Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

9-15-2015 12:00 AM

Interpretation of Mutations, Expression, Copy Number in Somatic Breast Cancer: Implications for Metastasis and Chemotherapy

Stephanie Dorman
The University of Western Ontario

Supervisor
Dr. Peter Rogan
The University of Western Ontario

Graduate Program in Biochemistry
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy
© Stephanie Dorman 2015

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Diagnosis Commons, Medical Genetics Commons, Neoplasms Commons, Oncology Commons, and the Therapeutics Commons

Recommended Citation

Dorman, Stephanie, "Interpretation of Mutations, Expression, Copy Number in Somatic Breast Cancer: Implications for Metastasis and Chemotherapy" (2015). *Electronic Thesis and Dissertation Repository*. 3227.
https://ir.lib.uwo.ca/etd/3227

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

INTERPRETATION OF MUTATIONS, EXPRESSION AND COPY NUMBER IN SOMATIC BREAST CANCER: IMPLICATIONS FOR METASTASIS AND CHEMOTHERAPY

(Thesis format: Integrated Article)

by

Stephanie N. Dorman

Graduate Program in Biochemistry

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

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

© Stephanie N. Dorman 2015

Abstract

Breast cancer (BC) patient management has been transformed over the last two decades due to the development and application of genome-wide technologies. The vast amounts of data generated by these assays, however, create new challenges for accurate and comprehensive analysis and interpretation. This thesis describes novel methods for fluorescence in-situ hybridization (FISH), array comparative genomic hybridization (aCGH), and next generation DNA- and RNA-sequencing, to improve upon current approaches used for these technologies. An *ab initio* algorithm was implemented to identify genomic intervals of single copy and highly divergent repetitive sequences that were applied to FISH and aCGH probe design. FISH probes with higher resolution than commercially available reagents were developed and validated on metaphase chromosomes. An aCGH microarray was developed that had improved reproducibility compared to the standard Agilent 44K array, which was achieved by placing oligonucleotide probes distant from conserved repetitive sequences.

Splicing mutations are currently underrepresented in genome-wide sequencing analyses, and there are limited methods to validate genome-wide mutation predictions. This thesis describes Veridical, a program developed to statistically validate aberrant splicing caused by a predicted mutation. Splicing mutation analysis was performed on a large subset of BC patients previously analyzed by the Cancer Genome Atlas. This analysis revealed an elevated number of splicing mutations in genes involved in NCAM pathways in basal-like and HER2-enriched lymph node positive tumours. Genome-wide technologies were leveraged further to develop chemosensitivity models that predict BC response to paclitaxel and gemcitabine. A type of machine learning, called support vector machines (SVM), was used to create predictive models from small sets of biologically-relevant genes to drug disposition or resistance. SVM models generated were able to predict sensitivity in two groups of independent patient data.

High variability between individuals requires more accurate and higher resolution genomic data. However the data themselves are insufficient; also needed are more insightful analytical methods to fully exploit these data. This dissertation presents both improvements in data quality and accuracy as well as analytical procedures, with the aim of detecting and

interpreting critical genomic abnormalities that are hallmarks of BC subtypes, metastasis and therapy response.

Keywords

genomic technology, breast cancer, nucleic acid hybridization, fluorescence in-situ hybridization, microarray, copy number changes, next generation sequencing, splicing mutations, NCAM, mutation validation, chemosensitivity, support vector machines, machine learning, paclitaxel, gemcitabine, formalin fixed paraffin embedded tissue

Co-Authorship Statement

Chapter 1 – Stephanie Dorman wrote the introduction/literature review and created all figures and tables, with the exception of Figure 1.4, which was adopted from previous work (as cited). Peter Rogan and Joan Knoll provided helpful comments and feedback.

For Chapters 2, 3, 4, and 5, Stephanie Dorman performed all research and analyses with the exceptions noted below. Stephanie Dorman and Peter Rogan conceived and designed the experiments. Stephanie Dorman wrote all Chapters with helpful comments and revisions from Peter Rogan and the other co-authors, unless stated otherwise.

Chapter 2 – Ben Shirley developed and ran the *ab initio* algorithm, and ran PICKY to design the genome-wide oligonucleotides. Natasha Caminsky designed and validated the *NOTCH1* and *CCND1* probes under the supervision of Stephanie Dorman, Peter Rogan and Joan Knoll.

Chapter 3 – Peter Rogan elaborated the problem and conceived of the analytic solution. Coby Viner implemented the algorithm, wrote, and tested the Veridical software and its accompanying Perl and R scripts. Coby Viner and Stephanie Dorman designed the methods used, generated, and interpreted mutation results. Ben Shirley wrote and tested the Shannon pipeline, which provided the predictions evaluated by Veridical. Coby Viner, Stephanie Dorman, and Peter Rogan wrote the manuscript, which has been approved by all of the authors.

Chapter 4 – Peter Rogan and Stephanie Dorman conceived of the study. Peter Rogan directed the project, and Stephanie Dorman and Coby Viner performed all analyses. Specifically, Coby Viner ran Strelka and Veridical, and generated the Circos plots and word clouds. Stephanie Dorman and Peter Rogan wrote and revised the chapter, with input from Coby Viner.

Chapter 5 – Peter Rogan, Joan Knoll, and Stephanie Dorman conceived of the study. Stephanie Dorman and Katherina Baranova performed all bioinformatics analysis. Stephanie Dorman completed all nucleic acid extractions and cDNA synthesis on the breast cancer formalin fixed paraffin embedded tissue samples. Stephanie Dorman and Katharina Baranova completed all copy number and gene expression measurements. Stephanie Dorman,

Katharina Baranova, and Peter Rogan wrote the manuscript, which was approved by all authors.

Chapter 6 – Stephanie Dorman wrote the discussion with helpful comments and feedback from Peter Rogan and Joan Knoll.

Acknowledgments

I would like to start off by thanking my supervisor, Dr. Peter Rogan, for his guidance and advice over the course of my graduate studies. I am truly grateful for the scientific training I have received and opportunities I was granted as a PhD student in his laboratory. I would also like to acknowledge Dr. Joan Knoll, who has been an invaluable mentor and advisor. I feel very privileged to have worked with two scientists who have shared so much of their vast knowledge, expertise, and stories of human genetics research with me, as well as their love for Italian gelato and pizza.

I would like to thank Dr. Rob Hegele, for all of the helpful discussions during committee meetings, as well as a number of other individuals who have helped with the studies presented in this thesis – the Vancouver Prostate Center Microarray Facility for performing the *HapMap* hybridizations, Dr. Brad Urquhart for his guidance with the growth inhibition studies, and Linda Jackson for her help preparing the FFPE tissue blocks.

I would like to thank all of the past and present Rogan/Knoll Lab members. From identifying chromosomes to programming code, I hope I have taught you as much about Excel as you have in all of your respective areas of expertise. I will miss morning coffees with John and Natasha, pasta bar lunches with Li, Kat, and Ben, and tennis with Wahab. I am also grateful for all of the friendships I have made throughout the Department and CaRTT. A huge thanks to Ben Kleinstiver, for all of his support in both work and life, and Tom and Jason for the Grad Club dates and group chat support over the years.

Last but not least, I would like to thank my mom and dad, for their unwavering love and support, and for being the best parents anyone could ask for. I would like to thank my sisters, Sarah and Katie, and brother-in-law Tim, for being my biggest cheerleaders and keeping me in check with my mixed up expressions. A special thanks to my best friend and partner, Brad - I can't wait for our next adventure. To all of my other friends from Western and London, you mean the world to me and my time here would not have been the same without you.

Table of Contents

Abstract ii

Co-Authorship Statement iv

Acknowledgments vi

Table of Contents vii

List of Tables xii

List of Figures xiv

List of Appendices xvi

List of Abbreviations xx

Chapter 1 1

1 Introduction 1

1.1 Breast Cancer Overview 1

1.1.1 Gene Expression Subtypes of Breast Cancer 1

1.1.2 Genomic Analyses of Breast Cancer Tumours 2

1.2 Genomic technologies used in breast cancer research and clinical management 8

1.2.1 Fluorescence in-Situ Hybridization 8

1.2.2 Array Comparative Genomic Hybridization 10

1.2.3 Gene Expression Microarrays 12

1.2.4 Next Generation Sequencing 13

1.3 DNA Variants 16

1.3.1 Protein Coding Mutations 16

1.3.2 Splicing Mutations 18

1.4 Gene expression signatures in breast cancer 25

1.4.1 Predicting prognosis and patient outcome 27

1.4.2 Selecting therapies and predicting treatment response 28

1.5 The Minimal Breast Cancer Genome and its Relevance to Chemotherapy 29

1.6 Thesis Scope and Objectives 32

References 35

Chapter 2 54

2 Expanding probe repertoire and improving reproducibility in human genomic hybridization 54

2.1 Introduction 54

2.2 Materials and methods 56

2.2.1 scFISH probe design 56

2.2.2 scFISH probe development and hybridization 58

2.2.3 Genome-wide aCGH 58

2.2.4 Locus-specific aCGH 60

2.3 Results 61

2.3.1 Genome-wide coverage of *ab initio* sc intervals 61

2.3.2 *Ab initio* scFISH probes 63

2.3.3 *Ab initio* aCGH 63

2.3.4 Probe parameters affecting CVs 70

2.3.5 Targeted chromosome 15q11.2q13 aCGH detects AS deletion 72

2.4 Discussion 78

2.5 References 82

Chapter 3 89

3 Validation of predicted mRNA splicing mutations using high-throughput transcriptome data 89

3.1 Introduction 89

3.2 Methods 91

3.3 Results 101

3.3.1 Leaky Mutations 101

3.3.2 Inactivating Mutations 102

3.3.3 Cryptic Mutations 109

3.3.4 Performance 109

3.4 Discussion 116

3.5 References 120

Chapter 4 124

4 Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer 124

4.1 Introduction 124

4.2 Methods 126

4.3 Results 130

4.3.1 Derivation of mutations 130

4.3.2 Significantly mutated genes 130

4.3.3 Validating predicted splicing mutations 132

4.3.4 Copy number analysis of mutated genes 133

4.3.5 Analysis of pathways enriched in mutant genes 139

4.3.6 Relationship of mutation spectra to clinical findings 139

4.3.7 Elevation of NCAM1-related gene pathway mutations in lymph node-positive tumours 142

4.3.8 Analysis of tumour subtypes 142

4.4 Discussion 145

4.5 References 149

Chapter 5 155

5 Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning 155

5.1 Introduction 155

5.2 Materials and Methods 157

5.2.1 Data Acquisition 157
5.2.2 Cell Lines 159
5.2.3 Multiple Factor Analysis (MFA) 159
5.2.4 Support Vector Classification 159
5.2.5 Applying the cell line SVM to patient data 160
5.2.6 Clustering cell lines and patients using expression values of the SVM gene subsets 161
5.3 Results 161
5.3.1 Multiple Factor Analysis 161
5.3.2 Support Vector Machine (SVM) Learning 162
5.3.3 Applying the cell line-trained SVM to patient data 167
5.3.4 Clustering cell line and patient data based on SVM gene subsets 172
5.3.5 Significance of SVM classification accuracy 176
5.3.6 Translation of signature to other cancer types 179
5.4 Discussion 181
5.5 References 186
Chapter 6 197
6 Contextual Insights of Findings in this Dissertation 197
6.1 Current limitations of genomic technology 197
6.2 Advances in genomic technology described in this thesis 199
6.2.1 Fluorescence in-situ hybridization 199
6.2.2 Chromosomal Microarrays 200
6.2.3 Next Generation Sequencing 203
6.3 Implications for breast cancer treatment 208
6.3.1 DNA mutations in metastasis 208
6.3.2 Predicting tumour sensitivity to paclitaxel and gemcitabine 212

6.4 Thesis impact on personalized medicine in breast cancer 216

6.5 References 221

Appendices 234

Curriculum Vitae 337

List of Tables

Table 1.1 Summary of large-scale breast cancer sequencing studies. 4

Table 1.2 Frequency of commonly mutated genes in breast cancer tumours 5

Table 1.3 DNA sequencing software 16

Table 1.4 Splicing mutation and splice site analysis software..... 24

Table 1.5 Splicing mutation analyses performed in previous sequencing studies..... 25

Table 1.6 Treatment recommendations according to tumour subtype and/or receptor status 30

Table 1.7 Gene expression signatures developed to predict therapy response. 31

Table 2.1: Comparison of coefficients of variation of replicate probes by platform: Mann–Whitney rank sum test 66

Table 2.2: Comparison of cofficients of variation of replicate probes by platform: Paired *t*-tests 67

Table 2.3: Principal components analysis of genomic and probe parameters with coefficients of variation in *HapMap* pedigrees 71

Table 2.4: Analysis of variation of coefficients of variations in Agilent and Affymetrix aCGH probe subsets..... 73

Table 3.1 Definitions used within Veridical to determine which reads are checked. *A* and *B* represent natural site positions, defined in Figure 3.1(B). 93

Table 3.2 Examples of variants validated by Veridical and their selected read types..... 113

Table 4.1: Single nucleotide variant summaries by mutation type 131

Table 4.2: Genes most commonly mutated with splicing mutations 135

Table 4.3: Multiple factor analysis of *NCAM1* related pathway mutations and clinical parameters per tumour 144

Table 5.1 Using the support vector machine to predict patient response from archived formalin fixed paraffin embedded tissue 172

Table 5.2 Support vector machine predictions on 319 patients treated with paclitaxel from Hatzis et al. (2011)..... 173

Table 5.3 Support vector machine performance using randomly selected genes based off 100,000 iterations..... 180

Table 6.1 Examples of clinically significant genomic alterations in cancer testable by fluorescence in-situ hybridization..... 201

List of Figures

Figure 1.1 Significantly mutated genes in breast cancer tumours. 6

Figure 1.2 Example of bacterial artificial chromosome end pairs and fluorescence in-situ hybridization clones overlapping ERBB2. 11

Figure 1.3 Basic schematic of pre-mRNA splicing. 20

Figure 1.4 Splicing 5' donor and 3' acceptor sequence logos and frequency of reported mutations. 21

Figure 1.5 Aberrant splicing patterns resulting from DNA variants. 23

Figure 2.1: Fluorescence in-situ hybridization validated single copy probes. 64

Figure 2.2: The effect of genomic context on hybridization signal intensity variability. 68

Figure 2.3: Primary hybridization signal intensity data from *ab initio* and Agilent probe sequences covering Angelman Syndrome chromosome deletion region 76

Figure 3.1 Diagram portraying the definitions used within Veridical to specify genic variant position and read coordinates. 94

Figure 3.2 The algorithm employed by Veridical to validate variants. 95

Figure 3.3 Illustrative examples of aberrant splicing detection. 97

Figure 3.4 Integrative Genomics Viewer images depicting a predicted leaky mutation 103

Figure 3.5 Histogram of read-abundance-based intron inclusion with embedded Q-Q plots of the predicted leaky mutation 105

Figure 3.6 Examples of validated mutations. 107

Figure 3.7 Integrative Genomics Viewer images and their corresponding histograms with embedded Q-Q plots depicting all six variant-containing files with a mutation 112

Figure 3.8 Profiling data for Veridical runtime. 115

Figure 4.1 mRNA of *ABL1, CBFB, GATA3* and *PALB2,* which each have validated cryptic splicing mutations confirmed using tumour-matched RNA-Seq data. 134

Figure 4.2: Splicing mutations in *TP53*, *KMT2C* and *CDH1*. 137

Figure 4.3: Circos plot of validated splicing mutations by tumour subtype. 138

Figure 4.4: Word Clouds demonstrating differences between overrepresented mutated pathways in lymph node-positive (a) and lymph node-negative (b) tumours. 140

Figure 4.5: Percent of tumours with mutations by pathway group and clinical factors. 143

Figure 5.1 Workflow to derive gene signatures. 158

Figure 5.2 Genes associated with paclitaxel and gemcitabine mechanism of action. 164

Figure 5.3 Effect of the removal of each gene on the percent of cell lines misclassified during the support vector machine feature selection process to determine the most predictive gene set. 166

Figure 5.4 Coverage and reads from sequencing of three formalin fixed paraffin embedded tumour samples using originally extracted and whole genome amplified DNA. 171

Figure 5.5 Expression heatmap of the paclitaxel and gemcitabine support vector machine derived genes for the tested cell lines. 175

Figure 5.6 A) Expression heatmap of the paclitaxel support vector machine derived genes for 319 tumour samples (Hatzis et al. 2011). 177

Figure 5.7 The proportion of misclassified cell lines (A/C) and hinge loss scores (B/D) were measured on SVMs derived using randomly selected gene sets. 178

Figure 6.1 Screenshot from GTEx Portal – *ACSBG1* Gene View. 206

Figure 6.2 Replicating stage and expression of NCAM pathway and significantly mutated genes. 210

List of Appendices

Appendix S1: ..... Copyright permission for Chapters 2-4.....234

Appendix S1.1..... Copyright permission for Chapter 2.....234

Appendix S1.2..... Copyright permission for Chapter 3.....234

Appendix S1.3..... Copyright permission for Chapter 4.....234

Appendix S1.4..... Copyright permission for Chapter 5.....235

Appendix S2: .....Supplementary Information for Chapter 2.....236

Appendix S2.1 Supplementary Methods: *Ab Initio* single copy (sc) sequence algorithm and implementation..... 236

Appendix S2.2.....Coordinates and PCR primers of validated scFISH probes.....238

Appendix S3: .....Supplementary Information for Chapter 3.....239

Appendix S3.1..... Veridical variant input format.....239

Veridical exome annotation input format 240

Appendix S3.2..... Veridical output.....240

Appendix S3.3.....Supplementary Figure 1.....242

Appendix S4: .....Supplementary Information for Chapter 4.....244

Appendix S4.1..... SomaticSniper Supplementary Materials.....244

Appendix S4.1.1..... Supplementary Methods – Variant Calling Methods.....244

Appendix S4.1.2. Supplementary Results – SomaticSniper Variant Calling Results.....244

Appendix S4.1.3..... Variant Summaries by Mutation Type.....246

Appendix S4.1.4.....SomaticSniper Variants Compared to TCGA Findings.....246

Appendix S4.2..... Filtering criteria for splicing mutations.....247

Appendix S4.3.....Supplementary Figure 1.....248

Appendix S4.4..... Variants compared to those previously published by TCGA.....249

Appendix S4.5..... Overrepresentation analysis of TCGA mutations missed by Strelka.....250

Appendix S4.6..... MuSiC Results Compared to Significantly Mutated Genes.....253

Appendix S4.7 ..... Validated Cryptic Splicing Mutations 255

Appendix S4.7.1 ..... Cryptic Splicing Mutation Details 255

Appendix S4.7.2 ..... The rate of GATA3 abnormal splicing in variant containing tumour and tumour/normal controls 258

Appendix S4.8 ..... Supplementary Figure 4 259

Appendix S4.9 ..... Pathway Analyses 260

Appendix S4.9.1 ..... Pathways Overrepresented by Protein Coding and Splicing Mutations 260

Appendix S4.9.2 ..... Pathways Overrepresented by Every Splicing Mutation Type (inactivating, leaky, cryptic) 263

Appendix S4.9.3 ..... Comparing Grouped Pathways Overrepresented between LN- and LN+ Tumour Mutations 264

Appendix S4.9.4 ..... Pathway Analysis of Deleterious Mutations in LN- and LN+ Tumours 268

Appendix S4.10 ..... Frequency of Mutations in NCAM1 Pathway Genes 271

Appendix S4.11 ..... Breast Cancer Mutations by Subtype 273

Appendix S4.11.1 ..... Number of Mutations by Subtype 273

Appendix S4.11.2 ..... Pathway Analysis of Mutations by Subtype and Lymph Node Status 274

Appendix S4.11.3 ..... Word clouds of overrepresented pathways by subtype. 275

Appendix S4.11.4 281

Appendix S4.12 ..... Appendix S4 References 281

Appendix S5: ..... Supplementary Information for Chapter 5 282

Appendix S5.1 ..... Cell Lines Used 282

Appendix S5.2 Genes Included in the study relevant to paclitaxel and gemcitabine drug disposition 284

Appendix S5.3 ..... Copy Number Calling Methods 286

Appendix S5.4 DNA Sequencing Analysis Pipeline– Variant Calling and Interpretation Methods 286

Appendix S5.5 ..... Reproducibility of Cell Line Data 288

Appendix S5.5.1..... GI50 Studies.....288

Appendix S5.5.2.....CytoScan HD Array.....288

Appendix S5.5.3..... Gene Capture and DNA Sequencing.....288

Appendix S5.6.....Support vector machine feature selection.....291

Appendix S5.7..... Partial-Least Squares Regression.....292

Appendix S5.8....Gene expression and Copy Number analyses on FFPE tumour blocks.....292

Appendix S5.9..... Reproducibility of Data.....299

Appendix S5.10..... MFA Criteria.....307

Appendix S5.11..... Multiple Factor Analysis.....307

Appendix S5.11.1...Cell Line Numbers in the Multiple Factor Analyses Individual Factor Maps: 307

Appendix S5.11.2.....Paclitaxel Example - MAPT.....308

Appendix S5.11.3..... Gemcitabine Example - DCTD.....309

Appendix S5.12..... Paclitaxel Multiple Factor Analysis Results by Gene.....310

Appendix S5.13.....Gemcitabine Multiple Factor Analysis Results by Gene.....311

Appendix S5.14.....Cell Line GI50 vs. SVM Classification Score.....312

Appendix S5.14.1.....Paclitaxel SVM.....312

Appendix S5.14.2..... Gemcitabine SVM.....313

Appendix S5.15..... Single Gene paclitaxel and gemcitabine SVMs using cell line data.....314

Appendix S5.16.....Multiple Factor Analysis– Entire and SVM Gene Sets.....316

Appendix S5.16.1..... Paclitaxel – SVM Gene Set.....316

Appendix S5.16.2..... Paclitaxel – Entire Gene Set.....318

Appendix S5.16.3.....Gemcitabine – SVM Gene Set.....319

Appendix S5.16.4.....Gemcitabine– Entire Gene Set.....321

Appendix S5.17..... FFPE Samples – Gene expression measurements summary.....322

Appendix S5.17.1.....Number of measurements by gene compared to GTEx expression levels 322

Appendix S5.17.2.. Year of tissue block compared to number of measurements per sample 323

Appendix S5.18.....Patient Clustering Supplementary Results.....324

Appendix S5.18.1.....FFPE Patient Samples.....324

Appendix S5.18.2.....Hatzis et al. Patient Data.....326

Appendix S5.19..... *MAPT* Expression Affects Prognosis in Luminal Tumours.....329

Appendix S5.20.....Creating SVM models using lung and hematopoietic cell lines.....331

Appendix S5.20.1.....Feature Selection Process – Lung Cancer Cell Lines.....331

Appendix S5.20.2..... Feature Selection Process – Hematopoietic and Lymphoid Tissue Cancer Cell Lines 332

Appendix S5.20.3.....Final SMV Gene Sets for Paclitaxel.....332

Appendix S5.20.4.....MFA Using Genes in SVM – Lung.....333

Appendix S5.20.5..... MFA Using Genes in SVM – Hematopoietic and Lymphoid Tissue 334

Appendix S5.21.....References for Appendix S5.....335

List of Abbreviations

(Py)n	polypyrimidine tract
A	adenine
aCGH	array comparative genomic hybridization
Agilent 44K	Agilent Technologies Human Catalog CGH 4 × 44K microarray
AJCC	American Joint Committee on Cancer
ANOVA	analysis of variance
AS	Angelman syndrome
ASSEDA	Automated Splice Site and Exon Definition Analyses
BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
bp	basepairs
BWA	Burrows-Wheeler Aligner
C	cytosine
cDNA	complementary DNA
CEF	cyclophosphamide, epirubicin and fluorouracil
CGH	comparative genomic hybridization
CMF	cyclophosphamide, methotrexate, and fluorouracil
CN	copy number
CNC	copy number change
CNV	copy number variant
COSMIC	catalogue of somatic mutations in cancer
CT	convolution
CV	coefficients of variation
DAPI	4',6-diamidino-2-phenylindole
DNA	deoxyribonucleic acid
DNA-Seq	DNA sequencing
ECM	extracellular matrix
ER	estrogen receptor
ES	effect size
FCPT	Fisher's combined P-value

FDA	United States Food and Drug Administration
FDR	false discovery rates
FFPE	formalin fixed paraffin embedded
FISH	fluorescent in situ hybridization
G	guanine
GATK	Genome Analysis Toolkit
GEO	Gene Expression Omnibus
GI50	growth inhibition values
GRCh37	human Feb. 2009 assembly reference sequence
HER2	human epidermal growth factor
hg19	human Feb. 2009 assembly reference sequence
IGV	Integrative Genomics Viewer
indels	insertions/deletions
ISCA	International Standards for Cytogenomic Arrays
kb	kilobases
LN	lymph node
LRT	likelihood ratio test
MATS	Multivariate Analysis of Transcript Splicing
mb	megabases
MFA	multiple factor analysis
MLPA	multiplex ligation-dependent probe amplification
MuSiC	mutational significance in cancer
NCAM	neural cell adhesion molecule
NCI-MATCH	NCI-Molecular Analysis for Therapy Choice
nt	nucleotides
PCA	principal component analysis
PCR	polymerase chain reaction
pCR	pathological complete response
PR	progesterone receptor
q-PCR	real time PCR
RCB	residual cancer burden
RD	recurrent disease

RefSeq	NCBI Reference Sequence Genes
RNA	ribonucleid acid
RNA-Seq	RNA sequencing
ROI	regions of interest
RT-PCR	reverse transcriptase PCR
sc	single copy
scFISH	single copy fluorescent in situ hybridization
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SVM	support vector machine
T	thymine
TCGA	The Cancer Genome Atlas
Tm	melting temperature
U	uracil
VAAST	Variant Annotation, Analysis & Search Tool
VUS	variants of unknown significance
YJ	Yeo-Johnson

“Education is the most powerful weapon which you can use to change the world.”

-Nelson Mandela

Chapter 1

1 Introduction

1.1 Breast Cancer Overview

Breast cancer is the most frequently diagnosed cancer worldwide (1). In Canada, 1 in 9 women are expected develop breast cancer in their lifetime, with 24,000 new cases (26% of all cancer cases) in 2014 (2). Advancements in prevention, screening, and treatment strategies over the past 20 years have led to a steady decrease in mortality rates from breast cancer, yet it still accounts for 14% of cancer deaths in Canada (2). These rates are similar to those of the United States and other economically developed countries (1).

After diagnosis, clinicians rely on multiple parameters to direct treatment strategies and predict prognosis, including clinical factors, such as patient age, lymph node status (positive or negative), tumour size, and histological grade, as well as the status of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) in the tumour.

1.1.1 Gene Expression Subtypes of Breast Cancer

Although all breast tumours are grouped under the umbrella of one disease, breast cancer is remarkably complex. The traditional markers used for tumour classification are not able to fully portray the biological variability observed among breast tumours (including genomic alterations, cellular composition, and response to treatment). With the advancement of microarray technology, gene expression profiling led to the sub classification of breast cancer into 5 categories: Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like (3,4). More recently, an additional subtype was identified, Claudin-low (5), to make up the 5 intrinsic subtypes of breast cancer, and the additional Normal Breast-like group. These subtypes are now well-characterized, and have distinct gene expression patterns (3), require different treatment regimens (6) and vary in prognoses (7).

Luminal A tumours make up approximately 30% of breast cancer cases, and have the longest relapse-free and overall survival, whereas Luminal B tumours have lower relapse-free survival, similar to the other subtypes (8). The large majority (at least >90%) of Luminal A and B tumours are ER+ and can be identified by their gene expression signatures characteristic of luminal epithelial cells. These genes include a group of transcription factors, including ER, which can be used to differentiate between Luminal A and B tumours, because this proliferation signature is expressed at higher levels in the Luminal B subtype (3,8). HER2-enriched tumours are characterized by the amplification of the HER2 gene and historically, they have had low relapse-free and overall survival (8). However, the development of Trastuzumab, a monoclonal antibody against HER2, improved response rate and reduced the risk of death for this subtype by 20% when used in conjunction with chemotherapy (9). Basal-like and Claudin-low subtypes are similar in that they have low expression in both the Luminal and HER2-enriched intrinsic expression signatures, but differ in at least two groups of genes. Unlike the Basal-like subtype, Claudin-low tumours show low expression in a gene cluster enriched with cell-to-cell adhesion proteins, and high expression of a group of genes enriched with immune system response genes (8). Both Basal-like and Cluadin-low subtypes have poorer prognoses compared to Luminal A tumours, and similar to the outcomes of Luminal B tumours (8). Normal-like tumours are those that cluster with normal breast tissue in gene expression profiling. They have expression of genes that are characteristic of basal epithelial and adipose cells, and low expression of genes usually observed in Luminal cells. The intrinsic subtypes of breast cancer only consider tumour differences at the gene expression level and do not fully portray the molecular complexity of tumours at the genomic level.

1.1.2 Genomic Analyses of Breast Cancer Tumours

Genome instability is one of the major mechanisms that allows cells to develop into cancer (10). The cellular characteristics that enable malignant growth are known as the hallmarks of cancer, and include: the evasion of apoptosis, self-sufficiency in growth signals, insensitivity to anti-growth signals, sustained angiogenesis, limitless replicative potential, tissue invasion and metastasis, reprogramming of energy metabolism and

evading immune destruction (10,11). Our understanding of tumour genomes, and the specific types of aberrations and mutations driving tumourigenesis, is increasing rapidly as next generation sequencing is becoming more advanced and affordable.

At least five major genomic studies have begun to elucidate commonly mutated genes that may be causing or perpetuating tumour development in breast cancer (12-16). Two of these studies focused on tumours with specific pathological markers: one assessed 37 Basal-like and 28 other triple negative breast cancer (ER, PR and HER2 are not expressed) (15), and the other study analyzed 77 ER positive (Luminal) tumours (13). The remaining three sequencing studies assessed either all intrinsic subtypes (12,14), or did not perform subtyping analysis (16). Between all five studies, whole genome or exome sequencing was performed on a total of 860 tumours, and reported a combined 46,167 mutations (See Table 1.1 for a summary/breakdown of each study).

These sequencing studies demonstrated that mutations in different tumour suppressor or oncogenes can lead to the same breast cancer phenotypes. A total of 55 genes were cited as frequently mutated, although many were mutated in less than 10% of tumours (see Table 1.2 for a list of all genes and their mutation frequencies). At least 33 genes were statistically significantly mutated in the breast cancer tumours assessed, and there was considerable overlap between the five studies (Figure 1.1). Not surprisingly, *PIK3CA* and *TP53* were both identified to be significantly mutated in breast tumours across all five studies. *TP53* was identified as a tumour suppressor gene more than two decades ago, and at that time, was observed to be the most commonly altered gene in tumours (17). Frequent mutations in *PIK3CA* in breast cancer were observed as early as 2004, where 25% of the tumours assessed contained somatic mutations in the gene (18). Additional genes that were highlighted in at least two of the five studies included known breast cancer genes (*GATA3, RB1, AKT1, CDH1, MAP3K1, MLL3, CDKN1B and PTEN*) and newly identified ones (*CBFB, RUNX1, TBX3* and *SF3B1*).

These sequencing studies highlight the genomic diversity of mutations among breast cancer tumours. Of particular interest are the 40 (or more) genes that were identified as potential breast cancer genes in only one of five the studies. Discordance between the

Table 1.1 Summary of large-scale breast cancer sequencing studies.

Paper	No. Tum.	Subtypes	No. Mut.	Ave No. Mut. / Tum.	Significantly Mutated Genes	Method to Identify Sig. Mutated Genes
Banerji[12] (B)	108	Lum A = 38 Lum B = 22 HER2 = 21 Basal = 13 Norm = 5	4985	46	**Known:** *PIK3CA, TP53, AKT1, GATA3, MAP3K1* **New:** *CBFB*	MutSig[19] Algorithm - FDR <0.1
Ellis[13] (E)	77	Lum A/B = 77	3208	42	**Known:** *PIK3CA, TP53, GATA3, CDH1, RB1, MLL3, MAP3K1, CDKN1B* **New:** *TBX3, RUNX1, LDLRAP1, STMN2, MYH9, AGTR2, SF3B1, CBFB, ATR*	MuSiC[20] - Convolution Test FDR <0.26
Shah[15] (Sh)	65	Basal = 37	2414	37	**Known**: *TP53, PIK3CA, RB1, PTEN* ***New:*** *MYO3A, GH1*	Considered background mutation rates q < 0.1[21]
Stephens[16] (St)	100	N/A	7241	72	**Known***: *PIK3CA, TP53, CDH1, GATA3, MLL3, AKT1* **New^**: *ARID1B, CASP8, MAP3K1, MAP3K13, NCOR1, SMARCD1, CDKN1B, AKT2, TBX3*	* cited as frequently mutated ^ Searched for non-random clustering of somatic mutations[22,23]
TCGA[14] (T)	510	Lum A = 225 Lum B = 126 HER2 = 57 Basal = 93	28319	56	**Known:** *PIK3CA, PTEN, AKT1, TP53, GATA3, CDH1, RB1, MLL3, MAP3K1, CDKN1B* **New:** *TBX3, RUNX1, CBFB, AFF2, PIK3R1, PTPN22, PTPRD, NF1, SF3B1, CCND3*	MuSiC[20] - Convolution and Likelihood Ratio Tests FDR <0.05

N/A = subtyping analysis was not reported.

Table 1.2 Frequency of commonly mutated genes in breast cancer tumours (refer to Table 1.1 for study abbreviations and citations)

	Gene	B	E	Sh	St	T		Gene	B	E	Sh	St	T
1.	***PIK3CA***	27%	58%	11%	30%	36%	29.	***SETD2***			2%	1%	
2.	***TP53***	27%	23%	54%	37%	37%	30.	***AFF2***					3%
3.	***CDH1***		10%	3%	7%	7%	31.	***AGTR2***		3%			
4.	***GATA3***	4%	10%		14%	11%	32.	***AKT2***				1%	
5.	***MAP3K1***	3%	17%		6%	8%	33.	***APC***				1%	
6.	***MLL3***		6%	3%	5%	7%	34.	***ARID1B***				3%	
7.	***RB1***		5%	8%	2%	2%	35.	***ASXL1***				1%	
8.	***AKT1***	6%			4%	2%	36.	***BRAF***			3%		
9.	***CBFB***	4%	3%			2%	37.	***BRCA1***				1%	
10.	***CDKN1B***		3%		1%	1%	38.	***CCND3***					<1%
11.	***NCOR1***			2%	3%	3%	39.	***COL6A3***			6%		
12.	***PTEN***			8%	3%	3%	40.	***ERBB3***			3%		
13.	***SF3B1***		4%		4%	2%	41.	***GH1***			5%		
14.	***TBX3***		5%		3%	3%	42.	***KRAS***				1%	
15.	***ARID1A***			2%	3%		43.	***LDLRAP1***		3%			
16.	***ARID2***			3%	1%		44.	***MAP3K13***				2%	
17.	***ATR***		8%	6%			45.	***MYO3A***			9%		
18.	***BAP1***			2%	1%		46.	***NRAS***			3%		
19.	***BRCA2***			5%	0%		47.	***PTPN22***					1%
20.	***CASP8***			3%	3%		48.	***SMAD4***				1%	
21.	***ERBB2***			3%	1%		49.	***SMARCD1***				1%	
22.	***MAP2K4***				4%	4%	50.	***STK11***				1%	
23.	***MLL2***			2%	1%		51.	***STMN2***		3%			
24.	***MYH9***		5%	2%			52.	***SYNE1***			6%		
25.	***NF1***				2%	3%	53.	***SYNE2***			5%		
26.	***PIK3R1***			3%		3%	54.	***UBR5***			6%		
27.	***PTPRD***			2%		2%	55.	***USH2A***			9%		
28.	***RUNX1***		5%			4%							

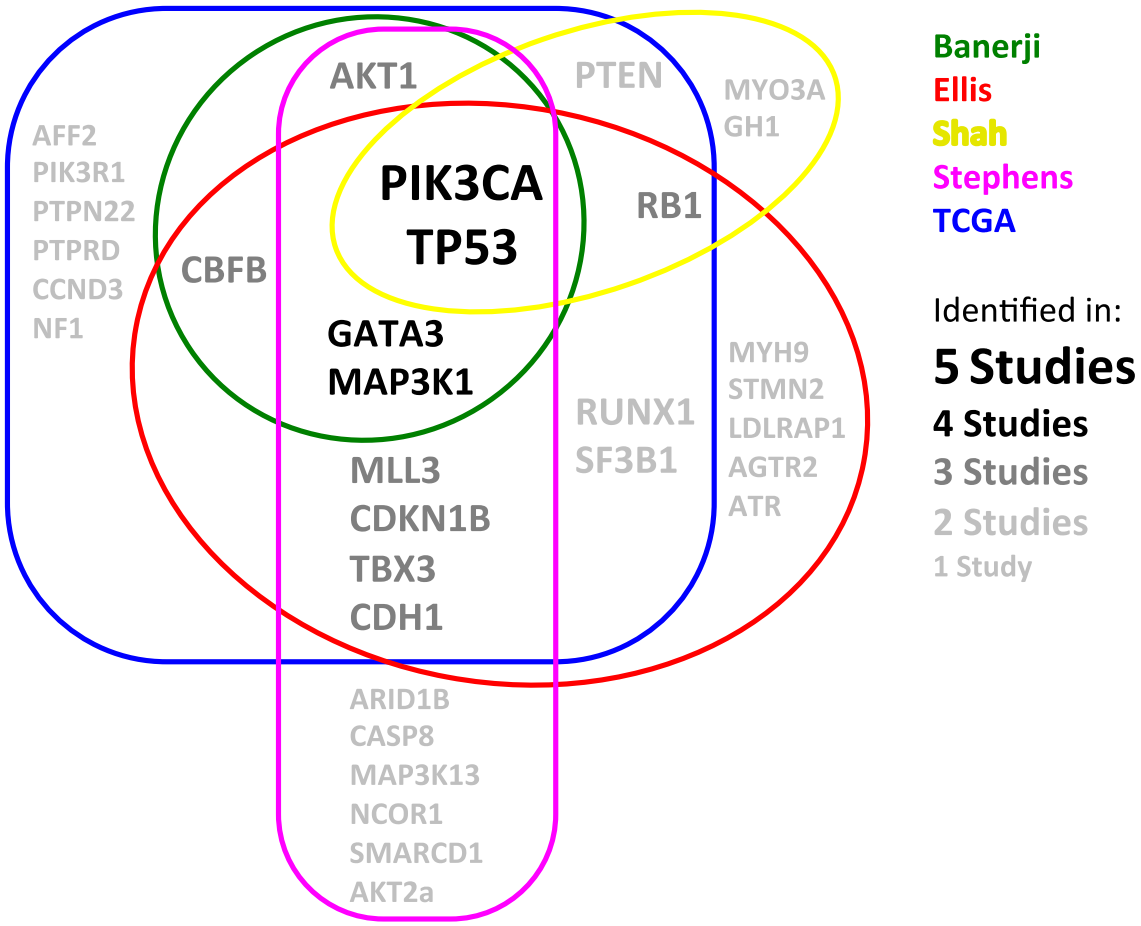


Figure 1.1 Significantly mutated genes in breast cancer tumours. Genes identified as significantly or commonly mutated were extracted from five major sequencing studies (found within each circle and colour coded according to the legend on the top right): Banerji et al. (2012), Ellis et al. (2012), Shah et al. (2012), Stephens et al. (2012), and the Cancer Genome Atlas (2012). The number of studies the gene was identified in is indicated by the bottom legend on the right.

studies may be due to low mutation rates, ranging from 1-9%, or differences in the methods of variant calling, variant filtering, and identifying common/significantly mutated genes. Of the five papers, four unique methods were used to identify which genes were "commonly" or "significantly" mutated, which in some cases lead to discrepancies as to which genes were included as significant. For example, *CASP8* was mutated in 3% of breast cancer tumours in two studies (Shah (15) and Stephens (16)), but only one study (Stephens) cited it as a potential new breast cancer gene. Shah et al., (2012), considered background mutation rates when identifying new breast cancer genes ($q < 0.1$) (21), whereas Stephens et al., (2012), searched for non-random clustering of somatic mutations (22,23).

Regardless of the differences between the five studies, the long list of potential driver genes created from these studies provides a new gene set to be explored and analyzed by the breast cancer community. Mutations in newly recognized genes may have implications in prognosis, treatment response, or provide the opportunity to identify new pathways for therapeutic targeting. For example, Stephens et al., (2012), identified 9 new potential driver genes that have not been previously noted in either breast or other cancer types. These genes are involved in pathways regulating the JUN kinases *MAP2K7* and *MAP2K8*. Mutations in the mitogen-activated protein kinase (MAPK) signaling pathway genes have been suggested to be associated with drug resistance (24), which could have implications for breast cancer treatment if tumours contain mutations in these genes.

Copy number analyses identified commonly deleted or amplified genes, including well known tumour suppressor or oncogenes (*TP53, PIK3CA, NRAS, EGFR, RB1,* and *ATM*), as well as new genes of interest that were not identified through DNA sequence analysis (*PRPS2*, *NRC31*, and four PKC-related genes) (15). These results were similar to the Cancer Genome Atlas (TCGA) study that confirmed previously reported copy number variations, and highlighted many of the same genes affected by copy number changes (including *PIK3CA, EGFR, FOXA1,* and *HER2* in amplified regions, as well as *MLL3, PTEN, and RB1* in deleted regions) (14). The TCGA study also identified five copy

number clusters that correlated with the gene expression subtypes, which had been observed before (25).

1.2 Genomic technologies used in breast cancer research and clinical management

It is possible that in order to achieve the greatest overall success when treating patients, tumours will need to be characterized at the genomic and/or transcriptomic level to guide treatment. This is the basis behind the NCI-Molecular Analysis for Therapy Choice (NCI-MATCH) trial that was recently announced in the United States, which aims to personalize drug selection based on analyzing patient's tumours for specific genetic abnormalities for which a targeted drug exists. There are a number of cytogenetic and molecular techniques that can be used in both research and clinical settings to analyze tumours for different types of mutations, guide diagnosis, predict prognosis and select treatment. Among the most common include fluorescence in-situ hybridization, genomic or gene expression microarrays, and next generation sequencing.

1.2.1 Fluorescence in-Situ Hybridization

Fluorescent in situ hybridization (FISH) uses fluorescently labeled nucleic acid probes to detect targeted genomic or transcriptomic sequences. FISH can be used to localize specific DNA sequences on interphase or metaphase chromosomes, or RNA sequences in cells or tissue samples. FISH was first reported in 1980, by a group that used 3' fluorescently labeled RNA to bind specific DNA sequences (26). Prior to the use of fluorophores, similar hybridization methods used radiolabelled probes, which was not optimal due to the instability of radiolabelled probes, low resolution, long exposure times, and the costs and risks associated with radioactive material (27). Before FISH was developed, conventional cytogenetic methods, such as karyotype analysis, were commonly used for disease research and diagnosis. Given its higher resolution, FISH can be used to detect structural rearrangements in chromosomes including translocations, inversions, insertions, and microdeletions, identify marker chromosomes, and delineate chromosomal breakpoints.

The first draft of the human genome provided the opportunity to develop thousands of DNA clones (primarily bacterial artificial chromosomes, or BACs) that contain genomic sequence tags, and have been mapped to specific chromosome bands (28). Libraries of BAC probes are commercially available, and can also be produced in the laboratory in high quantities using a polymerase with strand displacement activity (29). There have been numerous disease-specific FISH reagents and methods have been developed with proven clinical significance and higher resolution over conventional cytogenetic karyotyping (30,31). Although these BAC FISH probes are still commonly used, the majority of theses clones are greater than 100 kb, so their use is usually restricted to detecting larger rearrangements (Figure 1.2). For the majority of probe labeling and hybridization techniques, detecting small sequences (<10 kb) has been difficult, because smaller probes are often inconsistent and have low sensitivity (32). More recently, methods and techniques have been developed to improve the throughput, or resolution of FISH. For example, labeling probes using nick translation with an excess of DNA polymerase I has increased signal intensities of a 30 kb probe (32), using single copy DNA sequences has enabled FISH probe design where the exact DNA sequence and genomic location are known (33), and an automated analysis method using grid sampling was developed that reduced the time of analysis and evaluation of results down to 9 minutes per sample (34).

FISH is commonly used in clinical diagnosis for birth defects and developmental delay, prenatal testing, and acquired diseases. It is a main test for disorders caused by microdeletions (35) (such as Williams, Prader-Willi, Angelman, Miller-Dieker, DiGeorge, Wolf-Hirschhorn, Cri-du-chat, and Smith-Magenis Syndromes) or microduplications (35) (such as Charcot-Marie-tooth 1A and Pelizaeus-Merzbacher), and also has many different applications in oncology (36). For example, FISH is commonly used to detect specific gene fusions known to occur in certain types of cancers, such as the EML4-ALF fusion in non-small-cell lung cancer (37) and the BCR-ABL fusion (ie. the Philadelphia chromosome) in chronic myeloid leukemia (38). In breast cancer, the American Society of Clinical Oncology/College of American Pathologists has recommended that the *HER2* status (amplified or not) should be tested for all invasive breast cancer (39). They consider tumours to be *HER2*-positive if there are more than 6

copies of *HER2* per nucleus, or if the *HER2* gene signal to chromosome 17 ratio is more than 2.2. FISH results are typically used to confirm, or are confirmed with immunohistochemical assays, which were 92% concordant when assessing hundreds of samples (40). Even so, as much as 20% of *HER2* testing may be inaccurate (39), and probes with higher resolution to the *HER2* gene may be useful in improving the precision of these tests.

1.2.2 Array Comparative Genomic Hybridization

Comparative genomic hybridization (CGH) was developed for use on solid tumours in 1992, and was initially performed genome wide using metaphase chromosomes (41). The technique involves differentially labeling normal and tumour genomic DNA that are simultaneously hybridized to normal metaphase chromosomes in the presence of unlabeled Cot-1 DNA, which is used to block repetitive regions in the genome. Normal and tumour DNA are detected with red and green fluorophores, which allows quantification of the relative amount of normal versus tumour DNA using the ratio of green-to-red fluorescence. The resolution of CGH using metaphase chromosomes is low, detecting copy number changes greater than 20 megabases (Mb). However, at the time, it was still able to identify amplifications in tumours in regions containing oncogenes, including HER2 in breast cancer (41). The resolution of CGH was improved through it's application to microarrays, where targeted P1 phage or BAC clones were spotted on glass slides, hybridized to the sample and reference genome, and then imaged to derive fluorescence ratios of each clone (42). The approach was validated, in part, through more accurately detecting HER2 amplification in a breast cancer cell line and four breast cancer tumours. The resolution and genomic coverage of chromosomal microarrays, including array comparative genomic hybridization (aCGH) was developed further, using cDNA as probes (43), as well as oligonucleotides, which continue to be the current design today (44).

Chromosomal microarrays are now a first tier test to detect genomic aberrations associated with intellectual disability, autism, and many congenital disorders (45,46). aCGH can be used to determine major chromosomal aneuploidy as well as

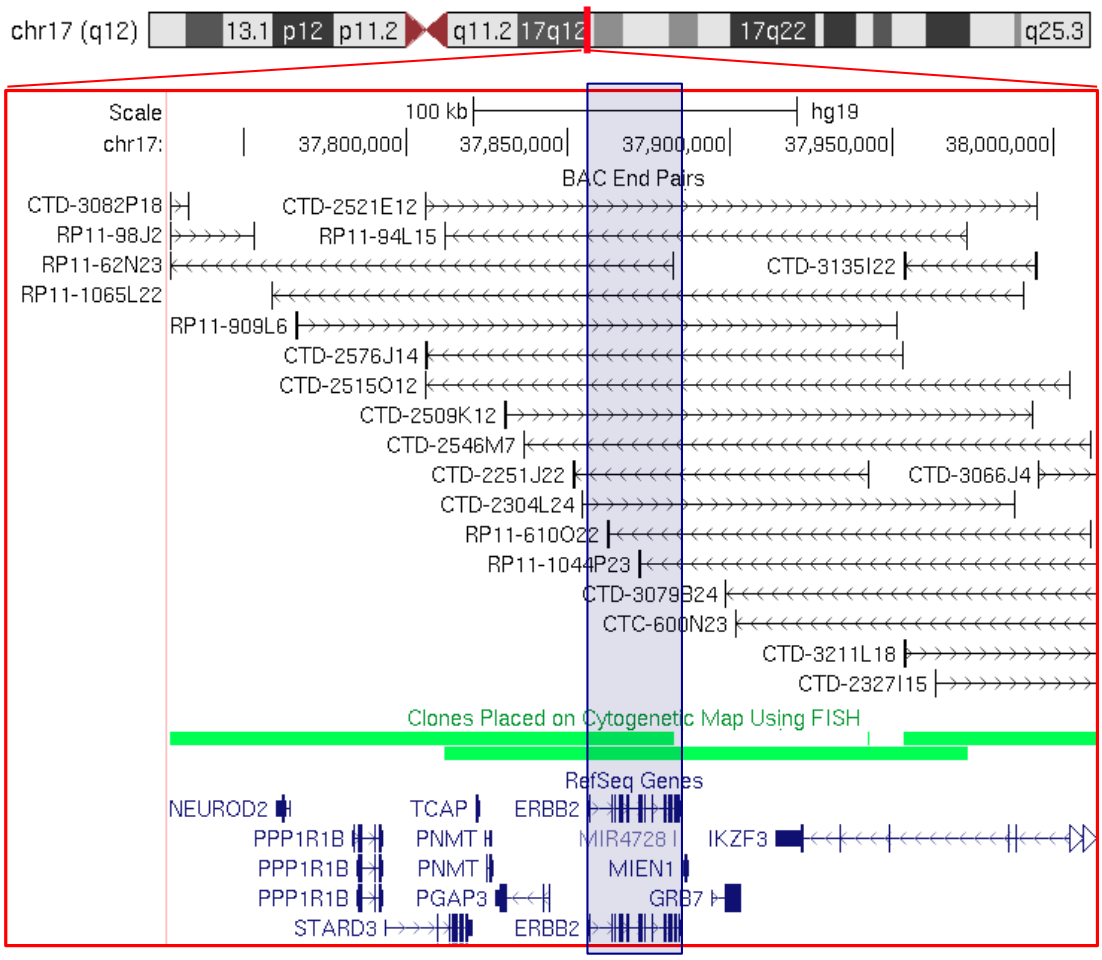


Figure 1.2 Example of BAC end pairs and FISH clones overlapping ERBB2. A screen shot of the UCSC Genome Browser displays the length of BAC probes (black with arrows and green bars) relative to small genes, such as ERBB2 (highlighted in navy blue). The chromosome and scale along the top depict the genomic chromosomal and genomic location in the region displayed.

submicroscopic duplications and deletions that can not be elucidated using conventional karyotyping. More recently, single polymorphism nucleotide (SNP) arrays were developed, which can provide similar chromosomal information, but also can identify genomic regions with loss of heterozygosity or mosaicism. The number of oligonucleotides on one array now ranges from hundreds of thousands to millions, depending on the commercial platform, which allows for reliable detection of copy changes as small as 25 kb. Reliable detection of small chromosomal gains and losses are important in clinical diagnosis, as it is estimated that submicroscopic deletions and duplications may be the underlying cause of up to 15% of genetic diseases (47).

Before the wide-spread adoption of next generation sequencing, aCGH in conjunction with gene expression data was used to segregate breast cancer tumours based on their copy number changes, and to identify likely "driver" or commonly dysregulated genes (48). Andre et al. (2009) found the number of copy number aberrations in any given tumour can range dramatically, from 1 to 318 copy number changes, and averaged 76. There were a total of 48 minimum common regions with frequent copy number changes (11 gains, 37 losses) that were found in >20% of samples. In addition, 20 genes were amplified in at least 10 cases, of which 15 genes were overexpressed at the mRNA level. Tumours were classified based their copy number profile, and there was partial overlap between the gene expression subtypes and aCGH-based classifications: basal-like tumours were more frequently class I (77%), 53% of Luminal A cancers were in class III, and 67% of HER2 tumours were class II (48).

1.2.3 Gene Expression Microarrays

Gene expression microarrays involve hybridizing fluorescently labelled complementary DNA (cDNA) to microarrays slides containing probes of mRNA sequences, and use similar principles as chromosomal microarrays. cDNA Microarrays used to analyze gene expression were first described by Schena and colleagues in 1995 (49). Forty-five cloned cDNA transcripts from *Arabidopsis*, a small flowering plant, were printed onto a glass slide for subsequent gene expression measurement using fluorescently labeled probes using reverse transcription of mRNA. Only one year later, microarrays containing 1,046

human cDNAs were described by the same group, representing one of the first parallel gene analyses that measured differential gene expression patterns under given experimental conditions (50). In this study, control treated (37°C) and heat-treated (43°C) human T (Jurkat) cells were fluorescently labeled with different fluorophores to identify gene expression changes in the heat shock response. The technique of measuring expression levels relative to a control sample is still widely used today. As with chromosomal microarrays, the resolution and capabilities of gene expression microarray have been significantly advanced with the application of oligonucleotide-spotted arrays that contain thousands of individual probes (51). Normalized gene expression values measured from signal intensities of the array are commonly clustered to visualize and quantitatively identify differences between two samples or states (52). Groups of genes that share biological function, chromosomal location or regulation within the differential gene sets can be determined, which helps infer the biological processes contributing to the two conditions measured (53). Gene expression microarrays have been particularly impactful in breast cancer, and have led to the identification of the intrinsic subtypes of breast cancer (3) (see section 1.1.1) and the development of gene expression signatures that are used for patient management and prognosis, and are described in detail in section 1.4.

1.2.4 Next Generation Sequencing

DNA sequencing was first described by Sanger et al. in 1977, where chain-terminating dideoxynucleotides were incorporated into DNA strands by DNA polymerase during *in vitro* DNA replication (54). The first human genome was published over twenty years later in 2001, which was the result of a decade-long international collaboration of 20 groups (55,56). The availability of this, and other, whole genome reference assemblies allow short DNA strands to be mapped, or aligned, to already known sequences in the genome. The possibility of short-read sequencing enabled the advancement of next-generation DNA sequencing, which has been rapidly developed in recent years. Many sequencers now take less than a week to complete a reaction (57), and consequently, the cost per reaction has fallen dramatically, making it accessible for both research and clinical applications.

Next generation sequencing involves a multi-step process where the sample is prepared, sequenced, and then analyzed. Initially, DNA template preparation is required to ready the DNA sample for the specific sequencing platform being used. Briefly, this involves shearing the DNA to a smaller fragment size, ligating common primers (adapters) to both ends of the DNA fragments, and amplifying the template being sequenced (most commonly through emulsion PCR or solid-phase amplification) (58). DNA samples are also often enriched for a target sequence, such as all coding regions (whole exome sequencing) (59) or specific genomic loci of interest using customized capture methods (60). There are currently multiple next-generation platforms that can be used to perform sequencing, including ion semiconductor (Ion Torrent sequencing), Pyrosequencing (454 Life Sciences), sequencing by synthesis (Illumina), and sequencing by ligation (SOLiD sequencing). Sequences are generated using the detection of individual nucleotides or oligonucleotides at sequential positions in the nucleic acid fragments being sequenced. For example, sequencing by synthesis employs nucleotides that are fluorescently modified with a reversible chain terminator, each nucleotide with a different colour (61), resulting in the addition of only a single nucleotide with DNA polymerase in a given cycle. The reaction is performed over millions of clusters, each containing many identical copies of a DNA fragment. Clusters are imaged during each cycle, and the colour the cluster emits indicates the nucleotide at that position. At the end of the cycle, the terminator is cleaved, allowing for the next nucleotide to be added.

Once the sequencing portion is complete, the DNA sequences obtained (or "reads") must be aligned to the human reference assembly to determine their specific genomic location. Reads can be single-end (one end of the DNA library is sequence), or paired-end (both ends of the DNA fragments are sequenced, meaning the sequences are in close chromosomal proximity to each other), and can range from 35-150 bp (ie. Illumina) to an average of 400 bp (ie. 454 Life Sciences). Mapping reads to the correct location of the approximately 3 billion nucleotides in the genome with high accuracy is an enormous task. For this reason, there have been over twenty sequence alignment software programs developed (62) (for example, Bowtie (63), Bowtie2 (64), SOAP2 (65), MAQ (66), BWA (67) and RMAP (68)). Many of these programs apply or improve upon the Burrows Wheeler Transformation, which is an algorithm that can be used to compress character

strings (or in this case the DNA sequence) using runs of similar characters (69). Each tool has strengths and caveats. Mapping quality, in many cases, is compromised for shorter runtimes through neglecting base quality scores, limiting the number of tolerated base mismatches, disabling gapped alignment or limiting gap length, and ignoring SNP information (62). A study comparing 6 common alignment programs found that most tools underestimate their mapping quality, and inaccurate alignments can be eliminated by removing reads with a mapping quality of less than 1 (70).

After sequencing reads have been aligned to the genome, DNA variants, which are nucleotides or sets of nucleotides that differ from the reference genome, can be detected (most commonly SNPs, insertions, and deletions). Similar to the abundance of sequence alignments programs that are available, there are over thirty different programs that perform variant calling (71,72). The Genome Analysis Toolkit (GATK) (71), which was developed at the Broad Institute in Cambridge, Massachusetts, has become one of the most common and recommended programs used for variant calling (73,74). However, many other programs have strengths and are useful for certain types of experiments. For example, determining somatic mutations in cancer can be performed more effectively with programs specifically designed to compare tumour and matched normal sequences (75-78).

One of the largest hurdles the genomics community will likely face over the next decade is the clinical interpretation of variants in genomes, exomes, and transcriptomes that result from next generation sequencing studies. Differentiating between non-pathogenic, natural variation and likely damaging mutations can be extremely difficult, and has significant implications for disease-related research. Once a variant list is compiled, which can contain thousands of variants per sample, there are a number of different software programs that can aid in variant interpretation. Usually, variants are assessed to determine whether they are common polymorphisms (79,80) (natural variation in the population), and whether they are likely to be pathogenic. Software programs have been designed for a number of different purposes, for example: to annotate whether the variant resides in an exon or within other genomic regions (promoters, splice sites, CpG islands) (81,82), to predict the effect of the variant on the protein product (83-85), and to assess

whether the variant is likely to cause defects in mRNA splicing (86,87). With the development of numerous software programs with overlapping functions, selecting which programs to use for sequencing analyses can be difficult. Between the numerous options for sequencing platforms, read alignment algorithms, and variant calling and interpretation software, there are hundreds of potential pipelines or combinations of analyses that can be performed (Table 1.3). For clinical laboratories, the American College of Medical Genetics does not recommend any specific software programs for next-generation sequencing analysis, rather, it is recommended to select programs based on what type of genomic variation you are expecting and the depth of sequencing coverage, and to explain any variant filtering criteria while clearly outlining limitations of the approach (89).

1.3 DNA Variants

Advancements in our technical ability to reliably detect mutations in thousands of genes in a given patient has greatly outpaced our ability to interpret and report on the data collected in a clinical setting (90). Single nucleotide variants (SNVs) or small insertions/deletions (indels) can be located in exons (protein coding regions), introns (between exons), or non-coding regions. A typical sequencing study usually does not analyze mutations in non-genic regions, given the low likelihood of pathogenicity and difficulty to predict its affect on cellular function.

1.3.1 Protein Coding Mutations

There are three possible amino acid consequences for a single nucleotide variant found in a coding region of a gene, and they can be classified as silent, missense, and nonsense (stop) mutations. Silent mutations arise when a single nucleotide is altered, but the mutated codon results in the incorporation of the wild-type amino acid into the protein. Conversely, missense mutations occur when the mutation leads to an alteration of the amino acid at the position of the variant. Nonsense mutations lead to a premature stop codon within the coding sequencing, which results in protein truncation. Small indels can

Table 1.3 DNA sequencing software

Program	Citations*	Reference**
		ALIGNMENT SOFTWARE
Bowtie	4176 / 5854	Langmead B. Genome Biol. 2009;10:R25.
BWA	3930 / 5715	Li H. Bioinformatics 2009;25:1754–1760.
MAQ	1367 / 1975	Li H. Genome Res. 2008;18:1851–1858.
Bowtie2	1173 / 1887	Langmead B. Nat. Methods 2012;9:357–359.
SOAP2	997 / 1438	Li R. Bioinformatics 2009;25:1966–1967.
BWA-SW	938 / 1412	Li H. Bioinformatics 2010;26:589–595.
SSAHA2	563 / 828	Ning Z. Genome Res. 2001;11:1725–1729.
BFAST	231 / 367	Homer N. PLoS ONE 2009;4:e7767.
Stampy	227 / 333	Lunter G. Genome Res. 2011;21:936–939.
ELAND	NA / NA	Cox AJ. Illumina. 2007
Novoalign	NA / NA	Novoalign. http://novocraft.com.
		VARIANT CALLERS
SAMtools	3953 / 5624	Li H. Bioinformatics 2009;25:2078–2079.
GATK	1207 / 1756	DePristo MA. Nat. Genet. 2011;43:491–498.
SOAP SNP	997 / 1438	Li R. Bioinformatics 2009;25:1966–1967.
IMPUTE2	701 / 997	Howie BN. PLoS Genet. 2009;5:e1000529.
VarScan 2	280 / 427	Koboldt DC. Genome Res. 2012;22:568–576.
Dindel	174 / 237	Albers CA. Genome Res. 2011;21:961–973.
CORTEX	83 / 143	Iqbal Z. Nat. Genet. 2012;44:226–232.
SomaticSniper	84 / 117	Larson DE. Bioinformatics 2012;28:311–317.
Beagle	73 / 105	Browning BL. Am. J. Hum. Genet. 2009;85:847–861.
Strelka	55 / 89	Saunders CT. Bioinformatics 2012 28: 1811-7.
CRISP	66 / 86	Bansal V. Bioinformatics 2010;26:i318–324.
Atlas 2	58 / 82	Challis D. BMC Bioinformatics 2012;13:8.
SliderII	24 / 43	Malhis N. Bioinformatics 2010;26:1029–1035.
Bambino	21 / 31	Edmonson MN. Bioinformatics 2011;27:865–866.
GSNP	2 / NA	Lu M. Proc. Int. Conf. Parallel Processing. 2011; 6047227, 592-601
MuTect	NA / NA	https://confluence.broadinstitute.org/display/CGATools/MuTect.
		VARIANT INTERPRETATION
PolyPhen	2312 / 3178	Adzhubei IA. Nat. Methods 2010;7:248–249.
SIFT	1226 / 1682	Kumar P. Nat Protoc 2009;4:1073–1081
ANNOVAR	910 / 1307	Wang K. Nucleic Acids Res. 2010;38:e164.
ESEfinder	851 / 1137	Cartegni L. Nucleic Acids Res. 2003;31:3568–3571.
PANTHER	845 / 1118	Thomas PD. Genome Res. 2003;13:2129–2141.
ESRSearch	621 / 890	Fairbrother WG. Science 2002;297:1007–1013.
HSF	439 / 581	Desmet F-O. Nucleic Acids Res. 2009;37:e67.
VEP	263 / 390	McLaren W. Bioinformatics 2010;26:2069–2070
SNAP	244 / 341	Bromberg Y. Nucleic Acids Res. 2007;35:3823–3835.
MutationAssessor	182 / 272	Reva B. Nucleic Acids Res. 2011;39:e118.
MutPred	173 / 251	Li B. Bioinformatics 2009;25:2744–2750.
dbNSEP	144 / 219	Liu X. Hum. Mutat. 2011;32:894–899.
ABSOLUTE	141 / 218	Carter SL. Nat. Biotechnol. 2012;30:413–421
GSITIC 2.0	156 / 212	Mermel CH. Genome Biol. 2011;12:R41.
CUPSAT	148 / 190	Parthiban V. Nucleic Acids Res. 2006;34:W239–242
Align-GVGD	114 / 137	Mathe E. 2006;34:1317–1325
SNPnexus	66 / 91	Chelala C. Bioinformatics 2009;25:655–661.

*number of times the paper has been cited in Scopus / Google Scholar as of May, 2015

** Many of the programs and references were extracted from Pabinger et al.[88]

also lead to missense or nonsense changes, insertions or deletions of one or a group of amino acids, or can result in a frameshift mutation, where the 3-nucleotide frame of the coding region is altered, leaving the portion of the protein after the mutation with an incorrect amino acid sequence.

It is generally accepted that frameshift and nonsense mutations are the coding mutations most likely to be pathogenic or damaging to a protein. The clinical relevance or interpretation of these variants depends on the protein the mutation is found in, and whether that protein has a known cellular function or is cited to play a role in the phenotype being assessed. For example, a germ-line nonsense mutation in *BRCA1* in an individual or family would be reported to the patient, as the mutation puts the individual at risk for developing breast or ovarian cancer (91). Correlating phenotype to genotype is much more difficult in the case of missense and silent mutations, as the effect on the protein's function, if any, is hard to predict. However, clear and easy to interpret mutations account for a very low number of the overall mutational load that is detected in patients. In breast cancer, only 5-10% of families with a strong history of ovarian or breast cancer ever learn what the causal mutation is (92). For this reason, there have been many computational approaches, both sequence- and structure-based, that have attempted to assess the pathogenicity of missense mutations (93). Programs predicting splicing mutations or assessing their transcriptional effect, however, have been much more limited.

1.3.2 Splicing Mutations

Before proteins are translated, genes are transcribed and modified in a number of different ways: pre-mRNA splicing joins coding regions to be used during protein translation, and a 5' cap and 3' poly A tail are added to promote translation, and enable transcript transport and stability. Splicing involves over 100 factors (94), and is a multi-step process that results in the removal of introns from RNA transcripts, adjoining neighbouring exons contained in the final mRNA. Splicing machinery, known as splicesomes, are made up of multiple proteins and recognize key sequences to delineate the intron/exon junctions. The 5' and 3' ends of the intron (known as 5' donor or 3'

acceptor sites) contain canonical dinucleotides “GT” and “AG” (95), respectively, which identify the intron boundaries (Figure 1.3). Within the intron there is a polypyrimidine tract, (Py)n, and an adenine (A), known as the “branch site” that is used for lariat formation. Briefly, splicing is carried out using two transesterification steps, whereby a 2’-hydroxyl group of the adenine residue at the branch site attacks the phosphate at the donor site, leading to cleavage of the 5’ exon-intron boundary and lariat formation, and then subsequent attachment to the 3’ exon (as depicted in Figure 1.3) (96). The sequences spanning the intron/exon boundaries are conserved, but do have natural variation among different splice sites, which can be displayed as sequence logos (Figure 1.4 A). These sequences are 28 (acceptor) and 10 (donor) nucleotides in length, and dictate the overall strength of the splice site (or the likelihood of the splicing machinery recognizing the site) (98).

Splicing is used in the cell as an additional level of protein diversity and regulation. Various protein isoforms can be produced from the same gene through the inclusion of different combinations of exons in the final mRNA transcript used for protein translation. Alternative splicing is suggested to be one of the most important components of the functional complexity of the human genome, and is estimated to affect 40-60% of all human genes (99). This natural alternative splicing is usually not pathogenic, as different transcripts of the same gene are often expressed in tissue specific patterns (100).

Because splicing machinery relies on the pre-mRNA sequence to correctly remove an intron, mutations in these regions can lead to aberrant splicing that can damage or alter protein function. For example, if any of the highly conserved “GT” (or U in RNA) or “AG” nucleotides were altered, splicing would not properly occur at that intron (101). Although less common, mutations beyond these highly conserved dinucleotides in the splice site sequences (donor and acceptor) can lead to aberrant splicing and pathogenicity (97). The number of deleterious SNV splicing mutations described in the literature generally relates to the information content at each position of the sequence logo (Figure 1.4 B).

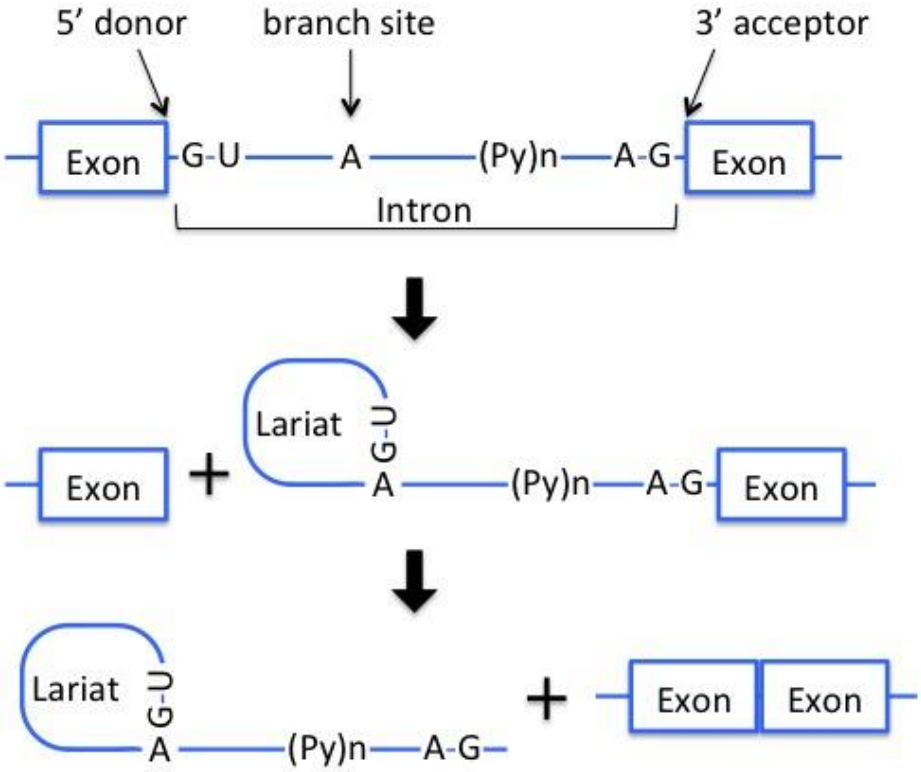


Figure 1.3 Basic schematic of pre-mRNA splicing. Diagram depicts an overview of mRNA splicing. Exons are indicated by the large blue-outlined boxes (as labeled), and introns are displayed as thin blue lines. Key nucleotides are labeled as “A”,”G”, and “U”, polypyrimidine tracts “(Py)n”.

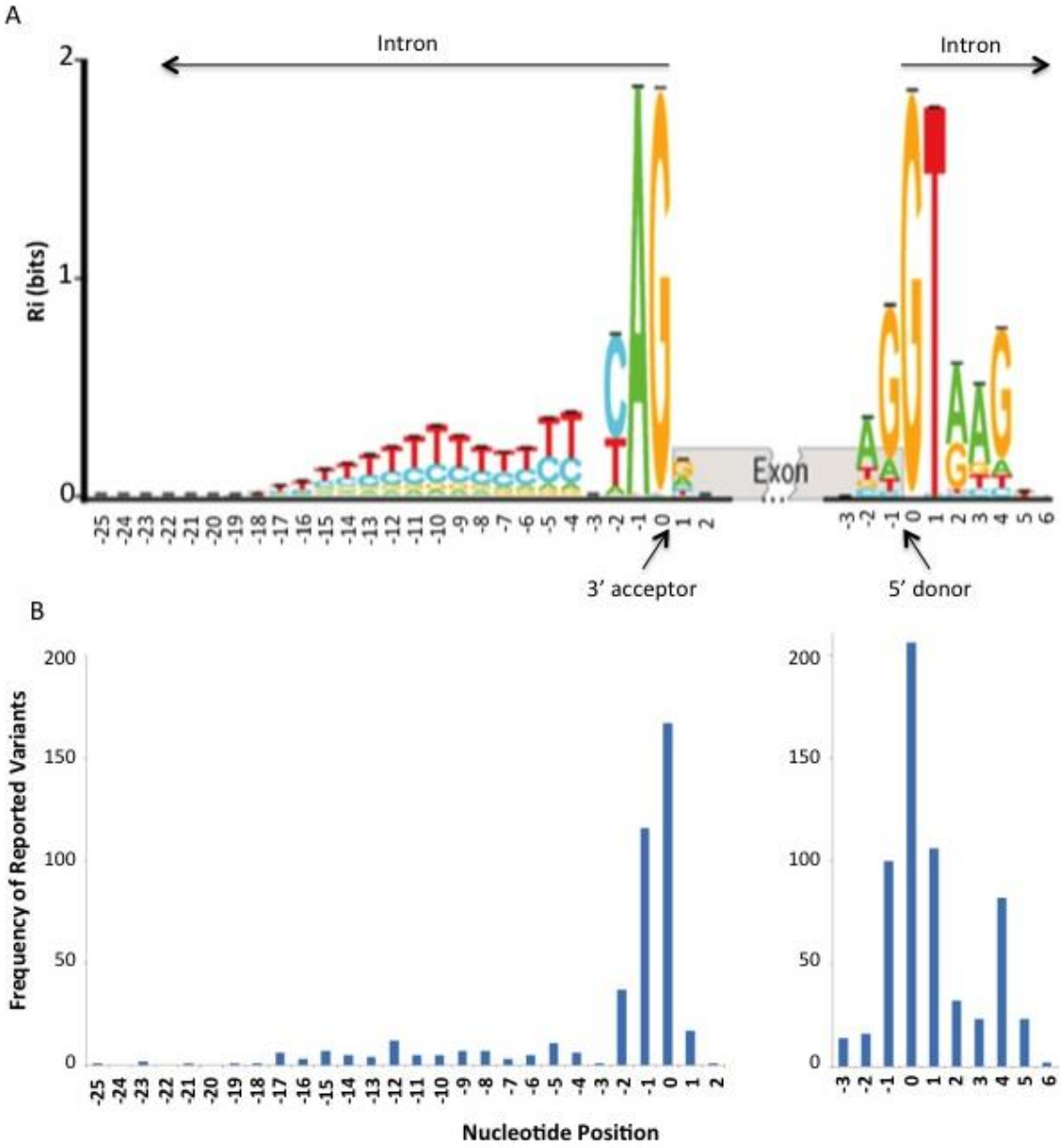


Figure 1.4 Splicing 5' donor and 3' acceptor sequence logos and frequency of reported mutations. A sequence logo for human acceptor (left) and donor (right) splice sites is displayed in A. The height of each nucleotide represents its frequency and the error bars indicate the standard deviation at that position. The distribution of deleterious single nucleotide variants reported in the literature to negatively affect splicing are displayed in B. This figure was adapted from Caminsky et al. (2014) (97) and Rogan et al. (2003) (98).

Splicing mutations can result in large changes to the final gene product, and hence, are commonly pathogenic. Up to 15% of all disease-causing mutations affect mRNA splicing (102), and this number is higher for certain genes, where splicing mutations can account for as many as 50% of the mutations reported (103). A number of different outcomes can arise from mutations that affect mRNA splicing. Mutations can inactivate a natural splice, which can result in the splicing machinery missing the corresponding donor or acceptor leading to intron retention (Figure 1.5 B), or the splicing machinery using a donor or acceptor from a neighbouring intron which would lead to exon skipping (Figure 1.5 C). An inactivating mutation at a natural splice site can also lead to the recognition of a weaker, so-called cryptic splice site in either the intron or exon, which would be recognized by the splicing machinery and result in the extension or truncation of the exon (termed cryptic splicing, Figure 1.5 D). A mutation may also activate a cryptic splice site, which would lead to a cryptic splicing phenotype (Figure 1.5 E).

Numerous software programs have been developed to analyze mutations and their potential effect on mRNA splicing. Commonly, splicing software programs require a DNA sequence containing the mutation as the input (Table 1.4). The program then determines the presence of splice sites or splicing regulatory factor binding sites, such as exonic splicing enhancers. The effect of the mutation on splicing can be determined by comparing the mutated sequence versus the wild-type sequence. Because of the nature of these programs, genome-wide capabilities are limited, and analysis of thousands of mutations filtered from next generation sequencing studies would be extremely laborious. There are also programs that analyze mutations, or lists of mutations, to determine if any affect splicing. Many of the common variant annotation and interpretation software programs are limited to identifying mutations that are likely to alter splicing by their location at the conserved dinucleotides (for example, ANNOVAR (81)), or within a limited splicing region (for example, Variant Effect Predictor looks as far as 8 nucleotides from the natural site (104)). The Automated Splice Site and Exon Definition Analyses (ASSEDA) (105), and the Shannon Human Splicing Pipeline (Shannon Pipeline) (86) software programs, however, employ information theory to extend the analysis to entire coding and non-coding regions of a gene. The application of information theory to DNA sequences was first proposed by Thomas D. Schneider in

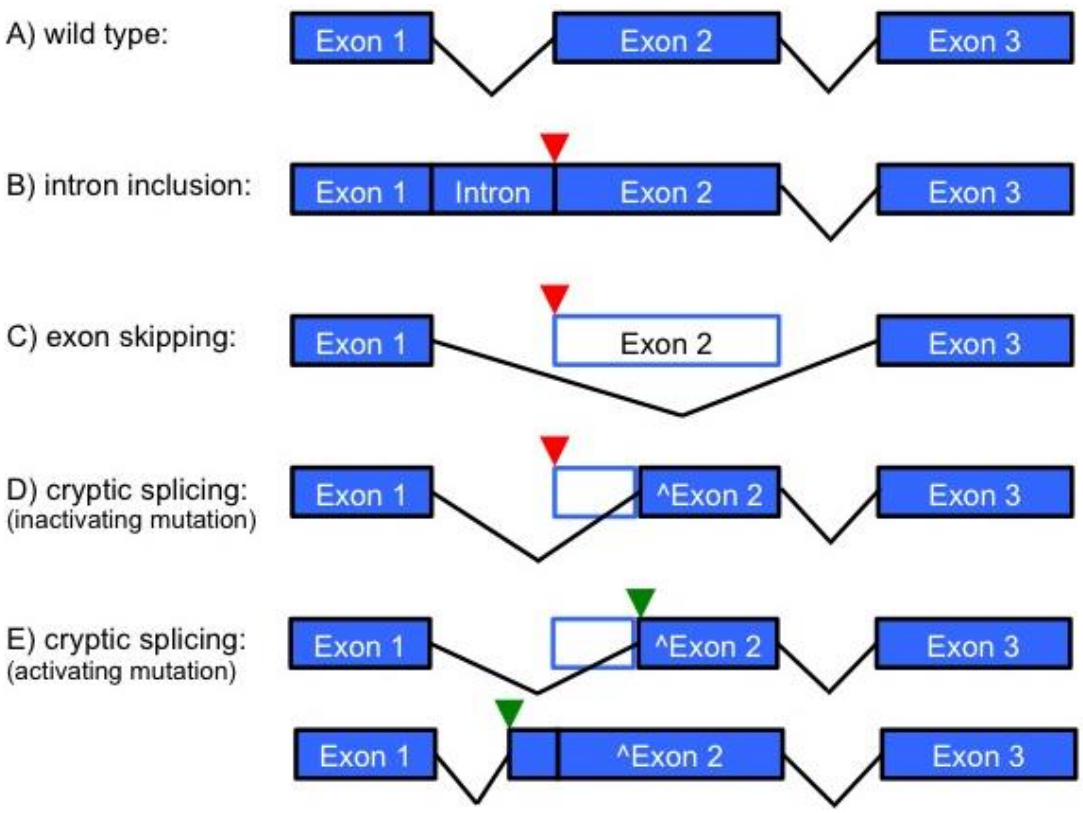


Figure 1.5 Aberrant splicing patterns resulting from DNA variants. Wildtype (A) and aberrantly spliced (B-D) transcripts are displayed to portray examples of the potential affect a mutation can have on mRNA splicing. Exons are indicated by the blue boxes, white filled in exons represents regions not maintained in the final transcript, and the black lines correspond to the sequences joined after splicing has occurred. Red arrows represent splice site-inactivating mutations, and green arrows represent activating mutations. A mutation can decrease the strength of a splice site, which can lead to intron retention in the final transcript (B), the affected exon being skipped and not retained in the final transcript, or a now-stronger neighbouring cryptic splice site to be used (resulting in part of the middle exon included in the final transcript). Alternatively, Exonic (E-top) or intronic (E-bottom) cryptic splices can be activated, resulting in the extension or truncation of the exon.

Table 1.4 Splicing mutation and splice site analysis software.

Program	Genome wide capability?	Analyzes mutations?	Analyzes sequences?
ANNOVAR[81]	yes	yes	no
Variant Effect Predictor[104]	yes	yes	no
ASSEDA[105]	no	yes	yes
Shannon Pipeline[86]	yes	yes	no
GeneSplicer[107]	yes	no	yes
Human Splice Finder[87]	no	yes	yes (max 2500 nts)
ESEfinder[108]	no	no	yes (max 5000 nts)
MaxEntScan[109]	no	no	yes (9 nt sequences)
Splice Site Prediction by Neural Network, NNSplice[110]	no	no	yes
NetGene2[111,112]	no	no	yes
SpliceView[113]	no	no	yes (max 31000 nts)
Splice Predictor[114]	no	no	yes
GenScan[115]	no	no	yes
Spliceman[116]	no	no	yes

1997 (106). The information theory-based approach is based on the formal relationship between information theory and the second law of thermodynamics. Each splice site is made up of information (measured in bits), which reflects the thermodynamic entropy and free energy of binding. The change in total information of the site is used to determine whether a mutation will strengthen or weaken the splice site.

The limited genome-wide capabilities and regions analyzed by most splicing software programs has led to an underrepresentation of splicing mutations in genome-wide studies. In the 5 major breast cancer sequencing studies (12-16), splicing mutations accounted for only 1.78-2.18% of all of the mutations reported (Table 1.5). This is likely due to the rudimentary approaches used to identify splicing mutations, which were limited to mutations located at the canonical dinucleotides at donor or acceptor sites. In addition, there are also currently limited efforts to attempt to validate the effect of splicing mutations on the mRNA transcript and protein product in large scale sequencing studies due to the large number of variants found, and efforts required for a single variant.

1.4 Gene expression signatures in breast cancer

The idea of personalized medicine is not new, however, clinical decision-making based on molecular profiling of individual tumours is still evolving. Early indications of personalized medicine date back to 1957, when 2 different papers suggested that genetic variation in enzymes may be linked to adverse drug response (117). Enzymes, such as cytochrome P450, can differ between individuals, which can determine how long and how much of the drug will remain active in the body. Characterizing an individual's metabolizing enzymes can dictate the dose required for effective response, therefore tailoring the treatment strategy for each patient. Personalized medicine has now extended well beyond analyzing drug metabolizing enzymes. Breast cancer, even more so than most cancers, is a mixture of several diseases, so it is intuitive that it would be ideal to tailor treatment and therapy selection on an individualized basis. Although it has been proposed for many years, it is now becoming feasible to determine the molecular

Table 1.5 Splicing mutation analyses performed in previous sequencing studies.

Paper	No. Splicing Mutations (Percent of all mutations)	Splicing mutation analysis	Validation approach to confirm affect on mRNA splicing.
Banerji[12]	97 (1.95%)	Oncotator – Used gene annotations to identify mutations at splice sites.	None
Ellis[13]	69 (2.15%)	Used gene structure to annotate "splice site" mutations.	None
Shah[15]	43 (1.78%)	Mutations were called using RNA sequencing - de novo splice sites were determined with HMMSplicer.	4 mutations were correlated to alternative splice junction usage in RNA sequencing data.
Stephens[16]	158 (2.18%)	Mutations mapped to essential splice sites.	None
TCGA[14]	506 (1.79%)	Annotated at "splice site" with gene annotation file.	None

profile of each tumour, and personalize each clinical decision based on certain characteristics. Consequently, there have been many studies applying gene expression analyses to individualize breast cancer management, including predicting prognosis, the benefit of adjuvant chemotherapy, tumour response to treatment, and development of new therapies.

1.4.1 Predicting prognosis and patient outcome

Research groups have been successful identifying and analyzing gene expression signatures in breast cancer that outperform conventional clinicopathologic criteria in predicting prognosis. These tests are effective in aiding to predict which patients are most likely to benefit from chemotherapy. The most common tests used today include Oncotype DX, MammaPrint, and PAM50. Oncotype DX is made up of a 21-gene assay, and provides a quantitative likelihood of disease recurrence (118). It was developed and commercialized based off of a study from 2004 that assessed the probability of breast cancer recurrence at 10 years using 668 Tamoxifen-treated, lymph-node negative, and estrogen-receptor positive tumours. The assay employs reverse-transcriptase polymerase chain reaction (RT-PCR), and measures the expression of the 21 selected genes to calculate a recurrence score (either low, intermediate, or high). Similarly, MammaPrint uses a 70-gene assay to identify early-stage breast cancer patients that are at risk of distant recurrence or metastasis following surgery (7). The assay differs in that it was developed independent of ER status or any prior treatment, contrary to Oncotype DX. The 70-gene signature was developed using DNA microarray analysis on primary tumours of 117 young patients, and stratifies patients that have “poor prognosis” and would likely benefit from adjuvant therapy. PAM50 is a 50-gene test that has been optimized to stratify tumours based on the intrinsic subtypes of breast cancer, which are then used to develop a risk of recurrence score (119). It was developed using both microarray and RT-PCR using 189 protoype samples, and then tested with an additional 761 patients to predict prognosis and 133 patients to predict complete pathological response to neoadjuvant chemotherapy. It is the only test of the three that directly leverages the intrinsic subtypes of breast cancer.

Gene expression signatures have an increased ability to recognize low-risk cases. This reduces the number of patients who receive adjuvant treatment, leads to a decrease in unnecessary toxicity, and lowers the cost of patient care (120). Regulatory bodies, such as the United States Food and Drug Administration (FDA), have recognized value and added benefit to patients by approving MammaPrint and PAM50 for clinical use, even though FDA approval is not required for laboratory-developed tests. Oncotype DX, which is currently the most commonly used test (121), is recommended by Cancer Care Ontario (122). The American Society of Clinical Oncology and the National Comprehensive Cancer Network also endorse these multigene assays to assist in treatment decisions for ER-positive cancer. Although these tests, and others, aid in deciding whether the patient would benefit from adjuvant therapy, clinicians still lack robust signatures that could indicate which specific treatments will be effective on a per patient basis (123,124).

1.4.2 Selecting therapies and predicting treatment response

Chemotherapy is currently recommended in cases where the benefit to the patient outweighs the risk of treatment. Conventional clinicopathological features indicating chemotherapy use for early breast cancer include histological grade 3 carcinomas, high Ki-67 levels, low hormone receptor status, *HER2* amplification or triple negative status, and tumours that have spread to three or more lymph nodes (125). Chemotherapy can be used in breast cancer treatment either before (neoadjuvant) or after (adjuvant) surgery. Adjuvant chemotherapy using cyclophosphamide, methotrexate, and fluorouracil (CMF) for lymph node positive breast cancer was first cited as an effective treatment strategy in 1976 (126), and was used until the substitution of methotrexate with epirubicin (CEF) (127) and then docetaxel (a taxane) (128) were later reported to be more successful combinations. Although clinical trials for many different adjuvant chemotherapy schedules have been conducted, there is ultimately still no consensus on which may be the most effective (129). Selection of the most effective adjuvant treatment is suggested to be individualized and should take into account clinical disease characteristics and patient-related factors (125,130,131). Treatment selection is already somewhat

personalized, profiling breast cancer tumours based on the intrinsic subtypes can direct recommendations in regard to endocrine, cytotoxic, and anti-HER2 therapies (Table 1.6).

Numerous studies have attempted to leverage genomic profiling in order to characterize or predict tumour response or patient outcome when treating with specific therapies (Table 1.7). Gene expression is most commonly used for this type of analysis, and signatures or indicators have ranged from including only a few genes to many. The majority of the studies performed to date are completed with a limited number of samples and/or patients. The availability of both training and test sets can be limited for specific therapies, but is increasing with dataset depositories such as the Gene Expression Omnibus (GEO). For example, in 2003, a 92-gene expression signature was created using 24 tumours, and was able to classify 10/11 sensitive tumours and 11/13 resistant tumours to neoadjuvant docetaxel in a leave-one-out analysis (132). No test set was used to validate the molecular profile, which was likely due to the limited availability of samples and/or the high costs for gene expression analyses at the time of the analysis. In contrast, a 20-gene signature was developed in 2014 to discriminate between chemoresistant and chemosensitive tumours to taxane-based therapies that used 160 tumours to develop the profile, and 659 datasets to test the method (133). Of the common prognostic gene signatures, the proliferation score from PAM50 is the only one that is able to identify patients that will benefit from a specific drug (low proliferation score predict weekly paclitaxel benefit) (134) or drug combinations (HER2-enriched tumours benefit from CEF over CMF) (135). Although our technological capabilities and access to data sets have greatly increased over the past decade, there is still no reliable genomic signature implemented in the clinic to select between chemotherapy agents on an individual basis.

1.5 The Minimal Breast Cancer Genome and its Relevance to Chemotherapy

Breast cancer studies to date have focused largely on genomic rearrangements, gene expression changes, and epigenetic alterations leading to the development and progression of the disease. Genomic regions in breast tumours that show high frequencies of abnormal rearrangements have been termed “saw-tooth” or “firestorm” regions (140),

Table 1.6 Treatment recommendations according to tumour subtype and/or receptor status

Subtype or receptor status	Type of therapy
Luminal A-like	endocrine therapy is the most critical, often used alone
Luminal B-like (HER2 -ve)	endocrine therapy for all, cytotoxic therapy for most
Luminal B-like (HER2 +ve)	cytotoxic therapy + anti-HER2 + endocrine therapy
HER2-positive	cytotoxic therapy + anti-HER2
Triple-negative	cytotoxic therapy

adopted from Schmidt et al. (2014) (129)

Table 1.7 Gene expression signatures developed to predict therapy response.

Study	Drug	Tumour	No. genes	No. training samples	No. test samples	Indication
Chang (2003)[132]	Docetaxel (neoadjuvant)^	Locally advanced	92	24 pre-operative core biopsies	N/A	Classifies tumours as sensitive or resistant (88% accuracy)
Ma (2004)[136]	Tamoxifen (adjuvant)*	Hormone receptor positive	2-gene ratio	60	N/A	Predictive of disease-free survival
Jansen (2005)[137]	Tamoxifen (first line treatment)*	ER-positive, advanced	44	46	66	Discriminate between patients with progressive disease and objective response
Hallet (2012)[138]	Chemotherapy regimens containing anthracycline and taxane drugs (neoadjuvant)^	N/A	2-gene index	488	N/A	Predicts complete pathological response
He (2014)[133]	Taxane-based therapies^	N/A	20	92 resistant / 68 sensitive	659 datasets	Discriminates between chemoresistant and chemosensitive individuals
Schmitt (2015)[139]	Trastuzumab and docetaxel (first line)*^	HER2-positive	8	79 frozen or FFPE	27 GEO datasets	Predicted response to treatment (76% accuracy)

*** hormone therapy, ^ chemotherapy**

and many genes have now been identified to be frequently mutated (14). However, studies that focus primarily on unaltered ("stable") regions in breast cancer have been limited to date.

Our laboratory has recently proposed that there is a minimal genome required for breast cancer cell survival (141). This minimal genome was derived by comparing independent data sets for regions of breast cancer genomes that are stable in copy number (140,142) with tumour gene expression levels that are similar to matched normal tissues (143,144). Genomic regions stable in copy number were obtained from two data sets that assessed a total of 243 (140) and 171 (142) primary breast tumours. The 812 derived "dually" stable regions (in both copy number and gene expression) contained a subset of 5,804 genes enriched for cellular metabolism, regulation of gene expression, DNA packaging, and regulation of apoptotic functions.

A selection of the stable genes identified are targets of existing anti breast-cancer therapies, including paclitaxel, estradiol, and topotecan. Growth inhibition of the breast cancer cell lines *MCF7*, *MDA-MB-231*, *HS578T*, and *T47D* has been demonstrated using therapeutic agents that target gene products of the stable regions (145). There was not, however, consistent drug sensitivity across all cell lines. The average GI50 values (drug concentrations are –log10M units) for paclitaxel, gemcitabine, and topotecan were found to be 8.07, 6.65, and 6.92 respectively. Cell lines with GI50 values lower than 1 unit from these averages are considered outliers, as this relates to a 10-fold increase in the concentration of drug required for 50% growth inhibition.

1.6 Thesis Scope and Objectives

In order to improve patient care through tailoring therapies based on disease characteristics, the field will require advancements in our ability to effectively interpret and analyze large genomic data. We hypothesize that improvements in genome-wide analyses - diagnostic tools and reagents for better detection of genetic abnormalities, mutation interpretation, and genomic signatures for chemotherapy response - can result in a more accurate understanding of tumour biology. This thesis introduces improvements in

both the design and analysis of experiments, and then applies these techniques to breast cancer. We describe the generation of new data, as well as leveraging existing data sets, with the same overall goal of validating the proposed methods and interpreting the results obtained.

Specifically, the main objectives of this thesis are to:

1) Improve the design and analysis of nucleic acid hybridization studies (specifically FISH and aCGH). We sought to develop a novel method to identify single copy regions in the genome that contain highly divergent repetitive elements, which we predicted to act as single copy sequences in optimized experimental conditions. We aimed to generate small, single copy FISH probes that contained divergent repetitive elements, to confirm their predicted behavior in metaphase FISH. In addition, we proposed that oligonucleotide placement throughout the genome (ie. distance to highly conserved repetitive elements) within these single copy regions can affect the variation observed in microarray signal intensities. Accordingly, we sought to develop an aCGH microarray to test this theory and compare the platform's reproducibility to a commonly used commercial platform. Both FISH and aCGH methods were validated on normal samples, as well as samples with known genomic alterations.

2) The large number of predicted DNA variants arising from genome-wide studies creates new challenges to validate the effect of any given variant on the mRNA transcript and protein product. Splicing mutations represent a unique set of variants that can be validated using mRNA sequences. We aimed to develop a software tool that can conduct genome-wide, statistically robust validation of predicted splicing mutations using sample-matched RNA sequencing data.

3) Splicing mutations are currently underpresented in large genomic studies, given that the majority of experiments only assess the canonical nucleotides at an intron/exon boundary. We aimed to carry out indepth splicing mutation analyses on a large set of breast cancer tumours using previously published data from the TCGA. We hoped to identify new underlying processes of tumour biology not

previously described by protein-coding dominated studies. Further, we planned to validate these mutations using the software described in objective 2.

4) Current selection of the specific cytotoxic agents to be used for breast cancer patient care does not consider analyzing the tumour for genes involved in drug disposition. In addition to the classical pathological features, these genes may be informative in identifying which tumours are the most likely to respond to certain therapies. This thesis aimed to use machine learning to develop predictive models of breast cancer tumour sensitivity to paclitaxel and gemcitabine. Rather than completing a genome-wide study, we sought to start with a much smaller set of biologically-relevant genes based on what is known about paclitaxel and gemcitabine drug mechanisms of action. In addition, a set of FFPE tumour samples was obtained from patients that were treated with paclitaxel and gemcitabine, and whose response to these drugs is known. In addition to previously published data sets, we planned to validate the predictive gene signatures on these FFPE samples through nucleic acid extraction and analysis. In turn, we wanted to determine whether high quality data could be obtained from FFPE samples, and how suitable their use would be in studies involving chemosensitivity predictions.

References

1. Jemal, A. *et al.* Global cancer statistics. *CA. Cancer J. Clin.* **61,** 69–90 (2011).

2. Canadian Cancer Society's Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2014. Toronto, ON: Canadian Cancer Society. (2014).

3. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).

4. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 10869–10874 (2001).

5. Herschkowitz, J. I. *et al.* Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol.* **8,** R76 (2007).

6. Goldhirsch, A. *et al.* Strategies for subtypes-dealing with the diversity of breast cancer: Highlights of the St Gallen international expert consensus on the primary therapy of early breast cancer 2011. *Ann. Oncol.* **22,** 1736–1747 (2011).

7. Van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–536 (2002).

8. Prat, A. & Perou, C. M. Deconstructing the molecular portraits of breast cancer. *Mol. Oncol.* **5,** 5–23 (2011).

9. Slamon, D. J. *et al.* Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2. *N. Engl. J. Med.* **344,** 783–792 (2001).

10. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).

11. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).

12. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486,** 405–409 (2012).

13. Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486,** 353–360 (2012).

14. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

15. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486,** 395–399 (2012).

16. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486,** 400–404 (2012).

17. Levine, A. J., Momand, J. & Finlay, C. A. The p53 tumour suppressor gene. *Nature* **351,** 453–456 (1991).

18. Bachman, K. E. *et al.* The PIK3CA gene is mutated with high frequency in human breast cancers. *Cancer Biol. Ther.* **3,** 772–775 (2004).

19. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer genes. *Nature* **499,** 214–218 (2013).

20. Dees, N. D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* **22,** 1589–1598 (2012).

21. Youn, A. & Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinforma. Oxf. Engl.* **27,** 175–181 (2011).

22. Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. Statistical Analysis of Pathogenicity of Somatic Mutations in Cancer. *Genetics* **173,** 2187–2198 (2006).

23. Varela, I. *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469,** 539–542 (2011).

24. Pritchard, A. L. & Hayward, N. K. Molecular pathways: mitogen-activated protein kinase pathway mutations and drug resistance. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **19,** 2301–2309 (2013).

25. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–352 (2012).

26. Bauman, J. G., Wiegant, J., Borst, P. & van Duijn, P. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA. *Exp. Cell Res.* **128,** 485–490 (1980).

27. Levsky, J. M. & Singer, R. H. Fluorescence in situ hybridization: past, present and future. *J. Cell Sci.* **116,** 2833–2838 (2003).

28. Cheung, V. G. *et al.* Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409,** 953–958 (2001).

29. Roohi, J., Cammer, M., Montagna, C. & Hatchwell, E. An improved method for generating BAC DNA suitable for FISH. *Cytogenet. Genome Res.* **121,** 7–9 (2008).

30. Goorha, S., Glenn, M. J., Drozd-Borysiuk, E. & Chen, Z. A set of commercially available fluorescent in-situ hybridization probes efficiently detects cytogenetic abnormalities in patients with chronic lymphocytic leukemia. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **6,** 48–53 (2004).

31. Mallo, M. *et al.* Fluorescence in situ hybridization improves the detection of 5q31 deletion in myelodysplastic syndromes without cytogenetic evidence of 5q-. *Haematologica* **93,** 1001–1008 (2008).

32. Kato, A., Albert, P. S., Vega, J. M. & Birchler, J. A. Sensitive fluorescence in situ hybridization signal detection in maize using directly labeled probes produced by high concentration DNA polymerase nick translation. *Biotech. Histochem. Off. Publ. Biol. Stain Comm.* **81,** 71–78 (2006).

33. Rogan, P. K., Cazcarro, P. M. & Knoll, J. H. M. Sequence-Based Design of Single-Copy Genomic DNA Probes for Fluorescence In Situ Hybridization. *Genome Res.* **11,** 1086–1094 (2001).

34. Alpár, D. *et al.* Automated FISH analysis using dual-fusion and break-apart probes on paraffin-embedded tissue sections. *Cytom. Part J. Int. Soc. Anal. Cytol.* **73,** 651–657 (2008).

35. Test and Technology Transfer Committee, American College of Medical Genetics, 9650 Rockville Pike, Bethesda, MD 20814-3998, United States. Technical and clinical assessment of fluorescence in situ hybridization: an ACMG/ASHG position statement. I. Technical considerations. Test and Technology Transfer Committee. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **2,** 356–361 (2000).

36. Wan, T. S. K. Cancer Cytogenetics: Methodology Revisited. *Ann. Lab. Med.* **34,** 413–425 (2014).

37. Shaw, A. T. *et al.* Clinical features and outcome of patients with non-small-cell lung cancer who harbor EML4-ALK. *J. Clin. Oncol.* **27,** 4247–4253 (2009).

38. Landstrom, A. P. & Tefferi, A. Fluorescent in situ hybridization in the diagnosis, prognosis, and treatment monitoring of chronic myeloid leukemia. *Leuk. Lymphoma* **47,** 397–402 (2006).

39. Wolff, A. C. *et al.* American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *J. Clin. Oncol.* **25,** 118–145 (2007).

40. Perez, E. A. *et al.* Immunohistochemistry and fluorescence in situ hybridization assessment of HER2 in clinical trials of adjuvant therapy for breast cancer (NCCTG

N9831, BCIRG 006, and BCIRG 005). *Breast Cancer Res. Treat.* **138,** 99–108 (2013).

41. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258,** 818–821 (1992).

42. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20,** 207–211 (1998).

43. Pollack, J. R. *et al.* Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23,** 41–46 (1999).

44. Barrett, M. T. *et al.* Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 17765–17770 (2004).

45. Miller, D. T. *et al.* Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.* **86,** 749–764 (2010).

46. Manning, M., Hudgins, L. & Professional Practice and Guidelines Committee. Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **12,** 742–745 (2010).

47. Vissers, L. E. L. M., Veltman, J. A., van Kessel, A. G. & Brunner, H. G. Identification of disease genes by whole genome CGH arrays. *Hum. Mol. Genet.* **14 Spec No. 2,** R215–223 (2005).

48. Andre, F. *et al.* Molecular Characterization of Breast Cancer with High-Resolution Oligonucleotide Comparative Genomic Hybridization Array. *Clin. Cancer Res.* **15,** 441–451 (2009).

49. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470 (1995).

50. Schena, M. *et al.* Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 10614–10619 (1996).

51. Hughes, T. R. *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* **19,** 342–347 (2001).

52. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 14863–14868 (1998).

53. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).

54. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467 (1977).

55. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

56. McPherson, J. D. *et al.* A physical map of the human genome. *Nature* **409,** 934–941 (2001).

57. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet. TIG* **24,** 133–141 (2008).

58. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).

59. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–276 (2009).

60. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27,** 182–189 (2009).

61. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).

62. Hatem, A., Bozdağ, D., Toland, A. E. & Çatalyürek, Ü. V. Benchmarking short sequence mapping tools. *BMC Bioinformatics* **14,** 184 (2013).

63. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

64. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

65. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinforma. Oxf. Engl.* **25,** 1966–1967 (2009).

66. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18,** 1851–1858 (2008).

67. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25,** 1754–1760 (2009).

68. Smith, A. D., Xuan, Z. & Zhang, M. Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* **9,** 128 (2008).

69. Burrows, M. & Wheeler, DJ. A block-sorting lossless data compression algorithm. *Digit. Equip. Corp.* (1994).

70. Ruffalo, M., LaFramboise, T. & Koyutürk, M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinforma. Oxf. Engl.* **27,** 2790–2796 (2011).

71. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

72. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25,** 2078–2079 (2009).

73. Pirooznia, M. *et al.* Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum. Genomics* **8,** 14 (2014).

74. Liu, X., Han, S., Wang, Z., Gelernter, J. & Yang, B.-Z. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE* **8,** (2013).

75. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–219 (2013).

76. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinforma. Oxf. Engl.* **28,** 311–317 (2012).

77. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinforma. Oxf. Engl.* **28,** 1811–1817 (2012).

78. Ding, J. *et al.* Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. *Bioinformatics* **28,** 167–175 (2012).

79. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).

80. Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* **28,** 352–355 (2000).

81. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

82. Kang, H. J., Choi, K. O., Kim, B.-D., Kim, S. & Kim, Y. J. FESD: a Functional Element SNPs Database in human. *Nucleic Acids Res.* **33,** D518–522 (2005).

83. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).

84. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al* **Chapter 7,** Unit7.20 (2013).

85. Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. & Casadio, R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* **30,** 1237–1244 (2009).

86. Shirley, B. C. *et al.* Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences. *Genomics Proteomics Bioinformatics* **11,** 77–85 (2013).

87. Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37,** e67 (2009).

88. Pabinger, S. *et al.* A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* **15,** 256–278 (2014).

89. Rehm, H. L. *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **15,** 733–747 (2013).

90. Anderson, M. W. & Schrijver, I. Next Generation DNA Sequencing and the Future of Genomic Medicine. *Genes* **1,** 38–69 (2010).

91. Easton, D. F., Ford, D. & Bishop, D. T. Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.* **56,** 265–271 (1995).

92. Howlader N, Noone AM, Krapcho M, Garshell J, Miller D, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z,Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2011, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2011/, based on November 2013 SEER data submission, posted to the SEER web site, April 2014.

93. Thusberg, J. & Vihinen, M. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum. Mutat.* **30,** 703–714 (2009).

94. Jurica, M. S. & Moore, M. J. Pre-mRNA Splicing: Awash in a Sea of Proteins. *Mol. Cell* **12,** 5–14 (2003).

95. Burset, M., Seledtsov, I. A. & Solovyev, V. V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28,** 4364–4375 (2000).

96. Pagani, F. & Baralle, F. E. Opinion: Genomic variants in exons and introns: identifying the splicing spoilers. *Nat. Rev. Genet.* **5,** 389–396 (2004).

97. Caminsky, N., Mucaki, E. J. & Rogan, P. K. Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. *F1000Research* **3,** (2014).

98. Rogan, P. K., Svojanovsky, S. & Leeder, J. S. Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. *Pharmacogenetics* **13,** 207–218 (2003).

99. Modrek, B. & Lee, C. A genomic view of alternative splicing. *Nat. Genet.* **30,** 13–19 (2002).

100. Brudno, M. *et al.* Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucleic Acids Res.* **29,** 2338–2348 (2001).

101. Wessagowit, V., Nalla, V. K., Rogan, P. K. & McGrath, J. A. Normal and abnormal mechanisms of gene splicing and relevance to inherited skin diseases. *J. Dermatol. Sci.* **40,** 73–84 (2005).

102. Krawczak, M., Reiss, J. & Cooper, D. N. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.* **90,** 41–54 (1992).

103. Ars, E. *et al.* Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. *Hum. Mol. Genet.* **9,** 237–247 (2000).

104. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).

105. Mucaki, E. J., Shirley, B. C. & Rogan, P. K. Prediction of Mutant mRNA Splice Isoforms by Information Theory-Based Exon Definition. *Hum. Mutat.* **34,** 557–565 (2013).

106. Schneider, T. D. Information content of individual genetic sequences. *J. Theor. Biol.* **189,** 427–441 (1997).

107. Pertea, M., Lin, X. & Salzberg, S. L. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* **29,** 1185–1190 (2001).

108. Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q. & Krainer, A. R. ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.* **31,** 3568–3571 (2003).

109. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **11,** 377–394 (2004).

110. Reese, M. G., Eeckman, F. H., Kulp, D. & Haussler, D. Improved splice site detection in Genie. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **4,** 311–323 (1997).

111. Hebsgaard, S. M. *et al.* Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* **24,** 3439–3452 (1996).

112. Brunak, S., Engelbrecht, J. & Knudsen, S. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* **220,** 49–65 (1991).

113. Rogozin, I. B. & Milanesi, L. Analysis of donor splice sites in different eukaryotic organisms. *J. Mol. Evol.* **45,** 50–59 (1997).

114. Brendel, V., Xing, L. & Zhu, W. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinforma. Oxf. Engl.* **20,** 1157–1169 (2004).

115. Burge, C. in *Computational Methods in Molecular Biology* 127–163 (Elsevier Science, 1998).

116. Lim, K. H. & Fairbrother, W. G. Spliceman--a computational web server that predicts sequence variations in pre-mRNA splicing. *Bioinforma. Oxf. Engl.* **28,** 1031–1032 (2012).

117. Marshall, A. Laying the foundations for personalized medicines. *Nat. Biotechnol.* **15,** 954–957 (1997).

118. Paik, S. *et al.* A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *N. Engl. J. Med.* **351,** 2817–2826 (2004).

119. Parker, J. S. *et al.* Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).

120. Desmedt, C. *et al.* Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **14,** 5158–5165 (2008).

121. Győrffy, B. *et al.* Multigene prognostic tests in breast cancer: past, present, future. *Breast Cancer Res.* **17,** 11 (2015).

122. Chang, M. *et al.* Comparison of Oncotype DX with Multi-gene Profiling Assays, (e.g., MammaPrint, PAM50) and Other Tests (e.g., Adjuvant! Online, Ki-67 and IHC4) in Early-stage Breast Cancer:Recommendations. *Recomm. Rep. MOAC 2* 1–39 (2013).

123. Cleator, S. & Ashworth, A. Molecular profiling of breast cancer: clinical implications. *Br. J. Cancer* **90,** 1120–1124 (2004).

124. Weigelt, B., Pusztai, L., Ashworth, A. & Reis-Filho, J. S. Challenges translating breast cancer gene signatures into the clinic. *Nat. Rev. Clin. Oncol.* **9,** 58–64 (2012).

125. Goldhirsch, A. *et al.* Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. ESMO* **24,** 2206–2223 (2013).

126. Bonadonna, G. *et al.* Combination chemotherapy as an adjuvant treatment in operable breast cancer. *N. Engl. J. Med.* **294,** 405–410 (1976).

127. Levine, M. N. *et al.* Randomized trial comparing cyclophosphamide, epirubicin, and fluorouracil with cyclophosphamide, methotrexate, and fluorouracil in premenopausal women with node-positive breast cancer: update of National Cancer Institute of Canada Clinical Trials Group Trial MA5. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **23,** 5166–5170 (2005).

128. Roché, H. *et al.* Sequential adjuvant epirubicin-based and docetaxel chemotherapy for node-positive breast cancer patients: the FNCLCC PACS 01 Trial. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **24,** 5664–5671 (2006).

129. Schmidt, M. Chemotherapy in early breast cancer: When, how and which one? *Breast Care* **9,** 154–160 (2014).

130. Cardoso, F. *et al.* Locally recurrent or metastatic breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **23,** vii11–vii19 (2012).

131. Oostendorp, L. J. M., Stalmeier, P. F. M., Donders, A. R. T., van der Graaf, W. T. A. & Ottevanger, P. B. Efficacy and safety of palliative chemotherapy for patients

with advanced breast cancer pretreated with anthracyclines and taxanes: a systematic review. *Lancet Oncol.* **12,** 1053–1061 (2011).

132. Chang, J. C. *et al.* Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet Lond. Engl.* **362,** 362–369 (2003).

133. He, D.-X., Xia, Y.-D., Gu, X.-T., Jin, J. & Ma, X. A 20-gene signature in predicting the chemoresistance of breast cancer to taxane-based chemotherapy. *Mol. Biosyst.* **10,** 3111–3119 (2014).

134. Martín, M. *et al.* PAM50 proliferation score as a predictor of weekly paclitaxel benefit in breast cancer. *Breast Cancer Res. Treat.* **138,** 457–466 (2013).

135. Cheang, M. C. U. *et al.* Responsiveness of Intrinsic Subtypes to Adjuvant Anthracycline Substitution in the NCIC.CTG MA.5 Randomized Trial. *Clin. Cancer Res.* **18,** 2402–2412 (2012).

136. Ma, X.-J. *et al.* A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* **5,** 607–616 (2004).

137. Jansen, M. P. H. M. *et al.* Molecular classification of tamoxifen-resistant breast carcinomas by gene expression profiling. *J. Clin. Oncol.* **23,** 732–740 (2005).

138. Hallett, R. M., Pond, G. & Hassell, J. A. A target based approach identifies genomic predictors of breast cancer patient response to chemotherapy. *BMC Med. Genomics* **5,** 16 (2012).

139. Schmitt, E. *et al.* Transcriptional expression of 8 genes predicts pathological response to first-line docetaxel + trastuzumab-based neoadjuvant chemotherapy. *BMC Cancer* **15,** (2015).

140. Hicks, J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* **16,** 1465–1479 (2006).

141. Park, N. I., Rogan, P. K., Tarnowski, H. E. & Knoll, J. H. M. Structural and genic characterization of stable genomic regions in breast cancer: relevance to chemotherapy. *Mol. Oncol.* **6,** 347–359 (2012).

142. Chin, S. F. *et al.* High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.* **8,** R215 (2007).

143. Turashvili, G. *et al.* Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer* **7,** 55 (2007).

144. Naderi, A. *et al.* A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene* **26,** 1507–1516 (2007).

145. Staunton, J. E. *et al.* Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 10787–10792 (2001).

Chapter 2

2 Expanding probe repertoire and improving reproducibility in human genomic hybridization

The work presented in this chapter is reproduced (with permission, Appendix S1) from:

Dorman, S.N., Shirley, B.C., Knoll, J.H.M., Rogan, P.K. (2013) Expanding probe repertoire and improving reproducibility in human genomic hybridization. *Nucleic Acids Research* 41(7): e81.

2.1 Introduction

Genome-derived nucleic acid hybridization probes are routinely used diagnostically to identify, detect or quantify specific DNA sequences. It has long been recognized that repetitive sequences in these probes can interfere with the detection of chromosome abnormalities through cross hybridization to multiple regions of the genome. This is because repetitive sequences comprise at least 50% of the human genome and consist of a diverse set of distinct families (1) with variable degrees of divergence, many of which are conserved throughout mammalian evolution (2,3). Elimination of these sequences is a key consideration in genomic probe and experimental design. These sequences can be sequestered away from unique sequences in labelled probes (4,5), 'blocked' with unlabelled Cot-1 DNA (6–8), or eliminated from the probe sequence by masking all elements related to known repetitive sequence families (9). We present an approach to improve the genomic resolution and reproducibility of fluorescent in situ hybridization (FISH) and microarray comparative genomic hybridization (aCGH). Inclusion of evolutionarily highly divergent repetitive elements increases genomic coverage without compromising the specificity of FISH and aCGH to the extent that conserved repetitive sequences would. Contextual effects of proximate, conserved repetitive sequences on probe design are also investigated.

FISH is an essential diagnostic tool for detection of contextual chromosome rearrangements. However, the diversity of relevant chromosomal abnormalities seen in patients with cancer or congenital diseases far exceeds the catalogue of available recombinant probes. Commercial FISH probes often include multiple genes, which reduces their specificity for targeting abnormalities confined to individual genes. The Cancer Genome Project (10) has identified translocations in 317 cancer genes implicated in oncogenesis, 177 of which are <100 kb. Single copy FISH (scFISH) involves sequence-based genomic DNA probes that are 100–500-fold smaller than commercial FISH probes (11), thus providing the higher resolution necessary for specific detection of contextual changes within small genes. Nevertheless, repeat-masked probes contain exclusively unique genomic sequences, which limit access in genomic regions densely populated with repetitive elements for scFISH.

aCGH determines copy number variation genome wide (12–14). It has been widely adopted in cancer research, disease gene discovery, prenatal diagnostics and has improved clinical diagnosis for patients with congenital and acquired diseases (15,16). aCGH has been recommended by the American and Canadian Colleges of Medical Genetics as a first-line test for individuals with development disabilities or congenital anomalies (17,18). Despite the ubiquity of this test, the accuracy and reproducibility of aCGH has recently been questioned (19–21). A study assessing 11 copy number variant (CNV) microarray platforms reported <50% similarity in CNV calls between software and analytical tools and <70% reproducibility in most replicate experiments (21). Multiple sources of data from different commercial platforms, analysed with the same software, call inconsistent copy number changes (CNC) (20), implicating the primary data as a significant contributor to this variability.

In FISH and aCGH, non-specific cross-hybridization to other genomic locations is most commonly prevented by sequestering repetitive sequences with excess unlabelled Cot-1 DNA (7,22). Addition of Cot-1 reduces consistency and increases variability in genomic hybridization to homologous targets, regardless of whether repetitive elements are present in the labelled DNA (23). Cot-1 DNA contains sc sequence impurities that increase variability in hybridizations. Probe sequences have also been designed to be

devoid of repetitive elements by synthesis of repeat-masked unique or sc intervals (9). However, the use of Cot-1 DNA in aCGH is unavoidable in order to prevent cross-hybridization between non-allelic repetitive regions in the labelled sample.

The proximity of repetitive elements to sc targets and the extent to which these sequences are conserved have not been considered in microarray probe design. We find that unique sequence microarray probes in close proximity to adjacent repetitive sequences, contribute to poor reproducibility of hybridization intensities, and the degree of repeat sequence divergence can affect the variability of hybridization intensities of these unique sequence probes. By mitigating these effects, it is possible to improve the genomic resolution and reproducibility of FISH and aCGH.

2.2 Materials and methods

2.2.1 scFISH probe design

We deduced a complete set of effectively sc regions using an *ab initio* divide-and-conquer search algorithm (24,25) directly from the reference human genome (GRCh37/hg19) (Appendix S2.1). This algorithm identified sc intervals without reliance on a catalogue of existing repetitive elements. The search constraints were tuned to include sequences containing highly divergent repetitive elements. Divergent copies of repetitive elements deviate sufficiently from conserved consensus sequences so as to preclude cross-hybridization to non-allelic genomic locations. A genome-wide set of *ab initio* sc intervals was derived and displayed as custom genome browser tracks. From these intervals, 15 scFISH probes >1.5 kb were designed to detect rearrangements within 10 small cancer-related onco- and tumour-suppressor genes (<50 kb; *CCND1*, *CDKN2A*, *CDKN2C*, *ERBB2*, *FGFR3*, *FLCN*, *KRAS*, *MYCN*, *NOTCH1*, *TP53*) designated by the Sanger Institute Cancer Genome Project (10). Regions of at least 2.5 kb for scFISH were used for primer design for long polymerase chain reaction (PCR) as previously described (9). Appendix S2.2 indicates the eight probes that were produced, their genomic coordinates, length and primer sequences.

Divergent repetitive elements included in each probe were localized by genome-wide Basic Local Alignment Search Tool (BLAST) and analysed for degree and extent of divergence from consensus sequences of the same repeat family or subfamily. To estimate stability of probe sequences, nick translation products of 300 nucleotides (nt) were simulated by windowing along the length of a probe. Melting temperatures (Tm) for each imperfect duplex were estimated (26) and then plotted for higher and lower stringency, post-hybridization experimental wash conditions (2X SSC, 37°C, 50% formamide; and 2X SSC, 42°C, 50% formamide). With more stringent post-hybridization washing conditions, the divergent repetitive elements were not expected to cross-hybridize to non-allelic genomic loci. Related, non-allelic sequences in the human genome were detected by BLAST analysis. All imperfect duplexes were estimated to exhibit predicted Tm at least 10°C lower than the homologous targets.

The performance of eight probes containing divergent repetitive elements was validated by scFISH to human metaphase cells with a normal karyotype. Primers for a genome-wide set of *ab initio* scFISH probes were designed using Primer 3 (27). Probe length and maximum Tm differences were optimized to produce the highest quality probes while maintaining genomic resolution. Primers were designed for intervals between 1.5–2 and 3.5–4 kb, with maximum Tm differences set at 0.5°C, 1°C and 2°C. scFISH probes produced with maximum Tm differences did not significantly vary; therefore, 0.5°C was used to ensure the highest quality PCR amplification. Primer3 parameters used to generate the 1500–2000 bp products were PRIMER_OPT_SIZE = 27, PRIMER_MAX_SIZE = 28, PRIMER_MIN_SIZE = 26, PRIMER_ PRODUCT_SIZE_RANGE = 1500–2000, PRIMER_PAIR_MAX_DIFF_TM = 0.5, PRIMER_OPT_TM = 63, PRIMER_MAX_TM = 65, and PRIMER_MIN_TM = 61. To generate 3500–4000 bp products, the parameters used were PRIMER_OPT_SIZE = 33, PRIMER_MAX_SIZE = 35, PRIMER_MIN_SIZE = 30, PRIMER_PAIR_ MAX_DIFF_TM = 0.5, PRIMER_PRODUCT_SIZE_RANGE = 3500–4000, PRIMER_OPT_TM = 64, PRIMER_MAX_TM = 66, PRIMER_MIN_TM = 62.

2.2.2 scFISH probe development and hybridization

Ab initio sc products were optimized by gradient thermal cycling, then amplified using long PCR with Platinum Pfx DNA Polymerase (Invitrogen™, CA). Amplicons were gel purified, extracted (QIAquick kit, Qiagen CA) and labelled by nick translation with digoxigenin-11-dUTP (Roche Diagnostics, ON, Can). Probes were hybridized on normal human lymphocyte metaphase chromosomes, detected with Cy3-conjugated anti-digoxin antibody (Cedarlane, CA), then washed and stained with 4',6-diamidino-2-phenylindole (DAPI) (28). At least 20 metaphases from cytogenetic preparations of control individuals were examined for each probe to confirm the chromosome location and hybridization efficiency. A probe from *CDKN2A,* which is abnormal in the preponderance of melanomas, was also hybridized to metaphase chromosomes of the melanoma cell line A-375 (29).

2.2.3 Genome-wide aCGH

A pool of suitable oligonucleotide probes from *ab initio* intervals was designed with PICKY (30), which matches melting temperatures to avoid complementarity between probes and stable hairpin formation. Default parameters were modified as follows: left selection boundary 200, right selection boundary 200, maximum oligonucleotide size 60, maximum match length 20, minimum match length 17 and probes per gene 5. PICKY-suggested 2 057 653 coordinate-defined probes from 513 689 *ab initio* sc intervals.

A subset of these probe sequences was selected to populate a custom genome-wide 4x44K array. To minimize cross-hybridization of *ab initio* probes to repetitive sequences within the labelled genomic sample, oligonucleotides were chosen complimentary to genomic targets whose distance to an adjacent conserved repetitive element exceeded the length of the labelled extension products. Products were <300 nt. Oligonucleotide targets and adjacent repeat elements were separated by at least 300 nt, for repetitive sequences with <30% divergence (higher divergence sequences were tolerated). For purposes of comparison, *ab initio* oligonucleotide targets were paired with Agilent Technologies Human Catalog CGH 4 × 44K microarray (Agilent 44K) genomic probe sequences in

closest genomic proximity to ensure similar distributions. Where possible, gene coverage was maximized. The Galaxy metaserver (https://main.g2.bx.psu.edu) was used to 'fetch' the closest non-overlapping feature for every interval, 'subtract' intervals present in the *ab initio* and Agilent 44K oligonucleotide sets and determine the base 'coverage' of all intervals. We first determined the distance in nt of the closest repeat masked repetitive element to each probe. Oligonucleotides within 300 nt of a repeat were subtracted from the set. The closest *ab initio* probe to a corresponding sequence on the Agilent 44K array was fetched. The distance between *ab initio* probes and adjacent repeat elements was then maximized on the custom designed microarray by selecting oligonucleotides central to each *ab initio* interval. Gene coverage, which was determined from the proximity of probes to known NCBI RefSeq gene sequences, demonstrated that the paired set of *ab initio* probes did not cover all known genes (31). Gene coverage in the custom microarray was improved by adding 1510 probes within or adjacent to the missing genes.

Ab initio normalization and replicate probes were also selected in close proximity of the corresponding Agilent probes. Both the custom designed *ab initio* 44K and commercial Agilent 44K microarrays were manufactured by Agilent. We hybridized them with genomic DNA from HapMap family trios (YRI: GM19143/GM19144/GM19415, and CEU: GM07019/GM07056/GM07022). DNA from the offspring (GM19145/GM07019) was used as the reference sample and co-hybridized with either the maternal (GM19143/GM07056) or paternal (GM19144/GM07022) sample on two replicate sectors of each array. To produce extension products <300 nt, DNA was subjected to heat fragmentation (98°C for 10′) before labelling and sized by electrophoresis. Pairs of genomic DNA samples (0.5 µg each) were individually enzymatically labelled using 5′-terminally labelled, fluorescent random nonamers (either Cy3 or Cy5 from IDT) with 5′→3′-exo- Klenow DNA polymerase (New England Biolabs), then mixed and co-hybridized according to the Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis Protocol (v6.2). Microarrays were scanned and quantified with Agilent Feature Extraction software (v10.5.1.1). Hybridization intensities of Agilent's non-human control sequences were used to correct for background fluorescence. The coefficients of variation [CV = |(Log2 ratio or signal intensity) standard deviation|/mean] were calculated from replicate spot intensities of each autosomal probe sequence on the same

microarray platform. Identical probe sequences were replicated within the same and on different sectors on the array, enabling comparisons of both inter- and intra-array reproducibility on each platform.

2.2.4 Locus-specific aCGH

Reusable 12K oligonucleotide microarrays were produced using a microarray DNA synthesizer in our laboratory (CustomArray, Bothell, WA). Duplicate arrays containing either *ab initio* sc probes or the published Agilent 44K array probe sequences were manufactured. These arrays were designed to contain a higher concentration of probes mapping within chromosome 15q11.2q13 to fully assess CNCs present in patient samples with chromosome abnormalities in this region. In all, 125 *ab initio* sc probes and 84 published Agilent 44K probes were replicated multiple times on each respective array. The remaining array content had genome-wide distribution which maximized gene coverage and minimized the distance between the pairs of Agilent and *ab initio* derived probe sequences.

Genomic DNA from WJK35, an Angelman syndrome (AS) patient cell line with a previously mapped chromosome 15 deletion (32) was used to assess reproducibility for calling copy number differences. DNA was labelled with random Cy5 nonamers as indicated earlier in the text. Each array was hybridized, washed and scanned, then stripped and re-hybridized with the same labelled DNA product. One of the microarrays could not be re-hybridized to a labelled DNA after the initial hybridization study because it failed a quality control test for intra-array reproducibility. For all of the other arrays, labelled genomic DNA was removed from the microarrays after the initial hybridization (Stripping Kit, CustomArray) and then re-imaged. Array performance was assessed for quality control by re-hybridizing a Cy5-labelled, random nonamer, which verifies probe integrity and consistency of signal intensity before subsequent re-hybridization. Custom microarrays were imaged with an Axon GenePix 4000 B microarray scanner (Molecular Devices US). CNV was analysed with Nexus 6.0 (Biodiscovery US) software.

2.3 Results

2.3.1 Genome-wide coverage of *ab initio* sc intervals

The density and coverage of unique sequences for hybridization studies in any genomic region is finite, and in some instances, underrepresented in regions associated with disease or relevant to gene regulation and expression. For example, more than one-fifth of RefSeq genes are covered >50% in gene lengths by repetitive elements (31). We implemented an *ab initio* algorithm, which does not require a catalogue of repetitive elements to locate all genomic intervals devoid of multicopy sequences (Appendix S2.1). The density and lengths of contiguous DNA sequences used for probe design were increased by tuning sequence alignment stringency to include divergent repetitive elements with hybridization kinetics similar to sc sequences, at the same time avoiding segmentally duplicated and self-chained alignments of close paralogues. Before selecting scFISH and microarray probes, the distribution of *ab initio* intervals was characterized among previously annotated genomic features. Overlapping, adjacent intervals were merged to generate contiguous sequences of maximal length, then compared with the complement of the collective set of annotated repetitive features with an exclusive disjunction (OR) operation (1,33–36). The coverage or sensitivity for the *ab initio* set of intervals comprised 87% of the complementing sequences. The specificity was 83%, indicating 17% contained multicopy sequences. However, alignments to human self-chained, paralogous sequence families comprised >90% of these false positive intervals, necessitating an additional filtering step to eliminate these potential probes.

The *ab initio* probe intervals were densely distributed along chromosomes, with >50% of intervals exceeding 1 kb. Less than 0.2% of all *ab initio* intervals were separated by >32 kb, with the majority (98%) occurring <8 kb apart. Gaps in the reference sequence assembly accounted for many of the widely separated *ab initio* regions. Gene coverage was assessed for *ab initio* intervals ≥50 nt to define potential targets for probe design of oligonucleotides for both aCGH and FISH. Genes with ≥50% coverage by *ab initio* intervals ranged from 5% of those on the Y chromosome to 84% of those on chromosome 18. On average, <8% of genes were completely missed by the *ab initio* algorithm (from

3% on chromosome 3 to 87% on the Y chromosome). Genes ≤20 kb comprised 90% of the genes without coverage. *Ab initio* intervals overlapped other genomic annotations (at genome.ucsc.edu), including 85% of CpG islands, 99% of Vista enhancers, 98% of transcribed, ultraconserved intergenic sequences and 97% of intragenic sequences. *Ab initio* sequence intervals covered the majority of disease-associated genes in the Catalogue of Somatic Mutations in Cancer (COSMIC) (84%), Gene Reviews (93%) and Pathogenic International Standards for Cytogenomic Arrays (ISCA) gene (95%) databases.

We then designed genome-wide sets of *ab initio* scFISH probes. PCR primer pairs were selected for 957 304 scFISH probes >1.5 kb from 194 795 unique genomic intervals (Supplementary Table 2.1, for all Supplementary Tables see the "Additional Files" electronic document). Of these, 455 978 of the scFISH probes overlap with known genes. Gene coverage varied from 48 to 58% for scFISH probes designed to be 1.5–2 kb and 3.5–4 kb, respectively. These two subsets of FISH probes together cover 71% of NCBI RefSeq genes. The median distance between adjacent scFISH probes is 6140 nt, with 89.5% of scFISH probes occurring within 25 kb of each other.

A set of oligonucleotides was designed for production of genome-wide and regionally targeted aCGH platforms. A total of 2 057 649 oligonucleotide sequences were derived, 756 235 of which were separated by at least 300 nt from the nearest conserved repetitive sequence (Supplementary Table 2.2). Oligonucleotide hybridization to these target sequences should reduce variability in signal intensities by minimizing cross-hybridization of labelled DNA to repetitive regions in non-target or Cot-1 DNA (23) and prevent sequestration of labelled sc sequences linked to cross-hybridizing adjacent repetitive sequences (37). The full oligonucleotide set covers 84.7% of known genes, whereas the reduced subset of well-separated sc targets covers 81.5%. The reduced subset of adjacent sc probes is separated from each other by ≤25 kb, with a median distance of 1.094 kb. Exceptionally long inter-probe intervals (>250 kb; n = 176) either occurred in centromeric regions, were enriched in multicopy sequences (i.e. paralogous self-chained alignments or segmental duplications), or were unsequenced.

2.3.2 *Ab initio* scFISH probes

Cytogenetic rearrangements involving small cancer genes (<50 kb) have been documented; however, large commercial FISH probes may not provide adequate specificity to resolve intragenic CNCs or delineate intragenic juxtaposition of sequences. *Ab initio* scFISH probe sequences containing divergent repetitive elements were used to detect small cancer genes (9,11) for *CCND1*, *CDKN2A*, *ERBB2*, *NOTCH1* and *TP53*. All scFISH probes hybridized to the correct chromosomal locations with high efficiency and specificity—17q21.1 (*ERBB2*), 9p21 (*CDKN2A*), 17p13.1 (*TP53*), 11q13 (*CCND1*) and 9q34.3 (*NOTCH1*). Representative hybridizations are shown in Figure 2.1. Inclusion of divergent repetitive elements in these probes did not produce any observed cross-hybridization with high stringency washing conditions. In addition, we hybridized *CDKN2A* Probe 1 to metaphase cells from a melanoma cell line (A-375). An aberrant hybridization pattern was observed on one chromosome 9p, with its hybridization signal telomeric relative to the normal chromosomal position (see Figure 2.1D). Inclusion of highly divergent repetitive elements significantly expands access to portions of the genome that were previously avoided by repeat masking sc sequences. A total of 95.6% (915 279) of these FISH probes overlap at least one divergent repetitive element. *Ab initio* scFISH probes consisting exclusively of sc sequences now comprise a minority of (3.7%; 35 658) of the genomic intervals.

2.3.3 *Ab initio* aCGH

Inclusion of divergent repetitive elements in genomic probes expands the regions accessible for probe development and the potential genomic resolution of aCGH. We have previously suggested that probe placement and, in particular, oligonucleotide targets in close proximity to conserved repetitive sequences may increase the variability in signal intensities observed in microarray hybridization (23). To test this idea, we selected oligonucleotide probes located greater than 300 nt away (the target size of the random primed DNA sample) from a conserved repetitive element. Hybridization results from our custom array design were directly compared with those obtained from the Agilent 44K

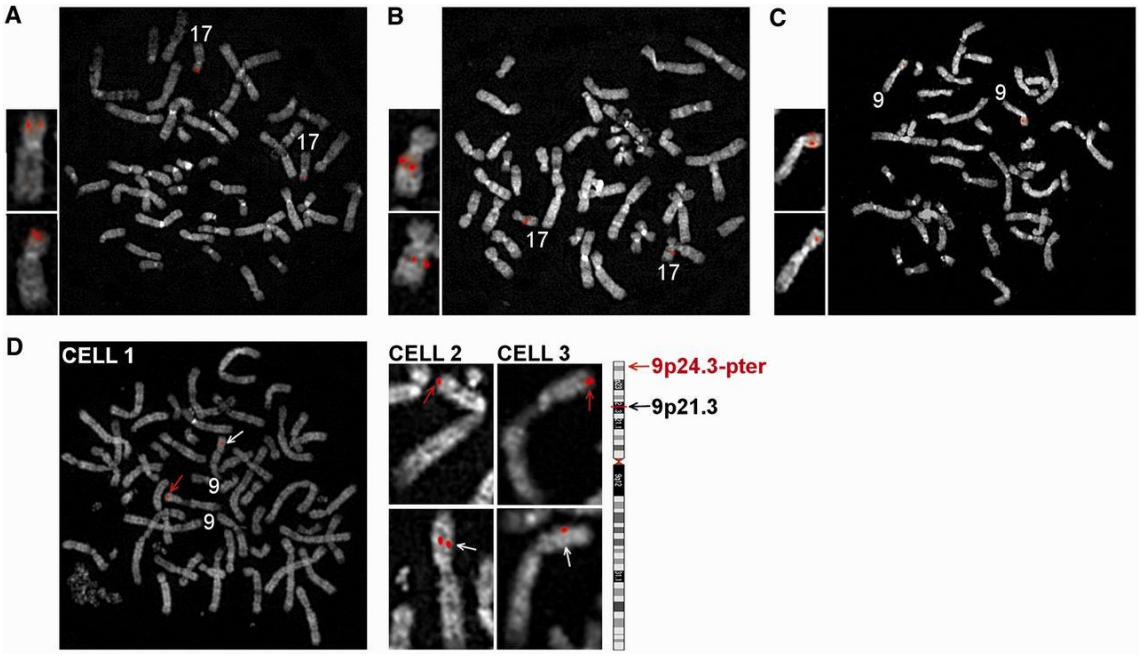


Figure 2.1: FISH validated sc probes. Normal metaphase chromosomes from three cells hybridized with probes targeting TP53 on chromosome 17p13.1 (A), *ERBB2* on 17q21.1 (B) and *CDKN2A* Probe1 on 9p21.3 (C) are shown. Hybridized chromosomes of each cell are enlarged and presented to the left of their respective metaphases. In panel (D), chromosome 9s from three different cells from melanoma A-375 cell line, hybridized to *CDKN2A* Probe 1, are presented. A complete metaphase is shown on the left and an ideogram of chromosome 9 on the right. One chromosome 9 in each cell shows hybridization as expected at 9p21.3 (white arrows), whereas the other homologue shows hybridization at the end of the chromosome (9p24.3-pter, red arrow). The aberrant location of the hybridization is likely due to a paracentric inversion between 9p21.3 and 9p24.3. Chromosomes are counterstained with DAPI. Note: The aberrant hybridization pattern is consistently seen on the chromosome 9 with the pale staining heterochromatin polymorphism in the q arm.

platform using the same labeled HapMap trio samples (i.e. healthy individuals). Reproducibilities of the *ab initio* and Agilent microarrays were compared from the CV of hybridization intensities of replicate oligonucleotide probes. The custom oligonucleotide array of genomic targets with this content exhibited lower variability in hybridization kinetics and increased consistency of signal intensities in aCGH. The median CVs of all probes in both replicates were lower in the *ab initio* custom array for both log2 ratio (17.8%) and proband (green) signal intensities (24.1%; Table 2.1; Mann–Whitney rank sum test; $P < 0.001$). Red signal intensities were excluded because they represented two different individuals (two sectors of each mother/father), which was insufficient to reliably compute CVs. The subset of probes contributing to higher variability in signal intensities in the Agilent platform exhibited lower reproducibility as a function of genomic location. CVs of different subsets of Agilent probes (all probes, probes within 300 nt of a repeat, and probes greater than 300 nt of the closest repeat) were compared with CVs for the closest *ab initio* probes. The mean CVs of the intensity $\log_2$ ratios of the *ab initio* probes were on average 48.3% below that of the corresponding Agilent genomic targets, when the corresponding Agilent probe was located within 300 nt of a conserved repetitive element (paired Student's t-test; $P < 0.05$; Table 2.2). The mean CVs after background correction for all probes, regardless of genomic context were 34% lower for one HapMap family ($P < 0.001$); however, the difference was not significant for the other family. For paired sets of *ab initio* and Agilent probes, CVs were not significantly different for Agilent probes separated from adjacent repetitive sequences by >300 nt. In probe pairs where the Agilent oligonucleotide was within 300 nt of a repeat, the CVs of the *ab initio* proband signal were lower in all instances, consistent with our previous analyses (23). We interpret these findings as follows: probes within 300 nt of a repetitive element have the potential to hybridize to a random-primed DNA extension product that contains both a sc target sequence as well as adjacent repetitive elements. Conserved repetitive elements present in hybridized DNA sample are susceptible to cross-hybridization with repeats in non-target labelled and C_0t-1 DNA. Figure 2.2A illustrates an example of this for a pair of probe sequences in TP53. Labelled random-primed (or nick translated) extension products containing a Tigger5 conserved repeat element

Table 2.1: Comparison of CV of replicate probes by platform: Mann–Whitney rank sum test

CVs tested	Log_2 Ratio		Proband	
Platform[a]	AG	AI	AG	AI
YRI DNA Samples				
Median CV	49.37	37.34	4.25	2.26
Interquartile range	85.62	66.51	3.18	1.65
P-value	<0.001		<0.001	
CEU DNA samples				
Median CV	88.69	78.70	3.51	3.46
Interquartile range	155.89	140.67	2.97	2.72
P-value	<0.001		<0.001	

Median CVs of the log_2 ratio and proband signal intensities ('Proband') were compared for both *HapMap* family DNA samples (YRI/CEU). Bolded values indicate CVs that were significantly lower in the *ab initio* platform compared with the corresponding Agilent data. Interquartile range demonstrates the larger range of CVs in the Agilent platform.

[a]AG = Agilent; number of probes = 42 492; AI = *Ab initio*; number of probes = 41 898; YRI = Yoruban HapMap trio; CEU = Caucasian HapMap trio.

Table 2.2: Comparison of CV of replicate probes by platform: Paired *t*-tests

CVs tested	Log_2 Ratio		
Platform[a]	AG	AI	*P*-value*
YRI DNA Samples			
All probes	328	216	**0.0019**
AG probes <300 nt	366	218	**0.0046**
AG probes >300 nt	260	213	0.0855
CEU DNA samples			
All probes	869	901	0.4655
AG probes <300 nt	1025	449	**0.0348**
AG probes >300 nt	594	1695	0.0975

Paired *t*-tests were performed for log_2 ratio CVs for all probe pairs, probe pairs where the Agilent oligonucleotide was within 300 nt of a repetitive element (AG probes <300 nt), and for probe pairs where the Agilent oligonucleotide probe was at least 300 nt from an adjacent repetitive element (AG probes >300 nt).

[a]AG = Agilent; number of probes = 42 492; AI = *Ab Initio*; number of probes = 41 898; YRI = Yoruban *HapMap* trio; CEU = Caucasian *HapMap* trio.

*Bolded values indicate $P < 0.05$.

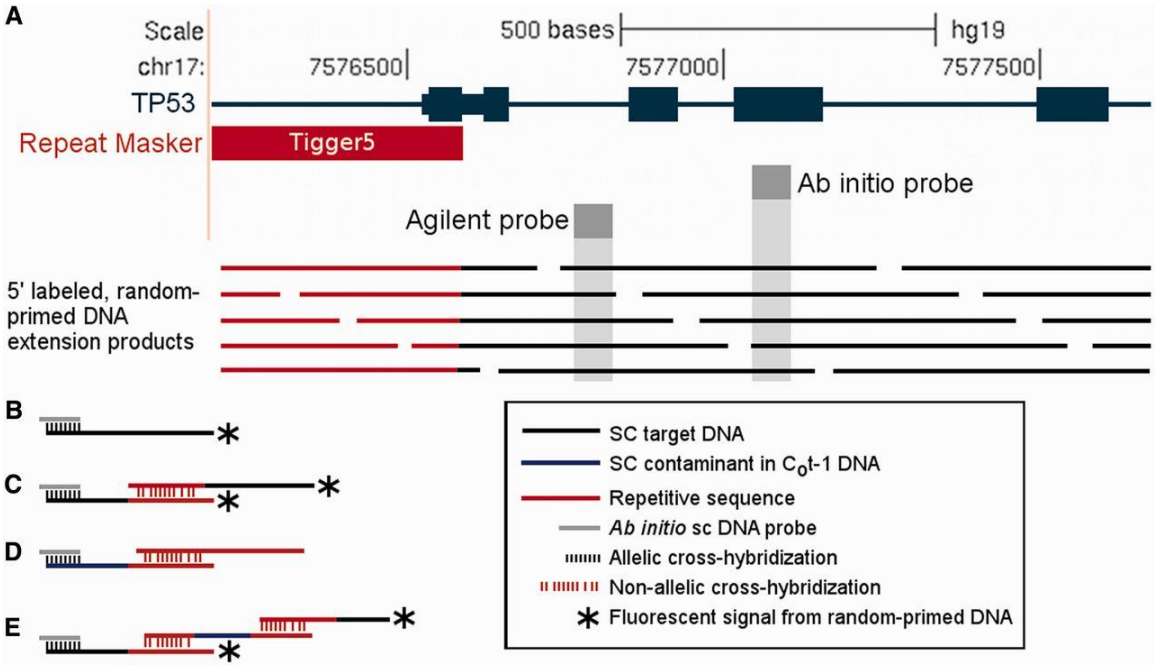


Figure 2.2: The effect of genomic context on hybridization signal intensity variability. (A) This panel demonstrates how the subtle differences in genomic location of *ab initio* and Agilent probes (dark grey; light grey vertical bars show target on extension products) may explain the higher CV in the Agilent platform. Simulated 5′ labelled, random-primed DNA extension products (of 300 nt) are windowed along the TP53 gene with the locations of a pair of Agilent and *ab initio* sc oligonucleotide probes. Increasing the distance between microarray probe sequences (in grey) and repetitive elements (in red) reduces the likelihood of hybridization to a labelled DNA product containing both the unique target (in black) and repetitive sequence. Extension products containing an adjacent Tigger5 repetitive element would be expected to hybridize to the Agilent probe located 179 nt away, but not to the *ab initio* sc probe situated 462 nt from the repeat, even though both are sc (black) probes. The average CV of this Agilent probe was 146, compared with the *ab initio* probe, which had a CV of 32. (B) Accurate hybridization signal intensity is achieved with sc target labelled DNA (black), exclusively hybridizing to probe sequence. Panels C and E depict how the presence of repetitive sequences in labelled target DNA can lead to higher than expected signal intensities. (C) Signals can be amplified by repeats (red) in close proximity to sc sequences (black), leading to non-allelic cross-hybridizations between repetitive elements adjacent to the labelled target DNA and other regions of the genome. (D) Unlabelled C_ot-1 DNA is known to be contaminated with sc sequences (blue), which can serve as microarray probe targets. These contaminants in C_ot-1 can suppress hybridization to desired target sequences by blocking the target labelled DNA from hybridizing to the probe sequences, reducing the overall fluorescent signal. (E) The major repetitive fraction in C_ot-1 DNA will hybridize to labelled,

random-primed DNA containing repetitive sequence (e.g. Tigger5 in this instance). This can result in an undesirable increase in signal intensity through bridging hybridization of labelled DNA target to other non-allelic repetitive sequences. This can be mediated by cross-hybridization to repetitive sequences in C_0t-1 DNA, which is usually added in stochiometric excess of the labelled sequence in microarray studies.

(11.5% divergent from the TcMar-Tigger consensus) cross-hybridized to the published Agilent probe sequence 179 nt away (CV = 146), but did not hybridize to the *ab initio* probe situated 462 nt from this repeat element (CV = 32). Calibration of the lengths of the labelled genomic DNA used in aCGH has been demonstrated to significantly improve microarray performance (38). Indeed, the observed CVs of these specific probes confirm the expected results.

2.3.4 Probe parameters affecting CVs

As the increased variability in microarray signal intensities can be attributed to proximate repetitive elements, we performed analysis of variance (ANOVA) and principal component analyses (PCA) to examine the characteristics of the oligonucleotide sequences that contribute to this source of noise. Genomic features (GC content, probe length, distance of nearest neighbouring repeat element and divergence) were determined for each set of paired probes and assessed by ANOVA for association with signal intensities and CVs. Repeat distance was associated with the $\log_2$ ratio CVs in both Agilent arrays ($P < 0.05$ and $P < 0.001$). In the second Agilent hybridization, repeat divergence ($P < 0.05$) was also associated with CVs. However, the CVs of $\log_2$ ratios were associated with neither repeat distance nor repeat divergence in either *ab initio* array ($P > 0.05$). PCA of data from both microarray platforms were consistent among replicate hybridizations for each platform; however, differences between Agilent and *ab initio* arrays were evident for two PCA eigenvectors (Table 2.3). The third component of the *ab initio* data was comprised of CV alone, with no significant interaction with the other factors, as expected from ANOVA. Differences in the Agilent data show that both the distance between probe and adjacent repetitive sequences, specifically within 300 nt, and the degree to which the repeat sequence is conserved, are not independent of the CVs of the probe signal intensities.

We then analysed the CVs of signal intensities from both the Agilent and Affymetrix (Santa Clara, US) microarrays for the same HapMap samples analysed previously. The CVs of four data sets (two Agilent, two Affymetrix) were compared within the same

Table 2.3: Principal components analysis of genomic and probe parameters with CV in *HapMap* pedigrees

Platform characteristics	YRI trio			CEU trio		
Eigenvectors	1	2	3	1	2	3
AB INITIO						
CV intensity	−0.0087	0.0734	**0.9970**	0.0038	−0.0723	**0.9959**
GC content	**0.4895**	**−0.4466**	0.0201	**0.4894**	**−0.4441**	−0.0742
Probe length	**−0.2562**	**0.6979**	−0.0689	**−0.2562**	**0.7002**	0.0195
Repeat distance	**0.6546**	**0.2000**	−0.0061	**0.6547**	**0.1987**	0.0268
Repeat divergence	**−0.5159**	**−0.5178**	0.0288	**−0.5159**	**−0.5174**	−0.0388
% Variance explained	26.9705	21.6464	19.9922	26.9700	21.6461	20.0012
AGILENT						
CV intensity	−0.0397	**−0.5311**	**0.8035**	0.0065	**0.5145**	**0.8554**
GC content	**−0.6950**	0.0436	0.0250	**−0.6957**	0.0444	−0.0118
Probe length	**0.6976**	−0.0016	−0.0149	**0.6979**	−0.0088	−0.0066
Repeat distance	**−0.1643**	**−0.2629**	**−0.4577**	**−0.1647**	**−0.3947**	**0.1772**
Repeat divergence	−0.0409	**0.8043**	**0.3796**	−0.0412	**0.7599**	**−0.4865**
% Variance explained	36.8101	20.1829	19.9547	36.7845	20.1786	19.9373

Principal component analysis was carried out to assess the relationship between probe CVs, GC content, probe length, distance of the closest repeat and its divergence from the consensus family sequence. In the *ab initio* probe set, the CV eigenvalues showed little or no interaction with other probe properties (compare eigenvectors 1 or 2 versus 3). In contrast, the corresponding eigenvalues were related to distance from and divergence of adjacent repetitive sequences in data from the Agilent platform. Bolded numbers indicate the parameter has a positive or negative effect of at least 15% overall.

hybridization. This eliminated the possibility that the observed results were derived from subtle differences in experimental conditions or labelling of genomic DNA. Probe CVs were calculated for the Agilent 44K array and the publically available Affymetrix Genome-Wide Human SNP Array 6.0 Sample Data Set (http://www.affymetrix.com/support/technical/sample_data/genomewide_snp6_data.affx) . The median CVs were compared using a Mann–Whitney Ranked Sum Test. Probes were categorized based on the repeat proximity (either within or beyond 300 nt) and level of divergence (±20% relative to the consensus repeat) of the repetitive element adjacent to a probe (Table 2.4). For both commercial data sources, probes within 300 nt of a repetitive element exhibit significantly higher CVs ($P < 0.001$), though the Affymetrix probes had lower CVs overall than those on the Agilent array. In the Affymetrix data, the level of repeat divergence contributes to probe signal intensity variability to a greater extent than the probe proximity to adjacent repetitive elements. In particular, the combination of low divergence and close proximity produces the highest probe CVs in both commercial microarray platforms. As expected, repeat divergence did not contribute to probe signal intensity CVs for probes at least 300 nt away from adjacent repetitive elements.

2.3.5 Targeted chromosome 15q11.2q13 aCGH detects AS deletion

Lower variability in signal intensities is desirable in aCGH to achieve more consistent calling of CNCs and accurate determination of copy number using fewer probes. To assess the reliability of *ab initio* probes in CNC detection, we performed aCGH on a sample with a documented chromosome deletion using custom-synthesized, targeted microarrays. A set of 12K oligonucleotide microarrays were produced with probes concentrated in the chromosome 15q11.2q13 region and genome-wide representation at other chromosomal locations. The arrays were simultaneously hybridized to random-primed DNA from a lymphoblastoid cell line derived from a patient with AS carrying a defined deletion of 5.01 Mb (32).

The same labelled sample was used in eight hybridizations: four containing identical

Table 2.4: Analysis of variation of CVs in Agilent and Affymetrix aCGH probe subsets

Repeat distance	Repeat divergence	No. probes	Median	P-value[a]	Repeat distance	Repeat divergence	No. probes	Median	P-value[a]
A. Affymetrix-GM07019					B. Affymetrix-GM19145				
<300	<20	576 831	0.0246	<0.001	<300	<20	576 363	0.0236	<0.001
>300	>20	276 461	0.0235		>300	>20	276 705	0.0223	
All	<20	840 370	0.0244	<0.001	All	<20	840 369	0.0235	<0.001
	>20	880 374	0.0237			>20	880 375	0.0224	
<300	All	1 180 744	0.0242	<0.001	<300	All	1 180 033	0.023	<0.001
>300		540 000	0.0238		>300		540 711	0.0227	
<300	<20	576 831	0.0246	<0.001	<300	<20	576 363	0.0236	<0.001
	>20	603 913	0.0238			>20	603 670	0.0224	
>300	<20	263 539	0.024	<0.001	>300	<20	264 006	0.0232	<0.001
	>20	276 461	0.0235			>20	276 705	0.0223	
C. Agilent-GM07019					D. Agilent-GM19145				
<300	<20	14 052	0.921	<0.001	<300	<20	14 052	0.503	<0.001
>300	>20	6 940	0.861		>300	>20	6 940	0.433	
All	<20	21 866	0.897	0.011	All	<20	21 866	0.484	<0.001
	>20	18 644	0.875			>20	18 644	0.449	
<300	All	25 756	0.901	<0.001	<300	All	25 756	0.482	<0.001
>300		14 754	0.862		>300		14 754	0.443	
<300	<20	14 052	0.921	0.007	<300	<20	14 052	0.503	<0.001
	>20	11 704	0.884			>20	11 704	0.457	
>300	<20	7 814	0.863	0.555	>300	<20	7 814	0.452	0.301
	>20	6 940	0.861			>20	6 940	0.433	

Comparison of probe CVs of Agilent and Affymetrix platforms based on proximity to and divergence level of neighbouring repetitive elements. Probe CVs were calculated for Affymetrix (panels A and B) and Agilent (panels C and D) data from hybridizations with the *HapMap* proband samples (panels A and C: GM07019, panels B and D: GM19145) used in this study. Median CVs of different groups of probes within each platform were compared using the Mann–Whitney rank sum test. Probe subsets were selected based on the distance to the closest repetitive element in nt

(either less or greater than 300 nt) and the divergence of the repetitive element from a consensus family sequence (less than or greater than 20%). [a]Mann–Whitney rank sum test.

probe content from the *ab initio* custom array and four containing published probe sequences from the Agilent 44K array. One of the arrays containing the Agilent probe design failed quality control owing to uneven oligonucleotide synthesis and was excluded from further analyses. The *ab initio* platform contained 125 probes and the Agilent platform contained 84 within the common AS deletion-breakpoint interval. Each probe was replicated on the array three times. The *ab initio* probes were distributed on average 52.54 kb apart throughout the CNC region, with a median distance between oligonucleotides of 18.01 kb. The Agilent probes were slightly more dispersed, with an average distance between oligonucleotides of 77.83 kb and a median distance of 52.11 kb. CNC detection was done by Rank Segmentation (39,40) and required at least five probes in a segment to assign a CNC.

Results from five of seven genomic microarrays called the AS deletion accurately: all four replicates of the *ab initio* probe set and one replicate containing Agilent probe sequences. Figure 2.3 indicates representative examples of primary signal intensities for the oligonucleotide probes spanning the deletion interval and flanking sequences for the *ab initio* and Agilent-based microarrays. The primary signal intensities of the *ab initio* probes displayed lower overall variability in the distributions of intensities in this genomic region. *Ab initio* probes within the deletion interval were then matched, based solely on genomic proximity, to the 76 Agilent probe sequences (excluding the breakpoint regions). Considering the matched probes alone, all four data sets from the *ab initio* platform were able to call the CNC, which was detectable on only a single array with Agilent probe content.

We tested the limits of sensitivity of the *ab initio* and Agilent microarrays to call CNCs by reducing the probe densities in this region by selecting one of two alternating probes (n = 37). All four replicates of the *ab initio* array still detected the AS deletion. Interestingly, one of the Agilent replicate arrays called the deletion, but it was a different microarray from the one indicated in the previous analysis that involved twice as many probes. The resolution and consistency of both array platforms of calling deletions was

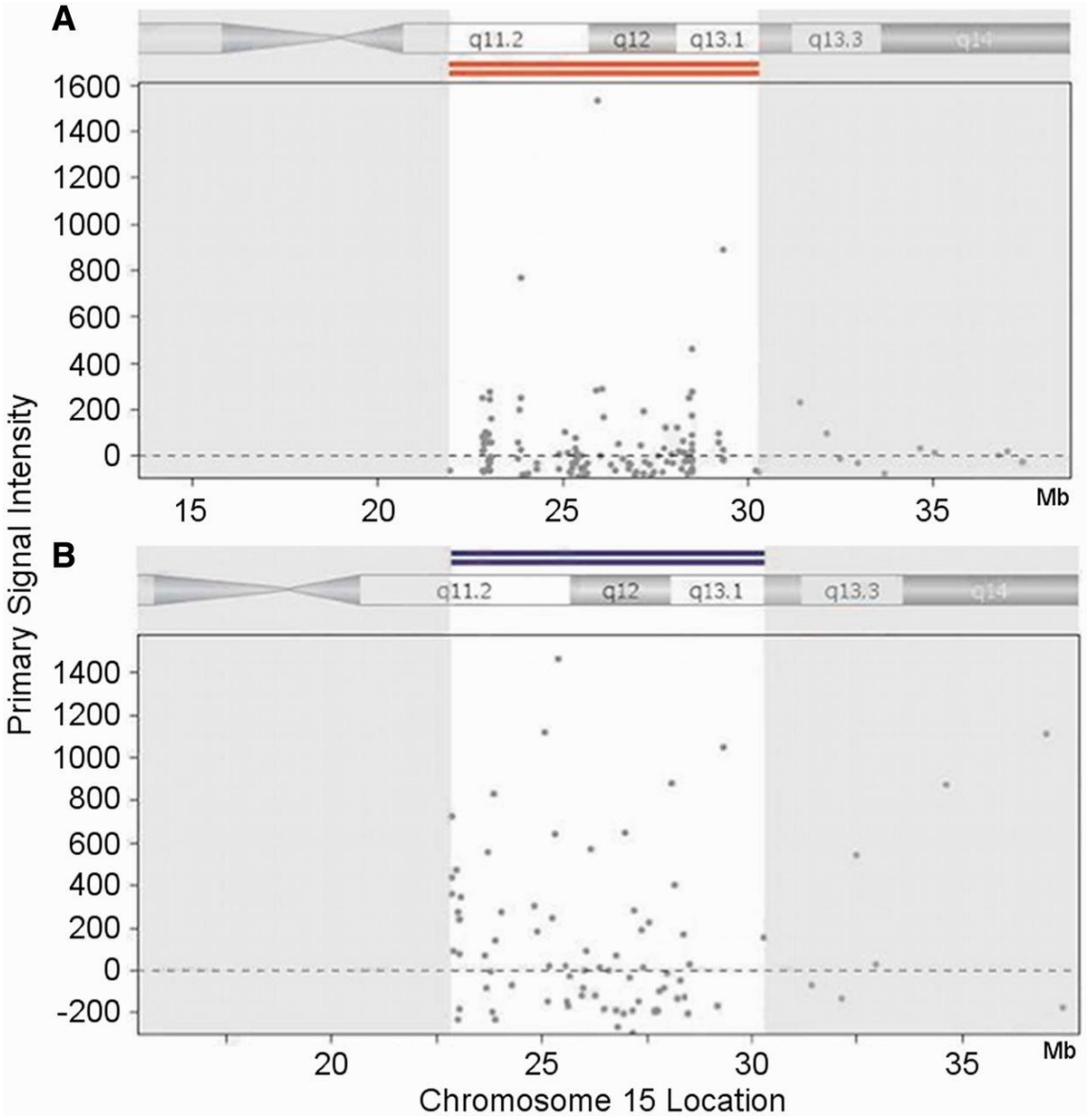


Figure 2.3: Primary hybridization signal intensity data from *ab initio* and Agilent probe sequences covering Angelman syndrome (AS) chromosome deletion region (chromosome 15q11.2q13.1). Primary signal intensity data are displayed from Nexus Biodiscovery software for one replicate each of the (**A**) *ab initio* and (**B**) Agilent probe sequences. Red and blue bars indicate copy number loss or gain, respectively. Details on the CNCs displayed were outputted as follows: (A) Deletion genome coordinate range called the following: 21 937 154–30 319 444, length: 8 362 290 nt, probe count: 123, probe signal intensity mean: 53.84, probe signal intensity median: −13.00. (B) Miscalled duplication coordinate range: 22 866 888–30 322 138, length: 7 455 250 nt, probe count: 73, probe signal intensity mean: 140.16, probe signal intensity median: 13.7. This figure demonstrates the greater variation in Agilent probe sequence signal intensities

compared with those from the *ab initio* array. The average standard deviation of the probe signal intensities between replicates in the *ab initio* CNC region (chr15: 21 937 154–30 319 444) is 138.08, whereas it is 238.04 (72% higher) for the Agilent probe sequences in the CNC region (chr15: 22 866 888–30 322 138).

unreliable when only 12 probes were scored (every third probe from the set of 37). A defined region within the deletion (*ab initio*—chr15:22 815 291–24 061 148 (hg19); Agilent—chr15:22 784 523–23 930 870) that spans the Angelman breakpoint 2 (BP2) (32) was called as a gain in one *ab initio* data set and all three Agilent data sets. By contrast, the region of the deletion distal to BP2 (*ab initio*—chr15:25 207 252–30 319 444; Agilent—chr15:25 143 144–30 322 138) is inferred as a copy number loss in all seven data sets. The mean CVs of all probes within BP2 that inconsistently called CNCs in both platforms were 34.87% (*ab initio*) and 17.75% (Agilent) higher than the other probes in the deletion interval. This is likely due to higher noise in the observed signal intensities. This may be related to interference of segmental duplicons in the hybridization, which are known to distort aCGH results (32). Segmental duplicons span 47% (*ab initio*) and 53% (Agilent) of the BP2 region. This is considerably higher compared with the genomic interval that was consistently called as a deletion and contains a smaller proportion of segmentally duplicated sequences (14%).

2.4 Discussion

Sequences of synthetic DNA probes used in genomic hybridization have been traditionally derived from unique sequences, or include repetitive elements that are sequestered during hybridization (4–9). The contextual effects of the genomic proximity of these sequences to repetitive elements have generally not been accounted for in assessing probe performance. Judicious selection of probes distant from adjacent conserved repetitive sequences can improve reproducibility of human genomic hybridization. Furthermore, probes incorporating divergent repetitive sequences do not adversely affect sc probe specificity. Under more stringent hybridization conditions, cross-hybridization catalysed by repetitive sequences is preventable. The inclusion of divergent repetitive elements expands genome-wide probe coverage, the outcome of which are increased lengths of scFISH probes in those regions and higher resolution in delineating novel genomic rearrangements by hybridization-based methods (such as genomic microarrays, multiplex ligation-dependent probe amplification (MLPA), PCR and others).

There are other established methods for producing short FISH probes. Software has been used to design smaller (10–100 kb) FISH probes (41), similar to our own scFISH products (9,11). Pools of labelled oligonucleotides have been used to visualize regions as small as 6.7 kb (42); however, the efficiency of detection with these pools is currently insufficient to be recommended for clinical use. Furthermore, both of these methods still require repeat-free regions for probe design. The *ab initio* scFISH probes presented here can reliably target small genes that are known to be commonly rearranged in cancer. By contrast, conventional, recombinant FISH probes extend well beyond the boundaries of these genes and often include neighbouring genes. Repeat-masked probes that lack divergent repetitive elements (9) within these genes are often too short to perform scFISH.

The coverage and level of specificity achieved by *ab initio* scFISH can confirm intragenic rearrangements or define small chromosomal aberrations detected by aCGH. Abnormalities that can be detected by these probes include small deletions (genes or exons), gene amplification, translocations and inversions involving the probe's genomic location. For example, *CCND1* at 11q13.3 is only 13.37 kb. A common translocation t(11;14)(q13,q32), which over-expresses this gene has been found in 20% of multiple myeloma cases (43,44) and 94% of mantle cell lymphoma patients (45). We have created two probes (<4 kb) targeting exons 3 (probe 1) and 5 (probe 2) of *CCND1*. In patients carrying this translocation, these probes will hybridize to the derivative chromosome 14. Commercial and cloned probes in this genomic region are considerably longer and would not detect rearrangements confined to this gene.

Despite the widespread application of aCGH for genome-wide copy number determination (46,47), the inter- and intra-platform reproducibility of both expression and copy number microarray data may be less than satisfactory (19–21,23,37,48–51). These previous studies have generally assumed that discrepancies resulted from stochastic noise in signal intensity measurements and have been attributed to algorithms used to call CNC analyses. Higher CVs of signal intensities have also been linked to probe length and composition, cross-, self- and perfect match hybridization free energies, melting temperatures, position within a target sequence, sequence complexity, potential

secondary structure and sequence information content (52). Nonetheless, these parameters have been described as insufficient for optimizing probe performance (53).

Our results suggest that the variability in aCGH studies does not originate solely from stochastic effects, but rather a systematic error introduced during probe design. We demonstrated that the genomic location of the probe relative to neighbouring conserved repetitive elements and the level of sequence divergence of the nearest repeat can account for 40% of the variance observed in the Agilent genomic microarray data. We were however not able to explain all of the variance in the signal intensity data. It has been recognized that self–self hybridization in solution may be responsible for variability by sequestering some of the labelled hybridizable sequences (37). We propose that formation of these duplexes is frequently catalysed by repeats in labelled DNA containing the sc target sequence. Repetitive sequences throughout the genome are of sufficiently high concentration for such events to be commonplace during hybridization. Other factors such as variation in the quantity of probe on the array and hybridization kinetics, could also account for the unexplained variance.

When expanding the oligonucleotide set with additional probes, it is important to consider the probe characteristics that are the most crucial to minimizing CVs. Probes within 300 nt of adjacent repetitive elements with <20% divergence from eponymic repeat family members have the poorest performance, with CVs on average 8.41% higher than those with greater separation from these elements. The variation of signal intensities is likely due to cross-hybridization to repetitive sequences present in the labelled target DNA as well as C_ot-1 DNA contaminated with the sc sequences detected by the probe (Figure 2.2). Figure 2.2B illustrates the expected hybridization pattern, when labelled sc target DNA hybridizes to the probe resulting in an accurate signal intensity. Figure 2.2C demonstrates the cross-hybridization that can occur when the microarray probe is located within 300 nt of a conserved repeat element (e.g. Agilent probe in panel 2A), resulting in an unexpected, higher signal intensity. In Figure 2.2D, reduced signal intensity can result from cross-hybridization of unlabelled sc sequences present in C_ot-1 DNA, which could block the labelled target sequences from hybridizing to the array. The signal can also be amplified when labelled DNA is bridged through non-allelic elements in unlabelled C_ot-1

DNA (Figure 2.2E). Increasing the genomic distance between sc target sequences used as probes on the microarray and conserved repetitive elements in the genome diminishes the likelihood of cross-hybridization to labelled target DNA products containing non-allelic repetitive sequences. We demonstrated that signal intensity CVs can be reduced by avoiding probe placement within 300 nt of a repeat element.

The reliability of calling CNCs is improved with probes that exhibit lower variation in primary signal intensities. Such probe sequences are of sufficient density in the genome that the same rearrangements analysed with commercial microarrays can be detected with greater reliability. The Agilent 44K array did not have sufficient probe density or low enough CVs to reliably detect a common chromosome 15q11.2q13 deletion, whereas a CNC based on 36 *ab initio*-designed probes was consistently called. Lowering CVs in microarray hybridization studies actually decreases the number of probes required for accurate CNC detection without significant loss in genomic resolution while still detecting small chromosome rearrangements. An implication of reliable detection of chromosome rearrangements with fewer probes is that it would facilitate increased multiplexing, with additional sectors on the same microarray allowing analysis of larger numbers of patient samples per array.

To overcome limitations in sensitivity, manufacturers have increased probe densities to perform copy number analysis by averaging CNC calling using the results of multiple probes. These probe densities partially compensate for loss of dynamic range that results from normalization (which statistically reduces noise). We have taken a different approach by populating the array with probes that have inherently lower susceptibility to noise. Future studies will determine the minimum number of *ab initio* probes required to call well-characterized CNCs for various clinically relevant genomic imbalances. Optimizing CNV calling algorithms will nevertheless continue to be a crucial factor in aCGH microarray experiments. Reliable detection of genomic abnormalities is crucial in diagnostic microarray studies, especially in situations where each patient sample is analysed with a single hybridization array.

2.5 References

1. Smit AF. The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. 1996;6:743-748.

2. Rogan PK, Pan J, Weissman SM. L1 repeat elements in the human ε-(G)γ-globin gene intergenic region: sequence analysis and concerted evolution within this family. Mol. Biol. Evol. 1987;4:327-342.

3. Mottez E, Rogan PK, Manuelidis L. Conservation in the 5′ region of the long interspersed mouse L1 repeat: implications of comparative sequence analysis. Nucleic Acids Res. 1986;14:3119-3136.

4. Lichter P, Cremer T, Borden J, Manuelidis L, Ward DC. Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. Hum. Genet. 1988;80:224-234.

5. Craig JM, Kraus J, Cremer T. Removal of repetitive sequences from FISH probes using PCR-assisted affinity chromatography. Hum. Genet. 1997;100:472-476.

6. Pinkel D, Landegent J, Collins C, Fuscoe J, Segraves R, Lucas J, Gray J. Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. Proc. Natl Acad. Sci. USA 1988;85:9138-9142.

7. Sealey PG, Whittaker PA, Southern EM. Removal of repeated sequences from hybridisation probes. Nucleic Acids Res. 1985;13:1905-1922.

8. Gray JW, Pinkel D. Molecular cytogenetics in human cancer diagnosis. Cancer 1992;69:1536-1542.

9. Rogan PK, Cazcarro PM, Knoll JH. Sequence-based design of single-copy genomic DNA probes for fluorescence in situ hybridization. Genome Res. 2001;11:1086-1094.

10. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat. Rev. Cancer 2004;4:177-183.

11. Knoll JH, Rogan PK. Sequence-based, in situ detection of chromosomal abnormalities at high resolution. Am. J. Med. Genet. A 2003;121A:245-257.

12. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat. Genet. 1998;20:207-211.

13. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat. Genet. 1999;23:41-46.

14. Barrett MT, Scheffer A, Ben-Dor A, Sampas N, Lipson D, Kincaid R, Tsang P, Curry B, Baird K, Meltzer PS, et al. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. Proc. Natl Acad. Sci. USA 2004;101:17765-17770.

15. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nat. Genet. 2005;37:S11-S17.

16. Shinawi M, Cheung SW. The array CGH and its clinical applications. Drug Discov. Today 2008;13:760-770.

17. Manning,M., Hudgins,L., and Professional Practice and Guidelines Committee. Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities. Genet. Med. 2010;12:742-745.

18. Duncan,A., Chodirker,B., and CCMG Clinical Practice, Cytogenetics and Prenatal Diagnosis Committees. Use of array genomic hybridization technology in constitutional genetic diagnosis in Canada. 2010. CCMG epub

19. Haraksingh RR, Abyzov A, Gerstein M, Urban AE, Snyder M. Genome-wide mapping of copy number variation in humans: comparative analysis of high resolution array platforms. PLoS One 2011;6:e27859.

20. Hester SD, Reid L, Nowak N, Jones WD, Parker JS, Knudtson K, Ward W, Tiesman J, Denslow ND. Comparison of comparative genomic hybridization technologies across microarray platforms. J. Biomol. Tech. 2009;20:135-151.

21. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat. Biotechnol. 2011;29:512-520.

22. Redon R, Fitzgerald T, Carter NP. Comparative genomic hybridization: DNA labeling, hybridization and detection. Methods Mol. Biol. 2009;529:267-278.

23. Newkirk HL, Knoll JH, Rogan PK. Distortion of quantitative genomic and expression hybridization by C_ot-1 DNA: Mitigation of this effect. Nucleic Acids Res. 2005;33:e191.

24. Rogan PK. 2010. US Patent 7,734,424, Dec. 30, 2005.

25. Rogan PK. 2012. US Patent 8,,209,129, Jun. 7, 2010.

26. Bolton ET, McCarthy BJ. A general method for the isolation of RNA complementary to DNA. Proc. Natl Acad. Sci. USA 1962;48:1390-1397.

27. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. 2000;132:365-386.

28. Knoll JH, Lichter P, Bakdounes K, Eltoum IE. In situ hybridization and detection using nonisotopic probes. Curr. Protoc. Mol. Biol. 2007. Chapter 14, Unit 14.7.

29. Giard DJ, Aaronson SA, Todaro GJ. In vitro cultivation of human tumors: Establishment of cell lines derived from a series of solid tumors. J. Natl Cancer Inst. 1973;51:1417-1423.

30. Chou HH. Shared probe design and existing microarray reanalysis using PICKY. BMC Bioinformatics 2010;11:196.

31. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005;33:D501-D504.

32. Khan WA, Knoll JH, Rogan PK. Context-based FISH localization of genomic rearrangements within chromosome 15q11.2q13 duplicons. Mol. Cytogenet. 2011;4:15.

33. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. Recent segmental duplications in the human genome. Science 2002;297:1003-1007.

34. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. 2001;11:1005-1017.

35. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl Acad. Sci. USA 2003;100:11484-11489.

36. Chiaromonte F, Yap VB, Miller W. Scoring pairwise genomic sequence alignments. Pac. Symp. Biocomput. 2002:115-126.

37. Lee Y, Ronemus M, Kendall J, Lakshmi B, Leotta A, Levy D, Esposito D, Grubor V, Ye K, Wigler M, et al. Reducing system noise in copy number data using principal components of self-self hybridizations. Proc. Natl Acad. Sci. USA 2012;109:E103-E110.

38. Craig JM, Vena N, Ramkissoon S, Idbaih A, Fouse SD, Ozek M, Sav A, Hill DA, Margraf LR, Eberhart CG, et al. DNA fragmentation simulation method (FSM) and fragment size matching improve aCGH performance of FFPE tissues. PLoS One 2012;7:e38881.

39. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004;5:557-572.

40. Darvishi K. Application of nexus copy number software for CNV detection and analysis. Curr. Protoc. Hum. Genet. 2010. 4.14.1–4.14.28.

41. Navin N, Grubor V, Hicks J, Leibu E, Thomas E, Troge J, Riggs M, Lundin P, Månér S, Sebat J, et al. PROBER: oligonucleotide FISH probe design software. Bioinformatics 2006;22:2437-2438.

42. Yamada NA, Rector LS, Tsang P, Carr E, Scheffer A, Sederberg MC, Aston ME, Ach RA, Tsalenko A, Sampas N, et al. Visualization of fine-scale genomic structure by oligonucleotide-based high-resolution FISH. Cytogenet. Genome Res. 2011;132:248-254.

43. Trakhtenbrot L, Hardan I, Koren-Michowitz M, Oren S, Yshoev G, Rechavi G, Nagler A, Amariglio N. Correlation between losses of IGH or its segments and deletions of 13q14 in t(11;14) (q13;q32) multiple myeloma. Genes Chromosomes Cancer 2010;49:17-27.

44. Kulkarni MS, Daggett JL, Bender TP, Kuehl WM, Bergsagel PL, Williams ME. Frequent inactivation of the cyclin-dependent kinase inhibitor p18 by homozygous deletion in multiple myeloma cell lines: ectopic p18 expression inhibits growth and induces apoptosis. Leukemia 2002;16:127-134.

45. Espinet B, Salaverria I, Beà Ś, Ruiz-Xivillé N, Balagué O, Salido M, Costa D, Carreras J, Rodríguez-Vicente AE, García JL, et al. Incidence and prognostic

impact of secondary cytogenetic aberrations in a series of 145 patients with mantle cell lymphoma. Genes Chromosomes Cancer 2010;49:439-451.

46. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, et al. Consensus statement: Chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am. J. Hum. Genet. 2010;86:749-764.

47. Ahn JW, Mann K, Walsh S, Shehab M, Hoang S, Docherty Z, Mohammed S, MacKie Ogilvie C. Validation and implementation of array comparative genomic hybridisation as a first line test in place of postnatal karyotyping for genome imbalance. Mol. Cytogenet. 2010;3:9.

48. Greshock J, Feng B, Nogueira C, Ivanova E, Perna I, Nathanson K, Protopopov A, Weber BL, Chin L. A comparison of DNA copy number profiling platforms. Cancer Res. 2007;67:10173-10180.

49. Curtis C, Lynch AG, Dunning MJ, Spiteri I, Marioni JC, Hadfield J, Chin S, Brenton JD, Tavaré S, Caldas C. The pitfalls of platform comparison: DNA copy number array technologies assessed. BMC Genomics 2009;10:588.

50. Cambon AC, Khalyfa A, Cooper NGF, Thompson CM. Analysis of probe level patterns in affymetrix microarray data. BMC Bioinformatics 2007;8:146.

51. Tan PK, Downey TJ, Spitznagel EL Jr., Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC. Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res. 2003;31:5676-5684.

52. Tulpan D. Recent patents and challenges on DNA microarray probe design technologies. Recent Pat. DNA Gene Seq. 2010;4:210-217.

53. Pozhitkov AE, Tautz D, Noble PA. Oligonucleotide microarrays: widely applied - poorly understood. Brief. Funct. Genomic. Proteomic. 2007;6:141-148.

54. Lin H, Ma X, Feng W, Samatova N. Coordinating computation and I/O in massively parallel sequence search. IEEE Trans. Parallel Distrib. Syst. 2010;22:529-543.

Chapter 3

3 Validation of predicted mRNA splicing mutations using high-throughput transcriptome data

The work presented in this chapter is reproduced (with permission, Appendix S1) from:

Viner, C., Dorman, S.N., Shirley, B.C., and Rogan, P.K. (2014) Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. [v2; ref status: Indexed, http://f1000r.es/378] *F1000Research* 3:8. DOI:10.12688/f1000research.3-8.v2

3.1 Introduction

DNA variant analysis of complete genome or exome data has typically relied on filtering of alleles according to population frequency and alterations in coding of amino acids. Numerous variants of unknown significance (VUS) in both coding and non-coding gene regions cannot be categorized with these approaches. To address these limitations, in silico methods that predict biological impact of individual sequence variants on protein coding and gene expression have been developed, which exhibit varying degrees of sensitivity and specificity (1). These approaches have generally not been capable of objective, efficient variant analysis on a genome-scale.

Splicing variants, in particular, are known to be a significant cause of human disease (2-5) and indeed have even been hypothesized to be the most frequent cause of hereditary disease (6). Computational identification of mRNA splicing mutations within DNA sequencing (DNA-Seq) data has been implemented to varying degrees of sensitivity, with most software only evaluating conservation solely at the intronic dinucleotides adjacent to the junction (i.e. (7)). Other approaches are capable of detecting significant mutations at other positions with constitutive, and in certain instances, cryptic, splice sites (5,8,9) which can result in aberrations in mRNA splicing. Presently, only information theory-based mRNA splicing mutation analysis has been implemented on a genome scale (10). Splicing mutations can abrogate recognition of natural, constitutive splice sites

(inactivating mutation), weaken their binding affinity (leaky mutation), or alter splicing regulatory protein binding sites that participate in exon definition. The abnormal molecular phenotypes of these mutations comprise: (a) complete exon skipping, (b) reduced efficiency of splicing, (c) failure to remove introns (also termed intron retention or intron inclusion), or (d) cryptic splice site activation, which may define abnormal exon boundaries in transcripts using non-constitutive, proximate sequences, extending or truncating the exon. Some mutations may result in combinations of these molecular phenotypes. Nevertheless, novel or strengthened cryptic sites can be activated independently of any direct effect on the corresponding natural splice site. The prevalence of these splicing events has been determined by ourselves and others (5,11-13). The diversity of possible molecular phenotypes makes such aberrant splicing challenging to corroborate at the scale required for complete genome (or exome) analyses. This has motivated the development of statistically robust algorithms and software to comprehensively validate the predicted outcomes of splicing mutation analysis.

Putative splicing variants require empirical confirmation based on expression studies from appropriate tissues carrying the mutation, compared with control samples lacking the mutation. In mutations identified from complete genome or exome sequences, corresponding transcriptome analysis based on RNA sequencing (RNA-Seq) is performed to corroborate variants predicted to alter splicing. Manually inspecting a large set of splicing variants of interest with reference to the experimental samples' RNA-Seq data in a program like the Integrative Genomics Viewer (IGV) (14), or simply performing database searches to find existing evidence would be time-consuming for large-scale analyses. Checking control samples would be required to ensure that the variant is not a result of alternative splicing, but is actually causally linked to the variant of interest. Manual inspection of the number of control samples required for statistical power to verify that each displays normal splicing would be laborious and does not easily lend itself to statistical analyses. This may lead to either missing contradictory evidence or to discarding a variant due to the perceived observation of statistically insignificant altered splicing within control samples. In addition, a list of putative splicing variants returned by variant prediction software can often be extremely large. The validation of such a

significant quantity of variants may not be feasible, for example, in certain types of cancer, in instances where the genomic mutational load is high and only manual annotation is performed. We have therefore developed Veridical, a software program that automatically searches all given experimental and control RNA-Seq data to validate DNA-derived splicing variants. When adequate expression data are available at the locus carrying the mutation, this approach reveals a comprehensive set of genes exhibiting mRNA splicing defects in complete genomes and exomes. Veridical and its associated software programs are available at: https://mutationforecaster.com.

3.2 Methods

The program Veridical was developed to allow high-throughput validation of predicted splicing mutations using RNA sequencing data. Veridical requires at least three files to operate: a DNA variant file containing putative mRNA splicing mutations, a file listing of corresponding transcriptome (RNA-Seq) BAM files, and a file annotating exome structure (Appendix S3.1-S3.3). A separate file listing RNA-Seq BAM files for control samples (i.e. normal tissue) can also be provided. Here, we demonstrate the capabilities of the software for mutations predicted in a set of breast tumours. Veridical compares RNA-Seq data from the same tumours with RNA-Seq data from control samples lacking the predicted mutation. However, in principle, potential splicing mutations for any disease state with available RNA-Seq data can be investigated. In each tumour, every variant is analyzed by checking the informative sequencing reads from the corresponding RNA-Seq experiment for non-constitutive splice isoforms, and comparing these results with the same type of data from all other tumour and normal samples that do not carry the variant in their exomes.

Veridical concomitantly evaluates control samples, providing for an unbiased assessment of splicing variants of potentially diverse phenotypic consequences. Note that control samples include all non-variant containing files (i.e. RNA-Seq files for those tumours without the variant of interest), as well any normal samples provided. Increasing the number of the set of control samples, while computationally more expensive, increases the statistical robustness of the results obtained.

For each variant, Veridical directly analyzes sequence reads aligned to the exons and introns that are predicted to be affected by the genomic variant. We elected to avoid indirect measures of exon skipping, such as loss of heterozygosity in the transcript, because of the possibility of confusion with other molecular etiologies (i.e. deletion or gene conversion), unrelated to the splicing mutations. The nearest natural site is found using the exome annotation file provided, based upon the directionality of the variant, as defined within Table 3.1. The genomic coordinates of the neighboring exon boundaries are then found and the program proceeds, iterating over all known transcript variants for the given gene. A diagram of this procedure is provided in Figure 3.1. The variant location, C, is specifically referring to the variant itself. JC refers to the variant-induced location of the predicted mRNA splice site, which is often proximate to, but distinct from the coordinate of the actual genomic mutation itself.

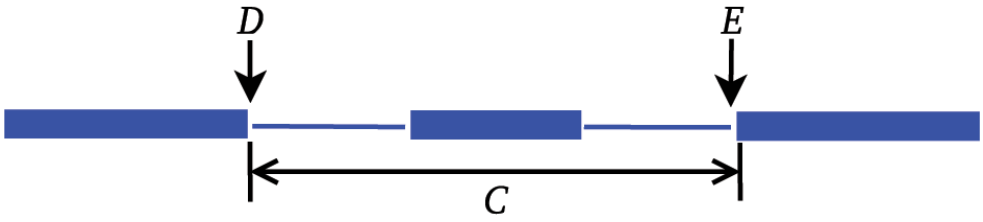
The program uses the BamTools API (15) to iterate over all of the reads within a given genomic region across experimental and control samples. Individual reads are then assessed for their corroborating value towards the analysis of the variant being processed, as outlined in the flowchart in Figure 3.2. Validating reads are based on whether they alter either the location of the splice junction (i.e. junction-spanning) or the abundance of the transcript, particularly in intronic regions (i.e. read-abundance). Junction-spanning reads contain DNA sequences from two adjacent exons or are reads that extend into the intron (Equation 1(e)). These reads directly show whether the intronic sequence is removed or retained by the spliceosome, respectively. Read-abundance validated reads are based upon sequences predicted to be found in the mutated transcript in comparison with sequences that are expected to be excised from the mature transcript in the absence of a mutation (Equation 1(f)). Both types of reads can be used to validate cryptic splicing, exon skipping, or intron inclusion. A read is said to corroborate cryptic splicing if and only if the variant under consideration is expected to activate cryptic splicing. Junction-spanning, cryptic splicing reads are those in which a read is exactly split from the cryptic splice site to the adjacent exon junction (Equation 1(a)). For read-abundance cryptic splicing, we define the concept of a read fraction, which is the ratio of the number of reads corroborating the cryptically spliced isoform and the number of reads that do not support the use of the cryptic splice site (i.e. non-cryptic corroborating) in the same

Table 3.1 Definitions used within Veridical to determine in which reads are checked. *A* and *B* represent natural site positions, defined in Figure 3.1(B).

$^{\alpha}$ – 5’ splice site $^{\beta}$ – 3’ splice site

Pertinent Splice Site			
A	*B*	Strand	Direction
Exonic	Donor$^{\alpha}$	+	➔
Exonic	Donor$^{\alpha}$	-	➔
Intronic	Acceptor$^{\beta}$	+	➔
Intronic	Acceptor$^{\beta}$	-	➔

A)



B)

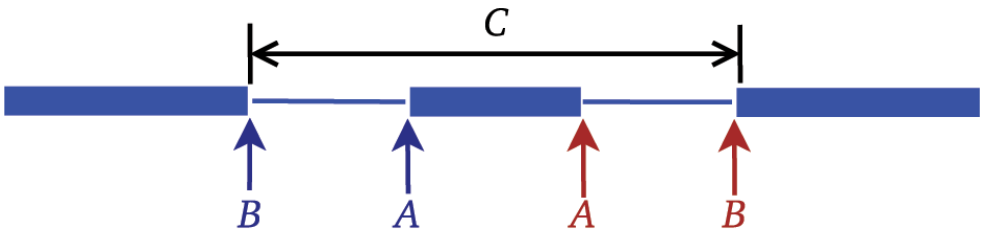


Figure 3.1 Diagram portraying the definitions used within Veridical to specify genic variant position and read coordinates. We employ the same conventions as IGV (14). Blue lines denote genes, wherein thick lines represent exons and thin lines represent introns. A) All reads overlapping or between *D* or *E* are extracted from the BAM files. We assume, for clarity of illustration, that the genome coordinate $D < E$. The variant, *C*, is contained somewhere within the middle exon or within one of its adjacent introns. B) Veridical searches for validating reads between *A* and *B*, the orientation of which is direction dependent. As indicated, the variant, *C*, is contained somewhere within the middle exon or within one of its adjacent introns. Depending upon the location of the variant, and the directionality (as described within Table 3.1), the interval boundaries may be delimited by either the blue or red set of labels.

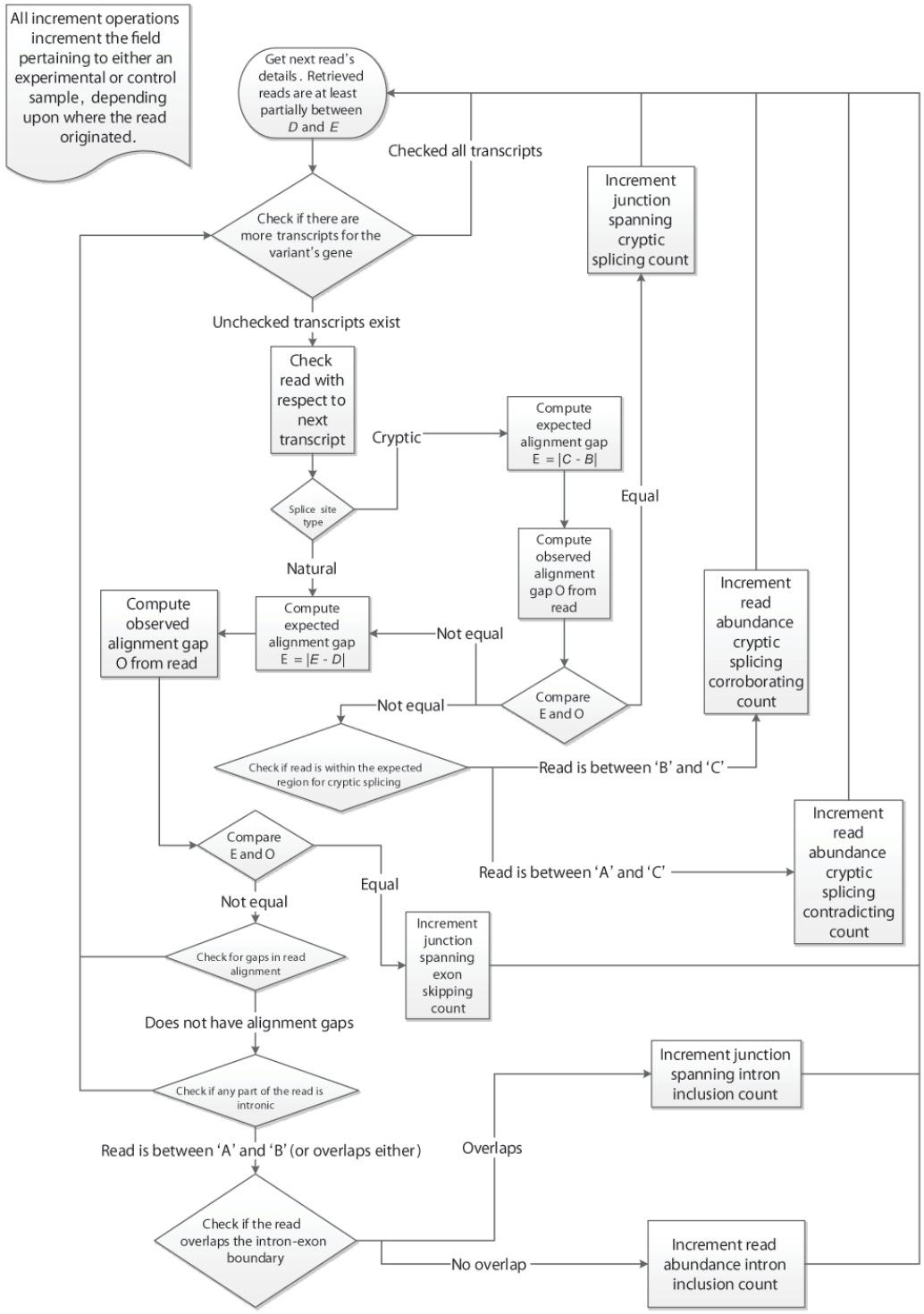


Figure 3.2 The algorithm employed by Veridical to validate variants. Refer to Table 3.1 for definitions concerning direction and Figure 3.1 for variable depictions. B is defined as follows: B (B site left (←) of A ⇒ B := D. B site right (→) of A ⇒ B := E.

genomic region of a sample. Cryptic corroborating reads are those which occur within the expected region where cryptic splicing occurs (i.e. spliced-in regions). This region is bounded by the variant splice site location and the adjacent (direction dependent) splice junction (Equation 1(a)). Non-cryptic corroborating reads, which we also termed "anti-cryptic" reads, are those that do not lie within this region, but would still be retained within the portion that would be excised, had cryptic splicing occurred (Equation 1(b)). To identify instances of exon skipping, Veridical only employs junction-spanning reads. A read is considered to corroborate exon skipping if the connecting read segments are split such that it connects two exon boundaries, skipping an exon in between (Equation 1(c)). A read is considered to corroborate intron inclusion when the read is continuous and either overlaps with the intron-exon boundary (and is then said to be junction-spanning) or if the read is within an intron (and is then said to be based upon read-abundance). We only consider an intron inclusion read to be junction spanning if it spans the relevant splice junction, A. Equation 1(d) formalizes this concept. We occasionally use the term "total intron inclusion" to denote that any such count of intron inclusion reads includes both those containing and not containing the mutation itself. Graphical examples of some of these validation events, with a defined variant location, are provided in Figure 3.3.

We proceed to formalize the above descriptions as follows. A given read is denoted by r, with start and end coordinates (r_s, r_e), if the read is continuous, or otherwise, with start and end coordinate pairs, (r_{s1}, r_{e1}) and (r_{s2}, r_{e2}) as diagrammed within Figure 3.3. Let ℓ be the length of the read. The set ζ denotes the totality of validating reads. The criterion for $r \in \zeta$ is detailed below. It is important to note that validating reads are necessary but not sufficient to validate a variant. Sufficiency is achieved only if the number of validating reads is statistically significant relative to those present in control samples. ζ itself is partitioned into three sets: ζ_c, ζ_e, and ζ_i for evidence of cryptic splicing, exon skipping, and intron inclusion, respectively. We allow partitions to be empty. Let J_C denote the adjacent splice junction, and let B denote the downstream natural site, as defined by Figure 3.3 and Table 3.1. Without loss of generality, we consider only the red (i.e.

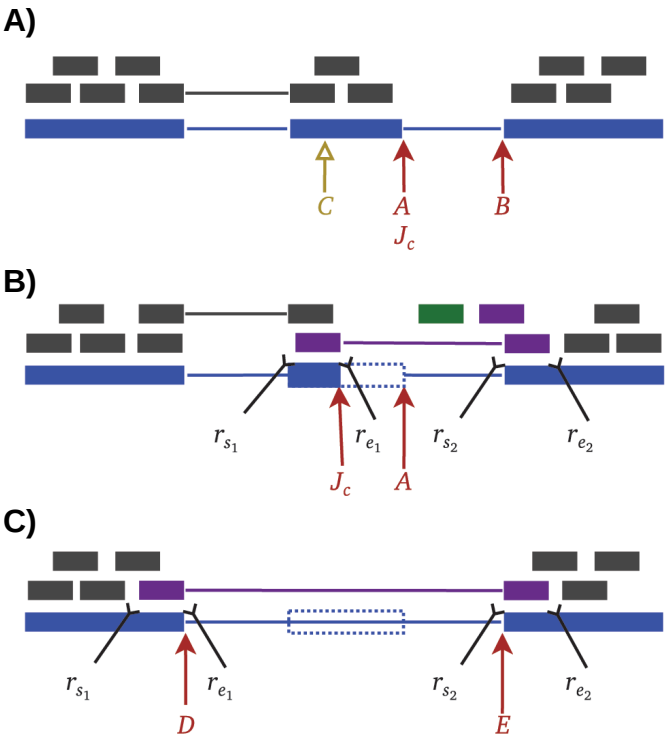


Figure 3.3 Illustrative examples of aberrant splicing detection. Grey lines denote reads, wherein thick lines denote a read mapping to genomic sequence and thin lines represent connecting segments of reads split across spliced-in regions (i.e. exons or included introns). Dotted blue rectangles denote portions of genes which are spliced out in a mutant transcript, but are otherwise present in a normal transcript. Mutant reads are purple if they are junction-spanning and green if they are read-abundance based. Start and end coordinates of reads with two portions are denoted by (r_{s1}, r_{e1}) and (r_{s2}, re2), while coordinates of those with only a single portion are denoted by (r_s, r_e). Refer to the caption of Figure 3.1 for additional graphical element descriptions. A) An example of a normally spliced transcript, assuming Veridical is validating a specific variant, *C*, shown in yellow. The adjacent intron-exon boundary, in this case, corresponds to both the adjacent splice junction, J_C, and the relevant natural site *A*. *B* is the downstream natural site. Veridical would not identify any aberrant splicing. B) An example of the variant causing the activation of a cryptic splice site. Additionally, there is intron inclusion present within the analysis region. Veridical would identify and report read counts for reads pertaining to the (junction-spanning, purple) cryptic splicing event and those pertaining to the observed (junction-spanning and read-abundance, green) intron inclusion. Since this pertains to a cryptic variant, the adjacent splice junction, J_C, is distinct from the relevant natural site A. C) An example of the variant causing the containing exon to be skipped. Veridical would report read counts for reads pertaining to the junction-spanning exon skipping event. These discontinuous reads are those, that like the one shown, span the variant containing exon.

direction is right) set of labels within Figure 3.1(B), as further typified by Figure 3.3. Then the (splice consequence) partitions of ζ are given by:

$$r \in \zeta_c \Leftrightarrow variant\ is\ cryptic\ \wedge (r_{S_2} - r_{e_1} = B - J_C \vee (r_S > J_C\ \wedge r_e < A)) \quad (1a)$$

$$r \notin \zeta_c \wedge variant\ is\ cryptic\ \wedge \neg\left(r_{S_2} - r_{e_1} = B - J_C\right) \Rightarrow r \in anti - cryptic \quad (1b)$$

$$r \in \zeta_e \Leftrightarrow (r_{e_1} = D \wedge r_{S_2} = E) \quad (1c)$$

$$r \in \zeta_i \Leftrightarrow (A \in [r_S, r_e]) \vee ((A \notin [r_S, r_e]) \wedge r_S > A - \ell \wedge r_e < B \wedge \neg(A \in [r_S, r_e])) \quad (1d)$$

We separately partition ζ by its evidence type, the set of junction-spanning reads, δ and read-abundance reads, α:

$$r \in \delta \Leftrightarrow (A \in [r_S, r_e]) \vee (r \in \zeta_C \wedge r_{S_2} - r_{e_1} = B - J_C) \quad (1e)$$

$$r \in \propto \Leftrightarrow r\ \notin\ \delta \quad (1f)$$

Once all validating reads are tallied for both the experimental and control samples, a p-v alue is computed. This is determined by computing a z-score upon Yeo-Johnson (YJ) (16) transformed data. This transformation, shown in Equation 2, ensures that the data is sufficiently normally distributed to be amenable to parametric testing.

$$\Psi(x, \lambda) = \begin{cases} \frac{(x+1)^\lambda}{\lambda}\ if\ x \geq 0 \wedge \lambda \neq 0 \\ log(x+1)\ if\ x \geq 0 \wedge \lambda = 0 \\ -\frac{(-x+1)^{2-\lambda}-1}{2-\lambda}\ if\ x < 0 \wedge \lambda \neq 2 \\ -\log(-x+1)\ if\ x < 0 \wedge \lambda = 2 \end{cases} \quad (2)$$

The transform is similar to the Box-Cox power transformation, but obviates the requirement of inputting strictly positive values and has more desirable statistical properties. Furthermore, this transformation allowed us to avoid the use of non-parametric testing, which has its own pitfalls regarding assumptions of the underlying data distribution (17). We selected λ = 12, because Veridical's untransformed output is skewed left, due to their being, in general, less validating reads in control samples and the fact that there are, by design, vastly more control samples than experimental samples. We

found that this value for λ generally made the distribution much more normal. A comparison of the distributions of untransformed and transformed data is provided in Appendix S3.4. We were not concerned about small departures from normality as a z-test with a large number of samples is robust to such deviations (18).

Thus, we can compute the p-value of the pairwise unions of the two sets of partitions of ζ, except the irrelevant $\zeta_e \cup \alpha = \emptyset$. We only provide p-values for these pairwise unions and do not attempt to provide p-values for the partitions for the different consequences of the mutations on splicing. While such values would be useful, we do not currently have a robust means to compute them. Our previous work provides guidance on interpretation of splicing mutation outcomes (3-5,10). Thus for $\zeta_x \in \{\zeta_c, \zeta_e, \zeta_i\}$, let Φ_Z (z) represent the cumulative distribution function of the one-sided (right-tailed — i.e. P[X > x]) standard normal distribution. Let N represent the total number of samples and let V represent the set of all ζ_x validations, across all samples. Then:

$$\mu = \frac{\sum_{j=1}^{N} V_j}{N} \qquad \sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(V_j - \bar{V})^2} \qquad z = \frac{|\zeta_x| - \mu}{\sigma} \qquad p = \Phi(\Psi\left(z, \frac{1}{2}\right))$$

The program outputs two tables, along with summaries thereof. The first table lists all validated read counts across all categories for experimental samples, while the second table does the same for the control samples. P-values are shown in parentheses within the experimental table, which refer to the column-dependent (i.e. the read type is given in the column header) p-value for that read type with respect to that same read type in control samples. The program produces three files: a log file containing all details regarding validated variants, an output file with the programs progress reports and summaries, and a filtered validated variant file. The filtered file contains all validated variants of statistical significance (set as $p < 0.05$, by default), defined as variants with one or more validating reads achieving statistical significance in a strongly corroborating read type. These categories are limited to all junction-spanning based splicing consequences and read-abundance total intron inclusion. For example, a cryptic variant for which $p = 0.04$ in the junction-spanning cryptic column would meet this criteria, assuming the default significance threshold.

The p-values given by Veridical are more robust when the program is provided with a large number of samples. The minimum sample size is dependent upon the desired power, α value, and the effect size (ES). The minimum samples size could be computed as follows: $N = \left\lceil \frac{\sigma^2 z^2}{ES^2} \right\rceil$. For α = 0.05 and β = 0.2 (for a power of 0.8): z = 2.4865 for the one-tailed test. Then, $N = \left\lceil \frac{\sigma^2 2.4865^2}{ES^2} \right\rceil$. Ideally, Veridical could be run with a trial number of samples.

Then, one would compute effect sizes from Veridical's output. The standard deviation in the above formula could also be estimated from one's data, although it should be transformed using Yeo-Johnson (such as via an appropriate R package) before computing this estimation.

We elected to use RefSeq (19) genes for the exome annotation, as opposed to, the more permissive exome annotation sets, UCSC Known Genes (20) or Ensembl (21). The large number of transcript variants within Ensembl, in particular, caused many spurious intron inclusion validation events. This occurred because reads were found to be intronic in many cases, when in actuality they were exonic with respect to the more common transcript variant. In addition, the inclusion of the large number of rare transcripts in Ensembl significantly increased program run-time and made validation events much more challenging to interpret unequivocally. The use of RefSeq, which is a conservative annotation of the human exome, resolves these issues. It is possible that some subset of unknown or Ensemble annotated intronic transcripts could be sufficiently prevalent to merit inclusion in our analysis. We do not attempt to perform the difficult task of deciding which of these transcripts would be worth using. Indeed, the task of confirming and annotating of such transcripts is already done by the more conservative annotation we employ.

We also provide an R program (22) which produces publication quality histograms displaying embedded Q-Q plots and p-values, to evaluate for normality of the read distribution and statistical significance, respectively. The R program performs the YJ transformation as implemented in the car package (23). The histograms generated by the program use the Freedman-Draconis (24) rule for break determination, and the Q-Q plots

use algorithm Type 8 for their quantile function, as recommended by Hyndman and Fan (25). This program is embedded within a Perl script, for better integration into our workflow. Lastly, a Perl program was implemented to automatically retrieve and correctly format an exome annotation file from the UCSC database (20) for use in Veridical. All data use hg19/GRCh37, however when new versions of the genome become available, this program can be used to update the annotation file.

3.3 Results

Veridical validates predicted mRNA splicing mutations using high-throughput RNA sequencing data. We demonstrate how Veridical and its associated R program are used to validate predicted splicing mutations in somatic breast cancer. Each example depicts a particular variant-induced splicing consequence, analyzed by Veridical, with its corresponding significance level. The relevant primary RNA-Seq data are displayed in IGV, along with histograms and Q-Q plots showing the read distributions for each example. The source data are obtained from controlled-access breast carcinoma data from The Cancer Genome Atlas (TCGA) (26). Tumour-normal matched DNA sequencing data from the TCGA consortium was used to predict a set of splicing mutations, and a subset of corresponding RNA sequencing data was analyzed to confirm these predictions with Veridical. Overall, 442 tumour samples and 106 normal samples were analyzed. Briefly, all variants used as examples in this manuscript came from running the matched TCGA exome files (to which the RNA-Seq data corresponds) through SomaticSniper (27) and Strelka (28) to call somatic mutations, followed by the Shannon Human Splicing Pipeline (10) to find splicing mutations, which served as the input to Veridical. Details of the RNA-Seq data can be found within the supplementary methods of the TCGA paper (26). Accordingly, the following examples demonstrate the utility of Veridical to identify potentially pathogenic mutations from a much larger subset of predicted variants.

3.3.1 Leaky Mutations

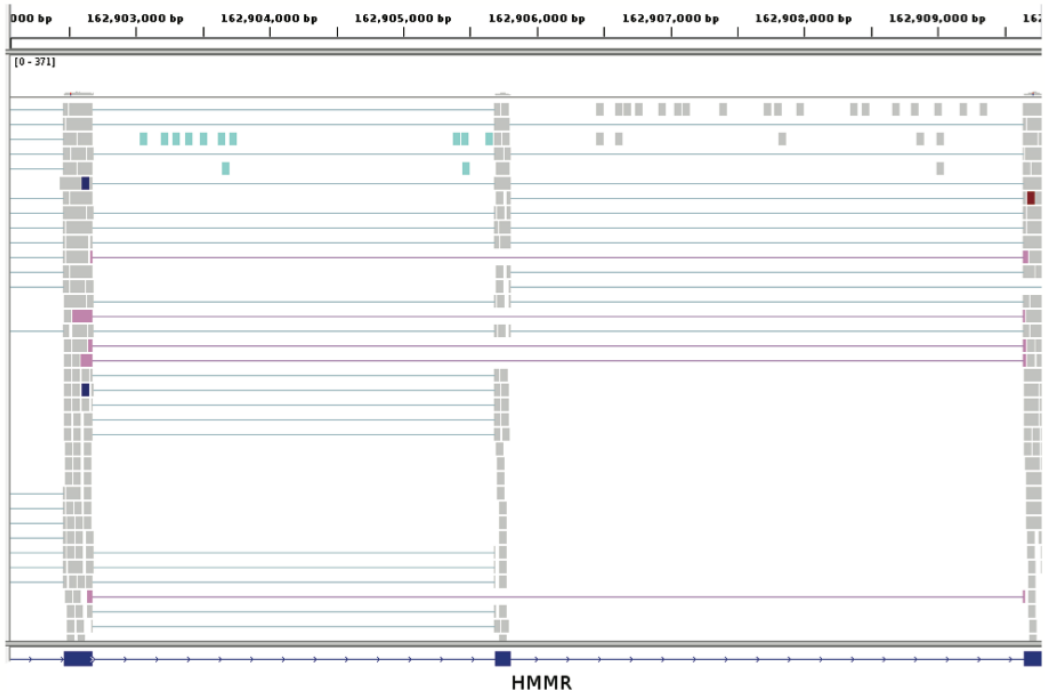
Mutations that reduce, but not abolish, the spliceosome's ability to recognize the intron/exon boundary are termed leaky (3). This can lead to the mis-splicing (intron

inclusion and/or exon skipping) of many but not all transcripts. An example, provided in Figure 3.4, displays a predicted leaky mutation (chr5:162905690G>T) in the HMMR gene in which both junction-spanning exon skipping ($p < 0.01$) and read-abundance-based intron inclusion ($p = 0.04$) are observed. We predict this mutation to be leaky because its final R_i exceeds 1.6 bits — the minimal individual information required to recognize a splice site and produce correctly spliced mRNA (4). Indeed, the natural site, while weakened by 2.16 bits, remains strong — 10.67 bits. This prediction is validated by the variant-containing sample's RNA-Seq data (Figure 3.4), in which both exon skipping (5 reads) and intron inclusion (14 reads, 12 of which are shown, versus an average of 4.051 such reads per control sample) are observed, along with 70 reads portraying wild-type splicing. Only a single normally spliced read contains the G→T mutation. These results are consistent with an imbalance of expression of the two alleles, as expected for a leaky variant. Figure 3.5 shows that for the distribution of read-abundance-based intron inclusion is marginally statistically significant ($p = 0.04$).

3.3.2 Inactivating Mutations

Variants that inactivate splice sites have negative final R_i values (3) with only rare exceptions (4), indicating that splice site recognition is essentially abolished in these cases. We present the analysis of two inactivating mutations within the PTEN and TMTC2 genes from different tumour exomes, namely: chr10:89711873A>G and chr12:83359523G>A, respectively. The PTEN variant displays junction-spanning exon skipping events ($p < 0.01$), while the TMTC2 gene portrays both junction-spanning and read-abundance-based intron inclusion (both splicing consequences with $p < 0.01$). In addition, all intron inclusion reads in the experimental sample contain the mutation itself, while only one such read exists across all control samples analyzed ($p < 0.01$). The PTEN variant contains numerous exon skipping reads (32 versus an average of 2.466 such reads per control sample). The TMTC2 variant contains many junction-spanning intron inclusion reads with the G→A mutation (all of its junction-spanning intron inclusion reads: 22 versus an average of 0.002 such reads per control sample). IGV screenshots for these variants are provided within Figure 3.6. This figure also shows an example of junction-spanning cryptic splice site activated by the mutation (chr1:985377C>T) within

A)



B)

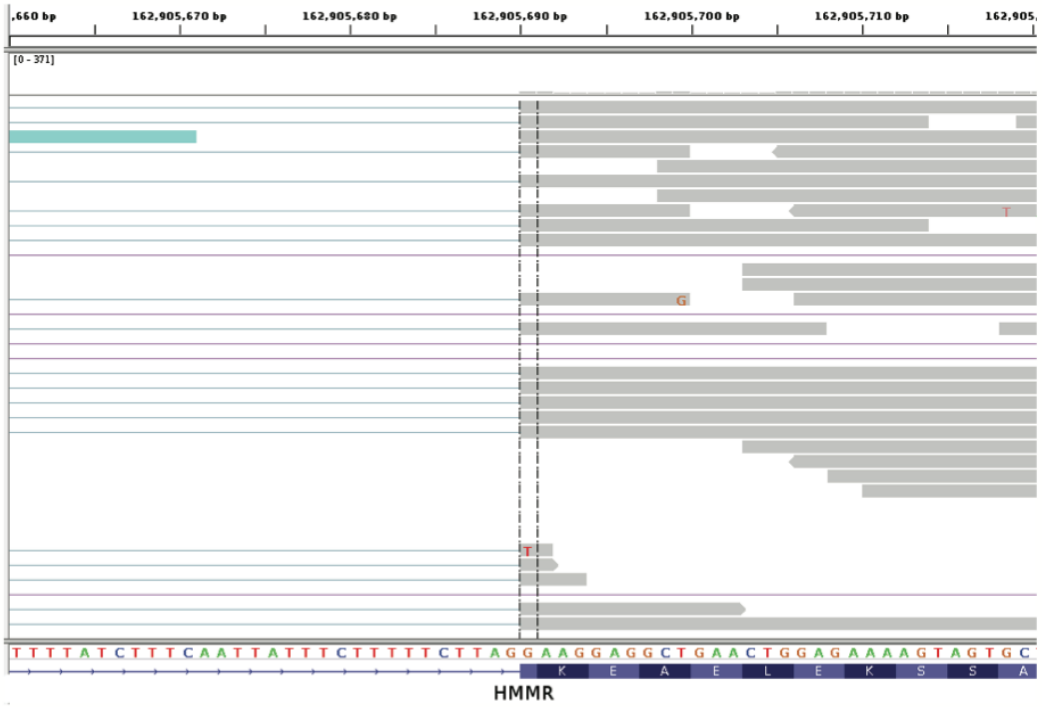


Figure 3.4 IGV images depicting a predicted leaky mutation (chr5:162905690G>T) within the natural acceptor site of exon 12 (162905689–162905806) of HMMR. This gene has four transcript variants and the given exon number pertains to isoforms a and b (reference sequences

NM_001142556 and NM_012484). RNA-Seq reads are shown in the centre panel. The bottom blue track depicts RefSeq genes, wherein each blue rectangle denotes an exon and blue connecting lines denote introns. In the middle panel, each rectangle (grey by default) denotes an aligned read, while thin lines are segments of reads split across exons. Red and blue coloured rectangles in the middle panel denote aligned reads of inserts that are larger or smaller than expected, respectively. Reads are highlighted by their splicing consequence, as follows: cryptic splicing (green), exon skipping (purple), junction-spanning intron inclusion (dark green), and read-abundance intron inclusion (cyan). (A) depicts a genomic region of chromosome 5: 162902054–162909787. The variant occurs in the middle exon. Intron inclusion can be seen in this image, represented by the reads between the first and middle exon (since the direction is left, as described within Table 1). These 14 reads are read-abundance-based, since they do not span the intron-exon junction. (B) depicts a closer view of the region shown in (A) — 162905660–162905719. The dotted vertical black lines are centred upon the first base of the variant-containing exon. The thin lines in the middle panel that span the entire exon fragment are evidence of exon skipping. These 5 reads are split across the exon before and after the variant-containing exon, as seen in (A).

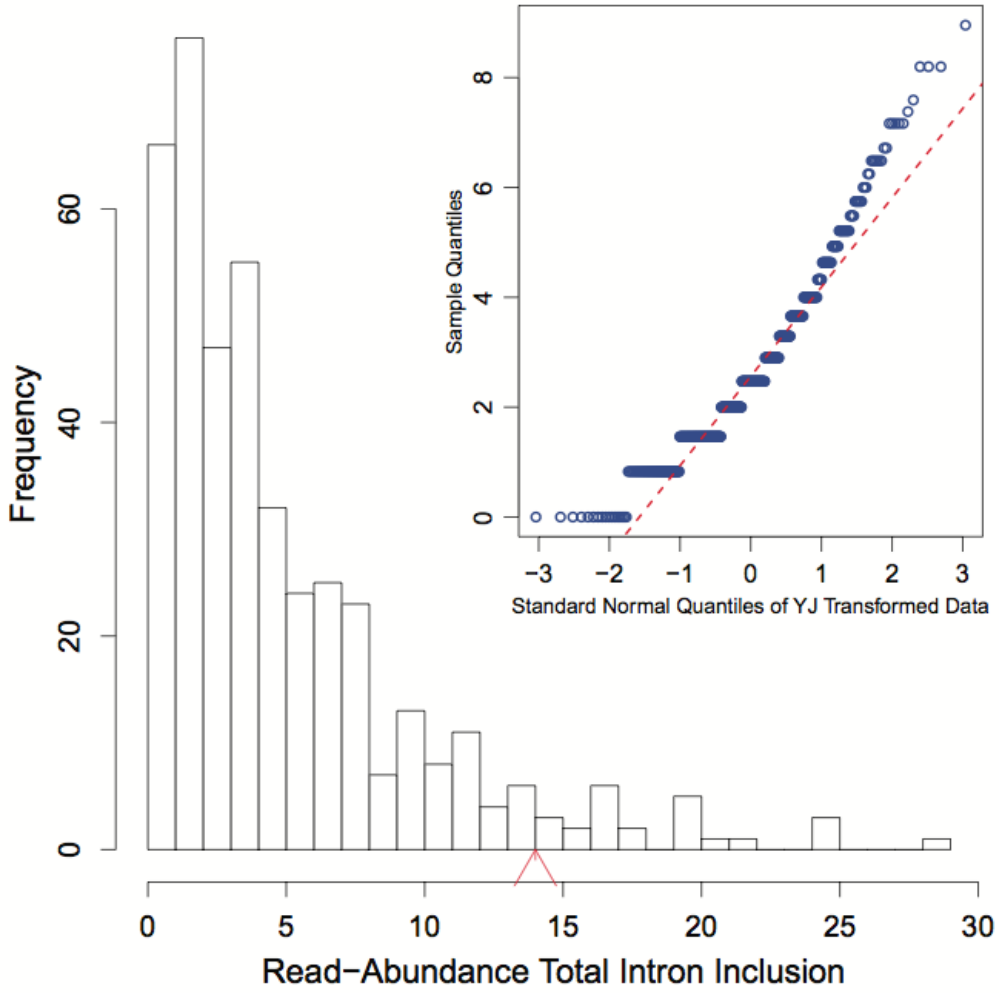
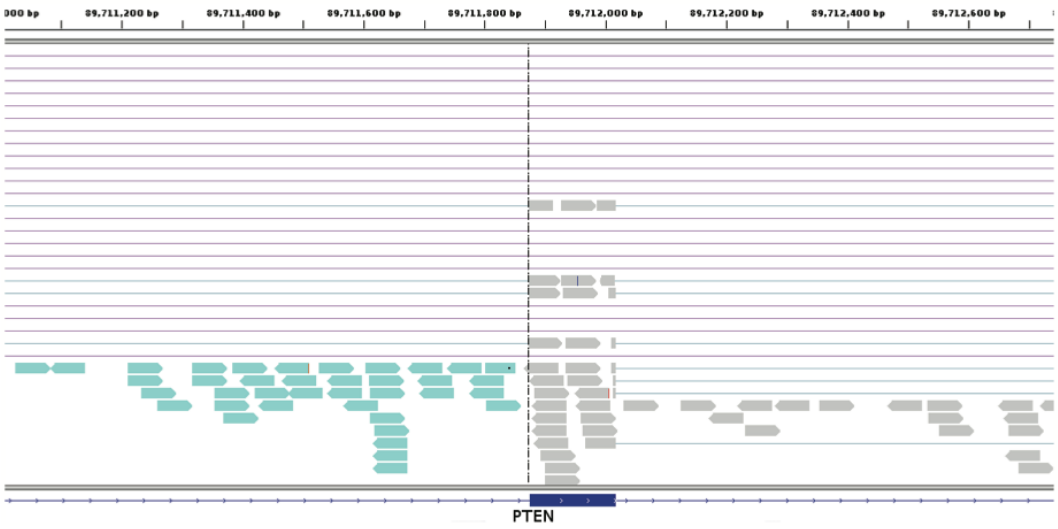
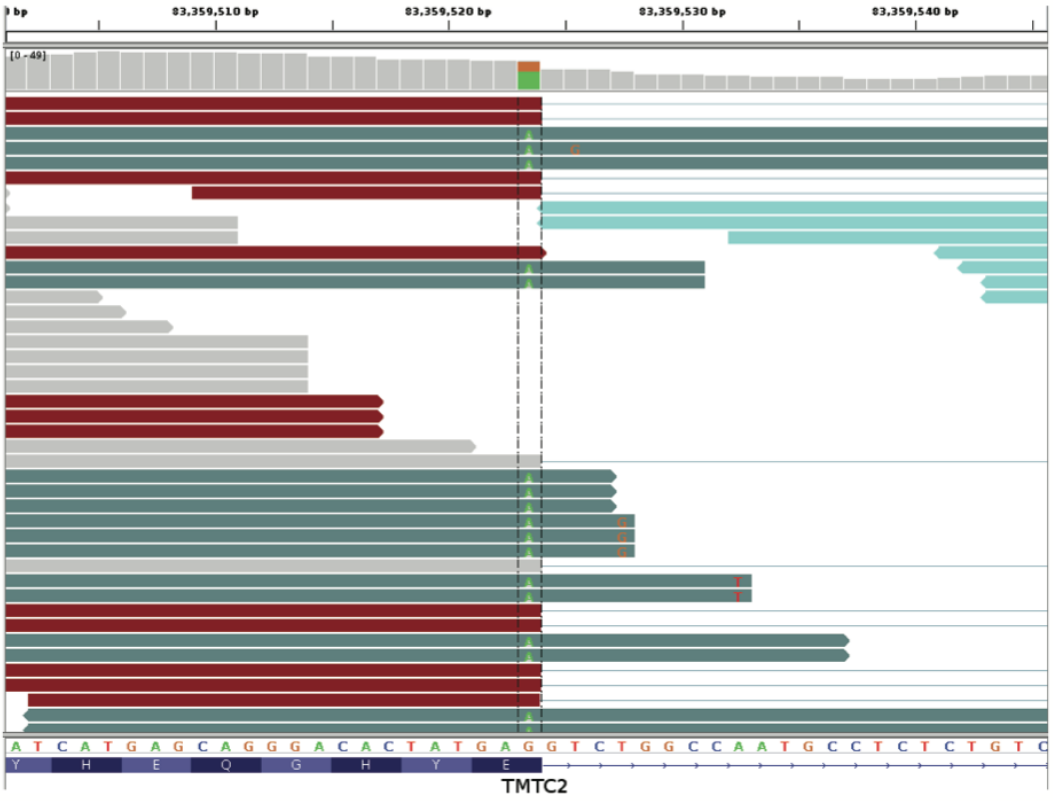


Figure 3.5 Histogram of read-abundance-based intron inclusion with embedded Q-Q plots of the predicted leaky mutation (chr5:162905690G>T) within HMMR, as shown in Figure 4. The arrowhead denotes the number of reads (14 in this case) in the variant-containing file, which is more than observed in the control samples (p = 0.04).

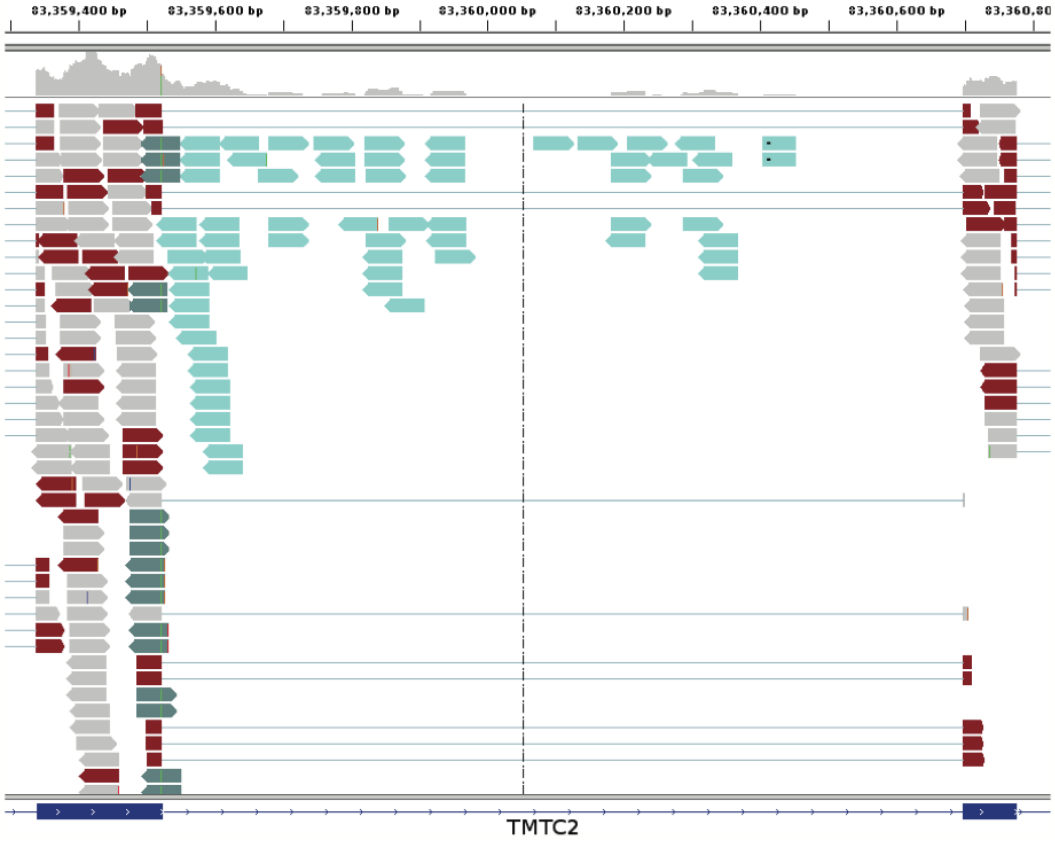
A)


89,711,200 bp
89,711,400 bp
89,711,600 bp
89,711,800 bp
89,712,000 bp
89,712,200 bp
89,712,400 bp
89,712,600 bp
PTEN

B)


83,359,510 bp
83,359,520 bp
83,359,530 bp
83,359,540 bp
[0 - 49]
A T C A T G A G C A G G G A C A C T A T G A G G T C T G G C C A A T G C C T C T C T G T C
Y H E Q G H Y E
TMTC2

C)



D)

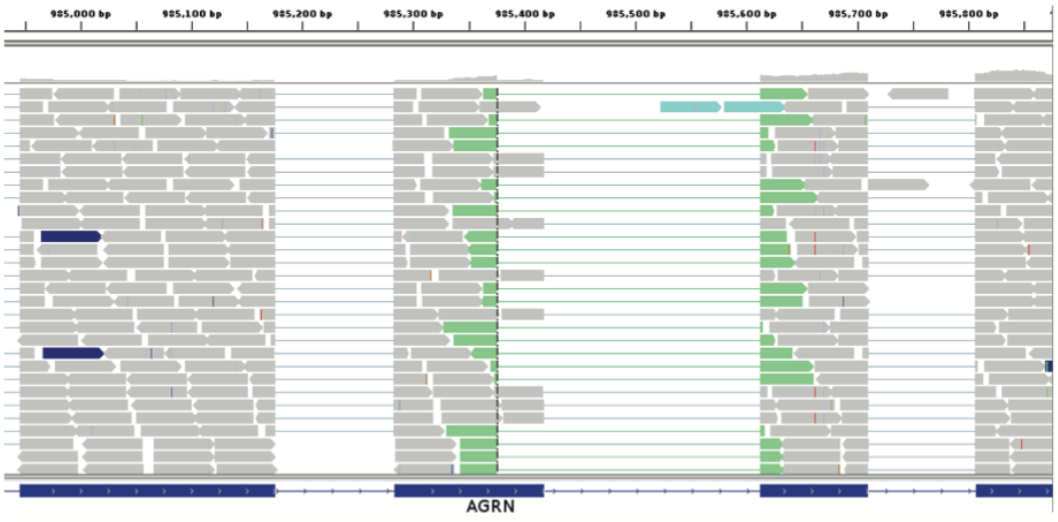


Figure 3.6 Examples of validated mutations. (A) depicts an inactivating mutation (chr10:89711873A>G) within the natural acceptor site of exon 6 (89711874–89712016) of PTEN.

The dotted vertical black line denotes the location of the relevant splice site. The region displayed is 89711004–89712744 on chromosome 10. Many of the 32 exon skipping reads are evident, typified by the thin lines in the middle panel that span the entire exon. There is also a substantial amount of read-abundance-based intron inclusion, shown by the reads to the left of the dotted vertical line. Exon skipping was statistically significant ($p < 0.01$), while read-abundance-based intron inclusion was not ($p = 0.53$). Panels (B) and (C) depict an inactivating mutation (chr12:83359523G>A) within the natural donor site of exon 6 (83359338–83359523) of TMTC2. (B) depicts a closer view (83359501–83359544) of the region shown in (C) and only shows exon 6. Some of the 22 junction-spanning intron inclusion reads can be seen. In this case, all of these reads contain the mutation, shown by the green adenine base in each read, between the two vertical dotted lines. (C) depicts a genomic region of chromosome 12: 83359221–83360885, TMTC2 exons 6–7. The variant occurs in the left exon. 65 read-abundance-based intron inclusion can be seen in this image, represented by the reads between the two exons. Panel (D) depicts a mutation (chr1:985377C>T) causing a cryptic donor to be activated within exon 27 (the second from left, 985282–985417) of AGRN. The region displayed is 984876–985876 on chromosome 1 (exons 26–29 are visible). Some of the 34 cryptic (junction-spanning) reads are portrayed. The dotted black vertical line denotes the cryptic splice site, at which cryptic reads end. The read-abundance-based intron inclusion, of which two reads are visible, was not statistically significant ($p = 0.68$). Refer to the caption of Figure 4 for IGV graphical element descriptions.

the AGRN gene. The concordance between the splicing outcomes generated by these mutations and the Veridical results indicates that the proposed method detects both mutations that inactivate splice sites and cryptic splice site activation.

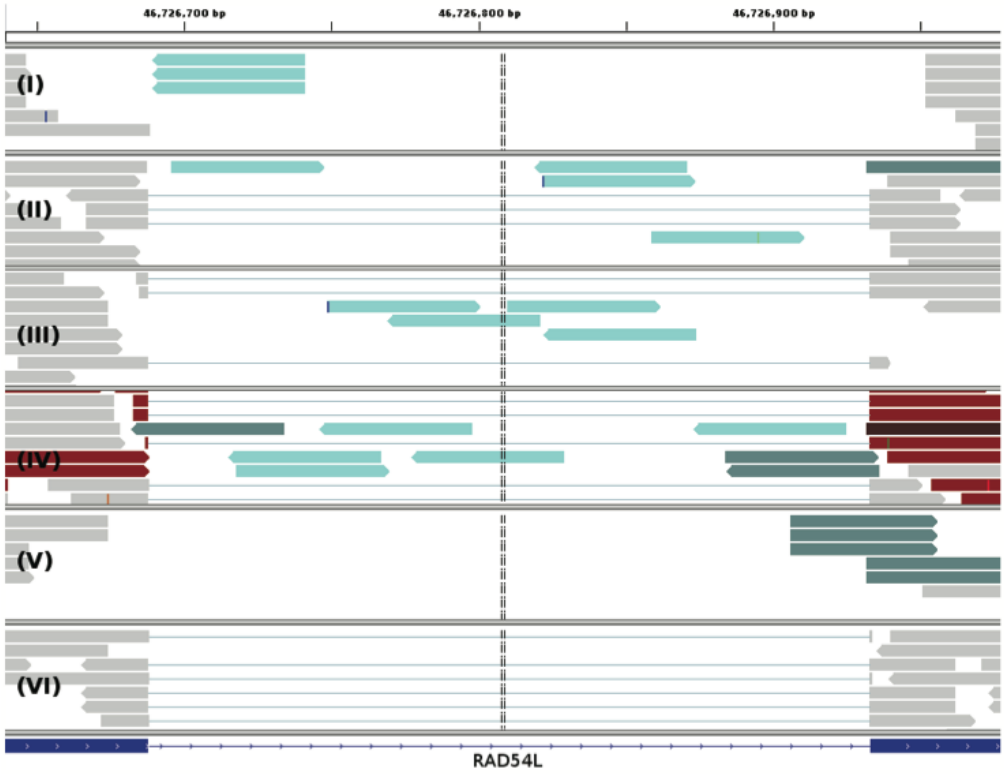
3.3.3 Cryptic Mutations

Recurrent genetic mutations in some oncogenes have been reported among tumours within the same, or different, tissues of origin. Common recurrent mutations present in multiple abnormal samples are recognized by Veridical. This avoids including a variant-containing sample among the control group, and outputs the results of all of the variant-containing samples. A relevant example is shown in Figure 3.7. The mutation (chr1:46726876G>T) causes activation of a cryptic splice site within RAD54L in multiple tumours. Upon computation of the p-values for each of the variant-containing tumours, relative to all non-variant containing tumours and normal controls, not all variant-containing tumours displayed splicing abnormalities at statistically significant levels. Of the six variant-containing tumours, two had significant levels of junction-spanning intron inclusion, and one showed statistically significant read-abundance-based intron inclusion. Details for all of the aforementioned variants, including a summary of read counts pertaining to each relevant splicing consequence, for experimental versus control samples, are provided in Table 3.2.

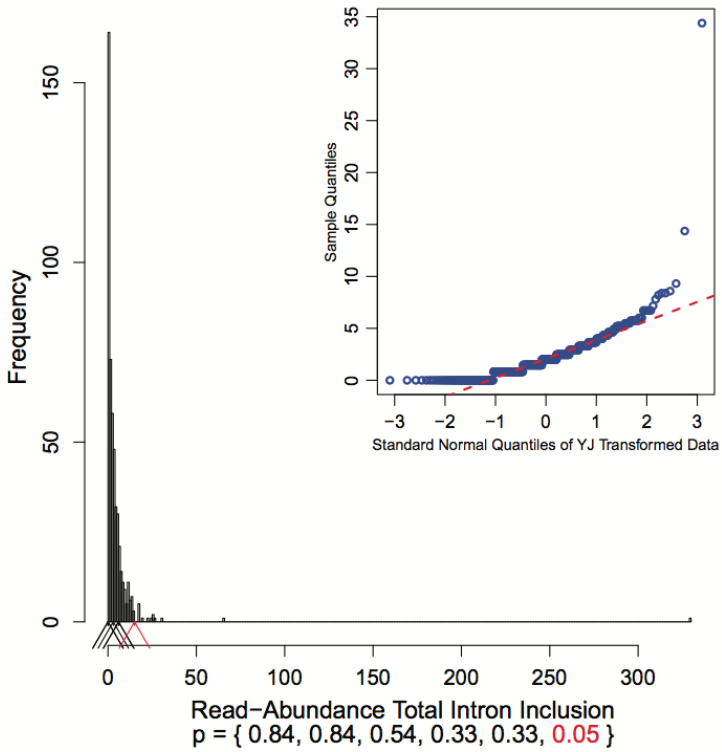
3.3.4 Performance

The performance of the software is affected by the number of predicted splicing mutations, the number of abnormal samples containing mutations and control samples and the corresponding RNA-Seq data for each type of sample. Veridical has the ability to analyze approximately 3000 variants in approximately 4 hours, assuming an input of 100 BAM files of RNA-Seq data. The relationship between time and numbers of BAM files and variants are plotted in Figure 3.8 for a 2.27 GHz processor. Veridical uses memory in linear proportion to the number and size of the input BAM files. In our tests, using RNA-Seq BAM files with an average size of approximately 6 GB, Veridical used approximately 0.7 GB for ten files to 1 GB for 100 files.

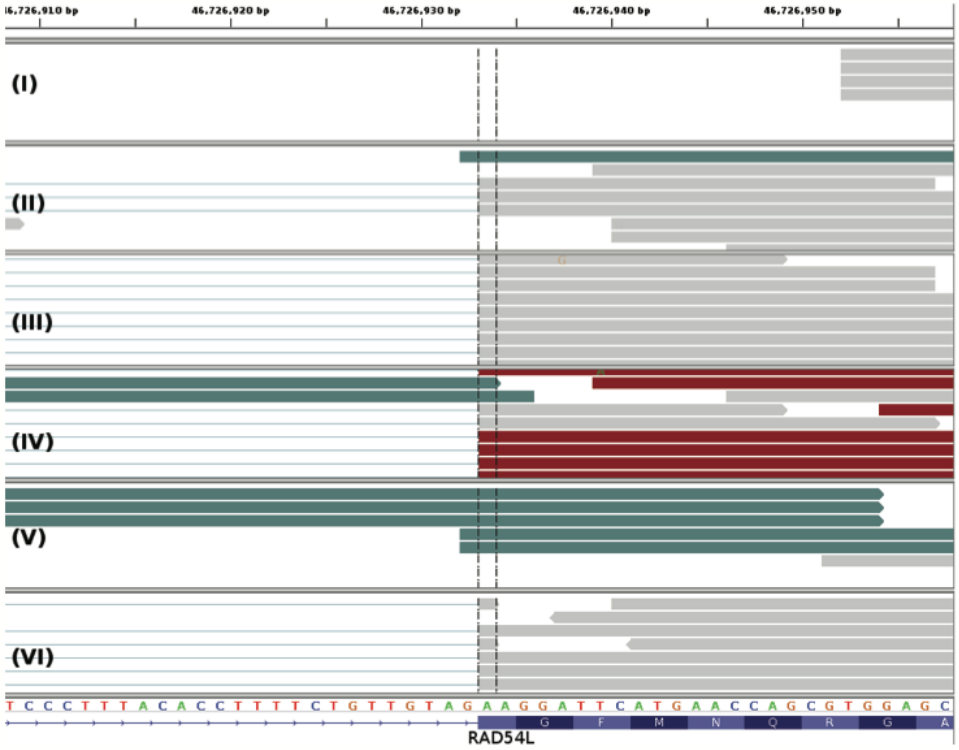
A)


46,726,700 bp
46,726,800 bp
46,726,900 bp
(I)
(II)
(III)
(IV)
(V)
(VI)
RAD54L

B)


Frequency
0
50
100
150
0
50
100
150
200
250
300
Read−Abundance Total Intron Inclusion
p = { 0.84, 0.84, 0.54, 0.33, 0.33, 0.05 }
Sample Quantiles
0
5
10
15
20
25
30
35
−3
−2
−1
0
1
2
3
Standard Normal Quantiles of YJ Transformed Data

C)


46,726,910 bp
46,726,920 bp
46,726,930 bp
46,726,940 bp
46,726,950 bp
(I)
(II)
(III)
(IV)
(V)
(VI)
RAD54L

D)

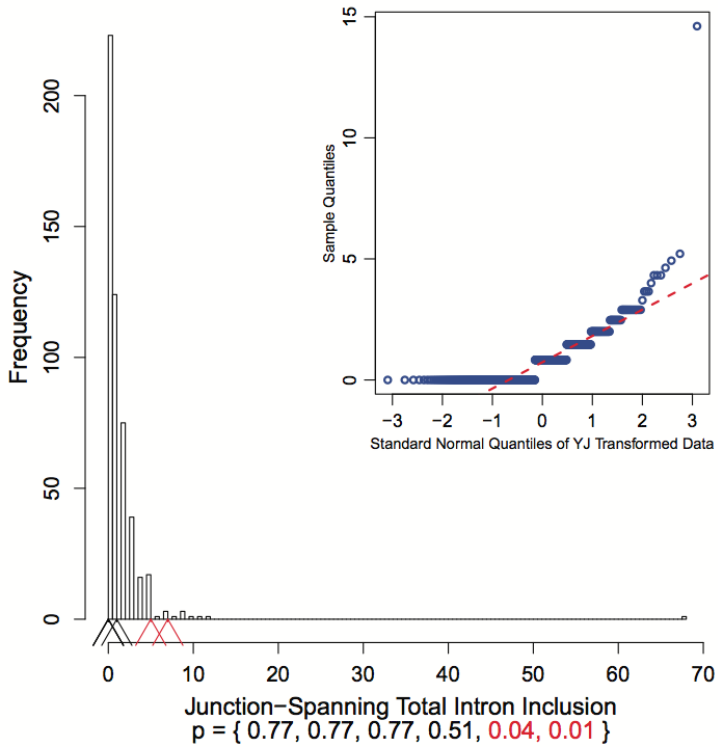

Frequency
0
50
100
150
200
0
10
20
30
40
50
60
70
Junction−Spanning Total Intron Inclusion
p = { 0.77, 0.77, 0.77, 0.51, 0.04, 0.01 }
Sample Quantiles
0
5
10
15
−3
−2
−1
0
1
2
3
Standard Normal Quantiles of YJ Transformed Data

Figure 3.7 IGV images and their corresponding histograms with embedded Q-Q plots depicting all six variant-containing files with a mutation (chr1:46726876G>T) which, in some cases, causes a cryptic donor to be activated within the intron between exons 7 and 8 of RAD54L. This results in the extension of the downstream natural donor (the 5′ end of exon 8). This gene has two transcript variants and the given exon numbers pertain to isoform a (reference sequence NM_003579). Only samples IV and V have statistically significant intron inclusion relative to controls. read-abundance-based intron inclusion can be seen in (A), between the two exons. The region displayed is on chromosome 1: 46726639–46726976. (B) depicts the corresponding histogram for the 15 read-abundance-based intron inclusion reads ($p = 0.05$) that are present in sample IV. The intron-exon boundary on the right is the downstream natural donor. (C) typifies some of the 13 junction-spanning intron inclusion reads that are a direct result of the intronic cryptic site's activation. In these instances, reads extending past the intron-exon boundary are being spliced at the cryptic site, instead of the natural donor. In particular, samples IV and V both have a statistically significant numbers of such reads, 7 ($p = 0.01$) and 5 ($p = 0.04$), respectively. This is further typified by the corresponding histogram in (D). (C) focuses upon exon 8 from (A) and displays the genomic positions 46726908–46726957. Refer to the caption of Figure 4 for IGV graphical element descriptions. In the histograms, arrowheads denote numbers of reads in the variant-containing files. The bottom of the plots provide p-values for each respective arrowhead. Statistically significant p-values and their corresponding arrowheads are denoted in red.

Table 3.2 Examples of variants validated by Veridical and their selected read types.

Gene	Chr	C_v	C_s	Variant	Type	Initial R_i	Final R_i	ΔRi	#	SC	ET	p-value	R_E	R_T	R_N	R_μ	Figure
HMMR	chr5	162905690	162905689	G/T	Leaky	12.83	10.67	-2.16		ES	JS	<0.01	5	11	0	0.02	3.4,3.5
										ES	RA	0.04	14	2133	103	4.051	
PTEN	chr10	89711873	89711874	A/G	Inactivating	12.09	-2.62	-14.71		ES	JS	<0.01	32	975	386	2.466	6(A)
TMTC2	chr12	83359523	83359524	G/A	Inactivating	1.74	-1.27	-3.01		ES	JS	<0.01	22	2241	383	4.754	6(B)
										ES	JSwM	<0.01	22	0	1	0.002	
										ES	RA	<0.01	65	7293	1395	15.739	6(C)
AGRN	chr1	985377	985376	C/T	Cryptic	-2.24	4.79	7.03		CS	JS	<0.01	34	97	23	0.217	6(D)
RAD54L	chr1	46726876	46726895	G/T	Cryptic	13.4	14.84	1.44	I	CS	JS	NA	0	645	58	1.274	7
										CS	RA	0.54	3	2171	290	4.458	
									II	CS	JS	0.51	1	645	58	1.274	
										CS	RA	0.33	6	2171	290	4.458	
									III	CS	JS	NA	0	645	58	1.274	
										CS	RA	0.33	6	2171	290	4.458	
									IV	CS	JS	0.01	7	645	58	1.274	
										CS	RA	0.05	15	2171	290	4.458	
									V	CS	JS	0.04	5	645	58	1.274	
										CS	RA	NA	0	2171	290	4.458	
									VI	CS	JS	NA	0	645	58	1.274	
										CS	RA	NA	0	2171	290	4.458	

Header abbreviations Chr, C_v, C_s, #, SC, and ET, denote chromosome, variant coordinate, splice site coordinate, sample number (where applicable), splicing consequence, and evidence type, respectively. Headers containing R with some subscript denote numbers of validated reads for the specified variant's splicing consequence(s) and evidence type(s). RE denotes reads within variant-containing tumour samples. RT and RN denote control samples, for tumours and normal cells, respectively. R_μ is the per sample mean of RT and RN. Splicing consequences: CS denotes

cryptic splicing, ES denotes exon skipping, and II denotes intron inclusion. Evidence types: JS denotes junction-spanning and RA denotes read-abundance.

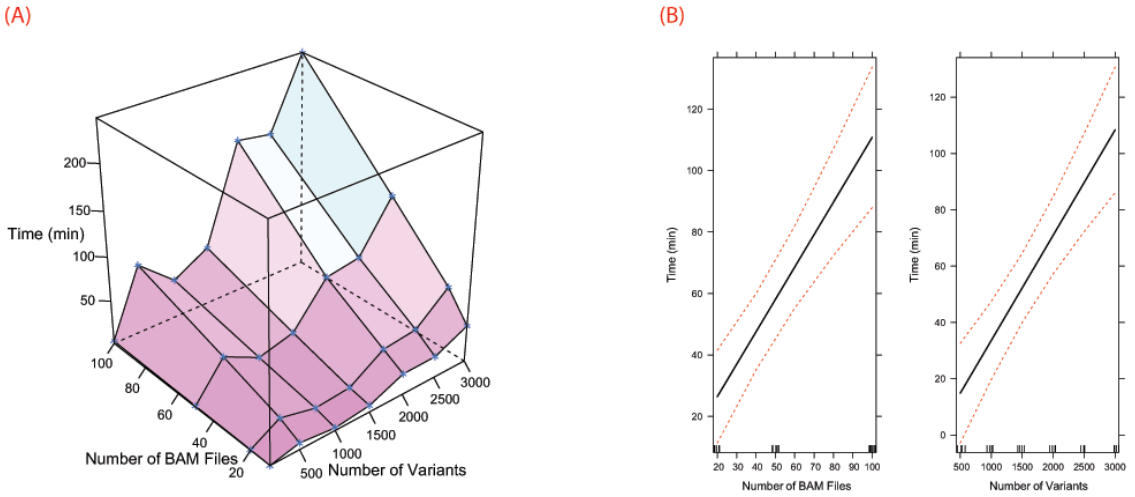


Figure 3.8 Profiling data for Veridical runtime. Tests were conducted upon an Intel Xeon @ 2.27 GHz. Visualizations were generated with R (22) using Lattice and Effects. A surface plot of time vs. numbers of BAM files and variants is provided in (A). Effect plots are given in (B) and demonstrate the effects of the numbers of BAM files and variants upon runtime. The effect plots were generated using a linear regression model (R^2 = 0.7525).

3.4 Discussion

We have implemented Veridical, a software program that automates confirmation of mRNA splicing mutations by comparing sequence read-mapped expression data from samples containing variants that are predicted to cause defective splicing with control samples lacking these mutations. The program objectively evaluates each mutation with statistical tests that determine the likelihood of and exclude normal splicing. To our knowledge, no other software currently validates splicing mutations with RNA-Seq data on a genome-wide scale, although many applications can accurately detect conventional alternative splice isoforms (i.e. (29)). Veridical is intended for use with large data sets derived from many samples, each containing several hundred variants that have been previously prioritized as likely splicing mutations, regardless of how the candidate mutations are selected. It is not practical to analyze all variants present in an exome or genome, rather only a filtered subset, due to the extensive computations required for statistical validation. As such, Veridical is a key component of an end-to-end, hypothesis-based, splicing mutation analysis framework that also includes the Shannon splicing mutation pipeline (10) and the Automated Splice Site Analysis and Exon Definition server (5). There is a trade-off between lengthy run-times and statistical robustness of Veridical, especially when there are either a large number of variants or a large number of RNA-Seq files. As with most statistical methods, those employed here are not amenable to small sample sets, but become quite powerful when a large number of controls are employed. In order to ensure that mutations can be validated, we recommend an excess of control transcriptome data relative to those from samples containing mutations ($> 5 : 1$), guided by the power analysis described in Methods. We do not recommend the use of a single nor a few control samples to corroborate a putative mutation. Not surprisingly, we have found that junction-spanning reads have the greatest value for corroborating cryptic splicing and exon skipping. Even a single such read is almost always sufficient to merit the validation of a variant, provided that sufficient control samples are used. For intron inclusion, both junction-spanning and read-abundance-based reads are useful and a variant can readily be validated with either,

provided that the variant-containing experimental sample(s) show a statistically significant increase in the presence of either form of intron inclusion corroborating reads.

Veridical is able to automatically process variants from multiple different experimental samples, and can group the variant information if any given mutation is present in more than one sample. The use of a large sample size allows for robust statistical analyses to be performed, which aid significantly in the interpretation of results. The main utility of Veridical is to filter through large data sets of predicted splicing mutations to prioritize the variants. This helps to predict which variants will have a deleterious effect upon the protein product. Veridical is able to avoid reporting splicing changes that are naturally occurring through checking all variant-containing and non-containing control samples for the predicted splicing consequence. In addition, running multiple tumour samples at once allows for manual inspection to discover samples that contained the alternative splicing pattern, and consequently, permits the identification of DNA mutations in the same location which went undetected during genome sequencing.

The statistical power of Veridical is dependent upon the quality of the RNA-Seq data used to validate putative variants. In particular, a lack of sufficient coverage at a particular locus will cause Veridical to be unable to report any significant results. A coverage of at least 20 reads should be sufficient. This estimate is based upon alternative splicing analyses in which this threshold was found to imply concordance with microarray and RT-PCR measurements (30-33). There are many potential legitimate reasons why a mutation may not be validated: (a) A lack of gene expression in the variant containing tumour sample, (b) nonsense-mediated decay may result in a loss of expression of the entire transcript, (c) the gene itself may have multiple paralogs and reads may not be unambiguously mapped, (d) other non-splicing mutations could account for a loss of expression, and (e) confounding natural alternative splicing isoforms may result in a loss of statistical significance during read mapping of the control samples. The prevalence of loci with insufficient data is dependent upon the coverage of the sequencing technology used. As sequencing technologies improve, the proportion of validated mutations is expected to increase. Such an increase would mirror that observed for the prevalence of alternative splicing events (34). In addition, mutated splicing factors

can disrupt splicing fidelity and exon definition (35). This effect could decrease Veridical's ability to validate splicing mutations affected by a disruption of the definition of the pertinent exon. Veridical does not currently form any equivalence between distinct variants affecting the same splice site. Such variants will be analyzed independently. Veridical is intended to be used with RNA-Seq data that not only corresponds to matched DNA-Seq data, but also only for sets of samples with comparable sequencing protocols, since the non-normalized comparisons performed rely upon the evening out of batch effects, due to a substantial number of control samples. It is important to note that acceptance of the null hypothesis, due to an absence of evidence required to disprove it, does not imply that the underlying prediction of a mutation at a particular locus is incorrect, but merely that the current empirical methods employed were insufficient to corroborate it.

"Validate," in the present context, refers to the condition where sufficient statistical evidence has been marshaled in support of a variant. However, the threshold for significance can vary so these analyses can also be thought of as strongly corroborating variants. Recent studies in Bayesian statistics have suggested that a p-value threshold of 0.05 does not correspond to strong support of the alternative hypothesis. Accordingly, Johnson (36) recommends the use of tests at the 0.005 or 0.001 level of significance.

We consider alternative splicing to be a different problem. Veridical does not aim to identify putatively pathogenic variants, but rather, to confirm existing in silico predictions thereof. We do infer exon skipping events (i.e. alternative splicing) *de novo*, but only to catalog dysregulated splicing "phenotypes" due to genomic sequence variants. This is not the first study to use a large control dataset. Indeed the Variant Annotation, Analysis & Search Tool (VAAST) (37) does this to search for disease-causing (non-splicing) variants and the Multivariate Analysis of Transcript Splicing (MATS) (29) tool (among others) can be used for the discovery of alternative splicing events. However, in our case, in most instances the distribution of reads in a single sample is compared to the distributions of reads in the control set, as opposed to a likelihood framework-based approach. We are suggesting that our approach be coupled to existing approaches to act as an a posteriori, hypothesis-driven, check on the veridicality of specific variants.

While there is considerable prior evidence for splicing mutations that alter natural and cryptic splice site recognition, we were somewhat surprised at the apparent high frequency of statistically significant intron inclusion revealed by Veridical. In fact, evidence indicates that a significant portion of the genome is transcribed (34), and it is estimated that 95% of known genes are alternatively spliced (30). Defective mRNA splicing can lead to multiple alternative transcripts including those with retained introns, cassette exons, alternate promoters/terminators, extended or truncated exons, and reduced exons (38). In breast cancer, exon skipping and intron retention were observed to be the most common form of alternative splicing in triple negative, non-triple negative, and HER2 positive breast cancer (39). In normal tissue, intron retention and exon skipping has been predicted to affect 2572 exons in 2127 genes and 50 633 exons in 12 797 genes, respectively (40). In addition, previous studies suggest that the order of intron removal can influence the final mRNA transcript composition of exons and introns (41). Intron inclusion observed in normal tissue may result from those introns that are removed from the transcript at the end of mRNA splicing. Given that these splicing events are relatively common in normal tissues, it becomes all the more important to distinguish expression patterns that are clearly due to the effects of splicing mutations — one of the guiding principles of the Veridical method.

Veridical is an important analytical resource for unsupervised, thorough validation of splicing mutations through the use of companion RNA-Seq data from the same samples. The approach will be broadly applicable for many types of genetic abnormalities, and should reveal numerous, previously unrecognized, mRNA splicing mutations in exome and complete genome sequences.

3.5 References

1. Rogan PK, Zou GY: Best practices for evaluating mutation prediction methods. Hum Mutat. 2013; 34(11): 1581–1582.

2. Krawczak M, Reiss J, Cooper DN: The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: Causes and consequences. Hum Genet. 1992; 90(1–2): 41–54.

3. Rogan PK, Faux BM, Schneider TD: Information analysis of human splice site mutations. Hum Mutat. 1998; 12(3): 153–171.

4. Rogan PK, Svojanovsky S, Leeder JS: Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. Pharmacogenetics. 2003; 13(4): 207–218.

5. Mucaki EJ, Shirley BC, Rogan PK: Prediction of mutant mRNA splice isoforms by information theory-based exon definition. Hum Mutat. 2013; 34(4): 557–565.

6. López-Bigas N, Audit B, Ouzounis C, et al.: Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett. 2005; 579(9): 1900–1903.

7. Wang K, Li M, Hakonarson H: ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38(16): e164.

8. Churbanov A, Vorechovský I, Hicks C: A method of predicting changes in human gene splicing induced by genetic variants in context of cis-acting elements. BMC Bioinformatics. 2010; 11(1): 22.

9. Pertea M, Lin X, Salzberg SL: GeneSplicer: A new computational method for splice site prediction. Nucleic Acids Res. 2001; 29(5): 1185–1190.

10. Shirley BC, Mucaki EJ, Whitehead T, et al.: Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences. Genomics Proteomics Bioinformatics. 2013; 11(2): 77–85.

11. Eswaran J, Cyanam D, Mudvari P, et al.: Transcriptomic landscape of breast cancers through mRNA sequencing. Sci Rep. 2012; 2: 264.

12. Eswaran J, Horvath A, Godbole S, et al.: RNA sequencing of cancer reveals novel splicing alterations. Sci Rep. 2013; 3: 1689.

13. Kwan T, Benovoy D, Dias C, et al.: Genome-wide analysis of transcript isoform variation in humans. Nat Genet. 2008; 40(2): 225–231.

14. Thorvaldsdóttir H, Robinson JT, Mesirov JP: Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform. 2013; 14(2): 178–192.

15. Barnett DW, Garrison EK, Quinlan AR, et al.: BamTools: A C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011; 27(12): 1691–1692.

16. Yeo IK, Johnson RA: A new family of power transformations to improve normality or symmetry. Biometrika. 2000; 87(4): 954–959.

17. Johnson DH: Statistical sirens: the allure of nonparametrics. Ecology. 1995; 76: 1998–2000.

18. Hubbard R: The probable consequences of violating the normality assumption in parametric statistical analysis. Area. 1978; 10(5): 393–398.

19. Pruitt KD, Tatusova T, Maglott DR: NCBI reference sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005; 33(Database Issue): D501–D504.

20. Hsu F, Kent JW, Clawson H, et al.: The UCSC known genes. Bioinformatics. 2006; 22(9): 1036–1046.

21. Hubbard T, Barker D, Birney E, et al.: The Ensembl genome database project. Nucleic Acids Res. 2002; 30(1): 38–41.

22. RDC Team R: A Language and Environment for Statistical Computing . Vienna, Austria: R Foundation for Statistical Computing, 2008.

23. Fox J, Weisberg S: An R Companion to Applied Regression, 2nd ed. Thousand Oaks CA: Sage, 2011.

24. Freedman D, Diaconis P: On the histogram as a density estimator: L2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete. 1981; 57(4): 453–476.

25. Hyndman RJ, Fan Y: Sample quantiles in statistical packages. American Statistician. 1996; 50(4): 361–365.

26. Koboldt DC, Fulton RS, McLellan MD, et al.: Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490(7418): 61–70.

27. Larson DE, Harris CC, Chen K, et al.: SomaticSniper: Identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012; 28(3): 311–317.

28. Saunders CT, Wong WSW, Swamy S, et al.: Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012; 28(14): 1811–1817.

29. Shen S, Park JW, Huang J, et al.: MATS: A bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. Nucleic Acids Res. 2012; 40(8): e61.

30. Pan Q, Shai O, Lee LJ, et al.: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008; 40(12): 1413–1415.

31. Griffith M, Griffith OL, Mwenifumbo J, et al.: Alternative expression analysis by RNA sequencing. Nat Methods. 2010; 7(10): 843–847.

32. Katz Y, Wang ET, Airoldi EM, et al.: Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. 2010; 7(12): 1009–1015.

33. Shen S, Lin L, Cai JJ, et al.: Widespread establishment and regulatory impact of Alu exons in human genes. Proc Natl Acad Sci U S A. 2011; 108(7): 2837–2842.

34. Kapranov P, Willingham AT, Gingeras TR: Genome-wide transcription and the implications for genomic organization. Nat Rev Genet. 2007; 8(6): 413–423.

35. Singh RK, Cooper TA: Pre-mRNA splicing in disease and therapeutics. Trends Mol Med. 2012; 18(8): 472–482.

36. Johnson VE: Revised standards for statistical evidence. Proc Natl Acad Sci U S A. 2013; 110(48): 19313–19317.

37. Yandell M, Huff C, Hu H, et al.: A probabilistic disease-gene finder for personal genomes. Genome Res. 2011; 21(9): 1529–1542.

38. Feng H, Qin Z, Zhang X: Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. Cancer Lett. 2013; 340(2): 179–191.

39. Eswaran J, Horvath A, Godbole S, et al.: RNA sequencing of cancer reveals novel splicing alterations. Sci Rep. 2013; 3: 1689.

40. Pal S, Gupta R, Davuluri RV: Alternative transcription and alternative splicing in cancer. Pharmacol Ther. 2012; 136(3): 283–294.

41. Takahara K, Schwarze U, Imamura Y, et al.: Order of intron removal influences multiple splice outcomes, including a two-exon skip, in a COL5A1 acceptor-site mutation that results in abnormal pro-a1 (V) N-propeptides and Ehlers-Danlos syndrome type I. Am J Hum Genet. 2002; 71(3): 451–465.

Chapter 4

4 Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer

The work presented in this chapter is reproduced (with permission, Appendix S1) from:

Dorman, S.N., Viner, C., Rogan, P.K. (2014) Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer. *Scientific Reports* 4:7063. DOI: 10.1038/srep07063

4.1 Introduction

Large-scale DNA sequencing studies have attempted to elucidate the genomic landscapes of breast cancer tumours to identify mutated genes and genomic variation that contribute to tumour development and progression (1-5). Typically, somatic mutations within gene coding regions are identified and then filtered for rare or novel variants predicted to affect protein structure or function (6-9). Frequently mutated genes are cataloged, with the goal of inferring defective genes that are more likely to contribute to tumour phenotypes. However, there does not appear to be a consistent set of somatic driver mutations in most breast cancer cases. For instance, in 100 cases, 73 different combinations of abnormal gene sequences were reported (4).

Some established cancer genes are enriched for mutations (i.e. *TP53*, *PIK3CA*, *PTEN*, *MAP3K1*, *AKT1*, *CDH1*, *GATA3*, *MLL3* and *RB1*), in addition to genes that were not previously associated with breast cancer (including *CBFB*, *RUNX1*, *TBX3*, *NF1* and *SF3B1*) (1-5). At least 49 genes (including known breast cancer genes) have been found to be significantly mutated, 16 of these reproducibly across multiple studies, and the majority were mutated in <10% of tumours.

Inconsistencies in mutation composition among different tumours present significant challenges to understanding the underlying etiology of tumour phenotypes. As a result of epistasis, mutations in genes with linked biochemical functions would be expected to reveal dysfunctional pathways in tumours (10). Focusing analyses to one molecular subtype of breast cancer can also be useful in delineating dysregulated pathways that define the basis of tumour phenotypes (3). Significant insight into tumour biology has come from selecting tumours with specific clinical identifiers, for example, by limiting mutation catalogs in metastatic tumours (10,11).

Somatic mutation analyses of tumour exomes have focused on alteration of amino acid sequences, or highly conserved dinucleotides adjacent to exons, which usually impact mRNA splicing. Since these variants most likely comprise only a fraction of the total mutational load, the pathways inferred to be dysregulated in these tumours may be incomplete. For example, in familial breast cancer, variants of unknown significance have been explained by both experimental validation and *in silico* predictions of defects in *BRCA 1/2* mRNA splicing (12,13). Typically, genomic studies have used tools that predict splicing mutations based on the highly conserved dinucleotide sequences at mRNA 5′ donor and 3′ acceptor sites (8,14). There are other well established methods that can identify splicing mutations beyond those directly at natural sites (15-17), but these approaches have not been applied to genome-scale cancer studies, until recently (18). Published studies have revealed only a small fraction of reported somatic mutations in cancer to be splicing mutations, accounting for only 2% of those reported (1-5). The present study considers the possibility that many somatic splicing mutations may be overlooked or are undetected by the conservative approaches currently used in analyses of tumour genomes.

Splicing mutations frequently lead to changes in the sequence and structure of the encoded protein, which are usually distinguishable from those generated by normal alternative splice isoforms. Constitutive splicing mutations are frequently deleterious and are a major cause of inherited and acquired diseases (19). In cancer, aberrant splicing (including alternative isoforms that are not a result of *cis* mutation) is known to cause or promote tumour propagation (20), and has been described as an additional hallmark of

the disease (21). RNA analyses can detect the effect of many splicing mutations directly (22,23). In this paper, we comprehensively analyze predicted splicing mutations in breast cancer tumours using DNA sequencing data from The Cancer Genome Atlas (TCGA) (5). We then use tumour-matched RNA sequencing data to statistically validate aberrant splicing patterns of expressed genes in these tumours that result from these mutations (24). We extended our splicing mutation analyses beyond molecular breast cancer subtypes and identified other clinical parameters associated with specific mutation pathways. We suggest that DNA sequencing analyses that incorporate in-depth splicing mutation studies reveal additional mutant genes and biochemical pathways, which may contribute to breast cancer etiology.

4.2 Methods

This study involved a reanalysis of controlled-access data from The Cancer Genome Atlas Project (NCBI dbGaP Project #988: Predicting common genetic variants that alter the splicing of human gene transcripts, PI: PK Rogan). DNA and RNA breast cancer sequencing data were obtained for 445 tumours from 442 patients (Supplementary Table 4.1; July, 2012 DNA-Seq download; July, 2013 RNA-Seq Download) (5). The tumour-normal pairs used mirrored those published by the TCGA in the Level 2 mutation data. Duplicate mutations in the same patient from two different tumour-normal pairs are reported, but were treated as one tumour for the mutation summaries reported by tumour. Somatic mutations were predicted from the same DNA sequencing data using two different algorithms: Strelka (v1.0.10) (6) and SomaticSniper (v1.0.2) (44) (See Appendix S4.1). Realignment was not necessary before running Strelka because of the program's internal realignment capabilities, so Strelka was run on the raw BAM files downloaded from TCGA. Default parameters were used with the provided Burrows-Wheeler Aligner (BWA) configuration file, since BWA was used in the initial exome alignments. Additionally, the isSkipDepthFilters configuration option was changed to true, since such depth filters are designed for use on whole-genome data and would erroneously filter out most data when used with exome sequencing data. Strelka's BWA quality control script was run to remove variants considered low quality. Variants that were found to be common SNPs, defined by those that were annotated with dbSNP135 in

over 1% of the population, were filtered out from the variant set before any subsequent analyses.

Somatic mutations, including single-nucleotide variants (SNVs) and insertion/deletions (indels) were used to predict the coding and non-coding genic effects of the variants. Annovar (August 23, 2013 release) (8) was used with default parameters to predict which variants are likely to affect amino acid sequence and splicing at the natural splice sites. The Shannon Human Splicing Pipeline Version 2.0 (Shannon Pipeline) (18) was used to complete a more in-depth analysis of splicing mutations, which predicts variants that will alter the binding affinity of the natural site or cause cryptic splicing (i.e. extension or truncation of an exon). The Shannon Pipeline results were subsequently filtered to prioritize which variants are most likely to have the greatest effect on mRNA splicing, using the filtering criteria outlined in Appendix S4.2.

Multiple factor analyses used the R package FactoMineR (version 1.25) (45). Clinical parameters were obtained from the TCGA including AJCC tumour staging (metastasis stage code, neoplasm disease lymph node stage, and neoplasm disease stage), receptor statuses (estrogen, progesterone, and *HER2*/neu immunohistochemistry receptor statuses) as well as patient status (neoplasm cancer status and vital status). These clinical parameters were input into FactoMineR as qualitative groups, as listed above, along with the number of *NCAM1* pathway mutations. Within the program, options were set to perform clustering after MFA, and to automatically determine the choice of the number of clusters. A second MFA was performed based on the number of *NCAM1* pathway mutations per tumour in genes present only in the *NCAM1* related pathways that were also not present in the collagen or extracellular matrix pathways.

Word Clouds were generated to portray the overrepresentation analysis of mutated pathway results generated with Reactome (29,30) and, in particular, the differences between lymph node-positive and -negative tumour samples. The primary input data for these graphics was the overrepresented pathways from Reactome, partitioned according to subtype and lymph status. Additional sets were composed of all subtypes and all subtypes with only pathways not found within both lymph status partitions. However, this

direct data was not suitable for plotting, as many pathways were vastly too specific and varied to portray any broader trends. Pathway abstraction was undertaken to mitigate these difficulties and allow for visual perception of trends in the data. The full Reactome human pathway hierarchy was downloaded, using the provided RESTful API (46). A query to abstract pathways was performed using the BaseX XML database engine (47). The abstraction was designed to generalize the pathways, while still maintaining sufficient specificity to confer biological meaning in this context. To accomplish this, corresponding pathways of specific depths were retrieved and abstracted by taking instead higher-order pathways in the hierarchy. Reactions or black box events that were four or five levels deep, as well as pathways that were four levels deep, were abstracted by taking the corresponding element of depth three (i.e. their parent or grandparent). Pathways one level higher in the hierarchy (i.e. the parent pathway) of all other pathways, reactions, or black box events (i.e. those not at the aforementioned depths) were retrieved. The resulting abstracted pathways were then used as input for the word clouds. They were generated using R (v3.0.2) with the RColorBrewer (v1.0.5 tm, and wordcloud packages (v2.4) (48). Parameters used to generate the word clouds were as follows: scale = c(wordFit,0.3), min.freq = 2, random.order = F, colors = brewer.pal(6, “Dark2”)[−1])), vfont = c(“serif”,“plain”).

The Mutational Significance in Cancer (v0.4) (MuSiC) (25) suite of tools was employed to identify genes significantly mutated in the breast cancer samples analyzed with the variant set derived in this study. Three tools from genome MuSiC were used with all default parameters: bmr calc-bmr, bmr calc-covg, and smg. NCBI Reference Sequence Genes release 62 (RefSeq) (49) were used as the regions of interest (ROI) file with the Human Feb. 2009 (GRCh37/hg19) assembly reference sequence for bmr calc-bmr and bmr calc-covg. All FDRs that we report pertaining to the MuSiC analysis used the Fisher's combined P-value (FCPT), convolution (CT) and likelihood ratio (LRT) statistical tests.

The software program Veridical (24) was used for *in silico* validation of all predicted splicing mutations using its default settings. At the time the program was run, Veridical rounded p-values to 2 decimal places. Validated results reported were filtered for cryptic

variants using reads demonstrating junction-spanning cryptic sites, junction-spanning exon skipping, or read-abundance intron inclusion, whereas reads for predicted natural splice site variants were filtered for all of the above evidence types, except for cryptic splice site-activating, junction-spanning reads. Variants were considered validated if at least one of the above categories for the indicated variant type were excluded from normal controls, but present in the transcriptome containing the predicted mutation ($p \leq 0.05$, after transformation of both sample and control read counts to a normal distribution and use of a parametric Z test). Validation was not always possible in instances where predicted mutations occurred in genes or exons with minimal cDNA coverage, resulting from either low expression in the breast tumours carrying the mutation (50), tissue-specificity of gene expression, or transcript instability from nonsense-mediated decay. Although Veridical provided experimental validation of predicted splicing mutations, the impact of these and protein coding mutations on tumour progression and biology could not be determined from the present analyses. Further laboratory studies with the original tumour tissues (which were not available), cell line or model organism studies would be required to prove biological significance.

RSeQC's (v2.3.7) ReadDist (51) script was used to generate the genome-wide intron inclusion data with the RefSeq gene annotation file to determine intronic genomic sequences. We ran BedTools multicov (v2.17.0) (52) upon the RefSeq (49) exome annotation BED file retrieved from the UCSC table browser (53) with a minimum map quality of 1. The returned coverage values were multiplied by the read length, and divided by the number of exonic bases. In cases of genes with more than one transcript, the shortest transcript was used such that the coverage values per exonic base were maximized, which is the most conservative assumption to adopt when excluding variants due to low coverage. The heat map, provided in Appendix S4.3, was generated by breast cancer subtype for this data using the R packages Hmisc (v3.14.3) and gplots (v2.12.1).

4.3 Results

4.3.1 Derivation of mutations

Somatic mutations in 472 breast cancer tumours from 445 breast cancer patients were called using matched tumour-normal DNA exome sequencing data from TCGA (5) (Supplementary Table 4.1). There were 149,959 single-nucleotide variants (SNVs) and 10,000 insertion/deletions (indels) detected using the variant caller, Strelka (6) (see Appendix S4.1 for results from an alternative variant caller and reasons for our selection of Strelka). Protein coding mutations were annotated by ANNOVAR (8) and splicing mutations with the Shannon Human Splicing Pipeline (18) (Table 4.1, see Supplementary Tables 4.2–4.4 for a list of all mutations). The Shannon Pipeline predicted significantly more splicing mutations than reported by TCGA, because the information-theoretic method employed enables analyses of variants beyond exon boundaries that alter mRNA splicing. 948 variants were found to affect both protein coding and splicing in 747 genes, among 319 tumours. *DYNC2H1*, *TP53* and *PASD* were the most commonly mutated of this group, containing 21, 11, and 9 exonic variants, respectively. Alteration of mRNA splicing was predicted as a result of 213 substitutions at synonymous codons among 139 tumours. Reanalysis of coding changes confirmed high concordance with the validated TCGA SNVs, however indels were less reproducible (Appendix S4.4). Overall, 82.1% (n = 21,041) of protein coding mutations, and 86.5% (n = 371) of splicing mutations reported by TCGA were confirmed. A small subset of protein coding TCGA substitutions that were missed occurred within genes commonly mutated in breast cancer (35 *TP53*, 13 *MLL3*, 22 *GATA3*, 25 *MAP3K1*, 11 *CDH1* and 10 *PIK3CA*; see Appendix S4.5), however all splicing-associated SNVs found by TCGA in cancer-related genes were detected.

4.3.2 Significantly mutated genes

Significantly mutated genes were identified with the Mutational Significance in Cancer (MuSiC) software suite (25). There were 225 genes with false discovery rates (FDR) of

Table 4.1: Single nucleotide variant summaries by mutation type

Type	Mutation Count
Protein Coding	
Synonymous	14,717
Nonsynonymous	40,649
Stop gain or loss	2,587
Total protein coding variants	**57,953**
Splicing	
Cryptic	1,130
Inactivating	1,355
Leaky	2,721
Total splicing variants	**5,206**
Protein coding mutations also predicted to affect splicing	
Synonymous	213
Nonsynonymous	664
Stop gain or loss	71
Total	**948**
Synonymous also splicing	1.4473%
Nonsynonymous also splicing	1.6335%
Stop gain or loss also splicing	2.7445%

<0.05, based on the Fisher's combined P-value (FCPT), convolution (CT) and likelihood ratio (LRT) tests. These results were compared with the 49 genes previously identified as significantly mutated (1-5) (Appendix S4.6). Among the previous genes reported by TCGA, *TP53, CDH1, MAP3K1*, and *MLL3* were significantly mutated in this study by all tests, and *AFF2, SF3B1, and CBFB* were significant for the CT and LRT tests only. We additionally identified *ARID1A* as significantly mutated, concordant with an independent, large-scale, breast cancer genomics study (4).

4.3.3 Validating predicted splicing mutations

Changes in mRNA splicing from the predicted mutations were validated with Veridical (24), which corroborates predicted, aberrant splice isoforms by assessing mutation-derived sequence reads in tumour RNA relative to their abundance in controls lacking the mutation. Controls comprised tumours lacking a particular mutation (usually, n = 414) plus additional normal samples (n = 106). Of all variants analyzed from the 415 tumours with RNA-Seq data (n = 4,952), 988 variants (~20%) in 819 genes caused one or more splicing aberrations at significantly higher levels than in controls ($p \leq 0.05$; i.e. intron inclusion, exon skipping, or cryptic splicing). Predicted natural splice site mutations (822 of 3,863, or 21.3%), were validated by abnormal mRNA isoforms more often than cryptic splice site mutations (166 of 1,089 or 15.2% variants). A total of 309 mutations were found to cause exon skipping, of which 163 (53%) led to expected frameshift mutations.

Sufficient expression levels for each gene, based on RNA-Seq coverage, were required for validation of mutations. An expression heat map, clustered by BC subtype, is shown in Appendix S4.3. Variants occurring within significantly expressed genes (defined as an average of ≥20 reads per base) were statistically validated for 862 (27%) of 3,156 variants ($p \leq 0.05$). Of 263 variants reported by TCGA in genes with at least this level of expression, 156 (59%) were validated by exon skipping (26 variants), by intron inclusion (80 variants), and by the combination of both types of evidence (50 variants, $p \leq 0.05$).

Predicted cryptic splicing mutations were confirmed based on the presence of unique junction-spanning reads corresponding the ectopically spliced isoforms in *GATA3*,

PALB2, *CBFB*, *ABL1*, *C2CD2L*, *ENSA*, *NASP*, *NOP9*, and *TFE3* (Appendix S4.7.1). Four of these genes have been linked to tumourigenesis: *ABL1*, an oncogene, *GATA3* and *PALB2*, which are associated with familial breast cancer (26,27), and *CBFB* has been recently implicated in breast cancer by TCGA (5) and others (1,2). These cryptic splicing mutations lead to short exonic deletions that alter the reading frame, and likely affect the activity of the gene products (Figure 4.1). The *GATA3* cryptic isoform is the only detectable transcript in the majority of controls, although it is substantially more abundant in the tumour sample (Appendix S4.7.2).

The most commonly mutated genes with splicing mutations were also found by MuSiC to be significantly mutated in these tumours ($n = 13$, FDR < 0.05), and at least one third of the mutations were validated with RNA-Seq data (Table 4.2). In *TP53*, which exhibited the highest density of splicing mutations (Figure 4.2), 18 of 23 (78%) predicted variants were validated to cause aberrant splicing ($p \leq 0.05$). All of the validated mutations exhibited statistically significant intron inclusion above normal controls, which was not observed genome wide (Appendix S4.8). In three instances, the variants also resulted in exon skipping.

4.3.4 Copy number analysis of mutated genes

The validated mutations are organized and segregated by tumour subtype on a Circos plot (28) (Figure 4.3). Copy number changes portray the genomic locations of deletions or amplifications that coincide with these variants. Validated splicing mutations exhibit a relatively uniform genomic distribution, except for significantly mutated genes, such as *TP53* on chromosome 17 and *HMCN1* on chromosome 1. We investigated variants in regions showing copy number losses, which may constitute the "second hit" in oncogenesis. Of the 49 genes found to be significantly mutated in breast cancer (1-5), five contained splicing mutations (*BRCA1* (2 tumours), *PTEN* (2 tumours), *MAP2K4* (4 tumours), *MAP3K1* (4 tumours) and *KMT2C* (7 tumours; also known as *MLL3*)) and also recurred within commonly deleted intervals. Of all genes with validated mutations in

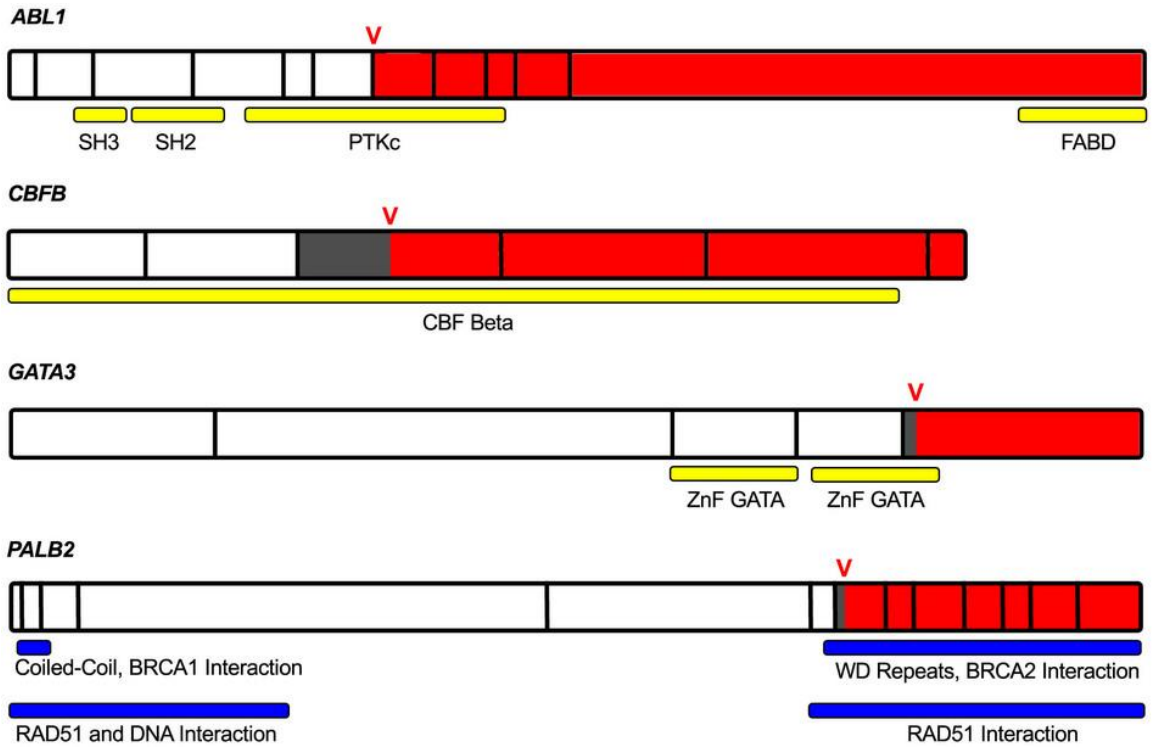


Figure 4.1 mRNA of *ABL1, CBFB, GATA3* and *PALB2*, which each have validated cryptic splicing mutations confirmed using tumour-matched RNA-Seq data. Full gene lengths are displayed with vertical black bars outlining exon boundaries. The location of the cryptic variant is denoted by the red V, and the variant consequence is highlighted by white (wild type), dark grey (exonic deletion), and red (frameshift mutation). Conserved domains and protein interactions are labeled by the yellow and blue horizontal bars, respectively. In *ABL1*, the catalytic and C-terminal F-actin binding domains are disrupted. In *PALB2*, the region that interacts with BRCA2 is truncated. In the *GATA3* aberrant transcript, the second zinc finger domain and a conserved motif crucial for DNA binding and protein function are affected by the altered reading frame.

Table 4.2: Genes most commonly mutated with splicing mutations

Gene Symbol*	# Splicing Mutations	# Validated	% Validated
TP53	24	18	75
HMCN1	19	9	47
KMT2C (MLL3)	19	7	37
FHAD1	12	4	33
RAB3GAP1	11	4	36
BCLAF1	11	3	27
ANKEF1	10	6	60
RRM1	8	4	50
RPRD1A	7	2	29
SCAMP5	7	2	29
CDH1	6	4	67
ACTR3	6	2	33

*FDR < 0.05 for all genes from MuSiC (Fisher's combined P-value, convolution and likelihood ratio tests).

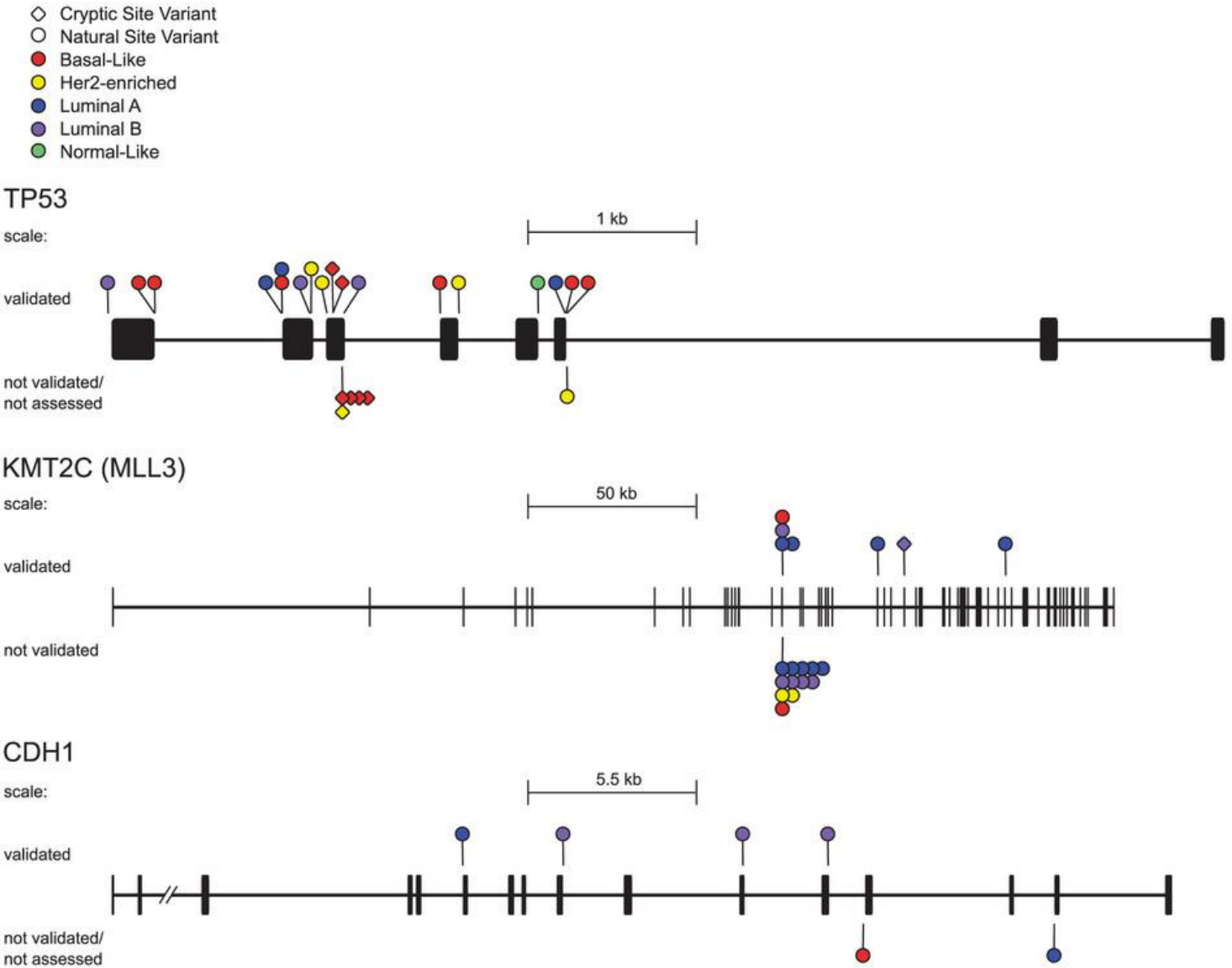

Cryptic Site Variant
Natural Site Variant
Basal-Like
Her2-enriched
Luminal A
Luminal B
Normal-Like
TP53
scale:
1 kb
validated
not validated/
not assessed
KMT2C (MLL3)
scale:
50 kb
validated
not validated
CDH1
scale:
5.5 kb
validated
not validated/
not assessed

Figure 4.2: Splicing mutations in *TP53*, *KMT2C* and *CDH1*. *TP53, KMT2C* and *CDH1* gene lengths are displayed with both exons (thick lines/boxes) and introns (thin horizontal lines), along with the location of all splicing mutations. Diamond markers denote cryptic mutations, natural splice site mutations are indicated by a circle and the colour of the marker corresponds with breast cancer tumour subtype. Mutations validated by Veridical are found above the gene, and those mutations not assessed or not validated are below.

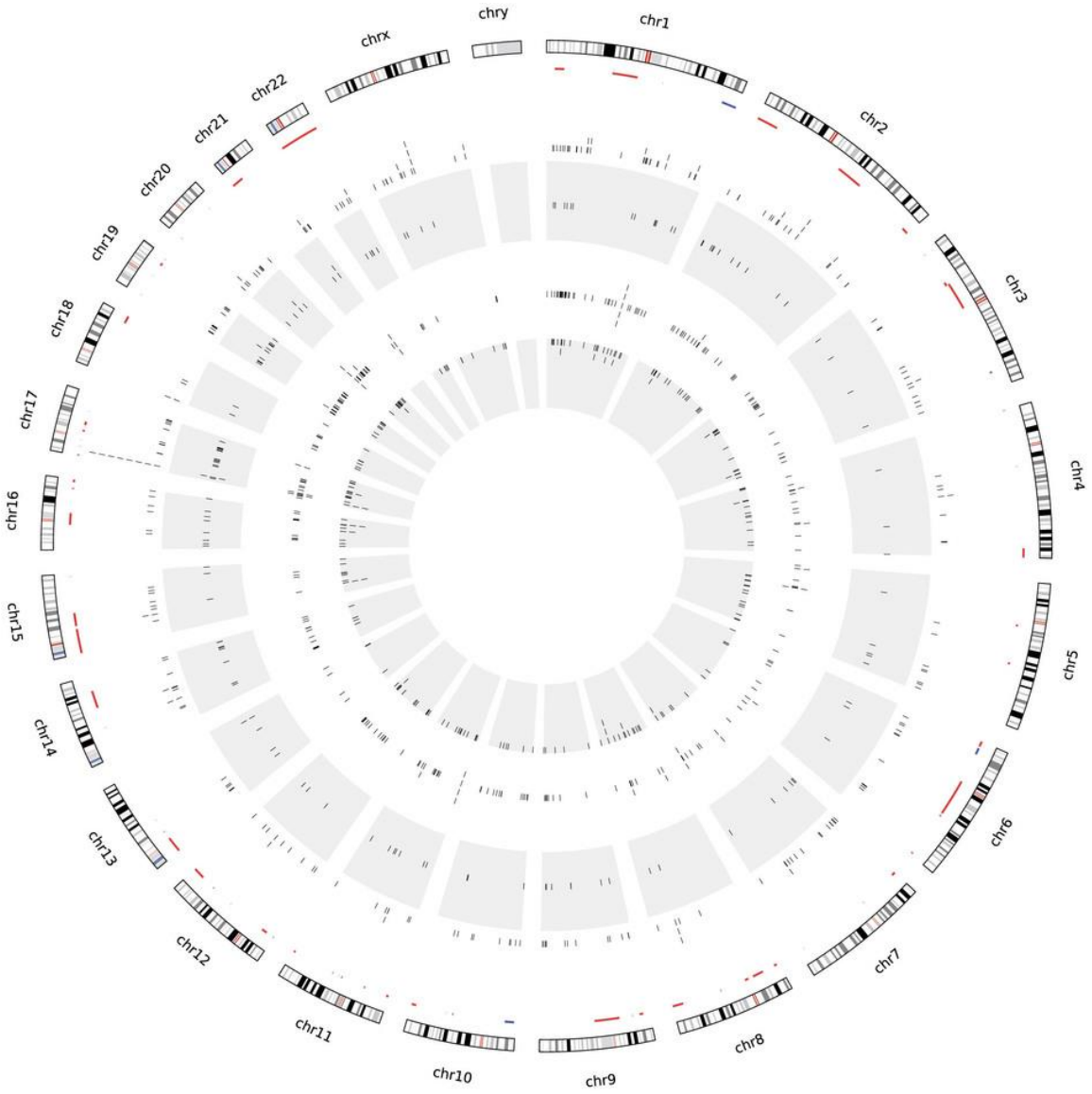


Figure 4.3: Circos plot of validated splicing mutations by tumour subtype. From the outermost ring in, chromosomes are labeled clockwise with copy number data inside them that displays deletions in red and amplifications in blue, mutations validated by Veridical (indicated by black ticks) are then plotted by subtype with basal-like in the outer white ring, *HER2*-enriched in the outer grey ring, then luminal A (inner white) and luminal B (inner grey).

deleted regions, 9 harbored more than 2 variants: 1 had three, 4 had four, and only *KMT2C* possessed more than 4 variants.

4.3.5 Analysis of pathways enriched in mutant genes

Mutated genes were clustered by pathway overrepresentation analysis (29) for protein coding (Supplementary Table 4.5, n = 202) and splicing mutations (Supplementary Table 4.6, n = 452). There were 100 pathways common to both mutation sets (Appendix S4.9.1). Pathways associated with all types of mRNA splicing mutations include those that affect collagen structural genes and enzymes that modify or metabolize collagen (n = 14, Appendix S4.9.2 #1–14), and several that involve the extracellular matrix (ECM, n = 4, Appendix S4.9.2 #15–18). Many of these pathways (n = 17, Appendix S4.9.2 #1–13,15–18) are also overrepresented by pathway analysis of protein coding mutations.

4.3.6 Relationship of mutation spectra to clinical findings

Segregating splicing mutations by patient lymph node status revealed significant differences in mutated pathways between the two groups. Biochemical pathways with overrepresented mutant genes in lymph node-negative (LN−) vs. lymph node-positive (LN+) tumours are indicated in Supplementary Tables 4.7 and 4.8, and compared in Supplementary Table 4.9. There are 94 pathways overrepresented in both LN+ and LN− (Supplementary Table 4.9 #421–514), including 17 collagen (Supplementary Table 4.9 #421–437), and 3 ECM (Supplementary Table 4.9 #438–440) pathways. Ontologically-related pathways (29,30) were grouped (Appendix S4.9.3) and visualized as Word Clouds (Figure 4.4). Pathway groups overrepresented ($p < 0.05$) in both tumour subsets included 17 pathways involving collagen-ECM protein phosphorylation pathways, metabolism, cell cycle, DNA repair, and cellular response to stress. However, 13 pathways involving collagen (Supplementary Table 4.9 #1–13), and 9 pathways involving *NCAM1* (Supplementary Table 4.9 #17–25) were overrepresented uniquely in LN+ tumors, but not in LN− tumours.

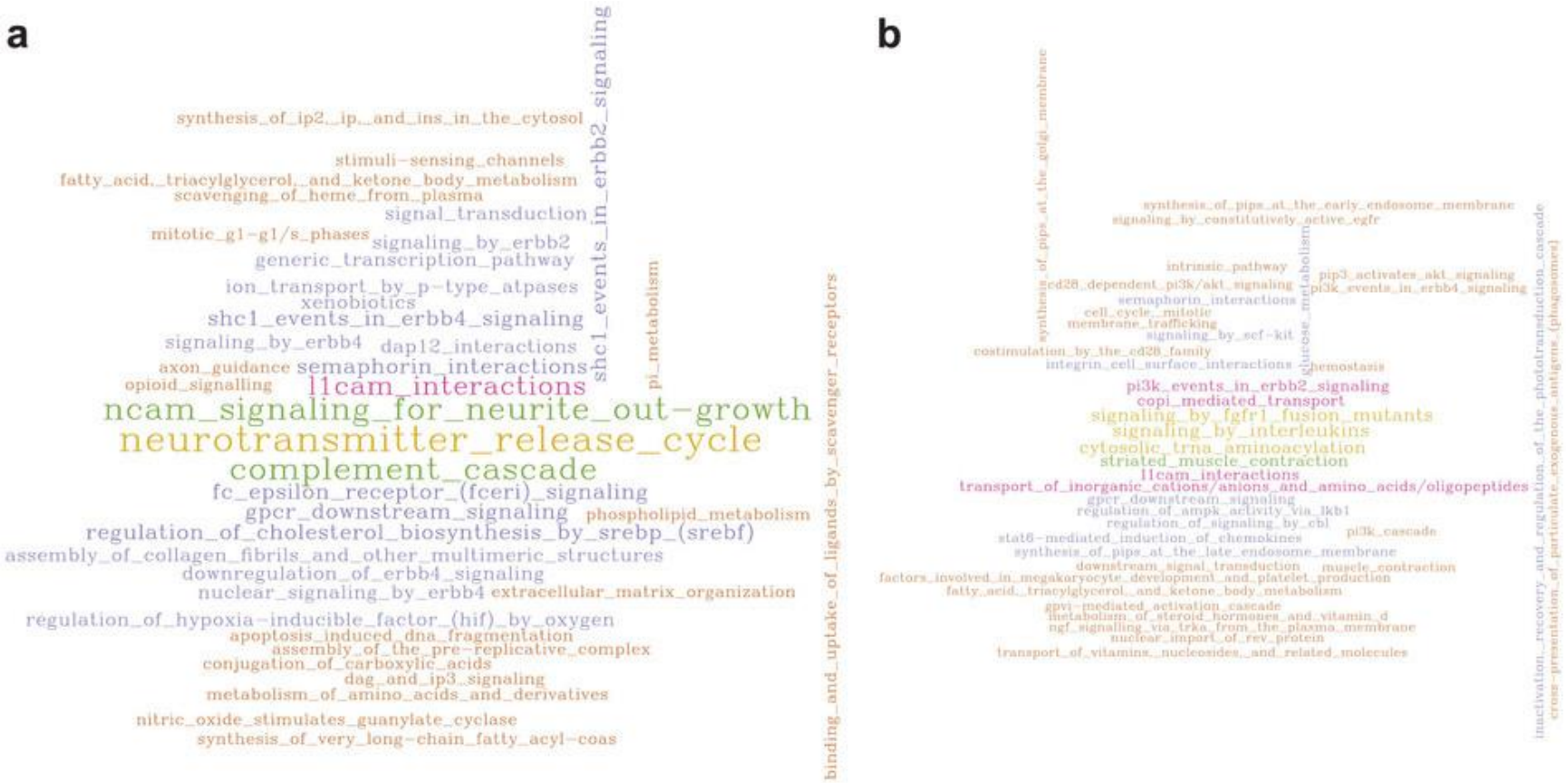


Figure 4.4: Word Clouds demonstrating differences between overrepresented mutated pathways in lymph node-positive (a) and lymph node-negative (b) tumours. The abstracted pathways (see methods) were plotted if present two or more times. The size of the words as well as the corresponding colours of the pathway names indicates the frequency of that abstracted pathway, and can be compared within and between the word clouds of each tumour subset.

NCAM1, or the neural cell adhesion molecule, is a member of the immunoglobulin super family with a role in cell-cell and cell-matrix interactions during development and cellular differentiation. Mutations in NCAM1 signaling genes for neurite outgrowth (Supplementary Table 4.10 #1) were still overrepresented in tumours with lymph node invasiveness, even after genes common to both tumour subsets were masked from the analysis, i.e. primarily collagen and ECM genes (Supplementary Tables 4.10 and 4.11). These include defects in NCAM1 interactions with FYN and GRB2, a ternary complex that participates in the conversion of RAS:GDP to RAS:GTP, which subsequently initiates the RAF/MAP kinase cascade.

We then reanalyzed these data after conservatively limiting the set of mutant genes to those containing the most deleterious mutations (Appendix S4.9.4; stop-gain, stop-loss, frameshift/indel mutations, and validated splicing mutations). Four of the 8 sub-pathways of NCAM1 signaling for neurite outgrowth were overrepresented solely in LN+ tumours. Autophosphorylation/dephosphorylation of NCAM1- bound Fyn, as well as NCAM1-interactions with collagens were overrepresented. The most commonly mutated genes within these pathways are *SPTA1*, *CACNA1D*, *COL6A5*, *NCAM1*, and *COL6A6* (Appendix S4.10). CACNA1D is a voltage-dependent Ca2+ channel (VDCC) that associates with NCAM1 in growth cones at the sites of NCAM1 clustering (29,30). In addition, 6 other channel genes that are expressed in breast tissue (31) were found to be frequently mutated (CACNA1C, CACNA1D, CACNA1G, CACNA1H, CACNB1, CACNB3). Mutations interrupting these VDCC interactions may alter the NCAM-dependent Ca2+ influx. Collagen VI is expressed as supramolecular aggregates of composite structures of different chains and is among the most abundant components of the ECM (32). Knockdown of NCAM significantly reduces expression of ECM components (33), including collagen, weakening the ECM. Mutations in these ECM components may also diminish matrix integrity, possibly resulting in more porous structures (34).

4.3.7 Elevation of NCAM1-related gene pathway mutations in lymph node-positive tumours

NCAM1, collagen, and ECM pathway mutations were assessed in tumours, stratifying by lymph node status and tumour stage (Figure 4.5). The percentage of tumours with NCAM1-related pathway splicing mutations was increased in N0 (110 localized tumours) and N1 (84 tumours with lymph node involvement), as well as Stage I (37) and II tumours (140). Advanced lymph node involvement and tumour stage were not associated with increased numbers of collagen and ECM pathway splicing mutations, but rather a decrease in the percent of tumours with these pathway mutations in advanced stages was observed. A multiple factor analysis (MFA; Table 4.3) was performed to assess contributions of the number of NCAM1-related pathway mutations per tumour (both protein coding and splicing), clinical parameters including stage (AJCC tumour stage, lymph node status and metastasis stage), receptor status (HER2, PR, and ER positivity), and patient outcome (relapsed, living/deceased). NCAM1-related pathway mutations were either absent (n = 213), harbored a single mutation (n = 117), or two or more mutations (n = 112) per tumour. The MFA components containing NCAM1-related pathway mutations were moderately correlated with both tumour stage and receptor status, and accounted for 11% of the variance.

4.3.8 Analysis of tumour subtypes

Splicing mutation analysis in different tumour subtypes revealed between 9–15 mutations per tumour, which generally accounted for 8–9% of all mutations detected (Appendix S4.11.1) and are similar levels to those previously reported (18). Pathway analyses for each subtype, stratified by lymph node status, indicated higher enrichment of NCAM1-related gene mutations in basal-like and *HER2*/*ERBB2*-enriched LN+ tumours (Appendix S4.11.2 & S4.11.3: see word clouds). LN+ basal-like and *HER2*-enriched tumours were the only tumours found to have significant enrichment in “NCAM signaling for neurite out-growth”, identifying those tumour subtypes and pathways that may play a role in tumour migration. No single gene was significantly mutated within the NCAM1

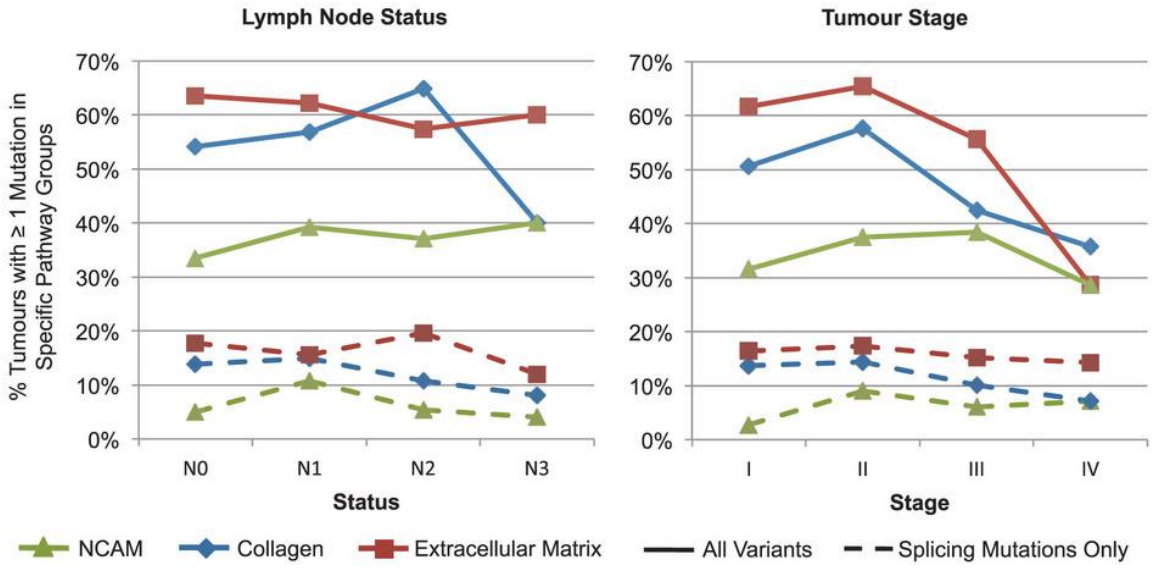


Figure 4.5: Percent of tumours with mutations by pathway group and clinical factors. The percent of tumours with NCAM1 (red square), collagen (blue diamond), and ECM (green triangle) pathway mutations were plotted by lymph node status and tumour stage for all mutations (solid lines), and splicing mutations alone (dashed line).

Table 4.3: Multiple factor analysis of *NCAM1* related pathway mutations and clinical parameters per tumour

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
A. No. Mutations in NCAM Pathways*	0.103	0.892	0.910	0.367	0.321
Stages	0.804	0.459	0.381	0.833	0.725
Receptor status	0.379	0.356	0.406	0.471	0.641
Patient status	0.868	0.159	0.050	0.106	0.159
% Variance explained	*7.618*	*5.699*	*5.635*	*4.944*	*4.694*
B. No. Mutations Unique to NCAM Pathways∧	0.264	0.899	0.894	0.304	0.300
Stage	0.791	0.413	0.380	0.877	0.752
Receptor status	0.389	0.427	0.411	0.429	0.610
Patient status	0.851	0.083	0.158	0.168	0.221
% Variance explained	*7.716*	*5.816*	*5.534*	*4.941*	*4.743*

*mutation count for all genes in NCAM pathways.

∧mutation count for genes unique to NCAM pathways, and not in collagen or ECM pathways.

pathways that were overrepresented in LN+ tumours. This suggests that a general defect in NCAM1-pathway signaling may be associated with lymph node metastasis in breast cancer.

4.4 Discussion

Breast carcinoma tumour exomes contain more deleterious mutations than previously recognized. Using Shannon information theory, we have predicted an expanded set of mutations that affect post-transcriptional mRNA processing that either reside in non-coding regions, or overlap known codons. We then employed Veridical (24), a high-throughput, genome-scale method, to statistically validate mRNA splicing consequences that result from the predicted variants. This study complements the analyses performed by TCGA (5), which comprehensively reported protein-coding mutations, along with gene expression, epigenetic, and copy number changes. Together with known deleterious coding sequence variants, the identification of such splicing mutations can refine and impact our understanding as to which biochemical pathways are dysregulated in these tumours.

Pathway overrepresentation analyses reproduced many of the same pathways identified by TCGA. In our analysis, a number of these attained or increased significance when genes with previously unrecognized splicing mutations were included. Both splicing mutations alone and the complete variant set from all tumours were enriched for genes in pathways known to play a role in tumour development and progression including signaling by growth factors, cell cycle, ECM organization, and cell-to-cell communication. Stratifying the tumours by lymph node status revealed that splicing mutations were enriched for genes within NCAM1 pathways in LN+ tumours, exclusively. Splicing mutations in these pathways were much rarer and sparsely distributed in LN− tumours, with 11 mutations in 92 LN− tumours and 25 mutations in 118 LN+ tumours. Interestingly, this enrichment was not observed when all protein coding substitutions were analyzed, but was significant when assessing all variants that were likely to be deleterious (i.e. validated splicing mutations, stop codon gain or losses

and frameshift substitutions). We did not attempt to differentiate loss versus gain of function, however splicing mutations and nonsense codons usually result in loss of function. The percent of tumours with NCAM1-related pathway mutations increased by 6% from lymph node stage N0 to N1 and N3 and by 7% from stage I to III. The lower fraction of tumours with collagen pathway mutations at higher lymph node stages (N3, N4), and with ECM-related mutations in tumour stages III and IV could be related to clonal selection of distinct metastatic phenotypes (35), however it is also possible that the decreases may not be significant due to the lower numbers of tumours in these categories.

Our results indicate that NCAM1 pathways are more likely to be dysregulated in tumours that have migrated to lymph nodes. We found the enrichment of NCAM1-related pathway splicing mutations in LN+ tumours was specifically present in *HER2*-enriched and basal-like tumours. Basal-like, specifically triple-negative, tumours have been associated with poor prognosis and survival (36). Early and metastatic *HER2* positive tumours were associated with poor prognoses (37) until the more recent introduction of *HER2*-targeted therapies (38). In these tumour subtypes, the presence of NCAM1-related pathway mutations may indicate a propensity to migrate and/or form distant metastases.

Dysregulated expression of NCAM1 has been suggested to contribute to tumour migration in other cancers: (i) gene silencing and localization studies have suggested that "NCAM is both necessary and sufficient to promote a migratory and invasive phenotype in EOC cells, with no major effect on cell proliferation" (34), (ii) overexpression of *NCAM1* has been linked to high ovarian carcinoma tumour grade (34) and greater metastatic potential in melanoma cells (39); (iii) preserved *NCAM1* expression in differentiated thyroid carcinoma has been cited as an indicator for tumours with as increased risk of forming distant metastases (40) and (iv) blocking NCAM1 function in murine lung tumour cells led to cell vulnerability to apoptosis. More generally, NCAM1 is known to play a role in apoptotic evasion and matrix degradation, and has potential roles in directional cell migration, cell polarity, extravasation and immunological escape (41). NCAM1-mediated stimulation of FGFR activity is causally linked to tumour malignancy, suggesting that this NCAM1-FGFR interaction may be an effective therapeutic target. It is notable that we find mutations in breast tumours that affect the

NCAM1-FGFR interaction occur in pathways that are overrepresented in LN+, but not LN− tumour genomes.

NCAM1 homophilic clusters form within lipid rafts on the cell membrane. Spectrin, an NCAM1-binding cytoskeletal protein, colocalizes with NCAM1 and is codistributed within lipid rafts (42). Frequent mutations in spectrin (*SPTA1*) may prevent its association with RPTPα, thereby impeding its subsequent association with the cytoplasmic NCAM1 domain, redistribution of NCAM1 and cluster formation. This could abrogate downstream interactions with FYN and GRB2, ultimately affecting activation of RAS. These findings merit further investigation into how dysregulation in these different partners (i.e. NCAM1, FGFR and the other interacting proteins), acting as an ensemble, may promote tumour metastasis.

The number of aberrant mRNA splicing mutations reported by TCGA (5) is <10% of those reported here, and the variants were not functionally validated in the previous study. We predict that 8% of all *cis*-activating point mutations detected in these tumours will significantly reduce the strength of the corresponding natural splice sites. The 5,206 splicing mutations reported here nearly double the number of mutations that lead to stop-gains or losses (2,587 variants in 1,907 genes), and the number of insertions/deletions leading to frameshift substitutions (2,707 variants in 1,848 genes) in this set of tumours. It is not surprising that these analyses revealed previously unrecognized pathways that may be dysregulated, in addition to those already known in these tumours.

Our analysis of significantly mutated genes based on the protein coding and splicing mutations reproduced many of the genes reported by TCGA, and revealed one additional gene, *ARID1A. ARID1A* has been implicated in breast cancer in a large-scale genomic study (4) and has also been mutated in 57% of ovarian clear-cell carcinoma tumours (43). Thirteen genes identified as significantly mutated in breast cancer by the TCGA did not reach statistical significance within our study (Supplementary Table 4.4). This can be explained by a number of different factors: differences in variant callers, variant annotation, the number of tumours analyzed and differences in the filtering of variants, once the gene set was derived. In addition, TCGA initially analyzed all variants (SNVs

and indels) by tumour subtype, unlike our study, which considered mutations in all tumours, then reanalyzed overrepresented pathways with mutations by subtype. Mutations that lead to a significant level of aberrant splicing can alter or improve genomic signatures, which are important when assessing potential biomarkers, diagnosis and prognosis, and metastatic or treatment-resistant tumour phenotypes.

4.5 References

1. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).

2. Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353–360 (2012).

3. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).

4. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).

5. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

6. Saunders, C. T. *et al.* Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).

7. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).

8. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

9. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **7**, 7.20.1-41; doi:10.1002/0471142905.hg0720s76 (2013).

10. Liu, X., Wang, J. & Chen, L. Whole-exome sequencing reveals recurrent somatic mutation networks in cancer. *Cancer Lett.* **340**, 270–276 (2013).

11. Ali, M. A. & Sjöblom, T. Molecular pathways in tumor progression: From discovery to functional understanding. *Mol. BioSyst.* **5**, 902–908 (2009).

12. Mucaki, E. J., Ainsworth, P. & Rogan, P. K. Comprehensive prediction of mRNA splicing effects of BRCA1 and BRCA2 variants. *Hum. Mutat.* **32**, 735–742 (2011).

13. Menéndez, M. *et al.* Assessing the RNA effect of 26 DNA variants in the BRCA1 and BRCA2 genes. *Breast Cancer Res. Treat.* **132**, 979–992 (2012).

14. Krawczak, M., Reiss, J. & Cooper, D. N. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: Causes and consequences. *Hum. Genet.* **90**, 41–54 (1992).

15. Mucaki, E. J., Shirley, B. C. & Rogan, P. K. Prediction of Mutant mRNA Splice Isoforms by Information Theory-Based Exon Definition. *Hum. Mutat.* **34**, 557–565 (2013).

16. Pertea, M., Lin, X. & Salzberg, S. L. GeneSplicer: A new computational method for splice site prediction. *Nucleic Acids Res.* **29**, 1185–1190 (2001).

17. Churbanov, A., Vorechovský, I. & Hicks, C. A method of predicting changes in human gene splicing induced by genetic variants in context of cis-acting elements. *BMC Bioinformatics* **11**, 1–12; doi:10.1186/1471-2105-11-22 (2010).

18. Shirley, B. C. *et al.* Interpretation, Stratification and Evidence for Sequence Variants Affecting mRNA Splicing in Complete Human Genome Sequences. *Genomics Proteomics Bioinformatics* **11**, 77–85 (2013).

19. López-Bigas, N., Audit, B., Ouzounis, C., Parra, G. & Guigó, R. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.* **579**, 1900–1903 (2005).

20. Venables, J. P. Aberrant and alternative splicing in cancer. *Cancer Res.* **64**, 7647–7654 (2004).

21. Ladomery, M. Aberrant alternative splicing is another hallmark of cancer. *Int. J. Cell Biol.* **2013**, 463786; doi:10.1155/2013/463786 (2013).

22. Coulombe-Huntington, J., Lam, K. C. L., Dias, C. & Majewski, J. Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS Genet.* **5**(12), e1000766 (2009).

23. Hatakeyama, K. *et al.* Identification of a novel protein isoform derived from cancer-related splicing variants using combined analysis of transcriptome and proteome. *Proteomics* **11**, 2275–2282 (2011).

24. Viner, C., Dorman, S. N., Shirley, B. C. & Rogan, P. K. Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. *F1000Res.* **3**, 8; doi:10.12688/f1000research.3-8.v2 (2014).

25. Dees, N. D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–1598 (2012).

26. Arnold, J. M. *et al.* Frequent somatic mutations of GATA3 in non-BRCA1/BRCA2 familial breast tumors, but not in BRCA1-, BRCA2- or sporadic breast tumors. *Breast Cancer Res. Treat.* **119**, 491–496 (2010).

27. Rahman, N. *et al.* PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.* **39**, 165–167 (2007).

28. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

29. Croft, D. *et al.* Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).

30. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42**, D472–D477 (2014).

31. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).

32. Fitzgerald, J., Holden, P. & Hansen, U. The expanded collagen VI family: New chains and new questions. *Connect. Tissue Res.* **54**, 345–350 (2013).

33. Håkansson, J. *et al.* Neural cell adhesion molecule-deficient ß-cell tumorigenesis results in diminished extracellular matrix molecule expression and tumour cell-matrix adhesion. *Tumor Biol.* **26**, 103–112 (2005).

34. Zecchini, S. *et al.* The adhesion molecule NCAM promotes ovarian cancer progression via FGFR signalling. *EMBO Mol. Med.* **3**, 480–494 (2011).

35. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

36. Shastry, M. & Yardley, D. A. Updates in the treatment of basal/triple-negative breast cancer. *Curr. Opin. Obstet. Gynecol.* **25**, 40–48 (2013).

37. Slamon, D. J. *et al.* Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235**, 177–182 (1987).

38. Jelovac, D. & Emens, L. A. HER2-directed therapy for metastatic breast cancer. *Oncology (Huntington, N. Y.)* **27**, 166–175 (2013).

39. Osborne, J. K. *et al.* NeuroD1 regulation of migration accompanies the differential sensitivity of neuroendocrine carcinomas to TrkB inhibition. *Oncogenesis* **2**, e63 (2013).

40. Yang, A. H., Chen, J. Y., Lee, C. H. & Chen, J. Y. Expression of NCAM and OCIAD1 in well-differentiated thyroid carcinoma: Correlation with the risk of distant metastasis. *J. Clin. Pathol.* **65**, 206–212 (2012).

41. Wai Wong, C., Dye, D. E. & Coombe, D. R. The role of immunoglobulin superfamily cell adhesion molecules in cancer metastasis. *Int. J. Cell Biol.* **2012**, 340296; doi:10.1155/2012/340296 (2012).

42. Leshchyns'ka, I., Sytnyk, V., Morrow, J. S. & Schachner, M. Neural cell adhesion molecule (NCAM) association with PKCβ2 via β1 spectrin is implicated in NCAM-mediated neurite outgrowth. *J. Cell Biol.* **161**, 625–639 (2003).

43. Jones, S. *et al.* Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science* **330**, 228–231 (2010).

44. Larson, D. E. *et al.* Somaticsniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).

45. Lê, S., Josse, J. & Husson, F. FactoMineR: An R package for multivariate analysis. *J. Stat. Software* **25**, 1–18 (2008).

46. Milacic, M. *et al.* Annotating cancer variants and anti-cancer therapeutics in Reactome. *Cancers* **4**, 1180–1211 (2012).

47. Grün, C., Gath, S., Holupirek, A. & Scholl, M. H. XQuery full text implementation in BaseX. *Lect. Notes Comput. Sci.* **569**, 114–128 (2009).

48. Feinerer, I., Hornik, K. & Meyer, D. Text Mining Infrastructure in R. *J. Stat. Software* **25**, 1–54 (2008).

49. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–5 (2007).

50. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

51. Wang, L., Wang, S. & Li, W. RSeQC: Quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).

52. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

53. Karolchik, D. *et al.* The UCSC table browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).

Chapter 5

5 Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning

The work presented in this chapter is reproduced (with permission, Appendix S1) from:

Dorman, S.N., Baranova, K., Knoll, J.H.M., Urquhart, B.L., Mariani, G., Carcangiu, M.L., Rogan, P.K. (2015) Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning. *Molecular Oncology*. DOI: 10.1016/j.molonc.2015.07.006.

5.1 Introduction

Chemotherapeutic agents, such as paclitaxel and gemcitabine, are recommended to patients with developed metastases, basal-like breast cancer, and high-risk indications (premenopausal, ER/PR-negative, HER2-status, large tumours, or node-positive) (1,2). There is currently no gold standard chemotherapy regimen (1,2). Treatment selection is suggested to be individualized and should take into account clinical disease characteristics, treatment history, patient-related factors, and patient preference. However, resistance is one of the major barriers to successful therapy. In a recent study, breast cancer patient response rates to paclitaxel and gemcitabine after 6 cycles of chemotherapy were found to be only 50.0% and 78.6% respectively (3). This has motivated a number of groups to develop gene signatures aimed at predicting therapeutic response to these drugs in breast cancer patients (4-6).

As in breast cancer patients, breast cancer cell lines show variable responses to growth inhibition by paclitaxel and gemcitabine (7,8). Cell lines mirror many of the pathological features of breast tumours, such as the intrinsic subtypes of breast cancer (9,10), and can be useful for testing anticancer therapy responses (11). Daemen et al. (2013) employed random forest machine learning to assess genomic information from 70 breast cancer cell lines (including DNA sequence, gene copy number, gene expression, promoter

methylation, protein expression, and the corresponding cell line response to 90 anti-cancer compounds) with the objective of establishing pretreatment signatures that predict response. The gene expression profile of the tumor subtype was found to be the most effective way to model response to therapy. However, many molecular signatures derived using genome-wide approaches are inconsistent between different data sets (12,13). This is partly due to the fact that deriving predictive gene models using thousands of genes risks overtraining, that is, fitting the noise rather than the actual gene signature in the data (12).

We recently defined a set of genes that are stable in gene expression and copy number in the majority (>90%) of breast cancer tumours (14). The stable gene set is composed of genes that are unmutated in the majority of tumours. Interestingly, many stable gene products were found to be targets of paclitaxel and gemcitabine. We examine the possibility that genomic differences in expression, copy number or mutation in these genes may be related to GI50. Rather than a genome-wide approach to predict sensitivity to paclitaxel and gemcitabine (eg. employed by Daemen et al. (2013)), we analyze stable and linked unstable genes in pathways that determine their disposition (Figure 5.1).

Gene panels were established based on biological and experimental studies of paclitaxel and gemcitabine metabolism. Paclitaxel binds to the β subunit of tubulin (*TUBB1*), inhibiting microtubule formation during mitosis (15). It also binds *BCL2,* which induces programmed cell death (16). Paclitaxel is now also recognized to target microtubule-associated proteins 2 (*MAP2*), 4 (*MAP4*) and Tau (*MAPT*) (17), as well as the xenobiotic receptor (*NRI12*, or PXR) (18). *SLCO1B3* transports paclitaxel into cells, and it is exported by *ABCB1* (P-glycoprotein), multidrug resistance-associated proteins *ABCC1* (19) and *ABCC10* (20), and the bile salt export pump *ABCB11* (21). Other genes previously implicated as contributing to paclitaxel resistance include *TMEM243* (22), *BCAP29* (23), *GBP1* (24), *TLR6* (25), *NFKB2* (26), *FGF2* (27), *BIRC5* (28), *TWIST1* (29), *FN1* (30), *OPRK1* (31), *CSAG2* (32), and *CNGA3* (31). Additionally, genes expressed in breast tissue involved in paclitaxel metabolism were included: *CYP2C8* and *CYP3A4* (33), as well as stable genes in pathways of known direct targets (14): *BAD*, *BBC3*, *BCL2L1*, *BMF*, *TUBB4A* (34), and *TUBB4B* (34).

Gemcitabine, a deoxycytidine analog, is transported into the cell by *SLC29A1* (35), *SLC29A2*, *SLC28A1* (36), and *SLC28A3* (37).The prodrug is then phosphorylated by *DCK*, *CMPK1*, and *NME1* to gemcitabine diphosphate and triphosphate (38). These active forms are incorporated into DNA, which halts replication and cell growth (39). Gemcitabine di- and triphosphate target ribonucleotide reductase (*RRM1*, *RRM2*, and *RRM2B*), and inhibit DNA synthesis (40). An alternative metabolite, difluorodeoxyuridine monophosphate, which is derived by cytidine deaminase (*CDA*) or dCMP deaminase (*DCTD*), inhibits thymidylate synthetase (*TYMS*), resulting in apoptosis (38).

We examine the hypothesis that genomic differences in genotypes, expression and copy number of these genes explain concentration-dependent growth inhibition by gemcitabine and paclitaxel. We then use machine learning to stratify the relative contributions of different genes to chemoresistance, by identifying corresponding genomic signatures at the transcriptional and genomic level in both cell line and patient data.

5.2 Materials and Methods

5.2.1 Data Acquisition

Growth inhibition (GI50), copy number, gene expression, and exome sequencing data were obtained from the supplementary data of Daemen et al. (2013). GI50s ($-\log_{10}M$, where M is the drug concentration required to inhibit cell line growth by 50%) for paclitaxel were available for 49 cell lines and GI50s for gemcitabine were available for 47 cell lines. Appendix S5.1 indicates the cell lines used and Appendix S5.2 indicates the gene, gene product names and their respective drug disposition functions. Appendix S5.3 & S5.4 describe copy number and variant calling, results of which are shown in Supplementary Tables 5.1 and 5.2. Log_2 normalized gene expression data were derived from Affymetrix Gene Chip Human Exon 1.0 ST arrays. Replication studies performed to re-measure and confirm GI50s, verify copy number and mutation data for a subset of the cell lines are outlined in Appendix S5.5. Figure 5.1 is an overview of the complete workflow used.

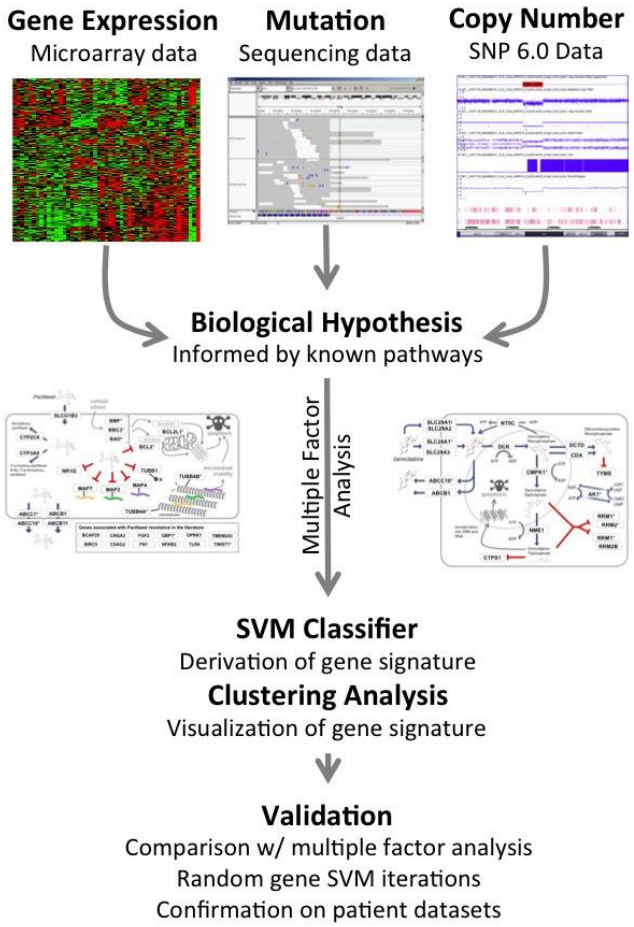


Figure 5.1 Workflow to derive gene signatures. Gene sets were derived for paclitaxel and gemcitabine based on known drug pathways, metabolism, and genes previously implicated in resistance. A multiple factor analysis was completed for each gene to determine which data types (gene expression, copy number, and mutation data) were correlated with the growth inhibitory values for paclitaxel and gemcitabine. Gene expression values were used to derive the paclitaxel SVM classifier, and both gene expression and copy number were used for the gemcitabine SVM. Cell lines were then clustered on optimized gene sets to visualize stratification of tumour subtype and sensitivity. The SVM classifiers were validated using random gene iterations to determine the significance of the classification accuracy, and patient data sets to ensure robustness of the models derived.

5.2.2 Cell Lines

Cell lines were composed of 10 basal, 9 claudin-low, 25 luminal, and 5 normal-like subtypes. Cell lines were designated resistant, if their GI50 was <8.0 for paclitaxel and <7.0 for gemcitabine, respectively. The threshold values for distinguishing sensitive from resistant cell lines were based on median GI50s for each particular drug (7.99 and 7.13, for paclitaxel and gemcitabine). Daemen et al. (2013) classified cell lines by comparing mean GI50s. We used median GI50, which is not impacted to the same extent by outlier cell lines.

5.2.3 Multiple Factor Analysis (MFA)

MFA was used to relate each cell line GI50 according to sets of genomic variables (41). The 44 (gemcitabine) or 45 (paclitaxel) breast cancer cell lines (Appendix S5.1) were treated as separate individuals. MFA was carried out with the R library "FactoMineR" (42), with GI50s, gene expression, copy number, mutation status (if the gene contained 1 or more mutations), and 31 and 18 genes associated with paclitaxel and gemcitabine activity, respectively, as input.

5.2.4 Support Vector Classification

A binary support vector machine (SVM) was trained with the Statistics Toolbox in MATLAB (Natick, MA) using `fitcsvm (linear kernel function)` and then tested with a leave-one-out cross-validation (using '`crossval'` and '`leaveout`' options). The SVM was trained on the cell lines and explanatory gene variables deemed relevant from the MFA: expression data for the paclitaxel SVM, and copy number and expression data for the gemcitabine SVM. The input data consisted of measurements from all genes used in the MFA. Sequential backward feature selection was performed for feature optimization (43) to minimize the percentage of misclassified cell lines (classification error) returned from the leave-one-out cross validation (Appendix S5.6). Genes that did not reduce or change the classification error were removed from the SVM (one at a time). This procedure was iterated until further gene removal lead to a higher

classification error (stopping criterion). By contrast with the SVM, a partial-least squares regression was not effective in relating genomic findings to paclitaxel response (Appendix S5.7).

The hinge loss was also determined for the subset of genes included in the final SVMs. Hinge loss applies a linear penalty for misclassified data according to their distance from the hyperplane. The loss function is represented by Equation 1 where y_j = {-1,1} and $f(X_j)$ is the score, i.e. hyperplane distance, for cell line j:

$$L = \max\left(0,1 - y_j f(X_j)\right) \quad (1)$$

5.2.5 Applying the cell line SVM to patient data

Formalin fixed, paraffin embedded (FFPE) tumour samples were obtained from the Fondazione IRCCS Istituto Nazionale dei Tumori (Milan, Italy), from leftover material available after diagnostic procedures in consented patients (44). Samples obtained were from patients that were first treated with paclitaxel (or in a small number of cases docetaxel) and carboplatin, and then subsequently gemcitabine, upon development of resistance. Clinical information was available as to whether the patients responded to each of the drugs (paclitaxel and gemcitabine). Tumour and control normal tissues were analyzed for expression and copy number of SVM genes, respectively, by real-time reverse-transcriptase polymerase chain reaction (qRT-PCR) and real time PCR (qPCR, methods described in Appendix S5.8). The cell line-based SVM models were used to predict patient sample drug responses in a blinded manner. Two SVM models were trained for paclitaxel and gemcitabine: one using the normalized gene expression values, and the other using expression values binned into 10 categories, using the Matlab function: `quantile(X,10)`. Binning was performed because amplifiable RNA template concentration in FFPE blocks is not known precisely, because it is subject to long term degradation and reactivity (45,46). Expression measurements were obtained for 11 genes from the paclitaxel SVM, and 6 genes for the gemcitabine SVM. The SVM was trained on the cell line data with these reduced gene sets. Predicted and actual responses were

compared, and odds risk ratios (contingency analysis) were calculated (GraphPad Prism, San Diego, California).

Patient data were also obtained from GEO Accession GSE25066, in which expression levels of tumours that were treated with taxane and anthracycline chemotherapy were reported (5). Expression levels for the paclitaxel SVM genes (except *BMF* and *CSAG2*, which were not measured) were extracted for those patients treated with paclitaxel (n = 319). In cases with multiple probe sets per gene, expression levels were averaged. The SVM predictions were then related to response to therapy and residual cancer burden class for each patient.

5.2.6 Clustering cell lines and patients using expression values of the SVM gene subsets

The unsupervised, hierarchical clustering function '`clustergram`' in Matlab was used to cluster cell lines and patient data (described in 2.5) according to gene expression values included in the optimized SVM. Expression values were normalized by row so the mean expression of each gene across individuals was 0, and the standard deviation was 1. Clustering was performed by individuals and genes, and dendrograms are displayed for each dimension that indicate relatedness based on their lengths and hierarchical branching.

5.3 Results

5.3.1 Multiple Factor Analysis

MFAs were performed using GI50s of 49 cell lines, and genomic measurements of 31 and 18 genes related to paclitaxel and gemcitabine activity from an existing data set (7). We re-confirmed measurements for a subset of the cell lines to ensure consistency between cell line sources (see Appendix S5.9). MFAs were assessed by statistics generated by the program, FactoMineR (42). Relationships were stratified by the correlation between the variable and GI50, the RV coefficient (a multivariate generalization of the squared Pearson correlation coefficient), the position of variables on

the correlation circle, and the representation quality of each variable group in the first two dimensions (cos2 values). These criteria were used to classify each gene as having a "strong relationship", "relationship", "possible relationship" or "no relationship" to GI50 (see Appendix S5.10 for the thresholds for each class). Examples of correlation circles and individual factor maps for *MAPT* (paclitaxel) and *DCTD* (gemcitabine) are illustrated in Appendix S5.11.

MFA revealed "strong relationships" between paclitaxel GI50 and copy number and/or gene expression for 11 genes, consisting of both negative relationships (diminished copy number and gene expression [-] for *CYP2C8, CYP3A4, NR1I2* (previously known as PXR), *TLR6,* and *TUBB1*) and positive relationships (increased copy number and gene expression [+] for *BBC3, BCL2L1, BMF, CNGA3, MAPT,* and *TUBB4B*) with increased chemoresistance (Appendix S5.12 lists all MFA measurements). The gemcitabine set revealed strong associations between resistance and *ABCB1* (+), *DCTD* (-), and *SLC28A1* (+) gene expression as well as strong relationships for *ABCC10* (+) and *CDA* (+) copy number (Appendix S5.13). The MFA results for paclitaxel (gene expression results only) and gemcitabine treatment (copy number and gene expression), in the respective pathway contexts, are summarized in Figure 5.2.

Point mutation status was based on 74 deleterious coding mutations (Supplementary Table 5.2) that were predicted to be damaging (47) or to affect mRNA splicing (48,49). Point mutations predicted to be damaging demonstrated strong relationships in *ABCB1* (n = 4, in 2 cell lines) to paclitaxel resistance and in *SLC28A3* (n = 3, in 2 cell lines) to increased sensitivity to gemcitabine. The limited number of cell lines with mutations in these genes cannot be effectively incorporated into machine learning models, and point mutation results were not included in these analyses.

5.3.2 Support Vector Machine (SVM) Learning

A binary SVM was employed to develop a predictive multigene classification of genomic signatures for resistance to these drugs (50). Based on MFA results, data types orthogonal to GI50 were excluded from the SVM (see section 2.5 for details). The classification error of the SVM model was minimized by removing genes, i.e. features, which did not

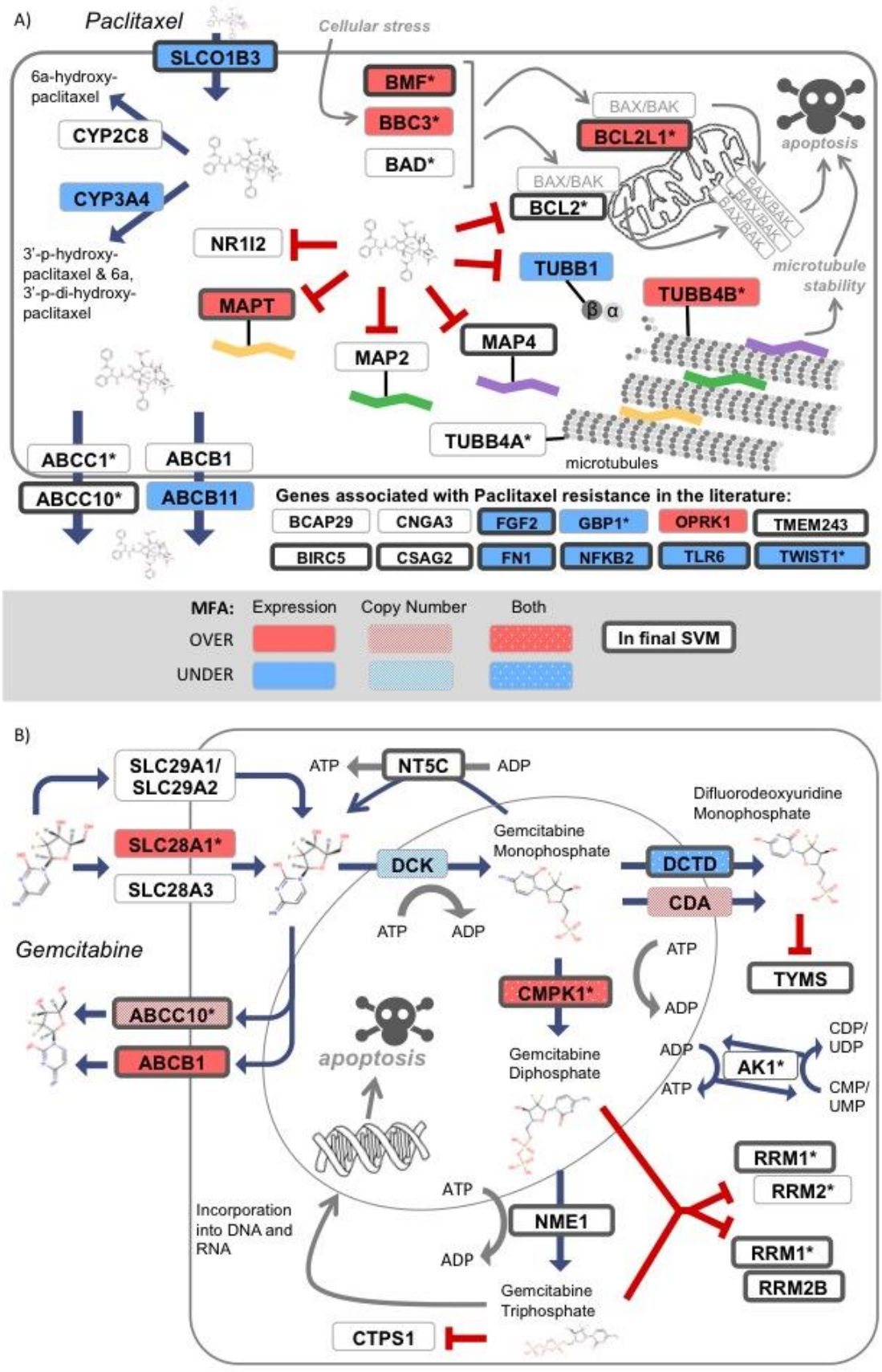

A)
Paclitaxel
Cellular stress
SLCO1B3
6a-hydroxy-paclitaxel
CYP2C8
CYP3A4
3'-p-hydroxy-paclitaxel & 6a, 3'-p-di-hydroxy-paclitaxel
BMF*
BBC3*
BAD*
BAX/BAK
BCL2L1*
BCL2*
apoptosis
NR1I2
MAPT
MAP2
MAP4
TUBB1
TUBB4B*
TUBB4A*
microtubule stability
microtubules
ABCC1*
ABCB1
ABCC10*
ABCB11
Genes associated with Paclitaxel resistance in the literature:
BCAP29
CNGA3
FGF2
GBP1*
OPRK1
TMEM243
BIRC5
CSAG2
FN1
NFKB2
TLR6
TWIST1*
MFA:
Expression
Copy Number
Both
OVER
UNDER
In final SVM
B)
SLC29A1/ SLC29A2
SLC28A1*
SLC28A3
Gemcitabine
ATP
NT5C
ADP
DCK
Gemcitabine Monophosphate
Difluorodeoxyuridine Monophosphate
DCTD
CDA
TYMS
CMPK1*
Gemcitabine Diphosphate
AK1*
CDP/ UDP
CMP/ UMP
ABCC10*
ABCB1
apoptosis
Incorporation into DNA and RNA
NME1
Gemcitabine Triphosphate
CTPS1
RRM1*
RRM2*
RRM1*
RRM2B

Figure 5.2 Genes associated with paclitaxel (A) and gemcitabine (B) mechanism of action (direct targets, metabolizing enzymes), genes previously associated with resistance, and stable genes in the biological pathways targets. Genes with an asterisk (*) are stable genes (Park et al., 2012). Genes highlighted in red showed a positive correlation (within dimension 1 and/or dimension 2) between gene expression or copy number, and resistance in the MFA, whereas genes highlighted in blue demonstrated a negative correlation. Genes outlined in dark grey are those included in the final predictive model that was derived using the SVM. Red T-shaped bars indicate the genes that paclitaxel directly binds/inhibits. Genes outlined in light grey (ie. BAX/BAK) were not included in the analysis because they were not stable genes in the BCL2 pathway.

improve accuracy by leave-one-out cross-validation (see section 2.5 for details). This feature selection process is illustrated in Figure 5.3-I. The optimized SVM was then trained, respectively, on 15 gene variables for paclitaxel (49 cell lines) and 10 variables for gemcitabine (44 cell lines). Gene expression values from *ABCC10*, *BCL2*, *BCL2L1*, *BIRC5*, *BMF*, *FGF2*, *FN1*, *MAP4*, *MAPT*, *NFKB2*, *SLCO1B3*, *TLR6*, *TMEM243*, *TWIST1*, and *CSAG2* comprised the final set of features used to train the SVM for classification of paclitaxel sensitivity. For gemcitabine, both gene expression values (from *ABCB1*, *ABCC10*, *CMPK1*, *DCTD*, *NME1*, *RRM1*, *RRM2B*) and copy number data (from *ABCC10*, *NT5C*, *TYMS*) were used in the final SVM. The distance of each cell line value from the SVM hyperplane that distinguishes the degree of sensitivity or resistance was plotted against the corresponding GI50 (Appendix S5.14). The trained SVMs misclassified 9 of 49 (18%) cell lines for paclitaxel and 7 of 44 (16%) for gemcitabine, which is comparable to, or more accurate than other approaches (51). Partitioning by histological subtype did not improve the classification accuracy; a single variable SVM model based on subtype misclassified 30% of cell lines for paclitaxel and 45% for gemcitabine (Appendix S5.15). The feature-optimized SVM outperformed the signature derived from the initial set of genes, which misclassified resistance/sensitivity of 36% of cell lines for paclitaxel and 64% for gemcitabine treatments. In addition, multi-gene MFA analyses of the final SVM gene sets demonstrate that the individual factor maps of the resistant and sensitive cell lines segregate to a greater degree than MFAs based on the initial gene sets, which were indistinguishable (Appendix S5.16). These differences were larger for gemcitabine than paclitaxel.

To assess the individual impacts of a gene on SVM accuracy, each gene remaining in the optimized SVM was removed, and the misclassification rate was redetermined (Figure 5.3A-II). *BCL2L1* and *MAPT* had the highest predictive value for paclitaxel sensitivity, with misclassification rates of 36% and 34%, respectively, when eliminated (compared to 21-30% for the other genes). It is notable that the MFA also showed strong associations with decreasing *MAPT* or *BLC2L1* expression and increasing paclitaxel sensitivity. *BCL2L1* is a member of the Bcl-2 family and is involved in regulation of apoptosis (16). Additional apoptotic regulators, such as *BMF* and *BCL2*, also appear in our SVM results,

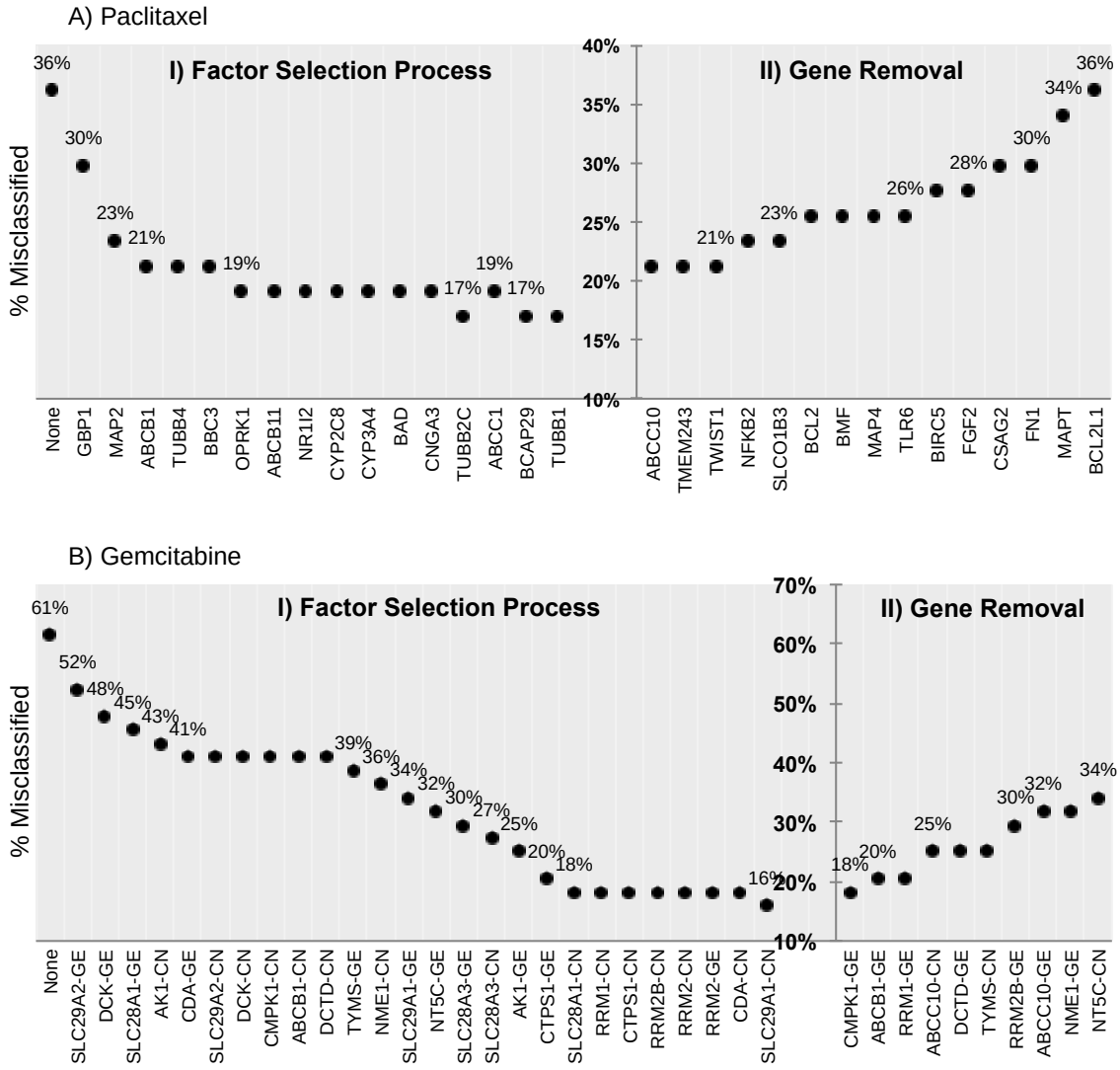


Figure 5.3 Effect of the removal of each gene on the percent of cell lines misclassified during the SVM feature selection process to determine the most predictive gene set (left panels AI and BI). The right panels (AII and BII) demonstrate the increase in the percent of cell lines misclassified when the expression of genes in the inferred, optimal gene set are subsequently eliminated from the SVM.

as paclitaxel is known to trigger apoptosis through these pathways (52). The loss of *MAPT* in breast cancer cells has been shown to sensitize those cells to the action of paclitaxel (53), which is supported by our analysis.

For gemcitabine, removing *NT5C* copy number, *NME1* gene expression, *ABCC10* gene expression, and *RRM2B* gene expression had the largest effects, by respectively increasing misclassification rates to 34%, 32%, 32%, and 30% (Figure 5.3B-II). *NT5C* is located on 17q25.1 a region associated with cancer (54). Allelic imbalances in *TYMS* have previously been hypothesized to be involved in drug resistance in renal cell carcinoma (55) and *ABCC10* has been associated with drug resistance (56). *NME1* is a known metastasis suppressor gene which may have great prognostic value (57). *RRM2B* and *RRM1* have been suggested to be associated with gemcitabine resistance (58) and have been shown to be overexpressed in a gemcitabine-resistant pancreatic cancer cell line (59).

5.3.3 Applying the cell line-trained SVM to patient data

Formalin fixed paraffin embedded (FFPE) tissue blocks were obtained from patients that were treated with paclitaxel and gemcitabine, and whose responses to both drugs are known. Gene expression measurements for 11 genes from the paclitaxel SVM, and gene expression (6 genes) and copy number (CN; 3 genes) from the gemcitabine SVM were obtained using qRT-PCR and qPCR (Supplementary Tables 5.3 and 5.4). Gene expression measurements were not obtained for *BMF*, *CSAG2*, *SLCO1B3*, *TWIST1* (paclitaxel), and *ABCB1* (gemcitabine), as no amplification was observed in these samples by 40 cycles. The absence of amplification in these genes was related to their low levels of expression in breast cancer tissue (Appendix S5.17.1). In cases where qRT-PCR showed no amplification for a specific sample out of the genes measured, the highest cycle run was used as the Ct value for that gene. Older samples, on average, had lower numbers of genes with successful measurements (Appendix S5.17.2).

An SVM was trained using the cell line data with a reduced set of 11 (paclitaxel – *ABCC10*, *BCL2*, *BCL2L1*, *BIRC5*, *FGF2*, *FN1*, *MAP4*, *MAPT*, *NFKB2*, *TLR6*, and

TMEM243) and 9 (gemcitabine – *ABCC10*, *CMPK1*, *DCTD*, *NME1*, *RRM1*, *RRM2B*, *ABCC10*-CN, *NT5C*-CN, and *TYSM*-CN) gene values, which corresponded to the measurements obtained from the FFPE tissue block studies. These SVMs were then applied to the FFPE tissue sample data to predict their sensitivity to paclitaxel and gemcitabine (see Supplementary Table 5.5 for full FFPE sample predictions). The paclitaxel SVM predicted drug sensitivity with 71% accuracy (Table 5.1), which was similar to a leave-one-out analysis on the cell line data, which classified cell lines with 70.2% accuracy (using the reduced 11-gene subset). Patients who were treated with docetaxel were excluded from this summary because the trained SVM only predicted cell line response to docetaxel with 57% accuracy (misclassified 19/44, based on GI50s). Docetaxel and paclitaxel GI50s for all cell lines were correlated only to a limited extent ($R^2 = 0.722$), consistent with the possibility that there might potentially be differences in mechanisms of drug metabolism and resistance between these drugs. The gemcitabine SVM did not perform as well on the patient sample data as it did on the cell line leave-one-out analysis, which was 79.6% accurate (using the reduced 9-gene subset). The gemcitabine SVM derived using binned expression values predicted patient response with 62% accuracy, however, 72% accuracy was achieved for samples with gene expression measurements available for at least 4 of the 6 genes.

Although DNA variants were not incorporated into the SVM models, we sequenced a subset of the FFPE samples to determine whether any potentially damaging mutations were present in paclitaxel/gemcitabine genes of interest, especially for genes that showed relationships between mutations and drug sensitivity in the MFA (Appendix S5.12 & S5.13). Native DNA from 8 samples (all tumours) and whole genome amplified (WGA) DNA from 16 samples (9 tumour and 7 matched normal tissue) were used for next generation sequencing that enriched for the genes of interest. WGA was required for 16 samples, because the amount of DNA extracted from the samples was not a sufficient starting quantity for the sequencing protocols used. Despite the fact that the samples had been qualified by PCR amplification from exons of several genes (including *BRCA1* and *BRCA2*), attempts to prepare NGS libraries for 2 of the original DNA samples were unsuccessful, presumably due to accumulated DNA damage during formaldehyde treatment and storage of the sample. Since spectrophotometric measurements indicated

Table 5.1 Using the SVM to predict patient response from archived FFPE tissue

	Paclitaxel		Gemcitabine	
	NORM	**10 bins**	**NORM**	**10 bins**
No. of accurate predictions	12	12	9	13
Total	17*	17	21	21
Percent accurate	71%	71%	43%	62%
Odds Ratio	5.83	6.00	3.00	3.33
P-value^	0.1534	0.1534	0.5333	0.3615

*4 patients were treated with docetaxel instead of paclitaxel, and were not included in this summary. ^Fisher's exact test. Gene expression values were either normalized (NORM) or binned into 10 categories (10 bins), as described in the methods. Please refer to Supplementary Table 5.5 for all FFPE clinical response/prediction data and the values used for binning.

that DNA was present in nearly all samples, WGA was used to recover the fraction of intact DNA present in the isolates that did not yield libraries by conventional procedures. The full methods used are described in Appendix S5.5.3, and the RNA sequences used for targeted DNA gene capture are listed in Supplementary Table 5.6.

DNA sequencing coverage was variable beween samples, ranging from 7-31 reads per base pair for the original DNA, and between 0-139 for the WGA DNA. DNA variants were detected in five of the original DNA samples (each with 6, 32, 46, 8, and 32 variants) and three of the WGA DNA samples (each with one variant). Of the variants residing in paclitaxel and gemcitabine genes of interest, 12 were predicted to be damaging (47) (two were novel with average heterozygosity <1% and 10 were known SNPs), and the remainder were predicted to be "tolerated" (4 novel, 108 known SNPs, see Supplementary Table 5.7 for full mutation list). There were very few (ie. 1 or none) variants detected in the WGA samples because these samples did not have uniform coverage throughout targeted genes. There was significant bias in the WGA DNA sequencing, where there were few regions with very high coverage (ie. as high as 4000 reads per bp), and the majority of regions with no coverage (Figure 5.4 – B/C). This was not the case with the original DNA samples that were sequenced, as coverage was more uniformly distributed among the genes of interest. This mirrors what we found in the gene expression experiments, where measurements were not obtained for every sample in every gene, suggesting that there are regions of the FFPE template DNA that are more difficult to amplify than others. In total, 5 (22%) out of 22 samples had acceptable, uniform coverage, which is in line with a previous study that found ~18% of FFPE samples pass quality control for subsequent next generation sequencing (81).

Ultimately, only 4 samples harbored potentially damaging mutations (47), including samples from patients 2 (in *SLC28A1,* two in *MAP4,* and *RRM2B*), 6 (in *NFKB2*), 8 (*ABCC1, SLC28A1,* and *RRM2B*), and 24 (three in *MAPT* and *BAD*). Of these genes, *ABCC1* mutations were associated with increased sensitivity to paclitaxel in the MFA (Appendix S5.12), and *RRM2B* mutations were associated with resistance (Appendix

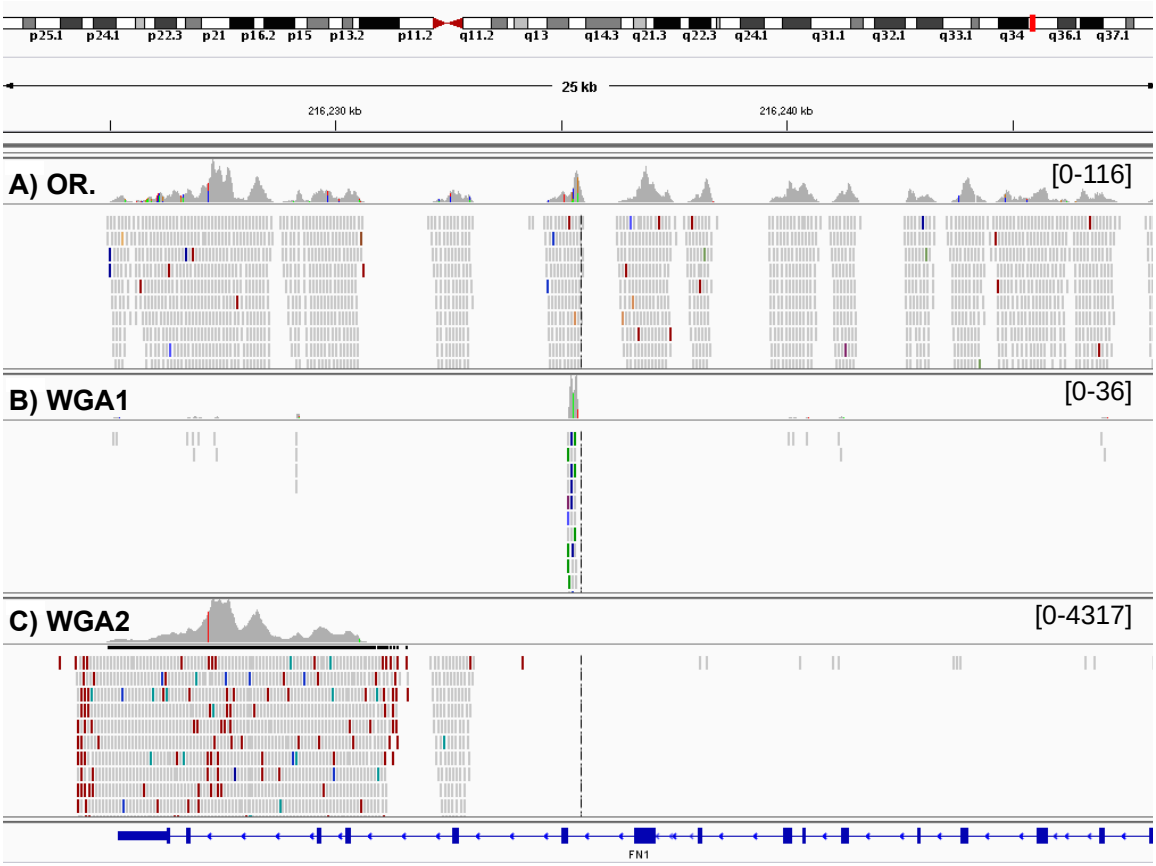


Figure 5.4 Coverage and reads from sequencing of three FFPE tumour samples using originally extracted (A) and whole genome amplified (B/C) DNA. An IGV screen shot covering a 25 kb portion of FN1 (gene displayed at the bottom in blue shows exons [thick bars] and introns [thin bars]) on chromosome 2 (specific band is indicated by the red bar in the top chromosome diagram). Coverage values for A-C are indicted in the top panel by the grey peaks, with the scale in square brackets on the right hand side. DNA sequencing reads are individually displayed as grey (and other colour) bars in the bottom panels of A-C.

S5.13). This corresponds with our patient response data: sample 2B was a gemcitabine non-responder and sample 8C initially responded to paclitaxel (and then subsequently developed resistance), and was resistant to gemcitabine. Although mutation data are a sparse data source that is not easily modeled by SVMs, it appears that mutations on an individual basis may provide insight into tumour response to paclitaxel or gemcitabine.

Gene expression measurements and clinical data were also obtained for 319 patient samples who were treated with paclitaxel and anthracycline chemotherapy (5). Gene expression data were not available for two genes from the paclitaxel SVM (*BMF* and *CSAG2*), which were two of the 4 genes that could not be measured in the FFPE samples. Consequently, the same 11-gene SVM used for the FFPE samples was applied to the data from Hatzis et al. (2011). SVM predictions were compared with the clinical outcome - whether the patient had recurrent disease (RD) or complete pathological response (pCR, see Table 5.2 for a summary, and Supplementary Table 5.8 for all predictions). The SVM predicted sensitivity in 52 of the 63 patients (84%) that showed pCR. All patients that showed complete pathological response exhibited no or minimal residual disease (residual cancer burden [RCB] class 0/1 (60), although some patients within this subset did not respond to therapy. This group of patients (RCB 0/1) may derive the greatest benefit from the paclitaxel SVM analysis. The SVM did not perform as well in predicting resistance, miscategorizing 135 patients of the 257 with RD (52.5%) as sensitive. However, performance of the SVM exceeded that of the 512-gene signature described in Hatzis et al. (2011) for both sensitive and resistant patients. The odds ratio of the 11-gene SVM was 4.484 (Fisher's exact test, $p<0.0001$), compared to the odds ratio of 3.181 of the predictive signature described in that study (Fisher's exact test, $p<0.0001$).

5.3.4 Clustering cell line and patient data based on SVM gene subsets

Two distinct groups emerge from unsupervised clustering using the SVM gene set for paclitaxel in the cell line data (Figure 5.5A). The left cluster (highlighted in light grey) corresponds with the luminal subtype, and the right corresponds to a mix of basal,

Table 5.2 SVM predictions on 319 patients treated with paclitaxel from Hatzis et al. (2011)

	Cell Line 11-gene SVM		Hatzis "Rx" Prediction	
	RD	pCR	RD	pCR
ALL RCB Classes				
Predicted Insensitive	119	10	186	28
Predicted Sensitive	138	52	71	34
Odds Ratio	4.484		3.181	
P-value^	<0.0001		<0.0001	
RCB Class 0/1 Only				
Predicted Insensitive	11	10	10	28
Predicted Sensitive	19	52	20	34
Odds Ratio	3.011		0.6071	
P-value^	0.0359		0.3673	

RD = recurrent disease (designated "insensitive" patient response), pCR = pathological complete response (designated as "sensitive" patient response), RCB = residual class burden (as described in Symmans et al. 2007). ^ p-values were determined using a Fisher's exact test. Please refer to Supplementary Table 5.8 for all predictions and patient information.

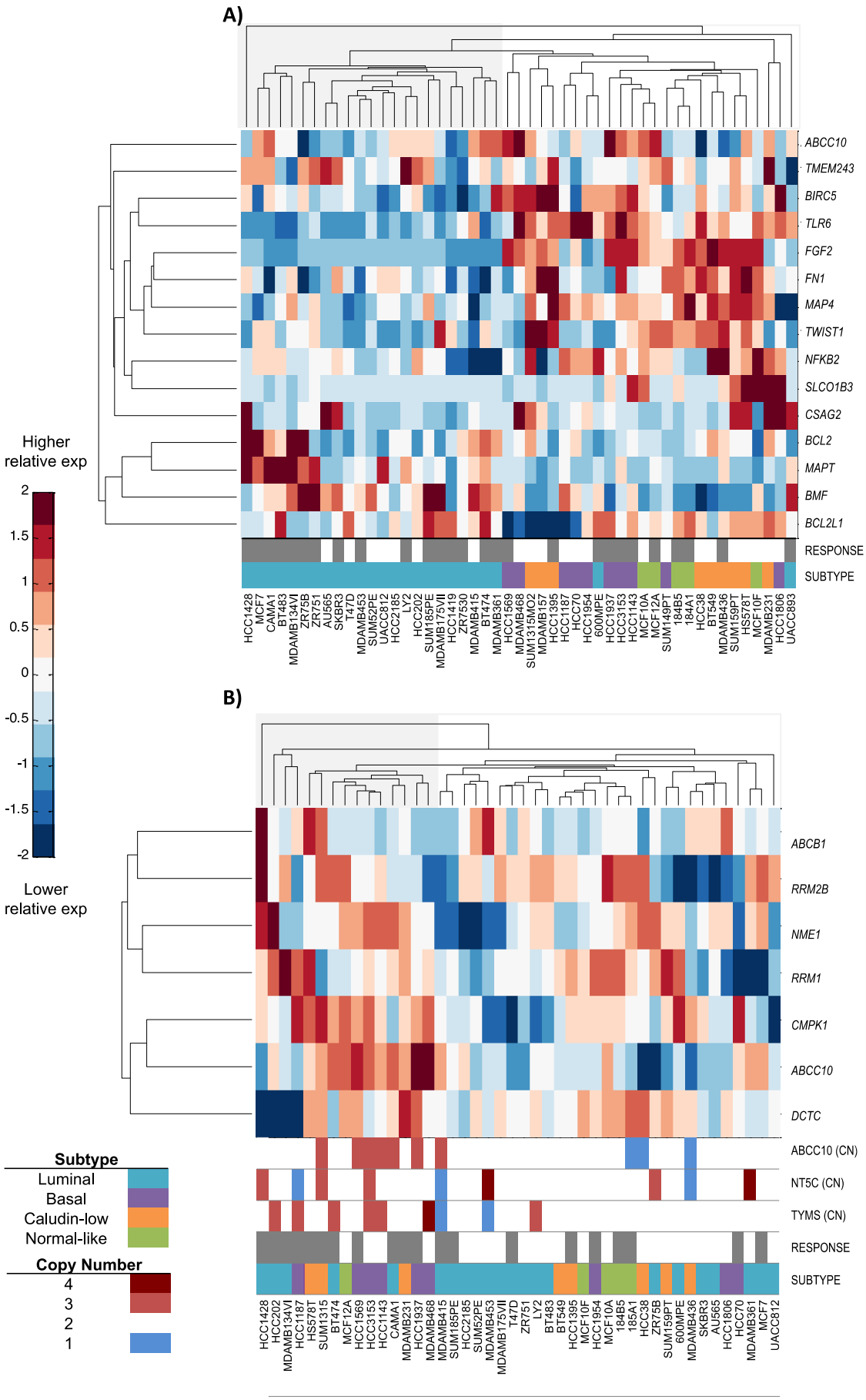

A)
ABCC10
TMEM243
BIRC5
TLR6
FGF2
FN1
MAP4
TWIST1
NFKB2
SLCO1B3
CSAG2
BCL2
MAPT
BMF
BCL2L1
RESPONSE
SUBTYPE
Higher relative exp
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
Lower relative exp
HCC1428 MCF7 CAMA1 BT483 MDAMB134VI ZR75B ZR751 AU565 SKBR3 T47D MDAMB453 SUM52PE UACC812 HCC2185 LY2 HCC202 SUM185PE MDAMB175VII HCC1419 ZR7530 MDAMB415 BT474 MDAMB361 HCC1569 MDAMB468 SUM1315MO2 MDAMB157 HCC1395 HCC1187 HCC70 HCC1954 600MPE HCC1937 HCC3153 HCC1143 MCF10A MCF12A SUM149PT 184B5 184A1 HCC38 BT549 MDAMB436 SUM159PT HS578T MCF10F MDAMB231 HCC1806 UACC893
B)
ABCB1
RRM2B
NME1
RRM1
CMPK1
ABCC10
DCTC
ABCC10 (CN)
NT5C (CN)
TYMS (CN)
RESPONSE
SUBTYPE
Subtype
Luminal
Basal
Caludin-low
Normal-like
Copy Number
4
3
2
1
HCC1428 HCC202 MDAMB134VI HCC1187 HS578T SUM1315 BT474 MCF12A HCC1569 HCC3153 HCC1143 CAMA1 MDAMB231 HCC1937 MDAMB468 MDAMB415 SUM185PE HCC2185 SUM52PE MDAMB453 MDAMB175VII T47D ZR751 LY2 BT483 BT549 HCC1395 MCF10F HCC1954 MCF10A 184B5 185A1 HCC38 ZR75B SUM159PT 600MPE MDAMB436 SKBR3 AU565 HCC1806 HCC70 MDAMB361 MCF7 UACC812

Figure 5.5 Expression heatmap of the paclitaxel and gemcitabine SVM derived genes for the tested cell lines. Each row represents a gene and each column a cell line. Red indicates higher expression and blue represents lower expression, as shown by the colour bar on the left. 'Resistant' cell lines are coloured grey and 'sensitive' cell lines are coloured white in the row labeled 'response'. Cell lines are labeled by subtype and copy number according to the legends. Clustering was done based on the similarity of each cell line's expression profile in the 1st (column) dimension and each gene's expression profile in the 2nd (row) dimension. The dendrograms on the top and left indicate the relatedness of each cell line and gene by the length and subdivision of the branches, with deeper branches indicating a stronger relationship and branches in the same 'tree' being more closely related to each other than data in other 'trees'. A) A section of the dendrogram for paclitaxel is shaded grey to indicate a cluster composed entirely of luminal cell lines and a higher proportion of resistant cell lines. The other section is white to indicate a cluster with very few luminal cell lines and a higher proportion of sensitive cell lines. B) A section of the dendrogram for gemcitabine is shaded grey to indicate a cluster composed of a higher proportion of resistant cell lines. The other section is white to indicate a cluster with a higher proportion of sensitive cell lines.

claudin-low, and normal-like subtypes. The proportions of resistant (71% of the left cluster) vs. sensitive (58% of the right cluster) cell lines are not statistically significant (χ^2 = 3.67, 1 degree of freedom, p = 0.056). Cell lines clustered using the gemcitabine SVM gene expression values display at least two distinct clusters that do not correspond to any subtype(s), but, stratify according to gemcitabine sensitivity (73%; left) or resistance (69%; right) (Figure 5.5B, chi-statistic = 10.75, p = 0.001, d.f. =1). Clustering of the FFPE derived samples was not as strong as a consequence of limited sample numbers and lack of expression measurements for every gene in every sample (Appendix S5.18.1). Nevertheless, clustering of expression in these samples mirrored the cell line data based on results for *MAPT* and *BCL2* (for paclitaxel) and *DCTD* (for gemcitabine).

Unsupervised clustering of expression data from Hatzis et al. (2011), using the paclitaxel SVM distinguished patients according to the proportions of those free of distant relapse (Figure 5.6 and Appendix S5.18.2). These clusters are partially distinguished by *MAPT* and *BCL2* expression (Figure 5.6A, the "low *MAPT*" cluster is indicated in purple, "high *MAPT*" in green). *MAPT* and *BCL2* are both components of the PAM50 Breast Cancer Intrinsic Classifier. Their expression patterns segregate into luminal and basal subtypes to a large extent. Low *MAPT* expressing luminal subtypes were observed to have significantly worse prognoses than higher *MAPT* expressing luminal tumours in the patient dataset ($p<0.05$, Appendix S5.19). The gene signature described by Hatzis et al. (2011) predicted treatment "sensitivity" and "insensitivity" accurately within the low *MAPT* cluster, where "sensitive" patients exhibit significantly longer times to distant relapse (Figure 5.6C, p = 0.0013, log rank test). However, this was not the case for the high *MAPT* cluster, as the proportion free of distant relapse between two predicted groups did not differ significantly (p = 0.10, log-rank test).

5.3.5 Significance of SVM classification accuracy

To assess the significance of the derived SVM, we selected 100,000 random sets of 15 genes from a set of expression values (to compare to the paclitaxel SVM) and 10 genes from a set of copy number and expression values (gemcitabine SVM) for 23,030 genes. Only 0.14% of paclitaxel and 0.01% of gemcitabine random gene combinations exceeded

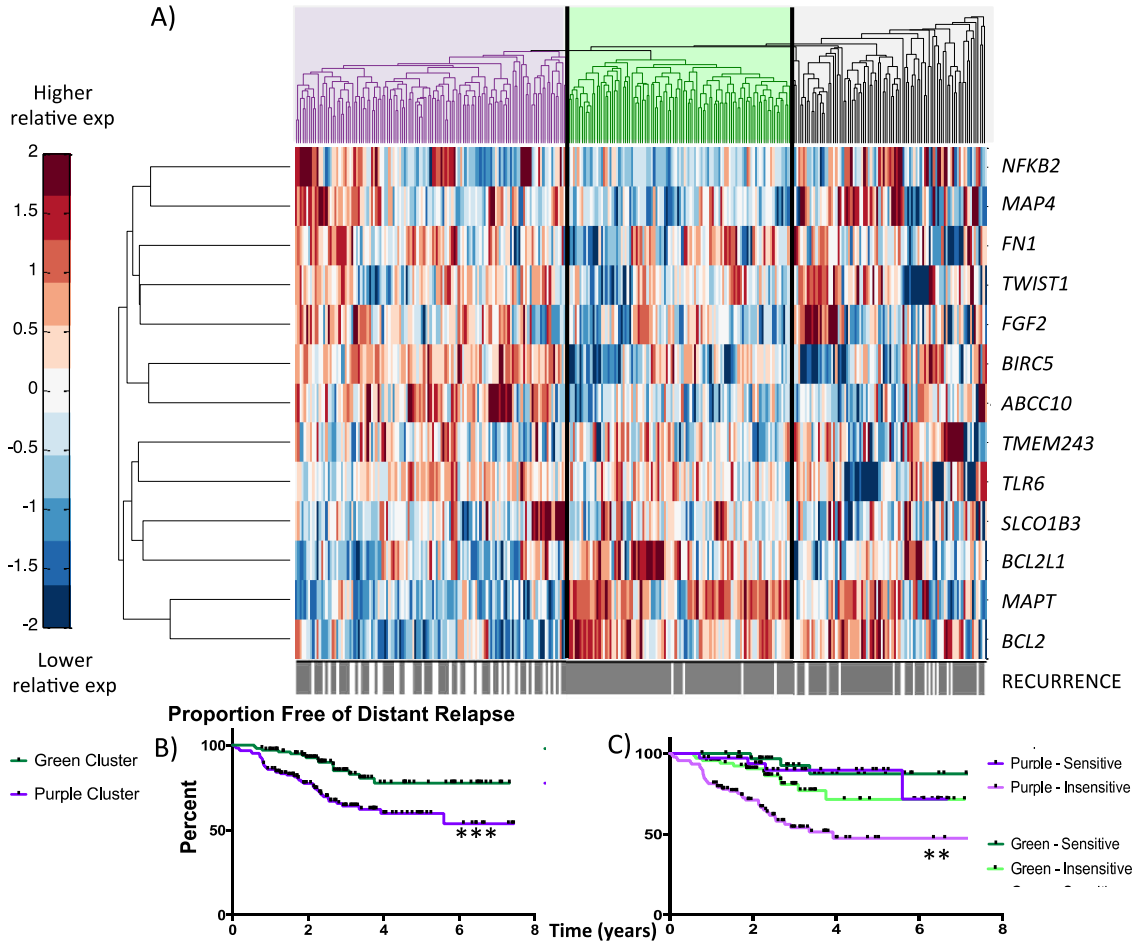


Figure 5.6 A) Expression heatmap of the paclitaxel SVM derived genes for 319 tumour samples (Hatzis et al. 2011). See Figure 5.5 for heat map labeling and diagram details. A section of the dendrogram on the top is shaded purple to indicate a cluster of tumours (83% luminal) with a significantly worse outcome assessed by the proportion free of distant relapse curves (shown in B). Another section is shaded green (63% basal) with significantly better outcomes. The cluster shaded gray (22% basal, 53% luminal) can be clustered independently with similar stratification by subtype and outcome (Supplemental Information VI). C) The Hatzis et al. (2011) gene signature performs very well in the purple cluster and poorly in the green, based on the Kaplan-Meier curves constructed on each subset using their published labels ("insensitive" and "sensitive").

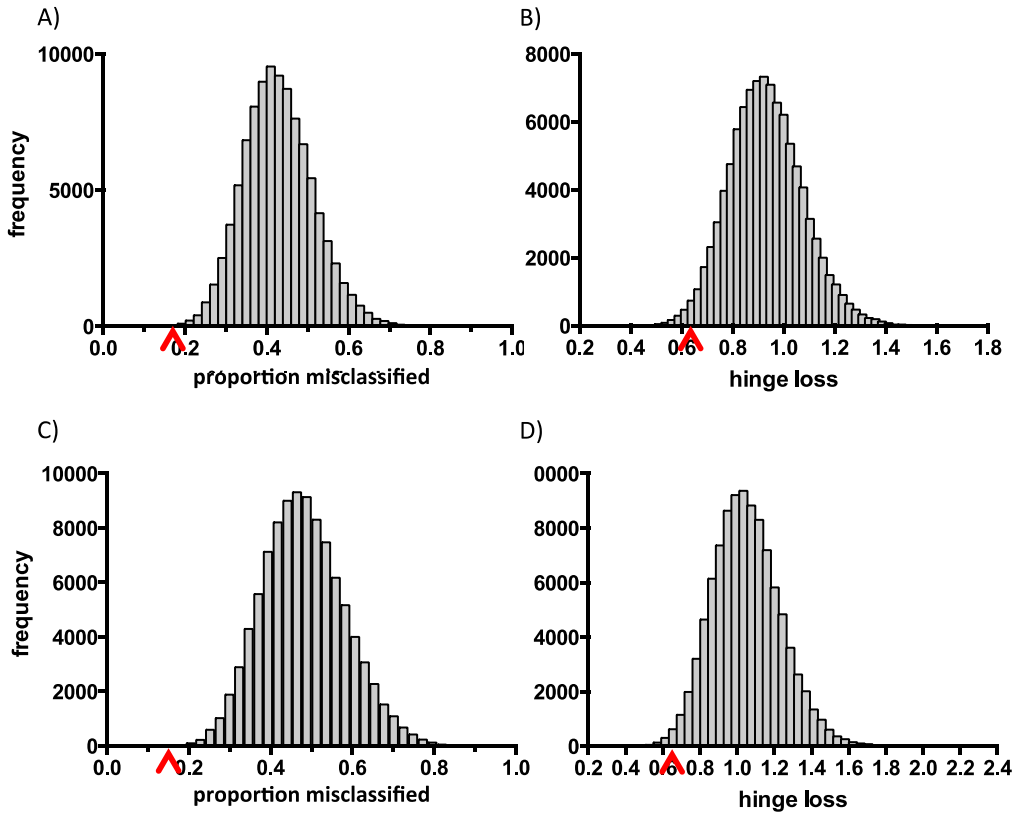


Figure 5.7 The proportion of misclassified cell lines (A/C) and hinge loss scores (B/D) were measured on SVMs derived using randomly selected gene sets. 15-gene (to compare to the paclitaxel SVM, A/B) or 10-gene (to compare to the gemcitabine SVM, C/D) values were randomly selected from an initial set of 23,030 genes and used to derive SVMs. The performance of 100,000 iterations of the random signatures are plotted in the above histograms. The hinge loss scores for the paclitaxel and gemcitabine final SVM gene subsets lie in the lowest 2nd (paclitaxel, z-score -2.0, $p < 0.05$ one-sided) and 1st percentiles (gemcitabine, $z = -2.16$, $p < 0.05$) of the data. Expression alone was used for the 15-gene sets (A/B). Copy number and expression were used for the 10-gene sets (C/D). The red arrow-heads indicate where the optimized paclitaxel and gemcitabine SVM gene signatures are found in the distribution.

the classification accuracy of the derived SVMs. (Figure 5.7 – A/C). The hinge loss, which increases based on the misclassified object's distance to the hyperplane, was 0.64 for the paclitaxel SVM and 0.66 for the gemcitabine SVM (optimal is close to zero). Among the random gene combinations, the likelihood of deriving SVMs with equal or lower scores was 1.45% and 0.83% for paclitaxel and gemcitabine, respectively (Figure 5.7 – B/D). Thus, the accuracy of the SVMs achieved for both drugs were not likely due to random chance ($p<0.05$ in all cases, Table 5.3).

Nearly all of the high performance random gene set combinations appear to be statistical artifacts. Analysis of 10,000 random gene selections found 18 combinations with lower paclitaxel misclassification response rates. All 18 signatures were unique (2 transcripts occurred twice) and transcript combinations were dominated (24%) by alternative splice variants and expressed pseudogenes. None of the random gene combinations were significantly associated with known biological pathways. Six of the random signatures contained $\geq$ 10 gene expression values in the patient data. None of these signatures predicted paclitaxel sensitivity, except one set containing *WWP1*, which has previously been suggested to be a prognostic indicator in breast cancer (61). This signature (and one based on *WWP1* expression alone) predicted more patients (5) to be sensitive to paclitaxel than our derived SVM. Similar numbers of patients predicted to be sensitive by both SVM models exhibited complete remission (52 vs. 55), however the *WWP1*-based SVM predicted sensitivity in a greater number of non-responders (n = 178) than our derived SVM (n= 138) and misclassified 41% of the cell lines. For the gemcitabine response, the SVM of a single random gene set had a lower misclassification rate than our derived SVM. The genes in this set were unrelated to gemcitabine metabolism, with 9 of 10 SVM variables exhibiting copy number changes, two of which involved non-coding RNA genes.

5.3.6 Translation of signature to other cancer types

To mitigate tissue-specific effects, we rederived SVM models specific to lung cancer (lung) and hematopoietic and lymphoid tissue cancer (hematopoietic) cell lines using

Table 5.3 SVM performance using randomly selected genes based off 100,000 iterations

	minimum	maximum	average	standard deviation	drug SVM	z-score	p-value	No. random SVMs ≤ drug SVMs[1]
percent misclassification of cell lines in leave-one-out analysis								
15-gene[2]	12.2%	83.7%	42.7%	8.8%	**18.4%**	-2.78	0.0027	141
10-gene[3]	12.2%	90.2%	48.0%	10.5%	**15.9%**	-3.06	0.0011	10
hinge loss score								
15-gene	0.39	1.66	0.93	0.14	**0.64**	-2.04	0.0207	1,453
10-gene	0.30	2.02	1.05	0.18	**0.66**	-2.16	0.0153	826

Misclassification rates and hinge loss scores were determined from SVMs derived using 100,000 random combinations of gene expression and copy number values from 23,0303 genes. The minimum, maximum, average, and standard deviations of each 100,000 iterations were determined, and compared to the paclitaxel and gemcitabine SVMs ("drug SVM"). [1]the number of random gene combinations with equal or lower misclassification rates or hinge loss scores compared to the drug SVMs, [2]random selection of 15 gene expression values were compared to the paclitaxel SVM, [3]random selection of 10 gene expression or copy number values were compared to the gemcitabine SVM.

expression data from the broad institute (www.broadinstitute.org/ccle/home; "CCLE_Expression_2012-09-29.res" and "CCLE_NP24.2009_profiling_2012.02.20.csv"). Lung and hematopoietic tissue types were chosen because they contained the highest number of cell lines with expression and paclitaxel GI50s. The final lung SVM contained 14 genes, and classified cell lines with 72% accuracy (Appendix S5.20.1). The final hematopoietic SVM was composed of 8 genes, and classified cell lines with 75% accuracy (Appendix S5.20.2). Four genes were present in all three (breast, lung and hematopoietic) cancer cell line SVMs (*BMF, FGF2, TMEM243,* and *TWIST1*), and 8 genes were eliminated from all of the SVMs (*ABCB11, BBC3, CNGA3, CYP2C8, CYP3A4, NR1I2*, *TUBB4A*, and *TUBB4B*; Appendix S5.20.3). MFAs using the Lung and Hematopoietic SVM gene sets do not show the same degree of segregation between resistant and sensitive cell lines as the breast SVM (Appendix S5.20.4 & S5.20.5).

5.4 Discussion

This paper describes the development of genomic signatures using support vector machines that can predict breast cancer tumour response to paclitaxel and gemcitabine. We used a biologically-driven approach to identify a meaningful group of genes whose expression levels and copy number may be useful in guiding selection of specific chemotherapy agents during patient treatment. Previous studies have derived associations between the genomic status of one or more genes and tumour response to certain therapies (5,51,62-65). Correlations between single gene expression and tumour resistance (32,62) do not take into account multiple mechanisms of resistance or assess interactions between multiple genes. ABC transporter overexpression has long been shown to confer resistance, but enzymatic or functional inhibition has not substantially improved patient response to chemotherapy (66).

Multi-gene analytical approaches have previously been successful in deriving prognostic gene signatures for metastatic risk stratification (Oncotype DX^{TM}, MammaPrint®), subtypes (PAM50), and efforts to predict chemotherapy resistance (67). Given the

complexity of genomic changes and the fundamental biological differences among the intrinsic subtypes of breast cancer (68,69), this approach has advantages over analysis of isolated genes. Reasonable gene signatures associated with breast cancer outcome can be obtained by chance alone (70), however our results show that such signatures are especially rare. Gene signatures derived without reference to the underlying mechanisms of chemotherapy response do not capture meaningful biological results (71).

Our approach started with a focused biologically-relevant initial gene set, rather than taking a genome-wide approach. The derived signatures were demonstrated to significantly outperform random selected combinations of genes in prediction of sensitivity and resistance. The random gene sets may be statistical artifacts, as they were not enriched for any biological relevant pathways, and included expressed pseudogenes. The compositions of these other gene sets were distinct from the set used to derive the SVM and another 20-gene signature for taxane sensitivity (6).

Our analysis highlights the importance of the expression of genes encoding microtubule-associated proteins and apoptotic regulators in paclitaxel resistance (17,72,73). *MAPT* expression was significantly correlated with drug resistance, and both *MAPT* and *MAP4* were components of the optimized paclitaxel SVM gene set. In clustering analysis of both cell lines and patients, *MAPT* was differentially expressed between tumour clusters stratified by subtype and outcome (Figures 5.5 and 5.6). Our results confirm that apoptosis-related proteins, particularly *BCL2L1*, but also *BCL2*, *BMF*, and *BIRC5*, contribute to paclitaxel sensitivity (74). *BCL2L1*, *BCL2* and *BMF* were found to be stable in breast cancer tumours, reinforcing the notion that alterations in stable genes contribute to drug resistance (14). Supplementary Table 5.9 describes genes analyzed in the context of their biological pathways and relevant literature.

The gemcitabine metabolic pathway has been well characterized (75), however the critical genes have not been treated as an ensemble in conferring resistance (see Supplementary Table 5.10 for interpretation of the MFA results for all genes). The MFA analyses indicated gemcitabine genes predominately contribute to drug resistance through overexpression. For *DCTD*, however, underexpression is associated with increased

resistance in the MFA analysis. *DCTD* deficiency causes an imbalance in the dNTP pool (76), which affects control of DNA replication. *DCTD* is inhibited by dFdCTP (a gemcitabine metabolite) through a mechanism by which gemcitabine exhibits self-potentiation (the reduction of competing natural metabolites) (77). Lower *DCTD* expression and as a consequence, activity would reduce gemcitabine self-potentiation by altering the dNTP pool. This state is related to drug resistance, which was noticeably lower in 4 cell lines with increased resistance (HCC1187, HCC1428, HCC202, and MDAMB134VI). Like *DCTD*, *CDA* also catalyzes the conversion of gemcitabine monophosphate to difluorodeoxyuridine monophosphate (Figure 5.2B), and accounts for 90% of this conversion in the cell (37). However, drug resistance was associated with CDA overexpression. Likewise, the ribonucleotide reductase subunits *RRM1* and *RRM2B* make significant contributions to the gemcitabine SVM. The *RRM1-RRM2B* complex is associated with mitochondrial genomic integrity (78) and *RRM2B* is necessary for nucleotide synthesis in DNA repair (79). Changes in *RRM2B* expression could be associated with mitochondrial dysfunction, or may result from loss of p53 expression, which usually induces *RRM2B* expression (80).

The 11-gene paclitaxel SVM was able to classify FFPE patient samples we obtained and measured in our lab with similar accuracy to that of the cell lines. In addition, the same SVM model was able to predict complete pathological response on a second patient data set, with greater accuracy than the originally reported gene signature (5). The SVM performed particularly well for predicting drug-sensitive tumours with low or no minimal residual disease (Table 5.2). The SVM gene signature proved to be resilient as a diagnostic marker, as the performance was not compromised by the lack of expression data for 4 genes.

Unlike paclitaxel, gemcitabine was not used to treat patients in the study by Hatzis et al. (2011) or other publically available data sets. The SVM analysis on RNA expression and DNA copy number from the FFPE-derived tumour punches appeared to predict response more accurately when expression values were obtained for most of the genes in the SVM. Obtaining high quality gene expression measurements from FFPE samples was especially difficult from older tissue blocks (Appendix S5.17.2) as previously noted (81).

Consequently, the SVM analysis may be better suited for fresh-frozen tumour tissue or more sensitive gene expression analyses (such as mRNA sequencing). Missing data appeared to impact the gemcitabine SVM to a greater extent than the paclitaxel SVM, which may be due to the smaller number of gene measurements required for this SVM.

Including gene expression subtype in the SVM did not improve the classification accuracy even though subtype is known to contribute to tumour biology (11). However, the two paclitaxel PAM50 genes (*MAPT* and *BCL2*) partially stratify the cell lines by subtype during unsupervised clustering (Figure 5.6). This is not the case in the gemcitabine gene set. In patient data, clustering by expression of the SVM genes also revealed statistically significant deterioration in outcome for low *MAPT* expressing luminal tumours (Appendix S5.19).

Machine learning may be a fruitful approach in the selection of other chemotherapy agents. Translating our results to the assessment of human tumour samples (4) confirmed our gene signature's relevance to predicting chemoresistance by SVM. A limitation of our work is that both SVMs were not integrated because cell lines were only treated with individual drugs, so predicting whether patient response to these drug interactions will be synergistic or antagonistic is not currently possible. In addition, while point mutations are well known contributors to chemoresistance of other drugs, this approach – for either SVM training or testing - is not conducive for prediction of chemosensitivity given the sparse number of observations for these types of mutations.

In cases without residual disease, the paclitaxel SVM was particularly effective in predicting which tumours would show complete pathological response. Docetaxel is prescribed somewhat interchangeably (5,82,83) and both paclitaxel and docetaxel act through similar biological pathways (84). However the performance of the paclitaxel SVM on patients treated with docetaxel was reduced. This SVM contains 8 paclitaxel resistance genes. Predictions of docetaxel sensitivity might be improved by rederiving a specific SVM using taxane pathway genes (84), and those known to be associated with resistance to doclitaxel (such as *CYP1B1* (85,56), *miR-141* or EIF4E (87), *DKK3* (88), *ABCB1* (89,90), *BIRC5* (91), *ABCC10* (92), *miR*-452 (93), and PAWR (94)). The

approach that we have introduced could aid in rational selection of other therapeutic regimens that evade or at least minimize the effects of chemoresistance.

5.5 References

1. Cardoso, F. *et al.* Locally recurrent or metastatic breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **23,** vii11–vii19 (2012).

2. Oostendorp, L. J. M., Stalmeier, P. F. M., Donders, A. R. T., van der Graaf, W. T. A. & Ottevanger, P. B. Efficacy and safety of palliative chemotherapy for patients with advanced breast cancer pretreated with anthracyclines and taxanes: a systematic review. *Lancet Oncol.* **12,** 1053–1061 (2011).

3. Lee, S.-Y. *et al.* Genetic polymorphisms of SLC28A3, SLC29A1 and RRM1 predict clinical outcome in patients with metastatic breast cancer receiving gemcitabine plus paclitaxel chemotherapy. *Eur. J. Cancer Oxf. Engl. 1990* **50,** 698–705 (2014).

4. Gąsowska-Bodnar, A. *et al.* Survivin expression as a prognostic factor in patients with epithelial ovarian cancer or primary peritoneal cancer treated with neoadjuvant chemotherapy. *Int. J. Gynecol. Cancer Off. J. Int. Gynecol. Cancer Soc.* **24,** 687–696 (2014).

5. Hatzis, C. *et al.* A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* **305,** 1873–1881 (2011).

6. He, D.-X., Xia, Y.-D., Gu, X.-T., Jin, J. & Ma, X. A 20-gene signature in predicting the chemoresistance of breast cancer to taxane-based chemotherapy. *Mol. Biosyst.* **10,** 3111–3119 (2014).

7. Daemen, A. *et al.* Modeling precision treatment of breast cancer. *Genome Biol.* **14,** R110 (2013).

8. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6,** 813–823 (2006).

9. Neve, R. M. *et al.* A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10,** 515–527 (2006).

10. Prat, A. *et al.* Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes. *Breast Cancer Res. Treat.* **142,** 237–255 (2013).

11. Heiser, L. M. *et al.* Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 2724–2729 (2012).

12. Ein-Dor, L., Zuk, O. & Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci.* **103,** 5923–5928 (2006).

13. Nilsson, R., Björkegren, J. & Tegnér, J. On reliable discovery of molecular signatures. *BMC Bioinformatics* **10,** 38 (2009).

14. Park, N. I., Rogan, P. K., Tarnowski, H. E. & Knoll, J. H. M. Structural and genic characterization of stable genomic regions in breast cancer: relevance to chemotherapy. *Mol. Oncol.* **6,** 347–359 (2012).

15. Jordan, M. A. & Wilson, L. Microtubules as a target for anticancer drugs. *Nat. Rev. Cancer* **4,** 253–265 (2004).

16. Ferlini, C. *et al.* Paclitaxel directly binds to Bcl-2 and functionally mimics activity of Nur77. *Cancer Res.* **69,** 6906–6914 (2009).

17. McGrogan, B. T., Gilmartin, B., Carney, D. N. & McCann, A. Taxanes, microtubules and chemoresistant breast cancer. *Biochim. Biophys. Acta* **1785,** 96–132 (2008).

18. Harmsen, S., Meijerman, I., Beijnen, J. H. & Schellens, J. H. M. Nuclear receptor mediated induction of cytochrome P450 3A4 by anticancer drugs: a key role for the pregnane X receptor. *Cancer Chemother. Pharmacol.* **64,** 35–43 (2009).

19. Heijn, M. *et al.* Anthracyclines modulate multidrug resistance protein (MRP) mediated organic anion transport. *Biochim. Biophys. Acta* **1326,** 12–22 (1997).

20. Chen, Z.-S. *et al.* Characterization of the transport properties of human multidrug resistance protein 7 (MRP7, ABCC10). *Mol. Pharmacol.* **63,** 351–358 (2003).

21. Lecureur, V. *et al.* Cloning and expression of murine sister of P-glycoprotein reveals a more discriminating transporter than MDR1/P-glycoprotein. *Mol. Pharmacol.* **57,** 24–35 (2000).

22. Duan, Z., Brakora, K. A. & Seiden, M. V. MM-TRAG (MGC4175), a novel intracellular mitochondrial protein, is associated with the taxol- and doxorubicin-resistant phenotype in human cancer cell lines. *Gene* **340,** 53–59 (2004).

23. Rao, P. S., Bickel, U., Srivenugopal, K. S. & Rao, U. S. Bap29varP, a variant of Bap29, influences the cell surface expression of the human P-glycoprotein. *Int. J. Oncol.* **32,** 135–144 (2008).

24. Duan, Z., Foster, R., Brakora, K. A., Yusuf, R. Z. & Seiden, M. V. GBP1 overexpression is associated with a paclitaxel resistance phenotype. *Cancer Chemother. Pharmacol.* **57,** 25–33 (2006).

25. Kaczanowska, S., Joseph, A. M. & Davila, E. TLR agonists: our best frenemy in cancer immunotherapy. *J. Leukoc. Biol.* **93,** 847–863 (2013).

26. Tantivejkul, K. *et al.* PAR1-mediated NFkappaB activation promotes survival of prostate cancer cells through a Bcl-xL-dependent mechanism. *J. Cell. Biochem.* **96,** 641–652 (2005).

27. Carmo, C. R., Lyons-Lewis, J., Seckl, M. J. & Costa-Pereira, A. P. A novel requirement for Janus kinases as mediators of drug resistance induced by fibroblast growth factor-2 in human cancer cells. *PloS One* **6,** e19861 (2011).

28. Lu, J. *et al.* Mitotic deregulation by survivin in ErbB2-overexpressing breast cancer cells contributes to Taxol resistance. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **15,** 1326–1334 (2009).

29. Hong, J. *et al.* Phosphorylation of serine 68 of Twist1 by MAPKs stabilizes Twist1 protein and promotes breast cancer cell invasiveness. *Cancer Res.* **71,** 3980–3990 (2011).

30. Xing, H. *et al.* Activation of fibronectin/PI-3K/Akt2 leads to chemoresistance to docetaxel by regulating survivin protein expression in ovarian and breast cancer cells. *Cancer Lett.* **261,** 108–119 (2008).

31. Duan, Z., Lamendola, D. E., Duan, Y., Yusuf, R. Z. & Seiden, M. V. Description of paclitaxel resistance-associated genes in ovarian and breast cancer cell lines. *Cancer Chemother. Pharmacol.* **55,** 277–285 (2005).

32. Duan, Z., Feller, A. J., Toh, H. C., Makastorsis, T. & Seiden, M. V. TRAG-3, a novel gene, isolated from a taxol-resistant ovarian carcinoma cell line. *Gene* **229,** 75–81 (1999).

33. Preissner, S. *et al.* SuperCYP: a comprehensive database on Cytochrome P450 enzymes including a tool for analysis of CYP-drug interactions. *Nucleic Acids Res.* **38,** D237–243 (2010).

34. Kavallaris, M. Microtubules and resistance to tubulin-binding agents. *Nat. Rev. Cancer* **10,** 194–204 (2010).

35. Marcé, S. *et al.* Expression of human equilibrative nucleoside transporter 1 (hENT1) and its correlation with gemcitabine uptake and cytotoxicity in mantle cell lymphoma. *Haematologica* **91,** 895–902 (2006).

36. Mackey, J. R. *et al.* Gemcitabine transport in xenopus oocytes expressing recombinant plasma membrane mammalian nucleoside transporters. *J. Natl. Cancer Inst.* **91,** 1876–1881 (1999).

37. Govindarajan, R. *et al.* Facilitated mitochondrial import of antiviral and anticancer nucleoside drugs by human equilibrative nucleoside transporter-3. *Am. J. Physiol. Gastrointest. Liver Physiol.* **296,** G910–922 (2009).

38. Ueno, H., Kiyosawa, K. & Kaniwa, N. Pharmacogenomics of gemcitabine: can genetic studies lead to tailor-made therapy? *Br. J. Cancer* **97,** 145–151 (2007).

39. Plunkett, W. *et al.* Gemcitabine: metabolism, mechanisms of action, and self-potentiation. *Semin. Oncol.* **22,** 3–10 (1995).

40. Mini, E., Nobili, S., Caciagli, B., Landini, I. & Mazzei, T. Cellular pharmacology of gemcitabine. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. ESMO* **17 Suppl 5,** v7–12 (2006).

41. Abdi, H. & Valentin, D. in *Encyclopedia of Measurement and Statistics* 657–663 (Sage, 2007).

42. Lê, S., Josse, J. & Husson, F. FactoMineR: An R package for multivariate analysis. *J. Stat. Softw.* **25,** 1–18 (2008).

43. Dash, M. & Liu, H. Feature Selection for Classification. *Intell. Data Anal.* 131–156 (1997). doi:10.3233/IDA-1997-1302

44. Musella, V. *et al.* Use of formalin-fixed paraffin-embedded samples for gene expression studies in breast cancer patients. *PloS One* **10,** e0123194 (2015).

45. Antonov, J. *et al.* Reliable gene expression measurements from degraded RNA by quantitative real-time PCR depend on short amplicons and a proper normalization. *Lab. Investig. J. Tech. Methods Pathol.* **85,** 1040–1050 (2005).

46. Fleige, S. & Pfaffl, M. W. RNA integrity and the effect on the real-time qRT-PCR performance. *Mol. Aspects Med.* **27,** 126–139 (2006).

47. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).

48. Shirley, B. C. *et al.* Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences. *Genomics Proteomics Bioinformatics* **11,** 77–85 (2013).

49. Viner, C., Dorman, S. N., Shirley, B. C. & Rogan, P. K. Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. *F1000Research* **3,** 8 (2014).

50. Furey, T. S. *et al.* Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinforma. Oxf. Engl.* **16,** 906–914 (2000).

51. Ma, X.-J. *et al.* A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* **5,** 607–616 (2004).

52. Kutuk, O. & Letai, A. Alteration of the mitochondrial apoptotic pathway is key to acquired paclitaxel resistance and can be reversed by ABT-737. *Cancer Res.* **68,** 7985–7994 (2008).

53. Bhat, K. M. R. & Setaluri, V. Microtubule-associated proteins as targets in cancer chemotherapy. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **13,** 2849–2854 (2007).

54. Fukino, K. *et al.* Frequent allelic loss at the TOC locus on 17q25.1 in primary breast cancers. *Genes. Chromosomes Cancer* **24,** 345–350 (1999).

55. Colavito, D. *et al.* Thymidylate synthetase allelic imbalance in clear cell renal carcinoma. *Cancer Chemother. Pharmacol.* **64,** 1195–1200 (2009).

56. Hopper-Borge, E. *et al.* Human multidrug resistance protein 7 (ABCC10) is a resistance factor for nucleoside analogues and epothilone B. *Cancer Res.* **69,** 178–184 (2009).

57. Shoushtari, A. N., Szmulewitz, R. Z. & Rinker-Schaeffer, C. W. Metastasis-suppressor genes in clinical practice: lost in translation? *Nat. Rev. Clin. Oncol.* **8,** 333–342 (2011).

58. Aye, Y., Li, M., Long, M. J. C. & Weiss, R. S. Ribonucleotide reductase and cancer: biological mechanisms and targeted therapies. *Oncogene* (2014). doi:10.1038/onc.2014.155

59. Wang, C. *et al.* Establishment of human pancreatic cancer gemcitabine‐resistant cell line with ribonucleotide reductase overexpression. *Oncol. Rep.* **33,** 383–390 (2015).

60. Symmans, W. F. *et al.* Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **25,** 4414–4422 (2007).

61. Nguyen Huu, N. S. *et al.* Tumour-promoting activity of altered WWP1 expression in breast cancer and its utility as a prognostic indicator. *J. Pathol.* **216,** 93–102 (2008).

62. Duan, Z. *et al.* Overexpression of MAGE/GAGE genes in paclitaxel/doxorubicin-resistant human cancer cell lines. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **9,** 2778–2785 (2003).

63. Glinsky, G. V., Berezovska, O. & Glinskii, A. B. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *J. Clin. Invest.* **115,** 1503–1521 (2005).

64. Rajput, S., Volk-Draper, L. D. & Ran, S. TLR4 is a novel determinant of the response to paclitaxel in breast cancer. *Mol. Cancer Ther.* **12,** 1676–1687 (2013).

65. Van ’t Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–536 (2002).

66. Samuels, B. L. *et al.* Modulation of vinblastine resistance in metastatic renal cell carcinoma with cyclosporine A or tamoxifen: a cancer and leukemia group B study. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **3,** 1977–1984 (1997).

67. Hess, K. R. *et al.* Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **24,** 4236–4244 (2006).

68. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

69. Dorman, S. N., Viner, C. & Rogan, P. K. Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer. *Sci. Rep.* **4,** 7063 (2014).

70. Venet, D., Dumont, J. E., Detours, V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* **7,** e1002240 (2011).

71. Drier, Y. & Domany, E. Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? *PLoS One*. **6,** e17795 (2011).

72. Tanaka, S. *et al.* Tau expression and efficacy of paclitaxel treatment in metastatic breast cancer. *Cancer Chemother. Pharmacol.* **64,** 341–346 (2009).

73. Wang, K. *et al.* Tau expression correlated with breast cancer sensitivity to taxanes-based neoadjuvant chemotherapy. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **34,** 33–38 (2013).

74. Flores, M. L. *et al.* Paclitaxel sensitivity of breast cancer cells requires efficient mitotic arrest and disruption of Bcl-xL/Bak interaction. *Breast Cancer Res. Treat.* **133,** 917–928 (2012).

75. Alvarellos, M. L. *et al.* PharmGKB summary: gemcitabine pathway. *Pharmacogenet. Genomics* **24,** 564–574 (2014).

76. Eriksson, S., Skog, S., Tribukait, B. & Jäderberg, K. Deoxyribonucleoside triphosphate metabolism and the mammalian cell cycle. Effects of thymidine on wild-type and dCMP deaminase-deficient mouse S49 T-lymphoma cells. *Exp. Cell Res.* **155,** 129–140 (1984).

77. Xu, Y. Z. & Plunkett, W. Modulation of deoxycytidylate deaminase in intact human leukemia cells. Action of 2',2'-difluorodeoxycytidine. *Biochem. Pharmacol.* **44,** 1819–1827 (1992).

78. Bourdon, A. *et al.* Mutation of RRM2B, encoding p53-controlled ribonucleotide reductase (p53R2), causes severe mitochondrial DNA depletion. *Nat. Genet.* **39,** 776–780 (2007).

79. Kuo, M.-L. *et al.* RRM2B suppresses activation of the oxidative stress pathway and is up-regulated by p53 during senescence. *Sci. Rep.* **2,** 822 (2012).

80. Tanaka, H. *et al.* A ribonucleotide reductase gene involved in a p53-dependent cell-cycle checkpoint for DNA damage. *Nature* **404,** 42–49 (2000).

81. Choudhary, A. *et al.* Evaluation of an integrated clinical workflow for targeted next-generation sequencing of low-quality tumor DNA using a 51-gene enrichment panel. *BMC Med. Genomics* **7,** 62 (2014).

82. Crown, J., O'Leary, M. & Ooi, W.-S. Docetaxel and paclitaxel in the treatment of breast cancer: a review of clinical experience. *The Oncologist* **9 Suppl 2,** 24–32 (2004).

83. O'Shaughnessy, J., Gradishar, W. J., Bhar, P. & Iglesias, J. Nab-paclitaxel for first-line treatment of patients with metastatic breast cancer and poor prognostic factors: a retrospective analysis. *Breast Cancer Res. Treat.* **138,** 829–837 (2013).

84. Oshiro, C. *et al.* Taxane pathway. *Pharmacogenet. Genomics* **19,** 979–983 (2009).

85. Chang, I. *et al.* Loss of miR-200c up-regulates CYP1B1 and confers docetaxel resistance in renal cell carcinoma. *Oncotarget* **6,** 7774–7787 (2015).

86. Cui, J. *et al.* Design and Synthesis of New α-Naphthoflavones as Cytochrome P450 (CYP) 1B1 Inhibitors To Overcome Docetaxel-Resistance Associated with CYP1B1 Overexpression. *J. Med. Chem.* **58,** 3534–3547 (2015).

87. Yao, Y.-S. *et al.* miR-141 confers docetaxel chemoresistance of breast cancer cells via regulation of EIF4E expression. *Oncol. Rep.* **33,** 2504–2512 (2015).

88. Tao, L., Huang, G., Chen, Y. & Chen, L. DNA methylation of DKK3 modulates docetaxel chemoresistance in human nonsmall cell lung cancer cell. *Cancer Biother. Radiopharm.* **30,** 100–106 (2015).

89. Hansen, S. N. *et al.* Acquisition of docetaxel resistance in breast cancer cells reveals upregulation of ABCB1 expression as a key mediator of resistance accompanied by discrete upregulation of other specific genes and pathways. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* (2015). doi:10.1007/s13277-015-3072-4

90. Kato, T. *et al.* Serum exosomal P-glycoprotein is a potential marker to diagnose docetaxel resistance and select a taxoid for patients with prostate cancer. *Urol. Oncol.* (2015). doi:10.1016/j.urolonc.2015.04.019

91. Ghanbari, P. *et al.* Inhibition of survivin restores the sensitivity of breast cancer cells to docetaxel and vinblastine. *Appl. Biochem. Biotechnol.* **174,** 667–681 (2014).

92. Domanitskaya, N. *et al.* Abcc10 status affects mammary tumour growth, metastasis, and docetaxel treatment response. *Br. J. Cancer* **111,** 696–707 (2014).

93. Hu, Q. *et al.* MicroRNA-452 contributes to the docetaxel resistance of breast cancer cells. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **35,** 6327–6334 (2014).

94. Pereira, M. C. *et al.* Prostate apoptosis response-4 is involved in the apoptosis response to docetaxel in MCF-7 breast cancer cells. *Int. J. Oncol.* **43,** 531–538 (2013).

Chapter 6

6 Contextual Insights of Findings in this Dissertation

It is estimated that in Canada, almost 24,000 deaths due to female breast cancer have been avoided since the mortality rate peaked in 1986 (1). Since then, the age-standardized mortality rate has fallen 43%, due to an increase in breast cancer screening and advancements in breast cancer treatment (1). The discoveries of the intrinsic subtypes of breast cancer and prognostic transcript profiles using gene expression microarrays have been instrumental in making breast cancer patient management to be more individualized. The ongoing advancements and reduction in cost of genomic technologies now provide even further opportunity to personalize breast cancer care. However, there are still gaps in genomic experiments, both in experimental design and interpretation of the data. In addition, there are currently no personalized genomic indicators for managing chemotherapy regimes that take into account drug resistance for breast cancer patients. The field has, and will, benefit from methods to improve upon current genomic analyses that detect cardinal abnormalities in driver genes, and predict metastatic progression and chemotherapy response. This thesis describes improvements for data quality and analysis for existing genomic technologies, with the aim of detecting and interpreting genomic abnormalities relevant to breast cancer metastasis and chemotherapy resistance.

6.1 Current limitations of genomic technology

Genome-wide assays, such as microarrays and next generation sequencing, have greatly improved our understanding of both normal and tumour genomes. The large amount of data generated from these experiments, however, creates new sets of challenges to ensure reproducible measurements, robust analyses, and meaningful interpretations.

The issue of low reproducibility, both between and within technology platforms, has not been fully resolved in genome-wide analyses (2-5) (Section 2.1 describes the variability observed in aCGH experiments in further detail). This is not surprising, because FISH, microarrays, and next generation sequencing all rely on the same stochastic events:

nucleic acid extraction, fragmentation, labeling, and hybridization. A recent study assessing replicate next generation sequencing experiments demonstrated concordant rates in single nucleotide variant calling ranged between 54-76% (4). In addition, batch effects occurring from laboratory-specific conditions can create major problems if the batch effect results in incorrect conclusions (6). Improving the reproducibility of these technologies has usually involved increasing the number of measurements obtained in a given experiment, whether through expanding the number of probes on a single microarray slide, or increasing the number of reads obtained from a sequencing experiment. However, genomic experiments are subject to both technical (i.e. experimental procedure) and biological (i.e. genetic) variation (7). Tumour heterogeneity makes the analysis of breast cancer particularly complicated (8). For this reason, single-cell genomic analyses (9) have been applied to cancer research (10,11). This thesis did not address biological variation to the same extent as technical variation, although it is nevertheless an extremely important aspect of tumour biology research.

With the abundance of different technology platforms, generated data, and computer software programs available, establishing robust genomic analysis pipelines remains challenging. This is true for both microarray and next generation sequencing analyses of DNA or RNA. For clinical applications, working groups, such as the American College of Medical Genetics (ACMG), have developed thorough guidelines for such analyses (12-14). For example, recommendations involving next generation sequencing for primary (production of sequence reads and assignment of base quality scores (12,15)), and secondary/tertiary (variant calling and interpretation (14,16)) analyses have been well documented. However, the main objective for clinical analysis (and the guidelines created) is to accurately report genomic variants that are likely relevant to a patient's diagnosis or health. This differs from research groups, who can tolerate greater difficulty interpreting the data and variants of unknown significance in exchange for more comprehensive results. There are numerous programs that can be used to discover and interpret data (Table 1.3), and this list is steadily increasing. Regardless, there is still an underrepresentation of non-coding variants in published genomic studies, such as in-depth splicing mutation analyses outlined in this thesis (Chapters 3 and 4). The

emergence of new software to predict or interpret non-coding variants (17-19) indicates the field is still working towards filling these gaps in current genomic analyses.

Interpreting DNA variants, how they affect cellular functions, and whether they are causing a certain phenoytype is still extremely difficult. Recent evidence presented in this thesis (Chapter 4) and others (20-24) show there is no single cause or set of abnormalities that account for these phenotypes. It is well thought that the interpretation of sequencing data from a full genome is now a much larger task than generating the data itself (25). Large data repositories, such as the International HapMap Project (26) and dbSNP (27), begin to allow us to understand which DNA variants are common among the population, and which variants are rare and potentially pathogenic. However, given the size of the human genome, the majority of variants observed in a given sample will be novel. Programs like SIFT (28) and PolyPhen (29) are able to provide some indication as to whether a mutation will be damaging to the protein's function, but have extremely low specificity (30). The genomic field will still greatly benefit from new programs to validate the predicted effects of a mutation on a genome-wide scale.

6.2 Advances in genomic technology described in this thesis

6.2.1 Fluorescence in-situ hybridization

FISH probes typically span a large genomic region along the chromosome, well beyond the length of a single gene. They have been very useful in delineating large pathogenic chromosomal aberrations, and have played an instrumental role in early gene and disease discovery. With the introduction of chromosomal microarrays, however, our ability to detect much smaller rearrangements has improved. In many cases, clinically significant findings from these high-resolution microarrays will require assays to confirm the suspected copy number change. Using *ab initio*-derived single copy intervals from the human genome sequence, high-resolution FISH probes were designed and validated for probes of small cancer genes. These FISH probes are small, usually less than 4 kb, and the exact genomic location of the probes is known. Further, we have automated the

design process, and have developed >450,000 primer pairs covering regions overlapping genes that could be developed into single copy probes. The advantages of this technology are that it can assess parts of genes and at small single copy regions that are embedded in highly repetitive regions. As with most methods involving nucleic acid hybridization, developing scFISH probes directly in highly conserved repetitive regions is not possible. However, scFISH probes have been used to delineate breakpoints within segmental duplicons (31) and telomeric regions (32). Although the scFISH probes are reproducible, the fluorescent signal is not as intense as traditional BAC probes, which recognize a much larger target on the chromosome. Developing probes with increased signal intensities could allow for easier analysis of interphase cells, as the *ab initio* probes developed in this thesis were only validated on metaphase chromosomes.

Although there are cases where genome-wide analysis is more suitable, FISH is a reliable and inexpensive method to assess specific genomic regions. Future work could include validating probes for specific actionable or clinically significant genomic alterations in oncology (Table 6.1), which would require the development of scFISH on solid tumour FFPE samples. scFISH probes are especially useful for cancer types in which chromosomal microarrays are not routinely used or effective (i.e. balanced translocations (32)). For example, it is now evident that tumours with *HER2* amplification, in addition to breast cancer for which it was originally developed, benefit from *HER2* targeted therapies (such as trastuzumab) (33). The *ERBB2*/*HER2* scFISH probe could be used as an inexpensive method to determine whether amplification is present in a tumour.

6.2.2 Chromosomal Microarrays

Chromosomal microarrays have been instrumental in advancing the evaluation of patients with constitutional abnormalities, and are now accepted as a first tier diagnostic test for patients with developmental delay, intellectual disability, congenital anomalies, and autism (34). However, using genome-wide approaches to detect copy number changes raises new limitations and regulatory challenges for clinical testing. There are still difficulties associated with accurately measuring copy number gains or losses, and the

Table 6.1 Examples of clinically significant genomic alterations in cancer testable by FISH.

Gene	Cancer type	Aberration	Clinical significance
APC	gastric	Decreased copy number/deletion	Significantly associated with lymph node invasion and metastasis[35]
HER2	breast, gastric	Gene amplification	Higher chance of success for treatment with *HER2* monoclonal antibody (ie. trastuzumab)[33,36]
EGFR	colorectal	Increased copy number	Higher chance of success for treatment with antiEGFR monocolonal antibody (ie. cetuximab and panitumumab)[37]
EGFR	non-small-cell lung	Increased copy number	Higher chance of success for treatment with gefitinib[38]
MET	squamous cell carcinoma (lung)	Increased copy number	Poor prognosis (shorter survival)[39]
E2F3	Urothelial carcinoma	Increased copy number	Higher frequency in metastasis[40]
ROS1 or ALK	non-small-cell lung	rearrangement/gene fusion	Treatment with crizotinib[41,42]

subsequent interpretation of the pathogenicity of any findings. The ACMG has approved a set of Standards and Guidelines for genomic copy number testing using microarrays (13,43,44). Microarray probes are suggested to be placed throughout the genome at regular intervals, to enable the detection of copy number changes of 400 kb or larger with 99% sensitivity. It is also recommended that there be an emphasis on probes targeting haploinsufficient genes with known phenotypic abnormalities (43), or regions known to be associated with unbalanced genomic alterations in cancer (44). In addition, it is desirable to be able to detect small rearrangements with high confidence and low false positive rates, to improve diagnosis of small clinically significant copy number variants (45).

Ab initio single copy intervals were used to design a genomic oligonucleotide microarray that demonstrated reduced noise in signal intensities compared to a common commercial platform. We suggest that genomic placement of oligonucleotides relative to repetitive elements can alter their susceptibility to cross hybridization, which increases variability in probe signal intensity. Historically, improved accuracy and resolution of commercial microarray platforms has been achieved by increasing the density of probes on the array (46). This thesis describes an alternative solution to overcoming noise: including a reduced set of oligonucleotides that demonstrate high reproducibility in signal intensity. This may offer a cost-effective solution for high throughput microarray testing by increasing the number of samples that can be processed per slide (through increased multiplexing with the same number of total probes).

These findings are not limited to microarray analysis, but rather apply to any nucleic acid hybridization experiment using genomic DNA. In next generation sequencing analysis, solution hybrid selection is becoming a useful method to enrich for targeted genomic sequences (47). This approach uses biotinylated RNA ‘bait’ that is hybridized to a sheared DNA sample, and then purified using streptavidin-coated beads to enrich for the target sequence. This thesis describes the application of *ab initio* sequences to design the RNA sequences (bait) used for DNA capture and subsequent sequencing (Appendix S5.5.3). Where possible, sequences were selected to be distant from conserved repetitive

sequences to minimize cross-hybridization and wasted coverage on unintended sequences. In addition, capture probes were designed in divergent repetitive elements to allow for greater coverage in some regions that would be excluded using repeat-masking (48). Capture probes resulted in enrichment of the targeted 45 gene sequences, with sufficient coverage to allow for multiplexing of 48 samples per sequencing experiment. Clinics or research groups with specific gene panels of interest could use this method as a cost-effective alternative to whole exome sequencing.

6.2.3 Next Generation Sequencing

With decreasing costs and the development of more user-friendly analysis software, next generation sequencing is becoming mainstream in both research and clinical settings. During mutation analyses, and especially when clinical decisions rely on the results of a study, it is important that we leverage the data to the best of our ability to obtain the most complete and accurate results. In this thesis, the Shannon Human Splicing Pipeline (49) was used to improve splicing mutation detection in 445 breast cancer tumours. Further, a software program was developed and described, named Veridical (50), to employ RNA sequencing data for validation of the predicted mutations' affect on mRNA splicing.

Veridical was the first published genome-wide tool that is able to directly link DNA mutations to aberrant mRNA splicing. Before the development of Veridical, validating splicing mutation could be fairly laborious. RT-PCR is the most common method used to confirm that a splicing mutation will cause abnormal splicing, either through measuring patient mRNA or a transfected cell line that expresses the mutation. Although this method is reliable for individual mutations, it would be difficult and time consuming to apply this technique to all predicted splicing mutations in a genome. For example, 5,206 splicing mutations were detected in 442 tumours (Chapter 4). Assuming patient mRNA is attainable, a very conservative estimate of 4-6 hours of hands-on time would be required to validate each mutation (to develop primers, set up and run the RT-PCR reaction, and analyze the results). This would amount to at least 2,600 8-hour workdays, or ~6 days per tumour to validate these results using a traditional approach.

In addition to its genome-wide capabilities, Veridical can compare the mutated sample to normal exome sequences or other controls to determine the corresponding frequency of the aberrant splicing pattern in samples that do not contain the variant of interest. Veridical is able to achieve high statistical power through comparing hundreds of controls, the extent of which would not be reasonable for a single-variant wet lab experiment. One additional benefit is that the RNA-Seq controls do not need to be generated by the group performing the study, due to the availability of data from online resources such as TCGA (https://tcga-data.nci.nih.gov/tcga/) and the International Cancer Genome Consortium (https://icgc.org).

Other software programs with similar objectives to Veridical have also been recently developed, including PVAAS (51) and SNPlice (52). PVAAS uses "spliced reads" (reads spanning two exons) from RNA sequencing data, and identifies non-canonical splicing, defined as splicing where the 5' and/or 3' splice site(s) are not known. It works in the reverse order of Veridical, identifying variants that are associated with the aberrant splicing after the non-canonical splicing reads are discovered. SNPlice finds RNA sequencing reads that contain a single nucleotide variant, and span into the intronic sequence. It highlights variants that preferentially occur in intron-containing molecules versus reads that are properly spliced, to implicate the variant in abnormal splicing.

The recent development of both PVAAS and SNPlice highlight the importance of identifying splicing mutations that cause aberrant splicing. They are potentially powerful tools that are especially useful in the absence of DNA sequencing data. However, they fail to address some key considerations that were incorporated into Veridical. Both approaches rely on associations between a variant and a splice form to potentially implicate the variant in abnormal splicing. There are two major flaws to this approach. First, the authors did not work with complete gene or genome data, and therefore all possible splicing variants (especially those deep in an intron) would not be present in the analysis. Because the true causal variant may not be detected or known, some atypical splicing transcripts may be miscalled as natural alternative splicing events. Second, a truly causal variant may be in linkage disequilibrium with the inferred variant, and therefore the cause of abnormal splicing is not explained correctly. Further, if two

variants reside in the same region, the programs may have difficulty determining which variant is affecting splicing. While both of these are serious drawbacks and the first is more likely a result of the authors' lack of complete genome or gene data, which means that their inferences were based on a minor fraction of genome variation (53).

Veridical differs in that it is hypothesis-driven, looking for aberrant splicing at the specific location of predicted splicing mutations rather than making post-hoc associations of variants to abnormal splicing, as in PVAAS and SNPlice. In addition, Veridical is able to perform robust statistical analyses against large sets of controls. This is important because it avoids mis-identifying naturally occurring abnormal splicing (i.e. the GATA3 cryptic splicing found in all controls in Appendix S2.2.3) or intron retention (i.e. the abundance of intron-spanning reads in both breast cancer samples and normal controls demonstrated in Appendix S2.2.4) as abnormal.

Veridical confirmed 19% of all splicing mutation predictions in a large subset of breast cancer tumours. That leaves the question, however, of why the other 81% of variants were not confirmed. The parameters outputted from the Shannon Pipeline (i.e. initial, final, or change in the splice site strength, distance to or strength of the nearest natural site) showed no obvious indications of whether the variant would be validated by RNA-Seq. This implies that it is not due to some variable of the algorithm underlying the Shannon Pipeline (which uses information theory), but rather related to the methods of validation.

The first of these issues is the fact that genes were not filtered based on breast tissue expression, so many of the genes harboring splicing mutations may be in genes not that are not expressed (and show minimal read coverage in the RNA-Seq data). For example, a donor mutation at a natural site in *ACSBG1* with a ΔR_i of -18.64 bits (inactivating the site) was not validated even though there was a dramatic decrease in the strength of the site. The GTEx (54) expression value, however, suggests this gene has very low expression in breast (mammary) tissue (Figure 6.1). Second, variants were only called and grouped within the tumour samples, so there was no information as to whether the normal breast samples contained the variant. If enough of the normal samples contained

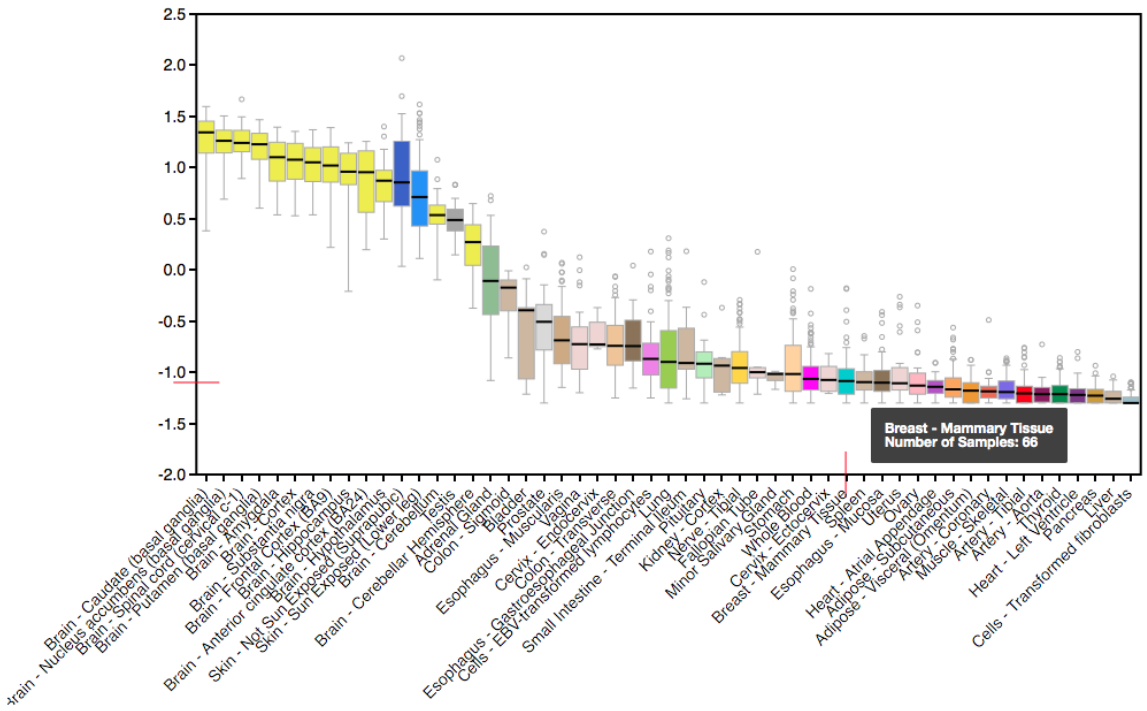


Figure 6.1 Screenshot from GTEx Portal – *ACSBG1* Gene View. Measured gene expression values of *ACSBG1* for different tissue types are listed along the x-axis. The vertical red bar indicates the location of the breast (mammary) tissue, which is filled in light blue. The horizontal red bar indicates the log(expression) value measured from 66 samples. Data Source: GTEx Analysis Release V4 (dbGaP Accession phs000424.v4.p1).

the variant, the abnormal splicing would not have enough statistical power to be observed as significant based on the p-value cutoffs applied. This situation is less likely, because common variants (in dbSNP present in more than 1% of the population) were filtered out of the analysis. Third, although from the same tumour, the DNA- and RNA-Seq data may represent genotypically-different cell populations due to tumour heterogeneity. Finally, some of the splicing variants may have been false positives (i.e. an artifact of the sequencing) or there was simply no evidence of aberrant splicing. Standard quality filters were used during variant calling, although this only reduces and does not fully eliminate false positives. In addition, predicting splicing mutations using information theory has been shown to have a sensitivity of 85% (18), so a minority of the predicted variants may not affect mRNA splicing.

In this thesis, the Shannon Pipeline and Veridical were applied to breast cancer tumours. Future studies could apply similar analyses (from Chapters 3 and 4) to other types of cancer using newly generated or previously published data (from groups like the Cancer Genome Atlas or International Cancer Genome Consortium). This would be particularly valuable in both heritable and somatic cancers where there has been either a lack of causal variants identified in a large portion of cases or where mutations in specific genes lead to clinical decisions. For example, our laboratory is applying splicing (among other non-coding) mutation detection to families with a strong history of breast and/or ovarian cancer that have tested negative for *BRCA1*/2 actionable mutations. *BRCA* testing primarily involves Sanger sequencing (55) of exons to assess mutations in coding regions, and so there are likely protein-damaging splicing mutations that are missed with standard techniques used in the clinic.

Efforts are currently underway in our laboratory to expand Veridical to incorporate additional types of analyses. For example, it could be used to detect whether any type of mutation (splicing or coding) is increasing nonsense-mediated decay (NMD). Transcript levels of both alleles could be detected, and the proportion of the transcript with versus without the mutation could indicate whether the mutated mRNA is susceptible to NMD. A similar type of analysis with different objectives (i.e. not assessing NMD)

comprehensively mapped genotype relationships with expression of specific transcripts using expression quantitative trait loci (eQTLs) in over 40 different tissue types (54,56). In addition to allele-specific expression, the RNA-Seq read coverage in the 5' end of the transcript (low) versus the 3' end of a transcript (high) may also indicate that NMD is occurring. Exon-exon junction protein complexes (EJC) are thought to be removed by the ribosome during the first round of protein translation. When there is a premature stop codon (and the ribosome is released), the 3' EJCs are not removed, and their presence on the transcript triggers the NMD process. Consequently, mutations in the last exon are often missed by NMD because they do not have any remaining EJCs. Veridical could also be applied to any other read-counting based analysis, such as detecting or quantifying non-coding or micro-RNAs in a disease sample or tissue type compared to controls.

6.3 Implications for breast cancer treatment

6.3.1 DNA mutations in metastasis

This thesis demonstrates that there are elevated numbers of NCAM pathway mutations in lymph node positive tumours. Lymph node involvement compared to tumour size can be a marker of the metastatic potential of a tumour independent of the tumour subtype (57). Therefore, NCAM pathway mutations may be an indicator for tumours most likely to metastasize.

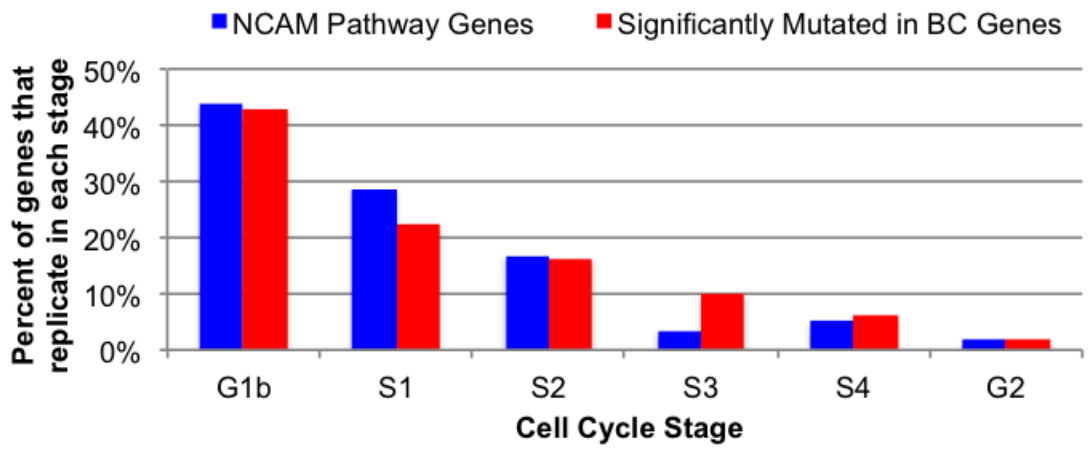
Cancer has long been proposed as a multistage process, both in tumour development (58,59) and advancement of the disease (60). Interestingly, the percent of tumours with NCAM pathway mutations drops off in later (stage IV) tumours. It is possible that NCAM pathway mutations increase metastatic potential in early tumour development, but are not clonally selected for once the tumour has spread. This would explain why these mutations are not present at high levels in advanced disease. If the NCAM pathway mutations were simply passenger mutations in breast cancer, it has been proposed that these genes would be low expressing (61) and late replicating (62-64), which have been associated with higher background mutation rates (65). The stage of replication and

expression levels were compared between NCAM pathway genes and genes significantly mutated in breast cancer, which were cited to be likely driver genes (20-24). We do not find any differences in both replication stage and expression levels between NCAM pathway and other significantly mutated genes (Figure 6.2), which supports excluding the possibility that NCAM pathway mutations are the result of bystander effects. The contributions of these defects to tumour metastasis would have to be demonstrated by functional studies (see below).

A high proportion of the breast cancer tumours assessed harbored extracellular matrix (ECM) and collagen mutations, although these mutations were found at similar levels in all tumours, regardless of their lymph node status. Clonal frequency was previously evaluated in a large set of breast tumours to segregate mutations as either early or later events, which delineated that mutations appear to be acquired later in tumour development in genes that play a role in cytoskeletal pathways, such as myosins, laminins, collagens, and integrins (21). In addition, the differential expression of ECM components has been used to classify breast cancer tumours into groups related to patient prognosis and tumour metastatic potential (66,67). These and other stromal signatures can have higher predictive power when combined with current pathogenic features (receptor status, tumour grade) (68). The ECM of tumours has been cited as a potential target for anti-cancer therapy, although it is challenging to identify which specific ECM component may serve as an effective therapeutic target (69).

Alternative ways to identify tumours that are likely to migrate to other tissues, beyond prognostic gene expression profiling, would be beneficial for many patients. Further work could be completed to confirm the hypothesis that NCAM pathway mutations are indicators for tumour migration. There are now effective, inexpensive ways to test a cell line's metastatic potential. For example, chick chorioallantonic membrane (CAM) assays in conjunction with multiple fluorescent imaging is a useful model to study angiogenesis, invasion, and metastasis (70-72). The assay involves measuring the level of intravasation and growth achieved by inoculated xenogenic tumour cells within the CAM of a chick embryo. Splicing mutations in NCAM pathway genes that were observed in this thesis (Section 4.3.7) could be introduced into breast cancer cell lines, and the cell line's ability

A)



B)

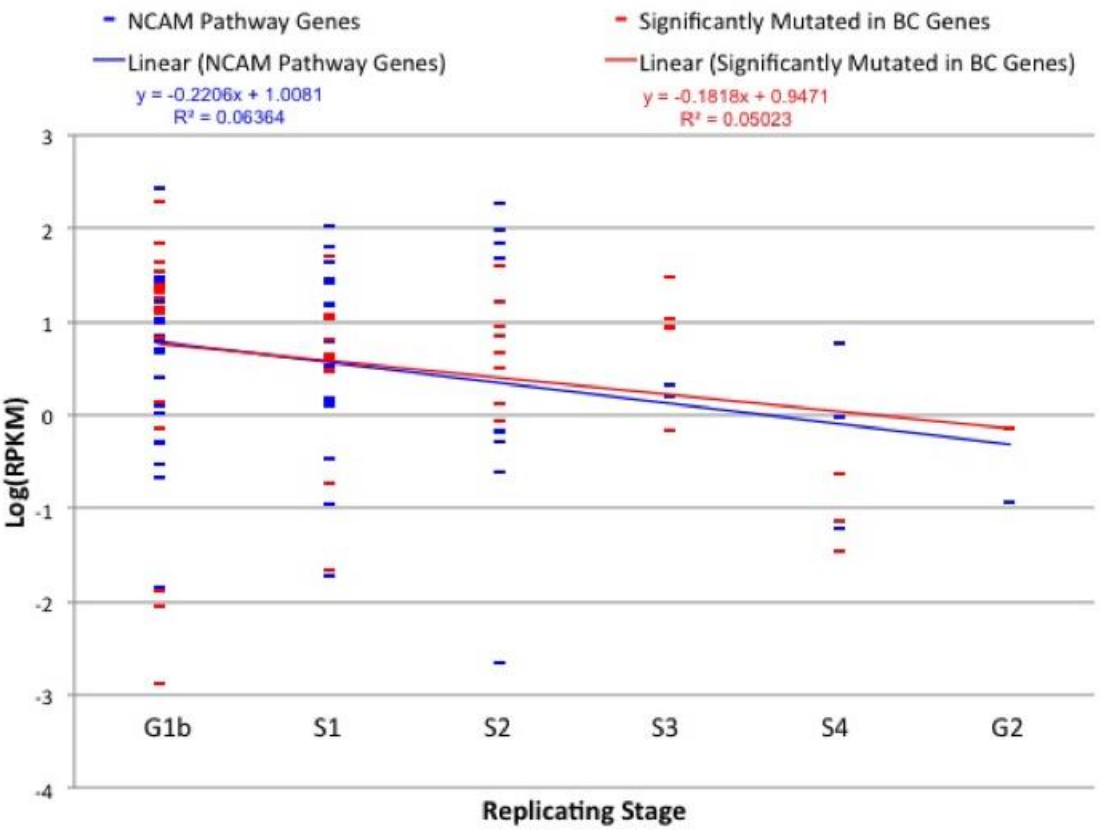


Figure 6.2 Replicating stage and expression of NCAM pathway and significantly mutated genes. A) Replicating stage was determined using the MCF7 cell line data from the UCSC Genome Browser track "Replication Timing by Repli-seq from ENCODE/University of Washington" (73,74). In cases where a gene was replicated during two stages equally, the earliest stage was used. Gene sets used can be found in Appendix S4.3.3 (significantly mutated

genes), and Appendix S4.3.8 (NCAM pathway genes). B) RPKM gene expression values were obtained for each gene (average of 66 normal breast mammary tissue samples) from the GTEx portal (http://www.gtexportal.org/home/). The log(RPKM) was plotted against replication stage for each gene, as in described in A.

to migrate through the CAM could indicate the affect of that mutation on a tumour metastatic potential. This type of study could also potentially identify specific genes within NCAM pathways that contribute the greatest to lymph node invasiveness. Other cell migration and metastatic potential *in vitro* assays that could be applied to study NCAM pathway mutations include scratch-wound assays (75) and Boyden chamber assays (76).

Our laboratory has recently proposed to carry out a prospective trial with basal-like and HER2-enriched breast cancer patients (Section 4.3.8 and Appendix S2.2.5) that would involve sequencing NCAM pathway genes at the point of diagnosis, surgery and/or relapse. Patients could be followed to determine whether those tumours with NCAM pathway mutations were more invasive than those that lacked mutations. A prospective trial would be required due to the fact that some patients with early stage tumours that contain NCAM mutations at initial diagnosis may have longer latency periods to metastasis. In addition, tumour dormancy (77,78) may significantly increase the time to distant metastasis, but the cell migration could still be due to NCAM pathway mutations in the primary tumour. Sequencing tumours that have already metastasized and have undergone further clonal selection will not necessarily harbor the same set of mutations as the primary tumour (21). This study would likely require several years, however would be non-invasive because it would not change the course of treatment for current breast cancer patients, meaning there would be limited downsides for patients enrolling in the study.

6.3.2 Predicting tumour sensitivity to paclitaxel and gemcitabine

Chemotherapy is widely used in breast cancer treatment, although selection of which specific agent to use is qualitative and variable due to patient-related factors. Developing robust genomic signatures to guide selection of chemotherapy agents would be particularly useful for triple negative (TNBC) and advanced breast cancer. In the case of TNBC, there are limited options for therapeutic treatment beyond conventional chemotherapy (79). TNBC (most commonly basal-like and Claudin-low subtypes) are usually aggressive and are more likely to become metastatic, however, women with

TNBC who have a complete pathological response to treatment have excellent outcomes (79). In advanced breast cancer, chemotherapy is used for palliative care and to improve quality of life given that the chance of survival and cure are low (80). Usually, a specific chemotherapy drug, or class of drugs, is only effective until the tumour develops resistance to the treatment. Therefore, it is advantageous to be able to identify those patients who would benefit from immediate treatment with cytotoxic therapies, and those for which surgery and radiation may be sufficient at the time of initial diagnosis. In addition, selecting the chemotherapy agent that is most likely to be effective early on may avoid periods of ineffective treatment and the corresponding unnecessary toxicity and side effects.

This thesis describes a novel approach that used machine learning to generate models that can predict breast cancer tumour sensitivity to paclitaxel and gemcitabine. Gene selection was driven by the biological understanding of these drugs, rather than employing a genome-wide approach that risks identifying un-meaningful signatures correlating to tumour response by chance (81,82). A reduced 11-gene signature for paclitaxel was able to predict tumour response in a set of formalin-fixed paraffin-embedded breast cancer tissue samples with similar accuracy to the cell line data. A reduced 9-gene signature was able to predict tumour response to gemcitabine in samples where at least 4 of the 6 gene expression measurements were obtained, however, it performed poorly on those with limited data. This result highlights the difficulties in working with FFPE tissue samples, where there can be variable and low preservation of nucleic acids (83,84). Measuring the FFPE samples using qRT-PCR was unsuccessful for some genes due to low expression and/or the differences in amplifiable template between samples.

The reduced 11-gene expression signature for paclitaxel was particularly effective in predicting patients with low residual cancer burden that will be have a complete pathological response to paclitaxel. It was not as effective at predicting tumours likely to show resistance, especially in advanced disease. This is not necessarily surprising, as primary and metastatic breast cancer tumours, both within and between lesions, are made up of multiple genetically diverse subpopulations of cancer cells (85). Recent data highlight that differences, in the case of both genomic aberrations and mutation

frequencies, have been observed between primary tumours and subsequent metastatic lesions (86,87). In multifocal breast cancer, it was found that the genetic differences of the lesions in each patient were significantly correlated with the physical differences between the tumours (88). Therefore, the gene signatures developed may only be relevant to a limited subset of the tumour populations related to primary breast cancer tumours, but not those of aggressive clonal isolates. In addition, it is likely that the SVM may only predict response to the specific lesion measured, and not to genetically differentiated lesions or metastases.

Recently, a 20-gene signature ("TAXSig") was developed that predicts chemoresistance to taxane-based therapies in breast cancer patients (89). This study included, but was not limited to, paclitaxel. There was no direct overlap in genes included in the TAXSig signature and the genes included in our SVM model, or randomly generated gene sets that had low misclassification rates from Figure 5.7. However, a pathway analysis using Reactome (90) revealed slight overlap in biological pathways between *FGFR1* from TAXSig, and a subset of the paclitaxel SVM genes. The 35 genes from both signatures are enriched for the innate immune system (p=0.034), as 7 genes (*FGF2, BCL2, BCL2L1, TLR6, NFKB2, FN1* from the SVM and *FGFR1* from TAXSig) are part of the 1,031 genes in this pathway. In addition, *FGFR1* (from TAXSig) interacts with *FGF2* and *FN1* in at least 53 and 31 additional specific signaling pathways, respectively. Although there is some overlap in biological pathways of the TAXSig and paclitaxel SVM gene sets, the majority of genes are unrelated. The taxane (TAXSig) resistance signature may be capturing a different mechanism (or mechanisms) of resistance, which may at least partially explain why chemosensitivity is not predicted with greater accuracy. The paclitaxel SVM was not predictive of docetaxel GI50s, further supporting the notion that they are unrelated processes. The paclitaxel SVM derived in this thesis was reliable in predicting tumours that will respond to the treatment, but nevertheless the phenotypes of patients or cell lines could not all be accurately predicted. One possible explanation might be that some of the features sensitizing a tumour to paclitaxel are independent from those leading to resistance. While we are not aware of any evidence that this occurs, such a hypothesis could explain why we are unable to predict the phenotypes of all cell lines and patients accurately.

Although previous studies have developed gene signatures to predict paclitaxel (or taxane) sensitivity, there has been limited work in using gene expression signatures to predict breast cancer sensitivity to gemcitabine. One study found that polymorphisms in *SLC28A3, SLC29A1,* and *RRM1* can predict metastatic breast cancer sensitivity to combination therapy with paclitaxel and gemcitabine (91). *RRM1* was included in the paclitaxel SVM, and *SLC28A3* mutations and GI50s were strongly related in the set of 44 cell lines assessed using a multiple factor analysis. A study assessing copy number changes in *RRM1* and *RRM2B* found that copy number aberrations of these genes were present in breast cancer tumours, but were not related to clinical outcome of patients treated with gemcitabine (92). During feature selection (used to generate the SVM), we found that copy number of both *RRM1* and *RRM2B* had no impact on the model's ability to predict gemcitabine sensitivity (Figure 5.3). Given that there is a need for models to predict gemcitabine, further work on large patient sets could be completed to validate or improve upon the gemcitabine SVM derived in this thesis. Although the SVM did not perform as well on a small number of FFPE tumour samples as it did in the cell lines, obtaining reliable gene expression measurements from these tumour blocks was challenging. Attempting a similar analysis on fresh-frozen tumours may provide further insight into the utility of the gemcitabine SVM in patient care.

Similar methods may be effective in generating models for other chemotherapy agents for which the biological mechanism of action is known. For example, pathways involved in the thiopurine class of drugs (including azathioprine, mercaptopurine, and 6-thioguanine) mechanisms of action and metabolism are well documented (93). As with gemcitabine, multiple enzymes (for example, *HPRT1, IMPDH1, GMPs,* and *TPMT*) are required to convert the drugs into their active metabolites before they are incorporated into RNA and DNA to exert cytotoxicity. Similarly, genes involved in the pathway (i.e. *NQO1, NOS3, XDH, TOP2A, NFKB1*) and transport (i.e. *ABCC1, ABCB1, RALBP1, SLC22A16*) of doxorubicin have also been previously described (94). These gene sets are strong candidates for use in the development of SVMs to predict chemosensitivity to the respective drugs (using their gene expression and/or copy number values), as the genes playing a role in drug disposition within the tumour itself. Therefore, it is reasonable to expect that changes in expression or copy number of the genes identified may predict the

effectiveness of thiopurine, doxorubicin, or other drugs with similar information. Conversely, tamoxifen metabolism largely takes place in the liver by multiple genes from the cytochrome P450 (CYP) and the UDP glucuronosyltransferases (UGT) families (95), and the exact downstream mechanism of action is not well documented. Measuring breast tumour expression or copy number of the CYP and UGT genes would not be informative, because this is not where the metabolism occurs for these drugs. We did not find the cytochrome P450 (CYP) genes to be informative for paclitaxel or gemcitabine, as they were not included in the final SVMs. Therefore, the approach described in this thesis may not be suitable for tamoxifen or other drugs with limited knowledge beyond the fact that their metabolism takes place in the liver. Eight genes were included in the final paclitaxel SVM that were not implicated in paclitaxel's disposition, but were previously implicated in resistance (*FGF2*, *TMEM243*, *BIRC5*, *CSAG2*, *FN1*, *NFKB2*, *TLR6*, *TWIST1*). These genes improved the accuracy of the SVM, indicating ancillary data would be useful in generating chemosensitivity models for other drugs (Figure 6.3). However, there were no additional genes in the gemcitabine analysis other than those directly in the drug pathway, indicating that they are not necessary for developing a successful model.

6.4 Thesis impact on personalized medicine in breast cancer

There are still a number of challenges that researchers and healthcare providers face regarding data analysis, management, and interpretation. This thesis describes improvements upon the techniques that are increasingly used for clinical care. Although this thesis focuses on leveraging genomic technologies to advance our knowledge in breast cancer, all of the techniques and methods described could be applied to other disease types.

There are many cases where point mutations in specific genes are relevant for cancer patient management in regard to predicting outcome or response to treatment. For example, Afatinib was found to be active in non-small-cell lung cancer in patients harboring uncommon *EGFR* mutations (96). The application of the Shannon Pipeline and Veridical for splicing mutation prediction and validation in the analysis of any tumour

type can expand current efforts to detect potentially damaging and relevant mutations. In breast cancer, this thesis found that a large subset of synonymous mutations identified by the TCGA to actually affect mRNA splicing. Synonymous mutations are usually not considered in downstream analyses (beyond variant detection), which would leave some potentially relevant or crucial mutations unreported due to a misidentification of their true effect on the protein product.

Machine learning is proving to be a robust tool in interpreting features and making predictions using large biological datasets (97,98). The biologically-driven machine learning approach described in this thesis could be employed for additional cancer types that are treated with generic chemotherapy agents. While there is no single recipe that will assure successful prediction of chemotherapy response, there are a number of key considerations that need to be accounted for in applying this approach. Specifically:

1) The quality of the tissues analyzed or data obtained should be verified before their application to this type of study. The availability of large genomic data sets with drug response information for a specific type of cancer are crucial for training and testing the predictive SVMs. Resources such as the Gene Expression Omnibus have greatly improved access to this type of data, which are usually made available from previous studies. However, this thesis and other studies (84) have described the level of degradation of nucleic acids in FFPE samples can be variable between tumours, and should be considered when conducting any study;

2) The training data needs to be representative of the tumour type as a whole, and contain roughly equal numbers of sensitive and resistant samples. In this thesis, we demonstrated that cell lines are both a practical and minimally invasive tool; one that can be used to generate gene signatures. However, SVMs perform the best when trained on equal (or close to equal) numbers of data sets in each of the binary classes (i.e. resistant or sensitive). We found that using 44-49 cell lines was sufficient, but when reducing this set by half, it was not adequate for the creation of robust models (data not shown). In addition, the dynamic ranges of GI50 observations did not appear to greatly affect SVM

performance, as paclitaxel GI50s were between 6.5-8.5, and gemcitabine GI50s were between 2.5-9.

3) Any SVM model generated (including the ones described in this thesis) would need to be validated on multiple independent patient data sets before they could be applied to patient treatment, where the outcome of the SVM may alter the course of therapy. For any biomarker, the FDA (or Health Canada) requires extensive analytical validation, clinical validation, and clinical qualification before it is approved to be used in the clinic (99). This level of validation was beyond the scope of this thesis, although we apply the paclitaxel derived SVMs to two different patient datasets. Ultimately, how these type of signatures perform in other patient groups would need to be determined before clinical adoption (as has been done for commercial diagnostic/prognostic assays (100,101).

4) SVM models would likely need to be derived for each tumour type separately. As described in Section 5.3.6, the genes distinguishing tissue specific expression classes dominated those associated with chemotherapy resistance in previous studies employing machine learning (102), but was not true for regression models (103). For example, tissue specific expression patterns were dominant when using genome-wide data with the random forest method (unsupervised machine learning) used by Daemen et al (2013). A benefit to the biologically-driven approach is that SVMs have a greater likelihood of success when using a limited number of attributes (i.e. gene parameters).

5) The genes selected should be relevant to chemotherapy response, and play a role in drug disposition within the tumour itself (as outlined in paragraph 5 of section 6.3.2). Pathways and genes that contribute to resistance in other less well-studied drugs may not be known, and the lack of these features in the SVM would lower prediction accuracy. There may be additional genes or biological functions involved in paclitaxel and/or gemcitabine mechanism of action that are not yet known, which may explain why the SVM is not able to predict drug sensitivity in 15-20% of cases. Alternatively, the these cases may habour point mutations in the present set of genes, or others, that are leading to chemosensitivity, which are not included in the current SVM models.

6) Unrelated prognostic indicators do not appear to be synergistic with our derived gene signatures. In section 5.3.5, this thesis demonstrated that gene signatures using randomly selected expression values (from 23,030 genes) that are able to predict cell line response to paclitaxel were not accurate in predicting response in a patient data set. Many of the genes included in these random signatures were pseudogenes and genes unrelated to the biology of a tumour or paclitaxel metabolism, which is an indication that the signatures derived from them could be statistical artifacts.

The only random signature, among the 10,000 that were derived, that was able to predict patient sensitivity in an external patient data set contained *WWP1*, which has been previously identified as a prognostic indicator for breast cancer (104). Adding *WWP1* to the paclitaxel SVM, however, greatly increased the misclassification rate of predicting cell line response (18% to 26%), and increased the number of patients predicted to be sensitive that were actually non-responsive. *WWP1* has not been previously identified as having a role in paclitaxel drug disposition, indicating that adding generic patient-outcome related genes that are not pertinent to biologically meaningful signatures of drug response, may not be an effective strategy to improve SVM accuracy.

7) One strategy worth considering for improving SVM performance is to stratify tumours by subtype (and/or receptor status) in concert with chemotherapy response. We showed in Chapter 4 of this thesis that different subtypes have diverse splicing mutation profiles, specifically that NCAM-related pathway mutations appear to be preferentially enriched in basal-like and HER2-enriched lymph node positive tumours (section 4.3.8). In our analysis of SVMs derived using random sets of genes, we found that one signature containing *WWP1* could be related to patient outcome. It has been suggested that *WWP1* plays a role in apoptosis in ER positive breast cancer (105), so although it did not improve the paclitaxel SVM for all tumour types, *WWP1* incorporation into an ER positive-specific SVM may possibly increase the classification accuracy for this subset of tumours. However, there were insufficient numbers of ER positive cell lines available for SVM training, and too few patients available with known ER phenotype to test its accuracy. Incorporating additional subtype-specific genes to the current SVM models could be one strategy that might increase the accuracies of these gene signatures.

8) In order to successfully incorporate any genomic signature for clinical application (whether for breast cancer or other tumour types), the expression and copy number studies would need to be performed within the clinically relevant time window either preceding or early on in chemotherapy treatment. For solid tumours, the assay would need to be completed in the timeframe between surgical removal, or biopsy of the tumour tissue, and the onset of treatment. Although this time frame will vary on a case-by-case basis, it would be advantageous and more feasible to accurately measure a small set (10-15) of expression and copy number values compared to performing larger scale (complete genome or exome) determination and analyses.

6.5 References

1. Canadian Cancer Society's Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2014. Toronto, ON: Canadian Cancer Society. (2014).

2. Draghici, S., Khatri, P., Eklund, A. C. & Szallasi, Z. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet. TIG* **22,** 101–109 (2006).

3. MAQC Consortium *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24,** 1151–1161 (2006).

4. Qi, Y. *et al.* Reproducibility of Variant Calls in Replicate Next Generation Sequencing Experiments. *PloS One* **10,** e0119230 (2015).

5. Pinto, D. *et al.* Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat. Biotechnol.* **29,** 512–520 (2011).

6. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11,** 733–739 (2010).

7. Molloy, M. P., Brzezinski, E. E., Hang, J., McDowell, M. T. & VanBogelen, R. A. Overcoming technical variation and biological variation in quantitative proteomics. *Proteomics* **3,** 1912–1919 (2003).

8. Barry, W. T. *et al.* Intratumor heterogeneity and precision of microarray-based predictors of breast cancer biology and clinical outcome. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **28,** 2198–2206 (2010).

9. Macaulay, I. C. & Voet, T. Single Cell Genomics: Advances and Future Perspectives. *PLoS Genet* **10,** e1004126 (2014).

10. Navin, N. E. Cancer genomics: one cell at a time. *Genome Biol.* **15,** 452 (2014).

11. Ni, X. *et al.* Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 21083–21088 (2013).

12. Rehm, H. L. *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med.* **15,** 733–747 (2013).

13. South, S. T. *et al.* ACMG Standards and Guidelines for constitutional cytogenomic microarray analysis, including postnatal and prenatal applications: revision 2013. *Genet. Med.* **15,** 901–909 (2013).

14. Gargis, A. S. *et al.* Good laboratory practice for clinical next-generation sequencing informatics pipelines. *Nat. Biotechnol.* **33,** 689–693 (2015).

15. Gargis, A. S. *et al.* Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat. Biotechnol.* **30,** 1033–1036 (2012).

16. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical

Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17,** 405–423 (2015).

17. Ward, L. D. & Kellis, M. Interpreting non-coding variation in complex disease genetics. *Nat. Biotechnol.* **30,** 1095–1106 (2012).

18. Mucaki, E. J., Shirley, B. C. & Rogan, P. K. Prediction of Mutant mRNA Splice Isoforms by Information Theory-Based Exon Definition. *Hum. Mutat.* **34,** 557–565 (2013).

19. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2015).

20. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

21. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486,** 395–399 (2012).

22. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486,** 405–409 (2012).

23. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486,** 400–404 (2012).

24. Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486,** 353–360 (2012).

25. Mardis, E. R. The $1,000 genome, the $100,000 analysis? *Genome Med.* **2,** 84 (2010).

26. International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).

27. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).

28. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).

29. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al* **Chapter 7,** Unit7.20 (2013).

30. Flanagan, S. E., Patch, A.-M. & Ellard, S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet. Test. Mol. Biomark.* **14,** 533–537 (2010).

31. Khan, W. A., Knoll, J. H. & Rogan, P. K. Context-based FISH localization of genomic rearrangements within chromosome 15q11.2q13 duplicons. *Mol. Cytogenet.* **4,** 15 (2011).

32. Knoll, J. H. M. & Rogan, P. K. Sequence-based, in situ detection of chromosomal abnormalities at high resolution. *Am. J. Med. Genet. A.* **121A,** 245–257 (2003).

33. Chua, C. *et al.* Phase II study of trastuzumab in combination with S-1 and cisplatin in the first-line treatment of human epidermal growth factor receptor HER2-positive advanced gastric cancer. *Cancer Chemother. Pharmacol.* (2015). doi:10.1007/s00280-015-2811-y

34. Miller, D. T. *et al.* Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.* **86,** 749–764 (2010).

35. Liang, L., Fang, J.-Y. & Xu, J. Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy. *Oncogene* (2015). doi:10.1038/onc.2015.209

36. Bang, Y.-J. *et al.* Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet Lond. Engl.* **376,** 687–697 (2010).

37. Moroni, M. *et al.* Gene copy number for epidermal growth factor receptor (EGFR) and clinical response to antiEGFR treatment in colorectal cancer: a cohort study. *Lancet Oncol.* **6,** 279–286 (2005).

38. Hirsch, F. *et al.* Combination of EGFR gene copy number and protein expression predicts outcome for advanced non-small-cell lung cancer patients treated with gefitinib. *Ann. Oncol.* **18,** 752–760 (2006).

39. Go, H. *et al.* High MET Gene Copy Number Leads to Shorter Survival in Patients with Non-small Cell Lung Cancer: *J. Thorac. Oncol.* **5,** 305–313 (2010).

40. Bambury, R. M. *et al.* DNA copy number analysis of metastatic urothelial carcinoma with comparison to primary tumors. *BMC Cancer* **15,** (2015).

41. Shaw, A. T. *et al.* Crizotinib in *ROS1* -Rearranged Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **371,** 1963–1971 (2014).

42. Shaw, A. T. *et al.* Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring ALK gene rearrangement: a retrospective analysis. *Lancet Oncol.* **12,** 1004–1012 (2011).

43. Kearney, H. M. *et al.* American College of Medical Genetics recommendations for the design and performance expectations for clinical genomic copy number microarrays intended for use in the postnatal setting for detection of constitutional abnormalities. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **13,** 676–679 (2011).

44. Cooley, L. D. *et al.* American College of Medical Genetics and Genomics technical standards and guidelines: microarray analysis for chromosome abnormalities in neoplastic disorders. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **15,** 484–494 (2013).

45. Asadollahi, R. *et al.* The clinical significance of small copy number variants in neurodevelopmental disorders. *J. Med. Genet.* **51,** 677–688 (2014).

46. Lenoir, T. & Giannella, E. The emergence and diffusion of DNA microarray technology. *J. Biomed. Discov. Collab.* **1,** 11 (2006).

47. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27,** 182–189 (2009).

48. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-3.0. (1996). at <http://www.repeatmasker.org>

49. Shirley, B. C. *et al.* Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences. *Genomics Proteomics Bioinformatics* **11,** 77–85 (2013).

50. Viner, C., Dorman, S. N., Shirley, B. C. & Rogan, P. K. Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. *F1000Research* **3,** 8 (2014).

51. Wang, L., Nie, J. J. & Kocher, J.-P. A. PVAAS: identify variants associated with aberrant splicing from RNA-seq. *Bioinformatics* **31,** 1668–1670 (2015).

52. Mudvari, P. *et al.* SNPlice: variants that modulate Intron retention from RNA-sequencing data. *Bioinforma. Oxf. Engl.* **31,** 1191–1198 (2015).

53. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

54. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).

55. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467 (1977).

56. Ellis, S. E. *et al.* RNA-Seq optimization with eQTL gold standards. *BMC Genomics* **14,** 892 (2013).

57. Yu, K.-D., Jiang, Y.-Z. & Shao, Z.-M. Difference between observed and expected number of involved lymph nodes reflects the metastatic potential of breast cancer independent to intrinsic subtype. *Oncotarget* (2015).

58. Armitage, P. & Doll, R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer* **8,** 1–12 (1954).

59. Greaves, M. Cancer causation: the Darwinian downside of past success? *Lancet Oncol.* **3,** 244–251 (2002).

60. Weiss, R. A. Multistage carcinogenesis. *Br. J. Cancer* **91,** 1981–1982 (2004).

61. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471,** 467–472 (2011).

62. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91,** 1033–1040 (2012).

63. Woo, Y. H. & Li, W.-H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat. Commun.* **3,** 1004 (2012).

64. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463,** 184–190 (2010).

65. Korthauer, K. D. & Kendziorski, C. MADGiC: a model-based approach for identifying driver genes in cancer. *Bioinformatics* **31,** 1526–1535 (2015).

66. Bergamaschi, A. *et al.* Extracellular matrix signature identifies breast cancer subgroups with different clinical outcome. *J. Pathol.* **214,** 357–367 (2008).

67. Triulzi, T. *et al.* Neoplastic and Stromal Cells Contribute to an Extracellular Matrix Gene Expression Profile Defining a Breast Cancer Subtype Likely to Progress. *PLoS ONE* **8,** e56761 (2013).

68. Triulzi, T., Orlandi, R. & Tagliabue, E. Stromal Responses among Carcinomas—Letter. *Clin. Cancer Res.* **20,** 1396–1396 (2014).

69. Jeney, A. & Harisi, R. Extracellular matrix as target for antitumor therapy. *OncoTargets Ther.* 1387 (2015). doi:10.2147/OTT.S48883

70. Leong, H. S., Chambers, A. F. & Lewis, J. D. Assessing cancer cell migration and metastatic growth in vivo in the chick embryo using fluorescence intravital imaging. *Methods Mol. Biol. Clifton NJ* **872,** 1–14 (2012).

71. Armstrong, P. B., Quigley, J. P. & Sidebottom, E. Transepithelial invasion and intramesenchymal infiltration of the chick embryo chorioallantois by tumor cell lines. *Cancer Res.* **42,** 1826–1837 (1982).

72. Deryugina, E. I. & Quigley, J. P. Chick embryo chorioallantoic membrane model systems to study and visualize human tumor cell metastasis. *Histochem. Cell Biol.* **130,** 1119–1130 (2008).

73. Thurman, R. E., Day, N., Noble, W. S. & Stamatoyannopoulos, J. A. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.* **17,** 917–927 (2007).

74. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 139–144 (2010).

75. Cory, G. Scratch-wound assay. *Methods Mol. Biol. Clifton NJ* **769,** 25–30 (2011).

76. Albini, A. *et al.* A rapid in vitro assay for quantitating the invasive potential of tumor cells. *Cancer Res.* **47,** 3239–3245 (1987).

77. Hedley, B. D. & Chambers, A. F. Tumor dormancy and metastasis. *Adv. Cancer Res.* **102,** 67–101 (2009).

78. Brackstone, M., Townson, J. L. & Chambers, A. F. Tumour dormancy in breast cancer: an update. *Breast Cancer Res. BCR* **9,** 208 (2007).

79. André, F. & Zielinski, C. C. Optimal strategies for the treatment of metastatic triple-negative breast cancer with currently approved agents. *Ann. Oncol.* **23,** vi46–vi51 (2012).

80. Jerusalem, G., Rorive, A. & Collignon, J. Chemotherapy options for patients suffering from heavily pretreated metastatic breast cancer. *Future Oncol. Lond. Engl.* **11,** 1775–1789 (2015).

81. Venet, D., Dumont, J. E. & Detours, V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* **7,** e1002240 (2011).

82. Drier, Y. & Domany, E. Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? *PloS One* **6,** e17795 (2011).

83. Do, H. & Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin. Chem.* **61,** 64–71 (2015).

84. Choudhary, A. *et al.* Evaluation of an integrated clinical workflow for targeted next-generation sequencing of low-quality tumor DNA using a 51-gene enrichment panel. *BMC Med. Genomics* **7,** 62 (2014).

85. Natrajan, R. C. Breast cancer heterogeneity: parallel evolution or conscious uncoupling? *J. Pathol.* (2015). doi:10.1002/path.4557

86. Eirew, P. *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518,** 422–426 (2015).

87. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461,** 809–813 (2009).

88. Desmedt, C. *et al.* Uncovering the genomic heterogeneity of multifocal breast cancer. *J. Pathol.* (2015). doi:10.1002/path.4540

89. He, D.-X., Xia, Y.-D., Gu, X.-T., Jin, J. & Ma, X. A 20-gene signature in predicting the chemoresistance of breast cancer to taxane-based chemotherapy. *Mol. Biosyst.* **10,** 3111–3119 (2014).

90. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42,** D472–477 (2014).

91. Lee, S.-Y. *et al.* Genetic polymorphisms of SLC28A3, SLC29A1 and RRM1 predict clinical outcome in patients with metastatic breast cancer receiving gemcitabine plus paclitaxel chemotherapy. *Eur. J. Cancer Oxf. Engl. 1990* **50,** 698–705 (2014).

92. Jørgensen, C. L. T. *et al.* Gene aberrations of RRM1 and RRM2B and outcome of advanced breast cancer after treatment with docetaxel with or without gemcitabine. *BMC Cancer* **13,** 541 (2013).

93. Zaza, G. *et al.* Thiopurine pathway. *Pharmacogenet. Genomics* **20,** 573–574 (2010).

94. Thorn, C. F. *et al.* Doxorubicin pathways: pharmacodynamics and adverse effects. *Pharmacogenet. Genomics* **21,** 440–446 (2011).

95. Klein, D. J. *et al.* PharmGKB summary: tamoxifen pathway, pharmacokinetics. *Pharmacogenet. Genomics* **23,** 643–647 (2013).

96. Yang, J. C.-H. *et al.* Clinical activity of afatinib in patients with advanced non-small-cell lung cancer harbouring uncommon EGFR mutations: a combined post-hoc analysis of LUX-Lung 2, LUX-Lung 3, and LUX-Lung 6. *Lancet Oncol.* **16,** 830–838 (2015).

97. Libbrecht, M.W., Noble, W.S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**(6), 321-32 (2015).

98. Kourou, K., *et al.* Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13**, 8-17 (2015).

99. Kelloff, G.J., *et al.* Biomarker development in the context of urologic cancers. *Urol. Oncol.* **33**(6), 295-301 (2015).

100. Van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–536 (2002).

101. Paik, S. *et al.* A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *N. Engl. J. Med.* **351,** 2817–2826 (2004).

102. Daemen, A. *et al.* Modeling precision treatment of breast cancer. *Genome Biol.* **14,** R110 (2013).

103. Geeleher, P., Cox, N. J. & Huang, R. S. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol.* **15,** R47 (2014).

104. Nguyen Huu, N. S. *et al.* Tumour-promoting activity of altered WWP1 expression in breast cancer and its utility as a prognostic indicator. *J. Pathol.* **216,** 93–102 (2008).

105. Zhou, Z., Liu, R. & Chen, C. The WWP1 ubiquitin E3 ligase increases TRAIL resistance in breast cancer. *Int. J. Cancer J. Int. Cancer* **130,** 1504–1510 (2012).

Appendices

Appendix S1: Copyright permission for Chapters 2-4

Appendix S1.1 Copyright permission for Chapter 2

Nucleic acids research

Order detail ID: 67483881
Order License Id: 3653800184419

Article Title: Expanding probe repertoire and improving reproducibility in human genomic hybridization.
Author(s): Dorman, Stephanie N ; et al
DOI: 10.1093/NAR/GKT048
Date: Apr 01, 2013
ISSN: 1362-4962
Publication Type: e-Journal
Volume: 41
Issue: 7
Start page: e81
Publisher: OXFORD UNIVERSITY PRESS
Author/Editor: Oxford University Press

Permission Status: Granted

Permission type: Republish or display content
Type of use: reuse in a thesis/dissertation

View details

Appendix S1.2 Copyright permission for Chapter 3

Copyright: © 2014 Viner C *et al.* This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Appendix S1.3 Copyright permission for Chapter 4

Scientific Reports

Order detail ID: 67483882
Order License Id: 3653800186241

ISSN: 2045-2322
Publication Type: e-Journal
Volume:
Issue:
Start page:
Publisher: Nature Publishing Group

Permission Status: Granted

Permission type: Republish or display content
Type of use: Republish in a thesis/dissertation

View details

Appendix S1.4 Copyright permission for Chapter 5

Molecular oncology

Order detail ID: 68373466
Order License Id: 3711970806556

ISSN: 1574-7891
Publication Type: Journal
Volume:
Issue:
Start page:
Publisher: ELSEVIER BV

Permission Status: Granted

Permission type: Republish or display content
Type of use: Thesis/Dissertation

View details

Appendix S2: Supplementary Information for Chapter 2

Appendix S2.1 Supplementary Methods: *Ab Initio* single copy (sc) sequence algorithm and implementation

The *ab initio* method eliminates the requirement to exclude sequences from a catalog of consensus-like repetitive elements. It exploits a state space search strategy in which a depth-limited search is run repeatedly, increasing the depth limit with each iteration, until it reaches the depth of the shallowest level, in order to determine the copy number of seed subsequences of a larger input sequence (e.g. a complete chromosome). In each iteration, progressively shorter sequences containing elements present in multiple copies in the genome are searched in a sequenced genome, at low stringency using BLAT (BLAST-like alignment tool), in parallel (using threaded jobs) on a cluster computer. To define the boundaries of sc segments, the above steps are recursively run on branched subsequences of repeat-containing intervals that occur adjacent to sc segments discovered in the previous step. In addition to finding known repeat sequence families, *ab initio* eliminates repeat elements, segmental duplicons, and conserved paralogs that are not filtered out by catalogue-based approaches. The algorithm is tuned to exclude highly and moderately conserved multicopy and/or repetitive sequences, but not highly divergent repetitive elements. The algorithm can be applied to any genome.

A secondary screen using multiprocessor BLAST analysis (54) filtered out any residual repetitive sequences. Parameters were selected to maximize speed without compromising sensitivity. The default parameters were modified to return 2 sequence alignments, using a word size set to 28, the number of best hits kept limited to 2, descriptions of 5 sequences retained, and an expected hit value threshold of 0.1. This threshold produced significant alignments ≤50 base pairs in length to genomic targets, when present. The parameters provided a reasonable level of genomic resolution and adequate sensitivity to detect nearly all conserved or moderately conserved repeat elements, while exhibiting performance suitable for genome-scale application. The average run time for the recursive BLAT runs, followed by filtering apparent sc results with mpiBLAST, using a 128 CPU Xeon-based compute cluster was 19 hours 20 minutes for a chromosome length of~130 Mb.

We then compared the *ab initio* genomic regions output with a deduced set of annotated, non-repetitive intervals to determine the sensitivity and specificity of the algorithm. The comparison set comprised the genomic complement of the combined set including segmental duplication, self-chained paralogous intervals, and repeat-masked sequences.

Appendix S2.2 Coordinates and PCR primers of validated scFISH probes

Gene Target	Genomic Coordinates	Probe Length (bp)	Primers	Hybridization Efficiency*
ERBB2	chr17: 37861155-37863542	2388	**L** GCATTGGGAGAATTAGTGTGTATTTATGTTG **R** GTTAGATGTTAGAAAGGACTTCCTGGTTGAG	96/68
CDKN2A (Probe 1)	chr9: 21991990-21995076	3087	**L** GTAAATGCACCAAGGTAGAAGTAACAAATCA **R** GTTTAGTTTAATTTCGCTTGTTTTCCAAATCT	100/79.8
CDKN2A (Probe 2)	chr9: 21981743-21985184	3442	**L** TAGTTCTACCACCTACTTTGTTACCCTGAAAA **R** TATATTTCATCAAGAAGTTGATTCCCTTGAGT	97.7/75.9
CDKN2A (Probe 3)	chr9: 21984688-21987911	3224	**L** TTTCACTGATAGGTTTAACACTGGTTTAGGAT **R** AATCTGCATTTTAAATAAACACTTGAAGGAGA	91.4/75.4
TP53	chr17: 7589527-7592796	3270	**L** CAAAGCTAGATAACAGGTAGATTGTTTTTCC **R** TAGAAGACACAAACTGCTAGATAAAATGTAAGC	95.7/70.3
CCND1 (Probe 1)	chr11: 69458658-69461950	3293	**L** ACGATTTCATTGAACACTTCCTCTCCAAAAT **R** CTGATGTAGCCCAACAATTCCAGTGACTT	100/94.1
CCND1 (Probe 2)	chr11: 69465465-69469037	3573	**L** ACATGGAGAGGTTAAGTCTGAAAAGGCTGA **R** CTCTCGATACACACAACATCCAGGACTTG	100/77.9
NOTCH1	chr9: 139435414-139438778	3365	**L** CCCAGCTCTCCTCAAAACAAAGAGAAAAA **R** TGACTACAGAACTCTGGGCAGAATGTTGA	100/73.9

scFISH probe primer design: Gene targets, genomic location, probe length and primers used for each validated probe. *Hybridization efficiencies are indicated as the percent of cells with both homologues clearly hybridized, preceded by the percent of cells that had at least one homologue hybridized to the correct chromosome band. Genomic coordinates are based on NCBI Build 37/hg19. L = left, R = right.

Appendix S3: Supplementary Information for Chapter 3

Appendix S3.1 Veridical variant input format

This input format most easily accepts formatted output from the Shannon Pipeline. In particular, all variants of interest should be concatenated into a single file. Once a, tab-delimited, concatenated file has been generated, it can easily be formatted correctly by using **FilterShannonPipelineResults.pl**. All file headers must precisely match their outlined schema. One can also manually ensure the following: the header line has no quotation marks or special characters, empty columns have been replaced by a period (.) and each variant line contains only a single gene (comma-delimited gene lists must be split such that there is only one gene per line). If one wishes Veridical to consider variants pertaining to more than one experimental sample, a comma-delimited list of experimental samples, in the form of BAM file names, must be provided as the **key** column. The **key** column must always contain at least one file name that is present as the base name of one of the files listed in the BAM file list that must be passed to Veridical.

Alternatively, one can prepare the input format as follows. The header must contain at least the following, case-insensitive, values to which the file's columns must adhere to: chromosome, splice&coordinate, strand, type, gene, location, location_type, heterozygosity, variant, input, key. The column headers need only contain the given text (i.e. a column labeled **gene_name** would be sufficient to satisfy the above requirement for a "gene" column). Column headers with ampersands (&) denote that all words joined by this symbol must be present for that column (i.e. **Splice_site_coordinate** satisfies the "splice&coordinate" requirement). The order of the columns is immaterial. The **input** column can contain any identifier for the variant and need not be unique. The **location** column specifies if the site is natural or cryptic. For Veridical, all that matters is that cryptic variants contain the word "**cryptic**" as part of their value in this column and that non-cryptic variants do not. The **location_type** column is only used for cryptic variants and specifies if the variant is intronic or exonic. It is not currently used by the program. This column must be present but can always be set to null (i.e).

A few rows from a sample variant file is provided below (text wrapped for readability):

```
Chromosome     Splice_site_coordinate      Strand     Ri-initial Ri-
final ΔRi Type  Gene_Name  Location Location_Type Loc._Rel._to_exon
Dist._from_nearest_nat._site  Loc._of_nearest_nat._site
Ri_of_nearest_nat Cryptic_Ri_rel._nat. rsID  Average_heterozygosity
Variant_coordinate    Input_variant      Input_ID RNASeqDirectory_ID
RNA_Seq_BAM_ID_KEY
chr10 89711874 + 12.09 -2.62 -14.71 ACCEPTOR PTEN     NATURALSITE . . .
. . . . . 89711873 A/G ID1  dir     file
chr10 89712017 + 5.18   -1.85  -7.03  DONOR  PTEN     NATURALSITE . . .
. . . . . 89712018 T/C ID1  dir     file
chrX   9621719 + -4.78  2.25  7.03   DONOR  TBL1X     CRYPTICSITE
EXONIC . 11  9621730  2.24  GREATER .     . 9621720 C/T ID1 dir file
```

Veridical exome annotation input format

This input format can be generated via **ConvertToExomeAnnotation.pl**. The file must be tab-delimited, excepting its header, which must be comma-delimited. It must have the following, case-insensitive, header columns, to which its data must adhere: transcript, chromosome, exon chr start, exon chr end, exon rank, gene. The column headers need only contain the given text (i.e. a column labeled **gene_name** would be sufficient to satisfy the above requirement for a "gene" column). The order of the columns is immaterial.

A few rows from a sample exome annotation file is provided below (text wrapped for readability):

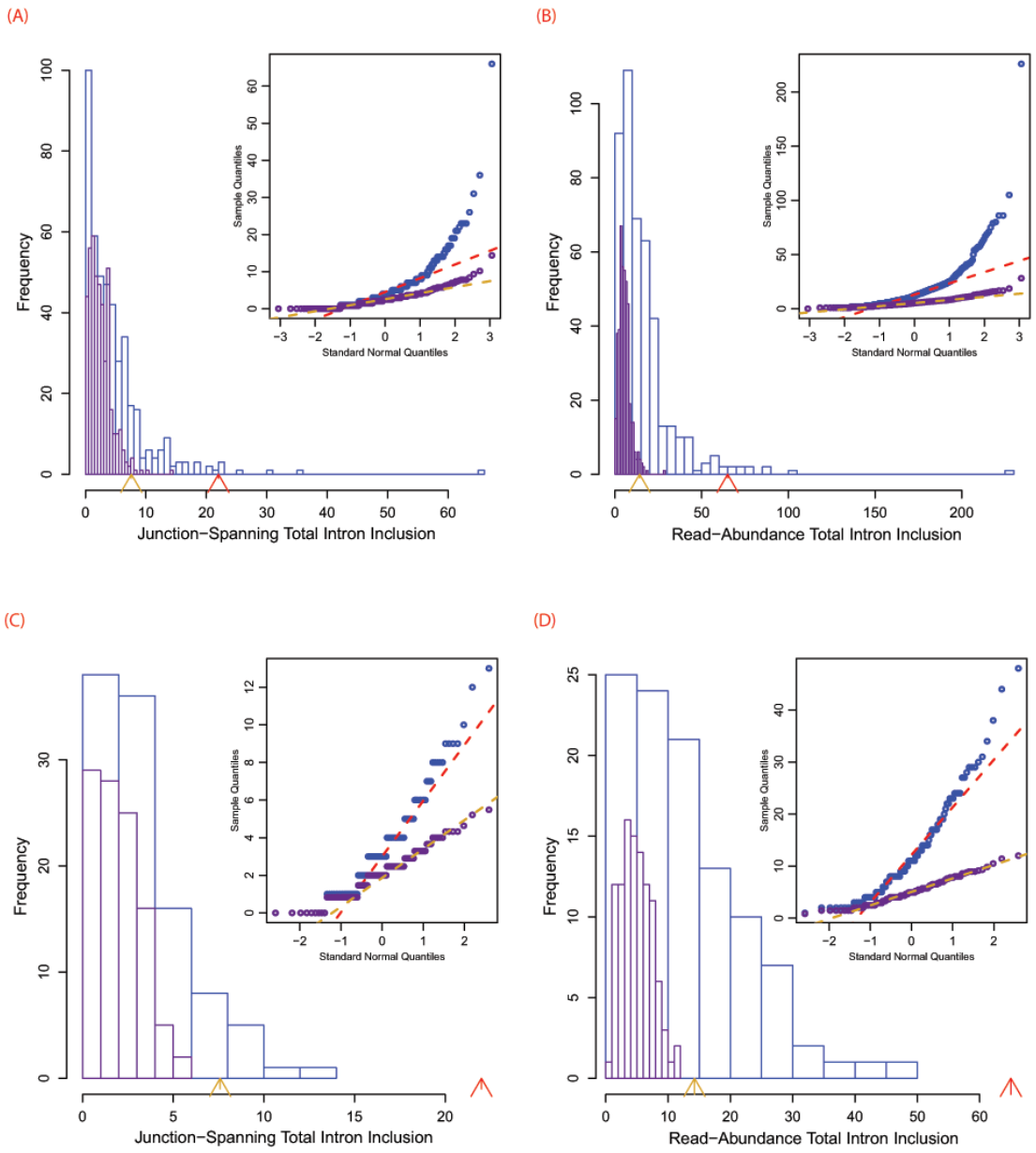
```
Transcript ID,ID,ID,Chromosome Name,Strand,     Exon Chr Start,Exon Chr
End,  Exon Rank in Transcript,Transcript Start, Transcript End,
Associated Gene Name
NM_213590 NM_213590   NM_213590 chr13  +  50571142    50571899 1
50571142   50592603 TRIM13
NM_213590 NM_213590   NM_213590 chr13  +  50586070    50592603 2
50571142   50592603 TRIM13
NM_198318 NM_198318   NM_198318 chr19  +  50180408    50180573 1
50180408   50191707 PRMT1
```

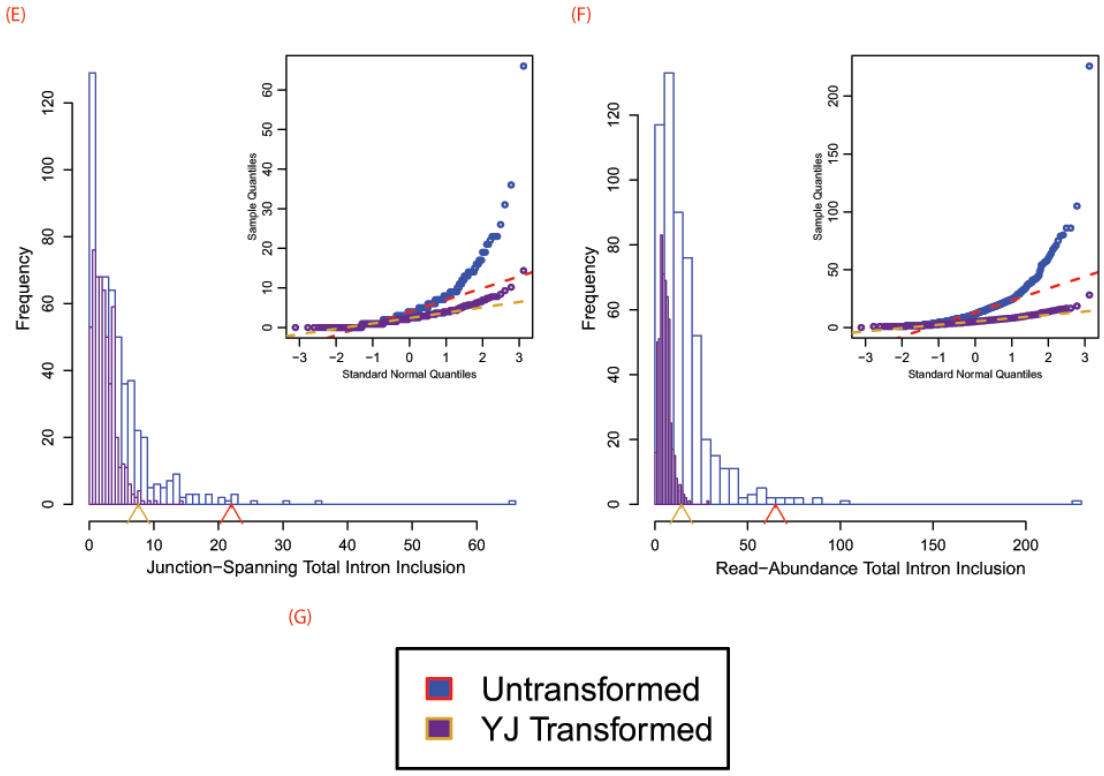
Appendix S3.2 Veridical output

If a variant contains any validating reads, Veridical outputs the variant in question, along with some summary information and a table specifying the numbers of each validating read type detected for both the experimental and control samples. Within the output of Veridical, the phrase: "Validated (*x*) variant *n* times" means that the variant was validated

mainly for splicing consequence *x* and has *n* validating reads. The variant will only appear within the ***.filtered** output file if the p-value for either junction-spanning or read-abundance-based reads for splicing consequence *x* was statistically significant (defined, by default, as: $p < 0.05$). After the variant being validated is provided, along with its primary predicted splicing consequence, the output is divided into two sections with identical contents: one for the experimental sample(s) and another for control samples. The summary enumerates the number of reads of each splicing consequence, partitioned by evidence type (junction-spanning or read-abundance-based), and by sample type (tumour or normal for control samples, and only tumour for experimental samples). A table describing the number of each read type for every file follows this summary. An example of this output, for the variant within *RAD54L*, as shown by Figure 7 and the last portion of Table 2, is provided. While Veridical outputs this as plain text, with the table in a tab-delimited format, we provide this output as an Excel document with descriptions of the meaning of each table heading, to clarify the presentation of the data. All input and output files for the five variants presented are provided. **VeridicalOutExample.xls** contains the output for the variant within *RAD54L*, along with descriptions of the terms used and the output format. **all.vin** contains the input variant file. **allTumoursBAMFileList.txt** and **allNormalsBAMFileList.txt** are the BAM file lists for tumour and normal samples, respectively. **all.vout** contains the Veridical output. The exome file can be retrieved using **ConvertToExomeAnnotation.pl**, available with the other programs at: www.veridical.org. The BAM file lists contain the TCGA file UUID, followed by a slash, followed by the file name. The RNA-Seq data itself can be downloaded from TCGA at: https://tcga-data.nci.nih.gov/tcga/.

Appendix S3.3 Supplementary Figure 1





Histogram and embedded Q-Q plots portraying the difference between untransformed and Yeo-Johnson (YJ) transformed data. The plots depict intron inclusion for the inactivating mutation (chr12:83359523G>A) within TMTC2, as shown in Figures 3.6(B) and 3.6(C). The arrowheads denote the number of reads in the variant-containing file, which is, in all cases, more than observed in the control samples ($p < 0.01$). The figure legend for all panels is provided in (G), which shows that blue and red plot elements correspond to untransformed data, while yellow and purple correspond to YJ transformed elements. Dotted lines in the Q-Q plots are lines passing through the first and third quantiles for a normal reference distribution. (A), (C), and (E) show junction-spanning based reads, while (B), (D), and (F) show read-abundance-based reads. (A/B) depict tumour sample distributions, (B/C) depict normal sample distributions, and (E/F) depict combined tumour and normal sample distributions.

Appendix S4: Supplementary Information for Chapter 4

Appendix S4.1 SomaticSniper Supplementary Materials

Appendix S4.1.1 Supplementary Methods – Variant Calling Methods

Two independent variant callers, Strelka (1) and SomaticSniper (2), were evaluated. The main analysis performed using results from Strelka, which has greater sensitivity and ability to detect subclonal mutations, by minimizing reporting of spurious variants and germline polymorphisms (3). Additionally, the SomaticSniper methods and results are reported below.

Before running SomaticSniper, all DNA sequencing BAM files were realigned using the Genome Analysis Toolkit (GATK) Indel Realigner program (4). In addition to default parameters, the knownAlleles parameter was used with the well-documented insertions/deletions (indels) files: `Mills_and_1000G_gold_standard.indels.b37.sites.vcf` (5) and `1000G_phase1.indels.b37.vcf` (6), available through the bioinformatic resource Galaxy (7, 8). SomaticSniper data was then post-processed to only include variants with both mapping and somatic qualities of at least 40 (equivalent to running it with `-Q 40 -q 40`).

Appendix S4.1.2 Supplementary Results – SomaticSniper Variant Calling Results

SomaticSniper variant predictions are summarized in Appendix S3.1.3. Notably, there were 1,208 variants from SomaticSniper that are predicted to affect both protein coding and splicing 594 genes. In the SomaticSniper data, mutations classified as both protein coding and splicing variants were found in 383 tumours, with 63 of these variants in *PASD1*, 61 in *PRSS3*, 52 in *NF1*. The variants in these genes, as well as others that were highly mutated, are the exact same genomic location and nucleotide change, suggesting that SomaticSniper reported higher numbers of SNPs (3) that were not annotated with dbSNP135 in >1% of the population, which was used to filter out common SNPs. There were 248 variants in 186 tumours from the SomaticSniper set that were classified as

silent amino acid changes from ANNOVAR, but were revealed to affect splicing from the Shannon Pipeline predictions.

There was relatively low concordance between the two variant callers, which reported variant lists with less than 50% similarity. There were 21,112 protein coding and 1,811 splicing variants common to both Strelka and SomaticSniper. The predicted variants were compared to the previously reported TCGA Level 2 somatic mutations (Appendix S3.1.4). Strelka showed the highest concordance with TCGA mutations, reporting 82.1% of protein coding mutations, and 86.5% of the splicing variants. Conversely, SomaticSniper predicted 73.4% protein coding and 75.3% splicing variants reported by TCGA.

Both of the somatic variant callers we employed utilize Baysian methods to elucidate somatic event probabilities. Strelka and SomaticSniper were found to be the two best variant callers in a comparison by Roberts et al 2013. Additionally, these two are a valuable combination, in that SomaticSniper is useful to generate "a variety of candidate SNV sites without any particular drawbacks", although with a fair amount of false positives, while Strelka is least prone to returning germ-line polymorphisms. The relative stringency of Strelka was our main reason for performing most of our analyses with it, along with the fact that many of its candidates (at probability 0.2) were also returned by other callers. It is worth mentioning that different callers have been found to have poor correlations at the same sites; in particular, Strelka and SomaticSniper were found to have a 0.21 Pearson correlation coefficient in the abovementioned study. Our use of Veridical to validate splicing variants with functional evidence of the mutation significantly resolves the inconsistency between somatic variant callers (for this type of mutation).

Appendix S4.1.3 Variant Summaries by Mutation Type

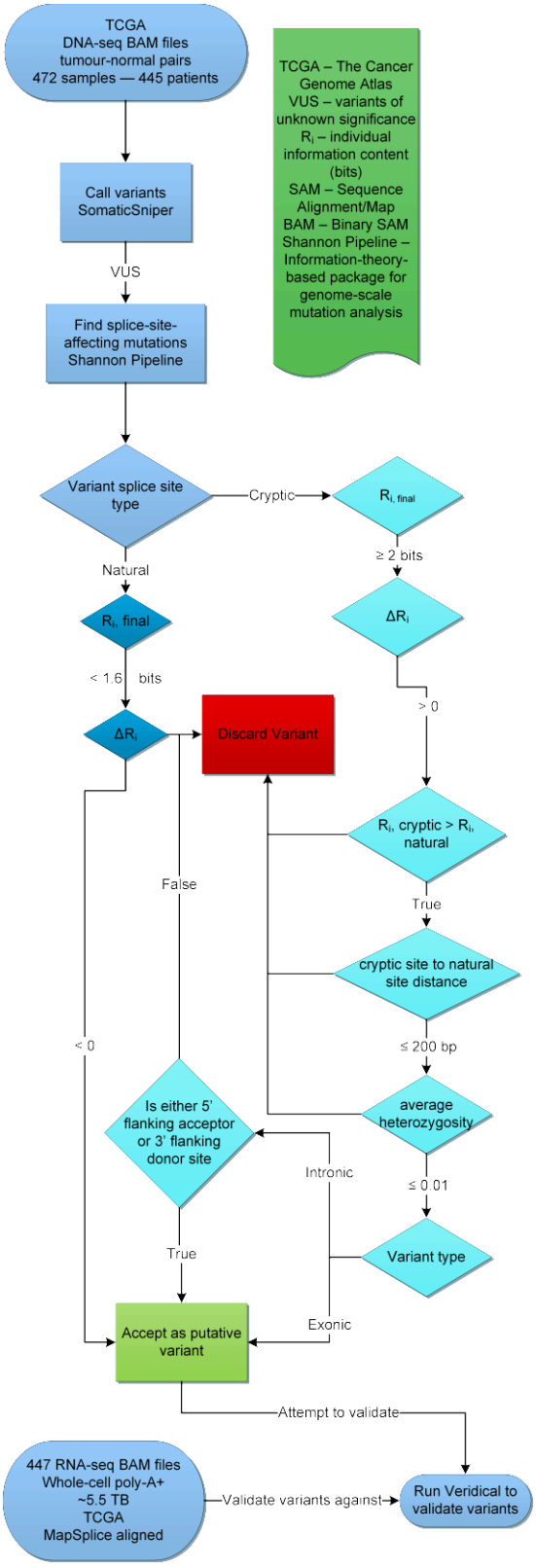
	Somatic Sniper
ANNOVAR protein coding variants	
Synonymous	23,458
Nonsynonymous	52,634
Stop gain or loss	2,127
Total protein coding variants	**78,219**
Shannon Pipeline splicing variants	
Cryptic	6,441
Inactivating	2,685
Leaky	10,648
Total splicing variants	**19,774**
Synonymous	248
Nonsynonymous	905
Stop gain or loss	55
Total	**1,208**
% Synonymous also splicing	1.0572%
% Nonsynonymous also splicing	1.7194%
% Stop gain or loss also splicing	2.5858%

Appendix S4.1.4 SomaticSniper Variants Compared to TCGA Findings

	Total TCGA	TCGA predicted by SomaticSniper
TCGA Protein Coding Variants		
SNVs Validated	5,557	4,365 (77.3%)
SNVs Not Validated	18,197	13,380 (72.2%)
Indels Validated	125	N/A
Indels Not Validated	1,758	N/A
Total	**25,637**	**17,745 (73.4%)**
TCGA Splicing Variants		
SNVs Validated	87	70 (80.5%)
SNVs Not Validated	342	253 (74.0%)
Total	**429**	**323 (75.3%)**

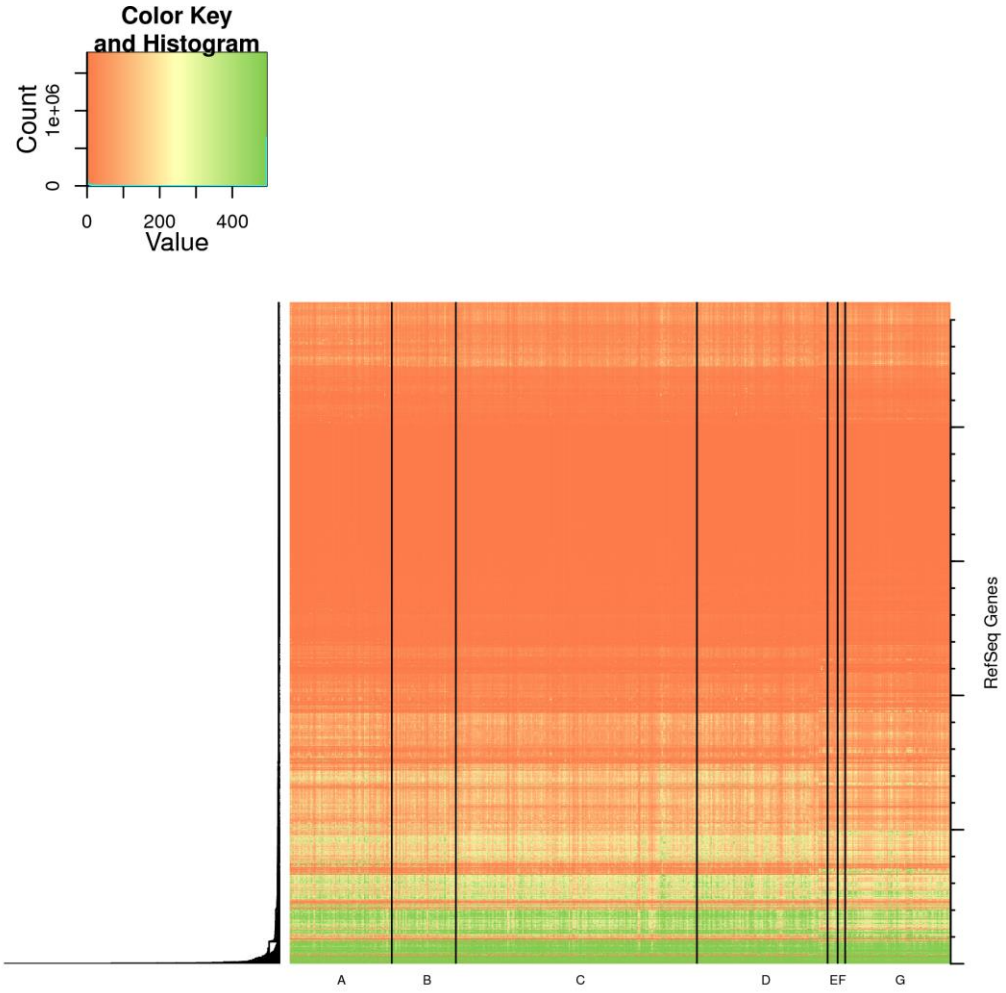
Appendix S4.2 Filtering criteria for splicing mutations

Supplementary Figure S6. Flowchart indicating procedure for filtering splicing mutation variants. Shannon pipeline splicing variants output was filtered using the steps shown in this flowchart to identify those variants that are likely to cause aberrant splicing. Upon identifying variants with Strelka (or Somatic Sniper), the VCF files were submitted to the Shannon splicing mutation pipeline, then categorized as either mutations affecting natural splice sites (3' acceptor, or 5' donor) or cryptic splice site strengths. In a small number of cases, both natural and cryptic splice sites were simultaneously altered. Natural sites that were predicted to be abolished were further considered. Predicted leaky splicing mutations were excluded from the present analysis, since the validation methods for such mutations has not yet been assessed. Aside from standard information theory-based mutation criteria, cryptic splicing mutation candidates were also filtered for proximity to the nearest neighboring natural splice site and population frequency. The filtered variant subset (n = 5,206) was used for all subsequent analyses.



Appendix S4.3 Supplementary Figure 1

RNA-Seq Coverage Heat Map by Subtype. Heatmap depicting coverage per exonic base of TCGA RNA-Seq tumour and normal data. Expression based on RNA-Seq datasets is shown along the x-axis, with tumours first, ordered by subtype, followed by matched normal breast tissues. These categories are demarcated within the heatmap by black vertical lines, which correspond to the sample types: (A) basal-like; (B) HER2-enriched; (C) luminal A; (D) luminal B; (E) tumour, subtype not available; (F) normal-like tumor; and (G) normal control samples. The y-axis consists of all RefSeq genes (with major and minor tick marks every 5,000 and 1,000 genes, respectively), clustered to form a dendrogram, which is visible on the left side of the graph. Genes with low nominal expression levels were below minimum threshold read counts for analysis by Veridical.



Appendix S4.4 Variants compared to those previously published by TCGA

	Total TCGA	No. TCGA mutations predicted
TCGA Protein Coding Variants		
SNVs Validated	5557	5085 (91.5%)
SNVs Not Validated	18197	15742 (86.5%)
Indels Validated	125	44 (35.2%)
Indels Not Validated	1758	170 (9.7%)
Total	**25637**	**21041 (82.1%)**
TCGA Splicing Variants		
SNVs Validated	87	80 (92.0%)*
SNVs Not Validated	342	291 (85.1%)^
Total	**429**	**371 (86.5%)**

*contains two variants that were filtered out based on our filtering criteria

^contains eight variants filtered out

Appendix S4.5 Overrepresentation analysis of TCGA mutations missed by Strelka

A pathway analysis using G:Profiler (Reimand et al. 2011) on the 4,654 TCGA variants, missed by Strelka, revealed 116 overrepresented pathways including development (20), morphogenesis (11), cellular processes (16), regulation (10), ion binding (6) and adhesion (5). Details below.

PATHWAY NAME	Pathway Group	Pathway Depth in Group	P-VALUE	# Genes mutated in pathway	Total # Genes in Pathway
cell adhesion	2	3	4.52E-15	200	1059
biological adhesion	2	2	5.61E-15	200	1061
multicellular organismal development	2	4	1.35E-14	626	4561
homophilic cell adhesion	2	5	4.26E-14	52	140
system development	2	5	5.93E-14	552	3942
single-organism cellular process	2	3	3.90E-13	1325	11241
developmental process	2	2	7.50E-13	685	5169
single-multicellular organism process	2	3	2.51E-12	813	6363
anatomical structure development	2	3	3.59E-12	614	4568
single-organism process	2	2	7.29E-12	1438	12461
multicellular organismal process	2	2	2.59E-11	833	6605
cellular process	2	2	6.19E-11	1668	14918
nervous system development	2	6	4.51E-10	296	1939
single-organism developmental process	2	3	6.29E-10	541	4033
cell-cell adhesion	2	4	8.99E-10	99	459
calcium ion binding	7	5	1.10E-09	139	737
organ development	2	6	7.55E-09	396	2824
anatomical structure morphogenesis	2	3	5.76E-08	337	2362
cellular component movement	2	4	1.07E-07	240	1570
neurogenesis	2	7	1.41E-07	204	1286
cell differentiation	2	4	2.72E-07	428	3176
cellular developmental process	2	3	3.58E-07	450	3375
cell development	2	4	3.66E-07	254	1704
generation of neurons	2	8	7.55E-07	192	1215
circulatory system development	2	6	2.25E-06	144	856
cardiovascular system development	2	6	2.25E-06	144	856
ion binding	7	2	5.39E-06	813	6765
cation binding	7	3	9.74E-06	565	4489
BioGRID interaction data	6	1	1.07E-05	933	7657
neuron differentiation	2	9	1.15E-05	176	1128
metal ion binding	7	4	1.74E-05	556	4424
cell communication	2	4	2.62E-05	693	5695
cell projection organization	2	2	2.68E-05	167	1069
organ morphogenesis	2	7	2.72E-05	145	896
cellular component morphogenesis	2	2	4.53E-05	175	1141

heart development	2	7	4.84E-05	83	438
signaling	2	2	5.57E-05	677	5571
single organism signaling	2	3	5.57E-05	677	5571
neuron development	2	10	6.08E-05	146	915
Factor: LRF; motif: VNNRMCCCC; match class: 3	3	2	6.81E-05	1542	12643
regulation of cellular process	2	3	1.02E-04	1040	9037
cell morphogenesis	2	3	1.12E-04	165	1075
biological regulation	2	2	1.15E-04	1157	10182
locomotion	2	2	1.87E-04	205	1410
calcium-dependent cell-cell adhesion	2	5	1.92E-04	15	31
biological_process	2	1	2.11E-04	1812	16892
binding	7	1	2.97E-04	1488	13523
cell-cell junction	25	1	5.93E-04	62	313
regulation of biological process	2	2	6.62E-04	1088	9577
cell morphogenesis involved in differentiation	2	4	7.48E-04	124	778
cytoskeleton	15	1	7.98E-04	273	2015
basement membrane	11	2	8.52E-04	27	93
Small cell lung cancer	14	1	9.47E-04	25	98
neuron projection development	2	3	9.98E-04	125	790
localization	2	2	1.24E-03	596	4926
tissue development	2	4	1.62E-03	226	1629
chordate embryonic development	4	2	1.62E-03	101	610
Factor: LRF; motif: VNNRMCCCC; match class: 2	3	3	1.63E-03	1374	11197
cytoskeletal part	15	1	2.10E-03	205	1457
system process	2	4	2.34E-03	258	1911
cytoskeleton organization	16	1	2.38E-03	141	931
embryo development ending in birth or egg hatching	4	1	2.77E-03	101	617
anatomical structure formation involved in morphogenesis	2	3	2.98E-03	144	959
MI:hsa-miR-940	22	1	3.63E-03	105	625
Factor: Sp1; motif: CCCCGCCCCN; match class: 3	5	2	4.64E-03	779	6025
cell morphogenesis involved in neuron differentiation	2	5	6.22E-03	101	628
cell projection morphogenesis	2	3	6.26E-03	119	770
cell projection	20	1	6.88E-03	194	1389
muscle structure development	2	4	7.08E-03	89	537
Factor: LRF; motif: VNNRMCCCC; match class: 4	3	1	7.75E-03	1600	13333
cellular response to growth factor stimulus	8	1	7.80E-03	102	639
plasma membrane part	9	1	8.10E-03	297	2282
negative regulation of biological process	2	3	9.91E-03	453	3686
axonogenesis	2	5	1.11E-02	92	566
Muscle contraction	13	1	1.11E-02	12	33
Striated Muscle Contraction	13	2	1.11E-02	12	33
Calcium Binds Troponin-C	13	3	1.11E-02	12	33
Myosin Binds ATP	13	3	1.11E-02	12	33

ATP Hydrolysis By Myosin	13	3	1.11E-02	12	33
Release Of ADP From Myosin	13	3	1.11E-02	12	33
response to growth factor stimulus	8	1	1.20E-02	103	653
neuron projection guidance	2	1	1.30E-02	68	386
axon guidance	2	2	1.30E-02	68	386
MI:hsa-miR-939	18	1	1.36E-02	104	637
Factor: Sp1; motif: CCCCGCCCCN; match class: 4	5	1	1.41E-02	881	6940
regulation of metabolic process	2	3	1.51E-02	660	5616
proteinaceous extracellular matrix	11	1	1.51E-02	71	410
Factor: LRF; motif: VNNRMCCCC; match class: 1	3	4	1.56E-02	925	7323
localization of cell	2	3	1.68E-02	150	1039
cell motility	2	4	1.68E-02	150	1039
cell part morphogenesis	2	3	1.70E-02	119	786
MI:hsa-miR-615-5p	10	1	1.72E-02	125	799
Pathways in cancer	21	1	1.80E-02	56	343
Factor: VDR; motif: GGGKNARNRRGGWSA; match class: 3	12	2	1.86E-02	1140	9204
neuron projection morphogenesis	2	4	2.15E-02	100	638
signal transduction	2	2	2.36E-02	596	5036
regulation of nucleobase-containing compound metabolic process	2	1	2.42E-02	473	3900
Factor: AP-2; motif: SNNNCCNCAGGCN; match class: 3	26	1	2.44E-02	753	5869
cell surface receptor signaling pathway	2	3	2.64E-02	361	2886
MI:hsa-miR-423-5p	23	1	2.68E-02	114	723
muscle organ development	2	7	2.76E-02	69	402
negative regulation of cytoskeleton organization	16	2	2.80E-02	23	86
in utero embryonic development	4	3	2.87E-02	66	380
positive regulation of cellular process	2	1	3.32E-02	451	3710
regulation of nitrogen compound metabolic process	2	4	3.36E-02	484	4013
cytoskeletal protein binding	24	1	3.41E-02	111	733
Factor: VDR; motif: GGGKNARNRRGGWSA; match class: 4	12	1	3.45E-02	1415	11684
Chronic myeloid leukemia	17	1	3.78E-02	18	75
transmission of nerve impulse	2	1	3.87E-02	120	808
blood vessel morphogenesis	2	4	3.89E-02	77	467
cellular component organization	2	1	4.33E-02	585	4958
cellular response to stimulus	2	1	4.44E-02	705	6085
negative regulation of cellular process	2	4	4.76E-02	411	3358
cellular response to epidermal growth factor stimulus	8	1	4.77E-02	8	14
MI:hsa-miR-675	1	1	4.95E-02	101	635
Phosphatidylinositol signaling system	19	1	4.99E-02	19	83

Appendix S4.6 MuSiC Results Compared to Significantly Mutated Genes

Gene Name	# Studies	Total Mutations	#Stop Gain/ Loss	# Missense	# Silent	# Splicing	# Validated Splicing	% Splicing Mutations Validated	MuSiC P-Value LRT	MuSiC P-Value CT	MuSiC FDR - LRT	MuSiC FDR - CT
PIK3CA*	5	181	0	3	177	1	0	0%	0.0226	0.0493	0.4088	1
TP53*	5	153	19	2	107	25	18	56%	0	0	0	0
GATA3*	4	10	0	2	7	1	1	100%	0.0075	0.0232	0.1926	0.7132
RB1*	4	19	5	1	12	1	0	0%	0.0610	0.0471	0.7410	1
AKT1*	3	6	0	1	5	0	0	NA	1	1	1	1
CBFB*	3	12	2	1	7	2	2	100%	0.0011	0.0001	0.0414	0.0122
CDH1*	3	20	5	1	5	9	4	22%	0	0	0	0
MAP3K1*	3	40	13	5	17	5	4	80%	0	0	0	0
KMT2C (MLL3)*	3	72	7	16	30	19	7	37%	0	0	0	0
PTEN*	3	11	4	0	5	2	2	100%	0.0116	0.0023	0.2677	0.1240
RUNX1*	3	8	1	1	6	0	0	NA	1	1	1	1
SF3B1*	3	19	1	6	11	1	0	0%	0	0.0006	0.0009	0.0418
CDKN1B*	2	1	0	0	1	0	0	NA	1	1	1	1
NF1*	2	27	3	3	17	4	2	50%	1	1	1	1
STMN2	2	1	0	0	1	0	0	NA	1	1	1	1
TBX3*	2	5	0	0	5	0	0	NA	1	1	1	1
AFF2*	1	20	3	5	11	1	0	0%	0.0006	0.0003	0.0257	0.0263
AGTR2	1	0	0	0	0	0	0	NA	1	1	1	1
APC	1	7	0	2	5	0	0	NA	1	1	1	1
ARID1A	1	32	6	5	19	2	1	50%	0	0	0	0.0005
ARID2	1	11	2	3	5	1	0	0%	0.4970	0.2598	1	1
ASXL1	1	8	1	3	4	0	0	NA	0.0025	0.0089	0.0806	0.3561
ATR	1	11	0	3	6	2	0	0%	1	1	1	1
BAP1	1	6	1	3	2	0	0	NA	0.1206	0.1182	1	1

BRAF	1	7	1	0	6	0	0	NA	1	1	1	1
BRCA1	1	15	2	2	7	4	2	50%	0.1109	0.0531	0.9945	1
BRCA2	1	15	2	3	9	1	1	100%	0.2360	0.2227	1	1
CCND3*	1	2	0	0	2	0	0	NA	1	1	1	1
COL6A3	1	15	1	3	11	0	0	NA	1	1	1	1
ERBB2	1	16	0	2	13	1	0	0%	0.1453	0.3469	1	1
ERBB3	1	16	1	2	11	2	0	0%	0.0355	0.3206	0.5458	1
GH1	1	1	0	1	0	0	0	NA	0.0893	0.4113	0.8910	1
KRAS	1	3	0	0	3	0	0	NA	1	1	1	1
LDLRAP1	1	0	0	0	0	0	0	NA	1	1	1	1
MAP2K4	1	16	2	0	10	4	4	100%	0.0536	0.0120	0.6904	0.4405
MLL2	1	24	1	10	13	0	0	NA	0.01427	0.02027	0.3081	0.6498
MYH9	1	15	0	2	13	0	0	NA	1	1	1	1
MYO3A	1	11	0	6	3	2	0	0%	0.0050	0.0009	0.1390	0.0571
NRAS	1	0	0	0	0	0	0	NA	1	1	1	1
PIK3R1*	1	7	0	1	5	1	0	0%	1	1	1	1
PTPN22*	1	5	0	0	5	0	0	NA	1	1	1	1
PTPRD*	1	25	3	4	18	0	0	NA	0.0335	0.0410	0.5258	1
SETD2	1	16	1	3	9	3	1	0%	0.0408	0.1420	0.5975	1
SMAD4	1	1	0	0	1	0	0	NA	1	1	1	1
STK11	1	1	0	0	1	0	0	NA	1	1	1	1
SYNE1	1	65	4	14	41	6	0	0%	1	1	1	1
SYNE2	1	57	0	10	44	3	0	0%	1	1	1	1
UBR5	1	29	0	10	18	1	1	100%	0.1179	0.0142	1	0.4993
USH2A	1	65	0	12	52	1	0	0%	1	1	1	1

* Identified by TCGA to be significantly mutated

Appendix S4.7 Validated Cryptic Splicing Mutations

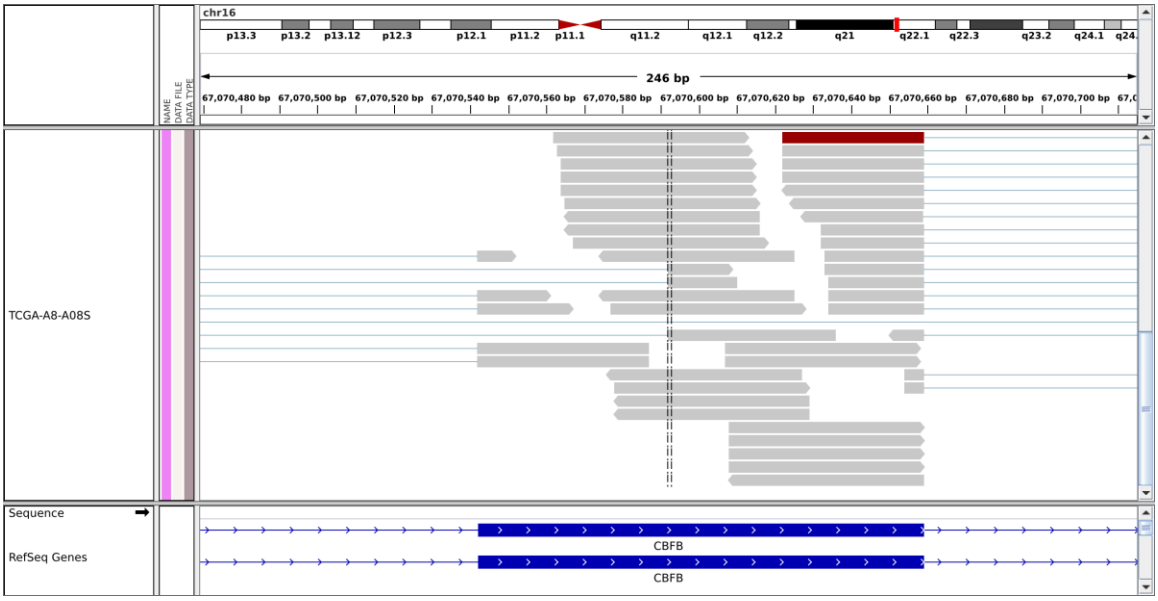
Appendix S4.7.1 Cryptic Splicing Mutation Details

Information theory based analysis and corresponding evidence demonstrating abnormal mRNA splicing in predicted mRNA splicing mutations. (A) Table indicates the TCGA sample identifier, variant, information analysis and statistical support for the mutation. (B) Screenshots from the Integrative Genomics Viewer (IGV) displaying junction-spanning reads that demonstrate cryptic splicing for mutations predicted by the Shannon Pipeline in the genes *CBFB, GATA3, PALB2,* and *ABL1*. The normal exonic structure is indicated by blue, with the thick bars representing exons, and the thin lines introns. RNA-Seq reads are shown in grey with the vertical dotted black lines demarcate the location of the cryptic splice site.

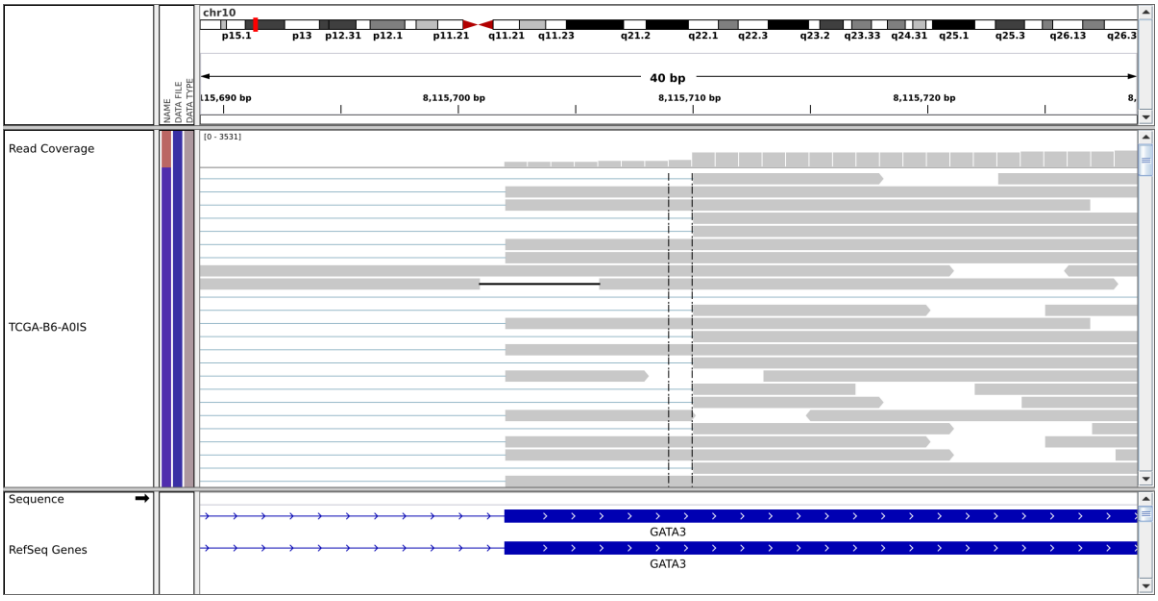
(A)

Patient	Gene	Splice Site Coordinate	Variant Coordinate	Ref/ Var	Ri-initial	Ri-final	Δ Ri	Cryptic Site Use P-Value	Exon Skipping P-Value
TCGA-A8-A08S	CBFB	chr16: 67070591	chr16: 67070577	G/T	5.6	7.5	1.9	< 0.005	0.12
TCGA-B6-A0I5	GATA3	chr10: 8115709	chr10: 8115702	A/C	4.2	5.9	1.7	< 0.005	NA
TCGA-B6-A0RT	PALB2	chr16: 23637694	chr16: 23637710	T/A	5.3	7.0	1.7	< 0.005	0.05
TCGA-B6-A0RV	ABL1	chr9: 133750256	chr9: 133750254	G/C	0.8	9.6	8.8	< 0.005	NA

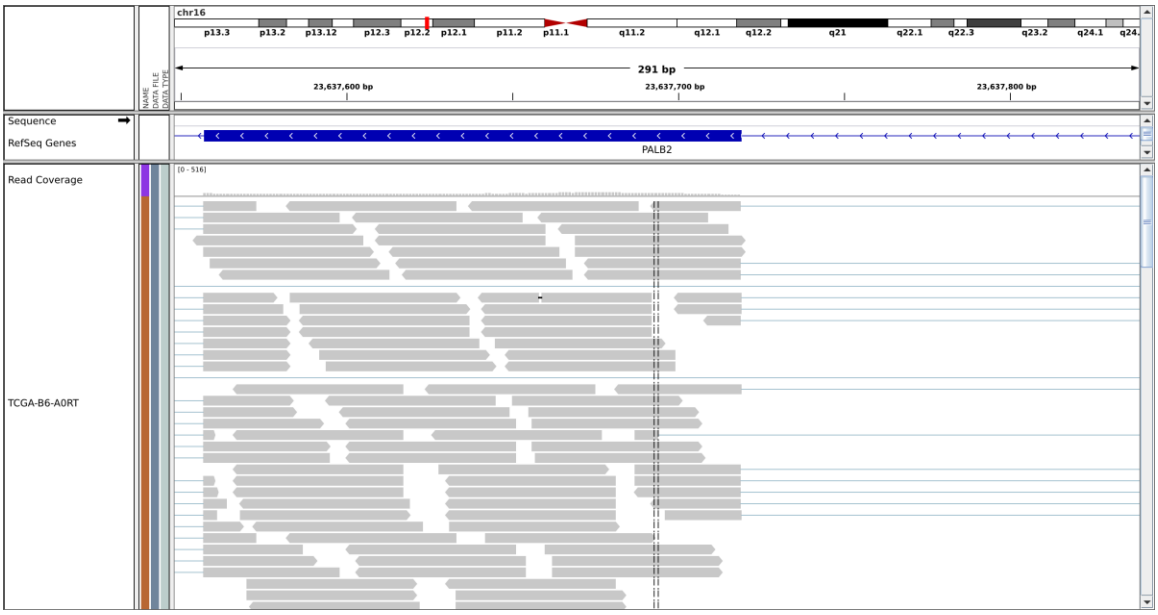
(B) *CBFB*



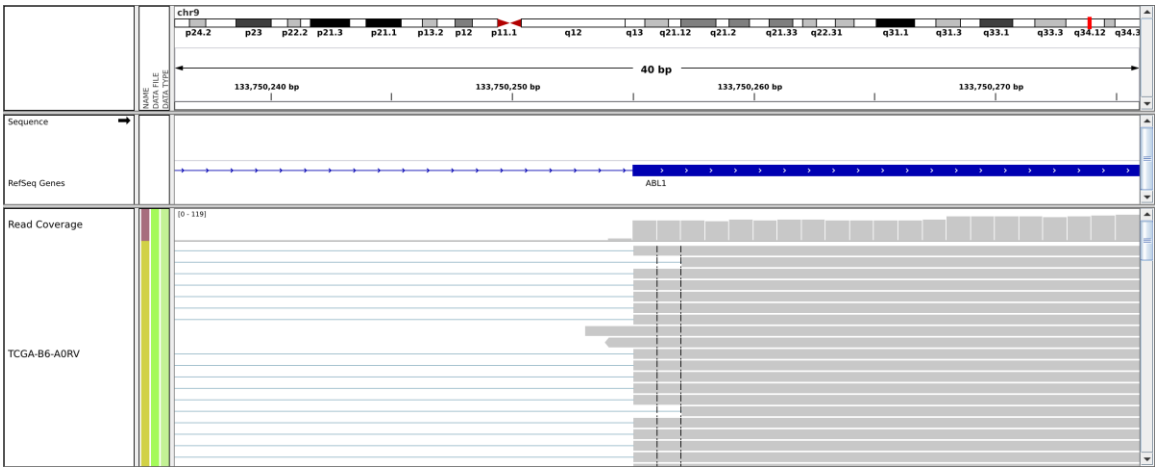
(C) *GATA3*



(D) *PALB2*

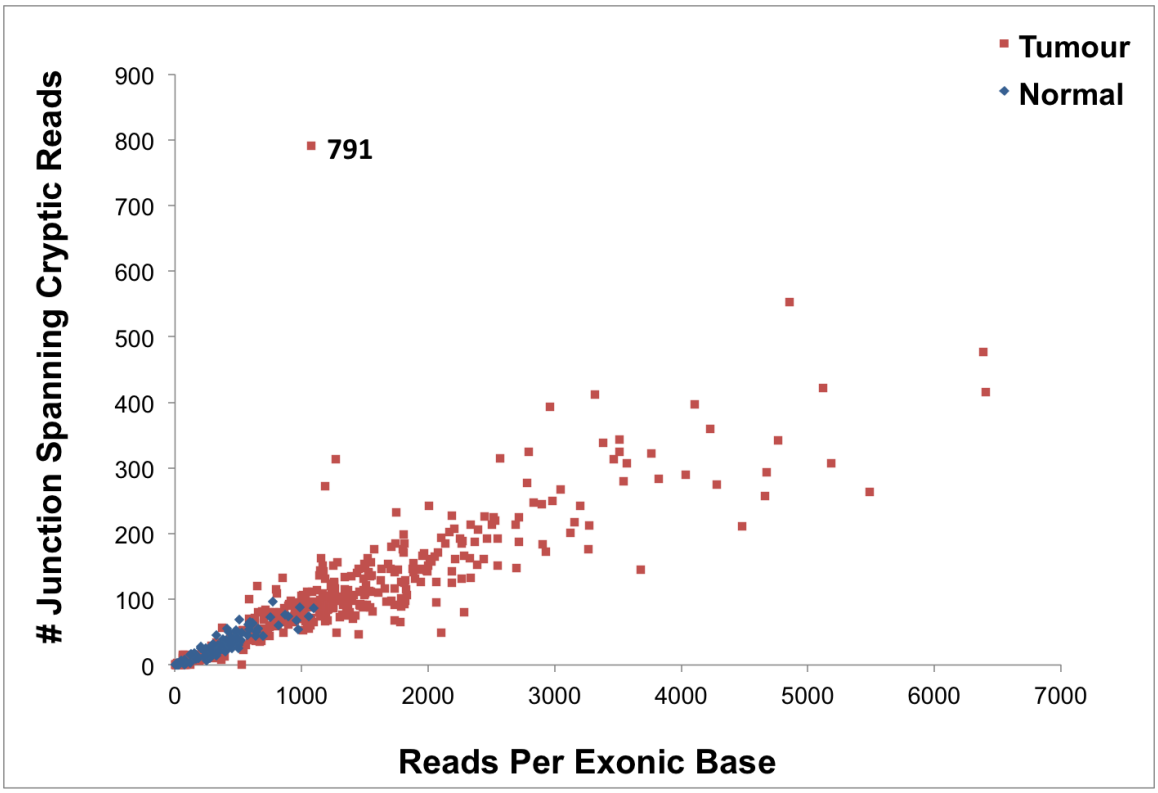


(E) *ABL1*



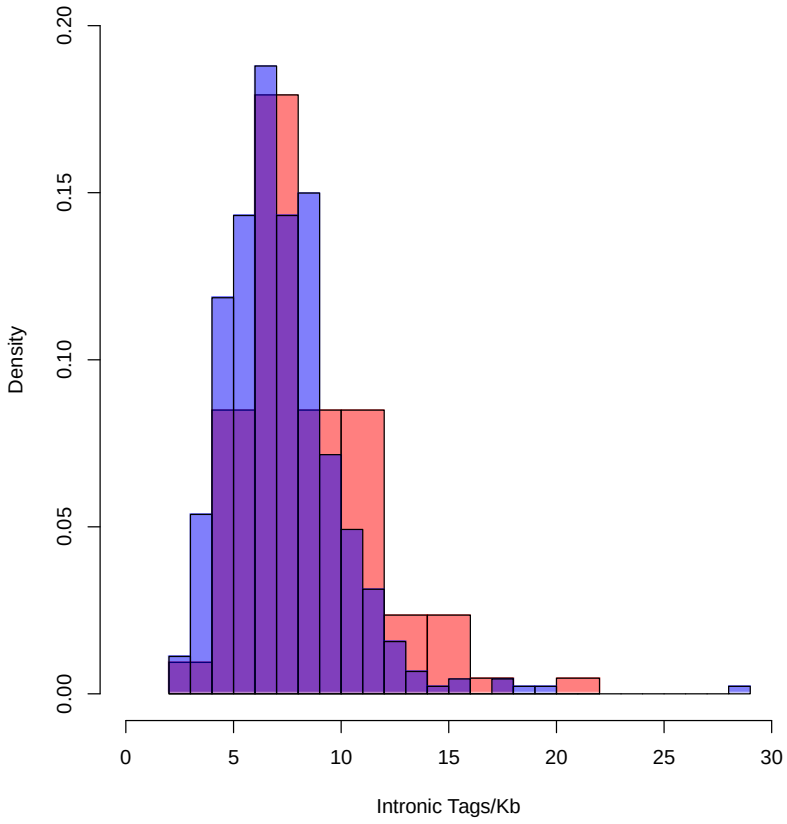
Appendix S4.7.2 The rate of GATA3 abnormal splicing in variant containing tumour and tumour/normal controls

Junction-spanning, cryptic splicing read counts for GATA3 mutation (chr10: g.8115702A>C). The number of RNA-Seq reads per exonic base were plotted against the number of reads demonstrating GATA3 cryptic splicing in the variant-containing tumours and controls. The variant containing tumour is indicated by the number of cryptic splicing reads (n = 791), tumours that do not contain this variant are in red, and normal controls are in blue. Cryptic splicing in the control samples likely occurs because the cryptic splice site (R_i = 4.2 bits) exceeds the strength of the natural splice site (R_i = 0.9 bits). However, the mutation further weakens the natural splice site (final R_i = 0.0 bits), while simultaneously strengthening the cryptic splice site (final R_i = 5.8 bits), which are consistent with the RNA-Seq analysis.



Appendix S4.8 Supplementary Figure 4

Intron inclusion in tumour and normal breast genomes, based on RNA-Seq evidence. Histogram of the density of intronic sequence reads for normal (blue) and tumour (red) RNA-Seq samples. Purple shading represents overlapping components of the two density distributions. Intron inclusion was calculated with RSeQC's ReadDist script and RefSeq's gene annotation. High levels of unspliced isoforms with intron inclusion were the most frequent outcome of mutations with significant effects on mRNA splicing. Nevertheless, when considering non-specific aberrant splicing across the transcriptome, the numbers of junction-spanning, intron inclusion reads present in normal and tumour samples did not significantly differ ($p > 0.1$). In fact, non-junction-spanning, intronic read-abundance reads of normal controls exceeded those of the tumour samples ($p < 0.01$). This suggests that validation events in these tumour samples are not due solely to intron inclusion and aberrant mRNA splicing known to be present in breast tumours (9). It is notable, however, that the levels of intronic inclusion for validated mutations significantly exceeded the read counts for all controls that did not contain these variants.



Appendix S4.9 Pathway Analyses

Appendix S4.9.1 Pathways Overrepresented by Protein Coding and Splicing Mutations

Pathway #	Pathways common to both Strelka splicing and protein coding mutations	Present in Pathway Analysis with both Protein Coding and Splicing Mutations?	
1	Anchoring fibril formation	YES	Col/ECM
2	Assembly of collagen fibrils and other multimeric structures	YES	Col/ECM
3	Association of procollagen chains	YES	Col/ECM
4	Collagen biosynthesis and modifying enzymes	YES	Col/ECM
5	Collagen formation	YES	Col/ECM
6	Collagen prolyl 3-hydroxylase converts proline to 3-hydroxyproline	YES	Col/ECM
7	Collagen prolyl 4-hydroxylase converts proline to 4-hydroxyproline	YES	Col/ECM
8	Collagen type VII binds laminin-322 and collagen IV	YES	Col/ECM
9	DDR1 binds collagens	YES	Col/ECM
10	ECM proteoglycans	YES	Col/ECM
11	Extracellular matrix organization	YES	Col/ECM
12	Galactosylation of collagen propeptide hydroxylysines by PLOD3	YES	Col/ECM
13	Glucosylation of collagen propeptide hydroxylysines	YES	Col/ECM
14	Interaction of NCAM1 with collagens	YES	Col/ECM
15	Non-integrin membrane-ECM interactions	YES	Col/ECM
16	PDI is a chaperone for collagen peptides	YES	Col/ECM
17	Procollagen lysyl hydrolases convert lysine to 5-hydroxylysine	YES	Col/ECM
18	Procollagen triple helix formation	YES	Col/ECM
19	Removal of fibrillar collagen C-propeptides	YES	Col/ECM
20	Removal of fibrillar collagen N-propeptides	YES	Col/ECM
21	Secretion of collagens	YES	Col/ECM
22	Formation of collagen fibres	NO	Col/ECM
23	Formation of collagen fibrils	NO	Col/ECM
24	Galactosylation of collagen propeptide hydroxylysines by procollagen galactosyltransferases 1, 2.	NO	Col/ECM
25	PDGF binds to extracellular matrix proteins	NO	Col/ECM
26	Cell Cycle, Mitotic	NO	Cancer
27	Cell-Cell communication	YES	Cancer
28	CBL, GRB2, FYN and PI3K p85 subunit are constitutively associated	NO	?
29	Signaling by FGFR1 fusion mutants	YES	Cancer
30	Indirect recruitment of PI3K to KIT via p(Y)-GAB2	YES	Cancer
31	Loss of Nlp from mitotic centrosomes	YES	?
32	Base Excision Repair	YES	Cancer

33	Mitotic Prometaphase	YES	Cancer
34	Signaling by FGFR in disease	YES	Cancer
35	PLCG1 events in ERBB2 signaling	YES	Cancer
36	Downstream signaling of activated FGFR	NO	Cancer
37	Signaling by FGFR1 mutants	NO	Cancer
38	Separation of sister chromatids	YES	?
39	Kinetochore assembly	YES	?
40	Recruitment of mitotic centrosome proteins and complexes	YES	?
41	Centrosome maturation	YES	?
42	Mitotic G2-G2/M phases	YES	Cancer
43	Signaling by ERBB2	NO	Cancer
44	Resolution of AP sites via the single-nucleotide replacement pathway	YES	Cancer
45	G2/M Transition	YES	Cancer
46	Transmembrane transport of small molecules	YES	Other
47	Ion channel transport	YES	Other
48	Axon guidance	YES	Other
49	Integrin cell surface interactions	YES	Other
50	Ion transport by P-type ATPases	YES	Other
51	Developmental Biology	YES	Other
52	NICD1 displaces co-repressor complex from RBPJ (CSL)	NO	Other
53	NICD1 PEST domain mutants displace co-repressor complex from RBPJ (CSL)	NO	Other
54	L1CAM interactions	YES	?
55	Transport of inorganic cations/anions and amino acids/oligopeptides	YES	Other
56	SLC-mediated transmembrane transport	YES	Other
57	cAMP degradation by Phosphodiesterases	YES	Other
58	CBL is tyrosine phosphorylated	YES	Other
59	Dystroglycan binds Laminins and Dystrophin	YES	Other
60	Signalling by NGF	YES	Other
61	P-type ATPases type IV transport external-facing APLs to internal side of the plasma membrane	YES	Other
62	P-type ATPases type IV transport internal-facing APLs to external side of the plasma membrane	YES	Other
63	NCAM signaling for neurite out-growth	YES	Other
64	NRAGE signals death through JNK	YES	?
65	p75NTR indirectly activates RAC and Cdc42 via a guanyl-nucleotide exchange factor	YES	Other
66	Signaling by Rho GTPases	YES	Other
67	Rho GTPase cycle	YES	Other
68	Stimuli-sensing channels	YES	Other
69	ABC-family proteins mediated transport	YES	Other
70	GEFs activate RhoA,B,C	NO	Other
71	Other semaphorin interactions	YES	Other
72	Signaling by PDGF	YES	Other
73	Semaphorin interactions	YES	Other
74	Signaling by Interleukins	YES	Other
75	Interaction between L1 and Ankyrins	YES	Other
76	Transmission across Chemical Synapses	YES	Other

77	Plk1-mediated phosphorylation of Nlp	YES	Other
78	Loss of proteins required for interphase microtubule organizationÃƒâ€šÃ‚ from the centrosome	NO	?
79	Loss of C-Nap-1 from centrosomes	YES	?
80	Dissociation of Phospho-Nlp from the centrosome	YES	?
81	Recruitment of Plk1 to centrosomes	YES	?
82	Resolution of Abasic Sites (AP sites)	YES	?
83	Platelet calcium homeostasis	YES	Other
84	Recruitment of additional gamma tubulin/ gamma TuRC to the centrosome	YES	?
85	Recruitment of CDK11p58 to the centrosomes	YES	?
86	Ankyrins link voltage-gated sodium and potassium channels to spectrin and L1	YES	Other
87	Translocation of Influenza A virus nonstructural protein 1 (NS1A) into the nucleus	YES	Other
88	Synthesis of PIPs at the early endosome membrane	YES	Other
89	The ABCC family mediates organic anion transport	NO	Other
90	PLC beta mediated events	YES	Other
91	Downstream signal transduction	YES	?
92	Phosphorylation of cohesin by PLK1 at centromeres	YES	?
93	PP2A-B56 dephosphorylates centromeric cohesin	YES	?
94	DAG and IP3 signaling	YES	Other
95	G-protein mediated events	YES	Other
96	Kinetochore capture of astral microtubules	YES	?
97	ESPL1 (Separase) cleaves centromeric cohesin	YES	Other
98	Recruitment of Grb2 to pFAK:NCAM1	YES	Other
99	2GABRA:2GABRB:GABRG:GABA transports extracellular Cl- to cytosol	YES	Other
100	GABA A receptor activation	YES	Other

Appendix S4.9.2 Pathways Overrepresented by Every Splicing Mutation Type (inactivating, leaky, cryptic)

			Overrepresented by protein coding mutation set?
1	Association of procollagen chains	Collagen	YES
2	Collagen biosynthesis and modifying enzymes		YES
3	Collagen formation		YES
4	Collagen prolyl 3-hydroxylase converts proline to 3-hydroxyproline		YES
5	Collagen prolyl 4-hydroxylase converts proline to 4-hydroxyproline		YES
6	DDR1 binds collagens		YES
7	Galactosylation of collagen propeptide hydroxylysines by PLOD3		YES
8	Galactosylation of collagen propeptide hydroxylysines by procollagen galactosyltransferases 1, 2.		YES
9	Glucosylation of collagen propeptide hydroxylysines		YES
10	PDI is a chaperone for collagen peptides		YES
11	Procollagen lysyl hydrolases convert lysine to 5-hydroxylysine		YES
12	Procollagen triple helix formation		YES
13	Secretion of collagens		YES
14	Degradation of collagen		NO
15	ECM proteoglycans	ECM	YES
16	Extracellular matrix organization		YES
17	Non-integrin membrane-ECM interactions		YES
18	Anchoring fibril formation		YES
19	Axon guidance	Other	YES
20	Cell Cycle, Mitotic		YES
21	Developmental Biology		YES
22	Integrin cell surface interactions		YES
23	L1CAM interactions		YES
24	Transmembrane transport of small molecules		YES
25	Activation of Chaperone Genes by XBP1(S)		NO
26	Activation of Chaperones by IRE1alpha		NO
27	Cell Cycle		NO
28	Hemostasis		NO

Appendix S4.9.3 Comparing Grouped Pathways Overrepresented between LN- and LN+ Tumour Mutations

RED = collagen, BLUE = extracellular matrix, GREEN = NCAM1 pathways, No. = number of pathways in group

Lymph Node Negative Tumours	No.	Lymph Node Positive Tumours	No.	Pathways in Both Lymph Node Positive and Negative Tumours	No. LN-	No. LN+
Signaling by FGFR1 fusion mutants	10	Neurotransmitter Release Cycle	10	Collagen biosynthesis and modifying enzymes	12	12
Cytosolic tRNA aminoacylation	9	Complement cascade	8	Semaphorin interactions	9	10
Striated Muscle Contraction	7	NCAM signaling for neurite out-growth	8	L1CAM interactions	8	8
PI3K events in ERBB2 signaling	6	SHC1 events in ERBB2 signaling	4	Signaling by Interleukins	12	1
COPI Mediated Transport	6	SHC1 events in ERBB4 signaling	4	Extracellular matrix organization	4	6
Glucose metabolism	4	Downregulation of ERBB4 signaling	3	GPCR downstream signaling	4	4
Synthesis of PIPs at the late endosome membrane	3	Generic Transcription Pathway	3	Regulation of Cholesterol Biosynthesis by SREBP (SREBF)	2	6
STAT6-mediated induction of chemokines	3	Nuclear signaling by ERBB4	3	Hemostasis	4	3
Signaling by SCF-KIT	3	Regulation of Hypoxia-inducible Factor (HIF) by Oxygen	3	Transport of inorganic cations/anions and amino acids/oligopeptides	5	1
Regulation of signaling by CBL	3	Signaling by ERBB4	3	Transmembrane transport of small molecules	3	3
Regulation of AMPK activity via LKB1	3	Xenobiotics	3	Condensation of Prometaphase Chromosomes	3	3
Transport of vitamins, nucleosides, and related molecules	2	Apoptosis induced DNA fragmentation	2	Signal Transduction	2	4
Synthesis of PIPs at the Golgi membrane	2	Assembly of the pre-replicative complex	2	Fc epsilon receptor (FCERI) signaling	1	5
Signaling by constitutively active EGFR	2	Binding and Uptake of Ligands by Scavenger Receptors	2	Recruitment of mitotic centrosome proteins and complexes	3	2
PIP3 activates AKT signaling	2	Conjugation of carboxylic acids	2	Factors involved in megakaryocyte development and platelet production	3	2
PI3K events in ERBB4 signaling	2	DAG and IP3 signaling	2	Synthesis of PIPs at the plasma membrane	2	3
PI3K Cascade	2	Metabolism of amino acids and derivatives	2	Non-integrin membrane-ECM interactions	2	3
Nuclear import of Rev protein	2	Mitotic G1-G1/S phases	2	Ion channel transport	2	3

Metabolism of steroid hormones and vitamin D	2	Nitric oxide stimulates guanylate cyclase	2	COPII (Coat Protein 2) Mediated Vesicle Transport	2	3
Intrinsic Pathway	2	Opioid Signalling	2	Cell Cycle	2	3
GPVI-mediated activation cascade	2	PI Metabolism	2	Assembly of collagen fibrils and other multimeric structures	1	4
Downstream signal transduction	2	Phospholipid metabolism	2	Muscle contraction	3	1
Cross-presentation of particulate exogenous antigens (phagosomes)	2	Scavenging of Heme from Plasma	2	Membrane Trafficking	3	1
Costimulation by the CD28 family	2	Signaling by ERBB2	2	Integrin cell surface interactions	3	1
Cell Cycle, Mitotic	2	Synthesis of IP2, IP, and Ins in the cytosol	2	Inactivation, recovery and regulation of the phototransduction cascade	3	1
CD28 dependent PI3K/Akt signaling	2	Synthesis of very long-chain fatty acyl-CoAs	2	Regulation of the Fanconi anemia pathway	2	2
Vitamin C (ascorbate) metabolism	1	Activated NOTCH1 Transmits Signal to the Nucleus	1	Nephrin interactions	2	2
VEGF ligand-receptor interactions	1	Amino Acid conjugation	1	Loss of proteins required for interphase microtubule organization√Ç¬†from the centrosome	2	2
tRNA Aminoacylation	1	Antiviral mechanism by IFN-stimulated genes	1	Loss of Nlp from mitotic centrosomes	2	2
Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds	1	Asparagine N-linked glycosylation	1	Fatty acid, triacylglycerol, and ketone body metabolism	2	2
Translocation of GLUT4 to the Plasma Membrane	1	CDC6 association with the ORC:origin complex	1	Developmental Biology	2	2
Tie2 Signaling	1	CDO in myogenesis	1	Collagen formation	2	2
TCR signaling	1	CDT1 association with the CDC6:ORC:origin complex	1	Stimuli-sensing channels	1	3
STING mediated induction of host immune responses	1	CREB phosphorylation through the activation of Adenylate Cyclase	1	Ion transport by P-type ATPases	1	3
Smooth Muscle Contraction	1	Calnexin/calreticulin cycle	1	ECM proteoglycans	1	3
Signaling by Rho GTPases	1	Chromosome Maintenance	1	DAP12 interactions	1	3
Signaling by FGFR mutants	1	Circadian Clock	1	Unfolded Protein Response	2	1
S6K1-mediated signalling	1	Conjugation of benzoate with glycine	1	Synthesis of PIPs at the early endosome membrane	2	1
S6K1 signalling	1	Conjugation of phenylacetate with glutamine	1	Platelet activation, signaling and aggregation	2	1

RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways	1	DNA Repair	1	NGF signalling via TRKA from the plasma membrane	2	1
Rho GTPase cycle	1	Disease	1	Cell-Cell communication	2	1
Response to elevated platelet cytosolic Ca2+	1	E2F-enabled inhibition of pre-replication complex formation	1	Signalling by NGF	1	2
Regulation of mRNA Stability by Proteins that Bind AU-rich Elements	1	EGFR interacts with phospholipase C-gamma	1	Regulation of Insulin Secretion	1	2
Rap1 signalling	1	ER Quality Control Compartment (ERQC)	1	p75 NTR receptor-mediated signalling	1	2
Polo-like kinase mediated events	1	Fanconi Anemia pathway	1	G2/M Checkpoints	1	2
Platelet Adhesion to exposed collagen	1	G-protein mediated events	1	Effects of PIP2 hydrolysis	1	2
PKB-mediated events	1	Interferon Signaling	1	Axon guidance	1	2
Phase 1 - Functionalization of compounds	1	Interleukin-2 signaling	1	Transmission across Chemical Synapses	1	1
Organic cation/anion/zwitterion transport	1	Lipoprotein metabolism	1	Stabilization of p53	1	1
NRAGE signals death through JNK	1	Lysine catabolism	1	SLC-mediated transmembrane transport	1	1
mTORC1-mediated signalling	1	MyD88 cascade initiated on plasma membrane	1	Signaling by NOTCH1 PEST Domain Mutants in Cancer	1	1
mTOR signalling	1	N-glycan trimming in the ER and Calnexin/Calreticulin cycle	1	Signaling by FGFR1 mutants	1	1
Metabolism of water-soluble vitamins and cofactors	1	NGF processing	1	Peroxisomal lipid metabolism	1	1
Metabolism of nucleotides	1	Neuronal System	1	NOTCH1 Intracellular Domain Regulates Transcription	1	1
Metabolism	1	PKA activation	1	Metabolism of proteins	1	1
Ligand-gated ion channel transport	1	Phase II conjugation	1	Meiosis	1	1
ISG15 antiviral mechanism	1	Platelet calcium homeostasis	1	M Phase	1	1
Integration of energy metabolism	1	Platelet homeostasis	1	Interleukin receptor SHC signaling	1	1
Inhibition of replication initiation of damaged DNA by RB1/E2F1	1	Platelet sensitization by LDL	1	Inositol phosphate metabolism	1	1
Homologous recombination repair of replication-independent double-strand breaks	1	Pyruvate metabolism and Citric Acid (TCA) cycle	1	ER to Golgi Transport	1	1
Golgi to ER Retrograde Transport	1	Signaling by EGFR	1	Depolarization of the Presynaptic Terminal Triggers the Opening of Calcium Channels	1	1
Gene Expression	1	Signaling by FGFR	1	Conversion from APC/C:Cdc20 to APC/C:Cdh1 in late anaphase	1	1
Gamma-carboxylation of protein precursors	1	Signaling by NOTCH2	1	Collagen degradation	1	1

E2F mediated regulation of DNA replication	1	Signaling by PDGF	1	ATM mediated phosphorylation of repair proteins	1	1
Cytokine Signaling in Immune system	1	Signaling by Robo receptor	1	Activation of Chaperones by IRE1alpha	1	1
Cyclin E associated events during G1/S transition	1	Signaling by the B Cell Receptor (BCR)	1			
Constitutive PI3K/AKT Signaling in Cancer	1	Syndecan interactions	1			
ChREBP activates metabolic gene expression	1	Synthesis and interconversion of nucleotide di- and triphosphates	1			
Cell surface interactions at the vascular wall	1	Synthesis of IP3 and IP4 in the cytosol	1			
Cell death signalling via NRAGE, NRIF and NADE	1	Transcriptional Regulation of White Adipocyte Differentiation	1			
CD28 co-stimulation	1					
Biological oxidations	1					
Antigen processing-Cross presentation	1					
AMPK inhibits chREBP transcriptional activation activity	1					
Activation of the AP-1 family of transcription factors	1					
Activation of Chaperones by ATF6-alpha	1					
Abacavir transmembrane transport	1					

Appendix S4.9.4 Pathway Analysis of Deleterious Mutations in LN- and LN+ Tumours

RED = collagen, BLUE = extracellular matrix, GREEN = NCAM1 pathways

Lymph Node Positive Tumour Mutations	Lymph Node Negative Tumour Mutations	Over Represented by Both Lymph Node Positive and Negative Tumour Mutations
Autophosphorylation of NCAM1 bound Fyn	Activation of Chaperones by ATF6-alpha	Assembly of collagen fibrils and other multimeric structures
Dephosphorylation of NCAM1 bound pFyn	Cargo, Sec31p:Sec13p, and v-SNARE recruitment	Association of procollagen chains
Interaction of NCAM1 with collagens	Cell-Cell communication	Collagen biosynthesis and modifying enzymes
NCAM signaling for neurite out-growth	Cleavage of ATF6-alpha by S1P	Collagen formation
NCAM1 interactions	COPII (Coat Protein 2) Mediated Vesicle Transport	Collagen prolyl 3-hydroxylase converts proline to 3-hydroxyproline
Formation of collagen fibres	ER to Golgi Transport	Collagen prolyl 4-hydroxylase converts proline to 4-hydroxyproline
Formation of collagen fibrils	factor VIII + von Willebrand factor multimer -> factor VIII:von Willibrand factor multimer	Galactosylation of collagen propeptide hydroxylysines by PLOD3
Removal of fibrillar collagen N-propeptides	factor VIII:von Willibrand factor multimer -> factor VIIIa + factor VIIIa B A3 acidic polypeptide + von Willibrand factor multimer	Galactosylation of collagen propeptide hydroxylysines by procollagen galactosyltransferases 1, 2.
Syndecan-1 binds collagen types I, III, V	FGFR1 fusions bind PLCgamma	Glucosylation of collagen propeptide hydroxylysines
Syndecan-1 binds collagen types I, III, V	Hemostasis	PDI is a chaperone for collagen peptides
Degradation of the extracellular matrix	Inhibition of integrin activation by sequestering PIP5KIgamma	Procollagen lysyl hydrolases convert lysine to 5-hydroxylysine
ECM proteoglycans	Interaction of integrin alphaEbeta7 with Cadherin-1	Procollagen triple helix formation
Non-integrin membrane-ECM interactions	Interleukin-1 receptor type 1 binds Interleukin 1	Removal of fibrillar collagen C-propeptides
PDGF binds to extracellular matrix proteins	Interleukin-2 signaling	Secretion of collagens
Activation of Adenylate Cyclase	Interleukin-7 signaling	Extracellular matrix organization
Activation of Chaperones by IRE1alpha	Na+-coupled HCO3- cotransport	Axon guidance
Activation of PPARA by Fatty Acid Ligands	NrCAM interactions	Cell Cycle
AGRN binds Laminins with gamma-1 subunit	p-PLCgamma dissociates from FGFR1 fusions	Developmental Biology
Ankyrins link voltage-gated sodium and potassium channels to spectrin and L1	Phosphorylation of STAT5 by FGFR1 fusions	Dissociation of Phospho-Nlp from the centrosome
Antiviral mechanism by IFN-stimulated genes	PI is phosphorylated to PI5P by PIKFYVE at the late endosome membrane	DOCKs bind to RhoGEFs

Association of MCM8 with ORC:origin complex

Binding of Beta-TrCP1 to phosphorylated PER proteins

Binding of IP3 to IP3 receptor

Ca2+ influx through voltage gated Ca2+ channels

Calcium Influx through Voltage-gated Calcium Channels

Calnexin/calreticulin cycle

Cell Cycle, Mitotic

Cell death signalling via NRAGE, NRIF and NADE

cGMP effects

Dephosphorylation of CK2-modified condensin I

Depolarization of the Presynaptic Terminal Triggers the Opening of Calcium Channels

Dystroglycan binds Laminins and Dystrophin

ER Quality Control Compartment (ERQC)

ERBB4:TAB2:NCOR1 complex translocates to the nucleus
ERBB4:TAB2:NCOR1 complex translocates to the nucleus
ERBB4s80 binds Tab2:Ncor1 complex
ERBB4s80 binds Tab2:Ncor1 complex
Fanconi Anemia pathway
Formation of the BRCA1-PALB2-BRCA2 complex
Interaction of L1 with Laminin-1
Interaction of nephrin with adherens junction-

PI is phosphorylated to PI5P by PIKFYVE at the late endosome membrane
PI(3,5)P2 is dephosphorylated to PI3P by FIG4 at the early endosome membrane
PI(3,5)P2 is dephosphorylated to PI3P by FIG4 at the Golgi membrane
PI(3,5)P2 is dephosphorylated to PI3P by FIG4 at the late endosome membrane
PI3P is phosphorylated to PI(3,5)P2 by Pikfyve at the early endosome membrane
PI3P is phosphorylated to PI(3,5)P2 by Pikfyve at the early endosome membrane
PI3P is phosphorylated to PI(3,5)P2 by PIKFYVE at the Golgi membrane
PI3P is phosphorylated to PI(3,5)P2 by PIKFYVE at the Golgi membrane
PI3P is phosphorylated to PI(3,5)P2 by PIKFYVE at the late endosome membrane
PI3P is phosphorylated to PI(3,5)P2 by PIKFYVE at the late endosome membrane

PLCgamma is phosphorylated by FGFR1-fusions

Plexin-A1-4 binds NRP1

Recruitment of additional gamma tubulin/ gamma TuRC to the centrosome

Release of platelet cytosolic components

Replication initiation regulation by Rb1/E2F1

Semaphorin interactions
Signaling by FGFR1 fusion mutants
SLC-mediated transmembrane transport

Stabilization of mRNA by HuR

Synthesis of IP3 and IP4 in the cytosol
Transcriptional activation of Acetyl-CoA

Integrin cell surface interactions

Interaction between L1 and Ankyrins

Loss of C-Nap-1 from centrosomes

Loss of Nlp from mitotic centrosomes

Loss of proteins required for interphase microtubule organizationÃ‚Â from the centrosome

Meiosis

Phosphorylation of MEK4 by MEKK1

Phosphorylation of p53 at ser-15 by ATM kinase

Plk1-mediated phosphorylation of Nlp

Recruitment of CDK11p58 to the centrosomes

Recruitment of Plk1 to centrosomes

Transport of inorganic cations/anions and amino acids/oligopeptides

associated proteins IP3 binds to the IP3 receptor, opening the endoplasmic reticulum Ca2+ channel ISG15 antiviral mechanism L1CAM interactions Mitotic Prometaphase N-glycan trimming in the ER and Calnexin/Calreticulin cycle NDP + reduced thioredoxin => dNDP + oxidized thioredoxin + H2O Neurofascin binds contactin-1:CASPR complex NICD1 displaces co-repressor complex from RBPJ (CSL) NICD1 displaces NCOR co-repressor complex from CSL NICD1 PEST domain mutants displace co-repressor complex from RBPJ (CSL) Nitric oxide stimulates guanylate cyclase NTN4 binds laminins with gamma-1, gamma-3 Opening of ER calcium channels by activated PKA p75 NTR receptor-mediated signalling Phosphorylation of FANCD2 by ATR/ATM Phosphorylation of FANCI by ATM/ATR Release of calcium from intracellular stores by IP3 receptor activation Signaling by ERBB4 Signaling by PDGF Signalling by NGF Syndecan interactions Transport of Ca++ from platelet dense tubular system to cytoplasm Unfolded Protein Response	carboxylase by ChREBP:MLX Transmembrane transport of small molecules Vesicle Budding Vesicle Uncoating Vesicular glutamate transport	

Appendix S4.10 Frequency of Mutations in NCAM1 Pathway Genes

Gene Name	All Mutations	Splicing Mutations	Indel Frameshift & Stop Gain/Loss Mutations
Total:	**425**	**37**	**35**
SPTA1	30	2	3
CACNA1D	22	3	7
COL6A5	19	1	0
NCAM1	17	2	1
COL6A6	16	1	1
COL6A3	15	0	3
CACNA1G	13	2	0
CACNA1I	13	1	0
COL4A1	13	4	0
CACNA1C	12	0	1
SPTBN1	12	2	1
COL3A1	11	0	2
SPTBN4	11	0	0
CACNB2	10	1	1
COL4A4	10	0	0
COL4A5	10	1	1
SPTB	10	0	2
CACNA1S	9	0	0
COL4A3	9	2	0
COL5A2	9	2	1
CNTN2	8	1	1
COL5A3	8	1	2
COL9A2	8	4	0
NCAN	8	1	1
SPTAN1	8	0	0
AGRN	6	0	0
COL5A1	6	0	2
COL6A2	6	0	0
PTPRA	6	0	0
CACNA1H	5	0	0
CACNB1	5	1	0
COL9A3	5	1	0
FGFR1	5	0	0
PRNP	5	0	0
SOS1	5	0	1
SPTBN5	5	0	0
COL9A1	4	0	0
CREB1	4	0	0

PTK2	4	0	0
CACNB3	3	0	0
COL4A2	3	0	0
KRAS	3	0	0
MAPK3	3	0	1
RAF1	3	1	0
RPS6KA5	3	0	0
CDK1	2	0	0
COL2A1	2	0	1
FYN	2	1	0
GDNF	2	0	0
GFRA1	2	0	0
GFRA2	2	0	0
GRB2	2	1	0
MAPK1	2	1	0
SPTBN2	2	0	0
SRC	2	0	0
ST8SIA2	2	0	1
HRAS	1	0	0
MAP2K2	1	0	0
YWHAB	1	0	1

Appendix S4.11 Breast Cancer Mutations by Subtype

Appendix S4.11.1 Number of Mutations by Subtype

Subtype*	No. Tumours	All Coding		Deleterious Coding^		Splicing	
		No.	Av.	No.	Av.	No.	Av.
Basal Like	81	15,383	190	1,350	16.7	1,288	15.9
HER2-enriched	51	8,633	169	708	13.9	729	14.3
Luminal A	192	22,634	118	1,889	9.8	1,786	9.3
Luminal B	104	14,501	139	1,166	11.2	1,209	11.6
"Normal Like"	6	1,105	184	105	17.5	78	13

*subtype not available for 8 tumours, ^ Frameshift indels, stop codon gain or loss, No. = total number of mutations, Av. = average number of mutations per tumour

Appendix S4.11.2 Pathway Analysis of Mutations by Subtype and Lymph Node Status

	Basal		Her2		Luminal A		Luminal B		Normal-Like	
	-	+	-	+	-	+	-	+	-	+
# Tumours	**50**	**31**	**18**	**32**	**89**	**96**	**43**	**59**	**1**	**5**
# Enriched Pathways Total	**115**	**115**	**132**	**120**	**122**	**142**	**185**	**147**	**8**	**7**
# NCAM Pathways	0	9	2	7	0	2	1	2	0	0
# Collagen Pathways	0	26	16	3	11	36	21	15	0	0
# ECM Pathways	0	4	3	0	1	5	4	4	0	0
# Mutations Total	**844**	**444**	**300**	**414**	**1013**	**698**	**548**	**639**	**29**	**49**
# NCAM Pathway Mutations	3	6	3	4	3	7	4	5	0	0
# Collagen Pathway Mutations	9	11	4	5	13	11	7	8	1	0

The “NCAM1 Interactions” and “Interaction of NCAM1 with collagens” pathways, were overrepresented in luminal B and HER2-enriched LN+ tumours. The NCAM1 interaction pathways contain a large number (n = 153) of different proteins where NCAM1 acts as a signal transducing receptor molecule. These functions are tangential to NCAM1’s role in neurite outgrowth and may explain why they are overrepresented in tumours that have not invaded the lymph nodes.

Appendix S4.11.3 Word clouds of overrepresented pathways by subtype.

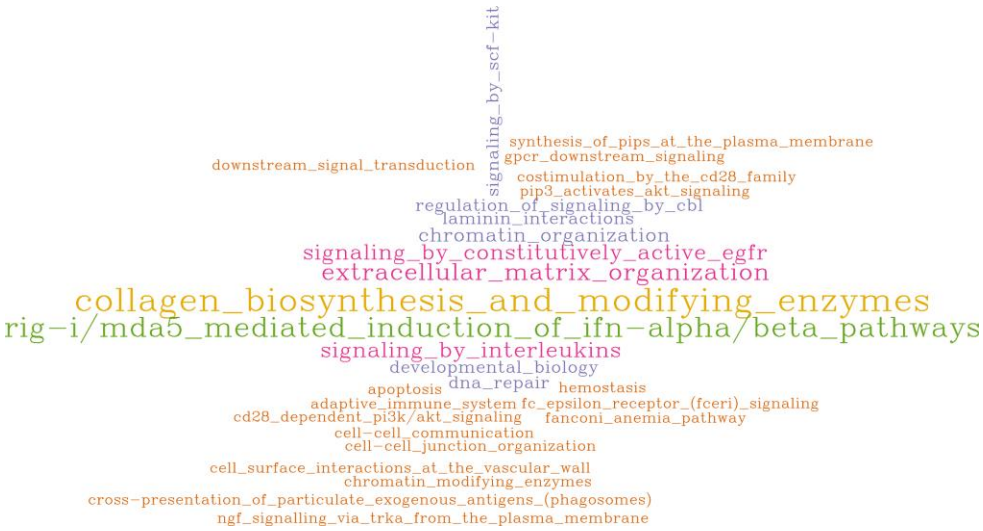
Word clouds of generalized overrepresented Reactome pathways for mutations stratified by lymph node status (positive or negative) and breast cancer subtype (basal-like (A), *HER2*-enriched (B), Luminal A (C), or Luminal B (D)). The size of each word is proportional to its frequency in the abstracted list of overrepresented pathways.

(A) Basal Like Lymph Node Positive

Basal Like Lymph Node Negative

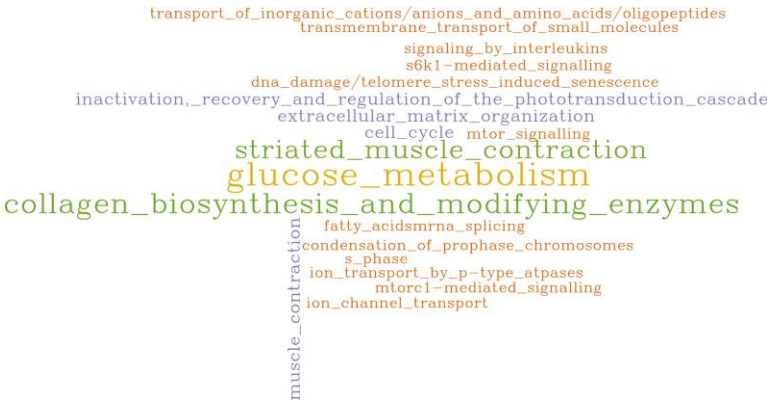
(B) *HER2*-enriched lymph node positive

HER2-enriched lymph node negative



(C) Luminal A lymph node positive

Luminal A lymph node negative



(D) Luminal B lymph node positive

Luminal B lymph node negative

Appendix S4.11.4

Appendix S4.12 Appendix S4 References

1. Saunders, C. T. et al. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28, 1811-1817 (2012).

2. Larson, D. E. et al. Somaticsniper: Identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28, 311-317 (2012).

3. Roberts, N. D. et al. A comparative analysis of algorithms for somatic SNV detection in cancer. Bioinformatics 29, 2223-2230 (2013).

4. McKenna, A. et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297-1303 (2010).

5. https://usegalaxy.org/librarycommon/ldda_info?library_id=f9ba60baa2e6ba6d&show_deleted=False&cntrller=library&folder_id=709338dd4741c085&use_panels=False&id=a78cd737488dc2c1.

6. https://main.g2.bx.psu.edu/library_common/ldda_info?library_id=f9ba60baa2e6ba6d&show_deleted=False&cntrller=library&folder_id=709338dd4741c085&use_panels=False&id=93fcddc692e0986f#.

7. Giardine, B. et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 15, 1451-1455 (2005).

8. Blankenberg, D. et al. Galaxy: a web-based genome analysis tool for experimentalists. Curr. Protoc. Mol. Biol. Chapter 19, Unit 19.10.1-21 (2010).

9. Eswaran, J. et al. RNA sequencing of cancer reveals novel splicing alterations. Scientific Reports 3 (2013).

Appendix S5: Supplementary Information for Chapter 5

Appendix S5.1 Cell Lines Used

Cell Line	Transcriptional subtype	Pac GI50	Gem GI50	Mut	CN	Exp	Pac SVM	Gem SVM	Pac MFA	Gem MFA
Total Count		**49**	**47**	**48**	**46**	**49**	**49**	**44**	**45**	**44**
184A1	Normal-like	7.35	6.16	1	1	1	1	1	1	1
184B5	Normal-like	7.74	6.14	1	1	1	1	1	1	1
600MPE	Luminal	7.51	7.64	1	1	1	1	1	1	1
AU565	Luminal (Her2+)	8.14	7.81	1	1	1	1	1	1	1
BT474	Luminal (Her2+)	7.99	4.88	1	1	1	1	1	1	1
BT483	Luminal	7.00	8.05	1	1	1	1	1	1	1
BT549	Claudin-low	8.16	8.08	1	1	1	1	1	1	1
CAMA1	Luminal	7.95	6.79	1	1	1	1	1	1	1
HCC1143	Basal	7.80	7.92	1	1	1	1	1	1	1
HCC1187	Basal	8.05	5.07	1	1	1	1	1	1	1
HCC1395	Claudin-low	7.71	6.47	1	1	1	1	1	1	1
HCC1419	Luminal (Her2+)	7.04	4.81	0	1	1	1	0	0	0
HCC1428	Luminal	7.58	3.58	1	1	1	1	1	1	1
HCC1569	Basal (Her2+)	7.95	6.76	1	1	1	1	1	1	1
HCC1806	Basal	8.11	8.72	1	1	1	1	1	1	1
HCC1937	Basal	7.81	5.97	1	1	1	1	1	1	1
HCC1954	Basal (Her2+)	8.15	4.51	1	1	1	1	1	1	1
HCC202	Luminal (Her2+)	8.10	4.82	1	1	1	1	1	1	1
HCC2185	Luminal	8.22	7.61	1	1	1	1	1	1	1
HCC3153	Basal	7.70	7.19	1	1	1	1	1	1	1
HCC38	Claudin-low	8.13	8.17	1	1	1	1	1	1	1
HCC70	Basal	8.03	4.58	1	1	1	1	1	1	1
HS578T	Claudin-low	8.38	5.66	1	1	1	1	1	1	1
LY2	Luminal	7.97	7.62	1	1	1	1	1	1	1
MCF10A	Normal-like	8.03	7.70	1	1	1	1	1	1	1
MCF10F	Normal-like	8.08	7.08	1	1	1	1	1	1	1
MCF12A	Normal-like	7.97	7.17	1	1	1	1	1	1	1
MCF7	Luminal	7.79	4.77	1	1	1	1	1	1	1
MDAMB134VI	Luminal	7.99	2.85	1	1	1	1	1	1	1
MDAMB157	Claudin-low	8.27	NA	1	1	1	1	0	1	0
MDAMB175VII	Luminal	7.74	8.12	1	1	1	1	1	1	1
MDAMB231	Claudin-low	8.37	5.93	1	1	1	1	1	1	1
MDAMB361	Luminal (Her2+)	7.79	8.23	1	1	1	1	1	1	1
MDAMB415	Luminal	8.18	6.05	1	1	1	1	1	1	1
MDAMB436	Claudin-low	7.65	7.49	1	1	1	1	1	1	1
MDAMB453	Luminal	7.99	7.85	1	1	1	1	1	1	1
MDAMB468	Basal	8.06	7.01	1	1	1	1	1	1	1
SKBR3	Luminal (Her2+)	7.94	7.97	1	1	1	1	1	1	1
SUM1315MO2	Claudin-low	8.29	6.91	1	1	1	1	1	1	1

SUM149PT	Basal	8.03	7.84	1	0	1	1	0	0	0
SUM159PT	Claudin-low	8.24	7.99	1	1	1	1	1	1	1
SUM185PE	Luminal	6.64	6.44	1	1	1	1	1	1	1
SUM52PE	Luminal	8.20	8.15	1	1	1	1	1	1	1
T47D	Luminal	8.02	6.02	1	1	1	1	1	1	1
UACC812	Luminal (Her2+)	8.08	7.75	1	1	1	1	1	1	1
UACC893	Luminal (Her2+)	7.93	3.54	1	0	1	1	0	0	0
ZR751	Luminal	7.76	7.45	1	1	1	1	1	1	1
ZR7530	Luminal (Her2+)	7.66	NA	1	0	1	1	0	0	0
ZR75B	Luminal	7.38	7.34	1	1	1	1	1	1	1

Pac = paclitaxel, Gem = gemcitabine, GI50 = -log(M), where M is the concentration of drug to inhibit cell growth by 50%, Mut = mutation data (exome sequencing), CN = copy number data (microarray), Exp = expression data (microarray), SVM = support vector machine, MFA = multiple factor analysis. For Mut, CN, and Exp, 1 indicates the data type was available for analysis, 0 indicates the data type was unavailable (from Daemen et al. 2013). For the SVM and MFA columns, 1 indicated that the cell line was included in the analysis, and 0 indicates that the cell line was not included (based on data availability).

Appendix S5.2 Genes Included in the study relevant to paclitaxel and gemcitabine drug disposition

Paclitaxel Genes	Full Gene/Protein Name	Drug Disposition	In Capture Array?
ABCB1	ATP-binding cassette, sub-family B (MDR/TAP), member 1	transporter (out of cell)	YES
ABCB11	ATP-binding cassette, sub-family B (MDR/TAP), member 11	transporter (out of cell)	YES
ABCC1	ATP-binding cassette, sub-family C (CFTR/MRP), member 1	transporter (out of cell)	YES
ABCC10	ATP-binding cassette, sub-family C (CFTR/MRP), member 10	transporter (out of cell)	YES
BAD	BCL2-associated agonist of cell death	in target pathway (*BCL2*)	YES
BBC3	BCL2 binding component 3	in target pathway	YES
BCAP29	B-cell receptor-associated protein 29	associated with resistance	YES
BCL2	B-cell CLL/lymphoma 2	direct target	YES
BCL2L1	BCL2-like 1	in target pathway (BCL2)	YES
BIRC5	baculoviral IAP repeat containing 5	associated with resistance	YES
BMF	Bcl2 modifying factor	in target pathway	YES
CNGA3	cyclic nucleotide gated channel alpha 3	associated with resistance	YES
CSAG2	CSAG family, member 2	associated with resistance	NO
CYP2C8	cytochrome P450, family 2, subfamily C, polypeptide 8	metabolizing enzyme	YES
CYP3A4	cytochrome P450, family 3, subfamily A, polypeptide 4	metabolizing enzyme	YES
FGF2	fibroblast growth factor 2	associated with resistance	YES
FN1	fibronectin 1	associated with resistance	YES
GBP1	guanylate binding protein 1, interferon-inducible	associated with resistance	YES
MAP2	microtubule-associated protein 2	direct target	YES
MAP4	microtubule-associated protein 4	direct target	YES
MAPT	microtubule-associated protein tau	direct target	YES
NFKB2	nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (p49/p100)	associated with resistance	YES
NR1I2	nuclear receptor subfamily 1, group I, member 2	direct target	YES
OPRK1	opioid receptor, kappa 1	associated with resistance	YES
SLCO1B3	solute carrier organic anion transporter family, member 1B3	transporter (into cell)	YES
TLR6	toll-like receptor 6	associated with resistance	YES
TMEM243	transmembrane protein 243, mitochondrial	associated with resistance	YES
TUBB1	tubulin, beta 1 class VI	direct target	YES
TUBB4A	tubulin, beta 4A class IVa	in target pathway (*TUBB1*)	YES

TUBB4B	tubulin, beta 4B class IVb	in target pathway (*TUBB1*)	YES
TWIST1	twist family bHLH transcription factor 1	associated with resistance	YES
Gemcitabine Genes			
ABCB1	ATP-binding cassette, sub-family B (MDR/TAP), member 1	transporter (out of cell)	YES
ABCC10	ATP-binding cassette, sub-family C (CFTR/MRP), member 10	transporter (out of cell)	YES
AK1	adenylate kinase 1	nucleotide metabolism	YES
CDA	cytidine deaminase	metabolizing enzyme	YES
CMPK1	cytidine monophosphate (UMP-CMP) kinase 1, cytosolic	direct target	YES
CTPS1	CTP synthase 1	direct target	NO
DCK	deoxycytidine kinase	metabolizing enzyme	YES
DCTD	dCMP deaminase	metabolizing enzyme	NO
NME1	NME/NM23 nucleoside diphosphate kinase 1	metabolizing enzyme	NO
NT5C	5', 3'-nucleotidase, cytosolic	metabolizing enzyme	NO
RRM1	ribonucleotide reductase M1	direct target	YES
RRM2	ribonucleotide reductase M2	in target pathway (RRM1)	YES
RRM2B	ribonucleotide reductase M2 B (TP53 inducible)	in target pathway (RRM1)	YES
SLC28A1	solute carrier family 28 (concentrative nucleoside transporter), member 1	transporter (into cell)	YES
SLC28A3	solute carrier family 28 (concentrative nucleoside transporter), member 3	transporter (into cell)	YES
SLC29A1	solute carrier family 29 (equilibrative nucleoside transporter), member 1	transporter (into cell)	YES
SLC29A2	solute carrier family 29 (equilibrative nucleoside transporter), member 2	transporter (into cell)	YES
TYMS	thymidylate synthetase	direct target	YES

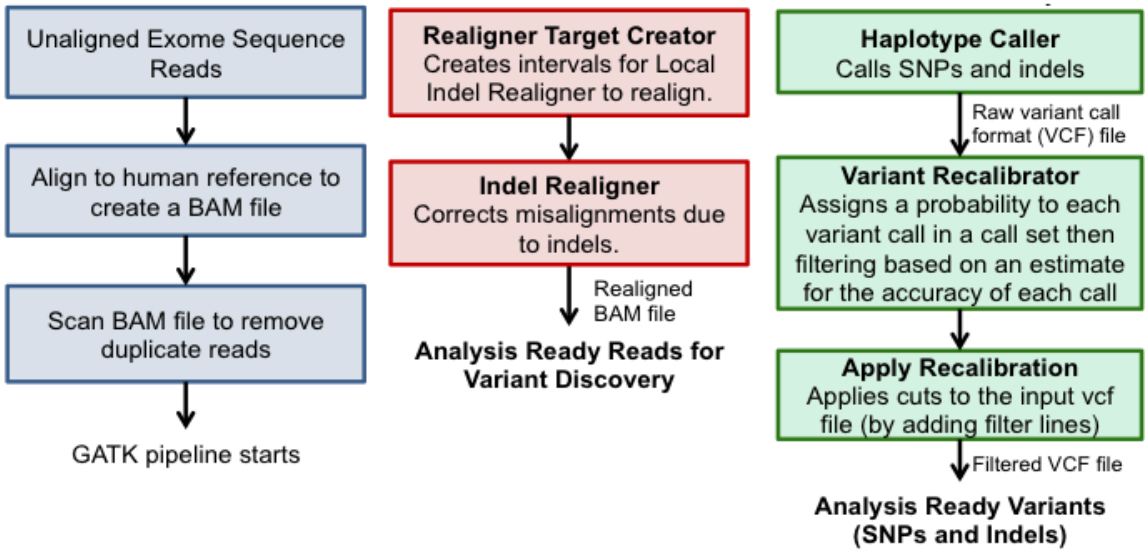
Appendix S5.3 Copy Number Calling Methods

Copy number data were available as CEL files from Affymetrix Genome-Wide Human SNP Array 6.0. CNV calls were generated with the PennCNV software[1] (2011 June 16 version) using the software pipeline and commands found at http://www.openbioinformatics.org/penncnv/penncnv_tutorial_affy_gw6.html. PennCNV output with copy number changes for all cell lines and genes can be found in Supplementary Table 5.1.

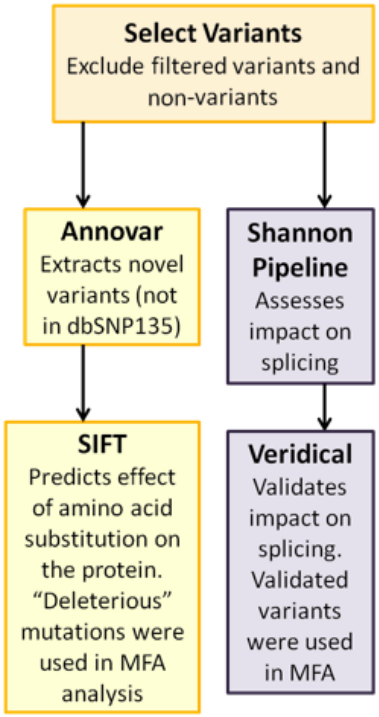
Appendix S5.4 DNA Sequencing Analysis Pipeline– Variant Calling and Interpretation Methods

Whole exome aligned sequencing data were available in the form of .bam files from Illumina Genome Analyzer IIx runs aligned to an hg19 genome build ("NCI60_WES_BAM_files:," n.d.). Variants were detected using the software workflow below (A-D). The Genome Analysis Toolkit (GATK)[2]was used for variant calling and filtering with default parameters (exceptions): Realigner Target Creator, IndelRealigner, Haplotype Caller, Variant Recalibrator (for indels, --minNumBadVariants was set to 5000 for LY2 and SUM159PT), and Apply Recalibration (ts_filter_level for indels was set at 99.0 and for SNPs at 99.9). VariantSelect was called to exclude non-variant loci and filtered loci with the default parameters for this purpose provided by GATK. Annovar[3] was used to annotate the variants (both single nucleotide changes and insertions/deletions) and filter variants present in dbSNP 135. SIFT[4] was used to predict which mutations (SNPs and indels) are likely damaging to the protein product, which were used in further analyses. Two software programs were used for splicing mutation analysis: Shannon Pipeline[5] was used to predict splicing mutations, and Veridical[6] was used to confirm aberrant splicing patterns in cell line-matched RNA-Seq data. In the Multiple Factor Analysis (MFA), mutation status was depicted with a binary variable in which the gene was assigned to be mutated or not. MFAs were also completed with total counts of likely deleterious mutation per cell line, which affected 10 genes, but did not alter the interpretation of the analysis.

A) Bam File Processing B) Bam File Realignment C) Variant Discovery



D) Variant Analysis



Appendix S5.5 Reproducibility of Cell Line Data

Appendix S5.5.1 GI50 Studies

Growth inhibition (GI_{50}) values represent the concentration of the chemical required to inhibit cell growth by 50% in comparison with untreated controls, and the study was carried out as previously described[7]. GI_{50} values were calculated using a sulforhodamine B (SRB) assay, which provides a sensitive method to measure cellular protein content. Cells were grown in 96 well plates for 24 hours, and then exposed to either paclitaxel or gemcitabine for 48 hours. We repeated triplicate GI50 measurements for 5 NCI-60 breast cancer cell lines: *SKBR3*, *HS578T*, *BT549*, *MDAMB231*, and *T47D*. Additionally, we quantified cell densities, and determined growth inhibition in order to resolve drug-induced cytotoxicity. Percent of cytotoxicity was calculated as 100 x (Cell Control – Experimental) ÷ (Cell Control). GI_{50} was then derived using Graph Pad Prism. Data were transformed using X = Log(X) and then a non-linear regression was performed using options: "dose-response inhibition" and "Log [inhibitor] vs. response (variable slope)."

Appendix S5.5.2 CytoScan HD Array

The re-measured microarray analyses for 5 cell lines in our laboratory (*MDAMB231, HS578T, MCF7, T47D*, and *SKBR3*) were completed using the CytoScan HD Array Kit and Reagent Kit Bundle (catalog #901835) following the recommended manufacturer's protocol (Affymetrix, Santa Clara, CA). The AffymetrixGeneChip Command Console Software was used with default options to analyze the .CEL files for copy number change calls, which were visualized and manually confirmed using the Chromosome Analysis Suite (version 2.1.0.16).

Appendix S5.5.3 Gene Capture and DNA Sequencing

Capture probes were designed (genomic coordinates of probes listed in Supplementary Table 5.6, and then produced on a cleavable microarray using Custom Array Microarray Synthesizer (Bothell, WA). Exons and 300 bp into the introns, for 44 of the 49 genes, were targeted. Genomic DNA was sheared to ~300bp fragments using the Covaris S220 Focused-ultrasonicator. Library preparation was carried out using the KAPA Biosystems

Standard High Throughput Library Preparation Kit and RNA bait from the capture array probes was used to enrich for the genes of interest[8]. DNA samples were quantified using qPCR (KAPA Library Quantification Kit for Illumina Platform) and then paired end reads (70 bp each side) were obtained using the standard Illumina Genome Analyzer IIX paired-end sequencing protocol.

Sequences of all exons (and 300 bp into each intron) for the 45 genes were selected using an *ab initio* approach[9]. Probe sequences were selected using PICKY 2.2 software[10] using the default settings with few exceptions (65ºC Tm, 30-70% GC content, 5 probes per sequence, 20 nt maximum overlap). MPI-BLAT was used to ensure the probes align only to the targeted sequence.

Generation, Cleavage and Purification Microarray Oligos

The selected sequences, with primer binding sites added to each end (5' ATCGCACCAGCGTGTN$_{36\text{-}70}$CACTGCGGCTCCTCA), were then synthesized onto two cleavable 12K microarray chips using a CustomArray Microarray Synthesizer (Bothell, WA). Probes were cleaved from the microarrays with concentrated (14.5N) ammonium hydroxide at 65ºC for 4 hours. Purified oligos were then amplified by 25 cycles of conventional PCR using KapaHiFi DNA Polymerase (KapaBiosystems). Biotin-labelled RNA bait was generated from this product with nested PCR on the amplified oligosusgin a MAXIscript SP6 *in vitro* transcription kit (Ambion) with a UTP to biotin-16-UTP (Roche) ratio of 4 to 1.

Sample Preparation

Genomic DNA (gDNA) from the MDAMB231 cell line was diluted to 100 ng/µL in a volume of 51 µL for S220 Focused-ultrasonicator (Covaris) shearing (150-300nt fragments generated with the following settings: Time 120 sec, Duty cycle 10%, Intensity 5, and Cycles per burst 200).

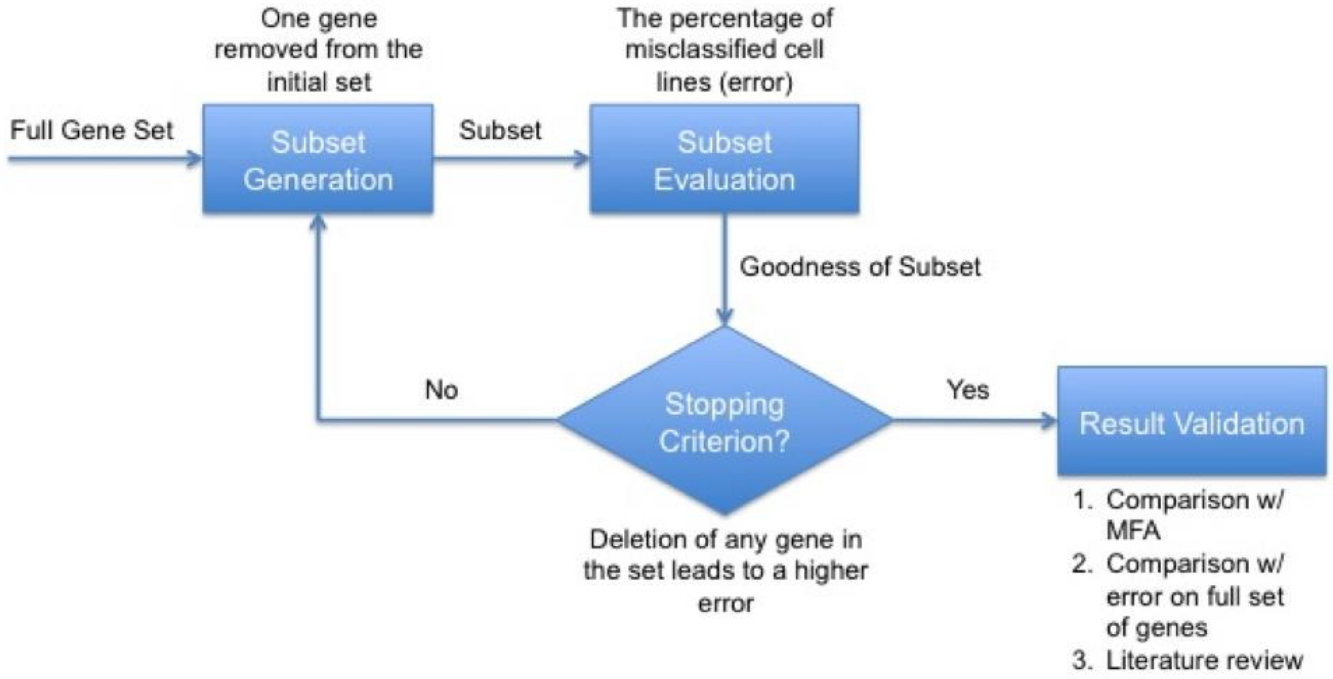
Library Preparation and Capture

The sheared samples were prepared using a KAPA Biosystems Standard (KK8200) and High Throughput (KK8234) Library Preparation kits, following the manufacturer's protocol (KapaBiosystems). Genes of interest were captured using the Tiled RNA bait, using the hybridization selection protocol from Gnirke et al (2009) with 1 to 2 ug of sample prep, 1.5 ug of RNA bait, and 75 uL of M-280 streptavidin Dynabeads (Invitrogen). DNA samples were quantified using qPCR following the protocol outlined by KAPA Library Quantification Kit for Illumina Platform (KAPA Biosystems, catalog# KK4824). Samples were then treated to standard Illumina paired-end sequencing on a Genome Analyzer IIx, with 70 bp, then a 7 bp index (used during multiplexing), and then 70 bp.

Bioinformatic Analysis

When sequencing was completed, data was demultiplexed (when necessary) and aligned to the human reference genome (hg19) using CASAVA v1.8.2 and CRAC (v1.3.0). BAM files were prepared for variant calling using Picard, and variant calling was performed on both sets of aligned sequences using the UnifiedGenotyper tool in the Genome Analysis Toolkit (GATK). Variants called outside of target regions were ignored. Variant analysis was completed as outlined in Supplementary Methods IID.

Appendix S5.6 Support vector machine feature selection



Adapted from Dash and Liu (1997)[1]

Appendix S5.7 Partial-Least Squares Regression

A partial-least squares regression (PLSR) was also performed to attempt to relate genomic findings to paclitaxel response, based on the fact that GI50 is a continuous variable. The predictive error of the model was measured by taking the absolute sum of the residuals (the actual GI50 minus the predicted GI50) of a leave-one-out cross-validation. One cell line at a time was left out of the analysis and its paclitaxel GI50 value was predicted using the beta values given by the regression line and then compared with its measured GI50 value.

Using the absolute sum of error as a measurement of predictive accuracy, we randomly selected subsets of genes ranging in number from 1 gene to 30 genes (out of a total of 31 genes) for 1,000 iterations each to attempt to find the most optimal number of genes. Of the 9 paclitaxel genes with the lowest error, two million model iterations were performed to find the best predictive subsets with the lowest error values. However, the lowest absolute sum of residual errors was ~10. The high residual means imply a lack of confidence that the genomic signature will reliably predict GI50. For this reason, we discontinued attempts to use PLSR to predict gemcitabine (or paclitaxel) chemosensitivity.

Appendix S5.8 Gene expression and Copy Number analyses on FFPE tumour blocks

Nucleic acids were extracted from the FFPE tissue samples using Qiagen'sAllPrep DNA/RNA FFPE Kit (Cat. No. 80234, Venlo, Limburg, Netherlands). The recommended protocol was used with the following exceptions: 1) Tissue used for the nucleic acid extraction was obtained using 1 mm Miltex Sterile Disposable Biopsy Punches (Cat. No. 33-31AA-P/25, Plainsboro, New Jersey), as opposed to using thin slices of the full block. Hematoxylin and eosin stained slides of each tissue block were marked by a pathologist to identify cancerous lesions and direct specific regions to punch. Using a biopsy punch allowed for targeted extractions, and minimized the amount of normal surrounding tissue used in the analysis. 2) 75 µl of mineral oil was used for tissue deparaffinization at 90°C

for 20 minutes, as previously described[12]. 3) The first proteinase K incubation was performed at 56°C for 2 hours.

cDNA was produced from tumour RNA using SuperScript II Reverse Transcriptase (Cat. No. 18064-014, Invitrogen, Carlsbad, CA, USA) and 250ng IDT ReadyMade random hexamer per reaction. (Cat. No. 51-01-18-25, San Jose, CA, USA). cDNA synthesis was carried out following the manufacturer's protocol, and purified using ethanol precipitation with 0.1X sodium acetate and 2.5X 100% anhydrous ethanol. Every cDNA sample used for gene expression measurement was then re-suspended in RNAse-free water, and diluted to 20 ng/µl for gene expression measurement. Purified DNA from the QiagenAllPrep columns were diluted to 9 ng/µl for copy number analysis.

qPCR (gemcitabine copy number genes only) and qRT-PCR (all expression genes) were performed with the SensiFast SYBR No-ROX kit (Cat. No. BIO-98020, Bioline, London, UK) using the recommended protocol. Primer pairs were designed using PrimerQuest (Integrated DNA Technologies, Coralville, Iowa), spanning exons when possible (qRT-PCR only). Each primer pair was optimized using duplicate 10 µl reactions for forward and reverse primer concentrations, and in some cases annealing temperatures. Primer sequences, annealing temperatures, and final concentrations used are listed in Table S5.6.1. *NT5C* qPCR was performed with a 10-second denaturation and 20-second extension in every cycle. All real-time PCR experiments used an Eppendorf Mastercycler realplex machine and followed the program: 95°C for 2 min, and 40 cycles of 5 s at 95°C, 10 s at 60°C (copy number and some gene expression primers) or 64.5°C (only gene expression primers), and 15 s at 72°C. A melting curve was measured for all reactions, and any measurements with abnormal melting curve (ie. at a lower temperature due to primer diming) were removed from any further analysis. Two 10 µl reactions were performed per primer pair, per sample.

Table S5.6.1					
Gene Name	**Amplicon Size**	**F/R**	**Sequence**	**Anneal. Temp.**	**Opt. Conc. (nM)**
PACLITAXEL - qRT-PCR					
ABCC10	109	F	TGGGAAGACATTTGATGCAC	64.5	400
		R	CTTCTCCCCCACCTCTGTCT		900
BCL2	90	F	CCTGTGGATGACTGAGTACCTGAA	64.5	400
		R	GGGCCGTACAGTTCCACAAAG		900
BCL2L1	94	F	CTTGGATGGCCACTTACCTGAATG	64.5	900
		R	GCATTGTTCCCATAGAGTTCCACAA		400
BIRC5	113	F	GCAGTTTGAAGAATTAACCCTTGGTG	64.5	900
		R	CCGCAGTTTCCTCAAATTCTTTCTTC		900
FGF2	86	F	AGAAGAGCGACCCTCACATCAA	64.5	900
		R	GTAACGGTTAGCACACACTCCTTTG		900
FN1	120	F	TTGGAGATTCATGGGAGAAGTATGTG	64.5	900
		R	CAGGACCACTTGAGCTTGGATAG		900
MAP4	91	F	TCCTCTCCTGGATGTTGATGAGAA	64.5	900
		R	AGATGGAGTATCTTCAATCTGGCTAGT		900
MAPT	93	F	GGCTCATTAGGCAACATCCATCATAA	64.5	900
		R	CTTCGACTGGACTCTGTCCTTGA		900
NFKB2	101	F	AGATGACATTGAGGTTCGGTTCTATG	64.5	400
		R	ACACAATGGCATACTGTTTATGCAC		400
TLR6	115	F	CCGACGGAAATGAATTTGCAGTAGAC	64.5	900
		R	AGCTCAGCGATGTAGTTCTGAGAC		900
TMEM243	104	F	AGGACTTTGCTACCAGGACCTAC	64.5	900
		R	GCTGCCAACAACTAAATTGATGATTCG		100
GEMCITABINE - qRT-PCR					
ABCC10	109	F	TGGGAAGACATTTGATGCAC	64.5	400
		R	CTTCTCCCCCACCTCTGTCT		900
CMPK1	84	F	GGGAAAGAGTAGTGGTAGGAGTGATG	64.5	900
		R	ATTGGCTTTGTTGACTGAAGGTAGG		400
DCTD	96	F	TACCATGATAGTGACGAGGCAACTG	64.5	400
		R	GACAATCTTGCTGCACTTCGGTATG		900
NME1	120	F	CCTTCATTGCGATCAAACCAGATG	64.5	400
		R	GATCTTCGGAAGCTTGCATGAATTT		400
RRM1	103	F	ACTATTTATTATGGTGCTCTGGAAGCC	64.5	400
		R	ACTGAAGAATTCCTTTGCTAACTGGAG		900
RRM2B	80	F	TCTGGCTAAAGAAGAGAGGTCTTATGC	64.5	400
		R	ACAGTGAAGTCCTTCATCTCTGCTG		400
GEMCITABINE - qPCR					
ABCC10	93	F	GAGAATAGTAGTAGCTTACCTTGTAG	60	400
		R	CATGTATTCAGAGCTTACTTTGTG		400
NT5C	126	F	CCTTGTCAGGATAATTCGTTCTAC	57.2	400
		R	CCAAGTCCCTATCCCTGAAT		400
TYMS	107	F	GTATGTCAGCCTTTCCCTTC	60	400
		R	CAGTGAACACGAGAAACAAATC		400
STANDARDS - qRT-PCR					
ACTB	101	F	TTGTTACAGGAAGTCCCTTGCC	64.5	400
		R	ATGCTATCACCTCCCCTGTGTG		400
B2M	86	F	TGCTGTCTCCATGTTTGATGTATCT	60	400
		R	TCTCTGCTCCCCACCTCTAAGT		400
GAPDH	87	F	TGCACCACCAACTGCTTAGC	60	400
		R	GGCATGGACTGTGGTCATGAG		900
STANDARDS - qPCR					
ACTB	101	F	TTGTTACAGGAAGTCCCTTGCC	64.5	400
		R	ATGCTATCACCTCCCCTGTGTG		400
RMND5A	99	F	GCCAGCTTCTGAATTATGGTCTTC	60	400
		R	GAAACTCAATGGAACCTTCTGTTTC		400

Expression values were normalized per sample based off of three genes: *ACTB*, *B2M*, and *GAPDH* using the equation (as previously described[13]):

$$\boldsymbol{expressionvalue} = \mathbf{2}^{-\Delta \boldsymbol{Ct}} = \mathbf{2}^{-(\boldsymbol{Ct\ of\ GOI - Ct\ of\ Average\ Standards})}$$

Gene expression values for the FFPE samples were clustered as described in main Methods.

For copy number, 5 or 6 dilutions from hgDNA (9, 3, 1, 0.33, 0.11, and in some cases 0.037 ng/µl) were used to construct a standard curve for each primer pair (Figure S5.6.1). Two reference genes (*ACTB* and *RNMD5A*) were used to normalize for sample variation. *ACTB* is a single copy gene (1 haploid gene), and *RMND5A* is a multicopy gene (3 haploid copies[14]). DNA from 9 lymph-node negative samples were used as normal controls to adjust for differences in primer efficiencies. Copy number calling was determined as previously described[15]:

Ct Values were measured using two 10 µl reactions for each sample and gene.

Raw copy values were derived from the equation of the standard curves for each gene (Figure S5.6.1), where y = Ct and x = log(Q).

Copy number calibration was performed per gene for each sample (both tumour and normal) by dividing the raw copy value call by the average of the copy value call of ACTB and RMND5A for that sample. This adjusted for differences in Ct values between samples.

$$calibrated\ copy\ number = \frac{raw\ copy\ value}{average\ (ACTB\ and\ RMND5A\ raw\ copy\ values)}$$

Final copy number values were determined by adjusting for the average calibrated copy numbers of the normal (lymph node negative) samples to adjust for differences in primers/gene measurements.

$$final\ copy\ number = \frac{calibrated\ copy\ number\ per\ sample}{average\ calibrated\ copy\ number\ of\ 9\ normals}$$

Copy number gains and losses were determined if the copy number call was at least 3 standard deviations (of *ACTB* and *RMND5A*) from the mean copy number for that gene (see Figure S5.6.2 for copy number changes highlighted in yellow, and Supplementary Table 5.3 for copy number calls). Because no copy number changes were expected for ACTB and RMND5A, the average standard deviation between the two, when calibrated against each other, was used (standard deviation = 0.06269). Any copy number gains were assumed to be a copy of 3, whereas losses were assumed to be 1.

Figure S5.6.1 – Copy number standard curves

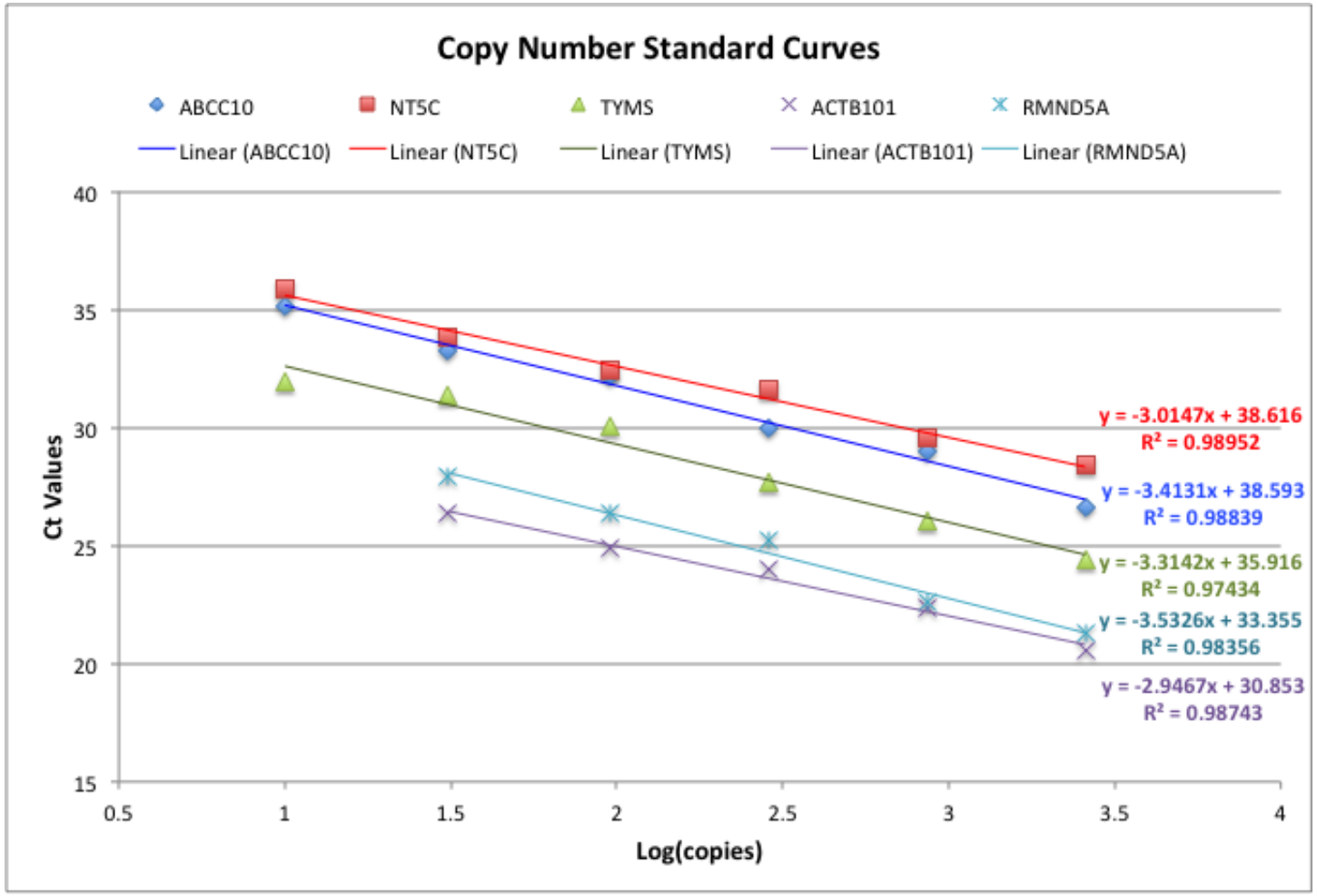
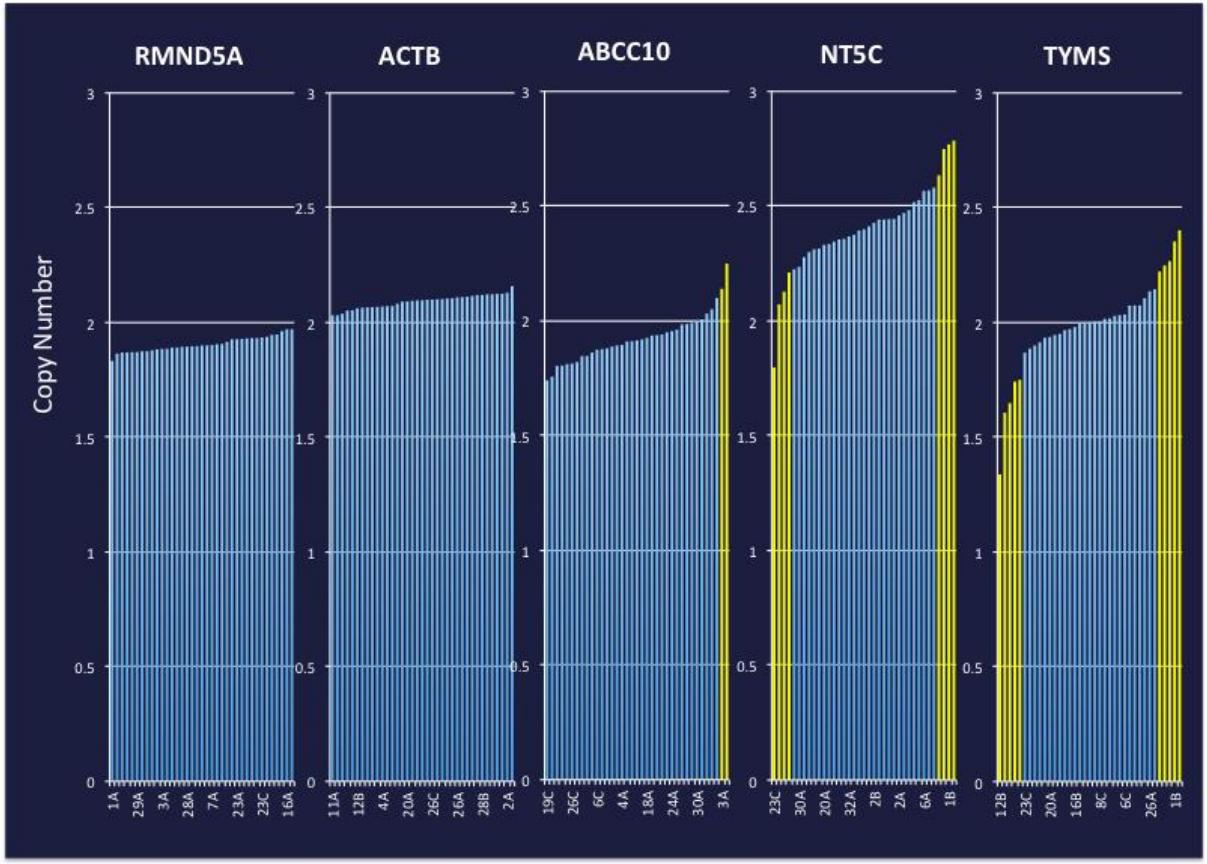


Figure S5.6.2 – Copy number gains and losses per gene



Note: For illustrative purporses only. Each gene lists samples (x-axis) in a different order, and not all samples are labeled. See Supplementary Table 5.3 for exact copy number calls per sample and gene.

Appendix S5.9 Reproducibility of Data

To assess reproducibility of the data used to derive the genomic signatures for paclitaxel and gemcitabine, we sought to determine the degree to which a sample of the cell line data was consistent between cell line sources. We obtained a subset of the cell lines used in the prior study by Daemon et al. from the Coriell Institute, and, redetermined their GI_{50} and copy number values as well as the variants present in the candidate gene sequences in one of the lines. Growth inhibition studies were carried out for 5 breast cancer cell lines (BT549, MDAMB231, HS578T, T47D, SKBR3) to determine the reproducibility of cell line sensitivity to paclitaxel and gemcitabine (Figures 5.7.1 and 5.7.2). Re-measured GI_{50} values were compared to GI_{50}s from Daemen et al. and those previously reported from Ring et al., 2008. The standard deviations of the GI_{50} values between studies were low for all measurements, except for the SKBR3 treated with gemcitabine (GI_{50} for SKBR3 was not determined by Ring et al., 2008). Although the differences in cell line growth inhibition were minimal (< 1 $\log_{10}$), our results were more similar to those reported by Daemen et al., 2013. The standard deviations between replicates from the Ring et al. study were more than twice our measurements for the same cell lines, except for BT549 paclitaxel and MDAMB231 gemcitabine GI_{50} values. In some instances, substitution of our GI_{50} values (or those obtained by Ring et al., 2008) for those determined by Daemen et al., 2013, could affect the subsequent classification of the cell line. For paclitaxel, triplicate assays of 4 of 5 lines (all but HS578T) exhibit GI_{50} values close to the median GI_{50} threshold for distinguishing sensitivity from resistance ($-\log_{10}M = 8$). For gemcitabine, a single cell line (SKBR3, $-\log_{10}M = 7$) was close to this threshold. This highlights the importance of conducting genomic analyses and GI_{50} studies on the same source line, given that clonal variation and genetic evolution can occur in cancer cell lines[16].

Copy number data of 5 cell lines (MDAMB231, T47D, MCF7, HS578T, and SKBR3) were measured using an AffymetrixCytoScan HD array and analyzed using the Affymetrix Chromosome Analysis Suite (ChAS; CytoScan HD data). The AffymetrixCytoScan HD array contains approximately 2.6 million copy probes and 750,000 SNP probes, whereas the Affymetrix SNP 6.0 array contains approximately

946,000 copy number probes and 906,600 SNP probes. DNA from MDAMB231 was extracted in May 2010 and again in February 2013 to compare different time points/passages of the cell line from the same batch. The copy number calls of the SNP 6.0 DNA data from the Daemen et al., 2013 study analyzed by PennCNV, and re-analyzed by ChAS, were compared with our CytoScan HD data (Supplementary Table 5.11). Copy number changes between the two time points of MDAMB231 were the same for all 49 genes. Copy number calls between the Daemen et al. and CytoScan HD data were largely concordant. Of the 49 genes and 5 cell lines (total of 245 copy number calls), 151 were the same (62%), and an additional 6 (2%) were concordant between our CytoScan HD data and PennCNV, but not the Daemen et al. data analyzed by ChAS. 33 (15%) of the copy number calls were different between our CytoScanHD data and Daemen data, but these appear to be real differences between the cell line karyotypes, because PennCNV and ChAS were consistent for the Daemen et al. data. Conversely, 33 (15%) copy number changes were inconsistently called between the PennCNV analyses, and the ChAS analyses of both data sets. In these cases, it is likely that PennCNV miscalled the copy number state. None of these copy number changes occurred in *NT5C, ABCC10,* and *TYMS,* which were present in the final SVM model for gemcitabine resistance. Another 22 (9%) copy number calls were inconsistent between our CytoScan HD data and PennCNV. Upon further analysis of the Daemen data set with ChAS, these differences appear to be due to noise in the SNP 6.0 data. One possible explanation is that SNP 6.0 probes neighboring conserved repetitive elements exhibit higher variation in signal intensities than probes in the Cytoscan HD, which are located further away from these sequences[9]. Inconsistencies may be also due to heterogeneous populations of mixtures of tumour cells each with different copy numbers within these populations. Concordant calls, the different noise levels in the data, and ambiguous copy number calls (ie. between a copy number of 1 and 2), and actual copy number differences are indicated in Figure 5.7.3.

The relevant gene sequences from MDAMB231 were derived using next generation sequencing with a custom oligonucleotide enrichment reagent that targeted 44 of the 49 genes (Supplementary Table 5.12; *CSAG2, CTPS1, DCTD, NME1,* and *NT5C* are not included). Results were compared with MDAMB231 exonic sequences

("NCI60_WES_BAM_files:," n.d.). In our analysis, which also includes newly determined intronic sequences flanking each exon (300 nt), 59 mutations were detected (Supplementary Table 5.12). Five variants were predicted to be damaging by SIFT and 37 were reproduced in both studies (36 SNPs and 1 insertion), of which 35 were known variants present in greater than 1% of the population, and 2 were novel. None of the damaging mutations were used in the MFA for MDAMB231, because the only likely damaging mutations between the two data sets were known, frequent variants.

GI50s (GI50 drug concentrations are in $-\log_{10}M$) were re-measured for paclitaxel and gemcitabine in cell lines BT549, HS578T, MDAMB231, T47D, and SKBR3, and then compared to 2 sets of previously published values[17,18]. The yellow bar indicated the GI50 threshold for resistant (below the line) and sensitive (above) cell lines.

Figure 5.7.1 – GI50s for Paclitaxel

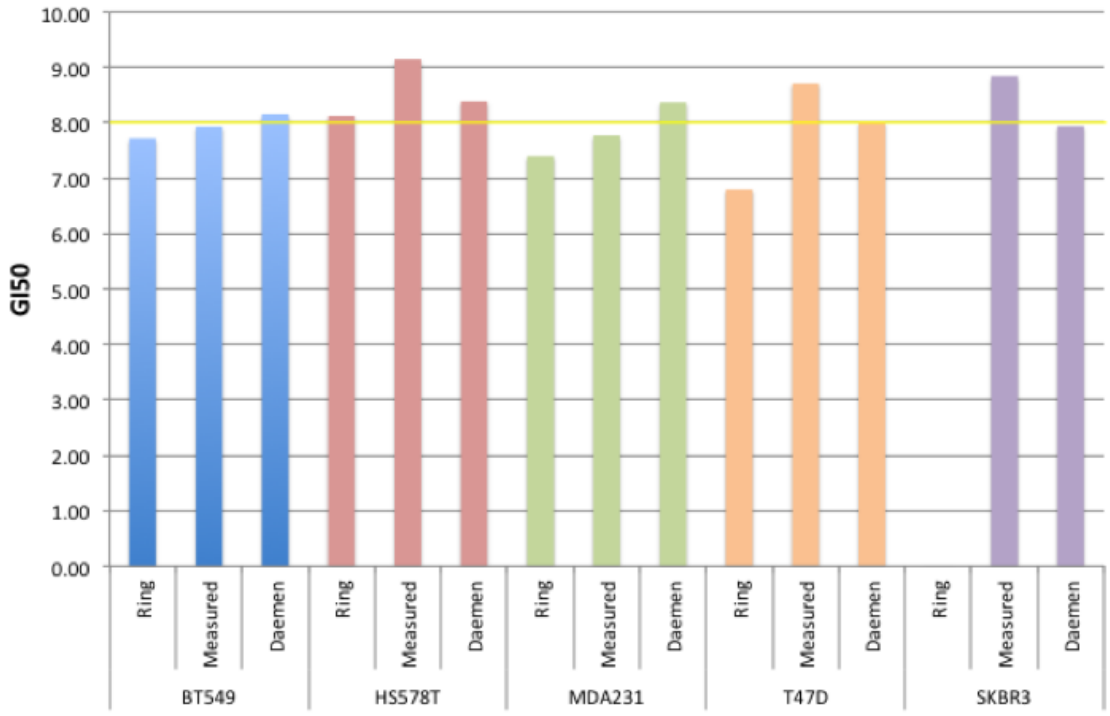


Figure 5.7.2 – GI50s for Gemcitabine

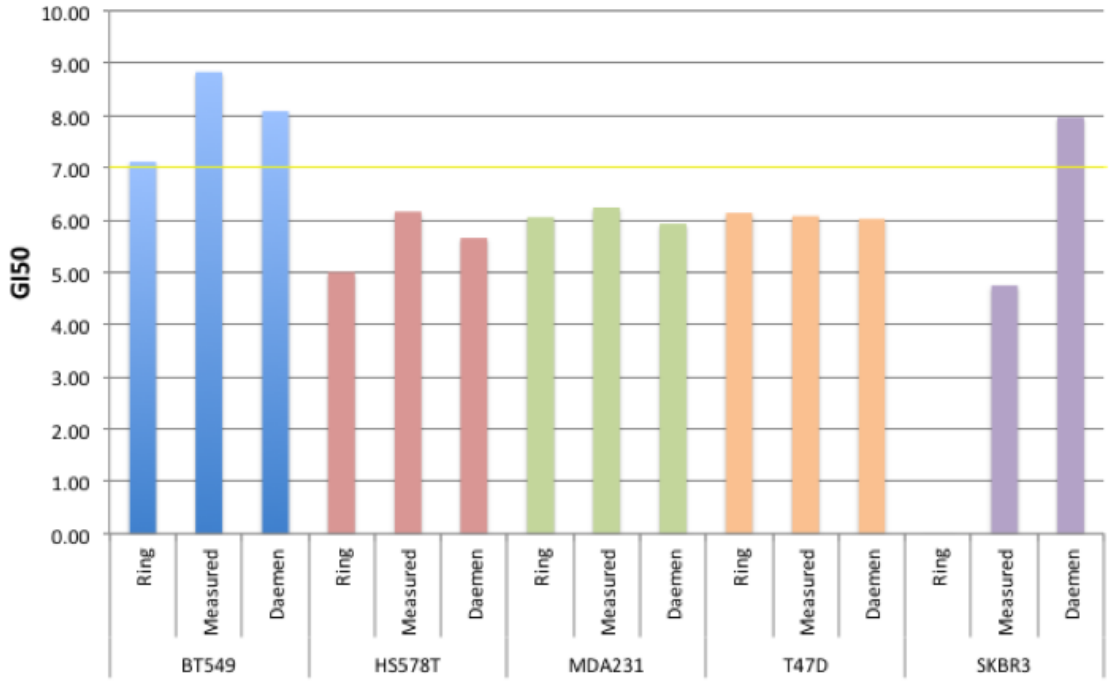


Figure 5.7.3 – MDAMB231 Copy Number Analysis

MDAMB231: Copy number analysis was performed using an AffymetrixCytoScan HD with DNA extracted from MDAMB231 in February of 2013 (dark blue in screen shots), and May of 2010 (pink). Both time points were compared to the Affymetrix SNP 6.0 data from Daemen et al. (light blue). Screen shots from ChAS are displayed for *ABCC10, NT5C, OPRK1, and TYMS.* Log 2 Ratios (green and red bars), copy number state, smooth signal, and genes are displayed for all three analyses (top to bottom). Log2 Ratios are displayed using a heat map between -1.5 and 1.5 (below).

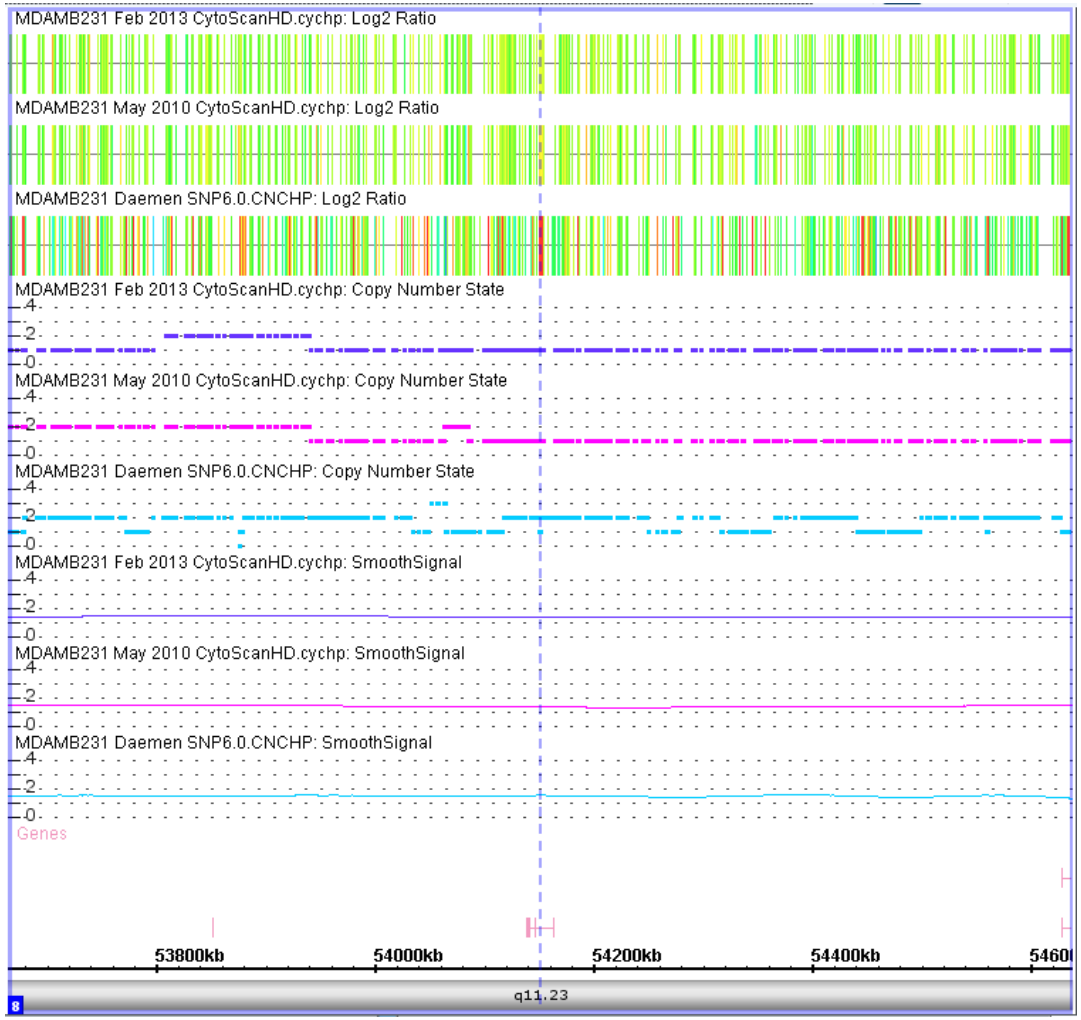
Log2 Ratio Heat Map:

-1.5 0 1.5

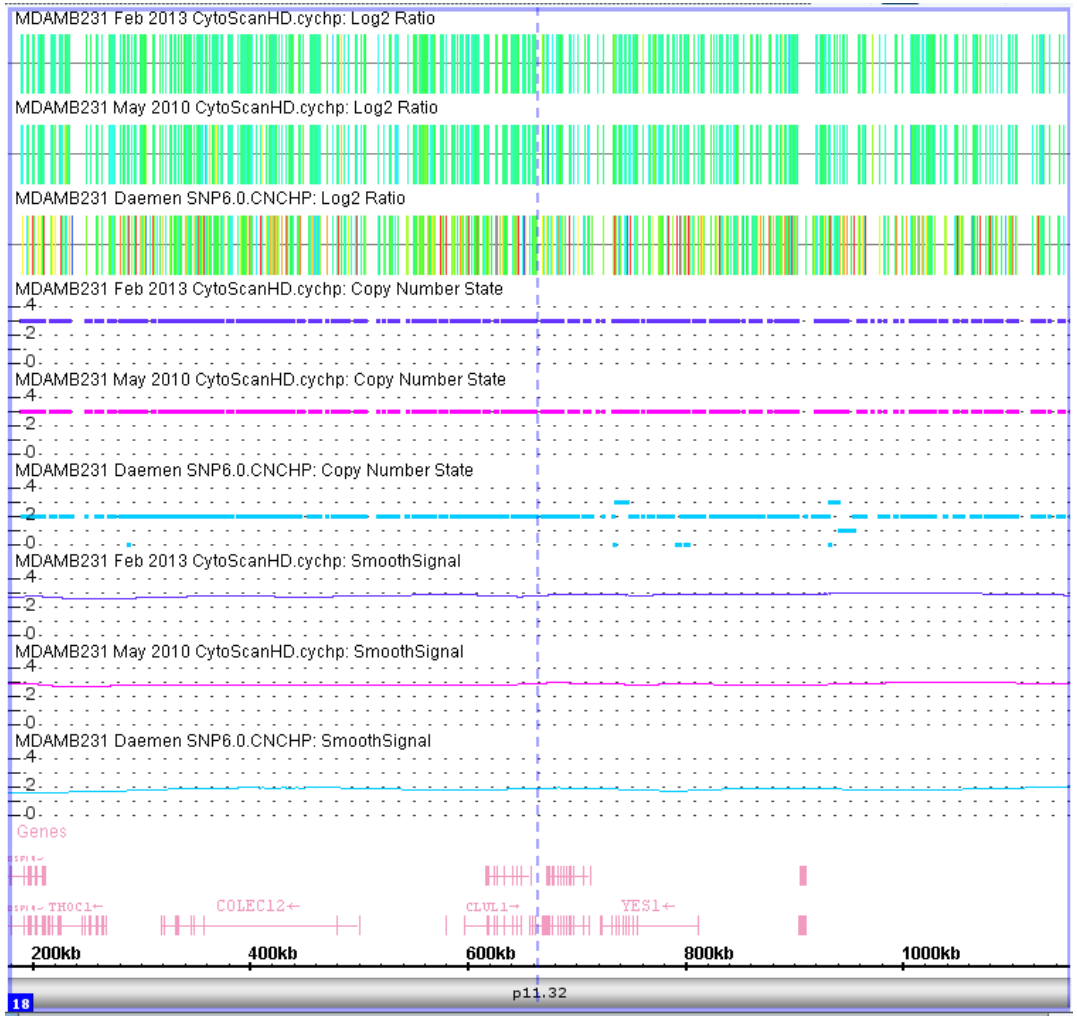
***ABCC10* (at dotted line),** chr6: 42,943,397 – 43,876,083, PennCNV call for Daemen data set: 2. A small deletion is detected by ChAS because the smooth signal drops below 1.5.It isn't clear whether it is a real copy number change because it is small, and the Log2Ratios are noisy (red and green bars).

NT5C (at dotted line), chr17: 72,625,852 – 73,628,356, PennCNV call for Daemen data set: 2. Normal copy number of 2 is seen in all three analyses. There is a larger range of log 2 ratios seen in the SNP 6.0 array.

OPRK1 (at dotted line), chr8: 53,664,288 – 54,638,171, PennCNV call for Daemen data set: 2. A deletion of 1 copy is evident for the two re-measured sets, and may be present in the Daemen et al. data. PennCNV called this region a copy number of 2, although the data is noisy and the smooth signal is between a copy of 1 and 2. This is an example where the noise in the data, or the cells are mosaic, may explain the discordant results between Daemen et al. and re-measured data.



TYMS (at dotted line), chr18: 177,496 – 1,153,659, PennCNV call for Daemen data set: 2. This demonstrates an example where the copy number is different between the Daemen data set and the re-measured data set. This appears to be a real change, as PennCNV and ChAS both clearly call a copy number of 2, but there is one extra copy detected in the re-measured data.



Appendix S5.10 MFA Criteria

Classification	Criteria *(all are required for any given classification)*				
	RV coefficient* of Factor	***cos2 value* of Factor***	***RV Coefficient* of GI50**	***cos2 value* of GI50**	**% variance explained**
Strong Relationship (Str Rel)^	>0.6	>0.4	>0.6	>0.4	>25%
Relationship (Rel)~	>0.5	>0.25	>0.5	>0.25	>25%
Possible Relationship (Pos)	>0.3	>0.1	>0.5	>0.25	>25%
*gene expression, copy number or mutation status, ^all dimension 1, ~dimensions 1 and 2					

Appendix S5.11 Multiple Factor Analysis

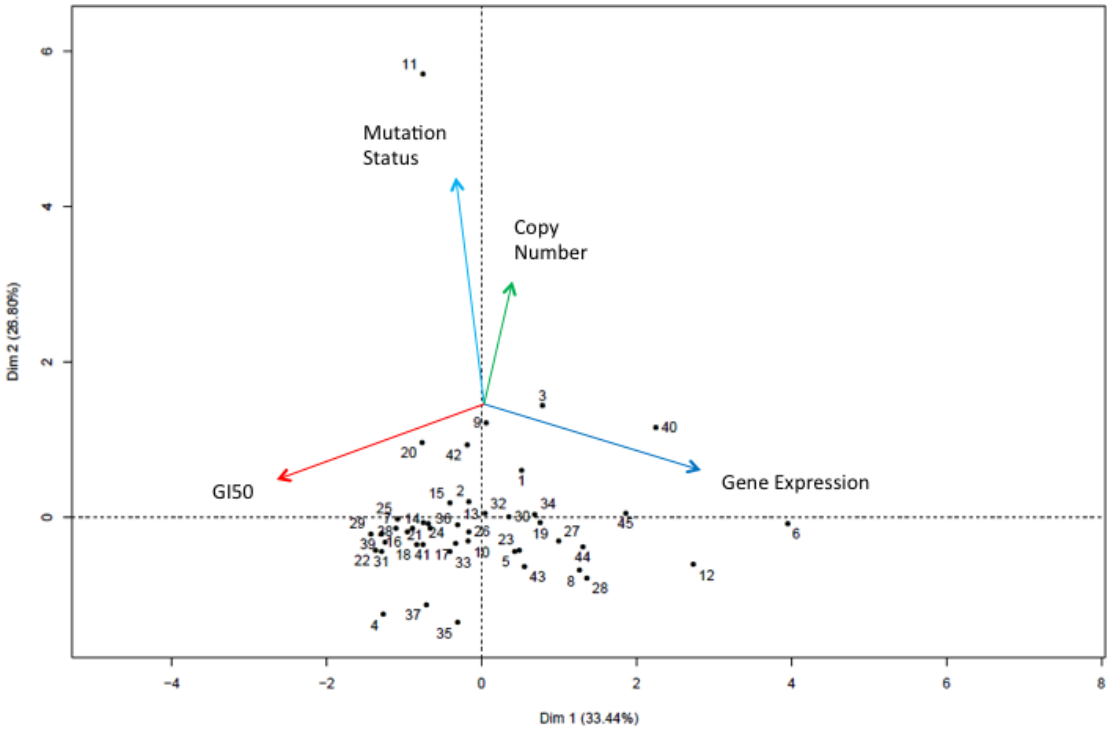
Appendix S5.11.1 Cell Line Numbers in the Multiple Factor Analyses Individual Factor Maps:

Cell Line	Pac MFA #	Gem MFA #	Cell Line	Pac MFA #	Gem MFA #	Cell Line	Pac MFA #	Gem MFA #
184A1	1		HCC1954	16		MDAMB231	31	30
184B5	2		HCC202	17		MDAMB361	32	31
600MPE	3		HCC2185	18		MDAMB415	33	32
AU565	4		HCC3153	19		MDAMB436	34	33
BT474	5		HCC38	20		MDAMB453	35	34
BT483	6		HCC70	21		MDAMB468	36	35
BT549	7		HS578T	22		SKBR3	37	36
CAMA1	8		LY2	23		SUM1315MO2	38	37
HCC1143	9		MCF10A	24		SUM159PT	39	38
HCC1187	10		MCF10F	25		SUM185PE	40	39
HCC1395	11		MCF12A	26		SUM52PE	41	40
HCC1428	12		MCF7	27		T47D	42	41
HCC1569	13		MDAMB134VI	28		UACC812	43	42
HCC1806	14		MDAMB157	29	N/A	ZR751	44	43
HCC1937	15		MDAMB175VII	30	29	ZR75B	45	44

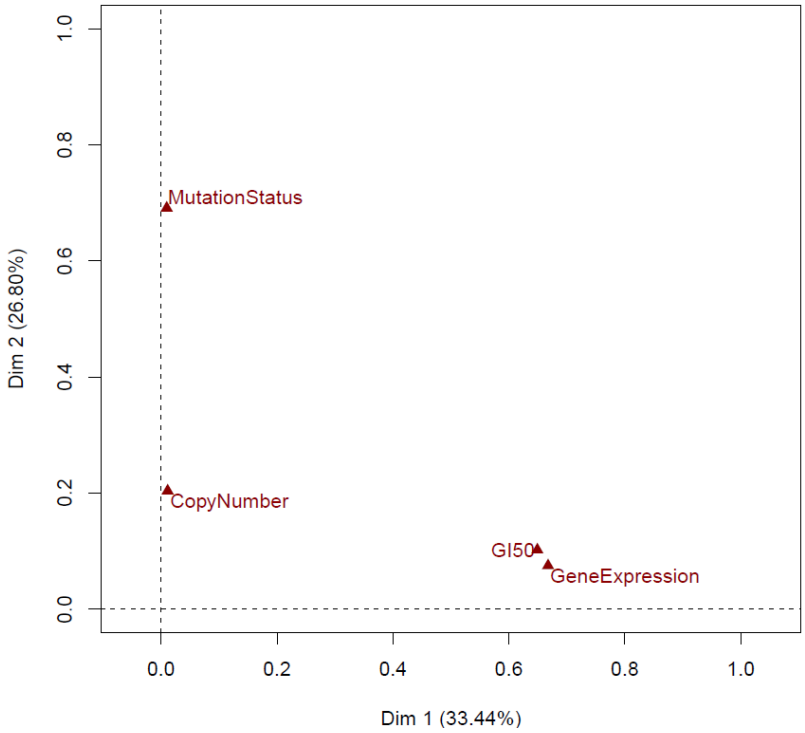
Individual factor maps, correlation circles, and groups representations are all formatted the same throughout the document. Factors (copy number, gene expression, mutation status and GI50) are labeled in the correlation circle arrows (overlaid on the individual factor map) and the groups representation. Cell lines are numbered according to the legend in the table of contents. Additional quantitative details of the MFAs can be found below.

Appendix S5.11.2 Paclitaxel Example - MAPT

Individual Factor Map – Dimensions 1 and 2 (% variance explained in brackets)

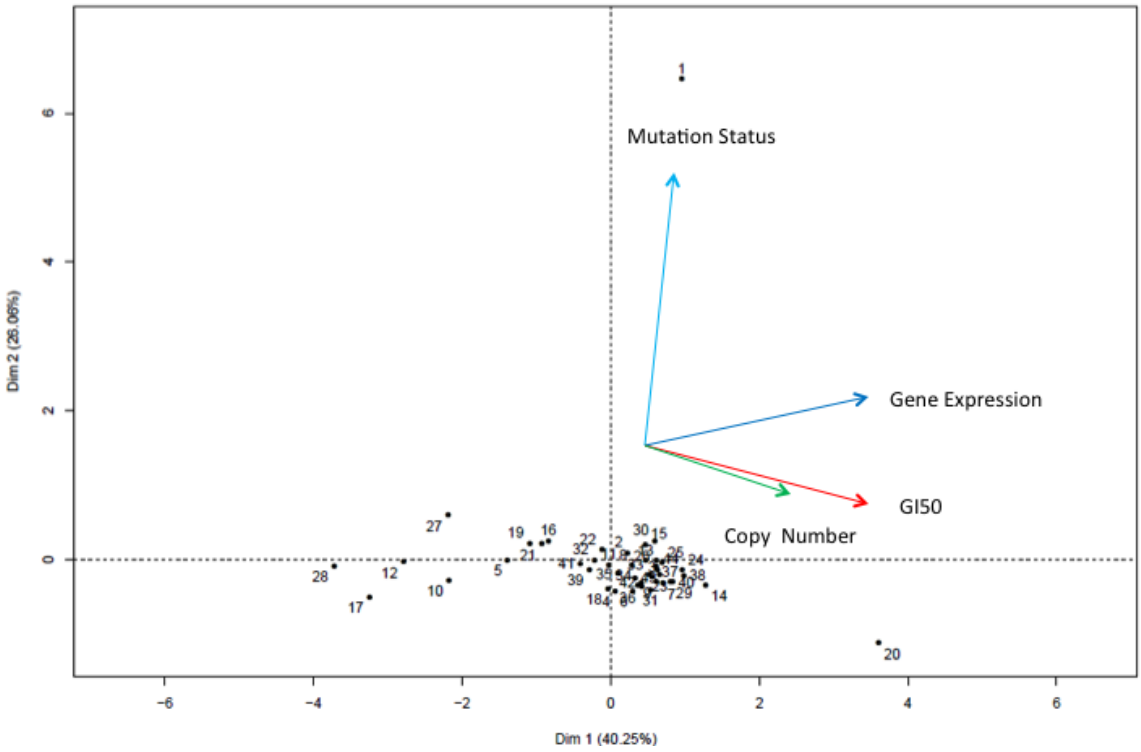


Groups Representation

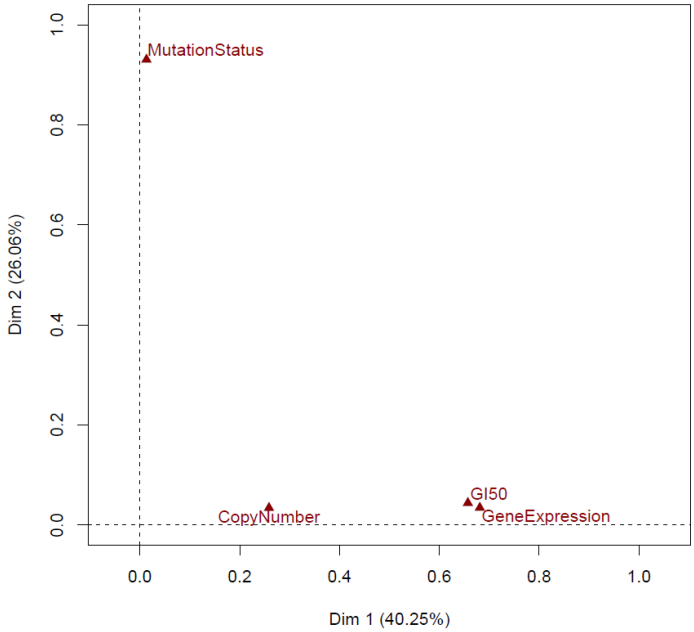


Appendix S5.11.3 Gemcitabine Example - DCTD

Individual Factor Map – Dimensions 1 and 2 (% variance explained in brackets)



Groups Representation



Appendix S5.12 Paclitaxel Multiple Factor Analysis Results by Gene

		Dim 1					Dim 2				Groups Representation - Dim 1				Groups Representation - Dim 2				GI50 Relationship ^			SMV*
Gene Name	***% VE***	**GI50**	**CN**	**GE**	**Mut**	***% VE***	**GI50**	**CN**	**GE**	**Mut**	**GI50**	**CN**	**GE**	**Mut**	**GI50**	**CN**	**GE**	**Mut**	**CN**	**GE**	**Mut**	
ABCB1	*31%*	0.77	0.08	0.17	0.78	*28%*	0.17	0.75	0.71	0.06	0.60	0.01	0.03	0.61	0.03	0.56	0.51	0.00			Str Rel (-)	
ABCB11	*41%*	0.54	0.61	0.76	NA	*32%*	0.75	0.63	0.02	NA	0.29	0.37	0.58		0.57	0.40	0.00		Pos (+)	Rel (+)		
ABCC1	*34%*	0.22	0.82	0.80	0.08	*27%*	0.72	0.02	0.10	0.76	0.05	0.67	0.64	0.01	0.51	0.00	0.01	0.57			Pos (+)	
ABCC10	*43%*	0.28	0.80	0.91	0.39	*26%*	0.38	0.48	0.04	0.81	0.08	0.65	0.83	0.15	0.15	0.23	0.00	0.65				21.28
BAD	*55%*	0.23	0.90	0.89	NA	*33%*	0.97	0.09	0.16	NA	0.05	0.81	0.80		0.95	0.01	0.02					
BBC3	*39%*	0.77	0.00	0.77	NA	*35%*	0.28	0.94	0.28	NA	0.59	0.00	0.59		0.08	0.88	0.08			Str Rel (-)		
BCAP29	*34%*	0.40	0.16	0.82	0.72	*26%*	0.31	0.93	0.19	0.24	0.16	0.02	0.66	0.52	0.10	0.86	0.04	0.06				
BCL2	*39%*	0.36	0.63	0.80	NA	*35%*	0.83	0.58	0.08	NA	0.13	0.39	0.63		0.70	0.33	0.01		Rel (+)			*25.53*
BCL2L1	*38%*	0.67	0.70	0.74	0.15	*32%*	0.38	0.52	0.32	0.88	0.45	0.49	0.55	0.02	0.15	0.27	0.10	0.77	Rel (-)	Str Rel (-)		36.17
BIRC5	*35%*	0.14	0.77	0.80	0.42	*28%*	0.89	0.00	0.39	0.43	0.02	0.59	0.63	0.18	0.79	0.00	0.15	0.19				27.66
BMF	*46%*	0.77	0.28	0.83	NA	*34%*	0.38	0.93	0.04	NA	0.59	0.08	0.70		0.14	0.87	0.00			Str Rel (-)		25.53
CNGA3	*43%*	0.72	0.75	0.46	NA	*31%*	0.35	0.21	0.88	NA	0.52	0.56	0.21		0.12	0.04	0.78		Str Rel (-)			
CYP2C8	*32%*	0.74	0.76	0.17	0.33	*25%*	0.30	0.14	0.95	0.13	0.55	0.57	0.03	0.11	0.09	0.02	0.89	0.02	Str Rel (+)			
CYP3A4	*29%*	0.76	0.11	0.68	0.34	*27%*	0.17	0.69	0.39	0.63	0.58	0.01	0.46	0.11	0.03	0.48	0.15	0.40		Str Rel (+)		
FGF2	*36%*	0.75	0.29	0.57	0.67	*26%*	0.18	0.89	0.17	0.44	0.56	0.08	0.33	0.45	0.03	0.79	0.03	0.19		Rel (+)	Rel (-)	27.66
FN1	*32%*	0.75	0.53	0.62	0.25	*27%*	0.25	0.30	0.39	0.87	0.56	0.28	0.39	0.06	0.06	0.09	0.15	0.76	Rel (+)	Rel (+)		29.79
GBP1	*52%*	0.54	0.75	0.84	NA	*30%*	0.81	0.49	0.09	NA	0.29	0.56	0.71		0.66	0.24	0.01		Rel	Rel (+)		
MAP2	*31%*	0.43	0.32	0.65	0.72	*28%*	0.55	0.73	0.48	0.22	0.19	0.10	0.43	0.52	0.31	0.54	0.23	0.05	Pos (+)			
MAP4	*43%*	0.38	0.74	0.81	0.60	*27%*	0.77	0.44	0.29	0.45	0.14	0.55	0.66	0.36	0.59	0.19	0.08	0.20				25.53
MAPT	*33%*	0.81	0.11	0.82	0.10	*27%*	0.32	0.45	0.27	0.83	0.65	0.01	0.67	0.01	0.10	0.20	0.07	0.69		Str Rel (-)		34.04
NFKB2	*31%*	0.68	0.72	0.52	0.06	*29%*	0.40	0.06	0.54	0.83	0.46	0.52	0.27	0.00	0.16	0.00	0.29	0.69	Rel (+)	Pos (+)		23.4
NR1I2	*38%*	0.74	0.74	0.21	NA	*33%*	0.12	0.16	0.98	NA	0.55	0.54	0.04		0.01	0.03	0.96		Str Rel (+)			
OPRK1	*33%*	0.57	0.82	0.57	0.16	*26%*	0.52	0.07	0.60	0.65	0.32	0.67	0.32	0.03	0.27	0.00	0.36	0.42	Pos	Pos(-)	Rel (+)	
SLCO1B3	*35%*	0.55	0.49	0.74	0.56	*30%*	0.68	0.70	0.26	0.39	0.30	0.24	0.55	0.31	0.47	0.50	0.07	0.16	Rel (-)	Rel (+)		23.4
TLR6	*32%*	0.69	0.44	0.70	0.35	*27%*	0.38	0.44	0.47	0.73	0.48	0.19	0.49	0.12	0.14	0.19	0.22	0.53	Pos	Str Rel (+)		25.53
TMEM243	*50%*	0.04	0.87	0.87	NA	*33%*	1.00	0.06	0.01	NA	0.00	0.75	0.76		1.00	0.00	0.00					*21.28*
TUBB1	*39%*	0.66	0.49	0.71	NA	*32%*	0.45	0.85	0.17	NA	0.43	0.24	0.50		0.20	0.73	0.03			Str Rel (+)		
TUBB4A	*43%*	0.30	0.78	0.76	NA	*33%*	0.95	0.11	0.27	NA	0.09	0.61	0.58		0.90	0.01	0.07					
TUBB4B	*36%*	0.75	0.33	0.64	NA	*34%*	0.05	0.87	0.50	NA	0.56	0.11	0.41		0.00	0.75	0.25		Pos	Str Rel (-)		
TWIST1	*38%*	0.75	0.53	0.53	NA	*33%*	0.00	0.71	0.70	NA	0.57	0.28	0.29		0.00	0.50	0.50		Rel (+)	Rel (+)		21.28
CSAG2																						29.79

Negative (-) = Higher predictive variable is associated with lower GI50 value (ie. Increased resistance), Positive = Higher predictive variable is associated with higher GI50 value (ie. Decreased resistance), CN = copy number, GE = gene expression, Mut = somatic

mutations, % VE = % variance explained, Dim = dimension, SVM = support vector machine, * = percent of misclassification if the gene is removed from the SVM, Str Rel = strong relationship to GI50, Rel = relationship to GI50, Pos = possibly a relationship to GI50. ^ blank boxes indicate no relationship.

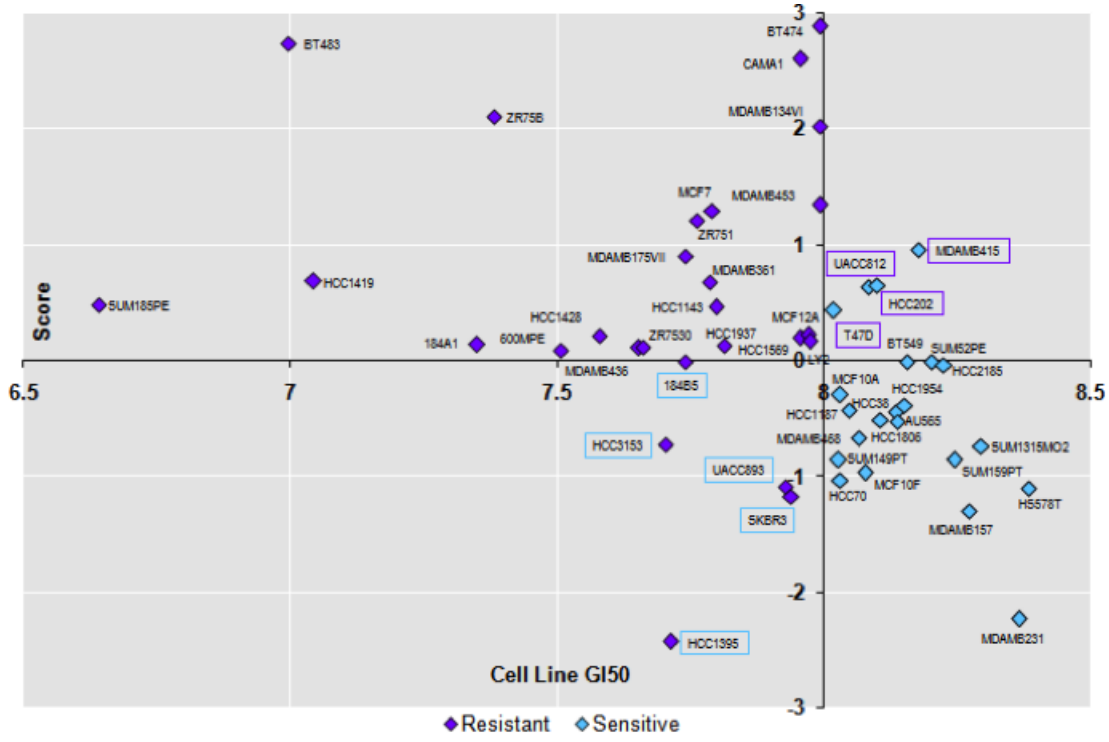
Appendix S5.13 Gemcitabine Multiple Factor Analysis Results by Gene

	Dim 1					Dim 2					Dim 1				Dim 2				GI50 Relationship^			SVM*
Gene Name	***% VE***	**GI50**	**CN**	**GE**	**Mut**	***% VE***	**GI50**	**CN**	**GE**	**Mut**	**GI50**	**CN**	**GE**	**Mut**	**GI50**	**CN**	**GE**	**Mut**	**CN**	**GE**	**Mut**	
ABCB1	0.32	0.76	0.31	0.77	0.15	0.26	0.26	0.59	0.17	0.79	0.58	0.10	0.60	0.02	0.07	0.35	0.03	0.62		**Str Rel (-)**		20.45 (Exp)
ABCC10	0.34	0.31	0.20	0.83	0.71	0.27	0.77	0.67	0.06	0.22	0.10	0.04	0.70	0.51	0.60	0.45	0.00	0.05	**Str Rel (-)**			31.82 (Exp), 25 (CN)
AK1	0.36	0.44	0.35	0.85	0.63	0.26	0.45	0.65	0.03	0.65	0.19	0.12	0.72	0.39	0.20	0.43	0.00	0.42	**Pos (-)**	**Pos (-)**		
CDA	0.42	0.77	0.79	0.16	NA	0.33	0.22	0.01	0.98	NA	0.60	0.63	0.03		0.05	0.00	0.95		**Str Rel (-)**			
CMPK1	0.37	0.55	0.63	0.82	0.29	0.26	0.10	0.42	0.07	0.91	0.31	0.40	0.67	0.08	0.01	0.18	0.00	0.83	**Rel (-)**	**Rel (-)**		18.18 (Exp)
CTPS1	0.35	0.52	0.77	0.70	0.18	0.27	0.51	0.33	0.22	0.83	0.27	0.59	0.49	0.03	0.26	0.11	0.05	0.68	**Pos**	**Pos (-)**		
DCK	0.31	0.83	0.57	0.43	0.15	0.28	0.04	0.31	0.73	0.71	0.69	0.32	0.18	0.02	0.00	0.09	0.53	0.51	**Rel (+)**			
DCTD	0.40	0.81	0.51	0.83	0.11	0.26	0.21	0.18	0.19	0.96	0.66	0.26	0.68	0.01	0.04	0.03	0.03	0.93	**Rel (+)**	**Str Rel (+)**		25 (Exp)
NME1	0.53	0.43	0.85	0.82	NA	0.31	0.90	0.15	0.31	NA	0.19	0.73	0.68		0.81	0.02	0.10		**Pos (-)**	**Pos (-)**		31.82 (Exp)
NT5C	0.55	0.46	0.84	0.87	NA	0.30	0.89	0.30	0.17	NA	0.21	0.70	0.75		0.79	0.09	0.03					34.09 (CN)
RRM1	0.45	0.5	0.82	0.9	0.24	0.25	0.19	0.31	0.07	0.92	0.25	0.67	0.81	0.06	0.03	0.1	0.01	0.85				20.45 (Exp)
RRM2	0.35	0.49	0.66	0.78	0.32	0.27	0.51	0.52	0.18	0.73	0.24	0.44	0.62	0.10	0.26	0.27	0.03	0.53	**Pos (-)**	**Pos (-)**		
RRM2B	0.43	0.37	0.86	0.82	0.41	0.25	0.64	0.12	0.19	0.72	0.13	0.74	0.67	0.16	0.41	0.01	0.04	0.52			**Rel (-)**	29.55 (Exp)
SLC28A1	0.37	0.36	0.73	0.48	0.76	0.27	0.75	0.25	0.63	0.20	0.13	0.53	0.23	0.58	0.56	0.06	0.40	0.04		**Str Rel (-)**		
SLC28A3	0.31	0.76	0.16	0.31	0.73	0.26	0.21	0.75	0.62	0.21	0.58	0.03	0.10	0.54	0.04	0.56	0.39	0.04			**Str Rel (+)**	
SLC29A1	0.54	0.02	0.90	0.90	NA	0.33	1.00	0.05	0.07	NA	0.00	0.81	0.81		1.00	0.00	0.00					
SLC29A2	0.46	0.34	0.87	0.76	0.62	0.25	0.84	0.08	0.14	0.53	0.12	0.76	0.57	0.38	0.71	0.01	0.02	0.28			**Pos (-)**	
TYMS	0.50	0.16	0.87	0.85	NA	0.33	0.98	0.00	0.18	NA	0.02	0.75	0.73		0.97	0.00	0.03					25 (CN)

Abbreviations for table listed in Appendix S5.12

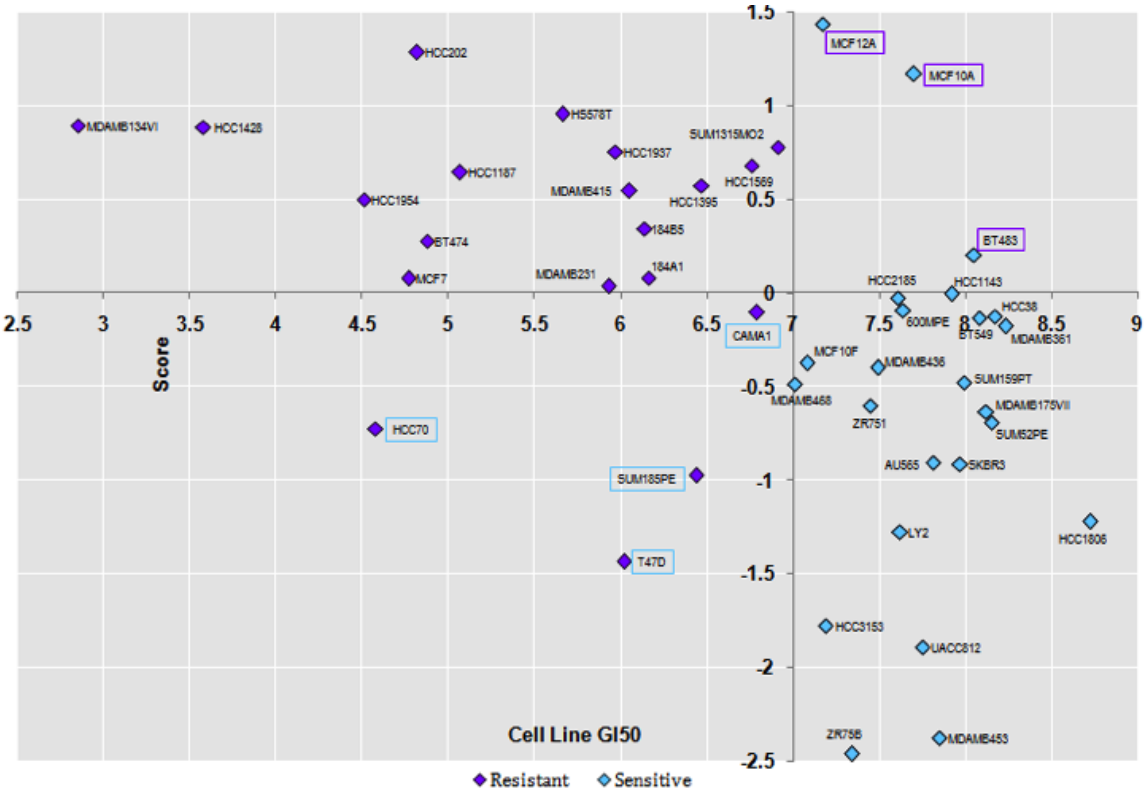
Appendix S5.14 Cell Line GI50 vs. SVM Classification Score

Appendix S5.14.1 Paclitaxel SVM



Support vector machine classification score plotted against the GI_{50} of the cell line for paclitaxel. The vertical axis crosses the horizontal axis at the median GI_{50} of all cell lines analyzed. Cell lines with scores >0 were classified as resistant, those with scores <0 are classified as sensitive. Cell lines outlined in a blue box are those classified as resistant, but are actually sensitive to the drug (false positives); cell lines outlined in purple box were misclassified as sensitive (false negatives).

Appendix S5.14.2 Gemcitabine SVM



Support vector machine classification score plotted against the GI_{50} of the cell line for gemcitabine. The vertical axis crosses the horizontal axis at the median GI_{50} of all cell lines analyzed. Cell lines with scores >0 were classified as resistant, those with scores <0 are classified as sensitive. Cell lines outlined in a blue box are those classified as resistant, but are actually sensitive to the drug (false positives); cell lines outlined in purple box were misclassified as sensitive (false negatives)

Appendix S5.15 Single Gene paclitaxel and gemcitabine SVMs using cell line data

	Paclitaxel	
SVM single variable	Percent misclassified	Hinge loss
Subtype	30.6%	0.69
ABCB1	44.9%	0.90
ABCB11	44.9%	0.90
ABCC1	44.9%	0.90
ABCC10	**44.9%**	**0.90**
BAD	46.9%	0.95
BBC3	34.7%	0.87
BCAP29	44.9%	0.90
BCL2	**44.9%**	**0.90**
BCL2L1	**44.9%**	**0.92**
BIRC5	**44.9%**	**0.90**
BMF	**42.9%**	**0.86**
CNGA3	44.9%	0.90
CSAG2	**36.7%**	**0.80**
CYP2C8	44.9%	0.90
CYP3A4	44.9%	0.90
FGF2	**42.9%**	**0.91**
FN1	**44.9%**	**0.85**
GBP1	36.7%	0.77
MAP2	40.8%	0.86
MAP4	**44.9%**	**0.90**
MAPT	**34.7%**	**0.81**
NFKB2	**32.7%**	**0.85**
NR1I2	44.9%	0.90
OPRK1	44.9%	0.90
SLCO1B3	**34.7%**	**0.78**
TLR6	**38.8%**	**0.81**
TMEM243	**44.9%**	**0.90**
TUBB1	44.9%	0.90
TUBB4A	46.9%	0.94
TUBB4B	44.9%	0.90
TWIST1	**44.9%**	**0.90**
15 Genes:	18%	0.64

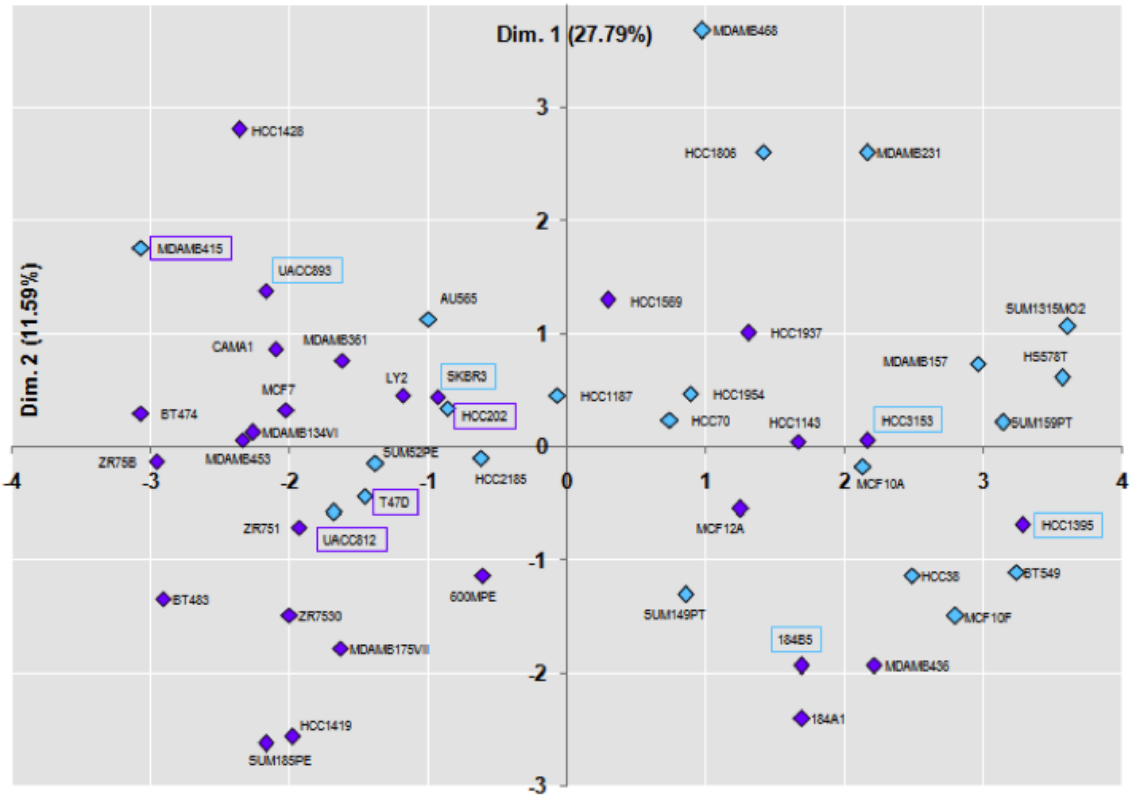
	Gemcitabine	
SVM single variable	Percent misclassified	Hinge loss
Subtype	45.5%	0.95
ABCB1-GE	45.5%	0.94
ABCC10-CN	47.7%	0.95
ABCC10-GE	36.4%	0.90
CMPK1-GE	40.9%	0.87
DCTD-GE	36.4%	0.90
NME1-GE	45.5%	0.91
NT5C-CN	47.7%	1.01
RRM1-GE	38.6%	0.95
RRM2B-GE	50.0%	0.98
TYMS-CN	45.5%	0.91
All Genes:	15%	0.66

Appendix S5.16 Multiple Factor Analysis– Entire and SVM Gene Sets

(for dimensions 1 and 2, % variance explained in brackets)

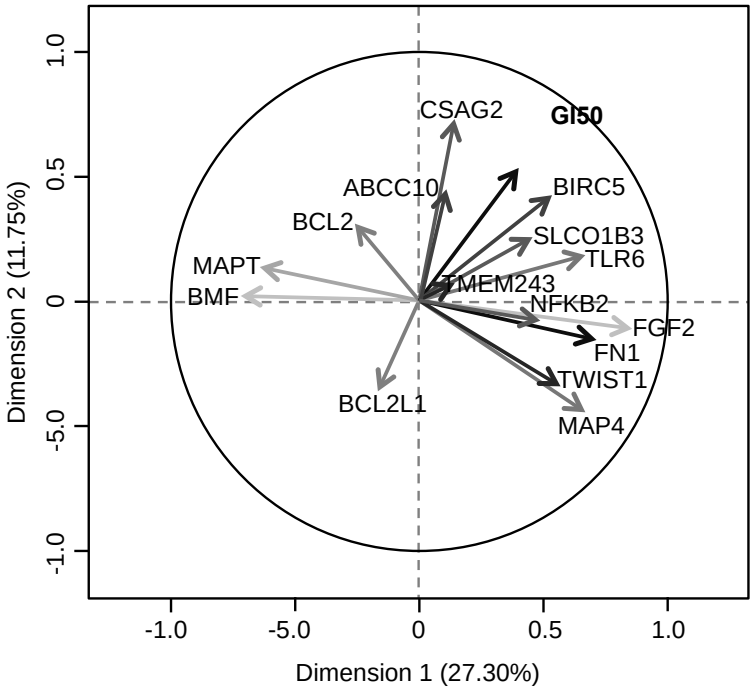
Appendix S5.16.1 Paclitaxel – SVM Gene Set

Individual Factor Map



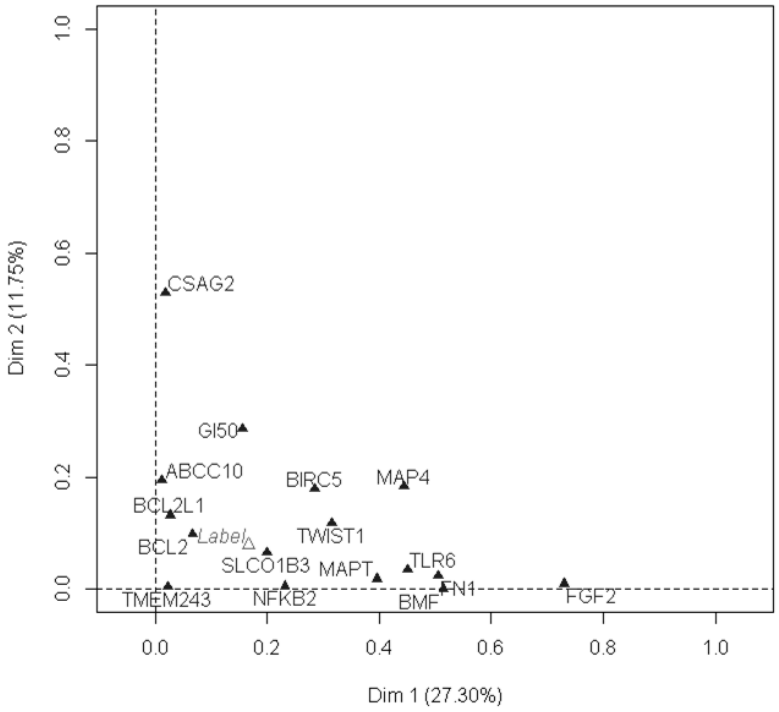
The individual factor maps generated with a multiple factor analysis using the gene set derived from the respective SVMs are displayed for paclitaxel. Purple points are resistant cell lines, blue points are sensitive cell lines. Cell lines outlined in a blue box are those classified as resistant, but are actually sensitive to the drug (false positives); cell lines outlined in purple box were misclassified as sensitive (false negatives). 9 of 49 cell lines were misclassified for paclitaxel.

Correlation Circle



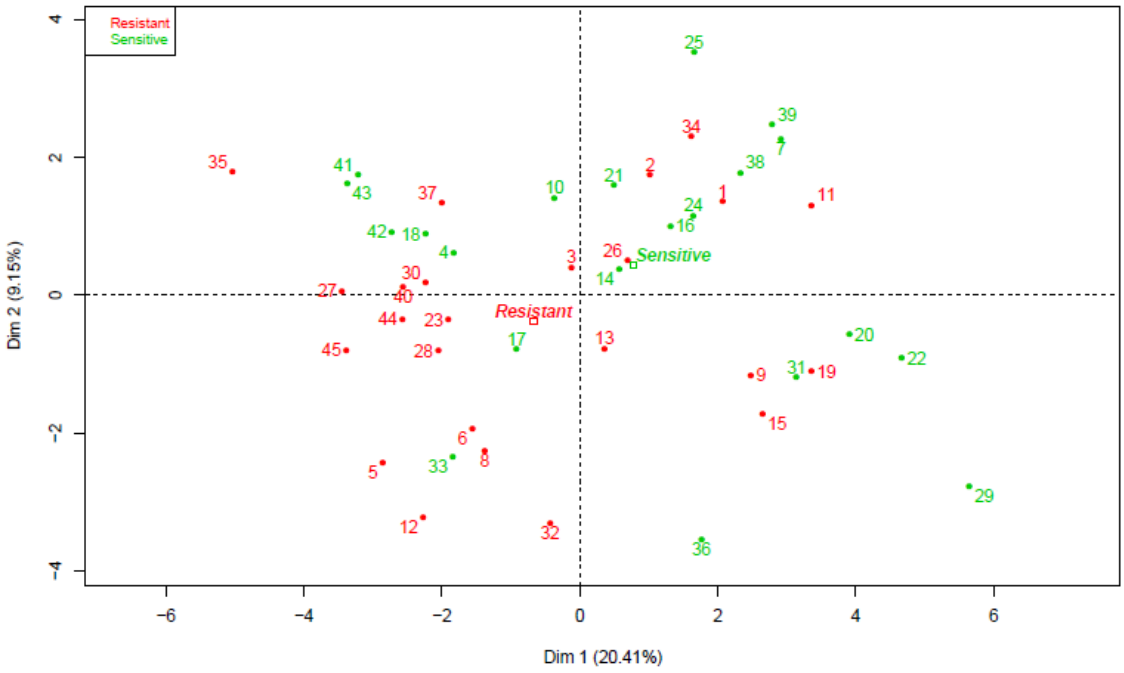
Factors (genes) used in the analysis are labeled and displayed as shaded grey arrows (correlation circle) or black triangles (groups representation) across dimensions 1 and 2. The percent variance explained by each dimension is indicated in brackets on all figures.

Groups Representation



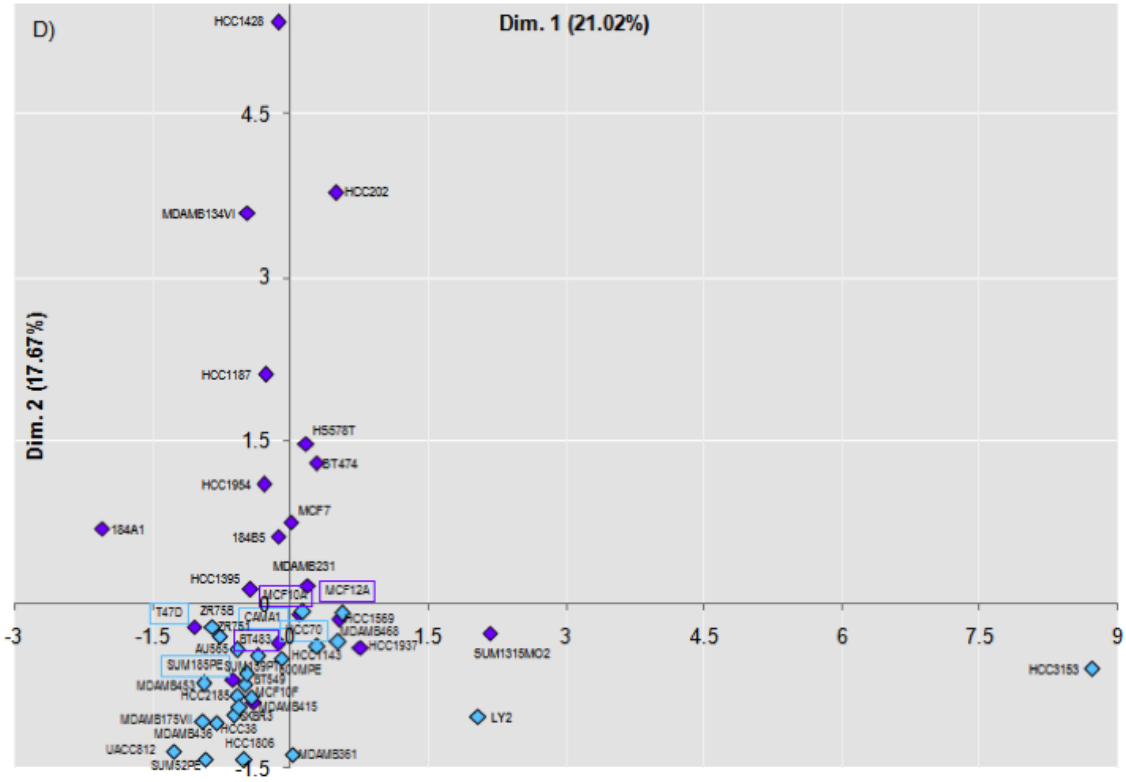
Appendix S5.16.2 Paclitaxel – Entire Gene Set

Individual Factor Map



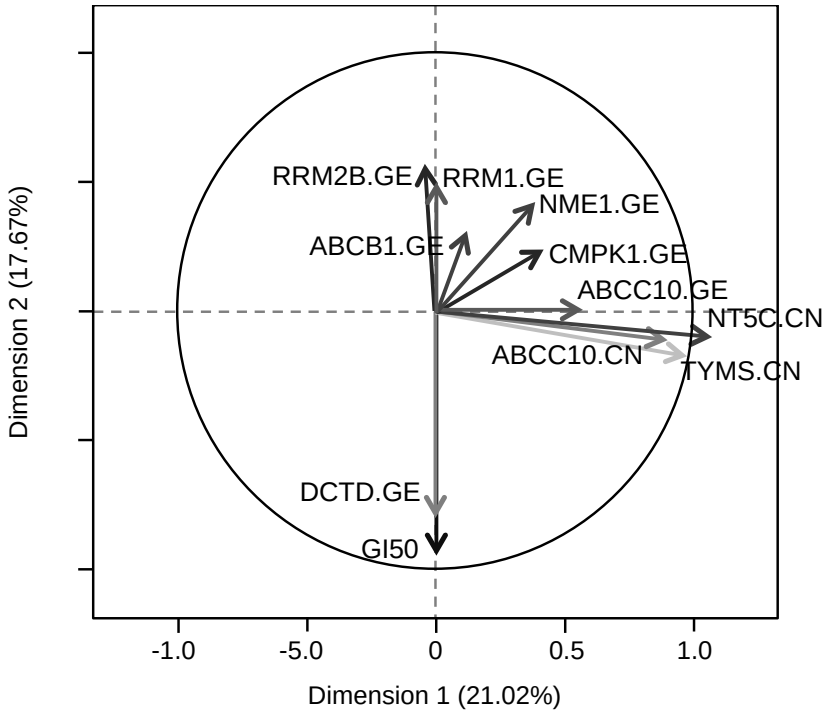
Appendix S5.16.3 Gemcitabine – SVM Gene Set

Individual Factor Map

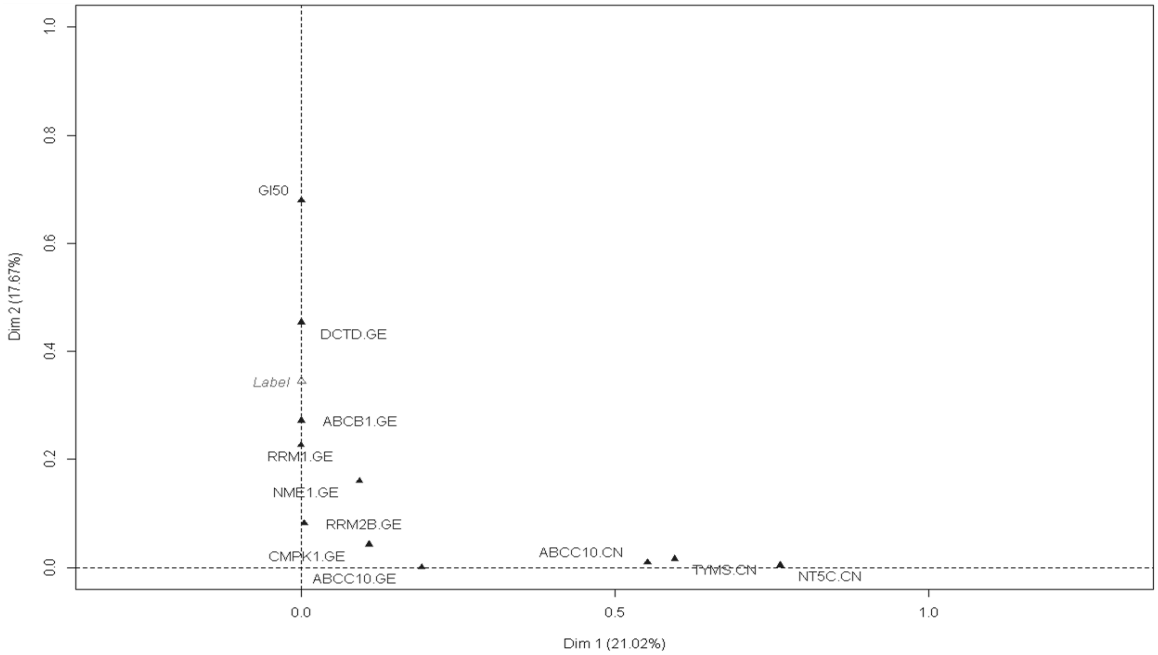


The individual factor maps generated with a multiple factor analysis using the gene set derived from the SVM are displayed for gemcitabine. Purple points are resistant cell lines, blue points are sensitive cell lines. Cell lines outlined in a blue box are those classified as resistant, but are actually sensitive to the drug (false positives); cell lines outlined in purple box were misclassified as sensitive (false negatives). 7 of 44 cell lines were misclassified for gemcitabine.

Correlation Circle

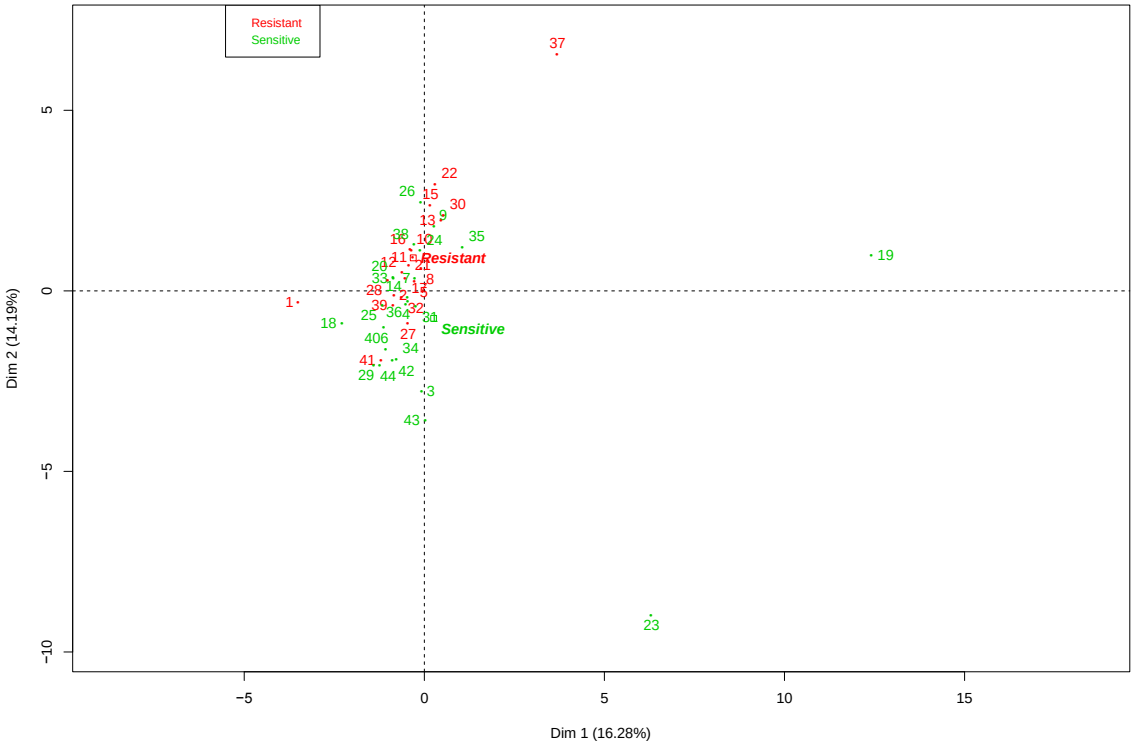


Groups Representation



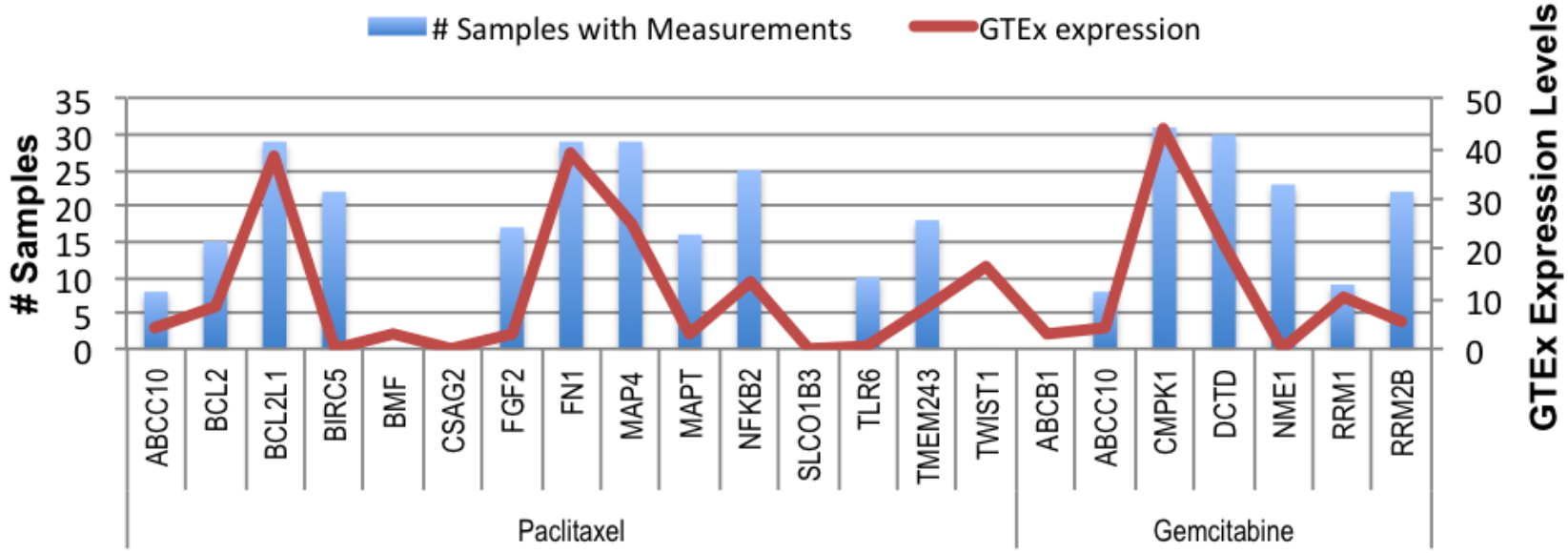
Appendix S5.16.4 Gemcitabine– Entire Gene Set

Individual Factor Map

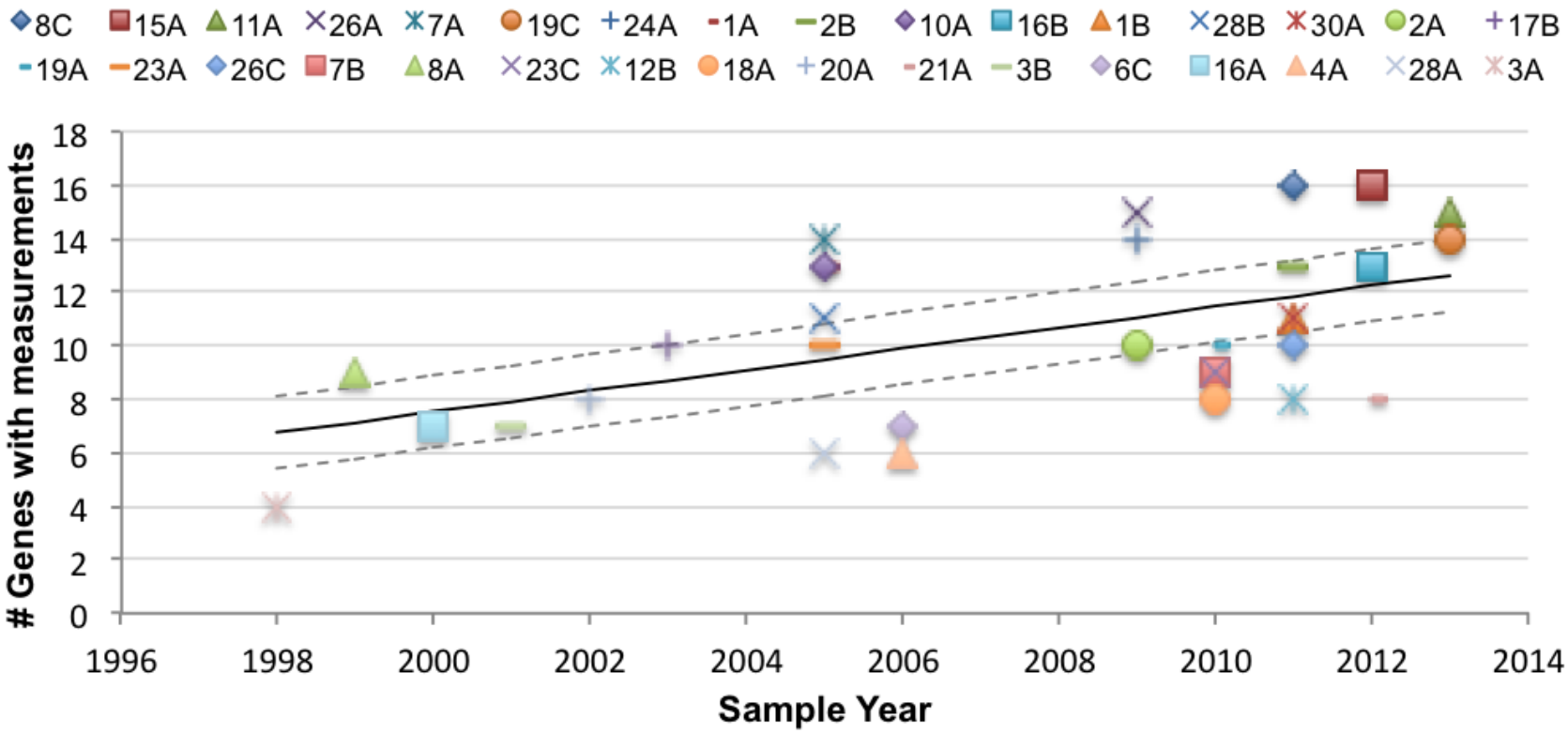


Appendix S5.17 FFPE Samples – Gene expression measurements summary

Appendix S5.17.1 Number of measurements by gene compared to GTEx expression levels



Appendix S5.17.2 Year of tissue block compared to number of measurements per sample



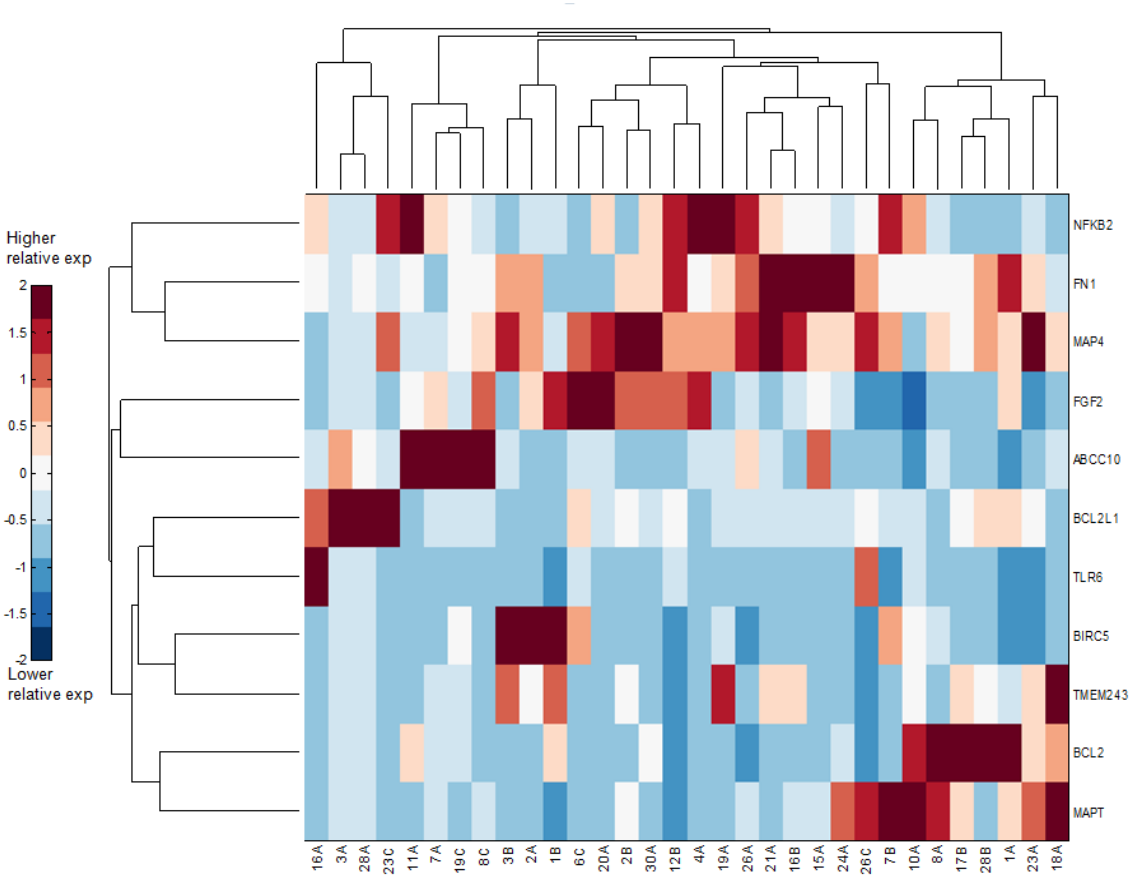
Appendix S5.18 Patient Clustering Supplementary Results.

Clustering was performed as in main Methods. Each cluster derived from the MD Anderson Patient Data was isolated and the tumours in each were summarized by subtype, number of distance recurrences ("events"), and mean time to distant recurrence (Tables S5.1-S5.3 – see below).

The 'grey' clusters were isolated and further clustered with similar stratification by gene expression and outcome (Supplementary Figure VI. 1).

Appendix S5.18.1 FFPE Patient Samples

Figure VI.1 – Paclitaxel FFPE Clustering Results



Expression heatmap of the paclitaxel SVM derived genes for our set of 32 FFPE samples, as measured by qPCR. Each row represents a gene and each column a tumour. Red indicates higher expression and blue represents lower expression, as shown by the colour bar on the left. Clustering was done based on the similarity of each tumour's and gene's expression profile. The dendrograms on the top and left indicate the relatedness of each tumour and gene by the length and subdivision of the branches, with deeper branches indicating a stronger relationship and branches in the same 'tree' being more closely related to each other than data in other 'trees'.

Figure VI.2 – Gemcitabine FFPE Clustering Results

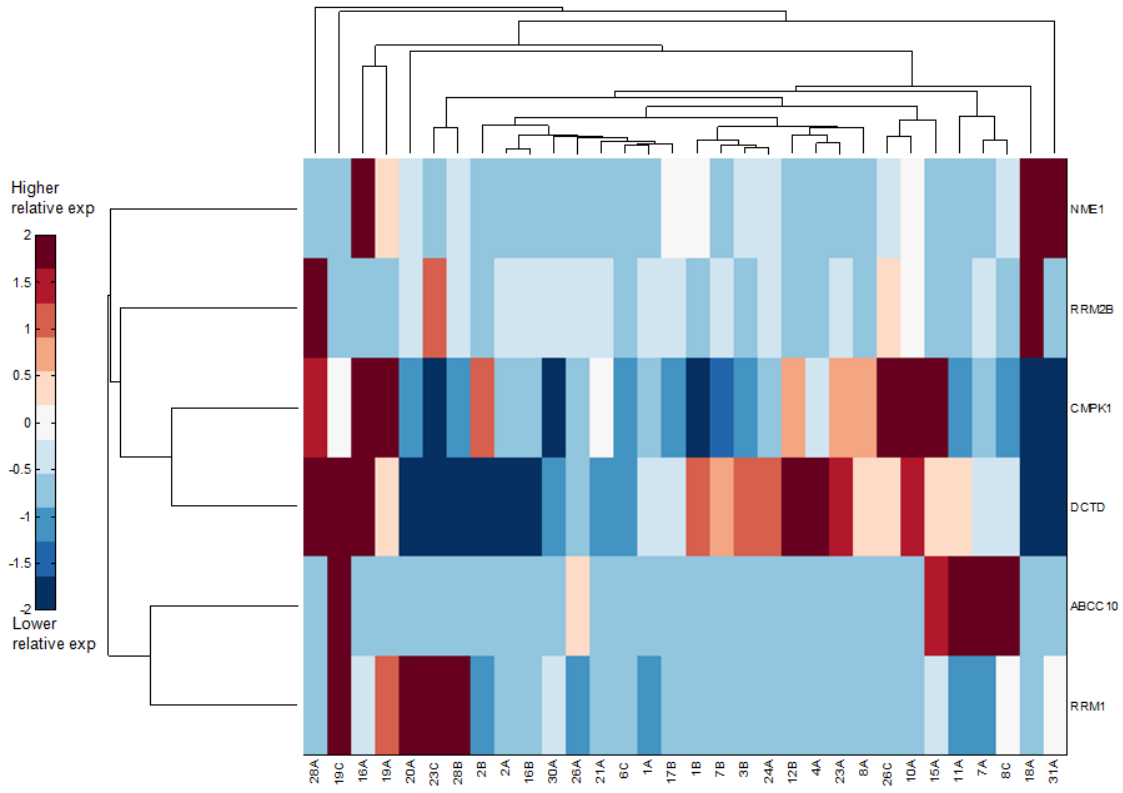
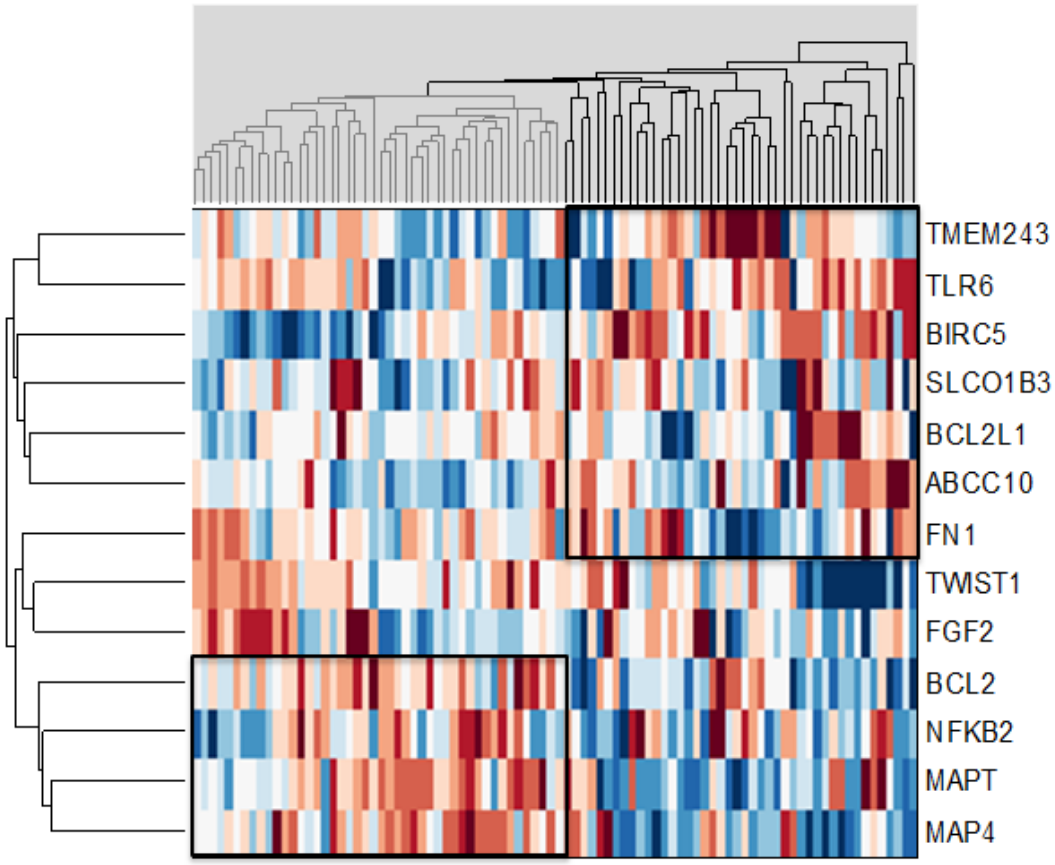


Figure legend as above (Figure VI. 3).

Note: sample 3A had an extremely high expression value for DCTD and distorted the row view for that gene. It has been removed in this figure for ease of visualization.

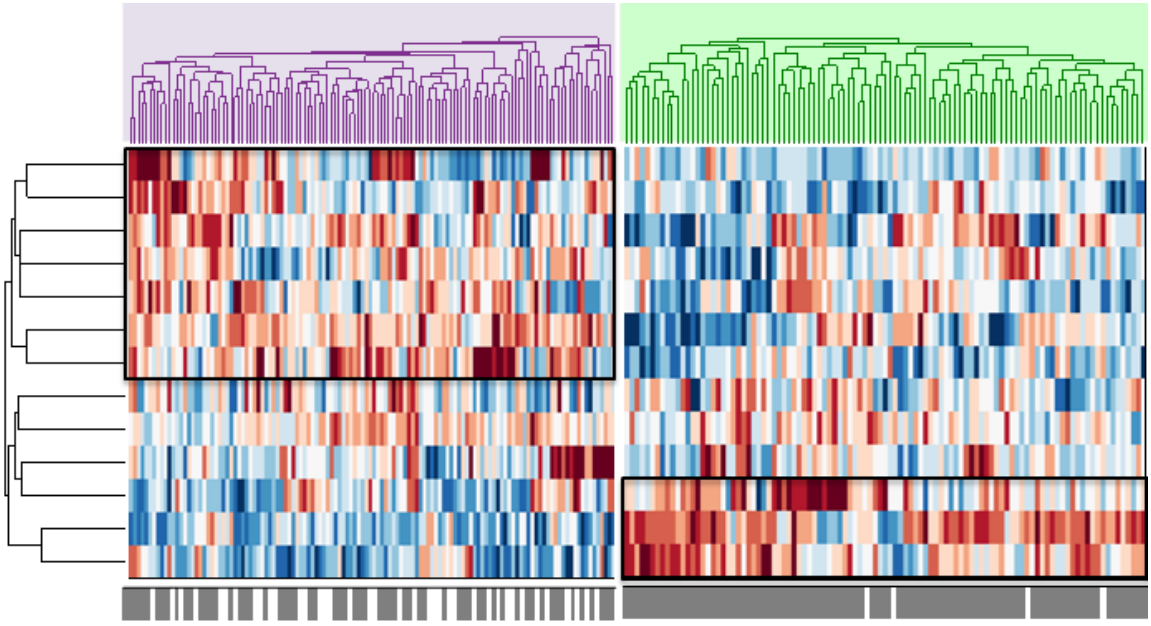
Appendix S5.18.2 Hatzis et al. Patient Data

Supplementary Figure VI.3 - Further Clustering of the 'Grey' Cluster



The 'grey' cluster from the previous clustering analysis was isolated and clustered further. The leftmost cluster (shaded a lighter grey) is composed of 70% luminal tumours with a mean survival time of 3.14 years. The rightmost cluster is composed of 43% basal tumours with a mean survival time of 2.45 years. The leftmost cluster also contains only 3 distant recurrences, with two of those being classified by the MD Anderson signature as "Sensitive". The 'light grey' cluster, meanwhile, is stratified very well on the basis of the MD Anderson signature (results not shown). This mirrors the results of the clustering analysis on the 'green' and 'purple' tumour clusters.

Supplementary Figure VI.4 - Zoom on the 'purple' and 'green' clusters.



The clusters from Figure 11 in the main paper were isolated from the main heatmap for easier visualization of the differential gene expression that distinguishes each cluster. Figure legend as in the main paper.

Table S5.1: Summary of tumours contained in each cluster.

"Green" Cluster								
RD/pCR	99	4			96%	4%		
Insensitive/ Sensitive	68	35			66%	34%		
Basal/LumA/ LumB/Normal	8	65	21	3	8%	63%	20%	3%
Num events:	15							
Average time:	3.15 years							
Total observations:	103							

Table S5.2: Summary of tumours contained in each cluster.

"Purple" Cluster								
RD/pCR	85	41			67%	33%		
Insensitive/ Sensitive	91	35			72%	28%		
Basal/LumA/ LumB/Normal	79	9	12	7	63%	7%	10%	6%
Num events:	41							
Average time:	2.64 years							
Total observations:	126							

Table S5.3: Summary of tumours contained in each cluster.

Remaining ("Grey") Clusters								
RD/pCR	73	17			81%	19%		
Insensitive/ Sensitive	55	35			61%	39%		
Basal/LumA/ LumB/Normal	20	29	19	17	22%	32%	21%	19%
Num events:	14							
Average time:	2.80 years							
Total observations:	90							

RD: recurrent disease **pCR**: pathological complete response

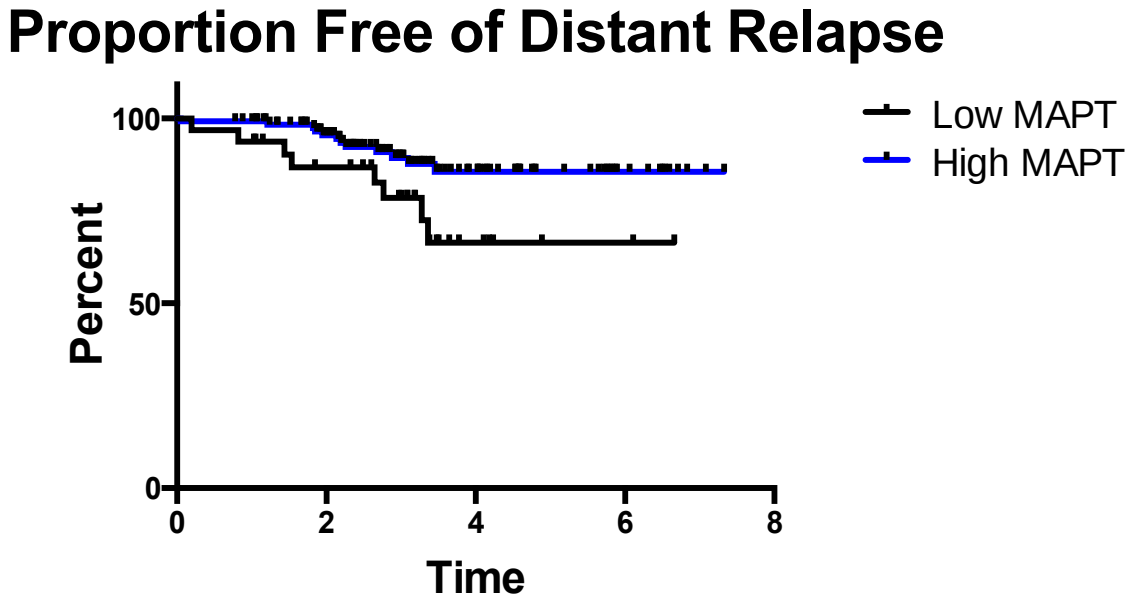
Insensitive/Sensitive as predicted by Hatzis et. al. (2011)

Events: distant relapse **Time**: time to distant relapse

Appendix S5.19 *MAPT* Expression Affects Prognosis in Luminal Tumours.

MAPT is part of the PAM50 and clearly segregates the data into luminal and basal subtype to a large extent. However, some luminal tumours express *MAPT* at a lower level than the majority. Low *MAPT* expressing luminal subtypes fall into the low *MAPT* expressing 'purple' cluster (Supplementary Figure VI. 4) and have a significantly worse prognosis than higher *MAPT* expressing luminal tumours in the patient dataset (Supplementary Figure VII. 1).

Supplementary Figure VII. 1 - Kaplan-Meier curves for low MAPT expressing luminal tumours vs. higher MAPT expressing luminal tumours.

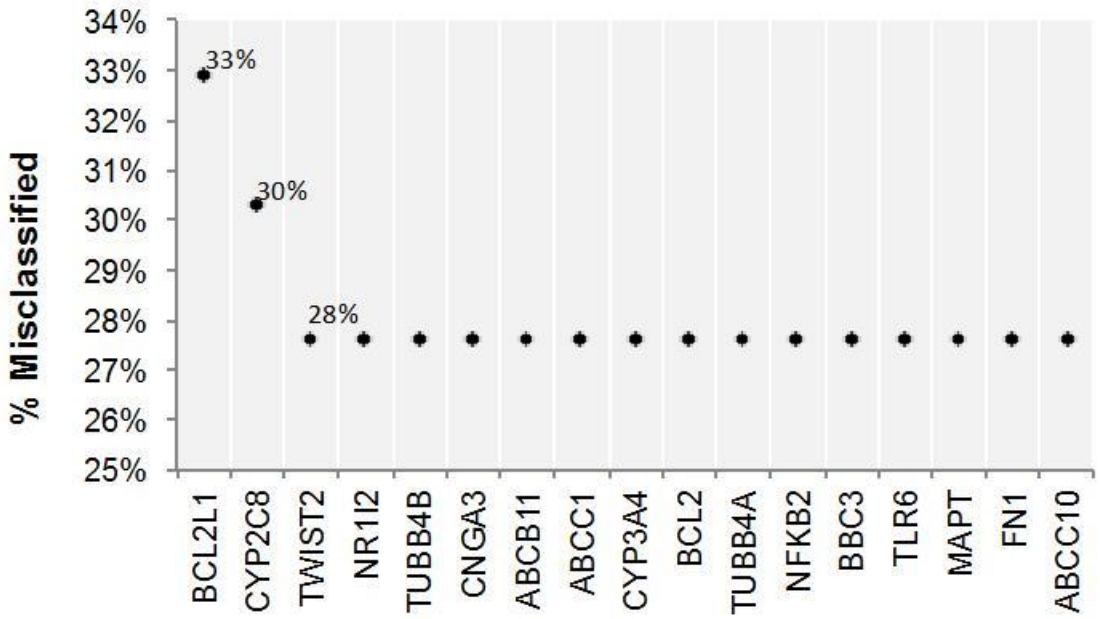


'Low' vs. 'high' expression was stratified by median *MAPT* expression across all tumours, regardless of subtype. Luminal tumours with expression values below the overall median were classified as 'low *MAPT*' and those with values above were classified as 'high *MAPT*'. There were 32 low *MAPT* expressing' luminal tumours in the low *MAPT* set and 123 high *MAPT* expressing luminal tumours. In the log-rank test, the Kaplan-Meier results are significant (p = 0.037). The log-rank hazard ratio is 2.503 (95% CI of ratio: 1.071 to 9.203).

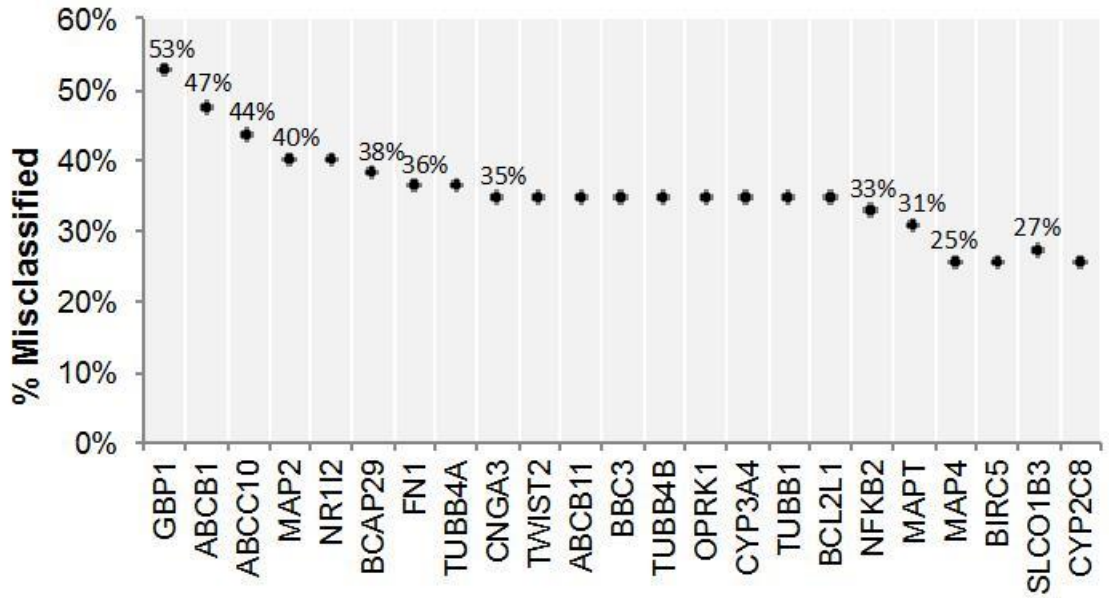
Appendix S5.20 Creating SVM models using lung and hematopoietic cell lines.

We initially investigated the possibility that the paclitaxel breast cancer SVM model could predict cell line sensitivity to this drug in 22 other cancer cell line types. The respective misclassification rates were higher than with the breast cancer cell lines. We attempted to classify resistance with the SVM model in other neoplastic tissues, including from autonomic ganglia (10 cell lines), biliary tract (1), bone (10), central nervous system (27), endometrium (17), hematopoietic and lymphoid tissue (55), kidney (8), large intestine (18), liver (15), lung (76), oesophagus (15), ovary (24), pancreas (25), pleura (7), prostate (3), salivary gland (1), skin (35), soft tissue (11), stomach (14), thyroid (3), upper aerodigestive tract (6), and urinary tract (12). As Daemon et al., 2013 reported, clustering of individual tissue types dominates the analysis of chemosensitivity. The tissue-specific gene expression program of the cell lines could explain why the breast cancer signature was not transferable.

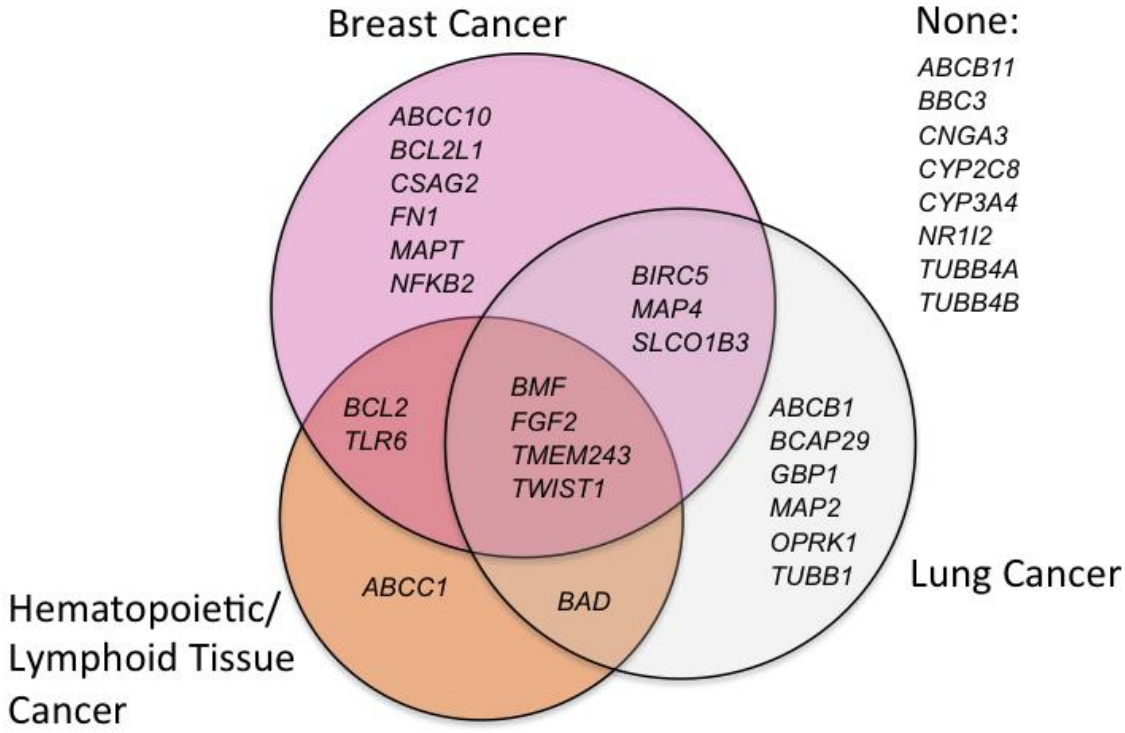
Appendix S5.20.1 Feature Selection Process – Lung Cancer Cell Lines



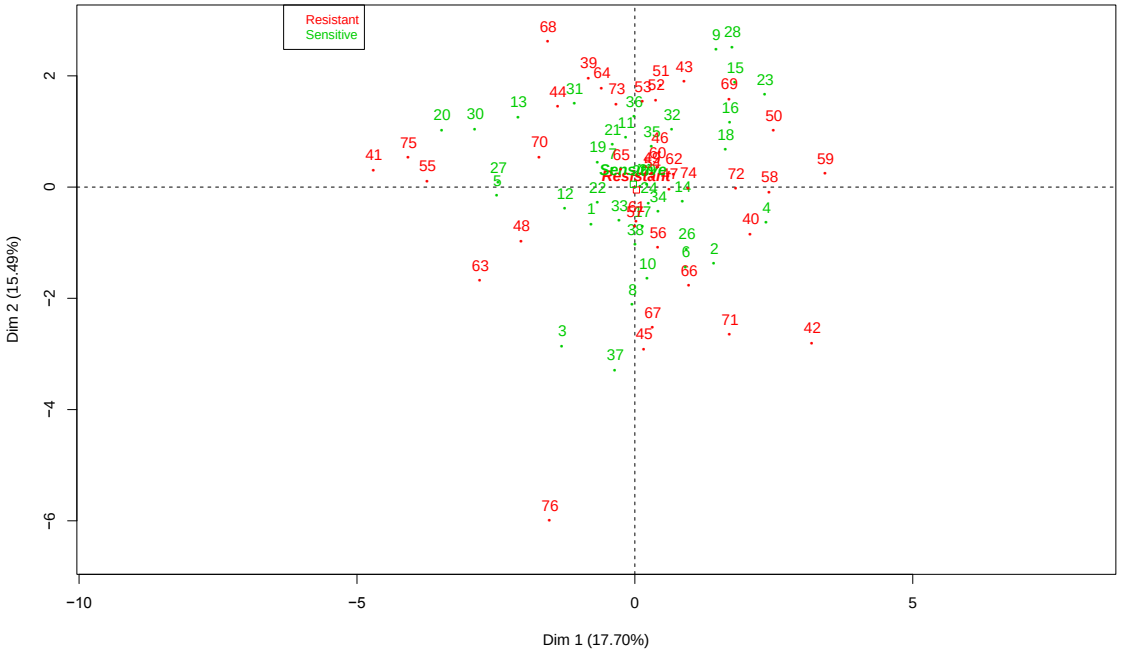
Appendix S5.20.2 Feature Selection Process – Hematopoietic and Lymphoid Tissue Cancer Cell Lines

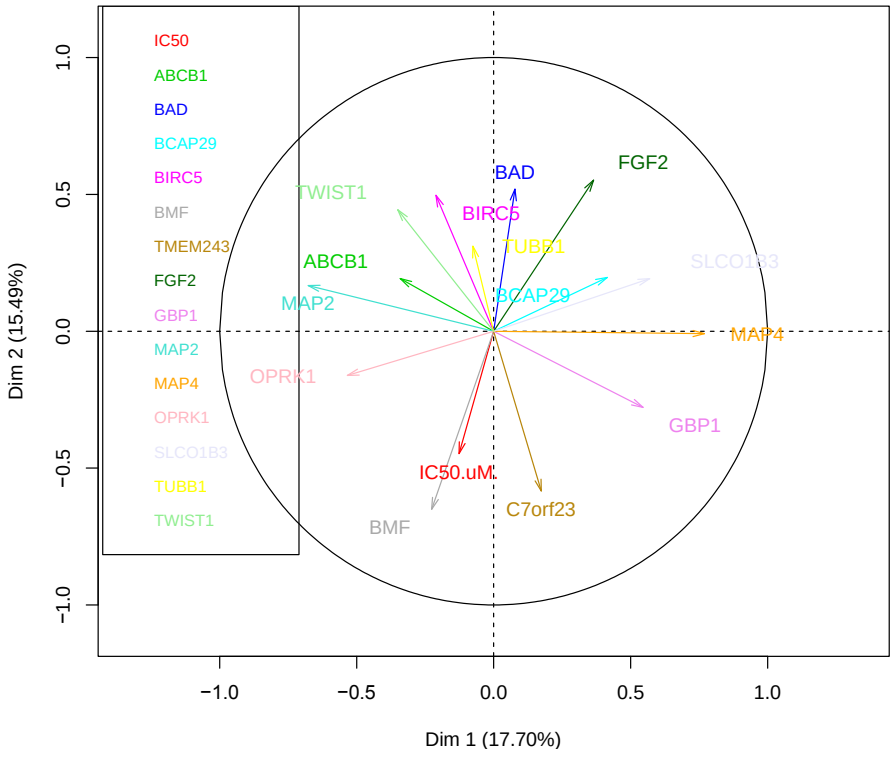


Appendix S5.20.3 Final SMV Gene Sets for Paclitaxel

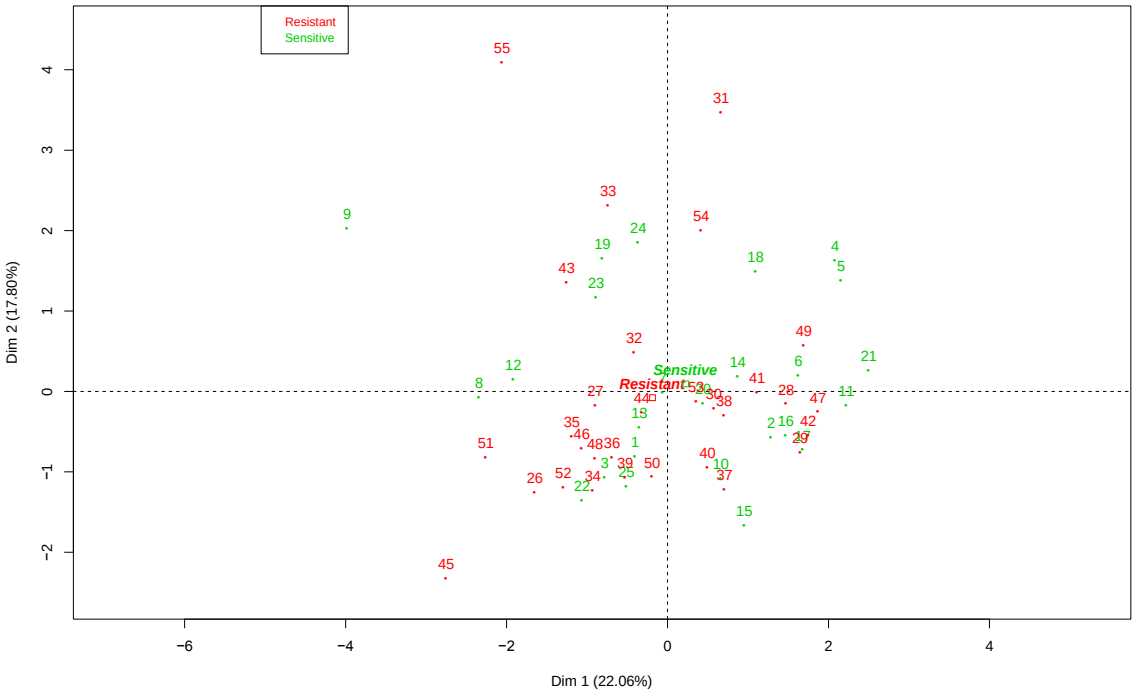


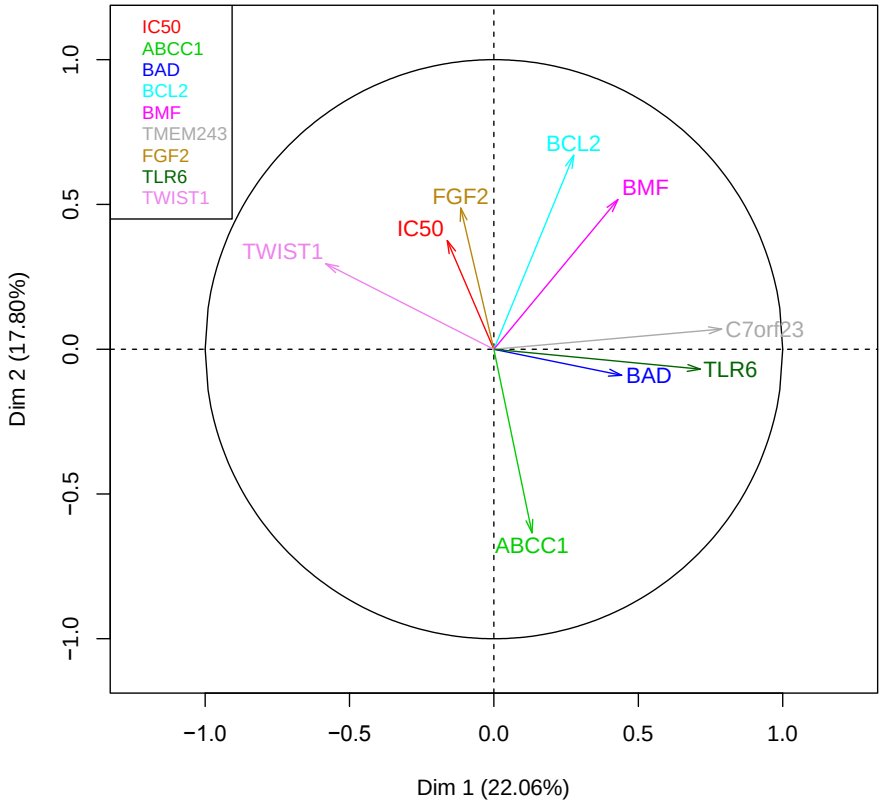
Appendix S5.20.4 MFA Using Genes in SVM – Lung





Appendix S5.20.5 MFA Using Genes in SVM – Hematopoietic and Lymphoid Tissue





Appendix S5.21 References for Appendix S5

1. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17,** 1665–1674 (2007).

2. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

3. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

4. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).

5. Shirley, B. C. *et al.* Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences. *Genomics Proteomics Bioinformatics* **11,** 77–85 (2013).

6. Viner, C., Dorman, S. N., Shirley, B. C. & Rogan, P. K. Validation of predicted mRNA splicing mutations using high-throughput transcriptome data. *F1000Research* **3,** 8 (2014).

7. Skehan, P. *et al.* New colorimetric cytotoxicity assay for anticancer-drug screening. *J. Natl. Cancer Inst.* **82,** 1107–1112 (1990).

8. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27,** 182–189 (2009).

9. Dorman, S. N., Shirley, B. C., Knoll, J. H. M. & Rogan, P. K. Expanding probe repertoire and improving reproducibility in human genomic hybridization. *Nucleic Acids Res.* **41,** e81 (2013).

10. Chou, H.-H., Hsia, A.-P., Mooney, D. L. & Schnable, P. S. Picky: oligo microarray design for large genomes. *Bioinforma. Oxf. Engl.* **20,** 2893–2902 (2004).

11. Dash, M. & Liu, H. Feature Selection for Classification. *Intell. Data Anal.* 131–156 (1997). doi:10.3233/IDA-1997-1302

12. Heikal, N., Nussenzveig, R. H. & Agarwal, A. M. Deparaffinization with mineral oil: a simple procedure for extraction of high-quality DNA from archival formalin-fixed paraffin-embedded samples. *Appl. Immunohistochem. Mol. Morphol. AIMM Off. Publ. Soc. Appl. Immunohistochem.* **22,** 623–626 (2014).

13. Cronin, M. *et al.* Measurement of gene expression in archival paraffin-embedded tissues: development and performance of a 92-gene reverse transcriptase-polymerase chain reaction assay. *Am. J. Pathol.* **164,** 35–42 (2004).

14. Park, N. I., Rogan, P. K., Tarnowski, H. E. & Knoll, J. H. M. Structural and genic characterization of stable genomic regions in breast cancer: relevance to chemotherapy. *Mol. Oncol.* **6,** 347–359 (2012).

15. Yu, J. *et al.* Copy-number analysis of topoisomerase and thymidylate synthase genes in frozen and FFPE DNAs of colorectal cancers. *Pharmacogenomics* **9,** 1459–1466 (2008).

16. Nugoli, M. *et al.* Genetic variability in MCF-7 sublines: Evidence of rapid genomic and RNA expression profile modifications. *BMC Cancer* **3,** (2003).

17. Ring, B. Z., Chang, S., Ring, L. W., Seitz, R. S. & Ross, D. T. Gene expression patterns within cell lines are predictive of chemosensitivity. *BMC Genomics* **9,** 74 (2008).

18. Daemen, A. *et al.* Modeling precision treatment of breast cancer. *Genome Biol.* **14,** R110 (2013).

Curriculum Vitae

Name:	Stephanie N. Dorman
Post-secondary Education and Degrees:	University of Western Ontario London, Ontario, Canada 2005-2010 B.M.Sc.
	University of Western Ontario London, Ontario, Canada 2005-2010 H.B.A.
	University of Western Ontario London, Ontario, Canada 2010-2015 Ph.D.
Honours and Awards:	Dr. Bishnu D. Sanwal Graduate Performance Award 2015
	Ontario Graduate Scholarship 2012-2013, 2013-2014, 2014-2015
	CIHR Strategic Training Program in Cancer Research and Technology Transfer 2011-2012, 2012-2014, 2014-2015
	Translational Breast Cancer Research Studentship 2011-2012, 2012-2014, 2014-2015
	Canadian Cancer Society Research Institute Travel Award 2014
	CIHR Institute Community Support Travel Award 2014
	Global Opportunities Award 2013
	Graduate Thesis Research Fund Award 2011-2012, 2012-2013
Related Work	Teaching Assistant, Biochemistry 2280

Experience The University of Western Ontario
2012-2015

Publications:

Dorman SN, Baranova K, Rogan PK. (2015) A genomic signature for Paclitaxel and Gemcitabine resistance in breast cancer. *Molecular Oncology*. DOI: 10.1016/j.molonc.2015.07.006.

Dorman SN, Viner C, Rogan PK. (2014) Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer. *Scientific Reports*. 4:7063, 1-9.

Viner C, Dorman SN, Shirley BC and Rogan PK. (2014) Validation of predicted mRNA splicing mutations using high-throughput transcriptome data *F1000Research*. 3:8. [v2; status: indexed, http://f1000r.es/378]

Dorman SN, Shirley BC, Knoll JHM, Rogan PK. (2013) Expanding probe repertoire and improving reproducibility in human genomic hybridization. *Nucleic Acids Research*. 41:7, e81.