Master of Biostatistics Workplace Portfolio Project

University of Sydney

Name: Max Yan

Student ID: 311298761

November 2019

PREFACE

Project title

Gene expression and survival analysis of breast cancer subtypes defined by immunohistochemistry

Location and dates

The project was conducted at the Systems Biology Initiative, School of Biotechnology and Biomolecular Science, University of New South Wales, between February and November 2019.

Relevant BCA Units

Survival Analysis (SVA)

Bioinformatics (BIF)

Context

I am a senior staff specialist at the Department of Anatomical Pathology, NSW Health Pathology – Randwick, Prince of Wales Hospital. I am also a conjoint lecturer at the Faculty of Medicine, University of New South Wales. I am routinely involved in the diagnosis, subtyping and staging of breast cancer as part of my work. Subtyping of breast cancer by immunohistochemistry is vital for prognostication and treatment in the clinical setting. This project seeks to understand the gene expression signatures underlying the breast cancer subtypes defined by immunohistochemistry. It also investigates the effect of subtyping on survival.

Dr Kenneth Beath (Department of Mathematics and Statistics, Macquarie University) is the statistical supervisor.

The bioinformatic analyses were supervised by Dr Susan Corley and Prof Marc Wilkins (Systems Biology Initiative, School of Biotechnology and Biomolecular Science, University of New South Wales)

Student's role

I am the principal investigator, and am responsible for the conception, design and conduct of the project. This includes:

- Obtaining RNA-seq and clinico-pathological data for 105 clinical breast cancer specimens from the National Cancer Institute online database, plus data entry into a format suitable for subsequent analyses.
- Analysis of RNA-seq data via edgeR, limma and R studio, including drafting and execution of the required R scripts.
- Analysis of survival data via STATA.
- Drafting of the manuscript, including the preparation of tables, heatmaps and figures.

Reflections on learning

Communication

Communication of statistical and bioinformatics results in a clear and concise manner was of great importance. Care was taken to ensure the language used in the results section was readily understandable by the intended target audience of clinicians and pathologists. Tables were formatted and annotated in a manner to ensure accurate and concise visual representation of data. Complex statistical data, not regarded as essential for understanding of the key findings, were referred to the appendix to facilitate interpretation of results.

Work patterns/planning

Time management and organizational skills were vital. Self-imposed deadlines for specific tasks were set throughout the 10 month period, to ensure that the project was completed in a timely manner. For the bioinformatics portion of the project, a large number of data files (> 150) were generated. These had to be organized into folders and labeled with the appropriate headings to ensure that they were readily accessible for bioinformatics analysis. Throughout the project I maintained regular contact with my supervisors, who provided guidance in regards to the direction of the project. They were also vital in helping me answer many of the statistical and bioinformatics questions that arose during the project. They also gave much needed assistance with the drafting of the manuscript.

Statistical issues

For RNA-seq analysis, issues included importing, organizing, annotating, transforming, filtering and normalizing data via the edgeR package, removing heteroscedascity from count data via the *voom* function, and fitting linear models for comparisons of interest. This project has been invaluable in teaching me the R programming language, the use of the various *bioconductor* packages and the organisation of information in R via data frames and matrices. I have also gained exposure to gene set testing via online databases such as DAVID. Issues arising from the multivariable survival analysis included testing for interactions and the proportional hazards assumption, plus the identification of influential observations. This project has consolidated the STATA programming skills I have learnt during the course.

Ethical considerations

The project was carried out in accordance with NHMRC ethics guidelines. The data for the survival analyses were collected by Prof Sandra O'Toole and Dr Ewan Millar at the Garvan Institute, with ethics approval provided by St Vincent's Hospital HREC. All data received were de-identified to maintain confidentiality. De-identified data for the PAM50 and RNA-seq cohorts were obtained from publicly available databases.

Student Declaration

I declare this project is evidence of my own work, with direction and assistance provided by my project supervisors. This work has not been previously submitted for academic credit.

Max Yan

November 2019

Supervisor Statement

I can confirm that Max conducted this work independently. This project incorporated a number of analyses both bioinformaticsand survival analysis based on the bioinformatics results. He has shown an excellent ability to construct the analyses and present them. He has also been rigorous in sticking to his time-line, on what has been a large project.

Dr Kenneth Beath

November 2019

PROJECT REPORT

Title

Gene expression and survival analysis of breast cancer subtypes defined by immunohistochemistry

Abstract

Introduction:

Breast cancer in the clinical setting may be divided into four subtypes (luminal A, luminal B, HER2 and basal) by immunohistochemical (IHC) analysis. This project aims to 1) compare subtype classification by IHC vs. the PAM50 gene classifier, 2) investigate the gene expression profiles of the IHC subtypes and 3) investigate the prognostic implications of the IHC subtypes.

Method:

The concordance between the IHC and PAM50 classifiers was assessed in a group of 405 breast cancers used in a previous study by Breffer *et al*. [1]. For RNA-seq analysis, data for a group of 105 cancers subtyped by IHC were obtained from The Cancer Genome Atlas Program (TCGA), National Cancer Institute (NCI) [2]. The data were analysed with limma and edgeR to obtain the gene expression profiles for the IHC subtypes. An exploratory unsupervised hierarchical cluster analysis was performed to help understand the usefulness of these gene sets in providing information on IHC cancer subtypes. Lastly, breast cancer specific overall survival among the IHC subtypes was assessed in a cohort of 292 cancers from the Garvan Insititute.

Results:

The concordance between subtyping by PAM50 and IHC among the 405 tumours was 59%. For cancers classified as HER2 by IHC (HER2 IHC cancers), 94% (16 out of 17) were also classified as HER2 by PAM50. For basal IHC cancers the concordance with PAM50 was 86% (54 out of 63). For luminal A and luminal B IHC cancers, the

concordance with PAM50 was 59% (148 out of 251) and 28% (21 out of 74) respectively. Linear modeling of RNA-seq data showed basal IHC cancers were enriched for genes associated with tumour hypoxia, activation of the beta-catenin pathway, progression through the cell cycle and stem cell proliferation. Whereas as luminal A cancers showed down-regulation of genes associated with progression through the cell cycle and cell migration. An exploratory unsupervised hierarchical cluster analysis by the PAM50 gene set supported the usefulness of these genes in understanding the differences between the subtypes. Survival analysis using the Cox regression model showed, compared to luminal A IHC cancers, basal (HR = 4.08, 95% CI: 2.02 – 5.77), and HER2 (HR = 4.08, 95% CI: 2.14 – 7.78) cancers were associated with shorter breast cancer specific overall survival (*p* = 0.018). A tendency for poorer survival was seen for luminal B IHC cancers (HR = 2.04, 95% CI: 0.98 – 4.25) was observed.

Conclusion:

Breast cancer IHC subtypes have distinctive gene expression profiles that relate to their biological behaviour. Subtyping by IHC yields important prognostic information that may aid the clinical management of breast cancers.

Glossary

Bioconductor package	Bioconductor package provides tools for the analysis of high-throughput genomic data. It uses the R statistical programming language, is open source and allows for open development.
camera	Correlation Adjusted MEan RAnk gene set test. This test for enrichment of a particular gene annotation category among a gene set (composed of genes that are differentially expressed). Unlike previous methods, camera does not assume the genes within the set are independent (i.e. not correlated).
cDNA microarray	A cDNA microarray allows the measurement of a large number of genes simultaneously. mRNA is reverse transcribed into cDNA. The cDNA, attached with a dye, binds to a matching sequence on a spot within the array, and is visualised. It has largely been replaced by RNA-seq.
Counts per million (cpm)	In RNA-seq analysis, cpm is the count of the sequenced fragment of interest divided by the total number of reads times one million.
DAVID	Database for Annotation, Visualization and Integrated Discovery. DAVID provides a set of functional annotation tools to understand biological meaning behind large list of genes.
Distant metastasis	Breast cancers that have spread to distant sites, beyond the breast and local axillary lymph nodes.
edgeR	Empirical Analysis of Digital Gene Expression Data in R. This is a Bioconductor package used for differential expression analysis of gene expression data.
Endocrine (hormone) therapy	Treatment that stop estrogen from attaching to its receptor. An example would be the drug Tamoxifen.
Estrogen receptor	Receptor for the hormone estrogen found on normal breast cells. In breast cancer cells, binding of estrogen to its receptor promotes cell proliferation.
Formalin fixed, paraffin embedded (FFPE) tissue	Method for preserving tissue in formalin and wax. It is a cost effective method for preserving tissue at room temperature. Disadvantages include denaturation of proteins, plus the degradation of DNA and RNA, thus limiting the number of genetic tests that may be performed.
Fresh frozen tissue	Tissue is frozen in liquid nitrogen and stored in a -80 C freezer. It is the method of choice for preserving the integrity of DNA, RNA and proteins. Disadvantages include the costs and logistics involved in freezing the tissue immediately and storage in a -80 C freezer. In addition, morphology is suboptimal due to freezing artefact, rendering the specimen more difficult to interpret for the pathologist.

Functional annotation clustering	This is a tool on DAVID which uses an algorithm to group similar, redundant annotation (GO terms) from different resources into a single group, facilitating biological analysis.
GO term	Gene Ontology term. This is a defined term used to represent the properties of a particular gene product. A GO term may relate to one of three domains: 1) Molecular function performed by the gene product (e.g. adenylate cyclase activity), 2) Cellular component (e.g. ribosome) and 3) Biological process, which is a larger process accomplished by multiple molecular activities (e.g. DNA repair)
HER2	Human epidermal growth receptor 2 is a protein that appears on the surface of some breast cancer cells. This protein is involved in breast cancer cell growth and survival. Testing for HER2 may be performed via immunohistochemistry or in situ hybridisation. HER2-positive breast cancers can benefit from therapies which directly targets the HER2 receptor, such as trastuzumab.
High throughput sequencing	DNA and RNA sequences have traditionally been elucidated using a low throughput technique called Sanger sequencing. High throughput sequencing technologies are capable of sequencing multiple DNA molecules in parallel, enabling hundreds of millions of DNA molecules to be sequenced at the same time.
Immunohistochemistry (IHC)	Immunohistochemistry uses antibodies to bind to and identify specific proteins in a tissue section. Antibodies bound to the antigen on a tissue section may then be visualised under a microscope.
In situ hybridisation	In situ hybridisation uses complementary DNA strands to bind to and identify a DNA sequence of interest in a tissue section. The bound DNA strand is then visualised under a microscope. In HER2 ISH testing, this is used count the number of HER2 DNA copies in a cell. HER2 is said to be amplified if there are too many copies of HER2 DNA. HER2 amplification leads to overproduction of HER2 protein.
Ki67 index	Ki67 is a protein expressed by proliferating cells. The Ki67 index is based on the proportion of tumour cells expressing Ki67. A Ki67 index of > 14% is associated with increased proliferation and a worse prognosis.
limma	Linear Models for Microarray Data is an R/Bioconductor software package that provides an integrated solution for analysing data from gene expression experiments.
Lymph node status	Lymph (tissue fluid) from the breast drains to lymph nodes in the axilla (arm pit). This is the first place breast cancer will spread to outside of the breast. Involvement of these lymph nodes is associated with a worse prognosis.
PAM50 assay	Prediction Analysis of Microarray 50. This assay looks at the activity of 50 genes to identify the breast cancer subtype and to estimate the risk of distant recurrence.

Progesterone receptor	Receptor for the hormone progesterone. Strong expression for both estrogen and progesterone receptors in breast cancer increases the effectiveness of anti-estrogen (endocrine) therapy.
RNA-seq	RNA-seq (RNA-sequencing) is a technique that can examine the quantity and sequences of RNA in a sample using high throughput (next generation) sequencing.
Tumour grade	Breast cancers are assigned a grade out of 3 based on their appearance under a microscope. Grade 1 cancers are well differentiated (i.e. most resembles normal tissue) and are associated with a better prognosis. Grade 3 cancers are poorly differentiated and are more likely to behave aggressively.
Tumour status	This describes the tumour's size and extent of local invasion. The four stages are - T1: < 2cm in size, T2: between 2cm and 5cm, T3: > 5cm, T4: Tumour involves skin or chest wall
voom	Voom stands for variance modeling at the observational level. Linear modelling in limma assumes variance is independent of the mean. This is not the case for log-cpm counts in RNA-seq. The voom method estimates the mean-variance trend of the log-cpm counts, and uses this mean-variance relationship to predict the variance of each log-cpm value. The predicted variance is then encapsulated as an inverse weight for the log-cpm value. Application of voom to the data results in the elimination of the mean-variance relationship.

Introduction

Breast cancer is a heterogenous disease that can be divided into four intrinsic subtypes (luminal A, luminal B, HER2 and basal) with important prognostic and therapeutic implications [1-3]. Early classifiers were based on gene expression data obtained from cDNA arrays. However, the difficulties and costs associated with obtaining fresh frozen tissue required for the arrays, have limited its use to the research setting. Currently all breast cancers undergo routine testing for ER (Estrogen receptor), PR (Progesterone receptor), HER2 and Ki67 proliferative marker expression via immunohistochemistry (IHC) performed on formalin fixed, paraffin embedded (FFPE) tissue. This IHC biomarker panel offers a cost-effective (< $100) and rapid (< 1 day) alternative for breast cancer subtyping. It has also been shown to be effective in predicting response to targeted therapy and prognosis [4].

In view of the difficulties associated with obtaining frozen tissue, the PAM50 panel, composed of 50 genes (appendix 1), was developed for the subtyping of tumours using FFPE tissue. The usage of FFPE tissue, which is readily available from the pathology laboratory, has allowed increased use of gene expression analysis in the clinical setting [5]. The PAM50 Prosigna® assay (Nanostring technologies™) was approved for clinical use by the FDA in 2013. There is evidence to suggest it may add additional prognostic information to the IHC panel [6]. Despite its utility, the cost of the assay ($2900) precluded its routine use in clinical practice [7]. This may be an important issue, as a recent study by Kim et al. suggests the discordance between IHC and PAM50 subtypes may be as high as 38.4% [8].

More recently, RNA-seq data have become available for breast cancer analyses. This has provided a further source of quantitative data for the analysis of biomarkers. It has become the method of choice for gene expression analysis in view of a number of technical advantages over microarrays, such as the ability to detect novel transcripts and superior detection of genes with low expression [9]. While RNA-seq has traditionally been performed on high quality RNA derived from frozen tissue, more recent studies have suggested it may also be accurately performed on FFPE tissue [10].

This study will therefore aim to:

1) Compare intrinsic subtype classification derived from IHC vs. the PAM50 gene classifier.

2) Investigate the gene expression profiles underlying the breast cancer subtypes defined by IHC, and define any classifier genes. This will be performed using RNA-seq data.

3) Investigate the overall survival of the intrinsic breast cancer subtypes defined by IHC.

Methods

Classification of immuniohistochemical (IHC) intrinsic subtypes by ER, PR and HER2

Cases were classified into four IHC intrinsic subtypes based on immunohistochemistry for ER and PR, plus HER2 ISH (in situ hybridization) [3, 4]: luminal A (ER+ and/or PR+ and HER2-), luminal B (ER+ and/or PR+ and HER2+), HER2 (ER- and PR-, HER2+) and basal (ER-, PR- and HER2-) [4]. Tumors were considered HER2 positive if there was 3+ staining for HER2 on IHC, and/or they were amplified on ISH using a HER2: chromosome 17 ratio higher than 2.2.

Comparison of intrinsic subtype classification via IHC and PAM50 gene classifier

IHC data (ER, PR, HER2 ISH) and intrinsic subtypes as defined by the PAM50 gene classifier were obtained for a cohort of 405 breast cancers, from a publically available database (GEO accession: GSE81538). The data were derived from a previous breast biomarker study by Brueffer *et al*. [1]. The patient characteristics of this cohort are as previously described.

Table 1. Clinicopathological characteristics of 105 breast cancers with RNA-Seq data, number and percentages (in brackets) unless otherwise stated.

Characteristic	All cases	Lum A	Lum B	HER2	Basal
Age, median (range, years)	56 (29 – 90)	56 (37 -90)	57.5 (29 – 90)	55.5 (43 – 80)	48 (40 – 66)
Tumour status					
T1	22 (21.0)	10 (22)	8 (27)	3 (15)	1 (10)
T2	65 (61.9)	24 (53)	19 (63)	14 (70)	8 (80)
T3	14 (13.3)	8 (18)	3 (10)	2 (10)	1 (10)
T4	3 (2.9)	2 (4)	0	1 (5)	0
Unknown	1 (1.0)	1 (2)	0	0	0
Total	105 (100)	45 (100)	30 (100)	20 (100)	10 (100)
Lymph node status					
N0	49 (46.7)	22 (49)	14 (47)	8 (40)	5 (50)
N0 (i+)	6 (5.7)	3 (7)	1 (3)	0	2 (20)
N1	28 (26.7)	11 (2)	8 (27)	7 (35)	2 (20)
N2	10 (9.5)	5 (11)	4 (13)	1 (5)	0
N3	9 (8.6)	3 (7)	2 (7)	3 (15)	1 (10)
Unknown	3 (2.9)	1 (2)	1 (3)	1 (5)	0
Total	105 (100)	45 (100)	30 (100)	20 (100)	10 (100)
Distant metastases					
M0	96 (91.4)	42 (94)	25 (83)	20 (100)	9 (90)
M1	3 (2.9)	1 (2)	1 (3)	0	1 (10)
Unknown	6 (5.7)	2 (4)	4 (13)	0	0
Total	105 (100)	45 (100)	30 (100)	20 (100)	10 (100)
TNM Stage					
I	17 (16.2)	9 (20)	6 (20)	1 (5)	1 (10)
IIa	38 (36.2)	12 (27)	11 (37)	9 (45)	6 (60)
IIb	28 (26.7)	14 (31)	6 (20)	6 (30)	2 (20)
III	18 (17.1)	8 (18)	6 (20)	4 (20)	0
IV	3 (2.9)	1 (2)	1 (3)	0	1 (10)
Unknown	1 (1.0)	1 (2)	0	0	0
Total	105 (100)	45 (100)	20 (100)	20 (100)	10 (100)
Relapse					
No	90 (85.7)	40 (89)	27 (90)	17 (85)	6 (60)
Yes	15 (14.3)	5 (11)	3 (10)	3 (15)	4 (40)
Total	105 (100)	45 (100)	30 (100)	20 (100)	10 (100)
Death					
No	95 (90.5)	43 (96)	28 (93)	17 (85)	7 (70)
Yes	10 (9.5)	2 (4)	2 (7)	3 (15)	3 (30)
Total	105 (100)	45 (100)	30 (100)	20 (100)	10 (100)
Median follow-up (range, months)	19.3 (0 – 125.5)	19.1 (0 – 125.6)	15.0 (0 – 125.6)	25.8 (0 – 67.5)	21.1 (8.3 – 42.3)

Gene expression analysis of IHC intrinsic subtypes via RNA-Seq

RNA-Seq data (in the form of counts per million (cpm)), ER, PR and HER2 status were obtained for a second cohort of 105 breast cancers (The Cancer Genome Atlas Program (TCGA), National Cancer Institute (NCI)) [2]. Details regarding the platforms used for high-throughput sequencing may be obtained from the TCGA website [5]. As defined by IHC, there were 45 luminal A, 30 luminal B, 20 HER2 and 10 basal cancers. The available clinic-pathological data for this cohort included age at diagnosis, lymph node status, distant metastases, TNM stage, relapse-free and overall survival. These are summarized in table 1.

Analysis of RNA-Seq gene expression data

Analysis of RNA-Seq data was performed using limma (release 3.9) and edgeR (release 3.8) as previously described by Law *et al*.[6], using RStudio (version 1.2.1335) and R (version 3.6.0). Genome wide annotation data was obtained from the Bioconductor Package org.Hs.eg.db, (release 3.9)[7]. Genes with low counts were filtered out by only retaining the rows where the cpm is at least 1 in at least ten samples. Normalisation was performed by the method of trimmed mean of M-values (TMM)[8]. The normalisation factors calculated were used as a scaling factor for the library sizes. A preliminary multidimensional scaling (MDS) plot was used to explore whether the samples cluster by intrinsic IHC subtypes.

Two different linear models were used to investigate differential gene expression:

Model 1. This is a linear model on log-cpm counts with the contrasts between the four IHC subtypes being the independent variables (predictors of gene expression). Contrasts for all

six pairwise comparisons between the four subtypes were entered into the contrast matrix. The contrast matrix obtained can be seen in Table 2.

Table 2. Contrast matrix used to obtain a gene set that may predict the IHC subtype of a sample

Levels	BasalvsLuma	Basalvslumb	BasalvsHer	LumavsLumb	LumavsHer	LumbvsHer
Basal	1	1	1	0	0	0
Luma	-1	0	0	1	1	0
Lumb	0	-1	0	-1	0	1
Her	0	0	-1	0	-1	-1

Model 2. This is a linear model on log-cpm counts with one particular IHC subtype (e.g. luminal A) being the independent variable. For each subtype, the contrast matrix is composed of a pairwise comparison between the subtype of interest versus all other subtypes. For example, to obtain the gene expression signature of luminal A cancers, luminal A cancers are compared to all other cancers (Table 3).

Table 3. Contrast matrix for obtaining the gene expression signature of luminal A cancers

Levels	LAvsnonLA
LA	1
nonLA	-1

In limma, linear modeling assumes log-cpm values are normally distributed [6]. This assumption is violated by raw log-cpm data obtained by high throughput sequencing, where the variance is not independent of the mean. To eliminate the mean-variance relationship, the voom "variance modeling at the observational level" function was used to calculate weights that were incorporated into the models [9]. Plots of the mean-variance trends were performed, before and after the application of voom, to ensure this source of heteroscedascity was removed. Linear modeling in limma was carried out using the lnFit

and contrasts.fit functions. These functions fit a separate model to the expression values for each gene. Empirical Bayes moderation is then carried out by borrowing information across all genes to obtain an estimate of gene-wise variability. This approach is equivalent to shrinking the estimated sample variances towards a pooled estimate, resulting in more stable inference when the number of arrays is small [10]. Heatmaps were generated, via the heatmap.2 function in R, to explore the usefulness of the gene sets in providing information on the IHC subtypes. In this method, Euclidean distances between the samples are calculated using the transformed cpm values. Clustering along both the x and y axis is then performed via the complete agglomeration method.

Gene set testing with camera and DAVID

Gene set testing was performed by applying the *camera* method [11], to the *c2* gene signatures and the c5 gene ontology sets from the Broad Institute's MSigDB c2 collection [12]. The c2 collection is composed of curated gene sets from online databases, publications from PUBMED and knowledge of domain experts. The *camera* function performs competitive gene testing to assess whether genes in a given set are highly ranked in terms of differential expression relative to genes that are not in the set. It uses limma's modeling framework and also incorporates weights derived from *voom*. Competitive gene set testing, adjusted for inter-gene correlation, is used to obtain a *p* value, corrected for multiple tests via the false discovery rate (FDR). Functional Annotation Clustering was performed in The Database for Annotation, Visualization and Integrated Discovery (DAVID version 6.8)[13, 14]. Fisher exact tests were performed to assess whether the GO terms were more enriched in the list of differentially expressed genes.

Table 4. Clinico-pathological characteristics of 271 breast cancers

Characteristic	Number (%)
Age (yrs)	
≤ 50	101 (37.4)
> 50	169 (62.6)
Tumour size	
≤ 20mm	160 (59.3)
> 20mm	110 (40.7)
Lymph node status	
Negative	149 (55.6)
Positive	119 (44.4)
Not available	2
Grade	
1	45 (16.7)
2	102 (37.8)
3	123 (45.6)
Estrogen receptor	
Negative	82 (30.5)
Positive	187 (69.5)
Not available	1
Progesterone receptor	
Negative	113 (41.9)
Positive	157 (58.1)
HER2 FISH	
Not-amplified	219 (81.1)
Amplified	51 (18.9)
Endocrine therapy	
No	134 (49.6)
Yes	136 (50.4)
Chemotherapy	
No	164 (60.7)
Yes	106 (39.3)

Patient cohort for survival analysis

Two hundred and ninety-two (292) invasive breast cancers, with survival and treatment data, were obtained from the Garvan Institute, courtesy of Prof Sandra O'Toole and A/Prof Ewan Millar. None of the patients had distant metastatic disease at the time of diagnosis. Twenty-one cases were excluded due to the absence of complete data for estrogen receptor (ER), progesterone receptor (PR) or HER2 amplification status. The final cohort was composed of 271 cancers. This study has ethics committee approval (HREC SVH H94/080 and 06336 H00036).

The clinico-pathological data available for survival analysis included: age at diagnosis (≤50 or > 50 years of age), tumour size (≤ 20mm or > 20mm), axillary lymph node status (positive or negative), Elston and Ellis grade (out of 3), endocrine therapy (Yes or No), chemotherapy (Yes or No), ER, PR and HER2 status (positive or negative), relapse free survival (defined by distant metastasis) and overall survival (defined as death from breast cancer). The median follow-up period was 87 months (defined as time from diagnosis to death or last follow-up). The clinico-pathological characteristics of the cohort are summarised in Table 4.

Statistical analysis of survival data

The distribution of the clinico-pathological characteristics between the four breast cancer subtypes was assessed via a chi square test. Kaplan Meier curves for overall survival, stratified by breast cancer subtypes, and by clinico-pathological characteristics were charted. A multivariable analysis, using the Cox proportional hazards model, was performed to assess differences in overall survival between the four breast cancer subtypes. For this analysis, luminal A was used as the reference subtype as previous published data suggests this subtype may have the most favourable prognosis [4]. Six other standard clinico-pathological variables in the multivariable model included age, tumour size, lymph node status, grade, endocrine therapy and chemotherapy. Interactions between breast cancer subtypes and clinic-pathological variables were assessed. Scaled Schoenfeld residuals were used to test for the validity of the proportional hazards assumption. Cox-Snell residuals were calculated and a plot of the Nelson-Aalen estimate of the cumulative hazard function vs. the Cox-Snell residuals was performed to test overall goodness of fit.

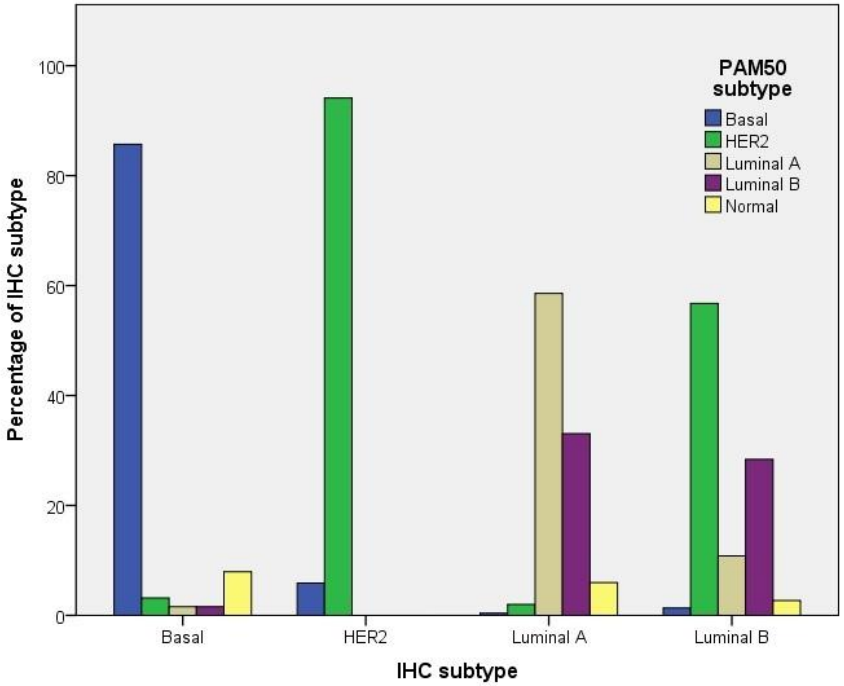
Influential observations were identified using DFBETA approximation of Cook's distances.

Results

PART 1. COMPARISON OF INTRINSIC SUBTYPE CLASSIFICATION VIA IHC AND THE PAM50 GENE CLASSIFIER

For the 405 tumours in the study by Brueffer *et al*. [1], the distribution of the PAM50 subtypes within each IHC subtype is shown in figure 1. Discordance between IHC and PAM50 subtyping was seen in 166 out of 405 cancers (41.0%). Agreement was greatest for HER2 cancers as defined by IHC, where 16 out of 17 (sensitivity = 94%, 95% CI: 71 - 99%) cancers were also defined as HER2 by PAM50. For basal cancers defined by IHC, agreement with the PAM50 subtype was seen in 54 out of 63 cancers (sensitivity = 86%, 95% CI: 75 – 93%). For cancers defined as luminal A by IHC, the majority of cases (148 out of 251 cancers, sensitivity = 59%, 95% CI: 52 – 65%) were concordant with the PAM50 classifier. Of the 103 cases defined as luminal A by IHC, but not by the PAM50 classifier, the majority (n = 83) were reclassified as luminal B by PAM50. Discordance was greatest for luminal B cancers as defined by IHC, where a minority of cases (21 out of 74, sensitivity = 29%, 95% CI: 19 – 40%) were concordant with the PAM50 classifier. Of the 53 discordant luminal B IHC cases, the majority were reclassified as HER2 (42 out of 53), followed by luminal A (8 out of 53) on the PAM50 classifier.

Figure 1. Distribution of PAM50 subtypes within each immunohistochemistry subtype



PART 2. RNA-SEQ ANALYSIS OF BREAST CANCER IHC SUBTYPES

Using RNA-Seq data, a preliminary multidimensional scaling (MDS) plot of all 105 cancers was performed to explore whether the samples cluster by intrinsic IHC subtypes (figure 2). The 10 basal type cancers were found to form a basal dominated cluster on the MDS plot. The majority of the luminal A cancers (30 out of 45) were also found to form a luminal A dominated cluster. The remaining 15 luminal A cancers overlapped with the other three subtypes. Similarly, the majority of the HER2 cancers (15 out of 20) formed a HER2 dominated cluster, with a minority (5 out of 15) overlapping with the basal cluster. For luminal B cancers, most of the cases (17 out of 20) formed a loose cluster interposed between the luminal A and HER2 clusters. There were two outlying luminal B cancers within the luminal A cluster (LumB-14 and 19) and one in the basal cluster (LumB-13).

Figure 2. Preliminary multidimensional scaling plot of all 105 cancers with RNA-Seq data, colour coded by IHC subtype, red = luminal A (n = 45), green = luminal B (n = 30), blue = HER2 (n = 20), black = basal (n = 10).

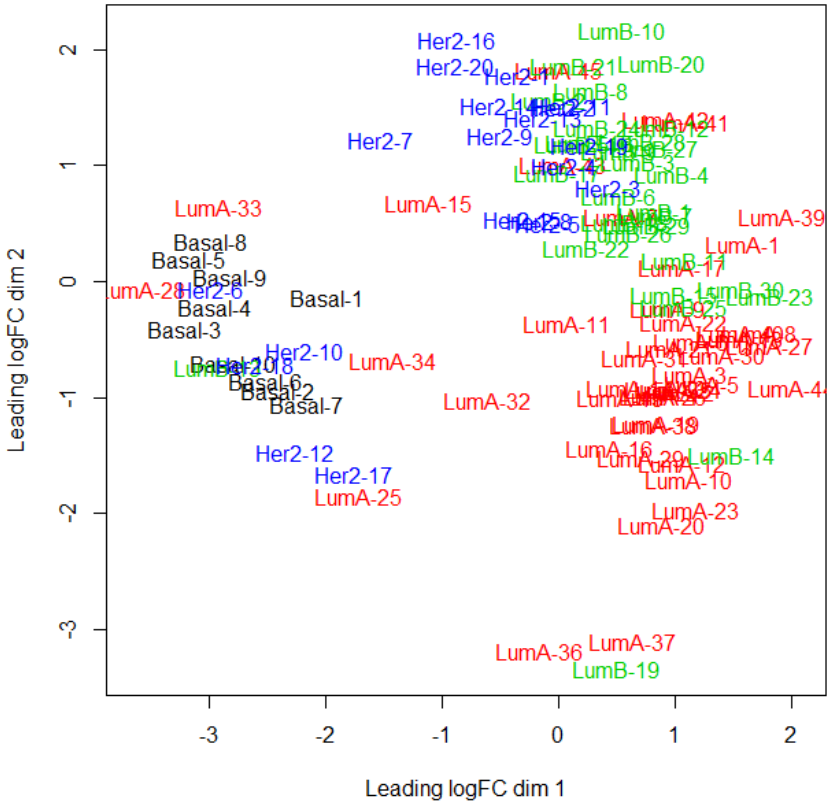
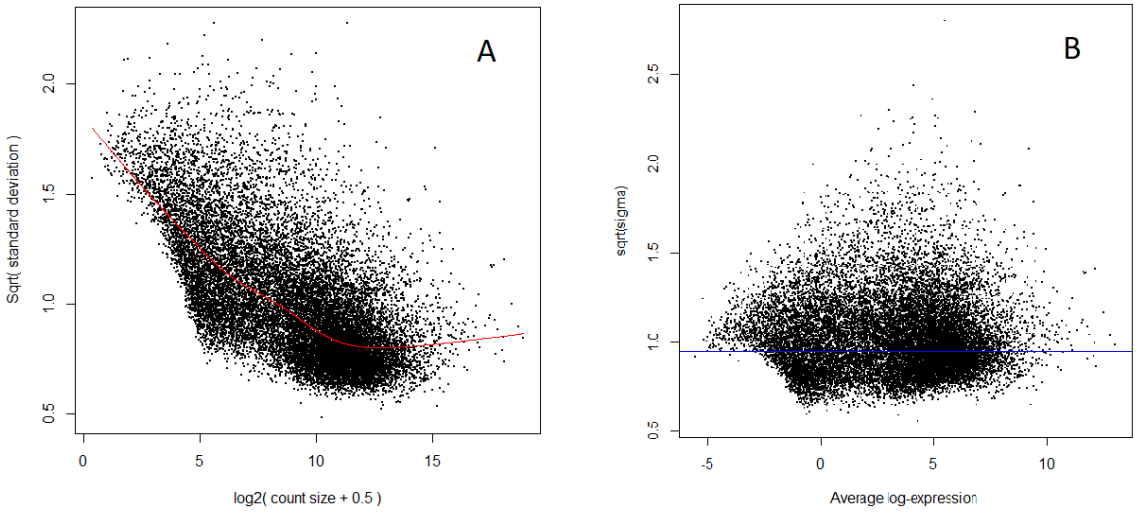


Figure 3. The residual variances of the log-CPM values in the linear model are plotted against the means for each gene. (A) Before voom is applied to the data. (B) After voom is applied to the data.



In limma, linear modelling assumes that log-cpm values are normally distributed [6]. The mean variance relationship of log-cpm values was therefore explored using a voom plot (figure 3). Prior to the application of voom, a distinct relationship existed between the means and variances. As can be seen in figure 3a, a decrease in the mean log-cpm count is associated with a rise in variance. The voom method estimates the mean-variance trend of the log-cpm counts, and uses this mean-variance relationship to predict the variance of each log-cpm value. The predicted variance is then encapsulated as an inverse weight for the log-cpm value [9]. As shown in figure 3b, application of voom to the data results in the elimination of the mean-variance relationship.

Figure 4. Heatmap of log-CPM values of all 105 breast cancers using the contrast matrix from model 1 (see text above), for 105 breast cancers. The differentially expressed genes were chosen based on their adjusted *p* value being < 0.001.

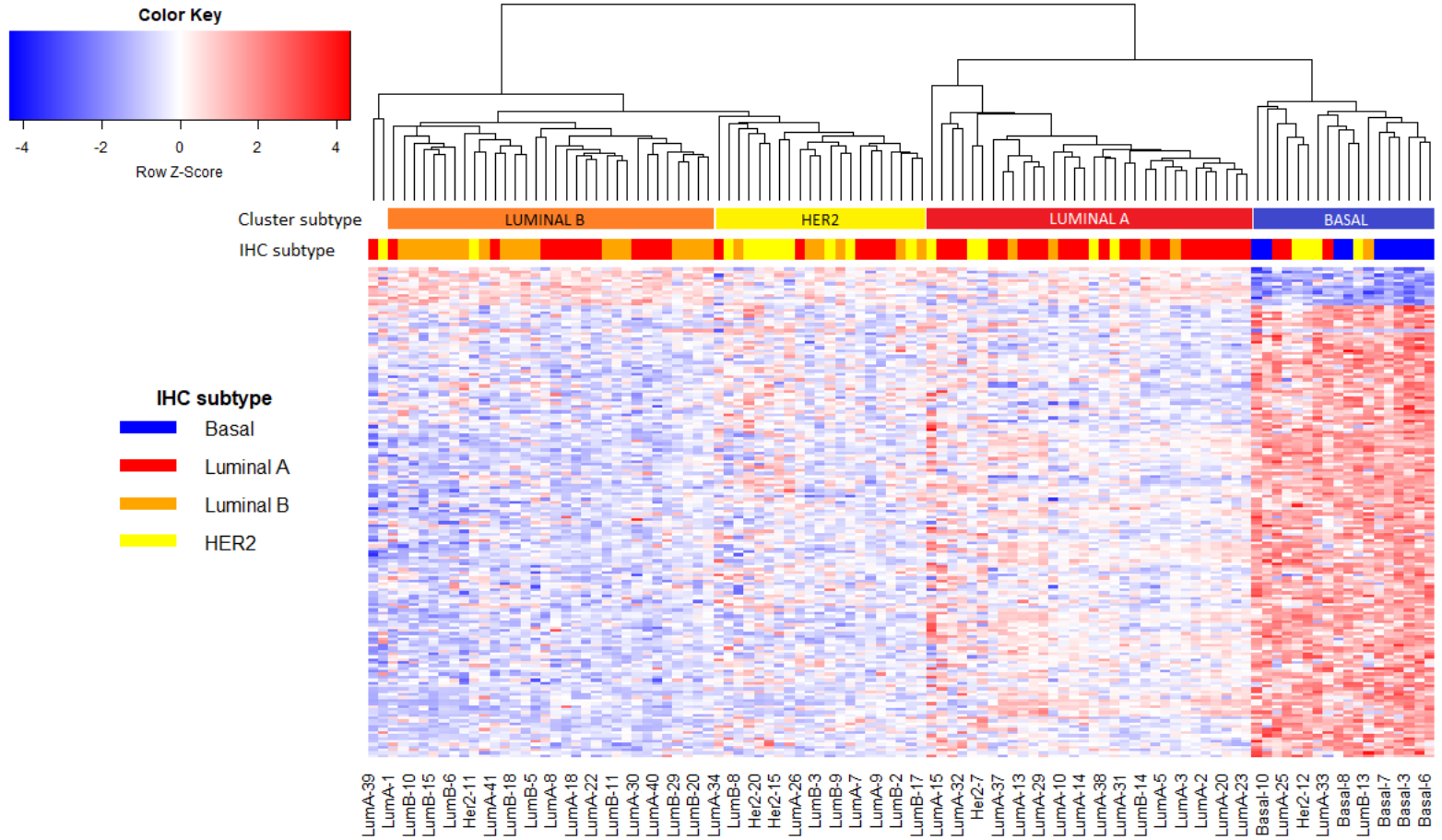
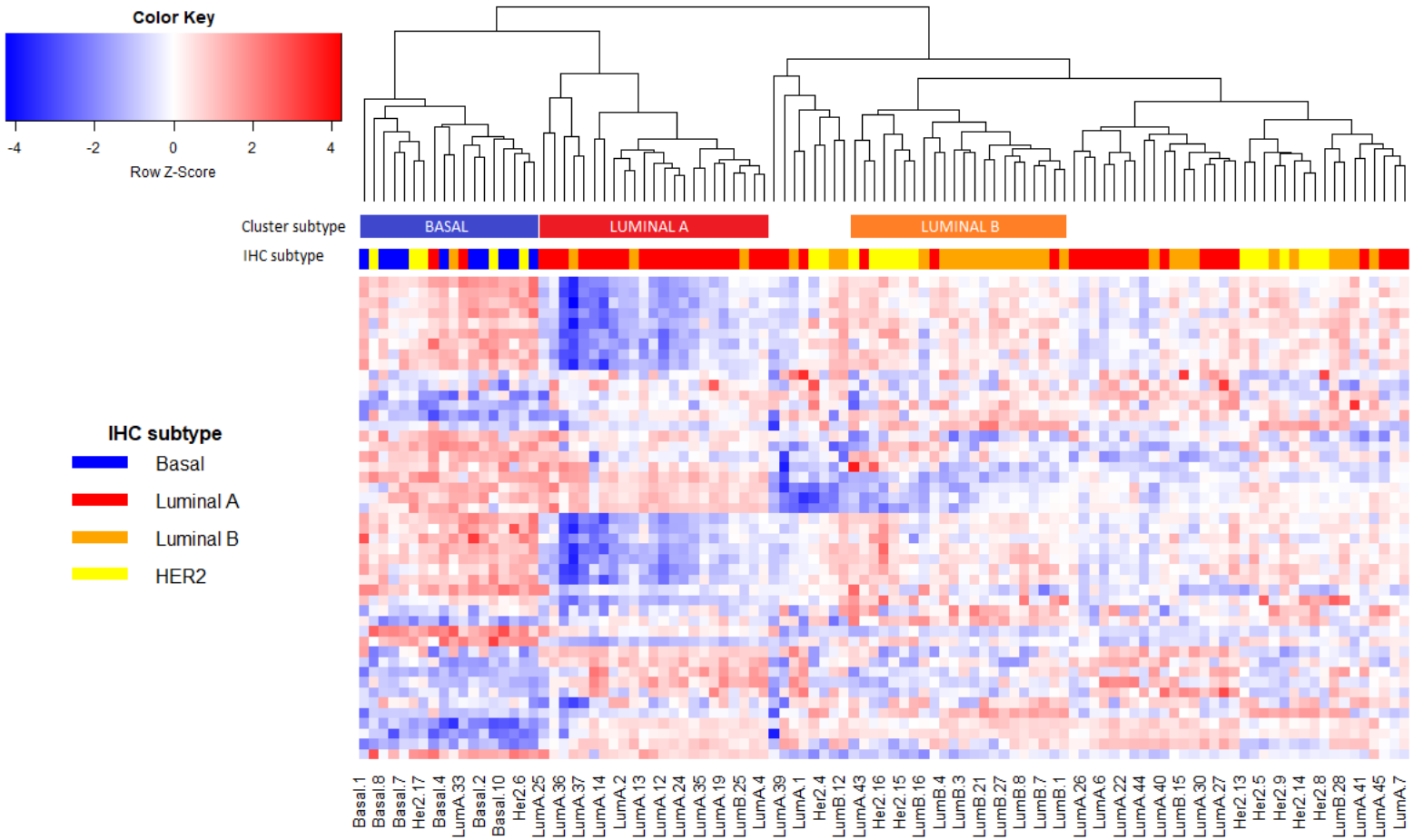


Figure 5. Heatmap of log-CPM values for the PAM50 gene set, for all 105 breast cancers.



Differentially expressed genes among the IHC subtypes

The number of significantly up- and down-regulated genes, with an adjusted *p* value of < 0.05, between the subtypes is listed in table 5.

Table 5. Numbers of up- and down-regulated genes, with an adjusted *p* value of < 0.05, between the subtypes.

	Basal vs Luminal A	**Basal vs Luminal B**	**Basal vs HER2**	**Lumina A vs Luminal B**	**Luminal A vs HER2**	**Luminal B vs HER2**
Down	2209	1637	183	54	1700	129
Not significant	13350	14476	17891	18601	15574	18451
Up	3140	2586	625	44	1425	119

168 differentially expressed genes were used to create a heatmap via an exploratory unsupervised hierarchical cluster analysis. These genes were chosen based on their adjusted *p* value being < 0.001. The usefulness of these gene sets is supported by the heatmap (figure 4). Of the 105 cancers, 103 were divided into 4 major clusters. There is a "basal" cluster (n = 18), containing all 10 basal IHC cancers, plus 3 luminal A, 1 luminal B and 4 HER2 cancers. The remaining cancers may be further divided into 3 clusters. There is a "luminal A" cluster (n = 32), which includes 23 out of the 45 luminal A cancers, and a "luminal B" cluster (n = 32), which includes 19 of the 30 luminal B cancers. Lastly there is a "HER2" cluster (n = 21), which includes 9 of the 20 HER2 cancers.

Comparison of differentially expressed genes between IHC subtypes versus the PAM50 gene set

After removal of duplicates, 3650 genes were found to be differentially expressed, with an adjusted p value of < 0.05, among the 4 subtypes. Of the 50 genes used by the PAM50 classifier, 43 (86%) were also found to be differentially expressed among the four IHC subtypes. An exploratory cluster analysis was then performed using the PAM50 gene set (figure 5). There were 5 major clusters. The "basal" cluster (n = 18) again included all 10 cancers designated as basal type by IHC. The "luminal A" cluster (n = 23) which included 20 cancers designated as luminal A cancers by IHC, and a "luminal B" cluster (n = 22), which included 13 cancers designated as luminal B by IHC. The remaining 2 clusters (n = 8 and 34) and were composed of a mixture of luminal A, luminal B and HER2 cancers. HER2 cancers did not appear to form a distinct group, and were somewhat evenly distributed among 4 of the 5 major clusters, including the luminal B and basal groups.

Gene expression signatures of the intrinsic IHC subtypes

Gene expression signatures for each intrinsic IHC subtype were obtained. For example, in order to obtain a luminal A gene expression signature, the luminal A cancers as defined by IHC were compared to all other cancers. Gene set testing via *camera* and DAVID was then performed. This was repeated for the luminal B, basal and HER2 IHC subtypes.

Luminal A IHC subtype gene signature

Nine hundred and sixty-six genes were found to be differentially expressed by the 45 luminal A IHC cancers compared to the 60 non-luminal A cancers (adj. $p < 0.05$). The top 40

differentially expressed genes, ranked by adjusted *p* value, are shown in appendix 2. Consistent with their IHC profile, luminal A cancers showed increased gene expression for ER (*ESR1*, fold change (fc) = 7.73, adj. $p = 8.1 \times 10^{-5}$) and PR (*PGR*, fc = 7.18, adj. $p = 5.6 \times 10^{-6}$), and reduced expression for HER2 (*ERBB2*, fc = 0.271, adj. $p = 5.0 \times 10^{-6}$). Selected gene sets that were highly ranked by applying the *camera* method to the C2 gene signatures are listed in table 6. These included gene sets that were associated with estrogen receptor expression [15], longer relapse-free survival [16] and well differentiated cancers [17]. While their *p* values were < 0.05, they were no longer significant after adjusting for multiple tests (adj. *p* (FDR) > 0.05).

Functional annotation clustering performed by both DAVID suggests luminal A cancers are enriched for gene ontology (GO) terms associated with regulation of mitotic activity/cell cycle and cell migration (table 7). Similar GO terms were enriched when the *camera* method was applied to the C5 gene sets, however they were not significant after adjusting for multiple tests (adj. *p* > 0.05). 172 genes linked to the GO terms associated with mitotic activity/cell cycle regulation were used to generate an exploratory heatmap (figure 6). This supported usefulness of these genes in understanding the pathogenesis of luminal A cancers. Cluster analysis divided the 105 cancer into 2 groups. The smaller group of 51 cancers included 39 (out of 45) luminal A cancers.

Table 6. Selected gene sets enriched in luminal A IHC subtype as determined by the *camera* method performed on the C2 gene sets.

Gene set	No. genes	Direction	p value	FDR	Ref.
YANG_BREAST_CANCER_ESR1_BULK_DN	5	Down	0.014	0.69	[14]
SMID_BREAST_CANCER_RELAPSE_IN_LIVER_DN	8	Up	0.017	0.74	[15]
RHODES_UNDIFFERENTIATED_CANCER	19	Down	0.016	0.74	[16]

Table 7. Selected gene ontology terms enriched in luminal A IHC subtype, as assessed by DAVID and the Camera method performed on the C5 gene sets

GO TERMS ASSOCIATED WITH MITOTIC ACTIVITY/CELL CYCLE				
DAVID	**Count**	**PValue**	**Fold Enrichment**	**Benjamini**
GO:0030071~regulation of mitotic metaphase/anaphase transition	10	9.70E-05	5.23	0.006
GO:1902099~regulation of metaphase/anaphase transition of cell cycle	10	1.15E-04	5.12	0.007
GO:0010965~regulation of mitotic sister chromatid separation	10	1.36E-04	5.02	0.008
Camera method C5 gene set	**Count**	**Direction**	***p* value**	**FDR**
GO_REGULATION_OF_MITOTIC_CELL_CYCLE	32	Down	0.017	0.66
GO_MITOTIC_SISTER_CHROMATID_SEGREGATION	18	Down	0.033	0.66

Table 7 (Cont.)

GO TERMS ASSOCIATED WITH CELL MIGRATION				
DAVID	**Count**	***p* value**	**Fold Enrichment**	**Benjamini**
GO:0030335~positive regulation of cell migration	30	0.002	1.87	0.046
Camera method C5 gene set	**Count**	**Direction**	***p* value**	**FDR**
GO_CEREBRAL_CORTEX_CELL_MIGRATION	3	Down	0.009	0.66
GO_REGULATION_OF_ENDOTHELIAL_CELL_MIGRATION	4	Down	0.019	0.66
GO_REGULATION_OF_EPITHELIAL_CELL_MIGRATION	6	Down	0.049	0.66
GO_REGULATION_OF_CELLULAR_COMPONENT_MOVEMENT	58	Down	0.015	0.66
GO_REGULATION_OF_ACTIN_FILAMENT_BASED_MOVEMENT	3	Down	0.017	0.66
GO_ESTABLISHMENT_OF_LOCALIZATION_BY_MOVEMENT_ALONG_MICROTUBULE	3	Up	0.040	0.66

Figure 6. Heatmap of log-CPM values for 172 genes differentially expressed between luminal A and non-luminal A cancers that are associated with mitotic activity/cell cycle progression.

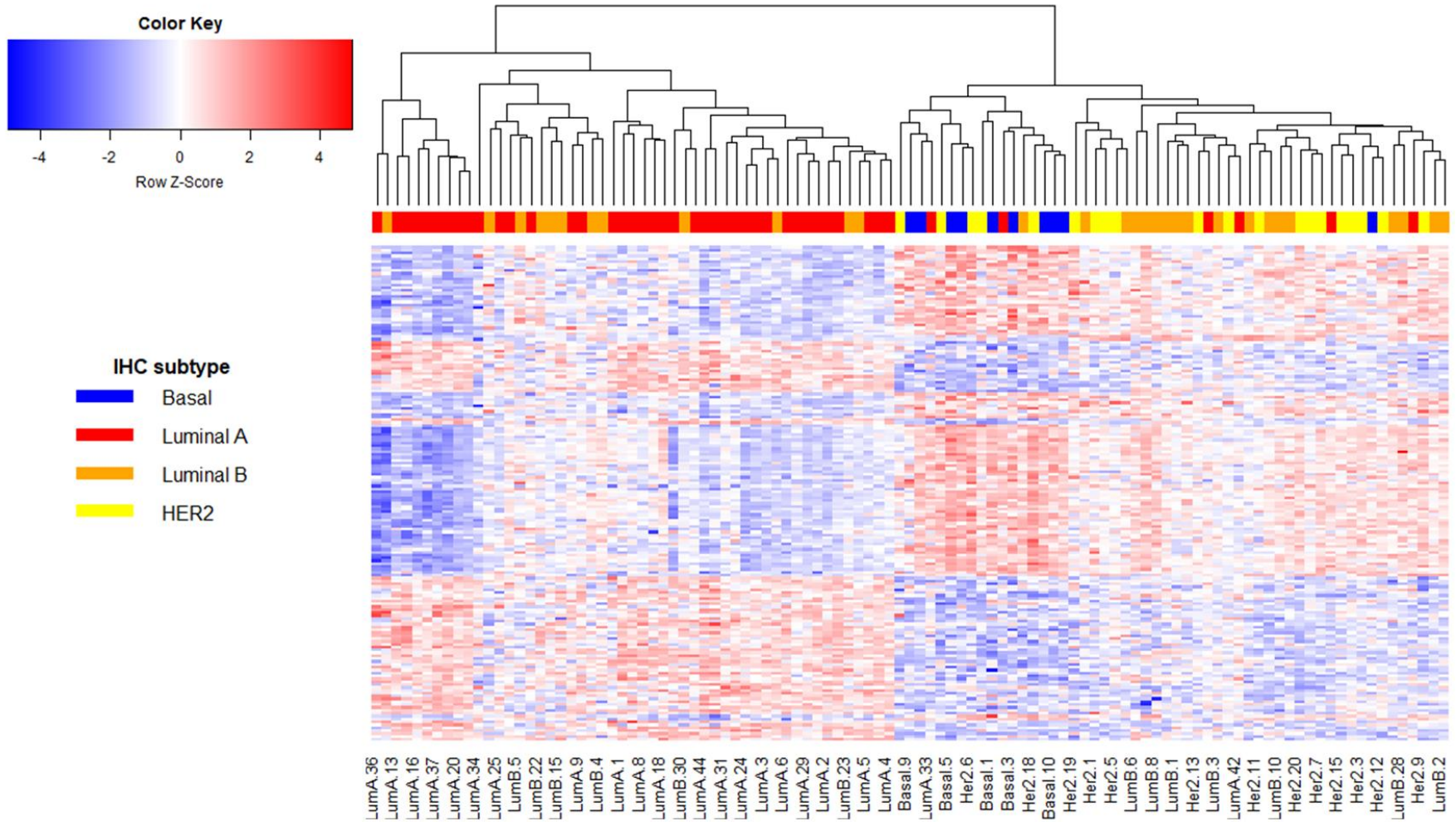
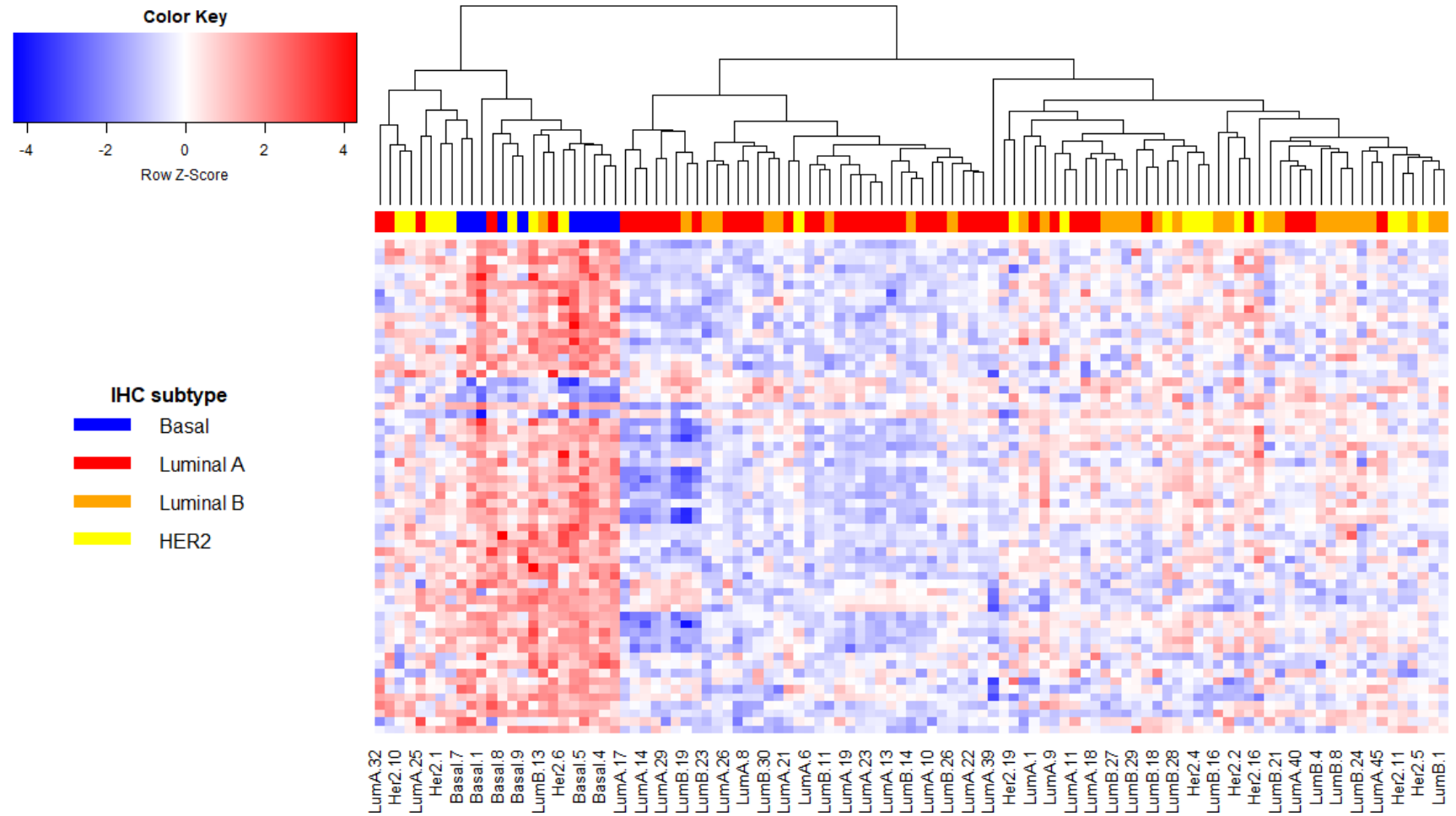


Figure 7. Heatmap of log-CPM values for 61 genes differentially expressed between basal and non-basal cancers that are associated with tumour hypoxia.



Basal IHC subtype gene signature

Basal IHC cancers have a distinctive gene signature. This is reflected by the large number of genes (n = 2095) that are differentially expressed by the 10 basal IHC cancers, compared to the 95 non-basal cancers (adj. p value < 0.05). Their distinctive gene signature is also clearly visible in the preliminary MDS plot (figure 2) and heatmap (figure 4). The top 40 differentially expressed genes in basal IHC cancers are listed in appendix 3. As previously mentioned, basal IHC cancers are negative for ER, PR and HER2 on immunohistochemistry/ISH. This is reflected by the low gene expression for estrogen receptor (*ESR1*, fc = 0.023, adj. *p* value = 0.007) and HER2 receptor (*ERBB2*, fc = 0.156, adj. $p = 0.002$) compared to the other subtypes. Interestingly progesterone receptor (*PGR*) was not significantly reduced. As expected, gene sets associated with tumour hypoxia/*HIF1-α* expression [18, 19] and beta-catenin pathway activation [20] and were highly ranked by the *camera* method performed on the C2 gene sets (table 8). Gene sets associated with poor prognosis [21], relapse in bone [16], metastases [22] and BRCA1 mutations [23, 24] were enriched, however they did not reach significance after adjusting for multiple tests (adj. $p > 0.05$). 31 genes associated with tumour hypoxia, differentially expressed between basal and non-basal cancers, were used to generate an exploratory heatmap on all 105 cancers (figure 7). On cluster analysis, there was a subset of 24 cancers, including all 10 basal cancers, which was enriched for genes associated with hypoxia.

Functional annotation clustering was performed via DAVID and the *camera* method applied to the C5 gene sets. DAVID showed basal IHC cancers were enriched for Gene

Ontology terms associated with mitotic activity/progression through the cell cycle, stem cell/neural precursor cell proliferation and mesenchymal/embryonic/neural development (table 9). An association with mitotic activity/progression through cell cycle and neural precursor cell proliferation was also seen in the C5 gene set, however this was not significant after adjusting for multiple tests (adj. $p > 0.05$).

Table 8. Selected gene sets enriched in basal IHC subtype as determined by the *camera* method performed on the C2 gene sets.

Gene Set	No. Genes	Direction	*p* value	FDR	Ref
Hypoxia					
ELVIDGE_HYPOXIA_BY_DMOG_UP	28	Up	0.028	0.99	[18]
ELVIDGE_HYPOXIA_UP	32	Up	0.017	0.99	[18]
FARDIN_HYPOXIA_9	2	Up	0.005	0.99	[19]
Prognosis					
NADERI_BREAST_CANCER_PROGNOSIS_DN	3	Down	0.078	0.99	[21]
SMID_BREAST_CANCER_RELAPSE_IN_BONE_UP	40	Up	0.11	0.99	[16]
ZUCCHI_METASTASIS_DN	10	Down	0.17	0.99	[22]
beta catenin pathway					
ST_WNT_BETA_CATENIN_PATHWAY	2	Up	0.048	0.99	[20]
BRCA1					
ONGUSAHA_BRCA1_TARGETS_UP	2	Up	0.065	0.99	[23]
VANTVEER_BREAST_CANCER_BRCA1_DN	5	Down	0.14	0.99	[24]

Table 9. Selected gene ontology terms enriched in basal IHC subtype, as assessed by DAVID and the Camera method performed on the C5 gene sets

GO TERMS ASSOCIATED WITH MITOTIC ACTIVITY/CELL CYCLE				
DAVID	**Count**	***p* value**	**Fold Enrichment**	**Benjamini**
GO:0007049~cell cycle	241	2.9E-15	1.63	1.2E-11
GO:0000278~mitotic cell cycle	161	1.7E-15	1.88	1.5E-11
GO:0000280~nuclear division	111	9.2E-15	2.16	2.6E-11
GO:1903047~mitotic cell cycle process	149	1.3E-14	1.89	2.8E-11
GO:0007067~mitotic nuclear division	89	4.7E-14	2.34	7.9E-11
Camera method C5 gene set	**Count**	**Direction**	***p* Value**	**FDR**
GO_POSITIVE_REGULATION_OF_CELL_CYCLE_PHASE_TRANSITION	13	Up	0.03	0.99
GO_NEGATIVE_REGULATION_OF_CELL_CYCLE_G2_M_PHASE_TRANSITION	8	Down	0.03	0.99
GO_MITOTIC_G2_M_TRANSITION_CHECKPOINT	7	Down	0.04	0.99
GO_MITOTIC_G2_DNA_DAMAGE_CHECKPOINT	4	Down	0.04	0.99

GO TERMS ASSOCIATED WITH STEM CELL/NEURAL PRECURSOR CELL PROLIFERATION				
DAVID	**Count**	***p* value**	**Fold Enrichment**	**Benjamini**
GO:2000648~positive regulation of stem cell proliferation	15	2.1E-06	4.47	2.5E-04
GO:0072091~regulation of stem cell proliferation	18	1.1E-05	3.40	9.6E-04
GO:0061351~neural precursor cell proliferation	28	1.9E-05	2.46	0.001
GO:0072089~stem cell proliferation	23	5.7E-05	2.58	0.004
Camera method C5 gene set	**Count**	**Direction**	***P* value**	**FDR**
GO_NEURAL_PRECURSOR_CELL_PROLIFERATION	16	Down	0.05	0.99

GO TERMS ASSOCIATED WITH MESENCHYMAL/EMBRYONAL/NEURAL DEVELOPMENT				
DAVID	**Count**	***p* value**	**Fold Enrichment**	**Benjamini**
GO:0035295~tube development	101	4.2E-11	1.97	2.5E-08
GO:0050767~regulation of neurogenesis	109	1.2E-09	1.81	4.6E-07
GO:0043009~chordate embryonic development	94	7.2E-09	1.84	2.2E-06
GO:0060485~mesenchyme development	48	6.7E-08	2.32	1.2E-05
GO:0048762~mesenchymal cell differentiation	40	8.0E-08	2.54	1.4E-05

HER2 IHC subtype gene signature

Two hundred and sixty-two genes were differentially expressed in the 20 HER2 cancers compared to the other 85 cancers, with an adjusted *p* value of < 0.05. This is much fewer than the 2096 genes observed for the 10 basal type cancers. As expected, HER2 IHC cancers showed increased expression for HER2 (*ERBB2*, fc = 3.93, adj. p = 0.028) and reduced expression for estrogen receptor (*ESR1*, fc = 0.071, adj. *p* = 0.014). No difference was seen in the expression for progesterone receptor (PGR, adj, *p* > 0.05). The top 40 differentially expressed genes are shown in appendix 4. Gene set analysis via *camera* showed enrichment for genes associated with amplification of 17q11-q21 [25]. This is concordant with the amplification of HER2 (located in 17q12) as assessed by in situ hybridization. Other enriched gene sets included those associated with a poor prognosis , RNA polymerase transcription [26] and telomere maintenance [27] (table 10). No GO terms were significantly enriched (adj. *p* > 0.05 for all terms) on DAVID or the C5 gene sets. This may be due to the small number of differentially expressed genes included for analysis.

Table 10. Selected gene sets enriched in HER2 IHC subtype as determined by the *camera* method.

Gene Set	NGenes	Direction	*p* value	FDR	Ref
NIKOLSKY_BREAST_CANCER_17Q11_Q21_AMPLICON	5	Up	0.045	0.88	[25]
REACTOME_RNA_POL_I_TRANSCRIPTION	8	Up	0.040	0.88	[26]
REACTOME_TELOMERE_MAINTENANCE	8	Up	0.040	0.88	[27]

Luminal B IHC subtype gene signature

Only 31 genes were differentially expressed by luminal B cancers compared to other cancers (appendix 5). This is not an unexpected finding, as luminal B cancers share phenotypic features with luminal A (estrogen receptor expression) and HER2 (HER2 amplification) IHC subtypes. This is also reflected by the preliminary multidimensional scaling plot, in which luminal B cancers form a loose cluster between, as well as overlying the luminal A and HER2 clusters. As mentioned, luminal B IHC cancers are defined by ER and/or PR expression, plus HER2 amplification. As expected, HER2 IHC cancers showed increased expression for HER2 (*ERBB2*, adj. $p = 0.041$, fc = 3.73). No difference in expression for ER (*ESR1*) or PR (*PGR*) was seen. Mostly one, and occasionally two, matching genes were found in common with the C2 gene sets via the *camera* method. No GO terms were significantly enriched (adj. $p > 0.05$ for all terms) on DAVID or the C5 gene sets.

PART 3. SURVIVAL ANALYSIS OF IHC SUBTYPES

Preliminary analysis of cohort with survival data

A comparison of the clinico-pathological variables between the four breast cancer subtypes is shown in table 11. There were significant differences in tumour size and grade between the four subtypes (both $p < 0.001$). Luminal A cancers were more likely to be ≤ 20mm in size (120/171, 70%), compared to luminal B (10/17, 37%), HER2 (11/24, 46%) and basal types (19/48, 40%). Luminal A cancers were also less likely to have grade of 3 out of 3 (42/171, 24%), compared to luminal B (21/27, 78%), HER2 (21/24, 87%) and basal types (39/48, 81%). No differences were found among the four subtypes for age, lymph node status, endocrine therapy or chemotherapy.

Table 11. Comparison of clinico-pathological variables between the four subtypes

Clinico-pathological variable	Luminal A n = 171	Luminal B n = 27	HER2 n = 24	Basal n = 48	*p* value
Age (yrs)					
≤ 50	67 (39%)	13 (48%)	7 (29%)	14 (29%)	0.30
> 50	104 (61%)	14 (52%)	17 (71%)	34 (71%)	
Tumour size					
≤ 20mm	120 (70%)	10 (37%)	11 (46%)	19 (40%)	< 0.001
> 20mm	51 (30%)	17 (63%)	13 (54%)	29 (60%)	
Lymph node status					
Negative	95 (56%)	15 (56%)	13 (54%)	26 (54%)	0.99
Positive	74 (44%)	12 (44%)	11 (46%)	22 (46%)	
Not available	2				
Grade					
1	44 (26%)	0	0	1 (2%)	< 0.001
2	85 (50%)	6 (22%)	3 (13%)	8 (17%)	
3	42 (24%)	21 (78%)	21 (87%)	39 (81%)	
Endocrine therapy					
No	76 (44%)	14 (52%)	14 (58%)	30 (63%)	0.12
Yes	95 (56%)	13 (48%)	10 (42%)	18 (37%)	
Chemotherapy					
No	112 (66%)	15 (56%)	11 (46%)	26 (54%)	0.17
Yes	59 (34%)	12 (44%)	13 (54%)	22 (46%)	

Figure 8. Kaplan-Meier curves of overall survival for age, size, lymph node status, grade, hormone therapy and chemotherapy (log rank test).

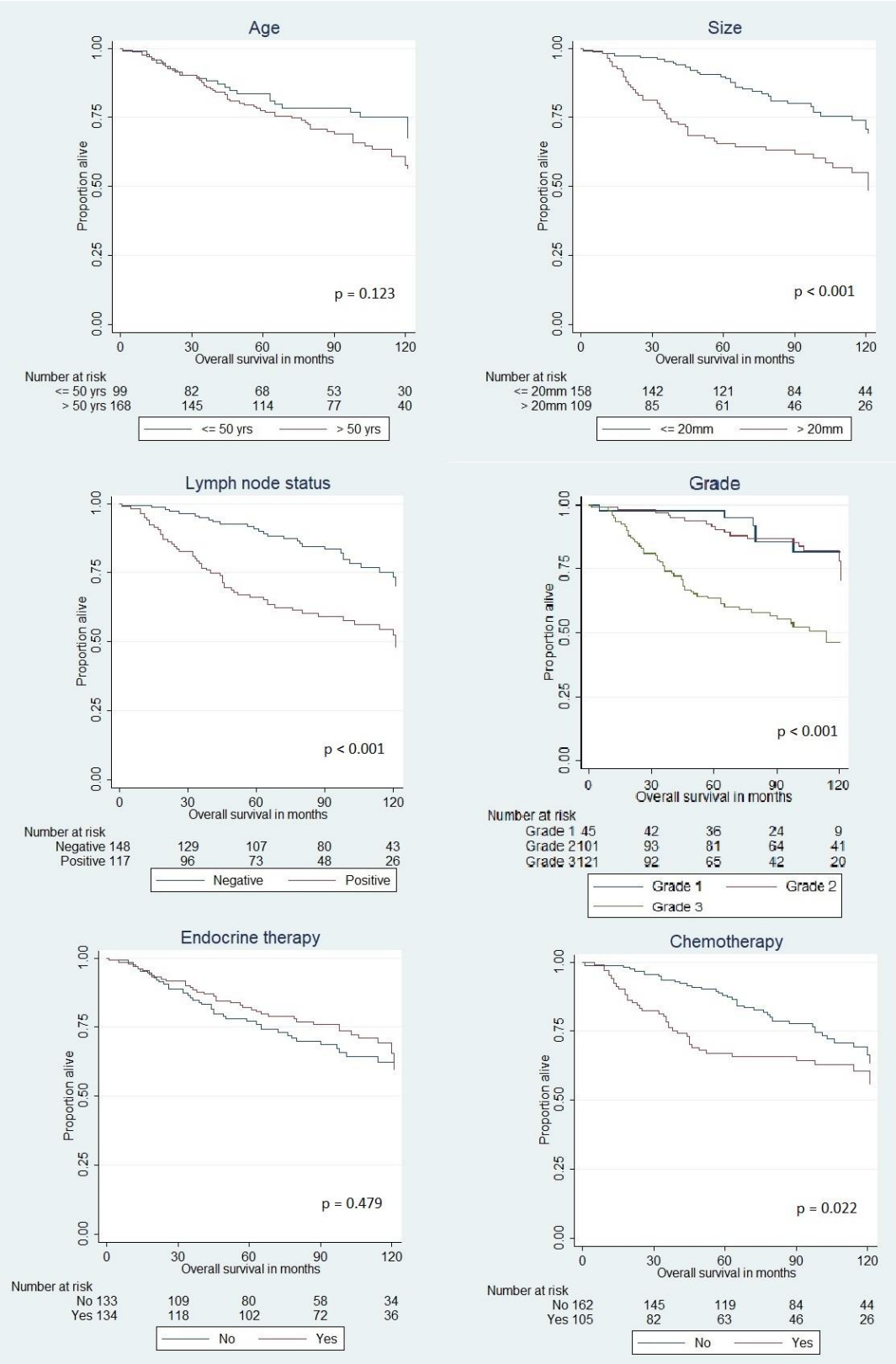


Table 12a. Univariable analysis and final multivariable Cox regression model (stratified by chemotherapy) for overall survival

Characteristic	Univariable Hazard ratio	Univariable 95% confidence interval	Univariable *p* value	Multivariable Hazard ratio	Multivariable 95% confidence interval	Multivariable *p* value
Tumour type						
Luminal A (reference)	1.00	-		1.00	-	
Luminal B	3.13	1.61 – 6.09	< 0.001	2.04	0.98 – 4.25	0.018
HER2	4.08	2.14 – 7.78		2.87	1.39 – 5.94	
Basal	3.41	2.02 – 5.77		2.11	1.17 – 3.81	
Age						
≤ 50 yrs (reference)	1.00	-	0.167	1.00	-	0.11
> 50 yrs	1.39	0.87 – 2.23		1.54	0.90 – 2.61	
Tumour size						
≤ 20mm	1.00	-	< 0.001	1.00	-	0.47
> 20mm	2.29	1.48 – 3.55		1.20	0.73 – 1.95	
Lymph node status						
Negative (reference)	1.00	-	< 0.001	1.00	-	< 0.001
Positive	2.59	1.65 – 4.08		3.50	1.97 – 6.20	
Grade of 3						
No (reference)	1.00	-	< 0.001	1.00	-	0.020
Yes	3.38	2.13 – 5.37		1.96	1.11 – 3.44	
Endocrine therapy						
No (reference)	1.00	-	0.416	1.00	-	0.008
Yes	0.84	0.54 – 1.29		0.50	0.29 – 0.83	
Chemotherapy						
No (reference)	1.00	-	0.030	NA*	NA*	NA*
Yes	1.61	1.05 – 2.49				

NA* Not available as model is stratified by chemotherapy

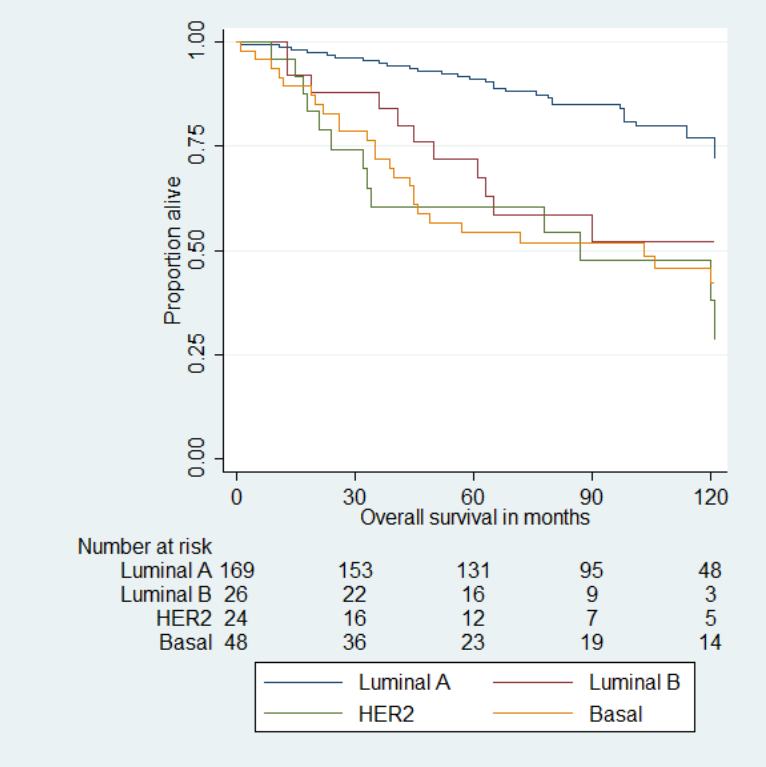
Table 12b. Preliminary multivariable Cox regression model (including chemotherapy) for overall survival

Characteristic	Hazard ratio	95% confidence interval	*p* value
Tumour type			
Luminal A (reference)	1.00	-	
Luminal B	2.15	1.03 – 4.47	0.010
HER2	3.06	1.48 – 6.35	
Basal	2.24	1.24 – 4.07	
Age			
≤ 50 yrs (reference)	1.00	-	0.068
> 50 yrs	1.64	0.96 – 2.79	
Tumour size			
≤ 20mm	1.00	-	0.37
> 20mm	1.25	0.77 – 2.01	
Lymph node status			
Negative (reference)	1.00	-	< 0.001
Positive	3.65	2.08 – 6.38	
Grade of 3			
No (reference)	1.00	-	0.030
Yes	1.86	1.06 – 3.26	
Endocrine therapy			
No (reference)	1.00	-	0.005
Yes	0.47	0.28 – 0.79	
Chemotherapy			
No (reference)	1.00	-	0.82
Yes	1.06	0.64 – 1.78	

Univariable analysis for overall survival

Kaplan-Meier curves for overall survival were drawn for the standard clinico-pathological variables. Significant differences in overall survival were observed for tumour size, lymph node status, grade and chemotherapy (log rank test, $p < 0.05$, figure 9). In view of their similarity in survival, grades 1 and 2 were combined into a single category for subsequent analyses. The univariable analyses of overall survival, for the clinic-pathological variables are shown in table 12. A shorter overall survival was associated with size > 20mm (HR 2.29, 95% CI: 1.48 – 3.55, $p < 0.001$), positive lymph node status (HR 2.59, 95% CI: 1.65 – 4.08, $p < 0.001$), a grade of 3 (HR 3.38, 95% CI: 2.13 – 5.37, $p < 0.001$) and chemotherapy (HR 1.61, 95% CI: 1.05 – 2.49, $p = 0.030$).

Figure 9. Kaplan-Meier curves for overall survival by breast cancer subtype (log rank test $p < 0.001$)



The univariable analysis of overall survival for the breast cancer subtypes is listed in table 12. Compared to luminal A cancers, luminal B (HR = 3.13, 95% CI: 1.61-6.09, $p = 0.001$), HER2 (HR = 4.08, 95% CI: 2.14-7.78, $p < 0.001$) and basal (HR = 3.41, 95% CI: 2.02-5.77, $p < 0.001$) cancers are associated with shorter overall survival. On the Kaplan-Meier curves, while luminal A cancers have a more favourable prognosis, no obvious differences in overall survival are seen among the luminal B, HER2 and basal cancers (figure 6).

Multivariable analysis for overall survival

The final multivariable model was stratified by chemotherapy to prevent violation of the proportional hazards assumption (table 12a). Diagnostic testing for the assumptions

underlying model, and the rationale for the final model are described in appendix 6. Compared to luminal A cancers, HER2 (HR 2.87, 95% CI: 1.39 – 5.94, $p = 0.004$) and basal (HR 2.11, 95% CI: 1.17 – 3.81, $p = 0.014$) cancers were associated with a shorter overall survival. There was a tendency for luminal B cancers to be associated with a shorter overall survival, however this p value was > 0.05 (HR 2.04, 95% CI: 0.98 – 4.25, $p = 0.056$). Lymph node positivity (HR 3.5, 95% CI: 1.97 – 6.20, $p < 0.001$) and a grade of 3 (HR 1.96, 95% CI 1.11 – 3.44, $p = 0.020$) were also associated with a poorer outcome. Treatment with endocrine therapy was associated with a favourable outcome (HR 0.50, 95% CI: 0.29 – 0.83, $p = 0.008$). A hazard ratio for chemotherapy was not available in the final model due to stratification. However in a preliminary model including chemotherapy (table 12b), when adjusted for other factors, chemotherapy no longer correlated with a shorter overall survival (HR 1.06, 95% CI: 0.64 – 1.78, $p = 0.817$).

Discussion

The importance of subtyping by immunohistochemistry is supported by the St Galen guidelines, which states that the adjuvant treatment of breast cancers should be based on pathological determination of ER, PR and HER2 immunohistochemistry (IHC) and HER2 in situ hybridization (ISH) [28]. Any ER expression on IHC (i.e. luminal A or B by IHC) warrants endocrine therapy with tamoxifen (or an aromatase inhibitor if the patient is postmenopausal). Over-expression of HER2 (i.e. luminal B or HER2 by IHC) justifies the use of anti-HER2 treatment (such as trastuzumab). Chemotherapy should be considered for all basal and HER2 cancers except for very low risk cancers. Guidelines for the use of chemotherapy in luminal A cancers (determined by IHC) is difficult to define, however relative indications include low ER/PR expression, grade of 3, high Ki67 proliferation index (>30%), lymphovascular invasion and a tumour size of > 5cm.

Subtyping by gene expression assays (e.g. PAM50 (Prosigna), Oncotype DX, MammaPrint) may be useful in luminal A IHC cancers, where the use of chemotherapy is uncertain after consideration of pathological data mentioned above. There is evidence to suggest gene expression analysis by PAM50 may yield additional prognostic data over conventional IHC subtypes in these instances [29]. The findings derived from the data of Brueffer *et al*. [1], may offer an explanation for this – 33% of cancers defined as luminal A by IHC were re-classified as luminal B (with a worse prognosis) by PAM50. Furthermore, 11% of cancers defined as luminal B by IHC were re-classified as luminal A (with a better prognosis) by PAM50. In this study, HER2 expression was used to distinguish between luminal A (HER2 negative) and luminal B (HER2 positive) IHC subtypes. Our analysis of

RNA-seq data has revealed additional biomarkers which may potentially be used to improve the classification of luminal A and B cancers by IHC. Cluster analysis with the PAM50 gene set suggests, in addition to HER2, luminal B cancers show increased expression of ANLN (Anillin), CCNB1 (Cyclin B1), MKI67 (Ki67) and MYBL2 (MYB Proto-Oncogene Like 2). Of these biomarkers, Ki67 is already in routine clinical use, with a Ki67 index of > 14% being associated with increased risk of relapse [30]. Cyclin B1 is not in routine clinical use, but its expression has been linked to poor prognosis in ER positive cancers [31]. Addition of these biomarkers may improve the identification of ER+ HER2- cancers that are at increased risk of progression.

Our overall rate of discordance between IHC and PAM50 was 41%, which was similar to 38% obtained by Kim *et al*. [32]. In practice, treatment decisions are driven by assessment of protein receptor expression by IHC. Nevertheless the discordance between subtyping by IHC and PAM50 raises clinically relevant issues regarding the potential under and over-treatment of a small proportion of breast cancers. For example:

- 6% of luminal A cancers defined by IHC were reclassified as HER2 by PAM50. Would these cancers benefit from the addition of anti-HER2 therapy?
- 57% of luminal B cancers defined by IHC were reclassified as HER2 by PAM50. Are these cases being overtreated with endocrine therapy?
- 7% of basal type cancers defined by IHC were reclassified as cancers where targeted therapy is available (Luminal A and B, HER2) by PAM50. It is uncertain whether these cases may respond to targeted therapy despite the absence of corresponding protein receptor expression on IHC.

In this study, cluster analysis was used as a supervised learning technique on a training dataset only (with no testing dataset). This was an exploratory analysis which sought to understand the usefulness of particular genes in providing information on IHC cancer subtypes. It should be noted that regular supervised learning techniques, beyond the scope of this study, would need to be employed to 1) understand the structure underlying the data, 2) predict IHC cancer subtypes in a testing dataset and 3) quantitively describe how well the gene information re-creates the IHC subtypes. Nevertheless, our findings suggest differences in gene expression profiles underlie the different IHC subtypes. Basal IHC is the most distinctive subtype, with the largest number of up and down-regulated genes (table 5 and fig 4). Of the 6 pair-wise comparisons between the IHC subtypes, the comparison between basal vs. luminal A yielded the largest number of differentially expressed genes (3140 up and 2209 down, table 5). The differences in gene expression between basal and luminal A cancers are reflected by their different clinico-pathological characteristics. Basal cancers are characterized by triple negativity for ER, PR and HER, higher grade, larger size and a poorer prognosis, whereas luminal A cancers are associated with ER expression, a lower grade, smaller size and better survival (tables 11, 12a and figure 8).

Gene set analysis provides an insight into the molecular mechanisms underlying the aggressive behavior of basal cancers. Our data shows basal cancers are enriched for genes linked to:

1) Tumour hypoxia: The link between basal cancers and hypoxia is supported by their microscopic morphology. Basal cancers often have central areas of necrosis as a consequence of tumour hypoxia [33]. Furthermore, activation of the hypoxia-inducible factor (HIF) pathway has been linked to the aggressive behaviour of basal cancers and chemotherapy resistance [34].

2) Beta-catenin pathway activation: Activation of b-catenin pathway in basal cancers has been linked to tumour proliferation, survival, matrix remodeling and a worse prognosis [35].

3) Stem cell phenotype: Basal cancer cells with a stem cell phenotype are responsible for driving intratumoral cellular heterogeneity, continued proliferation, resistance to therapy and metastasis [36].

4) Mesenchymal development: Basal breast cancers cells may lose their epithelial characteristics and polarity, resulting in a mesenchymal phenotype with increased migratory behavior [37].

The two most similar subtypes with the least number of differentially expressed genes were luminal A vs luminal B (44 up and 54 down). This is not an unexpected finding since both tumour types are driven by ER expression. Similarly, comparison between the two HER2 expressing subtypes (luminal B vs. HER2) yielded a relatively small number of differentially expressed genes (119 down and 129 up). Only 31 genes were differentially expressed between 30 luminal B and 75 non-luminal B cancers. No gene sets or GO terms were significantly enriched on DAVID or *camera* testing on C2 and C5 gene sets. This

may be due to luminal B cancers sharing phenotypic features with luminal A (ER expression) and HER2 (HER2 amplification) cancers.

Regarding the prognosis of the different breast cancer IHC subtypes, our findings are broadly in line with previous studies [3, 4, 38]. Luminal B, basal and HER2 cancers are associated with higher grade and larger tumour size (both $p < 0.001$). These three subtypes also have a worse prognosis compared to luminal A cancers in a univariable analysis. After adjusting for standard prognostic parameters such as grade and lymph node status, basal (HR 2.11, 95% CI: 1.17 – 3.81, $p = 0.014$) and HER2 (HR 2.87, 95% CI: 1.39 – 5.94, $p = 0.004$) IHC subtypes were independent predictors of shorter breast cancer specific overall survival. The luminal B IHC subtype was associated with a shorter overall survival; however this was not significant in a multivariable analysis (HR 2.04, 95% CI: 0.98 – 4.25, $p = 0.056$). Our findings justify the rationale for the use of adjuvant chemotherapy in these aggressive subtypes. In a univariable analysis, chemotherapy was associated with shorter overall survival ($p = 0.022$). This effect disappeared after adjusting for other clinico-pathological factors and IHC subtype (HR 1.06, 95% CI: 0.64 – 1.78, $p = 0.817$). The shorter survival in patients given chemotherapy is likely to be due to this treatment being administered to patients with poor prognostic variables, rather than an adverse effect from the treatment itself. Endocrine therapy was effective in improving overall survival (HR 0.50, 95% CI: 0.29 – 0.83, $p = 0.008$).

The main limitation of this study is the availability of only three IHC biomarkers (ER, PR, HER2) in the NCI cohort. Previous studies have suggested a five biomarker panel with the

addition of CK5/6 and EGFR IHC may be superior in defining basal phenotype. The addition of CK5/6 and EFGR may result in an improved specificity of 100%, and a sensitivity of 76%, when compared to the "gold standard" of gene expression analysis. This is supported by our RNA-Seq data which shows these genes are up-regulated in basal cancers (KRT5 and EGFR, see figure 5). The use of a three biomarker panel may potentially misclassify luminal and normal-like breast cancers as basal-like [30]. There is also data to suggest a six biomarker panel, with the addition of Ki67 IHC, may be helpful in identifying high risk luminal A cancers which should be reclassified as HER2 negative luminal B cancers [30].

In summary, breast cancer IHC subtypes have distinctive gene expression signatures. Subtyping by IHC also provides important data that may help guide prognostication and treatment decisions.

References

1. Brueffer C, Vallon-Christersson J, Grabau† D, Ehinger A, Häkkinen J, Hegardt C, Malina J, Chen Y, Bendahl P-O, Manjer J *et al*: **Clinical Value of RNA Sequencing–Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network—Breast Initiative**. *JCO Precision Oncology* 2018(2):1-18.
2. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM: **Toward a Shared Vision for Cancer Genomic Data**. *N Engl J Med* 2016, **375**(12):1109-1112.
3. Cheang MC, Voduc D, Bajdik C, Leung S, McKinney S, Chia SK, Perou CM, Nielsen TO: **Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype**. *Clin Cancer Res* 2008, **14**(5):1368-1376.
4. Millar EK, Graham PH, O'Toole SA, McNeil CM, Browne L, Morey AL, Eggleton S, Beretov J, Theocharous C, Capp A *et al*: **Prediction of local recurrence, distant metastases, and death after breast-conserving therapy in early-stage invasive breast cancer using a five-biomarker panel**. *J Clin Oncol* 2009, **27**(28):4701-4708.
5. **TCGA Molecular Characterization Platforms, accessed 15th July 2019** [https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/technology]
6. Law CW, Alhamdoosh M, Su S, Dong X, Tian L, Smyth GK, Ritchie ME: **RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR**. *F1000Res* 2016, **5**.
7. M C: **org.Hs.eg.db: Genome wide annotation for Human. R package version 3.8.2.** In.; 2019.
8. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome Biol* 2010, **11**(3):R25.
9. Law CW, Chen Y, Shi W, Smyth GK: **voom: Precision weights unlock linear model analysis tools for RNA-seq read counts**. *Genome Biol* 2014, **15**(2):R29.
10. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* 2004, **3**:Article3.
11. Wu D, Smyth GK: **Camera: a competitive gene set test accounting for inter-gene correlation**. *Nucleic Acids Res* 2012, **40**(17):e133.
12. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.
13. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**(1):44-57.
14. Huang da W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Res* 2009, **37**(1):1-13.

15. Yang F, Foekens JA, Yu J, Sieuwerts AM, Timmermans M, Klijn JG, Atkins D, Wang Y, Jiang Y: **Laser microdissection and microarray analysis of breast tumors reveal ER-alpha related genes and pathways**. *Oncogene* 2006, **25**(9):1413-1419.
16. Smid M, Wang Y, Zhang Y, Sieuwerts AM, Yu J, Klijn JG, Foekens JA, Martens JW: **Subtypes of breast cancer show preferential site of relapse**. *Cancer Res* 2008, **68**(9):3108-3114.
17. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression**. *Proc Natl Acad Sci U S A* 2004, **101**(25):9309-9314.
18. Elvidge GP, Glenny L, Appelhoff RJ, Ratcliffe PJ, Ragoussis J, Gleadle JM: **Concordant regulation of gene expression by hypoxia and 2-oxoglutarate-dependent dioxygenase inhibition: the role of HIF-1alpha, HIF-2alpha, and other pathways**. *J Biol Chem* 2006, **281**(22):15215-15226.
19. Fardin P, Barla A, Mosci S, Rosasco L, Verri A, Varesio L: **The l1-l2 regularization framework unmasks the hypoxia signature hidden in the transcriptome of a set of heterogeneous neuroblastoma cell lines**. *BMC Genomics* 2009, **10**:474.
20. Moon RT: **Wnt/beta-catenin pathway**. *Sci STKE* 2005, **2005**(271):cm1.
21. Naderi A, Teschendorff AE, Barbosa-Morais NL, Pinder SE, Green AR, Powe DG, Robertson JF, Aparicio S, Ellis IO, Brenton JD *et al*: **A gene-expression signature to predict survival in breast cancer across independent data sets**. *Oncogene* 2007, **26**(10):1507-1516.
22. Zucchi I, Mento E, Kuznetsov VA, Scotti M, Valsecchi V, Simionati B, Vicinanza E, Valle G, Pilotti S, Reinbold R *et al*: **Gene expression profiles of epithelial cells microscopically isolated from a breast-invasive ductal carcinoma and a nodal metastasis**. *Proc Natl Acad Sci U S A* 2004, **101**(52):18147-18152.
23. Ongusaha PP, Ouchi T, Kim KT, Nytko E, Kwak JC, Duda RB, Deng CX, Lee SW: **BRCA1 shifts p53-mediated cellular outcomes towards irreversible growth arrest**. *Oncogene* 2003, **22**(24):3749-3758.
24. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al*: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530-536.
25. Nikolsky Y, Sviridov E, Yao J, Dosymbekov D, Ustyansky V, Kaznacheev V, Dezso Z, Mulvey L, Macconaill LE, Winckler W *et al*: **Genome-wide functional synergy between amplified and mutated genes in human breast cancer**. *Cancer Res* 2008, **68**(22):9532-9540.
26. Comai L: **Mechanism of RNA polymerase I transcription**. *Adv Protein Chem* 2004, **67**:123-155.
27. **Reactome. Telomere Maintenance.**
28. Goldhirsch A, Ingle JN, Gelber RD, Coates AS, Thurlimann B, Senn HJ, Panel m: **Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009**. *Ann Oncol* 2009, **20**(8):1319-1329.

29. Lundberg A, Lindstrom LS, Harrell JC, Falato C, Carlson JW, Wright PK, Foukakis T, Perou CM, Czene K, Bergh J *et al*: **Gene Expression Signatures and Immunohistochemical Subtypes Add Prognostic Value to Each Other in Breast Cancer Cohorts**. *Clin Cancer Res* 2017, **23**(24):7512-7520.
30. Viale G, Hanlon Newell AE, Walker E, Harlow G, Bai I, Russo L, Dell'Orto P, Maisonneuve P: **Ki-67 (30-9) scoring and differentiation of Luminal A- and Luminal B-like breast cancer subtypes**. *Breast Cancer Res Treat* 2019.
31. Agarwal R, Gonzalez-Angulo AM, Myhre S, Carey M, Lee JS, Overgaard J, Alsner J, Stemke-Hale K, Lluch A, Neve RM *et al*: **Integrative analysis of cyclin protein levels identifies cyclin b1 as a classifier and predictor of outcomes in breast cancer**. *Clin Cancer Res* 2009, **15**(11):3654-3662.
32. Kim HK, Park KH, Kim Y, Park SE, Lee HS, Lim SW, Cho JH, Kim JY, Lee JE, Ahn JS *et al*: **Discordance of the PAM50 Intrinsic Subtypes Compared with Immunohistochemistry-Based Surrogate in Breast Cancer Patients: Potential Implication of Genomic Alterations of Discordance**. *Cancer Res Treat* 2019, **51**(2):737-747.
33. Fulford LG, Easton DF, Reis-Filho JS, Sofronis A, Gillett CE, Lakhani SR, Hanby A: **Specific morphological features predictive for the basal phenotype in grade 3 invasive ductal carcinoma of breast**. *Histopathology* 2006, **49**(1):22-34.
34. Tan EY, Yan M, Campo L, Han C, Takano E, Turley H, Candiloro I, Pezzella F, Gatter KC, Millar EK *et al*: **The key hypoxia regulated gene CAIX is upregulated in basal-like breast tumours and is associated with resistance to chemotherapy**. *Br J Cancer* 2009, **100**(2):405-411.
35. Khramtsov AI, Khramtsova GF, Tretiakova M, Huo D, Olopade OI, Goss KH: **Wnt/beta-catenin pathway activation is enriched in basal-like breast cancers and predicts poor outcome**. *Am J Pathol* 2010, **176**(6):2911-2920.
36. Thiagarajan PS, Sinyuk M, Turaga SM, Mulkearns-Hubert EE, Hale JS, Rao V, Demelash A, Saygin C, China A, Alban TJ *et al*: **Cx26 drives self-renewal in triple-negative breast cancer via interaction with NANOG and focal adhesion kinase**. *Nat Commun* 2018, **9**(1):578.
37. Sarrio D, Rodriguez-Pinilla SM, Hardisson D, Cano A, Moreno-Bueno G, Palacios J: **Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype**. *Cancer Res* 2008, **68**(4):989-997.
38. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, Hernandez-Boussard T, Livasy C, Cowan D, Dressler L *et al*: **Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma**. *Clin Cancer Res* 2004, **10**(16):5367-5374.

Appendix 1. List of PAM50 genes

ACTR3B	KNTC2
ANLN	KRT14
BAG1	KRT17
BCL2	KRT5
BIRC5	MAPT
BLVRA	MDM2
CCNB1	MELK
CCNE1	MIA
CDC20	MKI67
CDC6	MLPH
CDCA1	MMP11
CDH3	MYBL2
CENPF	MYC
CEP55	NAT1
CXXC5	ORC6L
EGFR	PGR
ERBB2	PHGDH
ESR1	PTTG1
EXO1	RRM2
FGFR4	SFRP1
FOXA1	SLC39A6
FOXC1	TMEM45B
GPR160	TYMS
GRB7	UBE2C
KIF2C	UBE2T

Appendix 2. Top 40 differentially expressed genes for luminal A IHC subtype

SYMBOL	logFC	AveExpr	t	*p* value	adj. *p* val	B
N4BP2L1	1.04	3.26	6.69	1.1E-09	1.2E-05	11.8
LINC00173	2.48	-1.09	6.65	1.3E-09	1.2E-05	11.0
JADE2	1.12	5.29	6.37	4.7E-09	1.5E-05	10.4
NAGS	1.60	1.18	6.39	4.3E-09	1.5E-05	10.4
TTC34	1.97	-1.21	6.45	3.3E-09	1.5E-05	10.1
AZU1	2.49	-2.63	6.42	3.7E-09	1.5E-05	9.27
C5AR2	2.12	2.74	6.27	7.7E-09	1.8E-05	9.96
TPRG1	2.58	3.06	6.19	1.1E-08	2.0E-05	9.60
ACBD4	1.03	2.86	6.18	1.2E-08	2.0E-05	9.55
NA	3.47	-3.78	6.21	1.0E-08	2.0E-05	7.69
MAPT-IT1	2.91	-3.45	6.14	1.4E-08	2.1E-05	7.66
MAPT	2.33	4.98	6.03	2.3E-08	2.7E-05	8.89
NA	1.41	0.94	6.04	2.3E-08	2.7E-05	8.84
DNA2	-1.01	3.08	-6.02	2.5E-08	2.7E-05	8.86
NA	2.27	-2.99	5.96	3.3E-08	3.2E-05	7.31
HMGB3	-1.21	6.03	-5.90	4.4E-08	3.9E-05	8.26
WNK4	2.96	2.29	5.82	6.1E-08	4.1E-05	8.01
PRX	1.12	2.41	5.81	6.5E-08	4.1E-05	7.94
MIR497HG	1.25	-0.33	5.86	5.2E-08	4.1E-05	7.91
MSL3P1	-2.03	-1.44	-5.86	5.2E-08	4.1E-05	7.43
MCM10	-1.54	3.35	-5.79	7.2E-08	4.2E-05	7.85
CEP55	-1.40	4.15	-5.79	7.2E-08	4.2E-05	7.84
ZNF516	1.29	3.93	5.75	8.7E-08	4.7E-05	7.66
EXO1	-1.56	3.35	-5.74	8.9E-08	4.7E-05	7.66
UGCG	1.23	6.31	5.75	8.7E-08	4.7E-05	7.59
SUSD3	2.34	3.61	5.73	9.5E-08	4.8E-05	7.58
PYY	2.67	-2.89	5.73	9.3E-08	4.8E-05	6.53
MELK	-1.48	4.04	-5.70	1.1E-07	4.9E-05	7.48
CCNE1	-1.70	2.35	-5.70	1.1E-07	4.9E-05	7.47
CBX7	1.04	4.42	5.70	1.1E-07	4.9E-05	7.43
LINC01843	2.38	-2.38	5.70	1.1E-07	4.9E-05	6.65
PER1	1.15	4.88	5.68	1.2E-07	5.1E-05	7.34
LAD1	-1.67	5.95	-5.67	1.2E-07	5.3E-05	7.26
HMGA1	-1.06	7.03	-5.65	1.4E-07	5.5E-05	7.16
GLIPR1L2	1.49	-0.26	5.64	1.4E-07	5.5E-05	7.05
TMEM229B	1.21	3.81	5.61	1.6E-07	6.0E-05	7.08
SOX11	-2.96	2.54	-5.61	1.6E-07	6.0E-05	7.07
ECE2	-1.13	3.53	-5.58	1.9E-07	6.3E-05	6.95
MYBL2	-1.71	5.45	-5.58	1.8E-07	6.3E-05	6.92
NA	2.24	-2.96	5.59	1.8E-07	6.3E-05	5.99

Appendix 3. Top 40 differentially expressed genes for basal IHC subtype

SYMBOL	logFC	AveExpr	t	*p* value	adj. *p* val	B
SRSF12	3.99	-0.95	11.79	4.4E-21	8.3E-17	37.0
UGT8	5.24	0.33	11.55	1.6E-20	1.1E-16	36.1
LINC02487	4.31	-2.68	11.52	1.8E-20	1.1E-16	35.3
ROPN1	6.24	-2.51	11.02	2.4E-19	9.0E-16	33.2
NA	5.41	-4.20	11.05	2.1E-19	9.0E-16	32.6
ART3	5.63	-1.61	10.08	3.2E-17	9.9E-14	28.6
EN1	4.65	1.52	9.84	1.2E-16	3.1E-13	27.4
SLC6A15	6.12	-3.03	9.66	2.9E-16	6.7E-13	26.3
VGLL1	5.81	-0.43	9.52	6.1E-16	1.1E-12	25.8
MAPK4	5.20	-1.71	9.53	5.7E-16	1.1E-12	25.7
AFAP1-AS1	4.51	-1.27	9.41	1.1E-15	1.9E-12	25.1
INA	5.01	-3.03	9.17	3.7E-15	5.8E-12	23.7
LOC1019294	4.43	-4.45	9.09	5.8E-15	8.3E-12	22.8
LEMD1	4.18	-2.95	9.06	6.8E-15	9.1E-12	23.1
NA	5.46	-2.94	9.04	7.3E-15	9.1E-12	23.2
PCDH8	4.63	-3.75	9.02	8.4E-15	9.8E-12	22.8
OPRK1	4.59	-3.70	8.91	1.5E-14	1.6E-11	22.2
C1QL2	5.04	-3.54	8.86	1.9E-14	2.0E-11	22.1
LINC02188	5.30	-0.88	8.67	5.1E-14	5.0E-11	21.5
C6orf15	4.51	-2.71	8.59	7.4E-14	6.9E-11	20.9
HPDL	3.49	0.06	8.58	8.0E-14	7.2E-11	21.0
RSPO4	3.99	-1.77	8.54	9.8E-14	8.3E-11	20.7
POLR2F	3.59	-2.21	8.48	1.3E-13	1.1E-10	20.4
NA	4.93	-2.69	8.45	1.5E-13	1.2E-10	20.3
DLL3	4.09	-2.68	8.39	2.1E-13	1.6E-10	19.9
CELF4	3.20	-1.14	8.35	2.6E-13	1.9E-10	19.8
KIF1A	6.24	0.66	8.33	3.0E-13	2.0E-10	19.7
FZD9	3.58	-1.37	8.26	4.1E-13	2.8E-10	19.4
CASC8	3.75	-1.87	8.24	4.5E-13	2.9E-10	19.2
LYAR	1.33	3.76	8.22	5.0E-13	3.0E-10	19.2
NKX1-2	5.02	-3.03	8.23	4.8E-13	3.0E-10	19.1
TAFA3	3.60	-1.94	8.21	5.2E-13	3.0E-10	19.1
CXCL5	4.09	-2.49	8.22	5.1E-13	3.0E-10	19.1
UPF3B	1.36	3.75	8.19	5.8E-13	3.2E-10	19.1
CCNE1	2.80	2.35	8.19	6.0E-13	3.2E-10	19.0
CDKN2A	3.12	3.22	8.17	6.7E-13	3.5E-10	18.9
CA9	5.42	-0.94	8.08	1.1E-12	5.3E-10	18.5
CRHR1	3.72	-1.68	8.03	1.4E-12	6.7E-10	18.2
COMMD2	1.20	5.09	8.01	1.5E-12	7.0E-10	18.1
CHODL	3.74	-1.24	7.94	2.2E-12	1.0E-09	17.8

Appendix 4. Top 40 differentially expressed genes for HER2 IHC subtype

SYMBOL	logFC	AveExpr	t	*p* value	adj. *p* val	B
NCALD	1.78	4.37	6.75	7.9E-10	1.5E-05	12.1
PRAC2	3.41	-3.81	6.28	7.5E-09	7.0E-05	8.17
GSDMC	2.75	0.49	5.95	3.4E-08	0.00021	8.40
IGF1R	-2.03	6.88	-5.69	1.1E-07	0.00043	7.41
KRT4	3.86	-2.11	5.67	1.2E-07	0.00043	6.84
NA	2.95	-3.13	5.64	1.4E-07	0.00043	6.21
PADI2	2.52	5.85	5.55	2.1E-07	0.00055	6.84
NCCRP1	3.00	1.74	5.51	2.5E-07	0.00056	6.66
SCCPDH	-1.53	5.91	-5.46	3.1E-07	0.00056	6.47
SLC6A11	3.04	0.21	5.45	3.3E-07	0.00056	6.32
SEL1L3	1.42	4.89	5.40	4.1E-07	0.00056	6.19
LOC10192731	3.27	-2.62	5.39	4.2E-07	0.00056	5.58
ZP2	3.39	-4.12	5.42	3.6E-07	0.00056	5.06
CRISP2	3.10	-4.12	5.41	3.8E-07	0.00056	4.91
NANOS1	1.83	2.53	5.34	5.3E-07	0.00065	5.99
TMEM65	1.04	5.24	5.31	6.0E-07	0.00070	5.83
SLC12A1	3.12	-1.02	5.28	6.9E-07	0.00076	5.51
SYT16	2.62	-2.59	5.22	8.8E-07	0.00091	4.88
KDM4B	-1.15	6.49	-5.19	1.0E-06	0.00098	5.32
AQP5	3.95	0.35	5.18	1.1E-06	0.00098	5.31
SOX11	2.68	2.54	5.13	1.3E-06	0.0010	5.16
SCEL	3.09	-2.87	5.14	1.3E-06	0.0010	4.55
LINC02159	2.24	-3.64	5.14	1.3E-06	0.0010	4.05
RNF145	1.16	5.40	5.06	1.8E-06	0.0013	4.81
SMCO4	1.10	3.50	4.98	2.4E-06	0.0016	4.55
LOC10536934	1.51	-0.04	4.98	2.4E-06	0.0016	4.45
A2ML1	3.05	0.49	4.96	2.7E-06	0.0017	4.43
NA	1.94	-1.94	4.93	3.0E-06	0.0019	3.96
LCT	2.46	-2.16	4.87	3.9E-06	0.0023	3.76
B3GNT7	1.68	2.40	4.83	4.6E-06	0.0026	3.98
CHRM1	2.35	0.97	4.81	4.9E-06	0.0026	3.90
NA	1.04	-0.36	4.74	6.7E-06	0.0034	3.49
HSD17B2	2.45	-0.13	4.73	6.9E-06	0.0034	3.54
EVL	-1.40	7.28	-4.71	7.3E-06	0.0035	3.45
HIST1H2BG	2.09	1.29	4.70	7.8E-06	0.0036	3.48
TUBA4A	1.35	5.01	4.70	7.9E-06	0.0036	3.40
ATP10B	2.20	-2.49	4.66	9.3E-06	0.0041	2.91
C1orf116	2.29	4.00	4.64	9.8E-06	0.0042	3.21
RERG	-2.18	4.80	-4.62	1.1E-05	0.0044	3.19
CDKN3	1.13	3.05	4.57	1.3E-05	0.0052	3.00

Appendix 5. 31 differentially expressed genes for luminal B IHC subtype

SYMBOL	logFC	AveExpr	t	*p* value	adj. *p* val	B
PLAAT2	1.91	0.90	5.15	1.2E-06	0.015	5.06
UNC5A	2.14	1.48	5.02	2.1E-06	0.015	4.63
AKAP5	1.42	2.60	4.92	3.1E-06	0.015	4.31
TINAGL1	-1.31	4.84	-4.89	3.6E-06	0.015	4.20
IYD	2.66	-0.17	4.87	3.9E-06	0.015	3.87
GSDMB	1.60	3.75	4.62	1.1E-05	0.024	3.18
ORMDL3	1.50	7.37	4.60	1.1E-05	0.024	3.10
CT62	2.74	0.22	4.60	1.2E-05	0.024	3.00
LRFN2	1.85	-1.02	4.60	1.2E-05	0.024	2.77
RDM1P5	1.28	-0.90	4.55	1.4E-05	0.027	2.60
ABCC12	3.00	-1.44	4.51	1.7E-05	0.028	2.43
FFAR2	1.81	0.85	4.48	1.9E-05	0.029	2.61
PPFIBP2	0.69	5.49	4.42	2.4E-05	0.030	2.43
RIMS1	2.15	-0.34	4.41	2.4E-05	0.030	2.28
NA	1.90	-1.01	4.43	2.3E-05	0.030	2.22
GDPD1	0.98	2.23	4.39	2.7E-05	0.032	2.34
TEKT5	1.31	-1.07	4.37	2.9E-05	0.032	1.98
SUCO	0.69	6.65	4.31	3.6E-05	0.036	2.03
NA	1.37	-0.66	4.31	3.6E-05	0.036	1.87
HEATR6	0.94	4.69	4.28	4.1E-05	0.039	1.92
TBC1D1	-0.66	5.79	-4.23	4.9E-05	0.041	1.75
ELAC1	-0.78	2.00	-4.21	5.3E-05	0.041	1.72
SLC50A1	0.89	6.46	4.22	5.1E-05	0.041	1.71
MYRFL	1.44	-0.32	4.23	5.0E-05	0.041	1.66
PGAP3	1.44	6.41	4.20	5.6E-05	0.041	1.61
ATG16L1	0.52	5.60	4.18	5.9E-05	0.041	1.57
ERBB2	1.90	9.42	4.17	6.1E-05	0.041	1.56
RHBG	1.45	-1.07	4.18	6.1E-05	0.041	1.39
NA	1.58	-2.68	4.16	6.4E-05	0.041	0.93
EPYC	2.31	-0.04	4.15	6.7E-05	0.041	1.46
NA	0.75	3.14	4.12	7.6E-05	0.046	1.42

Appendix 6. Statistical appendix for survival analysis

Preliminary multivariable model

A preliminary multivariable analysis of overall survival is shown in table 12b. Adjusted for other clinico-pathological variables, Luminal B (HR 2.15, 95% CI: 1.03 – 4.47, $p = 0.040$), HER2 (HR 3.06, 95% CI: 1.48 – 6.35, $p = 0.003$) and basal (HR 2.24, 95% CI: 1.24 – 4.07, $p = 0.008$) cancers are associated with a poorer overall survival. Shorter overall survival was also associated with positive lymph node status (HR 3.65, 95% CI: 2.08 – 6.38, $p < 0.001$) and a grade of 3 (HR 1.86, 95% CI: 1.06 – 3.26, $p = 0.030$). A more favourable overall survival was seen in patients treated with endocrine therapy (HR 0.47, 95% CI: 0.28 – 0.79, $p = 0.005$). When adjusted for other factors, chemotherapy no longer correlated with a shorter overall survival (HR 1.06, 95% CI: 0.64 – 1.78, $p = 0.817$).

Interactions between breast cancer subtypes and clinico-pathological variables

An interaction term composed of cancer type and age was added to the multivariable model. No significant interactions were seen between the cancer subtypes and age ($p > 0.05$). This analysis was repeated for the other five clinico-pathological variables (size, lymph node status, grade, endocrine therapy and chemotherapy). No significant interactions were found (all $p > 0.05$).

Testing of proportional hazards assumption

Scaled Schoenfeld residuals were used to test for the validity of the proportional hazards assumption in the preliminary model. This was tested for each co-variate and the whole model (table 13). The global test showed the model violated the proportional hazards assumption ($p = 0.031$). For the individual co-variates, lymph node status ($p = 0.124$) and

chemotherapy (p = 0.129) came closest to violating the assumption. For these two co-variates, the scaled Schoenfeld residuals were plotted against time (figure 9). For chemotherapy a pronounced downward slope was observed, which suggests the effect of chemotherapy on survival is not constant over time.

Table 13. Test of proportional hazards assumption for each co-variate and the whole preliminary model

Characteristic	***p* value**
Tumour type	
Luminal A	-
Luminal B	0.55
HER2	0.87
Basal	0.17
Age (≤ or > 50yrs)	0.85
Tumour size (≤ or > 20mm)	0.37
Lymph node status	0.12
Grade of 3	0.37
Endocrine therapy	0.42
Chemotherapy	0.13
Whole model	0.031

Figure 9. Plots of scaled Schoenfeld residuals over time for lymph node status and chemotherapy

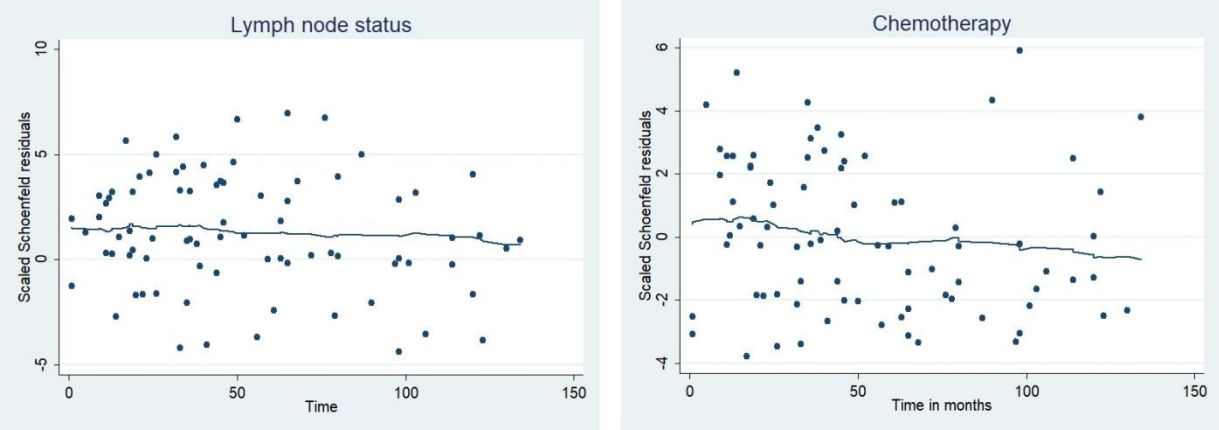
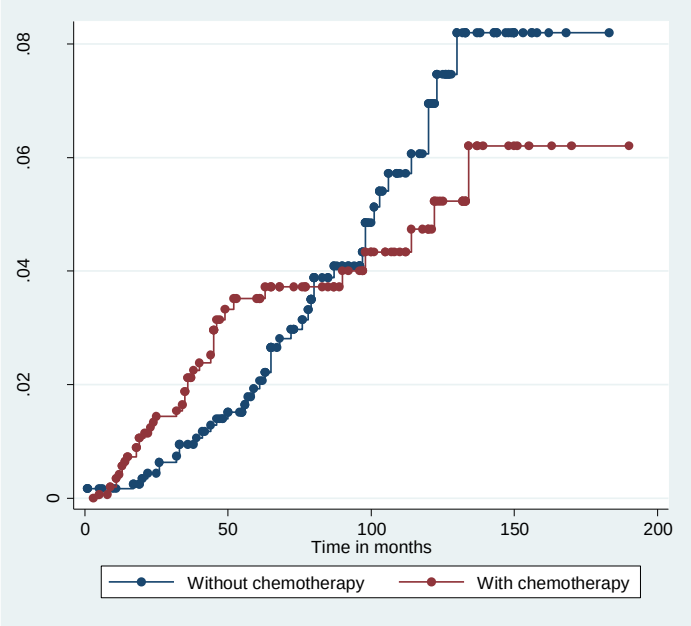


Table 14. Test of proportional hazards assumption for each co-variate and the whole model, multivariable model stratified by chemotherapy

Characteristic	*p* value
Tumour type	
Luminal A	-
Luminal B	0.55
HER2	0.92
Basal	0.14
Age (≤ or > 50yrs)	0.84
Tumour size (≤ or > 20mm)	0.36
Lymph node status	0.14
Grade of 3	0.39
Endocrine therapy	0.37
Whole model	0.25

Figure 10. Plot of cumulative baseline hazard, stratified by chemotherapy



The final multivariable analysis was modified to allow for stratification by chemotherapy, resulting in different baseline hazards for cases treated and not treated with chemotherapy (table 12a). Interpretation of the effect of chemotherapy may be problematic. However, this may not necessarily be an issue, as the preliminary multivariable model suggests chemotherapy does not have an effect on overall survival. The global test shows the

stratified model no longer violates the proportional hazards assumption (p = 0.245, table 14). Plots of the cumulative baseline hazards show the two groups (with and without chemotherapy) have different baseline hazards, vindicating the decision to stratify by chemotherapy (figure 10).

Test of overall goodness of fit and hypothesis testing

For the multivariable model stratified by chemotherapy, Cox-Snell residuals were calculated to assess the overall goodness of fit. A plot of the Nelson-Aalen estimate of the cumulative hazard function vs. the Cox-Snell residuals follows an approximately 45 degree line, which is supportive of a satisfactory overall model fit (figure 9). There was some divergence present for subjects with a longer overall survival. A likelihood ratio test for breast cancer subtype resulted in a p value of 0.018, compatible with breast cancer subtype having an effect on overall survival, and rejection of the null hypothesis.

Figure 9. Plot of the Nelson-Aalen estimate of the cumulative hazard function vs. the Cox-Snell residuals

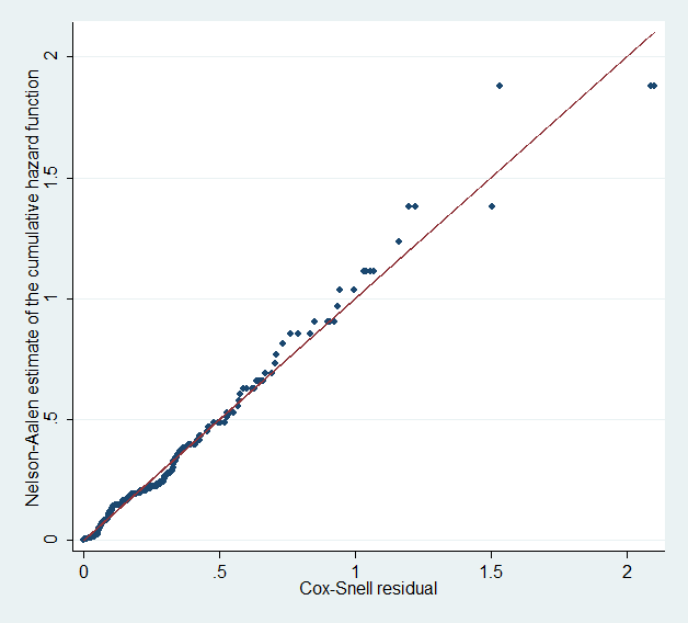


Figure 10. Plots of DFBETA vs. time for each of the variables in the multivariable model.

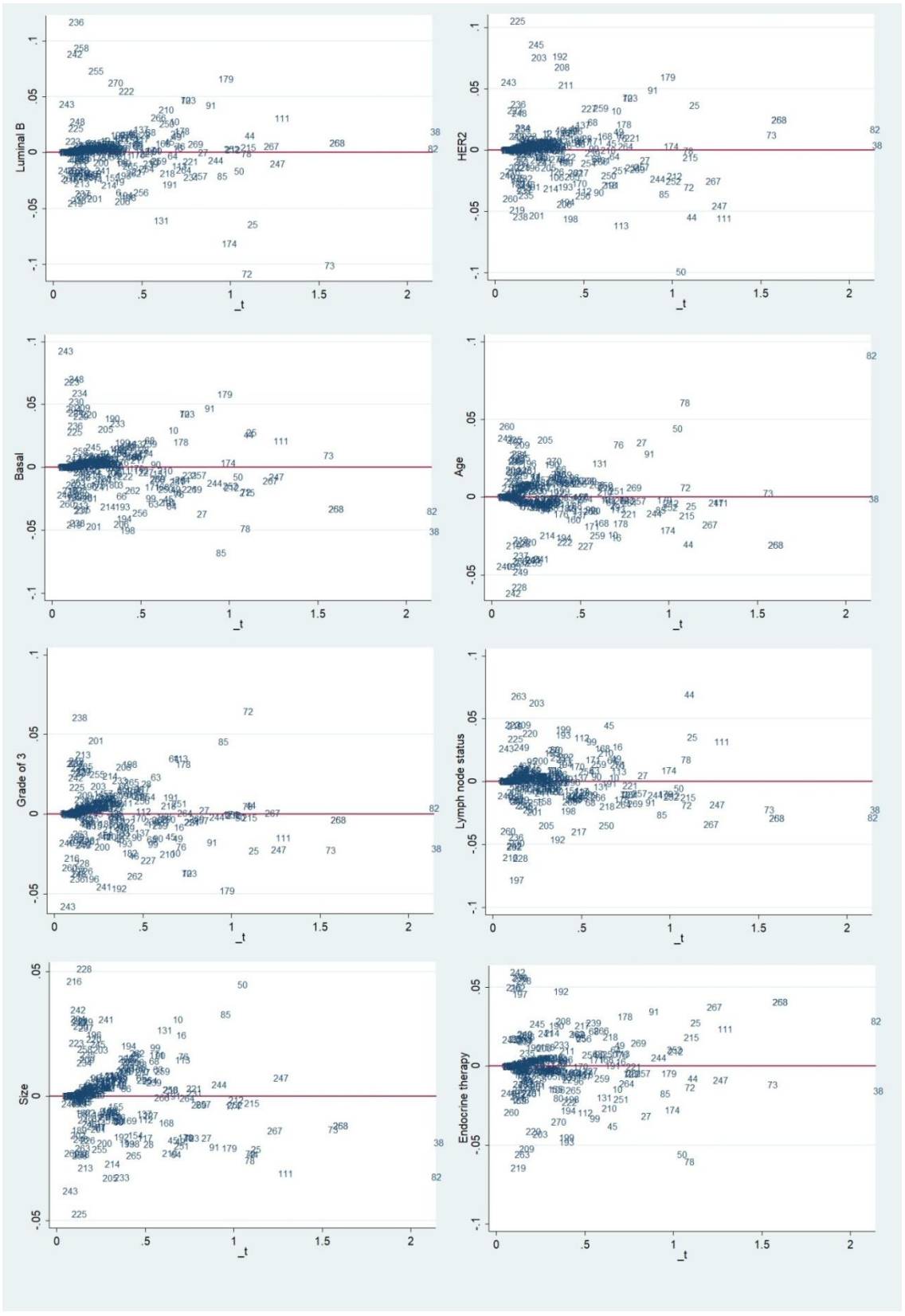


Table 15. List of influential cases, as determined by DFBETA, and their observations

ID	Size > 20mm	Lymph node positive	Age > 50 yrs	Grade 3	IHC type	Endocrine therapy	Chemotherapy	Survival (months)	Death
44	Yes	No	Yes	Yes	HER2	No	Yes	123	No
50	No	Yes	No	Yes	HER2	Yes	Yes	133	No
72	Yes	Yes	Yes	No	Luminal B	Yes	No	150	No
73	Yes	Yes	Yes	Yes	Luminal B	Yes	No	114	No
82	Yes	Yes	No	Yes	Basal	No	Yes	150	No
85	No	Yes	Yes	No	Basal	Yes	No	144	No
197	No	No	Yes	No	Luminal A	Yes	Yes	98	Yes
203	Yes	Yes	Yes	Yes	HER2	No	No	17	Yes
216	Yes	No	Yes	No	Luminal A	Yes	No	56	Yes
219	Yes	Yes	No	Yes	Luminal A	No	Yes	11	Yes
225	No	Yes	Yes	Yes	HER2	Yes	Yes	9	Yes
228	Yes	No	No	No	Luminal A	Yes	No	123	Yes
236	No	No	Yes	No	Luminal B	Yes	No	61	Yes
238	No	No	Yes	Yes	Luminal A	No	No	44	Yes
243	No	Yes	Yes	No	Basal	Yes	Yes	5	Yes
263	No	Yes	Yes	No	Luminal A	No	No	32	Yes

Identifying influential observations

DFBETA approximation of Cook's distances was used to estimate the influence of each observation on the regression estimate. For each variable, the residuals were plotted against time (figure 10). Sixteen influential cases were selected for further assessment (table 15). On review, subtyping based on ER, PR and HER2 status was correctly performed in all cases. In general, these influential observations fell into two groups. The first group was composed of cases with poor prognostic parameters (size > 20mm, positive lymph nodes, grade 3, non-luminal type) that were still alive after more than 10 years (no. 44, 50, 72, 73, 82 and 85). The second group was composed of cases with favorable prognostic parameters (size ≤ 20mm, negative lymph nodes, non-grade 3, luminal A type), but have an overall survival of less than 5 years (no. 203, 216, 219, 225, 238, 243, 263). While unusual, the observations were regarded as plausible, and were kept in the dataset.