enhanced or depressed activity.

In this study, tags were mapped to Unigene clusters and provided to the online DAVID resource. DAVID performs statistical significance testing to identify enrichments of GO terms and KEGG pathways (Huang da, 2007). Re-sampling based false discovery rate (FDR) estimates were chosen to correct for multiple testing. In addition, only GO term-gene associations with acceptable evidence codes were chosen, based upon the recommendation of the
developers of the GoMiner software package; specifically, the assignment must be made based on a traceable author statement (TAS) or inferred by: direct assay (IDA), mutant phenotype (IMP), genetic or physical interaction (IGI, IPI), sequence or structural similarity (ISS), or expression pattern (IEP) (Zeeberg, 2003).

4.2.7 Statistical analysis

4.2.7.1 Preprocessing

In order to reduce the dimensionality of the data to a level that allowed analysis to be undertaken in a reasonable time, tags were conservatively pre-filtered to include only those that were expressed at least 4 times in at least one library (19,147 tags). Tags that do not meet this criterion cannot be distinguished from the variability due to counting, and will not be statistically informative.

4.2.7.2 Feature selection

The counts for each tag were fit to a $k$-component Poisson mixture model (Zuyderduyn, 2007; see Chapter III). When it is determined that number of components in the mixture model is one ($k=1$), then the variance in counts cannot be distinguished from sampling variability and are excluded from consideration. When $k>1$, a test statistic is calculated to determine how well the components separate the different stages of progression into some meaningful groups of interest (e.g. pre-malignant versus malignant, non-invasive versus invasive, etc.). Let $x_i$ be the class label for library $i$ (i.e. 0 for epithelial brushing, 1 for bulk normal, 2 for metaplasia/dysplasia, 3 for CIS, 4 for invasive). Let $\omega_i$ be the binary group label of interest for library $i$ (i.e. 0 for pre-malignant, 1 for malignant). Let $\beta$ be an ordered vector containing the estimated coefficients for the $k$ mixture components (e.g. $\beta_1<\beta_2<...<\beta_K$) and let $\tau$ be an index value between 1 and $K-1$ that partitions the elements of $\beta$ into two groups. Let $P(k|Y_i,n_i)$ be the
posterior probability that the count \( Y_i \) observed for library \( i \) of size \( n_i \) arose from component \( k \).

Since each component is Poisson distributed, this will be Poisson\((Y_i,n_i\exp(\beta_k))\). For each class of libraries \( x \), the average probability \( p_x = \text{avg}[P(k \leq \tau|Y_i)] \) is calculated. The score reflecting whether a tag count corresponds to one or the other group is:

\[
\Theta_0 = \prod p_x, \text{ for each class } x \text{ where } \omega=0
\]
\[
\Theta_1 = \prod (1-p_x), \text{ for each class } x \text{ where } \omega=1
\]

The final test statistic is simply:

\[
\Theta = \Theta_0 \times \Theta_1
\]

This value can range from 0-1, exclusive; the value approaches 1 as the tag counts correspond to the components separating the two groups. A confidence calculation is made for each possible separation \( \tau \) and the highest score is taken. An example calculation is shown in Figure 4.1.

The value of the test statistic is used to rank the relative suitability of each tag in a list of candidates. In order to determine a threshold score that can be used to isolate a list of high-quality candidates, a Monte Carlo simulation was performed. The class labels were randomly re-assigned, and the procedure described above was re-run. This was repeated 1,000 times. The number of candidates identified at a given threshold score was averaged over the 1,000 rounds of simulation and then compared to the actual number identified at that threshold for the true class labels. For example, at a confidence of 0.70 the Monte Carlo simulation may identify an average of 10 tags, while the true class labels identify 100 tags. Thus, for this threshold, one can estimate the false discovery rate (FDR) as \( 10/100=10\% \) (i.e. candidates that are not truly associated with the sample types, but are a result of random variation). Threshold scores were chosen to achieve a 5\% FDR.
**Figure 4.1: A confidence score calculation for a tag.** This example uses the tag `CTGCACCTTAC` and group labels corresponding to a test for differential expression in malignant samples. A detailed explanation of the procedure is found in Section 4.2.7.2. The tag maps to MCM7 and is part of the optimal malignant signature discussed in Section 4.3.3.3.
4.2.7.3 Estimating the selectivity of a candidate tag list and the generation of optimal signatures

The sensitivity and specificity of a given candidate tag list was estimated using leave-one-out cross-validation (LOOCV). Candidate tags were generated as described above, except one library was left out. This was repeated for all libraries. A prediction of whether the library was from group zero \((\omega = 0)\) or group one \((\omega = 1)\) (e.g. non-malignant versus malignant) was made with the Poisson distribution using the parameters determined by the mixture model fit:

\[
P(\omega | Y_i, n_i, \beta, \tau) = \sum_j \text{Poisson}(Y_i, n_i \exp(\beta_j))
\]

These values were scaled to ensure that the probability that the sample came from any group is 100%. Finally, these values were averaged over all of the tags in the candidate list to arrive at a final membership probability. Let \(G = i\) denote the set of libraries where \(\omega = i\). The sensitivity and specificity were estimated as follows:

\[
sensitivity = \frac{\sum_{G=0} P(\omega = 0 | Y_i, n_i, \beta, \tau)}{\sum_{G=0} P(\omega = 0 | Y_i, n_i, \beta, \tau) + \sum_{G=1} P(\omega = 0 | Y_i, n_i, \beta, \tau)}
\]

\[
specificity = \frac{\sum_{G=1} P(\omega = 1 | Y_i, n_i, \beta, \tau)}{\sum_{G=0} P(\omega = 1 | Y_i, n_i, \beta, \tau) + \sum_{G=1} P(\omega = 1 | Y_i, n_i, \beta, \tau)}
\]

In addition, a test statistic value was selected to generate an optimal signature using a forward selection strategy. A cutoff score was established and then successively lowered, and candidate genes that met this threshold were added to the signature in each round of cross-validation until a maximum selectivity was reached as determined by a receiver operating characteristic (ROC) curve. This score was then used to select an optimal signature from the full dataset.

4.2.8 Tag to gene mapping

To increase the accuracy and confidence of tag to gene mapping, tags were assigned additional nucleotides as described in Chapter II. A combination of resources was used to
facilitate tag to gene mapping. For rapid mapping of large tag lists, the LongSAGE (21bp) version of the SAGE Genie database (Boon, 2002) was used to allow the matching of 15-16bp tags obtained by the additional nucleotide procedure. When performing a more comprehensive mapping (i.e. to determine the source of a small number of tags of high interest), the SAGE Genie mapping was supplemented with:

1) Exact-match BLASTN (Altschul, 1997) to the human genome (NCBI 36 assembly) using the web-based EnsEMBL BLAST resource (Flicek, 2008)

2) Exact-match BLASTN to the NCBI human EST collection (8,137,888 sequences) (Benson, 2008) using the web-based NCBI BLAST resource (Johnson, 2008)

3) Custom Perl scripts that match a specified tag to other tags in the library that differ by a single nucleotide or a single insertion or deletion. The tag was considered an artefact only if the closely related tag was expressed at a 10-fold greater level.

4) Match to a complete database of antisense tags and tags occurring at anchoring enzyme sites other than the 3’-most position from a meta-catalogue of sequences from Refseq (Pruitt, 2007) and Unigene (Wheeler, 2008) (developed by G. Vatcher)

4.2.9 Microarray validation

Validation of gene expression signatures was carried out using publicly available microarray profiles. Unless otherwise stated, the data values determined by the processing protocol of the original authors were used. Genes that were identified as having an effect on the bronchial epithelium with exposure to tobacco smoke were validated with the following datasets:
1) The Spira dataset, containing 75 samples (23 never smokers, 18 former smokers, 34 current smokers) (Spira, 2004).

2) The Carolan dataset, containing 44 samples (18 never smokers, 26 current smokers) (Carolan, 2006).

3) The Beane dataset, containing 102 samples (21 never smokers, 31 former smokers, 52 current smokers) (Beane, 2007).

Genes that were identified in the developmental stages of SCC were validated with the following datasets:

1) The Bhattacharjee dataset, containing 38 samples (17 normal, 21 SCC) (Bhattacharjee, 2001). The profiles were generated using the Affymetrix U133A oligonucleotide array which contains 22,283 different probe sets. The published microarray data was processed using now obsolete methods, so the original Affymetrix CEL files were obtained from the authors’ website and processed with updated methods using the Bioconductor (www.bioconductor.org) R analysis package (Gautier, 2004). Specifically, the Robust Multichip Average method, which applies background correction and a quantile normalization step, is used (Bolstad, 2003; Irizarry, 2003a; Irizarry, 2003b). This technique is consistent with the data processing used in the remaining microarray validation datasets.

2) The Erez/Dehan dataset (GEO accession: GSE1987), containing 25 samples (2 normal, 7 tumour-associated normal, 16 SCC) (Erez, 2004; Dehan, 2007). The profiles were generated using the Affymetrix U95A oligonucleotide array which contains 12,651 different probe sets.

3) The Wachi dataset (GEO accession: GSE3268), containing 10 matched same-patient samples (5 normal, 5 SCC) (Wachi, 2005). As with the Bhattacharjee
dataset, the profiles were generated using the Affymetrix U133A oligonucleotide array which contains 22,283 different probe sets.

4.2.10 Tissue samples

Immunohistochemistry was performed on sections of formalin-fixed, paraffin-embedded tissue blocks (gift from Jaclyn Hung). These included 10 samples of Stage I and Stage II squamous cell lung cancers obtained from surgical resection. A positive control was included using 5 samples of invasive breast carcinoma, where MMP11 is known to be expressed. In addition, two samples from normal lung were included (gift from Calum MacAulay).

4.2.11 MMP11 antibody

Detection of MMP11 was done using a mouse monoclonal antibody purchased from LabVision (Cat.#MS-1035-P1ABX), which has high sensitivity and specificity in both Western blots and formalin-fixed, paraffin-embedded tissues.

4.2.12 Immunohistochemistry

Tissue sections were de-waxed by heating at 60°C for 20 minutes, followed by two 5 minute washes with xylene, two 2 minute washes with 100% ethanol, two 2 minute washes with 95% ethanol, two 2 minute washes with 70% ethanol, and two 5 minutes washes with distilled water. Antigen retrieval was performed by placing slides in a container filled with citrate acid buffer (0.094g citric acid and 0.60g sodium citrate per 250mL H₂O, pH adjusted to 6.0) and microwaving for 2.5 minutes at high power, 4 minutes at 10% power, and 4 minutes at 20% power. The container was cooled for 30 minutes in running water after which the slides were removed and rinsed three times with distilled water. Quenching to block endogenous peroxidise activity was performed by washing slides in PBS (0.8g PBS, 0.8g potassium dihydrogen orthophosphate, 5.4g potassium chloride, 32g di-sodium hydrogen orthophosphate per 4L H₂O).

115
for 3 minutes, 3% H$_2$O$_2$ in a light protected container for 10 minutes, and twice with PBS for 5
minutes. Several drops of DAKO Universal Blocker were applied to each slide and incubated
for 15 minutes. The slides were each tapped on paper towel to remove excess blocker. The
MMP11 antibody was diluted 500-fold with 1% BSA-PBS (5.8mL 0.1% Triton-PBS and 200µL
30% BSA) and several drops applied to each slide. The slides were incubated for 60 minutes.
The slides were then washed three times with 0.1% Triton-PBS for 1 minute. Several drops of
DAKO Envision secondary antibody (polymer linker) were then applied to each slide and
incubated for 30 minutes. The slides were each tapped on paper to towel to remove excess
antibody and then washed three times with PBS for 1 minute. Several drops of DAB were
applied to each slide and incubated for 7 minutes in a light protected container. The reaction was
then stopped by rinsing the slides in running water. The slides were then counterstained with
hematoxylin.
4.3 RESULTS

Thirty-nine SAGE libraries were constructed using: brushings of normal bronchial epithelium from four never smokers, twelve former and eight current smokers; and bulk samples from two pools of normal lung parenchyma, one squamous metaplasia, one dysplasia, five carcinoma \textit{in situ} (CIS), and six invasive cancers (Table 4.1). In addition, a publicly available bulk normal lung library generated as part of a larger human transcriptome profiling effort (GEO Accession: GSM762) was included, bringing the size of the total dataset to 40 libraries. For brevity, the libraries are prefixed with the labels \textbf{NS, FS, CS} (i.e. never, former, current smoker); \textbf{N, M, D, C and I}, respectively (the public library is referred to by its existing accession). Sample and library preparation and data pre-processing are described in Section 4.2.1-4.2.2. The majority of the libraries were sequenced to exceptional depth (>100,000 tags). The combined dataset contains a total of 4,614,031 tags, representing 256,361 unique sequences. 123,635 (48.2\%) of these are observed more than once, and are therefore less likely to represent artefacts arising from PCR amplification or sequencing errors. The brushings and normal lung parenchyma libraries have been previously published with an accompanying analysis (Lonergan, 2006).

4.3.1 A global view of the transcriptome during the development of SCC

4.3.1.1 Multidimensional scaling analysis

An unsupervised clustering of profiles was accomplished using multi-dimensional scaling (MDS) (as described in Section 4.2.3). In order to reduce the effect of random sampling variation on the Pearson distance, only the 10,718 tags where a normalized count of 5/100,000 was observed in at least one library were considered. The similarity between all 40 libraries was projected into three dimensions (Figure 4.2). The clearest distinction appears between libraries
Table 4.1: Summary of SAGE libraries

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
<th>Tag Count</th>
<th>Tag Types</th>
<th>% Count</th>
<th>% Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS90</td>
<td>normal bronchial epithelium (never smoker)</td>
<td>138419</td>
<td>31768</td>
<td>14.1</td>
<td>61.6</td>
</tr>
<tr>
<td>NS92</td>
<td>normal bronchial epithelium (never smoker)</td>
<td>127477</td>
<td>29118</td>
<td>13.9</td>
<td>60.8</td>
</tr>
<tr>
<td>NS93</td>
<td>normal bronchial epithelium (never smoker)</td>
<td>132292</td>
<td>30920</td>
<td>14.6</td>
<td>62.5</td>
</tr>
<tr>
<td>NS97</td>
<td>normal bronchial epithelium (never smoker)</td>
<td>139356</td>
<td>33155</td>
<td>14.9</td>
<td>62.6</td>
</tr>
<tr>
<td>FS03</td>
<td>normal bronchial epithelium (former smoker)</td>
<td>118395</td>
<td>27798</td>
<td>14.9</td>
<td>63.5</td>
</tr>
<tr>
<td>FS20</td>
<td>normal bronchial epithelium (former smoker)</td>
<td>59831</td>
<td>18401</td>
<td>20.0</td>
<td>64.9</td>
</tr>
<tr>
<td>FS21</td>
<td>normal bronchial epithelium (former smoker)</td>
<td>135539</td>
<td>31944</td>
<td>14.5</td>
<td>61.6</td>
</tr>
<tr>
<td>FS24</td>
<td>normal bronchial epithelium (former smoker)</td>
<td>140290</td>
<td>31507</td>
<td>13.8</td>
<td>61.3</td>
</tr>
<tr>
<td>FS32</td>
<td>normal bronchial epithelium (former smoker)</td>
<td>71637</td>
<td>21341</td>
<td>19.9</td>
<td>66.8</td>
</tr>
<tr>
<td>FS34</td>
<td>normal bronchial epithelium (former smoker)</td>
<td>63766</td>
<td>18487</td>
<td>19.2</td>
<td>64.4</td>
</tr>
<tr>
<td>FS38</td>
<td>normal bronchial epithelium (former smoker)</td>
<td>50331</td>
<td>16045</td>
<td>22.3</td>
<td>70.1</td>
</tr>
<tr>
<td>FS44</td>
<td>normal bronchial epithelium (former smoker)</td>
<td>147304</td>
<td>32528</td>
<td>13.6</td>
<td>61.6</td>
</tr>
<tr>
<td>FS49</td>
<td>normal bronchial epithelium (former smoker)</td>
<td>121901</td>
<td>30072</td>
<td>15.5</td>
<td>62.6</td>
</tr>
<tr>
<td>FS50</td>
<td>normal bronchial epithelium (former smoker)</td>
<td>113315</td>
<td>28663</td>
<td>16.2</td>
<td>64.1</td>
</tr>
<tr>
<td>FS73</td>
<td>normal bronchial epithelium (former smoker)</td>
<td>144650</td>
<td>34744</td>
<td>15.3</td>
<td>63.6</td>
</tr>
<tr>
<td>FS06+07</td>
<td>normal bronchial epithelium (former smoker)</td>
<td>159687</td>
<td>45248</td>
<td>13.9</td>
<td>71.1</td>
</tr>
<tr>
<td>CS26</td>
<td>normal bronchial epithelium (current smoker)</td>
<td>74155</td>
<td>21279</td>
<td>18.5</td>
<td>64.6</td>
</tr>
<tr>
<td>CS28</td>
<td>normal bronchial epithelium (current smoker)</td>
<td>66327</td>
<td>18410</td>
<td>18.0</td>
<td>64.8</td>
</tr>
<tr>
<td>CS36</td>
<td>normal bronchial epithelium (current smoker)</td>
<td>60645</td>
<td>16831</td>
<td>19.9</td>
<td>64.4</td>
</tr>
<tr>
<td>CS42</td>
<td>normal bronchial epithelium (current smoker)</td>
<td>69997</td>
<td>20026</td>
<td>18.3</td>
<td>64.0</td>
</tr>
<tr>
<td>CS48</td>
<td>normal bronchial epithelium (current smoker)</td>
<td>91467</td>
<td>22911</td>
<td>15.5</td>
<td>61.9</td>
</tr>
<tr>
<td>CS75</td>
<td>normal bronchial epithelium (current smoker)</td>
<td>156777</td>
<td>33869</td>
<td>13.2</td>
<td>61.1</td>
</tr>
<tr>
<td>CS80</td>
<td>normal bronchial epithelium (current smoker)</td>
<td>130300</td>
<td>30739</td>
<td>14.5</td>
<td>61.4</td>
</tr>
<tr>
<td>CS94</td>
<td>normal bronchial epithelium (current smoker)</td>
<td>135783</td>
<td>32171</td>
<td>14.7</td>
<td>62.2</td>
</tr>
<tr>
<td>GSM762</td>
<td>normal lung</td>
<td>89607</td>
<td>24694</td>
<td>18.1</td>
<td>65.5</td>
</tr>
<tr>
<td>N13</td>
<td>normal lung parenchyma (4 pooled samples)</td>
<td>45127</td>
<td>11776</td>
<td>16.6</td>
<td>63.8</td>
</tr>
<tr>
<td>N16</td>
<td>normal lung parenchyma (4 pooled samples)</td>
<td>42784</td>
<td>13102</td>
<td>20.8</td>
<td>68.0</td>
</tr>
<tr>
<td>M51</td>
<td>squamous metaplasia</td>
<td>156100</td>
<td>35063</td>
<td>15.2</td>
<td>67.5</td>
</tr>
<tr>
<td>D101</td>
<td>dysplasia</td>
<td>136257</td>
<td>31194</td>
<td>14.0</td>
<td>61.4</td>
</tr>
<tr>
<td>C01</td>
<td>carcinoma in situ</td>
<td>121363</td>
<td>28482</td>
<td>15.9</td>
<td>67.7</td>
</tr>
<tr>
<td>C02</td>
<td>carcinoma in situ</td>
<td>134327</td>
<td>27341</td>
<td>13.1</td>
<td>64.6</td>
</tr>
<tr>
<td>C05</td>
<td>carcinoma in situ</td>
<td>152599</td>
<td>34719</td>
<td>15.0</td>
<td>66.0</td>
</tr>
<tr>
<td>C27</td>
<td>carcinoma in situ</td>
<td>129372</td>
<td>28787</td>
<td>13.9</td>
<td>62.7</td>
</tr>
<tr>
<td>C39</td>
<td>carcinoma in situ</td>
<td>154387</td>
<td>33249</td>
<td>13.4</td>
<td>62.4</td>
</tr>
<tr>
<td>I04</td>
<td>invasive carcinoma</td>
<td>138269</td>
<td>31164</td>
<td>14.5</td>
<td>64.3</td>
</tr>
<tr>
<td>I08</td>
<td>invasive carcinoma</td>
<td>133099</td>
<td>25736</td>
<td>12.0</td>
<td>62.1</td>
</tr>
<tr>
<td>I11</td>
<td>invasive carcinoma (4 pooled samples)</td>
<td>117768</td>
<td>25641</td>
<td>13.8</td>
<td>63.3</td>
</tr>
<tr>
<td>I12</td>
<td>invasive carcinoma (4 pooled samples)</td>
<td>103119</td>
<td>21373</td>
<td>13.0</td>
<td>62.7</td>
</tr>
<tr>
<td>I18</td>
<td>invasive carcinoma</td>
<td>150444</td>
<td>31817</td>
<td>13.2</td>
<td>62.2</td>
</tr>
<tr>
<td>I22</td>
<td>invasive carcinoma</td>
<td>159588</td>
<td>45713</td>
<td>19.9</td>
<td>69.3</td>
</tr>
</tbody>
</table>

Summary of forty (40) SAGE libraries constructed with samples of normal bronchial epithelium from never (4), former (12) and current (8) smokers; normal lung tissue (3); squamous metaplasia (1); dysplasia (1); carcinoma in situ (5); and invasive carcinoma (6). Twenty-eight (28) libraries consist of >100,000 SAGE tags.
Figure 4.2: Multidimensional scaling (MDS) analysis of 40 SAGE libraries from samples reflecting different stages of SCC development. Datapoints are shaped and colour-coded according to the sample type: bulk normal is denoted with a white circle, brushings from bronchial epithelium are green circles (dark green are never smokers, green are former smokers, and cyan are current smokers), metaplasia is denoted with a blue upwards triangle, dysplasia with a purple downwards triangle, carcinoma \textit{in situ} with a yellow diamond, and invasive carcinoma with a red square. The full 3D plot is displayed in the top left section. Cross-sections, clockwise from the top right, show the X-Y, Y-Z, and X-Z planes. The dashed circles on the X-Y plane highlight the separation of brushing and bulk tissue sample types.
taken from normal or smoke-exposed epithelial brushings and bulk tissue samples (including the bulk normal), showing that the largest source of variation arises from the method used to obtain the samples. Having identified this aspect of the data's structure, an additional MDS was performed separately on these two groups to identify substructure(s) that have a more relevant biological basis.

The MDS analysis of the 24 brushing samples reveals evidence of transcriptome changes in the epithelium of the lung due to smoke exposure, although this distinction is not particularly strong (Figure 4.3). Almost half of the libraries are organized into a single, tight cluster containing 4/4 never smokers, 6/12 former smokers, and 1/8 current smokers. The remaining libraries are positioned at varying distances from the main cluster, with the former smokers tending to lay closest and current smokers tending to be most distant. Although these data strongly suggest significant changes in gene expression as a result of smoke exposure (and particularly recent exposure), there appear to be other sources of variation that can have a substantial impact on the transcriptome. These factors remain to be elucidated, as an investigation of pack-years, lung function, or time since smoking cessation (Table 4.2) and a correlation with the relationship between sample transcriptomes was not found.

The MDS analysis of the 14 tissue samples representing different stages of SCC development reveals clearer differences (Figure 4.4). The three normal samples form a distinct cluster apart from both pre-malignant and malignant samples. A further partition of the latter can be identified that separates 3/5 CIS samples and 3/6 invasive carcinoma samples from the pre-malignant samples and the remainder of the malignant samples. This suggests that squamous differentiation, the hallmark of metaplasia, is the predominant source of variation in gene expression. Although there is a degree of delineation between pre-malignancy and malignancy, this phenotypic distinction is not a particularly strong feature of the data. Moreover, a distinct positioning of the invasive samples is not evident.
Table 4.2: Source patient information for bronchial epithelium brushings samples

<table>
<thead>
<tr>
<th>ID</th>
<th>Gender</th>
<th>Age</th>
<th>Pack-years</th>
<th>Cessation time (years)</th>
<th>Lung function (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS90</td>
<td>M</td>
<td>58</td>
<td>n/a</td>
<td>n/a</td>
<td>115</td>
</tr>
<tr>
<td>NS92</td>
<td>F</td>
<td>56</td>
<td>n/a</td>
<td>n/a</td>
<td>104</td>
</tr>
<tr>
<td>NS93</td>
<td>M</td>
<td>53</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>NS97</td>
<td>F</td>
<td>81</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>FS03</td>
<td>M</td>
<td>68</td>
<td>33</td>
<td>19</td>
<td>50</td>
</tr>
<tr>
<td>FS06+07</td>
<td>M</td>
<td>69</td>
<td>100</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>FS21</td>
<td>M</td>
<td>68</td>
<td>30</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>FS24</td>
<td>M</td>
<td>70</td>
<td>75</td>
<td>17</td>
<td>76</td>
</tr>
<tr>
<td>FS32</td>
<td>M</td>
<td>67</td>
<td>55</td>
<td>5</td>
<td>n/a</td>
</tr>
<tr>
<td>FS20</td>
<td>M</td>
<td>65</td>
<td>82</td>
<td>10</td>
<td>59</td>
</tr>
<tr>
<td>FS34</td>
<td>F</td>
<td>56</td>
<td>64</td>
<td>1.5</td>
<td>71</td>
</tr>
<tr>
<td>FS44</td>
<td>F</td>
<td>63</td>
<td>45</td>
<td>4.5</td>
<td>83</td>
</tr>
<tr>
<td>FS49</td>
<td>M</td>
<td>72</td>
<td>40</td>
<td>32</td>
<td>87</td>
</tr>
<tr>
<td>FS50</td>
<td>F</td>
<td>71</td>
<td>56</td>
<td>16</td>
<td>58</td>
</tr>
<tr>
<td>FS38</td>
<td>M</td>
<td>72</td>
<td>63</td>
<td>6</td>
<td>n/a</td>
</tr>
<tr>
<td>FS73</td>
<td>M</td>
<td>69</td>
<td>55</td>
<td>21</td>
<td>57</td>
</tr>
<tr>
<td>CS26</td>
<td>F</td>
<td>63</td>
<td>40</td>
<td>n/a</td>
<td>69</td>
</tr>
<tr>
<td>CS28</td>
<td>M</td>
<td>56</td>
<td>62</td>
<td>n/a</td>
<td>89</td>
</tr>
<tr>
<td>CS36</td>
<td>F</td>
<td>63</td>
<td>44</td>
<td>n/a</td>
<td>96</td>
</tr>
<tr>
<td>CS42</td>
<td>M</td>
<td>68</td>
<td>81</td>
<td>n/a</td>
<td>76</td>
</tr>
<tr>
<td>CS48</td>
<td>M</td>
<td>64</td>
<td>45</td>
<td>n/a</td>
<td>73</td>
</tr>
<tr>
<td>CS75</td>
<td>M</td>
<td>66</td>
<td>53</td>
<td>n/a</td>
<td>85</td>
</tr>
<tr>
<td>CS80</td>
<td>M</td>
<td>52</td>
<td>48</td>
<td>n/a</td>
<td>63</td>
</tr>
<tr>
<td>CS94</td>
<td>F</td>
<td>55</td>
<td>34</td>
<td>n/a</td>
<td>81</td>
</tr>
</tbody>
</table>
Figure 4.3: Multidimensional scaling (MDS) analysis of 24 SAGE libraries from brushings of bronchial epithelium with different levels of tobacco smoke exposure. Datapoints are colour-coded according to the sample type: dark green are never smokers, green are former smokers, and cyan are current smokers. The full 3D plot is displayed in the top left section. Cross-sections, clockwise from the top right, show the X-Y, Y-Z, and X-Z planes.
Figure 4.4: Multidimensional scaling (MDS) analysis of 16 SAGE libraries from bulk samples reflecting the different stages of SCC development. Datapoints are colour-coded according to the sample type: bulk normal is denoted with a white circle, metaplasia is denoted with a blue upwards triangle, dysplasia with a purple downwards triangle, carcinoma in situ with a yellow diamond, and invasive carcinoma with a red square. The full 3D plot is displayed in the top left section. Cross-sections, clockwise from the top right, show the X-Y, Y-Z, and X-Z planes. The dashed circles on the X-Y plot highlight the separation between normal, pre-malignant, and malignant samples.
These data strongly suggest that the majority of transcriptome change occurs at the earliest stages of SCC development. While the changes associated with squamous differentiation appear most consistent, changes associated with malignant transformation appear to be substantial but highly variable. Two factors may explain this observation: 1) cell type heterogeneity (e.g. cells of stromal origin, residual normal epithelium, or members of the immune system) may introduce considerable variability, and 2) the particular genes that show altered expression and the extent of this alteration may differ substantially from tumour to tumour. Finally, gene expression changes associated with the invasive phenotype appear to be relatively limited. From this analysis, it is impossible to speculate on the variability of such changes. Although within-stage similarity coupled with a separation of samples according to the stage of SCC development (i.e. invasive samples most distant from normal, CIS between normal and invasive, etc.) would be an ideal result showing a strong relationship between histological classification and the transcriptome, the observed result is not entirely inconsistent with existing knowledge. Notably, genetic heterogeneity is an accepted reality in lung tumours (i.e. candidate oncogenes always appear disrupted in a certain percentage of tumours, typically 20-50%), and the high percentage of CIS that eventually progress to a full-blown invasive carcinoma (90-100%) suggests the number of additional molecular changes required to facilitate this process is rather small.

### 4.3.1.2 Common gene expression patterns identified by k-means clustering

In order to explore the transcriptome further, a k-means clustering analysis was performed (as detailed in Section 4.2.4). A total of 2,222 tags with a mean expression of 5/100,000 tags were included. The elbow criterion suggests $k=9$ clusters are sufficient to capture the most prominent features of the data (Figure 4.5A). Of these tags, 2,087 (93.9%) cluster into one of four patterns of expression (Figure 4.5B). These clusters appear to represent distinct
Figure 4.5: Determination of cluster number for k-means analysis and the four major clusters. The top graph (A) is an elbow plot used to estimate a reasonable number of clusters. The x-axis is the number of clusters, and the y-axis is the total within-cluster dispersion ($S$) as defined in Cai et al. (2004). The chosen number ($k=9$) is indicated. The bottom graph (B) depicts the values of the four most populated clusters. The x-axis shows the sample and the y-axis shows the value of the cluster centre. The cluster centre values include a correction for the library size (see Section 4.2.4) and correspond to the ratio of expression (e.g. 0.10 means that 10% of the total observed expression appears in a given sample).
transcriptome types or states, as the four appear in equilibrium with one another.

Of particular note is the opposing relationship of Cluster A with Cluster B. This is a significant concern for the analysis of the bulk samples, where the contribution of these two clusters is highly variable. Given the observed pattern and the identity of these samples, a reasonable conclusion is that an increase in Cluster A tags indicates a greater number of stromal cells (e.g. fibroblasts, endothelial cells), while Cluster B indicates a greater number of the actual pre-malignant or malignant cells (e.g. squamous epithelium).

An increase in expression of tags in Clusters C and D are strongly associated with epithelial brushing samples. The two clusters are similar except for an increase in the expression of Cluster D tags in samples from current smokers. Given the purity of these samples, these two clusters likely correlate with the baseline expression of normal epithelium and an additional “stressed” state (i.e. as a result of current exposure to tobacco smoke), respectively. Furthermore, the values for the invasive sample I22 and, to a lesser extent, the dysplasia sample D101 seem to suggest some contamination with these cells. This manifests in the MDS analysis as a shift of the position of these samples towards the space occupied by the brushing samples (Figure 4.2). This may have resulted from the inadvertent capture of tissue surrounding the margin of these sites of interest.

A GO enrichment analysis was performed on each cluster to determine if certain biological themes are overrepresented (as detailed in Section 4.2.6). The tags in each cluster were mapped to Unigene entries and submitted to the online DAVID analysis resource (Table 4.3) (Huang da, 2007). When enriched GO terms were identified with an FDR ≤5%, several clusters could be associated with specific themes. Cluster A is enriched for genes encoding intracellular proteins and functions that could be associated with general cellular housekeeping, including primary metabolic processes, gene expression, and translation. Cluster B and Cluster C did not show any significant enrichments. Cluster D was associated with the axoneme and
Table 4.3: Description of clusters identified by $k$-means analysis

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Number Tags</th>
<th>Unique Unigene IDs</th>
<th>Unique DAVID IDs</th>
<th>Putative biological significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>948</td>
<td>624</td>
<td>534</td>
<td>housekeeping type I</td>
</tr>
<tr>
<td>B</td>
<td>571</td>
<td>281</td>
<td>220</td>
<td>housekeeping type II</td>
</tr>
<tr>
<td>C</td>
<td>458</td>
<td>307</td>
<td>243</td>
<td>housekeeping type III</td>
</tr>
<tr>
<td>D</td>
<td>110</td>
<td>76</td>
<td>60</td>
<td>ciliated epithelial cells</td>
</tr>
<tr>
<td>E</td>
<td>54</td>
<td>36</td>
<td>33</td>
<td>keratinocyte differentiation</td>
</tr>
<tr>
<td>F</td>
<td>44</td>
<td>30</td>
<td>26</td>
<td>extracellular matrix</td>
</tr>
<tr>
<td>G</td>
<td>15</td>
<td>7</td>
<td>4</td>
<td>immune system markers</td>
</tr>
<tr>
<td>H</td>
<td>12</td>
<td>2</td>
<td>1</td>
<td>secretoglobin</td>
</tr>
<tr>
<td>I</td>
<td>10</td>
<td>7</td>
<td>4</td>
<td>pulmonary surfactants</td>
</tr>
</tbody>
</table>
contained genes previously found in normal lung, consistent with their increased presence in the
ciliated epithelial cells collected by brushing. The minor clusters contain small numbers of genes
with obvious and specific roles (Figure 4.6). Cluster E contains genes involved in keratinocyte
differentiation. Cluster F contains genes that are secreted and carry out structural and regulatory
roles in the extracellular matrix. Cluster G contains immunoglobulins, indicating these samples
have an increased infiltration by immune cells. Cluster H contains secretoglobin, a gene known
to be expressed in the lung but whose function is not yet clear. Finally, Cluster I contains several
pulmonary surfactants which are important for maintaining the structural integrity and function
of the alveoli.

4.3.2 Transcriptional signatures of developmental stages

A supervised classification strategy was undertaken to identify sets of genes that are
highly correlated with the stages of SCC development (detailed in Section 4.2.7). The Poisson
mixture model (Chapter II) was particularly amenable for this dataset; indeed, it was partially
inspired by the challenges this SAGE data presented. Specifically, the notion of mixture
“components”, which essentially act as a discrete Bayesian prior distribution, can account for: a)
a wide range of expression, possibly the result of different mechanisms of dysregulation or the
extent to which they occur (e.g. the number of copies involved in copy number variation); and b)
some of the cellular heterogeneity revealed by the $k$-means analysis (Section 4.3.1.2). A
continuous, unimodal prior (such as the Beta or Gamma distribution), which is typically used in
studies similar to this, fits poorly in such cases and regularly provides questionable estimates of
significance (Zuyderduyn, 2007).

4.3.2.1 Changes associated with tobacco smoke exposure

The incidence of lung SCC is strongly associated with long-term exposure to tobacco
smoke. Some gene expression changes that occur as a result of smoking may have casual or
Figure 4.6: The five minor k-means clusters. The clustering algorithm was set to determine $k=9$ clusters. The top graph (A) depicts the values of the fifth and sixth most populated clusters. The bottom graph (B) depicts the values of the seventh to ninth most populated clusters. Both plots show the samples on the x-axis and the values of the cluster centres on the y-axis. The cluster centre values include a correction for the library size (see Section 4.2.4) and correspond to the ratio of expression (e.g. 0.10 means that 10% of the total observed expression appears in a given sample).
mitigating roles in the ultimate development of a tumour, while others may be a transient cellular response to this exposure that ceases upon smoking cessation. To explore the latter, an analysis of the subset of SAGE libraries from bronchial brushings was performed by searching for gene expression changes that were specific to one of the groups of never (NS), former (FS), and current (CS) smokers (Figure 4.7). Although substantial numbers of tags were identified that appear differentially expressed in the CS group and likely represent differences resulting from an acute response to tobacco exposure, none could be identified in the NS group. This supports the notion that the bronchial epithelium reverts to a near-normal state upon smoking cessation. However, a small number of changes were identified in the FS group that suggests the presence of a distinct and persistent gene expression signature in cells that have had substantial past exposure to tobacco smoke.

### 4.3.2.1.1 Acute response to tobacco smoke exposure

In the CS group, a set of 70 upregulated and 9 downregulated tags were identified (Figures 4.8-4.9). Based on LOOCV, the sensitivity and specificity of these signatures are 66.2% and 86.4%, respectively, for the upregulated tags and 64.1% and 80.6%, respectively, for the downregulated tags. Optimal signatures were constructed to maximize the selectivity. This resulted in a 7 gene upregulated signature (ADH7, NQO1, CYP1B1, ALDH3A1, GPX2, AKR1B10, TFF3) with a sensitivity and specificity of 79.7% and 95.8%, respectively. Correspondingly, a single downregulated gene (C3) was identified with a sensitivity and specificity of 96.4% and 99.4%, respectively. The upregulated signature showed excellent agreement with all 3 of the validation microarray datasets, with a combined sensitivity and specificity of 92.0% and 92.8%, respectively (Table 4.4). However, the downregulated signature showed inconsistent performance. Although the C3 gene performed well in classifying the Carolan dataset, specificity was poor on the Spira and Beane datasets. The combined sensitivity
Figure 4.7: Candidate smoke exposure tag selection plots. Each plot depicts the number of tags (bar plot, left y-axis) and estimated false discovery rate (FDR) (line plot, right y-axis) selected at a given threshold test statistic value (x-axis) (the thick vertical line marks the 5% FDR level). The number of tags selected at the 5% FDR level and the value of the test statistic (\(\Theta\)) are shown at the top-right of each plot.
Figure 4.8: Heatmap with sample-wise hierarchical clustering of 70 tags upregulated in bronchial epithelium from current smokers. Samples are bronchial brushings from never, former, and current smokers. The dendrogram is generated using complete-linkage with a Poisson-based distance metric based on Cai et al. (2004). Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key placed between the dendrogram and the heatmap denotes the sample type: never smokers are dark green, former smokers are light green, and current smokers are cyan. Additional nucleotides identified by the XBP-SAGE method (Chapter II) are shown. Tags prefixed with an asterisk (*) indicate that the additional nucleotides were needed for a successful mapping.
Figure 4.9: Heatmap with sample-wise hierarchical clustering of 9 tags downregulated in bronchial epithelium from current smokers. Samples are bronchial brushings from never, former, and current smokers. The dendrogram is generated using complete-linkage with a Poisson-based distance metric based on Cai et al. (2004). Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key placed between the dendrogram and the heatmap denotes the sample type: never smokers are dark green, former smokers are light green, and current smokers are cyan. Additional nucleotides identified by the XBP-SAGE method (Chapter II) are shown. Tags prefixed with an asterisk (*) indicate that the additional nucleotides were needed for a successful mapping.
Table 4.4: Microarray validation of smoke-exposure signatures

<table>
<thead>
<tr>
<th>Regulation</th>
<th>Dataset</th>
<th>Genes</th>
<th>Never smokers</th>
<th>Former smokers</th>
<th>Current smokers</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Up</strong></td>
<td>Spira</td>
<td>7</td>
<td>23/23</td>
<td>14/18</td>
<td>33/34</td>
<td>97.0%</td>
<td>90.2%</td>
</tr>
<tr>
<td></td>
<td>Carolan</td>
<td>6</td>
<td>18/18</td>
<td>n/a</td>
<td>26/26</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td>Beane</td>
<td>7</td>
<td>21/21</td>
<td>27/31</td>
<td>44/52</td>
<td>84.6%</td>
<td>92.3%</td>
</tr>
<tr>
<td><strong>Down</strong></td>
<td>Spira</td>
<td>1</td>
<td>20/23</td>
<td>14/18</td>
<td>21/34</td>
<td>92.1%</td>
<td>54.0%</td>
</tr>
<tr>
<td></td>
<td>Carolan</td>
<td>1</td>
<td>18/18</td>
<td>n/a</td>
<td>24/26</td>
<td>100.0%</td>
<td>90.0%</td>
</tr>
<tr>
<td></td>
<td>Beane</td>
<td>1</td>
<td>20/21</td>
<td>20/31</td>
<td>27/52</td>
<td>97.9%</td>
<td>35.7%</td>
</tr>
</tbody>
</table>

Classification performance of genes mapped to SAGE tags identified as upregulated in current smokers compared to never or former smokers. $k$-means clustering ($k=2$) was used to classify the samples. The optimal upregulated signature of 7 tags and optimal signature of 1 downregulated tag was used. Note: $AKR1B10$ was not present on the Carolan dataset microarrays.
and specificity was 96.4% and 51.3%, respectively.

The 70 upregulated tags were mapped to 37 unique Unigene identifiers and a GO enrichment analysis revealed the terms “electron transport” (GO:0055114; FDR<0.1%) and “oxidoreductase activity, acting on CH-OH group of donors” (GO:0016614; FDR=0.1%) as significant themes. Furthermore, a KEGG enrichment analysis revealed an overrepresentation of members of the pathways responsible for the “metabolism of xenobiotics by cytochrome P450” (KEGG:hsa00980; FDR<0.1%) and “glutathione metabolism” (KEGG:hsa00480; FDR<0.1%). Both of these pathways work in concert to drive the metabolism of a diverse number of chemicals, including those found in tobacco smoke. Indeed, these two pathways form the canonical Phase I and Phase II enzyme system that is the first line of defence against foreign compounds. The Phase I reactions, involving cytochrome P450 in particular, drive various modification reactions that enable Phase II reactions, which includes glutathione conjugation, to detoxify the substrate. Specifically, ADH7, AKR1C2, AKR1C3, ALDH3A1, CBR1, CYP1A1, and CYP1B1 are known Phase I enzymes, while GPX2, GPX4, GSR, GSTA2, GSTP1, NQO1, and TALDO1 are known Phase II enzymes. All but one of the genes present in the 7 gene optimal signature are known players in this system. The only exception, TFF3, is a component of the epithelial mucosa of several tissues, including the lung. It is thought to be responsive to epithelial damage and involved in affecting the repair of the mucosa (Wiede, 1999; Oertel, 2001; Hoffmann 2007).

4.3.2.1.2 Persistent response to tobacco smoke exposure

No significant differences were found that were specific to the NS group, which would correspond to changes that persist even after an individual has stopped smoking. This suggests that the majority of the cellular response to tobacco smoke is acute and returns to a predominantly normal state once the exposure is removed. However, a small number of genes
were identified in the FS group (Figure 4.10A-B). Such changes do not fit well with the notion of a dichotomous state that is influenced by past or present smoke exposure. Rather, these changes would correspond to a past exposure signature distinct from the acute response. A set of 4 upregulated tags had a sensitivity and specificity of 74.5% and 88.6%, respectively. However, this signature showed no consistent agreement with the Spira and Beane microarray datasets, which both contain former smokers. The single downregulated tag maps to \textit{EPHX1}, and the sensitivity and specificity of this gene in identifying a former smoker is 65.9% and 84.8%, respectively. Although this gene did not show significant change in the Spira dataset, the Beane dataset also displayed a noticeable downregulation of \textit{EPHX1} in former smokers, particularly in those who had quit for longer than 24 months (Figure 4.10C). The former smokers in the Spira dataset include a large number of individuals who smoked for a relatively short period of time (\(\leq 10\) pack-years) and/or quit for short period of time (<12 months) before the samples were taken. It is possible that \textit{EPHX1} is downregulated following cessation after long-term exposure to tobacco smoke. Such an event would be interesting because of its role in the Phase I enzyme system, which is so clearly affected by acute exposure. Moreover, the gene is prominent in a known mechanism of tobacco smoke-induced mutagenesis. Normally \textit{EPHX1} plays a protective role, but benzo[a]pyrene, the major mutagen in tobacco smoke, is a procarcinogen that must undergo modification by both CYP1A1 or CYP1B1 and \textit{EPHX1} to form benzo[a]pyrene-7,8-dihydrodiol-9,10-oxide, a potent carcinogen (Denissenko, 1996). A loss of \textit{EPHX1} activity following cessation may then render the lung tissue more susceptible to mutation. Although the totality of the data suggests there is a true effect, the duration and amount of exposure and the time since smoking cessation appear to be important variables that cannot be adequately modelled with the available data.
Figure 4.10: Heatmaps with sample-wise hierarchical clustering of 4 tags upregulated and 1 tag downregulated in former, but not current, smokers. A boxplot of the expression of the candidate downregulated gene *EPHX1* in the Beane microarray dataset is also shown. Samples are bronchial brushings from never, former, and current smokers. The dendrogram is generated using complete-linkage with a Poisson-based distance metric based on Cai *et al.* (2004). Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key placed between the dendrogram and the heatmap denotes the sample type: never smokers are dark green, former smokers are light green, and current smokers are cyan. Additional nucleotides identified by the XBP-SAGE method (Chapter II) are shown. Tags prefixed with an asterisk (*) indicate that the additional nucleotides were needed for a successful mapping.
4.3.2.2 Changes associated with SCC development

Gene expression differences associated with SCC development were explored by grouping the bulk samples into 4 groups: normal, pre-malignant (metaplasia/dysplasia), CIS, and invasive carcinoma. The brushing libraries were included as an additional group, and were always used to support the bulk normal lung samples during classification. Since the $k$-means analysis suggests that bulk normal samples are primarily composed of stromal cells, while the brushing samples are primarily composed of epithelial cells, a putative gene expression marker must show a significant dysregulation compared to both of these groups. This approach has some cost in terms of power, as some differences present in a pure population of pre-malignant or malignant cells may have been apparent when comparing to a pure population of epithelial cells. In this case, a change in expression must be apparent in a mixed population when compared to two relatively pure populations. Nevertheless, the candidates that are identified can be regarded with high confidence. Attempts to model the amount of cellular heterogeneity proved unsuccessful and, interestingly, resulted in the identification of similar candidates (data not shown).

SAGE tags were identified at a cutoff FDR of 5% for each stage of progression (plots of the FDR over a range of test scores are shown in Figure 4.11). In addition, candidates were identified for all other possible combinations of the four groups not consistent with the known progression of SCC (the number of candidates for each combination are summarized in Figure 4.12). 608/789 (80%) of differentially expressed tags correspond to the accepted order of SCC progression. 146/159 (92%) of the remaining tags are either upregulated in metaplasia, dysplasia, and CIS; or CIS alone. This would be expected in a situation where stromal cells are a considerable presence in the invasive samples, as suggested by the $k$-means analysis. These two categories likely correspond to genes that are, in reality, associated with squamous differentiation or malignancy, respectively, and have simply had their increase obscured. A GO/KEGG
Figure 4.11: Candidate squamous cell lung cancer progression tag selection plots. Each plot depicts the number of tags (bar plot, left y-axis) and estimated false discovery rate (FDR) (line plot, right y-axis) selected at a given threshold test statistic value (x-axis) (the thick vertical line marks the 5% FDR level). The number of tags selected at the 5% FDR level and the value of the test statistic ($\Theta$) are shown at the top-right of each plot. The plot for upregulated invasive tags (bottom, far-right) is overlayed with the results of the procedure with sample I22 excluded (bar plot in light green, and line plot in green).
Figure 4.12: Venn diagram of the number of candidate tags identified for different combinations of SCC progression sample types. The left Venn diagram contains the number of candidates, and the right Venn diagram is a legend indicating the biological significance of each of the diagram’s regions. The dashed outline indicates the bronchial brushing sample type, which is combined with the bulk normal sample type in all comparisons, except for a test of brushings only (1407).
enrichment analysis on each of these combinations of sample types supports this view. While
tags upregulated in metaplasia and later stages compared to normal map to genes that are
strongly associated with keratinocyte differentiation and cornified envelope formation, which is
consistent with a squamous cell phenotype, these terms are also enriched in the two anomalous
comparisons.

4.3.2.2.1 Changes associated with pre-malignant transformation

138 upregulated and 316 downregulated tags were associated with the transition from
The placement of the unusual sample I22 in hierarchical clustering is consistent with the notion
of contamination by normal epithelium suggested by the k-means analysis. The upregulated tags
mapped to 85 unique Unigene entries and a GO enrichment analysis of these genes revealed the
term “keratinocyte differentiation” (GO:0030216; FDR<0.1%), consistent with a squamous cell
phenotype. This is implied from the upregulation of the genes \textit{CSTA}, \textit{SDC1}, \textit{SPRR1A}, \textit{SPRR2A},
\textit{SFN}, and \textit{TP63} (Gibbs, 1993; Lee, 2000; Nakajima, 2003; Ojeh, 2008; Truong, 2006).
Moreover, \textit{KRT6A}, \textit{KRT6B}, and \textit{KRT14} are upregulated, consistent with keratinisation, a
histological hallmark of metaplasia and SCC. Keratins may also play an active role in
progression. For example, a recent study transfected mice with the human \textit{KRT14} gene, and
constitutively activated it in lung Clara cells (Dakir, 2008). This was sufficient to activate
squamous differentiation, as indicated by an increase in the expression of particular markers,
including \textit{SPRR1A}. However, it was insufficient to immediately initiate squamous maturation,
although these mice developed hyperplastic and metaplastic lesions with time.

Based on LOOCV, the sensitivity and specificity of these signatures are 72.4% and
81.6%, respectively, for the upregulated tags and 83.1% and 76.7%, respectively, for the
downregulated tags. Optimal signatures were constructed to maximize the selectivity. This
Figure 4.13.1: Heatmap with sample-wise hierarchical clustering of the first 69 of 138 tags upregulated in metaplasia and later stages of SCC development. The dendrogram is generated using average-linkage with a Poisson-based distance metric based on Cai et al. (2004). Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key placed between the dendrogram and the heatmap denotes the sample type: white is bulk normal; epithelial brushings from never, former, and current smokers shown in dark green, light green, and cyan, respectively; metaplasia in blue; dysplasia in purple; carcinoma in situ in yellow; and invasive carcinoma in red. Additional nucleotides identified by the XBP-SAGE method (Chapter II) are shown. Tags prefixed with an asterisk (*) indicate that the additional nucleotides were needed for a successful mapping.
Figure 4.13.2: Heatmap with sample-wise hierarchical clustering of the final 69 of 138 tags upregulated in metaplasia and later stages of SCC development. The dendrogram is generated using average-linkage with a Poisson-based distance metric based on Cai et al. (2004). Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key placed between the dendrogram and the heatmap denotes the sample type: white is bulk normal; epithelial brushings from never, former, and current smokers shown in dark green, light green, and cyan, respectively; metaplasia in blue; dysplasia in purple; carcinoma in situ in yellow; and invasive carcinoma in red. Additional nucleotides identified by the XBP-SAGE method (Chapter II) are shown. Tags prefixed with an asterisk (*) indicate that the additional nucleotides were needed for a successful mapping.
Figure 4.14.1: Heatmap with sample-wise hierarchical clustering of the first 79 of 316 tags downregulated in metaplasia and later stages of SCC development. The dendrogram is generated using average-linkage with a Poisson-based distance metric based on Cai et al. (2004). Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key placed between the dendrogram and the heatmap denotes the sample type: white is bulk normal; epithelial brushings from never, former, and current smokers shown in dark green, light green, and cyan, respectively; metaplasia in blue; dysplasia in purple; carcinoma \textit{in situ} in yellow; and invasive carcinoma in red. Additional nucleotides identified by the XBP-SAGE method (Chapter II) are shown. Tags prefixed with an asterisk (*) indicate that the additional nucleotides were needed for a successful mapping.
Figure 4.14.2: Heatmap with sample-wise hierarchical clustering of the second 79 of 316 tags downregulated in metaplasia and later stages of SCC development.

The dendrogram is generated using average-linkage with a Poisson-based distance metric based on Cai et al. (2004). Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key placed between the dendrogram and the heatmap denotes the sample type: white is bulk normal; epithelial brushings from never, former, and current smokers shown in dark green, light green, and cyan, respectively; metaplasia in blue; dysplasia in purple; carcinoma in situ in yellow; and invasive carcinoma in red. Additional nucleotides identified by the XBP-SAGE method (Chapter II) are shown. Tags prefixed with an asterisk (*) indicate that the additional nucleotides were needed for a successful mapping.
Figure 4.14.3: Heatmap with sample-wise hierarchical clustering of the third 79 of 316 tags downregulated in metaplasia and later stages of SCC development. The dendrogram is generated using average-linkage with a Poisson-based distance metric based on Cai et al. (2004). Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key placed between the dendrogram and the heatmap denotes the sample type: white is bulk normal; epithelial brushings from never, former, and current smokers shown in dark green, light green, and cyan, respectively; metaplasia in blue; dysplasia in purple; carcinoma in situ in yellow; and invasive carcinoma in red. Additional nucleotides identified by the XBP-SAGE method (Chapter II) are shown. Tags prefixed with an asterisk (*) indicate that the additional nucleotides were needed for a successful mapping.
Figure 4.14.4: Heatmap with sample-wise hierarchical clustering of the final 79 of 316 tags downregulated in metaplasia and later stages of SCC development. The dendrogram is generated using average-linkage with a Poisson-based distance metric based on Cai et al. (2004). Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key placed between the dendrogram and the heatmap denotes the sample type: white is bulk normal; epithelial brushings from never, former, and current smokers shown in dark green, light green, and cyan, respectively; metaplasia in blue; dysplasia in purple; carcinoma in situ in yellow; and invasive carcinoma in red. Additional nucleotides identified by the XBP-SAGE method (Chapter II) are shown. Tags prefixed with an asterisk (*) indicate that the additional nucleotides were needed for a successful mapping.
resulted in a 2 gene upregulated signature (\textit{S100A2} and \textit{KRT6A}) with a sensitivity and specificity of 93.4% and 96.3%, respectively. Correspondingly, a 2 gene downregulated signature (\textit{C16orf89} and \textit{C5orf32}) was identified with a sensitivity and specificity of 93.1% and 93.8%, respectively. The upregulation of both \textit{S100A2} and \textit{KRT6A} is strongly supported by the three SCC microarray datasets (Figure 4.15). However, neither \textit{C16orf89} nor \textit{C5orf32} are present on these arrays and their downregulation in early stage lesions is yet to be confirmed.

Both upregulated genes have well established roles in squamous cell differentiation and lung cancer. \textit{S100A2} has been previously shown to be upregulated in almost all cases of SCC, and the level of expression is concordant with the amount of abnormality observed in pre-malignant and early stage lesions (Smith, 2004). Moreover, \textit{S100A2} transcription has been shown to be positively regulated by direct binding of TP63, a TP53 paralogue, to the promoter (Hibi, 2003). \textit{TP63} overexpression is also strongly implicated by the SAGE data, and is ranked eighth in the list of pre-malignant tags (Figure 4.13.1). The functions of \textit{C16orf89} and \textit{C5orf32} are currently unknown, and the characterization of these genes and their potential involvement in SCC remains to be explored.

\subsection{4.3.2.2 Changes associated with malignancy}

A set of 143 upregulated and 26 downregulated tags was associated with the malignant phenotype (Figures 4.16.1-4.16.2 and 4.17, respectively). Again, the unusual invasive sample I22 is misclassified. A GO enrichment analysis did not reveal any statistically significant terms.

Based on LOOCV, the sensitivity and specificity of these signatures are 61.4% and 85.3%, respectively, for the upregulated tags and 75.2% and 92.1%, respectively, for the downregulated tags. Optimal signatures were constructed to maximize the selectivity. This resulted in a 4 gene upregulated signature (\textit{MCM7}, \textit{SLC6A8}, \textit{CKS1B}, \textit{ATP1B3}) with a sensitivity and specificity of 81.4% and 92.1%, respectively. Correspondingly, a 2 gene downregulated
Figure 4.15: Microarray validation of optimal metaplasia progression-associated upregulated gene signature. Each box plot shows the sample type (x-axis) and log-hybridization (y-axis). Each box-and-whisker pattern depicts, from top to bottom, the maximum observed value, upper quartile, median value, lower quartile, and minimum observed value. Additional circles indicate outliers. Note: KRT6A is not represented on the Wachi dataset array.
Figure 4.16.1: Heatmap with sample-wise hierarchical clustering of the first 72 of 143 tags upregulated in malignant stages of SCC development. The dendrogram is generated using average-linkage with a Poisson-based distance metric based on Cai et al. (2004). Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key placed between the dendrogram and the heatmap denotes the sample type: white is bulk normal; epithelial brushings from never, former, and current smokers shown in dark green, light green, and cyan, respectively; metaplasia in blue; dysplasia in purple; carcinoma in situ in yellow; and invasive carcinoma in red. Additional nucleotides identified by the XBP-SAGE method (Chapter II) are shown. Tags prefixed with an asterisk (*) indicate that the additional nucleotides were needed for a successful mapping.
Figure 4.16.2: Heatmap with sample-wise hierarchical clustering of the final 71 of 143 tags upregulated in malignant stages of SCC development. The dendrogram is generated using average-linkage with a Poisson-based distance metric based on Cai et al. (2004). Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key placed between the dendrogram and the heatmap denotes the sample type: white is bulk normal; epithelial brushings from never, former, and current smokers shown in dark green, light green, and cyan, respectively; metaplasia in blue; dysplasia in purple; carcinoma in situ in yellow; and invasive carcinoma in red. Additional nucleotides identified by the XBP-SAGE method (Chapter II) are shown. Tags prefixed with an asterisk (*) indicate that the additional nucleotides were needed for a successful mapping.
Figure 4.17: Heatmap with sample-wise hierarchical clustering of the 26 tags downregulated in the malignant stages of SCC development. The dendrogram is generated using average-linkage with a Poisson-based distance metric based on Cai et al. (2004). Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key placed between the dendrogram and the heatmap denotes the sample type: white is bulk normal; epithelial brushings from never, former, and current smokers shown in dark green, light green, and cyan, respectively; metaplasia in blue; dysplasia in purple; carcinoma in situ in yellow; and invasive carcinoma in red. Additional nucleotides identified by the XBP-SAGE method (Chapter II) are shown. Tags prefixed with an asterisk (*) indicate that the additional nucleotides were needed for a successful mapping.
signature (\textit{CRIP2}, \textit{TMPRSS2}) was identified with a sensitivity and specificity of 75.2\% and 98.3\%, respectively. The dysregulation of all 6 genes is strongly supported by the three SCC microarray datasets (Figures 4.18.1-4.18.2).

Both \textit{MCM7} and \textit{CKS1B} are key, highly conserved players in eukaryotic cell division. \textit{MCM7} is part of a heterohexamer helicase complex (consisting of MCM2-7) and is critical to the initiation of DNA replication (Maiorano, 2006). \textit{MCM7} has become increasingly prominent as an accurate tumour cell proliferation marker in a variety of cancers, outperforming the current gold standard marker Ki67 (Xue, 2003; Li, 2005; Facoetti, 2006a; Facoetti, 2006b; Feng, 2008; Nishihara, 2008). A direct role in SCC formation has been suggested from a mouse model where \textit{MCM7} was transfected with a \textit{KRT14} promoter; these mice developed SCC when exposed to chemical carcinogenesis, whereas wild-type mice did not (Honeycutt, 2006). \textit{CKS1B} plays a less direct, but still critical, role as a cell cycle regulator. The protein promotes SKP2-mediated degradation of the tumour-suppressors CDKN1A and CDKN1B, which normally bind to CCNE1-CDK2 and CCND1-CDK4 complexes to maintain a quiescent state (Ganoth, 2001; Spruck, 2001). The proposed role of \textit{CKS1B} and SKP2 in degrading CDKN1B has been demonstrated in oral SCC (Kitajima, 2004).

As \textit{MCM7} and \textit{CKS1B} activity requires close cooperation with a number of other molecules (Figure 4.19), the expression of tags mapped to the genes for the origin of replication complex (\textit{ORC1L}-\textit{ORC6L}), other members of the \textit{MCM} family (\textit{SRF}, \textit{MCM2-10}), known co-participants in the pre-replication complex (\textit{CDC6}, \textit{CDC7}, \textit{CDT1}, \textit{DBF4}, \textit{CDC45L}, \textit{GMNN}), and cyclin-CDK complex members (\textit{CCND1}, \textit{CCNE1}, \textit{CDK2}, \textit{CDK4}, \textit{CDKN1A}, \textit{CDKN1B}) were examined in detail. Many of these genes are not detected, but it is impossible to determine whether this is due to a complete lack of expression or because expression levels are too low to be captured by SAGE. Although individually the genes that are detected are not expressed at levels necessary to provide statistical significance, taken as a whole a clear association with
Figure 4.18.1: Microarray validation of optimal malignant progression-associated upregulated gene signature. Each box plot shows the sample type (x-axis) and log-hybridization (y-axis). Each box-and-whisker pattern depicts, from top to bottom, the maximum observed value, upper quartile, median value, lower quartile, and minimum observed value. Additional circles indicate outliers.
Figure 4.18.2: Microarray validation of optimal malignant progression-associated downregulated gene signature. Each box plot shows the sample type (x-axis) and log-hybridization (y-axis). Each box-and-whisker pattern depicts, from top to bottom, the maximum observed value, upper quartile, median value, lower quartile, and minimum observed value. Additional circles indicate outliers.
Figure 4.19: A model of MCM7 and CKS1B function based on current knowledge.

A) Role of CSK1B in activating the Cyclin-CDK complex by targeting CDKN1A/CDKN1B for ubiquitin-mediated degradation. B) Role of the MCM family and other key proteins in initiating DNA replication. The content and style of bottom figure (B) is adapted from Maiorano et al. (2006). The version shown here labels the biomolecules using official gene names, includes DBF4 in a complex with CDC7, includes CDC45L as a member of the pre-initiation complex, includes CCNE1 in a complex with CDK2, and makes reference to the specific members of the ORC and RPA gene families. ??? denotes unknown members.
malignant progression is present. In particular, several other members of the MCM heterohexamer complex are upregulated (MCM2, MCM3, MCM5), as are CDK2 and CDT1, which bridge the MCM complex to the ORC (Figure 4.20) (Tsuyama, 2005). Moreover, CDKN1B, a key negative regulator of cell division, shows some evidence of downregulation as early as pre-malignancy. It is enticing to speculate that the increase in CKS1B acts to further abolish the function of this TSG through the SCF(SKP2)-mediated ubiquination mechanism.

Possible roles for the remaining upregulated malignancy-associated genes are unclear. SLC6A8 is responsible for transporting creatine and creatine analogues both in and out of cells (Sora, 1994). However, none of the tags corresponding to known creatine kinase genes, which catalyze the conversion of creatine from or to creatine phosphate and provide a source of or reservoir for ATP, show evidence of dysregulation. ATP1B3 is one of a family of regulatory subunits that heterodimerize with a catalytic subunit (one of ATP1A1, ATP1A2, or ATP1A3) to form an active Na⁺/K⁺-ATPase (Lingrel, 1990). Only the tag for ATP1A1 is detected and is expressed in all samples, but there is no evidence of dysregulation. The significance of increased ATP1B3 in malignant progression remains to be investigated. However, this gene has been shown to play a role in regulating T and B lymphocyte proliferation and may simply be part of an immune response at the site of malignancy (Chiampanichayakul, 2002).

CRIP2 is a member of the diverse LIM domain-containing proteins that are characterized by a two tandem zinc fingers that appear to mediate protein-protein interactions, rather than facilitating DNA binding (Karim, 1996; Zheng, 2007). CRIP2 has not been extensively characterized. However, it has been shown to act as a bridge between SRF (a.k.a. MCM1) and several GATA proteins, initiating a transcriptional program that causes the differentiation of fibroblasts into smooth muscle cells (Chang, 2003). One hypothesis is, given the increase in the MCM family discussed above, the loss of CRIP2 results in the abolition of a transcriptional program that may act against the progression of pre-malignant lesions to SCC.
Figure 4.20: Expression of select members of the MCM family and related genes.

Each plot shows the normalized expression (tags/50,000) (y-axis) for all 40 SAGE libraries (x-axis). MCM2, MCM3, MCM5, CDK2, and CDT1 show evidence for upregulation in malignant samples. CDKN1B shows evidence for downregulation in both pre-malignant and malignant samples.
TMPRSS2 is a serine protease, a family of genes that performs diverse physiological roles by cleaving a variety of target proteins. The biological function of TMPRSS2 is yet to be determined. However, TMPRSS2 has gained recent attention in prostate cancer pathogenesis as it forms a fusion protein with a number of ETS-family transcription factors (e.g. ERG, ETV1, and ETV4) in the majority of cases (Tomlins, 2005; Tomlins, 2006). TMPRSS2 overexpression is androgen-regulated, and the fusion of the 5′-UTR to an oncogenic ETS partner may be an early causal event in prostate tumourigenesis. Although there is no evidence that TMPRSS2 fusion genes are present in lung cancer, its identification here suggests it may have a more general role as a tumour suppressor.

4.3.2.2.3 Changes associated with an invasive phenotype

Using the complete dataset, a set of 7 upregulated tags was associated with an invasive phenotype (no set of downregulated was statistically significant). However, due to the small number of invasive samples (6), the unusual sample I22 introduces a substantial degree of interference in the selection of candidates. Given the results of both the k-means analysis and candidate selection procedure for pre-malignant and malignant tags, the contamination of this sample by normal epithelium has been reasonably established. Therefore, an additional candidate selection run was performed with I22 omitted, resulting in a large improvement (Figure 4.21).

The modified sample set produced a set of 29 upregulated tags associated with the invasive phenotype (again, no set of downregulated tags was statistically significant) (Figure 4.21). The tags mapped to 26 unique Unigene entries and a GO enrichment analysis of these genes revealed the terms “extracellular matrix” (GO:0031012; FDR<0.1%) and “collagen” (GO:0005581; FDR=1.9%). This is implied from the upregulation of the genes COL1A2, COL3A1, COL6A3, LAMA1, MFAP2, MMP11, MMP12, and SPP1. These terms are consistent
Figure 4.21: Heatmap with sample-wise hierarchial clustering of 29 tags upregulated in the invasive stage of SCC development. The dendrogram is generated using complete-linkage with a Poisson-based distance metric based on Cai et al. (2004). Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key placed between the dendrogram and the heatmap denotes the sample type: white is bulk normal; epithelial brushings from never, former, and current smokers shown in dark green, light green, and cyan, respectively; metaplasia in blue; dysplasia in purple; carcinoma in situ in yellow; and invasive carcinoma in red. Additional nucleotides identified by the XBP-SAGE method (Chapter II) are shown. Tags prefixed with an asterisk (*) indicate that the additional nucleotides were needed for a successful mapping.
with the invasive phenotype, as the degradation of and interaction with the extracellular matrix and surrounding tissue is an essential feature of this process.

Based on LOOCV, the sensitivity and specificity of this signature is 52.5% and 91.4%, respectively. An optimal signature was constructed to maximize the selectivity. This resulted in a 3 gene upregulated signature (\textit{MMP11, GAPDH, SHMT2}) with a sensitivity and specificity of 64.5% and 91.5%, respectively. Again, there is excellent concordance with the microarray validation datasets (Figure 4.22).

\textit{MMP11}, a member of the matrix metalloprotease family, has been strongly implicated in invasion by its presence in the stromal cells surrounding the invasive front of most, and likely all, epithelial cancers (Basset, 1993; Rio, 2005). A recent study identified \text{COL6A3} as a specific \text{MMP11} substrate (Motrescu, 2008). Interestingly, an antisense artefact tag from \text{COL6A3} was the fourth ranked tag in the list of invasive candidate tags. The expression of the \text{MMP11} protein was further explored by immunochemistry. A commercial antibody was obtained and its activity confirmed by Western blot of a protein extract from an MCF cell line transfected with the \text{MMP11} gene. Non-transfected MCF protein extract was used as a negative control (both extracts were a gift from Ulrich auf dem Keller, Overall Lab, UBC). No band was observed in the MCF\textsuperscript{WT} sample, and the presence of a band at the expected 55kDa was seen in the MCF\textsuperscript{MMP11} sample (the \text{MMP11} precursor is 54,618Da) (Figure 4.23). The antibody was then used on archival tissue sections from normal lung, breast cancer, and SCC. As expected, staining was absent in the lung cancer negative control and very strong in cases of breast carcinoma, where \text{MMP11} is known to be expressed and where it was first implicated in invasion (Figure 4.24A-B) (Bassett, 1993). Dark staining was observed both in breast tumour cells, and surrounding fibroblasts and collagenous structure. No staining was observed in normal lung tissue (Figure 4.24C). In lung cancer, \text{MMP11} was present in stromal cells surrounding invasive tumour cells (Figure 4.24D-F). Staining was noticeably lighter, which may indicate that \text{MMP11} is present in
Figure 4.22: Microarray validation of optimal invasive progression-associated upregulated gene signature. Each box plot shows the sample type (x-axis) and log-hybridization (y-axis). Each box-and-whisker pattern depicts, from top to bottom, the maximum observed value, upper quartile, median value, lower quartile, and minimum observed value. Additional circles indicate outliers.
Figure 4.23: Western blot confirming correct activity of MMP11 antibody. The expected size of the MMP11 precursor is 54.6kDa. Lane 1: 5μL Benchmark Pre-stained Marker; Lane 2: 10μL wild-type MCF7 cell lysate; Lane 3: 10μL MMP11 transfected MCF7 cell lysate.
Figure 4.24: MMP11 detection by immunohistochemistry. (A) squamous cell lung cancer negative control, (B) breast carcinoma positive control, (C) normal lung, (D,E,F) squamous cell lung cancer from 3 different patients. The presence of MMP11 is indicated by brown staining. All photographs were taken at 20X magnification, except the normal lung section which was taken at 40X magnification. All sections are counterstained with hematoxylin. Representative areas of malignant cells (*) and stromal tissue (S) are marked.
lower levels in SCC, although this is difficult to state definitively given the qualitative nature of immunohistochemistry.

Changes in cellular metabolic processes are of well-established importance in tumour growth and survival. Both GAPDH and SHMT2 are important players in metabolic systems; the former is a key part of the glycolysis pathway, which drives anaerobic energy metabolism, and the latter is important for nucleotide metabolism, which is necessary for transcription, translation, and DNA repair. An increase in glycolytic energy production as a means of overcoming the hypoxic conditions present in tumours, known as the Warburg effect, has been known for decades (Warburg, 1956). Mitochondrial SHMT2, and its cytosolic isoenzyme SHMT1, catalyzes the simultaneous, reversible conversion of L-glycine to serine and 5,6,7,8-tetrahydrofolate (THF) to $N^5,N^{10}$-methylene-THF (Schirch, 1982). $N^5,N^{10}$-methylene-THF is an essential substrate for thymidine biosynthesis (Figure 4.25). Increased SHMT2 expression and activity has been observed in a variety of tumours, including the lung, presumably to support the rapid increase in the rate of mitosis (Tendler, 1987). Interestingly, the other two enzymes in the cyclic pathway that regenerates $N^5,N^{10}$-methylene-THF and drives thymidine production are major, long-standing chemotherapeutic targets. TYMS, which utilizes $N^5,N^{10}$-methylene-THF to produce thymidine and the byproduct DHF, is the target of uracil analog 5-fluorouracil (5-FU), a principal agent in treating colon and pancreatic cancers (Danenberg, 1977). DHFR, which catalyzes the reduction of DHF back to THF, is the target of the folate analog methotrexate, the first chemotherapeutic agent (Schweitzer, 1990). It is still widely used to treat a variety of cancers, including those found in the lung. For this reason, serine hydroxymethyltransferases, the remaining component of this cyclic reaction, have been proposed as a potential treatment target (Agrawal, 2003; Rimpi, 2007). Moreover, there is evidence that SHMT2, rather than SHMT1, is the primary driver of thymidine biosynthesis (Fu, 2001).
Figure 4.25: Role of serine hydroxymethyltransferase in thymidine biosynthesis. SHMT1 and SHMT2 catalyze the conversion of serine to glycine, transferring the single carbon to tetrahydrofolate (THF) to form $N^5,N^{10}$-methylene-THF. Thymidylate synthase (TYMS) catalyzes the conversion of uracil to thymidine, producing dihydrofolate (DHF). Dihydrofolate reductase (DHFR) catalyzes the return of DHF to THF, and the cycle begins again. Several chemotherapeutic inhibitors that target TYMS or DHFR are noted.
4.3.3 Comparison to existing squamous cell lung cancer profiles

In order to explore the increased information potential of fully delineated developmental stages, the tags corresponding to lung SCC genes identified by previous transcriptome profiling studies were examined in this dataset. A prior SAGE study identified 10 tags that are upregulated in SCC compared to a normal bronchial epithelium (NHBE) cell line (Nacht, 2001) (Figure 4.26.1A). Of the three microarray studies, only the Bhattacharjee dataset identified specific differentially expressed candidates (Bhattacharjee, 2001) (Figure 4.26.1B). For the Erez/Dehan and Wachi datasets, the log-transformed hybridization values were subjected to Student’s $t$-test to identify the top 10 upregulated and downregulated genes according to $p$-value (Figure 4.26.2). While there is general agreement between the SAGE dataset presented here and these four prior studies, three major issues are evident. First, genes that appear to be associated with smoke-exposure are evident in the Nacht dataset. This is likely due to the use of an NHBE cell line for comparison rather than smoke-exposed epithelium, which is present in the majority of lung cancer patients. Second, genes that are strongly associated with the pre-malignant transformation of bronchial epithelium to metaplasia are highly overrepresented. This is particularly evident in the Bhattacharjee candidates, although it occurs to a large extent in all four datasets. Third, genes expressed by non-epithelial cells present in the surrounding stroma are present in the Nacht dataset, again likely due to the use of the NHBE cell line, and especially in the downregulated candidates in the Erez/Dehan and Wachi datasets. This comparison highlights the potential difficulties in identifying candidate lung cancer genes by global transcriptome profiling, and are relevant issues in the study of other solid tumours.
Figure 4.26.1: Heatmap of SAGE tag expression of candidate SCC genes identified from the Nacht and Bhattacharjee datasets. The Nacht genes are the top heatmap (A) and the Bhattacharjee genes are the bottom heatmap (B). Both sets of genes are those identified by the original study authors. Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key at the top of the heatmap denotes the sample type: white is bulk normal; epithelial brushings from never, former, and current smokers shown in dark green, light green, and cyan, respectively; metaplasia in blue; dysplasia in purple; carcinoma in situ in yellow; and invasive carcinoma in red.
Figure 4.26.2: Heatmap of SAGE tag expression of candidate SCC genes identified from the Erez/Dehan and Wachi datasets. The Erez/Dehan candidates are the top heatmaps corresponding to downregulated (A) and upregulated (B) genes. The Wachi candidates are the bottom heatmaps corresponding to downregulated (C) and upregulated (D) genes. Genes were identified by selecting those with the lowest $p$-value as determined by Student’s $t$-test. Each cell of the heatmap is based on the percentage expression across all samples to provide consistent colouration. A colour key shows the range of colours from 0-10% expression, with values exceeding this shown in bright red. A colour key at the top of the heatmap denotes the sample type: white is bulk normal; epithelial brushings from never, former, and current smokers shown in dark green, light green, and cyan, respectively; metaplasia in blue; dysplasia in purple; carcinoma in situ in yellow; and invasive carcinoma in red.
4.4 DISCUSSION

This study presents an initial view of the global transcriptional changes that occur over the course of squamous cell lung cancer development. These data suggest that a great deal, and perhaps the majority, of the changes to gene expression during the development of SCC occur with the transformation of normal epithelium to the squamous cell type. Although this process is a pre-condition for the development of SCC – hence the name – and may result in a tumour permissive state, it is not itself a malignant phenotype. Unfortunately, this results in a conflation of squamous differentiation-associated changes and malignancy-associated changes when comparing normal epithelium and de facto carcinomas, as is typically done. Gene expression changes associated with the invasive phenotype appear to be relatively limited. This is consistent with the fact that a high number (>90%) of in situ carcinomas eventually become invasive. The variability in the cellular composition of the bulk invasive samples used in this analysis may disguise changes to some extent. However, based on what is known about the candidate genes, the notion that a substantial number of invasion-promoting genes may arise from the cellular milieu surrounding a tumour is strongly supported. In other words, invasion may depend largely on gene expression changes in non-tumour cells. Of course, the possibility of a number of distinct, mutually exclusive determinants of this phenotype, which would complicate the identification of a consistent invasive signature, cannot be ruled out.

A number of gene expression signatures were identified using computational strategies designed to maximize the information content of SAGE profiles. Both the improvement to tag to gene mapping presented in Chapter II and the statistical model presented in Chapter III were vital elements of the methodology to analyze this complex dataset and identify high quality signatures that correspond to the metaplastic, malignant, and invasive phenotypes that characterize SCC development. When compared to three separate gene expression studies using
normal lung tissue and SCC samples, excellent agreement was found for all of the genes that comprise these signatures. The results of this effort are sufficient to develop a convincing portrait of key gene expression changes associated with SCC development, including a number of novel genes that are promising targets of future study.

The metaplastic phenotype is associated with the upregulation of $S100A2$ and $KRT6A$, and the downregulation of $C16orf89$ and $C5orf32$. While both upregulated genes have been previously implicated in keratinocyte or squamous differentiation, the two downregulated genes are completely uncharacterized and are attractive targets for exploratory studies.

The malignant phenotype is associated with the upregulation of $MCM7$, $SLC6A8$, $CKS1B$, and $ATP1B3$, and the downregulation of $CRIP2$ and $TMPRSS2$. $MCM7$ and $CKS1B$ are key participants in regulating the initiation of DNA replication and may represent a common modality that initiates SCC malignancy – likely in establishing and maintaining uncontrolled cell division. The role of the remaining genes is unclear, but both $CRIP2$ and $TMPRSS2$ are particularly attractive as targets for further study given the former’s possible role in bridging transcription machinery to particular loci, and the recent discovery of the latter’s fundamental importance in prostate cancer initiation.

The invasive phenotype is associated with an upregulation of $MMP11$, $GAPDH$, and $SHMT2$. $MMP11$ is of particular interest given the role the matrix metalloproteinase family in extracellular matrix degradation and remodelling that is considered a requirement for successful tumour invasion. An examination of MMP11 in invasive carcinomas by immunohistochemistry supports this hypothesis. Given the role of $GAPDH$ in anaerobic energy production, a long established feature of cancer etiology and progression, this gene is more an affirmation of this study’s methodology than a target of future work. Finally, the close relationship of $SHMT2$ to thymidine biosynthesis and a pathway that has long been a target of conventional chemotherapy is alluring and may represent a drug target with the potential to improve an existing strategy of
In addition, an analysis of SAGE libraries obtained from bronchial epithelium from never, former, and current smokers revealed a distinct signature associated with recent exposure to tobacco smoke. These genes were overwhelmingly associated with the Phase I and Phase II enzyme system responsible for xenobiotic metabolism. Although no gene expression changes that persist once an individual has quit were found, downregulation of the Phase I enzyme \textit{EPHX1} to levels below that of a never smoker may represent a distinct effect of cessation following long-term exposure. Although EPHX1 activity is required to generate a mutagenic form of the known pro-carcinogen benzo[a]pyrene, the primary role of the enzyme is protective and a decrease in its activity may render the epithelium more susceptible to the harmful effects of environmental toxins.

Finally, the data presented represents one of the most comprehensive catalogues of gene expression change in the step-wise development of an invasive carcinoma from normal tissue. The data and statistical analysis are of significant value to lung cancer researchers in terms of defining the temporal contribution of particular genes in SCC development. The dataset is amenable to meta-analyses that combine these data with other high-throughput experiments to identify important progression-related events (e.g. copy number variations or aberrant methylation).
BIBLIOGRAPHY


5.1 OVERALL DISCUSSION AND CONCLUSION

The initial goal of this thesis was to determine the relationship of the transcriptome to the formation and progression of squamous cell lung cancer (SCC), a common tumor subtype that contributes to a large proportion of total cancer deaths (Canadian Cancer Society, 2008; Jemal, 2007). The use of serial analysis of gene expression (SAGE) to generate the necessary transcriptome profiles was appealing due to the unbiased and comprehensive nature of the technique (Velculescu, 1995). While a relatively straightforward strategy involving existing, and widely adopted, approaches to SAGE analysis was envisioned, these techniques quickly proved to be problematic. These challenges were exacerbated by the lack of additional biological material to fully validate findings. Indeed, a common approach in large-scale genomic studies is to sacrifice optimization at the data analysis phase of an investigation for the sake of rapid transition to more targeted follow-up study where, presumably, false findings can be quickly identified.

One of the first issues that became clear was that, without the ability to validate, SAGE tag to gene mapping would be an area of continuous frustration. The veracity of SAGE analysis hinges on the ability to determine the original transcript from which an observed tag arose. Often there are several transcripts that share the same theoretical SAGE tag, resulting in a situation where a differentially expressed gene cannot be unambiguously identified. This is typically addressed by either: a) simply accepting that some candidate tags are lost to this issue, or b) performing additional experiments (e.g. qPCR) to determine which of the possible candidates is differentially expressed. The former is acceptable when the number of candidates is relatively large and/or the study objective is more concerned with uncovering larger biological
themes, rather than specific genes. This proved not be the case in the dataset central to this thesis. The latter, of course, is only feasible in cases where biological material is not limiting. Chapter II describes the strategy developed to increase the effective length of tag sequences (usually by 20%) to reduce the problem of transcript ambiguity. However, it was also discovered that many of the tag to gene mappings that were previously considered unambiguous actually arose from other sources – most often as artefacts from highly expressed tags with close sequence similarity, or from a specific type of SAGE artefact that results in the capture of antisense sequences; but also novel transcript variants of known genes, or entirely novel loci. Applying this strategy proved invaluable in the analysis of the SCC dataset, as 35% of the candidate tags identified benefited from either reduced ambiguity or re-assignment to an improved source transcript.

The multiple developmental stages of SCC represented by the SAGE dataset presented challenges for statistical analysis. Standard hypothesis tests were inappropriate for comparisons involving multiple groups, and the heterogeneity of the sampled tissues represented an additional level of complexity not adequately addressed by simple approaches. Existing work by several researchers had begun to tackle the challenge of determining statistical significance when comparing multiple groups of SAGE libraries (Baggerly, 2003; Baggerly, 2004; Vencio, 2004; Lu, 2005). Although these methods provided an excellent framework with which to begin the analysis of the SCC dataset, it became clear that some of the existing assumptions used in these statistical models, although fundamentally sound, were inadequate. At 40 libraries, the sheer size of the SCC dataset was the largest factor in enabling these inaccuracies to be uncovered. The major issue was the assumption that additional variance could be adequately represented by a hierarchical model. Such models consist of a unimodal prior distribution to supplement the binomial or Poisson distribution that accounts for the variance that must arise from sampling a small subset of tags from the large initial pool of tags generated from biological material.
Initially, effort was spent on determining a better choice for such a distribution, which was met with some limited success. Chapter III describes the eventual development of a statistical approach that utilizes a Poisson mixture model, rather than the previously proposed hierarchical model, to assign significance to SAGE library comparisons. This model was inspired by the structure of the SCC dataset, and while initially appearing to draw its strength by accounting for the extensive sample heterogeneity, its strong performance on previously published datasets, several of which were generated from relatively pure biological material, revealed a general applicability to all SAGE data.

Chapter IV describes the final analysis of the SCC dataset. To date, it represents the most comprehensive profile of the transcriptome during the early development of a solid tumour. While the sample heterogeneity discussed above was a confounding factor throughout this research, a strict adherence to sound statistical methodology and the use of the techniques described in Chapters II and III resulted in the identification of several hundred genes associated with several steps in the development and progression of SCC.

A targeted analysis of the subset of samples drawn from brushings of bronchial epithelium from never, former, and current smokers was performed to gain insight into the role of tobacco exposure on the normal lung transcriptome. This analysis revealed a large response to such exposure, primarily from the Phase I and Phase II enzyme system. However, there was no evidence that any of these changes persist once an individual ceases exposure. One intriguing observation was a former smoker-specific decrease in the expression of *EPHX1*, a known player in generating a highly mutagenic product from the pro-carcinogen benzo[a]pyrene (Denissenko, 1996). A decrease in the activity of this enzyme in former smokers may leave the epithelium more susceptible to future mutagenesis by, for example, air-borne carcinogens. A larger, longitudinal study with careful control of variables such as the timeline of past exposure and cessation are required to determine if this is a true effect. Of interest was the lack of any gene
expression changes associated with the acute response persisting in later stages of SCC development. It appears that upon the formation of squamous metaplasia, the normal molecular response to tobacco smoke ceases altogether. This supports the notion that these early lesions are comprised of cells that are highly susceptible to further mutagenesis.

A more comprehensive analysis was then performed that focussed on the pre-malignant and malignant stages of SCC. A number of broad insights were gained. First, the largest and most consistent set of gene expression changes is associated with the development of pre-malignant squamous metaplasia. This has two major implications for future studies of SCC and, quite likely, other tumour types: 1) the use of normal bronchial epithelium as the “baseline” from which to assess gene expression changes in SCC is unwise, at least if the objective is to identify those associated with malignancy; and 2) the consistency of these pre-malignant changes means they are more likely to be identified as significant than malignant changes, which display a far more varied incidence and level of expression. Second, the number of changes associated with acquiring an invasive capability is remarkably small, even when considering the increased heterogeneity of the samples representing this phenotype.

Although not presented in this thesis, the development of the computational strategies involved many modifications and continual improvements. Access to the Westgrid resource, a high-performance computing (HPC) cluster containing over 1500 CPUs, allowed complex procedures to be applied to the large SCC dataset without major concerns about execution time. For example, the cross-validation and resampling strategies described in Chapter IV would take several months to execute on a typical desktop computer. The use of an HPC cluster allowed these procedures to be performed in a few days or less, and this allowed exceptional freedom to improve or re-run entire analysis pipelines when additional data became available or new ideas were formed. The difference in the quality of the first gene signatures developed at the beginning of this project and those presented in this thesis is enormous.
The final set of 13 genes most strongly associated with the complete transformation of bronchial epithelium to pre-malignant metaplastic lesions, to carcinoma in situ, and finally, the acquisition of the invasive phenotype represent a compelling set of targets for additional study. While the upregulated S100A2 and KRT6A genes are familiar participants in SCC and squamous differentiation, the downregulation of C16orf89 and C5orf32 in metaplasia is equally strong and presents an opportunity to investigate two entirely uncharacterized genes that are likely to play important roles in differentiation (Smith, 2004). MCM7 and CKS1B represent genes that play important and central roles in promoting DNA replication and their consistent increase upon progression to malignancy may represent a common outcome of the disruption of a range of TSGs and oncogenes (Ganoth, 2001; Spruck, 2001; Maiorano, 2006). However, the significance of the increased expression of SLC6A8, a creatine transporter, and ATP1B3, the regulatory subunit of Na⁺/K⁺-ATPase, is unclear and remains to be investigated. The decreased expression of CRIP2, a possible bridge between the general transcription machinery and specific transcription factors, may represent a central point at which the expression of many malignancy-associated genes is altered (Chang, 2003). A similar decrease in TMPRSS2, which is now established as a playing a critical role in prostate tumour initiation by driving the expression of certain oncogenes through gene fusion events, suggests this gene may have tumour-suppressing properties in its own right (Tomlins, 2005; Tomlins, 2006). The top candidate identified in association with invasive samples was MMP11. This gene has been established as an important player in tumour invasion in a wide variety of tumours, although the exact mechanism remains unclear (Basset, 1993; Rio, 2005). A preliminary examination of its protein expression by immunohistochemistry supports its importance in SCC. Ironically, a likely source of MMP11 is the stromal cells surrounding invasive cancer cells, and the impurity of the samples analyzed in this thesis may have facilitated its identification here. The finding of the metabolic gene GAPDH in association with invasion is unsurprising, as the increased reliance on glycolysis is a
long established property of cancer that is used to overcome the hypoxic environment faced by a growing tumour (Warburg, 1956). Finally, the increase in SHMT2 is interesting due to its role in driving thymidine biosynthesis in close cooperation with TMYS and DHFR, two of the longest standing targets of conventional chemotherapy (Danenberg, 1977; Schweitzer, 1990).

5.2 FUTURE PROSPECTS

The research described in this work presents the opportunity for further exploration in two major areas: 1) the SAGE method as a general approach for gene expression profiling and how to best utilize data captured by this technique; and 2) the molecular basis of squamous cell lung cancer progression and how this information can improve the prospects for individuals affected by this disease.

The XBP-SAGE approach presented in Chapter II demonstrates that additional sequence information can be extracted from SAGE data, resulting in a large increase in the fidelity of tag to gene mapping. This approach should be adopted as a standard method of data processing. Furthermore, a revisiting of the cost-benefit of adopting newer protocols, which produce longer tags, may be worthwhile for researchers that are considering the SAGE technique. For example, the cost of sequencing is almost 50% greater when utilizing the LongSAGE technique. If a proposed study plans to produce a number of profiles, the cost savings from using the shorter variant would allow the production and sequencing of libraries from additional biological replicates. Not only could additional libraries assist in the determination of extra nucleotides, but the statistical power gained may outweigh the slight loss in mapping fidelity.

The Poisson mixture model presented in Chapter III is demonstrably superior in determining the significance of gene expression changes when comparing SAGE libraries. However, a statistical model can only be shown to be an improvement over an existing one and there is always the possibility that an entirely novel approach exists to better describe a particular
type of data. Nevertheless, several avenues exist for incremental improvements to existing approaches. Developing strategies to incorporate multiple SAGE tags to determine significance, rather than assessing them individually, are likely to bear fruit. For example, as discussed in Chapter II, there are several mechanisms that result in the capture of multiple tags from a single transcript. Better estimates of significance could be determined by utilizing this property. A similar approach could be developed for tags arising from genes under common regulatory control.

The analysis of the changes in the transcriptome during the early development of SCC presented in Chapter IV provides a plethora of future work. The observation that most, and perhaps the majority, of changes occur during the development of pre-malignant squamous metaplasia invites additional gene expression profiling that compares tissue from a much larger set of these lesions to more advanced stages of tumour progression. Such a study would likely produce a much clearer picture of malignant transformation and reveal more dynamic aspects of transcriptome change not possible with the dataset used here. Nevertheless, the genes identified in Chapter IV represent excellent candidates for future study. An obvious next step is to confirm their specific association with a given stage of progression by performing qPCR validation on a larger panel of new samples representing the complete spectrum of SCC development. Western blot experiments or immunohistochemical staining can be used to confirm the presence of the corresponding protein product in cases where an antibody is available.

Assuming these experiments confirm their involvement in a given stage of progression, a number of general strategies can be envisioned for several candidates. In the case of EPHX1, which may undergo downregulation in the bronchial epithelium following smoking cessation, a longitudinal study examining the incidence of pre-malignant lesion formation or progression of existing lesions in individuals relative to this gene’s activity is a possibility. In the case of C16orf89 and C5orf32, which show a very consistent loss of expression with the formation of
metaplasia, almost nothing is currently known. Both are evolutionarily conserved in vertebrates, but there are no within-species similarities to other genes or recognizable protein domains, so their function is likely quite specialized (Wheeler, 2008). These genes could potentially be cloned to facilitate the isolation of their protein product, whereupon techniques like a pull-down assay using epithelial cell lysates may help identify potential interactions with proteins having known functions. A strategy like an siRNA knockdown assay may also identify a specific phenotype. A similar approach may also help elucidate the role of CRIP2 and TMPRSS2, which are downregulated during the progression to malignancy. In this case, a knockdown assay in a squamous cell line could be followed by looking for an increase in malignant phenotypes (e.g. increased mitotic rate). Experiments with more direct clinical relevance are appealing for genes found to be upregulated during the progression to malignancy or the acquisition of invasive capabilities. In the case of the former, MCM7 is particularly attractive, given its strong performance as a prognostic marker in a range of tumours, including other lung cancer subtypes (Xue, 2003; Li, 2005; Facoetti, 2006a; Facoetti, 2006b; Feng, 2008; Nishihara, 2008). In the case of the latter, MMP11 shares similar promise as a prognostic indicator and may have additional utility as a molecular indicator of invasiveness in very early tumours. For example, although carcinoma in situ appears locally confined and is highly treatable by surgery alone, the presence of MMP11 may identify a subset of patients who are at increased risk of recurrence and would benefit from adjuvant chemotherapy or aggressive surveillance.

In conclusion, the use of genome-wide experimental approaches such as global gene expression profiling, along with the use of sophisticated and carefully applied bioinformatic techniques, have the potential to increase our understanding of cancer and uncover new avenues of attack against this deadly disease. The continued application and improvement of these approaches are a cause for great optimism, and will undoubtedly benefit human health.
BIBLIOGRAPHY


APPENDIX I

LIKELIHOOD OF OBSERVING A DITAG GIVEN BOTH CONTRIBUTING TAG SEQUENCES

<table>
<thead>
<tr>
<th>CATG$\Phi_1$CATG</th>
<th>$\Phi_1$NN</th>
<th>$\Phi_1$NX</th>
<th>$\Phi_1$XX</th>
<th>$\times \ P(l_d=28)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Theta_{10}$NN</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$\Theta_{10}$NX</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$\Theta_{10}$XX</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CATG$\Phi_1$CATG</th>
<th>$\Phi_1$AN</th>
<th>$\Phi_1$AX</th>
<th>$\Phi_1$BN</th>
<th>$\Phi_1$BX</th>
<th>$\Phi_1$XX</th>
<th>$\times \ P(l_d=29)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Theta_{10}$AN</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.625</td>
<td></td>
</tr>
<tr>
<td>$\Theta_{10}$AX</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0.125</td>
<td></td>
</tr>
<tr>
<td>$\Theta_{10}$BN</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0.125</td>
<td></td>
</tr>
<tr>
<td>$\Theta_{10}$BX</td>
<td>0.625</td>
<td>0.625</td>
<td>0.125</td>
<td>0.125</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>$\Theta_{10}$XX</td>
<td>0.625</td>
<td>0.625</td>
<td>0.125</td>
<td>0.125</td>
<td>0.25</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CATG$\Phi_1$CATG</th>
<th>$\Phi_1$AN, $\Phi_1$AX</th>
<th>$\Phi_1$BA</th>
<th>$\Phi_1$BB</th>
<th>$\Phi_1$BX</th>
<th>$\Phi_1$XX</th>
<th>$\times \ P(l_d=30)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Theta_{10}$AA</td>
<td>0</td>
<td>$P(l_t=10)P(l_c=12)+P(l_c=11)^2$</td>
<td>$P(l_t=11)^2$</td>
<td>$P(l_t=10)P(l_t=12)\times0.2$</td>
<td>$P(l_t=10)P(l_t=12)\times0.0625$</td>
<td></td>
</tr>
<tr>
<td>$\Theta_{10}$AB</td>
<td>$P(l_t=12)P(l_c=10)$</td>
<td>$P(l_t=10)^2+P(l_t=11)^2$</td>
<td>$P(l_t=10)P(l_t=12)\times0.2$</td>
<td>$P(l_t=10)P(l_t=12)\times0.0625$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Theta_{10}$AX</td>
<td>$P(l_t=12)P(l_c=10)\times0.25$</td>
<td>$P(l_t=10)^2+P(l_t=11)^2$</td>
<td>$P(l_t=10)P(l_t=12)\times0.2$</td>
<td>$P(l_t=10)P(l_t=12)\times0.0625$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Theta_{10}$BN, $\Theta_{10}$BX</td>
<td>0</td>
<td>$P(l_t=10)P(l_c=12)$</td>
<td>0</td>
<td>$P(l_t=10)P(l_t=12)\times0.2$</td>
<td>$P(l_t=10)P(l_t=12)\times0.0625$</td>
<td></td>
</tr>
<tr>
<td>$\Theta_{10}$XX</td>
<td>$P(l_t=12)P(l_c=10)\times0.0625$</td>
<td>$P(l_t=10)^2+P(l_t=11)^2\times0.25$</td>
<td>$P(l_t=10)P(l_t=12)\times0.2$</td>
<td>$P(l_t=10)P(l_t=12)\times0.0625$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### $\text{CATG}_1 \text{ABC}_1 \text{CATG}$

<table>
<thead>
<tr>
<th>$\theta_{10 AB}$</th>
<th>$\phi_{10 CB}$</th>
<th>$\phi_{10 CA}$</th>
<th>$\phi_{10 CX}$</th>
<th>$\phi_{10 BN}$, $\phi_{10 BX}$</th>
<th>$\phi_{10 XX}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.625</td>
<td>0</td>
<td>0.15625</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0.125</td>
<td>0</td>
<td>0.03125</td>
<td></td>
</tr>
<tr>
<td>0.625</td>
<td>0.125</td>
<td>0.25</td>
<td>0</td>
<td>0.0625</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0.15625</td>
<td>0.03125</td>
<td>0.0625</td>
<td>0</td>
<td>0.015625</td>
<td></td>
</tr>
</tbody>
</table>

$\times P(l_d=31)$

### $\text{CATG}_1 \text{ABCD}_1 \text{CATG}$

<table>
<thead>
<tr>
<th>$\theta_{10 AB}$</th>
<th>$\phi_{10 DC}$</th>
<th>$\phi_{10 DB}$</th>
<th>$\phi_{10 DX}$</th>
<th>$\phi_{10 XX}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.25</td>
<td>0.0625</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>0.0625</td>
<td>0.015625</td>
<td></td>
</tr>
<tr>
<td>0.0625</td>
<td>0</td>
<td>0.015625</td>
<td>0.00390625</td>
<td></td>
</tr>
</tbody>
</table>

$\times P(l_d=32)$

$l_d$: ditag length (20-24, where e.g. CATGX$_{20}$CATG is $l_d=20$)

$\theta_{10}$: some arbitrary 10bp sequence

A, B: some arbitrary, known nucleotide

N: a known nucleotide, where the actual base is not relevant

X: an unknown nucleotide

$P(l_t=x)$: probability that a tag is length $x$

$P(l_d=x)$: probability that a ditag is length $x$
APPENDIX II

MODEL FITTING R SOURCE CODE

Sample Data

```r
# generate some sample data, replace with actual data
counts <- c( 9, 13, 11, 8, 9, 20, 16, 19, 18, 15 )
library.sizes <- rep( 100000, 10 )

# generate some class labels, replace with actual labels
# in this example, first 5 are class 0, last 5 are class 1
classes <- c( rep( 0, 5 ), rep( 1, 5 ) )
```

Log-linear (Poisson) regression model

```r
# perform the model fit
fit <- glm( counts ~ offset(log(library.sizes)) + classes,
            family=poisson(link="log") )

# get the beta coefficients
beta0 <- fit$coefficients[[1]]
betal <- fit$coefficients[[2]]

# get the expression (as a proportion) for each group
prop0 <- exp(beta0)
prop1 <- exp(beta0+betal)

# calculate the significance score for differential expression
# i.e. null hypothesis is that betal = 0
p.value <- 2 * pt( -abs(t.value), fit$df.residual )
```

Overdispersed log-linear regression model

```r
# requires the MASS library
library( MASS )

fit <- glm.nb( counts ~ offset(log(library.sizes)) + classes )

# get the beta coefficients and dispersion
beta0 <- fit$coefficients[[1]]
betal <- fit$coefficients[[2]]
dispersion <- 1 / fit$theta

dispersion

# get the expression (as a proportion) for each group
prop0 <- exp(beta0)
prop1 <- exp(beta0+betal)

# calculate the significance score for differential expression
# i.e. null hypothesis is that betal = 0
p.value <- 2 * pt( -abs(t.value), fit$df.residual )
```
Poisson mixture model

# requires the flexmix library
library( flexmix )

# set fitting control parameters to settings that work well
# for SAGE
custom.FLXcontrol <- list( iter.max=200,
    minprior=0,
    tolerance=1E-6,
    verbose=0,
    classify="hard",
    nrep=1 )
custom.FLXcontrol <- as( custom.FLXcontrol, "FLXcontrol" )

# specify the maximum number of model components
maxk <- 5

# specify the number of fit attempts per component
fit.attempts <- 5

# create objects to store fit for each k
fits <- list()
aic.fits <- rep( NA, maxk )

# increase the number of components until AIC decreases
for( k in 1:maxk ) {
  # don't bother fitting if there are fewer distinct values
  # than k
  if( length(unique((counts+1)/(sizes+2))) < k ) break

  # make an initial "good" guess of class membership
  # using k-means – helps avoid falling into a local
  # likelihood maximum
  cm <- rep( 1, length(counts) )
  if( k > 1 ) cm <- kmeans( (counts+1)/(sizes+2),
              centers=k )$cluster

  for( i in 1:fit.attempts ) {
    fit <- try( flexmix( counts ~ 1,
            k=k,
            model=FLXglm( family="poisson",
                        offset=log(sizes) ),
            control=custom.FLXcontrol,
            cluster=cm ), silent=TRUE )
    # if fitting failed (did not converge), try again
    if( "try-error" %in% class(fit) ) next
    if( is.na( aic.fits[[k]] ) ) {
      fits[[k]] <- fit
      aic.fits[k] <- AIC( fits[[k]] )
    } else {
      if( AIC(fit) < aic.fits[k] ) {

# this attempt was an improvement, so use it
fits[[k]] <- fit
aic.fits[k] <- AIC( fit )
}
}

# if the fit failed all attempts, do not continue trying
# to fit with an increasing number of components
if( is.na(fits[[k]]) ) break

# if the fit found less components than specified, do not
# continue to fit with an increasing number of components
if( max(clusters(fits[[k]])) > k ) break

# what is the optimal k? (minimum AIC)
k.optimal <- which( aic.fits == min( aic.fits, na.rm=TRUE ) )[1]
fit.optimal <- fits[[k.optimal]]

# get the theta parameters (component coefficients)
thetas <- array( dim=k.optimal )
for( i in 1:k.optimal ) thetas[i] <- parameters(fit.optimal, component=i)$coef

# get the pi parameters (mixing coefficients)
pis <- attributes(fit.optimal)$prior

# what is the test statistic score that the fitted components
# differentiate between groups?
confidence.up <- pmm.testStatistic(
    fit.optimal, k.optimal, classes, downreg=F )
confidence.down <- pmm.testStatistic(
    fit.optimal, k.optimal, classes, downreg=T )

Mixture model test statistic

"pmm.testStatistic" <- function( fit, k, classes, downreg=T, 
groups=NULL ) {

    if( is.null( groups ) ) groups <- classes

    if( k == 1 ) return 0

    # get mixture component coefficients
    coefs <- array( dim=k )
    for( i in 1:k ) coefs[i] <- parameters(fit, component=i)$coef

    # get posterior probabilities
    post.probs <- matrix( 
        ncol=k, 
        data=as.numeric(fit@posterior[["unscaled"]]) )
# get mixing coefficients
pis <- attributes(fit)$prior

# scale the posterior probabilities
post.probs <- post.probs / rowSums(post.probs)

# reorder the coefs and posterior probabilities
post.probs <- post.probs[, order(coefs, decreasing=downreg)]
coefs <- coefs[order(coefs, decreasing=downreg)]

scores <- rep(NA, k-1)
for( tau in 1:(k-1) ) {
    class.probs <- rep(1, length(
        unique(classes[which(!is.na(classes))])))
    post.probs2 <- post.probs[, c(1:tau)]
    if( tau > 1 ) post.probs2 <- rowSums(post.probs2)
    for( cls in unique(classes[which(!is.na(classes))]) ) {
        class.probs[cls+1] <- sum(post.probs2[which(classes==cls)])
    }
    p0 <- 1
    for( idx in groups[which(groups[,2]==0),1] ) {
        p0 <- p0*class.probs[idx+1]
    }
    p1 <- 1
    for( idx in groups[which(groups[,2]==1),1] ) {
        p1 <- p1*(1-class.probs[idx+1])
    }
    scores[tau] <- p0*p1
}

return( max( scores[tau] ) )
APPENDIX III

DERIVATION OF POISSON MIXTURE MODEL DIFFERENTIAL EXPRESSION CONFIDENCE SCORE

Given Bayes Theorem:

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

First, define A as the sample class (e.g. normal, cancer) and B as a mixture component(s). Therefore, re-write Bayes Theorem as:

\[ P(\omega|k) = \frac{P(k|\omega)P(\omega)}{P(k)} \]

The probability of sample type \( \omega \) arising from mixture component \( k \) is given as:

\[ P(k|\omega) = \frac{\sum_{j\in\omega} \tau_{jk}}{N_\omega} \]

where \( \tau_{jk} \) is the posterior probability that observation \( j \) (where \( j\in\omega \) is the subset of observations of type \( \omega \)) \( j \) arose from component \( k \) and \( N_\omega \) is the number of libraries of type \( \omega \).

The probability that a sample is of type \( \omega \) is simply:

\[ P(\omega) = \frac{N_\omega}{N} \]

where \( N \) is the total number of libraries.

The unconditional probability \( P(k) \) that a sample arose from mixture component \( k \) is the mixing coefficient \( \pi_k \).
Substituting terms, we arrive at:

\[
P(\omega|k) = \frac{\sum_{j \in \omega} \tau_{jk} N_{\omega}}{N_{\omega} \pi_k}
\]

Finally, by cancelling like terms, we have the expression:

\[
P(\omega|k) = \frac{\sum_{j \in \omega} \tau_{jk}}{N_{\omega} \pi_k}
\]