

# **Development of a simple artificial intelligence method to accurately subtype breast cancers based on gene expression barcodes**

Fanechka N. Esterhuysen



A thesis submitted in partial fulfilment of the requirements for the degree of  
MAGISTER SCIENTIAE (M.Sc.)  
South African National Bioinformatics Institute (SANBI)  
University of the Western Cape

Supervisor: Professor Junaid Gamiieldien

**December 2018**

## **KEYWORDS**

Microarray

RNA-Seq

Gene Expression Barcode

Feature Selection

Machine Learning

Support Vector Machine

Gene Signature

Classification



UNIVERSITY *of the*  
WESTERN CAPE

## ABBREVIATIONS

ANN	Artificial Neural Network
cDNA	Complementary DNA
DNA	Deoxyribonucleic Acid
DNA-Seq	DNA Sequencing
ER-positive	Estrogen Receptor Positive (breast cancer)
fRMA	Frozen Robust Multiarray Analysis
FS	Feature Selection
GC-RMA	GeneChip Robust Multiarray Analysis
GExB	Gene Expression Barcode
GTEx project	Genotype-Tissue Expression project
HER2-positive	Human Epidermal Growth Factor Receptor 2 Positive (breast cancer)
kNN	k-Nearest Neighbours
MC-OVO-SVM	Multiclass One-versus-One Support Vector Machine
MC-OVR-SVM	Multiclass One-versus-Rest Support Vector Machine
ML	Machine Learning
mRNA	Messenger RNA
NGS	Next Generation Sequencing
NN	Neural Network
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCR	Polymerase chain reaction
RMA	Robust Multiarray Analysis
RNA	Ribonucleic Acid
RNA-Seq	RNA Sequencing
rtPCR	Real-Time PCR
SOM	Self-Organizing Maps
SVM	Support Vector Machines
TCGA	The Cancer Genome Atlas
TN	Triple Negative (breast cancer)

## **ABSTRACT**

**INTRODUCTION:** Breast cancer is a highly heterogeneous disease. The complexity of achieving an accurate diagnosis and an effective treatment regimen lies within this heterogeneity. Subtypes of the disease are not simply molecular, i.e. hormone receptor over-expression or absence, but the tumour itself is heterogeneous in terms of tissue of origin, metastases, and histopathological variability. Accurate tumour classification vastly improves treatment decisions, patient outcomes and 5-year survival rates. Gene expression studies aided by transcriptomic technologies such as microarrays and next-generation sequencing (e.g. RNA-Sequencing) have aided oncology researcher and clinician understanding of the complex molecular portraits of malignant breast tumours. Mechanisms governing cancers, which include tumorigenesis, gene fusions, gene over-expression and suppression, cellular process and pathway involvement, have been elucidated through comprehensive analyses of the cancer transcriptome. Over the past 20 years, gene expression signatures, discovered with both microarray and RNA-Seq have reached clinical and commercial application through the development of tests such as Mammaprint®, OncotypeDX®, and FoundationOne® CDx, all which focus on chemotherapy sensitivity, prediction of cancer recurrence, and tumour mutational level.

The Gene Expression Barcode (GExB) algorithm was developed to allow for easy interpretation and integration of microarray data through data normalization with frozen RMA (fRMA) preprocessing and conversion of relative gene expression to a sequence of 1's and 0's. Unfortunately, the algorithm has not yet been developed for RNA-Seq data. However, implementation of the GExB with feature-selection would contribute to a machine-learning based robust breast cancer and subtype classifier.

**METHODOLOGY:** For microarray data, we applied the GExB algorithm to generate barcodes for normal breast and breast tumour samples. A two-class classifier for malignancy was developed through feature-selection on barcoded samples by selecting for genes with 85% stable absence or presence within a tissue type, and differentially stable between tissues. A multi-class feature-selection method was employed to identify genes with variable expression in one subtype, but 80% stable absence or presence in all other subtypes, i.e. 80% in  $n-1$

subtypes.

For RNA-Seq data, a barcoding method needed to be developed which could mimic the GExB algorithm for microarray data. A z-score-to-barcode method was implemented and differential gene expression analysis with selection of the top 100 genes as informative features for classification purposes.

The accuracy and discriminatory capability of both microarray-based gene signatures and the RNA-Seq-based gene signatures was assessed through unsupervised and supervised machine-learning algorithms, i.e., K-means and Hierarchical clustering, as well as binary and multi-class Support Vector Machine (SVM) implementations.

RESULTS: The GExB-FS method for microarray data yielded an 85-probe and 346-probe informative set for two-class and multi-class classifiers, respectively. The two-class classifier predicted samples as either normal or malignant with 100% accuracy and the multi-class classifier predicted molecular subtype with 96.5% accuracy with SVM.

Combining RNA-Seq DE analysis for feature-selection with the z-score-to-barcode method, resulted in a two-class classifier for malignancy, and a multi-class classifier for normal-from-healthy, normal-adjacent-tumour (from cancer patients), and breast tumour samples with 100% accuracy. Most notably, a normal-adjacent-tumour gene expression signature emerged, which differentiated it from normal breast tissues in healthy individuals.

CONCLUSION: A potentially novel method for microarray and RNA-Seq data transformation, feature selection and classifier development was established. The universal application of the microarray signatures and validity of the z-score-to-barcode method was proven with 95% accurate classification of RNA-Seq barcoded samples with a microarray discovered gene expression signature. The results from this comprehensive study into the discovery of robust gene expression signatures holds immense potential for further R&F towards implementation at the clinical endpoint, and translation to simpler and cost-effective laboratory methods such as qPCR-based tests.

## DECLARATION

I declare that *Development of a simple artificial method to accurately subtype breast cancers based on gene expression barcodes* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Full name: Fanechka Naomi Esterhuysen

Date: December 2018

Signature: *Esterhuysen*



# CONTENTS

KEYWORDS .....	i
ABBREVIATIONS .....	ii
ABSTRACT .....	iii
DECLARATION .....	v
CONTENTS .....	vi
List of Figures .....	xi
List of Tables.....	xii
Chapter 1 .....	1
Literature Review.....	1
1.1 Machine Learning.....	1
1.1.1 Different Machine Learning classifiers and algorithms .....	2
1.1.1.1 K-means clustering.....	2
1.1.1.2 Hierarchical Clustering .....	3
1.1.1.3 Neural Networks .....	5
1.1.1.4 Self-Organizing Maps (SOMs) .....	6
1.1.1.5 Support Vector Machines (SVMs).....	6
1.1.1.5.1 Binary SVMs.....	7
1.1.1.5.2 Multi-class SVMs.....	8
1.2 Application of Machine Learning (ML) in Biomedical Scenarios.....	8
1.3 Feature Selection .....	9
1.3.2 Dimension reduction of expression microarray data using feature selection.....	11
1.4 Microarrays and Gene-expression signatures.....	11
1.4.1 Microarray Technology.....	11
1.4.2 Expression Profiling.....	12
1.4.3 Microarray data analysis .....	13
1.4.4 Frozen Robust Microarray Analysis (fRMA) .....	14
1.4.5 The Gene Expression Barcode algorithm .....	15
1.4.6 Gene expression profiling in Breast Cancer.....	17
1.4.7 Prognostic gene expression profiling.....	17
1.5 Next-Generation Sequencing.....	18

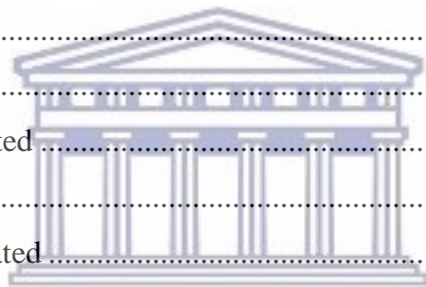
1.5.1	RNA-Seq Technology .....	18
1.5.2	RNA-Seq Data Analysis .....	19
1.5.3	Application of RNA-Seq within Cancer Studies .....	20
1.5.4	Prognostic and Diagnostic Gene Expression Profiling .....	20
1.6	Breast Cancer: Molecular Subtyping through Gene Expression Analysis .....	21
1.7	Research Rationale .....	21
1.8	Aims and Objectives .....	22
1.8.1	The application of the GExB to Micro-array data .....	23
1.8.2	The development of a barcoding method for RNA-sequencing data, comparable to the GExB algorithm .....	23
Chapter 2	.....	25
A Two-Class Breast Cancer Classifier for Malignancy	.....	25
ABSTRACT	.....	25
2.1	Introduction .....	26
2.1.1	Breast Cancer Classification .....	26
2.1.2	Gene Expression Profiling .....	27
2.1.3	Microarray Data Analysis simplified with The Gene Expression Barcode (GExB) algorithm .....	28
2.1.4	Feature Selection for Classification .....	29
2.1.5	Study Aims and Objectives .....	29
2.2	Methods and Materials .....	30
2.2.1	Data Curation .....	30
2.2.2	Gene Expression Barcode (GExB) implementation and data integration .....	32
2.2.3	Feature selection .....	33
2.2.4	Machine learning classifier based evaluation of the signatures .....	35
2.2.4.1	K-means and Hierarchical clustering .....	35
2.2.4.2	Support Vector Machines (SVM) .....	35
2.3	Results .....	36
2.3.1	Phase I: Preliminary Method Design .....	36
2.3.1.1	Gene Expression Barcode-Feature Selection paired method (GExB-FS) .....	36



2.3.1.2	Machine Learning: K-means, Hierarchical clustering, SVM.....	36
2.3.2	Phase II: Method Optimization .....	38
2.3.2.1	Gene Expression Barcode-Feature Selection paired method (GExB-FS).....	38
2.3.2.2	Machine Learning: K-means, Hierarchical clustering, SVM.....	39
2.4	Discussion .....	40
2.5	Conclusion.....	44
Chapter 3 .....		46
A Multi-Class Breast Cancer Classifier for Molecular Subtyping.....		46
ABSTRACT .....		46
3.1	Introduction .....	47
3.1.1	Breast Cancer and Personalized Medicine.....	47
3.1.2	Multi-class Classification and predictive modelling.....	49
3.1.3	Implementing Frozen Robust Multi-array Analysis (fRMA) and the Gene Expression Barcode (GExB) algorithm for microarray gene expression data.....	50
3.1.4	Machine Learning and Feature selection for Breast Cancer Classification .....	50
3.1.5	Aims and Objectives .....	51
3.2	Materials and Method.....	52
3.2.1	Data Curation .....	52
3.2.2	Gene Expression Barcode (GExB) implementation and data integration .....	53
3.2.3	Feature selection and application to datasets .....	54
3.2.3.1	Phase I: Preliminary Phase - Method Development .....	54
3.2.3.2	Phase II: Method Optimization .....	55
3.2.4	Machine Learning classifier evaluation .....	55
3.2.5	K-means and Hierarchical clustering.....	56
3.2.6	<i>k</i> -Nearest Neighbour classification .....	56
3.2.7	Multi-class Support Vector Machines (SVM) classification .....	57
3.3	Results .....	57
3.3.1	Phase I: Preliminary Method Development.....	57
3.3.2	Phase II: Method Optimization .....	59

3.4	Discussion .....	61
3.5	Conclusion.....	64
Chapter 4	.....	65
A Multi-Class Classifier for RNA-Seq Breast Cancer Data	.....	65
ABSTRACT	.....	65
4.1	Introduction .....	67
4.1.1	RNA-Sequencing and Cancer .....	67
4.1.2	Breast Cancer Transcriptomics .....	67
4.1.3	Public Transcriptomic Data .....	68
4.1.4	The Cancer Genome Atlas (TCGA).....	68
4.1.5	The Genotype-Tissue Expression (GTEx) project.....	69
4.1.6	Research Aims and Objectives.....	69
4.2	Materials and Methods .....	70
4.2.1	Data Curation .....	70
4.2.2	Discovery of Differentially Expressed Genes (DEGs) .....	72
4.2.3	Paired TCGA Samples .....	72
4.2.4	Integrated GTEx and TCGA datasets .....	73
4.2.5	Separation of GTEx normal from TCGA normal-adjacent-tumour	73
4.2.6	Z-Score Barcoding of RNA-Seq count data .....	74
4.2.7	Z-score Barcoding of unpaired TCGA and GTEx samples .....	74
4.2.8	Unsupervised Machine Learning: Hierarchical and K-means Clustering.....	75
4.2.9	Supervised Machine Learning: Support Vector Machines .....	76
4.3	Results .....	76
4.3.1	Feature set discovery in a paired TCGA normal-tumour dataset....	76
4.3.2	Feature set discovery in an integrated GTEx-TCGA dataset.....	82
4.3.3	Feature set discovery for multi-class classification of a GTEx normal, TCGA normal-adjacent-tumour and TCGA primary tumour integrated dataset .....	82
4.3.4	Z-scores and “Barcoding” RNA-Seq gene counts .....	83
4.3.5	Machine Learning classification .....	84
4.3.5.1	Clustering and SVM classification of TCGA data.....	84

4.3.5.2 Clustering and SVM classification of GTEx-TCGA integrated data.....	85
4.3.5.3 Multi-class classification of healthy breast, normal-adjacent-tumour (NAT) and primary tumour tissues.....	85
4.4 Discussion .....	89
4.5 Conclusions .....	93
CHAPTER 5 .....	94
Conclusions and Future Work.....	94
5.1 Conclusion.....	94
5.2 Discovered signatures are applicable across technologies .....	95
5.3 Future Work .....	96
6 References .....	97
7 Appendices.....	117
Appendix I.....	117
Appendix II.....	126
GTEx Data Curated .....	126
Appendix III.....	134
TCGA Data Curated .....	134



UNIVERSITY *of the*  
WESTERN CAPE

## List of Figures

Figure 1. 1: An example of K-means clustering where $k = 3$ (Jain, 2009) .....	4
Figure 1. 2: An example of a dendrogram generated from hierarchical clustering. (Jain, 2009).....	5
Figure 1. 3: An example of a binary, two-class SVM with hyperplane construction (Statnikov et al., 2005).....	7
Figure 2. 1: Example of expression calls from a micro-array converted to an absolute call by use of the Gene Expression Barcode algorithm. ....	33
Figure 2. 2: Hierarchical clustering of Training Data Set using 64 informative probe set .....	37
Figure 2. 3: Hierarchical clustering of Training Data Set using 85 informative probe set .....	39
Figure 3. 1. : Hierarchical clustering of Training Data Set using 346-gene signature.....	58
Figure 4. 1: Hierarchical clustering of “barcoded” TCGA Normal-Adjacent-Tumour (NAT) and TCGA Primary Tumour (Tumour) Unpaired RNA-Seq samples ( $n = 100$ ) yielded 98% accuracy when classified using the Top 100 DEG's discovered using 40 paired Normal-Adjacent-Tumour and Tumour samples.....	80
Figure 4. 2: Hierarchical clustering of “barcoded” TCGA Normal-Adjacent-Tumour (NAT) and TCGA Primary Tumour (Tumour) Unpaired RNA-Seq samples ( $n = 100$ ) yielded 100% accuracy when classified using the Top 100 DEG's discovered using 80 paired Normal-Adjacent-Tumour and Tumour samples.....	81
Figure 4. 3: Z-Score Heatmap of Paired Normal-Tumour Samples using Top 100 DEGs as a feature set .....	84
Figure 4. 4: Barcode-based hierarchical clustering of 300 GTEx and TCGA samples, yielded 98% accuracy when classified with the Top 59-overlapping DEG's (described in Sections 4.2.5 and 4.3.3). ....	86
Figure 4. 5: Barcode-based hierarchical clustering of 300 GTEx and TCGA samples, yielded 100% accuracy when classified with the Top 216-overlapping DEG's (described in Sections 4.2.5 and 4.3.3). ....	87
Figure 4. 6: Heatmap of 59 DEG's separating GTEx normal, TCGA NAT, and TCGA tumour tissues.....	89

## List of Tables

Table 2. 1: Summary of Breast Cancer Samples curated.....	32
Table 2. 2: Validation of Preliminary Two-class Classifier.....	38
Table 2. 3: Validation of Optimized Two-class Classifier.....	40
Table 3. 1: Summary of Breast Cancer Samples curated.....	53
Table 3. 2: Validation of Preliminary Multi-class Classifier .....	59
Table 3. 3: kNN Leave-Out-One Cross-Validation classification of Training dataset.....	60
Table 3. 4: One-versus-one Multi-Class SVM classification of Unseen Validation dataset.....	60
Table 4. 1: Summary of breast tissue samples curated from The Cancer Genome Atlas Data Repository.....	71
Table 4. 2: Top 50 DEG's extracted from differential expression analysis of 80 paired TCGA NAT and Primary Tumour samples .....	78
Table 4. 3: Classification results of TCGA Normal-Adjacent-Tumour (NAT) and TCGA Primary Tumour (Tumour) Unpaired RNA-Seq samples (n = 140) .....	85
Table 4. 4: Classification results of GTEx Normal (Normal) and TCGA Primary Tumour (Tumour) Test RNA-Seq samples (n = 100).....	85
Table 4. 5: Classification results of GTEx Normal (Normal), TCGA Normal-Adjacent-Tumour (NAT) and TCGA Primary Tumour (Tumour) RNA-Seq samples with Validation dataset.....	88
Table 4. 6: Overlap of Top DEG's: SVM classification with Validation dataset..	88
Table 7. 1: Microarray breast tissue samples curated.....	117
Table 7. 2: Normal breast tissue samples filtered from GTEx Version 7 Gene Counts file .....	126
Table 7. 3: Paired TCGA NAT and Primary Tumour Samples .....	134
Table 7. 4: Unpaired TGCA NAT Samples .....	136
Table 7. 5: Unpaired TCGA Tumour Samples .....	139

# Chapter 1

## Literature Review

The pairing of biological data and computational algorithms has contributed to new classification models of cancer. Past and current high throughput analysis of cells and tissues is revolutionizing biomedical and biological research. Completion of the whole human genome, discoveries of gene sequence and annotation along with the development of microarray technology, and more recently, next-generation sequencing (NGS) technologies, over the past 15 years has seen characterization of cells and tissues in greater depth. Although our knowledge of the human genome has improved vastly, genomic data does not provide enough information on the differentiation of cell types, while Transcriptomic data has proven to be more informative. Despite these advancements, there have been little to no big advances in diagnosis or treatment (McCall, Uppal, Jaffee, Zilliox, & Irizarry, 2011; Zilliox & Irizarry, 2007a).

The vast amount of publicly available gene expression data has seen a move towards classification models for cancer from gene expression profiling. The profiling entails examination of the differential expression of genes and their unique combinations in different states of the cancerous tissues and healthy tissue. Gene signatures have been developed which can predict cancer subtype and prognosis, such as Mammaprint® and Oncotype® DX. The robust nature of machine learning algorithms has accelerated and assisted the design of such signatures through application of the mathematical and data sciences to biomedical questions.

### 1.1 Machine Learning

Machine learning encompasses the design and application of algorithms that enable the use of existing data to establish models for pattern recognition, classification and prediction (Alpaydin, 2010). The aim of automatic model construction approaches is to minimize human biases and errors that could skew selection and performance of the algorithm, while enabling the discovery of subtle

patterns and associations between data points. Over the years, machine learning techniques have become more pliable and have been expanding together with mathematical frameworks for measuring reliability. The coupling has led to improving the efficiency and accuracy of discoveries made in biology and understanding complex biological data (Sommer & Gerlich, 2013; Tarca, Carey, Chen, Romero, & Drăghici, 2007).

Within machine learning, two exemplars exist; supervised and unsupervised learning. Supervised learning entails a sample or group using a feature set of attributes such as genes. The resultant classification scheme is a set of rules that designate objects based on the values of the features. The primary objective of supervised learning is to construct a system capable of accurately predicting the class “membership” of an object. Other than accurate classification of unknown objects, supervised machine learning also aims to be able to predict possible outliers in data; those instances that do not specifically match any of the predefined classes according to the features selected by the algorithms designed. An example of object-to-class assignments, in a biological setting, would be classification of tissue gene expression profiles to disease group (Libbrecht, Noble, & Genome, 2017).

Unsupervised learning, conversely, has no predefined class labels for the data to be studied. The aim instead is to simultaneously analyse the data and observe similarities between objects. The similarities observed, called clusters would define groups of objects. Hence, unsupervised learning's intention is to reveal naturally occurring groupings of objects based on the measurements of specific features in data (Yip, Cheng, & Gerstein, 2013)

### **1.1.1 Different Machine Learning classifiers and algorithms**

#### **1.1.1.1 K-means clustering**

Clustering algorithms are considered to be a form of unsupervised learning. Data instances that share similarities are grouped together. The algorithm can only access data about the features describing each object. However, in real



applications of clustering, the scientist usually has some knowledge about the dataset (Wagstaf, Cardie, Rogers, & Schroedl, 2001). Data clustering can be separated into two types, namely hierarchical and partitional clustering.

K-means clustering is classified as a partitional clustering algorithm. The method finds a partition that separates data (Jain & Dubes, 1988) by minimizing the squared error between the empirical mean of a cluster and the data instances, called points, of the said cluster. The main aim of the K-means algorithm is to minimize the squared error of all the clusters specified for a given dataset being investigated for classification (Drineas, Frieze, Kannan, Vempala, & Vinay, 2004).

The algorithm first selects  $k$  initial clusters, then for a specific data instance,  $x$ , assigns it to the closest related cluster centers. Every time a new data instance is added to the dataset, the cluster center is re-computed to be the average (mean) of its constituent data instances. K-means converges when there are no changes made to the clusters formed; the squared error of the cluster's mean is minimized and the cluster is centred (MacQueen, 1967).

Distance metrics are used to compute the distance between related samples or data instances in a cluster and also the distance between the different clusters. Euclidean distance, a metric based on the Pythagoras theorem of points in a dimensional plane, is applied in K-means clustering (Mao & Jain, 1996).

### **1.1.1.2 Hierarchical Clustering**

Hierarchical clustering is based on variants of primarily three algorithms; single-link (King, 1967), complete-link (Sneath & Sokal, 1962), and multi-variance (Murtagh, 1984; Ward, 1963). The two most broadly used algorithms in hierarchical clustering are single-link and complete-link, however, the two algorithms are distinctive in the manner in which they separate and characterize clusters of similarity. The hierarchy of clusters formed from the single-link algorithm can easily be used to construct dendograms; allowing the easy



visualization of clusters formed. The visualization of hierarchical clustering is depicted in a tree-like format, a dendrogram, and branches correlate to the data instances being clustered (D'haeseleer, 2005), with closely related data clustering together as one big branch. Dendrograms are particularly useful for classification applications within biological settings.

Nested clusters, data instances or samples branching from the main branches of the dendrogram, are found either in an agglomerative or a divisive manner. The agglomerative mode entails starting with each data point in its own cluster and merging the most similar clusters in successive order to form a hierarchy. The divisive mode separates and organizes all data points from a single large cluster into smaller clusters (Jain, 2010).

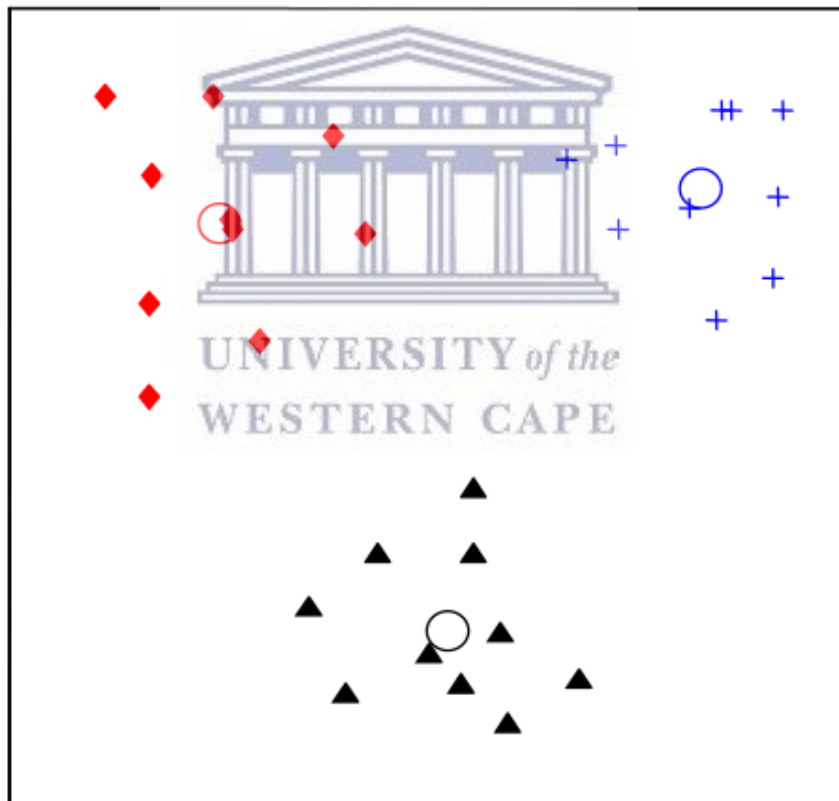


Figure 1. 1: An example of K-means clustering where  $k = 3$  (Jain, 2009).

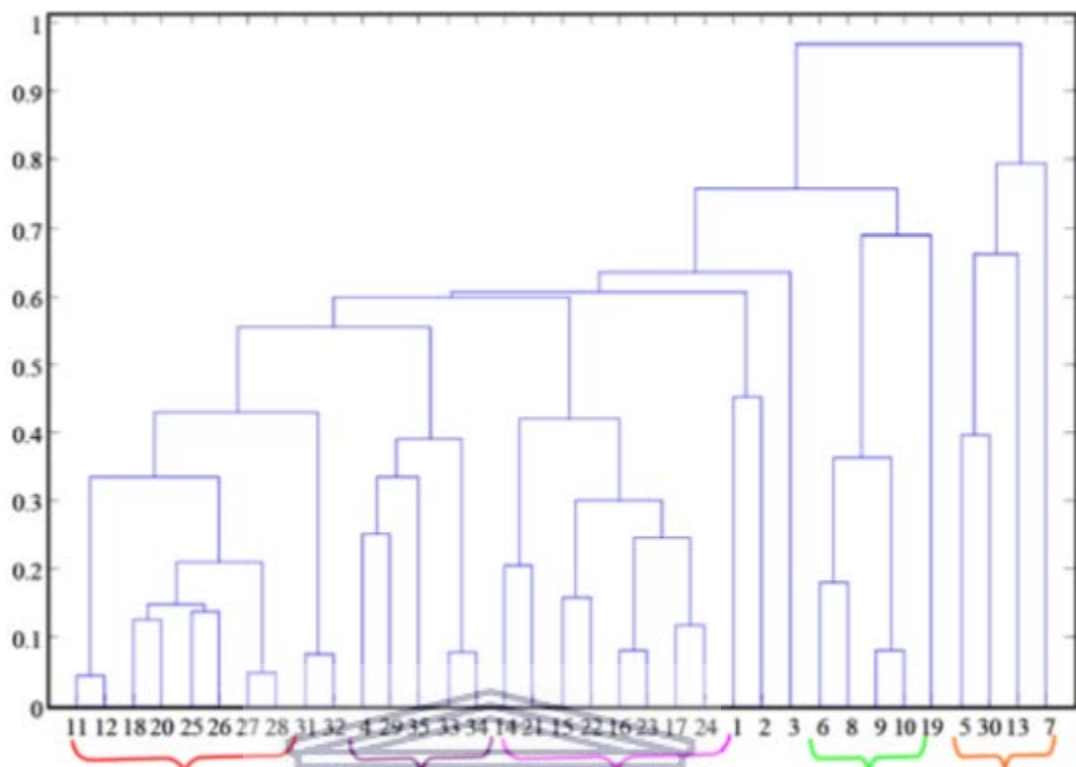


Figure 1. 2: An example of a dendrogram generated from hierarchical clustering. (Jain, 2009).

### 1.1.1.3 Neural Networks

Artificial Neural Networks (ANNs) are essentially mathematical models that were developed based on how biological nervous systems function and transmit signals and impulses. ANNs are a form of supervised learning, which is commonly categorised as semi-supervised learning. It uses a feed forward network; signals which can be represented by variables such as genes which are either mutated or gene expression levels of a specific cell, are inseminated through the layers of units. The units referred to mimic neurons, and are referred to as nodes (Abraham, 2005).

Usually three units make up the network; (1) an input layer, which is in most cases fed with gene expression data, (2) a hidden layer(s) of units, and (3) an output layer, one for each classification of tissue, in a biological instance (Mitchell, 1997). The connections formed between the layers are assigned weights, which are adjusted during the training phase of machine learning. With

back-propagation neural networks, the algorithm adjusts the weights by back-propagating the error between the units until the best fit for the training data is found (Statnikov, Aliferis, Tsamardinos, Hardin, & Levy, 2005). The weights are modelled on neuronal synapses and input signals are disseminated in a non-linear fashion as to simulate how signals are transmitted by neurons (Abraham, 2005).

Commonly used NN algorithms include Forward Propagation, Back Propagation, and Probabilistic Neural Networks (Mitchell, 1997).

#### **1.1.1.4 Self-Organizing Maps (SOMs)**

SOMs can be viewed as a derivative of Artificial Neural Networks. Data dimension reduction is the main objective of SOMs and is primarily a qualitative data visualization tool. The algorithm learns the classification, topology and distribution of input vectors. Neurons or nodes are assigned according to the amount of input vectors. Nodes in close proximity to each other learn to respond to similar input data.

The algorithm is designed such that data regularities and correlations are detected; resulting in future response being adapted accordingly. Data visualization aims to solve humans' inability to visualize high-dimensional data through mapping data in a 1- or 2-dimensional space. SOMs generate maps that plot data instance similarities into clusters (Abraham, 2005). The machine learning application itself is unsupervised.

#### **1.1.1.5 Support Vector Machines (SVMs)**

Support Vector Machines (SVM) was initially developed for two-class classification problems. The aim was to develop an algorithm capable of robust pattern recognition that would have high generalization ability with minimal errors in the training datasets. Polynomial and radial basis function equations were used to obtain optimal margins that would separate two classes within a training set (Cortes & Vapnik, 1995). SVM has been shown to classify data with superior accuracy to other supervised machine learning algorithms like early ANN's and

ensemble classification methods (Statnikov et al., 2005). The applications of SVM's vary from text-categorization technologies (Joachims, 1998), to facial recognition software (Osuna, Freund, & Girosit, 2000), to biological implementation in disease classification like cancer and bacterial infections (Su et al., 2001).

#### 1.1.1.5.1 Binary SVMs

Support Vector Machines are considered one of the most reliable forms of machine learning (Furey et al., 2000). Initially designed for binary, or two-class classification, the algorithm maps data instances to a dimensional space. A maximum-margin margin hyperplane is then identified to separate training instances (Vapnik, 1998). The set of training instances used to construct the boundary or hyperplane, are referred to as support vectors. When an unknown data sample is introduced, the algorithm will classify it based on the side of the hyperplane it falls into (Statnikov et al., 2005).

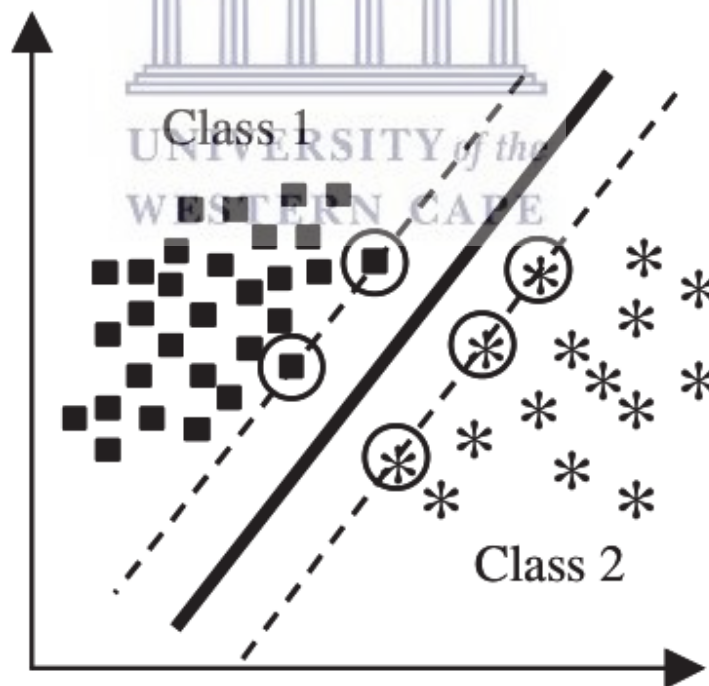


Figure 1. 3: An example of a binary, two-class SVM with hyperplane construction (Statnikov et al., 2005).

#### 1.1.1.5.2 Multi-class SVMs

Multi-class SVMs arose as the need for multi-category classification for disease and industry arose. The most commonly used amendments of the binary SVM algorithms are the *One-versus-Rest (OVR)* and the *One-versus-One (OVO)* adaptations (Ulrich, 1999). The OVR method constructs  $k$  binary SVM classifiers: class 1 (positive) versus all the other classes and proceeds to do the same for all  $k$  classes in the experiment. The combined decision function would correlate to the maximum value of  $k$  binary decision functions. The OVO method, builds binary classifiers for all pairs of classes. Subsequently, a binary problem is solved: a decision function assigns an instance to a class with the largest number of votes. Recent studies have revealed that the OVR multi-class SVM algorithm has superior classification performance which is further enhanced when feature-selection methods are applied to data preceding SVM classification (Statnikov et al., 2005).

#### 1.2 Application of Machine Learning (ML) in Biomedical Scenarios

Machine learning has the ability to solve classification problems in real world medical diagnosis. The development of algorithms such as Artificial Neural Networks (NN), Decision Trees,  $k$ -Nearest Neighbour ( $k$ NN), and Support Vector Machines (SVM) have assisted in both disease diagnosis and classification but also the interpretation of biological data from technologies like PCR, Microarray assays, and DNA- and RNA-sequencing data (Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2015).

Bioinformatics has been able to make progress in fields relating to disease diagnosis and mechanism through genomic and proteomic function prediction. Applications of machine learning to systems biology include: protein-coding genes, protein function prediction, protein-RNA interactions (Caragea & Honavar, 2009), and the impact of these genetic factors on cell regulation and function.

Machine learning has been applied extensively in cancer classification from gene expression profiles (Kourou et al., 2015).

Artificial intelligence has been used in cancer prediction and prognosis for more than 25 years; predominantly with applications of NN's and decision trees (Cichetti, 1992). The diagnosis of cancers is not only achieved through gene expression analysis, but also from tumour biopsy histopathological examinations, X-rays and CRT images. Machine learning algorithms have contributed to accurate classification of tumours using data from all of these technologies (Liotta & Petricoin, 2000; Zhou, Liu, & Wong, 2004). The accurate prediction of cancer susceptibility and diagnosis, which integrates both macro (physical) and microscopic (genetic) data, has vastly improved through the application of machine learning algorithms. Furthermore, ML has assisted in the identification of novel disease biomarkers and drug targets (Cruz & Wishart, 2006).

### **1.3 Feature Selection**

Irrelevant information is part of raw data generated from biological studies (Guyon, Weston, Stephen, & Vapnik, 2002). The need for Feature Selection (FS) techniques in bioinformatics has therefore grown in recent years, as it is now a requirement in the building of models for real-world applications. Originally, the designs for pattern recognition software were not built to manage large amounts of data. Due to the high dimensionality of biological data used in computational biology, dimension reduction is implemented to facilitate the interpretation of data. Feature selection offers dimension reduction without the loss of the original data representation, and merely selects a subset of the definitive properties of a data instance, e.g. genes expressed. FS can be applied to both supervised and unsupervised machine learning algorithms and classifiers (Liu & Yu, 2005).

The three main aims of FS approaches include: (1) to avoid over-fitting and improve model performance, (2) to provide faster and more cost-effective models, and (3) to gain deeper insight into the underlying processes that generated the data. Selection of features cannot be dependent of the parameters of the optimized machine learning algorithm applied or classification model under investigation. Idealistically, the optimal model parameters and optimal feature set are paired

(Daelemans & Hoste, 2002). Within classification schemes, there are three categories of FS methods; filter, wrapper and embedded methods. Each differ in how they are implemented with the construction of the classification model (Saeys, Inza, & Larrañaga, 2007).

### **1.3.1 Feature Selection techniques**

Filter techniques evaluate the relevance of features by taking only intrinsic properties of the data into account. The approach calculates a feature relevance score, and low-scoring features are removed from the original feature set (Saeys et al., 2007). Features selected must be relevant for prediction, but redundant features should be minimized. Relevance criteria measures how well a feature, e.g. a gene expressed or microarray chip probe, distinguishes between classes of data. Criteria like Symmetric Uncertainty (SU), Spearman rank correlation coefficient (CC), Value Difference Metric (VDM), Fit Criterion (FC) measure how useful a variable is for predicting the class of a data instance (Auffarth, 2010). Thereafter, the set of features selected are presented as input to the classification algorithm.

Wrapper techniques embed the model hypothesis search within the feature subset search. With wrapper methods, a search protocol in the space of possible feature subsets is defined, and various subsets of features are generated and evaluated. The evaluation of a specific subset of features is obtained by training and testing a specific classification model, rendering this approach tailored to a specific classification algorithm. To search the space of all feature subsets, a search algorithm is then ‘wrapped’ around the classification model.

With embedded techniques, the search for an optimal subset of features is built into the classifier construction itself, and can be seen as a search in the combined space of feature subsets and hypotheses. This method interacts with the classifier and is better computationally when compared to wrapper methods. The embedded approach is also capable of modelling feature dependencies. As with wrapper approaches, embedded approaches are thus specific to a given learning algorithm



(Saeys et al., 2007).

### **1.3.2 Dimension reduction of expression microarray data using feature selection**

Univariate filter techniques are most favoured in dimension reduction of microarray data. The method is fast and efficient, yet simple. In comparative studies of different classification algorithms paired with feature selection, the filter method is most prevalent in evaluation and investigation of DNA and mRNA microarray datasets (Dudoit, Fridlyand, & Speed, 2002; J. W. Lee, Lee, Park, & Song, 2005; Li, Zhang, & Ogihara, 2004). Reasons for this include; the output of feature ranking is easy to understand, the gene-ranking output fulfils the objectives of bio-domain experts that want to validate results in laboratories, and short computation time for data analysis (Saeys et al., 2007).

However, univariate approaches have restrictions, and in some instances lead to less accurate classifiers as they ignore gene-gene interactions. FS techniques using wrapper or embedded methods, can offer a way to perform multivariate gene subset selection (Saeys et al., 2007). Hybrid methods that incorporate univariate pre-selection with multivariate altered wrapper methods have also been proposed in the case of cancer classification (Ruiz, Riquelme, & Aguilar-Ruiz, 2006).

## **1.4 Microarrays and Gene-expression signatures**

### **1.4.1 Microarray Technology**

Microarray chips are designed to generate gene expression measures from cell and tissue samples by using cellular mRNA to elucidate gene up-regulation and down-regulation in different tissues; ranging from biological to agricultural settings. Nucleic acid microarrays make use of short oligonucleotides (15-25 nt), long oligonucleotides (50-120 nt), and PCR-generated complimentary DNA (cDNAs) (100-3000 base pairs) as array elements (Miller & Tang, 2009; Stears, Martinsky, & Schena, 2003).

Short oligonucleotides and cDNAs have both been shown to perform well for



expression analysis. However, each has its own drawback. Short oligonucleotides can lack single-gene specificity in complex hybridizations. On the other hand, PCR-generated cDNAs produce strong signals and high specificity (*Schena, 1996; Lockhart et.al., 1996; Yuen et.al., 2002*). Long oligonucleotides produce strong hybridization signals, good specificity and the ability to unambiguously identify transcripts within samples; but are dependent on the availability of genomic sequence information for each species under study (*Kane et.al., 2000*).

Expansion of traditional microarrays into exon arrays has allowed for larger coverage of exon regions of genes, and has been termed as whole transcript arrays. This is also largely due to an increase in array features, by decreasing the number of probes (Okoniewski & Miller, 2008). Gene alternative splicing through hybridization of variant transcript isoforms is detectable by exon arrays, along with expression levels of each exon independently (Bemmo et al., 2008; Kapur, Xing, Ouyang, & Wong, 2007).

#### **1.4.2 Expression Profiling**

Quantitative gene expression data is generated by transcript profiling. In order for profiling to take place, one- or two-colour fluorescent schemes are implemented (R. J. Cho et al., 1998), and the most broadly used and easily interpreted scheme is two-colour fluorescence. Each RNA sample is labelled with two different fluorescent tags prior to hybridization with cDNA. Visualization of genes that are “activated” or “repressed” is produced from two-colour graphical superimposition. The two-colour graphic representation of probes expressing genes at different levels, allows the separation and comparison of various tissues based on their respective expression profiles. This process allows for the throughput of high quality gene expression data (S. M. Y. Lee et al., 2002).

Detection of fluorescent probes (tags attached to oligonucleotides/genes) is achieved with instruments that contain confocal optics, photomultiplier tubes, and charge-coupled devices. The detection instruments render graphical images in tagged-image file format (TIFF), which are two-dimensional, 16-bit numerical

representations of microarray surfaces with intensity values assigned. These numerical values are then interpreted as expression values of genes (Carr, Somogyi, & Michaels, 1997). The data collected is raw data, and sequentially further data analysis which includes transformation and normalization of data is necessary for extrapolation of biologically significant knowledge from machine learning algorithm applications and microarray analysis software (Stears et al., 2003).

### **1.4.3 Microarray data analysis**

In order for microarray data to become useful in biological settings, a wide range of data analysis and processing is required. The two most important components of the analysis are design and pre-processing. Both are necessary steps preceding the classification of genes, cells and tissues, as well as validation of data (Allison, Cui, Page, & Sabripour, 2006).

Design: How the microarray experiments and the relevant study is designed impacts efficiency and validity of experiments. Within the design of a study, there are certain optimization steps that can be employed (Kerr, 2003). Firstly, biological replication is imperative. There are two forms of replication which can be applied to microarray experiments, which include technical and/or biological replication (Churchill, 2002; Yang, Buckley, & Speed, 2001). Secondly, the pooling of biological samples may further assist design optimization. This is due to the fact that when trying to ascertain and identify differential gene expression, high data variability can be eliminated from a study (Kendziorski, Irizarry, Chen, Haag, & Gould, 2005). And thirdly, avoiding confounding by extraneous factors is vital. When such factors vary with the independent variable of the experiment, it may yield confusing and erroneous conclusions of a study (Kerr, 2003).

Preprocessing: Image analysis and data normalization and transformation form part of pre-processing. These steps are required in order to remove systematic variation in the data. Normalization of data from different experiments and chip platforms is necessary to account not only for background noise (mismatched

probes), but also technical variance of fluorescence readings, which infer up- and down-regulation of gene expression, of microarray chips. Data transformation typically describes mathematical formulas being applied to data to change the format. Most often,  $\log_2$  is applied to numerical values produced from micro-array detection technologies (Allison et al., 2006).

The most broadly applied micro-array data normalisation algorithm used is called robust multi-array average (RMA), designed for use on Affymetrix and Nimblegen microarray platforms (Irizarry, Bolstad, et al., 2003). The algorithm corrects data for background noise by transforming the data. Normalization by the algorithm is performed with a formula that uses normal distribution and a linear model to estimate expression values on a log scale). An alteration to RMA is GCRMA, which corrects for the GC content of the oligonucleotides used in the initial microarray chip experiment (Bolstad, Irizarry, Åstrand, & Speed, 2003).

#### **1.4.4 Frozen Robust Microarray Analysis (fRMA)**

The use of gene expression microarray experiments has become broadly used for research in biological studies. Methods for data analysis have had to adapt to the various aspects that affect micro-array data, such as batch effects, noise and reproducibility of experiments. Micro-array analysis consists firstly, of preprocessing the probe-level fluorescent readings to gene-level expression estimates. This initial step requires algorithms to resolve multiple or batches of arrays together (Bolstad et al., 2003). Despite the robust nature of the RMA algorithm, the multi-array processing complicates and limits inquiry (Ramasamy, Mondry, Holmes, & Altman, 2008). To process individual array experiments is computationally expensive, and introducing data from single arrays cannot be combined without introducing noise. This is a real dilemma for applying microarray technologies to clinical settings; the requirement is to extract actionable information from a single sample as opposed to a batch set of samples from an isolated experiment.

The frozen Robust Micro-array Analysis (fRMA) algorithm was hence designed,

as it presented a method to pre-process individual array experiments, while retaining the advantages of batch array pre-processing (McCall, Bolstad, & Irizarry, 2010a). The basis of fRMA is simple; the parameter estimates are pre-computed on a massive and biologically diverse database of micro-array experiments, after which these parameters are frozen. This is then used to pre-process individual or low sample batches and later condensed for analysis.

#### **1.4.5 The Gene Expression Barcode algorithm**

The complexity of distinguishing tissues based on transcriptomic or microarray data is due to the use of relative expression of genes when reporting data, i.e. which genes are differentially expressed in one condition compared to others (Parkinson et al., 2009). Probe effects and noisy data obfuscate the correlation between observed probe intensity and actual expression of a transcript. Knowing absolute expression of genes, i.e. whether a gene is expressed or not, instead of relative expression of a gene in a tissue type would improve our understanding of systems and cellular biology, and provide a starting point for research targeting drug discovery and personalized medicine (McCall et al., 2011).

For the above reasons The Gene Expression Barcode project was established. The initial barcode algorithm, referred to as *Barcode 1.0* (McCall et al., 2011) was based on a basic detection method and distance calculation. The rationale behind the algorithm was to develop the first method that could clearly demarcate expressed from silenced genes; and in so doing, denominate a specific or unique gene expression barcode for each tissue type. Vast numbers of raw microarray data was curated from publicly available datasets in the Gene Expression Omnibus (GEO) and ArrayExpress data repositories and pre-processed with the same algorithm. Clinical data from three cancer studies and one Alzheimer's disease study was also collected. The aim was to evaluate which probe intensity relates to expression. Thereafter, the intensity distribution for each gene needed to be determined. Genes that are shown to be expressed would be classified as ones and silenced genes, as zeros. The sequence generated is referred to as the gene expression barcode (Zilliox & Irizarry, 2007a).

Due to the original barcode methodology only being able to provide absolute expression measures for a limited number of genes, the algorithm was extended to estimate transcriptomes (McCall et al., 2011). This is motivated by the fact that transcriptome data allows insight into what discriminates cell and tissue types, hence contributing to the classification of unknown biological samples. In order to clearly classify genes as silenced or expressed, one needs clear separation between high and low expression values. This is not the case in the majority of genes. The original barcode algorithm was further developed to determine a more extensive estimate of cell-type transcriptomes by calculating expression calls for all genes represented on the array. This was achieved by firstly, establishing a set of negative control experiments; secondly, by mass curation of publicly available microarray data from the Affymetrix Human Genome U133A (HG133a), U133 Plus 2.0 (HG133plus2) and Mouse Genome 430 2.0 (Mouse4302) platforms; and thirdly, applying the probability of expression (POE) model in a novel setting (Parmigiani, Garrett, Anbazhagan, & Gabrielson, 2002).

A new version of the algorithm resulted which produced standardized values; allowing for comparison across all genes. The standardized values may be translated into absolute expression calls; silenced or expressed genes by designation of a single threshold value. The resultant binary values correlating to expression calls is called the “barcode” (McCall et al., 2011). Although the Gene Expression Barcode Version 3.0 has been extended to include other sequencing platforms, a method for barcoding RNA-Seq expression or raw count data has not yet emerged (McCall et al., 2014).

One of the other differences in methodology that separates the Gene Expression Barcode from other absent/present call algorithms is the approach to microarray raw data pre-processing. The common analysis tool for micro-array data is RMA; Robust Micro-array Analysis, but the barcode algorithm implements an altered algorithm, called frozen Robust Micro-array Analysis (fRMA) (McCall, Bolstad, & Irizarry, 2009).

#### **1.4.6 Gene expression profiling in Breast Cancer**

DNA-microarray technologies have provided researchers with the ideal tools and opportunities to perform comprehensive molecular and genetic profiling of breast cancer (Trevino, Falciani, & Barrera-Saldaña, 2007). Microarray techniques provide insights into cell biology as well as developing clinically useful classification models. This has allowed clinicians to predict, amongst others, disease recurrence and response to different treatments, which promises to improve disease management of cancer patients (Cooper, 2001).

#### **1.4.7 Prognostic gene expression profiling**

Over the past years, several breast cancer research groups have conducted gene-expression profiling studies with the objective of improving on traditional prognostic markers. Researchers from the Netherlands Cancer Institute in Amsterdam (NKI) reported a 70-gene prognostic signature (Mammaprint™) developed on the Agilent platform (Straver et al., 2010).

The sample size consisted of 78 systemically untreated lymph-node-negative breast cancers of patients younger than 55 years of age. A year later, Mammaprint™ was validated on a larger set of 295 young patients, this time with a mixed sample set. The NKI provided proof that the 70-gene signature was the strongest predictor for distant metastasis-free survival, independent of adjuvant treatment, tumour size, histological grade and age, both in node-negative and node-positive cohorts.

A similar study was done by a group in Rotterdam; generating a 76-gene signature that was able to determine the development of distant metastases in untreated patients of all age groups with node-negative breast cancer (Y. Zhang et al., 2009). The main difference between the Amsterdam and Rotterdam studies was the microarray platform used and the study design used in the development of the classifiers. Both classifiers appeared to be good predictors of the development of distant metastases within the first 5 years, but showed a decreased prognostic ability with the increasing number of follow-up years.



## 1.5 Next-Generation Sequencing

Sanger sequencing emerged as a “first-generation” sequencing method, and was soon widely adopted (Sanger & Coulson, 1975). Next-generation sequencing (NGS) refers to second and third generation sequencing platforms which are able to simultaneously sequence millions to billions of sequence reads for transcriptome assemblies and analyses (Figueroa, Tang, & Taur, 2014). The past ten years has seen the rapid development of various platforms, with slightly differing techniques, for the high-throughput sequencing of genomes and transcriptomes (Levy & Myers, 2016).

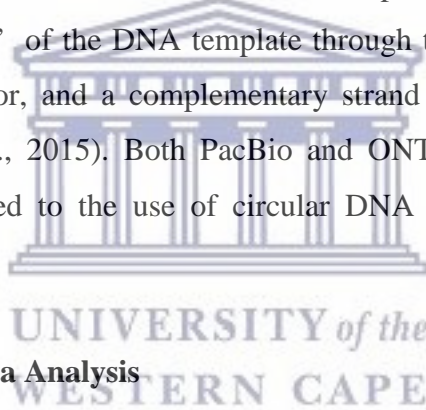
### 1.5.1 RNA-Seq Technology

RNA-sequencing (RNA-Seq) is an NGS technique which directly sequences RNA transcripts present within a cell or sample (Kukurba & Montgomery, 2015). The exploratory capabilities of RNA-Seq allows for the quantification and detection of not only protein-coding RNAs, but also non-coding RNA, miRNA, siRNA, and small RNA classes involved in RNA stability, protein translation, or chromatin state modulation (Han, Gao, Muegge, Zhang, & Zhou, 2015; Trapnell, Pachter, & Salzberg, 2009). As a whole RNA-Seq has allowed for whole transcriptome sequencing and analysis, but may also be applied to differing extents depending on the objectives of the research question.

Library preparation and sequencing comprises of multiple steps, which rely on biochemical interactions of synthetic nucleotides, and enzymes typically involved in *in vivo* DNA replication and/or RNA transcription and translation. Different technologies (different companies) achieve this through different techniques: (a) Illumina HiSeq/MiSeq technologies incorporate reversible terminator chemistry - sequencing by synthesis is achieved through reversible terminator nucleotides labelled with a different fluorescent dye, and subsequent imaging detects the positioning of these synthetic nucleotides to infer DNA sequence (Ansorge, 2009). (b) Life Technologies SOLiD sequencing utilises ligation of dinucleotide probes with DNA ligase enzymes – 16 different dinucleotide probes (labelled by four different colours) are hybridized to a template sequence (RNA fragment),

with ligation cycles resetting the primer end to successfully add the correct nucleotide complementary to the template (Ku & Roukos, 2013).

Third-generation sequencing (TGS) emerged 5 years ago in the form of Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) platform, and Oxford Nanopore technologies (ONT) following closely (Weirather et al., 2017). Both sequencing platforms use a similar technique of detecting clonally amplified DNA, as the Illumina platform (Levy & Myers, 2016). (a) PacBio differs from Illumina in that it captures a single DNA molecule, and uses circular DNA templates with hairpin adapters ligated so that the polymerase reaction synthesises a complementary circular strand (Rhoads & Au, 2015). (b) ONT implements a nanopore-based single molecule. Single-stranded DNA (ssDNA) is directly sequenced, and uses a similar circular DNA template as PacBio. Sequencing occurs by “threading” of the DNA template through the nanopore, addition of a ligated hairpin adaptor, and a complementary strand built via molecular motor proteins (Laver et al., 2015). Both PacBio and ONT produce continuous long reads (CLR) attributed to the use of circular DNA templates (Rhoads & Au, 2015).



### **1.5.2 RNA-Seq Data Analysis**

RNA-Seq produces thousands-to-millions of reads, i.e. sequence fragments, of varying lengths. Numerous Python and R packages have been developed specifically for the analysis of sequencing data. Prior to any biological investigation of the transcriptomic data generated, data pre-processing is performed. Quality assessment is the first step in bioinformatics RNA-Seq pipelines followed by mapping of the transcript fragments to a reference genome in order to ascertain the identity, location, and functions of the sequences (Han et al., 2015).

Following alignment of transcripts, gene expression is quantified by counting the number of transcript reads mapped to the respective reference genome location (Conesa et al., 2016). The gene counts generated from software like HTSeq-count



(Anders, Pyl, & Huber, 2015), or featureCounts (Liao, Smyth, & Shi, 2014) can be implemented for gene expression analysis, following normalization of raw count data. Furthermore, analysis of the aligned transcriptome can be employed to identify alternative splicing of genes, variant detection, pathway analysis through gene enrichment and discovery of gene co-expression networks (Han et al., 2015; Pereira, Imada, & Guedes, 2017).

### **1.5.3 Application of RNA-Seq within Cancer Studies**

Due to the ability of RNA-Seq to reveal a cell or tissue's entire transcriptome, integrative studies into cancer physiology have become possible. There exists a strong correlation between a tumour's transcriptome and phenotypic presentation. Deep sequencing permits a full view of the genetic regulatory and expression mechanisms governing tumorigenesis and pathophysiology of cancer (L. Wan, Pantel, & Kang, 2013). NGS has enabled the identification of gene mutations, oncogenic gene fusions (Byron, Van Keuren-Jensen, Engelthaler, Carpten, & Craig, 2016), methylation abnormalities, chromosomal rearrangements, and gene expression alterations within diseases (Ashwag Albukhari, Fawzi F. Bokhari, 2015). Interrogation of these genetic and transcriptomic cancer-specific traits may aid in the diagnosis and prognosis of different cancers and subtypes.

### **1.5.4 Prognostic and Diagnostic Gene Expression Profiling**

Gene expression profiling for diagnostic biomarker discovery has been successfully applied to a number of different cancers. Utilising RNA-Seq data, prognostic signatures for invasive lobular breast cancer (Ciriello et al., 2015), pancreatic adenocarcinoma (Kirby et al., 2016), lung adenocarcinoma (Shukla et al., 2017), as well as biomarker signatures for cancers of unknown origin (Wei, Shi, Jiang, Kumar-Sinha, & Chinnaiyan, 2014). The Cancer Genome Atlas consortium has also employed comprehensive analysis with integrative transcriptomic studies, through the application of different RNA-Seq and DNA-Seq platforms, for the discovery of molecular portraits of breast tumours (Koboldt et al., 2012), and lung adenocarcinomas (Collisson et al., 2014).

### **1.6 Breast Cancer: Molecular Subtyping through Gene Expression Analysis**

Breast cancer is most frequently diagnosed in women in Western countries and accounts for approximately 30% of all cancers diagnosed and 16% of cancer deaths (F. Bray et al., 2018). Breast cancer is a clinically, molecularly and pathologically heterogeneous disease. Gene expression profiling has allowed the identification of molecular breast cancer subtypes. Clinically, the disease has been categorized into three basic therapeutic groups. Estrogen positive (ER) breast cancer is the most diverse in presentation (Paik et al., 2004). The HER2 subtype, is characterized by the presence of HER2 gene, which implicates that the tumour is stimulated by elevated levels of growth hormones (Moasser, 2007). Triple-negative breast cancer tumours do not express any hormonal receptors and are essentially progesterone, estrogen and HER2 negative. Triple-negative cancers are viewed as the most difficult to treat with the poorest patient survival outcomes (Sorlie et al., 2003).

Various clinical and pathological factors, such as age, menopausal status, tumour size, histological grade, lymphovascular invasion, oestrogen receptor have been implicated as prognostic indicators of clinical course (Perou et al., 2000). Primary treatments consist of tumour excision and radiation or mastectomy with or without radiotherapy. Adjuvant therapies have been shown to improve the long-term survival of patients (Dinh, Sotiriou, & Piccart, 2007).

### **1.7 Research Rationale**

The burden of breast cancer incidence and prevalence in both developed and developing countries has motivated the continual research on treatment biomarkers and more accurate classification models. Heterogeneous diseases like breast cancer require investigation into the genetic differences between diseased and healthy states through gene expression profiling. Large public repositories exist, such as NCBI, GEO and Array express, containing thousands of mRNA and cDNA microarray data samples. This provides researchers with an abundance of reusable data from which novel biological insights and predictive diagnostics can be developed in a cost-effective manner. The key and associated challenge to

optimally exploiting the diversity of data available, however, is in integrating breast cancer microarray samples from different microarray platforms and study population. Studies have shown that increased sample numbers and diversity should increase statistical power and discovery of population-independent predictive signatures (Nevins et al., 2003; Rung & Brazma, 2013). In the current era, the advent of large-scale next-generation sequencing, and the advantages of RNA-Seq in complete cancer transcriptomic profiling, holds immense promise for more accurate diagnostic and prognostic signature discovery (Cieřlik & Chinnaiyan, 2018).

Prognostic and predictive gene signatures like Mammaprint™ and Oncotype DX™ (Buyse et al., 2006; Toole, Kidwell, & Van Poznak, 2014), using gene expression profiling with large microarray datasets indicates that gene-expression profiling has great potential for improving breast cancer management and increasing our understanding of disease biology. To date, only one clinically available gene signature is available developed using RNA-Seq data, FoundationOne Heme (Doebele et al., 2015), and focuses on gene fusion detection in soft tissue sarcomas (Byron et al., 2016).

Machine learning has been broadly applied to building breast cancer classifiers from gene expression data (Yue, Wang, Chen, Payne, & Liu, 2018). The simplicity of the Gene Expression Barcode (GExB), allows the integration of data from a diversity of experiments to develop accurate classifiers using machine learning algorithms. The absolute measures of expression, 1's and 0's, generated by the GExB, make implementation of a filter feature selection technique attractive; setting parameters for relevant variable (gene/probe) identification of differentiating features between diseased and healthy breast tissues. Applying these features to sophisticated algorithms, like SVM, holds promise for identifying robust and accurate gene signatures and the absence-presence nature of the signals would allow any finding to be easily migrated to simpler technology platforms such as RT-PCR.

### **1.8 Aims and Objectives**

This study was split into two parts: 1) The application of the GExB to Micro-array data and 2) The development of a barcoding method for RNA-sequencing data,

comparable to the GExB algorithm

### **1.8.1 The application of the GExB to Micro-array data**

Hypothesis: Integrating existing breast cancer microarray expression data using the Gene Expression Barcode concept will enable the discovery of easily assayable signatures for classifying breast cancer samples into subtypes.

Main aim: Develop a feature selection method to identify predictive signatures in *simplified* expression datasets and test classification accuracy on “real” clinical datasets.

The following objectives were identified for achieving the main aim:

- (1) Production of gene expression barcodes for breast cancer subtypes and development of a method for integrating barcodes from different chips.
- (2) Development of an automated feature selection pipeline for identifying a minimal set of expression features based on (1).
- (3) Evaluation and optimisation of the feature-selection method using a simple classifier.
- (4) Derivation of a variation of the feature-selection method for development of a multi-class classifier.

### **1.8.2 The development of a barcoding method for RNA-sequencing data, comparable to the GExB algorithm**

Main aim: To discover a novel method to convert gene counts in RNA-Seq data to absolute calls of expression, i.e. 1's and 0's, and therefore creating a “barcoding” method for NGS data

Objectives:

- (1) Development of a method for barcoding RNA-Seq data and application on breast cancer data from The Cancer Genome Atlas (TCGA)
- (2) Development of a two-class classifier for TCGA normal and tumour samples with feature selection based on best differentially expressed genes
- (3) Integration of RNA-Seq data from normal breast tissue samples, from the Genotype-Tissue Expression (GTEx) project to discover a

signature for multi-class classification capable of distinguishing between normal, normal-from-cancer-patient, and primary tumour samples.



UNIVERSITY *of the*  
WESTERN CAPE

## **Chapter 2**

### **A Two-Class Breast Cancer Classifier for Malignancy**

#### **ABSTRACT**

#### **INTRODUCTION:**

Breast cancer is a heterogeneous disease with an ever-growing increase in the biological subtypes being recognized. Along with molecular subtypes, are metastatic and primary cancers, where molecular profile, tumour histology and grade collectively contribute to subtype diversity. Accurate subtype classification has been shown to coincide with improved diagnosis, prognosis and aetiology; imparting a comprehensive patient status with strong correlation to clinical and 5-year survival outcomes. Histopathological examinations of tumours, which are mostly inaccurate, is unfortunately still the classification method of choice.

The use of gene expression profiling has been studied extensively for implementation in breast cancer subtyping. These profiles include classification, prognosis and in the case of MammaPrint™, chemotherapy sensitivity, and breast cancer recurrence with Oncotype DX™. Despite the success of MammaPrint™ and Oncotype™ DX, significant advances in diagnosis and treatment by gene expression profiling, diagnostic gene signatures need to be further explored.

The Gene Expression Barcode was developed to overcome the constraints of microarray expression data, such as probe effects, noisy data and the relationship between intensity and actual expression. The algorithm shows clear demarcation of low and high expression measurements to classify genes as silenced and expressed by means of a binary 'barcode'. As signatures derived from absolute expression calls would simplify implementation in a laboratory setting, we explored the potential of expression barcodes as features for machine learning based classification. We present a simple method, which combines biologically relevant feature selection with the K-means clustering algorithm to accurately classify breast tissue samples as being normal or malignant.

#### **METHODOLOGY:**

Carefully selected and curated normal and tumour samples were obtained from the NCBI's Gene Expression Omnibus database. We developed a filter to produce a minimal discriminating feature set/barcode by selecting probes which were



stably expressed within tissue types, yet differentially expressed between the two tissue types. K-means clustering of tissues based on the minimal feature set was performed to ensure that the barcode signature was able to correctly classify the training set. Unseen samples from both the original and unrelated experiments were then classified to ascertain the predictive accuracy of the signature and its ability to generalize to the classification of unseen samples.

#### RESULTS:

The optimized feature selection filter of binary data reduced the feature set from 22215 to 85 informative probes. K-means clustering showed clear separation of normal epithelial breast tissue and primary tumour samples. A 100% accuracy in tissue classification was observed, even for samples from tumour classes not represented in the training set.

#### DISCUSSION:

With a simplistic filtering and clustering technique, we were able to classify unseen normal breast and tumour samples with 100% accuracy based on gene expression data that has been converted to 'absence/presence calls'. We propose that such signatures, which may be easily translated to a PCR- or hybridization-based laboratory test, shows promise for reliable classification of tissues of ambiguous malignancy status. Furthermore, we predict that pairing our barcode and filtering approach with more powerful classification techniques such as multi-category support vector machines could produce robust expression-based classifiers that have potential for clinical application.

## **2.1 Introduction**

The prevalence of breast cancer incidence has risen to 8 million cases globally between 2007 and 2015, making it the leading cause of cancer deaths among women internationally and in South Africa (Ferlay et al., 2015; Siegel et al., 2012). This highlights the importance of early-detection and accurate classification of a biopsied tissue prior to any treatment decisions being made.

### **2.1.1 Breast Cancer Classification**

The development of prognostic and predictive breast cancer gene expression

signatures has been a decade-long aim of many gene-expression profiling and bioinformatics studies. Diagnosing a patient with breast cancer accurately from the molecular portrait of the biopsy cells improves the treatment choices made for the patient, and also the survival outcomes (Henderson & Patek, 1998). The complexity of making accurate diagnostic decisions, however, lies in the heterogeneous nature of the disease; which is comprised of distinct subtypes having varied clinical, pathological and molecular presentations. Initial diagnosis of breast cancer heavily relies upon histopathology examination of biopsied tissues; immunohistochemical (IHC) staining of biopsied cells which reveal the presence or expression of hormone receptors estrogen, progesterone and human epidermal growth factor 2 (Her2) (Patnayak et al., 2015). However, these examinations are poorly reproduced for a given breast cancer case, and thus cannot always be relied upon as informative enough for clinicians to make a diagnosis (Haibe-Kains, 2010). In a recent study on the reliability of IHC examinations, 83% of the molecular subtypes were shown to be misinterpreted (Jorns, Healy, & Zhao, 2013). This demonstrates that even though IHC in terms of hormone receptor presence is reliable, misinterpretation still negatively impacts the treatment decisions of clinicians, with far-reaching consequences.

### **2.1.2 Gene Expression Profiling**

Advances in microarray technology have granted biologists the ability to measure and assess the expression levels of thousands of genes in a single assay. Using the data and knowledge obtained, the discovery of molecular breast cancer subtypes has emerged. Gene expression profiling has been used to develop clinically relevant and implemented signatures for diagnostics and prognostics (Sotiriou & Piccart, 2007). The development of predictive and prognostic gene signatures such as MammaPrint™ has assisted clinicians with informed treatment decisions. MammaPrint™ is a 70-gene signature predictor of chemotherapy sensitivity of a patient. This clinical assay however has limitations in application, as only adjuvant drug therapy choices and consequent treatment course decisions, are informed. Oncotype DX™ is a PCR assay used in breast cancer prognostics; consisting of a 21-gene signature, which assigns a score for the likelihood of



recurrence of breast cancer in lymph-node negative, estrogen-positive patients. The assay is however limited to application in estrogen-positive breast cancer subtype cases (Toole et al., 2014). Despite the efforts of molecular biologists and bioinformaticians to discover a generically, globally applicable gene expression assay which can assess the multiple facets of the disease, the constraints of using traditional microarray raw data analysis has impacted the discovery of such signatures (Nevins et al., 2003).

### **2.1.3 Microarray Data Analysis simplified with The Gene Expression**

#### **Barcode (GExB) algorithm**

The development of Frozen Robust Microarray Analysis (fRMA) and the Gene Expression Barcode (GExB) has addressed the difficulty in using integrated microarray data from different experimental cohorts and different micro-chip platforms to generate molecular profiles of tissues (McCall et al., 2011). Robust Microarray Analysis (RMA), a broadly used normalization tool for raw data, is restricted to application of the experiment set under investigation (Irizarry, Hobbs, et al., 2003). Normalization parameters and threshold values for assigning up- or down-regulation of genes' expression are relative measures across the microchip signals being interpreted (McCall et al., 2011). fRMA conversely, has precomputed generalised values, from thousands of microarray samples, for data normalization, which allows raw data to be preprocessed identically. The meta-analysis of healthy and diseased tissues of the body has standardized parameters that mitigate the clouding factors of gene expression level values; probe effects, mismatched probes, noisy data (McCall, Bolstad, & Irizarry, 2010b). The GExB algorithm takes this continuous data, i.e. expression values, and assigns an absolute measure of gene presence or absence, represented by a "1" or "0" respectively (McCall et al., 2011; Zilliox & Irizarry, 2007b). The resultant sequence, or barcode, is similar to the Affymetrix MAS 5.0 absent-present algorithm, but is more robust and not limited to a single experiment set due to the fRMA preprocessing phase.

#### **2.1.4 Feature Selection for Classification**

Beyond the analysis of continuous data is discovering genes and/or microarray probes that are informative as distinct features to separate subtypes of a tissue. Feature selection is an integral part of designing a classifier; predominantly paired with a machine-learning algorithm. K-means clustering, an unsupervised clustering algorithm and Support Vector Machines (SVM) paired with feature-selection has proven to build efficient and accurate classifiers. SVM in particular has been used in classification systems for medical diagnosis (Akay, 2009). The success of the SVM, however, relies upon an optimal feature set and training dataset size. Finding a balance between informative features, microarray probes in this instance, and too many restrictions is key to avoid an overfitted classification model, which is still accurate (Domingos & Pedro, 2012).

#### **2.1.5 Study Aims and Objectives**

We therefore propose an integrated approach to develop a two-class signature that can accurately distinguish healthy breast tissue from malignant breast tumours, as a way to assess the utility of expression barcodes in tissue classification, and as a step towards developing a multi-class classifier. Public biomedical databases contain millions of dollars' worth of potentially reusable gene expression data that can be used to derive novel biological insights or to develop predictive diagnostics. The key lies in integrating and normalizing data from different technology platforms to make them comparable. Increased sample numbers and diversity is expected to increase statistical power and discovery of population-independent predictive signatures. The GExB data transformation procedure allows data integration, since data from different chip platforms is made numerically comparable. Feature selection is also simplified due to absolute calls being compared versus continuous data i.e. relative expression. Provided the feature set is small enough, the simplicity of "on/off" expression signals of the GExB allows any identified signature to be easily migrated to and assayed on simpler technology platforms such as real-time multiplex PCR.

We hypothesized that integrating existing breast cancer microarray expression

data using the Gene Expression Barcode concept will enable the discovery of easily assayable signatures for classifying breast cancer samples into subtypes.

Our main aims included:

- 1) Developing a method for integrating barcodes from different Affymetrix chips/platforms into a 'meta-dataset' with as many samples as possible.
- 2) Developing a feature selection pipeline for identifying a minimal set of discriminating expression features based on (1).
- 3) Producing gene expression signatures for normal and cancer breast tissue types
- 4) Development of a variation of the optimized feature-selection method for multi-class classification.

## **2.2 Methods and Materials**

### **2.2.1 Data Curation**

In order to build large and diverse training and test datasets with machine learning algorithms, data integration was imperative. Integrated datasets not only offer biological diversity to the classifier, but improves the likelihood of discovering informative microarray probe sets that can be generically applied to any of the Human Genome Array platforms and to many population groups.

During data collection from the Gene Expression Omnibus (GEO) repository, labelled samples were hand-curated. The GEO annotation of the sample file was very important, as the initial classification of the tissue has to be reliable in order to assemble a training set which was accurate and biologically correct. In particular, the healthy breast tissue samples were collected from cancer-free patients, so as to ascertain a true molecular portrait of normal breast tissue.

Raw micro-array data was curated from the NCBI Gene Expression Omnibus. Table 2.1 shows the variability of the data sources. Samples used as training data came from different experiment sets than data samples used for validation of the two-class classifier. Most notably, Her2-positive breast cancer samples were

obtained for validation, but were *not* included in the training set data and were instead intended to serve as a very difficult test case for the predictor.

Further sample source variability was introduced by integrating data from different Affymetrix Human Genome Array Platforms; GPL96\* ([HG-U133A] Affymetrix Human Genome U133A Array) and GPL570\* ([HG-U133\_Plus\_2] Affymetrix Human Genome U133 Plus 2.0 Array).



Table 2. 1: Summary of Breast Cancer Samples curated

Dataset	Tissue Type	GEO Series	GEO Platform (Affymetrix)
<b>Training</b>	Normal Epithelium	GSE20437	GPL96*
	Normal Duct	GSE5764	GPL570*
	Normal Lobe	GSE5764	GPL570
	Triple Negative Breast Tumour	GSE25065	GPL96
	Estrogen-Positive Breast Tumour	GSE25065	GPL96
	Primary Breast Tumour	GSE2990	GPL96
	Inflammatory Breast Cancer Tumour	GSE5847	GPL570
	<b>Test/ Validation</b>	Normal Epithelium	GSE9574
Triple Negative Breast Tumour		GSE31519	GPL96
Her2-Positive Breast Tumour		GSE42822	GPL96
Estrogen-Positive Breast Tumour		GSE23988; GSE22093	GPL96
Primary Breast Tumour		GSE21217; GSE5462	GPL96
Inflammatory Breast Cancer Tumour		GSE22597	GPL96

\*GPL96 - [HG-U133A] Affymetrix Human Genome U133A Array

\*GPL570 - [HG-U133\_Plus\_2] Affymetrix Human Genome U133 Plus 2.0 Array

### 2.2.2 Gene Expression Barcode (GExB) implementation and data integration

The raw microarray data collected was pre-processed with the fRMA algorithm. This ensured that the samples were identical with regard to the expression calls rendered from the varied experiment sources. Batches of micro-array samples were pre-processed according to tissue type (Normal or Tumour) and allocation to

either Training set or Validation/Test set.

The Gene Expression Barcode (GExB) algorithm was then applied to the pre-processed data to convert the raw expression calls rendered from the micro-array chips to an absolute call for probe, 1 or 0. Figure 1 shows how the algorithm converts the raw data to a barcoded sequence for each sample. The data from different platforms were then merged/integrated so as to form a training set of each tissue type with only absolute calls, a 1 or a 0 to represent the absence or presence of a gene expressed.

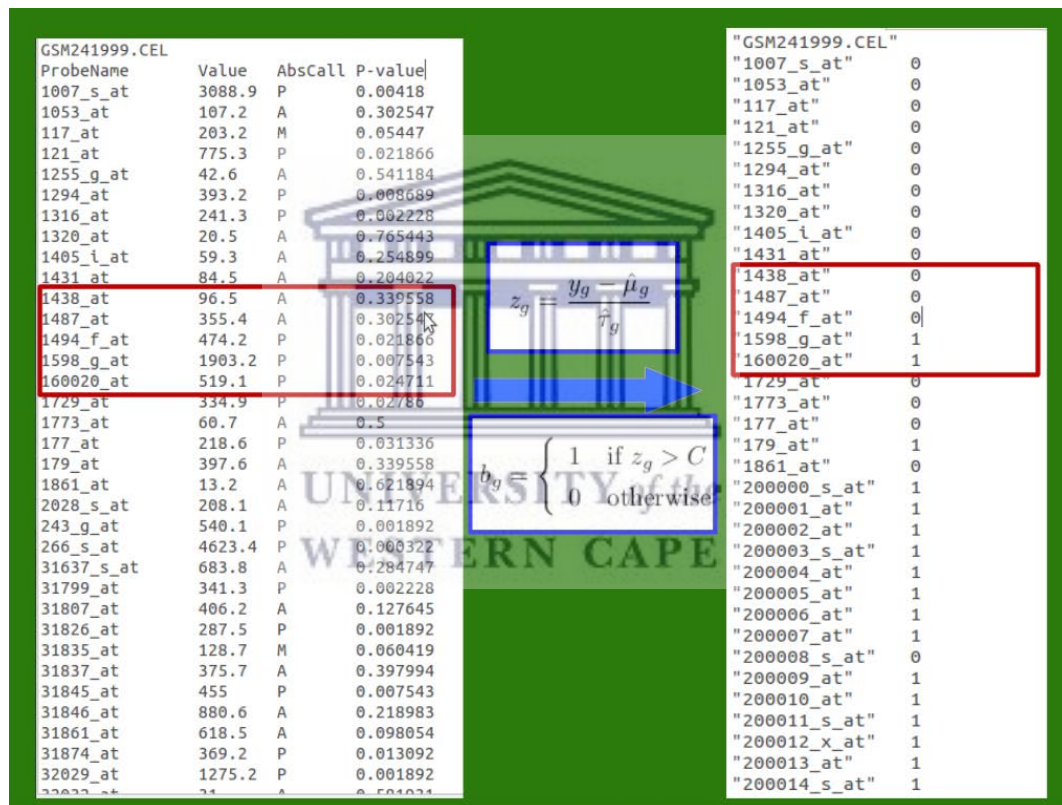


Figure 2.1: Example of expression calls from a micro-array (left side) converted to an absolute call of gene expression (right side) by use of the Gene Expression Barcode algorithm.

### 2.2.3 Feature selection

Feature selection was the first step in the protocol towards finding informative probes that reliably discriminate normal breast tissue from tumour (malignant) tissues. Using the GExB algorithm, along with a filter-feature selection approach,



the discovery of informative probes was a two-step method.

#### Phase I: Signature discovery

To ascertain if the approach validity for further exploration, 30 samples per tissue subtype, 60 in total were used as the training set. Criteria for filtering included:

- 1) 90% stable expression of a gene/probe within each tissue type (1 or 0) and;
- 2) Differential expression between the two tissue types, e.g. “1” in class A and “0” in class B.

60 unseen samples for every subtype within the two classes was used to validate whether the features selected were informative.

#### Phase II: Method Optimization

Once the results from Phase I proved the GExB-FS method capable of producing an informative probe set that could accurately discriminate between normal and tumour breast tissue samples, the next step was to optimize the informative probe set with a larger training data set. Feature selection required the criteria of the filter criteria to be adjusted. The training set now included 100 samples, 50 samples per tissue subtype, and 120 unseen test samples for validation of the two-class classifier. The main differences included:

- 1) 85% stability of a probe being absent or present in a tissue subtype;
- 2) Differential expression between subtypes; present (“1”) in class and absent (“0”) in class B or vice versa.

When the samples numbers were increased for the training set, the 90% stability parameter applied in Phase I proved restrictive, and too few informative probes were produced. Thus, the within-class stability was lowered to allow for slight variability in expression of genes/probes, yet still yielding a small and informative probe set capable of discriminating between the tissue subtypes. Although the second criteria remained the same with regard to differential expression between the subtypes, due to the variability of probe presence/absence permitted by the 85% stability parameter, the new features discovered would add a new dimension to how informative the probes selected would be by:



- 1) Probe “expressed” in A and “not-expressed” in B; or
- 2) Probe “expressed” in B and “not-expressed” in A; or
- 3) Probe “expressed” in A and “unstable” in B, and vice versa

#### **2.2.4 Machine learning classifier based evaluation of the signatures**

Machine learning algorithms were used to evaluate the ability of the features selected to successfully discriminate between healthy/normal and tumour/malignant breast tissues.

##### **2.2.4.1 K-means and Hierarchical clustering**

K-means clustering was employed as an initial unsupervised machine learning algorithm (performed with R, and visualized in RStudio) to assess if the feature selected to separate the tissue types into two clusters successfully. The algorithm was run at 1000 iterations for both the preliminary and final training sets. Hierarchical clustering (performed with R, and visualized in RStudio) was employed consequently to visualize how the samples were classified based on their relation to each other, i.e. how similar the barcodes of each sample were to one another, and if based upon these similarities within a class, could be separated from another class. This part of classifier design was part of the initial validation of the feature selection paired GExB protocol.

##### **2.2.4.2 Support Vector Machines (SVM)**

Subsequent to K-means clustering, the robustness of the features selected was further evaluated using a more sophisticated machine learning algorithm, in this case SVM. The *e1071* R package, which contains libraries for support vector machines (libsvm), was implemented. The training sets were used to train the machine to recognise a sample based on the pattern of probe absence or presence within a tissue subtype. Validation of the predictive capacity of the features selected was completed with unseen samples. Additionally, samples of a malignant breast tissue subtype not included in the training sets was also tested on the classifier, as a very difficult test case not usually performed in such research.

## **2.3 Results**

### **2.3.1 Phase I: Preliminary Method Design**

#### **2.3.1.1 Gene Expression Barcode-Feature Selection paired method (GExB-FS)**

The filter approach to selecting features using the barcode processed samples produced 64 informative probes. The 90% stability parameter was strict enough to rule out excessive variability between samples of the same tissue type. Differential expression analysis proved that there are definite differences of gene expression in diseased tissue compared to healthy tissue. The filter applied minimised the data significantly – from more than 22000 probes to just 64 informative probes. Reducing the high dimensionality of the data was achieved as less than 1% of the original data was used to discriminate between breast tissue subtypes.

#### **2.3.1.2 Machine Learning: K-means, Hierarchical clustering, SVM**

The dendrogram in Figure 2.2 shows a clear separation of Normal from Tumour breast tissue using the 64 informative probes. The two distinct branches within the dendrogram illustrate the robustness of the feature selection method applied alongside the GExB protocol in identifying probe signatures that can discriminate between tissue types.

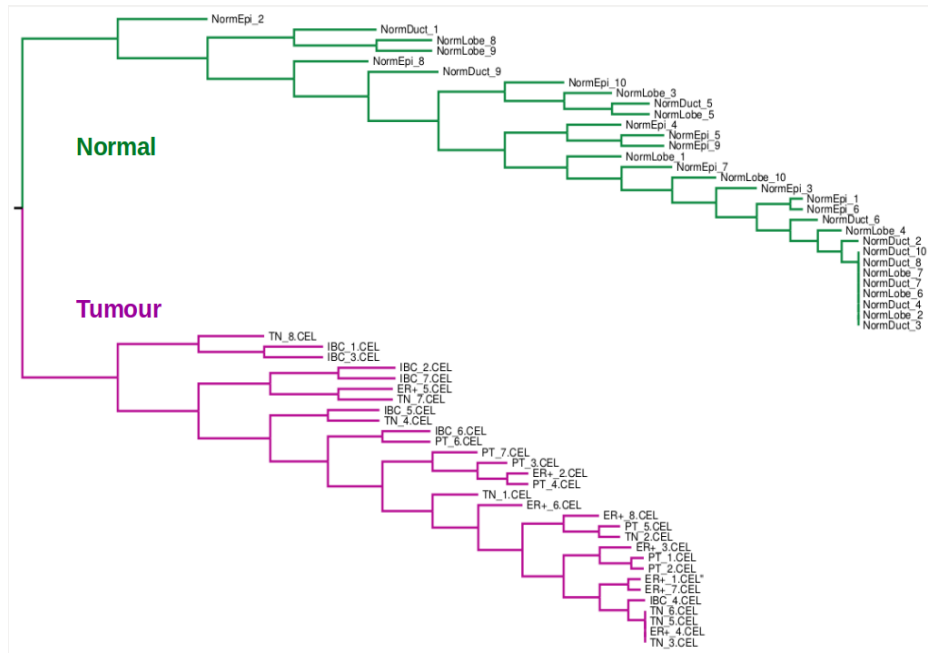


Figure 2. 1: Hierarchical clustering of Training Data Set using 64 informative probe set

Validation of the two-class classifier was confirmed when unseen samples were tested alongside the training data. Table 2.2 shows that both K-means clustering and SVM had a high classification accuracy of 95% and 100% respectively. Most notable was the unseen subtype of breast cancer, Her2-positive, which despite not being part of the training set, classified 100% accurately with the rest of the breast cancer samples.

Table 2. 2: Validation of Preliminary Two-class Classifier

Tissue Type	No. samples tested	Classification accuracy (K-means)	Classification accuracy (with SVM)
Normal	10	100%	100%
Primary Tumour	10	100%	100%
Estrogen Receptor Positive	10	100%	100%
Triple Negative	10	80%	100%
Inflammatory Breast Cancer	10	90%	100%
Her2-positive	10	100%	100%
<b>Total:</b>	<b>60</b>	<b>95%</b>	<b>100%</b>

### 2.3.2 Phase II: Method Optimization

#### 2.3.2.1 Gene Expression Barcode-Feature Selection paired method (GExB-FS)

From the results obtained in the preliminary phase, the classifier was further developed to ensure true validity and to assess the generic nature of the feature selection GExB paired method. However, when samples numbers were increased within the training set, the initial 90% stability (of gene expressed/unexpressed) parameter became restrictive. Too few informative probes were rendered to clearly demarcate tissue subtypes.

When the stability parameter was lowered to 85%, 85 informative probes remained after filtering. The lowered criteria did not compromise on the stable absence or presence of expression of a gene, but did permit slight variability of expression to be included. The new parameters brought about a new dimension to the features selected. An additional criterion was introduced to feature selection during method optimization. The differential expression of probes was no longer limited to the scenario of “on” in class A and “off” in class B, but allowed for a stable-but-varied expression measure to be introduced. Within a larger dataset, tumour heterogeneity would factor in, due to tumour stages and grading, and

molecular subtype. Lowering the stability criteria to 85% allowed probes which had a slight variance (15%) in expression due to the mixed nature of the samples within this study, to still be considered as the general up- or down-regulation of that gene. Probes which were expressed or not expressed less than 90% of the time were previously excluded in Phase I, lowering the stability cut-off parameter allowed a more informative probe set to be produced during Phase II.

### 2.3.2.2 Machine Learning: K-means, Hierarchical clustering, SVM

Figure 2.3 showed that when hierarchical clustering was applied to the data subsequent to K-means clustering, a clear separation of tissue types was clear. The dendrogram illustrates that the new parameters gave similarly high-accuracy results to that of the preliminary phase.

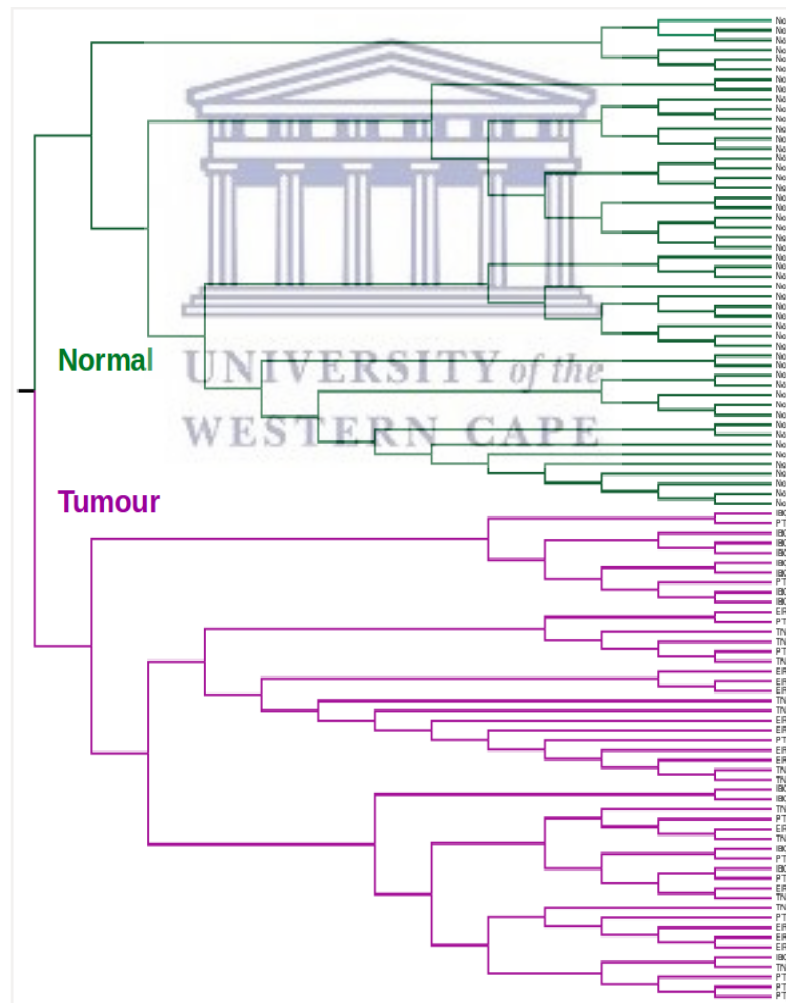


Figure 2. 2: Hierarchical clustering of Training Data Set using 85 informative probe set

Unseen data confirmed that the selected probe signature was robust and informative enough to enable exceptionally accurate classifications. The accuracy of the two-class classifier with K-means clustering improved from 95%, in the preliminary phase, to 100% (Table 2.3). Both K-means and SVM classified unseen microarray samples 100% accurately. The Her2-positive breast cancer samples, not initially part of the training set, classified 100% accurately again.

The new minimised 85 informative probe set thus proved to improve the accuracy of the two-class classifier with regard to both K-means and SVM classification. The improved accuracy could be attributed to the optimised probe set being more informative as it allowed for previously excluded probes which were not always present or absent in 90% of samples, but were discriminative between the two breast tissue subtypes.

Table 2. 3: Validation of Optimized Two-class Classifier

Tissue Type	Classification accuracy (K-means) <i>n</i> = 10	Classification accuracy (with SVM) <i>n</i> = 20
Normal	100%	100%
Primary Tumour	100%	100%
Estrogen Receptor Positive	100%	100%
Triple Negative	100%	100%
Inflammatory Breast Cancer	100%	100%
Her2-positive	100%	100%
<b>Total:</b>	<b>100%</b>	<b>100%</b>

## 2.4 Discussion

**GExB-FS method showed significant promise for classification with a small dataset.**

The results yielded in the preliminary phase proved that the Gene Expression Barcode (GExB) shows promise in classification, even when using a small training dataset. The binary expression measures simplified feature selection of informative probes that could reliably separate the healthy/normal breast cancer



samples from the malignant samples. When implementing a simple filter based purely on the stability of a probe's presence, a 1 or a 0 as allocated by the GExB algorithm, the features found do not require further analyses. This is due to biological and pathological relevance of the genes linked to the probes not being taken into account to avoid assumptions that may prematurely discard relevant features. The criteria for filtering raw microarray data to discover an informative feature set by minimisation of probes thus did not include biological measures; and the association of a probes to genes involved in disease, cancer, known biological pathways were ignored. Instead, the filter criteria purely selected for parameters related to stable expression and differential expression. In so doing, the informative probe set may include genes not yet associated with breast cancer, cancer, apoptosis or any of the known malignancy pathways. Thus, unknown genes may also be included in the informative feature set which would ordinarily be excluded. While not the aim of this study, these genes could be further explored as potentially being involved in tumorigenesis.

#### **Method optimization justified by preliminary phase results**

The credibility of the GExB-FS method applied to a machine learning training set has been proven by the results rendered from validation testing with both SVM and K-means algorithms. The SVM algorithm, originally designed to solve binary classification problems outperformed the K-means clustering algorithm trained with the same data and tested with the same validation set. The true measure of the protocol design was in testing the machine-learning algorithms with completely unseen samples in the form of Her2-positive breast cancer tissue, i.e. not part of the training tumour set at all. Surprisingly, both K-means and SVM classified the Her2-positive class as malignant with 100% accuracy, and SVM performed with 100% accuracy for all tissue types.

The compelling results of the preliminary phase gave justification towards further developing and optimizing the discovery of a barcode signature that could classify a tissue as healthy or malignant with a larger data set. Research has shown that larger datasets, trimmed with informative feature sets and applied to sophisticated

machine learning algorithms like SVMs produce more robust predictive gene signatures (Domingos & Hulten, 2003; Domingos & Pedro, 2012).

### **Larger training set improves classification**

The positive effect of training a machine-learning algorithm with a larger dataset was illustrated when comparing Tables 2.2 and 2.3. The larger dataset yielded 100% accuracy for both the unsupervised and supervised machine-learning algorithms, proving that the features selected were informative in that they were distinctive to which genes are differently expressed between healthy and tumour tissues.

The improved accuracy of the two-class classifier can be partially attributed to two main differences in the feature-selection phase of the protocol. Firstly, the stability parameters were lowered to 85%, allowing variability within the differentiating features; although stable in “A”, varied stability in “B”, instead of stable in both but differential. This allowed for a more informative probe set to be discovered. Secondly, there are 21 more probes selected as features with the larger training set, and these extra probes found, offer more tissue-discrimination potential. This was expected, as previous insights into machine-learning imply that more data the algorithm has to “learn” from, the easier it is to recognise an instance of similarity (Domingos & Pedro, 2012). An unsupervised machine learning algorithms ability to correctly classify samples and discriminate between different classes is boosted with more samples to train from; thereby finding similar features between samples to form distinct data clusters.

The Her2-positive breast tumour subtype, initially not part of the tumour training set continued to classify with 100% accuracy with both K-means clustering and SVM algorithms. Validation with an unseen sample and tissue subtype demonstrates the discriminative ability of the 85 informative probe set in correctly classifying healthy and malignant tissues. This strongly suggests that there is potential to identify and classify ambiguous breast tissue samples or apparently benign tumours that have as yet unexpressed malignancy potential and would

require higher-priority interventions.

### **SVM accurately discriminates GExB-FS processed data**

The GExB-FS method was shown to be a reliable discriminator between healthy and malignant breast tissues when training data, containing only the discovered features, is used to train an SVM. This is due to the absolute calls made by the GExB algorithm. Training a sophisticated supervised machine-learning algorithm, like SVM, on distinct discriminative features distilled from a large training set, achieves two of the prerequisites for optimal machine-learning classification; more training examples with distinctive features allows the machine, SVM, to make better informed decisions (Akay, 2009). Traditional methods of differential gene expression analysis of microarray are largely impacted by technical variance in datasets, specifically of the same tissue type, in the form of noise and batch effects. Batch effects are caused when samples are processed in different batches, resulting in experimental bias linked to the array and probe fluorescence readings (Scherer, 2009). This has previously been shown in studies finding molecular signatures of breast cancer having similar aims and approaches but yielding different outcomes despite using machine learning algorithms for classification (Ransohoff, 2005). The GExB algorithm addresses these biases as samples are preprocessed with fRMA, and barcoding may be executed on single samples, or for batches. Integration of samples from different platforms, and experiments is possible through comparative gene expression calls in the form of 1's and 0's.

### **GExB-FS method is reproducible**

Data integration is complex due to experiment cohorts utilizing different raw data preprocessing methods; i.e. Robust Microarray Analysis (RMA), Log2 intensities, Affymetrix MAS 5.0 Suite. Improving classification models and deriving gene expression signatures that are robust and accurate, however, relies upon data integration. The 85 informative probe set was developed by integrating raw microarray data from 5 experiment sets and validated with data from 8 experiment sets. To ensure data comparability, the data had to be processed identically (Table 2.1). As the GExB uses Frozen Robust Microarray Analysis (fRMA) to normalize

and pre-process data, this in itself ensures that each sample used it processed identically, regardless of which original experiment it was used in, and the data normalization method initially used. Thus, the GExB calls related to absence and presence are not only reproducible, but also integrative. The SVM can thus be trained on a diversity of samples and be validated by unseen samples, including those not included in the initial training set.

## **2.5 Conclusion**

By applying a novel paired method, the Gene Expression Barcode and Feature Selection (GExB-FS method), data from 13 different Affymetrix experiment sets processed 2 different chip platforms. The result was a set of more than 300 samples integrated to develop a two-class breast tissue classifier. Application of the GExB-FS led to the discovery of a minimised feature set which accurately discriminated between healthy and malignant breast tissue samples. An 85 informative probe set was produced as a signature for breast tissue subtype classification, with 100% accuracy.

The implications for such a reliable signature, is ease of translation into a simple laboratory testing protocol, such as RT-PCR. The small feature set, 85 probes, can also be assayed on a standard 96-well PCR plate, without the expense or complications of designing a new technology. Absolute calls, absence and presence of a probe or gene expressed, are much easier to assess and implement in a laboratory set up. Moreover, the importance of being able to classify a sample as malignant or normal is crucial in identifying cancers, since tissues may look normal according to microscopic and histopathological studies, but may in fact be cancerous (K. Graham et al., 2010). Clear and accurate classification of a tissue is the first step towards an accurate and informative diagnosis.

We have shown in this chapter that the GExB-FS method has the potential for use in developing a multi-class breast cancer classifier. As the method could identify discriminating features between healthy and diseased tissues with as yet unprecedented accuracy, it thus may be able to identify features that separate

multiple subtypes of a disease, which is the primary aim of the next chapter.



UNIVERSITY *of the*  
WESTERN CAPE

## Chapter 3

### A Multi-Class Breast Cancer Classifier for Molecular Subtyping

#### ABSTRACT

**INTRODUCTION:** The Gene Expression Barcode (GExB) method, which converts continuous expression levels into binary calls signifying genes as silenced or expressed, was previously employed as a way to enable integration of data from multiple experiments and across chip platforms for the purposes of machine-learning based classification of tumour samples. In combination with a simplistic feature selection method, a gene signature for the identification of malignant breast tissue samples was discovered in Chapter 2. Following the 100% accuracy of our two-class classifier for identification of healthy and/or malignant breast tissue, we explored whether our GExB + Feature Selection (GExB-FS) approach can be used to develop a multi-class classifier for breast cancer subtyping.

**METHODOLOGY:** We implemented a multi-class feature selection variation and tested it on samples from normal and several subtypes of malignant tumours. The 85-90% stability criteria was adjusted to 80% stable in  $n-1$  subtypes to identify a signature which could accurately classify healthy breast tissue and three molecular subtypes; Estrogen-Positive, Her2-Positive, and Triple Negative. The training set for the optimized multi-class classifier included 200 samples, with 80 samples for validation with  $k$ -Nearest Neighbour ( $k$ NN) and multi-class Support Vector Machines (MC-SVM).

**RESULTS:** The feature-selection filter yielded an expression barcode of 346 probes, which enabled clear separation of malignant breast tumour subtypes and unseen samples from entirely different origin than the training set, and classified with 90% accuracy using simple K-means clustering. Optimized classifier development, with implementation of the 346 –gene signature, classified unseen samples with 96% accuracy (MC-SVM).

**DISCUSSION:** The generated binary calls enabled us to develop a simple yet biologically-relevant feature selection/minimization method that simultaneously addressed the 'curse of dimensionality' and the sparsity of training samples, which are significant problems when using microarray data in machine-learning



applications. The ability of the GExB-FS approach to enable identification of signatures able to discriminate between breast cancer subtypes is illustrated with the high accuracy of MC-SVM classification results. We were able to derive an optimized variation of the feature selection method applied in two-class classification to identify a gene/probe signature capable of reliably classifying molecular breast cancer subtypes. While the 346 probe set can be probably be further trimmed to a much smaller core feature set, which was beyond the scope of this study, it would still be easy to implement the signatures on a mini-array or in a PCR array. This would enable, for example, assessment of the clinical validity in a trial across multiple population groups and of its potential for further development into “real-world” applications.

### **3.1 Introduction**

#### **3.1.1 Breast Cancer and Personalized Medicine**

The accurate classification of breast cancer greatly improves the survival outcomes of the patient. Correct diagnosis and insights into prognosis allow the clinicians to make informed decisions regarding treatment and tumour resection (Olopade, Grushko, Nanda, & Huo, 2008). Diagnostics and prognostics based on the molecular and gene expression profile of breast cancer subtypes translates into personalized cancer treatment. Personalized medicine greatly enhances the survival of the patient as treatments are tailored to the disease case presented (S.-H. Cho, Jeon, & Kim, 2012).

The classification of a breast cancer tumour varies on molecular, pathophysiology and clinical presentation of the disease. The cancer in itself can be either *in situ* (localized) or metastatic (spreading) in nature and can also be a primary, originating in the breast tissue, or secondary to another cancer site. Underlying biology of the tumour includes: tumour size, lymph node involvement & lymphovascular invasion, tumour grade. Molecular status, related to hormone receptor expression is the basis of molecular subtyping of breast cancer, namely, estrogen receptor (ER+), human epidermal growth factor receptor 2 (HER2), and Triple Negative which does not express

estrogen, progesterone or HER2 receptors (Alanko, Heinonen, Scheinin, Tolppanen, & Vihko, 1985; Chia et al., 2012; Kennecke et al., 2010).

The variation in hormone receptor status of breast cancer molecular subtypes indicates a difference in gene expression of proteins (hormone receptors) involved in the pathophysiology of the cancer subtypes. Studies aimed at assessing the expression of hormone receptors, estrogen, progesterone, and human epidermal growth factor (Her2), has led to gene expression profiling of breast cancer molecular subtypes (Kapp et al., 2006; Perou et al., 2000). In a 2008 study, gene expression data obtained either from cDNA or mRNA microarray chips was processed and analysed to ascertain if a pattern of gene expression, a signature exists for a specific breast cancer subtype and can be used a predictive measure for breast cancer diagnostics. Through integrating previously identified subtype signatures, they discovered that subtype prediction and prognosis were linked (Wirapati et al., 2008).

MammaPrint™ and Oncotype DX are two prognostic gene expression signatures which have been implemented in breast cancer diagnosis. MammaPrint™ is a 70-gene signature, developed on Agilent microarray data which classifies a patient as chemotherapy suitable or unsuitable. The 70-gene signature was validated with the MINDACT trial, the signature is able to assess chemotherapy sensitivity with genes associated with disease outcome and distant metastasis within 5 years (Mook, Van't Veer, Rutgers, Piccart-Gebhart, & Cardoso, 2007). The limitations of this signature include that the tumour tested needs to be a stage I or II cancer with no lymph node or metastases involved (Buyse et al., 2006). The Oncotype DX 21-gene signature assesses the prognosis of ER+ and DCIS (ductal carcinoma *in situ*) using a recurrence score on RT-PCR data (Toole et al., 2014). These two clinically implemented gene signatures, which predict the prognostic outcome of a patient with breast cancer and allow clinicians to make a more holistic diagnosis and informed treatment decisions (Marchionni et al., 2008; Sotiriou & Piccart, 2007). The same principle has been applied to other cancers

including prostate and colon cancers (Cruz & Wishart, 2006).

### **3.1.2 Multi-class Classification and predictive modelling**

Due to population genetics and dynamics, epigenetics and different breast cancer stages, validated breast cancer subtypes present with variable gene expression profiles within a particular subtype. Thus tumour subtype classifiers need to be as generically applicable as possible (Burrell, McGranahan, Bartek, & Swanton, 2013). Multi-class cancer classification based on molecular subtypes is vastly complex due to the predominant difference between the subtypes being hormone receptor expression. Genetically, this is based on the differential expression of a small set of genes and can make the discovery of signatures related to differential expression difficult. Feature selection approaches have aimed to solve the classification dilemma by filtering samples in a univariate manner using genes known to be involved in cancer pathophysiology and hormone receptor expression (Statnikov et al., 2005). Machine learning algorithms would then be used to confirm predictive capability of the signatures identified. Although the identification of clinically applicative gene signatures have been successful, these have been shown to be limited in application and population dependent, and are not generically applicable due to lack of data diversity (Creighton et al., 2006).

The analysis of gene expression data from cDNA and mRNA microarray experiments has led to class discovery in cancers (Golub et al., 1999) and subsequent subtype classification of leukaemia and other cancers. A classifier for leukaemia genetic subtype classification, based on classes identified by Golub and colleagues, was developed by applying an intrinsic gene set for feature selection and classification with the *k*-Nearest Neighbours algorithm (Andersson et al., 2007). Advances made in breast cancer subtype identification (Perou et al., 2000) and validation (Sorlie et al., 2003) has led to hierarchical clustering models developed for the classification of breast cancer based on the estrogen receptor status of a tumour (Sorlie et al., 2003).

### **3.1.3 Implementing Frozen Robust Multi-array Analysis (fRMA) and the Gene Expression Barcode (GExB) algorithm for microarray gene expression data**

Microarray data is considered to be highly dimensional (Hira & Gillies, 2015). Transcriptomic data generated for a single sample may include gene expression readings for more than 7000 genes, represented by 11 probes each (McCall et al., 2010b). When applied to gene expression studies to profile a particular tissue type or disease state, the data produced becomes exceedingly voluminous, given that often hundreds of samples are used in comprehensive transcriptomic analyses. Furthermore, expression data is continuous, and requires analysis of relative expression and relative differential measures. Studies based on relativity do not often perform well on other populations and are not easily reproduced (Haibe-Kains, 2010).

The Gene Expression Barcode (GExB) algorithm, which integrates frozen RMA (fRMA) pre-processing of microarray data, provides an easily implementable solution to high-dimensional continuous microarray data. The ability to assign a discrete value, 1 or 0, to infer up- (“on”) or down-regulation (“off”) simplifies differential gene expression analysis for classification of biological samples (McCall et al., 2014).

### **3.1.4 Machine Learning and Feature selection for Breast Cancer**

#### **Classification**

Machine learning algorithms have been extensively used in research studies to develop breast cancer multi-class classifiers and discovery of gene expression signatures (Hu et al., 2006). A comparison of machine learning algorithms has revealed the one-versus-one (OVO) and one-versus-rest (OVR) implementations of multi-class SVM (MC-SVM) to be most efficient and accurate (Saeys et al., 2007). However as robust ML algorithms are, when applied to microarray data, they are still struck by the “curse of dimensionality” (Bolón-Canedo, Sánchez-Marroño, & Alonso-Betanzos,

2016). This implies that the highly-dimensional data is too complex, with too many similarities between biological samples for the ML algorithm to differentiate between, thus hampering the development of robust classifiers.

In order to optimize ML-based classifiers, feature selection is employed. Feature selection may be defined as the process of eliminating non-relevant, redundant, or non-informative features in a data set (Blum & Rivest, 1992). Within microarray studies, this would translate to selecting genes or probes which are capable of differentiating between samples or tissues.

Gene selection filtering, has been shown to improve the accuracy of cancer classification when applied to different machine learning algorithms including Support Vector Machines (SVM), and artificial neural networks (ANN) (Golub et al., 1999). The most discriminating features (informative genes or probes) used when training an ML algorithm, will produce the most accurate classifier (Libbrecht et al., 2017).

### **3.1.5 Aims and Objectives**

Feature selection (FS) and machine learning (ML) have been paired in the development of classifiers for cancer, including breast cancer (Akay, 2009; Lu et al., 2005). As illustrated in Chapter 2, implementing fRMA for data pre-processing and the Gene Expression Barcode (GExB) simplifies both the integration and filtering of data to select for easily identifiable and discriminatory microarray probes. The combining of this approach for feature selection with well-established ML algorithms capable of handling multiple sample categories, holds the potential for the development of a robust multi-class breast cancer subtype classifier.

The multi-class phase of this project therefore aimed to:

- 1) Develop an automated feature selection pipeline for identifying a minimal set of expression features based on Expression Barcoded data from multiple tissue types.

- 2) Evaluate and optimise feature-selection method using simple machine learning classifier; K-means clustering, k-Nearest Neighbour.
- 3) Produce multi-class gene expression signatures for normal breast tissue and carcinoma subtypes.
- 4) Derive a variation of the optimized feature-selection method presented in Chapter 2, with Support Vector Machines (SVM), for breast cancer molecular subtype classification.

## **3.2 Materials and Method**

### **3.2.1 Data Curation**

Raw microarray data for building a multi-class classifier was collected in a similar manner to the two-class classifier (Chapter 2). 320 breast tissue samples were curated from NCBI Gene Expression Omnibus (GEO), varying in experiment sets and Affymetrix microarray platforms. Samples were curated for the four breast tissue subtypes by collecting data from annotated samples, with previous immunohistochemical identification of tissue subtype, or hormone receptor status (Table 3.1).

Most notably, in the Training dataset, the normal breast tissue set was curated from three different experiment sets, and the Her2-positive breast cancer subtype was curated from two different experiment sets. This would ensure diversity of data on two levels; 1) Different microarray assay platforms, and 2) Raw data samples from seven completely independent experiment sets for the Training dataset.



Table 3. 1: Summary of Breast Cancer Samples curated

Dataset	Tissue Type	Number of Samples	GEO Series	GEO Platform (Affymetrix)
<b>Training</b>	Normal (Epithelium, Duct, Lobe)	57	GSE20437; GSE5764	GPL96*; GPL570*
	Triple Negative Breast Tumour	50	GSE25065	GPL96
	Estrogen-Positive Breast Tumour	77	GSE25065	GPL96
	Her2-Positive Breast Tumour	53	GSE37946; GSE42822	GPL96
<b>Test/ Validation</b>	Normal Epithelium	20	GSE9574	GPL96
	Triple Negative Breast Tumour	21	GSE31519	GPL96
	Her2-Positive Breast Tumour	20	GSE22597	GPL96
	Estrogen-Positive Breast Tumour	22	GSE22093; GSE23988	GPL96

\*GPL96 - [HG-U133A] Affymetrix Human Genome U133A Array

\*GPL570 - [HG-U133\_Plus\_2] Affymetrix Human Genome U133 Plus 2.0 Array

### 3.2.2 Gene Expression Barcode (GExB) implementation and data integration

The raw data samples were preprocessed with the fRMA algorithm prior to gene expression barcode generation to ensure comparability of data from different experiment sets and different chip platforms. Absolute calls for gene expression continuous values were computed for each sample in batch form for each breast tissue and breast cancer subtypes. Thereafter, the barcodes generated were merged into an integrated dataset.

### 3.2.3 Feature selection and application to datasets

Feature selection for multi-class classification proved more complex, as expected, since molecular subtypes, Estrogen Positive, Her2-Positive, and Triple Negative are closely related. Identifying differentiating features, i.e. probes or genes that are differently expressed, is challenging due to the three molecular subtypes sharing similar core gene expression profiles.

During the development of the two-class breast cancer classifier, stability parameters of 90% and 85% absence or presence of a probe was used as feature selection criteria, in Phase I and II of classifier development, respectively. Here, the 85-90% stability criterion was too stringent to identify an informative feature set containing probes which were stable (absent or present) and differentially expressed between all the breast cancer subtypes. The parameters implemented during two-class classifier development were therefore adjusted to address the similarities between the molecular breast cancer subtypes.

#### 3.2.3.1 Phase I: Preliminary Phase - Method Development

Using a training set of 120 samples; 30 samples per breast tissue subtype, the following feature extraction criteria were applied:

- 1) 80% expression or non-expression stability in  $n-1$  subtypes;

A probe would have to be either absent or present 80% of the time in at least 3 of the 4 subtypes.

- 2) Differential expression between subtypes.

The 80% stable in  $n-1$  subtypes allowed probes that were absent or present at stable rate in 3 subtypes, but unstable in 1 of the subtypes to be accepted as informative. A probe that was stable in one subtype but unstable in another subtype could be regarded as a feature that distinguishes the two subtypes. Permutations of probe presence, absence, and varied absence/presence allowed an informative probe set to be identified that could separate subtypes from one another. This signature was then applied to filter the barcoded

training and test sample data.

### 3.2.3.2 Phase II: Method Optimization

During Phase II, the training set was increased to include 200 samples, 50 samples per tissue subtype. Feature selection with a larger training set, using the same stability parameters of 80% in  $n-1$  proved challenging due to extreme variability in the data. Thus the predictive performance of the initial 346-gene signature was tested on a larger training set.

During two-class classifier development, a larger training dataset produced a signature which improved classification accuracy from 95% to 100%; the more samples the learning algorithms K-means and SVM had to train on, the more efficiently an unseen sample could be labelled as healthy or malignant correctly. This motivated the application of the multi-class gene signature set to a larger dataset.

By applying the feature set discovered with a small training set to a larger training and validation set we aimed to:

- 1) Train the machine learning algorithms,  $k$ -Nearest Neighbour ( $k$ NN) and Multi-class Support Vector Machine (MC-SVM), with a larger set of data samples which would,
- 2) Provide the learning algorithms with a more heterogeneous gene expression barcode profile for each subtype.

### 3.2.4 Machine Learning classifier evaluation

Three different machine learning methods were used to assess the ability of the gene signature to successfully separate four different breast tissue subtypes. The test dataset included 10 samples per subtype, 40 samples in total, for Phase I and 20 samples per subtype, 80 samples in total, for Phase II.

### 3.2.5 K-means and Hierarchical clustering

K-means and hierarchical clustering (performed with R, and visualized in RStudio) was applied to both the training set and validation set in Phase I only. The unsupervised machine learning algorithms were applied to ascertain an initial indication of how well the signature could cluster samples into their respective known breast tissue subtypes. While the K-means clustering algorithm was initially applied to the larger set; however known constraints within the algorithm (Raykov, Boukouvalas, Baig, & Little, 2016), proved it to be unsuitable to accurately cluster the four breast tissue subtypes.

### 3.2.6 *k*-Nearest Neighbour classification

*k*-Nearest Neighbour (*k*NN) clustering was introduced as the initial machine learning algorithm to test the optimized multi-class classifier with the larger training set of 200 samples. *k*NN is a supervised instance-based learning algorithm, which places a sample closest to other samples that are similar based on the specified-identity features (Lopez de Mantaras & Armengol, 1998). Default *k*NN algorithms employ 5*k*NN – whereby Euclidean distance is used to measure the relation of a single sample to five other similar samples, and consequently place them in the same class (Coomans & Massart, 1982). Leave-Out-One Cross Validation (LOOCV) can be paired with *k*NN classification, where with each training iteration of algorithm, one sample is left out, which in turn verifies the correct allocation of a sample to its correct class (Saligan, Fernández-Martínez, de Andrés-Galiana, & Sonis, 2014).

The chosen classification parameters were that data be separated into four clusters, where each sample was related to five neighbouring samples, i.e. data instances within that specific cluster. LOOCV was performed on the training set, to ascertain if the signature was robust enough to separate the four tissue subtypes with reasonable accuracy.

### **3.2.7 Multi-class Support Vector Machines (SVM) classification**

Although SVM was designed to solve binary classification problems, multi-class SVM (MC-SVM) derivatives of the algorithm exist. The two most commonly used algorithms being One-versus-One (OVO) and One-versus-Rest (OVR) MC-SVM. OVO-SVM recognises each class separately from one another, and thus  $k > 2$ . OVR-SVM requires multiple iterations of a binary-SVM, with each class versus all other classes in various combinations.

Phase I used the OVR implementation of the algorithm, to ascertain if an SVM trained on the signature classified unseen samples correctly. OVR required four iterations, as each subtype had to classify against the three other subtypes, i.e. four combinations of the binary SVM classifier where subtype A versus BCD, B versus ACD, C versus ABD and, D versus ABC. The OVR classifier used default parameters and tested all of the learning kernels: linear, radial basis function (RBF), polynomial and sigmoid.

LIBSVM (Library for Support Vector Machines) (Chang & Lin, 2011) has a built in OVO MC-SVM module which simplified the implementation of the algorithm with the larger filtered dataset of 200 samples and validation with 80 samples. The polynomial kernel was chosen as it best fit the variability of the data and had previously performed well in Phase I.

## **3.3 Results**

### **3.3.1 Phase I: Preliminary Method Development**

346 informative probes that could distinguish between closely related molecular breast cancer subtypes and healthy breast tissue, and classify unseen samples correctly were discovered.

The dendrogram in Figure 3.1 shows the clear separation of Normal and Tumour breast cancer samples, using the 346-gene signature. Within the Tumour branch, there are three nested clusters; Estrogen-Positive and Her2-

Positive clusters branching from the Triple Negative clusters, as though two nested clusters are found within a larger cluster. This was expected as the three breast cancer tumour subtypes known to be molecularly similar.

The results illustrated below reveal that signatures derived from microarray data transformed using the GExB-FS method to be reliable and accurate in multi-class classification, when filtered data is applied to unsupervised machine learning algorithms, like hierarchical clustering. Disease subtype tissues that have slight differences are successfully grouped in their own clusters.

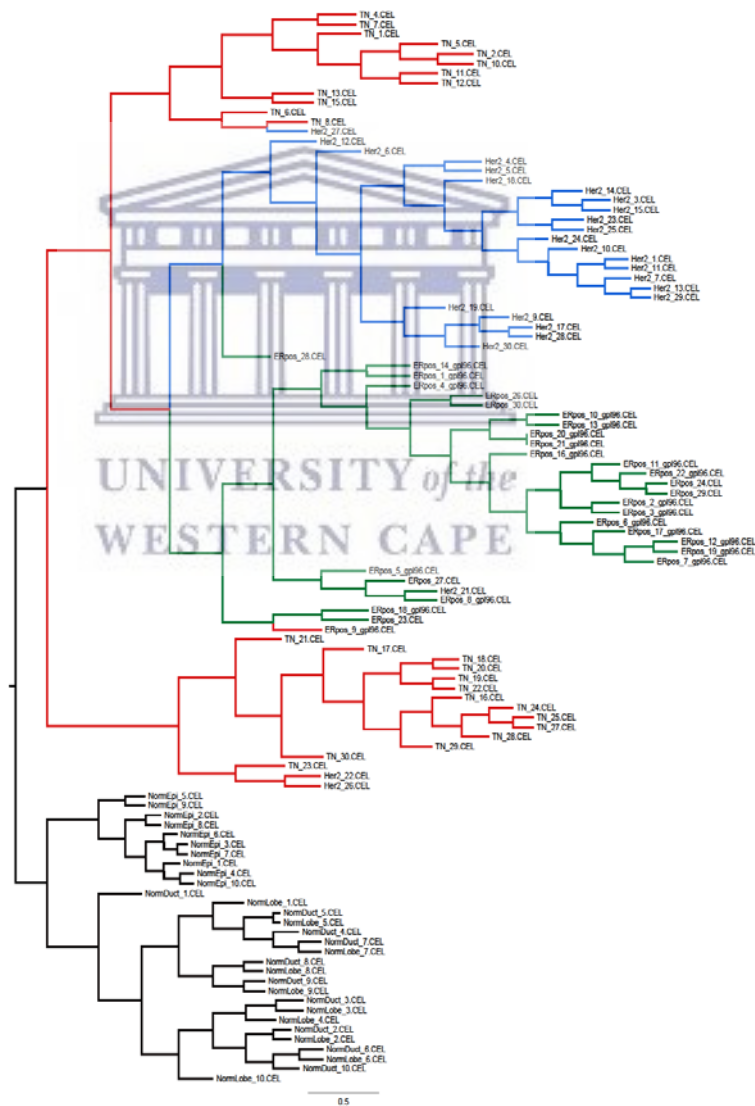


Figure 3. 1. : Hierarchical clustering of Training Data Set using 346-gene signature



During Phase I, a small set of samples was used to validate the ability of the 346-probe set to separate tissues of four different subtypes. K-means clustering performed well under a multi-class scenario, with 90% accuracy. SVM improved the accuracy of classification, with 95% accurate classification of validation data. The two subtypes, Estrogen- and Her2-Positive breast tumours, both had improved classification accuracy from 80% to 90%, with the implementation of the OVR MC-SVM algorithm (Table 3.2).

Table 3. 2: Validation of Preliminary Multi-class Classifier

Tissue Type	No. samples tested	Classification accuracy (K-means clustering)	Classification accuracy (with SVM)
Normal	10	100%	100%
Estrogen Receptor Positive	10	80%	90%
Triple Negative	10	100%	100%
Her2-Positive	10	80%	90%
<b>Total:</b>	<b>40</b>	<b>90%</b>	<b>95%</b>

### 3.3.2 Phase II: Method Optimization

As K-means clustering rendered poor clustering results with the larger training set, Leave-Out-One Cross Validation of the  $k$ NN algorithm was applied to ascertain if the informative feature set was still able to accurately separate samples into their four respective groups. LOOCV- $k$ NN classified training data with 87% accuracy.

Table 3. 3: *kNN Leave-Out-One Cross-Validation classification of Training dataset*

<b>Tissue Type</b>	<b>No. samples tested</b>	<b>Classification accuracy</b>
Normal	50	100%
Estrogen Receptor Positive	50	78%
Triple Negative	50	90%
Her2-Positive	50	80%
<b>Total:</b>	<b>200</b>	<b>87%</b>

In Table 3.4, the number of unseen samples used to validate the informative probe set as features for a multi-class classifier, were double in comparison to Table 3.2. The LIBSVM implementation OVO MC-SVM was trained on 200 samples filtered with the 346-gene signature. Multi-class SVM improved classification of unseen samples for the 4 breast cancer and tissue subtypes from 95% during the preliminary phase, to 96.25%.

Table 3. 4: *One-versus-one Multi-Class SVM classification of Unseen Validation dataset*

<b>Tissue Type</b>	<b>No. samples tested</b>	<b>Classification accuracy</b>
Normal	20	100%
Estrogen Receptor Positive	20	90%
Triple Negative	20	95%
Her2-Positive	20	100%
<b>Total:</b>	<b>80</b>	<b>96.25%</b>

### **3.4 Discussion**

#### **Constraints of Multi-Class Classifier Development**

During the preliminary phase of the multi-class classifier development, discovering an informative probe set capable of clearly distinguishing between different breast cancer molecular subtypes (Triple Negative, Her2-Positive and Estrogen-Positive) became a challenge. Although implementation of the GExB algorithm simplified the discovery of differentially expressed genes through generating 0's and 1's as an absolute measure for gene expression, the barcode expression profile of the three breast cancer subtypes remained largely similar.

The 85% stability parameter (absent or present 85% of the time) implemented with two-class classification, on the basis of differential expression, was too restrictive to identify a large enough probe set capable of discriminating between four classes (three subtypes and normal) of breast tissue. Taking this into account, the parameters for feature-selection of barcoded samples were relaxed. 80% stable in  $n-1$  subtypes satisfied two of the initial classification criteria of the two-class classifier, while it introduced an additional distinguishing criterion. Essentially, a probe was allowed to have varied absence or presence, below 80%, if it was absent or present 80% of the time in the other three subtypes/tissue types.

#### **GExB-FS method shows promise for developing a multi-class breast cancer classifier**

The results depicted in Figure 3.1 and Table 3.2 show that the GExB-FS approach, with a relaxed filtering criteria, identified 346 informative probes capable of discriminating four different breast tissues with 90-95% accuracy, despite biological and pathological relevance of the genes not being taken into account. This was achieved with a training set of 120 samples, 30 per subtype, and performance measured on a small validation set of 40 samples. The results obtained during this preliminary phase of classifier development, suggested that increasing the training data set, may yield either a more

discriminatory probe set, or improve the multi-class classifier accuracy. A larger dataset was expected to provide the learning algorithms with a heterogeneous portrait of the breast cancer subtypes to create a generic pattern against which to classify an unseen sample.

Application of the feature-selection criteria to a larger dataset ( $n = 200$ ), yielded an probe set which was not able to discriminate between the four classes as accurately as the initial 346 probe set, as only 79 informative probes were generated. This may have been due to expression profile similarities between the molecular subtypes, as although breast cancer tumours are highly heterogeneous, the intrinsic gene set which separates subtypes is still less than 500 genes (Perou et al., 2000). Similarly, when adjusting the criteria to 80% stable in  $n-2$  subtypes, thus still permitting variation of stable gene expression in at least one of the four subtypes, 2518 informative probes were identified. Although eight times the number of probes initially identified, the now larger feature set was too large for the machine learning algorithms to train effectively, and too unstable to allow accurate separation of different breast tissues. If a classification model's features are manipulated too much, the classifier becomes over fitted. Conversely, if the feature set is too large or too variable, the classifier is not discriminative enough to identify new samples accurately (Golub et al., 1999; Sima & Dougherty, 2006).

### **Larger training set improves classification**

The high accuracy of the multi-class classifier in the preliminary phase (Table 3.2) motivated further optimization with a larger training and validation dataset. Training a classifier on a larger dataset with limited discriminating features has been shown to improve the accuracy and reliability of a classifier (Yu & Liu, 2004). Indicated by Tables 3.3 and 3.4, the larger dataset, trained on the initial 346 informative probes generated in the preliminary phase, yielded a 87% and 96% accurate classification of validation samples by LOOCV kNN of the Training set and MC-SVM,

respectively. The signature identified in the preliminary phase thus proved to be robustly discriminative of multi-class data, in congruence with the theory that greater training data numbers yield better classification results on proven and validated discriminatory features (Fan & Fan, 2008).

We hypothesise that finding informative feature sets that differentiate between two subtypes at a time, and then combining signatures non-redundantly may be the key to discovering more informative features that offer more information regarding diagnostic criteria such as tumour staging and prognosis.

### **Robust Gene Signature discovered with GExB-FS Approach**

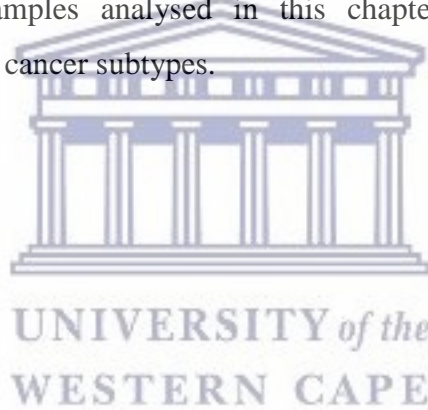
The particular feature selection filter technique was applied as a model-free method. Feature selection was performed completely independent of the machine learning algorithms and ignored feature dependencies. Although considered a disadvantage of univariate filter models (Saeys et al., 2007), the approach was beneficial when developing a multi-class classifier. The model-free approach has been considered attractive in microarray based gene expression profiling, as it is less stringent than making expression-distribution assumptions in complex biological scenarios where the underlying physiology is not yet fully understood (Troyanskaya, Garber, Brown, Botstein, & Altman, 2002).

The GExB-FS approach was able to perform with 96.25% accuracy with OVO MC-SVM classification. The result was beyond expectation due to the complexity and known difficulty of solving multi-class cancer scenarios where subtypes are so closely related. A previous study on multi-category classification methods for gene expression-based cancer diagnosis that used MC-SVM algorithms could not classify cancer samples above 95% accuracy (Statnikov et al., 2005). This proves that the GExB component of signature discovery has a positive effect on identifying features that strongly discriminate between tissue subtypes, when filtered data is classified with

MC-SVMs.

### **3.5 Conclusion**

The feature selection and classifier development approach we employed, the GExB-FS method, and the validation of the discovered gene signature, with unseen microarray gene expression data has proven the 346-gene signature effective and accurate in classifying breast tissue subtypes. The probes would be easily translated into a laboratory test, as standard RT-PCR 384 well plates would be able to replicate and ascertain the absence and presence of transcripts. Future investigations into the overlap of the two-class and multi-class gene signatures may allow the development a single signature capable of not only identifying the malignant status of a tumour, but also its the molecular subtype. This is already indicated by our ability to separate healthy breast tissue samples analysed in this chapter from the three known molecular breast cancer subtypes.





## **Chapter 4**

### **A Multi-Class Classifier for RNA-Seq Breast Cancer Data**

#### **ABSTRACT**

#### **INTRODUCTION:**

The advent of Next-Generation Sequencing (NGS) technologies has endowed cancer researchers with the ability to delve deeper into the genomics and transcriptomics governing cancer pathophysiology. RNA sequencing (RNA-Seq) is one such NGS platform which sequences a partial or complete transcriptome of a single cell or clusters of cells (tissue) and reveals the abundance and presence of absence of transcripts within a specific physiological state. Given the complex nature of breast cancer, with various molecular presentations of the disease, deep transcriptomic analysis allows for applications in gene expression studies, biomarker discovery, gene fusion and gene insertion-deletion events with potential to guide treatment and diagnosis.

Projects such as The Cancer Genome Atlas and The Genotype-Tissue Expression project, have aimed to use RNA-Seq to comprehensively examine gene expression in different healthy and cancerous human tissues with their data being publicly available. Potential gene expression signatures have also emerged from these studies for breast, prostate, colorectal, ovarian and endometrial cancers, amongst others.

The Gene Expression Barcode (GExB) algorithm introduced a sophisticated method for gene expression studies in microarray data, by assigning 1's and 0's as absent-present calls for genes in a sample. As demonstrated in Chapters 2 and 3, application of this algorithm enables successful integration of breast cancer microarray data originating from different studies and development of disease state and subtype classifiers when used alongside machine learning algorithms. However, no equivalent of the barcoding method exists for RNA-Seq data as yet. We thus aimed to develop a statistics-driven GExB-like method for RNA-Seq and to apply it in the discovery of a multi-class gene signature for classifying normal, normal-adjacent-tumour, and primary breast tumour samples from different public data repositories.

## METHODOLOGY:

We used a two-prong approach, which included first discovering highly differentially expressed genes (DEG's) in TCGA normal-adjacent-tumour (NAT) and primary tumour breast samples, as well as between normal samples from women that did not have breast cancer in GTEx and TCGA tumour samples. Secondly, in parallel, we developed and applied “z-score-to-barcode” method to raw RNA-Seq gene count data to i) calculate relative gene expression levels for transcriptomic data, and ii) convert these values to 1's and 0's in a GExB fashion. Following the establishment of the methods for RNA-Seq informative gene set discovery and barcoding of gene count data, the study was extended to compare top DEG's from TCGA and integrated dataset analysis, to determine whether GTEx normal tissues could be classified as distinct from TCGA tumour and NAT samples.

## RESULTS:

Barcoding of RNA-Seq data and application of a discovered expression signature enabled unseen samples in a test set to be labelled as the correct tissue type with above 95% accuracy when K-means clustering, Hierarchical Clustering and Support Vector Machines were employed for classification. In addition, we found that indicated that normal breast tissue from healthy women had a pattern of gene expression that is distinct from NAT tissues from breast cancer patients.

## DISCUSSION:

A potentially novel and robust method for barcoding and classification of breast cancer RNA-Seq samples was established as demonstrated by the accuracy of the two-class and multi-class classifiers built using our unique approach. This method also has the potential to be applied to other cancers and diseases. The detection of a unique transcriptomic portrait of NAT tissues suggest that the similarities between NAT and tumour tissues and the differences between NAT and healthy-normal tissues need to be taken into account during biomarker/diagnostic gene signature research. Possible insights into tumorigenesis and cancer metastases, along with robust discriminatory genes may be obscured or not revealed at all if NAT tissues are not considered a tissue subtype during cancer gene expression studies.

## **4.1 Introduction**

### **4.1.1 RNA-Sequencing and Cancer**

Next-generation sequencing (NGS) has allowed further exploration of the genomic and transcriptomic portraits of cells and tissues through high-throughput DNA and RNA sequencing technologies (Ng & Kirkness, 2010). DNA-Seq uncovers genomic aberrations via detection of genetic lesions, while RNA-Seq reveals the downstream consequences of these lesions, i.e. mutations, nucleotide insertions and deletions, exon-skipping and gene fusion (M. Wan, Wang, Gao, & Sklar, 2014). RNA-Seq achieves this by taking a snapshot of a cell or tissue's transcriptome for a given physiological state, and is capable of capturing all RNA's of the cell.

Transcriptomics allows for elucidation of the cellular state at the transcript level, and therefore the genes which are expressed or not, in a physiological condition. This has provided insight into those genes' involvement in a particular disease state (George, Ashokachandran, Paul, & Girijadevi, 2017), as it allows for the analysis of a continuously changing cellular environment.

In recent years, optimization and reduction in costs of RNA sequencing have provided researchers a deeper understanding of a cell or tissue's mechanisms of gene expression and genetics underlying diseases (A. Desai & Jere, 2012). The application of NGS to various cancers, including breast cancer (Koboldt et al., 2012), lung adenocarcinomas (Shukla et al., 2017)), and colorectal cancer (Cancer Genome Atlas Network, 2012) , have revealed gene expression signatures not previously detected with array technologies. The verification of these signatures can be achieved through targeted RNA sequencing of the relevant transcripts (Tewhey et al., 2009).

### **4.1.2 Breast Cancer Transcriptomics**

Breast cancer is a complex disease, with a multiplicity of clinical presentations differing in their histopathological and molecular portraits. The heterogeneity

of breast cancer, or cancer as a whole, can be attributed to differences in cell type origin, gene mutations, gene isoform expressed, indels, SNPs, or hormone receptors expressed (Turashvili & Brogi, 2017).

Predictive and prognostic gene expression signatures have arisen and been applied clinically, e.g. Mammaprint® and Oncotype DX®, however, they were designed using microarrays and the resultant laboratory assays report expression at “gene level”. The emergence of NGS, and its application in cancer research, now affords researchers the opportunity to look beyond these established gene signatures. Transcriptome profiling of the cancer cell can sequence deeper to the isoform level (A. N. Desai & Jere, 2015), as well as detect transcripts of mRNA's, non-coding RNA's , differences in gene isoform expression, lending distinct advantages in understanding breast cancer progression, metastasis, potentially leading to better and more accurate classification and diagnosis.

#### **4.1.3 Public Transcriptomic Data**

Due to the biomedical advantages of RNA sequencing, coupled with the biological complex landscape of cancer, and understanding it's genomic and transcriptomic position in relation to healthy tissues, several large projects and consortia arose to address these needs by producing data for analysis by the scientific community.

#### **4.1.4 The Cancer Genome Atlas (TCGA)**

In order to accelerate an extensive understanding of the cancer genome, The Cancer Genome Atlas (TCGA) was launched by the National Institute of Health (NIH) in 2005, with the International Cancer Genome Consortium (ICGC) launched in 2008 (Chin, Andersen, & Futreal, 2011). TCGA is a vast catalogue, containing thousands of RNA-Seq samples, with more than 30 malignant tumour types as well as normal tissue control samples. The TCGA network has executed large-scale studies on breast cancer, lung adenocarcinoma, glioblastoma, colorectal cancer, ovarian and endometrial

cancer, and pan-cancer studies to fully elucidate the comprehensive molecular portraits of these cancers (Tomczak, Czerwińska, & Wiznerowicz, 2015).

The available data types include RNA-Seq, microRNA sequencing (miRNAseq), DNA-Seq, SNP-based platforms, array-based DNA methylation sequencing, amongst others.

#### **4.1.5 The Genotype-Tissue Expression (GTEx) project**

The Genotype-Tissue Expression project was launched by the NIH in 2010. The aim of the project was to establish a database that would allow the study of differences in gene expression in human tissues (Lonsdale et al., 2013). Since its inception, the project has sequenced more than 10 000 samples (spanning 53 different tissues) from 714 donors. This in-depth analysis of multi-tissue transcriptomes has allowed molecular portraits for healthy or normal tissues to emerge, which can now aid the assessment of diseased tissues (Ardlie et al., 2015; Keen & Moore, 2015).

#### **4.1.6 Research Aims and Objectives**

The Gene Expression Barcode (GExB) for microarray data proved robust in developing a multi-class breast cancer subtype classifier. Unfortunately, since the release of the GExB version 3.0 in 2014 (McCall et al., 2014), a GExB algorithm for RNA-Seq data has not yet been developed. Owing to the accuracy of the microarray breast cancer classifiers developed using GExB for feature selection and data transformation; we aimed to develop a similarly applicable method for RNA-Seq breast cancer data.

Aided with the differential gene expression package for RNA-Seq data, *edgeR*, we aimed to develop a simplified method for discovering a possible gene expression signature for breast cancer classification. In order to achieve this, the following objectives have been established:

- 1) Discover method to convert gene counts in RNA-Seq data to absolute calls of expression, i.e. 1's and 0's, and therefore creating a “barcoding” method for NGS data
- 2) Applying the method of barcoding to RNA-Seq cancer data from The Cancer Genome Atlas (TCGA)
- 3) Develop a two-class classifier for TCGA normal and tumour samples with feature selection based on best differentially expressed genes
- 4) Integrate RNA-Seq normal breast tissue samples, from GTEx project to discover a signature for multi-class classification capable of distinguishing between normal, normal-from-cancer-patient, and primary tumour samples.

## **4.2 Materials and Methods**

### **4.2.1 Data Curation**

RNA-Seq data (raw gene counts) from breast tissue samples from The Cancer Genome Atlas (TCGA) repository was manually curated. The dataset shown in Table 4.1 is in whole or part based upon data generated by the TCGA Research Network (“The Cancer Genome Atlas Program - National Cancer Institute,” n.d.). The individual samples were curated using their assigned TCGA sample ID’s from the Genomic Data Commons Data Portal (“GDC Data Portal,” n.d.). Individual sample ID’s may be viewed in Appendix II, Table 7.3. Both normal-adjacent-tumour breast and primary breast tumour samples were curated. The correct molecular subtypes of tumour samples according PAM50 classification were obtained from supplementary materials of a TCGA research paper (The Cancer Genome Atlas Network, 2012).

For the purpose of discovering a robust set of differentially expressed genes in breast cancer, 40 paired normal/primary tumour samples were collected as a priority, but unpaired samples were also collected for downstream analysis. In Table 4.1, the tumour sample set in both paired and unpaired analysis consisted of the triple negative, estrogen-receptor positive and Her2-receptor positive molecular subtypes. The paired dataset was analyzed independently of the unpaired dataset to avoid bias in classification models.



Normal breast tissue RNA-Seq samples (raw gene counts) from healthy individuals were curated from the GTEx project's data portal ("The Genotype-Tissue Expression (GTEx) project Data Portal," n.d.) on 07/13/2018, to be integrated with the TCGA dataset. These samples were selected from version 7 of GTEx publicly available gene count data; which includes a total of 11688 samples which cover 53 different tissue types. The same dataset, with patient information can be obtained from dbGaP ("dbGaP | phs000424.v7.p2 | Common Fund (CF) Genotype-Tissue Expression Project (GTEx)," n.d.). Individual sample ID's may be viewed in Appendix II, Table 7.2.

*Table 4. 1: Summary of breast tissue samples curated from The Cancer Genome Atlas Data Repository*

<b>Dataset</b>	<b>Tissue Type</b>	<b>Data type</b>	<b>Data Repository/Project</b>
<b><i>Paired samples</i></b>	Normal-Adjacent-Tumour Breast Tissue	Raw RNA-Seq counts	TCGA*
	Triple Negative Primary Tumour		
	Her2-Positive Primary Tumour		
	Estrogen-Positive Primary Tumour		
<b><i>Unpaired samples</i></b>	Normal-Adjacent-Tumour Breast Tissue	Raw RNA-Seq counts	TCGA*
	Triple Negative Primary Tumour		
	Her2-Positive Primary Tumour		
	Estrogen-Positive Primary Tumour		
<b><i>Integrated samples</i></b>	Normal (Healthy) Breast Tissue	Raw RNA-Seq counts	GTEx**

\* TCGA – The Cancer Genome Atlas

\*\* GTEx – The Genotype-Tissue Expression project

#### 4.2.2 Discovery of Differentially Expressed Genes (DEGs)

#### 4.2.3 Paired TCGA Samples

The R package, *edgeR* (Law, Alhamdoosh, Su, Smyth, & Ritchie, 2016; Robinson, McCarthy, & Smyth, 2010), was applied for the discovery of differentially expressed genes between normal and tumour samples. An initial set of 40 paired samples (1 normal-adjacent-tumour and 1 primary tumour sample from a single patient, from 20 different patients) was used. This was done to ensure that a differential gene expression signal was indeed present, and to ensure correct implementation and application of the *edgeR* package, which was central to subsequent analyses.

The paired sample dataset was filtered to remove any genes which consistently had a zero value across both normal and tumour samples, which deemed  $\pm 14\%$  of the genes to be as non-informative. Data was then normalized using the “upper-quartile” normalization method, allowing the distribution of the data to be less skewed.

The generalized linear model (GLM) model, built into the *edgeR* package for more complex experimental designs of paired tissue samples, was implemented to identify differentially expressed genes.

Once the method for discovering DEG's was established, the same protocol was then applied to a larger dataset: 80 paired samples – 1 normal-adjacent-tumour and 1 primary tumour sample from 40 different patients. Although a slightly different output of DEG's was to be expected, the informative genes would later be evaluated with cluster analysis to ascertain which set of DEG's best discriminate between unpaired normal-adjacent-tumour and primary tumours. To this end, signatures representing the top 100, 75, 50, and 25 differentially expressed genes were selected to test their ability to classify unpaired samples.

#### **4.2.4 Integrated GTEx and TCGA datasets**

Before the edgeR package could be implemented for the discovery of DEG's between GTEx normal samples (from healthy individuals) and TCGA primary tumour samples, the RNA-Seq files needed to be comparable, considering their different gene library sizes. A shared gene library of 53197 genes that overlapped between the two different datasets was then used to filter the data prior to downstream analysis.

A similar protocol for the discovery of DEG's was implemented for the integrated datasets consisting of 100 GTEx normal samples and 100 TCGA primary tumour samples. Trimming of zero's discarded  $\pm 6\%$  of genes as non-informative. The resultant DEG's were extracted and the top 100, 75, 50, and 25 differentially expressed genes were selected for evaluation in classification.

#### **4.2.5 Separation of GTEx normal from TCGA normal-adjacent-tumour**

TCGA normal breast tissue samples are labelled as normal-adjacent-tumour (NAT), meaning that they are collected from patients who already have breast cancer. In an effort to elucidate if there were differences between normal breast tissue samples from healthy individuals (GTEx) versus from cancer patients (TCGA NAT), we extracted the top 500 and top 1000 DEG's from the two different DEG analysis iterations (TCGA normal versus tumour and GTEx normal versus TCGA tumour), and extracted the overlapping genes.

This was done to ascertain if there was indeed a difference between normal breast tissue (from healthy individuals) and normal breast tissue (adjacent-tumour) gene expression, and possibly discover a set of DEG's which could firstly characterize normal-adjacent-tumour samples, and secondly evaluate the ability of the discovered signature to accurately discriminate between normal, normal-adjacent-tumour, and primary tumour samples using a multi-class classification.

#### 4.2.6 Z-Score Barcoding of RNA-Seq count data

Negative and positive  $z$ -scores represent normalized relative gene expression level, and can be used to substitute mRNA level (Siegfried *et al.*, 2015) or raw HTSeq counts. In order to generate these  $z$ -scores for the normal and tumour RNA-Seq samples, the `scale()` command was used in R, which would convert raw gene counts to a  $z$ -scores within an individual sample. For each gene (of each sample), where the raw count of a gene (of a sample) =  $x$ , the mean of gene counts within that sample =  $mean$ , and standard deviation of that sample's gene counts =  $sd$ , then the  $z$ -scores for each gene within a sample could be calculated as:  $z = (x - mean)/sd$ .

The extracted  $z$ -scores could thus now be 'barcoded', where a '0' would be assigned to a negative  $z$ -score and a '1' to a positive  $z$ -score. The ability of the RNA-Seq barcode to definitively classify normal and tumour samples was evaluated and once confirmed, could then be applied to other tissue types or experimental designs.

#### 4.2.7 Z-score Barcoding of unpaired TCGA and GTEx samples

As with the TCGA paired sample dataset, RNA-Seq gene counts of unpaired TCGA primary tumour and normal-adjacent-tumour samples, along with GTEx normal samples were converted to  $z$ -scores using the `scale()` method in R. The  $z$ -scores were then converted to barcodes as described above.

Prior to assessment of the barcode for classification of breast tumour samples, the informative genes (Top DEG's) were extracted to simplify feature selection.

For each of the classification scenarios the following filtering method was applied:

- 1) For TCGA normal-adjacent-tumour versus TCGA tumour, the Top 100, 75, 50 and 25 differentially expressed genes (DEG's), discovered using paired samples, and were extracted.

- 2) For GTEx normal versus TCGA tumour, the Top 100, Top 100, 75, 50 and 25 differentially expressed genes (DEG's) were extracted.
- 3) For the integrated GTEx-TCGA dataset (to discover a set of DEG's which could firstly characterize normal-adjacent-tumour samples), the overlap of the Top 500 and Top 1000 DEG's of each DEG analysis iteration was extracted from GTEx, and TCGA datasets.

#### **4.2.8 Unsupervised Machine Learning: Hierarchical and K-means**

##### **Clustering**

Machine learning algorithms like Hierarchical clustering and K-means clustering are heuristic in nature and allow for initial evaluation of the strength and/or accuracy of a classification model.

Hierarchical clustering uses agglomerative clustering – where each sample is initially assigned to its own cluster, two neighbouring clusters (of 1 sample each) are then linked to each other based on similarities. Each iteration continues to link similar samples to each other until distinctive clusters are formed – the merged clusters creating a binary tree or hierarchy.

K-means clustering uses partitional clustering. The goal of the algorithm is to group similar samples together into  $k$  partitions (clusters). The Euclidean distance between samples is used to cluster similar samples together, where Euclidean distance inversely correlates to similarity. For each iteration of clustering, a sample is assigned to the closest cluster center. Each resulting group or partition will include samples of mutual similarity.

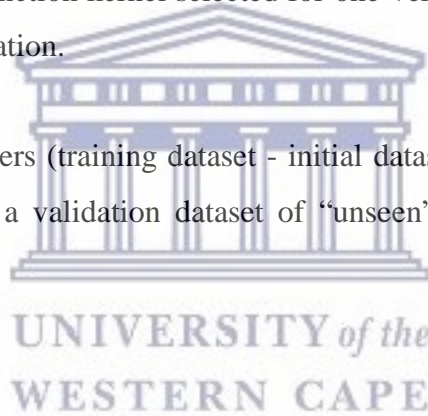
Hierarchical clustering was first applied to the gene sets detailed in 2.4.1, and once the feature sets were assessed as being informative enough to accurately classify barcoded RNA-Seq breast tissue samples (normal, normal-adjacent-tumour, and primary tumour), K-means clustering was applied all datasets for each of the informative gene sets.

#### 4.2.9 Supervised Machine Learning: Support Vector Machines

Support Vector Machines (SVM) was employed as a supervised machine learning algorithm after initial evaluation of the feature sets discovered. SVMs allow the user to input a training set of data where class, or tissue type can be specified. The algorithm then calculates a “margin” or hyperplane of separation between samples of two different classes.

Unseen validation datasets consisting of TCGA normal-adjacent-tumour and primary tumour, and GTEx normal samples that had been “barcoded” were classified using SVMs to assess if the feature sets could accurately assign samples to their correct known classes. The “*e1071*” R package was used for SVM classification, with the linear kernel selected for binary (two-class) and the radial basis function kernel selected for one-versus-one (multi-class OVO-SVM) C-classification.

The SVM classifiers (training dataset - initial datasets used in DEG analysis) were tested with a validation dataset of “unseen” samples for iterations as detailed in 2.4.1.



### 4.3 Results

#### 4.3.1 Feature set discovery in a paired TCGA normal-tumour dataset

The filtering of genes that were consistently lowly expressed lowered the number of genes from 60483 down to  $\pm 52\ 000$  genes. Following DE analysis, the top 100, 75, 50 and 25 differentially expressed genes were found to have fold changes greater than 2, with very low adjusted p-values (Table 4.2), suggesting that these feature set(s) would likely be able to definitively differentiate between normal-adjacent-tumour and tumour samples.

Prior to machine learning classification of unpaired normal-adjacent-tumour (NAT) and primary tumour samples was performed, hierarchical clustering was applied to the top 100 DEG's from each of the DEG analyses iterations; Figure 4.1 shows 140 unpaired samples clustered with informative genes discovered using 40



paired samples, and Figure 4.2 shows the same 140 unpaired samples clustered with informative genes discovered using 80 paired samples. The DEG's extracted from the 80 paired sample analysis was used in subsequent, downstream analyses and classification, as these genes were able to distinguish between tissue types more accurately.



Table 4. 2: Top 50 DEG's extracted from differential expression analysis of 80 paired TCGA NAT and Primary Tumour samples.

Differentially Expressed Genes	Log Fold-Change	p-Values	FDR (adjusted p-Values)
ENSG00000249669.6	-4.434694263	1.8373E-117	9.9443E-113
ENSG00000123500.8	7.434499304	7.8473E-113	2.1236E-108
ENSG00000230838.1	6.210239154	1.31508E-77	2.37259E-73
ENSG00000099953.8	6.134165335	3.41309E-77	4.39418E-73
ENSG00000167900.10	3.161174788	4.05936E-77	4.39418E-73
ENSG00000169241.16	2.370218283	1.2027E-75	1.08492E-71
ENSG00000269936.3	-3.774200534	4.38549E-75	3.39086E-71
ENSG00000137225.11	-2.864682714	1.28804E-73	8.71421E-70
ENSG00000203805.9	5.995143604	6.09783E-71	3.6671E-67
ENSG00000077152.8	3.345475916	1.15564E-70	6.25478E-67
ENSG00000154736.5	-2.880787494	1.07524E-67	5.2906E-64
ENSG00000122641.9	3.871265606	6.11414E-66	2.75768E-62
ENSG00000119771.13	-2.675690579	7.13479E-63	2.97049E-59
ENSG00000143549.18	1.842349082	5.66739E-61	2.19101E-57
ENSG00000123975.4	2.538594917	1.14512E-57	4.13191E-54
ENSG00000241684.4	-3.285083983	4.73071E-57	1.5936E-53
ENSG00000179796.10	-4.122695034	5.00539E-57	1.5936E-53
ENSG00000172318.5	-3.629498652	6.0258E-56	1.81189E-52
ENSG00000148053.14	-3.526384565	8.78702E-56	2.5031E-52
ENSG00000179094.12	-1.776648312	7.30914E-55	1.978E-51
ENSG00000158850.13	1.648201487	1.22839E-53	3.06004E-50
ENSG00000198932.11	-2.169099127	1.24383E-53	3.06004E-50
ENSG00000170312.14	3.356942048	2.78003E-53	6.54201E-50
ENSG00000083067.21	-2.926262212	3.887E-53	8.76582E-50
ENSG00000160753.14	1.893639404	5.68973E-53	1.2318E-49
ENSG00000117650.11	4.710060486	7.30671E-53	1.52103E-49
ENSG00000161888.10	3.284471119	8.00579E-53	1.60483E-49
ENSG00000090889.11	4.347664643	2.6363E-52	5.09596E-49
ENSG00000169258.6	2.926351502	3.00573E-52	5.60973E-49

ENSG00000025423.10	3.256130792	6.27624E-52	1.13232E-48
ENSG00000166803.9	3.648661362	9.33087E-52	1.58489E-48
ENSG00000143228.11	4.067274187	9.37042E-52	1.58489E-48
ENSG00000100526.18	3.243788354	1.04394E-51	1.71218E-48
ENSG00000143493.11	1.56345217	1.23769E-51	1.97026E-48
ENSG00000208035.1	-4.785676044	1.28835E-51	1.99231E-48
ENSG00000136158.9	-2.274222562	1.49651E-51	2.24992E-48
ENSG00000013810.17	2.436535656	2.07864E-51	3.04066E-48
ENSG00000134690.9	3.076288541	3.65671E-51	5.20831E-48
ENSG00000076382.15	2.646701137	8.46716E-51	1.17507E-47
ENSG00000108924.12	-3.178286978	9.55405E-51	1.29276E-47
ENSG00000188486.3	1.90014401	4.91226E-50	6.48467E-47
ENSG00000065534.17	-1.953704813	5.7853E-50	7.45533E-47
ENSG00000127564.15	4.258846765	6.82836E-50	8.59484E-47
ENSG00000154263.16	-3.702210309	7.62125E-50	9.37484E-47
ENSG00000079462.6	2.61127339	1.05422E-49	1.26797E-46
ENSG00000177628.14	1.662281779	1.78839E-49	2.06772E-46
ENSG00000168497.4	-3.659070393	1.79556E-49	2.06772E-46
ENSG00000135094.9	3.120126981	2.35637E-49	2.657E-46
ENSG00000149923.12	1.542231541	3.98288E-49	4.39937E-46
ENSG00000254986.6	1.755083595	6.51367E-49	7.05092E-46

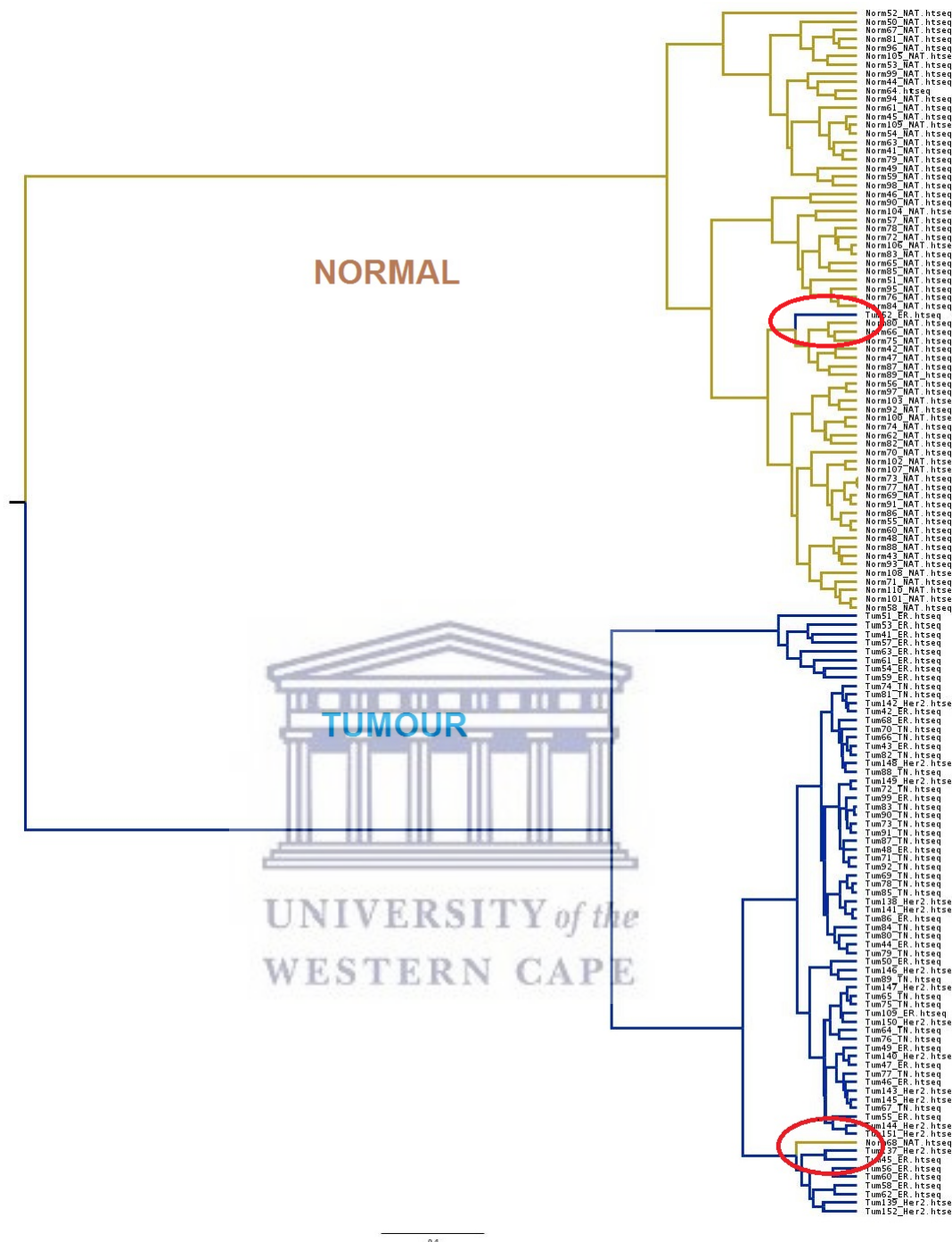


Figure 4. 1: Hierarchical clustering of “barcoded” TCGA Normal-Adjacent-Tumour (NAT) and TCGA Primary Tumour (Tumour) Unpaired RNA-Seq samples (n = 100) yielded 98% accuracy when classified using the Top 100 DEG's discovered using 40 paired Normal-Adjacent-Tumour and Tumour samples.

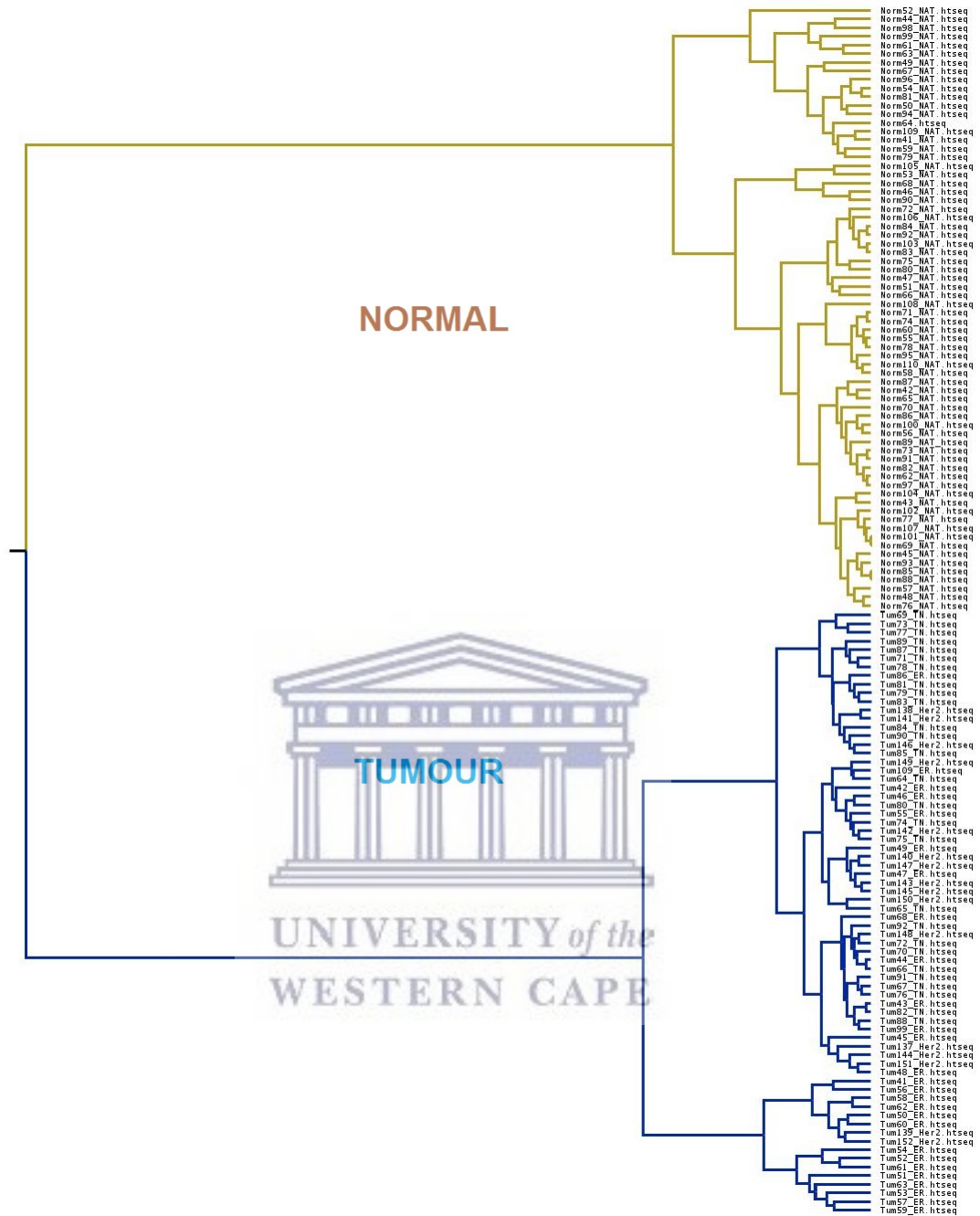


Figure 4. 2: Hierarchical clustering of “barcoded” TCGA Normal-Adjacent-Tumour (NAT) and TCGA Primary Tumour (Tumour) Unpaired RNA-Seq samples ( $n = 100$ ) yielded 100% accuracy when classified using the Top 100 DEG's discovered using 80 paired Normal-Adjacent-Tumour and Tumour samples.

### **4.3.2 Feature set discovery in an integrated GTEx-TCGA dataset**

Integration of GTEx and TCGA RNA-Seq gene count data resulted in a  $\pm 6\%$  and  $\pm 12\%$  data loss due to a difference in GTEx and TCGA library sizes; GTEx consisting of 56203 genes and TCGA consisting of 60843 genes – with a resulting overlap of 53196 genes. Filtering of lowly expressed genes resulted in a further reduction of uninformative data, decreasing the number of genes to 50758.

Although there was a 17% loss of TCGA primary tumour data, and 7% loss of GTEx normal data, the resultant top 100, 75, 50, and 25 DEG's, had very low adjusted p-values, and large log-fold changes (Table 4.2), suggesting that these DEG's could be used for feature selection prior to machine learning classification.

### **4.3.3 Feature set discovery for multi-class classification of a GTEx normal, TCGA normal-adjacent-tumour and TCGA primary tumour integrated dataset**

Subsequent to discovering informative genes (DEG's) capable of discerning between a) TCGA normal-adjacent-tumour and primary tumour samples, and b) GTEx normal (healthy) and TCGA primary tumour samples, the overlap of these two DEG analyses experimental designs yielded only 1 gene in common when comparing the top 100 DEG's.

Intersecting the top 500 and top 1000 statistically significant DEGs of each analysis yielded 59 and 216 genes, respectively.

These sets were used as features for multi-class classification modelling to determine the presence of a distinct gene expression signature for normal-adjacent-tumour samples from cancer patients when compared to normal breast from healthy women.



#### 4.3.4 Z-scores and “Barcoding” RNA-Seq gene counts

The heatmap in Figure 4.3 indicates that the top 100 DEG's discovered with edgeR accurately separate normal samples from tumour samples. The z-scores generated to produce the map allow the differences in relative gene expression levels between the two tissue types to be easily visualized. The TCGA normal-adjacent-tumour samples appear on the left-side of the heatmap, with TCGA primary tumour samples appearing on the right-side of the image. This indicated that these z-scores could be converted to 1's and 0's to generate a barcode for normal and tumour samples. In order to mimic the Gene Expression Barcode's (GExB) single sample algorithm (barcoding of a single microarray sample) (McCall et al., 2014, 2011), which would produce statistically derived absolute gene expression calls, the R *scale()* function was applied to each sample independently of other samples from the same tissue type, i.e. normal-adjacent-tumour or primary tumour, unlike typical application of z-score to cancer genomic data which scales the raw gene count data across the tissue (Colaprico, Olsen, supervisor, & Bontempi Biopark Charleroi, 2016). Scaling of data with similar statistical methods, has been shown to improve classification of TCGA data (Rahman et al., 2015). We hypothesized that z-scores and their conversion to absence/presence calls would perform similarly, as they were previously applied in RNA-Seq data to infer gene expression (Siegfried et al., 2015).

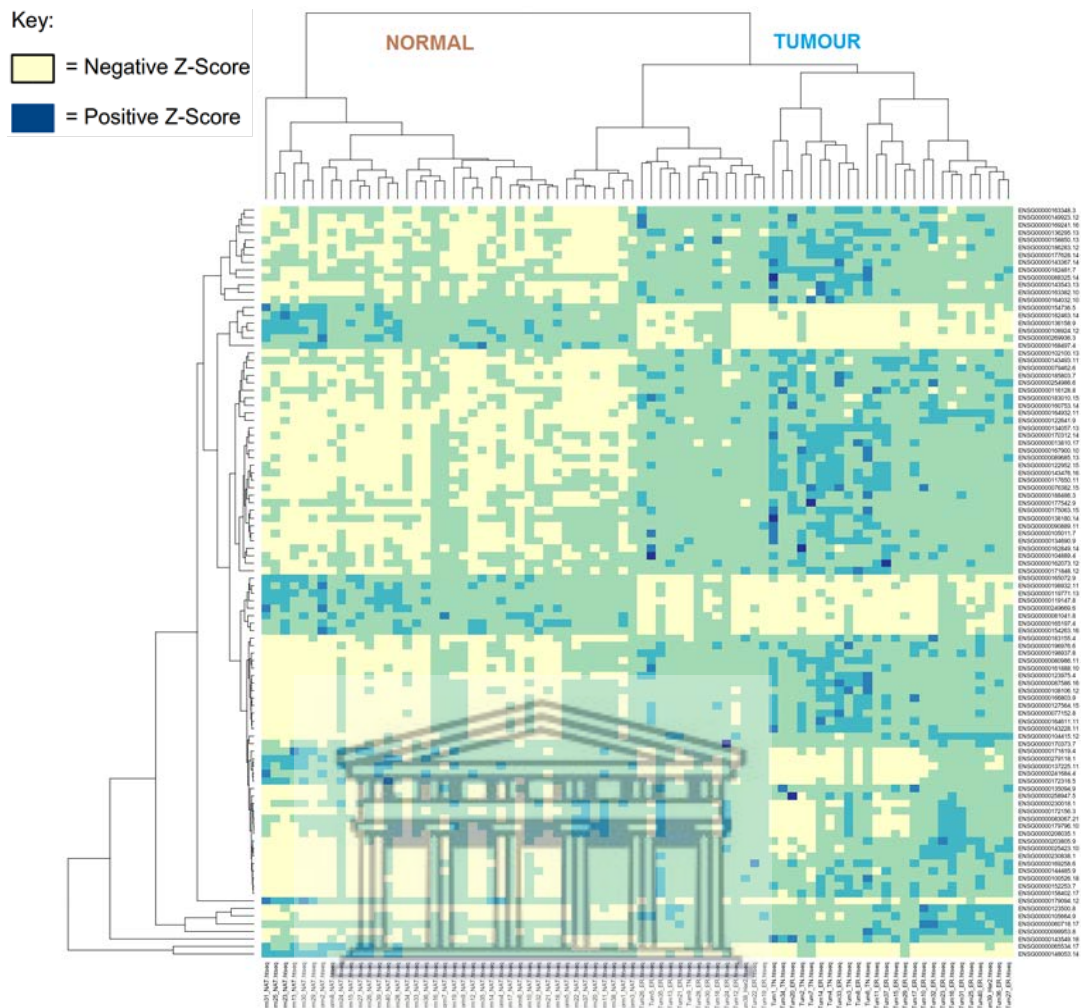


Figure 4. 3: Z-Score Heatmap of Paired Normal-Tumour Samples using Top 100 DEGs as a feature set

### 4.3.5 Machine Learning classification

#### 4.3.5.1 Clustering and SVM classification of TCGA data

K-means clustering was able to use the expression barcode generated from DEG's discovered on paired TCGA samples to classify 140 unpaired TCGA normal-adjacent-tumour and primary tumour samples with above 95% accuracy. Hierarchical clustering and SVM binary classification improved the accuracy to 100% (Table 4.3).

Table 4. 3: Classification results of TCGA Normal-Adjacent-Tumour (NAT) and TCGA Primary Tumour (Tumour) Unpaired RNA-Seq samples (n = 140)

	K-means		Hierarchical Clustering		SVM	
	NAT	Tumour	NAT	Tumour	NAT	Tumour
<i>Top100 DEG</i>	100%	98%	100%	100%	100%	100%
<i>Top75 DEG</i>	99%	98%	100%	100%	100%	100%
<i>Top50 DEG</i>	100%	97%	100%	100%	100%	100%
<i>Top25 DEG</i>	100%	98%	100%	100%	100%	100%

#### 4.3.5.2 Clustering and SVM classification of GTEX-TCGA integrated data

K-means clustering, hierarchical clustering and SVM were all able to use the barcode discovered using the dataset of 200 samples to classify 100 unseen GTEX normal and TCGA primary samples with 100% accuracy (Table 4.4).

Table 4. 4: Classification results of GTEX Normal (Normal) and TCGA Primary Tumour (Tumour) Test RNA-Seq samples (n = 100)

	K-means		Hierarchical Clustering		SVM	
	Normal	Tumour	Normal	Tumour	Normal	Tumour
<i>Top100 DEG</i>	100%	100%	100%	100%	100%	100%
<i>Top75 DEG</i>	100%	100%	100%	100%	100%	100%
<i>Top50 DEG</i>	100%	100%	100%	100%	100%	100%
<i>Top25 DEG</i>	100%	100%	100%	100%	100%	100%

#### 4.3.5.3 Multi-class classification of healthy breast, normal-adjacent-tumour (NAT) and primary tumour tissues

Clustering analysis using the barcodes generated for the 59 gene and 216 gene signatures described in 3.1.3 on 300 samples of GTEX normal, TCGA NAT, and TCGA primary tumour revealed that both were discriminatory features to accurately separate the three tissue types (Figures 4.4 and 4.5). NAT tissue was also shown to be a “subtype” of normal samples, distinct from breast tissue from the healthy individuals in GTEX.

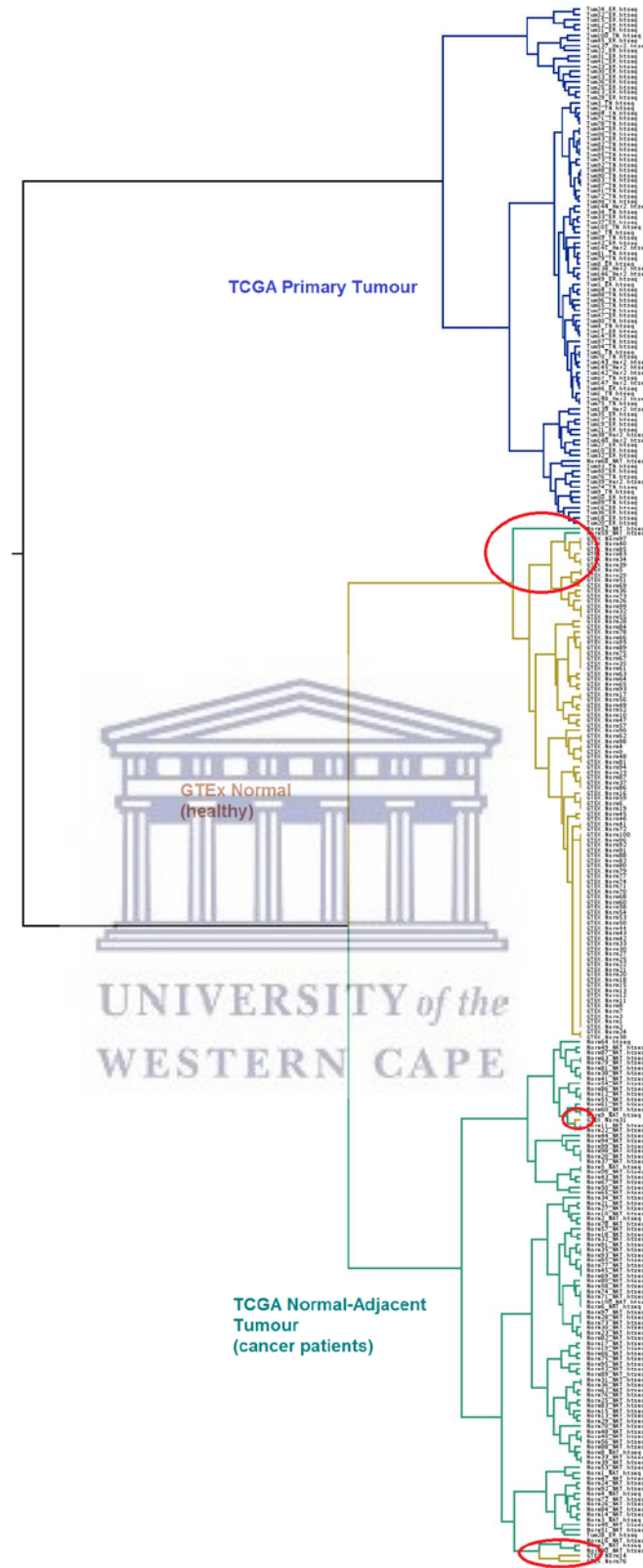


Figure 4. 4: Barcode-based hierarchical clustering of 300 GTEX and TCGA samples, yielded 98% accuracy when classified with the Top 59-overlapping DEG's (described in Sections 4.2.5 and 4.3.3).

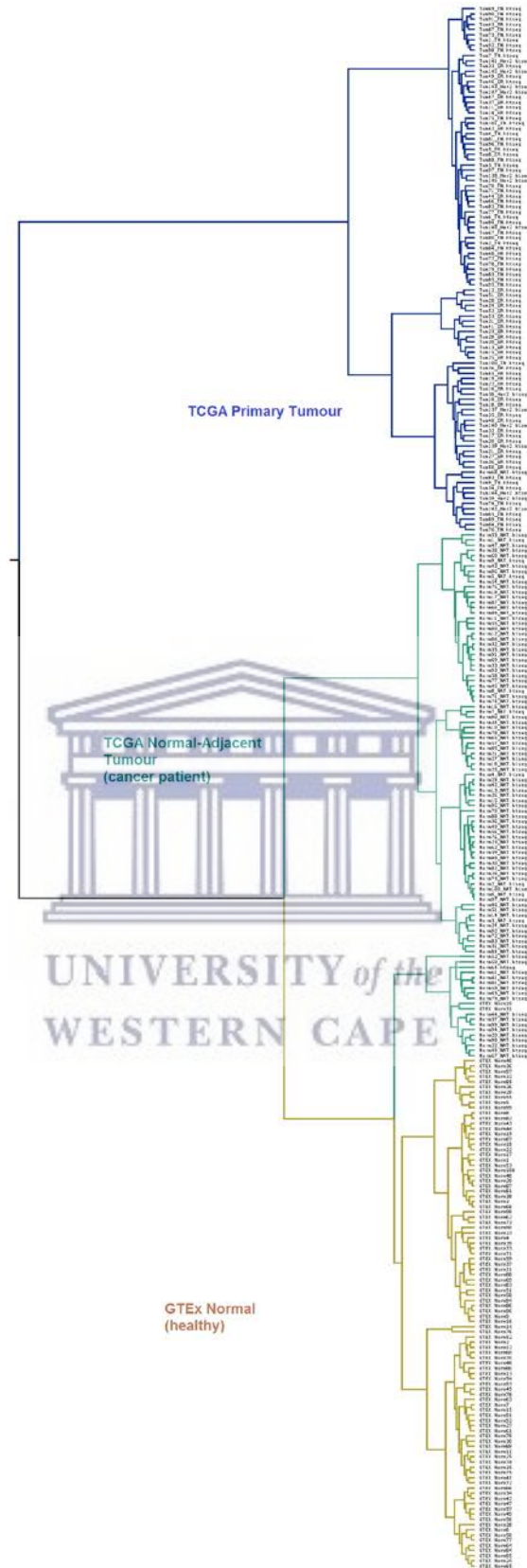


Figure 4. 5: Barcode-based hierarchical clustering of 300 GTEX and TCGA samples, yielded 100% accuracy when classified with the Top 216-overlapping DEG's (described in Sections 4.2.5 and 4.3.3).

Validation of these gene sets to correctly characterize a sample as normal (healthy), normal-adjacent-tumour (from a cancer patient), or primary breast tumour (cancer) was performed with a test dataset of unseen samples. K-means and hierarchical clustering results (Table 4.5) showed that unsupervised machine learning methods could classify samples with above 85% accuracy.

*Table 4. 5: Classification results of GTEX Normal (Normal), TCGA Normal-Adjacent-Tumour (NAT) and TCGA Primary Tumour (Tumour) RNA-Seq samples with Validation dataset*

	K-means			Hierarchical Clustering		
	GTEX Normal	TCGA NAT	TCGA Tumour	GTEX Normal	TCGA NAT	TCGA Tumour
59 DEG	100%	92%	90%	100%	92%	100%
216 DEG	100%	85%	86%	100%	92%	100%

A multi-class one-versus-one support vector machine (multi-class OVO-SVM) trained on the 300 samples hierarchically clustered in figures 4.4 and 4.5 above, and tested with 113 unseen samples (classified with K-means and hierarchical clustering in table 4.4 above) was able to distinguish between the three different tissue types with above 99% accuracy (Table 4.6).

*Table 4. 6: Overlap of Top DEG's: SVM classification with Validation dataset*

Tissue Type	No. samples tested	Classification accuracy	
		59 DEG's	216 DEG's
GTEX Normal	50	98% (49)	100% (50)
TCGA Normal-Adjacent-Tumour	13	100% (13)	100% (13)
TCGA Primary Tumour	50	100% (50)	100% (50)
<b>Total:</b>	<b>113</b>	<b>99.12%</b>	<b>100%</b>



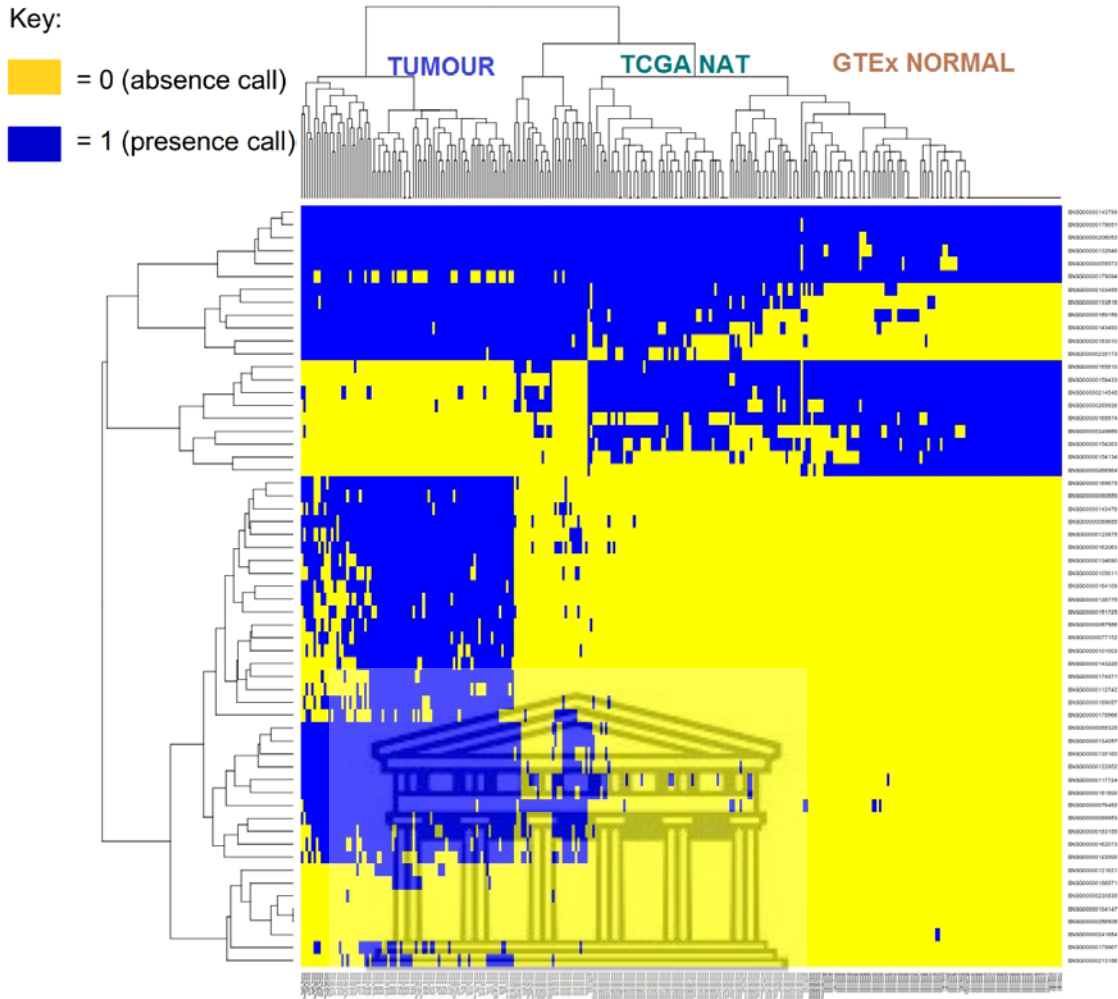


Figure 4. 6: Heatmap of 59 DEG's separating GTEX normal, TCGA NAT, and TCGA tumour tissues

#### 4.4 Discussion

##### RNA-Seq DE analysis and sample-level Z-Score data transformation produced phenotypically accurate segregating expression portraits

The *edgeR* package, when applied to TCGA paired normal-adjacent-tumour and tumour samples, as well as GTEX-TCGA integrated data reported robust sets of differentially expressed genes (DEG's), as evidenced by their very high log-fold changes and small adjusted p-values. Our application of the R *scale()* function to each sample independently of other samples from the same tissue type revealed a “transcriptomic portrait” of genes switched on and off in that sample, which could be integrated with other samples of the same type for feature selection.

The suitability of our strategy is illustrated by a heatmap of scaled data, i.e. z-

scores, in Figure 4.3, which shows mostly uniform within-class up or down regulation of extracted DEG's signatures across the tissue types.

### **RNA-Seq Barcodes of DEG's produces powerful classification signatures**

The conversion of these z-scores to a “barcode”, 1's and 0's proved robust for classification of TCGA data. Figure 4.2 and Table 4.2 demonstrated that using multi-tissue RNA-Seq barcodes of pre-identified DEG's as classification features was remarkably accurate in identifying a samples as malignant (primary tumour) or normal.

The degree of classification accuracy (Table 4.6) attained when applying the approach to the integrated RNA-Seq data from GTEx (breast tissue from healthy women) and TCGA (tumour and adjacent normal) was still more notable, given that multi-class classification is inherently more difficult and the fact that data was lost during integration. Thus our approach of DEG discovery coupled with z-score-to-barcode data transformation, proved to be a reliable way of identifying robust predictive gene signatures in integrated data from multiple sources.

In recent years, the emergence of in-depth transcriptome profiling has led to a few diagnostic and prognostic gene expression signatures being developed from RNA-Seq data. Most notably, a lung adenocarcinoma 4-gene prognostic signature (Shukla et al., 2017), and a colorectal cancer 12-gene prognostic signature (Sun et al., 2018), developed with TCGA RNA-Seq data and survival analysis algorithms. Studies which were aimed at building classifiers for subtyping and staging of cancers, with SVM implementation however, are more closely comparable to this study. A cancer diagnostic classifier based on gene expression (RNA-Seq data) of blood platelets was able to classify cancer subtypes with above 75% accuracy with SVM implementation (Y.-H. Zhang et al., 2017). A breast cancer staging classifier which could determine early or late stage cancer, implemented SVM recursive feature elimination (SVM-RFE) and was able to classify with tumours with above 98% accuracy (Yao, Zhang, Du, Liu, & Xu, 2015). In contrast, the two-class (normal versus tumour) and the multi-class classifiers developed were

able to classify unseen samples with 100% accuracy.

We thus propose that our method of sample-level z-score based barcoding of RNA-Seq data is analogous to the Gene Expression Barcode algorithm designed specifically for microarray data and has likely future utility in discovery of predictive signatures in other study scenarios. As several of the signature genes were previously described as novel RNA-Seq derived biomarkers (Wang, Gerstein, & Snyder, 2009), we further propose that the signatures may have potential for breast cancer diagnostics R&D, since the absence-presence calls can be readily and simply assessed using qPCR.

### **NAT tissues classify separately from healthy and tumour tissues**

When overlapping the top 500 and top 1000 DEG's from TCGA and GTEx-TCGA analysis, barcode signatures of 59 and 216 genes emerged which were able to distinguish between the three different tissue types, respectively. NAT tissue presents with a distinctly different transcriptome portrait when compared to normal samples acquired from healthy individuals (GTEx) and primary breast tumours from TCGA, as represented in Figure 4.6.

These observations are strongly supported by previous studies geared towards the elucidation of the molecular profiles of histologically normal tissues adjacent to malignant breast tumours. A study undertaken by Boston University School of Medicine revealed that 25-53% of the genes over-expressed or under-expressed in estrogen-receptor positive or negative breast tumours were shared with normal-adjacent-tumour tissue samples (Kelly Graham, Ge, de Las Morenas, Tripathi, & Rosenberg, 2011). These findings were echoed in a 2015 study which revealed that intrinsic tumour subtypes (PAM50 subtypes) were reflected in histologically normal-adjacent-tumour tissues, and suggested that the shared molecular portrait(s) may account for cancer recurrence and the derivation of biomarkers may be plausible (Casbas-Hernandez et al., 2015).

Aran and colleagues took the above into account and investigated the NAT transcriptome. Their analysis spanned a few different cancers and included the

integration of normal tissues from healthy individuals (GTEx project). It was revealed that not only did NAT tissue share a partial molecular profile with tumour samples, and a partial molecular profile with healthy tissues, but also possessed its own gene expression signatures (Aran et al., 2017). Furthermore, they hypothesize that the distinct molecular portrait of NAT tissues may in fact be attributed to the tumour's effect on the adjacent tumours due to particular genes being up-regulated in NAT tissues which are linked to molecular subtypes, and immune response pathways. Upon closer examination of Figure 4.5, within the 216 informative gene set, NAT tissues appear as a “subtype” of normal in relation to GTEx normal, but are more closely related to tumour samples. This is and can be deduced from Figure 4.6, where some genes of NAT tissues begin to exhibit.

The experimental design of cancer studies aimed at gene signature discovery thus needs to take these findings into account. More expansive research is required with larger sample numbers, to discover a “universal” signature which can definitively distinguish between normal and normal-adjacent-tumour samples. Although Table 4.3 reveals a possible robust gene signature to classify tumour samples accurately, there was no overlap between the Top 100 DEG's from TCGA normal versus tumours and the Top 100 DEG's from GTEx normal versus TCGA tumour. Researchers may be missing possible biomarkers/gene signatures which are more discriminative in cancer classification due to the overlap of genes expressed between NAT and tumour tissues (Figure and Table 4.6). In order to build true multi-class classifiers which can distinguish between normal, normal-adjacent-tumour (pre-cancer), and different molecular or intrinsic subtypes of breast cancer, NAT tissues must be treated as separate class from healthy-normal samples, and an integration of the two in biomarker discovery may prove beneficial in robust and accurate cancer classifier development.

### **Limitations to integrative RNA-Seq data analysis**

Comprehensive studies which can reveal novel insights into tumorigenesis, metastases, and progression of cancers, including breast cancer, is limited by not only the amount of data available for normal-adjacent-tumour tissues and normal

(healthy) samples but also by the steps necessary to integrate such data.

The integration of data from The GTEx and TCGA projects resulted in the loss of potentially informative genes due to differences in study designs, and gene definitions used. Although publicly available data from NCBI GEO may aid in increasing dataset size for DEG discovery and serve as validation sets, this too possesses constraints due to difference in sequencing platforms – Illumina HiSeq 2000 for GTEx and TCGA data, Illumina HiSeq 2500, 3000 & 4000 for the majority of transcriptomic data published in NCBI GEO. A possible solution to integrating RNA-Seq data generated on different platforms may also require accessing raw data, post assembly and alignment of transcripts, and generating HTSeq counts prior to differential expression analysis may better normalize the expression counts used.

#### **4.5 Conclusions**

The integration of GTEx and TCGA data allowed for the discovery of a very distinct NAT tissue gene expression profile. Each iteration of differential expression analysis revealed three different classifiers that all classified unseen data with 100% accuracy. Although a distinctly different molecular portrait of TCGA NAT tissues was revealed, the 10,000 genes not present in the GTEx gene library, contributed to a set of informative genes still capable of segregating tumour samples from normal samples. Thus NAT tissues may still serve as a control in cancer transcriptomic studies, along with the additional advantages of easy biospecimen accessibility during tumour biopsy collection.

The development of a barcoding method for RNA-Seq gene count data proved robust in transforming continuous data and enabled an “ease” of tissue discrimination to classifier development. The results obtained, albeit convincingly validated on an unseen dataset, could be further assessed with samples from other study consortia. Further investigation of NAT differentiating genes through function and pathway enrichment analysis may also ascertain their molecular roles in tumorigenesis and potential as both early breast cancer detection biomarkers and candidate drug targets.

## **CHAPTER 5**

### **Conclusions and Future Work**

The primary aim of our study was to integrate breast cancer microarray expression data using the GExB algorithm to discover easily assayable signatures for breast cancer subtypes. We also aimed to extend this to RNA-Seq data, with the implementation of our own RNA-Seq barcode method.

#### **5.1 Conclusion**

Transformation of expression data into barcodes simplifies discovery of features that are able to discriminate between sample types. Used in combination with machine learning and customised feature selection, gene expression barcodes produced signatures that clearly separate breast cancer subtypes.

Enrichment analysis of both the microarray and RNA-Seq gene signatures revealed that unbiased approaches to FS can greatly enhance the biologically relevant discoveries made in bioinformatics. Within both gene signatures, 306 known genes from microarray and 213 known genes from RNA-Seq, diseases in which these genes were involved included cancers of the skin (melanoma), lung, liver, kidneys, breast, endometrium (uterine), leukaemia, as well as illnesses with a known inflammatory nature such as arthritis, lupus erythematosus, and Crohn's disease. Moreover, the both sets were found to be statistically enriched for biological pathways and processes relevant to cancer, including: apoptosis, p53 signalling, and signalling pathways regulating pluripotency of stem cells.

The barcoding method developed for RNA-Seq data holds promise for implementation in biomarker discovery for cancer in the NGS era. The NAT specific profile discovered was easily detectable and visualized with data transformation from continuous data to discrete data. Interrogation of the informative gene sets, in particular GTE<sub>x</sub> normal versus TCGA primary



tumour, 36 novel genes were involved in an accurate classifier being developed. These results pose various questions to cancer researchers surrounding not only differential expression analysis of tumours' experimental design, but also the very nature of normal tissues surrounding tumours and the possible biological insights into cancer metastases (Pietras & Östman, 2010). Mechanisms of tumorigenesis and the tumour's interaction with its surrounding environment needs to be closely investigated to fully elucidate the unique portrait of NAT tissues (Grivennikov, Greten, & Karin, 2010; Terzić, Grivennikov, Karin, & Karin, 2010).

Haibe-Kains speculated in an article about classification models for breast cancer using gene expression could, “if widely used in a standardized fashion, could dramatically change the way in which patients are managed in a clinical setting and, hopefully, could lead to substantial improvements in outcome and survival” (Haibe-Kains, 2010). The potential to design and implement clinical assays, e.g. qPCR, from RNA-Seq discovered biomarkers is clearly illustrated through the methods developed, implemented and validated throughout this research study.



## **5.2 Discovered signatures are applicable across technologies**

As a final step in validating both the gene expression signatures discovered on microarray data, as well as the RNA-Seq z-score-to-barcode method, classification of RNA-Seq barcoded samples was performed with the two-class microarray feature set. The 85 microarray probes were mapped to Ensembl gene ID's used in Illumina HiSeq 2000 sequencing data. This resulted in 75 genes, which were then extracted from 300 barcoded RNA-Seq breast tissue samples – 100 GTEx normal, 100 TCGA NAT and 100 TCGA Tumour. Hierarchical clustering resulted in 95.33% accurate classification of barcoded RNA-Seq samples.

### **5.3 Future Work**

To discover a more informative and probe/gene set for multiclass breast cancer classification, data slicing should be considered. This would entail analysing the classes of breast cancer in a one-versus-one (OVO) rather than OVR fashion to identify genes or probes that distinguish one subtype from another. Intersecting the OVO-FS discovered probes could allow a smaller but more informative probe set to be determined.

Data curation, however extensive, was limited and could be extended to a far larger dataset. The poorly labelled samples and mislabelled samples could be considered as ambiguous and be further explored using the original 346-gene signature, then using those assigned classes to build a larger training dataset. While further evaluation of the RNA-Seq barcoding technique is necessary, our results point to its potential for application to other cancers, as well as other diseases or R&D applications that could benefit from accurate classification of clinical phenotypes, therapeutic responses and expected survival times, etc.



## 6 References

- Abraham, A. (2005). *Handbook of Measuring System Design. Handbook of Measuring System Design*. <https://doi.org/10.1007/BF00657309>
- Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 36(2), 3240–3247. <https://doi.org/10.1016/j.eswa.2008.01.009>
- Alanko, A., Heinonen, E., Scheinin, T., Tolppanen, E. M., & Vihko, R. (1985). Significance of estrogen and progesterone receptors, disease-free interval, and site of first metastasis on survival of breast cancer patients. *Cancer*, 56(7), 1696–1700. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/4027900>
- Allison, D. B., Cui, X., Page, G. P., & Sabripour, M. (2006). Microarray data analysis: From disarray to consolidation and consensus. *Nature Reviews Genetics*. <https://doi.org/10.1038/nrg1749>
- Alpaydin, E. (2010). *Introduction to Machine Learning Second Edition. Introduction to Machine Learning*. [https://doi.org/10.1007/978-1-62703-748-8\\_7](https://doi.org/10.1007/978-1-62703-748-8_7)
- Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btu638>
- Andersson, A., Ritz, C., Lindgren, D., Edén, P., Lassen, C., Heldrup, J., ... Fioretos, T. (2007). Microarray-based classification of a consecutive series of 121 childhood acute leukemias: prediction of leukemic and genetic subtype as well as of minimal residual disease status. *Leukemia*, 21, 1198–1203. <https://doi.org/10.1038/sj.leu.2404688>
- Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New Biotechnology*, 25(4), 195–203. <https://doi.org/10.1016/j.nbt.2008.12.009>
- Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskotsky, B., Krings, G., ... Butte, A. J. (2017). Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nature Communications*, 8(1), 1077. <https://doi.org/10.1038/s41467-017-01027-z>
- Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand,

- E. T., ... Dermitzakis, E. T. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235), 648–660. <https://doi.org/10.1126/science.1262110>
- Ashwag Albukhari, Fawzi F. Bokhari, and H. C. (2015). Next-Generation Sequencing in the Era of Cancer-Targeted Therapies: Towards the Personalised Medicine. In H. C. W. Wu (Ed.), *Next Generation Sequencing in Cancer Research* (2nd ed., pp. 39–55). Springer International Publishing Switzerland. [https://doi.org/10.1007/978-3-319-15811-2\\_3](https://doi.org/10.1007/978-3-319-15811-2_3)
- Auffarth, B. (2010). Clustering by a genetic algorithm with biased mutation operator. In *2010 IEEE World Congress on Computational Intelligence, WCCI 2010 - 2010 IEEE Congress on Evolutionary Computation, CEC 2010*. <https://doi.org/10.1109/CEC.2010.5586090>
- Bemmo, A., Benovoy, D., Kwan, T., Gaffney, D. J., Jensen, R. V., & Majewski, J. (2008). Gene expression and isoform variation analysis using affymetrix exon arrays. *BMC Genomics*. <https://doi.org/10.1186/1471-2164-9-529>
- Blum, A. L., & Rivest, R. L. (1992). Training a 3-node neural network is NP-complete. *Neural Networks*, 5(1), 117–127. [https://doi.org/10.1016/S0893-6080\(05\)80010-3](https://doi.org/10.1016/S0893-6080(05)80010-3)
- Bolón-Canedo, V., Sánchez-Marroño, N., & Alonso-Betanzos, A. (2016). Feature selection for high-dimensional data. *Progress in Artificial Intelligence*, 5(2), 65–75. <https://doi.org/10.1007/s13748-015-0080-y>
- Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/19.2.185>
- Burrell, R. A., McGranahan, N., Bartek, J., & Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467), 338–345. <https://doi.org/10.1038/nature12625>
- Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M., Glas, A. M., ... Straehle, C. (2006). Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute*. <https://doi.org/10.1093/jnci/djj329>

- Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D., & Craig, D. W. (2016). Translating RNA sequencing into clinical diagnostics: Opportunities and challenges. *Nature Reviews Genetics*, *17*(5), 257–271. <https://doi.org/10.1038/nrg.2016.10>
- Cancer Genome Atlas Network. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*(7407), 330–337. <https://doi.org/10.1038/nature11252>
- Caragea, C., & Honavar, V. (2009). Machine learning in Computational Biology. *Encyclopedia of Database Systems*, (Gm 066387), 1663–1667. <https://doi.org/10.1007/978-0-387-39940-9>
- Carr, D. B., Somogyi, R., & Michaels, G. (1997). Templates for Looking at Gene Expression Clustering. *Statistical Computing and Statistical Graphics Newsletter*, (1995), 20–29.
- Casbas-Hernandez, P., Sun, X., Roman-Perez, E., D’Arcy, M., Sandhu, R., Hishida, A., ... Troester, M. A. (2015). Tumor Intrinsic Subtype Is Reflected in Cancer-Adjacent Tissue. *Cancer Epidemiology and Prevention Biomarkers*, *24*(2), 406–414. <https://doi.org/10.1158/1055-9965.EPI-14-0934>
- Chang, C.-C., & Lin, C.-J. (2011). LIBSVM. *ACM Transactions on Intelligent Systems and Technology*. <https://doi.org/10.1145/1961189.1961199>
- Chia, S. K., Bramwell, V. H., Tu, D., Shepherd, L. E., Jiang, S., Vickery, T., ... Nielsen, T. O. (2012). A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *18*(16), 4465–4472. <https://doi.org/10.1158/1078-0432.CCR-12-0286>
- Chin, L., Andersen, J. N., & Futreal, P. A. (2011). Cancer genomics: from discovery science to personalized medicine. *Nature Medicine*, *17*(3), 297–303. <https://doi.org/10.1038/nm.2323>
- Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., ... Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*. [https://doi.org/10.1016/S1097-2765\(00\)80114-8](https://doi.org/10.1016/S1097-2765(00)80114-8)

- Cho, S.-H., Jeon, J., & Kim, S. Il. (2012). Personalized Medicine in Breast Cancer: A Systematic Review. *Journal of Breast Cancer*, 15(3), 265. <https://doi.org/10.4048/jbc.2012.15.3.265>
- Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*. <https://doi.org/10.1038/ng1031>
- Cichetti, D. V. (1992). Neural Networks and Diagnosis in the Clinical Laboratory: State of the Art. *Clinical Chemistry*, 38(1), 9–10.
- Cieřlik, M., & Chinnaiyan, A. M. (2018). Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics*, 19(2), 93–109. <https://doi.org/10.1038/nrg.2017.96>
- Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., ... Perou, C. M. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*, 163(2), 506–519. <https://doi.org/10.1016/j.cell.2015.09.033>
- Colaprico, A., Olsen, C., supervisor, A., & Bontempi Biopark Charleroi, G. (2016). *Biography Outline NGS overview Methodologies Acknowledgements Mining and analysis of genomic and epigenomic data (TCGA) using R*. Retrieved from <https://iimo.pl/img/workshop-1/slides5.pdf>
- Collisson, E. A., Campbell, J. D., Brooks, A. N., Berger, A. H., Lee, W., Chmielecki, J., ... Cheney, R. (2014). Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. *Nature*, 511(7511), 543–550. <https://doi.org/10.1038/nature13385>
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*. <https://doi.org/10.1186/s13059-016-0881-8>
- Coomans, D., & Massart, D. L. (1982). Alternative k-nearest neighbour rules in supervised pattern recognition : Part 1. k-Nearest neighbour classification by using alternative voting rules. *Analytica Chimica Acta*, 136, 15–27. [https://doi.org/10.1016/S0003-2670\(01\)95359-0](https://doi.org/10.1016/S0003-2670(01)95359-0)
- Cooper, C. S. (2001). Applications of microarray technology in breast cancer research. *Breast Cancer Research*. <https://doi.org/10.1186/bcr291>



- Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*.  
<https://doi.org/10.1023/A:1022627411411>
- Creighton, C. J., Cordero, K. E., Larios, J. M., Miller, R. S., Johnson, M. D., Chinnaiyan, A. M., ... Rae, J. M. (2006). Genes regulated by estrogen in breast tumor cells in vitro are similarly regulated in vivo in tumor xenografts and human breast tumors. *Genome Biology*, 7(4), R28.  
<https://doi.org/10.1186/gb-2006-7-4-r28>
- Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*.  
<https://doi.org/10.1177/117693510600200030>
- D'haeseleer, P. (2005). How does gene expression clustering work? *Nature Biotechnology*. <https://doi.org/10.1038/nbt1205-1499>
- Daelemans, W., & Hoste, V. (2002). Evaluation of machine learning methods for natural language processing tasks. In *LREC 2002: third international conference on language resources and evaluation*.
- dbGaP | phs000424.v7.p2 | Common Fund (CF) Genotype-Tissue Expression Project (GTEx). (n.d.). Retrieved March 11, 2019, from [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000424.v7.p2](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v7.p2)
- Desai, A., & Jere, A. (2012). Next-generation sequencing: ready for the clinics? *Clinical Genetics*, 81(6), 503–510. <https://doi.org/10.1111/j.1399-0004.2012.01865.x>
- Desai, A. N., & Jere, A. (2015). Next-Generation Sequencing for Cancer Biomarker Discovery. In *Next Generation Sequencing in Cancer Research, Volume 2* (pp. 103–125). Cham: Springer International Publishing.  
[https://doi.org/10.1007/978-3-319-15811-2\\_7](https://doi.org/10.1007/978-3-319-15811-2_7)
- Dinh, P., Sotiriou, C., & Piccart, M. J. (2007). The evolution of treatment strategies: Aiming at the target. *Breast*.  
<https://doi.org/10.1016/j.breast.2007.07.032>
- Doebele, R. C., Davis, L. E., Vaishnavi, A., Le, A. T., Estrada-Bernal, A., Keysar, S., ... Low, J. A. (2015). An oncogenic NTRK fusion in a patient with soft-tissue sarcoma with response to the tropomyosin-related kinase inhibitor

- LOXO-101. *Cancer Discovery*. <https://doi.org/10.1158/2159-8290.CD-15-0443>
- Domingos, P., & Hulten, G. (2003). A General Framework for Mining Massive Data Streams. *Journal of Computational and Graphical Statistics*, 12(4), 945–949. <https://doi.org/10.1198/1061860032544>
- Domingos, P., & Pedro. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78. <https://doi.org/10.1145/2347736.2347755>
- Drineas, P., Frieze, A., Kannan, R., Vempala, S., & Vinay, V. (2004). Clustering {L}arge {G}raphs via the {S}ingular {V}alue {D}ecomposition. *Machine Learning*, 56, 9–33.
- Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*. <https://doi.org/10.1198/016214502753479248>
- F. Bray, Jacques Ferlay, Isabelle Soerjomataram, Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*. <https://doi.org/10.3322/caac.21492>
- Fan, J., & Fan, Y. (2008). HIGH-DIMENSIONAL CLASSIFICATION USING FEATURES ANNEALED INDEPENDENCE RULES 1. *The Annals of Statistics*, 36(6), 2605–2637. <https://doi.org/10.1214/07-AOS504>
- Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., ... Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, 136(5), E359–E386. <https://doi.org/10.1002/ijc.29210>
- Figueroa, C. J., Tang, Y. W., & Taur, Y. (2014). *Principles and Applications of Genomic Diagnostic Techniques. Molecular Medical Microbiology: Second Edition* (Vol. 1–3). Elsevier Ltd. <https://doi.org/10.1016/B978-0-12-397169-2.00022-6>
- Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of

- cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10), 906–914. <https://doi.org/10.1093/bioinformatics/16.10.906>
- GDC Data Portal. (n.d.). Retrieved March 11, 2019, from <https://portal.gdc.cancer.gov/repository>
- George, B., Ashokachandran, V., Paul, A. M., & Girijadevi, R. (2017). Transcriptome Sequencing for Precise and Accurate Measurement of Transcripts and Accessibility of TCGA for Cancer Datasets and Analysis. In *Applications of RNA-Seq and Omics Strategies - From Microorganisms to Human Health*. InTech. <https://doi.org/10.5772/intechopen.70026>
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*. <https://doi.org/10.1126/science.286.5439.531>
- Graham, K., De Las Morenas, A., Tripathi, A., King, C., Kavanah, M., Mendez, J., ... Rosenberg, C. L. (2010). Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. *British Journal of Cancer*, 102(8), 1284–1293. <https://doi.org/10.1038/sj.bjc.6605576>
- Graham, K., Ge, X., de Las Morenas, A., Tripathi, A., & Rosenberg, C. L. (2011). Gene expression profiles of estrogen receptor-positive and estrogen receptor-negative breast cancers are detectable in histologically normal breast epithelium. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, 17(2), 236–246. <https://doi.org/10.1158/1078-0432.CCR-10-1369>
- Grivennikov, S. I., Greten, F. R., & Karin, M. (2010). Immunity, Inflammation, and Cancer. *Cell*. <https://doi.org/10.1016/j.cell.2010.01.025>
- Guyon, I., Weston, J., Stephen, B., & Vapnik, V. (2002). A gene selection method for cancer classification. *Machine Learning*. <https://doi.org/10.1155/2012/586246>
- Haibe-Kains, B. (2010). Classification models for breast cancer molecular subtyping: What is the best candidate for a translation into clinic? *Women's Health*, 6(5), 623–625. <https://doi.org/10.2217/whe.10.50>

- Han, Y., Gao, S., Muegge, K., Zhang, W., & Zhou, B. (2015). Advanced applications of RNA sequencing and challenges. *Bioinformatics and Biology Insights*, 9, 29–46. <https://doi.org/10.4137/BBI.S28991>
- Henderson, I. C., & Patek, A. J. (1998). The relationship between prognostic and predictive factors in the management of breast cancer. *Breast Cancer Research and Treatment*, 52(1–3), 261–288. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10066087>
- Hira, Z. M., & Gillies, D. F. (2015). A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics*, 2015, 198363. <https://doi.org/10.1155/2015/198363>
- Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., ... Perou, C. M. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, 7, 1–12. <https://doi.org/10.1186/1471-2164-7-96>
- Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4), e15. <https://doi.org/10.1093/nar/gng015>
- Irizarry, R. A., Hobbs, B., Collin, F., Beazer-barclay, Y. D., Antonellis, K. J., & Speed, T. P. (2003). Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data, (June), 249–264.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>
- Jain, A. K., & Dubes, R. C. (1988). Clustering Methods and Algorithms. In *Algorithms for Clustering Data*.
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. <https://doi.org/10.1007/s13928716>
- Jorns, J. M., Healy, P., & Zhao, L. (2013). Review of Estrogen Receptor, Progesterone Receptor, and HER-2/neu Immunohistochemistry Impacts on Treatment for a Small Subset of Breast Cancer Patients Transferring Care to

- Another Institution. *Arch Pathol Lab Med*, 137, 1660–1663.  
<https://doi.org/10.5858/arpa.2012-0670-OA>
- Kapp, A. V., Jeffrey, S. S., Langerød, A., Børresen-Dale, A.-L., Han, W., Noh, D.-Y., ... Tibshirani, R. (2006). Discovery and validation of breast cancer subtypes. *BMC Genomics*, 7, 231. <https://doi.org/10.1186/1471-2164-7-231>
- Kapur, K., Xing, Y., Ouyang, Z., & Wong, W. H. (2007). Exon arrays provide accurate assessments of gene expression. *Genome Biology*.  
<https://doi.org/10.1186/gb-2007-8-5-r82>
- Keen, J. C., & Moore, H. M. (2015). The genotype-tissue expression (GTEx) project: Linking clinical data with molecular analysis to advance personalized medicine. *Journal of Personalized Medicine*, 5(1), 22–29.  
<https://doi.org/10.3390/jpm5010022>
- Kendzioriski, C., Irizarry, R. A., Chen, K.-S., Haag, J. D., & Gould, M. N. (2005). On the utility of pooling biological samples in microarray experiments. *Proceedings of the National Academy of Sciences*.  
<https://doi.org/10.1073/pnas.0500607102>
- Kennecke, H., Yerushalmi, R., Woods, R., Cheang, M. C. U., Voduc, D., Speers, C. H., ... Gelmon, K. (2010). Metastatic behavior of breast cancer subtypes. *Journal of Clinical Oncology*, 28(20), 3271–3277.  
<https://doi.org/10.1200/JCO.2009.25.9820>
- Kerr, M. K. (2003). Design Considerations for Efficient and Effective Microarray Studies. *Biometrics*. <https://doi.org/10.1111/j.0006-341X.2003.00096.x>
- King, B. (1967). Step-Wise Clustering Procedures. *Journal of the American Statistical Association*. <https://doi.org/10.1080/01621459.1967.10482890>
- Kirby, M. K., Ramaker, R. C., Gertz, J., Davis, N. S., Johnston, B. E., Oliver, P. G., ... Myers, R. M. (2016). RNA sequencing of pancreatic adenocarcinoma tumors yields novel expression patterns associated with long-term survival and reveals a role for ANGPTL4. *Molecular Oncology*.  
<https://doi.org/10.1016/j.molonc.2016.05.004>
- Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., ... Palchik, J. D. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), 61–70.



<https://doi.org/10.1038/nature11412>

- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8–17. <https://doi.org/10.1016/j.csbj.2014.11.005>
- Ku, C. S., & Roukos, D. H. (2013). From next-generation sequencing to nanopore sequencing technology: Paving the way to personalized genomic medicine. *Expert Review of Medical Devices*. <https://doi.org/10.1586/erd.12.63>
- Kukurba, K. R., & Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols*. <https://doi.org/10.1101/pdb.top084970>
- Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*. <https://doi.org/10.1016/j.bdq.2015.02.001>
- Law, C. W., Alhamdoosh, M., Su, S., Smyth, G. K., & Ritchie, M. E. (2016). RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research*, 5, 1408. <https://doi.org/10.12688/f1000research.9005.2>
- Lee, J. W., Lee, J. B., Park, M., & Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics and Data Analysis*. <https://doi.org/10.1016/j.csda.2004.03.017>
- Lee, S. M. Y., Li, M. L. Y., Tse, Y. C., Leung, S. C. L., Lee, M. M. S., Tsui, S. K. W., ... Wayne, M. M. Y. (2002). Paeoniae Radix, a Chinese herbal extract, inhibit hepatoma cells growth by inducing apoptosis in a p53 independent pathway. *Life Sciences*. [https://doi.org/10.1016/S0024-3205\(02\)01962-8](https://doi.org/10.1016/S0024-3205(02)01962-8)
- Levy, S. E., & Myers, R. M. (2016). Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics*, 17(1), 95–115. <https://doi.org/10.1146/annurev-genom-083115-022413>
- Li, T., Zhang, C., & Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/bth267>
- Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features.



- Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btt656>
- Libbrecht, M. W., Noble, W. S., & Genome, E. (2017). HHS Public Access, *16*(6), 321–332. <https://doi.org/10.1038/nrg3920>.Machine
- Liotta, L., & Petricoin, E. (2000). Molecular profiling of human cancer. *Nature Reviews Genetics*. <https://doi.org/10.1038/35049567>
- Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*. <https://doi.org/10.1109/TKDE.2005.66>
- Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., ... Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. <https://doi.org/10.1038/ng.2653>
- Lopez de Mantaras, R., & Armengol, E. (1998). Machine learning from examples: Inductive and Lazy methods. *Data & Knowledge Engineering*, *25*(1–2), 99–123. [https://doi.org/10.1016/S0169-023X\(97\)00053-0](https://doi.org/10.1016/S0169-023X(97)00053-0)
- Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., ... Golub, T. R. (2005). MicroRNA expression profiles classify human cancers. *Nature*, *435*(7043), 834–838. <https://doi.org/10.1038/nature03702>
- MacQueen, J. (1967). Some Methods for classification and Analysis of Multivariate Observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability 1967*. <https://doi.org/citeulike-article-id:6083430>
- Mao, J., & Jain, A. K. (1996). A self-organizing network for hyperellipsoidal clustering (HEC). *IEEE Transactions on Neural Networks*. <https://doi.org/10.1109/72.478389>
- Marchionni, L., Wilson, R. F., Wolff, A. C., Marinopoulos, S., Parmigiani, G., Bass, E. B., & Goodman, S. N. (2008). Systematic review: gene expression profiling assays in early-stage breast cancer. *Annals of Internal Medicine*, *148*(5), 358–369. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/18252678>
- Mccall, M. N., Bolstad, B. M., & Irizarry, R. A. (2009). FROZEN ROBUST MULTI-ARRAY ANALYSIS ( fRMA ) FROZEN ROBUST MULTI-ARRAY ANALYSIS ( fRMA ). *Public Health*, (May 2009). <https://doi.org/10.1186/1471-2105-12-369>

- McCall, M. N., Bolstad, B. M., & Irizarry, R. A. (2010a). Frozen robust multiarray analysis (fRMA). *Biostatistics*, *11*(2), 242–253. <https://doi.org/10.1093/biostatistics/kxp059>
- McCall, M. N., Bolstad, B. M., & Irizarry, R. A. (2010b). Frozen robust multiarray analysis (fRMA). *Biostatistics*, *11*(2), 242–253. <https://doi.org/10.1093/biostatistics/kxp059>
- McCall, M. N., Jaffee, H. A., Zelisko, S. J., Sinha, N., Hooiveld, G., Irizarry, R. A., & Zilliox, M. J. (2014). The Gene Expression Barcode 3.0: Improved data processing and mining tools. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkt1204>
- McCall, M. N., Uppal, K., Jaffee, H. A., Zilliox, M. J., & Irizarry, R. A. (2011). The gene expression barcode: Leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research*, *39*(SUPPL. 1), 1011–1015. <https://doi.org/10.1093/nar/gkq1259>
- Miller, M. B., & Tang, Y. W. (2009). Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical Microbiology Reviews*. <https://doi.org/10.1128/CMR.00019-09>
- Mitchell, T. (1997). Does machine learning really work? *AI Magazine*. <https://doi.org/10.1609/aimag.v18i3.1303>
- Moasser, M. M. (2007). The oncogene HER2: Its signaling and transforming functions and its role in human cancer pathogenesis. *Oncogene*. <https://doi.org/10.1038/sj.onc.1210477>
- Mook, S., Van't Veer, L. J., Rutgers, E. J. T., Piccart-Gebhart, M. J., & Cardoso, F. (2007). Individualization of therapy using mammaprint<sup>®</sup><sup>™</sup>: From development to the MINDACT trial. *Cancer Genomics and Proteomics*, *4*(3), 147–156.
- Murtagg, F. (1984). Complexities of Hierarchical Clustering Algorithms State Of The Art. *Computational Statistics*. <https://doi.org/10.1002/anie.201402890>
- Nevins, J. R., Huang, E. S., Dressman, H., Pittman, J., Huang, A. T., & West, M. (2003). Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Human Molecular Genetics*, *12*(suppl 2), R153–

- R157. <https://doi.org/10.1093/hmg/ddg287>
- Ng, P. C., & Kirkness, E. F. (2010). Whole genome sequencing. *Methods in Molecular Biology*. <https://doi.org/10.1007/978-1-60327-367-1-12>
- Okoniewski, M. J., & Miller, C. J. (2008). Comprehensive analysis of affymetrix exon arrays using bioConductor. *PLoS Computational Biology*, 4(2), 2–7. <https://doi.org/10.1371/journal.pcbi.0040006>
- Olopade, O. I., Grushko, T. A., Nanda, R., & Huo, D. (2008). Advances in Breast Cancer: Pathways to Personalized Medicine. *Clinical Cancer Research*, 14(24), 7988–7999. <https://doi.org/10.1158/1078-0432.CCR-08-1211>
- Osuna, E., Freund, R., & Girosit, F. (2000). Training support vector machines: an application to face detection. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (November 2014), 130–136. <https://doi.org/10.1109/CVPR.1997.609310>
- Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., ... Wolmark, N. (2004). Oncotype DX- 1st Paik paper (A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England Journal of Medicine*. <https://doi.org/10.1056/NEJMoa041588>
- Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., ... Brazma, A. (2009). ArrayExpress update - From an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkn889>
- Parmigiani, G., Garrett, E. S., Anbazhagan, R., & Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*. <https://doi.org/10.1111/1467-9868.00358>
- Patnayak, R., Jena, A., Rukmangadha, N., Chowhan, A. K., Sambasivaiah, K., Phaneendra, B. V., & Reddy, M. K. (2015). Hormone receptor status (estrogen receptor, progesterone receptor), human epidermal growth factor-2 and p53 in South Indian breast cancer patients: A tertiary care center experience. *Indian Journal of Medical and Paediatric Oncology: Official Journal of Indian Society of Medical & Paediatric Oncology*, 36(2), 117–122. <https://doi.org/10.4103/0971-5851.158844>

- Pereira, M. A., Imada, E. L., & Guedes, R. L. M. (2017). RNA-seq: Applications and Best Practices. *Applications of RNA-Seq and Omics Strategies - From Microorganisms to Human Health*. <https://doi.org/10.5772/intechopen.69250>
- Perou, C. M., Sørli, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C., ... Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, *406*(6797), 747–752. <https://doi.org/10.1038/35021093>
- Pietras, K., & Östman, A. (2010). Hallmarks of cancer: Interactions with the tumor stroma. *Experimental Cell Research*. <https://doi.org/10.1016/j.yexcr.2010.02.045>
- Rahman, M., Jackson, L. K., Johnson, W. E., Li, D. Y., Bild, A. H., & Piccolo, S. R. (2015). Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics (Oxford, England)*, *31*(22), 3666–3672. <https://doi.org/10.1093/bioinformatics/btv377>
- Ramasamy, A., Mondry, A., Holmes, C. C., & Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Medicine*. <https://doi.org/10.1371/journal.pmed.0050184>
- Ransohoff, D. F. (2005). Bias as a threat to the validity of cancer molecular-marker research. *Nature Reviews Cancer*. <https://doi.org/10.1038/nrc1550>
- Raykov, Y. P., Boukouvalas, A., Baig, F., & Little, M. A. (2016). What to do when K-means clustering fails: A simple yet principled alternative algorithm. *PLoS ONE*, *11*(9), 1–28. <https://doi.org/10.1371/journal.pone.0162259>
- Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*. <https://doi.org/10.1016/j.gpb.2015.08.002>
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Ruiz, R., Riquelme, J. C., & Aguilar-Ruiz, J. S. (2006). Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognition*. <https://doi.org/10.1016/j.patcog.2005.11.001>
- Rung, J., & Brazma, A. (2013). Reuse of public genome-wide gene expression

- data. *Nature Reviews Genetics*, 14(2), 89–99.  
<https://doi.org/10.1038/nrg3394>
- Saeyns, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507–2517.  
<https://doi.org/10.1093/bioinformatics/btm344>
- Saligan, L. N., Fernández-Martínez, J. L., de Andrés-Galiana, E. J., & Sonis, S. (2014). Supervised classification by filter methods and recursive feature elimination predicts risk of radiotherapy-related fatigue in patients with prostate cancer. *Cancer Informatics*, 13(January), 141–152.  
<https://doi.org/10.4137/CIN.S19745>
- Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*. [https://doi.org/10.1016/0022-2836\(75\)90213-2](https://doi.org/10.1016/0022-2836(75)90213-2)
- Scherer, A. (2009). Variation, Variability, Batches and Bias in Microarray Experiments: An Introduction. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*.  
<https://doi.org/10.1002/9780470685983.ch1>
- Shukla, S., Evans, J. R., Malik, R., Feng, F. Y., Dhanasekaran, S. M., Cao, X., ... Chinnaiyan, A. M. (2017). Development of a RNA-seq based prognostic signature in lung adenocarcinoma. *Journal of the National Cancer Institute*.  
<https://doi.org/10.1093/jnci/djw200>
- Siegel, R., Naishadham, D., Jemal, A., Lockwood, W. W., Siegel, R., Naishadham, D., ... Bally, M. (2012). Cancer statistics, 2012. *CA: A Cancer Journal for Clinicians*, 62(1), 10–29.  
<https://doi.org/10.3322/caac.20138>. Available
- Siegfried, J. M., Lin, Y., Diergaard, B., Lin, H.-M., Dacic, S., Pennathur, A., ... Stabile, L. P. (2015). Expression of PAM50 Genes in Lung Cancer: Evidence that Interactions between Hormone Receptors and HER2/HER3 Contribute to Poor Outcome. *Neoplasia*, 17(11), 817–825.  
<https://doi.org/10.1016/J.NEO.2015.11.002>
- Sima, C., & Dougherty, E. R. (2006). What should be expected from feature selection in small-sample settings. *Bioinformatics*, 22(19), 2430–2436.



- <https://doi.org/10.1093/bioinformatics/btl407>
- Sneath, P. H. A., & Sokal, R. R. (1962). Numerical taxonomy. *Nature*.  
<https://doi.org/10.1038/193855a0>
- Sommer, C., & Gerlich, D. W. (2013). Machine learning in cell biology – teaching computers to recognize phenotypes. *Journal of Cell Science*, *126*(24), 5529–5539. <https://doi.org/10.1242/jcs.123604>
- Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., ... Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, *100*(14), 8418–8423. <https://doi.org/10.1073/pnas.0932692100>
- Sotiriou, C., & Piccart, M. J. (2007). Taking gene-expression profiling to the clinic: When will molecular signatures become relevant to patient care? *Nature Reviews Cancer*. <https://doi.org/10.1038/nrc2173>
- Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., & Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, *21*(5), 631–643. <https://doi.org/10.1093/bioinformatics/bti033>
- Stears, R. L., Martinsky, T., & Schena, M. (2003). Trends in microarray analysis. *Nature Medicine*. <https://doi.org/10.1038/nm0103-140>
- Straver, M. E., Glas, A. M., Hannemann, J., Wesseling, J., Van De Vijver, M. J., Rutgers, E. J. T., ... Rodenhuis, S. (2010). The 70-gene signature as a response predictor for neoadjuvant chemotherapy in breast cancer. *Breast Cancer Research and Treatment*, *119*(3), 551–558. <https://doi.org/10.1007/s10549-009-0333-1>
- Su, A. I., Schultz, P. G., Welsh, J. B., Sapinoso, L. M., Kern, S. G., Dimitrov, P., ... Frierson, H. F. (2001). Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Research*. <https://doi.org/doi:10.1063/1.4772594>
- Sun, D., Chen, J., Liu, L., Zhao, G., Dong, P., Wu, B., ... Dong, L. (2018). Establishment of a 12-gene expression signature to predict colon cancer prognosis. *PeerJ*, *6*, e4942. <https://doi.org/10.7717/peerj.4942>
- Tarca, A. L., Carey, V. J., Chen, X. wen, Romero, R., & Drăghici, S. (2007).



- Machine learning and its applications to biology. *PLoS Computational Biology*, 3(6). <https://doi.org/10.1371/journal.pcbi.0030116>
- Terzić, J., Grivennikov, S., Karin, E., & Karin, M. (2010). Inflammation and Colon Cancer. *Gastroenterology*. <https://doi.org/10.1053/j.gastro.2010.01.058>
- Tewhey, R., Nakano, M., Wang, X., Pabón-Peña, C., Novak, B., Giuffrè, A., ... Frazer, K. A. (2009). Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biology*, 10(10), R116. <https://doi.org/10.1186/gb-2009-10-10-r116>
- The Cancer Genome Atlas Program - National Cancer Institute. (n.d.). Retrieved March 11, 2019, from <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga?redirect=true>
- The Genotype-Tissue Expression (GTEx) project Data Portal. (n.d.). Retrieved from <https://gtexportal.org/home/datasets>
- Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Współczesna Onkologia*, 1A, A68–A77. <https://doi.org/10.5114/wo.2014.47136>
- Toole, M. J., Kidwell, K. M., & Van Poznak, C. (2014). Oncotype Dx results in multiple primary breast cancers. *Breast Cancer: Basic and Clinical Research*. <https://doi.org/10.4137/BCBCR.S13727>
- Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btp120>
- Trevino, V., Falciani, F., & Barrera-Saldaña, H. (2007). DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research. *Molecular Medicine*. <https://doi.org/10.2119/2006-00107.Trevino>
- Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., & Altman, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics (Oxford, England)*, 18(11), 1454–1461. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12424116>
- Turashvili, G., & Brogi, E. (2017). Tumor Heterogeneity in Breast Cancer. *Tumor Heterogeneity in Breast Cancer. Front. Med*, 4, 227.

<https://doi.org/10.3389/fmed.2017.00227>

- Ulrich, H. G. K. (1999). Pairwise classification and support vector machines. In *Advances in kernel methods*.
- Vapnik, V. (1998). *Statistical learning theory*. 1998. Complexity. <https://doi.org/10.1002/cplx.10094>
- Wagstaf, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained K-means Clustering with Background Knowledge. *Eighteenth International Conference on Machine Learning*, 577–584. <https://doi.org/10.1109/TPAMI.2002.1017616>
- Wan, L., Pantel, K., & Kang, Y. (2013). Tumor metastasis: Moving new biological insights into the clinic. *Nature Medicine*. <https://doi.org/10.1038/nm.3391>
- Wan, M., Wang, J., Gao, X., & Sklar, J. (2014). RNA Sequencing and its Applications in Cancer Diagnosis and Targeted Therapy. *North American Journal of Medicine and Science*, 7(4), 156–162. <https://doi.org/10.7156/najms.2014.0704156>
- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 57–63. <https://doi.org/10.1038/nrg2484>
- Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*. <https://doi.org/10.1080/01621459.1963.10500845>
- Wei, I. H., Shi, Y., Jiang, H., Kumar-Sinha, C., & Chinnaiyan, A. M. (2014). RNA-Seq Accurately Identifies Cancer Biomarker Signatures to Distinguish Tissue of Origin. *Neoplasia*, 16(11), 918–927. <https://doi.org/10.1016/j.neo.2014.09.007>
- Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., ... Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, 6(1), 100. <https://doi.org/10.12688/f1000research.10571.2>
- Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B.,

- ... Delorenzi, M. (2008). Meta-analysis of gene expression profiles in breast cancer: Toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Research*, 10(4), 1–11. <https://doi.org/10.1186/bcr2124>
- Yang, Y. H., Buckley, M. J., & Speed, T. P. (2001). Analysis of cDNA microarray images. *Briefings in Bioinformatics*. <https://doi.org/10.1093/bib/2.4.341>
- Yao, F., Zhang, C., Du, W., Liu, C., & Xu, Y. (2015). Identification of gene-expression signatures and protein markers for breast cancer grading and staging. *PLoS ONE*, 10(9), 1–17. <https://doi.org/10.1371/journal.pone.0138213>
- Yip, K. Y., Cheng, C., & Gerstein, M. (2013). Machine learning and genome annotation: A match meant to be? *Genome Biology*, 14(5). <https://doi.org/10.1186/gb-2013-14-5-205>
- Yu, L., & Liu, H. (2004). *Efficient Feature Selection via Analysis of Relevance and Redundancy*. *Journal of Machine Learning Research* (Vol. 5). Retrieved from <http://www.jmlr.org/papers/volume5/yu04a/yu04a.pdf>
- Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. *Designs*, 2(2), 13. <https://doi.org/10.3390/designs2020013>
- Zhang, Y.-H., Huang, T., Chen, L., Xu, Y., Hu, Y., Hu, L.-D., ... Kong, X. (2017). Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets. *Oncotarget*, 8(50), 87494–87511. <https://doi.org/10.18632/oncotarget.20903>
- Zhang, Y., Sieuwerts, A. M., McGreevy, M., Casey, G., Cufer, T., Paradiso, A., ... Foekens, J. A. (2009). The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy. *Breast Cancer Research and Treatment*. <https://doi.org/10.1007/s10549-008-0183-2>
- Zhou, X., Liu, K. Y., & Wong, S. T. C. (2004). Cancer classification and prediction using logistic regression with Bayesian gene selection. *Journal of Biomedical Informatics*. <https://doi.org/10.1016/j.jbi.2004.07.009>
- Zilliox, M. J., & Irizarry, R. A. (2007a). A gene expression bar code for

microarray data. *Nature Methods*, 4(11), 911–913.  
<https://doi.org/10.1038/nmeth1102>

Zilliox, M. J., & Irizarry, R. A. (2007b). A gene expression bar code for microarray data. *Nature Methods*. <https://doi.org/10.1038/nmeth1102>



UNIVERSITY *of the*  
WESTERN CAPE

## 7 Appendices

### Appendix I

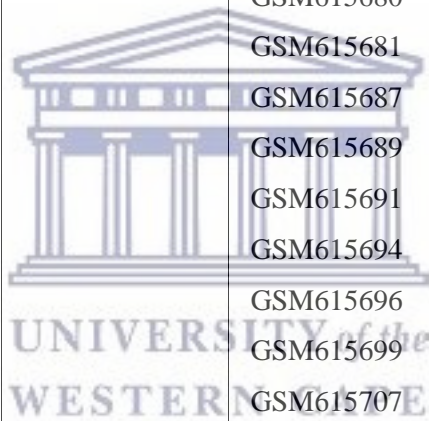
Table 7. 1: Microarray breast tissue samples curated

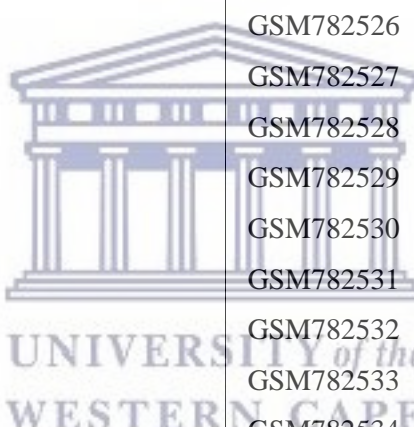
Tissue Type	GEO* Series	GEO Accession Numbers	GEO Platform (Affymetrix™)
Normal Epithelium	GSE20437	GSM512539 GSM512540 GSM512541 GSM512542 GSM512543 GSM512544 GSM512545 GSM512546 GSM512547 GSM512548 GSM512549 GSM512550 GSM512551 GSM512552 GSM512553 GSM512554 GSM512555 GSM512556	GPL96**
	GSE9574	GSM241999 GSM242000 GSM242001 GSM242002 GSM242003 GSM242004 GSM242005 GSM242006 GSM242007 GSM242008	

		GSM242009 GSM242010 GSM242011 GSM242012 GSM242013 GSM242014 GSM242015 GSM242016 GSM242017 GSM242018	
Normal Duct	GSE5764	GSM134584 GSM134588 GSM134687 GSM134690 GSM134693 GSM134696 GSM134699 GSM134702 GSM134705 GSM134708	GPL570***
Normal Lobe	GSE5764	GSM134586 GSM134589 GSM134688 GSM134691 GSM134694 GSM134697 GSM134700 GSM134703 GSM134706 GSM134709	GPL570
Triple Negative Breast Tumour	GSE25065	GSM615637 GSM615639 GSM615640 GSM615644	GPL96



GSM615650  
GSM615651  
GSM615657  
GSM615658  
GSM615660  
GSM615661  
GSM615666  
GSM615667  
GSM615668  
GSM615671  
GSM615672  
GSM615674  
GSM615677  
GSM615680  
GSM615681  
GSM615687  
GSM615689  
GSM615691  
GSM615694  
GSM615696  
GSM615699  
GSM615707  
GSM615712  
GSM615714  
GSM615715  
GSM615716  
GSM615727  
GSM615733  
GSM615739  
GSM615742  
GSM615744  
GSM615746  
GSM615757  
GSM615762  
GSM615764



		GSM615766 GSM615769 GSM615773 GSM615776 GSM615780 GSM615784 GSM615785 GSM615823 GSM615824 GSM615827	
	GSE31519	 GSM782523 GSM782524 GSM782525 GSM782526 GSM782527 GSM782528 GSM782529 GSM782530 GSM782531 GSM782532 GSM782533 GSM782534 GSM782535 GSM782536 GSM782537 GSM782538 GSM782539 GSM782540 GSM782541 GSM782542 GSM782543	
Estrogen-Positive Breast Tumour	GSE25065	GSM615638 GSM615648 GSM615656	GPL96

	GSM615669 GSM615688 GSM615701 GSM615702 GSM615703 GSM615708 GSM615709 GSM615718 GSM615725 GSM615728 GSM615730 GSM615732 GSM615736 GSM615737 GSM615741 GSM615748 GSM615754 GSM615761 GSM615767 GSM615774 GSM615782 GSM615783 GSM615796 GSM615809 GSM615819 GSM615822
GSE23988	GSM590841 GSM590843 GSM590844 GSM590845 GSM590846 GSM590847 GSM590849 GSM590854

		GSM590856 GSM590857 GSM590859 GSM590861	
	GSE22093	GSM549241 GSM549247 GSM549258 GSM549259 GSM549261 GSM549264 GSM549266 GSM549269 GSM549270 GSM549272	
Her2-Positive Breast Tumour	GSE42822	GSM105051 GSM105051 GSM105060 GSM105068 GSM105068 GSM105062 GSM105064 GSM105065 GSM105068 GSM105062 GSM105065 GSM105067 GSM105069 GSM105060 GSM105063 GSM105064 GSM105065 GSM105067	GPL96
	GSE37946	GSM930525 GSM930526	

		GSM930527 GSM930528 GSM930530 GSM930531 GSM930533 GSM930534 GSM930535 GSM930539 GSM930541 GSM930542 GSM930543 GSM930544 GSM930546 GSM930547 GSM930549 GSM930551 GSM930552 GSM930555 GSM930557 GSM930558 GSM930559 GSM930560 GSM930561 GSM930563 GSM930564 GSM930566 GSM930568 GSM930569 GSM930571 GSM930574	
Primary Breast Tumour	GSE21217	GSM530556 GSM530557 GSM530558 GSM530559	GPL96

		GSM530560 GSM530561 GSM530562 GSM530563 GSM530564 GSM530565 GSM530566	
	GSE5462	GSM125123 GSM125125 GSM125127 GSM125129 GSM125131 GSM125133 GSM125135 GSM125137 GSM125139 GSM125141	
Inflammatory Breast Tumour	GSE5847	GSM136373 GSM136374 GSM136375 GSM136376 GSM136377 GSM136378 GSM136379 GSM136380 GSM136381 GSM136382 GSM136383 GSM136384 GSM136385	GPL570
	GSE22597	GSM560663 GSM560664 GSM560665 GSM560666	GPL96



GSM560667

GSM560668

GSM560669

GSM560670

GSM560671

GSM560672

GSM560673

GSM560674

GSM560675

GSM560676

GSM560677

GSM560678

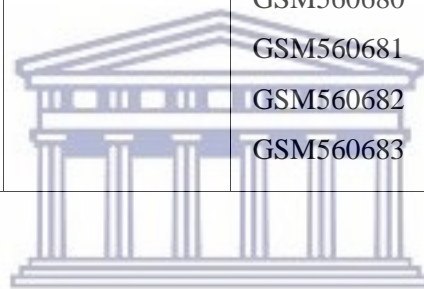
GSM560679

GSM560680

GSM560681

GSM560682

GSM560683



UNIVERSITY *of the*  
WESTERN CAPE

## Appendix II

### GTEEx Data Curated

Table 7. 2: Normal breast tissue samples filtered from GTEEx Version 7 Gene Counts file

Training Data – GTEEx Sample ID	Sample Name
GTEX-1117F-2826-SM-5GZXL	GTEEx-Norm1
GTEX-111YS-1926-SM-5GICC	GTEEx-Norm2
GTEX-1122O-1226-SM-5H113	GTEEx-Norm3
GTEX-117XS-1926-SM-5GICO	GTEEx-Norm4
GTEX-117YX-1426-SM-5H12H	GTEEx-Norm5
GTEX-1192X-2326-SM-5987X	GTEEx-Norm6
GTEX-11DXW-0626-SM-5N9ER	GTEEx-Norm7
GTEX-11DXY-2326-SM-5GICW	GTEEx-Norm8
GTEX-11DXZ-1926-SM-5GZZL	GTEEx-Norm9
GTEX-11DZ1-0326-SM-5N9BN	GTEEx-Norm10
GTEX-11EI6-0626-SM-5985T	GTEEx-Norm11
GTEX-11EM3-1326-SM-5N9C6	GTEEx-Norm12
GTEX-11EMC-2026-SM-5A5JV	GTEEx-Norm13
GTEX-11EQ9-1826-SM-5Q5AJ	GTEEx-Norm14
GTEX-11GS4-2126-SM-5A5KR	GTEEx-Norm15
GTEX-11GSO-1926-SM-5A5K3	GTEEx-Norm16
GTEX-11I78-2226-SM-5PNYA	GTEEx-Norm17
GTEX-11LCK-2426-SM-5HL5F	GTEEx-Norm18
GTEX-11NSD-0926-SM-5N9DR	GTEEx-Norm19
GTEX-11NV4-2026-SM-5N9DG	GTEEx-Norm20
GTEX-11O72-2126-SM-5N9FO	GTEEx-Norm21
GTEX-11OF3-1926-SM-59889	GTEEx-Norm22
GTEX-11ONC-2126-SM-5HL6E	GTEEx-Norm23
GTEX-11P7K-0726-SM-5EGKX	GTEEx-Norm24
GTEX-11P81-1926-SM-5BC53	GTEEx-Norm25
GTEX-11P82-1326-SM-5HL62	GTEEx-Norm26
GTEX-11PRG-0826-SM-5EQ6A	GTEEx-Norm27
GTEX-11TT1-2126-SM-5GU5Y	GTEEx-Norm28
GTEX-11TUW-1826-SM-5BC5D	GTEEx-Norm29
GTEX-11WQC-1726-SM-5GU4W	GTEEx-Norm30
GTEX-11WQK-2426-SM-5GU5C	GTEEx-Norm31
GTEX-11ZTT-2326-SM-5EQLG	GTEEx-Norm32
GTEX-11ZUS-0826-SM-5FQUY	GTEEx-Norm33

GTEX-1211K-1926-SM-5EQLB	GTEX-Norm34
GTEX-1269C-2426-SM-5FQSN	GTEX-Norm35
GTEX-12BJ1-1826-SM-5HL9N	GTEX-Norm36
GTEX-12KS4-0126-SM-5Q5A5	GTEX-Norm37
GTEX-12WSK-2226-SM-5GCO5	GTEX-Norm38
GTEX-12WSM-1726-SM-5BC6J	GTEX-Norm39
GTEX-12WSN-1326-SM-5GCNT	GTEX-Norm40
GTEX-12ZZX-1126-SM-5EGKB	GTEX-Norm41
GTEX-13113-1726-SM-5GCOO	GTEX-Norm42
GTEX-1313W-0826-SM-5EQ4T	GTEX-Norm43
GTEX-1314G-1226-SM-5BC6D	GTEX-Norm44
GTEX-131XW-0726-SM-5EGK3	GTEX-Norm45
GTEX-131YS-0626-SM-5EGKL	GTEX-Norm46
GTEX-132AR-0826-SM-5EGK6	GTEX-Norm47
GTEX-132NY-0826-SM-5K7Y7	GTEX-Norm48
GTEX-133LE-1726-SM-5K7VQ	GTEX-Norm49
GTEX-1399U-1826-SM-5PNZ1	GTEX-Norm50
GTEX-139T6-1626-SM-5PNYZ	GTEX-Norm51
GTEX-139T8-0826-SM-5L3DE	GTEX-Norm52
GTEX-139TU-0626-SM-5KM3X	GTEX-Norm53
GTEX-13CF2-2026-SM-5K7VI	GTEX-Norm54
GTEX-13CF3-2126-SM-5IFJP	GTEX-Norm55
GTEX-13D11-1026-SM-5IJFB	GTEX-Norm56
GTEX-13FHO-0826-SM-5L3E8	GTEX-Norm57
GTEX-13FTW-1426-SM-5LZWZ	GTEX-Norm58
GTEX-13FTX-1126-SM-5N9EN	GTEX-Norm59
GTEX-13FTY-2226-SM-5J1ND	GTEX-Norm60
GTEX-13N11-1726-SM-5J1OJ	GTEX-Norm61
GTEX-13N1W-0626-SM-5MR4U	GTEX-Norm62
GTEX-13NZ8-0126-SM-5IJCT	GTEX-Norm63
GTEX-13NZ9-1026-SM-5MR5K	GTEX-Norm64
GTEX-13NZB-2126-SM-5MR4Y	GTEX-Norm65
GTEX-13O3O-0826-SM-5K7WE	GTEX-Norm66
GTEX-13O3P-0826-SM-5L3DH	GTEX-Norm67
GTEX-13O3Q-2226-SM-5KM4O	GTEX-Norm68
GTEX-13O61-1826-SM-5KM4I	GTEX-Norm69
GTEX-13OW5-2226-SM-5L3HC	GTEX-Norm70

GTEX-13OW8-0226-SM-5K7UP	GTEEx-Norm71
GTEX-13PL6-2926-SM-5L3I2	GTEEx-Norm72
GTEX-13PVQ-1026-SM-5KM3M	GTEEx-Norm73
GTEX-13PVR-2226-SM-7DHKP	GTEEx-Norm74
GTEX-13QIC-2326-SM-5LU5N	GTEEx-Norm75
GTEX-13QJ3-0826-SM-7DHKK	GTEEx-Norm76
GTEX-13SLW-2526-SM-62LDQ	GTEEx-Norm77
GTEX-13SLX-2326-SM-5ZZWE	GTEEx-Norm78
GTEX-13VXU-2826-SM-664MA	GTEEx-Norm79
GTEX-13W3W-1226-SM-5LU4H	GTEEx-Norm80
GTEX-13W46-0826-SM-5LU3H	GTEEx-Norm81
GTEX-144GL-2026-SM-5LU3O	GTEEx-Norm82
GTEX-144GM-0926-SM-5O994	GTEEx-Norm83
GTEX-145LT-0726-SM-5S2VM	GTEEx-Norm84
GTEX-145ME-1526-SM-5Q5F2	GTEEx-Norm85
GTEX-145MF-2226-SM-7EPIR	GTEEx-Norm86
GTEX-145MN-1926-SM-5SIAI	GTEEx-Norm87
GTEX-145MO-0826-SM-5NQBL	GTEEx-Norm88
GTEX-146FH-0826-SM-5SI8T	GTEEx-Norm89
GTEX-14753-2426-SM-5LU8U	GTEEx-Norm90
GTEX-147F4-2826-SM-5NQBN	GTEEx-Norm91
GTEX-14A5I-0726-SM-5TDEB	GTEEx-Norm92
GTEX-14AS3-1626-SM-5S2OY	GTEEx-Norm93
GTEX-14B4R-1226-SM-5TDDT	GTEEx-Norm94
GTEX-14BMU-1626-SM-5TDE7	GTEEx-Norm95
GTEX-14BMV-0626-SM-793AU	GTEEx-Norm96
GTEX-14DAR-1326-SM-7DUEG	GTEEx-Norm97
GTEX-14E6C-1326-SM-62LEQ	GTEEx-Norm98
GTEX-14E6E-1326-SM-5S2NR	GTEEx-Norm99
GTEX-14E7W-0826-SM-62LEJ	GTEEx-Norm100
<b>Test Data – GTEEx Sample ID</b>	<b>Sample Name</b>
GTEX-14H4A-2526-SM-5YYAY	GTEEx-Norm101
GTEX-14ICK-2426-SM-6EU27	GTEEx-Norm102
GTEX-14LLW-0626-SM-62LFC	GTEEx-Norm103
GTEX-14PHY-1926-SM-5YY95	GTEEx-Norm104
GTEX-14PJ4-2126-SM-6ETZJ	GTEEx-Norm105

GTEX-14PJO-0726-SM-69LO8	GTEX-Norm106
GTEX-14PKU-0426-SM-6EU1P	GTEX-Norm107
GTEX-14PN4-0626-SM-62LFP	GTEX-Norm108
GTEX-15DCZ-0726-SM-69LOV	GTEX-Norm109
GTEX-15EOM-5019-SM-793DK	GTEX-Norm110
GTEX-15ER7-1626-SM-6PAMZ	GTEX-Norm111
GTEX-15ETS-0626-SM-7KUMX	GTEX-Norm112
GTEX-15FZZ-0726-SM-7KUFZ	GTEX-Norm113
GTEX-15G19-2126-SM-6M48J	GTEX-Norm114
GTEX-15RJE-2626-SM-7KFT1	GTEX-Norm115
GTEX-15SHW-1326-SM-6PAL8	GTEX-Norm116
GTEX-15UF6-0126-SM-6PAMB	GTEX-Norm117
GTEX-15UF7-0726-SM-6M46D	GTEX-Norm118
GTEX-16BQI-1026-SM-7KUEA	GTEX-Norm119
GTEX-16NGA-0826-SM-718AF	GTEX-Norm120
GTEX-16YQH-2826-SM-6PAMY	GTEX-Norm121
GTEX-17EUY-1926-SM-7DUF6	GTEX-Norm122
GTEX-17EVP-0226-SM-79OND	GTEX-Norm123
GTEX-17EVQ-0426-SM-7LG57	GTEX-Norm124
GTEX-17F96-2426-SM-7IGLN	GTEX-Norm125
GTEX-17F97-2526-SM-7EWDV	GTEX-Norm126
GTEX-17F98-0526-SM-79OK5	GTEX-Norm127
GTEX-17GQL-0326-SM-7LG5U	GTEX-Norm128
GTEX-17HG3-0126-SM-7IGNH	GTEX-Norm129
GTEX-17HGU-1326-SM-79OKB	GTEX-Norm130
GTEX-17HHE-1426-SM-7EPH4	GTEX-Norm131
GTEX-17HHY-0926-SM-793C1	GTEX-Norm132
GTEX-17JCI-0726-SM-7EPH1	GTEX-Norm133
GTEX-17KNJ-2026-SM-7LG53	GTEX-Norm134
GTEX-17MF6-0326-SM-7EPH5	GTEX-Norm135
GTEX-17MFQ-0926-SM-7LG4S	GTEX-Norm136
GTEX-183FY-1126-SM-7DHLJ	GTEX-Norm137
GTEX-183WM-0726-SM-7LTAA	GTEX-Norm138
GTEX-18465-2026-SM-718AP	GTEX-Norm139
GTEX-18A6Q-0926-SM-7LG4N	GTEX-Norm140
GTEX-18A7A-0726-SM-7LTAI	GTEX-Norm141
GTEX-18A7B-2626-SM-7LG55	GTEX-Norm142

GTEX-18D9A-1526-SM-7LG4J	GTEEx-Norm143
GTEX-18QFQ-0826-SM-718AX	GTEEx-Norm144
GTEX-1A3MV-1626-SM-731C1	GTEEx-Norm145
GTEX-1A3MX-2726-SM-718B6	GTEEx-Norm146
GTEX-1A8G7-2426-SM-731AK	GTEEx-Norm147
GTEX-1AMEY-1026-SM-718AA	GTEEx-Norm148
GTEX-1AX8Z-0926-SM-731AW	GTEEx-Norm149
GTEX-1AX9I-0726-SM-73KWV	GTEEx-Norm150
GTEX-1AX9J-1126-SM-731B7	GTEEx-Norm151
GTEX-1B8KE-1226-SM-73KWK	GTEEx-Norm152
GTEX-1B8KZ-1526-SM-7DUG7	GTEEx-Norm153
GTEX-1B932-0826-SM-73KXG	GTEEx-Norm154
GTEX-1B933-2526-SM-7IGO5	GTEEx-Norm155
GTEX-1B97I-0426-SM-79OL7	GTEEx-Norm156
GTEX-1B97J-0426-SM-79OLQ	GTEEx-Norm157
GTEX-1BAJH-0826-SM-7EWEF	GTEEx-Norm158
GTEX-1C64O-0726-SM-7DUFU	GTEEx-Norm159
GTEX-1C6VQ-0426-SM-79OOX	GTEEx-Norm160
GTEX-1C6VS-0726-SM-7EPHF	GTEEx-Norm161
GTEX-1CAMR-1426-SM-793BO	GTEEx-Norm162
GTEX-1CAMS-1426-SM-7IGPM	GTEEx-Norm163
GTEX-1CB4E-0226-SM-79OLW	GTEEx-Norm164
GTEX-1CB4G-2326-SM-79OOI	GTEEx-Norm165
GTEX-1CB4J-1826-SM-7EWF9	GTEEx-Norm166
GTEX-1E2YA-2726-SM-7IGPW	GTEEx-Norm167
GTEX-1EKGG-2626-SM-7IGPY	GTEEx-Norm168
GTEX-1EU9M-2826-SM-7EWFH	GTEEx-Norm169
GTEX-PSDG-1626-SM-48TCQ	GTEEx-Norm170
GTEX-Q2AG-0326-SM-48U1O	GTEEx-Norm171
GTEX-QDT8-0626-SM-48TYW	GTEEx-Norm172
GTEX-QEG5-0726-SM-4R1JQ	GTEEx-Norm173
GTEX-QEL4-2126-SM-447AE	GTEEx-Norm174
GTEX-QMRM-1626-SM-4R1KV	GTEEx-Norm175
GTEX-QVJO-1826-SM-447C9	GTEEx-Norm176
GTEX-R3RS-0626-SM-48FE1	GTEEx-Norm177
GTEX-R53T-1526-SM-48FEK	GTEEx-Norm178
GTEX-R55D-0826-SM-48FEA	GTEEx-Norm179



GTEX-REY6-2426-SM-48FF5	GTEEx-Norm180
GTEX-RUIJ-0626-SM-4WAWY	GTEEx-Norm181
GTEX-RU72-0626-SM-46MUI	GTEEx-Norm182
GTEX-RUSQ-2026-SM-4GIAK	GTEEx-Norm183
GTEX-RWS6-1926-SM-47JXY	GTEEx-Norm184
GTEX-S32W-2026-SM-4AD6E	GTEEx-Norm185
GTEX-S33H-0326-SM-4AD6N	GTEEx-Norm186
GTEX-S341-1526-SM-4AD6K	GTEEx-Norm187
GTEX-S4P3-1326-SM-4AD6V	GTEEx-Norm188
GTEX-S4Q7-1126-SM-4AD6R	GTEEx-Norm189
GTEX-S4UY-0726-SM-4AD6X	GTEEx-Norm190
GTEX-S7SE-0826-SM-4AT4D	GTEEx-Norm191
GTEX-SE5C-2126-SM-4BRUJ	GTEEx-Norm192
GTEX-T2IS-1526-SM-32QPR	GTEEx-Norm193
GTEX-T2YK-2226-SM-32QPT	GTEEx-Norm194
GTEX-T5JC-2126-SM-32PMO	GTEEx-Norm195
GTEX-T5JW-2026-SM-4DM63	GTEEx-Norm196
GTEX-T6MN-0726-SM-32PML	GTEEx-Norm197
GTEX-T6MO-0326-SM-32QOK	GTEEx-Norm198
GTEX-TKQ1-0226-SM-33HB5	GTEEx-Norm199
GTEX-TKQ2-1826-SM-33HB2	GTEEx-Norm200
GTEX-TML8-1226-SM-32QON	GTEEx-Norm201
GTEX-TMMY-0726-SM-33HBE	GTEEx-Norm202
GTEX-U3ZH-1426-SM-4DXSR	GTEEx-Norm203
GTEX-U3ZN-1926-SM-4DXSG	GTEEx-Norm204
GTEX-U412-1826-SM-4DXTJ	GTEEx-Norm205
GTEX-U8XE-0826-SM-4E3J1	GTEEx-Norm206
GTEX-UJHI-1426-SM-3DB9C	GTEEx-Norm207
GTEX-UPK5-2326-SM-3P5Z8	GTEEx-Norm208
GTEX-UTHO-1026-SM-3GAF7	GTEEx-Norm209
GTEX-V955-2026-SM-3GAFA	GTEEx-Norm210
GTEX-VJWN-0726-SM-3GIJ8	GTEEx-Norm211
GTEX-VUSG-2226-SM-4KKZO	GTEEx-Norm212
GTEX-W5X1-2326-SM-3GIL6	GTEEx-Norm213
GTEX-WFON-1826-SM-3GILG	GTEEx-Norm214
GTEX-WI4N-1426-SM-3LK7H	GTEEx-Norm215
GTEX-WOFL-0826-SM-3MJG1	GTEEx-Norm216

GTEX-WRHU-0326-SM-3MJFY	GTEEx-Norm217
GTEX-WXYG-2226-SM-4E3IM	GTEEx-Norm218
GTEX-WY7C-2726-SM-3NB3P	GTEEx-Norm219
GTEX-WYBS-0926-SM-3NM94	GTEEx-Norm220
GTEX-WYJK-1326-SM-3NB2T	GTEEx-Norm221
GTEX-WYVS-1726-SM-3NMAY	GTEEx-Norm222
GTEX-X15G-1626-SM-3NMB3	GTEEx-Norm223
GTEX-X261-0626-SM-3NMD9	GTEEx-Norm224
GTEX-X4EP-2926-SM-3P5YQ	GTEEx-Norm225
GTEX-X4XY-0926-SM-4E3JD	GTEEx-Norm226
GTEX-XBED-1626-SM-47JYN	GTEEx-Norm227
GTEX-XGQ4-0926-SM-4AT4U	GTEEx-Norm228
GTEX-XMD1-0826-SM-4AT52	GTEEx-Norm229
GTEX-XMD2-0926-SM-4WWEF	GTEEx-Norm230
GTEX-XMK1-1126-SM-4IHJ8	GTEEx-Norm231
GTEX-XOT4-0726-SM-4GIAW	GTEEx-Norm232
GTEX-XQ3S-1326-SM-4BOPQ	GTEEx-Norm233
GTEX-XQ8I-2426-SM-4WAXY	GTEEx-Norm234
GTEX-XUW1-2326-SM-4BOO5	GTEEx-Norm235
GTEX-XUZC-1626-SM-4BRVP	GTEEx-Norm236
GTEX-XV7Q-2326-SM-4BRVZ	GTEEx-Norm237
GTEX-XYKS-1326-SM-4BRUN	GTEEx-Norm238
GTEX-Y111-2026-SM-4SOJA	GTEEx-Norm239
GTEX-Y114-2026-SM-4TT7L	GTEEx-Norm240
GTEX-Y3I4-1526-SM-4TT7K	GTEEx-Norm241
GTEX-Y3IK-2326-SM-4WWDT	GTEEx-Norm242
GTEX-Y5LM-1726-SM-4VDSX	GTEEx-Norm243
GTEX-Y5V5-2126-SM-4WWFO	GTEEx-Norm244
GTEX-Y5V6-2126-SM-4WWFX	GTEEx-Norm245
GTEX-Y8E4-1626-SM-5S2MW	GTEEx-Norm246
GTEX-Y8LW-1626-SM-5IFHX	GTEEx-Norm247
GTEX-Y9LG-1426-SM-5IFJZ	GTEEx-Norm248
GTEX-YB5E-1726-SM-5IFJ3	GTEEx-Norm249
GTEX-YB5K-1626-SM-5IFIN	GTEEx-Norm250
GTEX-YEC3-1026-SM-5IFI5	GTEEx-Norm251
GTEX-YFC4-1426-SM-5IFJG	GTEEx-Norm252
GTEX-YFCO-1826-SM-4W1YH	GTEEx-Norm253

GTEX-YJ8O-2226-SM-5IFHW	GTEEx-Norm254
GTEX-ZA64-1526-SM-5CVMD	GTEEx-Norm255
GTEX-ZAB4-2526-SM-5HL8M	GTEEx-Norm256
GTEX-ZAJG-0626-SM-5HL8X	GTEEx-Norm257
GTEX-ZC5H-2626-SM-5J2MG	GTEEx-Norm258
GTEX-ZDTT-2126-SM-5S2OJ	GTEEx-Norm259
GTEX-ZDXO-0126-SM-5S2ND	GTEEx-Norm260
GTEX-ZDYS-1126-SM-5K7UB	GTEEx-Norm261
GTEX-ZEX8-2226-SM-57WC6	GTEEx-Norm262
GTEX-ZF29-1926-SM-5S2P1	GTEEx-Norm263
GTEX-ZF2S-2026-SM-5E461	GTEEx-Norm264
GTEX-ZF3C-2326-SM-5S2MZ	GTEEx-Norm265
GTEX-ZLFU-2126-SM-4WWEV	GTEEx-Norm266
GTEX-ZLV1-1426-SM-4WWES	GTEEx-Norm267
GTEX-ZPIC-1126-SM-5BC7F	GTEEx-Norm268
GTEX-ZQG8-0726-SM-5P9H9	GTEEx-Norm269
GTEX-ZQUD-1926-SM-51MSA	GTEEx-Norm270
GTEX-ZT9W-2026-SM-51MRA	GTEEx-Norm271
GTEX-ZTTD-1026-SM-51MRD	GTEEx-Norm272
GTEX-ZTX8-1226-SM-4YCE9	GTEEx-Norm273
GTEX-ZU9S-1926-SM-5NQBP	GTEEx-Norm274
GTEX-ZUA1-1526-SM-59HLS	GTEEx-Norm275
GTEX-ZV6S-1826-SM-5NQ8D	GTEEx-Norm276
GTEX-ZV7C-1826-SM-5NQ83	GTEEx-Norm277
GTEX-ZVE2-1226-SM-5NQ8R	GTEEx-Norm278
GTEX-ZVT2-1826-SM-5NQ8W	GTEEx-Norm279
GTEX-ZVT4-1026-SM-57WC4	GTEEx-Norm280
GTEX-ZVTK-0326-SM-51MRR	GTEEx-Norm281
GTEX-ZVZQ-0826-SM-51MRF	GTEEx-Norm282
GTEX-ZWKS-2826-SM-5NQ74	GTEEx-Norm283
GTEX-ZXES-0826-SM-5E43C	GTEEx-Norm284
GTEX-ZY6K-1626-SM-5GZWV	GTEEx-Norm285
GTEX-ZYFC-0826-SM-5E44K	GTEEx-Norm286
GTEX-ZYT6-0126-SM-5E45J	GTEEx-Norm287
GTEX-ZYW4-0826-SM-5GIDG	GTEEx-Norm288
GTEX-ZZ64-1226-SM-5E43R	GTEEx-Norm289
GTEX-ZZPU-0626-SM-5E43T	GTEEx-Norm290

### Appendix III

#### TCGA Data Curated

Table 7. 3: Paired TCGA NAT and Primary Tumour Samples

Complete TCGA ID	Sample Name	Sample Name	Molecular Subtype
TCGA-BH-A18V	Norm1_NAT	Tum1_TN	Triple Negative
TCGA-BH-A1EV	Norm10_NAT	Tum10_ER	Estrogen Positive
TCGA-BH-A18P	Norm11_NAT	Tum11_ER	Estrogen Positive
TCGA-BH-A1ET	Norm12_NAT	Tum12_ER	Estrogen Positive
TCGA-BH-A0DP	Norm13_NAT	Tum13_ER	Estrogen Positive
TCGA-BH-A0E1	Norm14_NAT	Tum14_ER	Estrogen Positive
TCGA-BH-A0BJ	Norm15_NAT	Tum15_ER	Estrogen Positive
TCGA-BH-A0H7	Norm16_NAT	Tum16_ER	Estrogen Positive
TCGA-BH-A0BC	Norm17_NAT	Tum17_ER	Estrogen Positive
TCGA-BH-A0BA	Norm18_NAT	Tum18_ER	Estrogen Positive
TCGA-BH-A0DH	Norm19_NAT	Tum19_ER	Estrogen Positive
TCGA-BH-A18Q	Norm2_NAT	Tum2_TN	Triple Negative
TCGA-BH-A0H9	Norm20_NAT	Tum20_ER	Estrogen Positive
TCGA-BH-A0BV	Norm21_NAT	Tum21_ER	Estrogen Positive
TCGA-BH-A0B8	Norm22_NAT	Tum22_ER	Estrogen Positive
TCGA-BH-A0AZ	Norm23_NAT	Tum23_ER	Estrogen Positive
TCGA-BH-A0BM	Norm24_NAT	Tum24_ER	Estrogen Positive
TCGA-BH-A0BQ	Norm25_NAT	Tum25_ER	Estrogen Positive
TCGA-BH-A0BT	Norm26_NAT	Tum26_ER	Estrogen Positive
TCGA-BH-A0DG	Norm27_NAT	Tum27_ER	Estrogen Positive
TCGA-BH-A0DO	Norm28_NAT	Tum28_ER	Estrogen Positive
TCGA-BH-A0DT	Norm29_NAT	Tum29_ER	Estrogen Positive
TCGA-BH-A0E0	Norm3_NAT	Tum3_TN	Triple Negative
TCGA-BH-A0H5	Norm30_NAT	Tum30_ER	Estrogen Positive
TCGA-BH-A0HA	Norm31_NAT	Tum31_ER	Estrogen Positive
TCGA-BH-A18J	Norm32_NAT	Tum32_ER	Estrogen Positive
TCGA-BH-A18L	Norm33_NAT	Tum33_ER	Estrogen Positive
TCGA-BH-A1EW	Norm34_NAT	Tum34_TN	Triple Negative
TCGA-BH-A0AY	Norm35_NAT	Tum35_ER	Estrogen Positive

TCGA-BH-A0AU	Norm36_NAT	Tum36_ER	Estrogen Positive
TCGA-BH-A0B5	Norm37_NAT	Tum37_ER	Estrogen Positive
TCGA-BH-A1EN	Norm38_NAT	Tum38_Her2	Her2-Positive
TCGA-BH-A1FU	Norm39_NAT	Tum39_Her2	Her2-Positive
TCGA-A7-A0CE	Norm4_NAT	Tum4_TN	Triple Negative
TCGA-BH-A18K	Norm40_NAT	Tum40_ER	Estrogen Positive
TCGA-A7-A13E	Norm5_NAT	Tum5_ER	Estrogen Positive
TCGA-BH-A0B3	Norm6_NAT	Tum6_TN	Triple Negative
TCGA-BH-A0BW	Norm7_NAT	Tum7_TN	Triple Negative
TCGA-BH-A0DL	Norm8_NAT	Tum8_ER	Estrogen Positive
TCGA-E2-A158	Norm9_NAT	Tum9_TN	Triple Negative



UNIVERSITY *of the*  
WESTERN CAPE

Table 7. 4: Unpaired TCGA NAT Samples

Complete TCGA ID	Sample Name	Molecular Subtype
TCGA-E9-A1RH	Norm41_NAT	Normal - Solid Tissue
TCGA-BH-A0DQ	Norm42_NAT	Normal - Solid Tissue
TCGA-E2-A15M	Norm43_NAT	Normal - Solid Tissue
TCGA-E2-A1LS	Norm44_NAT	Normal - Solid Tissue
TCGA-BH-A209	Norm45_NAT	Normal - Solid Tissue
TCGA-BH-A0DZ	Norm46_NAT	Normal - Solid Tissue
TCGA-AC-A2FM	Norm47_NAT	Normal - Solid Tissue
TCGA-BH-A1FR	Norm48_NAT	Normal - Solid Tissue
TCGA-E9-A1RC	Norm49_NAT	Normal - Solid Tissue
TCGA-A7-A0DB	Norm50_NAT	Normal - Solid Tissue
TCGA-BH-A0DK	Norm51_NAT	Normal - Solid Tissue
TCGA-GI-A2C8	Norm52_NAT	Normal - Solid Tissue
TCGA-E2-A15K	Norm53_NAT	Normal - Solid Tissue
TCGA-E9-A1RD	Norm54_NAT	Normal - Solid Tissue
TCGA-BH-A203	Norm55_NAT	Normal - Solid Tissue
TCGA-BH-A0C0	Norm56_NAT	Normal - Solid Tissue
TCGA-BH-A18U	Norm57_NAT	Normal - Solid Tissue
TCGA-BH-A1FJ	Norm58_NAT	Normal - Solid Tissue
TCGA-E9-A1RF	Norm59_NAT	Normal - Solid Tissue
TCGA-BH-A1F2	Norm60_NAT	Normal - Solid Tissue
TCGA-E2-A15I	Norm61_NAT	Normal - Solid Tissue
TCGA-BH-A0C3	Norm62_NAT	Normal - Solid Tissue
TCGA-A7-A13G	Norm63_NAT	Normal - Solid Tissue
TCGA-E9-A1RI	Norm64_NAT	Normal - Solid Tissue
TCGA-E9-A1RB	Norm65_NAT	Normal - Solid Tissue
TCGA-A7-A0CH	Norm66_NAT	Normal - Solid Tissue
TCGA-BH-A1EO	Norm67_NAT	Normal - Solid Tissue
TCGA-E9-A1N5	Norm68_NAT	Normal - Solid Tissue
TCGA-A7-A13F	Norm69_NAT	Normal - Solid Tissue
TCGA-E9-A1N9	Norm70_NAT	Normal - Solid Tissue
TCGA-AC-A23H	Norm71_NAT	Normal - Solid Tissue
TCGA-E2-A1LB	Norm72_NAT	Normal - Solid Tissue



TCGA-E9-A1N6	Norm73_NAT	Normal - Solid Tissue
TCGA-BH-A1EU	Norm74_NAT	Normal - Solid Tissue
TCGA-A7-A0DC	Norm75_NAT	Normal - Solid Tissue
TCGA-E2-A153	Norm76_NAT	Normal - Solid Tissue
TCGA-BH-A0BZ	Norm77_NAT	Normal - Solid Tissue
TCGA-BH-A1FE	Norm78_NAT	Normal - Solid Tissue
TCGA-E9-A1R7	Norm79_NAT	Normal - Solid Tissue
TCGA-BH-A18S	Norm80_NAT	Normal - Solid Tissue
TCGA-E9-A1NF	Norm81_NAT	Normal - Solid Tissue
TCGA-BH-A208	Norm82_NAT	Normal - Solid Tissue
TCGA-E9-A1N4	Norm83_NAT	Normal - Solid Tissue
TCGA-E2-A1L7	Norm84_NAT	Normal - Solid Tissue
TCGA-BH-A1F8	Norm85_NAT	Normal - Solid Tissue
TCGA-BH-A18R	Norm86_NAT	Normal - Solid Tissue
TCGA-AC-A2FB	Norm87_NAT	Normal - Solid Tissue
TCGA-BH-A1FC	Norm88_NAT	Normal - Solid Tissue
TCGA-BH-A0HK	Norm89_NAT	Normal - Solid Tissue
TCGA-AC-A2FF	Norm90_NAT	Normal - Solid Tissue
TCGA-E2-A1IG	Norm91_NAT	Normal - Solid Tissue
TCGA-E2-A1LH	Norm92_NAT	Normal - Solid Tissue
TCGA-GI-A2C9	Norm93_NAT	Normal - Solid Tissue
TCGA-BH-A0DD	Norm94_NAT	Normal - Solid Tissue
TCGA-BH-A1FN	Norm95_NAT	Normal - Solid Tissue
TCGA-E9-A1NG	Norm96_NAT	Normal - Solid Tissue
TCGA-BH-A1FB	Norm97_NAT	Normal - Solid Tissue
TCGA-A7-A0D9	Norm98_NAT	Normal - Solid Tissue
TCGA-E2-A1BC	Norm99_NAT	Normal - Solid Tissue
TCGA-BH-A1FM	Norm100_NAT	Normal - Solid Tissue
TCGA-BH-A1FH	Norm101_NAT	Normal - Solid Tissue
TCGA-BH-A1FD	Norm102_NAT	Normal - Solid Tissue
TCGA-BH-A0BS	Norm103_NAT	Normal - Solid Tissue
TCGA-BH-A18M	Norm104_NAT	Normal - Solid Tissue
TCGA-E9-A1ND	Norm105_NAT	Normal - Solid Tissue
TCGA-BH-A0B7	Norm106_NAT	Normal - Solid Tissue

TCGA-BH-A1F0	Norm107_NAT	Normal - Solid Tissue
TCGA-BH-A1FG	Norm108_NAT	Normal - Solid Tissue
TCGA-BH-A204	Norm109_NAT	Normal - Solid Tissue
TCGA-BH-A1F6	Norm110_NAT	Normal - Solid Tissue
TCGA-BH-A0DV	Norm111_NAT	Normal - Solid Tissue
TCGA-BH-A18N	Norm112_NAT	Normal - Solid Tissue
TCGA-E9-A1NA	Norm113_NAT	Normal - Solid Tissue



UNIVERSITY *of the*  
WESTERN CAPE

Table 7. 5: Unpaired TCGA Tumour Samples

<b>Complete TCGA ID</b>	<b>Sample Name</b>	<b>Molecular Subtype</b>
TCGA-BH-A0HL	Tum135	Estrogen-Positive
TCGA-BH-A0RX	Tum64	Triple Negative
TCGA-AO-A0J4	Tum133	Triple Negative
TCGA-A7-A0DA	Tum134	Triple Negative
TCGA-D8-A142	Tum65	Triple Negative
TCGA-BH-A0HN	Tum41	Estrogen-Positive
TCGA-A2-A0T0	Tum66	Triple Negative
TCGA-A2-A0YE	Tum67	Triple Negative
TCGA-A2-A0YJ	Tum68	Estrogen-Positive
TCGA-A2-A0D0	Tum69	Triple Negative
TCGA-A2-A04U	Tum70	Triple Negative
TCGA-AO-A0J6	Tum71	Triple Negative
TCGA-A2-A0YM	Tum72	Triple Negative
TCGA-A2-A04Q	Tum73	Triple Negative
TCGA-A2-A0D2	Tum74	Triple Negative
TCGA-A2-A0SX	Tum75	Triple Negative
TCGA-AO-A0JL	Tum76	Triple Negative
TCGA-AO-A12F	Tum77	Triple Negative
TCGA-BH-A0B9	Tum78	Triple Negative
TCGA-A2-A04T	Tum79	Triple Negative
TCGA-B6-A0RT	Tum80	Triple Negative
TCGA-AO-A128	Tum81	Triple Negative
TCGA-AO-A129	Tum82	Triple Negative
TCGA-AO-A124	Tum83	Triple Negative
TCGA-B6-A0RU	Tum84	Triple Negative
TCGA-B6-A0IQ	Tum85	Triple Negative
TCGA-B6-A0IJ	Tum86	Estrogen Positive
TCGA-B6-A0X1	Tum87	Triple Negative
TCGA-B6-A0RE	Tum88	Triple Negative
TCGA-A2-A0ST	Tum89	Triple Negative
TCGA-AR-A0TP	Tum42	Estrogen Positive
TCGA-A1-A0SO	Tum90	Triple Negative

TCGA-A8-A07C	Tum91	Triple Negative
TCGA-A8-A07O	Tum92	Triple Negative
TCGA-A8-A08H	Tum93	Triple Negative
TCGA-A8-A08R	Tum94	Triple Negative
TCGA-AN-A04D	Tum95	Triple Negative
TCGA-AN-A0AL	Tum96	Triple Negative
TCGA-AN-A0AR	Tum97	Triple Negative
TCGA-AN-A0AT	Tum98	Triple Negative
TCGA-AN-A0FJ	Tum99	Estrogen Positive
TCGA-AN-A0FL	Tum100	Triple Negative
TCGA-AN-A0FX	Tum101	Triple Negative
TCGA-AN-A0G0	Tum102	Triple Negative
TCGA-AN-A0XU	Tum103	Triple Negative
TCGA-AR-A0TS	Tum104	Triple Negative
TCGA-AR-A0TU	Tum105	Triple Negative
TCGA-AR-A0U0	Tum106	Triple Negative
TCGA-AR-A0U4	Tum107	Triple Negative
TCGA-AR-A1AH	Tum43	Estrogen Positive
TCGA-AR-A1AI	Tum108	Triple Negative
TCGA-AR-A1AJ	Tum109	Estrogen Positive
TCGA-AR-A1AQ	Tum110	Triple Negative
TCGA-AR-A1AY	Tum111	Triple Negative
TCGA-BH-A0AV	Tum112	Triple Negative
TCGA-BH-A0BG	Tum113	Triple Negative
TCGA-BH-A0BL	Tum114	Triple Negative
TCGA-BH-A0WA	Tum115	Triple Negative
TCGA-BH-A18G	Tum116	Triple Negative
TCGA-C8-A12K	Tum117	Triple Negative
TCGA-C8-A12V	Tum118	Triple Negative
TCGA-C8-A131	Tum119	Triple Negative
TCGA-C8-A134	Tum120	Triple Negative
TCGA-D8-A147	Tum121	Triple Negative
TCGA-E2-A14N	Tum122	Triple Negative
TCGA-E2-A14R	Tum123	Triple Negative

TCGA-E2-A14X	Tum124	Triple Negative
TCGA-E2-A150	Tum125	Triple Negative
TCGA-E2-A159	Tum126	Triple Negative
TCGA-E2-A1AZ	Tum127	Triple Negative
TCGA-E2-A1B5	Tum128	Estrogen Positive
TCGA-A8-A08L	Tum44	Estrogen Positive
TCGA-B6-A0IK	Tum129	Triple Negative
TCGA-A8-A08J	Tum130	Estrogen Positive
TCGA-A2-A0T1	Tum137	Her2-Positive
TCGA-AO-A0J2	Tum131	Triple Negative
TCGA-BH-A0EE	Tum138	Her2-Positive
TCGA-A2-A0CY	Tum45	Estrogen Positive
TCGA-AO-A12D	Tum139	Her2-Positive
TCGA-AO-A0JE	Tum140	Her2-Positive
TCGA-A2-A0EQ	Tum141	Her2-Positive
TCGA-AO-A03L	Tum46	Estrogen Positive
TCGA-A8-A075	Tum47	Estrogen Positive
TCGA-A8-A081	Tum48	Estrogen Positive
TCGA-A8-A08X	Tum142	Her2-Positive
TCGA-A8-A094	Tum49	Estrogen Positive
TCGA-C8-A12P	Tum143	Her2-Positive
TCGA-C8-A12Z	Tum144	Her2-Positive
TCGA-C8-A137	Tum145	Her2-Positive
TCGA-E2-A14P	Tum146	Her2-Positive
TCGA-E2-A1B0	Tum147	Her2-Positive
TCGA-A2-A0CU	Tum50	Estrogen Positive
TCGA-BH-A18T	Tum132	Triple Negative
TCGA-B6-A0X4	Tum51	Estrogen Positive
TCGA-BH-A0EA	Tum52	Estrogen Positive
TCGA-BH-A18N	Tum53	Estrogen Positive
TCGA-BH-A1EU	Tum54	Estrogen Positive
TCGA-BH-A1EO	Tum63	Estrogen Positive
TCGA-B6-A0WS	Tum62	Estrogen Positive
TCGA-BH-A1ES	Tum55	Estrogen Positive

TCGA-B6-A0X0	Tum56	Estrogen Positive
TCGA-BH-A0DS	Tum57	Estrogen Positive
TCGA-AO-A0JA	Tum58	Estrogen Positive
TCGA-AO-A0JF	Tum59	Estrogen Positive
TCGA-A7-A0CD	Tum60	Estrogen Positive
TCGA-D8-A145	Tum61	Estrogen Positive
TCGA-AN-A0XW	Tum136	Estrogen Positive
TCGA-B6-A1KF	Tum148	Her2-Positive
TCGA-A2-A1G1	Tum149	Her2-Positive
TCGA-AR-A24U	Tum150	Her2-Positive
TCGA-C8-A1HK	Tum151	Her2-Positive
TCGA-E2-A1LB	Tum152	Her2-Positive



UNIVERSITY *of the*  
WESTERN CAPE