Breast Cancer Intrinsic Subtypes: A Critical Conception in Bioinformatics

Heloisa Helena Zaccaron Milioli

B.Sc. in Biological Sciences

M.Sc. in Genetics

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

The University of Newcastle

Faculty of Science and Information Technology

School of Environmental and Life Sciences

Callaghan, NSW

Australia

September, 2016

Statement of Originality

The thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to the final version of my thesis being made available worldwide when deposited in the University's Digital Repository[1]*, subject to the provisions of the Copyright Act 1968.*

September, 2016

_____

Heloisa Helena Zaccaron Milioli

_____

Prof. Pablo Moscato

[1] Unless an Embargo has been approved for a determined period

Statement of Authorship

I hereby certify that the work embodied in this thesis contains a published paper/s/scholarly work of which I am a joint author. I have included as part of the thesis a written statement, endorsed by my supervisor, attesting to my contribution to the joint publication/s/scholarly work.

September, 2016

_____

Heloisa Helena Zaccaron Milioli

_____

Prof. Pablo Moscato

Acknowledgements

I would like to express my deep gratitude to Prof Pablo Moscato. I appreciated the guidance and encouragement he has provided throughout my time at Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine (CIBM). I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly. I would like to thank my co-supervisors, A/Prof Regina Berretta and Dr Jannette Sakoff, for their professional advices and useful critiques.

Special thanks should be given to Dr Carlos Riveros for his patient assistance and unconditional support. For all the extra hours he spent working with me, the constructive criticism and friendly advice during my PhD. I am sincerely grateful for sharing his truthful and illuminating views on a number of issues related to the breast cancer project. I am also grateful to Dr Renato Vimieiro for the valuable help in data management, extensive collaboration and magic proofreading. In particular, I thank Inna Tishchenko for her precious effort and intelligence in the analysis of the data.

I would like to extend my thanks to all CIBM students and collaborators who contributed with valuable discussions and enthusiastic encouragements: Ademir Cristiano Gabardo, Ahmed Shamsul Arefin, Amer Abu Zaher, Amir Salehipour, Chloe Warren, Claudio Sanhueza, Francia Jimenez, Leila Moslemi Naeni, Luke Mathieson, Mohammad Nazmul Haque, Nasimul Noman, Natalie de Vries, Nisha Puthiyedth, Shannon Fenn. Thanks for sharing the experience, either positive or negative. I also acknowledge the Hunter Medical Research Institute (HMRI) and University of Newcastle (UoN) staffs for sharing an amazing productive environment.

I express my warm thanks to Jennie Thomas for her enthusiastic support of students and researchers through a number of grants and scholarships. Being part of your *family* is a great honour and an enormous pleasure. Thanks for believing in my research and for funding my dreams. I am also grateful to A/Prof David Wild for guiding me throughout my visit to Bloomington, US.

Special thanks to my beloved family for their unconditional support!

Words cannot express how grateful I am to my mother, father, aunt, brother and nephew for all of the sacrifices that you've made on my behalf. Your positive energies have sustained me thus far. I would also like to thank my in-laws for striving towards my goal. Finally, I would like express appreciation to my beloved husband, Jorge André Martins. I would not be here without him, his patience and love.

To the best grandmother,

Helena Mafioleti Zaccaron

(Wherever you are)

Table of Contents

List of Figures

List of Tables

List of Equations

Abbreviations

AACR	Australasian Association of Cancer Registries
ACS	American Cancer Society
AIHW	Australian Institute of Health and Welfare
AJCC	American Joint Committee on Cancer
AR	Androgen receptor
ARI	Adjusted Rand Index
BL1	Basal-like 1
BL2	Basal-like 2
BLBC	Basal-like breast cancer
BLIA	Basal-like immune-activated
BLIS	Basal-like immune-suppressed
ChIP-chip	Chromatin immunoprecipitation on chip
CIBEX	Center for information biology gene expression database
CIBM	Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine
CGH	Comparative genomic hybridization
CNA	Copy number aberration
CNV	Copy number variation
CTD	Comparative Toxicogenomic Database
DamID	DNA adenine methyltransferase identification
DAVID	Database for Annotation, Visualization and Integrated Discovery
DDBJ	DNA Data Bank of Japan
DNA	Deoxyribonucleic acid
EBI	European Bioinformatics Institute
EGA	European Genome-Phenome Archive
EpCAM	Epithelial cell adhesion molecule
ER	Oestrogen receptor
FGED	Functional Genomics Data Society
FOIPPA	Freedom of Information and Protection of Privacy Act
FS	Feature Selection
GEO	Gene Expression Omnibus
HER2	Human epidermal growth factor receptor 2
HREC	Human Research Ethics Committee

HTC	High content screening
HTS	High-throughput screening
ICGC	International Cancer Genomics Consortium
IDC	Invasive ductal carcinoma
IHC	Immunohistochemical
IHGSC	International Human Genome Sequencing Consortium
ILC	Invasive lobular carcinoma
IM	Immunomodulatory
JS	Jensen Shannon
Ki-67	Antigen identified by monoclonal antibody Ki-67
***k*NN**	*k* nearest neighbours
LAR	Luminal androgen receptor
lincRNA	long intergenic non-coding RNA
MA	Memetic algorithm
MCC	Matthews' Correlation Coefficient
MDL	Minimum Description Length Principle
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
MIAME	Minimum Information About a Microarray Experiment
microRNA	miRNA
MGED	Microarray Gene Expression Data Society
MS	Menopausal status
MST	Minimum Spanning Tree
NCBI	National Center for Biotechnology Information
NOS	Not otherwise specified
NPI	Nottingham prognostic score
NSC	Nearest Shrunken Centroids
NST	No special type
ORF	Open reading frame
PIPA	Personal Information Protection Act
PIPEDA	Personal Information Protection and Electronic Documents Act
PR	Progesterone receptor
PRC	Priority Research Centres
RHD	Research Higher Degree
RNA	Ribonucleic acid
ROCK	Research Online Cancer Knowledgebase
RT-PCR	Reverse-transcriptase Polymerase chain reaction

SAM	Sentrix® Array Matrix
SCM	Subtype Classification Model
SNP	Single nucleotide polymorphism
SSP	Single Sample Predictor
TCGA	The Cancer Genome Atlas
TEND	Trends in the Exploration of Novel Drug targets
TNBC	Triple-negative breast cancer
TNM	Tumour size, nodes, metastasis
TTD	Therapeutic Target Database
UCSC	University of California Santa Cruz
WEKA	Waikato Environment for Knowledge Analysis

Achievements

During my PhD, I applied for grants; submitted manuscripts for publication; and attended workshops, conferences and seminars. The relevant achievements are listed as follows:

Grants Awarded

- Hunter Medical Research Institute, 2014.

JENNIE THOMAS MEDICAL RESEARCH TRAVEL GRANT (AUD $10,000)

- Hunter Cancer Research Alliance, 2015.

HCRA TRAVEL GRANT (AUD $1,000)

- Hunter Cancer Research Alliance, 2016.

HCRA PhD Research Award 2016 (AUD $5,000).

- EMBL Australia PhD, 2016.

Travel Grant to attend the 18th EMBL PhD Symposium (AUD $3,000).

- XII ELAG Course Fellowship (USD $700)

Instituto Genética Para Todos – Brazil (unable to attend)

Papers Published in Journals

MILIOLI, H.H.; VIMIEIRO, R.; RIVEROS, C.; TISHCHENKO, I.; BERRETTA, R.; MOSCATO, P. (2015) The discovery of novel biomarkers improves breast cancer intrinsic subtype prediction and reconciles the original PAM50 labels in the METABRIC data set. *PLoS One*; 10(7): 0129711. doi: 10.1371/journal.pone.0129711

MILIOLI, H.H. (2015). The IMPAKT of breast cancer research: fundamental science and clinical medicine. *Future Science OA*; (0). doi: 10.4155/fso.15.69

MILIOLI, H.H.; VIMIEIRO, R.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. (2016) Iteratively refining breast cancer intrinsic subtypes in the METABRIC dataset *BioData Mining*; 9:2. doi: 10.1186/s13040-015-0078-9

TISHCHENKO, I.; **MILIOLI, H.H.**; RIVEROS, C.; MOSCATO, P. (2016) Extensive Transcriptomic and Genomic Analysis Provides New Insights about Luminal Breast Cancers. *PLoS One; 11*(6): e0158259. doi: 10.1371/journal.pone.0158259

MILIOLI, H.H. Life as an early career researcher: interview with Heloisa Helena Milioli**.** *Future Science OA*; 1(4) (2016). doi: 10.4155/fsoa-2016-0033.

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. *BMC Med Genomics*; 10(1):19 (2017). doi: 10.1186/s12920-017-0250-9.

MILIOLI, H.H.; RIVEROS, C.; VIMIEIRO, R.; BERRETTA, R.; MOSCATO, P. Meta-features modelling gene expression imbalances: an innovative strategy for breast cancer subtype prediction. Manuscript in preparation to be submitted for publication at Genomics, Proteomics and Bioinformatics (GPB).

Abstracts Published

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; SAKOFF, J.; BERRETTA, R.; MOSCATO, P. Consensus on breast cancer cell lines classification for an effective and efficient clinical decision-making. *IMPAKT 2015 Breast Cancer Conference. Annals of Oncology* 26 (suppl 3):iii32-iii33 (2015). doi: 10.1093/annonc/mdv121.08

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Molecular classification of basal-like breast cancer subtypes based on predictive survival markers. *IMPAKT 2015 Breast Cancer Conference. Annals of Oncology.* 26 (suppl 3):iii17-iii18 (2015). doi: 10.1093/annonc/mdv117.11

MILIOLI, H.H., TISHCHENKO, I., RIVEROS, C., BERRETTA, R. & MOSCATO, P. (2015) Basal-like breast cancer subgroups uncovered by genomic and transcriptomic profiles and overall survival outcomes. *Hunter Cancer Research Alliance Annual Symposium*. *Asia-Pacific Journal of Clinical Oncology* 11(Suppl. 5):6-19 (2015). doi: 10.1111/ajco.12444

TISHCHENKO, I., **MILIOLI, H.H.**, RIVEROS, C. & MOSCATO, P. How intrinsic are luminal breast cancer subtypes? *Hunter Cancer Research Alliance Annual Symposium*. *Asia-Pacific Journal of Clinical Oncology* 11(Suppl. 5):6-19 (2015). doi: 10.1111/ajco.12444

MILIOLI, H.H., SANHUEZA, C., BERRETTA, R. & MOSCATO, P. (2015) ABSTRACT P40 Breast Cancer Molecular Portraits of Intrinsic Subtypes and Integrative Clusters in the METABRIC Data Set. *Hunter Cancer Research Alliance Annual Symposium*. *Asia-Pacific Journal of Clinical Oncology* 12(Suppl. 6):13-34 (2016). doi: 10.1111/ajco.12618

Oral Presentations

MILIOLI, H.H.; VIMIEIRO, R.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Refining the breast cancer molecular subtypes in the METABRIC data set. *World Congress on Controversies in Breast Cancer (CoBRA), 2015. Melbourne, AU.*

MILIOLI, H.H.; SANHUEZA, C.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Breast cancer molecular portraits of intrinsic subtypes and integrative clusters in the METABRIC data set. **Young Scientist Award** *2nd World Congress on Controversies in Breast Cancer (CoBrCa) 2016. Barcelona, Spain.*

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Basal-like breast cancers uncovered by genomic and transcriptomic profiles and patients' overall survival. *Sydney Cancer Conference (SCC) 2016. Sydney, AU.*

Poster Sessions

MILIOLI, H.H.; VIMIEIRO, R.; RIVEROS, C.; SAKOFF, J.; BERRETTA, R.; MOSCATO, P. Breast Cancer Subtypes Individuation Driving Novel Drug Targets for Tailored Therapies. *Translational Cancer Research Conference, 2013. Newcastle, AU.*

MILIOLI, H.H.; VIMIEIRO, R.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Identification of novel biomarkers for predicting breast cancer intrinsic subtypes. *ASMR Satellite Scientific Meeting, 2014. Newcastle, AU.*

MILIOLI, H.H.; RIVEROS, C.; VIMIEIRO, R.; MOSCATO, P. Meta-features as predictors of breast cancer intrinsic subtype in the METABRIC gene expression data set. **Best Poster Award (Bronze Medal)** *International Conference on Bioinformatics, 2014. Sydney, AU.*

RIVEROS, C.; **MILIOLI, H.H.**; VIMIEIRO, R.; BERRETTA, R.; MOSCATO, P. Discovery of gene interactions by GPU-enabled computation of pairwise expression level metafeatures. *International Conference on Bioinformatics, 2014. Sydney, AU.*

MILIOLI, H.H.; RIVEROS, C.; VIMIEIRO, R.; TISHCHENKO, I.; BERRETTA, R.; MOSCATO, P. Using an iterative approach to reclassify sample subtypes in the METABRIC breast cancer data set. **Best Poster Award (Third place)** *BioInfoSummer, 2014. Melbourne, AU.*

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Basal-like breast cancer subsets revealed by survival predictor genes. *ASMR Satellite Scientific Meeting, 2015. Newcastle, AU.*

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Molecular classification of basal-like breast cancer subtypes based on predictive survival markers. *IMPAKT 2015 Breast Cancer Conference. Brussels, BE.*

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; SAKOFF, J.; BERRETTA, R.; MOSCATO, P. Consensus on breast cancer cell lines classification for an effective and efficient clinical decision-making. *IMPAKT 2015 Breast Cancer Conference. Brussels, BE.*

MILIOLI, H.H.; RIVEROS, C.; VIMIEIRO, R.; MOSCATO, P. Meta-features predicting gene expression imbalances across breast cancer intrinsic subtypes. *EMBL Australia PhD Symposium, 2015. Melbourne, AU.*

TISHCHENKO, I., **MILIOLI, H.H.**, RIVEROS, C. & MOSCATO, P. How intrinsic are luminal breast cancer subtypes? *Hunter Cancer Research Alliance Annual Symposium, 2015. Newcastle, AU.*

MILIOLI, H.H., TISHCHENKO, I., RIVEROS, C., BERRETTA, R. & MOSCATO, P. Basal-like breast cancer subgroups uncovered by genomic and transcriptomic profiles and overall survival outcomes. *Hunter Cancer Research Alliance Annual Symposium, 2015. Newcastle, AU.*

NAENI, L., **MILIOLI, H.H.**, TISHCHENKO, BERRETTA, R. & MOSCATO, P. (2015) A New Clustering Approach Identifies Candidate Biomarkers for Breast Cancer Subtyping. *BioInfoSummer, 2015*. *Sydney, AU.*

MILIOLI, H.H.; RIVEROS, C.; VIMIEIRO, R.; MOSCATO, P. Meta-features predicting gene expression imbalances across breast cancer intrinsic subtypes. **Best Poster Presentation** *BioInfoSummer, 2015. Sydney, AU.*

Other Presentations

Confirmation Year Presentation

Faculty of Science and IT. The University of Newcastle, 2013.

RHD candidates are required to submit the 'Confirmation Year Report' and present the research overview. In August 2013, I presented the preliminary results in the Faculty of Science and IT as an open seminar.

HCRA, ECR and PhD Student (HEAPS) Seminar Series

Hunter Medical Research Institute. The University of Newcastle, 2014 and 2015.

The HEAPS seminar series are organised by the Hunter Cancer Research Alliance (HCRA) for RHD students and supervisors. It is an opportunity for researchers to practice presenting (and critiquing) work in a local and highly supportive environment. In 2014 and 2015, I presented and discussed the results of my research as well as supported other researchers' work.

HUBS3302 Bioinformatics Mini-Conference

Faculty of Health and Medicine. The University of Newcastle, 2014 and 2015.

The purpose of this event is to inspire students in the field and, specially, in their final project for the discipline. In the 2014 and 2015 Bioinformatics Mini-Conference, organised by Belinda Goldie, I presented my research on breast cancer.

Science and Engineering Challenge

Faculty of Engineering and Built Environment. The University of Newcastle, 2014, 2015 and 2016.

The 'Science and Engineering Challenge' organise a number of events aimed at challenging students of all different ages in Science and Engineering. As part of the team, I coordinated activities in Tamworth (2014), Muswellbrook (2014), Dubbo (2015), Newcastle (2015), Central Coast (2016) and Narrabri (2016), and presented my research to the Rotary International (Australian Rotary Districts) in Tamworth and Dubbo.

Faculty Progress Seminar

Faculty of Science and IT. The University of Newcastle, 2015.

Students in the Faculty of Science and IT are required to present a Progress Seminar after completing 2 to 3 years of a PhD. In June 2015, I discussed the overall aims and results of my research and outlined my thesis to fellow RHD candidates and academics in the school.

Google Computer Science for High Schools

Faculty of Engineering and Built Environment. The University of Newcastle, 2015 and 2016.

The University of Newcastle's Computer Science 4 High Schools (CS4HS) is an introductory workshop for in-service and pre-service teachers (both at primary and secondary level), and career advisors focused on developing competencies included in the recently approved Digital Technologies curriculum and is accredited by BOSTES. In three events, I had the opportunity to explain the relevance of computer science to analyse biological/medical data.

Relevant Activities

Course: **Winter School in Mathematical and Computational Biology**

University of Queensland (UQ), Brisbane, 2013.

The winter school introduced mathematical and computational biology and bioinformatics to advanced undergraduate and postgraduate students, postdoctoral researchers and others working in the field. Important topics, such as mathematics, statistics, computer science, information technology, biology, chemistry and medical sciences and engineering, were selected for each day. Lectures and interactive discussions were ministered by national and international authorities.

Course: **European Molecular Biology Laboratory (EMBL) Australia PhD Course**

Australian National University (ANU), Canberra, 2014.

EMBL Australia offered to sixty students a unique introduction to research with the annual EMBL Australia PhD Course. The two-week program shows students how their research fits into the bigger picture of science, and introduces a range of fields including: bioinformatics, developmental biology, genomics, systems biology and regenerative medicine.

Course: **European Molecular Biology Laboratory (EMBL) Australia PhD Course**

Welcome Genome Campus, Hinxton, UK, 2016.

This course introduced a wide range of post-genome techniques including practical experience in performing (1) high-throughput RNAi screening, (2) microarray gene expression analysis and interpretation, using a range of commercial and academic software tools, (3) next-generation sequencing and alignment; (4) protein-protein interaction networks and integration with other data sources, and (5) pathway analysis. Laboratory work was based on the training of state-of-the-art methods and complementary approaches to address biological and medical questions.

Training: **Collaborative Research Training in Human Genetics and Bioinformatics**

Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine (CIBM). The University of Newcastle, 2014.

The CIBM established a research-training program in 2014 that contributed to improve the capacity of young investigators to conduct human genetics and bioinformatics research. The training promoted scientific collaborations between the University of Newcastle and international (undergraduate) students. The proposed program provided opportunities to generate expertise that could contribute to the

long-term goal of harnessing genetic knowledge and bioinformatics skills to diagnose, prevent, or treat diseases. Training activities were coordinated, facilitated and monitored by Prof. Pablo Moscato, A/Prof Regina Berretta and PhD student Heloisa Helena Milioli.

Short-term Exchange Program: **Cheminformatics and Chemogenomics Research Group (CCRG)**

Indiana University (IU), Bloomington USA, 2015.

Further investigation on cheminformatics and toxicogenomics has been developed in collaboration with A/Prof. David J. Wild (May/June 2015), at the School of Informatics and Computing in Bloomington (USA). These approaches were used to delineate drug-targets for basal-like breast cancer, one of the most aggressive subtypes with limited therapy response. Further research, however, is required to design and perform in vitro tests.

Organising Committee: **Australian Society for Medical Research (ASMR) Satellite Scientific Meeting**

Hunter Medical Research Institute (HMRI), Newcastle, 2015.

This event showcases the recent research achievements of Hunter scientists, encourages postgraduate and student interactions and fosters collaboration between researchers within the Faculty of Health and Medicine, HMRI and the international community. In the 2015 edition, I was member of the committee.

Abstract

Breast cancers have been uncovered by high-throughput technologies that allow the investigation at the genomic, transcriptomic and proteomic levels. In the early 2000s, the gene expression profiling has led to the classification of five intrinsic subtypes: luminal A, luminal B, HER2-enriched, normal-like and basal-like. A decade later, the spectrum of copy number aberrations has further expanded the heterogeneous architecture of this disease with the identification of 10 integrative clusters (IntClusts). The referred classifications aim at explaining the diverse phenotypes and independent outcomes that impact clinical decision-making. However, intrinsic subtypes and IntClusts show limited overlap. In this context, novel methodologies in bioinformatics to analyse large-scale microarray data will contribute to further understanding the molecular subtypes. In this study, we focus on developing new approaches to cover multi-perspective, highly dimensional, and highly complex data analysis in breast cancer. Our goal is to review and reconcile the disease classification, underlying the differences across clinicopathological features and survival outcomes. For this purpose, we have explored the information processed by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC); one of the largest of its type and depth, with over 2000 samples. A series of distinct approaches combining computer science, statistics, mathematics, and engineering have been applied in order to bring new insights to cancer biology. The translational strategy will facilitate a more efficient and effective incorporation of bioinformatics research into laboratory assays. Further applications of this knowledge are, therefore, critical in order to support novel implementations in the clinical setting; paving the way for future progress in medicine.

Keywords

Breast cancer, Intrinsic subtypes, Integrative clusters, IntClusts, Microarray, Gene expression, Copy number aberration, MicroRNA, METABRIC, Feature selection, Data mining, Ensemble learning, Prediction models, Classification

CHAPTER 1

1. INTRODUCTION AND OVERVIEW

Chapter 1 is the thesis prospectus that will assist the reader in understanding the context of this study on breast cancer. The first two topics, **1.1 Breast Cancer: an Overview** and **1.2 Bioinformatics Resources and Tools**, contextualise the significance of investigating this disease using promising bioinformatics approaches. Next, the **1.3 Research Motivation** enlightens the main points underlying this study and the most important questions to be addressed. The main goals and the specific aims are defined, for each chapter, in **1.4 Research Aims and Thesis Structure**, which summarizes the thesis content and the headings choice. Additionally, the achievements obtained during the research higher degree (RHD) candidature are listed in **Achievements** and the corresponding work developed at The University of Newcastle, between July 2012 and July 2016, in **Relevant Activities**. The last section,

1.5 References, cites the relevant publications supporting this introductory section.

1.1 Breast Cancer: an Overview

Breast cancer is the second most common type of cancer overall, with a high rate of incidence among women worldwide (Ferlay et al., 2015; Siegel et al., 2014). In Australia, it is the most frequently reported cancer in females, excluding non-melanoma skin cancers. The Australian Institute of Health and Welfare estimated that more than 15270 women were diagnosed with breast cancer and 3000 died from this disease in 2014 (AIHW & AACR, 2014a, 2014b). The lifetime risk of developing breast cancer is 1 in 11 before the age of 75, and 1 in 8 before the age of 85 (AIHW & AACR, 2014b). Moreover, a projection of cancer incidence in Australia for 2020 predicts 17210 new cases of breast cancer among women (AIHW, 2012). Despite the increasing incidence, reductions in mortality have been reported, corresponding with advances in screening policies and treatment protocols (AIHW & AACR, 2014b).

As with most cancers, the causes of breast cancer are not completely understood. The risks of developing breast cancer have been related to a range of aspects including age, race, ethnicity, lifestyle and environment. Hormonal and reproductive factors are particularly important and underlie the aspects of menarche, menopause, parity, breastfeeding, oral contraceptive intake and menopausal hormone replacement therapy (Barnard et al., 2015; Forman et al., 2015). Familial history is another important risk factor. Patients with germ line mutations in *BRCA1* or *BRCA2* genes show an increased predisposition to breast and ovarian carcinomas (Miki et al., 1994; Wooster et al., 1995). Two other genes associated with rare cancer syndromes, *TP53* (Li–Fraumeni syndrome) (Engreitz et al., 2011) and *PTEN* (Cowden syndrome) (Guenard et al., 2007; Lynch et al., 1997), also contribute to the rise in breast cancer cases (Lalloo & Evans, 2012). Low-to-moderate penetrant genes/loci may be also involved, such as *ATM*, *BRIP1*, *CDH1*, *CHEK2*, *NBS1*, *PALB2*, *RAD51* and *STK11* (Hollestelle et al., 2010; Nevanlinna & Bartek, 2006; Shuen & Foulkes, 2011). The number of contributing genes,

however, remains under investigation (Duffy et al., 2017; Oldenburg et al., 2007; Stratton & Rahman, 2008; Topalian et al., 2016).

Sporadic mutations, on the other hand, vary markedly between individual tumours (Stephens et al., 2012). Genes previously implicated in breast cancer (*PIK3CA*, *AKT1*, *GATA3*, *RB1*, *MLL3*, *MAP3K1* and *CDKN1B*), and a number of novel significantly mutated genes have been identified, including *AFF2*, *CBFB*, *NF1*, *RUNX1*, *PIK3R1*, *PTPN22*, *PTPRD*, *TBX3*, *SF3B1* and *CCND3* (TCGA, 2012). The presence of multiple drivers has been associated with cancer molecular heterogeneity and subclonal evolution. Changes in gene expression levels have also emerged as biomarkers for breast cancer subtyping, such as hormone receptors (*ESRI*, *PGR*, *ERBB2*), basal cytokeratins (*KRT5*, *KRT6*, *KRT17*), markers of proliferation (*AURKA*, *MELK*, *MKI67*, *PCNA*), and growth factor receptors (*EGFR*, *VEGFR*) (Rakha et al., 2008). Overall, these markers have shown potential in predicting the disease behaviour, patients' outcome, and are able to guide clinical decision-making.

Multi-gene lists and predictor models have been used to reduce the multidimensional complexity of breast cancers. The signatures have been reported within the molecular patterns strongly correlated to clinical prognosis (Fan et al., 2011; Wang et al., 2005), disease progression (Seoane et al., 2014; Venet et al., 2011), and patient survival (Naderi et al., 2006). Mammaprint® (Agendia, Huntington Beach, CA) and Oncotype DX® (Genome Health Inc, Redwood City, CA), two commercial assays, are standard examples of genome supervised predictors (Glas et al., 2006; Paik et al., 2004). The main purpose is to either delineate treatment or anticipate the patient's outcome (van't Veer & Bernards, 2008) by estimating the likelihood of distant recurrence in the five years following diagnosis and the risk of metastasis, respectively. Alternatively, the PAM50 method has been proposed to classify tumour subtypes according to the correlation with expression values of 50 genes, defined as centroids (Parker et al., 2009). New concepts underlying subtype prediction are based on risk models that incorporate molecular signatures shared among tumours with analogous behaviour, for a group-based treatment design.

In the post-genomic era, microarray integrated applications have enabled the identification of relevant markers for a range of distinct purposes: early detection, disease prognosis, drug target and tailored therapy (Kulasingam et al., 2010). The genes encountered in different studies, nonetheless, are still highly variable, non-overlapping, and generally require specialised investigative technologies (Borrebaeck, 2017; Dolled-Filhart et al., 2006), . Despite the clear impact molecular profiling has made in improving the way breast cancer is now perceived as multiple entities, there is still a great deal of work ahead. The advent of high-

throughput technologies and the massive amount of information produced offer scientists a unique opportunity for uncovering new portraits of breast cancer and determining the mechanisms behind the clinical heterogeneity of this disease. Mining putative biomarkers from 'Big Data' sources is therefore a valuable strategy and a great challenge in the field. These sources need to be well defined and, ultimately, translated and transformed into laboratory assays and clinical applications. The overarching goal is to delineate more effective treatments and optimal response in patient care.

1.2 Bioinformatics Resources and Tools

High-throughput biological data can now be analysed and interpreted using the interdisciplinary field of bioinformatics. Bioinformatics combines computer science, statistics, mathematics, and engineering for the development of software tools and data analytics methods (Yigitoglu et al., 2015). These subjects are strengthened by the concepts and processes involving DNA, RNA and proteins; complex molecules engaged in dynamic and interactive systems. In this setting, bioinformatics has firmly established itself as a new discipline in response to the accelerating demand for a flexible and intelligent means of storing, managing and querying and, most importantly, understanding large amounts of biological information. This subject has, therefore, an endless potential (in the 21st century) to evolve as it faces new perspectives that simultaneously emerge with data collection and sample analysis (Carrey & Stodden, 2010).

The release of the public draft of the human genome was the culmination of a pivotal bioinformatics and biological endeavour. It brought with it the promises of improving our understanding of diverse aspects of molecular biology and clinical medicine (Attwood et al., 2011; Lander et al., 2001). Initially the main concern of bioinformatics was the creation and maintenance of databases for storing biological information such as nucleotide and amino acid sequences. Lately, the emphasis has shifted towards actionable insights for the analysis and interpretation of data, involving entire cohorts stored across different databases (Hood, 2003). These databases were designed with independent interfaces whereby researchers could access existing files, submit new data or revise the stored data. The major sequence database is GenBank, maintained by the National Center for Biotechnology Information (NCBI) at the National Institutes of Health, which comprises an annotated collection of DNA and protein sequences. Alternative portals used to browse the human genome (and other sequence-based) data are the UCSC Genome Browser, developed at the University of California Santa Cruz

(UCSC), and Ensembl, a framework of the European Bioinformatics Institute. The databases are constantly upgraded in the integrated and interconnected online environment (Baxevanis, 2009).

In the early 2000's, microarrays have gone from obscurity to being almost ubiquitous in biological research. Members of the Microarray Gene Expression Data Society (MGED), now known as Functional Genomics Data Society (FGED), emphasised the importance of databanks to store and share the data within public domains. Peer reviewed repositories also support academic and industry standards by promoting wide applications in comprehensive and expanded investigations to facilitate biological and biomedical discoveries. Examples of databanks are: Gene Expression Omnibus (GEO) stored at the NCBI, ArrayExpress from the European Bioinformatics Institute (EMBL-EBI) and CIBEX maintained by the DNA Data Bank of Japan (DDBJ). In order to improve the transparency of microarray studies, authors have to supply details of samples, protocols and platforms according to the Minimum Information About a Microarray Experiment (MIAME) guidelines (Ball et al., 2004; Engreitz et al., 2011). For instance, research consortia such as the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), The Cancer Genome Atlas (TCGA) and International Cancer Genomics Consortium (ICGC) have published large collections in breast cancer (Verhaak & Mills, 2012). However, more robust data sources are yet to be shared (Hayden, 2014).

Methodologies for microarray analysis have progressed from simple visual assessments of results to a weekly deluge of papers that describe novel algorithms for validating changes in the gene expression profile. High levels of data analysis, consequently, require some key components – design, pre-processing, inference, classification and validation – to address important concepts where consensus has emerged, or in areas yet obscure (Allison et al., 2006). Although the available procedures might be bewildering to biologists, bioinformaticians can recognise the competence among the different methods used to deal with multiple data sets (Altman & Miller, 2011). The main challenge is the efficient analysis of microarray data generated by different research groups across distinct platforms and technologies (Moreau et al., 2003); alongside the integration of information sources in genomics, transcriptomics, proteomics and epigenomics (Su et al., 2012).

Array technologies such as comparative genomic hybridisation (CGH), single nucleotide polymorphism (SNP) detection, gene expression profiling, chromatin immunoprecipitation on chip (ChIP-on-chip) and DNA adenine methyltransferase identification (DamID) are promising new tools in medical research (Sims, 2009). In particular, high-resolution DNA copy number aberration (CNA) and variation (CNV) have shown a potential role in breast cancer research (Xu et al., 2012), impacting expression levels and protein structure

(Krepischi et al., 2012; Stankiewicz & Lupski, 2010). These high-throughput measurements enhance the molecular investigation and allow a faster transition from laboratory findings to novel applications in the clinical practice. By modelling and simulating these measures using innovative approaches, researchers expect to make rapid progress in science and uncover dynamic and complex biological systems (Yao, 2002).

Bioinformatics is widely used for the identification of candidate genes and proteins, from normal cellular activities and altered states in different diseases. It has led to the better understanding of intrinsic mechanisms and molecular pathways driving the phenotype of a variety of diseases (Huang et al., 2011). Ultimately, the interdisciplinary field allows robust data analysis to create more reliable global perspectives from which unifying principles in bioinformatics can be discerned to yield a healthier future with personalised medicine (Yulug & Gur-Dedeoglu, 2008).

1.3 Research Motivation

Breast cancer is a common and heterogeneous disease affecting women of all ages. This heterogeneity poses significant challenges not only in breast cancer management, but in understanding the biology of tumours and the course of this disease (Dawson et al., 2013). Breakthroughs in molecular biology, however, have influenced clinical decision-making. In particular, bioinformatics has allowed researchers to inquire more deeply into the nature of breast cancer. A possible direction for future research is to look into different groups of patients with similar behaviour, focusing on group-based intervention strategies in applied medicine (Weigelt et al., 2012).

In practice, the large scale collection of data raises the urgent need to integrate and utilise these robust resources for biomarker discovery and biomedical applications. Relatively few methods, however, have shown the capacity of dealing with ‘Big Data’ information sources (Dutta et al., 2012). Therefore, novel approaches are valuable for strengthening investigations into breast cancer, adding to the breadth of known medicine (Colombo et al., 2011). These facts frame not only the study motivation but also the big challenges explored throughout my thesis.

1.3.1 Research Questions

Assuming the intrinsic heterogeneity of human breast cancers and the lack of consensus for the disease classification, several important questions need to be answered. Before inquiring into those, it is important to stress that a proper classification of breast tumours, especially with regard to the analysis of molecular profiles, would lead to the identification of accurate biomarkers and to the definition of robust prediction models. The current breast cancer subtypes – or the distinct molecular diseases – need to be further explored and validated, and remain the objects of medical research. That said, the questions supporting this study are:

"How many groups or different subtypes could be clearly identified in breast cancer disease using gene expression microarray data? Are they molecularly and clinically well defined?"

"Which genes or signatures are able to individualise the different breast cancer subtypes? Are these genes relevant targets for tailored treatment?"

"How could molecular data, including genome and transcriptome microarrays, be better combined or integrated to improve the understanding of the disease or the subtypes' classification?"

"Is it possible to link cell line profiles with the breast cancer subtypes in order to provide consistent information for 'in vitro' drug tests?"

1.4 Research Aims and Thesis Structure

This is an integrated investigation of breast cancer disease, concentrated on large-scale "Big Data" analytics. The overall research objective is to improve the understanding of breast cancer molecular architecture by applying sophisticated bioinformatics algorithms. Breast tumours are evaluated by the correlations of gene expression, microRNA profiling and CNA patterns, and by using methods based in computer science, statistics, mathematics and engineering. Our ultimate goal is to be able to link transcript variants and pathways into particular subgroups with individual biomarkers for diagnosis, prognosis and treatment of breast carcinomas.

With regard to the structure of the document, it is important to stress that every chapter is an independent section containing a thorough description of methods, results and discussion, in the context of the literature. Ultimately, a major conclusion summarises the whole body of my research, pointing to overall remarks of each section. To comprehend the thesis structure, the appropriate content and specific aims are defined for each chapter as follow:

Chapter 2 – Breast Cancer: Current Status and Perspectives

Basic concepts covering breast cancer incidence, classification and subtyping are provided in this chapter. The extensive literature review embraces the actual data, guidelines and protocols used to direct clinical decision-making and to guide future research outcomes. Breast cancer related studies were reviewed to establish a consensus on the knowledge of this disease.

Chapter 3 – Microarray Technologies and 'Omics' Data Sets

The gene expression microarray methodology is carefully described and compared against the well-established platforms Illumina and Affymetrix. Public microarray data sets in breast cancer are later detailed in the context of this study. More information on the ethics application and approval is also provided in this chapter.

Chapter 4 – Identification of Novel Biomarkers for Breast Cancer Subtyping

This chapter refers to the manuscript published in *PLoS One*[2]. Here I introduce a valuable strategy to deal with the challenges of identifying and predicting breast cancer intrinsic subtypes. In this chapter, I aim to:

- *Identify novel biomarkers for subtype individuation by exploring the competence of a newly proposed method named CM1 score, and;*
- *Improve class prediction by applying an ensemble learning approach, as opposed to the use of a single classifier.*

[2] Milioli, H. H., Vimieiro, R., Riveros, C., Tishchenko, I., Berretta, R., & Moscato, P. (2015). The Discovery of Novel Biomarkers Improves Breast Cancer Intrinsic Subtype Prediction and Reconciles the Labels in the METABRIC Data Set. *PLoS One, 10*(7), e0129711.

Chapter 5 – Iteratively Refining the METABRIC Subtype Labels

The content of this chapter is published in *BMC BioData Mining*[3]. It complements the previous analysis by filling the gaps identified in the first paper. Towards the development of more robust and reliable strategies, I aim to:

- *Consistently discriminate the breast cancer intrinsic subtypes and improve class prediction in the METABRIC data set, using an iterative approach.*

Chapter 6 – Meta-features for Predicting Breast Cancer Intrinsic Subtypes

This chapter is board at *Genomics, Proteomics & Bioinformatics*[4]. It contains a novel systematic approach based on mathematical modelling, feature selection and data mining that considers pairwise probes, meta-features, to explore new 'constructs' for distinguishing breast cancer subtypes. With this novel approach, I aim to:

- *Identify meta-features at a minimum template able to predict and explain the breast cancer intrinsic subtypes.*

Chapter 7 – Basal-Like Breast Cancer Subtype

This chapter is also published as an article at *BMC Medical Genomics*[5]. It contains integrative data from basal-like breast cancers, including the gene expression, miRNA profiles, copy number aberrations and survival outcomes. These data are used to understand the subtype contradictory behaviour and limited therapy response, with the aim to:

- *Identify survival markers that are able to stratify basal-like breast cancers with distinct molecular profiles, clinical features and disease outcomes.*

By centring attention on public databases, I investigate the connection of biomarkers and drug-targets in breast cancer disease (in **7.7 Supporting Information**), with the aim to:

- *Provide putative drug targets that may help to select drug combinations to inform future lab experiments.*

[3] Milioli, H.H.; Vimieiro, R.; Tishchenko, I.; Riveros, C.; Berretta, R.; Moscato, P. (2016). Iteratively refining breast cancer intrinsic subtypes in the METABRIC dataset. *BioData Mining*; 9:2.

[4] Milioli, H.H.; Riveros, C.; Vimieiro, R.; Tishchenko, I.; Berretta, R.; Moscato, P. Meta-features modelling gene expression imbalances: an innovative strategy for breast cancer subtype prediction. Manuscript submitted to *Genomics, Proteomics & Bioinformatics*.

[5] Milioli, H.H.*; Tishchenko, I.*; Riveros, C.; Berretta, R.; Moscato, P. Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. *BMC Med Genomics*; 10(1):19 *co-authorship.

Chapter 8 – Concluding Remarks

In the last chapter, I summarise the conclusions of all previous chapters based on the research questions defined in *Chapter 1*. Furthermore, I point out the future directions of the breast cancer investigations, in the context of the present study. Briefly, researchers should look into different subtypes, focusing on molecular cause and effect, for novel tailored intervention strategies and clinical applications.

1.5 References

AIHW. (2012). *Australian Institute of Health and Welfare. Cancer incidence projections: Australia, 2011 to 2020*. (Cat. no. CAN 62). Canberra: AIHW.

AIHW, & AACR. (2014a). *Australian Institute of Health and Welfare & Australasian Association of Cancer Registries. Cancer in Australia*. Canberra: AIHW.

AIHW, & AACR. (2014b). *Australian Institute of Health and Welfare & Australasian Association of Cancer Registries. Cancer in Australia: an overview.* Canberra: AIHW.

Allison, D. B., Cui, X., Page, G. P., & Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet, 7*(1), 55-65.

Altman, R. B., & Miller, K. S. (2011). 2010 translational bioinformatics year in review. *J. Am. Med. Inform. Assoc., 18*(4), 358-366.

Attwood, T., Gisel, A., Bongcam-Rudloff, E., & Eriksson, N. (2011). *Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective*: INTECH Open Access Publisher.

Ball, C. A., Brazma, A., Causton, H., Chervitz, S., Edgar, R., Hingamp, P., et al. (2004). Submission of microarray data to public repositories. *PLoS Biol., 2*(9), E317.

Barnard, M. E., Boeke, C. E., & Tamimi, R. M. (2015). Established breast cancer risk factors and risk of intrinsic tumor subtypes. *Biochim. Biophys. Acta, 1856*(1), 73-85.

Baxevanis, A. D. (2009). The importance of biological databases in biological discovery. *Curr Protoc Bioinformatics, Chapter 1*, Unit 1 1.

Borrebaeck, C. A. K. (2017). Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer. *Nat. Rev. Cancer, 17*(3), 199-204.

Carey, V. J., & Stodden, V. (2010). *Reproducible Research Concepts and Tools for Cancer Bioinformatics*. US: Springer.

Colombo, P., Milanezi, F., Weigelt, B., & Reis-Filho, J. S. (2011). Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction. *Breast Cancer Res., 13*(3), 212.

Dawson, S. J., Rueda, O. M., Aparicio, S., & Caldas, C. (2013). A new genome-driven integrated classification of breast cancer and its implications. *EMBO J., 32*(5), 617-628.

Dolled-Filhart, M., Ryden, L., Cregger, M., Jirstrom, K., Harigopal, M., Camp, R. L., et al. (2006). Classification of breast cancer using genetic algorithms and tissue microarrays. *Clin. Cancer Res., 12*(21), 6459-6468.

Duffy, M. J., Harbeck, N., Nap, M., Molina, R., Nicolini, A., Senkus, E., et al. (2017). Clinical use of biomarkers in breast cancer: Updated guidelines from the European Group on Tumor Markers (EGTM). *Eur. J. Cancer, 75*, 284-298.

Dutta, B., Pusztai, L., Qi, Y., Andre, F., Lazar, V., Bianchini, G., et al. (2012). A network-based, integrative study to identify core biological pathways that drive breast cancer clinical subtypes. *Br. J. Cancer, 106*(6), 1107-1116.

Engreitz, J. M., Chen, R., Morgan, A. A., Dudley, J. T., Mallelwar, R., & Butte, A. J. (2011). ProfileChaser: searching microarray repositories based on genome-wide patterns of differential expression. *Bioinformatics, 27*(23), 3317-3318.

Fan, C., Prat, A., Parker, J., Liu, Y., Carey, L. A., Troester, M. A., et al. (2011). Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Med. Genomics, 4*(1), 3.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., et al. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer, 136*(5), E359-386.

Forman, M. R., Winn, D. M., Collman, G. W., Rizzo, J., & Birnbaum, L. S. (2015). Environmental exposures, breast development and cancer risk: Through the looking glass of breast cancer prevention. *Reprod. Toxicol., 54*, 6-10.

Glas, A. M., Floore, A., Delahaye, L. J., Witteveen, A. T., Pover, R. C., Bakx, N., et al. (2006). Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics, 7*(1), 278.

Guenard, F., Labrie, Y., Ouellette, G., Beauparlant, C. J., Bessette, P., Chiquette, J., et al. (2007). Germline mutations in the breast cancer susceptibility gene PTEN are rare in high-risk non-BRCA1/2 French Canadian breast cancer families. *Fam. Cancer, 6*(4), 483-490.

Hayden, E. C. (2014). Cancer-gene data sharing boosted: efforts to get more breast-cancer gene variants into public databases are gaining ground. *Nature, 510*(7504), 198-199.

Hollestelle, A., Wasielewski, M., Martens, J. W., & Schutte, M. (2010). Discovering moderate-risk breast cancer susceptibility genes. *Curr. Opin. Genet. Dev., 20*(3), 268-276.

Hood, L. (2003). Systems biology: integrating technology, biology, and computation. *Mech. Ageing Dev., 124*(1), 9-16.

Huang, L., Zhao, S., Frasor, J. M., & Dai, Y. (2011). An integrated bioinformatics approach identifies elevated cyclin E2 expression and E2F activity as distinct features of tamoxifen resistant breast tumors. *PLoS One, 6*(7), e22274.

Krepischi, A. C., Achatz, M. I., Santos, E. M., Costa, S. S., Lisboa, B. C., Brentani, H., et al. (2012). Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer Res., 14*(1), R24.

Kulasingam, V., Pavlou, M. P., & Diamandis, E. P. (2010). Integrating high-throughput technologies in the quest for effective biomarkers for ovarian cancer. *Nat. Rev. Cancer, 10*(5), 371-378.

Lalloo, F., & Evans, D. G. (2012). Familial breast cancer. *Clin. Genet., 82*(2), 105-114.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921.

Lynch, E. D., Ostermeyer, E. A., Lee, M. K., Arena, F., Ji, H., Dann, J., et al. (1997). Inherited mutations in PTEN that are associated with breast cancer, cowden disease, and juvenile polyposis. *Am. J. Hum. Genet., 61*, 1254-1260.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., et al. (1994). A Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1. *Science, 266*(5182), 66-71.

Moreau, Y., Aerts, S., Moor, B. D., Strooper, B. D., & Dabrowski, M. (2003). Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet., 19*(10), 570-577.

Naderi, A., Teschendorff, A. E., Barbosa-Morais, N. L., Pinder, S. E., Green, A. R., Powe, D. G., et al. (2006). A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene, 26*(10), 1507-1516.

Nevanlinna, H., & Bartek, J. (2006). The CHEK2 gene and inherited breast cancer susceptibility. *Oncogene, 25*(43), 5912-5919.

Oldenburg, R. A., Meijers-Heijboer, H., Cornelisse, C. J., & Devilee, P. (2007). Genetic susceptibility for breast cancer: how many more genes to be found? *Crit. Rev. Oncol. Hematol., 63*(2), 125-149.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., et al. (2004). A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *N. Engl. J. Med., 351*(27), 2817-2826.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol., 27*(8), 1160-1167.

Rakha, E. A., El-Sayed, M. E., Reis-Filho, J. S., & Ellis, I. O. (2008). Expression profiling technology: its contribution to our understanding of breast cancer. *Histopathol, 52*(1), 67-81.

Seoane, J. A., Day, I. N. M., Gaunt, T. R., & Campbell, C. (2014). A pathway-based data integration framework for prediction of disease progression. *Bioinformatics, 30*(6), 838-845.

Shuen, A. Y., & Foulkes, W. D. (2011). Inherited mutations in breast cancer genes - risk and response. *J. Mammary Gland Biol. Neoplasia, 16*(1), 3-15.

Siegel, R., Ma, J., Zou, Z., & Jemal, A. (2014). Cancer statistics, 2014. *CA Cancer J. Clin., 64*(1), 9-29.

Sims, A. H. (2009). Bioinformatics and breast cancer: what can high-throughput genomic approaches actually tell us? *J. Clin. Pathol., 62*(10), 879-885.

Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annu. Rev. Med., 61*, 437-455.

Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature, 486*(7403), 400-404.

Stratton, M. R., & Rahman, N. (2008). The emerging landscape of breast cancer susceptibility. *Nat. Genet., 40*(1), 17-22.

Su, J., Huang, D., Yan, H., Liu, H., & Zhang, H. (2012). Advances in Bioinformatics Tools for High-Throughput Sequencing Data of DNA Methylation. *Hereditary Genet, 1*(107), 2161-1041.

TCGA. (2012). Comprehensive molecular portraits of human breast tumours. *Nature, 490*(7418), 61-70.

Topalian, S. L., Taube, J. M., Anders, R. A., & Pardoll, D. M. (2016). Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nat. Rev. Cancer, 16*(5), 275-287.

van't Veer, L. J., & Bernards, R. (2008). Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature, 452*(7187), 564-570.

Venet, D., Dumont, J. E., & Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol., 7*(10), e1002240.

Verhaak, R. G. W., & Mills, G. B. (2012). Regulation of mRNA expression in breast cancer - a cis-tematic trans-action. *Breast Cancer Res., 14*(5), 322.

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet, 365*(9460), 671-679.

Weigelt, B., Pusztai, L., Ashworth, A., & Reis-Filho, J. S. (2012). Challenges translating breast cancer gene signatures into the clinic. *Nat. Rev. Clin. Oncol., 9*(1), 58-64.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., et al. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature, 378*(6559), 789-792.

Xu, Y., Duanmu, H., Chang, Z., Zhang, S., Li, Z., Li, Z., et al. (2012). The application of gene co-expression network reconstruction based on CNVs and gene expression microarray data in breast cancer. *Mol. Biol. Rep., 39*(2), 1627-1637.

Yao, T. (2002). Bioinformatics for the genomic sciences and towards systems biology. Japanese activities in the post-genome era. *Prog. Biophys. Mol. Biol., 80*(1), 23-42.

Yigitoglu, B., Uctepe, E., Yigitoglu, R., Gunduz, E., & Gunduz, M. (2015). Bioinformatics in Breast Cancer Research.

Yulug, I. G., & Gur-Dedeoglu, B. (2008). Functional genomics in translational cancer research: focus on breast cancer. *Brief Funct Genomic Proteomic, 7*(1), 1-7.

CHAPTER 2

2. BREAST CANCER: CURRENT STATUS AND PERSPECTIVES

The purpose of *Chapter 2* is to contextualise breast cancer from origin to classification, as defined in **2.1 Breast Carcinogenesis** and **2.2 The Breast Tumour Classification**, respectively. The molecular characterisation of subtypes is introduced in **2.3 Intrinsic Subtypes** and further detailed in the subsections devoted to each of the five main groups: luminal A, luminal B, HER2-enriched, basal-like and normal-like. The most recent classification of breast tumours, containing a unique interpretation of the genome and transcriptome profiles, is described in **2.4 Novel Integrative Clusters**. The following section, **2.5 Predicting Molecular Subtypes**, discuss important concepts that are inherent to prediction models for differentiating breast cancer intrinsic subtypes. A review of current methods and applications is also included in the thesis to highlight the progress of translational research, from fundamental science to applied medicine. The literature presented to substantiate *Chapter 2* is listed in **2.6 References**.

2.1 Breast Carcinogenesis

Understanding the human breast gland development is a prerequisite for capturing the critical steps involved in breast tissue morphogenesis – the distinct cell types – and breast cancer origin (Bertos & Park, 2011). The breast gland is a specific type of apocrine gland that evolves as an appendage from the epidermis. The mammary gland development occurs through distinctive main stages throughout embryonic, pubertal and adult life. At each stage, different signals are required to induce changes in both the epithelium and the surrounding mesenchyme/stroma. Hormones and growth factors, for instance, play a central role in different stages of gland development and are also involved in breast carcinogenesis. The control of the gland morphogenesis, nonetheless, remains one of the most challenging issues in developmental biology (Watson & Khaled, 2008).

During embryogenesis, the gland development is dependent on heterotypic interactions that conduct the epithelial cells to proliferate and invade the underlying stroma, originating a branch of rudimentary epithelial ducts (Sternlicht et al., 2006). The residual structures enter in a quiescent allometric phase that lasts until puberty; a process essentially identical in human males and females (Russo & Russo, 2004). At puberty, the branching morphogenesis occurs in females by reason of systematic released hormones, oestrogen and progesterone. The development and differentiation of epithelial structures – in terminal ducts and lobules – occurs concomitant with the surrounding expansion of mesenchymal cells, including adipocytes, fibroblasts, blood vessels and immune cells. In adults, the breast gland is under continuous remodelling, through constant cell turnover, during each menstrual period, substantiated by cell proliferation, differentiation and apoptosis. The cycle remains until the ovary function declines at menopause stage (Lanigan et al., 2007). During pregnancy, in particular, an increased production of ovarian hormones results in further expansion of epithelial compartment that develops the functional lactating breast (Russo & Russo, 2004).

The mammary gland is composed by a variety of stem cells that are essential for the organ development and tissue homeostasis. These stem cells originate the mature epithelium – luminal and basal – of essentially luminal or myoepithelial lineage, or via a series of lineage-restricted intermediates. The luminal lineage can be further subdivided into ductal and alveolar cells that form the ducts and the alveolar units, respectively. Myoepithelial cells, in contrast, are specialised contractile cells located at the basal surface of the epithelium adjacent to the basement membrane (Visvader, 2009). The non-epithelial or stromal elements in the mammary

gland are: fibroblasts, endothelial cells, macrophages, and adipocytes; collectively the mammary fat pad, modulating tissue specificity of the normal breast as well as the growth, survival, polarity, and invasive behaviour of breast cancer cells (Polyak & Kalluri, 2010).

The existence of a continuum of stem cells active at different points in mammary development had been identified by the various histological, immunochemical and molecular approaches (Sims et al., 2007). Importantly, the self-renewal allows the hierarchy of progenitors and their inter-relationship to be determined in main stages of the gland transformation and breast carcinogenesis. In spite of the obvious disparity between the highly regulated process of development and the less organised environment of invasive cancer, many identical mechanisms and signalling pathways regulate both activities. The surrounding stromal is strikingly similar in normal epithelial populations and in invasive tumour cells. Accordingly, the environment interactions are important in the two conditions, and also implicate in the cellular aetiology underpinning breast cancer heterogeneity (Lanigan et al., 2007; Visvader, 2009).

The initiation of breast cancers is due to genetic and epigenetic transforming events that occur in a single cell (Polyak, 2007). Stem cells are slow-dividing, long-lived cells that by nature are exposed to damage and accumulate mutations over the years (Dontu et al., 2003). As a result, a cancer starts, normally, with mutations in a stem cell or in their lineage-restricted progeny (transit amplifying cells or committed differentiated cells) leading to a multistep evolution that contribute to cancer progress (Stingl & Caldas, 2007). The natural history of breast cancer involves the expansion of transformed cells through hyperproliferative stages, and subsequent in situ and invasive carcinomas, and finally to metastatic disease (Petersen & Polyak, 2010). The linear path of succession nevertheless oversimplifies the reality of breast cancer (Hanahan & Weinberg, 2000).

In principle, the development proceeds via a process formally analogous to Darwinian evolution, in which genetic changes confer one or another type of adaptive advantage. Intrinsic factors and/or random mutations act in the primary tumour, therefore, selecting multiple cells according to acquired abilities in a particular microenvironment (Hanahan & Weinberg, 2000). The current course of breast tumour progression, in fact, presents complex variables involving new clones and high heterogeneity. Genomes of tumours often become unstable and new transformations increase at significant rates in each generation. Furthermore, the cell plasticity continually changes at random, exceeding the ability of Darwin selection to eliminate clones genetically less suitable. Consequently, the breast tumour mass contains a large number of distinct sectors or populations with distinct subclones. The clonal cells may, yet, evolve from

primary directly to metastatic stages – depending on the mutation site –, without passing through intermediate conditions (Polyak, 2011).

In the context of a malignant cell transformation, Hanahan and Weinberg (2000) suggested a set of manifestations that collectively dictate tumour growth: sustaining proliferative signalling, evading growth-inhibitory signals, resisting programmed cell death (apoptosis), enabling replicative immortality, inducing angiogenesis and activating invasion and metastasis. Underlying these hallmarks are the genome instability, which generates the genetic diversity, and inflammation, which fosters multiple hallmark functions. In the last decade, two emerging hallmarks have been added to the list: reprogramming of energy metabolism and evading immune destruction (Hanahan & Weinberg, 2011). Most likely these potential mechanisms concomitantly occur in different cancers, and their relative contribution varies according to tumour type and progression stage (Polyak, 2007).

Finally, the breast cancer is a result of direct and indirect molecular perturbations that may provoke expression changes of greater amplitude in downstream genes or entire pathways. The dramatic changes in gene expression patterns of the tumour-associated luminal, myoepithelial and stromal cells determine the differences in mammary carcinogenesis. Notably, there exist enough evidences suggesting that deregulation in pathways and cancer initiation versus promotion are events clearly divergent, complex and heterogeneous, which difficult the disease biological understanding and the breast cancer classification (Petersen & Polyak, 2010; Polyak, 2011).

2.2 The Breast Tumour Classification

The breast cancer classification is defined by virtual categories according to different criteria and serving a different purpose (Eusebi, 2010). The major categories currently involves the assessment of histological aspects – incorporating morphology-based (Ellis et al., 1992) and histological grade (*Elston-Ellis modification* of *Scarff*-Bloom-Richardson grading system) –, and staging pathological parameters: tumour size (T), axillary lymph-node involvement (N), and the presence or absence of distant metastases (M) (Brierley et al., 2016). Additionally, immunohistochemical (IHC) markers such as the hormone receptors oestrogen (ER) and progesterone (PR), and the overexpression and/or amplification of the human epidermal growth

factor receptor 2 (*ERBB2*, also termed *HER2* or *HER2/neu*), provide a therapeutic predictive value (Harris et al., 2007). Classifying the features is crucial not only to select the most effective treatment for each tumour type, but also to delineate patient prognosis (Dawson et al., 2013).

According to the histology, the vast majority of breast carcinomas (50-80%) are derived from the epithelium lining of the ducts, designated as invasive ductal carcinomas (IDC) not otherwise specified (NOS), also named IDC of no special type (NST). The second most common type is the invasive lobular carcinoma (ILC) which comprises of 5%-15% of all cases. Furthermore, other variants of invasive breast carcinomas, recognised as "histological special types", accounted together up to 25% of all breast cancers (e.g. tubular, cribriforme, mucinous, papillary, apocrine, neuroendocrine, medullary, secretory, adenoid cystic carcinoma, acinic, metaplastic) (Weigelt; Geyer; et al., 2010).

Grade is an assessment of the degree of differentiation (i.e. tubule formation and nuclear pleomorphism) and proliferative activity (i.e. mitotic index) of a tumour, and reproduces its aggressiveness. The tumour grade was determined by the Scarf-Bloom-Richardson Grading system and modified by Elston and Ellis (1991). For instance, cells are distinguished as well differentiated (low grade), moderately differentiated (intermediate grade), and poorly differentiated (high grade), as they progressively lose the features seen in normal breast cells. The rarity of many of these neoplasms linked with the lack of standardised criteria for their diagnosis and the low inter-observer reproducibility, however, limit randomised studies to define optimal treatments (Nagao et al., 2012; Yerushalmi et al., 2009).

Furthermore, the classification of malignant tumours, including malignant lesions of the breast, using the TNM system evaluates priority classification by anatomic extent; where *T* refers to the extent of the primary tumour (**Table 2.1**), *N* the absence or presence of regional lymph-node metastasis in the armpits, neck, and inside the chest (**Table 2.2**), and *M* the absence or presence of metastasis spread to a more distant part of the body (e.g. brain, lung, liver, bone) (**Table 2.3**). Once the T, N, and M are determined, a stage of 0, I, II, III, or IV is assigned, with stage 0 being in situ, stage I being early stage invasive cancer, and stage IV being the most advanced (**Table 2.4**). Finally, the staging remains an important feature to evaluate prognosis and recurrence, besides treatment decision (Brierley et al., 2016; Edge & Carlson, 2011).

Table 2.1 Primary Tumour (T)

TX	**Primary tumour cannot be assessed.**
T0	**No evidence of primary tumour.**
Tis	**Carcinoma *in situ*.** DCIS – Ductal carcinoma *in situ*; LCIS – Lobular carcinoma *in situ* Paget disease of the nipple NOT associated with invasive carcinoma and/or carcinoma *in situ* (DCIS and/or LCIS) in the underlying breast parenchyma. Carcinomas in the breast parenchyma associated with Paget disease are categorised based on the size and characteristics of the parenchymal disease, although the presence of Paget disease should still be noted.
T1	**Tumour ≤20 mm in greatest dimension.** **T1mi** – Tumour ≤1 mm in greatest dimension. **T1a** – Tumour >1 mm but ≤5 mm in greatest dimension. **T1b** – Tumour >5 mm but ≤10 mm in greatest dimension. **T1c** – Tumour >10 mm but ≤20 mm in greatest dimension.
T2	**Tumour >20 mm but ≤50 mm in greatest dimension.**
T3	**Tumour >50 mm in greatest dimension.**
T4	**Tumour of any size with direct extension to the chest wall and/or to the skin (ulceration or skin nodules)[a]** **T4a** – Extension to the chest wall, not including only pectoralis muscle adherence/invasion. **T4b** – Ulceration and/or ipsilateral satellite nodules and/or oedema (including peau d'orange) of the skin, which do not meet the criteria for inflammatory carcinoma. **T4c** – Both T4a and T4b. **T4d** – Inflammatory carcinoma.

Note: [a] Invasion of the dermis alone does not qualify as T4.
Data obtained from AJCC: Breast. In: Edge SB, Byrd DR, Compton CC, et al., eds.: AJCC Cancer Staging Manual. 7th ed. New York, NY: Springer, 2010, pp 347-76.

Table 2.2 Regional Lymph Nodes (N)

NX	**Regional lymph nodes cannot be assessed (e.g., previously removed).**
N0	**No regional lymph node metastases.**
N1	**Metastases to movable ipsilateral level I, II axillary lymph node(s).**
N2	**Metastases in ipsilateral level I, II axillary lymph nodes, clinically fixed or matted. OR** **Metastases in clinically detected[b] ipsilateral internal mammary nodes in the *absence* of clinically evident axillary lymph node metastases.** **N2a** – Metastases in ipsilateral level I, II axillary lymph nodes fixed to one another (matted) or to other structures. **N2b** – Metastases only in clinically detected[b] ipsilateral internal mammary nodes and in the *absence* of clinically evident level I, II axillary lymph node metastases.
N3	**Metastases in ipsilateral infraclavicular (level III axillary) lymph node(s) with or without level I, II axillary lymph node involvement. OR** **Metastases in clinically detected[a] ipsilateral internal mammary lymph node(s) with clinically evident level I, II axillary lymph node metastases. OR** **Metastases in ipsilateral supraclavicular lymph node(s) with or without axillary or internal mammary lymph node involvement.** **N3a** – Metastases in ipsilateral infraclavicular lymph node(s). **N3b** – Metastases in ipsilateral internal mammary lymph node(s) and axillary lymph node(s). **N3c** – Metastases in ipsilateral supraclavicular lymph node(s).

Note: [a] Clinically detected is defined by imaging studies (excluding lymphoscintigraphy) or by clinical examination and having characteristics highly suspicious for malignancy or a presumed pathologic macrometastasis based on fine needle aspiration biopsy with cytologic examination.
Data obtained from AJCC: Breast. In: Edge SB, Byrd DR, Compton CC, et al., eds.: AJCC Cancer Staging Manual. 7th ed. New York, NY: Springer, 2010, pp 347-76.

Table 2.3 Distant Metastasis (M)

M0	**No clinical or radiographic evidence of distant metastases.**
cM0(i+)	**No clinical or radiographic evidence of distant metastases, but deposits of molecularly or microscopically detected tumour cells in circulating blood, bone marrow, or other non-regional nodal tissue that are ≤0.2 mm in a patient without symptoms or signs of metastases.**
M1	**Distant detectable metastases as determined by classic clinical and radiographic means and/or histologically proven >0.2 mm.**

Data obtained from AJCC: Breast. In: Edge SB, Byrd DR, Compton CC, et al., eds.: AJCC Cancer Staging Manual. 7th ed. New York, NY: Springer, 2010, pp 347-76.

Table 2.4 Anatomic stage/prognostic groups

Stage 0	**Tis**	**N0**	**M0**
Stage IA	T1*	N0	M0
Stage IB	T0	N1mi	M0
	T1*	N1mi	M0
Stage IIA	T0	N1**	M0
	T1*	N1**	M0
	T2	N0	M0
Stage IIB	T2	N1	M0
	T3	N0	M0
Stage IIIA	T0	N2	M0
	T1*	N2	M0
	T2	N2	M0
	T3	N1	M0
	T3	N2	M0
Stage IIIB	T4	N0	M0
	T4	N1	M0
	T4	N2	M0
Stage IIIC	Any T	N3	M0
Stage IV	Any T	Any N	M1

Note: *T1 includes T1mi. ** T0 and T1 tumours with nodal micrometastases only are excluded from Stage IIA and are classified Stage IB.
Data obtained from AJCC: Breast. In: Edge SB, Byrd DR, Compton CC, et al., eds.: AJCC Cancer Staging Manual. 7th ed. New York, NY: Springer, 2010, pp 347-76.

Combined with histopathological assessment and TNM classification, the standard evaluation of breast cancer also includes the IHC characterisation of ER, PR and *HER2* status. The hormone oestrogen and progesterone are important regulators of cell proliferation and differentiation and are crucial to guide endocrine-based therapies. Hormone receptor-positive breast cancers account for 75-80% of all cases; around 65% express both ER and PR, 10% are

ER-positive and PR-negative, 5% are ER-negative and PR-positive. Furthermore, *HER2* represents an additional predictive marker in routine use. Approximately 10-15% of breast cancers present *HER2* over-expression and/or amplification (Dawson et al., 2013). In this case, patients are candidates to receive target therapies with the humanised monoclonal antibody trastuzumab (Herceptin®), or other HER2-targeted therapy. Despite the value of ER and PR, their ability to direct the most appropriate systemic therapy remains defective; likewise, only part of the *HER2*-positive patients respond to treatment (Viale, 2012).

The first major breakthrough in applied investigation of breast cancers occurred with the innovative molecular methodologies of gene expression arrays (Portier et al., 2012). At the molecular level, histological features appear more homogenous, which allow further stratification of tumours according to intrinsic characteristics (Rakha et al., 2010). By using multidimensional variation and a hierarchical clustering analysis of gene expression profiling, Perou et al. (2000) and Sørlie et al (2001) provided early insights into the molecular heterogeneity of the disease. Five distinct subtypes were identified based on the gene expression information: luminal A, luminal B, HER2-enriched, basal-like and normal-like breast tissue. Accordingly, differences in gene expression – reflecting basic alterations in the tumour cell biology – were associated with significant variation in clinical (Hu et al., 2006; Sørlie et al., 2003). In 2011, at 12th St. Gallen International Breast Cancer Conference, for the first time experts have suggested the use of intrinsic biological subtypes for therapeutic decisions over early breast cancers, and for dealing with the diversity of tumours (Harbeck et al., 2013).

2.3 Intrinsic Subtypes

Molecular profiling has fundamentally changed our understanding of breast cancer since the initial landmark study by Perou et al (2000) and Sørlie et al. (2001), when a new taxonomy of breast cancers was proposed, based on the expression patterns of the so-called ‘intrinsic genes’. Intrinsic genes were defined as genes with a great variation in expression levels among different tumours (Strehl et al., 2011). In this context, the classification of the five intrinsic subtypes is mainly driven by the expression of oestrogen receptor (ER) and ER-related genes, separating ER-positive (luminal A and luminal B) from the ER-negative tumours (HER2-enriched and basal-like). Furthermore, this classification is extended to the high proliferation of HER2 and

related genes, mapping to the region of the HER2 amplicon on chromosome 17, amongst other clinical markers (Ki67) (Reis-Filho & Pusztai, 2011).

2.3.1 Luminal A and B

Cancers of luminal-type are characterised by the expression of genes similar to that observed in normal breast luminal epithelial cells. In addition, this type typically express luminal cytokeratins 8 (KRT8) and 18 (KRT18) (Strehl et al., 2011). At the molecular level, however, luminal A and B are considered as two different subtypes with independent intrinsic features and distinct clinical behaviour (Ciriello et al., 2013; Sørlie et al., 2003). The major differences between luminal A and luminal B are marked by the levels of proliferation (*MKI67* and *BIRC5*) and cell cycle-associated (*CCNB1* and *MYBL2*) genes, which are more pronounced in luminal B breast cancers (M. C. Cheang et al., 2009; Creighton, 2012; Wirapati et al., 2008). In clinical practice, the distinction of A and B subtypes is of high interest; luminal A breast carcinomas are typically at a lower risk for relapse, whereas luminal B generally carries a worse prognosis (Strehl et al., 2011) and higher recurrence scores (Fernández et al., 2015; Inic et al., 2014).

Luminal A composes about 40% of all breast cancers and exhibits the best prognosis of all breast cancer subtypes (Perou et al., 2000; Sørlie et al., 2001; Sotiriou & Pusztai, 2009). Usually have high expression of ER-related genes, low expression of the *HER2* cluster of genes, and low expression of proliferation genes (Hu et al., 2006). Studies performed with luminal A tumours confirmed markedly gene-expression deregulation of *LIV1*, *HNF3A* or *FOXA1*, *XBP1*, *GATA3* (Sørlie, 2004; Sørlie et al., 2001), *ESR1*, *TFF3* (Weigelt; Baehner; et al., 2010). Integrated analysis of gene expression microarrays and copy number variation (CNVs) and aberration (CNAs) have revealed further complexity and diversity within the breast cancer subtypes (Ciriello et al., 2013). The DNA copy number profile of Luminal A is correlated with low-grade tumours, frequently displaying 1q gain and 16q loss.

Luminal B, less common, accounts for 20% and have relatively lower expression of ER-related genes, variable expression of the *HER2*, and higher expression of proliferation genes. In comparison to luminal A, type B tumours show worse breast cancer prognosis and have high recurrence scores (Strehl et al., 2011). Luminal B was further distinguished from luminal A by the high expression of a particular set of genes such as *GGH, NSEP1*, *CCNE1* (Sorlie, 2004; Sotiriou & Pusztai, 2009), *SQLE* (Weigelt; Baehner; et al., 2010), *FGFR1* and *ZIC3* (Reis-Filho & Pusztai, 2011). Furthermore, the copy number analysis of Luminal B

tumours revealed a more complex genomic profile with amplifications in 8p11, 8q21, 11q13, 17q12 (HER2 locus) and 20q13, associated with a poor outcome (Bergamaschi et al., 2006; Chin et al., 2006; Cornen et al., 2014). Additionally, losses have been reported in Luminal B tumours in comparison to other breast cancer subtypes: 6q14, 9p21, 18p11 (Cornen et al., 2014). These changes may play a role in the tumour development and hormonal therapy resistance.

2.3.2 HER2-enriched

The HER2-enriched subtype, about 10% of all breast cancers, is defined by the overexpression of *HER2/ERBB2* and proliferation markers, and low expression of luminal genes. Hence, these tumours are typically negative for the hormone receptors ER and PR. Notably, HER2-enriched subtype is characterised by high expression of several genes in the *HER2* amplicon at 17q22.24 including *HER2 (ERBB2), GRB7, TRAP100* (Reis-Filho & Pusztai, 2011; Sorlie, 2004). However, not all breast cancers defined as *HER2*-positive (20 to 30%) are classified as HER2-enriched by the molecular profiling. HER2-enriched subtype comprises only about half of clinically *HER2*-positive breast cancer; the other part express both the HER2 and luminal gene clusters, and fall in a luminal subtype. The HER2-enriched subtype has been affected by a prognostic disadvantage, not responding to traditional chemotherapy and also presenting a significant resistance to treatment with *HER2* target therapy (Sotiriou & Pusztai, 2009; Strehl et al., 2011). Patients diagnosed within this subtype have also relatively high rates of metastasis to brain, liver, bone, and lung sites (Kennecke et al., 2010).

2.3.3 Basal-like

The basal-like subtype accounts for approximately 15% of invasive breast cancers. Notably, there is an association between the basal-like subtype and patient race and age. Population-based studies revealed the subtype prevalence in young (< 50) African-American women (Parker et al., 2009; Strehl et al., 2011). This group is characterised by the expression of cytokeratins (*KRT5/6, KRT14* and *KRT17*), typically found in the basal cells of normal mammary gland epithelium, and high expression of proliferation genes (Badve et al., 2011; M. C. U. Cheang et al., 2008). Tumours also exhibit altered expression of *ANXA8, CX3CL1*, *TRIM29* (Sorlie, 2004) *FABP7*, *LAMC2*, *ID4* (Weigelt; Baehner; et al., 2010), *FGFR2*, *CDKN2A* and *RB1* (Reis-Filho & Pusztai, 2011). Most of the basal-like samples are ER-, PR -, and *HER2*-negative; so called triple-negative. Although the terminologies are used

interchangeably, basal-like tumours are considered more homogeneous than triple-negatives breast cancers (Bertucci et al., 2012; Cleator et al., 2007).

Furthermore, *BRCA1* mutation-associated breast carcinomas strongly resemble basal-like tumours and might be regarded as a special subgroup within this intrinsic subtype. It has also been reported that triple negative and basal-like tumours have the highest frequency of copy number alterations, gains and losses, in comparison to other breast cancer subtypes (Engebraaten et al., 2013; Weigman et al., 2012). According to the copy number landscape, several features were observed in basal-like tumours including widespread genomic instability and common gains of 1q, 3q, 8q and 12p, and loss of 4q, 5q and 8p (TCGA, 2012).

Overall, patients diagnosed with basal-like tumours have a poor prognosis with worse outcome and decreased overall survival (Banerjee et al., 2006). Patients, however, have a relatively divergent disease outcome and varying overall survival (Carey et al., 2010; Rakha et al., 2008). Many individuals have shown high mortality and recurrence rates in the first 3-5 years, and others survivability of over 10 years following the diagnosis. In the last case, the prognosis is better than those of luminal breast cancer subtype (M. C. U. Cheang et al., 2008; Mulligan et al., 2008). Thus, the unpredictable behaviour and the refractory nature of these tumours have an impact on clinical assessments (Kreike et al., 2007; Rakha et al., 2008); with tumours less responsive to chemotherapy, and more sensitive to neoadjuvant chemotherapy (Banerjee et al., 2006).

Recent studies on triple-negative breast cancers (Burstein et al., 2015; Jézéquel et al., 2015; Lehmann et al., 2011) pointed to the existence of intrinsic basal-like subtypes. The classification proposed by Lehmann et al. (2011) revealed molecular groups of triple-negative tumours, including Basal-like 1 (BL1), Basal-like 2 (BL2) and Immunomodulatory (IM), overlapping basal-like samples. Alternatively, Burstein and colleagues (2015) defined the Basal-Like Immune-Suppressed (BLIS) and Basal-Like Immune-Activated (BLIA) subtypes. More recently, Jézéquel et al. (2015) pointed to two other groups: a basal-like with low immune response and high M2-like macrophages, and basal-enriched with high immune response and low M2-like macrophages. All studies, however, focused on investigating the molecular heterogeneity of triple-negative breast cancers; partially supporting each other. In fact, the classification of triple-negatives is not ideal for defining basal-like entities, and further analyses are required in the field.

2.3.4 Normal-like

Normal-like is typified by similar gene expression pattern as normal breast cells, and remains an enigmatic subtype. A normal breast tissue shows the highest expression of many genes, such as *PIK3R1* and *AKR1C1,* known to be characteristic of adipose tissues and other non-epithelial cell types (Calza et al., 2006). Overall, normal breast-like tumours are part of the ER-negative branch; but also part of the ER-positive in other studies. It is therefore unclear whether these tumours represent poorly sampled tumour tissue – contaminated with surrounding normal breast tissue – or a distinct important group. This subtype remains a significant issue to be solved in both research and clinical approach (Strehl et al., 2011).

2.3.5 Other groups

Following the initial identification of the intrinsic molecular subtypes, gene expression studies have evolved and further sub-classification of breast cancers into new molecular entities have been proposed. Herschkowitz et al. (2007) and Prat et al. (2010) have identified a new breast cancer intrinsic subtype known as Claudin-low. The newly described subtype comprises of non-basal triple-negative breast cancers, characterised by low to absent expression of genes involved in tight junctions and cell-cell adhesion (claudin 3, 4 and 7, E-cadherin), differentiated luminal cell surface markers (EpCAM and *MUC1*) and enrichment for epithelial-to-mesenchymal transition markers and immune response involving CD44 and CD24, *ALDH1A1*, *IL6*, *CXCL2*, *CDH1* (Herschkowitz et al., 2007; Prat & Perou, 2011; Reis-Filho & Pusztai, 2011). Noteworthy, claudin-low subset lack classical basal-like markers such as cytokeratins 5 and 6 (*KRT5* and *KRT6*) and epidermal growth factor receptor (*EGFR*) (Strehl et al., 2011). Clinically, the majority of Claudin-low tumours have a poor prognosis with a high frequency of metaplastic and medullary differentiation (Prat & Perou, 2011).

Molecular apocrine, another distinct molecular subgroup of breast tumours, is marked by ER-negative and AR-positive. *HER2* amplification is commoner in the molecular apocrine than the other groups. The apocrine tumours express *AR*, *HMGCR*, *GHR*, *PRLR* and *EGFR* (Farmer et al., 2005), *FAS*, *XBP1*, *ERBB2* (Weigelt; Baehner; et al., 2010). In the molecular classification setting, the variety on the gene expression information in the molecular apocrine subtype difficult its integration with previous microarray schemes for breast cancer subtyping (Farmer et al., 2005). In sum, the classifications herein mentioned have all some merits and several limitations. The taxonomy of breast cancers is, therefore, an open field to be explored with the purpose of understanding the mechanisms driving the disease course (Viale, 2012).

2.4 Novel Integrative Clusters

Through intricate analysis, METABRIC has proposed novel breast cancer subtypes with unique interpretations of the genome and transcriptome profile. The molecular profiling was defined across tumors of nearly two thousand women, for whom histopathological and clinical information had been meticulously recorded. This cohort performed an integrative clustering framework (iCluster), described by Shen et al. (2009), to identify not only gene expression patterns, but distinct loci that contribute to the disease phenotype. In this respect, cis-acting genes that exhibited significant associations with CNAs across the entire cohort of tumours influenced variation among groups. Finally, the new patterns and 'clusters' (IntClust 1 to IntClust 10) in the data led to the early conclusion that 'breast cancer' is in fact at least ten different diseases; each containing its own molecular fingerprint (Curtis et al., 2012). Despite this new interpretation, breast cancer disease remains poorly understood, molecularly inconsistent classified and beyond pronounced improvements the clinical practice.

Tumours within each cluster were compared across several attributes, such as clinicopathological features and survival outcomes. Tumours were yet stratified according to grade, tumour size, number of lymph nodes, age at diagnosis, and integrative cluster membership. As a result, molecular characteristics were observed for each of the 10 clusters in both data sets discovery and validation, demonstrating subtype reproducibility (Curtis et al., 2012). These findings have potential implications for the individualisation of treatment approaches, providing insights for a personalised breast cancer management (Dawson et al., 2013). The complete classification, nonetheless, should integrate information beyond the genomic landscape and transcriptomic approach. In this context, information of abnormalities in DNA methylation, microRNA expression and proteins offer other opportunities to further characterise the molecular architecture of breast cancer (TCGA, 2012). The cellular and molecular heterogeneity of breast tumours and the large number of plays potentially involved in controlling cell growth, death, and differentiation emphasise the importance of studying multiple alterations in concert (Sørlie et al., 2001).

2.5 Predicting Molecular Subtypes

Microarray technologies and gene expression profiling have been widely explored in breast cancer research. In the direction of developing useful tools to delineate the disease behaviour, a number of prognostic signatures have been proposed, such as Mammaprint® (Agendia, Huntington Beach, CA USA), Oncotype DX® (Genome Health Inc, Redwood City, CA USA), Veridex 76-gene LLC (Johnson & Johnson Company, San Diego, CA USA), MapQuant Dx (IPSOGEN SA, Marseilles, FR and New Haven, CT USA) and Breast Cancer Index (bioTheranostics Inc, San Diego, CA USA). Despite differences in the genes that integrate each of the signatures, the coverage of proliferation-related genes has led to the identification of similar groups of patients having poor prognosis. Due to the genes association to proliferative state, most of the signatures show great prognostic value for ER-positive patients; however, have limited capacity to delineate ER-negative tumours. Overall signatures have shown virtual prognostic information with potential to uncover other tumour features that are beyond what is offered by semi-quantitative assessment of ER, PR, HER2, and Ki67.

New concepts involving prediction models incorporate intrinsic molecular features shared among tumours with analogous behaviour. These features have been used to classify breast cancers into the five main subtypes: luminal A, luminal B, HER2-enriched, normal-like and basal-like (Herschkowitz et al., 2007; Hu et al., 2006; Perou et al., 2000; Prat et al., 2010; Sørlie et al., 2001; Sørlie et al., 2003). Parker et al. (2009) proposed a Single Sample Predictor (SSP) method to classify tumour subtypes according to the correlation with Nearest Shrunken Centroids (NSC) (Tibshirani et al., 2002). The so-called PAM50 method uses a 50 gene set as centroids. These genes are mainly involved in cell proliferation and are highly correlated with breast cancer subtypes. In the same direction, Haibe-Kains et al. (2012) attempted to simplify the subtypes prediction by using a Subtype Classification Model (SCM) based on three key genes: estrogen receptor 1 (*ESR1*), erb-b2 receptor tyrosine kinase 2 (*ERBB2*), and aurora kinase A (*AURKA*). Overall, the main goal of the disease subtyping is to define sets of patients at risk more likely to respond to selective drugs in a group-based tailored therapy.

Predictor models in breast cancer research have brought new insights to translational science and applied medicine, and are of unquestionable value to clinical practice. There is an agreement in the outcome predictions for ER-positive patients, even though different – in size and shape – gene sets are used for breast cancer prognostication (Fan et al., 2006). The independent sets result from the diversity in high-dimensional data sets and feature selection

approaches that frequently lead to the identification of distinct features (Ein-Dor et al., 2005). In addition, limiting the number of features may impact the classifiers performance, whereas selecting a high number may result in overfitting decision rules. Other important issues may arise with random sample collection, gene expression analysis and microarray technology. Hence, a range of different gene lists are selected (Popovici et al., 2010). The weaknesses of these methods lie in the analysis of multiple data sources and serial approaches.

Subtyping prediction methods, on the other hand, showed only a *moderate agreement* between sample labelling across distinct studies (Weigelt; Mackay; et al., 2010), intrinsic errors (Ebbert et al., 2011), wide confidence intervals (Michiels et al., 2005) and independent predictive value (Prat et al., 2012). Lusa et al. (2007) yet reported the limitations of using SSPs models across data sets due to the fact that the centroids values should be in the same scale of external cohorts. It is noteworthy to mention that sample size, in this case, causes a profound impact on the labels assignment, induced by sample stratification. On the other hand, the molecular subtypes of breast cancer are not completely understood and need additional research (Wirapati et al., 2008). For instance, it has been observed that proliferation in luminal samples forms a continuum and any division into luminal A and B is somewhat arbitrary (Pfeffer, 2013). Basal-like samples were also further divided into subsets of divergent molecular characterisation and clinical outcome (Bertucci et al., 2012; Lehmann et al., 2011).

Stringent standardisation of data sets and methodologies are therefore required to improve breast cancer classification and subtype prediction. Alternatively, the development of novel bioinformatics approaches – independent of data composition – will contribute to the analysis of independent data sets and combined technologies (Paquet & Hallett, 2015). Innovative strategies are therefore mandatory towards the interpretation of more robust and complex data sets prior the translating fundamental medical research into clinical applications (Michiels et al., 2011).

2.6 References

Badve, S., Dabbs, D. J., Schnitt, S. J., Baehner, F. L., Decker, T., Eusebi, V., et al. (2011). Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. *Mod. Pathol., 24*(2), 157-167.

Banerjee, S., Reis-Filho, J. S., Ashley, S., Steele, D., Ashworth, A., Lakhani, S. R., et al. (2006). Basal-like breast carcinomas: clinical outcome and response to chemotherapy. *J. Clin. Pathol., 59*(7), 729-735.

Bergamaschi, A., Kim, Y. H., Wang, P., Sørlie, T., Hernandez-Boussard, T., Lonning, P. E., et al. (2006). Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes, Chromosomes Cancer, 45*(11), 1033-1040.

Bertos, N. R., & Park, M. (2011). Breast cancer - one term, many entities? *J. Clin. Invest., 121*(10), 3789-3796.

Bertucci, F., Finetti, P., & Birnbaum, D. (2012). Basal breast cancer: a complex and deadly molecular subtype. *Curr. Mol. Med., 12*(1), 96.

Brierley, J. D., Gospodarowicz, M. K., & Wittekind, C. (2016). *TNM classification of malignant tumours*: John Wiley & Sons.

Burstein, M. D., Tsimelzon, A., Poage, G. M., Covington, K. R., Contreras, A., Fuqua, S. A., et al. (2015). Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin. Cancer Res., 21*(7), 1688-1698.

Calza, S., Hall, P., Auer, G., Bjohle, J., Klaar, S., Kronenwett, U., et al. (2006). Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. *Breast Cancer Res., 8*(4), R34.

Carey, L. A., Winer, E. P., Viale, G., Cameron, D., & Gianni, L. (2010). Triple-negative breast cancer: disease entity or title of convenience? *Nat. Rev. Clin. Oncol., 7*(12), 683-692.

Cheang, M. C., Chia, S. K., Voduc, D., Gao, D., Leung, S., Snider, J., et al. (2009). Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J. Natl. Cancer Inst., 101*(10), 736-750.

Cheang, M. C. U., Voduc, D., Bajdik, C., Leung, S., McKinney, S., Chia, S. K., et al. (2008). Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clin. Cancer Res., 14*(5), 1368-1376.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell, 10*(6), 529-541.

Ciriello, G., Sinha, R., Hoadley, K. A., Jacobsen, A. S., Reva, B., Perou, C. M., et al. (2013). The molecular diversity of Luminal A breast tumors. *Breast Cancer Res. Treat., 141*(3), 409-420.

Cleator, S., Heller, W., & Coombes, R. C. (2007). Triple-negative breast cancer: therapeutic options. *Lancet Oncol., 8*(3), 235-244.

Cornen, S., Guille, A., Adélaïde, J., Addou-Klouche, L., Finetti, P., Saade, M.-R., et al. (2014). Candidate luminal B breast cancer genes identified by genome, gene expression and DNA methylation profiling. *PLoS One, 9*(1), e81843.

Creighton, C. J. (2012). The molecular profile of luminal B breast cancer. *Biologics, 6*, 289-297.

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature, 486*(7403), 346-352.

Dawson, S. J., Rueda, O. M., Aparicio, S., & Caldas, C. (2013). A new genome-driven integrated classification of breast cancer and its implications. *EMBO J., 32*(5), 617-628.

Dontu, G., Al-Hajj, M., Abdallah, W. M., Clarke, M. F., & Wicha, M. S. (2003). Stem cells in normal breast development and breast cancer. *Cell Prolif., 36*(s1), 59-72.

Ebbert, M., Bastien, R. R., Boucher, K. M., Martin, M., Carrasco, E., Caballero, R., et al. (2011). Characterization of uncertainty in the classification of multivariate assays: application to PAM50 centroid-based genomic predictors for breast cancer treatment plans. *J Clin Bioinform, 1*(1), 37.

Edge, S. B., & Carlson, R. W. (2011). Breast Cancer Staging: Predicting Outcome and Response to Treatment. *Breast Surgical Techniques and Interdisciplinary Management*, 269-285.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., & Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics, 21*(2), 171-178.

Ellis, I. O., Galea, M., Broughton, N., Locker, A., Blamey, R. W., & et al. (1992). Pathological prognostic factors in breast cancer: II. Histological type: Relationship with survival in a large study with long term follow up. *Histopathol, 20*(6), 479-489.

Elston, C. W., & Ellis, O. (1991). Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathol, 19*(15), 403-410.

Engebraaten, O., Vollan, H. K. M., & Børresen-Dale, A.-L. (2013). Triple-negative breast cancer and the need for new therapeutic targets. *Am. J. Pathol., 183*(4), 1064-1074.

Eusebi, V. (2010). Classifications and prognosis of breast cancer: from morphology to molecular taxonomy. *Breast J, 16*(1), S15-16.

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., et al. (2006). Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med., 335*(6), 560-569.

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., et al. (2005). Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene, 24*(29), 4660-4671.

Fernández, A. G., Chabrera, C., Garcia Font, M., Fraile, M., Lain, J. M., Gonzalez, S., et al. (2015). Differential patterns of recurrence and specific survival between luminal A and luminal B breast cancer according to recent changes in the 2013 St Gallen immunohistochemical classification. *Clin. Transl. Oncol., 17*(3), 238-246.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., et al. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *J. Natl. Cancer Inst., 104*(4), 311-325.

Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell Press, 100*, 57-70.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell, 144*(5), 646-674.

Harbeck, N., Thomssen, C., & Gnant, M. (2013). St. Gallen 2013: Brief Preliminary Summary of the Consensus Discussion. *Breast Care, 8*(2), 102-109.

Harris, L., Fritsche, H., Mennel, R., Norton, L., Ravdin, P., Taube, S., et al. (2007). American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *J. Clin. Oncol., 25*(33), 5287-5312.

Herschkowitz, J. I., Simin, K., Weigman, V. J., Mikaelian, I., Usary, J., Hu, Z., et al. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol., 8*(5), R76.

Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics, 7*(1), 96.

Inic, Z., Zegarac, M., Inic, M., Markovic, I., Kozomara, Z., Djurisic, I., et al. (2014). Difference between Luminal A and Luminal B Subtypes According to Ki-67, Tumor Size, and Progesterone Receptor Negativity Providing Prognostic Information. *Clin. Med. Insights Oncol., 8*, 107-111.

Jézéquel, P., Loussouarn, D., Guérin-Charbonnel, C., Campion, L., Vanier, A., Gouraud, W., et al. (2015). Gene-expression molecular subtyping of triple-negative breast cancer tumours: importance of immune response. *Breast Cancer Res., 17*(1), 43.

Kennecke, H., Yerushalmi, R., Woods, R., Cheang, M. C., Voduc, D., Speers, C. H., et al. (2010). Metastatic behavior of breast cancer subtypes. *J. Clin. Oncol., 28*(20), 3271-3277.

Kreike, B., van Kouwenhove, M., Horlings, H., Weigelt, B., Peterse, H., Bartelink, H., et al. (2007). Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas. *Breast Cancer Res., 9*(5), R65.

Lanigan, F., O'Connor, D., Martin, F., & Gallagher, W. M. (2007). Molecular links between mammary gland development and breast cancer. *Cell. Mol. Life Sci., 64*(24), 3161-3184.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., et al. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest., 121*(7), 2750-2767.

Michiels, S., Koscielny, S., & Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet, 365*(9458), 488-492.

Michiels, S., Kramar, A., & Koscielny, S. (2011). Multidimensionality of microarrays: statistical challenges and (im)possible solutions. *Mol. Oncol., 5*(2), 190-196.

Mulligan, A. M., Pinnaduwage, D., Bull, S. B., O'Malley, F. P., & Andrulis, I. L. (2008). Prognostic effect of basal-like breast cancers is time dependent: evidence from tissue microarray studies on a lymph node-negative cohort. *Clin. Cancer Res., 14*(13), 4168-4174.

Nagao, T., Kinoshita, T., Hojo, T., Tsuda, H., Tamura, K., & Fujiwara, Y. (2012). The differences in the histological types of breast cancer and the response to neoadjuvant chemotherapy: the relationship between the outcome and the clinicopathological characteristics. *Breast, 21*(3), 289-295.

Paquet, E. R., & Hallett, M. T. (2015). Absolute assignment of breast cancer intrinsic molecular subtype. *J. Natl. Cancer Inst., 107*(1), dju357.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol., 27*(8), 1160-1167.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature, 406*(6797), 747-752.

Petersen, O. W., & Polyak, K. (2010). Stem cells in the human breast. *Cold Spring Harb. Perspect. Biol., 2*(5), a003160.

Pfeffer, U. (2013). *Cancer genomics: molecular classification, prognosis and response prediction*: Springer Science & Business Media.

Polyak, K. (2007). Breast cancer: origins and evolution. *J. Clin. Invest., 117*(11), 3155-3163.

Polyak, K. (2011). Heterogeneity in breast cancer. *J. Clin. Invest., 121*(10), 3786-3788.

Polyak, K., & Kalluri, R. (2010). The role of the microenvironment in mammary gland development and cancer. *Cold Spring Harb. Perspect. Biol., 2*(11), a003244.

Popovici, V., Chen, W. Y., Gallas, B., Hatzis, C., Shi, W., Samuelson, F., et al. (2010). Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res., 12*(1), R5.

Portier, B. P., Gruver, A. M., Huba, M. A., Minca, E. C., Cheah, A. L., Wang, Z., et al. (2012). From morphologic to molecular: established and emerging molecular diagnostics for breast carcinoma. *N Biotechnol, 29*(6), 665-681.

Prat, A., Parker, J. S., Fan, C., & Perou, C. M. (2012). PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer. *Breast Cancer Res. Treat., 135*(1), 301-306.

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., et al. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res., 12*(5), R68.

Prat, A., & Perou, C. M. (2011). Deconstructing the molecular portraits of breast cancer. *Mol. Oncol., 5*(1), 5-23.

Rakha, E. A., Reis-Filho, J. S., Baehner, F., Dabbs, D. J., Decker, T., Eusebi, V., et al. (2010). Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Res., 12*(207), 1-12.

Rakha, E. A., Reis-Filho, J. S., & Ellis, I. O. (2008). Impact of basal-like breast carcinoma determination for a more specific therapy. *Pathobiology, 75*(2), 95-103.

Reis-Filho, J. S., & Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet, 378*(9805), 1812-1823.

Russo, J., & Russo, I. H. (2004). Development of the human breast. *Maturitas, 49*(1), 2-15.

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics, 25*(22), 2906-2912.

Sims, A. H., Howell, A., Howell, S. J., & Clarke, R. B. (2007). Origins of breast cancer subtypes and therapeutic implications. *Nat. Clin. Pract. Oncol., 4*(9), 516-525.

Sorlie, T. (2004). Molecular portraits of breast cancer: tumour subtypes as distinct disease entities. *European journal of cancer, 40*(18), 2667-2675.

Sørlie, T. (2004). Molecular portraits of breast cancer: tumour subtypes as distinct disease entities. *Eur. J. Cancer, 40*(18), 2667-2675.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A., 98*(19), 10869-10874.

Sørlie, T., Tibshirani, R., Parker, J. S., Hastie, T., Marron, J. S., Nobel, A., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U. S. A., 100*(14), 8418-8423.

Sotiriou, C., & Pusztai, L. (2009). Gene-Expression Signatures in Breast Cancer. *N. Engl. J. Med., 360*(8), 790-800.

Sternlicht, M. D., Kouros-Mehr, H., Lu, P., & Werb, Z. (2006). Hormonal and local control of mammary branching morphogenesis. *Differentiation, 74*(7), 365-381.

Stingl, J., & Caldas, C. (2007). Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. *Nature, 7*(10), 791-799.

Strehl, J. D., Wachter, D. L., Fasching, P. A., Beckmann, M. W., & Hartmann, A. (2011). Invasive Breast Cancer: Recognition of Molecular Subtypes. *Breast Care, 6*(4), 258-264.

TCGA. (2012). Comprehensive molecular portraits of human breast tumours. *Nature, 490*(7418), 61-70.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U. S. A., 99*(10), 6567-6572.

Viale, G. (2012). The current state of breast cancer classification. *Ann. Oncol., 23*(10), x207-210.

Visvader, J. E. (2009). Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis. *Genes Dev., 23*(22), 2563-2577.

Watson, C. J., & Khaled, W. T. (2008). Mammary development in the embryo and adult: a journey of morphogenesis and commitment. *Development, 135*(6), 995-1003.

Weigelt, B., Baehner, F. L., & Reis-Filho, J. S. (2010). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *J. Pathol., 220*(2), 263-280.

Weigelt, B., Geyer, F. C., & Reis-Filho, J. S. (2010). Histological types of breast cancer: how special are they? *Mol. Oncol., 4*(3), 192-208.

Weigelt, B., Mackay, A., A'hern, R., Natrajan, R., Tan, D. S. P., Dowsett, M., et al. (2010). Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol., 11*(4), 339-349.

Weigman, V. J., Chao, H.-H., Shabalin, A. A., He, X., Parker, J. S., Nordgard, S. H., et al. (2012). Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. *Breast Cancer Res. Treat., 133*(3), 865-880.

Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., et al. (2008). Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res., 10*(4), R65.

Yerushalmi, R., Hayes, M. M., & Gelmon, K. A. (2009). Breast carcinoma - rare types: review of the literature. *Ann. Oncol., 20*(11), 1763-1770.

CHAPTER 3

3. MICROARRAY TECHNOLOGIES AND 'OMICS' DATA SETS

Chapter 3 contains an overview of gene expression microarrays which includes definitions and implications for breast cancer research. This topic is introduced in **3.1 Microarray technologies**. In conducting research, the selection of state-of-the-art data sets from the public domain determines the quality of the analysis provided and, consequently, the later achievements. Section **3.2 The METABRIC Breast Cancer Data Set** shows one of the most comprehensive data sets available in the field, containing over 2000 samples. The molecular profile and clinicopathological information that comes along with this data is also crucial in supporting the application of distinct bioinformatics approaches. For external validation across platforms (Illumina and Affymetrix), a second data set is assessed: ROCK. The most important details for this data set are discussed in **3.3 ROCK: Integrative Breast Cancer Data**. It is noteworthy that the high quality of both data sets, compared to other available data sets, exposes the urgent need for developing and applying novel strategies to uncover breast cancer subtypes.

3.1 Microarray technologies

Bioinformatics have developed following the technological improvements of large-scale microarray data (Yigitoglu et al., 2015). Despite the fact that microarray is relatively novel – the "gene chip" industry started to grow in the 1990's with exponential improvements in the methodology – there are thousands of publications relying on this approach. Microarrays have showed a profound impact on a range of studies and have significantly accelerated the rate of scientific discoveries (Ball et al., 2004). It is a valuable technology in the sense that arrays allow the investigation of thousands of small molecules simultaneously, at one time. In addition, the analysis of microarrays permits comparisons of expression levels between different cells or tissues, such as diseased and normal profiles or treated versus non-treated samples (Villasenor-Park & Ortega-Loayza, 2013).

The development of arrays was possible due to innovations in micro-engineering, molecular biology and bioinformatics (Heller, 2002). With this integration, the technology of DNA microarrays (cDNA microarrays, oligonucleotide microarrays and SNP microarrays) has become the most sophisticated and the most widely used in the last decades. In the same line, other types of microarrays have emerged, including the RNA, miRNA, Protein, Peptide, Tissue, Cellular, Antibody, Carbohydrate, etc. Besides the variety of target molecules, this standard methodology offers a comprehensive qualitative and quantitative assessment of probe arrays into microchips, created by robotic machines (van Bakel & Holstege, 2004).

DNA or oligonucleotides microarrays (Figure 3.1) are an orderly arrangement of nucleotide sequences attached to a solid surface – usually made of glass or silicon – by a covalent bond to a chemical matrix via epoxy-silane, amino-silane, lysine and polyacrylamide. The fixed elements are used as probes – selected from GenBank, dbEST, and RefSeq – to hybridise a cDNA or cRNA from test samples, prepared under appropriate conditions (Nguyen et al., 2002). Firstly, mRNA molecules are selected from a target of investigation and a reference sample. Then, complementary molecules of the mRNA are produced and labelled with either a fluorescent dye (“fluorophore”) or a radioactive isotope. Binding fluorophores, commonly Cy3 (which fluoresces green) and Cy5 (which fluoresces red), for example, facilitate direct parallel comparison between different cell or tissue types. Finally, the labelled molecules are hybridised to the probes, during which process the targets competitively bind to the corresponding array probe (Villasenor-Park & Ortega-Loayza, 2013).

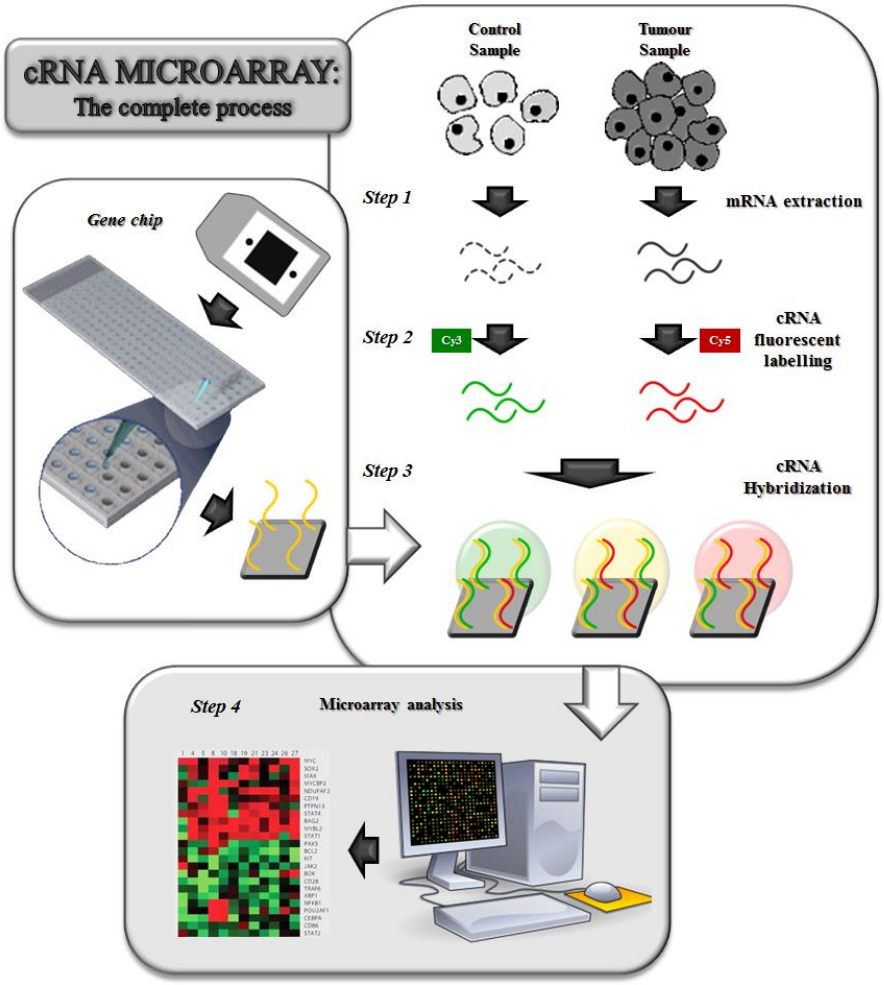


Figure 3.1 Conceptual view of a cRNA microarray processing.

In the gene chip, individual probes are immobilised on the array surface and spotted along the probe cells. Each probe cell contains millions of copies of the same oligonucleotide, or probe. For the hybridisation procedure, the methodology is divided on four main steps. *Step1* defines the mRNA extraction from control (non-tumour) and tumour tissue samples that are subsequently copied into fluorescent labelled cRNA fragments in *Step 2*. *Step 3* is marked by the cRNA hybridisation over the array surface, in a competitive probe-target interaction. The array then undergoes a series of washing and staining phases. In *Step 4*, each probe cell is scanned by a laser to quantify the levels of hybridisation obtained with the intensity measurement at the probe location. The probe intensity is adjusted to overcome possible defects in the array and the data is normalised and processed using computer resources.

Hybridisation is identified based on fluorescence detection of fluorophore-labelled to determine the relative abundance of nucleic acid sequences in the sample. A special scanner detected and recorded the fluorescent intensity for each spot/areas on the microarray slide. The level of expression is subsequently measured in a semi-quantitative manner by comparing the level of messenger RNA (mRNA) of thousands equivalent genes amongst the target and

reference sample (Strehl et al., 2011). If a particular gene is very active (overexpressed), it produces many molecules of mRNA, thus, more labelled complementary molecules, which hybridise to the probe on the microarray slide and generate a very bright fluorescent area. Genes that are somewhat "less active" (under-expressed) produce fewer mRNAs, which results in dimmer fluorescent spots. The average of signal from Cy3 and Cy5 are result of gene expression competition between distinct samples, which means that specific gene is more/less expressed in one sample than in other when the colour is green or red, and equally expressed when the signal emitted is yellow. Ultimately, if there is no fluorescence, none of the messenger molecules have hybridised, indicating that the gene is inactive in that sample (Villasenor-Park & Ortega-Loayza, 2013).

Although microarrays are capable of generating a large amount of significant data, the data-intensive nature of microarray technology has created an unprecedented informatics and analytical challenge. Inappropriate selection of "normal" or control specimens and experimental samples may not yield relevant results. The type of data created also depends on several other variables such as diverse methods of generating labelled material, experimental design, data standardisation, image acquisition and analysis, normalisation, statistical significance inference, biological exploratory data analysis, class prediction and validation (Allison et al., 2006; Leung & Cavalieri, 2003).

Determining consistency is complicated if all aspects are to be assessed in a non-arbitrary way across the different platforms and their variants. In addition, reliability is a sensitive issue for each group that provide the technology: Illumina, Affymetrix, Agilent Technologies, NimbleGen Systems (van Bakel & Holstege, 2004). Verification of data generated from microarray experiments using quantitative RT-PCR (reverse-transcriptase Polymerase Chain Reaction), northern blot analysis, or RNase protection assays may also be important (Villasenor-Park & Ortega-Loayza, 2013). Apart from quality microarray issues, data interpretation is currently the main bottleneck in microarray analyses.

In particular, the automated integration of complementary information in analysis algorithms is not yet well established. The main reasons for that are the lack of a common nomenclature and the difficulty of querying data in a single format. The Gene Ontology Consortium and similar initiatives have taken on the hard task of providing a cohesive connected framework. Although not intrinsic to microarray technology, these efforts are fundamental for the success of the technology (Hoheisel, 2006). Moreover, open-source initiatives such as Bioconductor (http://www.bioconductor.org), which are written in the R statistical programming language (http://www.r-project.org), provide a means for developing,

testing and disseminating new algorithms. To highlight, comprehensive expert systems that carry out data interpretation automatically are currently under development, likely to be available in the future (Reimers, 2010).

Recently, a PAM50 assay based on the expression of 50 genes has been suggested to classify breast cancer samples into each of the five intrinsic subtypes (Nielsen et al., 2010; Parker et al., 2009). On the other hand, microarray data has also been used by several groups to identify distinct prognostic signatures of breast cancer (Paik et al., 2004; van't Veer et al., 2005; Van De Vijver et al., 2002). Two main signatures, for instance, have been approved for clinical use and are now being tested in randomised clinical trials (Cardoso et al., 2008; Sparano & Paik, 2008). MammaPrint® (a microarray-based assay of the Amsterdam 70-gene breast cancer signature) categorises tumours as either high or low risk, in patient lymph node negative, whereas Onco*type* DX™ (an assay of a panel of 21 genes designed for use in ER positive tumours) reports a Recurrence Score (RS), where a higher RS is associated with a worse prognosis. Other tests have also been performed such as Mammostrat®, BreastOncPx™, MapQuant Dx™ and eXagenBC (Reis-Filho & Pusztai, 2011; Ross et al., 2008).

3.1.1 Illumina Approach

The Illumina, Inc. (San Diego, CA, USA) company developed and markets integrated array-based systems and assays for a broad range of applications including genotyping, gene expression and epigenetics. Illumina offers basically two microarray platforms: the Sentrix® Array Matrix (SAM) and the Sentrix® BeadChip; both containing hundreds of thousands of copies of covalently attached oligonucleotide probes (Steemers & Gunderson, 2005). The custom microarray Illumina SAM is configured with 96 fiber-optic bundles, each comprising an individual array. This unique format allows fast, simple and simultaneous analysis of samples on 96 arrays. For users with more moderate throughput demands, Illumina has introduced the BeadChip format. In general, the properties are similar to SAM, however BeadChip are used to process up to 16 samples per chip (http://www.illumina.com).

The company's proprietary BeadArray™ technology – now used in leading genomics centres around the world – provides the throughput, cost effectiveness and flexibility necessary to enable researchers in the life sciences and pharmaceutical industries to perform the billions of tests. This technology is based on 3-micron silica beads randomly arranged, with each bead binding many identical copies of a gene-specific probe. The BeadArray™ is constructed so that

there are roughly 30 randomly positioned replicates in microwells on either of two substrates: fiber optic bundles or planar silica slides. When randomly assembled on one of these two substrates, the beads have a uniform spacing of ~5.7 microns. Each bead is covered with hundreds of thousands of copies of a specific oligonucleotide that act as the capture sequences in one of Illumina's assays. BeadArray™ technology is utilised in Illumina's iScan System for a broad range of DNA and RNA analysis applications (Steemers & Gunderson, 2005). Importantly, the design yields higher confidence calls and more robust estimations of microarrays (Du et al., 2008).

Illumina has developed a spectrum of proprietary assays for application on microarray platform. The BeadChip format is currently used in Illumina's Infinium™ Genotyping, DASL™ Gene Expression, and Focused Arrays applications (GoldenGate®). These assays have been successfully employed, including collection of the majority of the Phase I genotyping data for the International HapMap Project. Illumina is focusing on extending the applications of the above assays and developing new assays for future products. Illumina's goal is to deliver high-performance, high-throughput solutions that enable researchers to expand experimental scale while reducing the cost of large-scale research (Steemers & Gunderson, 2005).

3.1.2 Affymetrix Platforms

Affymetrix (Santa Clara, CA, USA) is a company that manufactures high quality DNA microarrays. The Affymetrix GeneChip® System uses arrays fabricated by direct synthesis of thousands of oligonucleotides probes on the glass surface using the photolithographic technology (a process of using light to control the manufacture of multiple layers of material). This direct synthesis approach improves the accuracy of the arrays and avoids the need for probe sequence verification. In addition, the probe sets are given different suffixes to describe their uniqueness or their ability to bind different genes or splice variants (Bumgarner, 2013).

Affymetrix GeneChip® micro arrays are current the most commonly used (McCall & Almudevar, 2012). In this technology, probes for messenger (mRNA) and long intergenic non-coding RNA transcripts (lincRNA) are distributed randomly across the chip to nullify any region specific bias. Samples for hybridisation (targets) are antisense copy RNA (cRNA) made in vitro using T7 RNA polymerase in the presence of biotinylated ribonucleotides Bio-CTP and Bio-UTP. Moreover, control and experimental samples are hybridised into separate chips and a comparison is performed to determine the differential expression levels. After hybridisation, the

chip is stained and read with a confocal scanner. Once scanned, the software computes cell intensity data (CEL files) based on captured image file. It contains a single intensity value for each probe cell delineated by the grid, estimated by the Cell Analysis algorithm (http://www.affymetrix.com).

Because of the variety of GeneChip® available for the Affymetrix platform, the compatible applications are divided into various classes: Whole Transcript Expression (GeneChip® Sample HTA 2.0 Array Data, Exon 1.0 ST Array Sample Data, Gene 1.0 ST Array Sample Data); SNP & CNV DNA (Genome-Wide Human SNP Arrays and Genotype and Allele Frequency Data); Regulation & Tiling (ChIP-on-chip); 3' RNA (High Throughput Plate Data and 3' Expression Sample Data); and Integrated Genomics (Human Tiling 1.0 & Human U133 Expression Data). As the applications for GeneChip® Arrays continue to grow in number, the available software tools also evolve in variety and scope. It clearly demonstrates high performance and provides a power physical coverage of the elements in the cellular system, thereby significantly increasing the overall picture of genomics, transcriptomics and proteomics, besides epigenomics variables.

3.2 The METABRIC Breast Cancer Data Set

3.2.1 Biospecimen Collection and Ethics Approval

METABRIC has described a collection of primary fresh frozen breast cancer specimens and a subset of normal tissues, selected from tumour banks in the United Kingdom and Canada. The consortium integrated genomic and transcriptomic data, composed by a subset of 2136 gene expression arrays (Illumina_Human_WG-v3) and 2477 genotyping arrays (Affymetrix SNP 6.0) (**Table 3.1**). In turn, gene expression and genotyping values are detailed for primary tumours in the discovery (997) and validation (995) sets, with comprehensive patient long-term clinical and pathological outcomes. High quality cRNA data is also available for 144 normal samples, derived from adjacent normal breast tissue (non-tumour); and DNA information for 485 normal samples, extracted from adjacent tissue or peripheral blood. Importantly, the mentioned normal subset (controls) matches the tumours in the discovery set (Curtis et al., 2012).

Table 3.1 METABRIC microarray data description

Microarray Data Type	Number of Samples		Data Description
Gene Expression Data set (cRNA microarray)	2136	997	Tumour samples – Discovery set
		995	Tumour samples – Validation set
		144	Normal breast samples (Controls)
Genotyping Data Set (SNP, CNA, CNV)	2477	997	Tumour samples – Discovery set
		995	Tumour samples – Validation set
		485	Normal samples (Controls)

The related gene expression and genotyping data are hosted by the European Bioinformatics Institute (EBI) and deposited in the European Genome-Phenome Archive (EGA) at http://www.ebi.ac.uk/ega/, under accession number EGAS00000000083 (**Table 3.2**). The microRNA information is also available under EGAS00000000122 (**Table 3.3**). An agreement among members of the consortium and funders, nevertheless, governs the terms for accessing the METABRIC data, besides the conditions for archiving the array files (Curtis et al., 2012b). To support this information, the paperwork comprising the data access application may be downloaded from http://www.combio.group.cam.ac.uk/Resources/METABRIC.html. As regards the application, the document "Data Access Application Form" was submitted in December/2012, with a project and/or purpose following the rules and procedures respectively established in "Data Access Agreement" and "Guidelines and Information". After the review by the METABRIC Data Access Committee, the permission for downloading the microarray files was granted in February/2013.

Primary invasive breast cancer and normal breast tissue were obtained with appropriate ethical consent from the relevant institutional review board. The METABRIC study protocol, detailing the molecular profiling methodology, was approved by the ethics committees in Cambridge and Vancouver (Addenbrooke's Hospital, Cambridge, United Kingdom; Guy's Hospital, London; Nottingham; Vancouver; Manitoba), the two sites responsible for the molecular analysis of the samples (Curtis et al., 2012). The data is protected and subjected to applicable international laws, which include the UK Data Protection Act 1998 the Personal Information Protection and Electronic Documents Act (Canada) ("PIPEDA"), the Freedom of Information and Protection of Privacy Act, R.S.B.C. 1996 c. 165 ("FOIPPA") and the Personal Information Protection Act, 2003, S.B.C., c. 63 ("PIPA").

Table 3.2 Data accession – gene expression and genotyping information

Data set ID	Technology	NS[a]	Description
EGAD00010000164	Affymetrix SNP 6.0	1992	Affymetrix 6.0 CEL files
EGAD00010000162	Illumina HT 12	2136	Illumina HT 12 IDATS
EGAD00010000210	Illumina HT 12	997	Normalised expression data; Discovery set
EGAD00010000211	Illumina HT 12	995	Normalised expression data; Validation set
EGAD00010000212	Illumina HT 12	144	Normalised expression data; Normals
EGAD00010000213	Affymetrix SNP 6.0	997	Segmented (CBS[b]) copy number aberrations (CNA); Discovery set
EGAD00010000214	Affymetrix SNP 6.0	997	Segmented (CBS) copy number variants (CNV); Discovery set
EGAD00010000215	Affymetrix SNP 6.0	997	Segmented (CBS) copy number aberrations (CNA); Validation set
EGAD00010000216	Affymetrix SNP 6.0	997	Segmented (CBS) copy number variants (CNV); Validation set
EGAD00010000217	Affymetrix SNP 6.0	997	Segmented (HMM[c]) copy number aberrations (CNA); Discovery set

Note: [a]NS – number of samples; [b]CBS – circular binary segmentation; [c]Hidden Markov Model.

Human research projects conducted at the University of Newcastle by staff and students also require approval from the University's Human Research Ethics Committee (HREC). According to HREC, the project nominated *"An investigation on the consensus between different genomic and transcriptomic results in breast cancer"* ensures compliance with regulatory and legislative requirements and policies relating to human research. The use of this data set was approved by the HREC of The University of Newcastle, Australia, (approval number: H-2013-0277). This confirms the protection of the welfare and rights of participants in research, which is compulsory on all institutions and organisations that receive research funding from the Australian government.

Table 3.3 Data accession – microRNA expression information

Data set ID	Technology	NS[a]	Description
EGAD00010000434	Illumina HT 12	1302	Normalised mRNA expression
EGAD00010000436	Illumina HT 12	1302	Illumina HT 12 IDAT files
EGAD00010000438	Agilent ncRNA 60k	1480	Normalised miRNA expression data
EGAD00010000440	Affymetrix SNP 6.0 raw	1302	Segmented copy number data
EGAD00010000442	Affymetrix SNP 6.0 raw	1302	Affymetrix SNP 6.0 CEL files
EGAD00010000444	Agilent ncRNA 60k	1480	Agilent ncRNA 60k txt files

Note: [a]NS – number of samples.

3.2.2 Gene Expression Data Description

The METABRIC data transcriptome profiling was performed using the Illumina Totalprep RNA amplification kit (Ambion, Warrington, UK) and hybridised onto the Illumina HT-12 v3 Expression Beadchips per the manufacturer's instructions. The R language and environment was applied to process BeadChips, once scanning was complete and raw data were available. Processing included the generation of quality assessment information and adjustment for spatial artefacts. As a result, the data were summarised as a series of matrix – 48803 probes as rows, and sample ids representing the number of columns – containing values of log2 intensities and standard errors. For quality-control, arrays were then normalised to remove partly the probe-level data artefacts. As a result, the normalisation method makes the chips have identical intensity distribution (Curtis et al., 2012).

3.2.3 Genotype Calling

DNAs were hybridised with the Affymetrix SNP 6.0 arrays per the manufacturer's instructions (Affymetrix, Santa Clara/ CA) and analysis of copy number and genotyping were performed on the Affymetrix SNP 6.0 platform. In this scenery, the SNP CEL files are available

for download and processing. For instance, after the METABRIC managed the files, each probe was flagged as an inherited CNV when the sequences in the tumour samples within a region matched in the HapMap or in the Normals list. On the other hand, the frequencies of CNAs were obtained after removing these CNVs from the data and after matching with correspondent normal breast samples or blood, when available (Curtis et al., 2012; Dunning et al., 2010).

Furthermore, summaries germline CNV alterations were computed as LOSS and GAIN; and CNA represented by loss homozygous and hemizygous (HOMD and HETD, respectively), neutral (NEUT), gain and amplification (GAIN and AMP, respectively), and high-level amplification (HLAMP). Note that germline and somatic events were treated separately, however in some cases more than one event can occur within different segments of the same gene. For instance, a sample may exhibit a gain and neutral segment or, alternatively, a region of loss and gain within a gene. The entire information of CNV and CNA from this data set can be assessed in Supplementary Information in Curtis et al. (2012).

3.2.4 The Breast Cancer Cohort

The genomic and transcriptomic analysis of breast cancer performed by Curtis et al. (2012) defined the integrative cluster groups: IntClust 1 to IntClust 10. Other relevant details concerning the comprehensive investigation of breast tumours include: age of diagnosis; menopausal status; survival analysis; tumour grade, size and stage; lymph nodes metastasis; histological type; hormonal ER and PR condition; HER2 amplification; and P53 mutation (Curtis et al., 2012). The PAM50 intrinsic subtypes (luminal A, luminal B, HER2-enriched, normal-like and basal-like) (Parker et al., 2009) are also provided as well as the list of genes and annotated probes on the Illumina HT-12 v3 BeadChip used for classification (Curtis et al., 2012).

3.3 ROCK: Integrative Breast Cancer Data

The second data set integrates the Research Online Cancer Knowledgebase (ROCK) online interface (Sims et al., 2010; Ur-Rehman et al., 2013) and is publicly available at Gene

Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/), under data source access GSE47561. This source integrates ten data studies (Table 3.4) performed on the Affymetrix Human Genome U133A Array (HG-U133A) platform, from GEO and EBI. The matrix contains log2 RMA re-normalised gene expression data in a unique comprehensive report of 1570 samples. Thus, the GSE47561 data set was used as a second validation set to test our method. In brief, both METABRIC and ROCK data sets have information on patients' long-term clinical and pathological outcomes, including the sample assignment into intrinsic subtypes (luminal A, luminal B, HER2-enriched, normal-like, and basal-like) according to the PAM50 method (Parker et al., 2009). The METABRIC data set has a more comprehensive description of patient clinical features, whereas the ROCK data set contains limited survival information across the ten different studies.

Table 3.4 Overview of the ten data sets in the ROCK online portal

Samples	Microarray Technology	Microarray Platform	Rep.	Accession	Reference
286	Affymetrix	HG-U133A	GEO	GSE2034	(Wang et al., 2005)
200	Affymetrix	HG-U133A	GEO	GSE11121	(Schmidt et al., 2008)
230	Affymetrix	HG-U133A	GEO	GSE20194	(Popovici et al., 2010) (MAQC Consortium, 2010)
159	Affymetrix	HG-U133A/B	GEO	GSE1456	(Pawitan et al., 2005)
96	Affymetrix	HG-U133A	GEO	GSE2603	(Minn et al., 2005)
149	Affymetrix	HG-U133A/B	GEO	GSE6532	(Loi et al., 2007) (Loi et al., 2008)
42	Affymetrix	HG-U133A	GEO	GSE20437	(Graham et al., 2010)
115	Affymetrix	HG-U133A	EBI	E-TABM-185	(Lukk et al., 2010) (Wu et al., 2013)
198	Affymetrix	HG-U133A	GEO	GSE7390	(Desmedt et al., 2007) (Patil et al., 2015)
95	Affymetrix	HG-U133A	GEO	GSE5847	(Boersma et al., 2008)

Note: Rep. – Repository.

3.4 References

Allison, D. B., Cui, X., Page, G. P., & Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet, 7*(1), 55-65.

Ball, C. A., Brazma, A., Causton, H., Chervitz, S., Edgar, R., Hingamp, P., et al. (2004). Submission of microarray data to public repositories. *PLoS Biol., 2*(9), E317.

Boersma, B. J., Reimers, M., Yi, M., Ludwig, J. A., Luke, B. T., Stephens, R. M., et al. (2008). A stromal gene signature associated with inflammatory breast cancer. *Int. J. Cancer, 122*(6), 1324-1332.

Bumgarner, R. (2013). Overview of DNA microarrays: types, applications, and their future. *Curr. Protoc. Mol. Biol.*, 22-21.

Cardoso, F., Van't Veer, L., Rutgers, E., Loi, S., Mook, S., & Piccart-Gebhart, M. J. (2008). Clinical application of the 70-gene profile: the MINDACT trial. *J. Clin. Oncol., 26*(5), 729-735.

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature, 486*(7403), 346-352.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., et al. (2007). Strong Time Dependence of the 76-Gene Prognostic Signature for Node-Negative Breast Cancer Patients in the TRANSBIG Multicenter Independent Validation Series. *Clin. Cancer Res., 13*(11), 3207-3214.

Du, P., Kibbe, W. A., & Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics, 24*(13), 1547-1548.

Dunning, M. J., Curtis, C., Barbosa-Morais, N. L., Caldas, C., Tavaré, S., & Lynch, A. G. (2010). The importance of platform annotation in interpreting microarray data. *Lancet Oncol., 11*(8), 717.

Graham, K., de las Morenas, A., Tripathi, A., King, C., Kavanah, M., Mendez, J., et al. (2010). Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. *Br. J. Cancer, 102*(8), 1284-1293.

Heller, M. J. (2002). DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng, 4*(1), 129-153.

Hoheisel, J. D. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet, 7*(3), 200-210.

Leung, Y. F., & Cavalieri, D. (2003). Fundamentals of cDNA microarray data analysis. *Trends Genet., 19*(11), 649-659.

Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A. M., Gillet, C., et al. (2007). Definition of Clinically Distinct Molecular Subtypes in Estrogen Receptor–Positive Breast Carcinomas Through Genomic Grade. *J. Clin. Oncol., 25*(10), 1239-1246.

Loi, S., Haibe-Kains, B., Desmedt, C., Wirapati, P., Lallemand, F., Tutt, A., et al. (2008). Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics, 9*(1), 239.

Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., Goncalves, A., Huber, W., et al. (2010). A global map of human gene expression. *Nat. Biotechnol., 28*(4), 322-324.

MAQC Consortium. (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotech, 28*(8), 827-838.

McCall, M. N., & Almudevar, A. (2012). Affymetrix GeneChip microarray preprocessing for multivariate analyses. *Brief Bioinform, 13*(5), 536-546.

Minn, A. J., Gupta, G. P., Siegel, P. M., Bos, P. D., Shu, W., Giri, D. D., et al. (2005). Genes that mediate breast cancer metastasis to lung. *Nature, 436*(7050), 518-524.

Nguyen, D. V., Arpat, B., Wang, N., & Carroll, R. J. (2002). DNA Microarray Experiments: Biological and Technological Aspects. *Biometrics, 58*(4), 701-717.

Nielsen, T. O., Parker, J. S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., et al. (2010). A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin. Cancer Res., 16*(21), 5222-5232.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., et al. (2004). A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *N. Engl. J. Med., 351*(27), 2817-2826.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol., 27*(8), 1160-1167.

Patil, P., Bachant-Winner, P.-O., Haibe-Kains, B., & Leek, J. T. (2015). Test set bias affects reproducibility of gene signatures. *Bioinformatics, 31*(14), 2318-2323.

Pawitan, Y., Bjohle, J., Amler, L., Borg, A. L., Egyhazi, S., Hall, P., et al. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res., 7*(6), R953-964.

Popovici, V., Chen, W. Y., Gallas, B., Hatzis, C., Shi, W., Samuelson, F., et al. (2010). Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res., 12*(1), R5.

Reimers, M. (2010). Making Informed Choices about Microarray Data Analysis. *PLoS Comput. Biol., 6*(5), e1000786.

Reis-Filho, J. S., & Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet, 378*(9805), 1812-1823.

Ross, J. S., Hatzis, C., Symmans, W. F., Pusztai, L., & Hortobagyi, G. N. (2008). Commercialized multigene predictors of clinical outcome for breast cancer. *Oncologist, 13*(5), 477-493.

Schmidt, M., Böhm, D., von Törne, C., Steiner, E., Puhl, A., Pilch, H., et al. (2008). The Humoral Immune System Has a Key Prognostic Impact in Node-Negative Breast Cancer. *Cancer Res., 68*(13), 5405-5413.

Sims, D., Bursteinas, B., Gao, Q., Jain, E., MacKay, A., Mitsopoulos, C., et al. (2010). ROCK: a breast cancer functional genomics resource. *Breast Cancer Res. Treat., 124*(2), 567-572.

Sparano, J. A., & Paik, S. (2008). Development of the 21-gene assay and its application in clinical practice and clinical trials. *J. Clin. Oncol., 26*(5), 721-728.

Steemers, F. J., & Gunderson, K. L. (2005). Illumina, Inc. *Pharmacogenomics, 6*(7), 777-782.

Strehl, J. D., Wachter, D. L., Fasching, P. A., Beckmann, M. W., & Hartmann, A. (2011). Invasive Breast Cancer: Recognition of Molecular Subtypes. *Breast Care, 6*(4), 258-264.

Ur-Rehman, S., Gao, Q., Mitsopoulos, C., & Zvelebil, M. (2013). ROCK: a resource for integrative breast cancer data analysis. *Breast Cancer Res. Treat., 139*(3), 907-921.

van't Veer, L. J., Paik, S., & Hayes, D. F. (2005). Gene expression profiling of breast cancer: a new tumor marker. *J. Clin. Oncol., 23*(8), 1631-1635.

van Bakel, H., & Holstege, F. C. (2004). In control: systematic assessment of microarray performance. *EMBO Rep, 5*(10), 964-969.

Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. M., Voskuil, D. W., et al. (2002). A gene-expession signature as a predictor of survival in breast cancer. *N. Engl. J. Med., 347*(24), 1999-2009.

Villasenor-Park, J., & Ortega-Loayza, A. G. (2013). Microarray technique, analysis, and applications in dermatology. *J. Invest. Dermatol., 133*(4), e7.

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet, 365*(9460), 671-679.

Wu, M., Liu, L., Hijazi, H., & Chan, C. (2013). A multi-layer inference approach to reconstruct condition-specific genes and their regulation. *Bioinformatics, 29*(12), 1541-1552.

Yigitoglu, B., Uctepe, E., Yigitoglu, R., Gunduz, E., & Gunduz, M. (2015). Bioinformatics in Breast Cancer Research *A Concise Review of Molecular Pathology of Breast Cancer* (pp. 175-185): InTech.

CHAPTER 4

4. IDENTIFICATION OF NOVEL BIOMARKERS FOR BREAST CANCER SUBTYPING

Chapter 4 refers to the first work on the METABRIC data set and outlines the challenges involved with identifying breast cancer intrinsic subtypes. The content is structured as a research paper – **4.1 Introduction**, **4.2 Methods**, **4.3 Results**, **4.4 Discussion**, **4.5 Conclusion**, **4.6 References** and **4.7 Supporting Information** – consistent with our publication in *PLoS One*[6]. The purpose of this analysis is to: a) identify novel biomarkers for subtype individuation by exploring the competence of the CM1 score, and b) apply ensemble learning, as opposed to the use of a single classifier, for sample subtype assignment. To achieve this, we select probes with highly discriminative patterns of expression across samples for each intrinsic subtype. We further assessed the ability of these probes on assigning correct subtype labels using an ensemble learning approach. Accordingly, I portray well-established genes and novel biomarkers for predicting breast cancer intrinsic subtypes. These subtypes are compared with clincopathological information – current clinical markers ER, PR and HER2 – and survival data in the METABRIC and ROCK data sets.

[6] Milioli, H. H., Vimieiro, R., Riveros, C., Tishchenko, I., Berretta, R., & Moscato, P. (2015). The Discovery of Novel Biomarkers Improves Breast Cancer Intrinsic Subtype Prediction and Reconciles the Labels in the METABRIC Data Set. *PLoS One, 10*(7), e0129711.

4.1 Introduction

Breast cancer has been perceived as several distinct diseases characterised by intrinsic aberrations, heterogeneous behaviour and divergent clinical outcomes (Reis-Filho & Pusztai, 2011). The classification of breast cancer in discernible molecular subtypes has motivated translational researchers in the past decades towards the design of patient prognosis and the development of tailored treatments (Portier et al., 2012). In this scenario, the analysis of breast tumours using microarray data has significantly improved the disease taxonomy and the discovery of new biomarkers for implementation in clinical practice (Dowsett et al., 2013; Kelly et al., 2012; Prat et al., 2012; van't Veer et al., 2002). In the early 2000s, five intrinsic subtypes were proposed: luminal A, luminal B, HER2-enriched, normal-like and basal-like breast tumours (Perou et al., 2000; Sørlie et al., 2001; Sørlie et al., 2003). Following this initial molecular taxonomy, further sub-classifications of breast cancer in distinct entities have been suggested (Herschkowitz et al., 2007; Lehmann et al., 2011; Prat et al., 2010).

The transcriptomic patterns observed across subtypes has given us insight into the molecular complexity and inherent alterations in tumour cells modelling the breast cancer heterogeneity and unpredicted outcome (Nielsen et al., 2010; Weigelt; Baehner; et al., 2010). Strikingly, intrinsic gene lists have been explored to reliably assign breast tumour samples into formal molecular subtypes, survival rate and treatment outline (Bastien et al., 2012; Herschkowitz et al., 2007; Hu et al., 2006; Parker et al., 2009; Perou et al., 2000; Sørlie et al., 2001; Sørlie et al., 2003; van De Vijver et al., 2002). Recently, Parker et al. (2009) proposed a list of 50 genes that together with the Prediction Analysis for Microarrays (PAM) classification algorithm (Tibshirani et al., 2002) aimed at identifying subtypes and enlarging the prognostic information with high potential for validation in clinical settings (Parker et al., 2009; Perou et al., 2010; Weigelt; Mackay; et al., 2010). The resulting technique, called the PAM50 method, has been widely applied to categorise tumours into one of the five classical intrinsic subtypes.

Although independent cohorts attempted to identify molecular subtypes, the chosen microarray-based Single Sample Predictor (SSP) model revealed unreliable assignments and modest agreement between studies (Haibe-Kains et al., 2012; Weigelt; Mackay; et al., 2010). In fact, the perceived inability of some analytical methods to deal with the challenges of processing high-dimensional data, in addition to the difficulties on validating independent/unpaired technologies may limit the precise characterisation of the subtypes (Sotiriou & Pusztai, 2009; Weigelt; Mackay; et al., 2010; Weigelt & Reis-Filho, 2009).

Therefore, novel methods are urgently needed in order to provide better tumour stratification and accurate biomarkers identification (Colombo et al., 2010; Weigelt et al., 2012). In this scenario, the high quality of the microarray gene expression data set processed by METABRIC, with over 2000 samples (Curtis et al., 2012), offers a unique opportunity to refine and expand the list of transcripts that best discriminate intrinsic subtypes. A precise classification of breast tumours, consequently, would lead to improvements in the valuation of the disease, currently guided by oestrogen and progesterone receptor (ER and PR) status, and HER2 amplification (Ambs, 2010; Weigelt & Reis-Filho, 2009).

In this report, we focus on the use of a ranking feature method based on the CM1 score (Marsden et al., 2013) to identify probe sets that appear naturally from the METABRIC breast cancer data set. For doing so, we use the entire set of 48803 probes as an alternative to the selection from pre-existing literature as performed by other authors. Moreover, the quality of the probes for predicting subtypes is carefully appraised in the METABRIC data set (Illumina BeadArray) and further validated in different studies (Affymetrix GeneChip) accessed through the ROCK interface (Ur-Rehman et al., 2013). However, instead of relying on a single method to assign sample subtype, as suggested by Parker et al. (2009) with the PAM50 method, we explore ensemble learning. Our analysis is based on the performance of a large set of classification models from the Weka software suite (Witten et al., 2016); a technique previously recommended by Gómez-Ravetti and Moscato (2008). The classifiers are used in combination with the list of probes selected using CM1 score and, alternatively, with the 50 genes from the PAM50 commercial assay. We also compute several statistical measures to determine the power of both lists on predicting breast cancer subtypes. Ultimately, we correlate the study outcomes within current clinical information and survival analysis.

4.2 Methods

4.2.1 Study Design and Computing Resources

In this study, we propose a systematic approach that aims at improving breast cancer subtype prediction. The systematic approach is built based on feature selection and data mining concepts. We first compute the CM1 score – using the microarray mRNA expression values – to rank the whole set of probes based on their discriminative power across breast cancer subtypes.

We then select the top 10 probes that best represent each intrinsic subtype. The quality of this selection is assessed using a set of classifiers from the Weka software suite with the METABRIC and ROCK data sets, followed by the statistical analysis. The process flow is depicted in **Figure 4.1**, and further explained in the remainder of this section.

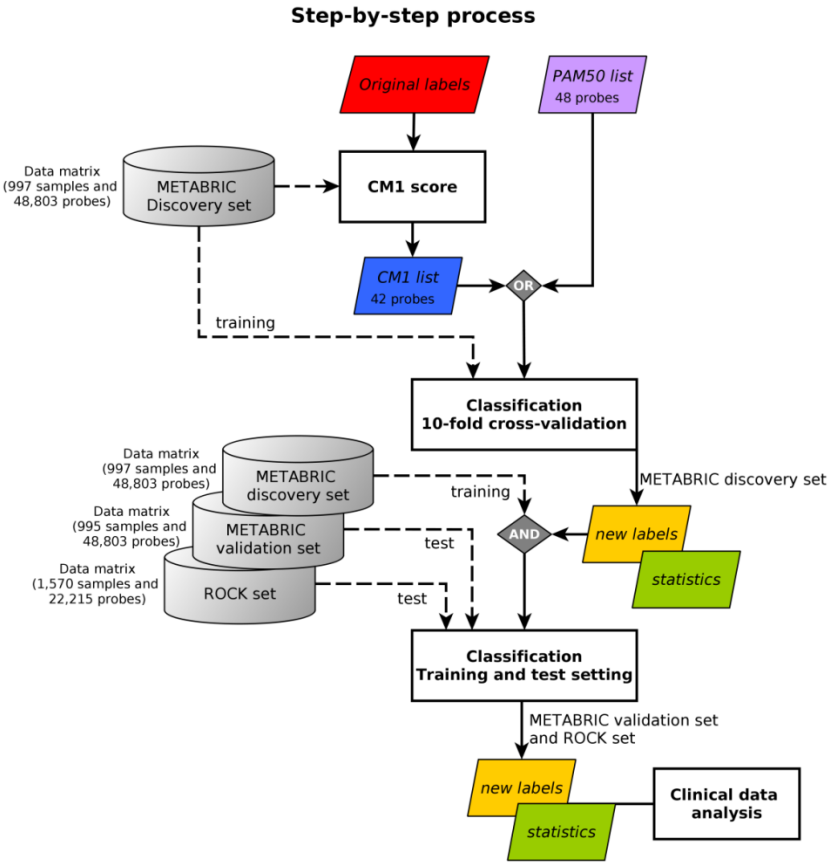


Figure 4.1 The step-by-step process

The image shows the method steps based on CM1 score and ensemble learning. The METABRIC discovery set is used to compute the CM1 score, based on the original labels previously assigned with the PAM50 method. This step has an output of 42 discriminative probes selected, the CM1 list. The following step involves the sample subtype classification based on a 10-fold cross-validation. Samples in the METABRIC discovery set are considered to train 24 classifiers using the CM1 list and, alternatively, the PAM50 list. The samples are partitioned into ten folds; then a model is built using 90% of samples, which is used to predict the labels of the remaining 10%. After the ten turns are finished, the level of association between the predicted and original METABRIC labels is computed using several statistics. In the training-test setting, labels of samples in the METABRIC validation set and ROCK set are predicted with the models built in the discovery. Statistics measurements are again computed to assess the model performance on predicting breast cancer subtypes. In both classification steps, the new labels are attributed based on the consensus of the majority of the classifiers. Finally, the results or new labels are compared against the clinical data, the current markers ER, PR and HER2, and survival curves.

4.2.2 Selection of Biomarkers Using the CM1 Score

The CM1 score is a supervised univariate method used to measure the difference in expression levels of samples in two different classes (Marsden et al., 2013). In this study, it is used as a ranking feature to select a subset of highly discriminative probes for each breast cancer intrinsic subtype. Let *X* and *Y* be a partition of a set of samples into two classes, with *X* the 'class of interest' and *Y* the 'remaining classes'. A sample either belongs to *X* or to *Y*. For each probe *i* we compute the CM1 score (**Equation 4.1**) as:

Equation 4.1 CM1 score

$$CM1_i(X,Y) = \frac{\bar{x}_i - \bar{y}_i}{1 + (\max\{y_i\} - \min\{y_i\})}$$

where $\bar{x}_i$ is the average expression value of the probe *i* for samples in class *X*, $\bar{y}_i$ is the average expression value of the probe *i* for samples in class *Y* ; max(y_i) and min(y_i) are the maximum and minimum expression values of the probe *i* for samples in the class *Y*, respectively. **Equation 4.1** can be interpreted as the normalised difference between the averages of expression values in the class *X* and *Y*. The normalisation is proportional to the range of values in *Y*.

To define the most discriminative probes for each breast cancer subtype (luminal A, luminal B, HER2-enriched, normal-like and basal-like), we computed the CM1 score for each of 48803 probes taking the subtype of interest and the remaining ones. This results in 5 lists of 48803 CM1 scores.

Considering the fact that (Parker et al., 2009) were able to define the five breast cancer classes based on 50 genes, for each subtype we chose the 10 most important probes (5 with the greatest positive CM1 score values – indicating up-regulated probes relative to the other subtypes –, and 5 with the smallest negative values – representing down-regulation). This set is referred to as the *balanced top ten* in this paper. Collecting the balanced top ten lists of all subtypes leads to a new set of 42 unique Illumina probes, meaning that 8 probes appear in multiple subtypes. This list is hereafter called the *CM1 list*.

4.2.3 The Quality of CM1 List Based on Ensemble Learning

The quality of the CM1 list for distinguishing subtypes was assessed using a list of well-known classifiers available in the Weka data mining software suite (Witten et al., 2016). It uses different types of classifiers such as bayesian, functions, lazy, meta, rule-based and decision trees. Each classifier was trained with a subset of the data comprising all samples in the METABRIC discovery set and the 42 probes in the CM1 list using both 10-fold cross-validation and training-test setting. In the 10-fold cross-validation, the samples are first partitioned into ten folds; then a model is built using 90% of samples, which is thereafter used to predict the labels of the remaining 10%. After the ten turns are finished, the level of association between the predicted and original METABRIC labels is computed using Cramer's V (Liebetrau, 1983). In the training-test setting, labels of samples in the METABRIC validation set and ROCK data are predicted using models built with the samples in the discovery set. The new labels were attributed based on the consensus of the majority of the classifiers (i.e. more than 50% percent), and whenever such condition was not achieved samples were marked as inconsistent (INC).

A similar approach was performed with the PAM50 list to serve as baseline for comparing the results obtained with the 42 probes from the CM1 list. The 50 genes identified by (Parker et al., 2009) were mapped to Illumina probes by Curtis et al. (2012), following strict criteria. Only genes and corresponding probe with perfect annotation (Dunning et al., 2010) on the Illumina HT-12 v3 BeadChip were considered. Probes containing SNPs, multiple targets or mismatches, or lying in repeat-masked regions were discarded. Finally, a total of 48 probes corresponding to genes in the PAM50 list were selected to conduct the classification experiments as described for the CM1 list. For Affymetrics HG-U133A, the CM1 and PAM50 lists were mapped according to *genefu* R package, using Entrez Gene ID as reference. For instance, the 42 probes from the CM1 list matched 33 probes, whereas the 48 from PAM50 list paired 43 probes in the Affymetrix platform. In case of multiple mappings the probe with the most variation was selected according to the *genefu* instructions. Before testing the classifiers in ROCK data set, the Affymetrix and Illumina expression levels were min-max normalised.

4.2.4 Statistical Analysis

Cramer's V. Given a *r* x *c* contingency table, with “r” rows and “c” columns, describing the association between the original labels and those predicted by the majority of classifiers, respectively, Cramer's V measures the level of association between those two nominal variables.

The statistic ranges from 0, representing no association between the two variables, to 1, representing complete association. Cramer's V is computed using **Equation 4.2**.

Equation 4.2 Cramer's V

$$\emptyset = \sqrt{\frac{X^2}{N\ min\{r-\ 1, c-1\}}}$$

where *N* is the number of samples in the data set, and X^2 is Pearson's chi-squared value.

Average sensitivity (AS). The average sensitivity (Witten et al., 2016) was also computed to assess the performance of classifiers with both lists. The AS is the average proportion of accurately classified samples of each subtype. Considering a *r* x *c* contingency table associating initial and predicted labels, the average sensitivity of a classifier is given by **Equation 4.3**.

Equation 4.3 Average sensitivity

$$AS = \frac{1}{r}\sum \frac{n_{ii}}{n_{i\bullet}}$$

where *r* is the number of classes (subtypes), n_{ii} is the number of samples of class *i* correctly predicted as *i*, and $n_{i\bullet}$ is the number of samples of class *i* (row marginal).

Fleiss' kappa. The consensus of the different classification methods concerning the samples' labels was measured by the popular interrater reliability metric Fleiss' kappa (Fleiss, 1971; Fleiss et al., 2004). The statistic was used to gauge not only the agreement among classifiers trained with the different probe sets, but also between the labels assigned by the majority of classifiers and the original METABRIC labels. It also quantifies the agreement between predicted labels using the CM1 and PAM50 lists.

Assuming a *r* x *c* contingency table informing how many times each of the classes were assigned to each of the *s* samples in the *k* different sample labelling, the Fleiss' kappa statistic is computed as defined by **Equation 4.4**.

Equation 4.4 Fleiss' kappa

$$\kappa = \frac{\sum\sum n_{ij}^2 - sk\,[1 + (k-1)\sum p_j^2]}{sk\,(k-1)\,(1-\sum p_j^2)}$$

where n_{ij} contains the number of times sample i was assigned label j, $\sum_j n_{ij} = k$, and $p_j = (\sum_j n_{ij})/sk$ is the probability with which the label j is assigned to a sample.

Kappa values range from $[-\sum p_j^2 /(1-\sum p_j^2)]$ to + 1, which, according to Landis and Koch (1977), can be interpreted in the following manner: (1) values below zero show *poor agreement*; (2) $0 \leq \kappa \leq 0.20$, *slight agreement*; (3) $0.21 \leq \kappa \leq 0.40$, *fair agreement*; (4) $0.41 \leq \kappa \leq 0.60$, *moderate agreement*; (5) $0.61 \leq \kappa \leq 0.80$, *substantial agreement*; and (6) $0.81 \leq \kappa \leq 1$, *almost perfect agreement*.

Adjusted Rand Index. The agreement between pairs of sample labellings was also quantified using this metric. It ranges between 0 to 1, where 1 indicates an almost perfect concordance between the two compared bipartitions, and 0 a complete discordance between them. The Adjusted Rand Index is a version of Rand index corrected for chance when the partitions are picked at random (Hubert & Arabie, 1985; Vinh et al., 2009). Given a *r* x *c* contingency table between two labelling *R* and *C*, it can be measured by (**Equation 4.5**):

Equation 4.5 Adjusted Rand Index

$$\mathrm{ARI(R,C)} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{n_{i\bullet}}{2}\sum_j \binom{n_{\bullet j}}{2}]/\binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\bullet}}{2} + \sum_j \binom{n_{\bullet j}}{2}\right] - \sum_i \binom{n_{i\bullet}}{2}\sum_j \binom{n_{\bullet j}}{2}]/\binom{N}{2}}$$

where $1 \leq i \leq r$, $1 \leq j \leq c$, and n_{ij} is an entry of the contingency table representing the number of samples that are in class R_i in the partition *R* and C_j in the partition *C*, $n_{i\bullet}$ and $n_{\bullet j}$ are the table's marginals.

4.2.5 Survival Analysis

The survival analysis for each breast cancer subtype is performed using Cox proportional hazards model from the package survival in the R software (Kalbfleisch & Prentice, 2011; Therneau & Grambsch, 2000). Only patients who either died due to the disease or are still alive are considered for model estimation. The clinical parameters relevant for the survival study are chosen in correspondence with Curtis et al. (2012): age at the time of diagnosis, tumour size, tumour grade, the number of positive lymph nodes and ER status according to immunohistochemistry. Since the probability model based on the observations available at certain time points becomes less and less reliable with the increasing time, the median survival lines based on the last 10 observations are plotted in dash. Due to the compilation of ten different studies and the existence of significant gaps in patients' clinical information, the survival curves in the ROCK data set are not representative across subtypes. In particular, the number of patients with information about overall survival and disease free survival is limited to only 405, with no specification on the cause of death (i.e. if due to disease or not).

4.3 Results

4.3.1 Section Description and Resources

To understand the results described in this section, we introduce the sequence of our approach which combines the CM1 score and ensemble learning. First, we detail the selection of discriminative probes ranked according to the CM1 score; calculated for each of the five breast cancer subtypes. Second, we show the quality of our probes by using 24 classification models based on a 10-fold cross-validation and training-test setting in the METABRIC and ROCK data sets. The same approach is also performed with the list of 50 genes used in the PAM50 method. In addition, statistical analysis is reported to determine the power of both lists on predicting breast cancer subtypes. Finally, we demonstrate the consistency between the new labels assigned with current clinical markers ER, PR and HER2, and survival curves. The step-by-step approach is detailed in the Materials and Methods section.

4.3.2 Using the CM1 List to Differentiate Breast Cancer Subtypes

The CM1 score was applied to rank the set of 48803 probes for each of the five subtypes in the METABRIC discovery data set (**Supporting Information Table 4.9**). It is important to remark that this method used the original PAM50 subtypes attributed to samples in the METABRIC discovery set. The purpose of doing so is to provide a better molecular characterisation of each class using the wealth of the METABRIC transcriptomic data, besides improving the breast cancer subtype prediction. The probes with the top five negative and top five positive CM1 scores were selected for each subtype. Here, we aimed at obtaining 50 probes that appear naturally from a rich and unique data set. We would then be able to compare such a list with the list of 50 genes embedded in the PAM50 method – the PAM50 list. The final list comprising the union of the top ranked probes is displayed in Table 4.1, and their CM1 scores and ranks in each subtype in **Table 4.2**. Some of the 50 probes selected, however, discriminate more than one subtype and resulted in a list of 42 unique elements, the CM1 list. Our selection includes 30 novel biomarkers, while the remaining 12 genes are common with the PAM50 list.

The effectiveness of the CM1 list for segregating the five subtypes is depicted in **Figure 4.2**. The figure shows the expression values of the top five negative and top five positive ranked probes for each subtype across 997 samples in the METABRIC discovery set. For instance, the ten probes selected for the basal-like subtype – the most representative class – expose a consistent separation between samples from this class and the remaining ones. The second heat map in **Figure 4.3** illustrates the expression levels of unique probes from the CM1 list in the Illumina platform, in which rows represent probes and columns represent samples. Rows and columns were ordered according to gene expression similarity using a memetic algorithm. This image also exposes the overall discriminative power of our list for distinguishing samples of the five subtypes.

A detailed description of our 42 probes in the context of the literature can be found in **Supporting Information – Text 4.1**. Among them we highlight seven, targeting the following transcripts: *AURKB*, *CCL15*, *C6orf211*, *GABRP*, *IGF2BP3*, *PSAT1*, and *TFF3*. **Figure 4.4** shows the box plot of their expression levels across intrinsic subtypes in the METABRIC discovery and validation sets, and the ROCK set. We emphasised these transcripts due to the remarkable differential expression behaviour across the five classes. Besides, they are novel potential markers for breast cancer subtyping, not considered by Parker et al. (2009). Box plots of expression levels for all transcripts in the CM1 list in the METABRIC discovery and validation and ROCK data sets are provided in **Supporting Information – Figure 4.11**. Even

though those probes were selected from the METABRIC discovery set only, their variation across subtypes in the validation set and ROCK test set are also impressive.

Table 4.1 CM1 List

Probe ID	Gene name	Gene Symbol and Aliases
ILMN_1684217	Aurora kinase B	***AURKB**; AIK2, AIM1, ARK2, AurB, IPL1, STK5, AIM-1, STK12, PPP1R48, aurkb-sv1, aurkb-sv2*
ILMN_1683450	Cell division cycle associated 5	***CDCA5**; SORORIN*
ILMN_1747016	Centrosomal protein 55kDa	***CEP55**; CT111, URCC6, C10orf3*
ILMN_2212909	Maternal embryonic leucine zipper kinase	***MELK**; HPK38*
ILMN_1714730	Ubiquitin-conjugating enzyme E2C	***UBE2C**; UBCH10, dJ447F3.2*
ILMN_1796059	Ankyrin repeat domain 30A	***ANKRD30A**; NY-BR-1, RP11-20F24.1*
ILMN_1651329	Long intergenic non-protein coding RNA 993	*LINC00993*
ILMN_2310814	Microtubule-associated protein tau	***MAPT**; TAU, MSTD, PPND, DDPAC, MAPTL, MTBT1, MTBT2, FTDP-17*
ILMN_1728787	Anterior gradient 3	***AGR3**; HAG3, hAG-3, BCMP11, PDIA18*
ILMN_1688071	N-acetyltransferase 1	***NAT1**; AAC1, MNAT, NATI, NAT-1*
ILMN_1729216	Crystallin, alpha B	***CRYAB**; MFM2, CRYA2, CTPP2, HSPB5, CMD1II, CTRCT16*
ILMN_1666845	Keratin 17	***KRT17**; PC, K17, PC2, PCHC1*
ILMN_1786720	Prominin 1	***PROM1**; RP41, AC133, CD133, MCDR2, STGD4, CORD12, PROML1, MSTP061*
ILMN_1753101	V-set domain containing T cell activation inhibitor 1	***VTCN1**; B7X, B7H4, B7S1, B7-H4, B7h.5, VCTN1, PRO1291, RP11-229A19.4*
ILMN_1798108	Chromosome 6 orf 211	*C6orf211*
ILMN_1747911	Cyclin-dependent kinase 1	***CDK1**; CDC2, CDC28A, P34CDC2*
ILMN_1666305	Cyclin-dependent kinase inhibitor 3	***CDKN3**; KAP, CDI1, CIP2, KAP1*
ILMN_1678535	Estrogen receptor 1	***ESR1**; ER, ESR, Era, ESRA, ESTRR, NR3A1*
ILMN_2149164	Secreted frizzled-related protein 1	***SFRP1**; FRP, FRP1, FrzA, FRP-1, SARP2*
ILMN_1788874	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3	***SERPINA3**; ACT, AACT, GIG24, GIG25*
ILMN_1785570	Sushi domain containing 3	***SUSD3***
ILMN_1803236	Chloride channel accessory 2	***CLCA2**; CACC, CACC3, CLCRG2, CaCC-3*
ILMN_2161820	Glycine-N-acyltransferase-like 2	***GLYATL2**; GATF-B, BXMAS2-10*
ILMN_1810978	Mucin-like 1	***MUCL1**; SBEM*
ILMN_1773459	SRY (sex determining region Y)-box 11	***SOX11***

ILMN_1674533	Transient receptor potential cation channel, suamily V, member 6	***TRPV6****; CAT1, CATL, ZFAB, ECAC2, ABP/ZF, LP6728, HSA277909*
ILMN_1687235 ILMN_2358760	Hepsin	***HPN****; TMPRSS1*
ILMN_1655915	Matrix metallopeptidase 11 (stromelysin 3)	***MMP11****; ST3, SL-3, STMY3*
ILMN_1711470	Ubiquitin-conjugating enzyme E2T (putative)	***UBE2T****; PIG50, HSPC150*
ILMN_1740609	Chemokine (C-C motif) ligand 15	***CCL15****; LKN1, NCC3, SY15, HCC-2, LKN-1, MIP-5, NCC-3, SCYL3, MIP-1D, MRP-2B, SCYA15, HMRP-2B, MIP-1 delta*
ILMN_1789507	Collagen, type XI, alpha 1	***COL11A1****; STL2, COLL6, CO11A1*
ILMN_1651282	Collagen, type XVII, alpha 1	***COL17A1****; BP180, BPA-2, BPAG2, LAD-1, BA16H23.2*
ILMN_1723684	Duffy blood group, atypical chemokine receptor	***DARC****; FY, Dfy, GPD, GpFy, ACKR1, CCBP1, CD234, WBCQ1*
ILMN_1809099	Interleukin 33	***IL33****; DVS27, IL1F11, NF-HEV, NFEHEV, C9orf26, RP11-575C20.2*
ILMN_1766650	Forkhead box A1	***FOXA1****; HNF3A, TCF3A*
ILMN_1811387	Trefoil factor 3 (intestinal)	***TFF3****; ITF, P1B, TFI*
ILMN_1738401	Forkhead box C1	***FOXC1****; ARA, IGDA, IHG1, FKHL7, IRID1, RIEG3, FREAC3, FREAC-3*
ILMN_1689146	Gamma-aminobutyric acid (GABA) A receptor, pi	***GABRP***
ILMN_1807423	Insulin-like growth factor 2 mRNA binding protein 3	***IGF2BP3****; CT98, IMP3, KOC1, IMP-3, VICKZ3*
ILMN_1692938	Phosphoserine aminotransferase 1	***PSAT1****; PSA, EPIP, PSAT*
ILMN_1668766	Rhophilin associated tail protein 1	***ROPN1****; CT91, ODF6, ROPN1A, RHPNAP1, ropporin*

Table 4.2 Scores and ranks for the CM1 list

Probe ID	Score	rank	score	Rank	score	rank	score	rank	score	Rank
ILMN_1728787	0.20	5	0.14	5	-0.31	2		54	-0.46	3

ILMN_1796059	0.22	3		8730		1434		3666	-0.39	5
ILMN_1684217	-0.20	1		74		497		146		97
ILMN_1798108		1980	0.16	2		68		405		179
ILMN_1740609		476		43		970	0.252	3		2776
ILMN_1747911		80	0.14	4		2080		194		1496
ILMN_1683450	-0.20	3		30		306		79		166
ILMN_1666305		16	0.15	3		438		167		917
ILMN_1747016	-0.20	5		88		362		73		127
ILMN_1803236		1875		354	0.32	3		688		13483
ILMN_1789507		12176		5363		1820	-0.155	3		9245
ILMN_1651282		915		16		4821	0.244	4		12205
ILMN_1729216		6657	-0.15	5		3008		52		45
ILMN_1723684		456		14		2830	0.255	2		4215
ILMN_1678535		8	0.18	1	-0.36	1		7	-0.44	4
ILMN_1766650		70		85		12522		216	-0.48	2
ILMN_1738401		1047		10		2254		226	0.44	1
ILMN_1689146		1177		13		1833		283	0.41	2
ILMN_2161820		310		270	0.33	1		791		1479
ILMN_1687235		79		1942		58	-0.157	2		211
ILMN_2358760		105		1941		73	-0.152	4		284
ILMN_1807423		1269		2087		21820		11567	0.41	3
ILMN_1809099		3400		141		6282	0.275	1		23413
ILMN_1666845		8365	-0.19	2		3879		35		29
ILMN_1651329	0.22	1		2481		1149		1159		20
ILMN_2310814	0.22	2		8776		33		1131		23
ILMN_2212909	-0.20	4		137		501		92		65
ILMN_1655915		5274		3486		3832	-0.166	1		4148
ILMN_1810978		20520		9	0.33	2		6		1495
ILMN_1688071	0.22	4		902	-0.26	5		24		19
ILMN_1786720		988	-0.17	3		273		465		20
ILMN_1692938		68		343		93		1864	0.39	5
ILMN_1668766		721		62		1415		368	0.41	4
ILMN_1788874		148		4633	-0.26	4		1961		1462
ILMN_2149164		11497	-0.20	1		1697	0.244	5		40
ILMN_1773459		185		621	0.29	5		10046		483
ILMN_1785570		11		2499	-0.31	3		438		82
ILMN_1811387		26		64		1263		661	-0.52	1
ILMN_1674533		643		605	0.30	4		2756		1819
ILMN_1714730	-0.20	2		9		318		43		353
ILMN_1711470		56		7		1732	-0.145	5		1113
ILMN_1753101		474	-0.15	4		2424		3373		1522

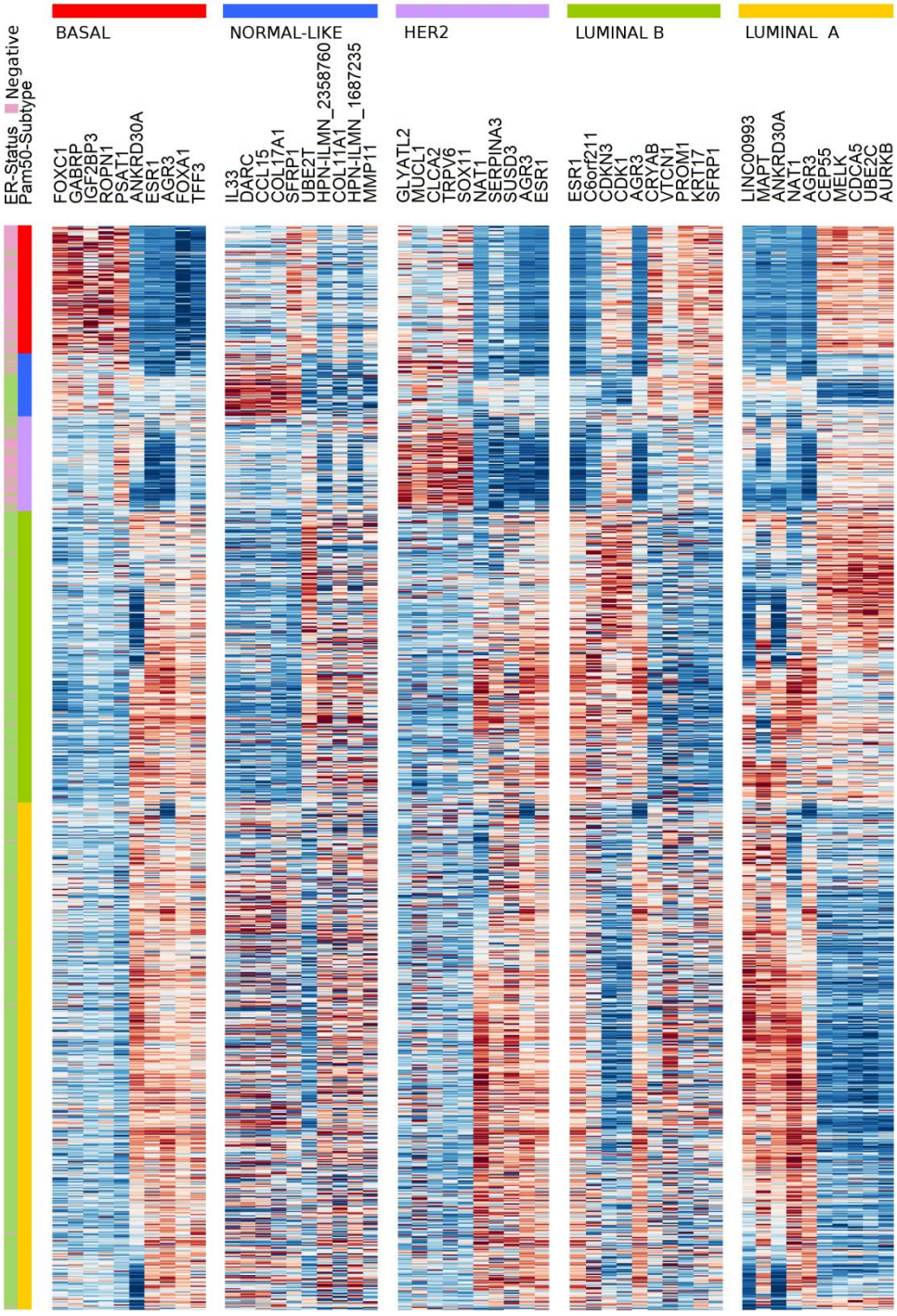


Figure 4.2 The gene expression profile of the balanced top ten probes selected for each of the five breast cancer intrinsic subtypes across 997 samples from the discovery set.

The annotated genes are defined for each subtype as an intrinsic, highly discriminative, signature. Samples were ordered according to the gene expression similarities in each breast cancer subtype. Colours represent the selected genes and sample subtypes: luminal A (yellow), luminal B (green), HER2-enriched (purple), normal-like (blue), and basal-like (red).

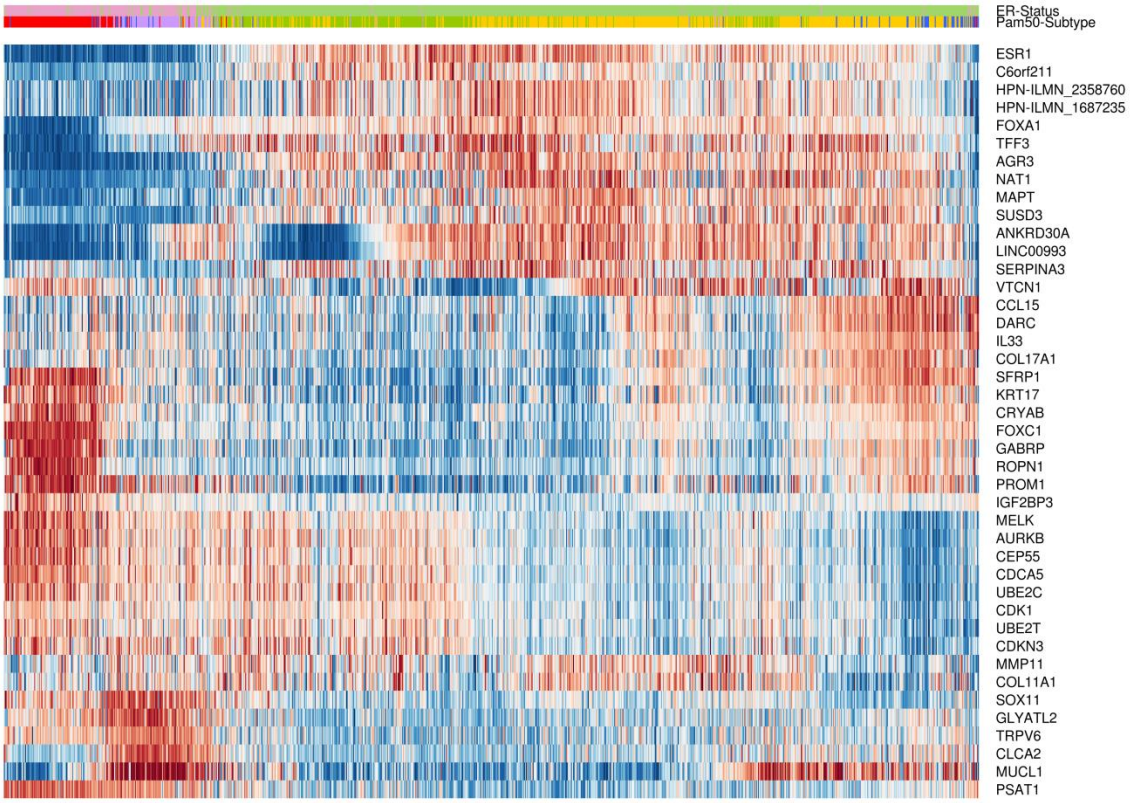


Figure 4.3 Gene expression patterns of the 42 probes selected using the CM1 score

The heat map diagram exhibit 42 probes (rows) and 997 samples (columns) from the discovery set ordered according to gene expression similarity, based on a memetic algorithm [27]. The labels highlighted on top show the sample distribution according to the ER positive and negative status. It also illustrates the original PAM50 subtypes luminal A (yellow), luminal B (green), HER2-enriched (purple), normal-like (blue), and basal-like (red) in the METABRIC discovery set. Two probes in the CM1 list refer to the same gene, HPN, which was then appended with the corresponding Illumina probe ID.

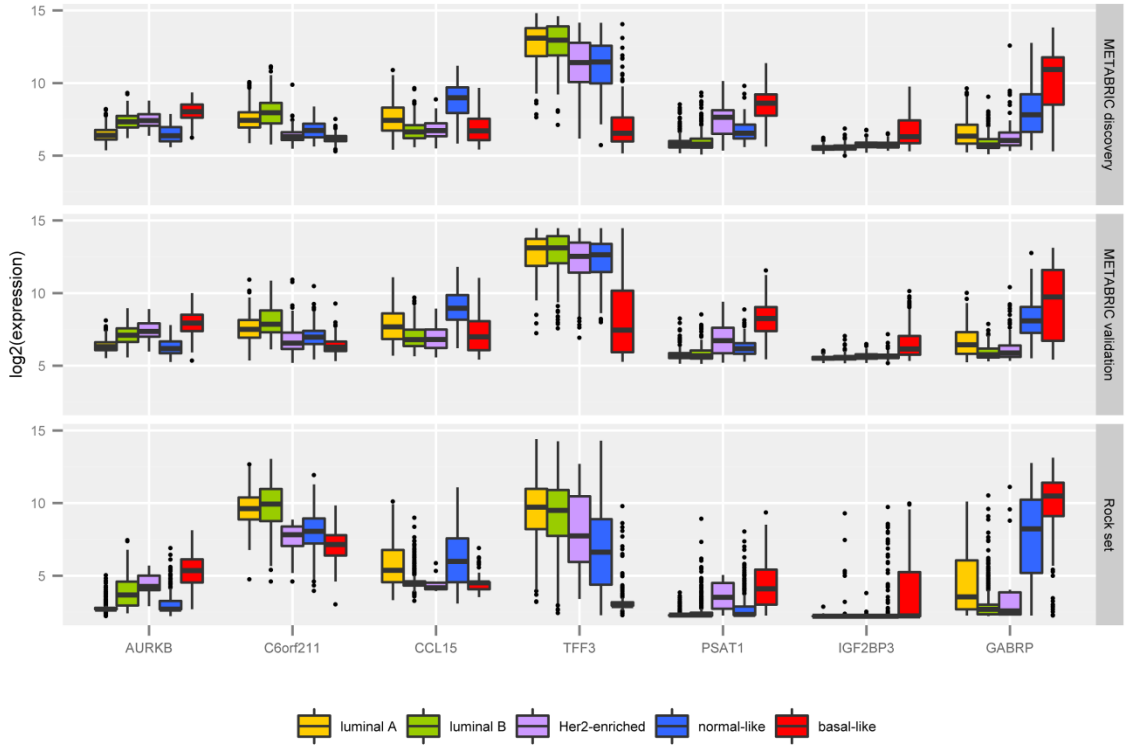


Figure 4.4 The mRNA log2 normalised expression values of 7 novel highly discriminative biomarkers across the five intrinsic subtypes

The box plot uncover the values of 997 samples in the METABRIC discovery set, 989 in the validation set, and 1570 in the ROCK test set.

4.3.3 The High Levels of Agreement Between CM1 and PAM50 Lists

After applying the ensemble learning, several statistical measures were computed as referred in Materials and Methods. The main purpose of the statistics is to determine the performance of the 24 classification methods from the Weka software suite. In other words, we investigate the consistency of intrinsic subtype labels attributed by the majority of classifiers having as input either the CM1 or PAM50 lists. The quality of both lists was estimated according to the Cramer's V statistic and the Average Sensitivity. Additionally, we computed the popular interrater reliability metric Fleiss' kappa to establish the consensus of sample labelling across different classifiers. This metric was used to gauge the agreement among classifiers trained with CM1 and PAM50 lists against the original labels in the data sets, and between the labels assigned by the majority of classifiers using both lists. Ultimately, we applied the Adjusted Rand Index to quantify the agreement between pairs of samples that are either in the same class or in different classes according to both lists.

Average Cramer's V statistic and Average Sensitivity to measure the performance of individual classifiers.

We determined the performance of the ensemble learning (**Supporting Information – Table 4.10 The performance of the classifiers using the CM1 list**

Table describing the performance of each classifier on the METABRIC discovery and validation sets, and ROCK test set using the CM1 list is summarised below (**Table 4.10s**). The original published file shows the percentage of correctly, incorrectly and not classified samples, Fleiss Kappa index, Cramer's V, Average Sensitivity, and other values for classification. The 24 classifiers from the Weka software suite are also listed. In addition, it contains the labels predicted by each classifier. Count of predicted labels was obtained with the consensus of the majority of classifiers.

Available online: doi:10.1371/journal.pone.0129711.s004

Supporting Information – Table 4.11

Table 4.11 The performance of the classifiers using the PAM50 list

Table describing the performance of each classifier on the METABRIC discovery and validation sets, and ROCK test set using the PAM50 list is summarised below (**Table 4.11s**). The original published file shows the percentage of correctly, incorrectly and not classified samples, Fleiss Kappa index, Cramer's V, Average Sensitivity, and other values for classification. The 24 classifiers from the Weka software suite are also listed. In addition, it contains the labels predicted by each classifier. Count of predicted labels was obtained with the consensus of the majority of classifiers.

Available online: doi:10.1371/journal.pone.0129711.s005

Table 4.10 **and Table 4.11**) with two measures: Cramer's V statistic and Average Sensitivity (**Table 4.3**). Cramer's V is used to measure the strength of association among variables in the row and column, given a contingency table (**Table 4.4, Table 4.5 and Table 4.6**). The rows represent the original PAM50 labels and the columns the subtypes assigned by the majority of the classifiers in the ensemble. For instance, Cramer's V statistic showed an average association between original and predicted subtypes of 0.73 ± 0.06 and 0.63 ± 0.04 in the METABRIC discovery and validation sets respectively with the CM1 list; and 0.75 ± 0.06 and 0.64 ± 0.04 with PAM50 list. Expanding the validation process using the ROCK test set, Cramer's V ranged from 0.57 ± 0.06 with the CM1, and 0.59 ± 0.05 using PAM50 list.

Table 4.3 The ensemble learning overall performance on assigning labels to samples in the METABRIC discovery and validation sets, and ROCK test set

	CM1 list		PAM50 list	
Data set	**CV**	**AS**	**CV**	**AS**
METABRIC discovery	0.73 ± 0.06	0.76 ± 0.06	0.75 ± 0.06	0.78 ± 0.07
METABRIC validation	0.63 ± 0.04	0.64 ± 0.04	0.64 ± 0.04	0.65 ± 0.05
ROCK test set	0.57 ± 0.06	0.67 ± 0.08	0.58 ± 0.05	0.69 ± 0.08

Note: Values are given as average ± std. deviation. CV- Cramer's V; AS- Average Sensitivity.

Table 4.4 Contingency tables for predicted labels using classifiers trained with the CM1 list

	METABRIC discovery						**METABRIC validation**						**ROCK test set**					
	LA	**LB**	**H**	**N**	**B**	**I**	**LA**	**LB**	**H**	**N**	**B**	**I**	**LA**	**LB**	**H**	**N**	**B**	**I**
LA	435	19	2	2	0	8	252	2	0	0	0	1	452	122	2	0	0	17
LB	24	234	0	0	0	10	62	156	0	0	0	6	18	371	42	0	2	14
H	4	4	67	0	2	10	23	45	71	2	2	10	0	1	13	0	0	0
N	13	0	8	31	0	6	80	0	0	59	0	5	115	8	36	74	56	50
B	0	0	10	2	103	3	6	7	22	19	142	17	0	0	0	7	166	4

Note: Rows contain labels assigned by the majority of classifiers trained with the CM1 list, while columns contain the original METABRIC labels assigned using the PAM50 method. In this table, LA corresponds to luminal A, LB corresponds to luminal B, H to HER2-enriched, N to normal-like, and B to basal-like. Labels marked as I refer to inconsistent assignments; situations where the classifiers did not achieve the majority on attributing a subtype label.

Table 4.5 Contingency tables for predicted labels using classifiers trained with the PAM50 list

	METABRIC discovery						**METABRIC validation**						**ROCK test set**					
	LA	**LB**	**H**	**N**	**B**	**I**	**LA**	**LB**	**H**	**N**	**B**	**I**	**LA**	**LB**	**H**	**N**	**B**	**I**
LA	440	17	1	1	0	7	254	0	0	0	0	1	530	46	2	0	0	15
LB	25	239	0	0	0	4	56	162	0	0	0	6	53	327	34	0	3	30
H	0	5	72	0	1	9	21	39	80	0	0	13	0	0	12	0	0	2
N	9	0	2	34	1	12	82	0	0	55	0	7	105	4	18	92	67	53
B	0	0	7	1	103	7	4	7	20	14	145	23	0	0	3	0	172	2

Note: Rows contain labels assigned by the majority of classifiers trained with the PAM50 list, while columns contain the original METABRIC labels assigned using the PAM50 method. In this table, LA corresponds to luminal A, LB corresponds to luminal B, H to HER2-enriched, N to normal-like, and B to basal-like. Labels marked as I refer to inconsistent assignments; situations where the classifiers did not achieve the majority on attributing a subtype label.

Table 4.6 Contingency tables for predicted labels using classifiers trained with CM1 and PAM50 lists

	METABRIC discovery						METABRIC validation						ROCK test set					
	LA	**LB**	**H**	**N**	**B**	**I**	**LA**	**LB**	**H**	**N**	**B**	**I**	**LA**	**LB**	**H**	**N**	**B**	**I**
LA	450	15	0	4	0	7	390	14	1	4	0	14	550	8	0	10	0	17
LB	20	235	0	0	0	2	12	185	8	0	0	5	112	361	0	0	0	29
H	0	0	75	2	1	9	0	1	83	0	1	8	0	4	67	0	8	21
N	0	0	0	28	0	7	6	0	0	61	1	12	0	0	0	67	0	7
B	0	0	2	0	101	2	0	0	1	0	140	3	0	0	0	2	219	3
I	4	11	5	2	3	12	9	8	7	4	3	8	26	4	2	13	15	25

Note: Rows contain the labels assigned by the majority of classifiers trained with the CM1 list, while columns contain labels assigned by the majority of classifiers trained with PAM50 list. In this table, LA corresponds to luminal A, LB corresponds to luminal B, H to HER2-enriched, N to normal-like, and B to basal-like. Labels marked as I refer to inconsistent assignments; situations where the classifiers did not achieve the majority on attributing a subtype label.

The Average Sensitivity statistic was used to characterise the average proportion of accurately labelled samples in each subtype. Considering the analysis with CM1 list, the measure was 0.76 ± 0.06 in the METABRIC discovery set and 0.64 ± 0.04 in the validation set; and with PAM50 list was 0.78 ± 0.07 and 0.65 ± 0.05, respectively. Likewise, the average sensitivity calculated for the ROCK test set was 0.67 ± 0.08 using the CM1 and 0.69 ± 0.08 with PAM50 list. A complete table containing the performance of all individual classification methods is available in the (**Supporting Information – Table 4.10 The performance of the classifiers using the CM1 list**

Table describing the performance of each classifier on the METABRIC discovery and validation sets, and ROCK test set using the CM1 list is summarised below (**Table 4.10s**). The original published file shows the percentage of correctly, incorrectly and not classified samples, Fleiss Kappa index, Cramer's V, Average Sensitivity, and other values for classification. The 24 classifiers from the Weka software suite are also listed. In addition, it contains the labels predicted by each classifier. Count of predicted labels was obtained with the consensus of the majority of classifiers.

Available online: doi:10.1371/journal.pone.0129711.s004

Supporting Information – Table 4.11

Table 4.11 The performance of the classifiers using the PAM50 list

Table describing the performance of each classifier on the METABRIC discovery and validation sets, and ROCK test set using the PAM50 list is summarised below (**Table 4.11s**).

The original published file shows the percentage of correctly, incorrectly and not classified samples, Fleiss Kappa index, Cramer's V, Average Sensitivity, and other values for classification. The 24 classifiers from the Weka software suite are also listed. In addition, it contains the labels predicted by each classifier. Count of predicted labels was obtained with the consensus of the majority of classifiers.

Available online: doi:10.1371/journal.pone.0129711.s005

Table 4.10 **and Table 4.11**).

The levels of agreement explained by interrater reliability metric Fleiss' kappa.

Fleiss' kappa was computed to assess the reliability of agreement between two raters, as displayed in **Table 4.7**. We initially compared the agreement *Among classifiers* which indicates the overall performance of classifiers alone. We then compared *Predicted vs Original*, that is, the agreement between subtypes assigned by the majority of classifiers using CM1 and PAM50 lists compared to the original PAM50 labels in the METABRIC discovery and validation sets, and ROCK test set. We also calculated the kappa between labels attributed by the majority of classifiers using both lists, *CM1 vs PAM50*. We refer to the Materials and Methods section for an interpretation of κ values. For instance, the high levels of agreement between two raters reflect more than what would be expected by chance.

Considering the agreement of the ensemble of classifiers, there was a *substantial agreement* in both METABRIC discovery and validation sets, and ROCK test set (**Table 4.7**). Fleiss' kappa was 0.73, 0.75 and 0.63 with the CM1 list for METABRIC discovery, validation and ROCK data sets, respectively. Values obtained with the PAM50 list were 0.72, 0.73 and 0.59, respectively. By comparing the subtypes predicted by the majority of classifiers and original PAM50 labels, there was an *almost perfect agreement* with CM1 (κ = 0.81) and PAM50 (κ = 0.84) lists in the discovery set. In the validation and ROCK sets, on the other hand, labels showed only a *moderate agreement* for both lists ($\kappa \sim 0.6$). Strikingly, the Fleiss' kappa between subtypes predicted using the CM1 and PAM50 lists (κ = 0.86, 0.83, and 0.80 in the METABRIC discovery, validation, and ROCK sets, respectively) revealed an *almost perfect agreement*. This statistical measure confirms our visual analysis of the contingency tables as they find strong relationship across the subtype labels in each data set. A detail of the agreement among classifiers by intrinsic subtype is shown in (**Supporting Information – Table 4.12**).

Table 4.7 Agreement of the 24 classifiers on assigning labels using Fleiss' kappa statistic

	METABRIC	**ROCK**

		Discovery	Validation	Test set
Among classifiers	**CM1**	0.73	0.75	0.63
	PAM50	0.72	0.73	0.59
Predicted vs. Original	**CM1**	0.81	0.60	0.59
	PAM50	0.84	0.62	0.64
CM1 vs. PAM50		0.86	0.83	0.80

Note: Rows entitled *Among classifiers* indicate agreement of classifiers. *Predicted vs. Original* shows the agreement between the mostly predicted and initial labels. Finally, the rows *CM1 vs. PAM50* contain the agreement between predicted labels using the CM1 and PAM50 lists.

The agreement according to the Adjusted Rand Index

The agreement between the different sample labelling was also scrutinised using the Adjusted Rand Index measure (**Table 4.8**). The values obtained with the CM1 list were 0.76 in the METABRIC discovery and 0.43 in the validation sets, and 0.45 in the ROCK test set. For PAM50 list, the values were 0.79, 0.46 and 0.51, respectively. Similar to Fleiss' kappa, the agreement between labels predicted with CM1 and PAM50 lists is higher than the agreement with the original labels. The Adjusted Rand Index values were 0.82, 0.79 and 0.64 for the three data sets, respectively. The numbers obtained with this measure also revealed remarkable concordance of CM1 and PAM50 lists assigned labels.

Table 4.8 Agreement measured by the Adjusted Rand Index between different labelling

	METABRIC		**ROCK**
	Discovery	**validation**	**test set**
CM1	0.76	0.43	0.45
PAM50	0.79	0.46	0.51
CM1-PAM50	0.82	0.79	0.64

Note: This table contains the agreement between the original and predicted labels of samples in the discovery and validation sets. *CM1-METABRIC* refers to agreement between the labels predicted by the majority of classifiers trained with the CM1 list and the original METABRIC labels; *PAM50-METABRIC* is the agreement between labels predicted by the majority of classifiers trained with the PAM50 list and original METABRIC labels; and *CM1-PAM50* is the agreement between predicted labels using both lists.

4.3.4 The Use of an Ensemble Learning with the CM1 List Improves the Subtype Distribution in the METABRIC and ROCK Data Sets

The number of samples in each original PAM50 subtype is markedly different across the METABRIC sets (**Figure 4.5**). In the discovery set, there is a clear abundance of luminal A

and B subtypes, precisely 73.62% of all samples. In contrast, the proportion of luminals in the validation set is only 48.14%. The ratio of luminal A to luminal B samples changed from 1.74 in the discovery to 1.14 in the validation set. However, when the CM1 or PAM50 lists are used in conjunction with the ensemble of classifiers, samples in the discovery and validation sets are more homogeneously distributed. The percentage of samples in the discovery set labelled as luminal A and B using CM1 and PAM50 lists are 73.53% and 73.72%, respectively. These proportions match the original number (73.62%). On the other hand, in the validation set the CM1 and PAM50 lists assigned a total of 64% and 63.19% luminal samples, against the 48.14% previously mentioned. The distribution of subtypes also becomes more similar to the discovery set. Likewise, ROCK test set also changed the pattern of class distribution after the performance of the ensemble of classifiers. The differences in class distributions might not be attributed to the randomisation procedure used by the studies as the performance of the ensemble of classifiers with both lists reconcile the distribution of subtypes.

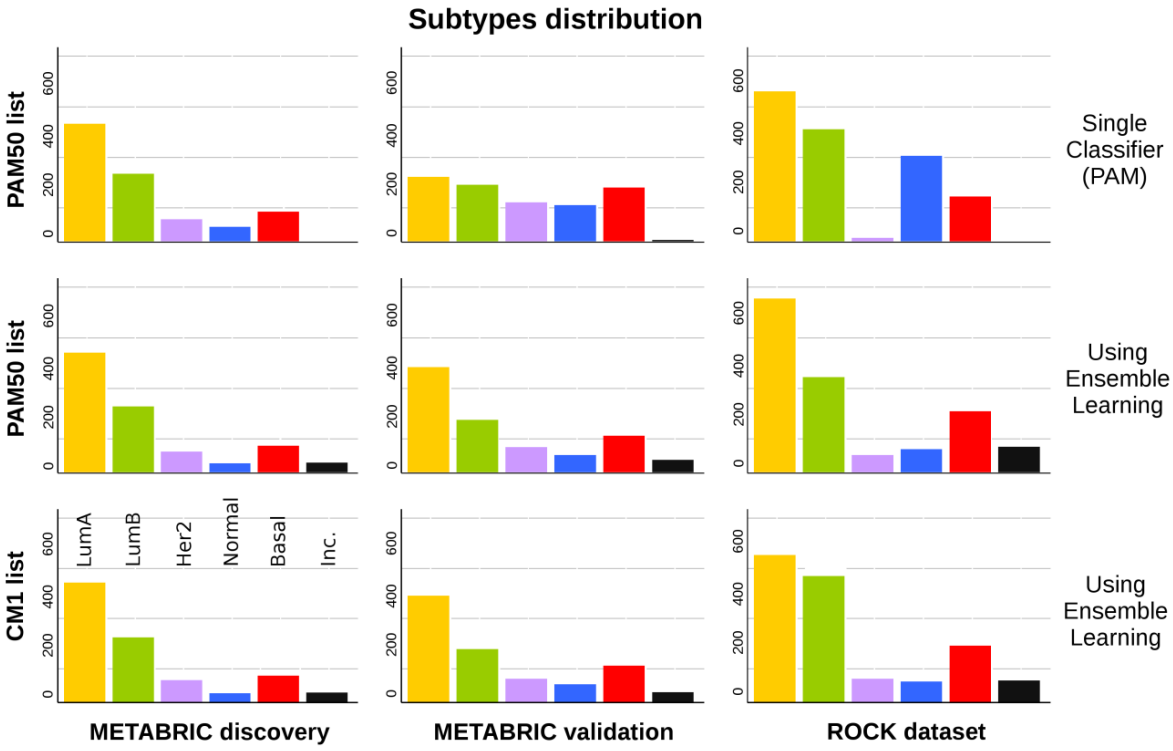


Figure 4.5 Class distribution in the METABRIC discovery and validation, and ROCK set

The bars represent the number of samples in each breast cancer subtype: luminal A (yellow), luminal B (green), HER2-enriched (purple), normal-like (blue) and basal-like- (red). In the first row, the labels refer to the original assignment using the PAM50 method. The following rows show the new labels attributed using an ensemble of 24 classifiers with PAM50 and CM1 lists, respectively. Samples were classified as inconsistent (black) if there was no consensus between the majorities of classifiers as to what should be the correct subtype.

We summarise the similarities and differences in subtypes distribution (graphically displayed in **Figure 4.5**) by computing the square root of the Jensen-Shannon divergence (Berretta & Moscato, 2010). This is a true metric of distance between probability distributions. Its plot in **Figure 4.6** shows the similarity between all possible pairs of data sets based on their distribution of subtype labels (**Supporting Information – Figure 4.9**). It can be observed that the original labels are the most divergent ones, especially in the METABRIC validation and ROCK test sets. The high similarity of samples distribution among subtypes based on the assignments with CM1 or PAM50 lists is evident. Such similarity was not expected for the ROCK set as the ensemble of classifiers was trained with METABRIC discovery (Illumina platform data) and tested in the ROCK set (Affymetrix platform data). The limited number of probes matching Illumina and Affymetrix in both lists (as described in Materials and Methods) seems not to affect the performance of the ensemble learning. Yet the divergences in the original class distributions might not be attributed to the randomisation procedure used by the consortium. These results point out to the relative strength and robustness of a set of classifiers compared to single methods to predict breast cancer subtype labels. They also indicate that there is an issue to be considered by researchers when using the original PAM50 labels from the METABRIC study for analysing data and building predictive models.

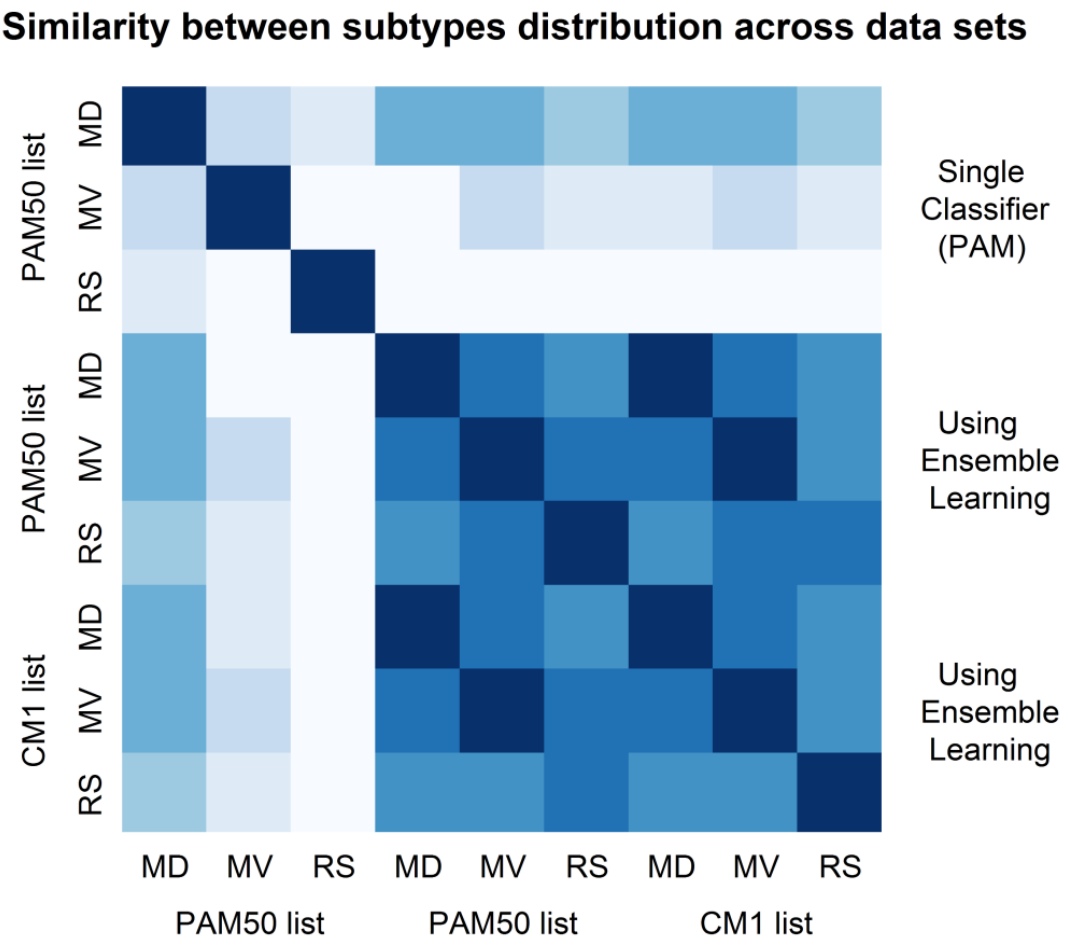


Figure 4.6 Similarity between subtypes distribution in the METABRIC discovery and validation sets, and in the ROCK set

The image shows the similarity between the subtypes distribution for METABRIC discovery (MD) and validation (MD) sets, and ROCK test set (RS). The similarity is measured using the square root of the Jensen-Shannon divergence. Darker shades represent more similar distributions, while lighter shades refer to divergent patterns. The diagonal shows the darkest colour as each data set is the closest to itself. According to this image, labels assigned using ensemble learning with CM1 and PAM50 lists are highly similar, and both exhibit lower levels of agreement with the original labels assigned using the PAM50 method.

4.3.5 Breast Cancer Intrinsic Subtypes Defined by Clinical Markers and Survival Curves

Given the heterogeneity among breast cancer patients and the intricate assignment of PAM50 labels in the original METABRIC data set, we further investigated whether significant

differences exist in the analysis of current clinical markers (ER, PR and HER2). **Figure 4.7, Figure 4.8** and **Figure 4.9** show, respectively, the distribution of the ER, PR and HER2 across intrinsic subtypes in the METABRIC discovery and validation sets, considering the original PAM50 labels and the labels assigned by ensemble of classifiers using CM1 and PAM50 lists. The new subtype labelling markedly improves the status of the clinical markers in the METABRIC data set. For instance, the ER marker distribution across subtypes shows an important decrease in the number of HER2-enriched and basal-like samples that are ER-positive according to the original PAM50 labels. The PR marker, likewise, varies the distribution when predicted labels based on the ensemble of classifiers using either CM1 and PAM50 list are compared with the original labels. HER2 amplification has a particular behaviour across all subtypes. Under the new subtype labels, the distribution of the three clinical markers becomes more consistent with what is expected according to the literature for each class: luminal A (ER+ and/or PR+, HER2-); luminal B (ER+ and/or PR+, HER2±); HER2-enriched (ER-, PR- and HER2+); and basal-like (ER-, PR-, HER2-) (de Kruijf et al., 2014).

Subsequently, we illustrate the survival curves for all breast cancer subtypes using Cox proportional hazards model, as described in Materials and Methods. The curves were plotted based on the original PAM50 labels and those assigned by the majority of classifiers. For generating the survival curves, we included the most relevant clinical variables as covariates: grade, size, age at diagnosis, number of lymph nodes positive, and ER status (immunohistochemistry) (Curtis et al., 2012). This analysis revealed different curves in the METABRIC discovery and validation sets (**Figure 4.10**). For instance, luminal B and basal-like subtypes show a similar pattern across data sets. Luminal A, HER2-enriched and normal-like, on the other hand, have a more consistent survival pattern when the CM1 and PAM50 lists are used in conjunction with the ensemble learning. Taken as a whole, the results of this section support the increased robustness of labels assigned by the ensemble of classifiers with the CM1 or PAM50 lists, and point out to inconsistencies in the original subtype assignment in the METABRIC study.

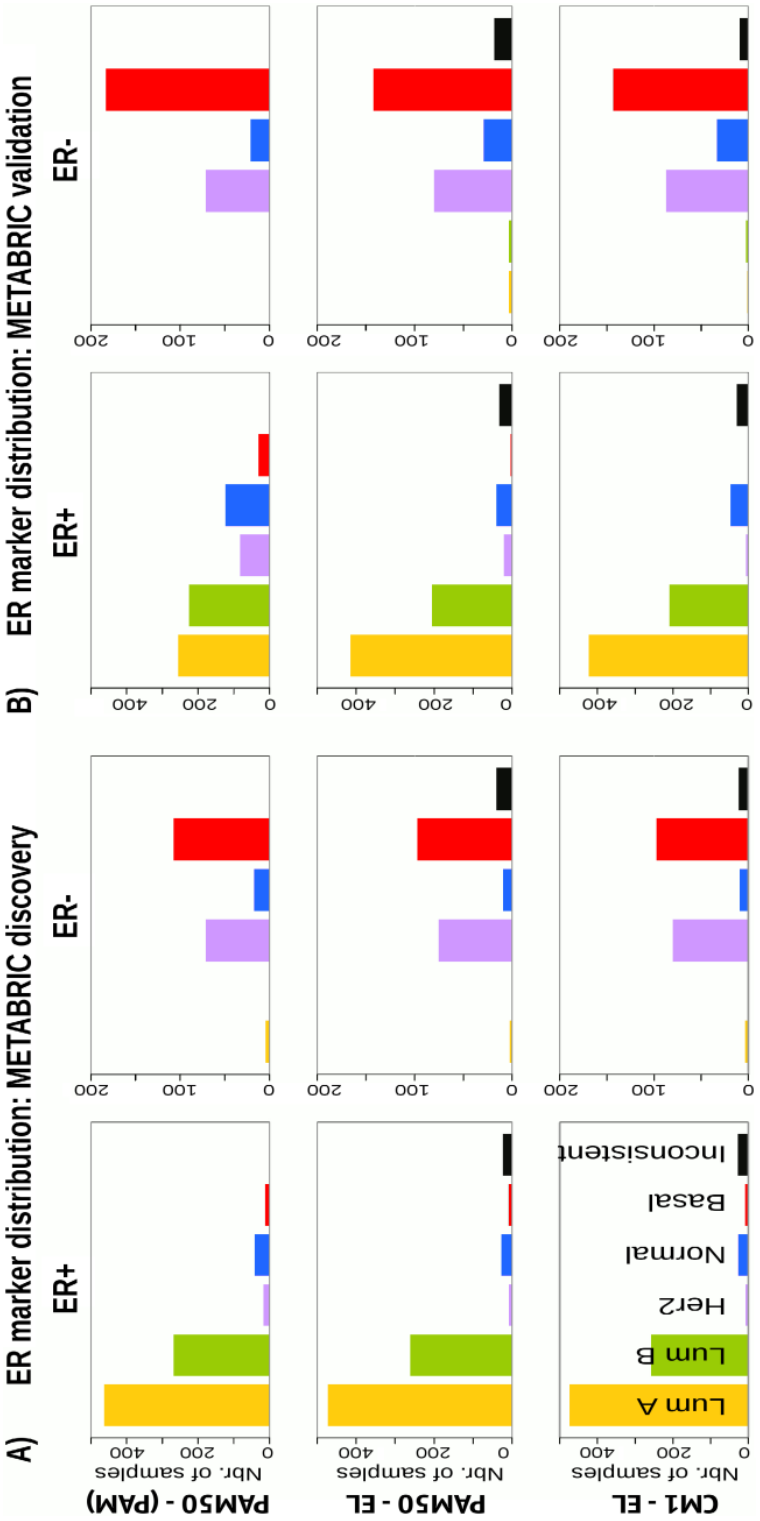


Figure 4.7 ER marker distribution across subtypes in the METABRIC data sets

(A) Discovery and (B) Validation. The bars represent the number of samples with ER positive and negative in the five intrinsic subtypes, based on the patients' clinical information. The top row is based on the original subtype labels obtained with the PAM50 list and a single classifier (PAM). Middle and bottom rows are based on the labels obtained by Ensemble Learning using the PAM50 and CM1 lists, respectively.

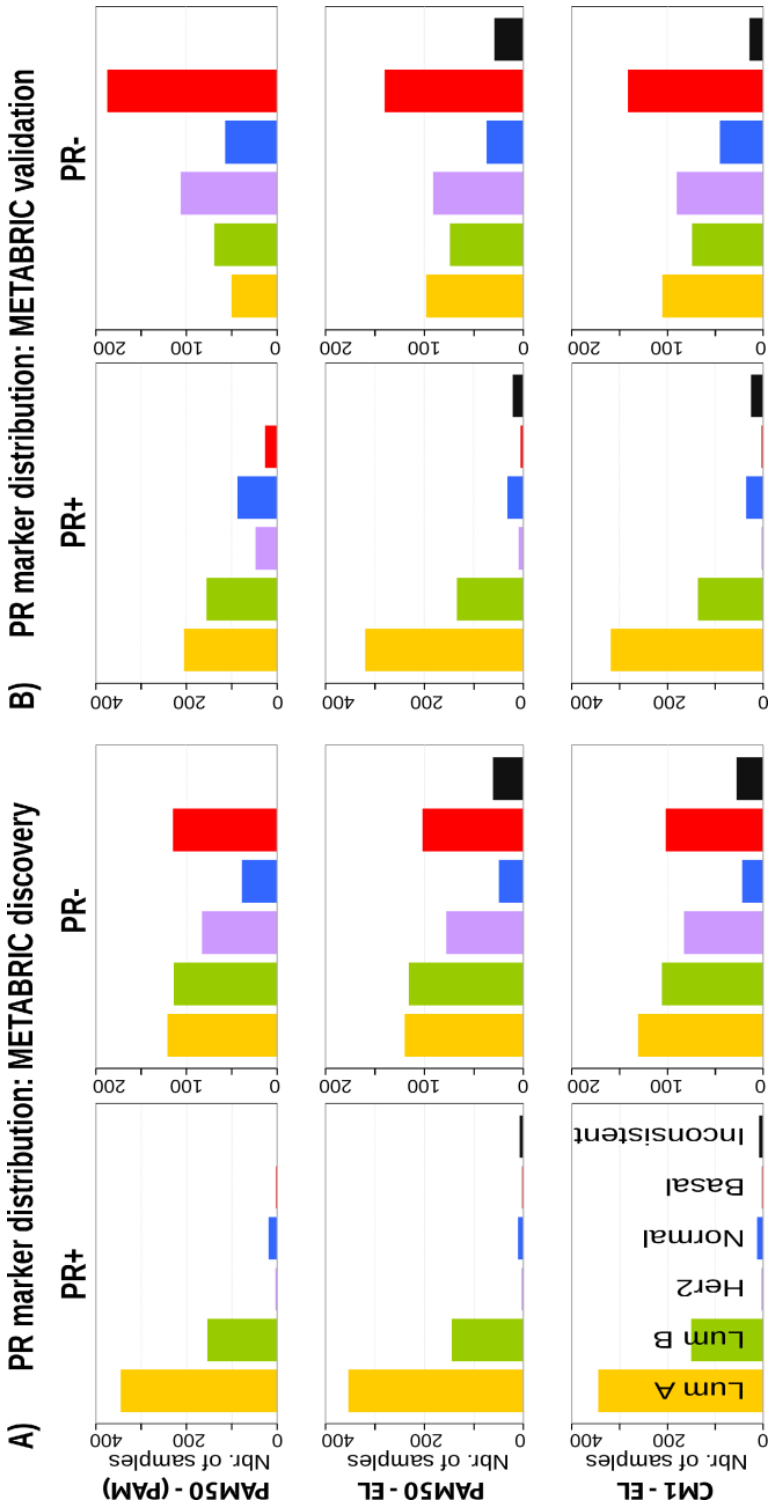


Figure 4.8 PR marker distribution across subtypes in the METABRIC data sets

(A) Discovery and (B) Validation. The bars represent the number of samples with PR positive and negative distributed in the five intrinsic subtypes, based on the patients' clinical information. The top row is based on the original subtype labels obtained with the PAM50 list and a single classifier (PAM). Middle and bottom rows are based on the labels obtained by Ensemble Learning using the PAM50 and CM1 lists, respectively.

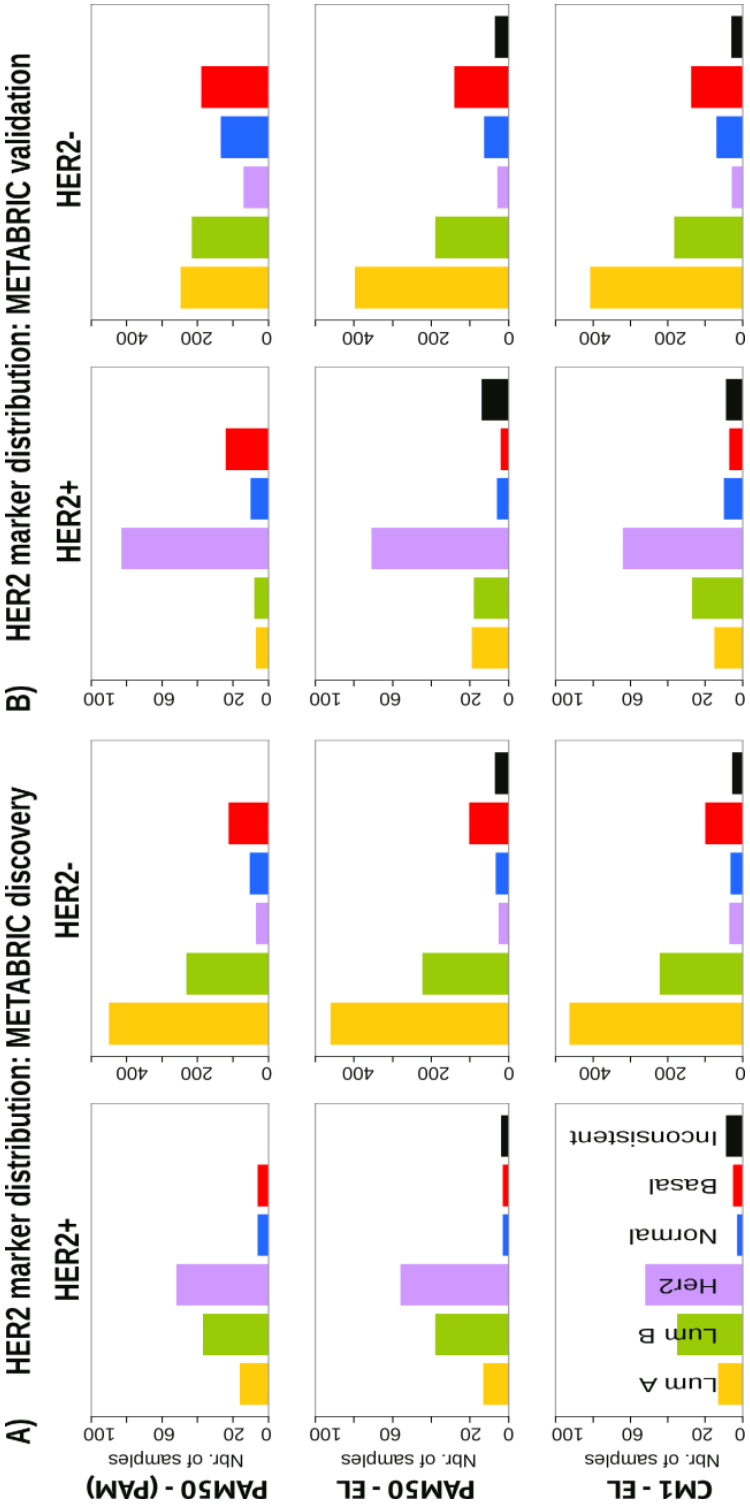


Figure 4.9 HER2 distribution across subtypes in the METABRIC data sets

Discovery and (B) Validation. The bars represent the number of samples with HER2 amplification (positive or negative) for each intrinsic subtype based on the patients' clinical information. The top row is based on the original subtype labels obtained with the PAM50 list and a single classifier (PAM). Middle and bottom rows are based on the labels obtained by Ensemble Learning using the PAM50 and CM1 lists, respectively.

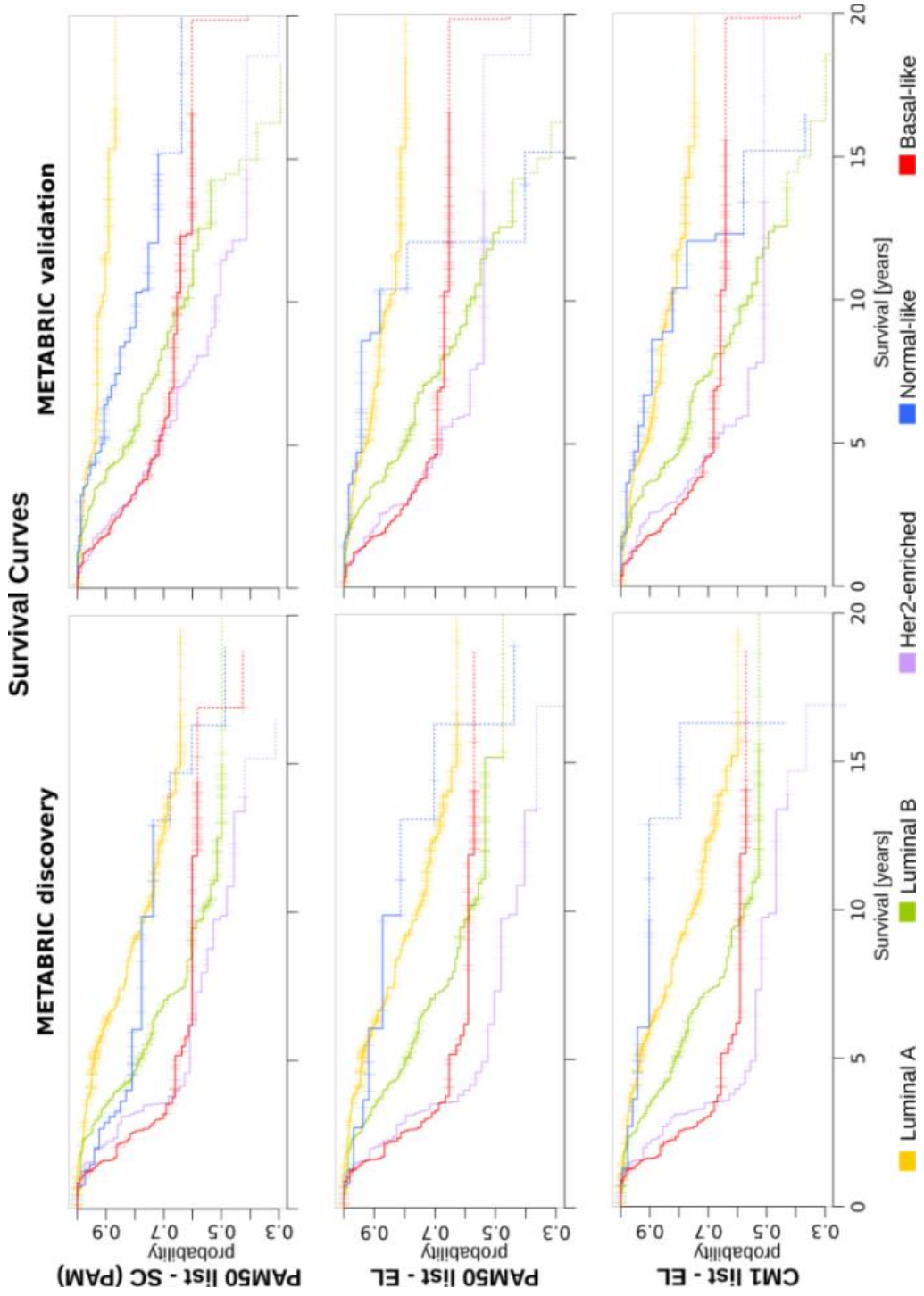


Figure 4.10 The survival curves for METABRIC discovery and validation sets

The survival curves for each breast cancer subtype are generated using Cox proportional hazards model based on the grade and size of the tumour, patient's age, number of lymph nodes positive and ER status. Each curve represents the survival probability at a certain time after the diagnosis. Ticks on the curve correspond to the observations of patients who are still alive, while drops indicate the death. The probability curves based on the last 10 observations are plotted in dash.

4.4 Discussion

In this study, we exposed the power of the CM1 list for improving the breast cancer subtype prediction in the METABRIC and ROCK data sets. The CM1 score portrayed 30 novel genes as potential biomarkers, along with 12 well-established markers shared between CM1 and PAM50 lists. The 42 biomarkers have a great potential to differentiate breast cancer intrinsic subtypes. Among them, *AGR3*, *HPN*, *ANKRD30A*, *AURKB*, *PROM1*, *VTCN1*, *CRYAB*, *CDK1*, *CDKN3*, *SERPINA3*, *SOX11*, *TRPV6*, *CLCA2*, *MUCL1*, *COL11A1*, *DARC*, *TFF3*, *IGF2BP3*, *IL33*, *SUSD3*, *PSAT1*, and *GABRP* are reported in different studies associated with breast cancer; however not in the context of subtype differentiation. Noteworthy, the CM1 list revealed a set of probes for which little literature exists in relation to breast cancer subtypes: *CDCA5*, *CCL15*, *COL17A1*, *GLYATL2*, *ROPN1*, LINC00993 and C6orf211. Their expression levels vary across different subtypes, and are yet a new avenue to be explored. We also emphasise the 12 common genes (*CEP55*, *ESR1*, *FOXA1*, *FOXC1*, *KRT17*, *MAPT*, *MELK*, *MMP11*, *NAT1*, *SFRP1*, *UBE2C*, and *UBE2T*) due to their important role for breast cancer disease and intrinsic subtyping.

Within the application of an ensemble of classifiers, CM1 and PAM50 lists showed concordant predictive power for disease subtyping. In fact, there was an almost perfect agreement between the labelling obtained with the majority of classifiers using both lists; however different from the original labels. In this study, we want to highlight the weakness of relying in a single method to assign subtypes labels, as opposed to the power and robustness of ensemble learning. We therefore discourage label assignments based on a single classifier and also suggest a thorough review of those intrinsic subtypes given the importance of such data sets to breast cancer research. The results indicate that there is an issue to be considered by researchers when using the original PAM50 labels for analysing data. The use of incorrect labels would lead to a plethora of misguided and misleading results by other investigators that use METABRIC or ROCK data sets.

In spite of luminals sharing the same origin and large molecular commonalities (Nguyen et al., 2008; Polyak, 2011), the ensemble of classifiers accurately predicted luminal samples in the METABRIC data set, and showed some ambiguity on assigning the subtype A or B for a small number of samples, especially in the ROCK data set. This may be a consequence of the reduced number of probes matching across Illumina and Affymetrix platforms. HER2-enriched notably improved label consistency in the ROCK data. Furthermore, the normal-like

tumours received more often contradictory and inaccurate subtype labelling among both data sets. The poor overall outcome for this subtype is supported by the discussion that normal-like is an artefact of sample processing with high contamination of normal breast tissue (Parker et al., 2009; Peppercorn et al., 2008; Weigelt; Baehner; et al., 2010); however, still crucial to be elucidated. Ultimately, the basal-like subtype maintained the classification with a unique profile, markedly divergent from other subtypes (Haibe-Kains et al., 2012; Mackay et al., 2011; Weigelt; Mackay; et al., 2010); even though this group has recently been partitioned into new fundamental classes (Herschkowitz et al., 2007; Prat et al., 2010).

4.5 Conclusion

Overall, the new intrinsic subtype labels based on the CM1 list and ensemble learning revealed more accurate distributions of clinical markers (ER, PR and HER2) and survival curves, when compared to the original PAM50 labels in the METABRIC cohort and ROCK test set. Interestingly, the CM1 list shows *ESR1* (ER) among the 42 probes, but brings other independent genes that are also relevant for overall predictions. Robust data sets like METABRIC have contributed to the understanding of breast cancer disease in terms of its molecular complexity and intrinsic alterations. The main limitation of the research in the field, nevertheless, is the uncertainty in the exact classification of intrinsic subtypes; over and above the discovery of molecular signatures and standard clinical biomarkers. Under consideration, a consistent taxonomy needs yet to be established prior to implementation in clinical practice. Additional research involving the genome, transcriptome, proteome, and epigenome, will lastly portray a true landscape of subtypes and contribute to breast cancer management.

4.6 References

Ambs, S. (2010). Prognostic significance of subtype classification for short-and long-term survival in breast cancer: survival time holds the key. *PLoS Med., 7*(5), e1000281.

Bastien, R. R., Rodríguez-Lescure, Á., Ebbert, M. T., Prat, A., Munárriz, B., Rowe, L., et al. (2012). PAM50 Breast Cancer Subtyping by RT-qPCR and Concordance with Standard Clinical Molecular Markers. *BMC Med. Genomics, 5*(1), 44.

Berretta, R., & Moscato, P. (2010). Cancer Biomarker Discovery: The Entropic Hallmark. *PLoS One, 5*(8), e12262.

Colombo, P.-E., Milanezi, F., Weigelt, B., & Reis-Filho, J. (2010). Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction. *Breast cancer research : BCR, 13*(212), 1-15.

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature, 486*(7403), 346-352.

de Kruijf, E. M., Bastiaannet, E., Rubertá, F., de Craen, A. J. M., Kuppen, P. J. K., Smit, V. T. H. B. M., et al. (2014). Comparison of frequencies and prognostic effect of molecular subtypes between young and elderly breast cancer patients. *Mol. Oncol., 8*(5), 1014-1025.

Dowsett, M., Sestak, I., Lopez-Knowles, E., Sidhu, K., Dunbier, A. K., Cowens, J. W., et al. (2013). Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. *J. Clin. Oncol., 31*(22), 2783-2790.

Dunning, M. J., Curtis, C., Barbosa-Morais, N. L., Caldas, C., Tavaré, S., & Lynch, A. G. (2010). The importance of platform annotation in interpreting microarray data. *Lancet Oncol., 11*(8), 717.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull., 76*(5), 378-382.

Fleiss, J. L., Levin, B., & Paik, M. C. (2004). The Measurement of Interrater Agreement *Statistical Methods for Rates and Proportions* (pp. 598-626). New York: John Wiley & Sons, Inc.

Gómez-Ravetti, M., & Moscato, P. (2008). Identification of a 5-Protein Biomarker Molecular Signature for Predicting Alzheimer's Disease. *PLoS One, 3*(9), e3111.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., et al. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *J. Natl. Cancer Inst., 104*(4), 311-325.

Herschkowitz, J. I., Simin, K., Weigman, V. J., Mikaelian, I., Usary, J., Hu, Z., et al. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol., 8*(5), R76.

Hu, Z., Fan, C., Oh, D. S., Marron, J., He, X., Qaqish, B. F., et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics, 7*(1), 96.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193-218.

Kalbfleisch, J. D., & Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*: John Wiley & Sons.

Kelly, C. M., Bernard, P. S., Krishnamurthy, S., Wang, B., Ebbert, M. T., Bastien, R. R., et al. (2012). Agreement in risk prediction between the 21-gene recurrence score assay (Oncotype DX(R)) and the PAM50 breast cancer intrinsic Classifier in early-stage estrogen receptor-positive breast cancer. *Oncologist, 17*(4), 492-498.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., et al. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest., 121*(7), 2750-2767.

Liebetrau, A. M. (1983). *Measures of association* (Vol. 32). Beverly Hills, CA: SAGE Publications, Inc.

Mackay, A., Weigelt, B., Grigoriadis, A., Kreike, B., Natrajan, R., A'Hern, R., et al. (2011). Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. *J. Natl. Cancer Inst., 103*(8), 662-673.

Marsden, J., Budden, D., Craig, H., & Moscato, P. (2013). Language Individuation and Marker Words: Shakespeare and His Maxwell's Demon. *PLoS One, 8*(6), e66813.

Nguyen, P. L., Taghian, A. G., Katz, M. S., Niemierko, A., Abi Raad, R. F., Boon, W. L., et al. (2008). Breast cancer subtype approximated by estrogen receptor, progesterone receptor, and HER-2 is associated with local and distant recurrence after breast-conserving therapy. *J. Clin. Oncol., 26*(14), 2373-2378.

Nielsen, T. O., Parker, J. S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., et al. (2010). A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin. Cancer Res., 16*(21), 5222-5232.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol., 27*(8), 1160-1167.

Peppercorn, J., Perou, C. M., & Carey, L. A. (2008). Molecular subtypes in breast cancer evaluation and management: divide and conquer. *Cancer Invest., 26*(1), 1-10.

Perou, C. M., Parker, J. S., Prat, A., Ellis, M. J., & Bernard, P. S. (2010). Clinical implementation of the intrinsic subtypes of breast cancer. *Lancet Oncol., 11*(8), 718-719.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature, 406*(6797), 747-752.

Polyak, K. (2011). Heterogeneity in breast cancer. *J. Clin. Invest., 121*(10), 3786-3788.

Portier, B. P., Gruver, A. M., Huba, M. A., Minca, E. C., Cheah, A. L., Wang, Z., et al. (2012). From morphologic to molecular: established and emerging molecular diagnostics for breast carcinoma. *N Biotechnol, 29*(6), 665-681.

Prat, A., Ellis, M. J., & Perou, C. M. (2012). Practical implications of gene-expression-based assays for breast oncologists. *Nat. Rev. Clin. Oncol., 9*(1), 48-57.

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., et al. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res., 12*(5), R68.

Reis-Filho, J. S., & Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet, 378*(9805), 1812-1823.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A., 98*(19), 10869-10874.

Sørlie, T., Tibshirani, R., Parker, J. S., Hastie, T., Marron, J. S., Nobel, A., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U. S. A., 100*(14), 8418-8423.

Sotiriou, C., & Pusztai, L. (2009). Gene-Expression Signatures in Breast Cancer. *N. Engl. J. Med., 360*(8), 790-800.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer Science & Business Media.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U. S. A., 99*(10), 6567-6572.

Ur-Rehman, S., Gao, Q., Mitsopoulos, C., & Zvelebil, M. (2013). ROCK: a resource for integrative breast cancer data analysis. *Breast Cancer Res. Treat., 139*(3), 907-921.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*(6871), 530-536.

van De Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med., 347*(25), 1999-2009.

Vinh, N. X., Epps, J., & Bailey, J. (2009). *Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?* Paper presented at the Proceedings of the 26th Annual International Conference on Machine Learning.

Weigelt, B., Baehner, F. L., & Reis-Filho, J. S. (2010). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *J. Pathol., 220*(2), 263-280.

Weigelt, B., Mackay, A., A'hern, R., Natrajan, R., Tan, D. S. P., Dowsett, M., et al. (2010). Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol., 11*(4), 339-349.

Weigelt, B., Pusztai, L., Ashworth, A., & Reis-Filho, J. S. (2012). Challenges translating breast cancer gene signatures into the clinic. *Nat. Rev. Clin. Oncol., 9*(1), 58-64.

Weigelt, B., & Reis-Filho, J. S. (2009). Histological and molecular types of breast cancer: is there a unifying taxonomy? *Nat. Rev. Clin. Oncol., 6*(12), 718-730.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann.

4.7 Supporting Information

Supporting Information – Table 4.9

Table 4.9 The CM1 score calculated for each breast cancer subtype

Table listing the CM1 score used to rank the set of 48803 probes for each of the five breast cancer subtypes in the METABRIC discovery data set. In each case, we selected the top 10 highly discriminative probes (5 with the greatest positive CM1 score values – indicating up-regulated probes, and 5 with the smallest negative values – down-regulated).

Available online: doi:10.1371/journal.pone.0129711.s003

Supporting Information – Table 4.10 The performance of the classifiers using the CM1 list

Table describing the performance of each classifier on the METABRIC discovery and validation sets, and ROCK test set using the CM1 list is summarised below (**Table 4.10s**). The original published file shows the percentage of correctly, incorrectly and not classified samples, Fleiss Kappa index, Cramer's V, Average Sensitivity, and other values for classification. The 24 classifiers from the Weka software suite are also listed. In addition, it contains the labels predicted by each classifier. Count of predicted labels was obtained with the consensus of the majority of classifiers.

Available online: doi:10.1371/journal.pone.0129711.s004

Supporting Information – Table 4.11

Table 4.11 The performance of the classifiers using the PAM50 list

Table describing the performance of each classifier on the METABRIC discovery and validation sets, and ROCK test set using the PAM50 list is summarised below (**Table 4.11s**). The original published file shows the percentage of correctly, incorrectly and not classified samples, Fleiss Kappa index, Cramer's V, Average Sensitivity, and other values for classification. The 24 classifiers from the Weka software suite are also listed. In addition, it contains the labels predicted by each classifier. Count of predicted labels was obtained with the consensus of the majority of classifiers.

Available online: doi:10.1371/journal.pone.0129711.s005

Table 4.10

Table 4.10 The performance of the classifiers using the CM1 list

Table describing the performance of each classifier on the METABRIC discovery and validation sets, and ROCK test set using the CM1 list is summarised below (**Table 4.10s**). The original published file shows the percentage of correctly, incorrectly and not classified samples, Fleiss Kappa index, Cramer's V, Average Sensitivity, and other values for classification. The 24 classifiers from the Weka software suite are also listed. In addition, it contains the labels predicted by each classifier. Count of predicted labels was obtained with the consensus of the majority of classifiers.

Available online: doi:10.1371/journal.pone.0129711.s004

Supporting Information – Table 4.11

Table 4.11 The performance of the classifiers using the PAM50 list

Table describing the performance of each classifier on the METABRIC discovery and validation sets, and ROCK test set using the PAM50 list is summarised below (**Table 4.11s**). The original published file shows the percentage of correctly, incorrectly and not classified samples, Fleiss Kappa index, Cramer's V, Average Sensitivity, and other values for classification. The 24 classifiers from the Weka software suite are also listed. In addition, it contains the labels predicted by each classifier. Count of predicted labels was obtained with the consensus of the majority of classifiers.

Available online: doi:10.1371/journal.pone.0129711.s005

Table 4.10 Summary performance of the classifiers using the CM1 list

		Cramer's V		
Classifiers	**Type**	**METABRIC Discovery**	**METABRIC Validation**	**ROCK Validation**
Bayes	BayesNet	0.77	0.68	0.65
	NaiveBayes	0.78	0.69	0.65
	NaiveBayesUpdateable	0.77	0.69	0.65
Functions	Logistic	0.71	0.65	0.62
	MultilayerPerceptron	0.78	0.66	0.58
	SimpleLogistic	0.81	0.66	0.59
	SMO	0.80	0.66	0.62
Lazy	IBk	0.68	0.62	0.58
	KStar	0.67	0.57	0.43
Meta	AttributeSelectedClassifier	0.69	0.62	0.56
	Bagging	0.75	0.63	0.55

	ClassificationViaRegression	0.76	0.65	0.54
	LogitBoost	0.75	0.63	0.56
	MultiClassClassifier	0.72	0.63	0.58
	RandomCommittee	0.75	0.63	0.59
Rules	DecisionTable	0.59	0.56	0.52
	JRip	0.66	0.60	0.48
	PART	0.73	0.60	0.58
Trees	HoeffdingTree	0.78	0.69	0.65
	J48	0.70	0.61	0.56
	LMT	0.81	0.66	0.59
	RandomForest	0.75	0.61	0.57
	RandomTree	0.63	0.58	0.44
	REPTree	0.70	0.63	NA

Table 4.11 Summary performance of the classifiers using the PAM50 list

Classifiers	Type	Cramer's V		ROCK
		METABRIC Discovery	METABRIC Validation	
bayes	BayesNet	0.78	0.69	0.67
	NaiveBayes	0.79	0.69	0.62
	NaiveBayesUpdateable	0.79	0.69	0.62
functions	Logistic	0.74	0.67	0.57
	MultilayerPerceptron	0.85	0.70	0.64
	SimpleLogistic	0.85	0.69	0.62
	SMO	0.85	0.68	0.64
lazy	IBk	0.73	0.66	0.62
	KStar	0.68	0.58	0.50
meta	AttributeSelectedClassifier	0.70	0.62	0.58
	Bagging	0.75	0.63	0.59
	ClassificationViaRegression	0.79	0.64	0.53
	LogitBoost	0.78	0.65	0.58
	MultiClassClassifier	0.75	0.61	0.58
	RandomCommittee	0.78	0.62	0.59
rules	DecisionTable	0.61	0.55	0.47
	JRip	0.68	0.64	0.51
	PART	0.70	0.62	0.52
trees	HoeffdingTree	0.78	0.69	0.63
	J48	0.74	0.61	0.55
	LMT	0.84	0.69	0.62
	RandomForest	0.77	0.62	0.58
	RandomTree	0.66	0.58	0.48
	REPTree	0.69	0.61	0.57

Supporting Information – Table 4.12

Table 4.12 The agreement between sample labelling with Fleiss' Kappa measure and the Jensen-Shannon divergence of two probability distributions

	METABRIC Discovery				
	Classifiers CM1	Classifiers PAM50	Majority CM1	Majority PAM50	CM1 PAM50
Luminal A	0.73	0.73	0.87	0.89	0.92
Luminal B	0.74	0.73	0.88	0.88	0.91
Her2	0.70	0.68	0.80	0.89	0.97
Normal	0.49	0.48	0.70	0.82	0.90
Basal	0.86	0.85	0.93	0.95	0.98
Overall	0.73	0.72	0.81	0.84	0.86
	METABRIC Validation				
	Classifiers CM1	Classifiers PAM50	Majority CM1	Majority PAM50	CM1 PAM50
Luminal A	0.76	0.74	0.60	0.62	0.92
Luminal B	0.75	0.72	0.65	0.69	0.89
Her2	0.70	0.66	0.55	0.62	0.93
Normal	0.64	0.59	0.48	0.48	0.91
Basal	0.85	0.86	0.80	0.84	0.99
Overall	0.75	0.73	0.60	0.62	0.83
	ROCK test set				
	Classifiers CM1	Classifiers PAM50	Majority CM1	Majority PAM50	CM1 PAM50
Luminal A	0.60	0.58	0.63	0.70	0.78
Luminal B	0.62	0.63	0.68	0.73	0.76
Her2	0.58	0.44	0.16	0.24	0.82
Normal	0.40	0.37	0.31	0.38	0.77
Basal	0.81	0.79	0.79	0.76	0.92
Overall	0.63	0.59	0.59	0.64	0.80

Note: Table containing the Fleiss' Kappa agreement of labels for the METABRIC discovery and validation sets, and ROCK test set. It shows the overall agreement (average values) *Among classifiers* using CM1 and PAM50 lists, as well as the agreement for each subtype. The *predicted—original* are described in the table and contain the agreement between the mostly predicted and initial labels of samples; whereas the *CM1—PAM50* show agreement between labels assigned by the majority of classifiers using CM1 and PAM50 lists.

Supporting Information –Table 4.13

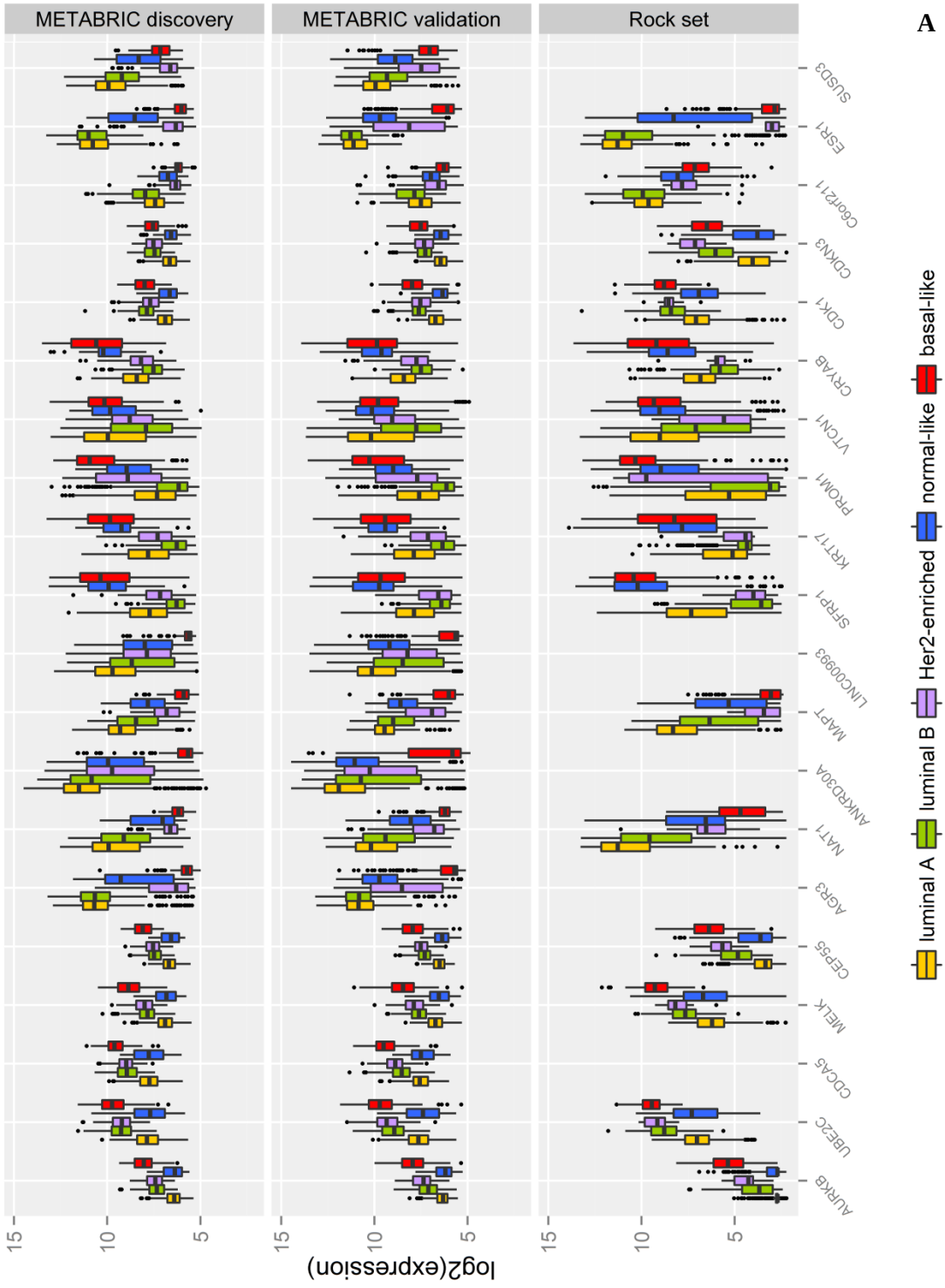
Table 4.13 The Jensen-Shannon divergence of two probability distributions

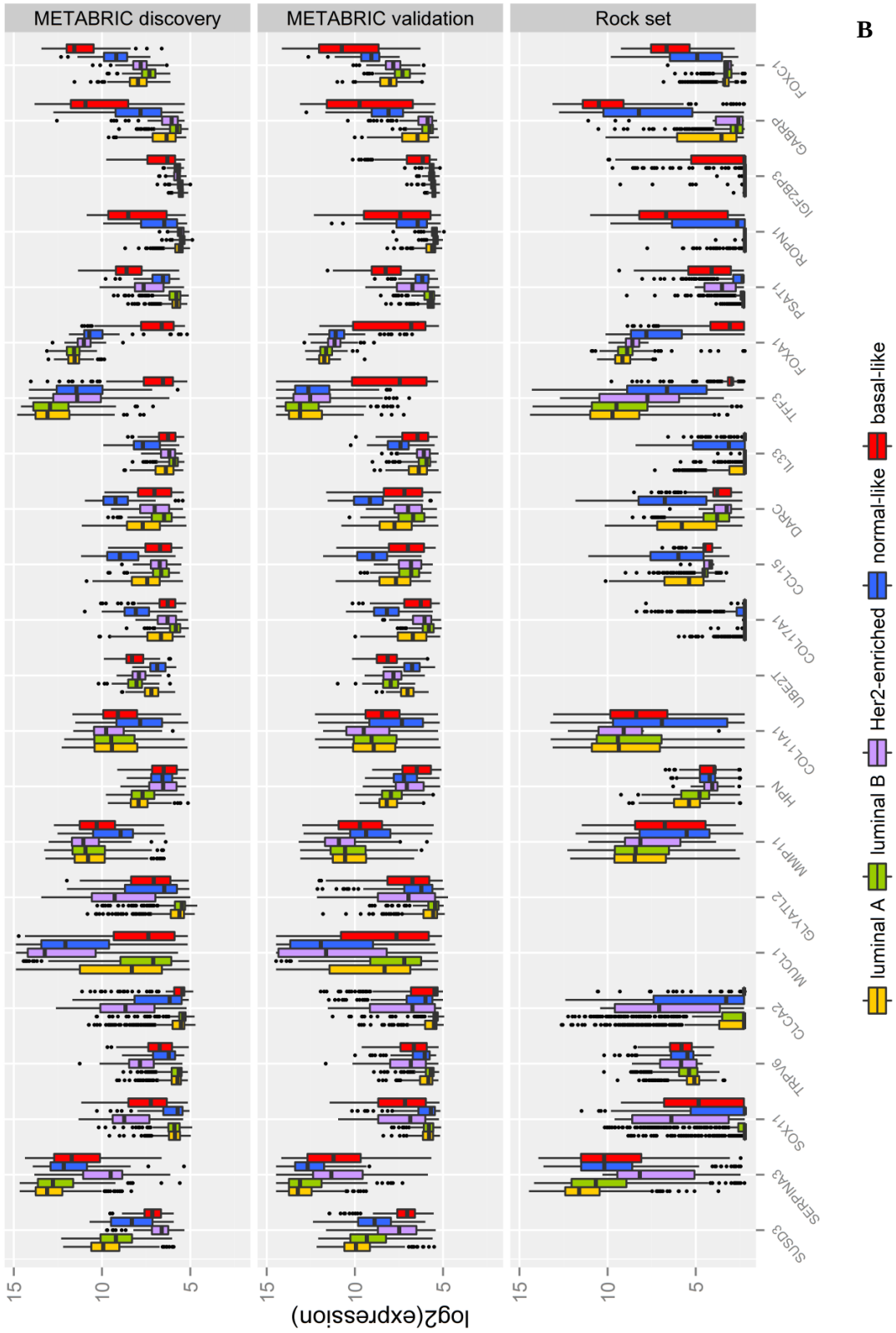
		Original labels			**CM1 list**			**PAM50 list**			
		discovery	validation	ROCK set	discovery	validation	ROCK set	discovery	validation	ROCK set	
PAM50 list	discovery	0	0.20	0.21	0.12	0.13	0.16	0.12	0.15	0.17	**Single Classifier (PAM)**
	validation	0.20	0	0.25	0.24	0.18	0.22	0.24	0.18	0.23	
	ROCK set	0.21	0.25	0	0.27	0.24	0.24	0.26	0.25	0.24	
CM1 list	discovery	0.12	0.24	0.27	0	0.08	0.10	0.01	0.01	0.10	**Ensenble Learning**
	validation	0.130	0.18	0.24	0.08	0	0.10	0.08	0.02	0.08	
	ROCK set	0.157	0.22	0.24	0.10	0.10	0	0.09	0.10	0.07	
PAM50 list	discovery	0.123	0.24	0.26	0.01	0.08	0.09	0	0.09	0.09	**Ensenble Learning**
	validation	0.145	0.18	0.25	0.09	0.02	0.10	0.09	0	0.08	
	ROCK set	0.167	0.23	0.24	0.10	0.08	0.07	0.09	0.08	0	

Note: The file also has the Jensen-Shannon divergence between two probability distributions. Numbers represent the similarity between subtypes' distribution for METABRIC discovery and validation sets, and ROCK test set. The similarity is measured using the square root of the Jensen-Shannon divergence.

Supporting Information – Figure 4.11

Figure 4.11 The mRNA log2 normalised expression values of 42 probes (A and B) in the CM1 list across the five intrinsic subtypes in the METABRIC discovery and validation, and ROCK




B
METABRIC discovery
METABRIC validation
Rock set
FOXC1
GABRP
IGF2BP3
ROPN1
PSAT1
FOXA1
TFF3
IL33
DARC
CCL15
COL17A1
UBE2T
COL11A1
HPN
MMP11
GLYATL2
MUCL1
CLCA2
TRPV6
SOX11
SERPINA3
SUSD3
15
10
5
log2(expression)
basal-like
normal-like
Her2-enriched
luminal B
luminal A

Supporting Information – Text 4.1

Text 4.1. CM1 list and literature review

The document shows the CM1 probe list along with an extensive literature review. The 42 CM1 biomarkers revealed a great potential to differentiate breast cancer intrinsic subtypes in the METABRIC and ROCK data sets. The 30 novel markers and 12 well-established genes vary the expression levels across different subtypes. The vast majority has been associated with breast cancer disease, either included or not in the subtyping context.

Illumina Probe / Gene Symbol	**Gene and Protein Review**
ILMN_1684217 AURKB	Aurora kinase B (*AURKB*; also known as *Aik2, AIK2, AIM1, AIM-1, ARK2, ARK-2, AurB, aurkb-sv1, aurkb-sv2, Aurora/IPL1-related kinase 2, Aurora-and IPL1-like midbody-associated protein 1, Aurora kinase B, Aurora-related kinase 2, IPL1, Serine/threonine-protein kinase 12, Serine/threonine-protein kinase aurora-B, STK-1, STK12, STK5*) encodes a protein member of the aurora kinase subfamily of serine/threonine kinases that function as a regulator of the centrosome cycle and mitotic spindle assembly. The protein AURKB and interacting proteins play an important role in chromosome condensation, segregation and cytokinesis and, consequently, in ploidy maintenance during cell division. Mitotic deregulations may contribute significantly to cell division errors and development of aggressive tumour cells (Ciriello et al., 2013; Hegyi et al., 2012). AURKB signalling was also linked to breast cancer and associated to poor prognosis (Ahn et al., 2013). This gene has been a target of inactivation via different studies and mechanisms (Arbitrario et al., 2010; Bush et al., 2013; Fiskus et al., 2012; Gully et al., 2010; Hardwicke et al., 2009; Kalous et al., 2013; Romanelli et al., 2012; Sanchez-Bailon et al., 2012; Soncini et al., 2006; T. Ueki et al., 2008).
ILMN_1683450 CDCA5	Cell division cycle associated 5 (*CDCA5*; also known as *Cell division cycle-associated protein 5, MGC16386, p35, Sororin, SORORIN*) is important for sister chromatid cohesion during mitosis, stabilizing proper chromatin association during G2 phase. The protein is also needed for efficient repair of DNA double-strand breaks and for stable presence of normal amounts of chromatin-bound cohesin population (Mertsch et al., 2008; W. Zhang et al., 2010). Carretero et al (Carretero et al., 2013) reported that "the reduced accumulation of AURKB at the inner centromere in cells that lack PDS5B impairs its error correction function, promoting chromosome mis-segregation and aneuploidy". Although systematic studies showed the up-regulation of this gene in a great majority of lung cancers, its involvement in breast cancer disease requires further investigation. The protein CDCA5 confers a potential diagnostic molecule and therapeutic target for promising strategies in new drug development (Nguyen et al., 2010).
ILMN_1747016 CEP55	Centrosomal protein 55kDa (*CEP55*; also known as *C10orf3, Centrosomal protein of 55 kDa, Cep55, CT111, FLJ10540, Up-regulated in colon cancer 6, URCC6*) encodes a mitotic phosphoprotein that acts in mitotic exit and cytokinesis. Both down- and up-regulation of *CEP55* causes a cytokinesis defect (Murphy et al., 2010). The gene expression variance, nonetheless, is not affected by hormone receptors such as oestrogen and progesterone or expression patterns of *ERBB2*. The centrosomal protein is detected in a wide variety of tumour cell lines and is considered as a novel breast tumour-associated antigen (Inoda et al., 2009). It was also reported that *CEP55* have genomic alternation in a comparison of pre-invasive ductal carcinoma in situ (DCIS) to invasive ductal carcinoma (IDC) by Colak et al. (Colak et al., 2013) and it was predictive of prognosis in ER-positive patients (Martin et al., 2008).

Illumina Probe / Gene Symbol	**Gene and Protein Review**
ILMN_2212909 MELK	The maternal embryonic leucine zipper kinase (*MELK*; also known as *hMELK, hPK38, HPK38, KIAA0175, Maternal embryonic leucine zipper kinase, Protein kinase PK38*) acts as a regulator of various processes such as cell cycle control, self-renewal of stem cells, apoptosis, and splicing regulation. The encoded protein physically interacts, phosphorylates and inhibits BCL2L14, repressing a pro-apoptotic member of the Bcl-2 family. The protein also mediates phosphorylation of CDC25B, regulating the entry into mitosis. In addition, MELK inhibits the spliceosome assembly during mitosis by phosphorylating ZNF622 and contributes to induce other apoptosis signalling regulation. The protein kinase is a promising molecular target for the treatment of breast cancer (Lin et al., 2007) as the up-regulation is linked to poor prognosis (Agnati et al., 2007; Canevari et al., 2013; Hebbard et al., 2010; Mahasenan & Li, 2012; Pickard et al., 2009). Finally, it has been suggested that paclitaxel may attenuate the expression of MELK (Warsow et al., 2013).
ILMN_1714730 UBE2C	Ubiquitin-conjugating enzyme E2C (*UBE2C*; also known as *dJ447F3.2, UbcH10, UBCH10, Ubiquitin carrier protein C, Ubiquitin-conjugating enzyme E2 C, Ubiquitin-protein ligase C*) encodes a member of the E2 ubiquitin-conjugating enzyme family. The ubiquitin modification in proteins is an important cellular mechanism of homoeostasis and fate (Loussouarn et al., 2009). UBE2C is required for cell cycle progression and checkpoint control through targeted degradation of short-lived proteins, the mitotic cyclins. Aberrations in this pathway is implicated in cancer progression and, importantly, in the pathogenesis of breast cancer (Psyrri et al., 2012; Rawat et al., 2013). The *UBE2C* up-regulation is also normally linked to high tumour grade and poor prognosis (Parris et al., 2014; Taylor et al., 2010).
ILMN_1796059 ANKRD30A	Ankyrin repeat domain 30A (*ANKRD30A*; also known as *Ankyrin repeat domain-containing protein 30A, NY-BR-1, Serologically defined breast cancer antigen NY-BR-1*) is an antigen expressed in mammary glands, primary and metastatic breast carcinomas (Jäger et al., 2007; Varga et al., 2006; Woodard et al., 2011). Interestingly, *ANKRD30A* is almost expressed exclusively in breast epithelium; with exception of testis and sweat gland tumours. Despite the insufficient knowledge about the biology and function of the gene, the tissue specificity may be useful for the diagnosis of breast carcinomas (Balafoutas et al., 2013; Giger et al., 2010; Jäger et al., 2007; Seil et al., 2007); and a potential target for treatment (immunotherapy) (J.-P. Theurillat et al., 2007; J. P. Theurillat et al., 2008).
ILMN_1651329 LINC00993	The long intergenic non-protein coding RNA 993 (LINC00993) matches a region in the chromosome 10 very close to *ANKRD30A*; and contains a SNP (rs77587276) variant. The region requires further investigation as it covers relevant probes associated with breast cancer disease; markedly, up-regulated in luminal subtype.
ILMN_2310814 MAPT	The microtubule-associated protein tau (*MAPT*; also known as *DDPAC, FLJ31424, FTDP-17, MAPTL, MGC138549, Microtubule-associated protein tau, MSTD, MTBT1, MTBT2, Neurofibrillary tangle protein, Paired helical filament-tau, PHF-tau, PPND, tau, TAU*) undergoes alternative splicing, originating several mRNA transcripts. The isoforms differ by having variant conserved repeat motifs in the microtubule-binding domain, and insertions in the N-terminal projection domain. Although the function of each isoform is unknown, the protein binds to both the outer and inner surfaces of microtubules, organizing the tubulin assembly and microtubule stabilisation (Ikeda et al., 2010). In breast cancer, *MAPT* expression is high in ER-positive low grade compared to ER-negative high grade tumours (Valet et al., 2013). Additionally, the gene up-regulation is correlated with favourable prognosis (Kotoula et al., 2013) and, at the same time, associated with resistance to taxanes, paclitaxel and docetaxel (Fountzilas et al., 2013; Ikeda et al., 2010; Mihály et al., 2013; Spicakova et al., 2010; Tanaka et al., 2009; K. Wang et al., 2013).

Illumina Probe / Gene Symbol	**Gene and Protein Review**
ILMN_1728787 AGR3	The anterior gradient 3 (*AGR3*; also known as *AG3, AG-3, Anterior gradient protein 3 homolog, BCMP11, Breast cancer membrane protein 11, HAG3, hAG-3, PDIA18, UNQ642/PRO1272*) functionality has been defined in breast cancer cells as involved in hormone responsiveness, cell adhesion, migration, and metastasis. This gene encodes a membrane protein with a potential role in tumorigenesis by interacting with metastasis-associated genes (Fletcher et al., 2003; Persson et al., 2005).
ILMN_1688071 NAT1	The enzyme encoded by N-acetyltransferase 1 (*NAT1*; also known as *AAC1, Arylamide acetylase 1, Arylamine N-acetyltransferase 1, MNAT, Monomorphic arylamine N-acetyltransferase, N-acetyltransferase type 1, NAT-1, NATI*) acts metabolizing drugs and other xenobiotics, and functions in folate catabolism (S. J. Kim; Kang; et al., 2008; Sim et al., 2008). Kim et al. (2008) reported the hypomethylation of the *NAT1* promoter region resulting in aberrant mRNA expression levels, with overexpression of the gene in breast carcinomas. Likewise, new insights into the associations of SNPs in the coding and control regions of *NAT1* have been described and suggested as a potential susceptibility biomarker for the disease (Sim et al., 2008).
ILMN_1729216 CRYAB	Crystallin, alpha B (*CRYAB*; also known as *Alpha(B)-crystallin, Alpha-crystallin B chain, CRYA2, CTPP2, Heat shock protein beta-5, HspB5, HSPB5, Renal carcinoma antigen NY-REN-27, Rosenthal fiber component*) is a member of the small heat shock protein (sHSP; also known as the HSP20) family, all of which share a common C terminal motif – the alpha crystallin domain. The protein CRYAB acts as molecular chaperones induced by ubiquitous stress and up-regulated by heat, radiation, oxidative stress and anticancer drugs. Other additional functions of alpha crystallins are the autokinase activity, participation in the intracellular architecture, and the control of large soluble protein aggregates. Additionally, CRYAB shows redundancy in interacting with various apoptosis pathways at multiple levels (Kabbage et al., 2012); besides plays a critical role in vasculature homeostasis and angiogenesis (Ruan et al., 2011). The protein is expressed widely in many tissues and organs (Campbell-Lloyd et al., 2013). In breast cancer, *CRYAB* is usually high differentially expressed in invasive tumours when compared to normal breast tissue specimens (Kabbage et al., 2012) and might be involved in chemotherapy response (Cortesi et al., 2009).
ILMN_1666845 KRT17	The protein encoded by keratin 17 (*KRT17*; also known as *39.1, CK-17, Cytokeratin-17, K17, Keratin, type I cytoskeletal 17, Keratin-17, PC, PC2, PCHC1*) plays a role in the formation and maintenance of various epidermal appendages, as the nail bed, hair follicle, and sebaceous glands. KRT17 regulates other protein synthesis and epithelial cell growth. In addition, the protein is a marker of basal cell differentiation as an attribute of a certain type of "stem cells". In the context, immunohistochemical studies revealed that basal-like breast tumours present *KRT17* up-regulation, with levels associated with a poor clinical outcome (Van De Rijn et al., 2002). This gene is also up-regulated in primary *BRCA1* mutant breast tumours; fact consistent with the reported connection between *BRCA1* mutation and basal-like subtypes (Gorski et al., 2010).
ILMN_1786720 PROM1*	Prominin 1 (*PROM1*; also known as *AC133, Antigen AC133, CD133, CORD12, MCDR2, MSTP061, Prominin-1, Prominin-like protein 1, PROML1, RP41, STGD4*) encodes a transmembrane glycoprotein, often expressed on adult stem cells. The protein plays an essential role in maintaining stem cell properties by suppressing differentiation. Aberrant expression *PROM1* is also associated with several types of cancer. In breast cancer, *PROM1* overexpression is positively related to tumour size, stage, and lymph node metastasis in invasive tumours (Q. Liu et al., 2009). Moreover, there is an important association with p53 mutation, mammary cell dedifferentiation, and the concomitant acquisition of stemlike properties (Coradini et al., 2012). In particular, basal-like subtype shows high p53 mutation and *PROM1* up-regulation, which improve tumour cells aggressiveness (Bertheau et al., 2007) due to activation of angiogenesis and metastasis (N. Liu et al., 2012); besides chemoresistance (Nadal et al., 2013).

Illumina Probe / Gene Symbol	**Gene and Protein Review**
ILMN_1753101 VTCN1	V-set domain containing T cell activation inhibitor 1 (*VTCN1*; also known as *B7h.5, B7H4, B7-H4, B7S1, B7X, FLJ22418, Immune costimulatory protein B7-H4, PRO1291, Protein B7S1, T-cell costimulatory molecule B7x, UNQ659/PRO1291, VCTN1, V-set domain-containing T-cell activation inhibitor 1*) is found on the surface of antigen-presenting cells and interacts with ligands attached to receptors on the surface of T cells. The protein negatively regulates the immune response of T cells by reducing the production of cytokines and controlling the cell cycle progression (Qian et al., 2011; Suh et al., 2006). High levels of the *VTCN1* mRNA and the related protein are associated with a number of cancers, including ovarian and breast cancers (Salceda et al., 2005). The up-regulation is also associated with tumour progression and poor prognosis (Heinonen et al., 2008). Although *VTCN1* detection is observed in PR- / HER2- tumours, the expression is independent of grade and stage (Tringler et al., 2005).
ILMN_1798108 C6orf211	The chromosome 6 open reading frame 211 (C6orf211) is mapped in a region close to *ESR1* and other genes (*AKAP12* and *CCDC170*), suggesting further investigation of a possible connection with the *ESR1* transcription and the luminal subtype in breast cancer disease.
ILMN_1747911 CDK1	The cyclin-dependent kinase 1 (*CDK1*; also known as *CDC2, CDC28A, Cell division control protein 2 homolog, Cell division protein kinase 1, Cyclin-dependent kinase 1, DKFZp686L20222, MGC111195, P34CDC2, p34 protein kinase*) plays a key role in the control of the cell cycle by modulating the centrosome as well as mitotic onset; promotes G2-M transition, and regulates G1 progress and G1-S transition associated with multiple interphase cyclins. The kinase activity of this protein is controlled by cyclin accumulation and destruction through the cell cycle. In addition, CDK1 complexes phosphorylate several substrates that trigger centrosome separation, Golgi dynamics, nuclear envelope breakdown, chromosome condensation, and apoptosis. An abnormal phosphorylation occurs in cancer cell lines, as well as in primary breast tissues and lymphocytes. Moreover, high *CDK1* activity was linked to the absence of a full DNA damage response in mitotic cells (Wei Zhang et al., 2011). Although the impact of this gene in breast cancers remains controversial, there is a significant association with unfavourable clinicopathologic feature such as high histologic grade, large tumour size, lymph node metastases and PR-negative tumours (S. J. Kim; Nakayama; et al., 2008). Ultimately, *CDK1* may be used as a predictive factor to identify patient's response to neoadjuvant chemotherapy (S. J. Kim et al., 2012; Torikoshi et al., 2013; Xia et al., 2014).
ILMN_1666305 CDKN3	The protein encoded by cyclin-dependent kinase inhibitor 3 (*CDKN3*; also known as *CDI1, CDK2-associated dual-specificity phosphatase, CIP2, Cyclin-dependent kinase inhibitor 3, Cyclin-dependent kinase-interacting protein 2, Cyclin-dependent kinase interactor 1, FLJ25787, KAP, KAP1, Kinase-associated phosphatase, MGC70625*) belongs to the dual specificity protein phosphatase family, active toward substrates containing either phosphotyrosine or phosphoserine residues. CDKN3 is a cyclin-dependent kinase inhibitor, and interacts / dephosphorylates CDK2 kinase, thereby reducing its ability to phosphorylate the retinoblastoma protein (RB). Non-phosphorylated RB binds transcription factor E2F1 and prevents the G1-S transition. *CDKN3* was reported to be deleted, mutated, or overexpressed in several types of cancers, including breast tumours (Yu et al., 2010).
ILMN_1678535 ESR1	Estrogen receptor 1 (*ESR1*; also known as *DKFZp686N23123, ER, Era, ER-alpha, ESR, ESRA, Estradiol receptor, Estrogen receptor, NR3A1, Nuclear receptor subfamily 3 group A member 1*) encodes a protein receptor, a ligand-activated transcription factor composed of several domains important for hormone binding, DNA binding, and activation of transcription. Oestrogen and its receptors are central regulators of breast cancer disease and are associated with response to endocrine therapy. Down-regulation of *ESR1*, or eventual mutations, may indicate intrinsic resistance to tamoxifen, increased risk of tumour recurrence (Aguilar et al., 2010; C. Kim et al., 2011; Stossi et al., 2012); even though the mechanisms by which oestrogen receptor dictates tumour status are poorly understood (Dunbier et al., 2011).

Illumina Probe / Gene Symbol	**Gene and Protein Review**
ILMN_2149164 SFRP1	Major gene expression changes occur during progression of neoplastic cells, including down regulation of secreted frizzled-related protein 1 (*SFRP1*; also known as *FRP, FRP1, FRP-1, FrzA, SARP2, SARP-2, Secreted apoptosis-related protein 2, Secreted frizzled-related protein 1, sFRP-1*) (Vargas et al., 2012a). SFRP1 functions as a negative regulator of Wnt/β-catenin pathway, implicated in several human cancers, including breast tumours and respective cell lines (Dahl et al., 2007; Gauger et al., 2011; Gostner et al., 2011; Matsuda et al., 2009; Mukherjee et al., 2012; Shulewitz et al., 2006; Suzuki et al., 2008; Ugolini et al., 2001). Reduced levels of *SFRP1* results in hyperplastic lesions and its loss may be a critical event in cancer initiation (Dumont et al., 2009). In breast carcinomas, *SFRP1* showed significant differences in methylation patterns between ER-negative and ER-positive tumours (Park et al., 2012). The hypermethylation of the *SFRP1* promoter and gene down-regulation has been widely reported in breast cancer (Browne et al., 2011; Vargas et al., 2012b; Yang et al., 2009) and associated with tumour invasion and decreased survival (Gauger & Schneider, 2014; Klopocki et al., 2004; Martin-Manso et al., 2011; Veeck et al., 2006). A potential combinatorial treatment - romidepsin and decitabine - has recently been administered in cell lines, promoting *SFRP1* reexpression with consequently proliferation inhibition and cell death induction via apoptosis (Cooper et al., 2012).
ILMN_1788874 SERPINA3	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3 (*SERPINA3*; also known as *AACT, ACT, alpha-1-antichymotrypsin, Alpha-1-antichymotrypsin, Cell growth-inhibiting gene 24/25 protein, GIG24, GIG25, MGC88254, Serpin A3*) encodes a plasma protease inhibitor and member of the serine protease inhibitor class. *SERPINA3* regulates the activity of neutrophil cathepsin G and is an oestrogen-induced gene. In breast cancer, the mRNA increased expression was reported as an indicator of good prognosis in oestrogen receptor positive breast cancer (Cimino et al., 2008; Sano et al., 2012; Yamamura et al., 2004). It is a maker of oestrogen regulation (Miller & Larionov, 2010); besides a predictor of tumour response to neoadjuvant chemotherapy (Sano et al., 2012).
ILMN_1785570 SUSD3*	The sushi domain containing 3 (*SUSD3*; also known as *MGC26847, Sushi domain-containing protein 3, UNQ9387/PRO34275*) down-regulated in breast carcinomas is associated with a malignant phenotype (short-term overall survival, endocrine insensitivity, triple-negative status, poor tumour differentiation). *SUSD3* is highly expressed in ER-positive breast tumours and the treatment with oestradiol may increase the gene expression in cancer cells (Moy et al., 2014). The *SUSD3* abnormal mRNA levels and the protein function, however, are still unclear and require urgent investigation (Cimino et al., 2008).
ILMN_1803236 CLCA2	The protein encoded by chloride channel accessory 2 (*CLCA2*; also known as *CACC, CACC3, CaCC-3, Calcium-activated chloride channel family member 2, Calcium-activated chloride channel protein 3, Calcium-activated chloride channel regulator 2, CLCRG2, FLJ97885, hCaCC-3, hCLCA2*) belongs to the calcium sensitive chloride conductance protein family. The protein plays a role in modulating chloride current across the plasma membrane in a calcium-dependent mode. CLCA2 is also involved in basal cell adhesion and/or stratification of squamous epithelia. In addition, the molecule is involved in the p53 network and may act as a tumour suppressor in breast and colorectal cancer, inhibiting cancer cell migration and invasion (Sasaki et al., 2012; Vijay Walia et al., 2009; V. Walia et al., 2012). The mechanisms behind the silencing of *CLCA2* in luminal breast cancers and the up-regulation in HER2-enriched and basal-like subtypes, however, have not been elucidated. Ultimately, cell lines *CLCA2*-negative treated with demethylating agents restored the expression of the gene, suggesting an epigenetic control (X. Li et al., 2004).

Illumina Probe / Gene Symbol	**Gene and Protein Review**
ILMN_2161820 GLYATL2*	The enzyme encoded by glycine-N-acyltransferase-like 2 (*GLYATL2*; also known as *Acyl-CoA:glycine N-acyltransferase-like protein 2, BXMAS2-10, GATF-B, Glycine N-acyltransferase-like protein 2, MGC24009*) conjugates medium- and long-chain saturated and unsaturated acyl-CoA esters to glycine, resulting in the production of N-oleoyl glycine and also N-arachidonoyl glycine. N-Oleoyl glycine and N-arachidonoyl glycine are identified as signalling molecules that regulate the perception of pain and body temperature, and also have anti-inflammatory properties (Waluk et al., 2012). *GLYATL2* is up-regulated in salivary gland and trachea, and detected also in spinal cord and skin fibroblasts. In addition, the high levels of the gene in skin and lung may indicate a role in barrier function/immune response and lipid signalling (Waluk et al., 2010).
ILMN_1810978 MUCL1*	The mucin-like 1 (*MUCL1*; also known as *Mucin-like protein 1, Protein BS106, SBEM, Small breast epithelial mucin, UNQ590/PRO1160*) encodes a 90 amino acids glycoprotein that exhibits characteristics of members of the mucin family. The presence of a hydrophobic signal peptide within the protein sequence suggests that MUCL1 is a secreted and subjected to proteolytic processing. The putative gene is expressed only in mammary and salivary glands and it is promising as a new biomarker with high tissue specificity (L. Liu et al., 2013), besides with a great potential for predicting metastasis and response to neoadjuvant chemotherapy (Z. Z. Liu et al., 2010). In breast cancer, the protein is more frequently observed in ER-negative than in ER-positive cancers, and positively associated with *HER2* overexpression. The evaluation of MUCL1 expression, nonetheless, needs to consider also the heterogeneity and different molecular subtypes (Valladares-Ayerbes et al., 2009). In general, increased *MUCL1* is associated with high tumour grades, lymph node metastasis (Weigelt et al., 2004) and reduced survival (Miksicek et al., 2002; Skliris et al., 2008).
ILMN_1773459 SOX11*	SRY (sex determining region Y)-box 11 (*SOX11*; also known as *Transcription factor SOX-11*) encodes a protein involved in the regulation of embryonic development and in the determination of the cell fate. The protein acts as a transcriptional regulator after modelling a complex with other proteins. SOX11 functions in the developing nervous system and play a role in tumorigenesis. In breast cancer patients, the gene up-regulation might contribute to a proliferative genotype which may be linked to poor prognosis and therefore worse overall survival (Lopez et al., 2012). Interestingly, *SOX11* levels were higher in basal-like and HER2-enriched breast cancers compared with other subtypes (Zvelebil et al., 2013).
ILMN_1674533 TRPV6	Transient receptor potential cation channel, subfamily V, member 6 (*TRPV6*; also known as *ABP/ZF, Calcium transport protein 1, CaT1, CAT1, CATL, CaT-L, CaT-like, ECaC2, ECAC2, Epithelial calcium channel 2, HSA277909, LP6728, Transient receptor potential cation channel subfamily V member 6, TrpV6, ZFAB*) encodes a member of a family of multipass membrane proteins that functions as calcium channels. The up-regulation of *TRPV6* is observed in several tumours such as breast, prostate, colon, thyroid and ovary; and in various tumour cell lines (Bolanz et al., 2008; Bowen et al., 2013). In particular, increased TRPV6 expression is a feature of ER-negative breast tumours (HER2-enriched and basal-like subtypes) and has been associated to patient decreased survival (Peters et al., 2012). The mechanism underlying the *TRPV6*-mediated regulation of cancer progression and its downstream signalling, however, remain poorly understood (S. Y. Kim et al., 2013). Inhibitors of the TRPV6 channel have been investigated as potential targets for diagnosis, prognosis and/or therapeutic approaches in cancers (Dhennin-Duthille et al., 2011; Landowski et al., 2011).
ILMN_1687235 HPN	The gene hepsin (*HPN*; also known as *Serine protease hepsin, TMPRSS1, Transmembrane protease serine 1*) encodes a type II transmembrane serine protease that is involved in diverse cellular functions, including cell growth and maintenance of cell morphology. The protein is cleaved into a catalytic serine protease chain and a non-catalytic scavenger receptor. The expression of the encoded protein is associated with the progression of several types of malignancies, nevertheless little is known about its clinical and biological significance in breast cancer. *HPN* is up-regulated in breast tumours; besides significantly associated with tumour stage, lymph node metastasis, oestrogen receptor positivity, and progesterone receptor positivity (Xing et al., 2011).

Illumina Probe / Gene Symbol	**Gene and Protein Review**
ILMN_1655915 MMP11	The matrix metallopeptidase 11 (stromelysin 3) (*MMP11*; also known as *Matrix metalloproteinase-11, MMP-11, SL-3, ST3, STMY3, Stromelysin-3*) is a member of the matrix metalloproteinase (MMP) family of proteases. These proteins are constituent of the extracellular matrix and act on the epithelial/connective interface in embryogenesis, wound healing, tissue involution, and reproduction. *MMP11* expressed in fibroblasts near areas of invasive carcinoma lead to the gain of metastatic potential for spread of tumour cells, with patterns of subsequent invasion and migration for different types of solid tumours and cell lines (DeLassus et al., 2011; DeLassus et al., 2008; Kasper et al., 2007; Kwon et al., 2011; K.-W. Min et al., 2013). In breast cancer, higher expression level of *MMP11* is correlated with patients having poorly differentiated tumours, increased invasiveness, node metastasis, and worse prognosis (Cheng et al., 2010; Eiro et al., 2013; Eiseler et al., 2009; Garcia et al., 2010; Tan et al., 2013). *MMP11* gene expression analysis may also be used in clinical applications for breast cancer diagnosis, management and therapy (Hegedüs et al., 2008; Selvey et al., 2004). Ultimately, the up-regulation of this gene is linked to other markers such as p53, ER and *HER2* (K. W. Min et al., 2012).
ILMN_1711470 UBE2T	The ubiquitin-conjugating enzyme E2T (putative) (*UBE2T*; also known as *Cell proliferation-inducing gene 50 protein, HSPC150, PIG50, Ubiquitin carrier protein T, Ubiquitin-conjugating enzyme E2 T, Ubiquitin-protein ligase T*) accepts ubiquitin from the E1 complex and catalyses its covalent attachment to other proteins. The covalent conjugation of ubiquitin to proteins regulates diverse cellular pathways and proteins. Ubiquitin is transferred to a target protein through a concerted action of ubiquitin-activating enzyme (E1), ubiquitin-conjugating enzyme (E2), and ubiquitin ligase (E3). UBE2T acts as a specific E2 ubiquitin-conjugating enzyme for the Fanconi anemia complex and contribute to ubiquitination and degradation of BRCA1. The enzyme is up-regulated in different types of cancer including breast, bladder, lung, and prostate cancers; playing essential role in cell proliferation. In breast tumours, the gene up-regulation cause the decrease of the BRCA1 levels, however, major pathways involving UBE2T are still poorly understood (Tomomi Ueki et al., 2009).
ILMN_1789507 COL11A1	The collagen, type XI, alpha 1 (COL11A1; also known as *CO11A1, COLL6, Collagen alpha-1(XI) chain, STL2*) encodes one of the two alpha chains of type XI collagen, a minor fibrillar collagen. Type XI collagen is a heterotrimer and play an important role in fibrillogenesis by controlling lateral growth of collagen fibrils. *COL11A1* is expressed by both the epithelial and stromal compartments and its expression is deregulated in a range of cancers, such as breast and colon. In particular, molecules related to extracellular matrix remodelling (e.g. COL11A1) are differentially expressed in breast tumours 'in situ' and invasive; enriched in metastatic tumour cells (Ellsworth et al., 2009; H. Kim et al., 2010; Vargas et al., 2012a).
ILMN_1740609 CCL15	The chemokine (C-C motif) ligand 15 (*CCL15*; also known as *C-C motif chemokine 15, Chemokine CC-2, HCC-2, HMRP-2B, Leukotactin-1, LKN1, Lkn-1, LKN-1, Macrophage inflammatory protein 5, MIP-1d, MIP-1D, MIP-1 delta, MIP5, MIP-5, Mrp-2b, MRP-2B, NCC3, NCC-3, SCYA15, SCYL3, Small-inducible cytokine A15, SY15*) encodes a secreted protein characterised by two adjacent cysteines, further processed into numerous smaller functional peptides. The protein has chemotactic factor that attracts T cells and monocytes; acts through C-C chemokine receptor type 1 (CCR1) and also binds to type 3 (CCR3). In hepatocellular carcinoma, the up-regulation of *CCL15* promotes cell migration and invasion (Y. Li et al., 2013). High levels of *CCL15* also increase the expression of matrix metalloproteinase and induce angiogenesis (Itatani et al., 2013).

Illumina Probe / Gene Symbol	**Gene and Protein Review**
ILMN_1651282 COL17A1*	The collagen, type XVII, alpha 1 (*COL17A1*; also known as *180 kDa bullous pemphigoid antigen 2, BA16H23.2, BP180, BPAG2, Bullous pemphigoid antigen 2, Collagen alpha-1(XVII) chain, FLJ60881, KIAA0204, LAD-1*) encodes the alpha chain of type XVII collagen, a transmembrane protein or a soluble form generated by proteolytic processing of the full length form. The protein is a structural component of hemidesmosomes, multiprotein complexes at the dermal-epidermal basement membrane zone that mediate adhesion of basal keratinocytes to the underlying membrane. Hemidesmosomal components are also implicated in signal transduction and thereby are able to influence cell growth, motility and differentiation. In neoplastic tissue, *COL17A1* aberrant expression depends on the stage of the tumour, down-regulated in mild dysplasia and up-regulation as the tumour further evolves (van Zalen et al., 2006).
ILMN_1723684 DARC	Duffy blood group, atypical chemokine receptor (*DARC*; also known as *CCBP1, CD234, Dfy, Duffy antigen/chemokine receptor, FY, Fy glycoprotein, Glycoprotein D, GPD, GpFy, Plasmodium vivax receptor, WBCQ1*) encodes a glycosylated membrane protein and a non-specific receptor for several chemokines. Polymorphisms in this gene are the basis of the Duffy blood group system. It is reported that DARC plays a negative regulatory role in human breast cancer. Overexpression of DARC protein in breast cancer cells leads to significant inhibition of tumorigenesis and metastasis (Bandyopadhyay et al., 2006; J. Wang et al., 2013; Zeng et al., 2011). DARC is also correlated with breast cancer incidence, axillary lymph node metastasis and overall survival (X. F. Liu et al., 2012).
ILMN_1809099 IL33*	Interleukin 33 (*IL33*; also known as *C9orf26, DKFZp586H0523, DVS27, IL1F11, IL-1F11, IL-33, Interleukin-1 family member 11, Interleukin-33, NFEHEV, NFHEV, NF-HEV, Nuclear factor from high endothelial venules, RP11-575C20.2*) is a member of the IL1 family that induces production of T helper-2 (Th2) associated cytokines. The protein acts as a chemoattractant tor Th2 cells, and amplifies immune responses during tissue injury. IL33 also functions as a chromatin-associated nuclear factor with transcriptional repressor properties. In breast cancer cells, a frequent overexpression is observed, though the gene deregulation is not clearly understood in the disease (Wu et al., 2012).
ILMN_1766650 FOXA1	The forkhead box A1 (*FOXA1*; also known as *Forkhead box protein A1, Hepatocyte nuclear factor 3-alpha, HNF3A, HNF-3A, HNF-3-alpha, MGC33105, TCF3A, TCF-3A, Transcription factor 3A*) encodes a member of the forkhead class of DNA-binding proteins. The nuclear factor is a transcriptional activator involved in embryonic development, establishment of tissue-specific gene expression and regulation of gene expression in differentiated tissues. The protein is also implicated in the development of multiple organs such as liver, pancreas, thyroid, prostate and breast. Basically, it modulates the transcriptional activity of nuclear hormone receptors (Bernardo & Keri, 2012). FOXA1 acts in both androgen receptor (AR) and oestrogen receptor (ER), directing the binding location, and therefore the transcriptional activity (Augello et al., 2011; Ni et al., 2011; Robinson et al., 2011). In breast cancer, FOXA1 plays a pivotal role in mammary ductal morphogenesis (Bernardo et al., 2010; Fu et al., 2011), tumour early stage, drug response, and metastatic disease (Robinson & Carroll, 2012). Mutation and SNP variation located in enhancer regions may alter FOXA1 binding affinity and affect breast cancer risk (Cowper-Sal et al., 2012; Katika & Hurtado, 2013; Meyer & Carroll, 2012; Robinson et al., 2013). In addition, high expression of FOXA1 is correlated with luminal A subtype and it is a significant predictor of survival in patients with ER-positive tumours (Badve et al., 2007; Bernardo & Keri, 2012; Mehta et al., 2012; Yamaguchi et al., 2008), and a marker of good prognosis (Albergaria et al., 2009; Habashy et al., 2008; Hisamatsu et al., 2012; Hisamatsu et al., 2015). Ultimately, genome analysis of ER-FOXA1 interactions is required to understand the molecular mechanisms of ER activity (Hurtado et al., 2011; Kong et al., 2011; Magnani & Lupien, 2014; Naderi et al., 2012).

Illumina Probe / Gene Symbol	**Gene and Protein Review**
ILMN_1811387 TFF3	The trefoil factor 3 (intestinal) (*TFF3*; also known as *hITF, HITF, hP1.B, Intestinal trefoil factor, ITF, P1B, Polypeptide P1.B, TFI, Trefoil factor 3*) gene is translated in a stable secretory protein having at least one copy of the trefoil motif and a domain with three conserved disulphides. TFF3 functions as 'luminal epithelium guardian', involved in the maintenance and repair of the mucosa after damage. Besides, promotes the mobility of epithelial cells in healing processes. Up-regulation of *TFF3* is observed in various neoplastic diseases, including breast cancer, where the gene has been target as a biomarker (Lasa et al., 2013). TFF3 is induced by hormones such as oestrogen, and is usually combined with TFF1 expression in ER-positive malignant breast tumour cells (Ahmed et al., 2012; Lacroix, 2006). Moreover, the *TFF3* levels are close to that of *ESR1*, yet reduce after tamoxifen treatment (Fenne et al., 2013; Taylor et al., 2010). These genes are components of the 'luminal epithelial' signature defining a well-differentiated, low-grade intrinsic subtype of breast cancer: the luminal A (Lacroix, 2006). Basal-like and claudin-low breast cancer subtypes showed frequent hypermethylation of the TFF3 promoter region (Roll et al., 2013; Sandhu et al., 2014).
ILMN_1738401 FOXC1	The forkhead box C1 (*FOXC1*; also known as *ARA, FKHL7, Forkhead box protein C1, Forkhead-related protein FKHL7, Forkhead-related transcription factor 3, FREAC3, FREAC-3, IGDA, IHG1, IRID1, RIEG3*) is part of the forkhead family of transcription factors, characterised by a common DNA-binding domain. All the mechanisms through *FOXC1* are not yet determined; however, the gene plays important roles in cell growth, survival, differentiation, and migration. *FOXC1* is identified as a functionally important biomarker of breast cancer aggressiveness, particularly associated with basal-like breast cancer subtype (Sizemore & Keri, 2012; Wang et al., 2012). The gene up-regulation in breast tumour cells induces epithelial-mesenchymal transition, drug resistance, and increased cell proliferation and invasion (Tkocz et al., 2012).
ILMN_1689146 GABRP	Gamma-aminobutyric acid (GABA) A receptor, pi (*GABRP*; also known as *GABA(A) receptor subunit pi, Gamma-aminobutyric acid receptor subunit pi, MGC126386, MGC126387*) encodes a transmembrane protein composed by multisubunit in the chloride channel that mediates synaptic transmission in the central nervous system. The gene is expressed in several non-neuronal tissues including breast and ovaries. In breast, *GABRP* is mainly expressed in myoepithelial/basal cells, and the function is related to tissue contractility (Lacroix, 2006). In breast cancer cells, the gene is normally up-regulated among ER-negative patients (Andres et al., 2013; Symmans et al., 2005).
ILMN_1807423 IGF2BP3*	The protein encoded by insulin-like growth factor 2 mRNA binding protein 3 (*IGF2BP3*; also known as *CT98, DKFZp686F1078, hKOC, IGF2 mRNA-binding protein 3, IGF-II mRNA-binding protein 3, IMP3, IMP-3, Insulin-like growth factor 2 mRNA-binding protein 3, KH domain-containing protein overexpressed in cancer, KOC1, VICKZ3, VICKZ family member 3*) is primarily located in the nucleolus and belongs to a conserved family of RNA-binding proteins. The RNA-binding factor recruits target transcripts to cytoplasmic protein-RNA complexes (mRNPs), modulating the rate and location of endonuclease attacks or microRNA-mediated degradation. The protein, nonetheless, is involved not only in RNA synthesis and metabolism, but in various important aspects of cell function, such as cell polarisation, migration, morphology, metabolism, proliferation and differentiation. IGF2BP3 is largely absent in adult tissues but *de novo* synthetised or severely up-regulated in various tumours and tumour-derived cells. In breast cancer, the up-regulation enhances tumour growth, angiogenesis and metastasis, resulting in poorer survival (Bell et al., 2013; Fadare et al., 2013; Lochhead et al., 2012), and chemoresistance (Samanta et al., 2013). High expression has also been associated with triple-negative breast carcinomas (Samanta et al., 2012; Walter et al., 2009; Won et al., 2013).

Illumina Probe / Gene Symbol	Gene and Protein Review
ILMN_1692938 PSAT1	The phosphoserine aminotransferase 1 (*PSAT1*; also known as *EPIP, MGC1460, Phosphohydroxythreonine aminotransferase, Phosphoserine aminotransferase, PSA, PSAT*) encodes a member of the class-V pyridoxal-phosphate-dependent aminotransferase family. Mutations in this gene are associated with phosphoserine aminotransferase deficiency. *PSAT1* methylation and aberrant expression are strongly correlated with specific clinical and pathologic features of breast cancer. Notably, the *PSAT1* hypermethylation is associated with low-grade, low-proliferation, hormone receptor ER-positive, lymph node positive breast cancer in post-menopausal women (Bu et al., 2013); besides it is an indicator of response to tamoxifen endocrine therapy (Martens et al., 2005). On the other hand, high expression of *PSAT1* is associated with decreased relapse-free and overall survival of patients, and linked to malignant phenotypic features of breast cancer (Bu et al., 2013).
ILMN_1668766 ROPN1	The rhophilin associated tail protein 1 (*ROPN1*; also known as *Cancer/testis antigen 91, CT91, DKFZp434B1222, ODF6, Rhophilin-associated protein 1A, RHPNAP1, ROPN1A, ropporin, Ropporin-1A*) is an important reproduction related gene. The protein is involved in sperm maturation, motility, capacitation, hyperactivation and acrosome reaction. Other important functions such as cAMP-dependent protein kinase regulator activity, protein binding activity, phosphorylation and signal transduction regulation were reported; even though ROPN1 requires further investigation (Lan et al., 2012). Recently, *ROPN1* was validated with diagnostic significance in basal-like breast cancer cells as one of the conserved elements of the *SOX10* signature (Ivanov et al., 2013).

Note: *Elements named according to Human HT (Illumina HT-12 v3) and matching different regions of the genome with more than one annotation in UCSC and iHOP.

Supporting References

Agnati, L. F., Genedani, S., Leo, G., Forni, A., Woods, A. S., Filaferro, M., et al. (2007). Abeta peptides as one of the crucial volume transmission signals in the trophic units and their interactions with homocysteine. Physiological implications and relevance for Alzheimer's disease. *J Neural Transm, 114*(1), 21-31.

Aguilar, H., Sole, X., Bonifaci, N., Serra-Musach, J., Islam, A., Lopez-Bigas, N., et al. (2010). Biological reprogramming in acquired resistance to endocrine therapy of breast cancer. *Oncogene, 29*(45), 6071-6083.

Ahmed, A. R., Griffiths, A. B., Tilby, M. T., Westley, B. R., & May, F. E. (2012). TFF3 is a normal breast epithelial protein and is associated with differentiated phenotype in early breast cancer but predisposes to invasion and metastasis in advanced disease. *Am J Pathol, 180*(3), 904-916.

Ahn, S. G., Lee, H. M., Lee, H. W., Lee, S. A., Lee, S. R., Leem, S. H., et al. (2013). Prognostic discrimination using a 70-gene signature among patients with estrogen receptor-positive breast cancer and an intermediate 21-gene recurrence score. *Int J Mol Sci, 14*(12), 23685-23699.

Albergaria, A., Paredes, J., Sousa, B., Milanezi, F., Carneiro, V., Bastos, J., et al. (2009). Expression of FOXA1 and GATA-3 in breast cancer: the prognostic significance in hormone receptor-negative tumours. *Breast Cancer Res., 11*(3), R40.

Andres, S. A., Brock, G. N., & Wittliff, J. L. (2013). Interrogating differences in expression of targeted gene sets to predict breast cancer outcome. *BMC Cancer, 13*(326), 1-18.

Arbitrario, J. P., Belmont, B. J., Evanchik, M. J., Flanagan, W. M., Fucini, R. V., Hansen, S. K., et al. (2010). SNS-314, a pan-Aurora kinase inhibitor, shows potent anti-tumor activity and dosing flexibility in vivo. *Cancer Chemother Pharmacol, 65*(4), 707-717.

Augello, M. A., Hickey, T. E., & Knudsen, K. E. (2011). FOXA1: master of steroid receptor function in cancer. *Embo J, 30*(19), 3885-3894.

Badve, S., Turbin, D., Thorat, M. A., Morimiya, A., Nielsen, T. O., Perou, C. M., et al. (2007). FOXA1 expression in breast cancer—correlation with luminal subtype A and survival. *Clin. Cancer Res., 13*(15), 4415-4421.

Balafoutas, D., zur Hausen, A., Mayer, S., Hirschfeld, M., Jaeger, M., Denschlag, D., et al. (2013). Cancer testis antigens and NY-BR-1 expression in primary breast cancer: prognostic and therapeutic implications. *BMC Cancer, 13*(1), 271.

Bandyopadhyay, S., Zhan, R., Chaudhuri, A., Watabe, M., Pai, S. K., Hirota, S., et al. (2006). Interaction of KAI1 on tumor cells with DARC on vascular endothelium leads to metastasis suppression. *Nat. Med., 12*(8), 933-938.

Bell, J. L., Wächter, K., Mühleck, B., Pazaitis, N., Köhn, M., Lederer, M., et al. (2013). Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? *Cell. Mol. Life Sci., 70*(15), 2657-2675.

Bernardo, G. M., & Keri, R. A. (2012). FOXA1: a transcription factor with parallel functions in development and cancer. *Biosci Rep, 32*(2), 113-130.

Bernardo, G. M., Lozada, K. L., Miedler, J. D., Harburg, G., Hewitt, S. C., Mosley, J. D., et al. (2010). FOXA1 is an essential determinant of ERalpha expression and mammary ductal morphogenesis. *Development, 137*(12), 2045-2054.

Bertheau, P., Turpin, E., Rickman, D. S., Espié, M., De Reyniès, A., Feugeas, J.-P., et al. (2007). Exquisite Sensitivity of TP53 Mutant and Basal Breast Cancers to a Dose-Dense Epirubicin − Cyclophosphamide Regimen. *PLoS Med., 4*(3), e90.

Bolanz, K. A., Hediger, M. A., & Landowski, C. P. (2008). The role of TRPV6 in breast carcinogenesis. *Mol. Cancer Ther., 7*(2), 271-279.

Bowen, C. V., DeBay, D., Ewart, H. S., Gallant, P., Gormley, S., Ilenchuk, T. T., et al. (2013). In vivo detection of human TRPV6-rich tumors with anti-cancer peptides derived from soricidin. *PLoS One, 8*(3), e58866.

Browne, E. P., Punska, E. C., Lenington, S., Otis, C. N., Anderton, D. L., & Arcaro, K. F. (2011). Increased promoter methylation in exfoliated breast epithelial cells in women with a previous breast biopsy. *Epigenetics, 6*(12), 1425-1435.

Bu, D., Lewis, C. M., Sarode, V., Chen, M., Ma, X., Lazorwitz, A. M., et al. (2013). Identification of Breast cancer DNA methylation Markers optimized for fine-needle aspiration samples. *Cancer Epidemiol., Biomarkers Prev., 22*(12), 2212-2221.

Bush, T. L., Payton, M., Heller, S., Chung, G., Hanestad, K., Rottman, J. B., et al. (2013). AMG 900, a small-molecule inhibitor of aurora kinases, potentiates the activity of microtubule-targeting agents in human metastatic breast cancer models. *Mol Cancer Ther, 12*(11), 2356-2366.

Campbell-Lloyd, A. J., Mundy, J., Deva, R., Lampe, G., Hawley, C., Boyle, G., et al. (2013). Is alpha-B crystallin an independent marker for prognosis in lung cancer? *Heart, lung & circulation, 22*(9), 759-766.

Canevari, G., Re Depaolini, S., Cucchi, U., Bertrand, J. A., Casale, E., Perrera, C., et al. (2013). Structural insight into maternal embryonic leucine zipper kinase (MELK) conformation and inhibition toward structure-based drug design. *Biochemistry, 52*(37), 6380-6387.

Carretero, M., Ruiz-Torres, M., Rodriguez-Corsino, M., Barthelemy, I., & Losada, A. (2013). Pds5B is required for cohesion establishment and Aurora B accumulation at centromeres. *Embo J, 32*(22), 2938-2949.

Cheng, C.-W., Yu, J.-C., Wang, H.-W., Huang, C.-S., Shieh, J.-C., Fu, Y.-P., et al. (2010). The clinical implications of MMP-11 and CK-20 expression in human breast cancer. *Clin. Chim. Acta, 411*(3), 234-241.

Cimino, D., Fuso, L., Sfiligoi, C., Biglia, N., Ponzone, R., Maggiorotto, F., et al. (2008). Identification of new genes associated with breast cancer progression by gene expression analysis of predefined sets of neoplastic tissues. *Int. J. Cancer, 123*(6), 1327-1338.

Ciriello, G., Sinha, R., Hoadley, K. A., Jacobsen, A. S., Reva, B., Perou, C. M., et al. (2013). The molecular diversity of Luminal A breast tumors. *Breast Cancer Res Treat, 141*(3), 409-420.

Colak, D., Nofal, A., Albakheet, A., Nirmal, M., Jeprel, H., Eldali, A., et al. (2013). Age-specific gene expression signatures for breast tumors and cross-species conserved potential cancer progression markers in young women. *PLoS One, 8*(5), e63204.

Cooper, S. J., Von Roemeling, C. A., Kang, K. H., Marlow, L. A., Grebe, S. K., Menefee, M. E., et al. (2012). Reexpression of tumor suppressor, sFRP1, leads to antitumor synergy of combined HDAC and methyltransferase inhibitors in chemoresistant cancers. *Mol. Cancer Ther., 11*(10), 2105-2115.

Coradini, D., Fornili, M., Ambrogi, F., Boracchi, P., & Biganzoli, E. (2012). TP53 mutation, epithelial-mesenchymal transition, and stemlike features in breast cancer subtypes. *BioMed Res Int, 2012*, 1-12.

Cortesi, L., Barchetti, A., De Matteis, E., Rossi, E., Della Casa, L., Marcheselli, L., et al. (2009). Identification of protein clusters predictive of response to chemotherapy in breast cancer patients. *J Proteome Res, 8*(11), 4916-4933.

Cowper-Sal, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., et al. (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet., 44*(11), 1191-1198.

Dahl, E., Wiesmann, F., Woenckhaus, M., Stoehr, R., Wild, P. J., Veeck, J., et al. (2007). Frequent loss of SFRP1 expression in multiple human solid tumours: association with aberrant promoter methylation in renal cell carcinoma. *Oncogene, 26*(38), 5680-5691.

DeLassus, G. S., Cho, H., & Eliceiri, G. L. (2011). New signaling pathways from cancer progression modulators to mRNA expression of matrix metalloproteinases in breast cancer cells. *J. Cell. Physiol., 226*(12), 3378-3384.

DeLassus, G. S., Cho, H., Park, J., & Eliceiri, G. L. (2008). New pathway links from cancer-progression determinants to gene expression of matrix metalloproteinases in breast cancer cells. *J. Cell. Physiol., 217*(3), 739-744.

Dhennin-Duthille, I., Gautier, M., Faouzi, M., Guilbert, A., Brevet, M., Vaudry, D., et al. (2011). High expression of transient receptor potential channels in human breast cancer epithelial cells and tissues: correlation with pathological parameters. *Cell. Physiol. Biochem., 28*(5), 813-822.

Dumont, N., Crawford, Y. G., Sigaroudinia, M., Nagrani, S. S., Wilson, M. B., Buehring, G. C., et al. (2009). Human mammary cancer progression model recapitulates methylation events associated with breast premalignancy. *Breast Cancer Res., 11*(6), R87.

Dunbier, A. K., Anderson, H., Ghazoui, Z., Lopez-Knowles, E., Pancholi, S., Ribas, R., et al. (2011). ESR1 is co-expressed with closely adjacent uncharacterised genes spanning a breast cancer susceptibility locus at 6q25.1. *PLoS genetics, 7*(4), e1001382.

Eiro, N., Fernandez-Garcia, B., Gonzalez, L. O., & Vizoso, F. J. (2013). Cytokines related to MMP-11 expression by inflammatory cells and breast cancer metastasis. *Oncoimmunology, 2*(5), e24010.

Eiseler, T., Doppler, H., Yan, I. K., Goodison, S., & Storz, P. (2009). Protein kinase D1 regulates matrix metalloproteinase expression and inhibits breast cancer cell invasion. *Breast Cancer Res., 11*(1), R13.

Ellsworth, R. E., Seebach, J., Field, L. A., Heckman, C., Kane, J., Hooke, J. A., et al. (2009). A gene expression signature that defines breast cancer metastases. *Clin Exp Metastasis, 26*(3), 205-213.

Fadare, O., Liang, S. X., Crispens, M. A., Jones, H. W., Khabele, D., Gwin, K., et al. (2013). Expression of the oncofetal protein IGF2BP3 in endometrial clear cell carcinoma: assessment of frequency and significance. *Hum. Pathol., 44*(8), 1508-1515.

Fenne, I. S., Helland, T., Flageng, M. H., Dankel, S. N., Mellgren, G., & Sagen, J. V. (2013). Downregulation of steroid receptor coactivator-2 modulates estrogen-responsive genes and stimulates proliferation of mcf-7 breast cancer cells. *PLoS One, 8*(7), e70096.

Fiskus, W., Hembruff, S. L., Rao, R., Sharma, P., Balusu, R., Venkannagari, S., et al. (2012). Co-treatment with vorinostat synergistically enhances activity of Aurora kinase inhibitor against human breast cancer cells. *Breast Cancer Res Treat, 135*(2), 433-444.

Fletcher, G., Patel, S., Tyson, K., Adam, P., Schenker, M., Loader, J., et al. (2003). hAG-2 and hAG-3, human homologues of genes involved in differentiation, are associated with oestrogen receptor-positive breast tumours and interact with metastasis gene C4. 4a and dystroglycan. *Br. J. Cancer, 88*(4), 579-585.

Fountzilas, G., Kotoula, V., Pectasides, D., Kouvatseas, G., Timotheadou, E., Bobos, M., et al. (2013). Ixabepilone administered weekly or every three weeks in HER2-negative metastatic breast cancer patients; a randomized non-comparative phase II trial. *PLoS One, 8*(7), e69256.

Fu, X., Huang, C., & Schiff, R. (2011). More on FOX News: FOXA1 on the horizon of estrogen receptor function and endocrine response. *Breast Cancer Res., 13*(2), 307.

Garcia, M. F., Gonzalez-Reyes, S., Gonzalez, L. O., Junquera, S., Berdize, N., Del Casar, J. M., et al. (2010). Comparative study of the expression of metalloproteases and their inhibitors in different localizations within primary tumours and in metastatic lymph nodes of breast cancer. *Int J Exp Pathol, 91*(4), 324-334.

Gauger, K. J., Chenausky, K. L., Murray, M. E., & Schneider, S. S. (2011). SFRP1 reduction results in an increased sensitivity to TGF-beta signaling. *BMC Cancer, 11*(1), 59.

Gauger, K. J., & Schneider, S. S. (2014). Tumour supressor secreted frizzled related protein 1 regulates p53-mediated apoptosis. *Cell Biol. Int., 38*(1), 124-130.

Giger, O., Caduff, R., O'Meara, A., Diener, P. A., Knuth, A., Jager, D., et al. (2010). Frequent expression of the breast differentiation antigen NY-BR-1 in mammary and extramammary Paget's disease. *Pathol Int, 60*(11), 726-734.

Gorski, J. J., James, C. R., Quinn, J. E., Stewart, G. E., Staunton, K. C., Buckley, N. E., et al. (2010). BRCA1 transcriptionally regulates genes associated with the basal-like phenotype in breast cancer. *Breast Cancer Res. Treat., 122*(3), 721-731.

Gostner, J. M., Fong, D., Wrulich, O. A., Lehne, F., Zitt, M., Hermann, M., et al. (2011). Effects of EpCAM overexpression on human breast cancer cell lines. *BMC Cancer, 11*(1), 45.

Gully, C. P., Zhang, F., Chen, J., Yeung, J. A., Velazquez-Torres, G., Wang, E., et al. (2010). Antineoplastic effects of an Aurora B kinase inhibitor in breast cancer. *Mol. Cancer, 9*(1), 42.

Habashy, H. O., Powe, D. G., Rakha, E. A., Ball, G., Paish, C., Gee, J., et al. (2008). Forkhead-box A1 (FOXA1) expression in breast cancer and its prognostic significance. *Eur. J. Cancer, 44*(11), 1541-1551.

Hardwicke, M. A., Oleykowski, C. A., Plant, R., Wang, J., Liao, Q., Moss, K., et al. (2009). GSK1070916, a potent Aurora B/C kinase inhibitor with broad antitumor activity in tissue culture cells and human tumor xenograft models. *Mol Cancer Ther, 8*(7), 1808-1817.

Hebbard, L. W., Maurer, J., Miller, A., Lesperance, J., Hassell, J., Oshima, R. G., et al. (2010). Maternal embryonic leucine zipper kinase is upregulated and required in mammary tumor-initiating cells in vivo. *Cancer Res., 70*(21), 8863-8873.

Hegedüs, L., Cho, H., Xie, X., & Eliceiri, G. L. (2008). Additional MDA-MB-231 breast cancer cell matrix metalloproteinases promote invasiveness. *J. Cell. Physiol., 216*(2), 480-485.

Hegyi, K., Egervari, K., Sandor, Z., & Mehes, G. (2012). Aurora kinase B expression in breast carcinoma: cell kinetic and genetic aspects. *Pathobiology, 79*(6), 314-322.

Heinonen, H., Nieminen, A., Saarela, M., Kallioniemi, A., Klefstrom, J., Hautaniemi, S., et al. (2008). Deciphering downstream gene targets of PI3K/mTOR/p70S6K pathway in breast cancer. *BMC Genomics, 9*(1), 348.

Hisamatsu, Y., Tokunaga, E., Yamashita, N., Akiyoshi, S., Okada, S., Nakashima, Y., et al. (2012). Impact of FOXA1 expression on the prognosis of patients with hormone receptor-positive breast cancer. *Ann. Surg. Oncol., 19*(4), 1145-1152.

Hisamatsu, Y., Tokunaga, E., Yamashita, N., Akiyoshi, S., Okada, S., Nakashima, Y., et al. (2015). Impact of GATA-3 and FOXA1 expression in patients with hormone receptor-positive/HER2-negative breast cancer. *Breast Cancer, 22*(5), 520-528.

Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D., & Carroll, J. S. (2011). FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat. Genet., 43*(1), 27-33.

Ikeda, H., Taira, N., Hara, F., Fujita, T., Yamamoto, H., Soh, J., et al. (2010). The estrogen receptor influences microtubule-associated protein tau (MAPT) expression and the selective estrogen receptor inhibitor fulvestrant downregulates MAPT and increases the sensitivity to taxane in breast cancer cells. *Breast Cancer Res., 12*(3), R43.

Inoda, S., Hirohashi, Y., Torigoe, T., Nakatsugawa, M., Kiriyama, K., Nakazawa, E., et al. (2009). Cep55/c10orf3, a tumor antigen derived from a centrosome residing protein in breast carcinoma. *J Immunother, 32*(5), 474-485.

Itatani, Y., Kawada, K., Fujishita, T., Kakizaki, F., Hirai, H., Matsumoto, T., et al. (2013). Loss of SMAD4 from colorectal cancer cells promotes CCL15 expression to recruit CCR1+ myeloid cells and facilitate liver metastasis. *Gastroenterol, 145*(5), 1064-1075.

Ivanov, S. V., Panaccione, A., Nonaka, D., Prasad, M. L., Boyd, K. L., Brown, B., et al. (2013). Diagnostic SOX10 gene signatures in salivary adenoid cystic and breast basal-like carcinomas. *Br J Cancer, 109*(2), 444-451.

Jäger, D., Filonenko, V., Gout, I., Frosina, D., Eastlake-Wade, S., Castelli, S., et al. (2007). NY-BR-1 is a differentiation antigen of the mammary gland. *Appl. Immunohistochem. Mol. Morphol., 15*(1), 77-83.

Kabbage, M., Trimeche, M., Ben Nasr, H., Hammann, P., Kuhn, L., Hamrita, B., et al. (2012). Expression of the molecular chaperone alphaB-crystallin in infiltrating ductal breast carcinomas and the significance thereof: an immunohistochemical and proteomics-based strategy. *Tumour biology, 33*(6), 2279-2288.

Kalous, O., Conklin, D., Desai, A. J., Dering, J., Goldstein, J., Ginther, C., et al. (2013). AMG 900, pan-Aurora kinase inhibitor, preferentially inhibits the proliferation of breast cancer cell lines with dysfunctional p53. *Breast Cancer Res Treat, 141*(3), 397-408.

Kasper, G., Reule, M., Tschirschmann, M., Dankert, N., Stout-Weider, K., Lauster, R., et al. (2007). Stromelysin-3 over-expression enhances tumourigenesis in MCF-7 and MDA-MB-231 breast cancer cell lines: involvement of the IGF-1 signalling pathway. *BMC Cancer, 7*(1), 12.

Katika, M. R., & Hurtado, A. (2013). A functional link between FOXA1 and breast cancer SNPs. *Breast Cancer Res., 15*(1), 303.

Kim, C., Tang, G., Pogue-Geile, K. L., Costantino, J. P., Baehner, F. L., Baker, J., et al. (2011). Estrogen receptor (ESR1) mRNA expression and benefit from tamoxifen in the treatment and prevention of estrogen receptor–positive breast cancer. *J. Clin. Oncol., 29*(31), 4160-4167.

Kim, H., Watkinson, J., Varadan, V., & Anastassiou, D. (2010). Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. *BMC Med. Genomics, 3*(1), 51.

Kim, S. J., Kang, H., Chang, H. L., Jung, Y. C., Sim, H., Lee, K. S., et al. (2008). Promoter hypomethylation of the N-acetyltransferase 1 gene in breast cancer. *Oncol. Rep., 19*(3), 663-668.

Kim, S. J., Nakayama, S., Miyoshi, Y., Taguchi, T., Tamaki, Y., Matsushima, T., et al. (2008). Determination of the specific activity of CDK1 and CDK2 as a novel prognostic indicator for early breast cancer. *Annals of oncology, 19*(1), 68-72.

Kim, S. J., Nakayama, S., Shimazu, K., Tamaki, Y., Akazawa, K., Tsukamoto, F., et al. (2012). Recurrence risk score based on the specific activity of CDK1 and CDK2 predicts

response to neoadjuvant paclitaxel followed by 5-fluorouracil, epirubicin and cyclophosphamide in breast cancers. *Annals of oncology, 23*(4), 891-897.

Kim, S. Y., Yang, D., Myeong, J., Ha, K., Kim, S. H., Park, E. J., et al. (2013). Regulation of calcium influx and signaling pathway in cancer cells via TRPV6-Numb1 interaction. *Cell Calcium, 53*(2), 102-111.

Klopocki, E., Kristiansen, G., Wild, P. J., Klaman, I., Castanos-Velez, E., Singer, G., et al. (2004). Loss of SFRP1 is associated with breast cancer progression and poor prognosis in early stage tumors. *Int. J. Oncol., 25*, 641-649.

Kong, S. L., Li, G., Loh, S. L., Sung, W. K., & Liu, E. T. (2011). Cellular reprogramming by the conjoint action of ERα, FOXA1, and GATA3 to a ligand-inducible growth state. *Mol. Syst. Biol., 7*(1), 526.

Kotoula, V., Kalogeras, K. T., Kouvatseas, G., Televantou, D., Kronenwett, R., Wirtz, R. M., et al. (2013). Sample parameters affecting the clinical relevance of RNA biomarkers in translational breast cancer research. *Virchows Arch., 462*(2), 141-154.

Kwon, Y. J., Hurst, D. R., Steg, A. D., Yuan, K., Vaidya, K. S., Welch, D. R., et al. (2011). Gli1 enhances migration and invasion via up-regulation of MMP-11 and promotes metastasis in ERalpha negative breast cancer cell lines. *Clin Exp Metastasis, 28*(5), 437-449.

Lacroix, M. (2006). Significance, detection and markers of disseminated breast cancer cells. *Endocr. Relat. Cancer, 13*(4), 1033-1067.

Lan, J., Zhao, J., & Liu, Y. (2012). Molecular cloning, sequence characterization, polymorphism and association analysis of porcine ROPN1 gene. *Mol. Biol. Rep., 39*(3), 2739-2743.

Landowski, C. P., Bolanz, K. A., Suzuki, Y., & Hediger, M. A. (2011). Chemical inhibitors of the calcium entry channel TRPV6. *Pharm. Res., 28*(2), 322-330.

Lasa, A., Garcia, A., Alonso, C., Millet, P., Cornet, M., Ramon y Cajal, T., et al. (2013). Molecular detection of peripheral blood breast cancer mRNA transcripts as a surrogate biomarker for circulating tumor cells. *PLoS One, 8*(9), e74079.

Li, X., Cowell, J. K., & Sossey-Alaoui, K. (2004). CLCA2 tumour suppressor gene in 1p31 is epigenetically regulated in breast cancer. *Oncogene, 23*(7), 1474-1480.

Li, Y., Wu, J., Zhang, W., Zhang, N., & Guo, H. (2013). Identification of serum CCL15 in hepatocellular carcinoma. *Br. J. Cancer, 108*(1), 99-106.

Lin, M.-L., Park, J.-H., Nishidate, T., Nakamura, Y., & Katagiri, T. (2007). Involvement of maternal embryonic leucine zipper kinase (MELK) in mammary carcinogenesis through interaction with Bcl-G, a pro-apoptotic member of the Bcl-2 family. *Breast Cancer Res., 9*(1), R17.

Liu, L., Liu, Z., Qu, S., Zheng, Z., Liu, Y., Xie, X., et al. (2013). Small breast epithelial mucin tumor tissue expression is associated with increased risk of recurrence and death in triple-negative breast cancer patients. *Diagn. Pathol., 8*(1), 71.

Liu, N., Yu, Q., Liu, T. J., Gebreamlak, E. P., Wang, S. L., Zhang, R. J., et al. (2012). P-cadherin expression and basal-like subtype in breast cancers. *Med Oncol, 29*(4), 2606-2612.

Liu, Q., Li, J. G., Zheng, X. Y., Jin, F., & Dong, H. T. (2009). Expression of CD133, PAX2, ESA, and GPR30 in invasive ductal breast carcinomas. *Chin Med J (Engl), 122*(22), 2763-2769.

Liu, X. F., Li, L. F., Ou, Z. L., Shen, R., & Shao, Z. M. (2012). Correlation between Duffy blood group phenotype and breast cancer incidence. *BMC Cancer, 12*(1), 374.

Liu, Z. Z., Xie, X. D., Qu, S. X., Zheng, Z. D., & Wang, Y. K. (2010). Small breast epithelial mucin (SBEM) has the potential to be a marker for predicting hematogenous micrometastasis and response to neoadjuvant chemotherapy in breast cancer. *Clin Exp Metastasis, 27*(4), 251-259.

Lochhead, P., Imamura, Y., Morikawa, T., Kuchiba, A., Yamauchi, M., Liao, X., et al. (2012). Insulin-like growth factor 2 messenger RNA binding protein 3 (IGF2BP3) is a marker of unfavourable prognosis in colorectal cancer. *Eur. J. Cancer, 48*(18), 3405-3413.

Lopez, F.-J., Cuadros, M., Cano, C., Concha, A., & Blanco, A. (2012). Biomedical application of fuzzy association rules for identifying breast cancer biomarkers. *Med. Biol. Eng. Comput., 50*(9), 981-990.

Loussouarn, D., Campion, L., Leclair, F., Campone, M., Charbonnel, C., Ricolleau, G., et al. (2009). Validation of UBE2C protein as a prognostic marker in node-positive breast cancer. *Br. J. Cancer, 101*(1), 166-173.

Magnani, L., & Lupien, M. (2014). Chromatin and epigenetic determinants of estrogen receptor alpha (ESR1) signaling. *Mol. Cell. Endocrinol., 382*(1), 633-641.

Mahasenan, K. V., & Li, C. (2012). Novel inhibitor discovery through virtual screening against multiple protein conformations generated via ligand-directed modeling: a maternal embryonic leucine zipper kinase example. *J. Chem. Inf. Model., 52*(5), 1345-1355.

Martens, J. W., Nimmrich, I., Koenig, T., Look, M. P., Harbeck, N., Model, F., et al. (2005). Association of DNA methylation of phosphoserine aminotransferase with response to endocrine therapy in patients with recurrent breast cancer. *Cancer Res., 65*(10), 4101-4117.

Martin-Manso, G., Calzada, M. J., Chuman, Y., Sipes, J. M., Xavier, C. P., Wolf, V., et al. (2011). sFRP-1 binds via its netrin-related motif to the N-module of thrombospondin-1 and blocks thrombospondin-1 stimulation of MDA-MB-231 breast carcinoma cell adhesion and migration. *Arch. Biochem. Biophys., 509*(2), 147-156.

Martin, K. J., Patrick, D. R., Bissell, M. J., & Fournier, M. V. (2008). Prognostic breast cancer signature identified from 3D culture model accurately predicts clinical outcome across independent datasets. *PLoS One, 3*(8), e2994.

Matsuda, Y., Schlange, T., Oakeley, E. J., Boulay, A., & Hynes, N. E. (2009). WNT signaling enhances breast cancer cell motility and blockade of the WNT pathway by sFRP1 suppresses MDA-MB-231 xenograft growth. *Breast Cancer Res., 11*(3), R32.

Mehta, R. J., Jain, R. K., Leung, S., Choo, J., Nielsen, T., Huntsman, D., et al. (2012). FOXA1 is an independent prognostic marker for ER-positive breast cancer. *Breast Cancer Res. Treat., 131*(3), 881-890.

Mertsch, S., Schmitz, N., Jeibmann, A., Geng, J. G., Paulus, W., & Senner, V. (2008). Slit2 involvement in glioma cell migration is mediated by Robo1 receptor. *J Neurooncol, 87*(1), 1-7.

Meyer, K. B., & Carroll, J. S. (2012). FOXA1 and breast cancer risk. *Nat. Genet., 44*(11), 1176.

Mihály, Z., Kormos, M., Lánczky, A., Dank, M., Budczies, J., Szász, M. A., et al. (2013). A meta-analysis of gene expression-based biomarkers predicting outcome after tamoxifen treatment in breast cancer. *Breast Cancer Res. Treat., 140*(2), 219-232.

Miksicek, R. J., Myal, Y., Watson, P. H., Walker, C., Murphy, L. C., & Leygue, E. (2002). Identification of a novel breast-and salivary gland-specific, mucin-like gene strongly expressed in normal and tumor human mammary epithelium. *Cancer Res., 62*(10), 2736-2740.

Miller, W. R., & Larionov, A. (2010). Changes in expression of oestrogen regulated and proliferation genes with neoadjuvant treatment highlight heterogeneity of clinical resistance to the aromatase inhibitor, letrozole. *Breast Cancer Res., 12*(4), R52.

Min, K.-W., Kim, D.-H., Do, S.-I., Pyo, J.-S., Kim, K., Chae, S. W., et al. (2013). Diagnostic and prognostic relevance of MMP-11 expression in the stromal fibroblast-like cells adjacent to invasive ductal carcinoma of the breast. *Ann. Surg. Oncol., 20*(3), 433-442.

Min, K. W., Kim, D. H., Do, S. I., Pyo, J. S., Kim, K., Chae, S. W., et al. (2012). Diagnostic and Prognostic Relevance of MMP-11 Expression in the Stromal Fibroblast-Like Cells Adjacent to Invasive Ductal Carcinoma of the Breast. *Annals of surgical oncology, 3*, S433-442.

Moy, I., Todorovic, V., Dubash, A. D., Coon, J. S., Parker, J. B., Buranapramest, M., et al. (2014). Estrogen-dependent sushi domain containing 3 regulates cytoskeleton organization and migration in breast cancer cells. *Oncogene, 34*(3), 323-333.

Mukherjee, N., Bhattacharya, N., Alam, N., Roy, A., Roychoudhury, S., & Panda, C. K. (2012). Subtype-specific alterations of the Wnt signaling pathway in breast cancer: Clinical and prognostic significance. *Cancer Sci., 103*(2), 210-220.

Murphy, K. J., Ter Horst, J. P., Cassidy, A. W., DeSouza, I. E., Morgunova, M., Li, C., et al. (2010). Temporal dysregulation of cortical gene expression in the isolation reared Wistar rat. *J Neurochem, 113*(3), 601-614.

Nadal, R., Ortega, F. G., Salido, M., Lorente, J. A., Rodríguez-Rivera, M., Delgado-Rodríguez, M., et al. (2013). CD133 expression in circulating tumor cells from breast cancer patients: potential role in resistance to chemotherapy. *Int. J. Cancer, 133*(10), 2398-2407.

Naderi, A., Meyer, M., & Dowhan, D. H. (2012). Cross-regulation between FOXA1 and ErbB2 signaling in estrogen receptor-negative breast cancer. *Neoplasia, 14*(4), 283-296.

Nguyen, M.-H., Koinuma, J., Ueda, K., Ito, T., Tsuchiya, E., Nakamura, Y., et al. (2010). Phosphorylation and activation of cell division cycle associated 5 by mitogen-activated protein kinase play a crucial role in human lung carcinogenesis. *Cancer Res., 70*(13), 5337-5347.

Ni, M., Chen, Y., Lim, E., Wimberly, H., Bailey, S. T., Imai, Y., et al. (2011). Targeting androgen receptor in estrogen receptor-negative breast cancer. *Cancer Cell, 20*(1), 119-131.

Park, S. Y., Kwon, H. J., Choi, Y., Lee, H. E., Kim, S. W., Kim, J. H., et al. (2012). Distinct patterns of promoter CpG island methylation of breast cancer subtypes are associated with stem cell phenotypes. *Modern pathology, 25*(2), 185-196.

Parris, T. Z., Kovács, A., Aziz, L., Hajizadeh, S., Nemes, S., Semaan, M., et al. (2014). Additive effect of the AZGP1, PIP, S100A8 and UBE2C molecular biomarkers improves outcome prediction in breast carcinoma. *Int. J. Cancer, 134*(7), 1617-1629.

Persson, S., Rosenquist, M., Knoblach, B., Khosravi-Far, R., Sommarin, M., & Michalak, M. (2005). Diversity of the protein disulfide isomerase family: identification of breast tumor induced Hag2 and Hag3 as novel members of the protein family. *Mol. Phylogenet. Evol., 36*(3), 734-740.

Peters, A. A., Simpson, P. T., Bassett, J. J., Lee, J. M., Da Silva, L., Reid, L. E., et al. (2012). Calcium Channel TRPV6 as a Potential Therapeutic Target in Estrogen Receptor–Negative Breast Cancer. *Mol. Cancer Ther., 11*(10), 2158-2168.

Pickard, M. R., Green, A. R., Ellis, I. O., Caldas, C., Hedge, V. L., Mourtada-Maarabouni, M., et al. (2009). Dysregulated expression of Fau and MELK is associated with poor prognosis in breast cancer. *Breast Cancer Res., 11*(4), R60.

Psyrri, A., Kalogeras, K. T., Kronenwett, R., Wirtz, R. M., Batistatou, A., Bournakis, E., et al. (2012). Prognostic significance of UBE2C mRNA expression in high-risk early breast cancer. A Hellenic Cooperative Oncology Group (HeCOG) Study. *Annals of oncology, 23*(6), 1422-1427.

Qian, Y., Shen, L., Cheng, L., Wu, Z., & Yao, H. (2011). B7-H4 expression in various tumors determined using a novel developed monoclonal antibody. *Clin. Exp. Med., 11*(3), 163-170.

Rawat, A., Gopal, G., Selvaluxmy, G., & Rajkumar, T. (2013). Inhibition of ubiquitin conjugating enzyme UBE2C reduces proliferation and sensitizes breast cancer cells to radiation, doxorubicin, tamoxifen and letrozole. *Cell. Oncol., 36*(6), 459-467.

Robinson, J. L., & Carroll, J. S. (2012). FoxA1 is a key mediator of hormonal response in breast and prostate cancer. *Front. Endocrinol. (Lausanne), 3*, 68.

Robinson, J. L., Holmes, K. A., & Carroll, J. S. (2013). FOXA1 mutations in hormone-dependent cancers. *Front. Oncol., 3*, 20.

Robinson, J. L., Macarthur, S., Ross-Innes, C. S., Tilley, W. D., Neal, D. E., Mills, I. G., et al. (2011). Androgen receptor driven transcription in molecular apocrine breast cancer is mediated by FoxA1. *Embo J, 30*(15), 3019-3027.

Roll, J. D., Rivenbark, A. G., Sandhu, R., Parker, J. S., Jones, W. D., Carey, L. A., et al. (2013). Dysregulation of the epigenome in triple-negative breast cancers: basal-like and claudin-low breast cancers express aberrant DNA hypermethylation. *Exp Mol Pathol, 95*(3), 276-287.

Romanelli, A., Clark, A., Assayag, F., Chateau-Joubert, S., Poupon, M. F., Servely, J. L., et al. (2012). Inhibiting Aurora Kinases Reduces Tumor Growth and Suppresses Tumor Recurrence after Chemotherapy in Patient-Derived Triple-Negative Breast Cancer Xenografts. *Mol. Cancer Ther., 11*(12), 2693-2703.

Ruan, Q., Han, S., Jiang, W. G., Boulton, M. E., Chen, Z. J., Law, B. K., et al. (2011). αB-crystallin, an effector of unfolded protein response, confers anti-VEGF resistance to breast cancer via maintenance of intracrine VEGF in endothelial cells. *Mol. Cancer Res., 9*(12), 1632-1643.

Salceda, S., Tang, T., Kmet, M., Munteanu, A., Ghosh, M., Macina, R., et al. (2005). The immunomodulatory protein B7-H4 is overexpressed in breast and ovarian cancers and promotes epithelial cell transformation. *Exp. Cell Res., 306*(1), 128-141.

Samanta, S., Pursell, B., & Mercurio, A. M. (2013). IMP3 protein promotes chemoresistance in breast cancer cells by regulating breast cancer resistance protein (ABCG2) expression. *The Journal of biological chemistry, 288*(18), 12569-12573.

Samanta, S., Sharma, V. M., Khan, A., & Mercurio, A. M. (2012). Regulation of IMP3 by EGFR signaling and repression by ERbeta: implications for triple-negative breast cancer. *Oncogene, 31*(44), 4689-4697.

Sanchez-Bailon, M. P., Calcabrini, A., Gomez-Dominguez, D., Morte, B., Martin-Forero, E., Gomez-Lopez, G., et al. (2012). Src kinases catalytic activity regulates proliferation, migration and invasiveness of MDA-MB-231 breast cancer cells. *Cellular signalling, 24*(6), 1276-1286.

Sandhu, R., Rivenbark, A. G., Mackler, R. M., Livasy, C. A., & Coleman, W. B. (2014). Dysregulation of microRNA expression drives aberrant DNA hypermethylation in basal-like breast cancer. *Int. J. Oncol., 44*(2), 563-572.

Sano, H., Wada, S., Eguchi, H., Osaki, A., Saeki, T., & Nishiyama, M. (2012). Quantitative prediction of tumor response to neoadjuvant chemotherapy in breast cancer: novel marker genes and prediction model using the expression levels. *Breast Cancer, 19*(1), 37-45.

Sasaki, Y., Koyama, R., Maruyama, R., Hirano, T., Tamura, M., Sugisaka, J., et al. (2012). CLCA2, a target of the p53 family, negatively regulates cancer cell migration and invasion. *Cancer Biol. Ther., 13*(14), 1512-1521.

Seil, I., Frei, C., Sültmann, H., Knauer, S. K., Engels, K., Jäger, E., et al. (2007). The differentiation antigen NY-BR-1 is a potential target for antibody-based therapies in breast cancer. *Int. J. Cancer, 120*(12), 2635-2642.

Selvey, S., Haupt, L. M., Thompson, E. W., Matthaei, K. I., Irving, M. G., & Griffiths, L. R. (2004). Stimulation of MMP-11 (stromelysin-3) expression in mouse fibroblasts by cytokines, collagen and co-culture with human breast cancer cell lines. *BMC Cancer, 4*(1), 40.

Shulewitz, M., Soloviev, I., Wu, T., Koeppen, H., Polakis, P., & Sakanaka, C. (2006). Repressor roles for TCF-4 and Sfrp1 in Wnt signaling in breast cancer. *Oncogene, 25*(31), 4361-4369.

Sim, E., Walters, K., & Boukouvala, S. (2008). Arylamine N-acetyltransferases: from structure to function. *Drug Metab. Rev., 40*(3), 479-510.

Sizemore, S. T., & Keri, R. A. (2012). The forkhead box transcription factor FOXC1 promotes breast cancer invasion by inducing matrix metalloprotease 7 (MMP7) expression. *J. Biol. Chem., 287*(29), 24631-24640.

Skliris, G. P., Hube, F., Gheorghiu, I., Mutawe, M. M., Penner, C., Watson, P. H., et al. (2008). Expression of small breast epithelial mucin (SBEM) protein in tissue microarrays (TMAs) of primary invasive breast cancers. *Histopathol, 52*(3), 355-369.

Soncini, C., Carpinelli, P., Gianellini, L., Fancelli, D., Vianello, P., Rusconi, L., et al. (2006). PHA-680632, a novel Aurora kinase inhibitor with potent antitumoral activity. *Clin. Cancer Res., 12*(13), 4080-4089.

Spicakova, T., O'Brien, M. M., Duran, G. E., Sweet-Cordero, A., & Sikic, B. I. (2010). Expression and silencing of the microtubule-associated protein Tau in breast cancer cells. *Mol. Cancer Ther., 9*(11), 2970-2981.

Stossi, F., Madak-Erdogan, Z., & Katzenellenbogen, B. S. (2012). Macrophage-elicited loss of estrogen receptor-alpha in breast cancer cells via involvement of MAPK and c-Jun at the ESR1 genomic locus. *Oncogene, 31*(14), 1825-1834.

Suh, W.-K., Wang, S., Duncan, G. S., Miyazaki, Y., Cates, E., Walker, T., et al. (2006). Generation and characterization of B7-H4/B7S1/B7x-deficient mice. *Mol. Cell. Biol., 26*(17), 6403-6411.

Suzuki, H., Toyota, M., Caraway, H., Gabrielson, E., Ohmura, T., Fujikane, T., et al. (2008). Frequent epigenetic inactivation of Wnt antagonist genes in breast cancer. *Br. J. Cancer, 98*(6), 1147-1156.

Symmans, W., Fiterman, D., Anderson, S., Ayers, M., Rouzier, R., Dunmire, V., et al. (2005). A single-gene biomarker identifies breast cancers associated with immature cell type and short duration of prior breastfeeding. *Endocr. Relat. Cancer, 12*(4), 1059-1069.

Tan, J., Buache, E., Alpy, F., Daguenet, E., Tomasetto, C. L., Ren, G. S., et al. (2013). Stromal matrix metalloproteinase-11 is involved in the mammary gland postnatal development. *Oncogene, 33*(31), 4050-4059.

Tanaka, S., Nohara, T., Iwamoto, M., Sumiyoshi, K., Kimura, K., Takahashi, Y., et al. (2009). Tau expression and efficacy of paclitaxel treatment in metastatic breast cancer. *Cancer Chemother. Pharmacol., 64*(2), 341-346.

Taylor, K. J., Sims, A. H., Liang, L., Faratian, D., Muir, M., Walker, G., et al. (2010). Dynamic changes in gene expression in vivo predict prognosis of tamoxifen-treated patients with breast cancer. *Breast Cancer Res., 12*(3), R39.

Theurillat, J.-P., Zürrer-Härdi, U., Varga, Z., Storz, M., Probst-Hensch, N. M., Seifert, B., et al. (2007). NY-BR-1 protein expression in breast carcinoma: a mammary gland differentiation antigen as target for cancer immunotherapy. *Cancer Immunol. Immunother., 56*(11), 1723-1731.

Theurillat, J. P., Zürrer-Härdi, U., Varga, Z., Barghorn, A., Saller, E., Frei, C., et al. (2008). Distinct expression patterns of the immunogenic differentiation antigen NY-BR-1 in normal breast, testis and their malignant counterparts. *Int. J. Cancer, 122*(7), 1585-1591.

Tkocz, D., Crawford, N. T., Buckley, N. E., Berry, F. B., Kennedy, R. D., Gorski, J. J., et al. (2012). BRCA1 and GATA3 corepress FOXC1 to inhibit the pathogenesis of basal-like breast cancers. *Oncogene, 31*(32), 3667-3678.

Torikoshi, Y., Gohda, K., Davis, M. L., Symmans, W. F., Pusztai, L., Kazansky, A., et al. (2013). Novel functional assay for spindle-assembly checkpoint by cyclin-dependent kinase activity to predict taxane chemosensitivity in breast tumor patient. *J Cancer, 4*(9), 697-702.

Tringler, B., Zhuo, S., Pilkington, G., Torkko, K. C., Singh, M., Lucia, M. S., et al. (2005). B7-h4 is highly expressed in ductal and lobular breast cancer. *Clinical cancer research, 11*(5), 1842-1848.

Ueki, T., Nishidate, T., Park, J. H., Lin, M. L., Shimo, A., Hirata, K., et al. (2008). Involvement of elevated expression of multiple cell-cycle regulator, DTL/RAMP (denticleless/RA-

regulated nuclear matrix associated protein), in the growth of breast cancer cells. *Oncogene, 27*(43), 5672-5683.

Ueki, T., Park, J.-H., Nishidate, T., Kijima, K., Hirata, K., Nakamura, Y., et al. (2009). Ubiquitination and downregulation of BRCA1 by ubiquitin-conjugating enzyme E2T overexpression in human breast cancer cells. *Cancer Res., 69*(22), 8752-8760.

Ugolini, F., Charafe-Jauffret, E., Bardou, V. J., Geneix, J., Adelaide, J., Labat-Moleur, F., et al. (2001). WNT pathway and mammary carcinogenesis: loss of expression of candidate tumor suppressor gene SFRP1 in most invasive carcinomas except of the medullary type. *Oncogene, 20*(41), 5810-5817.

Valet, F., de Cremoux, P., Spyratos, F., Servant, N., Dujaric, M. E., Gentien, D., et al. (2013). Challenging single- and multi-probesets gene expression signatures of pathological complete response to neoadjuvant chemotherapy in breast cancer: experience of the REMAGUS 02 phase II trial. *Breast, 22*(6), 1052-1059.

Valladares-Ayerbes, M., Iglesias-Díaz, P., Díaz-Prado, S., Ayude, D., Medina, V., Haz, M., et al. (2009). Diagnostic accuracy of small breast epithelial mucin mRNA as a marker for bone marrow micrometastasis in breast cancer: a pilot study. *J. Cancer Res. Clin. Oncol., 135*(9), 1185-1195.

Van De Rijn, M., Perou, C. M., Tibshirani, R., Haas, P., Kallioniemi, O., Kononen, J., et al. (2002). Expression of cytokeratins 17 and 5 identifies a group of breast carcinomas with poor clinical outcome. *Am. J. Pathol., 161*(6), 1991-1996.

van Zalen, S., Nijenhuis, M., Jonkman, M. F., & Pas, H. H. (2006). Two major 5'-untranslated regions for type XVII collagen mRNA. *J. Dermatol. Sci., 43*(1), 11-19.

Varga, Z., Theurillat, J.-P., Filonenko, V., Sasse, B., Odermatt, B., Jungbluth, A. A., et al. (2006). Preferential nuclear and cytoplasmic NY-BR-1 protein expression in primary breast cancer and lymph node metastases. *Clin. Cancer Res., 12*(9), 2745-2751.

Vargas, A. C., McCart Reed, A. E., Waddell, N., Lane, A., Reid, L. E., Smart, C. E., et al. (2012a). Gene expression profiling of tumour epithelial and stromal compartments during breast cancer progression. *Breast Cancer Res. Treat., 135*(1), 153-165.

Vargas, A. C., McCart Reed, A. E., Waddell, N., Lane, A., Reid, L. E., Smart, C. E., et al. (2012b). Gene expression profiling of tumour epithelial and stromal compartments during breast cancer progression. *Breast cancer research and treatment, 135*(1), 153-165.

Veeck, J., Niederacher, D., An, H., Klopocki, E., Wiesmann, F., Betz, B., et al. (2006). Aberrant methylation of the Wnt antagonist SFRP1 in breast cancer is associated with unfavourable prognosis. *Oncogene, 25*(24), 3479-3488.

Walia, V., Ding, M., Kumar, S., Nie, D., Premkumar, L. S., & Elble, R. C. (2009). hCLCA2 Is a p53-Inducible Inhibitor of Breast Cancer Cell Proliferation. *Cancer Res., 69*(16), 6624-6632.

Walia, V., Yu, Y., Cao, D., Sun, M., McLean, J. R., Hollier, B. G., et al. (2012). Loss of breast epithelial marker hCLCA2 promotes epithelial-to-mesenchymal transition and indicates higher risk of metastasis. *Oncogene, 31*(17), 2237-2246.

Walter, O., Prasad, M., Lu, S., Quinlan, R. M., Edmiston, K. L., & Khan, A. (2009). IMP3 is a novel biomarker for triple negative invasive mammary carcinoma associated with a more aggressive phenotype. *Hum. Pathol., 40*(11), 1528-1533.

Waluk, D. P., Schultz, N., & Hunt, M. C. (2010). Identification of glycine N-acyltransferase-like 2 (GLYATL2) as a transferase that produces N-acyl glycines in humans. *FASEB Journal, 24*(8), 2795-2803.

Waluk, D. P., Sucharski, F., Sipos, L., Silberring, J., & Hunt, M. C. (2012). Reversible lysine acetylation regulates activity of human glycine N-acyltransferase-like 2 (hGLYATL2): implications for production of glycine-conjugated signaling molecules. *The Journal of biological chemistry, 287*(20), 16158-16167.

Wang, J., He, Q., Shao, Y. G., & Ji, M. (2013). Chemokines fluctuate in the progression of primary breast cancer. *European Review for Medical and Pharmacological Sciences, 17*, 596-608.

Wang, J., Ray, P. S., Sim, M. S., Zhou, X. Z., Lu, K. P., Lee, A. V., et al. (2012). FOXC1 regulates the functions of human basal-like breast cancer cells by activating NF-kappaB signaling. *Oncogene, 31*(45), 4798-4802.

Wang, K., Deng, Q. T., Liao, N., Zhang, G. C., Liu, Y. H., Xu, F. P., et al. (2013). Tau expression correlated with breast cancer sensitivity to taxanes-based neoadjuvant chemotherapy. *Tumour biology, 34*(1), 33-38.

Warsow, G., Struckmann, S., Kerkhoff, C., Reimer, T., Engel, N., & Fuellen, G. (2013). Differential network analysis applied to preoperative breast cancer chemotherapy response. *PLoS One, 8*(12), e81784.

Weigelt, B., Verduijn, P., Bosma, A. J., Rutgers, E. J., Peterse, H. L., & van't Veer, L. J. (2004). Detection of metastases in sentinel lymph nodes of breast cancer patients by multiple mRNA markers. *Br. J. Cancer, 90*(8), 1531-1537.

Won, J. R., Gao, D., Chow, C., Cheng, J., Lau, S. Y., Ellis, M. J., et al. (2013). A survey of immunohistochemical biomarkers for basal-like breast cancer against a gene expression profile gold standard. *Modern pathology, 26*(11), 1438-1450.

Woodard, A. H., Yu, J., Dabbs, D. J., Beriwal, S., Florea, A. V., Elishaev, E., et al. (2011). NY-BR-1 and PAX8 immunoreactivity in breast, gynecologic tract, and other CK7+ carcinomas: potential use for determining site of origin. *Am J Clin Pathol, 136*(3), 428-435.

Wu, J., Liu, S., Liu, G., Dombkowski, A., Abrams, J., Martin-Trevino, R., et al. (2012). Identification and functional analysis of 9p24 amplified genes in human breast cancer. *Oncogene, 31*(3), 333-341.

Xia, Q., Cai, Y., Peng, R., Wu, G., Shi, Y., & Jiang, W. (2014). The CDK1 inhibitor RO3306 improves the response of BRCA-pro fi cient breast cancer cells to PARP inhibition. *Int. J. Oncol., 44*(3), 735-744.

Xing, P., Li, J.-g., Jin, F., Zhao, T.-t., Liu, Q., Dong, H.-t., et al. (2011). Clinical and Biological Significance of Hepsin Overexpression in Breast Cancer. *J. Investig. Med., 59*(5), 803-810.

Yamaguchi, N., Ito, E., Azuma, S., Honma, R., Yanagisawa, Y., Nishikawa, A., et al. (2008). FoxA1 as a lineage-specific oncogene in luminal type breast cancer. *Biochem. Biophys. Res. Commun., 365*(4), 711-717.

Yamamura, J., Miyoshi, Y., Tamaki, Y., Taguchi, T., Iwao, K., Monden, M., et al. (2004). mRNA expression level of estrogen-inducible gene, alpha 1-antichymotrypsin, is a predictor of early tumor recurrence in patients with invasive breast cancers. *Cancer Sci., 95*(11), 887-892.

Yang, Z. Q., Liu, G., Bollig-Fischer, A., Haddad, R., Tarca, A. L., & Ethier, S. P. (2009). Methylation-associated silencing of SFRP1 with an 8p11-12 amplification inhibits canonical and non-canonical WNT pathways in breast cancers. *Int. J. Cancer, 125*(7), 1613-1621.

Yu, Y.-N., Yip, G. W.-C., Tan, P.-H., Thike, A. A., Matsumoto, K., Tsujimoto, M., et al. (2010). Y-box binding protein 1 is up-regulated in proliferative breast cancer and its inhibition deregulates the cell cycle. *Int. J. Oncol., 37*(2), 483.

Zeng, X.-H., Ou, Z.-L., Yu, K.-D., Feng, L.-Y., Yin, W.-J., Li, J., et al. (2011). Coexpression of atypical chemokine binders (ACBs) in breast cancer predicts better outcomes. *Breast Cancer Res. Treat., 125*(3), 715-727.

Zhang, W., Kawanishi, M., Miyake, K., Kagawa, M., Kawai, N., Murao, K., et al. (2010). Association between YKL-40 and adult primary astrocytoma. *Cancer, 116*(11), 2688-2697.

Zhang, W., Peng, G., Lin, S.-Y., & Zhang, P. (2011). DNA Damage Response Is Suppressed by the High Cyclin-dependent Kinase 1 Activity in Mitotic Mammalian Cells. *J. Biol. Chem., 286*(41), 35899-35905.

Zvelebil, M., Oliemuller, E., Gao, Q., Wansbury, O., Mackay, A., Kendrick, H., et al. (2013). Embryonic mammary signature subsets are activated in Brca1 -/- and basal-like breast cancers. *Breast Cancer Res., 15*(2), R25.

CHAPTER 5

5. ITERATIVELY REFINING THE METABRIC SUBTYPE LABELS

As discussed in *Chapter 4*, a thorough review of the five intrinsic subtypes is essential to investigators in the field, given the importance of the METABRIC data set to breast cancer research. Then, *Chapter 5* consists in the application of an iterative approach to improve class prediction and label assignments across samples. The content is available as a 'short report' in *BMC BioData Mining*[7] and presented here in sections **5.1 Introduction**, **5.2 Methods**, **5.3 Results and Discussion**, **5.4 Conclusion**, **5.5 References** and **5.6 Supporting Information**. This analysis is based on an ensemble learning technique to assign the sample subtype label using a robust iterative approach. The new labelling is also compared with clinicopathological markers and patients' overall survival. The refinement of the METABRIC sample labels improves the source of fundamental science overall leading to more accurate outcomes for future clinical applications in medicine.

[7] Milioli, H.H.; Vimieiro, R.; Tishchenko, I.; Riveros, C.; Berretta, R.; Moscato, P. (2016). Iteratively refining breast cancer intrinsic subtypes in the METABRIC dataset. *BioData Mining*; 9:2.

5.1 Introduction

Translational research aims at bringing basic scientific discoveries into outcomes that help improve clinical decision-making. The PAM50 Breast Cancer Intrinsic Classifier (Parker et al., 2009) has lately been used to assign the molecular subtypes (luminal A, luminal B, HER2-enriched, basal-like and normal-like) (Hu et al., 2006; Perou et al., 2000; Sørlie et al., 2001; Sørlie et al., 2003) based on shrunken centroids of gene expression profiles (Tibshirani et al., 2002). It uses a Single Sample Predictor (SSP) model with an embedded 50-gene assay. In spite of the relevance of this method for clinical management, there are limited investigations in the literature that support the classification approach. Comparison with other methods showed only moderate agreement between subtype labels assigned, as well as independent clinical prognostic information (Ebbert et al., 2011; Haibe-Kains et al., 2012; Weigelt et al., 2010).

Other multi-gene signatures have also been reported within the molecular patterns strongly correlated to clinical prognosis (Fan et al., 2011; Wang et al., 2005), disease progression (Seoane et al., 2014; Venet et al., 2011), and patient survival (Naderi et al., 2006). Different methods, however, highlight a variety of gene lists of distinct size due to the analysis of diverse microarray data and platform technologies. Additionally, the methods currently applied also bring a pragmatic concern of using SSP models for predicting disease subtypes. Multiple classifiers or ensemble learning model, on the other hand, have compensated for poor learning algorithms by performing extra computation (Gómez-Ravetti & Moscato, 2008). Therefore, there is an urgent need for translating these novel strategies to provide more accurate predictions of clinicopathological outcome.

In 2012, METABRIC disclosed a rich gene expression cohort widely used for investigating breast cancer diseases (Curtis et al., 2012). In spite of the quality of this data set, there are some inconsistencies with regards to the subtype labels assigned in the original cohort. In our previous study (Milioli et al., 2015), a thorough review of the intrinsic subtypes was suggested and is, therefore, mandatory given the importance of this data set to breast cancer research. For this report, we then propose a more robust approach to iteratively refine the labels in the METABRIC data set, based on ensemble learning. The new labels are yet correlated to well-established clinicopathological markers and patient overall survival.

5.2 Methods

5.2.1 Transcriptomic Data Set

We use the transcriptomic data set disclosed by METABRIC (EGAS00000000083), which contains cDNA microarray profiling of about 2000 breast cancer samples performed on the Illumina HT-12 v3 platform (Illumina Human WG-v3) (Curtis et al., 2012). The samples were originally partitioned into two subsets: Discovery (997 samples) and Validation (989 samples), respectively used as training and test sets in our analysis. In this cohort, tumour samples were originally assigned on the five intrinsic subtypes (luminal A, luminal B, HER2-enriched, basal-like and normal-like) according to the PAM50 method (Parker et al., 2009).

5.2.2 The Refinement Method

The overview of the refinement method applied on the METABRIC data set is shown in **Figure 5.1**. The process is initialised with the discovery set and the original PAM50 labels as defined in Curtis et al. (2012). After computing the CM1 score (more details in *Chapter 4*, **Equation 4.1**), the top 10 highly discriminative probes (5 with the greatest positive CM1 score values – indicating up-regulated probes relative to the other subtypes, and 5 with the smallest negative values – representing down-regulation) are chosen for each class. The set of new features is used to train the 24 classifiers from the Weka software suite (Witten et al., 2016), where a 10-fold cross-validation is performed. If the majority of the classifiers agree on the same label, the sample is assigned with the corresponding subtype; otherwise it is marked as inconsistent and not further considered in the process. The stopping criterion is reached when there are no changes in the sample labels and feature set, or when the desired Fleiss' kappa value ($\kappa = 0.92$) is achieved between the previous and the current iteration steps. Values between 0.81 - 1 show *almost perfect agreement*, thus 0.92 is above the average for this interval.

When the stopping condition is fulfilled, the new list of features and sample labels are used for the training-test setting. Samples from the validation data set or previously marked as inconsistent are then classified by training the classifiers in the refined discovery set. However, in the training-test setting, at least two thirds of classifiers in the ensemble must agree on the same label for it to be assigned to a sample. As a larger data set is expected to provide more robustness, all the re-classified samples are run through the same refinement procedure again. The final outcome of this process is the set of refined features and the new labels.

Since many classifiers tend to perform best when trained on classes of equal sample size, we adjusted the number of patients in each subtype by looking at the minimum number of samples in one of the subgroups. The normal-like subtype is represented by only 58 samples; thus, the total number of samples used in the training is 290. For each other subtype, 58 samples are randomly chosen from the data set. The whole process is run ten times due to the interchangeable sample selection that weight the different gene expression information used for training purposes.

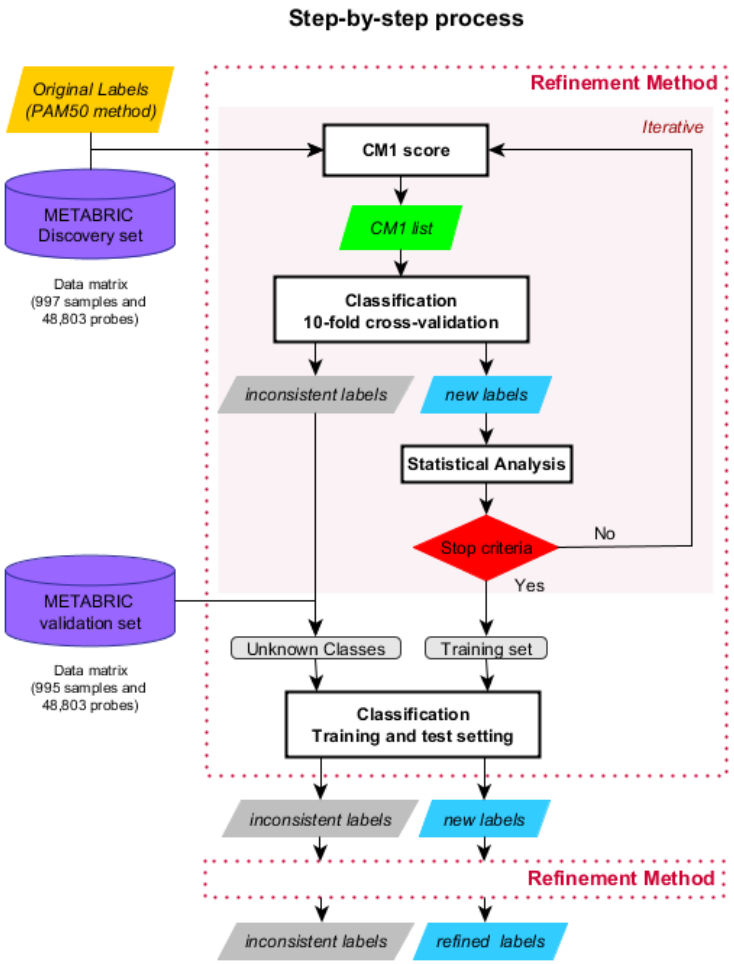


Figure 5.1 Refinement Method

The process is initialised with labels assigned using the PAM50 method. After computing the CM1 score, the top 10 highly discriminative probes are selected for each subtype. This set of features is used to train the 24 distinct classifiers for a 10-fold cross-validation classification. Samples are relabelled (eventually with the same label) if the classifiers agree in at least 50% of the cases; otherwise they are marked as inconsistent and not further considered in the iteration process. The stopping criterion is reached when there are no more changes in the sample labels or selected feature set, or when the desired Fleiss' kappa is achieved. After stopping, the final feature set and sample labels are used to classify the samples previously marked as inconsistent or from the validation data set. These samples are run through the same refinement procedure; inconsistent samples are reclassified and labels are refined.

5.2.3 The CM1 Score

The CM1 score, previously defined in *Chapter* 4, is a supervised method used to rank the variation of gene expression levels across samples from two different classes (Marsden et al., 2013; Milioli et al., 2015). The measure helps to identify the most discriminative features for each of the five breast cancer intrinsic subtypes: luminal A, luminal B, HER2-enriched, basal-like and normal-like. For a given subtype, we compute the CM1 score for each of the 48803 probes and select the 10 most discriminative ones. This happens iteratively in the refinement process each time the classifiers attribute a new label to a sample.

5.2.4 Statistical Analysis

Statistical measures have been computed in order to assess the quality of our results. We created a contingency table *r* x *c*, with "r" rows and "c" columns, comparing the predicted labels (rows) and labels from the previous refinement step (columns). Considering this table, we performed three tests, as follows:

- **Cramer's V** (Liebetrau, 1983) is used to measure the level of association between sample original and predicted labels (more details in *Chapter 4*, **Equation 4.2**).
- **Fleiss' kappa** (Fleiss, 1971; Fleiss et al., 2004) is a popular interrater reliability metric used to gauge the agreement between the original PAM50 labels and the labels assigned by the majority of classifiers (more details in *Chapter 4*, **Equation 4.4**).
- **Adjusted Rand Index (ARI)** (Hubert & Arabie, 1985; Vinh et al., 2009) measures the agreement between pairs of samples that are labelled either in the same class or in different classes (more details in *Chapter 4*, **Equation 4.5**).

5.2.5 Clinical Data and Survival Curves

The clinical markers oestrogen and progesterone receptors (ER and PR) and the human epidermal growth factor receptor 2 (HER2) are compared between original METABRIC labels and refined labels. Survival analysis was also performed, using Cox proportional hazards model from the package survival in the R software (Therneau & Grambsch, 2000). The *p*-value, used to test the null hypothesis that the curves stratified by subtype are identical in the overall population, is calculated using the log-rank test.

5.3 Results and Discussion

5.3.1 Discriminative Probes Used to Assign Intrinsic Subtype Labels in the Refinement Process

Samples were assigned into the five intrinsic subtypes based on the majority voting of classifiers (**Supporting Information – Table 5.3**), supported by their consistent performance across the ten runs (**Supporting Information – Table 5.4 and Table 5.5, Figure 5.4 and Figure 5.5**). During this procedure, 74 discriminative probes appeared (**Supporting Information – Table 5.6**) and, among then, 35 were recurrently selected (**Figure 5.2**). Overall, the association between the initial labels and those predicted using the ensemble learning (**Table 5.1**) was on average 0.95 according to Cramer's V. The consensus of sample labelling across different classifiers measured using Fleiss' kappa was 0.92. The ARI (1.00) also showed a maximum agreement between pairs of samples that are labelled either in the same or in different classes.

Table 5.1 Contingency table for predicted labels vs. initial subtypes (rows and columns, respectively)

Subtypes	Lum A	Lum B	HER2	Basal	Normal	Summary
Lum A	563	94	11	2	58	728
Lum B	102	383	77	19	19	600
HER2	7	1	149	59	18	234
Basal	0	0	0	230	3	233
Normal	33	0	1	15	95	144
Inconsistent	16	14	2	6	9	47
Summary	721	492	240	331	202	1986

Note: Columns represent the initial subtype labels, while rows contain the predicted labels. Breast cancer subtypes: Lum A – luminal A; Lum B – luminal B; HER2 – HER2-enriched; Basal – basal-like; Normal – normal-like.

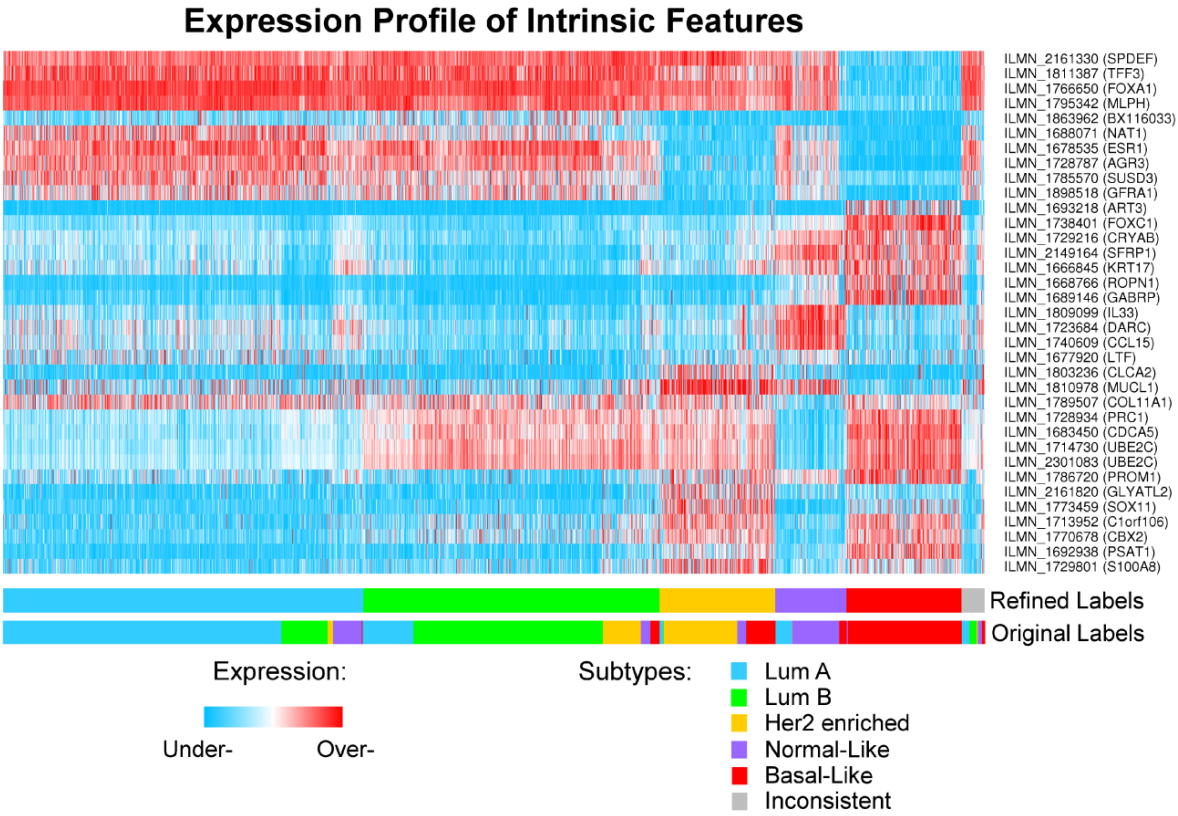


Figure 5.2 The heat map of refined intrinsic features selected using CM1 score

The heat map diagram exhibit 35 probes (rows) and 1992 samples (columns) from the discovery and validation sets ordered according to gene expression similarity. For visualisation, the expression values are normalised across the probes using a two-sided threshold of 1% (for under- and over-expression). The bars on the bottom show the sample distribution according to the refined and original labels assigned to the METABRIC cohort.

5.3.2 New Subtype Labels Reveal More Reliable Distribution of Clinical Markers and Survival Outcomes

We correlated the METABRIC and predicted labels with the current clinical markers ER, PR and *HER2*. **Table 5.2** shows the changes in number of samples across subtypes, labelled with the PAM50 method and refined labels, respectively. The refinement process improved the overall distribution to what is expected for each class: luminal A (ER+, PR+, HER2-), luminal B (ER+, PR±, HER2±), HER2-enriched (ER-, PR-, HER2+) and basal-like (ER-, PR-, HER2-); especially for HER2-enriched and basal-like subtypes. Samples labelled as inconsistent in our study may also reflect the heterogeneity of the disease and a hint to as-yet improperly characterised molecular subtypes.

Furthermore, the patient's overall survival significantly improved across subtypes when the original and refined labels are used to plot the curves for the METABRIC discovery and validation sets (**Figure 5.3**). The groups have a well-defined separation after the refinement process (p value 2.8 x 10^{-26}) compared to the original labels (p value 5.4 x 10^{-18}). These results also support a better characterisation of the intrinsic groups after the iterative approach.

Table 5.2 Number of samples for each clinical marker in the METABRIC data set according to the PAM50 method and refinement process

PAM50 method						
Class\Marker	**PR+**	**PR-**	**ER+**	**ER-**	**HER2+**	**HER2-**
Lum A	550	171	717	4	23	698
Lum B	309	183	492	0	45	447
HER2	51	189	98	142	135	105
Basal	29	302	41	290	30	301
Normal	106	96	164	38	16	186
Refinement process						
Class\Marker	**PR+**	**PR-**	**ER+**	**ER-**	**HER2+**	**HER2-**
Lum A	558	170	726	2	14	714
Lum B	358	242	599	1	83	517
HER2	11	223	19	215	139	95
Basal	7	226	9	224	4	229
Normal	85	59	115	29	4	140
Inconsistent	26	21	44	3	5	42
Summary	1045	941	1512	474	249	1737

Note: Clinical markers: PR – progesterone receptor; ER – estrogen receptor; HER2 – human epidermal growth factor receptor 2. Breast cancer subtypes: Lum A – luminal A; Lum B – luminal B; HER2 – HER2-enriched; Basal – basal-like; Normal – normal-like.

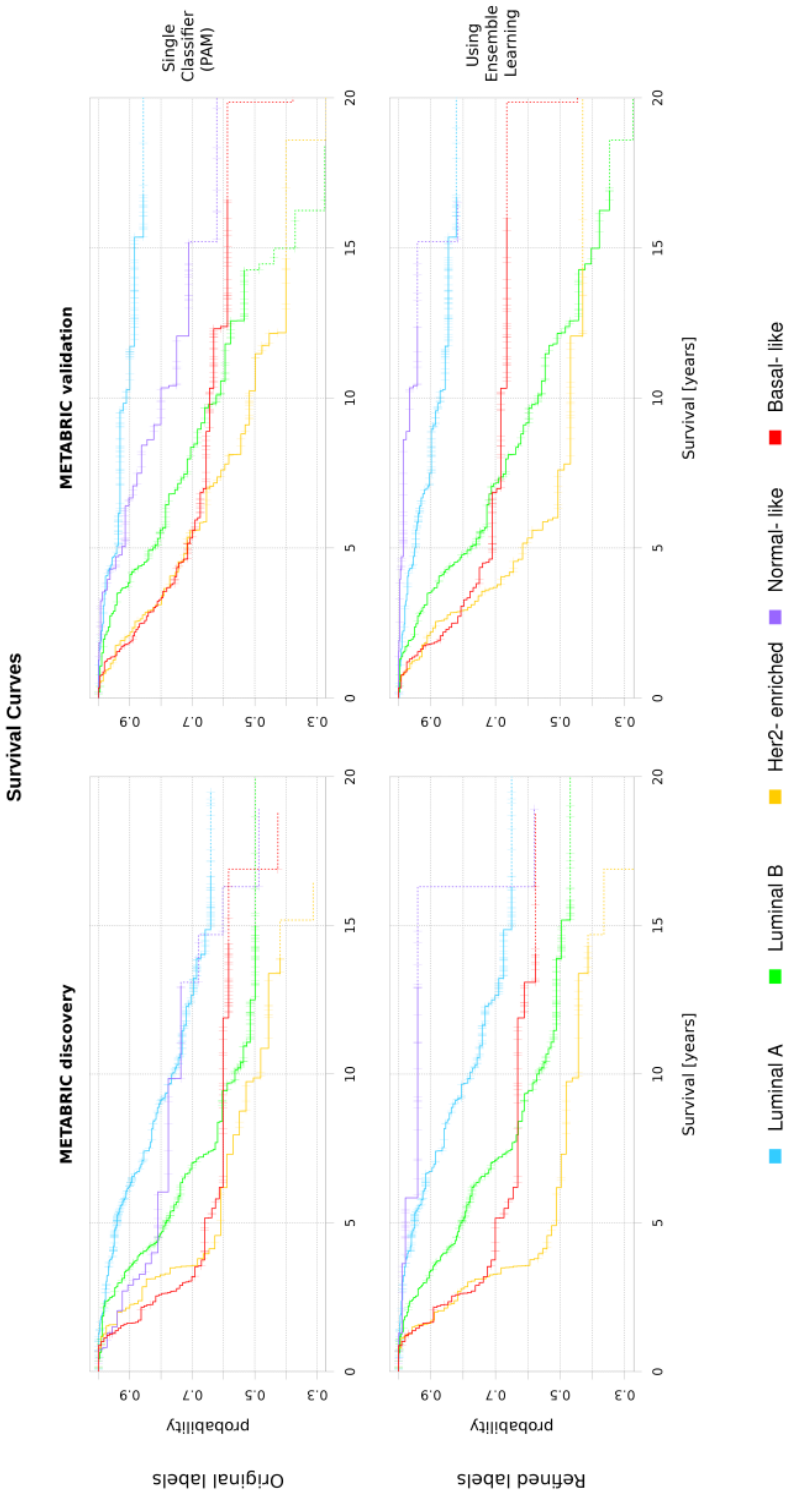


Figure 5.3 The survival curves for original and refined labels in the METABRIC discovery and validation sets

The survival curves for original and refined labels in the METABRIC discovery and validation sets. The survival curves for each breast cancer subtype are generated using Cox proportional hazard model. Each curve represents the survival probability at a certain time after the diagnosis. Drops in the curve indicate death. The probabilities of the last 10 observations are plotted in dash.

5.4 Conclusion

The iterative approach using CM1 score and ensemble learning has shown a great potential for predicting more accurate sample subtypes in the METABRIC breast cancer data set. The refined labels are of great value to breast cancer research and future clinical translational science (**Supporting Information – Text 1**). Given the relevance of accurate subtype assignments, we encourage researchers to consider the proposed refined labels when analysing the METABRIC data set.

5.5 References

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature, 486*(7403), 346-352.

Ebbert, M., Bastien, R. R., Boucher, K. M., Martin, M., Carrasco, E., Caballero, R., et al. (2011). Characterization of uncertainty in the classification of multivariate assays: application to PAM50 centroid-based genomic predictors for breast cancer treatment plans. *J Clin Bioinform, 1*(1), 37.

Fan, C., Prat, A., Parker, J., Liu, Y., Carey, L. A., Troester, M. A., et al. (2011). Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Med. Genomics, 4*(1), 3.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull., 76*(5), 378-382.

Fleiss, J. L., Levin, B., & Paik, M. C. (2004). The Measurement of Interrater Agreement *Statistical Methods for Rates and Proportions* (pp. 598-626). New York: John Wiley & Sons, Inc.

Gómez-Ravetti, M., & Moscato, P. (2008). Identification of a 5-Protein Biomarker Molecular Signature for Predicting Alzheimer's Disease. *PLoS ONE, 3*(9), e3111.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., et al. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *J. Natl. Cancer Inst., 104*(4), 311-325.

Hu, Z., Fan, C., Oh, D. S., Marron, J., He, X., Qaqish, B. F., et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics, 7*(1), 96.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193-218.

Liebetrau, A. M. (1983). *Measures of association* (Vol. 32). Beverly Hills, CA: SAGE Publications, Inc.

Marsden, J., Budden, D., Craig, H., & Moscato, P. (2013). Language Individuation and Marker Words: Shakespeare and His Maxwell's Demon. *PLoS One, 8*(6), e66813.

Milioli, H. H., Vimieiro, R., Riveros, C., Tishchenko, I., Berretta, R., & Moscato, P. (2015). The Discovery of Novel Biomarkers Improves Breast Cancer Intrinsic Subtype Prediction and Reconciles the Labels in the METABRIC Data Set. *PLoS One, 10*(7), e0129711.

Naderi, A., Teschendorff, A. E., Barbosa-Morais, N. L., Pinder, S. E., Green, A. R., Powe, D. G., et al. (2006). A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene, 26*(10), 1507-1516.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol., 27*(8), 1160-1167.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature, 406*(6797), 747-752.

Seoane, J. A., Day, I. N. M., Gaunt, T. R., & Campbell, C. (2014). A pathway-based data integration framework for prediction of disease progression. *Bioinformatics, 30*(6), 838-845.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A., 98*(19), 10869-10874.

Sørlie, T., Tibshirani, R., Parker, J. S., Hastie, T., Marron, J. S., Nobel, A., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U. S. A., 100*(14), 8418-8423.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer Science & Business Media.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U. S. A., 99*(10), 6567-6572.

Venet, D., Dumont, J. E., & Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol., 7*(10), e1002240.

Vinh, N. X., Epps, J., & Bailey, J. (2009). *Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?* Paper presented at the Proceedings of the 26th Annual International Conference on Machine Learning.

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet, 365*(9460), 671-679.

Weigelt, B., Mackay, A., A'hern, R., Natrajan, R., Tan, D. S. P., Dowsett, M., et al. (2010). Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol., 11*(4), 339-349.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann.

5.6 Supporting Information

Supporting Information – Table 5.3

Table 5.3 Refined subtype labels in the METABRIC data set

The refined breast cancer subtype labels defined for each sample in the METABRIC data set are listed. *Available online: doi:10.1186/s13040-015-0078-9.*

Supporting Information – Table 5.4

Table 5.4 List of the 24 classifiers used in the ensemble learning

Family	Classifier	Software Author
Bayes	BayesNet	Remco Bouckaert
	NaiveBayes	Len Trigg, Eibe Frank
	NaiveBayesUpdateable	Len Trigg, Eibe Frank
Functions	Logistic	Xin Xu
	MultilayerPerceptron	Malcolm Ware
	SimpleLogistic	Niels Landwehr, Marc Sumner
	SMO	Eibe Frank, Shane Legg, Stuart Inglis
Lazy	IBk	Eibe Frank, Len Trigg, Stuart Inglis
	KStar	Len Trigg, Abdelaziz Mahoui
Meta	AttributeSelectedClassifier	Mark Hall
	Bagging	Eibe Frank, Len Trigg, Richard Kirkby
	ClassificationViaRegression	Eibe Frank, Len Trigg
	LogitBoost	Len Trigg, Eibe Frank
	MultiClassClassifier	Eibe Frank, Len Trigg, Richard Kirkby
	RandomCommittee	Eibe Frank
Rules	DecisionTable	Mark Hall
	JRip	Xin Xu, Eibe Frank
	PART	Eibe Frank
Trees	HoeffdingTree	Richard Kirkby,

		Mark Hall
	J48	Eibe Frank
	LMT	Niels Landwehr, Marc Sumner
	RandomForest	Richard Kirkby
	RandomTree	Eibe Frank, Richard Kirkby
	REPTree	Eibe Frank

The table shows the family and implementation authors for each method algorithm. The set of classifiers used in this work correspond to a diverse group of classifier families, as implemented in the Weka 3.7.12 software package. The corresponding references are *available online: doi:10.1186/s13040-015-0078-9.*

Supporting Information – Table 5.5

Table 5.5 Average agreement of classifiers per subtype

Subtype	**Agreement**	**Agreement (no Inc.)**
Luminal A	0.84	0.90
Luminal B	0.88	0.92
HER2-enriched	0.94	0.99
Basal-like	0.96	0.99
Normal-like	0.79	0.90
Average	**0.88**	**0.94**

Note: The numbers represent the average agreement calculated across ten runs, with relation to the final labels. The “no Inc", in the second column, excludes samples labelled “Inconsistent" from the calculation, while in the first column all samples are taken.

Supporting Information – Table 5.6

Table 5.6 Probe appearance after ten iterative processes and the respective annotation based on Dunning et al. (2010) and Illumina array data

IlluminaID	Probe Quality	Entrez Reannotated	Symbol Reannotated	Ensembl Reannotated	Probe App.
ILMN_2326273	Perfect	1117	**CHI3L2**	ENSG00000064886	**1**
ILMN_1775235	Perfect	3899	**AFF3**	ENSG00000144218	**1**
ILMN_1796059	Perfect****	91074	**ANKRD30A**		**1**
ILMN_1835913	Bad	NA	**CD108903**	NA	**1**
ILMN_1673320	Perfect	374864	**C18orf34**	ENSG00000166960	**1**
ILMN_1766914	Perfect	4239	**MFAP4**	ENSG00000166482	**1**
ILMN_1663390	Perfect	991	**CDC20**	ENSG00000117399	**1**
ILMN_1769849	Perfect	84072	**HORMAD1**	ENSG00000143452	**1**
ILMN_1773006	Perfect	2167	**FABP4**		**1**
ILMN_1660114	Bad	22915	**MMRN1**	ENSG00000138722	**1**
ILMN_2108735	Perfect	1917	**EEF1A2**	ENSG00000101210	**1**
ILMN_1726720	Perfect	51203	**NUSAP1**	ENSG00000137804	**1**
ILMN_1753196	Perfect	9232	**PTTG1**		**1**
ILMN_1815184	Perfect	259266	**ASPM**	ENSG00000066279	**1**
ILMN_1689111	Perfect	6387	**CXCL12**	ENSG00000107562	**2**
ILMN_1715991	Perfect	8436	**SDPR**	ENSG00000168497	**2**
ILMN_1716407	Perfect	8470	**SORBS2**	ENSG00000154556	**2**
ILMN_1655915	Perfect	4320	**MMP11**	ENSG00000099953	**2**
ILMN_1722489	Perfect	7031	**TFF1**	ENSG00000160182	**2**
ILMN_1684217	Perfect	9212	**AURKB**		**2**
ILMN_1888901	Perfect***	NA	**BX106902**	NA	**2**
ILMN_1695658	Perfect	10112	**KIF20A**	ENSG00000112984	**2**
ILMN_2334359	Perfect	2674	**GFRA1**	ENSG00000151892	**2**
ILMN_1726204	Perfect	11341	**SCRG1**	ENSG00000164106	**2**
ILMN_1700337	Perfect****	10024	**TROAP**		**2**
ILMN_2174805	Bad	146894	**CD300LG**	ENSG00000161649	**2**
ILMN_1651282	Perfect	1308	**COL17A1**	ENSG00000065618	**2**
ILMN_1696243	Perfect	401236	**FLJ23152**	NA	**2**
ILMN_1716925	Perfect	161835	**FSIP1**	ENSG00000150667	**2**
ILMN_2246956	Perfect	596	**BCL2**	ENSG00000171791	**3**
ILMN_1725276	Perfect****	260436	**C4orf7**		**3**
ILMN_2092077	Perfect	5304	**PIP**	ENSG00000159763	**3**
ILMN_1789463	Perfect	5348	**FXYD1**	ENSG00000221857	**3**
ILMN_1764309	Perfect	124	**ADH1A**	ENSG00000187758	**4**
ILMN_1685916	Perfect	11004	**KIF2C**	ENSG00000142945	**4**
ILMN_2310814	Perfect	4137	**MAPT**	ENSG00000186868	**5**
ILMN_2382942	Perfect	771	**CA12**	ENSG00000074410	**5**
ILMN_1695397	Perfect	388743	**CAPN8**	ENSG00000203697	**5**

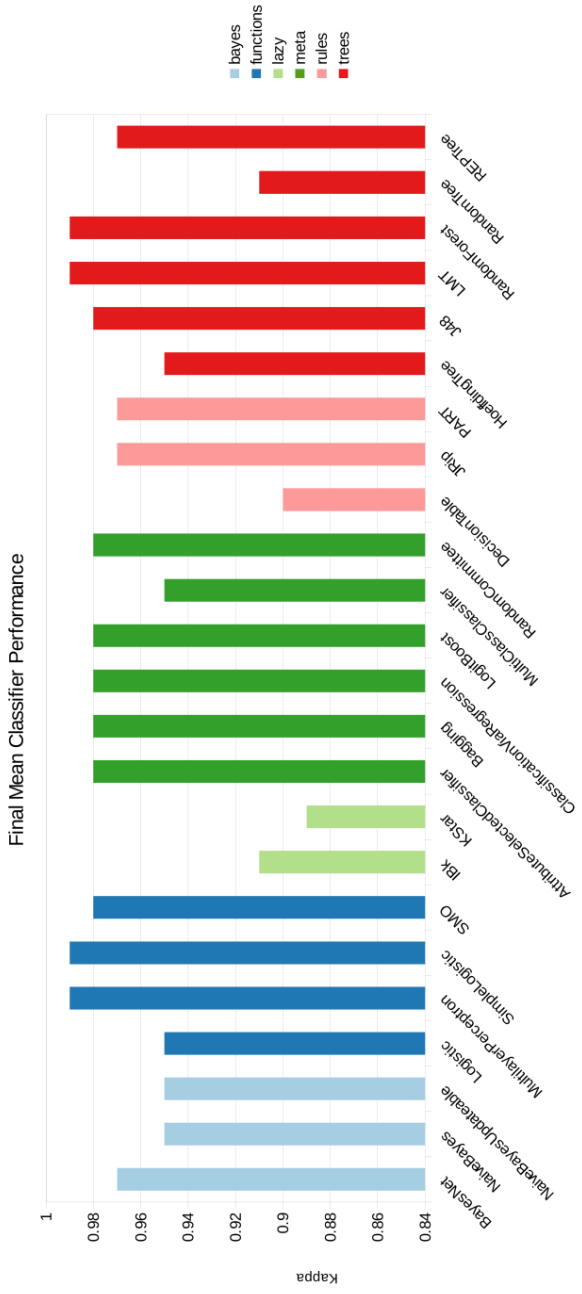
ILMN_1779416	Perfect	57758	**SCUBE2**	ENSG00000175356	**5**
ILMN_1668766	Perfect	54763	**ROPN1**	ENSG00000065371	**6**
ILMN_1770678	Perfect	84733	**CBX2**	ENSG00000173894	**6**
ILMN_1789507	Perfect	1301	**COL11A1**	ENSG00000060718	**6**
ILMN_1729216	Perfect	1410	**CRYAB**	ENSG00000109846	**6**
ILMN_1713952	Perfect	55765	**C1orf106**	ENSG00000163362	**6**
ILMN_1692938	Perfect	29968	**PSAT1**	ENSG00000135069	**8**
ILMN_1677920	Perfect	4057	**LTF**	ENSG00000012223	**8**
ILMN_1666845	Perfect	3872	**KRT17**	ENSG00000173801	**8**
ILMN_1693218	Perfect	419	**ART3**	ENSG00000156219	**8**
ILMN_1863962	Bad	NA	**BX116033**	NA	**9**
ILMN_1683450	Perfect	113130	**CDCA5**	ENSG00000146670	**9**
ILMN_1795342	Perfect	79083	**MLPH**	ENSG00000115648	**9**
ILMN_1809099	Perfect	90865	**IL33**	ENSG00000137033	**9**
ILMN_1723684	Perfect	2532	**DARC**	ENSG00000213088	**9**
ILMN_1811387	Perfect	7033	**TFF3**	ENSG00000160180	**10**
ILMN_1773459	Perfect	6664	**SOX11**	ENSG00000176887	**10**
ILMN_1729801	Perfect	6279	**S100A8**	ENSG00000143546	**10**
ILMN_1803236	Perfect	9635	**CLCA2**	ENSG00000137975	**10**
ILMN_2301083	Perfect****	11065	**UBE2C**		**10**
ILMN_2149164	Perfect	6422	**SFRP1**	ENSG00000104332	**10**
ILMN_2161330	Perfect	25803	**SPDEF**	ENSG00000124664	**10**
ILMN_1688071	Perfect	9	**NAT1**	ENSG00000171428	**10**
ILMN_1728787	Perfect	155465	**AGR3**	ENSG00000173467	**10**
ILMN_1689146	Perfect	2568	**GABRP**	ENSG00000094755	**10**
ILMN_1714730	Perfect****	11065	**UBE2C**		**10**
ILMN_1785570	Good	203328	**SUSD3**	ENSG00000157303	**10**
ILMN_1786720	Perfect****	8842	**PROM1**	ENSG00000007062	**10**
ILMN_2161820	Good	219970	**GLYATL2**	ENSG00000156689	**10**
ILMN_1678535	Perfect	2099	**ESR1**	ENSG00000091831	**10**
ILMN_1766650	Perfect	3169	**FOXA1**	ENSG00000129514	**10**
ILMN_1738401	Perfect	2296	**FOXC1**	ENSG00000054598	**10**
ILMN_1728934	Perfect	9055	**PRC1**	ENSG00000198901	**10**
ILMN_1898518	Perfect	2674	**GFRA1**	ENSG00000151892	**10**
ILMN_1810978	Bad	118430	**MUCL1**	ENSG00000172551	**10**
ILMN_1740609	Perfect	6358	**CCL15**	ENSG00000161574	**10**

Note: Probe App. – Probe Appearance. More Details on the annotation in Dunning et al. (2010).

Supporting Information – Figure 5.4

Figure 5.4 Mean Final Classifier Performance, as measured by Fleiss' κ against the final ensemble learning labels of all samples, across the 10 different refinement runs

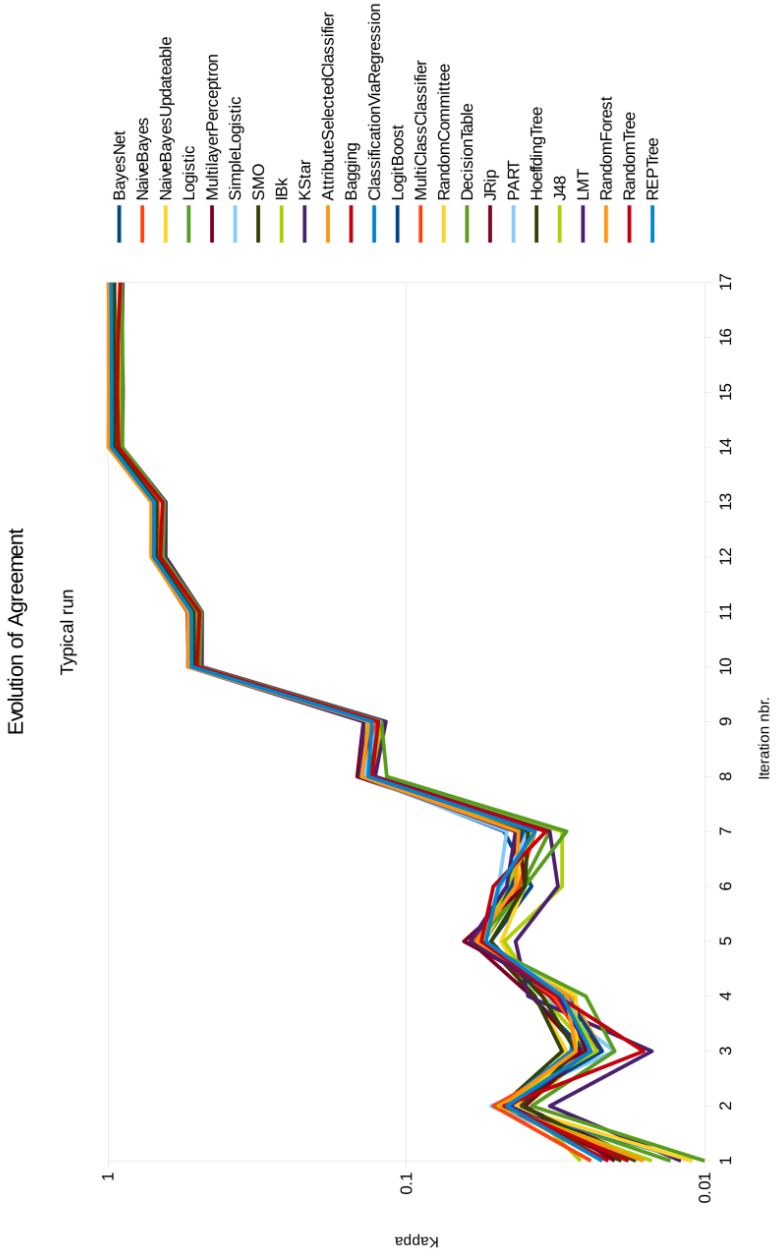
Classifiers are used with their default values, and experiments are repeated 10 times with different random seeds to provide an estimate of true value. The average mean performance of each classifier is shown in this figure. As can be observed, all classifiers attain a Kappa value greater than 0.89, which is considered an *almost perfect agreement*.



Supporting Information – Figure 5.5

Figure 5.5 Evolution of performance of classifiers along iterations in a typical refinement run. The κ values are measured against final ensemble learning labels

During the course of the refinement iterations, agreement among classifiers increases significantly, and more importantly, in a consistent manner. The evolution of the agreement, as measured by κ versus the final set of labels, for a typical iteration run.



Supporting Information – Text 5.1

Text 5.1 The MST-*k*NN clustering approach employed to the METABRIC data set.

Proximity graphs, instead of common methods of hierarchical clustering, have revealed hidden structures present in robust data sets. A combination of then is even more powerful (Jain et al., 1999). Accordingly, González-Barrios and Quiroz (2003) suggested a graph partition algorithm based on an intersection of two proximity graphs: Minimum Spanning Tree (MST) and *k* Nearest Neighbors (*k*NN). The modification on the MST-*k*NN clustering algorithm further introduced by Inostroza-Ponta (2008) is also able to deal with complex systems, establishing connection between close objects in a connective tree. This approach was then applied on the METABRIC data set to cluster related samples based on their gene expression profile. The major goal is to highlight sets of samples that cluster together according to molecular similarities and phenotypical characteristics – using the square root of Jensen Shannon divergence (Berretta & Moscato, 2010) – on breast cancer samples.

All breast tumours (997 from Discovery and 995 from Validation) and control samples (144) were independently clustered and compared against the original METABRIC PAM50 labels (**Figure 5.6**), the refined labels proposed in this chapter (**Figure 5.7**), and the novel IntClust classification defined by Curtis et al. (2012) (**Figure 5.8**). Labels were considered for relevant comparisons in terms of reliability and accuracy of breast cancer classification. The cluster analysis indicated that the five intrinsic subtypes have clear similarities based on their gene expression profiles, and are clustered together, but separated from controls. Noteworthy, the refined labels (**Figure 5.7)** show a more consistent distribution of the refined labels than the PAM50 subtypes, in (**Figure 5.6**), across samples in the METABRIC data set. Although these labels are comparable with the MST-*k*NN clustering approach, there is a visual inconsistency on the classes obtained using hierarchical clustering and MST-*k*NN approach.

Overall, there are shared features connecting luminal A and B samples; revealing the close relationship between tumours with common origin. Patients basal-like, on the other hand, are clearly distinct from other breast cancer subtypes and appear clustered together. IntClust labels (**Figure 5.8**), however, are not comparable with this approach due to the means of stratification based not only on the gene expression profile but also on the genomic aberrations across patients (more details below). Undoubtedly, other clustering approaches and complementary molecular information will contribute to improve the breast cancer subtypes classification and the disease understanding.

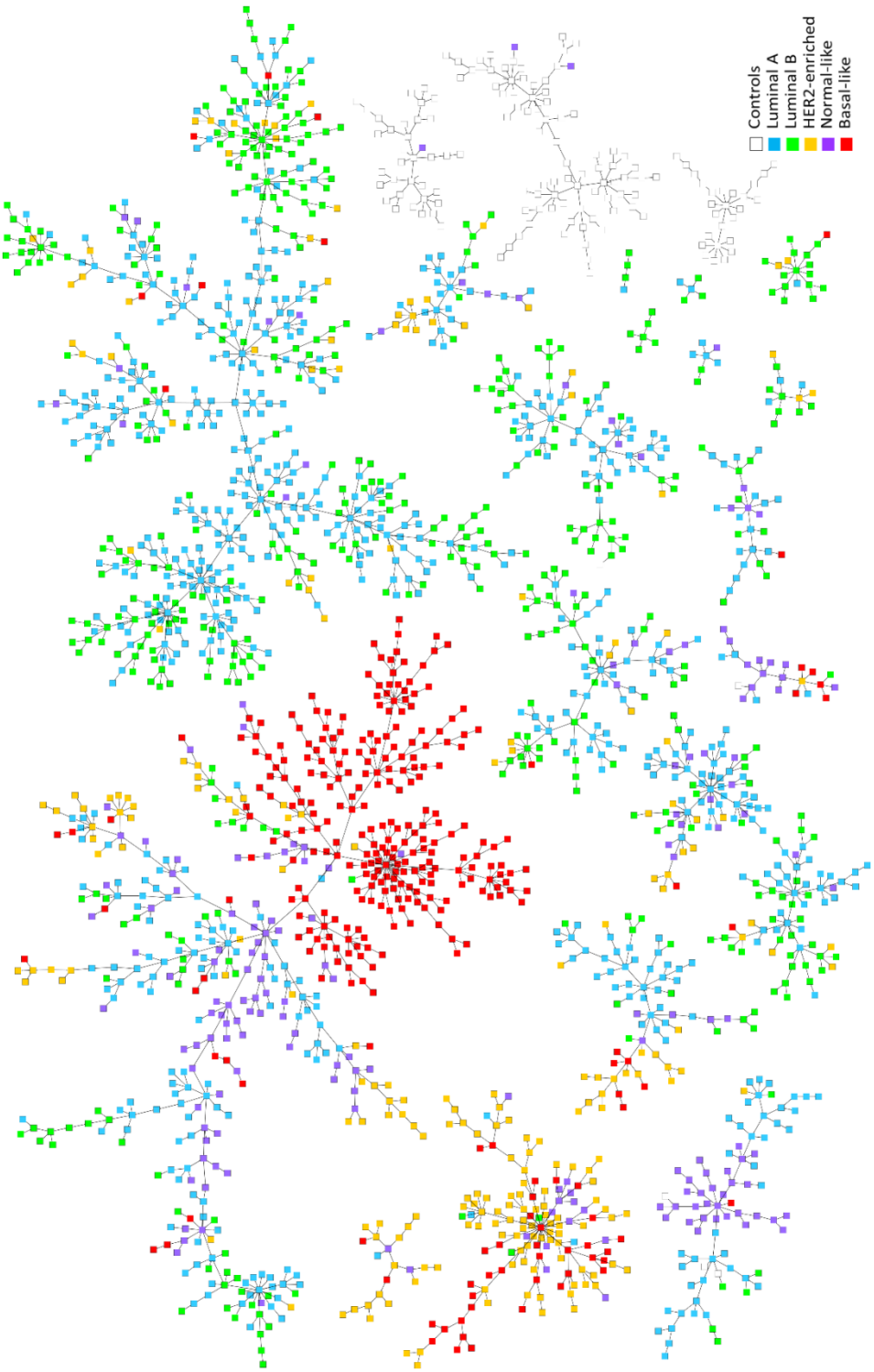
The varied molecular landscape of breast carcinomas is not entirely captured using histopathological or gene expression analysis (Dawson et al., 2013). An expanded portrait has

been obtained from studying the spectrum of copy number aberrations underlying the genomic architecture associated with intrinsic subtypes (Curtis et al., 2012). The integrated analysis of both genomic and transcriptomic METABRIC data sets revealed the impact of genomic aberrations on the transcriptomic set. Clustering analysis of the integrated data showed the existence of 10 molecular subgroups, the 10 integrative clusters (IntClust 1-10) (Dawson et al., 2013). In this analysis, we provide an overview of the 10 subtypes in comparison to the classification using the PAM50 labels (**Table 5.7**) and summarise the new insights gained with the refined labels (**Table 5.8)**. The later classification using the iterative approach has defined subtypes in better agreement with the recently proposed IntClusts than the original PAM50 labels.

Supporting Information – Figure 5.6

Figure 5.6 MST-*k*NN clustering, coloured according to the original METABRIC labels defined by the PAM50 method

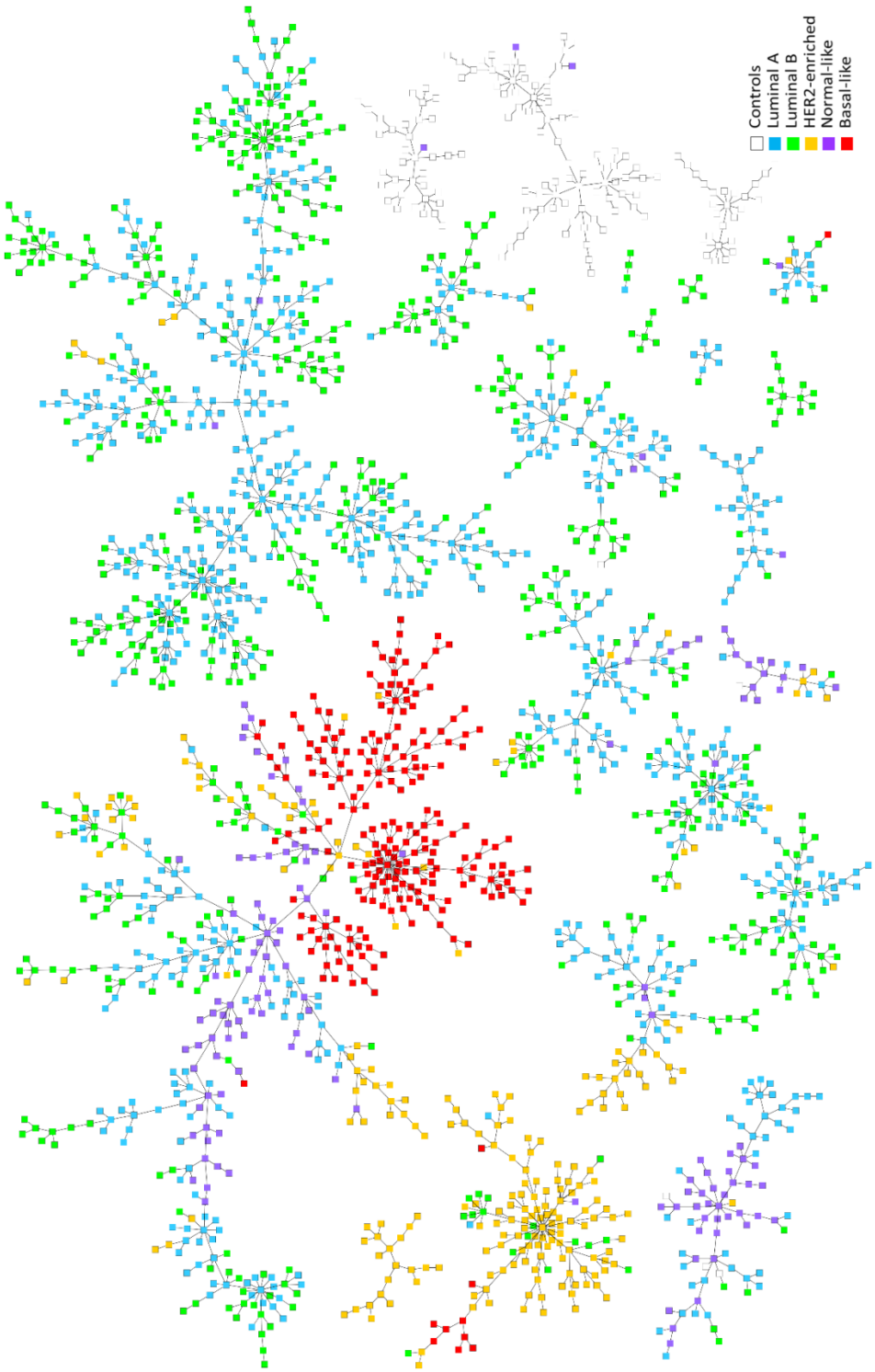
The MST-*k*NN clustering was used to group samples based on the similarities of their gene expression profile, establishing connections in a connective tree. The PAM50 labels were considered for further comparisons of reliability and accuracy of breast cancer classification.



Supporting Information – Figure 5.7

Figure 5.7 MST-*k*NN clustering, coloured according to the refined labels using an iterative process

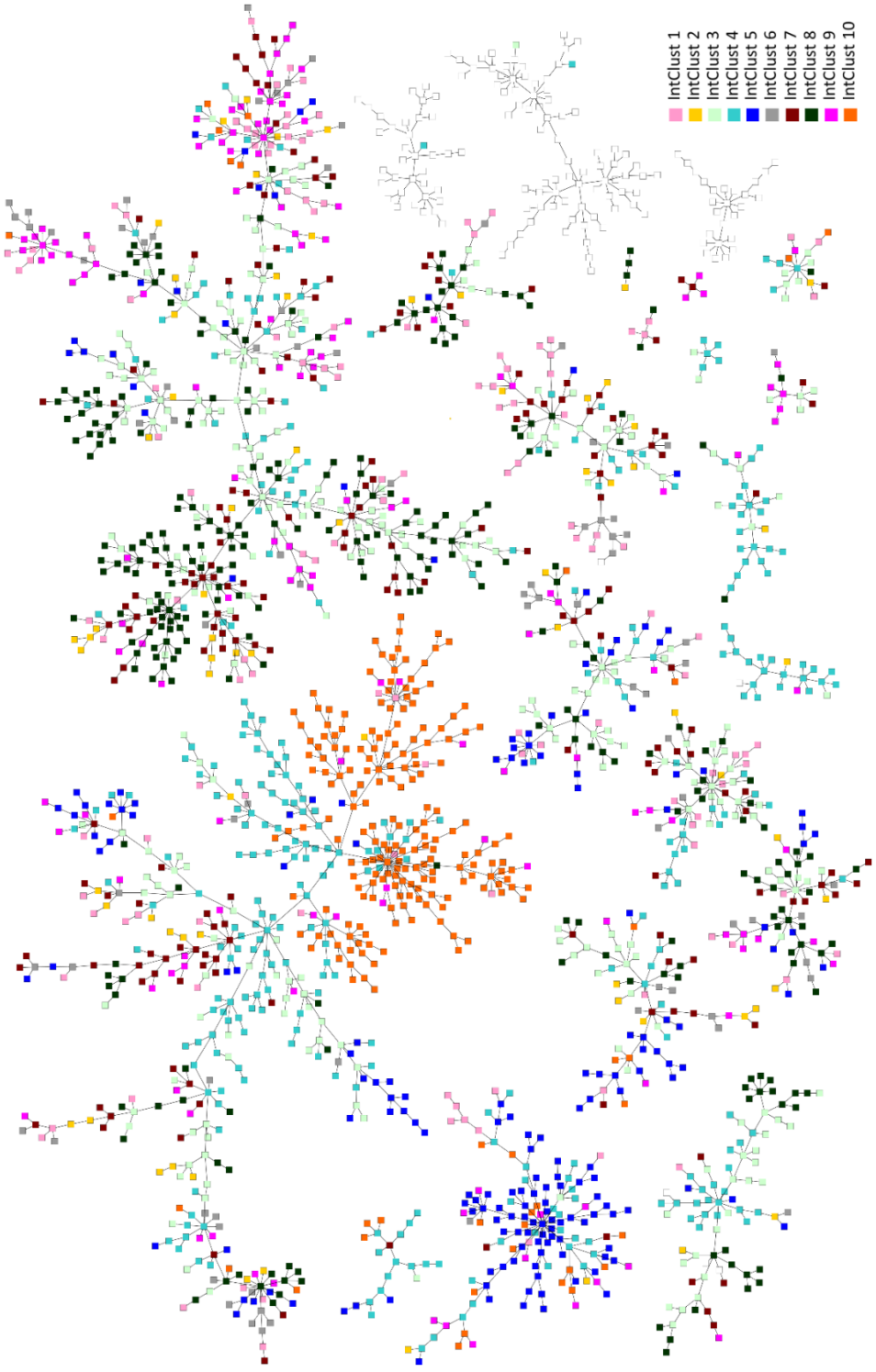
The MST-*k*NN clustering was used to group samples based on the similarities of their gene expression profile, establishing connections in a connective tree. The refined labels were considered for further comparisons of reliability and accuracy of breast cancer classification.



Supporting Information – Figure 5.8

Figure 5.8 MST-*k*NN clustering, coloured according to the IntClust classification proposed by Curtis et al. (2012)

The MST-*k*NN clustering was used to group samples based on the similarities of their gene expression profile, establishing connections in a connective tree. The IntCl;usts labels were considered for further comparisons of reliability and accuracy of breast cancer classification.



Supporting Information – Table 5.7

Table 5.7 The percentage of PAM50 labels matching integrative clusters (IntClust 1-10) in the METABRIC study

PAM50 Subtypes	Integrative Clusters										Summary
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Luminal A	7.90%	34.70%	66.30%	30.80%	9.90%	27.90%	64.20%	64.00%	16.40%	0.40%	322.70%
Luminal B	65.00%	50.00%	15.30%	8.40%	17.30%	50.00%	21.80%	29.70%	47.90%	6.20%	311.60%
Her2-enriched	15.00%	8.30%	3.10%	9.90%	56.50%	11.60%	4.70%	3.00%	17.80%	3.50%	133.50%
Normal-like	5.70%	4.20%	12.60%	31.70%	5.20%	7.00%	7.30%	3.00%	3.40%	0.40%	80.50%
Basal-like	6.40%	2.80%	1.70%	18.90%	11.00%	3.50%	1.60%	0.30%	13.70%	89.40%	149.30%
Summary	100%	100%	100.%	100.%	100%	100%	100%	100%	100%	100%	1000%

Note: The comparison of molecular subtypes in this table follow the approach published by Dawson et al (2013).

Supporting Information – Table 5.8

Table 5.8 The percentage of Refined labels matching integrative clusters (IntClust 1-10) in the METABRIC study

Refined Labels	Integrative Clusters										Summary
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Luminal A	9.60%	30.00%	71.20%	35.40%	5.30%	32.50%	64.90%	66.20%	14.20%	0.40%	329.90%
Luminal B	76.30%	60.00%	13.20%	12.00%	32.10%	66.30%	31.40%	31.40%	69.50%	5.80%	397.90%
Her2-enriched	8.90%	4.30%	3.90%	16.20%	61.50%	0.00%	1.60%	0.70%	7.10%	10.70%	114.80%
Normal-like	1.50%	4.30%	11.70%	28.20%	0.50%	1.20%	2.10%	1.40%	0.00%	0.90%	51.80%
Basal-like	3.70%	1.40%	0.00%	8.10%	0.50%	0.00%	0.00%	0.30%	9.20%	82.20%	105.60%
Summary	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	1000%

Note: The comparison of molecular subtypes in this table follow the approach published by Dawson et al (2013).

Supporting Information – Text 5.2

Text 5.2 This text reviews the ten integrative clusters, underlying differences in the disease outcome

IntClust 1

Integrative cluster 1 is predominantly classified into the luminal B (65% PAM50 vs 76.3% Refined) intrinsic subtype. It comprises a significant proportion of higher proliferation ER-positive/luminal B tumours, and are characterised by relatively high levels of genomic instability (Curtis et al., 2012). Genomic and transcriptomic features of IntClust 1 involve the amplification of the 17q23 locus and the highest prevalence of *GATA3* mutations. Amplification of 17q23 is associated with cis-driven overexpression of several adjacent genes including *RPS6KB1*, *PPM1D*, *PTRH2* and *APPBP2* (Dawson et al., 2013). These genes play an important role as key genomic drivers within the subtypes for disease stratification.

IntClust 2

Integrative cluster 2 is comprised of both luminal A (34.7% PAM50 vs 30% Refined) and luminal B (50% vs 60%) tumours. Remarkably, this subgroup has the worst prognosis among all ER-positive tumours (Dawson et al., 2013). The characteristic feature of IntClust 2 is the amplification of the 11q13/14 regions, containing several known and putative driver genes involved in breast cancer, including *CCND1*, *EMSY*, *RSF1*, C11orf67 and *PAK1*. Patients in this subgroup have amplifications involving multiple genes, suggesting a complex network where the combinations of drivers are likely to be more important than a single gene. The enrichment of genes involved in cell-cycle regulation, as exemplified by *CCND1* support the association with aggressive tumour behaviour in this cluster.

IntClust 3

Integrative cluster 3 is defined primarily by luminal A (66.3% PAM50 vs 71.2% Refined) tumours. Individuals within this subtype show small size, low-grade tumours and a low rates of regional lymph-node involvement; and the best prognosis among all the 10 IntClusts (Dawson et al., 2013). Identifying this cluster, with the majority of luminal A intrinsic subtype, is relevant as the patients usually respond to common hormonal therapy. Tumours, overall, show high frequency of *PIK3CA*, *CDH1* and *RUNX1* mutations.

IntClust 4

Integrative cluster 4 incorporates a mixture of intrinsic subtypes, including ER-posive and negative samples. Similarly to IntClust 3, IntClust4 is characterised by low levels of genomic instability. However, tumours within this subgroup show evidence of lymphocytic infiltration and deletions in T-cell receptor loci – on chromosomes 7 (TRG) and 14 (TRA) –, linked to somatic rearrangements in the infiltrating T cells. The presence of lymphocytes in these tumours are associated to a cancer immunological response that may potentially be exploited in the development of future therapeutics (Dawson et al., 2013).

IntClust 5

Integrative cluster 5 is associated to *HER2* (*ERBB2*) amplification, including mainly HER2-enriched ER-negative (56.5% PAM50 vs 61.5% Refined) and luminal B ER-positive (17.3% vs 32.1%) tumours. The amplification of the *HER2* locus, at 17q12, is frequently observed, as well as TP53 mutations and intermediate levels of genomic instability. Patients within this group show the worst survival in 10 years, high-grade tumours and involvement of regional lymph nodes (Dawson et al., 2013). Accordingly, these individuals might benefit from *HER2*-related targeted therapy.

IntClust 6

Integrative cluster 6 is a distinct cluster of ER-positive tumours, comprising both luminal A (27.9% PAM50 vs 32.5% Refined) and luminal B (50% vs 66.3%) cases. In this subtype, the defining molecular features are the low levels of *PIK3CA* mutations and the amplification of the 8p12 locus (Dawson et al., 2013). This region is commonly amplified in ER-positive breast cancers and encompasses the oncogenic driver *ZNF703*, involved in cancer cell differentiation, proliferation and invasion (Holland et al., 2011). The identification of more aggressive tumours (ER-positive/HER2-negative) within IntClust 6 may improve the disease management and the stratification of outcomes.

IntClust 7

Integrative cluster 7 is comprised predominately of ER and PR-positive luminal A (64.2% PAM50 vs 64.9% Refined) and luminal B (21.8% vs 31.4%) cases. Individuals within this cluster present low-grade, well-differentiated tumours; and the second best prognosis among all subgroups. Copy number aberrations in IntClust 7, differentiating from IntClust 3, are

specific 16p gain and 16q loss, as well as a higher frequency of 8q amplification (Dawson et al., 2013). Notably, tumours also show the highest frequency of *MAP3K1* and *CTCF* mutations across all clusters.

IntClust 8

Integrative cluster 8 is predominantly composed by ER-positive tumours of luminal A (64% PAM50 vs 29.7% Refined) and B (66.2% vs 31.4%) subtypes. Likewise IntClust 7, patients within IntClust 8 also present low-grade, well-differentiated tumours, and good prognosis. This subgroup, however, is characterised by 1q gain/16q loss that corresponds to a common unbalanced translocation event in Invasive Ductal Carcinomas (Russnes et al., 2010). High levels of *PIK3CA*, *GATA3* and *MAP2K4* mutations are also observed. Tumours previously grouped with the luminal A subtype label, in fact, are separated into three distinct IntClusts (IntClust 3, 7 and 8), containing independent genomic aberrations (Dawson et al., 2013).

IntClust 9

Integrative cluster 9 is a mixture of intrinsic subtypes, with a greater number of ER-positive luminal B (47.9% PAM50 vs 69.5% Refined) intrinsic subtype. Likewise IntClust 6, tumours in IntClust 9 have an intermediate prognosis and high levels of genomic instability. The main molecular characteristics within IntClust 9 include 8q alterations and 20q amplification. On chromosome 8p, *PP2R2A* deletions affect several signal transduction pathways, common in luminal B intrinsic subtype (Dawson et al., 2013). Additionally, mutations and methylation silencing of *PPP2R2A* have been reported in other solid malignancies (McConechy et al., 2011), suggesting the possible role of this gene as an important tumour suppressor.

IntClust 10

Integrative cluster 10 embraces mostly triple-negative tumours of basal-like (89.4% PAM50 vs 82.2% Refined) intrinsic subtype. Despite displaying intermediate levels of genomic instability, these tumours have the highest rates of *TP53* mutations. IntClust 10 is characterised by aberrations involving 5q loss and gains at 8q, 10p and 12p. In particular, 5q deletions are associated with the basal-like subgroup by modulating the landscape of genomic instability within these tumours. In this region, important genes regulate the cell-cycle, DNA repair and apoptosis, such as *AURKB*, *BCL2*, *BUB1*, *CDCA3*, *CDCA4*, *CDC20*, *CDC45*, *CHEK1*, *FOXM1*, *HDAC2*, *IGF1R*, *KIF2C*, *KIFC1*, *RAD51* and *UBE2C* (Dawson et al., 2013). The transcriptional changes observed within this subgroup are crucial to delineate the basal-like tumours, usually

less responsive to chemotherapy and sensitive to neoadjuvant chemotherapy (Banerjee et al., 2006).

Supporting References

Banerjee, S., Reis-Filho, J. S., Ashley, S., Steele, D., Ashworth, A., Lakhani, S. R., et al. (2006). Basal-like breast carcinomas: clinical outcome and response to chemotherapy. *J. Clin. Pathol., 59*(7), 729-735.

Berretta, R., & Moscato, P. (2010). Cancer Biomarker Discovery: The Entropic Hallmark. *PLoS One, 5*(8), e12262.

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature, 486*(7403), 346-352.

Dawson, S. J., Rueda, O. M., Aparicio, S., & Caldas, C. (2013). A new genome-driven integrated classification of breast cancer and its implications. *EMBO J., 32*(5), 617-628.

González-Barrios, J. M., & Quiroz, A. J. (2003). A clustering procedure based on the comparison between the k nearest neighbors graph and the minimal spanning tree. *Stat Probabil Lett, 62*(1), 23-34.

Holland, D. G., Burleigh, A., Git, A., Goldgraben, M. A., Perez-Mancera, P. A., Chin, S. F., et al. (2011). ZNF703 is a common Luminal B breast cancer oncogene that differentially regulates luminal and basal progenitors in human mammary epithelium. *EMBO Mol. Med., 3*(3), 167-180.

Inostroza-Ponta, M. (2008). *Thesis: An integrated and scalable approach based on combinatorial optimization techniques for the analysis of microarray data.* The University of Newcastle.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. *ACM Computing Surveys, 31*(3).

McConechy, M. K., Anglesio, M. S., Kalloger, S. E., Yang, W., Senz, J., Chow, C., et al. (2011). Subtype-specific mutation of PPP2R1A in endometrial and ovarian carcinomas. *The Journal of pathology, 223*(5), 567-573.

Russnes, H. G., Vollan, H. K. M., Lingjærde, O. C., Krasnitz, A., Lundin, P., Naume, B., et al. (2010). Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Sci. Transl. Med., 2*(38), 38ra47-38ra47.

CHAPTER 6

6. META-FEATURES FOR PREDICTING BREAST CANCER INTRINSIC SUBTYPES

There are a number of different ways of dealing with complex high dimensional data sets in breast cancer. As an alternative to the iterative approach for predicting subtypes (*Chapter 5*), *Chapter 6* introduces a systematic methodology for distinguishing samples across the intrinsic groups by looking at pairwise probes, termed *meta-features*, at a minimum template. This method is based on mathematical modelling, feature selection methods and data mining. The computational framework is motivated by the widespread interest in conducting and reporting robust methods for building accurate predictor models. The content is structured as a methodology article, submitted to *Genomics, Proteomics & Bioinformatics*[8], and here divided into sections **6.1 Introduction**,**6.2 Methods**, **6.3 Results and Discussion**, **6.4 References** and **6.5 Supporting Information**. This novel strategy underscores applied research in breast cancer for leveraging the utility of pairwise probes for covering both the intrinsic signature and the subtype prediction. It also delineates molecular imbalances across subtypes and supports the breast cancer group-based definition in the clinical setting.

[8] Milioli, H.H.; Riveros, C.; Vimieiro, R.; Tishchenko, I.; Berretta, R.; Moscato, P. Meta-features modelling gene expression imbalances: an innovative strategy for breast cancer subtype prediction. Manuscript submitted to *Genomics, Proteomics & Bioinformatics*.

6.1 Introduction

Microarray technologies and gene expression profiling have been widely explored in medical research. In the direction of developing useful tools to delineate the breast cancer behaviour, researchers have published a number of predictive models based on multi-gene signatures. The computational methods have shown gene expression values strongly correlated to clinical prognosis (Fan et al., 2011; Loi et al., 2008), disease progression (Seoane et al., 2014; Venet et al., 2011; Wang et al., 2005) and patient survival (Chang et al., 2005; Naderi et al., 2006). The main purpose is to either inform or anticipate the patient's outcome, and guide treatment decision-making (van't Veer et al., 2002; van De Vijver et al., 2002). Mammaprint® (Agendia, Huntington Beach, CA) and Oncotype DX® (Genome Health Inc, Redwood City, CA), two commercial assays, are standard examples of genome supervised predictors (Buyse et al., 2006; Glas et al., 2006; S. Paik et al., 2004). Based on the Amsterdam 70-gene signature, Mammaprint was designed to estimate the likelihood of distant recurrence in the five years following diagnosis. This investigation is also decisive for guiding systemic adjuvant therapy (Drukker et al., 2014; Kok et al., 2012). Similarly, Oncotype DX uses a panel of 21 genes to determine the risk of metastasis in women with early-stage hormone oestrogen receptor (ER) positive breast cancer. The test is assessed through the Recurrence Score and outlines the benefits of chemotherapy (Albain et al., 2010; Chen et al., 2013; Soonmyung Paik et al., 2006).

Gene expression cohorts were imperative to classify breast cancers into intrinsic subtypes: luminal A, luminal B, HER2-enriched, normal-like and basal-like (J. I. Herschkowitz et al., 2007; Hu et al., 2006; Perou et al., 2000; Prat et al., 2010; Sørlie et al., 2001; Sørlie et al., 2003). The new concept underlying subtype prediction is based on risk models that incorporate molecular signatures shared among tumours with analogous behaviour. In 2009, Parker et al. (2009) proposed a Single Sample Predictor (SSP) model to classify tumour subtypes according to the correlation with Nearest Shrunken Centroids (NSC) (Tibshirani et al., 2002). The so-called PAM50 method uses a 50 gene set as centroids. These genes are mainly involved in cell proliferation and are highly correlated with breast cancer subtypes. In the same direction, another research group attempted to simplify the subtypes prediction by using a Subtype Classification Model (SCM) based on three key genes: oestrogen receptor 1 (*ESR1*), erb-b2 receptor tyrosine kinase 2 (*ERBB2*), and aurora kinase A (*AURKA*) (Haibe-Kains et al., 2012).

Overall, the main goal of the disease subtyping is to define sets of patients more likely to respond to selective drugs in a group-based tailored therapy. The substantial impact of

predictor models in breast cancer research have brought new insights to translational science and applied medicine, and are of unquestionable value to clinical practice. Even though different gene sets are used for breast cancer prognostication, there is a significant agreement in the outcome predictions for individual patients (Fan et al., 2006). On the other hand, subtyping methods showed only a *moderate agreement* between sample labelling across distinct studies (Weigelt; Mackay; et al., 2010), besides intrinsic errors (Ebbert et al., 2011) and independent predictive value (Prat et al., 2012). The weaknesses of these methods lie in the analysis of multiple data sources and the distinct approaches. For instance, important issues may arise with independent sample collection, gene expression analysis and microarray technology. Hence, a range of gene lists of different size and shape are selected (Popovici et al., 2010). Stringent standardisation of data sets and methodologies are therefore required to improve breast cancer classification and subtype prediction. Novel strategies are also required for a robust analysis of complex data sets before translating medical research into clinical application.

In this study, we designed a novel systematic approach to define a robust pairwise set able to predict the breast cancer intrinsic subtypes. We hypothesise that combined gene expression data brings more reliable information than single gene assessments. By expanding the original METABRIC transcriptomic data set, we compute the relative pairwise differences of gene expression levels for all 48803 probes in each sample. The pairwise differences are henceforth named meta-features. The main strategy relies on mathematical modelling, feature selection and data mining approaches, making use of the following well-established methods: CM1 score, (α,β)-*k*-Feature set, and ensemble learning. Statistical measures (Cramer's V, Average sensitivity, Fleiss' kappa, and Adjusted Rand Index) are used to define the association between the original subtype labels and predicted ones. The applied research stresses pairwise intrinsic signature to explain both the genomic imbalance and the subtype prediction in breast cancer.

6.2 Methods

6.2.1 Ethics Statement and Data Description

Samples in the METABRIC data set (Curtis et al., 2012) were assigned into the five intrinsic subtypes (luminal A, luminal B, HER2-enriched, normal-like and basal-like) according

to the PAM50 method (Parker et al., 2009). Due to some inconsistencies in the original labelling, the subtypes used in this report follow the recent assignment provided by Milioli et al. (2015). Samples were partitioned back in the two original METABRIC subsets: Discovery (855 samples) and Validation (831 samples), respectively referred to as training and test sets in our analysis. Inconsistent samples (306) were discarded.

6.2.2 Study Design and Computing Resources

In this study, we propose a novel approach based on mathematical modelling, feature selection and data mining to improve the breast cancer subtype prediction. Based on pairwise expression values, we first create a robust matrix of meta-features – defined as the difference between pairs of probes – from the METABRIC data set. The mathematical operation, thus, embeds additional information to this data set. Using the computed matrix, we apply the feature selection methods, CM1 score and (α,β)-*k*-Feature set, to extract the most representative biomarkers for each subtype. Additionally, the quality of the meta-features selection is assessed using a set of classifiers from Weka, followed by the statistical analysis. All steps are shown in **Figure 6.1** and detailed in the remainder of this section.

Generating meta-features.

First, we computed the absolute difference between expression values of all possible pairs of probes from the original set of 48803, for each sample. This results in a robust matrix of new meta-features containing useful new information. The matrix is symmetric since the absolute difference is a symmetric operator (|F1 - F2| equal to |F2 - F1|). Each meta-feature is a combination of individual features, so that the simple mathematical model attributes a particular weight or function to the pair (Rocha de Paula et al., 2011). Using the pairwise strategy, we highlight the relationship between probes and explicit changes in tumour behaviour across groups of patients. The objective is to scrutinise the molecular patterns correlated to subtypes.

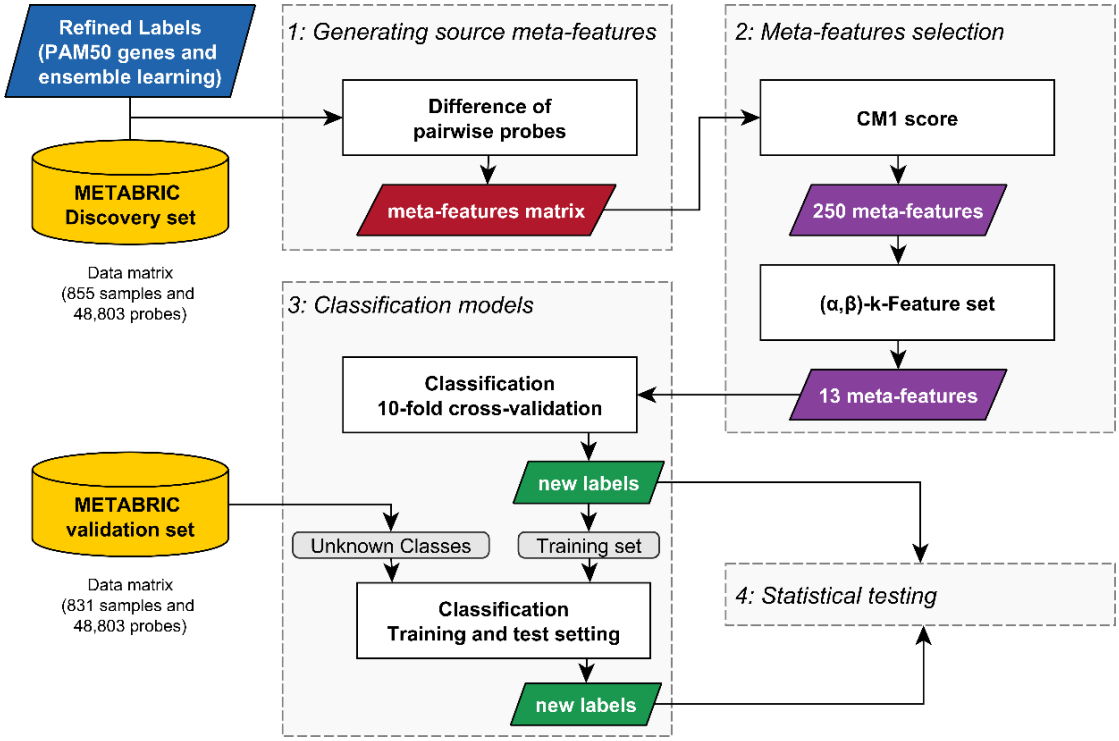


Figure 6.1 Summary systematic approach

The process is initialised with the new labels recently attributed to the METABRIC study, based on PAM50 genes and ensemble learning (Milioli et al., 2015). A robust matrix is generated after computing the absolute difference of pairwise probes for all samples. The CM1 score was used to select the top 50 highly discriminative meta-features for each subtype, resulting in a filter of 250. From this set of meta-features, 13 represent the size of the smallest subset able to discriminate the sample subtypes according to the (α,β)-k-Feature set. These meta-features are then used to train the classifiers in the ensemble learning. The labels in the validation set are predicted using the models built in the discovery set (10-fold cross-validation), in a training-test setting. The predictive power of the classifiers is determined by a range of statistical measures.

Meta-feature selection using CM1 score and (α,β)-k-Feature set.

The second step involves two well-established methods, the CM1 score (Marsden et al., 2013; Milioli et al., 2015) and (α,β)-*k*-Feature set (Cotta & Moscato, 2003; Cotta et al., 2004; Gómez-Ravetti et al., 2009), to define the most representative probes. The CM1 score (described in *Chapter 4*, **Equation 4.1**) is a supervised approach used to rank the features (in this case, the meta-features) according to their discriminative power for each intrinsic subtype or class. For computing the CM1 score we use the new labels provided by Milioli et al. (2015). Samples from the five intrinsic subtypes were alternately taken as the target set. In each case, the CM1 score was computed for all meta-features. The meta-features were then ranked from

the maximum to the minimum CM1 values. Ultimately, we selected 50 meta-features best ranked for each intrinsic subtype, comprising a total of 250. This number was arbitrarily defined for the first filtering. Although there is an overlap between meta-features across subtypes, we have maintained the denomination of 250 meta-features.

The (α,β)-*k*-Feature set problem is a feature selection approach based in combinatorial optimisation (**Supporting Information – Text 1**). The rationale behind this supervised method is to find the minimum subset of features that better discriminates two classes of samples by solving a combinatorial optimisation problem. For a given instance, the decision problem aims to find a *k* feature set in which every pair of samples from different classes can be "explained" by at least α features; and any pair of samples from the same class (identical values) by at least β features The process requires finding a minimum cardinality set of features that satisfies the requirement of being a *k*-Feature Set. The optimisation version of this problem has been formulated as an integer programming model (Berretta et al., 2008; Berretta et al., 2007). In this study, we endeavour to select the essential signature, from the list of 250 meta-features that is able to explain the breast cancer intrinsic subtypes.

Building classification models to assess the quality of meta-features.

In the third step, we used several machine learning algorithms to build a classification model based on the former set of representative meta-features. This assay relies on an ensemble of 22 classifiers from the Weka software suite (Witten et al., 2016). Each of the classifiers is trained with a subset of the data containing the list of meta-features for all samples in the discovery set in a 10-fold cross-validation setting, which involves randomly partitioning the original data set into 10 equally sized subsets. Each subset in turn is left out and the learning method is trained on the union of all the remaining subsets. The results of all 10 judgements, one for each member of the data set, are averaged, and the average represents the final error estimate. The labels in the validation set are then predicted using the models built in the discovery set, in a training-test setting. The purpose of performing ensemble learning is to assess the quality of the meta-features for the prediction of breast cancer intrinsic subtypes. In this context, the power and robustness of a set of classifiers has proven to be far superior than single predictors (Gómez-Ravetti & Moscato, 2008; Milioli et al., 2015).

6.2.3 Statistical Analysis

Statistical measures are used to assess the consistency of the subtype predictions. The quality of meta-features is estimated according to Cramer's V statistic. The consensus of sample labelling across different methods is defined by the popular interrater reliability metric Fleiss' kappa. This statistic is calculated to gauge the agreement, not only among classifiers trained with the set of meta-features, but also between the initial METABRIC subtype labels (Milioli et al., 2015) and the labels assigned by the majority of classifiers. Finally, we performed the Adjusted Rand Index to quantify the agreement between samples that are either in the same class or in different classes after a given partition. The measures are detailed as follows:

Cramer's V: measures the strength of association among variables of the rows and columns given a contingency table (Liebetrau, 1983). To calculate this metric, we assumed a r x c contingency table describing the association between the initial METABRIC labels and those predicted by the majority of classifiers using ensemble learning. More details in *Chapter 4*, **Equation 4.2**.

Fleiss' kappa: defines the reliability of agreement among the labels defined by different classifiers trained using the meta-features; and between the initial METABRIC labels and new labels predicted based on ensemble learning (Fleiss, 1971; Fleiss et al., 2004). Assuming a r x c contingency table informing how many times each of the classes were assigned to each of the samples in the k different sample labelling. More details in *Chapter 4*, **Equation 4.4**.

Adjusted Rand Index: quantifies the similarity between two sample labelling. It is a version of the Rand index corrected for chance when the partitions are picked at random (Hubert & Arabie, 1985). More details in *Chapter 4*, **Equation 4.5**.

6.3 Results and Discussion

6.3.1 Thirteen Meta-features Define Breast Cancer Intrinsic Subtypes

In this study, we investigated one of the most comprehensive data sets in breast cancer, expanding the ~2000 samples and 48803 probes to build functional pairwise constructs. The proposed approach is robust with regard to the mathematical computation of all possible probe combinations. It provides further information on modelling meta-features to predict intrinsic subtypes. Our method differs from previous studies in many aspects. We search for co-expression patterns whereas other groups have recognised individual gene levels to create molecular signatures for prediction. Moreover, we rely on a large data collection processed by METABRIC which is prominent in terms of microarray technology and genomic mapping. Here, we provide a transparent example of the manner in which a genomic-based predictor can be developed with significant detail and documentation to allow the process to be replicated by other researchers and in a range of fields. The results as follow introduce the utility of the information carried with the pairwise probes to improve the prediction of breast cancer intrinsic subtypes.

Defining pairwise patterns across intrinsic subtypes, however, does not mean that the probes are directly related in common network. Each individual transcript can be viewed as separate molecular underpinning, potentially independent in the mechanistic biology or as major system regulators. For instance, the first filter based on the CM1 score resulted in 250 meta-features (50 for each subtype), ranked according to the highest values for each group. Some meta-features, however, are able to discriminate more than one subtype and appeared repeatedly connected with a range of other transcripts. Then, a total of 153 unique probes connected with the respective co-expressed pair is displayed in **Figure 6.2**. Remarkable genes, such as *ESR1*, *FOXA1, KRT17, MLPH, SFRP1* and *UBE2C* were recognised as central among the pairwise probes across the five intrinsic subtypes in the METABRIC discovery set. These genes are well-established in the literature as to their involvement with breast cancer progression and intrinsic subtypes (Bastien et al., 2012; McCafferty et al., 2009; Parker et al., 2009; Reis-Filho & Pusztai, 2011; Weigelt; Baehner; et al., 2010). Novel transcripts, however, appear naturally linked to these genes, pointing to genomic imbalances across the subtypes. We believe that the co-expressed related features indicate that a larger and more complex construct could be uncovered from the data in order to explain the molecular arrangement of this heterogeneous disease.

The final set of 13 meta-features reflect the complementary selection based on (α,β)-k-Feature set (**Table 6.1**). This filter simplifies the complexity of the intrinsic subtypes by selecting the essential pairs able to differentiate the subtypes. Accordingly, the 13 meta-features are able to explain why samples are assigned in the same class or in different classes. The heat maps (**Figure 6.3**) generated for the discovery and validation sets support the hypothesis that pairwise probes distinguish the five intrinsic groups. The pairwise expression levels define the plot in which rows represent the meta-features and columns represent the samples. **Figure 6.4** depicts the coordinated patterns characterising subtypes. This plot shows reduced variability outside the upper and lower quartiles in comparison to the analysis of individual features (**Figure 6.5**). The results, therefore, emphasise the usefulness of the information carried within the functional constructs, in a minimum set. For instance, we suggest only 13 meta-features (23 features) to label the samples when the commercial PAM50 assay utilises a panel of 50 genes. On the other hand, our method introduces novel potential markers not previously considered to predict the breast cancer subtypes, such as: *CSN3, GAS1, KIF20A, LAMC3, MAD2L1, NCAPG, NUSAP1, ROPN1B, SULT1C2P1, SPDEF, TBC1D9* and *ZMYND10*.

Table 6.1 List of meta-features selected with CM1 score and (α,β)-k Feature set

Probe IDs	RefSeqGene
ILMN 1792400 - ILMN 1766650	CSN3 - FOXA1
ILMN 1668766 - ILMN 1703891	ROPN1 - TBC1D9
ILMN 1689146 - ILMN 2161330	GABRP - SPDEF
ILMN 1678720 - ILMN 1678535	SULT1C2P1 - ESR1
ILMN 2125763 - ILMN 1683450	ZMYND10 - CDCA5
ILMN 1726720 - ILMN 1786720	NUSAP1 - PROM1
ILMN 2301083 - ILMN 2163723	UBE2C - KRT7
ILMN 1777564 - ILMN 2149164	MAD2L1 - SFRP1
ILMN 1688642 - ILMN 1747016	LAMC3 - CEP55
ILMN 1651282 - ILMN 1695658	COL17A1 - KIF20A
ILMN 1772910 - ILMN 1789507	GAS1 - COL11A1
ILMN 1651282 - ILMN 2301083	COL17A1 - UBE2C
ILMN 1688642 - ILMN 1751444	LAMC3 - NCAPG

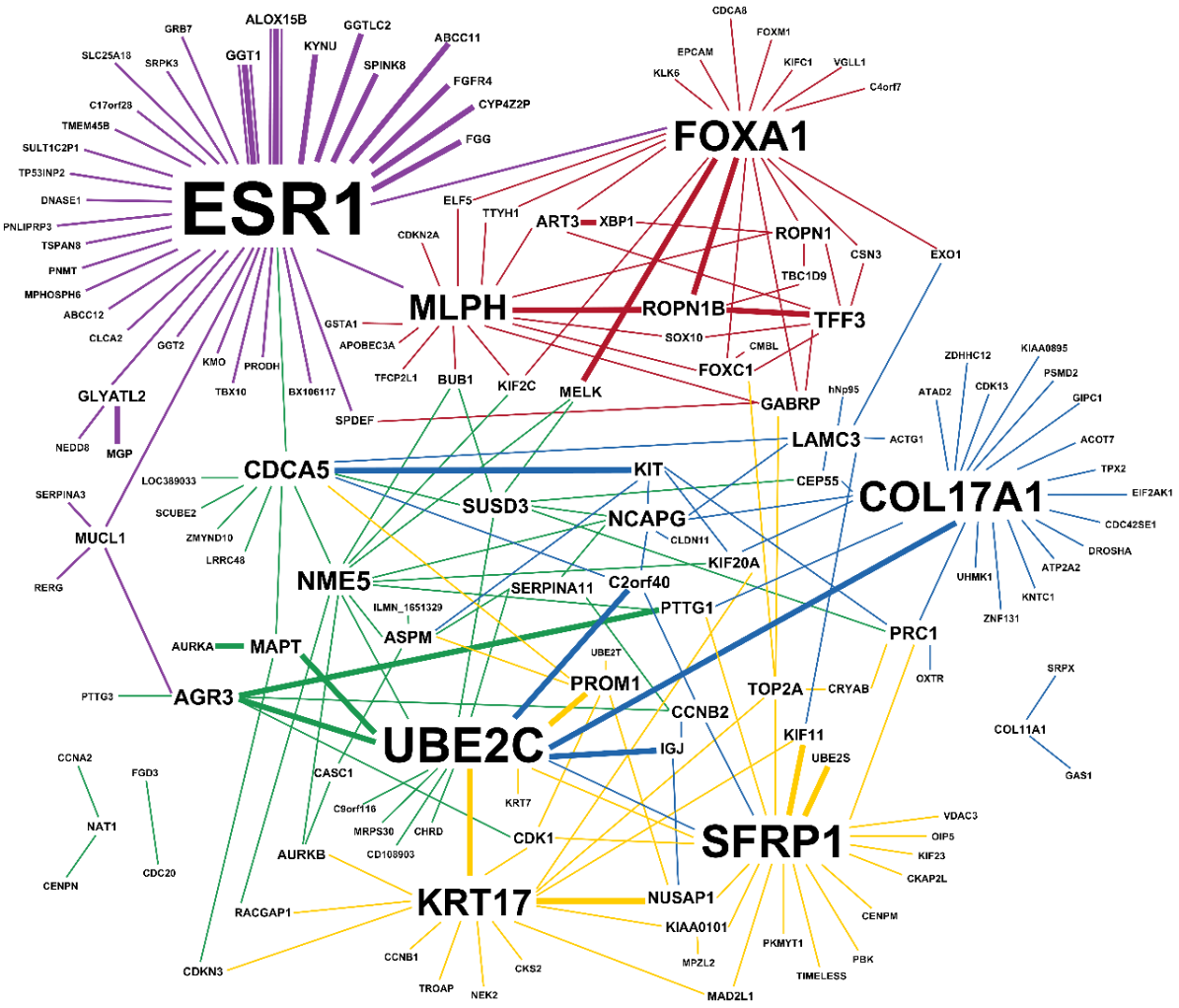


Figure 6.2 Meta-features selected with the CM1 score in the METABRIC discovery set

The image shows discriminative meta-features ranked with the CM1 score using 855 samples from the discovery set. Pairwise expression values were computed for each of the five intrinsic subtypes. Nodes represent the respective probe annotated and edges are the connections between features. The meta-feature connections are coloured according to the subtypes: luminal A (yellow), luminal B (green), HER2-enriched (purple), normal-like (blue), and basal-like (red).

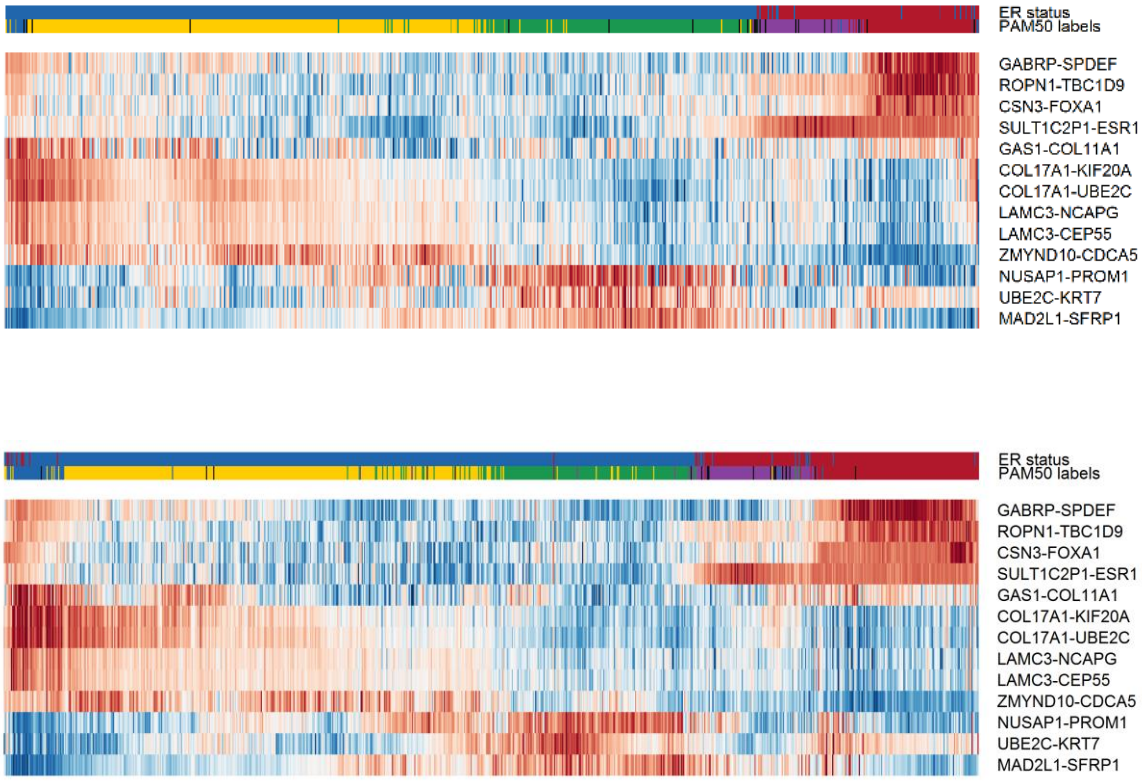


Figure 6.3 Gene expression patterns of the 13 meta-features selected using the CM1 score and (α,β)-k-Feature set

The heat map diagram exhibits 13 meta-features, in rows, and (A) 855 samples, in columns, in the discovery set and (B) 831 samples in the validation set. The images are based on meta-features selected in the discovery set, ordered according to the gene expression similarity using memetic algorithm. Labels highlighted on top show the sample distribution according to the ER positive and negative status. It also illustrates the initial subtypes (Milioli et al., 2015) as follow: luminal A (yellow), luminal B (green), HER2-enriched (purple), normal-like (blue), and basal-like (red).

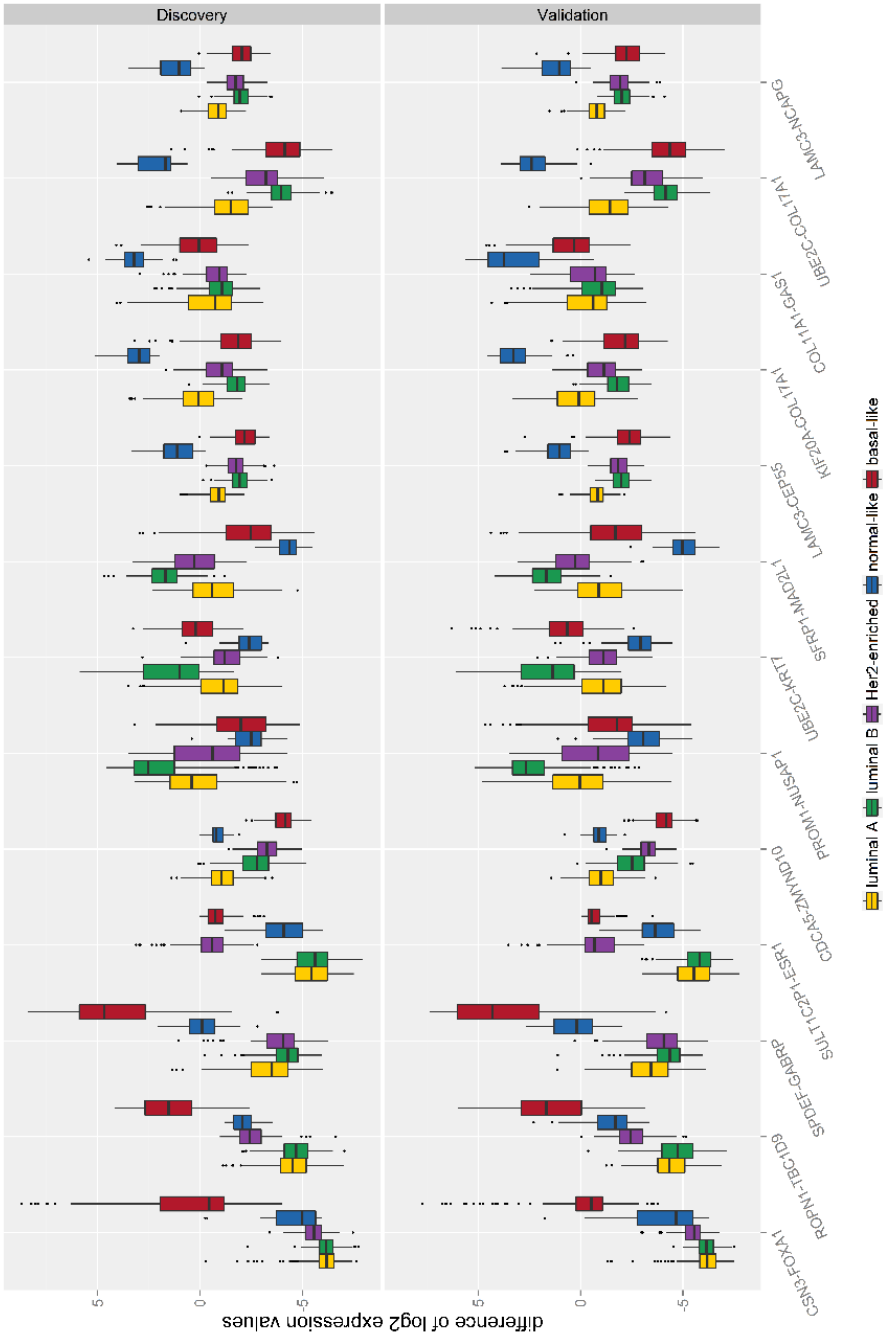


Figure 6.4 Pairwise expression patterns across intrinsic subtypes in the METABRIC discovery and validation sets

The box plot shows the pairwise expression patterns of 13 meta-features across the five breast cancer subtypes (luminal A, luminal B, HER2-enriched, normal-like and basal-like) in the discovery (855 samples) and validation (831 samples) sets.

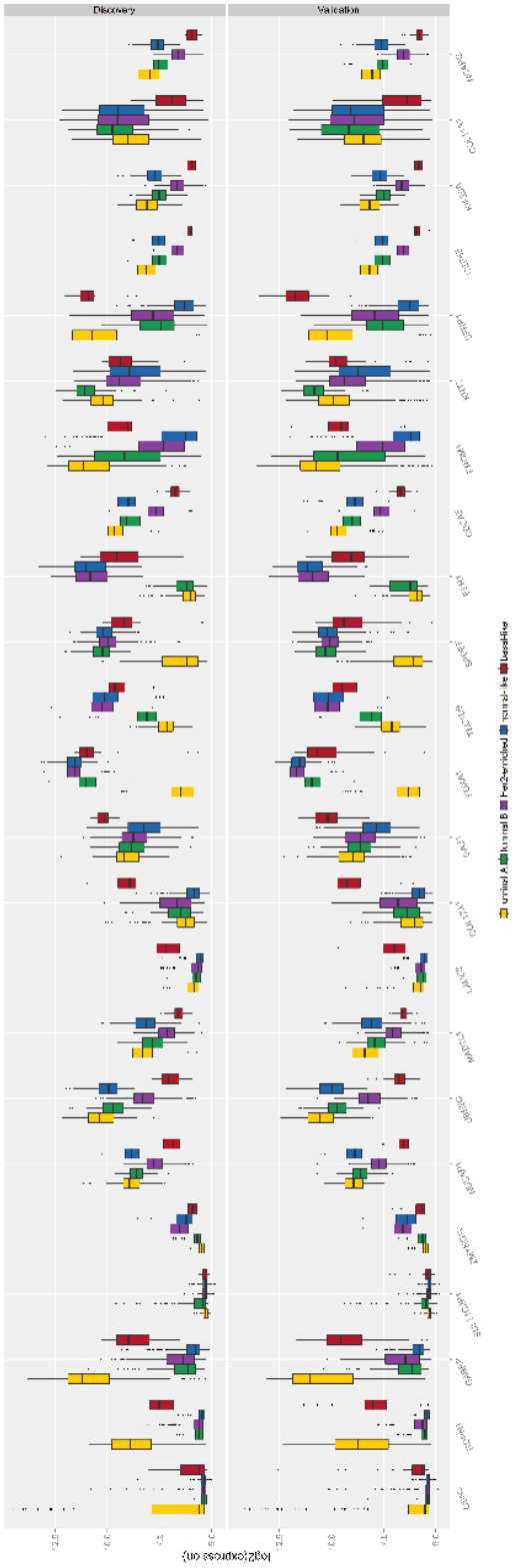


Figure 6.5 Individual expression patterns across intrinsic subtypes in the METABRIC discovery and validation sets

The box plot shows the expression patterns of 23 individual features across the five breast cancer subtypes (luminal A, luminal B, HER2-enriched, normal-like and basal-like) in the discovery (855 samples) and validation (831 samples) sets.

6.3.2 An Ensemble Learning Approach Validates the Quality of Meta-features for Predicting Subtypes

An ensemble learning based on 22 classification models was performed to further evaluate the quality of the 13 meta-features for predicting breast cancer intrinsic subtypes. Several statistical measures were computed to assess the consistency of our analysis: Cramer's V, Fleiss' kappa and Adjusted Rand Index. The results of each of these underlying metrics are described below. This step basically defines the reliability of the subtype labels assigned by the majority of the classifiers when compared to the initial METABRIC labels proposed by Milioli et al. (2015).

Cramer's V statistic reveals great overall performance of independent classification models.

Cramer's V statistic was used to compute the performance of the ensemble learning using the meta-features as an input in the training (discovery) and test (validation) sets. It measures the strength of association between two labelling given a contingency table (**Table 6.2**). In this case, rows represent the initial METABRIC labels and columns represent the subtypes assigned by the majority of the classifiers in the ensemble. The Cramer's V statistic showed in **Table 6.3** determines an average association of 0.91 ± 0.03 and 0.91 ± 0.03 in the discovery and validation sets respectively. Our results suggest a strong association between initial and predicted labels. The high values obtained in this study indicate that the proposed 'functional constructs' or meta-features together with an ensemble learning have great potential to predict the breast cancer subtypes. Correct assignments lead to a more precise prognostic as per low or high cancer-related genes expression and support clinical decision-making.

Table 6.2 Contingency tables for predicted labels using ensemble learning trained with 13 meta-features Discovery set Validation set

	Discovery set					***Validation set***				
	LA	**LB**	**H2**	**NL**	**BL**	**LA**	**LB**	**H2**	**NL**	**BL**
LA	389	9	1	1	0	344	17	0	1	1
LB	10	232	0	0	0	9	176	3	0	0
H2	0	0	94	1	1	0	0	95	0	0
NL	0	0	0	16	0	0	0	0	41	0
BL	0	0	1	1	99	0	0	2	2	140

Note: Rows contain the new sample labels in METABRIC (Milioli et al., 2015), while columns contain labels assigned by the majority of classifiers using the 13 meta-features. In this table, LA corresponds to luminal A, LB luminal B, H2 HER2-enriched, NL normal-like, and BL basal-like breast subtype.

Table 6.3 Performance of 22 Weka classifiers on predicting labels in the METABRIC discovery and validation sets

Cramer's V			
Type	**Classifier**	**Discovery**	**Validation**
bayes	BayesNet	0.93	0.93
	NaiveBayes	0.90	0.91
	NaiveBayesUpdateable	0.90	0.91
Functions	Logistic	0.94	0.92
	SimpleLogistic	0.95	0.94
	SMO	0.95	0.93
Lazy	IBk	0.94	0.91
meta	AttributeSelectedClassifier	0.89	0.89
	Bagging	0.88	0.90
	ClassificationViaRegression	0.91	0.90
	LogitBoost	0.93	0.94
	MultiClassClassifier	0.88	0.90
	RandomCommittee	0.94	0.94
Rules	DecisionTable	0.85	0.81
	JRip	0.87	0.93
	PART	0.92	0.91
trees	HoeffdingTree	0.90	NA
	J48	0.90	0.89
	LMT	0.95	0.94
	RandomForest	0.94	0.95
	RandomTree	0.88	0.87
	REPTree	0.86	0.85
Average		**0.91**	**0.91**
Standard Deviation		**0.03**	**0.03**

The almost perfect agreement on sample labelling defined by the interrater reliability metric Fleiss' kappa.

The Fleiss' kappa was computed to assess the agreement between classifiers and between samples labelling (**Table 6.4**). The first measurement indicates an overall agreement among individual classifiers of 0.957 for the discovery set and 0.941 for the validation set. The qualitative descriptions associated with intervals (described in *Chapter 4*) reveal an almost

perfect agreement for this case. The classifiers are attributing the same label to samples and it reflects more than would be expected by chance.

Table 6.4 Fleiss' kappa values and Adjusted Rand Index for the discovery and validation sets

Statistics	Fκ (1)	Fκ (2)	ARI
Discovery set	0.96	0.90	0.92
Validation set	0.94	0.88	0.89
Average	0.95	0.89	0.90
Standard Deviation	0.01	0.02	0.02

Note: ARI - Adjusted Rand Index; Fκ - Fleiss' kappa; (1) among classifiers; and (2) between refined METABRIC labels and predicted labels based on ensemble learning and meta-features.

The second statistic using Fleiss' kappa compared initial and predicted subtypes assigned by the majority of classifiers using meta-features in both METABRIC data sets. The results were 0.90 and 0.88 for the discovery and validation sets, respectively, also defining an almost perfect agreement between labelling. These results therefore confirm the relationship between subtypes displayed in the contingency table (**Table 6.2**), with the highest numbers on the diagonal. Although the highly concordant assignment, there is a small number for which a discrepancy was observed between labels marked with this approach and the previous study. These assignments, however, refer to samples with molecular ambiguities across subtypes, especially luminal A and B. Luminal tumours have the same tissue origin and share constituent similarities (Polyak, 2011).

The high agreement of predicted and initial labels according to the Adjusted Rand Index.

In order to infer a more consistent statistical analysis, the agreement between the different sample labelling was further scrutinised using the Adjusted Rand Index (**Table 6.4**). The value calculated for the METABRIC discovery set was 0.92 and for the validation was 0.89. These agreements between initial and predicted labels is significantly high.

6.3.3 Expanding Prediction Models Based on Microarray Data

It is reasonable to assume that the usage of meta-features introduce a striking generalisation of the breast cancer gene expression profile. Yet, the new 'constructs' reveal hidden insights in data that help explaining the breast cancer subtypes. Our approach embraces the dynamicity of the genome and the pairwise imbalances of breast tumours to predict more accurately the molecular subtypes. Additionally, the meta-features analysis would lead to improvements in the evaluation of the disease, currently guided by clinical markers such as oestrogen and progesterone receptor (ER and PR) status, and HER2 amplification (Ambs, 2010; Weigelt & Reis-Filho, 2009). The association of predicted labels and clinical information in the METABRIC data is well described in Milioli et al. (2015) and support our findings, as the concordance with our approach is high. The approach match with the next generation of medical applications that aim at covering molecular panels able to explain group of patients that likely have similar prognosis and survival (Perou et al., 2010).

Possible limitations of this study were also identified. Despite of the quality of the METABRIC data collection, there is a lack of matching between Illumina probes and other technologies, such as Affymetrix or Agilent, as per the annotation of (Dunning et al., 2010). The genetic background of each data set may result in distinct pairwise probes and independent molecular signatures for subtype assignments. Furthermore, the complexity imposed by different prediction models is another impediment to accurate subtype labelling. With advances in breast cancer research, there is a large number of flaws and errors in prediction methods to ascertain the sample subtype across different studies (Marchionni et al., 2013). However, there is no biological or mathematical reason to infer that a particular classification method is better than another since a range of distinct solutions is possible in the multidimensional gene expression space (Michiels et al., 2011). The last and major limitation is the uncertainty in the number of breast cancer subtypes (Jason I Herschkowitz et al., 2007; Hu et al., 2006; Lehmann et al., 2011). The true classification of the disease remains obscure, even though the description of the five intrinsic subtypes has had a substantial impact on the way how breast cancer is perceived. Prediction models and algorithms have, consequently, been affected by the fragmentary molecular taxonomy.

In this study, we introduced a novel strategy for subtype prediction by expanding the analysis of the METABRIC transcriptome data set. Simple mathematical modelling combined with well-established methodologies of feature selection and data mining revealed striking pairwise genome imbalances. Representative meta-features across intrinsic subtypes showed an extensive predictive power on labelling samples, in agreement with the new recently corrected

METABRIC labels. One advantage of this approach is the usage of fewer genes in comparison to the commercial PAM50 assay. Furthermore, the authenticity of the current systematic approach and the accuracy of its results demonstrated that it is a promising tool to predict intrinsic subtypes. The simplicity of our model provides an opportunity for wide application using a variety of data types with potential for progressing to clinical applications.

6.4 References

Albain, K. S., Barlow, W. E., Shak, S., Hortobagyi, G. N., Livingston, R. B., Yeh, I., et al. (2010). Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *Lancet Oncol., 11*(1), 55-65.

Ambs, S. (2010). Prognostic significance of subtype classification for short-and long-term survival in breast cancer: survival time holds the key. *PLoS Med., 7*(5), e1000281.

Bastien, R. R., Rodríguez-Lescure, Á., Ebbert, M. T., Prat, A., Munárriz, B., Rowe, L., et al. (2012). PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. *BMC Med. Genomics, 5*(1), 44.

Berretta, R., Costa, W., & Moscato, P. (2008). Combinatorial optimization models for finding genetic signatures from gene expression datasets. *Bioinformatics: Structure, Function and Applications* (Vol. 2, pp. 363-377): Humana Press.

Berretta, R., Mendes, A., & Moscato, P. (2007). Selection of Discriminative Genes in Microarray. Experiments Using Mathematical Programming. *Journal of Research and Practice in Information Technology, 39*(4), 287-299.

Buyse, M., Loi, S., Van't Veer, L., Viale, G., Delorenzi, M., Glas, A. M., et al. (2006). Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J. Natl. Cancer Inst., 98*(17), 1183-1192.

Chang, H. Y., Nuyten, D. S., Sneddon, J. B., Hastie, T., Tibshirani, R., Sørlie, T., et al. (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl. Acad. Sci. U. S. A., 102*(10), 3738-3743.

Chen, C., Dhanda, R., Tseng, W., Forsyth, M., & Patt, D. A. (2013). Evaluating use characteristics for the oncotype dx 21-gene recurrence score and concordance with chemotherapy use in early-stage breast cancer. *J Oncol Practice, 9*(4), 182-187.

Cotta, C., & Moscato, P. (2003). The k-Feature Set problem is W [2]-complete. *Journal of Computer and System Sciences, 67*(4), 686-690.

Cotta, C., Sloper, C., & Moscato, P. (2004). Evolutionary Search of Thresholds for Robust Feature Set Selection: Application to the Analysis of Microarray Data. *In Workshops on Applications of Evolutionary Computation, Springer-Verlag Berlin Heidelberg*, 21-30.

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature, 486*(7403), 346-352.

Drukker, C. A., van Tinteren, H., Schmidt, M. K., Rutgers, E. J., Bernards, R., van de Vijver, M. J., et al. (2014). Long-term impact of the 70-gene signature on breast cancer outcome. *Breast Cancer Res. Treat., 143*(3), 587-592.

Dunning, M. J., Curtis, C., Barbosa-Morais, N. L., Caldas, C., Tavaré , S., & Lynch, A. G. (2010). The importance of platform annotation in interpreting microarray data. *The Lancet Oncology, 11*(8), 717.

Ebbert, M., Bastien, R. R., Boucher, K. M., Martin, M., Carrasco, E., Caballero, R., et al. (2011). Characterization of uncertainty in the classification of multivariate assays: application to PAM50 centroid-based genomic predictors for breast cancer treatment plans. *J Clin Bioinform, 1*(1), 37.

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., et al. (2006). Concordance among gene-expression--based predictors for breast cancer. *N Engl J Med, 355*(6), 560-569.

Fan, C., Prat, A., Parker, J., Liu, Y., Carey, L. A., Troester, M. A., et al. (2011). Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Med. Genomics, 4*(1), 3.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull., 76*(5), 378-382.

Fleiss, J. L., Levin, B., & Paik, M. C. (2004). The Measurement of Interrater Agreement *Statistical Methods for Rates and Proportions* (pp. 598-626). New York: John Wiley & Sons, Inc.

Glas, A., Floore, A., Delahaye, L., Witteveen, A., Pover, R., Bakx, N., et al. (2006). Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC genomics, 7*(1), 278.

Gómez-Ravetti, M., Berretta, R., & Moscato, P. (2009). *Novel Biomarkers for Prostate Cancer Revealed by (α,β)-k-Feature Sets* (Vol. 5): Springer-Verlag Berlin Heidelberg.

Gómez-Ravetti, M., & Moscato, P. (2008). Identification of a 5-Protein Biomarker Molecular Signature for Predicting Alzheimer's Disease. *PLoS One, 3*(9), e3111.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., et al. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *J. Natl. Cancer Inst., 104*(4), 311-325.

Herschkowitz, J. I., Simin, K., Weigman, V. J., Mikaelian, I., Usary, J., Hu, Z., et al. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome biology, 8*(5), R76.

Herschkowitz, J. I., Simin, K., Weigman, V. J., Mikaelian, I., Usary, J., Hu, Z., et al. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol., 8*(5), R76.

Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics, 7*(1), 96.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193-218.

Kok, M. d., Koornstra, R. H., Mook, S., Hauptmann, M., Fles, R., Jansen, M. P., et al. (2012). Additional value of the 70-gene signature and levels of ER and PR for the prediction of outcome in tamoxifen-treated ER-positive breast cancer. *The Breast, 21*(6), 769-778.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., et al. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest., 121*(7), 2750-2767.

Liebetrau, A. M. (1983). *Measures of association* (Vol. 32). Beverly Hills, CA: SAGE Publications, Inc.

Loi, S., Haibe-Kains, B., Desmedt, C., Wirapati, P., Lallemand, F., Tutt, A., et al. (2008). Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics, 9*(1), 239.

Marchionni, L., Afsari, B., Geman, D., & Leek, J. (2013). A simple and reproducible breast cancer prognostic test. *BMC Genomics, 14*(1), 336.

Marsden, J., Budden, D., Craig, H., & Moscato, P. (2013). Language Individuation and Marker Words: Shakespeare and His Maxwell's Demon. *PLoS One, 8*(6), e66813.

McCafferty, M. P. J., Healy, N. A., & Kerin, M. J. (2009). Breast cancer subtypes and molecular biomarkers. *Diagn. Histopathol., 15*(10), 485-489.

Michiels, S., Kramar, A., & Koscielny, S. (2011). Multidimensionality of microarrays: statistical challenges and (im)possible solutions. *Mol. Oncol., 5*(2), 190-196.

Milioli, H. H., Vimieiro, R., Riveros, C., Tishchenko, I., Berretta, R., & Moscato, P. (2015). The Discovery of Novel Biomarkers Improves Breast Cancer Intrinsic Subtype Prediction and Reconciles the Labels in the METABRIC Data Set. *PLoS One, 10*(7), e0129711.

Naderi, A., Teschendorff, A. E., Barbosa-Morais, N. L., Pinder, S. E., Green, A. R., Powe, D. G., et al. (2006). A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene, 26*(10), 1507-1516.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., et al. (2004). A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *N. Engl. J. Med., 351*(27), 2817-2826.

Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W., et al. (2006). Gene Expression and Benefit of Chemotherapy in Women With Node-Negative, Estrogen Receptor Positive Breast Cancer. *Journal of Clinical Oncology, 24*(23), 3726-3734.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol., 27*(8), 1160-1167.

Perou, C. M., Parker, J. S., Prat, A., Ellis, M. J., & Bernard, P. S. (2010). Clinical implementation of the intrinsic subtypes of breast cancer. *Lancet Oncol., 11*(8), 718-719.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature, 406*(6797), 747-752.

Polyak, K. (2011). Heterogeneity in breast cancer. *J. Clin. Invest., 121*(10), 3786-3788.

Popovici, V., Chen, W. Y., Gallas, B., Hatzis, C., Shi, W., Samuelson, F., et al. (2010). Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res., 12*(1), R5.

Prat, A., Parker, J. S., Fan, C., & Perou, C. M. (2012). PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer. *Breast cancer research and treatment, 135*(1), 301-306.

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., et al. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res., 12*(5), R68.

Reis-Filho, J. S., & Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet, 378*(9805), 1812-1823.

Rocha de Paula, M., Gómez-Ravetti, M., Berretta, R., & Moscato, P. (2011). Differences in abundances of cell-signalling proteins in blood reveal novel biomarkers for early detection of clinical Alzheimer's disease. *PLoS One, 6*(3), e17481.

Seoane, J. A., Day, I. N. M., Gaunt, T. R., & Campbell, C. (2014). A pathway-based data integration framework for prediction of disease progression. *Bioinformatics, 30*(6), 838-845.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A., 98*(19), 10869-10874.

Sørlie, T., Tibshirani, R., Parker, J. S., Hastie, T., Marron, J. S., Nobel, A., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U. S. A., 100*(14), 8418-8423.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U. S. A., 99*(10), 6567-6572.

van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*(6871), 530-536.

van De Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med., 347*(25), 1999-2009.

Venet, D., Dumont, J. E., & Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol., 7*(10), e1002240.

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet, 365*(9460), 671-679.

Weigelt, B., Baehner, F. L., & Reis-Filho, J. S. (2010). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *J. Pathol., 220*(2), 263-280.

Weigelt, B., Mackay, A., A'hern, R., Natrajan, R., Tan, D. S. P., Dowsett, M., et al. (2010). Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol., 11*(4), 339-349.

Weigelt, B., & Reis-Filho, J. S. (2009). Histological and molecular types of breast cancer: is there a unifying taxonomy? *Nat. Rev. Clin. Oncol., 6*(12), 718-730.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann.

6.5 Supporting Information

Supporting Information – Text 6.1

Text 6.1 The (*α,β*)-k-Feature Set Problem

A combinatorial approach based on the mathematical model (*α,β*)-*k*-Feature Set Problem has been applied to select the best subset of features (a 'signature') able to discriminate two given classes of samples (Cotta et al., 2004). For the hypothetical matrix defined in **Table 6.5**, representing a microarray data set, consider a set of *m* samples (*Sample 1*, *2*, *3*, *4*, *5* and *6*), labelled for one of two possible classes, *F* or *G*. Each sample contains a boolean value for the *n* set of features (*Gene A*, *B*, *C*, *D*, and *E*); represented by 0 or 1, for *False* and *True*, respectively.

Table 6.5 An example of numerical matrix with five features and six samples belonging to class *F* or *G*.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Gene A	0	0	0	0	1	0
Gene B	1	1	1	1	0	1
Gene C	1	1	1	0	1	0
Gene D	1	1	1	1	0	0
Gene E	0	0	1	0	0	0
ClassLabel	F	F	F	G	G	G

Five features (Genes A, B, C, D and E) and six samples (Samples 1, 2, 3, 4, 5, and 6). Data adapted from: Berretta et al. (2008).

According to the data in **Table 6.5**, the decision problem *k*-Feature Set assesses if there is a set of *k* features (from the *n* set of features given) that can collectively explains each pair of samples belonging to the same class or to different classes (**Figure 6.6, a and b**). The problem can also be described using a graph; where we have a node for each feature, and a node for each pair of samples that belong to a different class and for each pair of samples that belong to the same class. Following the hypothetic data defined in **Table 6.5**, *Gene A* has the same value 0 for *Sample 1* and *Sample 2*, both from the same class *F*; thus, an edge from node *A* to node (1,2) is

added in a bipartite graph. In this case, feature *A* covers or explains the pair of samples (1,2) belonging to the same class. A similar approach is then used to assess the pair of samples from different classes. Again, *Gene A* has a distinct value for *Sample 1* and *Sample 5*, respectively, from classes *F* and *G*. Thus, there is an edge connecting both *A* and (1,5) nodes.

The generalisation of explaining pairs of samples in the same class (β) or in different classes (α) leads to the (α,β)-*k*-Feature Set problem, with two positive valued parameters: $\alpha \geq 1$ and $\beta \geq 0$ (Cotta et al., 2004). The goal is to find a *k* feature set in which every pair of samples from different classes can be "explained" by at least α features (**Figure 6.6**, **b**); and any pair of samples from the same class (identical values) by at least β features (**Figure 6.6**, **a**). The optimisation version of this problem has been formulated as an integer programming model (Berretta et al., 2008). A feasible solution, for $\alpha = 1$ and $\beta = 1$, is illustrated in (**Figure 6.7**).

The method intends to draw more robust signatures that contribute towards an improvement of research approaches and clinical applications. Furthermore, the *(α,β)-k*-Feature Set Problem is widely applicable to a range of biological data, such as genomics (SNPs, CNAs and CNVs), transcriptomics (gene expression) and proteomics information (Gómez-Ravetti et al., 2009).

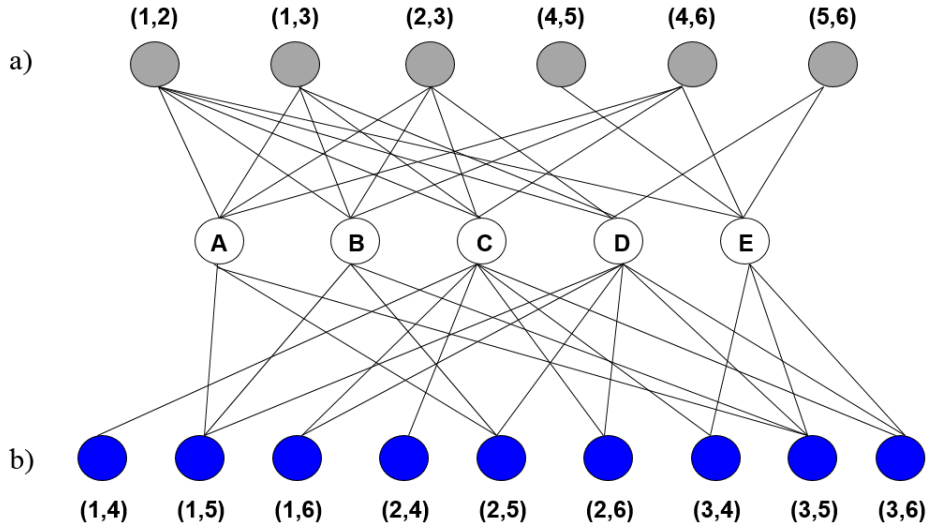


Figure 6.6 Graph representing an instance of the (*α*,*β*)-k-Feature Set; as per the data defined in Table 6.5.

In grey, are the nodes representing the pair of samples (*p*, *q*) from the same class; white nodes, a feature *i*; and blue nodes, the pair of samples (*p*, *q*) from different classes (*F* and *G*). *Figure adapted from Berretta et al. (2008).*

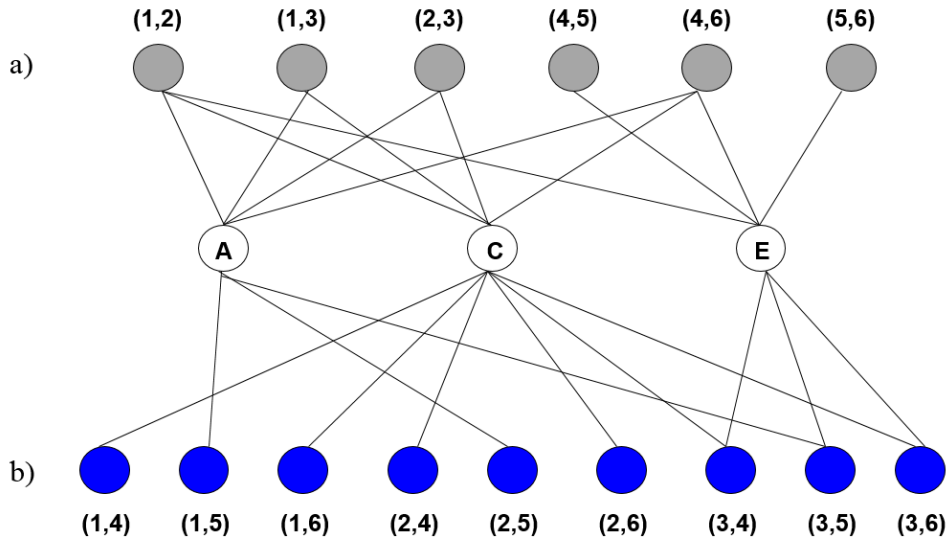


Figure 6.7 Graph containing a feasible solution for the (*α,β*)-k-Feature Set problem; as per the data defined in Table 6.5.

In grey, are the nodes representing the pair of samples (*p, q*) from the same class; white nodes, a feature *i*; and blue nodes, the pair of samples (*p, q*) from different classes (*F* and *G*). *Figure adapted from Berretta et al. (2008).*

Supporting References

Berretta, R., Costa, W., & Moscato, P. (2008). Combinatorial optimization models for finding genetic signatures from gene expression datasets. *Bioinformatics: Structure, Function and Applications* (Vol. 2, pp. 363-377): Humana Press.

Cotta, C., Sloper, C., & Moscato, P. (2004). Evolutionary Search of Thresholds for Robust Feature Set Selection: Application to the Analysis of Microarray Data. *In Workshops on Applications of Evolutionary Computation, Springer-Verlag Berlin Heidelberg*, 21-30.

Gómez-Ravetti, M., Berretta, R., & Moscato, P. (2009). *Novel Biomarkers for Prostate Cancer Revealed by (α,β)-k-Feature Sets* (Vol. 5): Springer-Verlag Berlin Heidelberg.

CHAPTER 7

7. BASAL-LIKE BREAST CANCER SUBTYPE

Among the five breast cancer intrinsic subtypes, the basal-like subtype is further explored in *Chapter 7*. These tumours form an important clinical group characterised by aggressive behaviour, poor prognosis and limited therapy response. However, the outcome of patients diagnosed within the basal-like subtype is contradictory. Some patients show increased risk of death within 5 years while others have a long-term survival of over 10 years. In this chapter, I investigate the genomic and transcriptomic signatures of 351 samples from the METABRIC and ROCK data sets to identify survival markers driving the disease outcomes. The content is available as a research paper at *BMC Medical Genomics*[9] and is presented here in sections **7.1 Introduction**, **7.2 Methods**, **7.3 Results**, **7.4 Discussion**, **7.5 Conclusion**,

[9] Milioli, H.H.*; Tishchenko, I.*; Riveros, C.; Berretta, R.; Moscato, P. Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. *BMC Med Genomics*; 10(1):19. *co-authorship.

7.6 References and

7.7 Supporting Information. In the breast cancer field, recognizing the disease's aggressive state is relevant to improving clinical decision-making, with the administration of effective tailored therapy for high-risk patients while avoiding aggressive treatments for low risk patients.

7.1 Introduction

Approximately 15% of all breast cancer cases are of basal-like subtype, often aggressive and highly recurrent lesions (Cleator et al., 2007; Lund et al., 2009; Millikan et al., 2008). Basal-like breast cancers (BLBCs) are defined by the lack of expression of the hormone receptors oestrogen (ER) and progesterone (PR), and the human epidermal growth factor receptor-2 (HER2) (Prat et al., 2013; Rody et al., 2011). Histologically, these tumours show high grade, high mitotic indices, presence of central necrotic or fibrotic zones, pushing borders of invasion, lymphocytic infiltrate and atypical medullary features (Putti et al., 2005). The breast basal cell layer is also characterised by high expression of cytokeratins (CK5/6, CK14, and CK17) and epidermal growth factor receptor (EGFR), amongst other markers (Badve et al., 2011; Cheang et al., 2008; Hallett et al., 2012; Nielsen et al., 2004; Valentin et al., 2012). All these features contribute to the limited therapeutic response and therefore impact in the refractory nature of these tumours (Kreike et al., 2007; Rakha et al., 2008). Thus, patients diagnosed with BLBC have a poor prognosis and a short-term disease-free and overall survival (Banerjee et al., 2006). A better understanding of the pathophysiology and molecular basis of basal-like tumours is necessary to delineate patient outcomes.

At the molecular level, basal-like tumours are considered more homogeneous than the immunohistochemically defined triple-negative breast cancers (TNBCs), even though the terminologies are used interchangeably (Bertucci et al., 2012; Cleator et al., 2007). Despite the relative molecular homogeneity, patients within this group still show divergent disease outcomes: some patients show high mortality and recurrence rates within the first 3-5 years, in contrast to others who survive over 10 years – with no recurrence – following the diagnosis (Banerjee et al., 2006; Carey et al., 2010; Rakha et al., 2008). For the latter group, the prognosis is better than those of luminal breast cancer subtype (Cheang et al., 2008; Mulligan et al., 2008). These observations suggest that BLBCs may be composed of at least two clinically distinct groups, with poor or excellent survival (Hallett et al., 2012). The molecular characterisation of these basal-like tumours is of particular interest in medicine since it may bring new insights to the disease understanding and management. Identifying markers and mechanisms involved in the differentiation of BLBCs is therefore an essential progression towards this end. Moreover, it would allow the development of tailored treatments with more effective individual response, leading to more personalised and conservative interventions for breast cancers (Fadare & Tavassoli, 2008).

Recent investigation of TNBCs pointed to the existence of intrinsic basal-like subtypes, with distinct molecular patterns (Burstein et al., 2014; Jézéquel et al., 2015; Lehmann et al., 2011). The stratification performed and described by Lehmann et al. (2011) (Lehmann et al., 2011) revealed the involvement of enriched cell cycle and cell division components in Basal-like 1 (BL1); growth factor signalling, glycolisis and gluconeogenesis pathways in Basal-like 2 (BL2); and immune cell processes in Immunomodulatory (IM). The authors also determined two other groups partially overlapping the basal-like subtype defined by the PAM50 classifier (Parker et al., 2009): Mesenchymal (M) and Mesenchymal stem-like (MSL). Alternatively, Burstein and colleagues (Burstein et al., 2014) defined the Basal-Like Immune-Suppressed (BLIS) and Basal-Like Immune-Activated (BLIA) subtypes. The former tumour type is characterised by multiple SOX family transcription factors, while the latter is described by Stat signal transduction molecules and cytokines. More recently, Jézéquel et al. (2015) (Jézéquel et al., 2015) pointed to two other groups: a basal-like with low immune response and high M2-like macrophages, and a basal-enriched with high immune response and low M2-like macrophages. All studies above described have focused on investigating the molecular heterogeneity of TNBCs, partially supporting each other.

Multi-gene models have also been applied to predict breast cancer subtype (Haibe-Kains et al., 2012; Parker et al., 2009), recurrence (Paik et al., 2004) and survival (Buyse et al., 2006; Glas et al., 2006). The selection of genes across samples has generally been associated with hormonal expression levels and proliferation modules. Since BLBCs and TNBCs are hormone receptor (ER and PR) negative and highly proliferative, the prediction power of markers to further separate patients at risk within these groups is of limited value in the current models (Liu et al., 2014). Clinical assays independently modelling triple-negative samples have revealed superior ability in predicting outcomes of early stage tumours (Yau et al., 2010; Yau et al., 2013). These assays and most approaches, however, have focused on the immunohistochemically defined TNBCs (Hallett et al., 2012; Sabatier et al., 2011; Teschendorff et al., 2007). A more robust approach for characterising BLBC outcomes is yet to be developed. Accordingly, a proper investigation of BLBCs remains mandatory and determinant for patients diagnosed within this subtype (Badve et al., 2011).

As the classification of TNBCs is not an ideal surrogate for defining BLBCs entities, a characterization of basal-like tumours at the genomic and transcriptomic levels is an urgent need. In this contribution, we aim at identifying markers associated with patients' survival using larger breast cancer cohorts from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis et al., 2012) and Research Online Cancer Knowledgebase (ROCK) (Ur-Rehman et al., 2013). Through the determination of this signature, our objective is

to stratify 351 tumours into basal-like subgroups, with varying clinical features and survival outcomes, and further describe each of them. Accordingly, we plan to explore the microarray data – including gene (mRNA) and microRNA (miRNAs) expression values, and copy number aberration (CNA) measurements – to expand the molecular characterisation of BLBCs, which to our knowledge has not yet been performed. The assessment of more comprehensive profiles of BLBCs is relevant for defining groups-at-risk in clinical settings and, more importantly, for improving therapy response.

7.2 Methods

7.2.1 Breast Cancer Data Sets

The METABRIC genomic and transcriptomic data sets were downloaded from the European Genome Phenome Archive (EGA) (http://www.ebi.ac.uk/ega), under the accession numbers EGAS00000000083 and EGAS00000000122. These publicly available collections contain genotyping (Affymetrix SNP 6.0), log2 normalised gene expression (Illumina Human WG-v3) and miRNA expression (Agilent ncRNA 60k) arrays for over 2000 breast tumours and 144 control (non-tumour) breast samples (Curtis et al., 2012). The original METABRIC study was approved by the ethics Institutional Review Boards in the UK and Canada (Addenbrooke's Hospital, Cambridge, United Kingdom; Guy's Hospital, London; Nottingham; Vancouver; Manitoba). Further analysis on this data was approved by the Human Research Ethics Committee (HREC) at the University of Newcastle, Australia (approval number: H-2013-0277).

The METABRIC cohort has a comprehensive description of patients long-term clinical and pathological outcomes. Tumour samples were assigned to a breast cancer subtype (luminal A, luminal B, HER2-enriched, normal-like, or basal-like) using an ensemble learning approach (Milioli et al., 2015), employing the set of 50 genes defined by Parker et al. (2009) (Parker et al., 2009). This approach has been previously shown to improve the samples classification and subtypes' assignment in METABRIC data set, and has revealed more consistency in terms of clinical features and survival outcomes (Milioli et al., 2015). Based on these labels, a subset of 250 basal-like tumours was selected for analysis in this study. For training and test purposes, this subset was randomly split into two sets of equal size (125) to avoid possible bias from the original cohort. The sets are hereafter referred to as the training and validation sets.

For additional validation across platforms, we used the ROCK data set obtained at Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/), under data source number GSE47561 (Sims et al., 2010; Ur-Rehman et al., 2013). This data set integrates ten different studies (GSE2034, GSE11121, GSE20194, GSE1456, GSE2603, GSE6532, GSE20437, GSE7390, GSE5847 and E-TABM-185) performed on the Affymetrix HG-U133A technology.

The compiled matrix contains log2 RMA renormalised gene expression values for 1570 tumour samples, of which are of basal-like subtype. The ROCK data set includes representative information for survival analysis, however, it lacks standard clinicopathological data which therefore has not been considered in this study.

7.2.2 Probe Selection Approach

Since the first aim of our study is to identify markers driving survival among basal-like patients, we designed a filtering technique to select a representative probe signature and reduce the bias arising from the high number of probes (48803) and low number of samples (125) in the training set. We defined two relevant criteria to select probes, which are involved in tumour initiation and/or progression, and are also correlated to survival, as detailed below.

The *Differential filter* (Tishchenko et al., 2016) was employed to select probes exhibiting distinct expression levels between tumours and controls. The underlying assumption is that probes truly correlated with breast cancer are linked to genomic changes or variations from healthy to cancerous tissue. We applied the Differential filter to each of the 48803 probes to test their separation power between the 125 tumours and 144 controls. This filter tests for three feasible cases: the expression levels in tumours are (a) *lower than*, (b) *higher than*, or (c) *lower and higher than* in control samples. The last case refers to genes that are up-regulated in some tumours and down-regulated in others, while the expression levels of controls lie between these two groups. To calculate a *p*-value for this case, we mirrored all expression levels on one side with respect to the mean value of controls. The separation power of each probe was defined as the minimal Wilcoxon test *p*-value calculated for the three cases. To determine the number of probes passing the *Differential filter*, we plotted an ordered $-\log_{10}$ normalised *p*-values against the corresponding probe ranks. The threshold was set approximately at the point of the highest curvature of this function. This threshold is based on the naturally emerging systemic behaviour and does not require an external definition. Probes passing this filter are referred to as the *differential probe set*.

The *Survival* filter (Tishchenko et al., 2016) – was used to further identify probes for which the expression levels are associated with patients' survival. This filter employs the Kaplan-Meier estimator to compute survival probabilities. The stratification power of each probe is calculated using the Log-rank test applied to two groups of samples corresponding to quantiles with the lowest and the highest expression values, respectively. We defined these quantiles by ordering all samples by their expression values of a probe and selected samples in the first and last thirds (the quantile from 0% to 33% in the relatively under-expressed and from 67% to 100% in the relatively over-expressed group). This analysis was performed in *R* using the *package survival* (Therneau, 2015). Since the survival information is not provided for all samples, this calculation was based on 115 basal-like tumour samples (from the total of 125) in the METABRIC training set. To determine the number of probes passing the Survival filter we used a similar threshold definition as for the Differential approach, i.e. by ordering the $-\log_{10}$ normalised *p*-values that emerged from the Log-rank test. These probes are further referred to as the *survival probe set*.

7.2.3 Clustering Basal-like Breast Cancer Samples

The second aim of our study is to identify and characterise basal-like subgroups with varying disease outcomes. To this end, we performed a hierarchical clustering of samples based on the previously defined *survival probe set*. This procedure exploits the assumption that probes showing most variations in expression and co-expression among each other are involved in similar biological mechanisms and have a high impact on the groups' delineation. To calculate the dissimilarity between the 115 samples from the METABRIC training set, for which the survival information is provided, we used the square root of the Jensen-Shannon divergence (Berretta & Moscato, 2010; Grosse et al., 2002; Merkin et al., 2012). We then generated the hierarchical clustering with the Ward's criterion that minimises the variance within clusters, using the *R package stats* (Murtagh & Legendre, 2013).

We further examined which probes from the survival probe set contribute the most to the separation of basal-like subgroups using the Wilcoxon test. We then ordered the $-\log_{10}$ normalised *p*-values to determine the probes that significantly differentiate between the subgroups by using the same threshold criterion as for the *Differential filter*. The purpose of this procedure is to refine the probes that best segregate basal-like subgroups of distinct disease outcome. These probes are further referred to as the *probe signature* and expose striking genes and cell mechanisms involved in the subgroups differentiation.

7.2.4 Validation across Data Sets

The basal-like entities were first matched to the METABRIC validation set by means of centroids computed based on the previously defined *probe signature*. Samples in this data set were then assigned to a subgroup according to the minimal Euclidean distance to a centroid.

An external validation was conducted on the ROCK data set, for which the centroids were mapped across technologies – from Illumina to Affymetrix – using the gene annotation packages *hgu133a.db* and *illuminaHumanv3.db* in *R* Bioconductor (Carlson, 2016a; Dunning et al., 2015; Gentleman, 2003). Since the mRNA level measurement and normalisation differ between METABRIC (Illumina) and ROCK (Affymetrix) data sets, we standardised the calculated centroid absolute values with respect to the average expression levels computed for all basal-like samples. This procedure is depicted in **Equation 7.1**, where $s_{i,j}$ is the expression value of probe *j* for sample *i*, and *N* is the total number of basal-like samples (*N* is equal to 115 in the METABRIC training set).

Equation 7.1 Normalisation

$$s_{i,j} = \frac{s_{i,j}}{\frac{1}{N}\sum_{i=1}^{N} s_{i,j}}$$

Following the centroids' normalisation, an analogous transformation of Affymetrix gene expression values was necessary to enable their direct application. Thus, we applied the same formula (Equation 7.1) to the ROCK data set, where the number *N* of total samples is 101. The assignment to subgroups was based on the minimal Euclidean distance to a standardised centroid.

7.2.5 Network Analysis

With the purpose to identify key players within the probe signature and their relation to each other, we generated and plotted a network graph using the Minimum Spanning Tree (MST) (Cormen, 2009). The distance $d(x, y)$ between two probes x, y is defined as $d(x, y) = \frac{1}{|\rho s(x,y)|}$ where $|\rho s(x, y)|$ is the value of the Spearman correlation between the probe expression calculated for 125 tumour samples from the training set.

To quantify the network analysis, we computed the betweenness centrality and node degree of each node (probe) using the *R* package *igraph* (Csardi & Nepusz, 2006). Generally, nodes with high betweenness centrality and degree values represent potential key players within the network. With regards to the centrality values, the most representative entities are highly connected to the rest of the tree; leaf-nodes have a betweenness centrality value of 0, while the most traversed nodes are assigned with the highest values (normalised up to 1). Node degree, on the other hand, is indicative of the number of direct neighbours of a node. Thus, probes with high degrees are also central (representative) for local groups with a relatively strong probe co-expression.

7.2.6 MicroRNA Differential Expression

To uncover the miRNAs differentiating the most between the basal-like subgroups, we applied the Wilcoxon test to expression values of each of the 853 probes available in the METABRIC data set. We considered those miRNAs with the emerging *p*-values smaller than 0.01 in both training and validation sets as relevant for the separation between basal-like subgroups. Both data sets were used due to the limited number of samples (146 in total) for which the miRNA expression profiles were provided. The miRNA probes were further investigated for possible target genes within the probe signature using *R* Bioconductor (*RmiR.Hs.miRNA* (Favero, 2013)) across five databases: miRBase, TarBase, PicTar, MirTarget2 and miRanda. For the miRNA and gene annotation we used the packages *hgug4112a.db* (Carlson, 2016b) and *lluminaHumanv3.db* (Dunning et al., 2015), respectively.

7.2.7 Copy Number Aberration Profiles

To quantise the CNA we employed the cytobands defined in the *hg18* data base that corresponds to the METABRIC platform. Aberrations were divided into two categories: losses (originally denoted as homozygous and heterozygous deletions) and gains (gains and amplifications). For each basal-like subgroup we then calculated the occurrence rates of gains and losses per cytoband, and applied the Binomial test to examine the hypothesis that the CNA distributions were the same among patient subgroups.

We further calculated the Percent Genome Altered (PGA) for each of the basal-like subgroups and applied the *Wilcoxon test* to these rates to obtain a significance value of the

difference between them. The aim of this approach is to identify stable/unstable genome profiles associated with the patient subgroups defined by our *probe signature* and to statistically describe whether they are consistently diverging.

7.3 Results

7.3.1 Survival-related Probes Defining Basal-like Subgroups

With the application of the *Differential* and *Survival filters* in the METABRIC training set we identified 15000 and 400 probes related to cancer initiation and/or progression, and patients survival, respectively. The corresponding probes in the differential probe set with distinct expression levels between tumours and controls showed significant *p*-values ranging from $2.36 \cdot 10^{-45}$ to $1.53 \cdot 10^{-7}$. The reduced number of probes in the survival probe set related to the individual survival had significant *p*-values ranging from $1.11 \cdot 10^{-4}$ to 0.04. These probes, ultimately, comprise a representative signature driving the outcome of basal-like patients in the METABRIC breast cancer cohort.

The hierarchical clustering of 115 basal-like samples based on the survival probe set has revealed two major subgroups: Basal I and Basal II, as shown in **Supporting Information – Figure 7.6**. A separation into more than two subgroups – in the next and subsequent hierarchical divisions in the dendrogram – was not supported due to the high similarity of subgroups in terms of their molecular profile and clinical outcome. The application of the Wilcoxon test has defined the probe signature containing the top 80 probes, with significant p-values ranging from $1.75 \cdot 10{-}13$ to $3.77 \cdot 10{-}4$, differentiating the most between the two basal-like groups at the transcriptomic (mRNA) level. A heat map of the 80-probe signature for the training set is plotted in **Figure 7.1**, where samples are ordered within each subgroup by their Euclidean distance to the corresponding centroids (**Supporting Information – Table 7.6, Table 7.7 and Table 7.8)**.

To characterise the 80-probe signature with respect to their cellular function, we clustered the probes by their mutual correlation into three groups – G1, G2 and G3 (

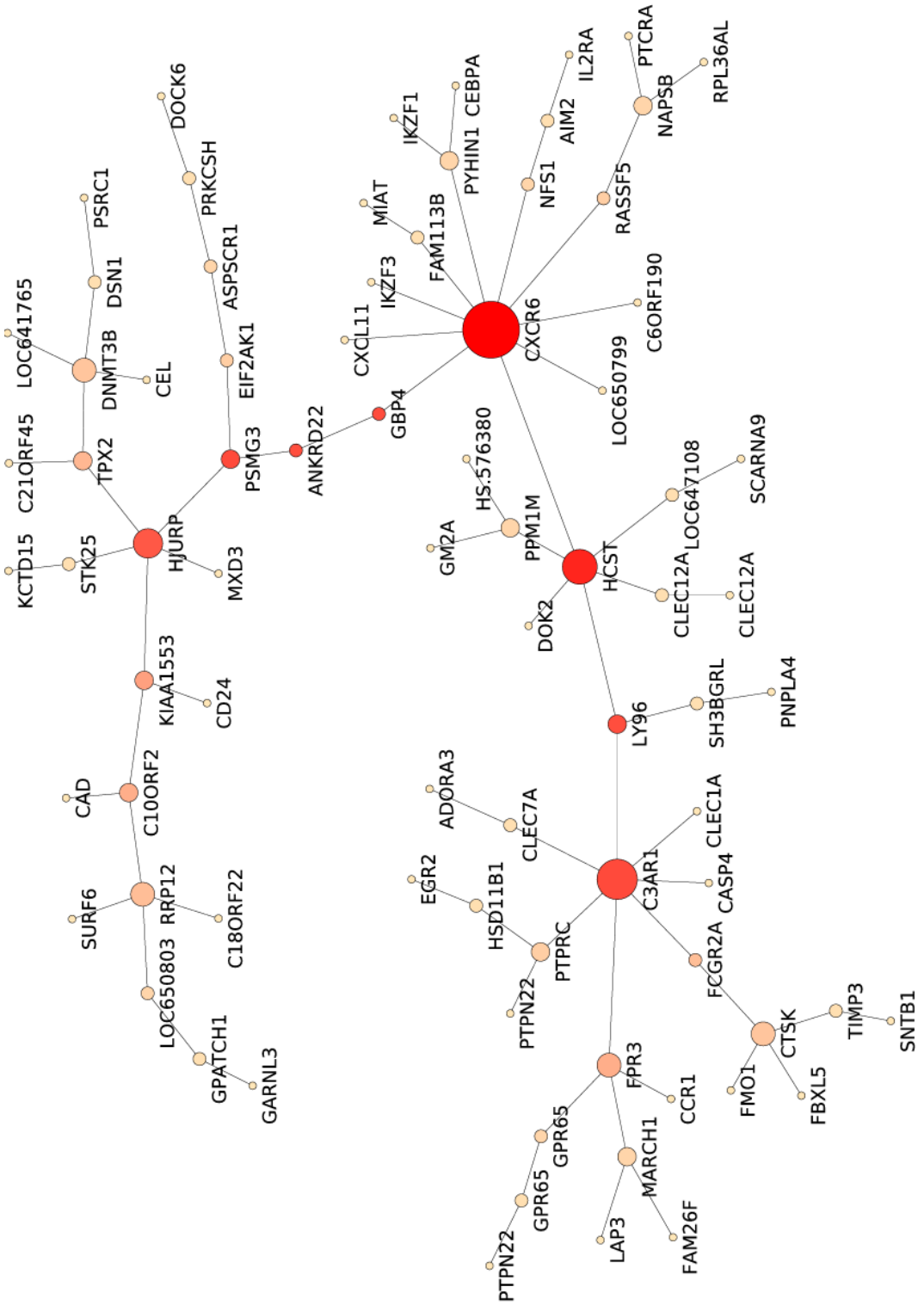


Figure 7.2 Minimum Spanning Tree of the 80-probe signature

The MST graph was generated for the 80 probes in the training set. Only probes with high correlation values between their expression levels are connected to a network. The size of each node is proportional to the computed node degree value (number of connections). The colour of

each node is reflective of the betweenness centrality value ranging between low (light pink) and high (red).

Table 7.1) – and functionally annotated using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (**Supporting Information – Table 7.9, Table 7.10 and Table 7.11)**. This analysis revealed that G1 probes are strongly associated with cell cycle control and cell division; they are over-expressed in Basal II subgroup. G2 showed relation to immune system and inflammatory response. Remarkably, the expression levels of G2 probes in Basal II are similar to that observed in controls, but much higher in Basal I, suggesting an intratumoral infiltration by lymphocytes in this subgroup. In the last group, G3, probes indicate an association (not significant) with metal-binding processes; they are under-expressed in Basal II when compared to Basal I and control samples.

The betweenness centrality and node degree analysis of the 80-probe signature (**Figure 7.2**) further outlined important genes differentiating between Basal I and Basal II subgroups (

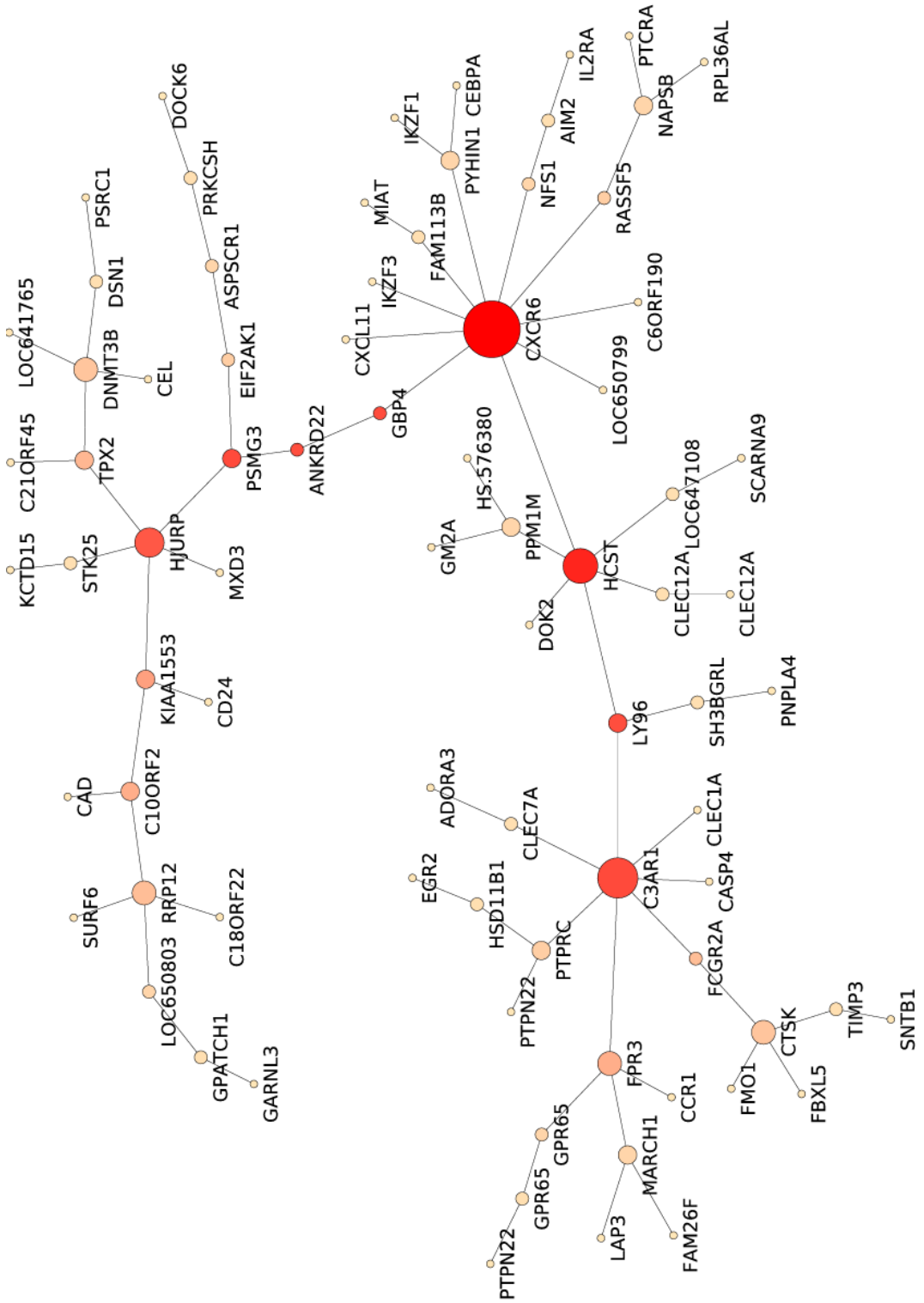


Figure 7.2 Minimum Spanning Tree of the 80-probe signature

The MST graph was generated for the 80 probes in the training set. Only probes with high correlation values between their expression levels are connected to a network. The size of each node is proportional to the computed node degree value (number of connections). The colour of

each node is reflective of the betweenness centrality value ranging between low (light pink) and high (red).

Table 7.1). The genes with the highest centrality values (B ≥ 0.1) in G1 are *PSMG3*, *HJURP*, *BEND3*, C10orf2, *TPX2*, *RRP12* and *DNMT3B*; in G2, *CXCR6*, *HCST*, *C3AR1*, *GBP4*, *LY96*, *ANKRD22*, *FPR3* and *FCGR2A*; and in G3, *CTSK*. Within this set, the genes *HJURP*, *RRP12*, *DNMT3B*, *CXCR6*, *HCST*, *C3AR1*, *FPR3* and *CTSK* also showed high node degree values (ND ≥ 4), representative for probe co-expression, corroborating with their key role on the differentiation of basal-like carcinomas.

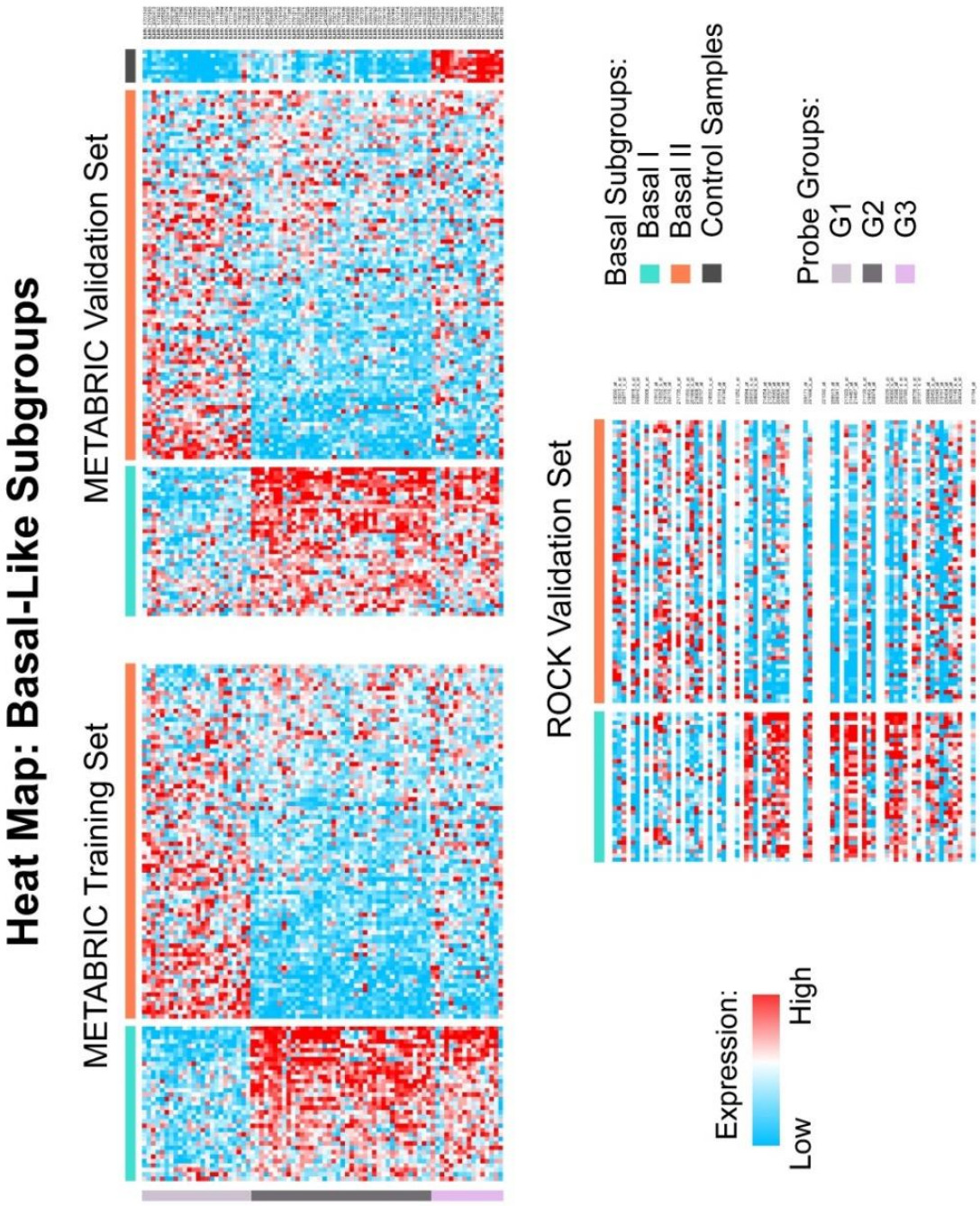


Figure 7.1 Heat map of the 80-genes signature in METABRIC training set

The figure displays 80 survival-related genes ordered by their overall rank within each basal subgroup. The probes expression levels in the METABRIC validation set were defined after the computation of centroids in the training set. Samples in each basal subgroup are ordered by their overall rank and the expression values are normalised across individuals. In the ROCK data set based on the Affymetrix identifier, all samples are ordered by their overall rank of 55 probes and split into two groups using the rank value range. The 55 Affymetrix probes correspond to 80 Illumina features defined in the METABRIC data set.

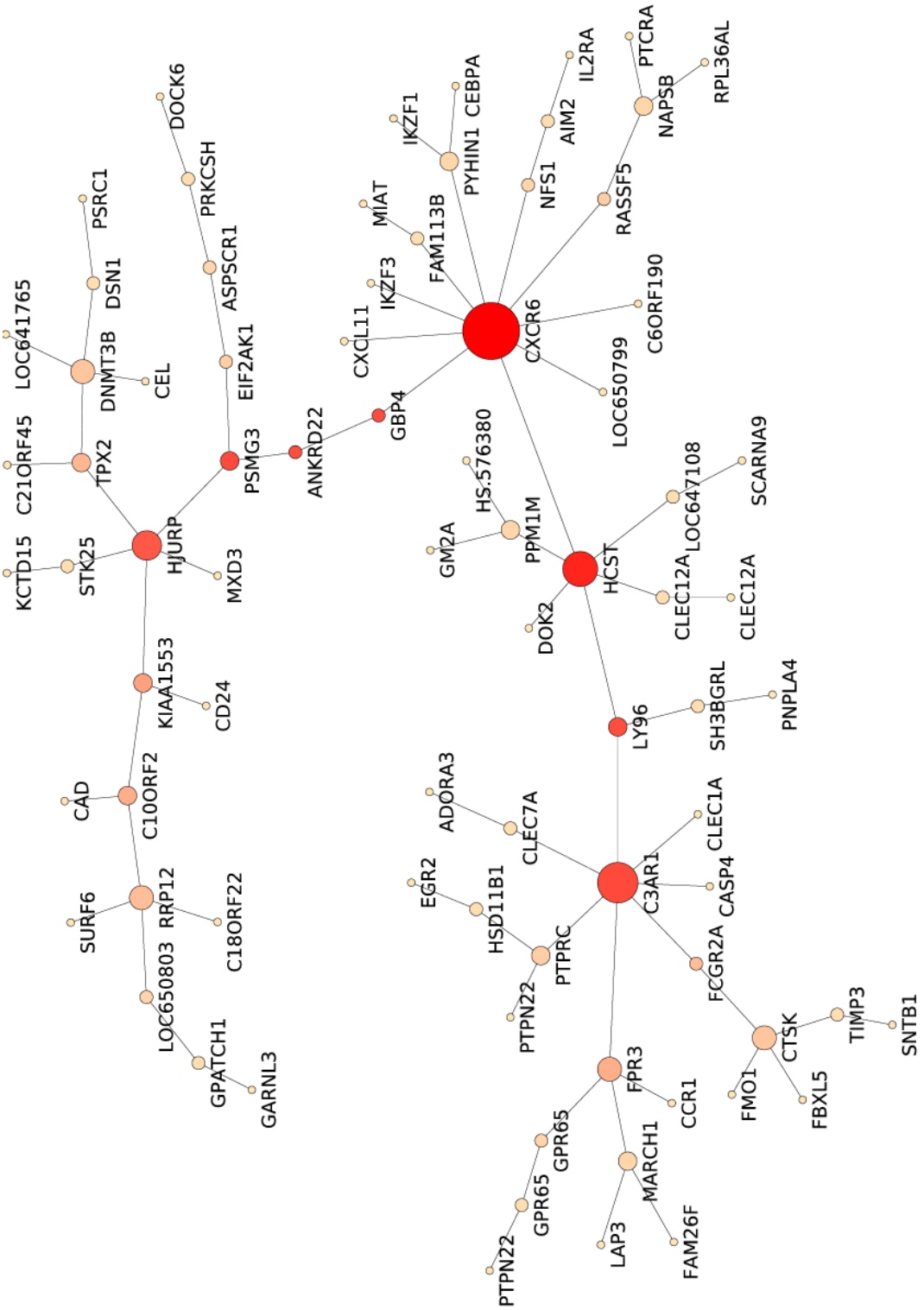


Figure 7.2 Minimum Spanning Tree of the 80-probe signature

The MST graph was generated for the 80 probes in the training set. Only probes with high correlation values between their expression levels are connected to a network. The size of each node is proportional to the computed node degree value (number of connections). The colour of each node is reflective of the betweenness centrality value ranging between low (light pink) and high (red).

Table 7.1 The 80-genes signature related to survival

Groups	Gene	Illumina_ID	B	ND
G1	C10orf2	ILMN_1701243	0.17	3
	RRP12	ILMN_1767253	0.12	4
	CD24	ILMN_2060413	0	1
	SURF6	ILMN_1778032	0	1
	GPATCH1	ILMN_1655625	0.03	2
	CEL	ILMN_1723418	0	1
	LOC641765	ILMN_1692198	0	1
	DNMT3B	ILMN_2328972	0.1	4
	MIS18A	ILMN_1712386	0	1
	DSN1	ILMN_1715905	0.03	2
	TPX2	ILMN_1796949	0.14	3
	HJURP	ILMN_1703906	0.42	5
	CAD	ILMN_1810992	0	1
	BEND3	ILMN_2375032	0.21	3
	EIF2AK1	ILMN_2156267	0.07	2
	PSMG3	ILMN_1802627	0.47	3
	MXD3	ILMN_1711904	0	1
	PSRC1	ILMN_2315964	0	1
	ASPSCR1	ILMN_1660749	0.05	2
	PRKCSH	ILMN_1777794	0.03	2
	LOC650803	ILMN_1803510	0.05	2
	KCTD15	ILMN_1786326	0	1
	RBFA	ILMN_1736130	0	1
	STK25	ILMN_1668090	0.03	2
G2	PYHIN1	ILMN_1742026	0.05	3
	THEMIS	ILMN_1684040	0	1
	PCED1B	ILMN_1712431	0.03	2
	PTCRA	ILMN_2091920	0	1
	HCST	ILMN_2396991	0.57	6
	LY96	ILMN_1724533	0.45	3
	CASP4	ILMN_1678454	0	1
	SNTB1	ILMN_1793410	0	1
	GBP4	ILMN_1771385	0.46	2
	DOK2	ILMN_1791211	0	1
	GM2A	ILMN_2221046	0	1
	FPR3	ILMN_2203271	0.17	4
	C3AR1	ILMN_1787529	0.47	7

	FCGR2A	ILMN_1666932	0.12	2
	CCR1	ILMN_1678833	0	1
	LOC647108	ILMN_1774206	0.03	2
	CLEC12A	ILMN_2403228	0	1
	CLEC12A	ILMN_1663142	0.03	2
	ADORA3	ILMN_1730710	0	1
	CLEC7A	ILMN_1700610	0.03	2
	LOC650799	ILMN_1715436	0	1
	MIAT	ILMN_1864900	0	1
	IKZF3	ILMN_2300695	0	1
	ANKRD22	ILMN_2132599	0.45	2
	AIM2	ILMN_1681301	0.03	2
	IL2RA	ILMN_1683774	0	1
	MARCH1	ILMN_2094942	0.05	3
	LAP3	ILMN_1683792	0	1
	GPR65	ILMN_2232121	0.03	2
	GPR65	ILMN_1734740	0.05	2
	FAM26F	ILMN_2066849	0	1
	CXCL11	ILMN_2067890	0	1
	NFS1	ILMN_1761314	0.05	2
	CXCR6	ILMN_1674640	0.68	10
	RASSF5	ILMN_2362902	0.07	2
	NAPSB	ILMN_1723043	0.05	3
	IKZF1	ILMN_1676575	0	1
	PTPN22	ILMN_1715885	0	1
	PTPRC	ILMN_1653652	0.07	3
	PTPN22	ILMN_2246328	0	1
G3	RPL36AL	ILMN_2189936	0	1
	GARNL3	ILMN_1779347	0	1
	PNPLA4	ILMN_1664348	0	1
	SH3BGRL	ILMN_1702835	0.03	2
	HS.576380	ILMN_1848030	0	1
	FMO1	ILMN_1684401	0	1
	CTSK	ILMN_1758895	0.1	4
	EGR2	ILMN_1743199	0	1
	CLEC1A	ILMN_1691339	0	1
	HSD11B1	ILMN_2389501	0.03	2
	CEBPA	ILMN_1715715	0	1
	TIMP3	ILMN_1701461	0.03	2

FBXL5	ILMN_1673370	0	1
SCARNA9	ILMN_1805064	0	1
PPM1M	ILMN_1657810	0.05	3
DOCK6	ILMN_1801226	0	1

Note: The 80 annotated Illumina probes distinguishing the basal-like subgroups are listed in this table, in the same order as in the heat map. The Gene Symbol and Illumina probe IDs are defined for each probe **Group**. This table also contains the betweenness centrality (**B**) and node degree (**ND**) values calculated for each probe in the basal-like training set.

7.3.2 Basal I and Basal II Validated across Independent Data Sets and Microarray Platforms

The quality of the 80-probe signature was evaluated using centroids calculated for the training set and applied to the METABRIC and ROCK validation sets. In ROCK, 55 annotated probes matched from Illumina to Affymetrix and were validated across the microarray platforms. The corresponding heat maps, in **Figure 7.1**, showed the existence of two main basal-like subgroups, Basal I and Basal II, in both METABRIC and ROCK validation sets. The two subgroups are consistent with regards to the population size and mRNA expression levels (in G1, G2 and G3) and further support the quality of the 80-probe signature. The definition of more than two subgroups in the hierarchical clustering would lead to the separation of entities with highly similar molecular profiles.

7.3.3 Clinical Features and Survival Outcomes Supporting the Basal-like Subgroups

The analysis of clinicopathalogical markers revealed a significant correlation between the basal-like subgroups defined in this study and tumour histology (Invasive Ductal Carcinoma versus medullary type), tumour size and p53 status (**Table 7.2**). According to histological classification, the medullary type is more common among Basal I patients. On the other hand, the Basal II subgroup is characterised by larger tumours (in size) and a higher frequency of p53 mutation. Clinical features, such as age, menopausal status (MS), grade, Nottingham Prognostic Index (NPI) and lymph nodes, did not show statistically significant variations across the two basal-like subgroups.

The survival analysis revealed significant differences in patients' outcome between Basal I and Basal II. Basal I showed a better prognosis in comparison to Basal II in all data sets (**Figure 7.3**), with the Log-rank test *p*-values of 0.0097, 0.017 and 0.043 for the METABRIC training, validation and ROCK data sets, respectively.

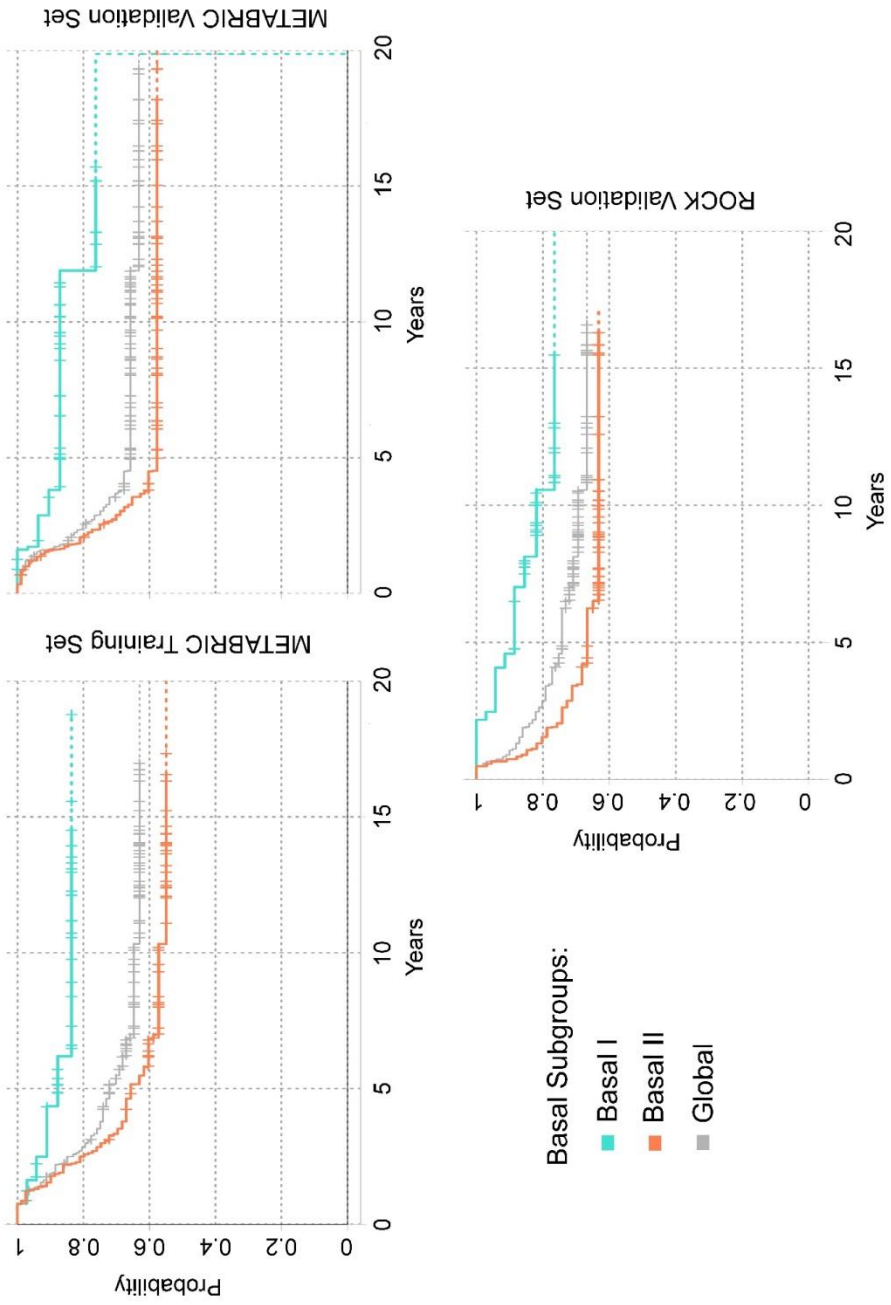


Figure 7.3 Survival curves in the METABRIC and ROCK data sets

The survival analysis was performed using the Kaplan-Meier curves. The grey line shows the disease specific survival of the basal-like subtype in the training and validation sets. Basal I subgroup is shown in turquoise, while Basal II in coral. Ticks represent sensors of patients who are alive and drops denote deaths. Lines based on the last ten observations are plotted in dash.

Table 7.2 Clinical information of patients and tumour samples in the METABRIC data set

		Training Set		**Validation Set**	
		Basal I	**Basal II**	**Basal I**	**Basal II**
Age [y.]	≤ 40	7	18	4	17
	41 to 50	11	18	10	21
	51 to 60	8	20	9	16
	> 60	9	24	13	35
	mean	50.6	52.5	54.7	54.1
	p-value	0.46		0.8	
MS	pre/post	18/17	36/43	15/21	37/52
	pre/post (%)	48.60%	54.40%	58.30%	58.40%
	mean	0.31		1	
Size [cm]	≤ 2 cm	15	30	17	32
	> 2 cm	20	50	19	55
	mean	2.35	2.97	2.26	2.9
	p-value	0.01		0.005	
Grade	grade 2	2	8	5	3
	grade 3	33	71	30	85
	mean	2.9	2.9	2.9	3
	p-value	0.4		0.092	
NPI	≤ 2.4	0	1	1	1
	2.4 to 3.4	1	6	3	2
	3.4 to 5.4	28	62	27	77
	> 5.4	6	11	5	9
	mean	4.7	4.6	4.5	4.6
	p-value	0.43		0.7	
LN	neg/pos	16/19	37/43	17/19	47/42
	neg/pos (%)	45.7%	46.2%	47.2%	52.8%
	p-value	1		0.34	
Histology	IDC	28	71	23	84
	ILC	0	3	1	2
	IDC-med	7	5	9	3
	others	0	0	3	1
	medullary (%)	20%	6.25%	25%	3.37%
	p-value	0.001		$5.4 \cdot 10^{-4}$	
P53	mut/wild	1/15	11/14	2/11	12/17
	mut/wild (%)	6.25%	44%	15.40%	41.40%
	p-value	$1.1 \cdot 10^{-7}$		$7 \cdot 10^{-4}$	
Population size		**35**	**80**	**36**	**89**

Note: The patients Age at diagnosis and menopausal status (MS) are listed for each subgroup. Median values were calculated for some of the variables. The clinicopathological characteristics described for the tumours are: Size, Grade, Nottingham prognostic index (NPI) and Lymph node (LN) status. The total number of lymph node positives and collected among patients are first detailed; followed by the ratio of patients with node negative and positive, in absolute values and percentage. The P53 wild and mutation status is also defined in absolute values and percentage. Tumour Histology is described for samples diagnosed with Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC), medullary carcinoma and others (tubular, mucinous and phyllodes tumours). The number of patients in each group is indicated in Population size.

7.3.4 MicroRNAs Differentially Expressed between Basal I and Basal II

We identified 17 miRNAs and 2 putative probes differentially expressed between the two basal-like subgroups (**Table 7.3**), with the *Wilcoxon test p*-values smaller than 0.01 in both METABRIC data sets (**Supporting Information - Table 7.12**). The probes hsa-miR-155, -342-5p and -150 showed the lowest p-values and an overexpression in Basal I, when compared to Basal II and control samples. The transcripts hsa-miR-19b-1*, -17* and -200c*, on the other hand, were over-expressed in Basal II tumours relative to Basal I and controls. The expression levels of all probes are depicted in **Figure 7.4**. Additionally, the identified miRNAs were matched against the 80-probe signature revealing a set of 50 gene-targets across five distinct databases, as listed in **Table 7.4** and further detailed for Basal I and Basal II in **Supporting Information – Table 7.12, Table 7.13** and **Table 7.14**. Among the gene-targets, C10orf2, *HSD11B1*, *EGR2*, *FBXL5*, *CLEC7A*, *DNMT3B*, *FMO1*, *CTSK* and *PYHIN1* were present in at least two databases. A comparison between miRNA and gene expression levels across subgroups showed significant correlations of hsa-miR-142-5p and *RASSF5*, hsa-miR-142-5p and *TIMP3*, hsamiR-150 and *MIAT*, and hsa-miR-22 and *TIMP3* in both Basal I and Basal I.

Table 7.3 MicroRNAs differentiating basal-like breast cancer subgroups

SGs	miRNA	Probe_IDs	*p*-value
BI	hsa-put-miR-92597	CRINCR2000005427	$2.8 . 10^{-4}$
	hsa-miR-361-3p	A_25_P00012305	$2.8 . 10^{-4}$
	hsa-miR-342-3p	A_25_P00012357	$4 . 10^{-4}$
	hsa-miR-140-3p	A_25_P00012177	$1.3 . 10^{-4}$
	hsa-miR-34a	A_25_P00012086	$4.9 . 10^{-3}$
	hsa-miR-22	A_25_P00010204	$6.3 . 10^{-3}$
	hsa-miR-142-5p	A_25_P00014844	$2 . 10^{-4}$
	hsa-miR-142-3p	A_25_P00011016	$2.2 . 10^{-3}$
	hsa-miR-155	A_25_P00012271	$6.3 . 10^{-6}$
	hsa-miR-342-5p	A_25_P00012354	$2 . 10^{-7}$
	hsa-miR-150	A_25_P00014847	$8.7 . 10^{-6}$
	hsa-put -miR-4391	CRINCR2000005084	$1.2 . 10^{-4}$
	hsa-miR-29c	A_25_P00012274	$6.7 . 10^{-3}$
	hsa-miR-29c*	A_25_P00013484	$5.6 . 10^{-4}$
	hsa-miR-29a	A_25_P00012013	$4.8 . 10^{-3}$

BII	hsa-miR-19b-1*	A_25_P00013163	$5.3 . 10^{-4}$
	hsa-miR-17*	A_25_P00013151	$5 . 10^{-4}$
	hsa-miR-17	A_25_P00013841	$1.9 . 10^{-3}$
	hsa-miR-200c*	A_25_P00013469	$1.8 . 10^{-4}$

Note: The list of miRNAs distinguishing Basal I (BI) and Basal II (BII) subgroups, with the corresponding *p*-value in the METABRIC training set. Probe IDs and annotated miRNA from Agilent platform are defined for each basal-like subgroup.

Table 7.4 MicroRNAs and corresponding target genes

miRNA	**Targets**
hsa-miR-361-3p	*C3AR1, CEBPA, GM2A, MIAT, SURF6,TIMP3*
hsa-miR-342-3p	*MXD3, PSMG3, PTCRA, PTPRC, TIMP3*
hsa-miR-140-3p	*C10orf2 , CXCL11, KCTD15, PNPLA4, PRKCSH, RRP12, STK25*
hsa-miR-34a	*CXCL11, DSN1, FCGR2A, GPR65, IKZF3, PNPLA4*
hsa-miR-22	*DOK2, GM2A, HSD11B1, MXD3, PNPLA4, STK25, TIMP3*
hsa-miR-142-5p	*C10orf2, CD24, CEBPA, EGR2, FBXL5, FPR3, HSD11B1, RASSF5, TIMP3*
hsa-miR-142-3p	*CD24, EGR2, PNPLA4, SH3BGRL*
hsa-miR-155	*PSRC1, RBFA*
hsa-miR-342-5p	*ASPSCR1, CASP4, IKZF1, PSRC1*
hsa-miR-150	*CCR1, EGR2, FBXL5, MIAT*
hsa-miR-29c	*CLEC7A, DNMT3B, FCGR2A, FMO1, KCTD15, MIAT, TPX2*
hsa-miR-29c*	*GARNL3, HJURP, MIS18A*
hsa-miR-29a	*CLEC7A, DNMT3B, FCGR2A, FMO1,KCTD15, MIAT, TPX2*
hsa-miR-19b-1*	*CXCR6, FCGR2A, HSD11B1, MXD3*
hsa-miR-17	*AIM2, BEND3, CEL, CTSK, EGR2, FBXL5, PNPLA4, PYHIN1, SNTB1, TIMP3*
hsa-miR-200c*	*DOK2, HJURP, IL2RA, PSRC1, RRP12*

Note: Differentially expressed miRNAs and corresponding target genes in the 80-probe signature. The matching targets were listed in five databases: miRBase, TarBase, PicTar, MirTarget2 and miRanda. Target genes that were present in at least two databases are underlined.

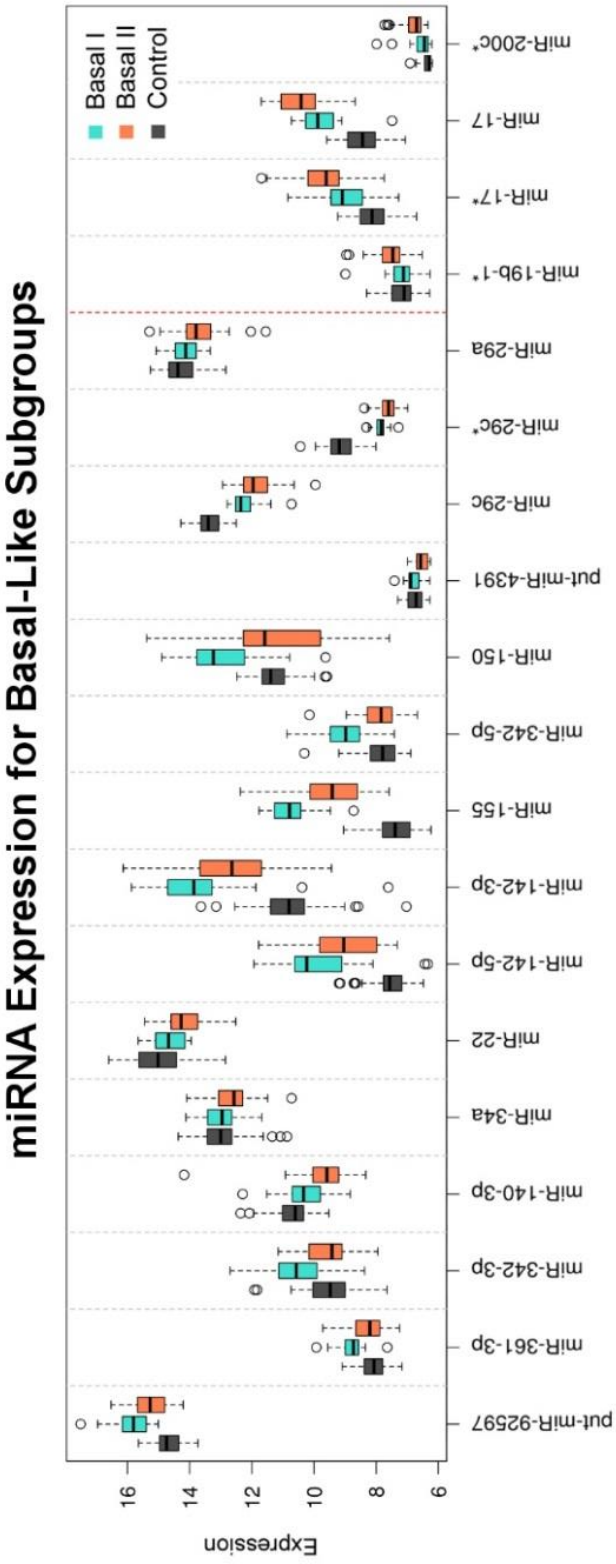


Figure 7.4 The box plot of miRNAs differentiating Basal I and Basal II subgroups

The image shows the expression levels of 19 miRNAs across basal-like subgroups and control samples in the METABRIC data set.

7.3.5 Copy Number Aberration Profiles Further Differentiating Basal-like Subgroups

The integrated analysis of CNA has revealed an increasing number of genomic changes from Basal I to Basal II subgroup (**Figure 7.5**) and uncovered cytobands with significant aberrations (binomial test *p*-values < 0.15) in both METABRIC training and validation sets (**Table 7.5**). Accordingly, critical gains/amplifications were detected on chromosomes 1q, 3q, 8q, 10p and 17q, and losses/deletions on 4q, 5q, 8p, Xp and Xq. Several of these aberrations have been previously associated with primary breast tumours and cell lines in BLBCs and/or TNBCs studies (Burstein et al., 2014; Engebraaten et al., 2013; Kao et al., 2009; Loo et al., 2011; Weigman et al., 2012). Notably, the percent of the genome being altered in the training set for Basal I was 2.74% for gains and 0.23% for losses; in Basal II it was 9.06 and 1.03%, respectively. The *Wilcoxon test* showed significant heterogeneity among the subgroups for the gains (*p*-value = $1.91 \cdot 10^{-6}$) and for losses (*p*-value = $9.55 \cdot 10^{-4}$). The same pattern was observed in the validation set for Basal I (3.58% for gains and 0.13%) and Basal II (10.46% for gains and 2.54%), also highly significant (*Wilcoxon test*: *p*-value = $1.11 \cdot 10^{-6}$ for gains and *p*-value = $5.37 \cdot 10^{-6}$ for losses). The increasing genome instability represented by increasing PGA, plotted in **Figure 7.5**, occurred consistently, from Basal I to Basal II, with the decreasing rates of patients' survival.

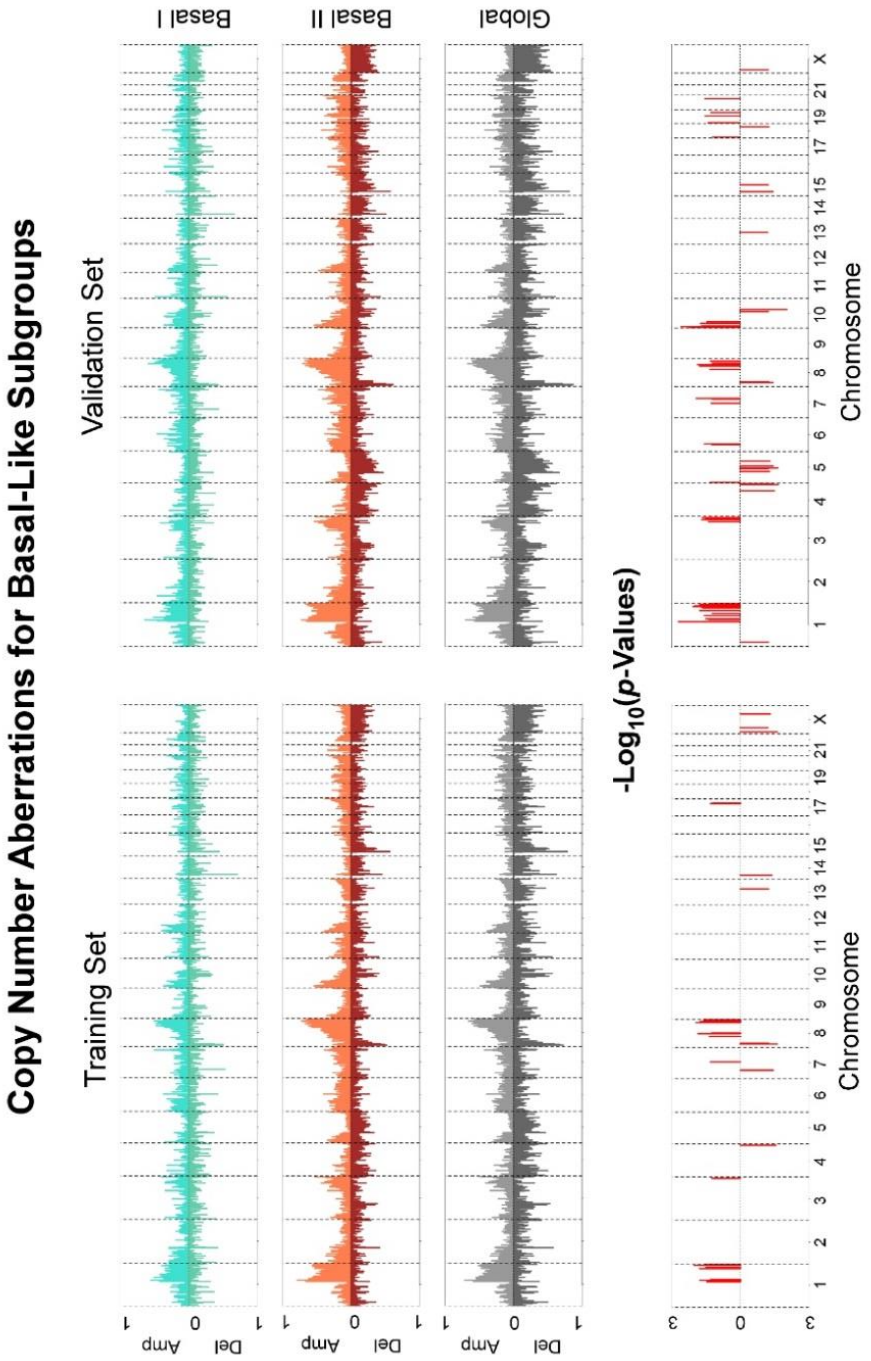


Figure 7.5 Copy number aberration of basal subgroups in METABRIC data set

The x-axis corresponds to 23 chromosomes including the X chromosome, while the y-axis represents the percentage of the population with amplifications or deletions in certain cytobands of these chromosomes. Positive values correspond to amplifications, while the negative ones to deletions.

Table 7.5 Cytobands associated with significant CNA acquisitions

Type	Cytoband	*p*-value (T)	*p*-value (V)
Gain	1q21.1	0.055	0.0012
	1q22	0.012	0.033
	1q23.1	0.038	0.024
	1q24.1	0.123	0.064
	1q32.3	0.099	0.14
	1q42.11	0.012	0.017
	1q42.12	0.08	0.015
	1q42.13	0.023	0.0059
	1q42.3	0.12	0.03
	1q43	0.0063	0.012
	1q44	0.021	0.049
	3q28	0.044	0.016
	8q13.2	0.044	0.054
	8q21.13	0.14	0.037
	8q22.1	0.092	0.06
	8q22.2	0.097	0.065
	8q23.2	0.12	0.0096
	8q24.11	0.075	0.049
	8q24.21	0.05	0.039
	8q24.22	0.012	0.086
	10p15.3	0.1	0.004
	10p12.32	0.12	0.013
	17q25.1	0.06	0.1
Loss	4q35.1	0.021	0.015
	5q12.2	0.15	0.04
	5q14.3	0.06	0.1
	8p21.2	0.046	0.027
	8p21.1	0.085	0.043
	Xp22.13	0.066	0.046
	Xp21.2	0.049	0.059
	Xq13.3	0.066	0.053
	Xq21.2	0.14	0.12
	Xq21.32	0.066	0.053

Note: Chromosomes and cytobands are defined for varying (type) copy number aberration, gains and losses, among basal-like tumours. The columns T and V represent the METABRIC training and validation sets, respectively.

7.4 Discussion

7.4.1 Survival-related Probes Defining the Molecular Signature of Basal-like Breast Cancer Subgroups

The basal-like subgroups defined in this study show distinct patterns in terms of tumour molecular profiles, clinicopathological features and patients survival outcomes. The characterisation of BLBCs, considering the two major entities Basal I and Basal II, is supported by the identification of the 80-probe signature, validated across Illumina and Affymetrix platforms in the METABRIC and ROCK cohorts. The importance of this signature, genes and gene-families, is defined by their functionality for each set: G1, G2 and G3. The annotated probes revealed their association with cell cycle and cell division components, immune/inflammatory regulation and metal binding, respectively, and defined Basal I (Immune Active) and Basal II (High Proliferative) subgroups. In Basal I, the over-expression of G2 probes suggests an immune activation and lymphocytic infiltration, particularly regulating tumour growth and patients' survival. This role has been previously associated with a better prognosis and therapy response (Andre et al., 2013), and has the potential to stratify basal-like breast cancers. On the other hand, the over-expression of G1 cell cycle-related genes and under-expression of G3 metal binding genes in Basal II impact on cell proliferation rates and energy metabolism. In this case, the cells reproduce at a rate far beyond the common bounds of a controlled cell cycle, concomitantly with other molecular changes in metabolic processes.

The G1 genes *PSMG3*, *HJURP*, *BEND3*, *TPX2*, *RRP12* and *DNMT3B* exhibited the highest centrality values and were over-expressed in the Basal II subgroup. *HJURP*, for instance, plays a central role in the maintenance of newly replicated centromeres and mitotic regulation. Increased levels of this gene in primary tumours and breast cancer cell lines have been previously correlated to decreased disease-free and overall survival (Z. Hu et al., 2010). Also involved in the mitotic spindle assembly, *TPX2*, when over-expressed, has been associated with proliferation networks and metastasis enhancement, holding a prognostic value for breast cancer patients (Geiger et al., 2014). Additionally, the hyperactivity of the DNA methyltransferase enzymes, or the over-expression of *DNMT3B*, has been further reported in BLBCs and TNBCs, where the hypermethylation events were more frequent than in other breast cancer subtypes (Roll et al., 2013). Hypermethylated tumours also presented decreased levels of regulatory miRNAs, including hsa-miR-29a and -29b. In particular, the under-expression of hsa-miR-29c has been marked as characteristic of BLBCs, segregating them into two subsets

(Sandhu et al., 2014), which has been supported by our findings. More studies, however, are required to investigate the biological role of other representative genes, such as *PSMG3*, *BEND3* and *RRP12* in G1.

A number of G2 genes are key regulators of the basal-like tumorigenesis, such as *CXCR6*, *HCST*, *C3AR1*, *GBP4*, *LY96*, *ANKRD22*, *FPR3* and *FCGR2A*. These genes show the highest betweenness centrality and node degree among tumours, and appeared over-expressed in Basal I. In other reports, the *CXCR6* over-expression has been linked to TNBCs, with distinct roles in autoimmunity and cancer (Chaturvedi et al., 2014). The coexpression of *CXCR6* and *CXCL16*, a chemokine ligand and receptor, has been associated with inflammatory response and cell migration (Darash-Yahana et al., 2009; Xiao et al., 2015). In addition, high levels of *HCST* (Hyka-Nouspikel et al., 2007; Hyka-Nouspikel & Phillips, 2006), *C3AR1* (S. W. Wu et al., 2015), *GBP4* (Y. Hu et al., 2011), *LY96* (Deguchi et al., 2016), *ANKRD22* (Caba et al., 2014), *FPR3* (Prevete et al., 2015) and *FCGR2A* (Nimmerjahn & Ravetch, 2008), have also been related to immune activation and/or inflammatory response in tumours; however, their role in basal-like breast malignancies are yet to be uncovered. In our study, the increased expression levels of these probes, among others genes in the signature, has brought new insights on the basal-like tumour origin and progression, and Basal I and Basal II differentiation.

Standard clinical variables such as tumour size, histology and p53 status have also corroborated with the existence of the two basal-like subgroups. Basal I showed the highest frequency of medullary type, whereas Basal II exhibits the largest average of tumour size and highest frequency of p53 mutation. The interpretation of these features, in practice, support the better outcome of patients within Basal I subgroup, when compared to Basal II. Patients' age, post-menopausal status, tumour grade, NPI and lymph node invasion, on the other hand, are of a limited value for distinguishing the subgroups. Most of these variables reflect the overall tumour aggressiveness and the subtype poor prognosis.

7.4.2 MicroRNA Expression Levels Differentiating Basal I from Basal II

This work is the first instance of miRNA data coverage yielding the analysis of basal-like subgroups, which includes patients with matched genomic, transcriptomic and long-term survival data (Dvinge et al., 2013). The miRNAs have showed an important value for differentiating Basal I (15) and Basal II (4). In Basal I, hsa-miR-361-3p, -342-3p, -140-3p, -34a,

-22, -142-5p, -142-3p, -155, -342-5p, -150, -29c and -29a presented increased expression relative to Basal II. Overall, hsa-miR-361-3p has been found over-expressed in TNBCs with respect to other subtypes and healthy controls (Shin et al., 2015); and used to discriminate *BRCA1/2* mutation carriers and non-carriers tumours (Tanic et al., 2015). Greater levels of this miRNA, however, have been associated with a protective value in tumour progression (Roth et al., 2012) and further linked to inflammatory response (Guo et al., 2015). In line with our findings, these results contain additional information for the better understanding of basal-like subgroups. Additionally, high levels of hsa-miR-342-5p (Leivonen et al., 2013; Pérez-Rivas et al., 2014) and -34a (Hargraves et al., 2015; M. Y. Wu et al., 2014) have been correlated to breast cancer decreased recurrence and increased survival; whereas low levels have been associated with cell death inhibition and therapy resistance. The hsa-miR-22 (Chen et al., 2015; Kong et al., 2014) and members of the hsa-miR-29 family (-29a, -29b and -29c) (Nygren et al., 2014; Sandhu et al., 2014) – previously identified as tumour suppressors – have also been implicated in increased survival (Nygren et al., 2014) and pointed out as promising prognostic biomarkers (Chen et al., 2015; Kang et al., 2015).

In Basal II, hsa-miR-19b-1, -17 and -200c presented higher expression levels relative to Basal I and control samples. Tumour cells with enhanced expression of hsa-miR-19 (-19a and -19b-1) have been shown to trigger epithelial-mesenchymal transition (Li et al., 2015). Notably, members of the hsa-miR-200 family have been described as major regulators of this biological process. High levels of hsa-miR-200c and -200b have been observed in circulating tumour cells from patients with metastatic breast cancers (Le et al., 2014), indicating the prognostic significance of this biological marker (Erbes et al., 2015; Tuomarila et al., 2014). Consistent with these observations, our results demonstrated the recurrent over-expression of hsa-miR-19b-1 and -200c in Basal II, with the worst disease outcome among the two basal-like subgroups. Ultimately, high levels of hsa-miR-17 has been commonly detected in TNBCs (Chang et al., 2015), associated with cell migration in vitro and metastasis in vivo (Vimalraj et al., 2013).

The above described miRNAs matched 50 gene-targets from the 80-probe signature. In our study, hsa-miR-200c* and -29c have been associated with *HJURP* expression levels in G1, hsa-miR-19b-1* with *CXCR6* in G2, and hsa-miR-17 with *CTSK* in G3, which are among the most important genes in the signature. None of these associations, however, have been reported in the literature. On the other hand, studies have demonstrated hits on the gene regulation between hsa-miR-142-5p and *CD24* (Venkatesan et al., 2015), hsa-miR-29 and *DNMT3B* (Morita et al., 2013; Nguyen et al., 2011), hsa-miR-142-3p and *EGR2* (Lagrange et al., 2013), hsa-miR-150 and *EGR2* (Q. Wu et al., 2010), hsa-miR-34a and *IKZF3* (Rodriguez-Ubreva et al., 2014), hsa-miR-150 and *MIAT* (Zhu et al., 2016), hsa-miR-342-3p and *PSMG3* (Czimmerer et

al., 2016; Wang et al., 2016), hsa-miR-17 and *TIMP3* (Yang et al., 2013). Our results further suggested an important correlation between miRNAS and gene expression values in both Basal I and Basal II, identified by this in silico approach. These and other correlations are, however, highly complex and not fully understood. Additional analysis using in vitro and in vivo models are required to validate our achievements.

7.4.3 Genomic Aberrations Further Characterise Basal II and Basal I Subgroups

Basal-like and triple-negative tumours exhibit the highest frequencies of genomic gains and losses in comparison to other breast cancer subtypes (Engebraaten et al., 2013). Significant aberrations observed in this study confirmed the genomic instability among basal-like and further differentiated the two subgroups. The most common aberrations delineating Basal II, with respect to Basal I, occurred on the chromosomes 1, 3, 4, 5, 8, 10, 17 and X.

Gains in 1q, 3q, 8q, 10p and 17q have been identified in our analysis and previously reported in triple-negative tumours (Engebraaten et al., 2013; Loo et al., 2011). Overall, gains on chromosome 1q are the most frequent CNAs detected in breast carcinomas and are normally complex and discontinuous (Yu et al., 2009). Amplicons of 1q, 8p and 10p have been also described. These amplicons have contributed to the molecular understanding of this disease and, specially, of basal-like intrinsic subtype (Vincent-Salomon et al., 2007). For instance, amplifications in 8q21 have been associated with high tumour grade, high levels of Ki67 and other proliferation markers, including *MYC*, *MDM2* and *CCND1* (Choschzick et al., 2010). Gains in 10p have further differentiated triple-negative cancers (Loo et al., 2011), and in 17q25 have distinguished *BRCA1*-mutated tumours (Toffoli et al., 2014).

Losses in 4q, 5q, 8p, Xp and Xq have been defined as key aberrations within basal-like tumours in our analysis and among other breast cancer studies (Burstein et al., 2014; Weigman et al., 2012). Frequent losses in 4q and 5q in *BRCA1*-mutated tumours have distinguished them from sporadic neoplasms. In particular, the loss in 5q has impacted the expression of several *BRCA1*-dependent genes involved in DNA repair, such as *RAD17* and *RAD51* (Johannsdottir et al., 2006). High incidence rates of gains in 5q14 have also been associated with a poor prognosis in BLBCs (Thomassen et al., 2013). Other evidence suggests that aberrations on the X chromosome are common to both *BRCA1*-mutated and sporadic tumours (Richardson et al., 2006).

Overall, these aberrations yielded an additional characterisation of Basal I and Basal II. The increasing *PGA*, or genome instability, from one subgroup to the other complemented the 80-probe signature via the transcriptomic assessment, which is still considered more representative of cellular processes at the proteomic scale (Tyanova et al., 2016). Although the identified CNA did not show a direct correlation with the 80 probes' expression levels, generally it may lead to widespread disruptions beyond the proposed signature. Ultimately, the above described gains and losses in cytobands – supported by a range of distinct approaches in the literature – further corroborate the differentiation of basal-like subgroups with divergent clinical features and survival outcomes.

7.4.4 Consensus on the Analysis of Basal-like Breast Cancer Subtypes: a Literature Overview

In this section, we further established a consensus on the description of basal-like subgroups (Basal I and Basal II) by comparing our results with other achievements across the literature (Burstein et al., 2014; Jézéquel et al., 2015; Lehmann et al., 2011; Sabatier et al., 2011), as per the focus of each study. Notably, most of them have centred on the classification of triple-negative entities, a more heterogeneous group than basal-like. For instance, among the six intrinsic TNBC subtypes defined by Lehmann et al. (2011), three were considered relevant for further comparisons against the proposed basal-like subgroups: the basal-like (BL1 and BL2) and the immunomodulatory (IM). The groups were described based on cell cycle regulation, DNA damage response and immunomodulatory related-genes, respectively. These genes hint to the involvement of similar mechanisms differentiating between Basal I and Basal II, indicating that both classifications are somehow related. Genes (G1) with high node centrality values in Basal II, such as *HJURP* and *TPX2* have been linked to aberrant proliferation networks, cell invasion and metastasis in breast cancer, in line with the definition of BL1 (Lehmann et al., 2011). In addition, genes (G2) defining the Basal I subgroup, including *CXCR6*, *HCST*, *C3AR1*, *GBP4*, *LY96*, *ANKRD22*, *FPR3* and *FCGR2A*, have association with immune activation and inflammatory response, closer to IM (Lehmann et al., 2011). Major regulations involving these genes support the existence of the two subgroups, even though the pool of samples was considerably distinct, BLBCs and TNBCs.

In the recent classification of TNBCs performed by Burstein et al. (2014) (Burstein et al., 2014), two groups were described: the basal-like immune-activated (BLIA) and immune-

suppressed (BLIS) subtypes, corresponding to the best and worst prognosis, respectively. In BLIA, tumours display an over-expression of Stat signal transduction molecules and cytokines; in BLIS, high levels of the immunosuppressing molecule VTCN1. The mechanisms defining BLIA follow the characteristics of Basal I, and BLIS follows Basal II. For example, Basal I and BLIA (Burstein et al., 2014) contain common genes and/or genes belonging to the same family, such as *CXCL9/10/11/13*, *GBP4/5* and *CD2/24*. Similarly, Jézéquel et al. (2015) identified two relevant subtypes: basal-like with low immune response and high M2-like macrophages (C2), and basal-enriched with high immune response and low M2-like macrophages (C3). The defined basal-like and basal-enriched groups shared evident similarities with Basal II and Basal I, respectively, and corroborated with our study in terms of probe signature and functionality. With regards to the TNBC classification, however, Lehmann et al. (2011), Burstein et al. (2014) and Jézéquel et al. (2015) partially support each other.

An alternative approach to differentiating two subgroups of basal-like – associated with either a low or high risk of disease relapse – has been tested by Hallett et al. (2012), using a 14-gene signature. Among the genes in the signature, *RPL3* and *GPR27* were listed as key markers of relapse, while *RPL36AL* and *GPR65* appeared as variants in the 80 survival-related probes. In the same direction, Sabatier et al. (2011) identified a 28-kinase metagene signature – associated with disease-free survival and immune response – used to divide the BLBCs into two groups: 'Immune High' and 'Immune Low'. This approach revealed key genes, including *IL2RG/B*, *GBP2*, *CCR5/7*, *CXCR3/5/6* and *CXCL9/13*, related to their family members in our signature, such as *IL2RA*, *GBP4*, *CCR1*, *CXCR6* and *CXCL11*. These genes appeared over-expressed in 'Immune High' and in Basal I subgroup, when compared to 'Immune Low' and Basal II (Sabatier et al., 2011).

Integrating these observations, there is a clear consensus on the segregation of basal-like breast cancers into at least two subgroups. Basal I (Immune Active) show molecular overlaps and phenotypic similarities with BLIA (Burstein et al., 2014), IM (Lehmann et al., 2011) and C3 (Jézéquel et al., 2015); Basal II (High Proliferative) matched with BLIS (Burstein et al., 2014) and C2 (Jézéquel et al., 2015). The comprehensive genomic and transcriptomic characterisation of the two subgroups, provided in this study, will lead to the better understanding of the mechanisms involved in basal-like tumours and to the identification of groups of patients with distinct disease outcome, supported by additional survival features (Hallett et al., 2012; Sabatier et al., 2011). The latter is crucial for improving the clinical decision-making and for helping tailor treatments that are focused on the immune system manipulation and the cell cycle pathway intervention. In general, tumours with activated immune response have shown a favourable prognosis (Bertucci et al., 2012) and are likely to

respond to chemotherapy (Sabatier et al., 2011), whereas the high proliferative ones have revealed increased risk of metastasis and recurrence (Fadare & Tavassoli, 2008). In this context, patients at a low risk should follow more conservative therapies and those at high risk should receive more effective drugs for improving individual response, towards a more personalised medicine.

7.5 Conclusion

Studies have demonstrated that the heterogeneity of BLBCs extends beyond the classic immunohistochemistry. Although several clinicopathological features have been used to discriminate between low- and high-risk patients, the identification of novel biomarkers with prognostic value remains an urgent need for improving breast cancer management. The 80-probe signature defined in this study, associated with varying survival outcomes, contains putative markers of disease progression and represents a promising asset for clinical applications. The integrated assessment of miRNA expression and CNA information, ultimately, contributes towards the definition of more comprehensive profiles of basal-like tumours. The importance of defining groups-at-risk of BLBCs is reflected in the impact of survival-related features in clinical settings and, more importantly, in therapy response.

7.6 References

Andre, F., Dieci, M. V., Dubsky, P., Sotiriou, C., Curigliano, G., Denkert, C., et al. (2013). Molecular pathways: involvement of immune pathways in the therapeutic response and outcome in breast cancer. *Clin. Cancer Res., 19*(1), 28-33.

Badve, S., Dabbs, D. J., Schnitt, S. J., Baehner, F. L., Decker, T., Eusebi, V., et al. (2011). Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. *Mod. Pathol., 24*(2), 157-167.

Banerjee, S., Reis-Filho, J. S., Ashley, S., Steele, D., Ashworth, A., Lakhani, S. R., et al. (2006). Basal-like breast carcinomas: clinical outcome and response to chemotherapy. *J. Clin. Pathol., 59*(7), 729-735.

Berretta, R., & Moscato, P. (2010). Cancer Biomarker Discovery: The Entropic Hallmark. *PLoS One, 5*(8), e12262.

Bertucci, F., Finetti, P., & Birnbaum, D. (2012). Basal breast cancer: a complex and deadly molecular subtype. *Curr. Mol. Med., 12*(1), 96.

Burstein, M. D., Tsimelzon, A., Poage, G. M., Covington, K. R., Contreras, A., Fuqua, S. A., et al. (2014). Comprehensive Genomic Analysis Identifies Novel Subtypes and Targets of Triple-negative Breast Cancer. *Clin. Cancer Res., 21*(7), 1688-1698.

Buyse, M., Loi, S., Van't Veer, L., Viale, G., Delorenzi, M., Glas, A. M., et al. (2006). Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J. Natl. Cancer Inst., 98*(17), 1183-1192.

Caba, O., Prados, J., Ortiz, R., Jiménez-Luna, C., Melguizo, C., Álvarez, P. J., et al. (2014). Transcriptional profiling of peripheral blood in pancreatic adenocarcinoma patients identifies diagnostic biomarkers. *Dig. Dis. Sci., 59*(11), 2714-2720.

Carey, L. A., Winer, E. P., Viale, G., Cameron, D., & Gianni, L. (2010). Triple-negative breast cancer: disease entity or title of convenience? *Nat. Rev. Clin. Oncol., 7*(12), 683-692.

Carlson, M. (2016a). hgu133a.db: Affymetrix Human Genome U133 Set annotation data (chip hgu133a). R package version 2.14.0.

Carlson, M. (2016b). hgug4112a.db: Agilent Human Genome, Whole annotation data (chip hgug4112a). R package version 3.1.3.

Chang, Y.-Y., Kuo, W.-H., Hung, J.-H., Lee, C.-Y., Lee, Y.-H., Chang, Y.-C., et al. (2015). Deregulated microRNAs in triple-negative breast cancer revealed by deep sequencing. *Mol. Cancer, 14*(1), 36.

Chaturvedi, P., Gilkes, D. M., Takano, N., & Semenza, G. L. (2014). Hypoxia-inducible factor-dependent signaling between triple-negative breast cancer cells and mesenchymal stem cells promotes macrophage recruitment. *Proc. Natl. Acad. Sci. U. S. A., 111*(20), E2120-E2129.

Cheang, M. C. U., Voduc, D., Bajdik, C., Leung, S., McKinney, S., Chia, S. K., et al. (2008). Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clin. Cancer Res., 14*(5), 1368-1376.

Chen, B., Tang, H., Liu, X., Liu, P., Yang, L., Xie, X., et al. (2015). miR-22 as a prognostic factor targets glucose transporter protein type 1 in breast cancer. *Cancer Lett., 356*(2), 410-417.

Choschzick, M., Lassen, P., Lebeau, A., Marx, A. H., Terracciano, L., Heilenkotter, U., et al. (2010). Amplification of 8q21 in breast cancer is independent of MYC and associated with poor patient outcome. *Mod. Pathol., 23*(4), 603-610.

Cleator, S., Heller, W., & Coombes, R. C. (2007). Triple-negative breast cancer: therapeutic options. *Lancet Oncol., 8*(3), 235-244.

Cormen, T. H. (2009). *Introduction to algorithms*: MIT press.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *Int J Complex Systems, 1695*(5), 1-9.

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature, 486*(7403), 346-352.

Czimmerer, Z., Varga, T., Kiss, M., Vázquez, C. O., Doan-Xuan, Q. M., Rückerl, D., et al. (2016). The IL-4/STAT6 signaling axis establishes a conserved microRNA signature in human and mouse macrophages regulating cell survival via miR-342-3p. *Genome Med., 8*(1), 63.

Darash-Yahana, M., Gillespie, J. W., Hewitt, S. M., Chen, Y.-Y. K., Maeda, S., Stein, I., et al. (2009). The Chemokine CXCL16 and Its Receptor, CXCR6, as Markers and Promoters of Inflammation-Associated Cancers. *PLoS One, 4*(8), e6695.

Deguchi, A., Tomita, T., Ohto, U., Takemura, K., Kitao, A., Akashi-Takamura, S., et al. (2016). Eritoran inhibits S100A8-mediated TLR4/MD-2 activation and tumor growth by changing the immune microenvironment. *Oncogene, 35*(11), 1445-1456.

Dunning, M., Lynch, A., & Eldridge, M. (2015). illuminaHumanv3.db: Illumina HumanHT12v3 annotation data (chip illuminaHumanv3). R package version 2.0.

Dvinge, H., Git, A., Gräf, S., Salmon-Divon, M., Curtis, C., Sottoriva, A., et al. (2013). The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature, 497*(7449), 378-382.

Engebraaten, O., Vollan, H. K. M., & Børresen-Dale, A.-L. (2013). Triple-negative breast cancer and the need for new therapeutic targets. *Am. J. Pathol., 183*(4), 1064-1074.

Erbes, T., Hirschfeld, M., Rücker, G., Jaeger, M., Boas, J., Iborra, S., et al. (2015). Feasibility of urinary microRNA detection in breast cancer patients and its potential as an innovative non-invasive biomarker. *BMC Cancer, 15*(1), 1.

Fadare, O., & Tavassoli, F. A. (2008). Clinical and pathologic aspects of basal-like breast cancers. *Nat. Clin. Pract. Oncol., 5*(3), 149-159.

Favero, F. (2013). RmiR.Hs.miRNA: Various databases of microRNA Targets. R package version 1.0.7.

Geiger, T. R., Ha, N.-H., Faraji, F., Michael, H. T., Rodriguez, L., Walker, R. C., et al. (2014). Functional analysis of prognostic gene expression network genes in metastatic breast cancer models. *PLoS One, 9*(11), e111813.

Gentleman, R. (2003). annotate: Annotation for microarrays. R package version 1.54.0.

Glas, A. M., Floore, A., Delahaye, L. J., Witteveen, A. T., Pover, R. C., Bakx, N., et al. (2006). Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics, 7*(1), 278.

Grosse, I., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J., & Stanley, H. E. (2002). Analysis of symbolic sequences using the Jensen-Shannon divergence. *Physical Review E, 65*(4), 041905.

Guo, Z., Wu, R. C., Gong, J., Zhu, W., Li, Y., Wang, Z., et al. (2015). Altered microRNA expression in inflamed and non-inflamed terminal ileal mucosa of adult patients with active Crohn's disease. *J. Gastroenterol. Hepatol., 30*(1), 109-116.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., et al. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *J. Natl. Cancer Inst., 104*(4), 311-325.

Hallett, R. M., Dvorkin-Gheva, A., Bane, A., & Hassell, J. A. (2012). A gene signature for predicting outcome in patients with basal-like breast cancer. *Proceedings: AACR 103rd Annual Meeting*, 3663-3663.

Hargraves, K. G., He, L., & Firestone, G. L. (2015). Phytochemical regulation of the tumor suppressive microRNA, miR-34a, by p53-dependent and independent responses in human breast cancer cells. *Mol. Carcinog.*

Hu, Y., Wang, J., Yang, B., Zheng, N., Qin, M., Ji, Y., et al. (2011). Guanylate Binding Protein 4 Negatively Regulates Virus-Induced Type I IFN and Antiviral Response by Targeting IFN Regulatory Factor 7. *The Journal of Immunology, 187*(12), 6456-6462.

Hu, Z., Huang, G., Sadanandam, A., Gu, S., Lenburg, M. E., Pai, M., et al. (2010). The expression level of HJURP has an independent prognostic impact and predicts the sensitivity to radiotherapy in breast cancer. *Breast Cancer Res., 12*(2), R18.

Hyka-Nouspikel, N., Lucian, L., Murphy, E., McClanahan, T., & Phillips, J. H. (2007). DAP10 Deficiency Breaks the Immune Tolerance against Transplantable Syngeneic Melanoma. *The Journal of Immunology, 179*(6), 3763-3771.

Hyka-Nouspikel, N., & Phillips, J. H. (2006). Physiological roles of murine DAP10 adapter protein in tumor immunity and autoimmunity. *Immunol. Rev., 214*(1), 106-117.

Jézéquel, P., Loussouarn, D., Guérin-Charbonnel, C., Campion, L., Vanier, A., Gouraud, W., et al. (2015). Gene-expression molecular subtyping of triple-negative breast cancer tumours: importance of immune response. *Breast Cancer Res., 17*(1), 43.

Johannsdottir, H. K., Jonsson, G., Johannesdottir, G., Agnarsson, B. A., Eerola, H., Arason, A., et al. (2006). Chromosome 5 imbalance mapping in breast tumors from BRCA1 and BRCA2 mutation carriers and sporadic breast tumors. *Int. J. Cancer, 119*(5), 1052-1060.

Kang, L., Mao, J., Tao, Y., Song, B., Ma, W., Lu, Y., et al. (2015). MicroRNA-34a suppresses the breast cancer stem cell-like characteristics by downregulating Notch1 pathway. *Cancer Sci., 106*(6), 700-708.

Kao, J., Salari, K., Bocanegra, M., Choi, Y.-L., Girard, L., Gandhi, J., et al. (2009). Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLoS One, 4*(7), e6146.

Kong, L.-M., Liao, C.-G., Zhang, Y., Xu, J., Li, Y., Huang, W., et al. (2014). A regulatory loop involving miR-22, Sp1, and c-Myc modulates CD147 expression in breast cancer invasion and metastasis. *Cancer Res., 74*(14), 3764-3778.

Kreike, B., van Kouwenhove, M., Horlings, H., Weigelt, B., Peterse, H., Bartelink, H., et al. (2007). Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas. *Breast Cancer Res., 9*(5), R65.

Lagrange, B., Martin, R. Z., Droin, N., Aucagne, R., Paggetti, J., Largeot, A., et al. (2013). A role for miR-142-3p in colony-stimulating factor 1-induced monocyte differentiation into macrophages. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research, 1833*(8), 1936-1946.

Le, M. T., Hamar, P., Guo, C., Basar, E., Perdigão-Henriques, R., Balaj, L., et al. (2014). miR-200 - containing extracellular vesicles promote breast cancer cell metastasis. *J. Clin. Invest., 124*(12), 5109.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., et al. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest., 121*(7), 2750-2767.

Leivonen, S.-K., Sahlberg, K. K., Makela, R., Kallioniemi, O., Borresen-Dale, A.-L., & Perala, M. (2013). High-throughput screens identify microRNAs essential for HER2-positive breast cancer cell growth. *Cancer Res., 73*(8 Supplement), 1956-1956.

Li, J., Yang, S., Yan, W., Yang, J., Qin, Y.-J., Lin, X.-L., et al. (2015). MicroRNA-19 triggers epithelial - mesenchymal transition of lung cancer cells accompanied by growth inhibition. *Lab. Invest., 95*(9), 1056-1070.

Liu, Z., Zhang, X.-S., & Zhang, S. (2014). Breast tumor subgroups reveal diverse clinical prognostic power. *Sci. Rep., 4*, 4002.

Loo, L. W., Wang, Y., Flynn, E. M., Lund, M. J., Bowles, E. J. A., Buist, D. S., et al. (2011). Genome-wide copy number alterations in subtypes of invasive breast cancers in young white and African American women. *Breast Cancer Res. Treat., 127*(1), 297-308.

Lund, M. J., Trivers, K. F., Porter, P. L., Coates, R. J., Leyland-Jones, B., Brawley, O. W., et al. (2009). Race and triple negative threats to breast cancer survival: a population-based study in Atlanta, GA. *Breast Cancer Res. Treat., 113*(2), 357-370.

Merkin, J., Russell, C., Chen, P., & Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science, 338*(6114), 1593-1599.

Milioli, H. H., Vimieiro, R., Riveros, C., Tishchenko, I., Berretta, R., & Moscato, P. (2015). The Discovery of Novel Biomarkers Improves Breast Cancer Intrinsic Subtype Prediction and Reconciles the Labels in the METABRIC Data Set. *PLoS One, 10*(7), e0129711.

Millikan, R. C., Newman, B., Tse, C. K., Moorman, P. G., Conway, K., Dressler, L. G., et al. (2008). Epidemiology of basal-like breast cancer. *Breast Cancer Res. Treat., 109*(1), 123-139.

Morita, S., Horii, T., Kimura, M., Ochiya, T., Tajima, S., & Hatada, I. (2013). miR-29 represses the activities of DNA methyltransferases and DNA demethylases. *Int. J. Mol. Sci., 14*(7), 14647-14658.

Mulligan, A. M., Pinnaduwage, D., Bull, S. B., O'Malley, F. P., & Andrulis, I. L. (2008). Prognostic effect of basal-like breast cancers is time dependent: evidence from tissue microarray studies on a lymph node-negative cohort. *Clin. Cancer Res., 14*(13), 4168-4174.

Murtagh, F., & Legendre, P. (2013). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification, 31*(3), 274-295.

Nguyen, T., Kuo, C., Nicholl, M. B., Sim, M.-S., Turner, R. R., Morton, D. L., et al. (2011). Downregulation of microRNA-29c is associated with hypermethylation of tumor-related genes and disease outcome in cutaneous melanoma. *Epigenetics, 6*(3), 388-394.

Nielsen, T. O., Hsu, F. D., Jensen, K., Cheang, M. C. U., Karaca, G., Hu, Z., et al. (2004). Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin. Cancer Res., 10*(16), 5367-5374.

Nimmerjahn, F., & Ravetch, J. V. (2008). Fcγ receptors as regulators of immune responses. *Nature Reviews Immunology, 8*(1), 34-47.

Nygren, M., Tekle, C., Ingebrigtsen, V., Mäkelä, R, Krohn, M., Aure, M., et al. (2014). Identifying microRNAs regulating B7-H3 in breast cancer: the clinical impact of microRNA-29c. *Br. J. Cancer, 110*(8), 2072-2080.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., et al. (2004). A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *N. Engl. J. Med., 351*(27), 2817-2826.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol., 27*(8), 1160-1167.

Pérez-Rivas, L. G., Jerez, J., M, Carmona, R., de Luque, V., Vicioso, L., Claros, M. G., et al. (2014). A microRNA signature associated with early recurrence in breast cancer. *PLoS One, 9*(3), e91884.

Prat, A., Adamo, B., Cheang, M. C. U., Anders, C. K., Carey, L. A., & Perou, C. M. (2013). Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. *Oncologist, 18*(2), 123-133.

Prevete, N., Liotti, F., Visciano, C., Marone, G., Melillo, R. M., & de Paulis, A. (2015). The formyl peptide receptor 1 exerts a tumor suppressor function in human gastric cancer by inhibiting angiogenesis. *Oncogene, 34*(29), 3826-3838.

Putti, T. C., El-Rehim, D. M. A., Rakha, E. A., Paish, C. E., Lee, A. H., Pinder, S. E., et al. (2005). Estrogen receptor-negative breast carcinomas: a review of morphology and immunophenotypical analysis. *Mod. Pathol., 18*(1), 26-35.

Rakha, E. A., Reis-Filho, J. S., & Ellis, I. O. (2008). Impact of basal-like breast carcinoma determination for a more specific therapy. *Pathobiology, 75*(2), 95-103.

Richardson, A. L., Wang, Z. C., De Nicolo, A., Lu, X., Brown, M., Miron, A., et al. (2006). X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell, 9*(2), 121-132.

Rodriguez-Ubreva, J., van Oevelen, C., Parra, M., Graf, T., & Ballestar, E. (2014). C/EBPa-mediated activation of MicroRNAs 34a and 223 inhibits lef1 expression to achieve efficient reprogramming into macrophages. *Mol. Cell. Biol., 34*(6), 1145-1157.

Rody, A., Karn, T., Liedtke, C., Pusztai, L., Ruckhaeberle, E., Hanker, L., et al. (2011). A clinically relevant gene signature in triple negative and basal-like breast cancer. *Breast Cancer Res., 13*(5), R97.

Roll, J. D., Rivenbark, A. G., Sandhu, R., Parker, J. S., Jones, W. D., Carey, L. A., et al. (2013). Dysregulation of the epigenome in triple-negative breast cancers: basal-like and claudin-low breast cancers express aberrant DNA hypermethylation. *Exp. Mol. Pathol., 95*(3), 276-287.

Roth, C., Stückrath, I., Pantel, K., Izbicki, J. R., Tachezy, M., & Schwarzenbach, H. (2012). Low levels of cell-free circulating miR-361-3p and miR-625* as blood-based markers for discriminating malignant from benign lung tumors. *PLoS One, 7*(6), e38248.

Sabatier, R., Finetti, P., Mamessier, E., Raynaud, S., Cervera, N., Lambaudie, E., et al. (2011). Kinome expression profiling and prognosis of basal breast cancers. *Mol. Cancer, 10*(86), 24.

Sandhu, R., Rivenbark, A. G., Mackler, R. M., Livasy, C. A., & Coleman, W. B. (2014). Dysregulation of microRNA expression drives aberrant DNA hypermethylation in basal-like breast cancer. *Int. J. Oncol., 44*(2), 563-572.

Shin, V., Siu, J., Cheuk, I., Ng, E., & Kwong, A. (2015). Circulating cell-free miRNAs as biomarker for triple-negative breast cancer. *Br. J. Cancer, 112*(11), 1751-1759.

Sims, D., Bursteinas, B., Gao, Q., Jain, E., MacKay, A., Mitsopoulos, C., et al. (2010). ROCK: a breast cancer functional genomics resource. *Breast Cancer Res. Treat., 124*(2), 567-572.

Tanic, M., Yanowski, K., Gómez-López, G., Rodriguez-Pinilla, M. S., Marquez-Rodas, I., Osorio, A., et al. (2015). MicroRNA expression signatures for the prediction of BRCA1/2 mutation-associated hereditary breast cancer in paraffin-embedded formalin-fixed breast tumors. *Int. J. Cancer, 136*(3), 593-602.

Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O., & Caldas, C. (2007). An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biol., 8*(8), R157.

Therneau, T. M. (2015). A Package for Survival Analysis in S. R package version 2.38.

Thomassen, M., Tan, Q., Burton, M., & Kruse, T. A. (2013). Gene Expression Meta-Analysis Identifies Cytokine Pathways and 5q Aberrations Involved in Metastasis of ERBB2 Amplified and Basal Breast Cancer. *Cancer Inform., 12*, 203-219.

Tishchenko, I., Milioli, H. H., Riveros, C., & Moscato, P. (2016). Extensive Transcriptomic and Genomic Analysis Provides New Insights about Luminal Breast Cancers. *PLoS One, 11*(6), e0158259.

Toffoli, S., Bar, I., Abdel-Sater, F., Delrée, P., Hilbert, P., Cavallin, F., et al. (2014). Identification by array comparative genomic hybridization of a new amplicon on chromosome 17q highly recurrent in BRCA1 mutated triple negative breast cancer. *Breast Cancer Res., 16*(6), 466.

Tuomarila, M., Luostari, K., Soini, Y., Kataja, V., Kosma, V.-M., & Mannermaa, A. (2014). Overexpression of MicroRNA-200c Predicts Poor Outcome in Patients with PR-Negative Breast Cancer. *PLoS One, 9*(10), e109508.

Tyanova, S., Albrechtsen, R., Kronqvist, P., Cox, J., Mann, M., & Geiger, T. (2016). Proteomic maps of breast cancer subtypes. *Nature communications, 7*.

Ur-Rehman, S., Gao, Q., Mitsopoulos, C., & Zvelebil, M. (2013). ROCK: a resource for integrative breast cancer data analysis. *Breast Cancer Res. Treat., 139*(3), 907-921.

Valentin, M. D., da Silva, S. D., Privat, M., Alaoui-Jamali, M., & Bignon, Y.-J. (2012). Molecular insights on basal-like breast cancer. *Breast Cancer Res. Treat., 134*(1), 21-30.

Venkatesan, N., Deepa, P. R., Khetan, V., & Krishnakumar, S. (2015). Computational and in vitro investigation of miRNA-gene regulations in retinoblastoma pathogenesis: miRNA mimics strategy. *Bioinform. Biol. Insights, 9*, 89.

Vimalraj, S., Miranda, P., Ramyakrishna, B., & Selvamurugan, N. (2013). Regulation of breast cancer and bone metastasis by microRNAs. *Dis. Markers, 35*(5), 369-387.

Vincent-Salomon, A., Gruel, N., Lucchesi, C., MacGrogan, G., Dendale, R., Sigal-Zafrani, B., et al. (2007). Identification of typical medullary breast carcinoma as a genomic sub-group of basal-like carcinomas, a heterogeneous new molecular entity. *Breast Cancer Res., 9*(2), R24.

Wang, S.-H., Ma, F., Tang, Z.-h., Wu, X.-C., Cai, Q., Zhang, M.-D., et al. (2016). Long non-coding RNA H19 regulates FOXM1 expression by competitively binding endogenous miR-342-3p in gallbladder cancer. *J. Exp. Clin. Cancer Res., 35*(1), 160.

Weigman, V. J., Chao, H.-H., Shabalin, A. A., He, X., Parker, J. S., Nordgard, S. H., et al. (2012). Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. *Breast Cancer Res. Treat., 133*(3), 865-880.

Wu, M. Y., Fu, J., Xiao, X., Wu, J., & Wu, R. C. (2014). MiR-34a regulates therapy resistance by targeting HDAC1 and HDAC7 in breast cancer. *Cancer Lett., 354*(2), 311-319.

Wu, Q., Jin, H., Yang, Z., Luo, G., Lu, Y., Li, K., et al. (2010). MiR-150 promotes gastric cancer proliferation by negatively regulating the pro-apoptotic gene EGR2. *Biochem. Biophys. Res. Commun., 392*(3), 340-345.

Wu, S. W., Fan, J., Hong, D., Zhou, Q., Zheng, D., Wu, D., et al. (2015). C3aR1 gene overexpressed at initial stage of acute myeloid leukemia-M2 predicting short-term survival. *Leuk. Lymphoma, 56*(7), 2200-2202.

Xiao, G., Wang, X., Wang, J., Zu, L., Cheng, G., Hao, M., et al. (2015). CXCL16/CXCR6 chemokine signaling mediates breast cancer progression by pERK1/2-dependent mechanisms. *Oncotarget, 6*(16), 14165-14178.

Yang, X., Du, W. W., Li, H., Liu, F., Khorshidi, A., Rutnam, Z. J., et al. (2013). Both mature miR-17-5p and passenger strand miR-17-3p target TIMP3 and induce prostate tumor growth and invasion. *Nucleic Acids Res., 41*(21), 9688-9704.

Yau, C., Esserman, L., Moore, D. H., Waldman, F., Sninsky, J., & Benz, C. C. (2010). A multigene predictor of metastatic outcome in early stage hormone receptor-negative and triple-negative breast cancer. *Breast Cancer Res., 12*(5), R85.

Yau, C., Sninsky, J., Kwok, S., Wang, A., Degnim, A., Ingle, J. N., et al. (2013). An optimized five-gene multi-platform predictor of hormone receptor negative and triple negative breast cancer metastatic risk. *Breast Cancer Res., 15*(5), R103.

Yu, W., Kanaan, Y., Baed, Y.-K., & Gabrielson, E. (2009). Chromosomal changes in aggressive breast cancers with basal-like features. *Cancer Genet. Cytogenet., 193*(1), 29-37.

Zhu, X., Yuan, Y., Rao, S., & Wang, P. (2016). LncRNA MIAT enhances cardiac hypertrophy partly through sponging miR-150. *Eur. Rev. Med. Pharmacol. Sci., 20*(17), 3653-3660.

7.7 Supporting Information

Supporting Information – **Table 7.6, Table 7.7 and Table 7.8**

Basal-like samples classification into Basal I and Basal II, and the centroids defining them. Tables 7.6 and 7.7 list the sample IDs for each basal-like subgroup, Basal I and Basal II; centroids are provided in Table 7.8. *Available online: doi:10.1186/s12920-017-0250-9*

Table 7.6 Basal-like samples classification for the validation set

Table 7.7 Basal-like samples classification for the validation set

Table 7.8 The centroids computed for differentiating Basal I and Basal II

Supporting Information – **Table 7.9, Table 7.10 and Table 7.11**

The annotation is based on the Database for Annotation, Visualization and Integrated Discovery (DAVID). *Available online: doi:10.1186/s12920-017-0250-9*

Table 7.9 The functional annotation of G1 probes according to DAVID

Table 7.10 The functional annotation of G2 probes according to DAVID

Table 7.11 The functional annotation of G3 probes according to DAVID

Supporting Information – **Table 7.12**

This table contains the *p*-values computed for the difference in expression levels between basal-like subgroups and control samples.

Table 7.12 MicroRNAs differentiating Basal I and Basal II

MicroRNA	Training Set	Validation Set
CRINCR2000005427	2.80E-04	1.10E-03
A_25_P00012305	2.80E-04	6.80E-03
A_25_P00012357	4.00E-04	2.70E-03
A_25_P00012177	1.30E-04	6.30E-03
A_25_P00012086	4.90E-03	1.60E-04
A_25_P00010204	6.30E-03	9.20E-03
A_25_P00014844	2.00E-04	4.10E-06
A_25_P00011016	2.20E-03	5.50E-05
A_25_P00012271	6.30E-06	4.70E-04
A_25_P00012354	2.00E-07	1.50E-05
A_25_P00014847	8.70E-06	2.90E-04
CRINCR2000005084	1.20E-04	4.40E-03
A_25_P00012274	6.70E-03	2.00E-03
A_25_P00013484	5.60E-04	8.00E-03
A_25_P00012013	4.80E-03	9.50E-03
A_25_P00013163	5.30E-04	1.90E-03
A_25_P00013151	5.00E-04	2.00E-03
A_25_P00013841	1.90E-03	4.20E-04
A_25_P00013469	1.80E-04	9.50E-03

Supporting Information – Table 7.13

This table contains miRNAs and gene targets with the respective expression levels in Basal I.

Table 7.13 MicroRNAs and gene targets in Basal I

Gene ID	Mature miRNA	miRNA Exprression	Gene Symbol	Gene Expression
56652	hsa-miR-140-3p	10.25586	C10orf2	6.337445
6373	hsa-miR-140-3p	10.25586	CXCL11	7.046886
79047	hsa-miR-140-3p	10.25586	KCTD15	6.504304
8228	hsa-miR-140-3p	10.25586	PNPLA4	5.694018
5589	hsa-miR-140-3p	10.25586	PRKCSH	8.599771
23223	hsa-miR-140-3p	10.25586	RRP12	7.704964
10494	hsa-miR-140-3p	10.25586	STK25	8.297879
1E+08	hsa-miR-142-3p	13.66798	CD24	12.73802
1959	hsa-miR-142-3p	13.66798	EGR2	8.11905
8228	hsa-miR-142-3p	13.66798	PNPLA4	5.694018
6451	hsa-miR-142-3p	13.66798	SH3BGRL	9.514529
56652	hsa-miR-142-5p	9.940696	C10orf2	6.337445
1E+08	hsa-miR-142-5p	9.940696	CD24	12.73802
1050	hsa-miR-142-5p	9.940696	CEBPA	8.541949
1959	hsa-miR-142-5p	9.940696	EGR2	8.11905
26234	hsa-miR-142-5p	9.940696	FBXL5	6.106371
2359	hsa-miR-142-5p	9.940696	FPR3	8.746832
3290	hsa-miR-142-5p	9.940696	HSD11B1	6.887209
83593	hsa-miR-142-5p	9.940696	RASSF5	8.677596
7078	hsa-miR-142-5p	9.940696	TIMP3	8.607555
1230	hsa-miR-150	12.74988	CCR1	6.603287
1959	hsa-miR-150	12.74988	EGR2	8.11905
26234	hsa-miR-150	12.74988	FBXL5	6.106371
440823	hsa-miR-150	12.74988	MIAT	6.040073
84722	hsa-miR-155	10.56577	PSRC1	5.536997
79863	hsa-miR-155	10.56577	RBFA	6.772597
9447	hsa-miR-17	9.7715	AIM2	7.817965
57673	hsa-miR-17	9.7715	BEND3	6.275244
1056	hsa-miR-17	9.7715	CEL	6.260176
1513	hsa-miR-17	9.7715	CTSK	10.41549
1959	hsa-miR-17	9.7715	EGR2	8.11905

26234	hsa-miR-17	9.7715	FBXL5	6.106371
8228	hsa-miR-17	9.7715	PNPLA4	5.694018
149628	hsa-miR-17	9.7715	PYHIN1	6.295934
6641	hsa-miR-17	9.7715	SNTB1	7.336399
7078	hsa-miR-17	9.7715	TIMP3	8.607555
10663	hsa-miR-19b-1*	7.189607	CXCR6	6.882271
2212	hsa-miR-19b-1*	7.189607	FCGR2A	8.539242
3290	hsa-miR-19b-1*	7.189607	HSD11B1	6.887209
83463	hsa-miR-19b-1*	7.189607	MXD3	6.601417
9046	hsa-miR-200c*	6.543946	DOK2	6.324262
55355	hsa-miR-200c*	6.543946	HJURP	7.261004
3559	hsa-miR-200c*	6.543946	IL2RA	6.139334
84722	hsa-miR-200c*	6.543946	PSRC1	5.536997
23223	hsa-miR-200c*	6.543946	RRP12	7.704964
9046	hsa-miR-22	14.65232	DOK2	6.324262
2760	hsa-miR-22	14.65232	GM2A	7.595457
3290	hsa-miR-22	14.65232	HSD11B1	6.887209
83463	hsa-miR-22	14.65232	MXD3	6.601417
8228	hsa-miR-22	14.65232	PNPLA4	5.694018
10494	hsa-miR-22	14.65232	STK25	8.297879
7078	hsa-miR-22	14.65232	TIMP3	8.607555
64581	hsa-miR-29a	14.15218	CLEC7A	6.128213
1789	hsa-miR-29a	14.15218	DNMT3B	5.943194
2212	hsa-miR-29a	14.15218	FCGR2A	8.539242
2326	hsa-miR-29a	14.15218	FMO1	7.385919
79047	hsa-miR-29a	14.15218	KCTD15	6.504304
440823	hsa-miR-29a	14.15218	MIAT	6.040073
22974	hsa-miR-29a	14.15218	TPX2	7.51496
64581	hsa-miR-29c	12.26873	CLEC7A	6.128213
1789	hsa-miR-29c	12.26873	DNMT3B	5.943194
2212	hsa-miR-29c	12.26873	FCGR2A	8.539242
2326	hsa-miR-29c	12.26873	FMO1	7.385919
79047	hsa-miR-29c	12.26873	KCTD15	6.504304
440823	hsa-miR-29c	12.26873	MIAT	6.040073
22974	hsa-miR-29c	12.26873	TPX2	7.51496
84253	hsa-miR-29c*	7.856357	GARNL3	5.846762
55355	hsa-miR-29c*	7.856357	HJURP	7.261004
54069	hsa-miR-29c*	7.856357	MIS18A	6.656334

83463	hsa-miR-342-3p	10.24132	MXD3	6.601417
84262	hsa-miR-342-3p	10.24132	PSMG3	7.054041
171558	hsa-miR-342-3p	10.24132	PTCRA	5.784044
5788	hsa-miR-342-3p	10.24132	PTPRC	6.505452
7078	hsa-miR-342-3p	10.24132	TIMP3	8.607555
79058	hsa-miR-342-5p	8.809464	ASPSCR1	9.870704
837	hsa-miR-342-5p	8.809464	CASP4	9.267788
10320	hsa-miR-342-5p	8.809464	IKZF1	7.596667
84722	hsa-miR-342-5p	8.809464	PSRC1	5.536997
6373	hsa-miR-34a	13.13989	CXCL11	7.046886
79980	hsa-miR-34a	13.13989	DSN1	6.882759
2212	hsa-miR-34a	13.13989	FCGR2A	8.539242
26191	hsa-miR-34a	13.13989	GPR65	6.040714
22806	hsa-miR-34a	13.13989	IKZF3	5.625264
8228	hsa-miR-34a	13.13989	PNPLA4	5.694018
719	hsa-miR-361-3p	8.76375	C3AR1	7.144413
1050	hsa-miR-361-3p	8.76375	CEBPA	8.541949
2760	hsa-miR-361-3p	8.76375	GM2A	7.595457
440823	hsa-miR-361-3p	8.76375	MIAT	6.040073
6838	hsa-miR-361-3p	8.76375	SURF6	8.59528
7078	hsa-miR-361-3p	8.76375	TIMP3	8.607555

Supporting Information – Table 7.14

This table contains miRNAs and gene targets with the respective expression levels in Basal II.

Table 7.14 MicroRNAs and gene targets in Basal II

Gene ID	Mature miRNA	miRNA Exprression	Gene Symbol	Gene Expression
56652	hsa-miR-140-3p	9.640101	C10orf2	6.677458
6373	hsa-miR-140-3p	9.640101	CXCL11	6.155282
79047	hsa-miR-140-3p	9.640101	KCTD15	6.723827
8228	hsa-miR-140-3p	9.640101	PNPLA4	5.594859
5589	hsa-miR-140-3p	9.640101	PRKCSH	8.906658
23223	hsa-miR-140-3p	9.640101	RRP12	8.093349
10494	hsa-miR-140-3p	9.640101	STK25	8.560094
1E+08	hsa-miR-142-3p	12.54444	CD24	13.56376
1959	hsa-miR-142-3p	12.54444	EGR2	7.264601
8228	hsa-miR-142-3p	12.54444	PNPLA4	5.594859
6451	hsa-miR-142-3p	12.54444	SH3BGRL	8.942829
56652	hsa-miR-142-5p	8.828202	C10orf2	6.677458
1E+08	hsa-miR-142-5p	8.828202	CD24	13.56376
1050	hsa-miR-142-5p	8.828202	CEBPA	7.954701
1959	hsa-miR-142-5p	8.828202	EGR2	7.264601
26234	hsa-miR-142-5p	8.828202	FBXL5	5.901917
2359	hsa-miR-142-5p	8.828202	FPR3	7.741393
3290	hsa-miR-142-5p	8.828202	HSD11B1	6.081242
83593	hsa-miR-142-5p	8.828202	RASSF5	7.743423
7078	hsa-miR-142-5p	8.828202	TIMP3	7.92518
1230	hsa-miR-150	11.01249	CCR1	6.077411
1959	hsa-miR-150	11.01249	EGR2	7.264601
26234	hsa-miR-150	11.01249	FBXL5	5.901917
440823	hsa-miR-150	11.01249	MIAT	5.6395
84722	hsa-miR-155	9.509899	PSRC1	5.721769
79863	hsa-miR-155	9.509899	RBFA	7.037016
9447	hsa-miR-17	10.43675	AIM2	6.815321
57673	hsa-miR-17	10.43675	BEND3	6.554452
1056	hsa-miR-17	10.43675	CEL	6.667653
1513	hsa-miR-17	10.43675	CTSK	9.360128
1959	hsa-miR-17	10.43675	EGR2	7.264601
26234	hsa-miR-17	10.43675	FBXL5	5.901917

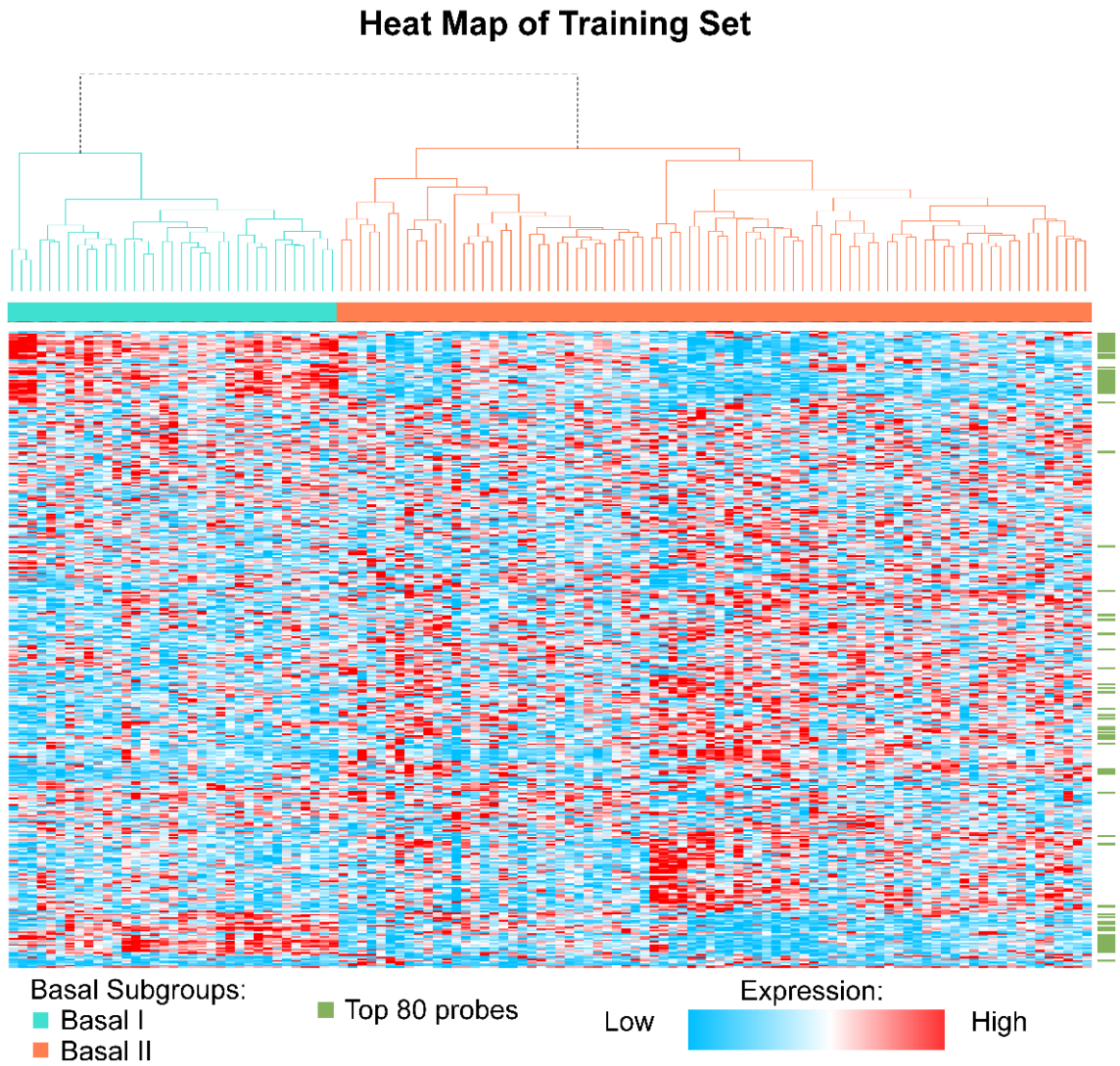
8228	hsa-miR-17	10.43675	PNPLA4	5.594859
149628	hsa-miR-17	10.43675	PYHIN1	5.734864
6641	hsa-miR-17	10.43675	SNTB1	7.038167
7078	hsa-miR-17	10.43675	TIMP3	7.92518
10663	hsa-miR-19b-1*	7.597889	CXCR6	5.992908
2212	hsa-miR-19b-1*	7.597889	FCGR2A	7.721801
3290	hsa-miR-19b-1*	7.597889	HSD11B1	6.081242
83463	hsa-miR-19b-1*	7.597889	MXD3	6.78071
9046	hsa-miR-200c*	6.780434	DOK2	5.848977
55355	hsa-miR-200c*	6.780434	HJURP	7.789458
3559	hsa-miR-200c*	6.780434	IL2RA	5.801549
84722	hsa-miR-200c*	6.780434	PSRC1	5.721769
23223	hsa-miR-200c*	6.780434	RRP12	8.093349
9046	hsa-miR-22	14.10687	DOK2	5.848977
2760	hsa-miR-22	14.10687	GM2A	6.933033
3290	hsa-miR-22	14.10687	HSD11B1	6.081242
83463	hsa-miR-22	14.10687	MXD3	6.78071
8228	hsa-miR-22	14.10687	PNPLA4	5.594859
10494	hsa-miR-22	14.10687	STK25	8.560094
7078	hsa-miR-22	14.10687	TIMP3	7.92518
64581	hsa-miR-29a	13.6678	CLEC7A	5.788349
1789	hsa-miR-29a	13.6678	DNMT3B	6.369691
2212	hsa-miR-29a	13.6678	FCGR2A	7.721801
2326	hsa-miR-29a	13.6678	FMO1	6.631723
79047	hsa-miR-29a	13.6678	KCTD15	6.723827
440823	hsa-miR-29a	13.6678	MIAT	5.6395
22974	hsa-miR-29a	13.6678	TPX2	8.119814
64581	hsa-miR-29c	11.81329	CLEC7A	5.788349
1789	hsa-miR-29c	11.81329	DNMT3B	6.369691
2212	hsa-miR-29c	11.81329	FCGR2A	7.721801
2326	hsa-miR-29c	11.81329	FMO1	6.631723
79047	hsa-miR-29c	11.81329	KCTD15	6.723827
440823	hsa-miR-29c	11.81329	MIAT	5.6395
22974	hsa-miR-29c	11.81329	TPX2	8.119814
84253	hsa-miR-29c*	7.570919	GARNL3	6.125192
55355	hsa-miR-29c*	7.570919	HJURP	7.789458
54069	hsa-miR-29c*	7.570919	MIS18A	7.022054
83463	hsa-miR-342-3p	9.43096	MXD3	6.78071
84262	hsa-miR-342-3p	9.43096	PSMG3	7.342099
171558	hsa-miR-342-3p	9.43096	PTCRA	5.464495

5788	hsa-miR-342-3p	9.43096	PTPRC	5.768135
7078	hsa-miR-342-3p	9.43096	TIMP3	7.92518
79058	hsa-miR-342-5p	7.905677	ASPSCR1	10.02142
837	hsa-miR-342-5p	7.905677	CASP4	8.496007
10320	hsa-miR-342-5p	7.905677	IKZF1	6.690869
84722	hsa-miR-342-5p	7.905677	PSRC1	5.721769
6373	hsa-miR-34a	12.65928	CXCL11	6.155282
79980	hsa-miR-34a	12.65928	DSN1	7.176337
2212	hsa-miR-34a	12.65928	FCGR2A	7.721801
26191	hsa-miR-34a	12.65928	GPR65	5.66744
22806	hsa-miR-34a	12.65928	IKZF3	5.405269
8228	hsa-miR-34a	12.65928	PNPLA4	5.594859
719	hsa-miR-361-3p	8.301586	C3AR1	6.422074
1050	hsa-miR-361-3p	8.301586	CEBPA	7.954701
2760	hsa-miR-361-3p	8.301586	GM2A	6.933033
440823	hsa-miR-361-3p	8.301586	MIAT	5.6395
6838	hsa-miR-361-3p	8.301586	SURF6	8.828021
7078	hsa-miR-361-3p	8.301586	TIMP3	7.92518

Supporting Information – Figure 7.6

Figure 7.6 The heat map of 400 probes in METABRIC training set

This heat map shows the hierarchical clustering of 115 basal-like samples based on the probes expression values. There are two major clusters representing Basal I (turquoise) and Basal II (coral). From these features, the top 80 best discriminating between the major groups the most are denoted in orange. The red and blue colours represent relative over- and under-expressions respectively. The expression values are normalised across the samples.



Supporting Information – Text 7.1

Text 7.1 Basal-like biomarkers and drug-targets

Among all breast cancers, basal-like and triple-negative are the greatest challenges for both oncologists and patients due to their unpredicted behaviour, high rates of recurrence and mortality. Although tumours within these types show similar clinicopathological features, they exhibit highly variable therapy response and disease outcome (Bosch et al., 2010). There is no baseline protocol for treating BLBCs or TNBCs (Prat et al., 2013). The current standard management usually consists of surgery, radiotherapy and chemotherapy, with the administration of cytotoxic drugs, alone or in combination (Toft & Cryns, 2011). The aggressive regimens are highly favourable to patient's response as they show immediate effect on cell proliferation and tumour growth. On the other hand, these regimens increase the toxicity – with the destruction of healthy normal cells – and lead to mild or severe adverse effects (Crown et al., 2012). Drug resistance and disease remission are, yet, additional issues to be considered (De Laurentiis et al., 2010).

The most common chemotherapeutic approach to treating advanced BLBCs or TNBCs is based on anthracycline and taxane combinations in the first line, followed by capecitabine as disease progresses (Oakman et al., 2010). The use of platinum-derived agents has also impacted the management of *BRCA1*-mutated tumours (Drost & Jonkers, 2014). Recent tests suggest the use of carboplatin as oppose to cisplatin, which was previously the most effective agent against triple-negative breast cancers (Carmo-Pereira et al., 1989; Kolarić & Vukas, 1991; Martin et al., 1991). However, there is a lack of evidence from random assignment trials to support the preferred use of platinum compounds over standard cytotoxic agents for TNBC, especially for early-stage disease. Although several clinical trials are currently ongoing in this population, further investigation of novel drugs is required, regardless of the stage at diagnosis (Crown et al., 2012).

Targeted therapies are used to precisely identify and attack cancer cells by affecting molecular pathways (Orlando et al., 2010). Drugs may directly interact with growth factor receptors, DNA-repair and apoptosis regulators, and angiogenesis mediators blocking cancer progression. Despite the potential of targeted therapies, clinically the drugs are rarely administered as a single agent due to the limited therapeutic effectiveness and onset of resistance. This approach is not likely to replace cytotoxic drugs in the foreseeable future; it will rather be used in combination. Within the next decades, the emergence of multi-targeting drugs is expected, not only for cancer but for other diseases of polygenetic nature. The concept of multi-targeting has emerged with the application of computational resources and network-based approach to define strong compounds for the different stages of clinical trials (Brandl et al., 2014; Soldi et al., 2013). In this context, medical advances have led to the identification of a

variety of potential targets and the development of anticancer drugs and new therapy combinations for breast cancer (Zhou et al., 2009).

The traditional way of drug discovery – also referred to as *de novo* drug discovery – is a complex, time-consuming and expensive process that has currently a high rate of failure (Strittmatter, 2014). From the initial identification of a compound to determining its pharmacological and toxicological activity *in vitro* (preclinical models) and *in vivo* (clinical trials) usually requires an average of 13 years of research, before its approval and commercialization (Gupta et al., 2013). Although the billions invested by pharmaceutical companies on drug discovery, development and marketing, the number of new approved chemical compound is significantly lower in comparison to that of failed drugs. The continued rising costs of this market have driven the industry towards the exploration of new strategies in the field.

The process of drug rediscovery – also known as redirection, repurposing, repositioning and reprofiling –, or finding new uses for existing compounds outside the scope of the original indication, is rather promising (Ashburn & Thor, 2004; Langedijk et al., 2015). It requires significantly less time for approval due to the wealth of preclinical and clinical information relating to toxicity, pharmacokinetic and pharmacodynamics effects. Furthermore, the development of drugs should cost significantly less compared to rationally designed new drugs. Drug rediscovery brings new opportunities to cancer research for exploiting alternative mechanisms of thousands of approved drugs, generics, and late-stage development agents, and promoting an open-source drug discovery along with the pharmaceutical industry (Napper & Mucke, 2015). Databases aim to convert fundamental information into meaningful and valuable knowledge and broad the horizons to clinical applications. In sheer volume, however, databases present qualitative and quantitative challenges. A step further involves the drug reformulation, dosage, delivery mechanisms and combination therapies.

High-throughput screening (HTS), high content screening (HTC), Chemoinformatics, Bioinformatics, as well as Network and Systems Biology have been used in conjunction with available information on known targets, drugs, biomarkers and pathways. These sources have been designed to accelerate the timelines of rediscovery and development of candidate drugs. In particular, great effort has been devoted to generate well-annotated repositories such as DrugBank, Therapeutic Target Database (TTD), Cancer Commons, Clearity Foundation, ChEMBL, Clinical Trail, Cancer Resource, Comparative Toxicogenomic Database (CTD), Kegg Drug, MetaDrug, My Cancer Genome, PharmGKB, PubChem, SuperTarget, Trends in the Exploration of Novel Drug targets (TEND) and IUPHAR/BPS Guide to Pharmacology. These platforms have linked chemical compounds with their molecular function, mechanisms of action

and adverse events. This information supports and expands the biological understanding of dynamic variables for planning viable and creative *in silico* models.

An example of 'gene-target' search, covering large collections from multiple databases, is detailed in **Table 7.15** . The biomarkers defining basal-like subtype were obtained from *Chapter 4* and *Chapter 7* and the respective compounds were defined across databases. The drug-target relations emerged from nine public databases: DrugBank, TTD, Clearity Foundation, Clinical Trail, My Cancer Genome, PharmGKB, SuperTarget, TEND and IUPHAR. We further demonstrate the application of a kernelization for (α,β,d)-Hitting Set to multiple drug selection for cancer therapy (**Figure 7.7**), indicating that this problem is readily scalable to large datasets (Mellor et al., 2010). The proposed approach is critical towards decision of future *in vitro* tests in cell lines. New strategies and techniques for drug discovery and development, however, are essential for advancing translational science. It is also imperative to identify breast cancer subtypes likely to benefit from a treatment and patients at high risk of toxicity.

Supporting Information – Table 7.15

Table containing potential compounds for treating basal-like and triple-negative breast cancers.

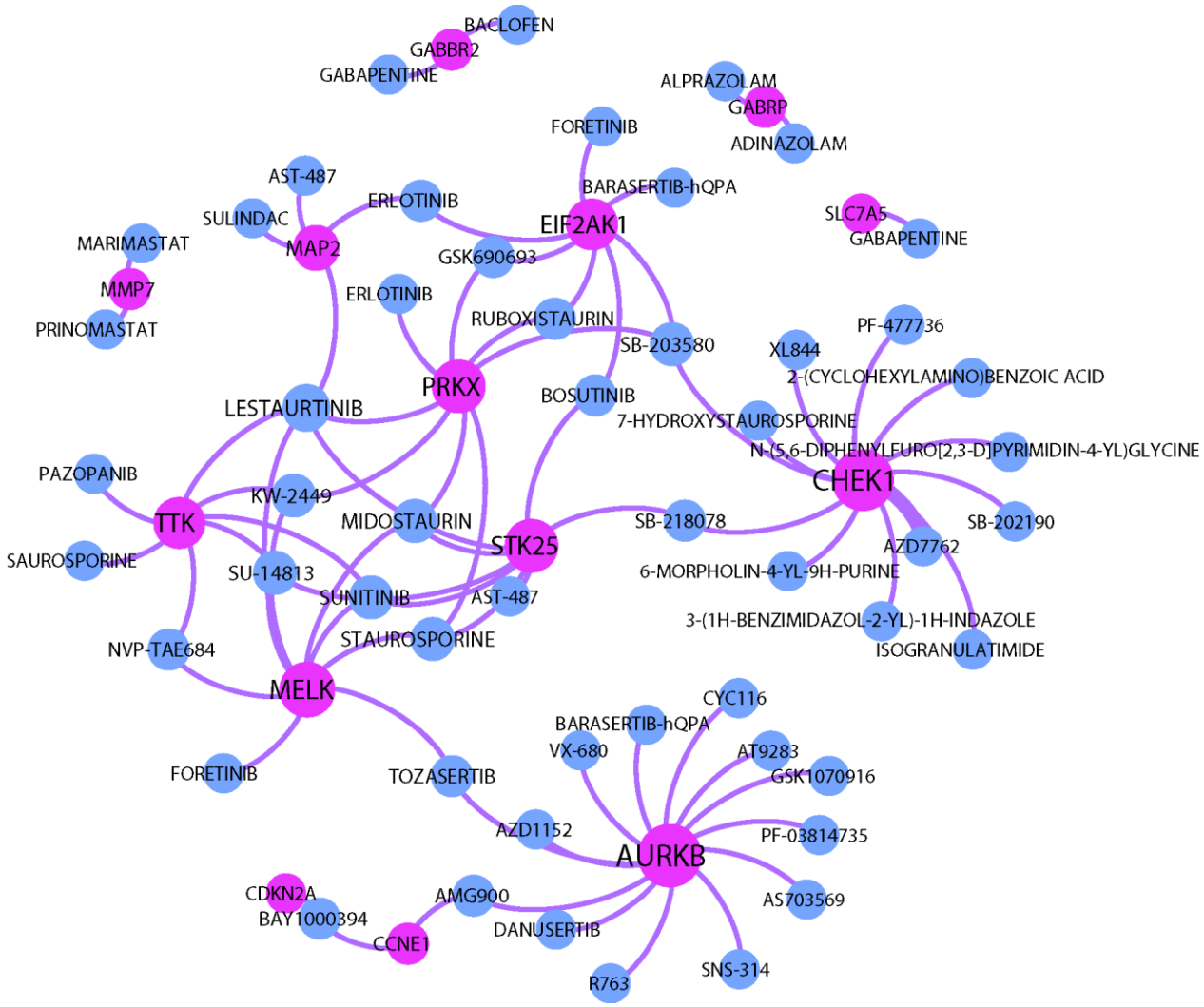
Table 7.15 Summary gene targets and corresponding drugs

Gene Symbol	Drugs
AURKB	AT9283, AMG900, AZD1152, CYC116, GSK1070916, PF-03814735, R763, SNS-314, TOZASERTIB, VX-680, AS703569, DANUSERTIB, BARASERTIB-hQPA
CCNE1	BAY1000394, AMG900
CDKN2A	BAY1000394
CHEK1	2-(CYCLOHEXYLAMINO)BENZOIC ACID, 3-(1H-BENZIMIDAZOL-2-YL)-1H-INDAZOLE, 6-MORPHOLIN-4-YL-9H-PURINE, 7-HYDROXYSTAUROSPORINE, AZD7762, AZD7762, ISOGRANULATIMIDE, N-(5,6-DIPHENYLFURO[2,3-D]PYRIMIDIN-4-YL)GLYCINE, PF-477736, SB-202190, SB-203580, XL844, AZD7762, SB-218078
EIF2AK1	BARASERTIB-hQPA, BOSUTINIB, ERLOTINIB, FORETINIB, GSK690693, RUBOXISTAURIN, SB-203580
GABBR2	BACLOFEN, GABAPENTINE
GABRP	ADINAZOLAM, ALPRAZOLAM,
MAP2	SULINDAC, AST-487, ERLOTINIB, LESTAURTINIB,
MELK	FORETINIB, KW-2449, LESTAURTINIB, MIDOSTAURIN, NVP-TAE684, STAUROSPORINE, SU-14813, SUNITINIB, TOZASERTIB
MMP7	MARIMASTAT, PRINOMASTAT
PRKX	ERLOTINIB, GSK690693, KW-2449, LESTAURTINIB, MIDOSTAURIN, RUBOXISTAURIN, SB-203580, STAUROSPORINE
SLC7A5	GABAPENTINE
STK25	AST-487, BOSUTINIB, LESTAURTINIB, MIDOSTAURIN, SB-218078, STAUROSPORINE, SU-14813, SUNITINIB
TTK	AST-487, KW-2449, LESTAURTINIB, NVP-TAE684, PAZOPANIB, SAUROSPORINE, SU-14813, SUNITINIB

Supporting Information – Figure 7.7

Figure 7.7 Network analysis of multiple drug targets for breast cancer therapy

The image was generated using the Gephi software. Genes and drugs are defined as nodes, connected by edges that show association between genes and putative drug targets. Genes are represented by pink nodes while drugs are represented by blue nodes (a contribution from *Ademir Gabardo*). Alternatively, this association may be expanded by applying hitting set hits to all selected targets at least once based on the kernelisation approach for a (α,β,d)-Hitting Set (Mellor et al., 2010).



Supporting References

Ashburn, T. T., & Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov, 3*(8), 673-683.

Bosch, A., Eroles, P., Zaragoza, R., Vina, J. R., & Lluch, A. (2010). Triple-negative breast cancer: molecular features, pathogenesis, treatment and current lines of research. *Cancer Treat. Rev., 36*(3), 206-215.

Brandl, M. B., Pasquier, E., Li, F., Beck, D., Zhang, S., Zhao, H., et al. (2014). Computational analysis of image-based drug profiling predicts synergistic drug combinations: Applications in triple-negative breast cancer. *Mol. Oncol., 8*(8), 1548-1560.

Carmo-Pereira, J., Olivera-Costa, F., & Henriquez, E. (1989). *Carboplatin as primary chemotherapy for disseminated breast carcinoma: a phase II study.* Paper presented at the 5th European Conference on Clinical Oncology.

Crown, J., O'Shaughnessy, J., & Gullo, G. (2012). Emerging targeted therapies in triple-negative breast cancer. *Ann. Oncol., 23*(suppl 6), vi56-vi65.

De Laurentiis, M., Cianniello, D., Caputo, R., Stanzione, B., Arpino, G., Cinieri, S., et al. (2010). Treatment of triple negative breast cancer (TNBC): current options and future perspectives. *Cancer Treat. Rev., 36S3*, S80–S86.

Drost, R., & Jonkers, J. (2014). Opportunities and hurdles in the treatment of BRCA1-related breast cancer. *Oncogene, 33*(29), 3753-3763.

Gupta, S. C., Sung, B., Prasad, S., Webb, L. J., & Aggarwal, B. B. (2013). Cancer drug discovery by repurposing: teaching new tricks to old dogs. *Trends Pharmacol. Sci., 34*(9), 508-517.

Kolarić, K., & Vukas, D. (1991). Carboplatin activity in untreated metastatic breast cancer patients—results of a phase II study. *Cancer Chemother. Pharmacol., 27*(5), 409-412.

Langedijk, J., Mantel-Teeuwisse, A. K., Slijkerman, D. S., & Schutjens, M.-H. D. B. (2015). Drug repositioning and repurposing: terminology and definitions in literature. *Drug Discovery Today, 20*(8), 1027-1034.

Martin, M., Diaz-Rubio, E., Casado, A., & López, V. J. (1991). Phase II study of carboplatin in advanced breast cancer: preliminary results. *Semin. Oncol., 18*(1 Suppl 2), 23-27.

Mellor, D., Prieto, E., Mathieson, L., & Moscato, P. (2010). A kernelisation approach for multiple d-Hitting Set and its application in optimal multi-drug therapeutic combinations. *PLoS One, 5*(10), e13055.

Napper, A. D., & Mucke, H. A. (2015). A Special Focus on Drug Repurposing, Rescue, and Repositioning. *Assay Drug Dev Technol, 13*(6), 293.

Oakman, C., Viale, G., & Di Leo, A. (2010). Management of triple negative breast cancer. *The Breast, 19*(5), 312-321.

Orlando, L., Schiavone, P., Fedele, P., Calvani, N., Nacci, A., Rizzo, P., et al. (2010). Molecularly targeted endocrine therapies for breast cancer. *Cancer Treat. Rev., 36*, S67–S71.

Prat, A., Adamo, B., Cheang, M. C., Anders, C. K., Carey, L. A., & Perou, C. M. (2013). Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. *Oncologist, 18*(2), 123-133.

Soldi, R., Cohen, A. L., Cheng, L., Sun, Y., Moos, P. J., & Bild, A. H. (2013). A genomic approach to predict synergistic combinations for breast cancer treatment. *Pharmacogenomics J., 13*(1), 94-104.

Strittmatter, S. M. (2014). Overcoming Drug Development Bottlenecks With Repurposing: Old drugs learn new tricks. *Nat. Med., 20*(6), 590-591.

Toft, D. J., & Cryns, V. L. (2011). Minireview: Basal-like breast cancer: from molecular profiles to targeted therapies. *Mol. Endocrinol., 25*(2), 199-211.

Zhou, B. B., Zhang, H., Damelin, M., Geles, K. G., Grindley, J. C., & Dirks, P. B. (2009). Tumour-initiating cells: challenges and opportunities for anticancer drug discovery. *Nat Rev Drug Discov, 8*(10), 806-823.

CHAPTER 8

8. CONCLUDING REMARKS

Chapter 8, the final chapter, draws on the conclusions of the previous chapters and addresses the research hypotheses and questions raised in *Chapter 1*. The **8.1 Final Statements** summarizes the impact of the devised bioinformatics methods on the analysis of the breast cancer disease, which prompt a more stringent assignment of intrinsic subtype labels in the METABRIC data set. The variety of prediction models and measures using machine learning algorithms indicates that the relation between molecular signatures subtype-specific and overall classification can be highly complex and counterintuitive. Intrinsic subtypes are not sufficiently well understood. After more than a decade of research conducted in molecular breast cancer classification, there still is no consensus on either the number or definition of intrinsic subtypes. Novel methods and applications are mandatory to support the constantly evolving high-throughput 'omics' technologies, towards the elucidation of the mechanisms underlying breast cancer. The course of breast cancer research is further delineated in **8.2 Future Directions** and emphasize that ‘divide and conquer’ schemes consist in the best temporary solution for effective personalised medicine. Finally, **8.3 Closing Note** summarizes the major contribution of this thesis to fundamental and clinical research.

8.1 Final Statements

Microarray technologies, at the time of their inception, were considered a new prospect in cancer research and oncology practice with the hope that within a decade diseases would be deeply understood. The robust analysis and interpretation of microarray gene expression data have provided a plethora of unique accomplishments and challenges. The high-throughput molecular profiles showed the immense potential of this technique for breast cancer subtyping and prediction. Notably, it became clear that breast cancer is not a single disease, but a collection of independent entities with significant differences in gene expression patterns and clinical outcomes. These entities, or breast cancer subtypes, have been thoroughly investigated in the last decades, and in this thesis from both theoretical and practical perspectives. The standard methods here applied, nevertheless, were exclusive in the form and reportage for the disease investigation; used to address the questions posed in *Chapter 1*.

"How many groups or different subtypes could be clearly identified in breast cancer disease using gene expression microarray data? Are they molecularly and clinically well defined?"

For addressing these questions, the present study considered exploring the use of standard clustering methods, as an alternative to common hierarchical clustering approaches, to improve the classification of breast cancer subtypes. MST-*k*NN clustering was applied to illustrate the complexity of this disease and, at the same time, investigate the consistency of current intrinsic subtypes. These applications were motivated by the dearth of attempts using unconventional methods in identifying homogeneous clusters. Early experiments – later introduced in *Chapter 5* (**Supporting Information – Figure 5.6, Figure 5.7, Figure 5.8, Table 5.7 and Table 5.8**) revealed connections between samples with similar gene expression profiles in the METABRIC data set. However, the same clustering methods have exposed major groups (IntClusts), not comparable with those from hierarchical approaches; consequently, not comparable with the current breast cancer classification (or intrinsic subtypes).

The MST-*k*NN clusters showed single connections as well as large and complete clusters generated from robust data. This data also indicated that the five intrinsic subtypes (luminal A, luminal B, HER2-enriched, basal-like and normal-like) assigned using the PAM50

method have a regular distribution on the METABRIC cluster. For this reason, we stand for the well-established subtypes and further investigate each one. In *Chapter 4*, we portrayed novel biomarkers explaining the five subtypes by exploring the ability of the CM1 score. Additionally, in *Chapter 5*, we refined the original labels and improved class prediction in the METABRIC and ROCK data sets using an iterative approach. The new labelling showed more reliable clinicopathological features and more consistent survival outcomes across the intrinsic subtypes, making them molecularly and clinically better defined (Figure **8.1** and **Figure 8.2**). The new labels are of great value to breast cancer research and future translational science, as inconsistent assignments may lead to misguided information in the field.

"Which gene or signatures are able to individualise the different breast cancer subtypes? Are these genes relevant targets for tailored treatment?"

In *Chapter 4*, we identified 30 novel biomarkers and 12 well-known genes for subtype individuation. These genes showed highly discriminative patterns of expression across samples for each intrinsic group. We further assessed the ability of these 42 probes in assigning the correct subtype labels using 24 different classifiers from the Weka software suite. For comparison, the same method was applied to the list of 50 genes from the PAM50 method. Towards the development of more reliable strategies, in *Chapter 5*, we designed an iterative approach to select probes that consistently discriminate breast cancer subtypes. Furthermore, in *Chapter 6*, we proposed a novel approach for leveraging the utility of pairwise probes in covering both the intrinsic signatures and the subtype individuation. All proposed methods have delineated molecular imbalances among subtypes and further support the group-based definition use in medical practice. The models devised in this study are promising prediction tools, widely applicable to a variety of data types.

Overall, the signatures we provide are more consistent and in better agreement with the distribution of clinical markers (ER, PR and HER2) and patients' overall survival than those defined by the PAM50 method. Considering single biomarkers from the signatures, they are capable to individualise subtypes, however, may or may not be a surrogate therapeutic targets. By exploring drug-target databases, in *Chapter 7*, we selected putative drugs from distinct databases to guide lab experiments for treating basal-like breast cancer, one of the most aggressive subtypes. *In vitro* tests are, however, required to determine effective markers for improving treatment response of an individual or a group of patients.

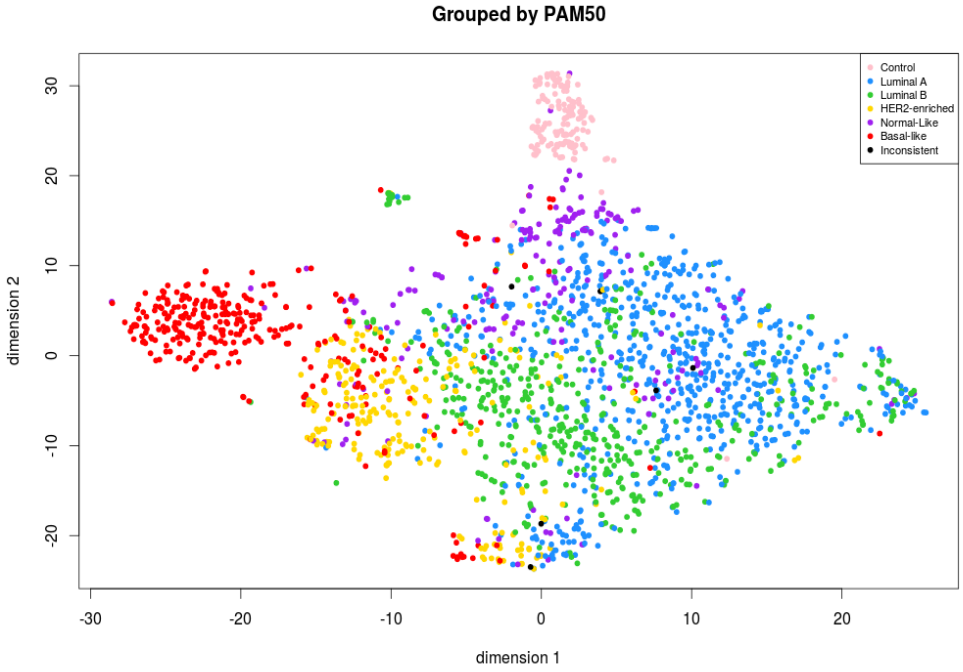


Figure 8.1 t-SNE graph of METABRIC samples coloured according to PAM50

The t-SNE graphs were plotted based on the distance between all METABRIC tumour samples, considering the set of 48803 probes, coloured with the respective PAM50 labels.

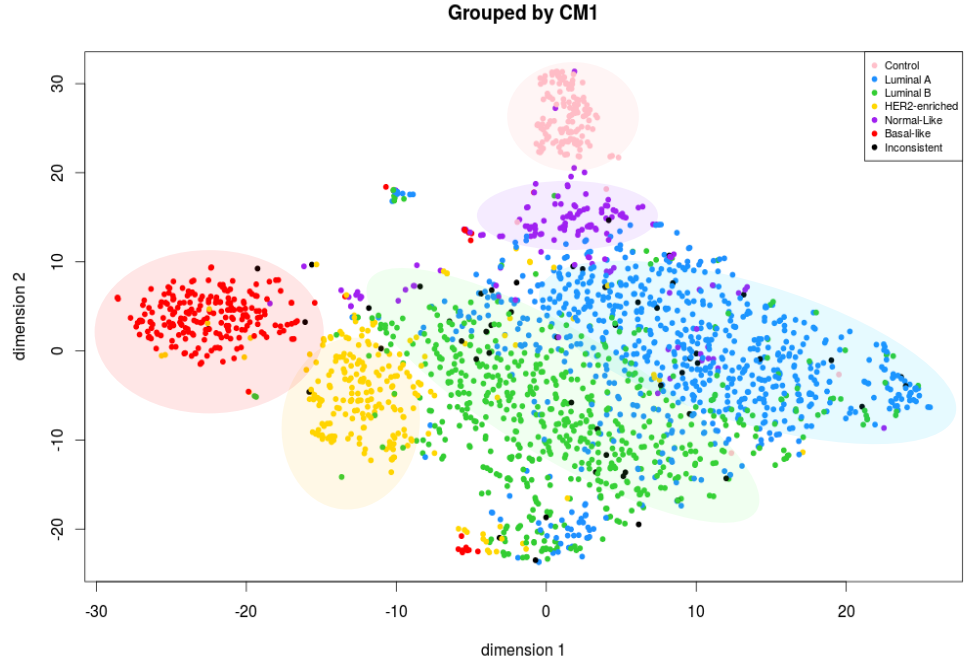


Figure 8.2 t-SNE graph of METABRIC samples coloured using the refined labels

The t-SNE graphs were plotted based on the distance between all METABRIC tumour samples, considering the set of 48803 probes, coloured with the new refined labels.

"How could molecular data, including genome and transcriptome microarrays, be better combined or integrated to improve the understanding of the disease or the subtypes' classification?"

The first data combination is described in *Chapter 6* with the introduction of a novel structure based on pairwise probes, expanding the information contained in the METABRIC transcriptome set. Simple mathematical modelling and well-established methodologies of feature selection and data mining were used to uncover molecular imbalances across subtypes. The computational framework was motivated by the widespread interest in conducting and reporting robust methods to build accurate predictor models. Accordingly, the most representative meta-features exhibited extensive predictive power for labelling samples. The current systematic approach prove that it is a promising tool for improving the understanding of the disease, especially for underlying breast cancer intrinsic subtypes.

In the second data integration, *Chapter 7*, we utilised transcriptomic (gene and miRNA expression) and genomic (copy number aberration) information to perform a comprehensive analysis of basal-like tumours. Thus, we provided an 80-probe signature associated with varying survival outcomes, including putative markers of disease progression and promising asset for clinical applications. This signature was able to distinguish between two basal-like subgroups (Basal I and Basal II) with divergent molecular profiles, clinical features and survival outcomes. Furthermore, miRNAs transcripts also correlated with the basal-like subgroups. The genomic analysis further differentiated Basal I and Basal II on the percentages of gains/amplifications and losses/deletions across samples. These results have demonstrated the heterogeneity of basal-like tumours beyond the classical immunohistochemistry.

The innovative assessment of genomic and transcriptomic data presented in this study contributes towards a more robust definition of breast cancer. The importance of defining groups-at-risk within subtypes is projected on the impact of breast cancer management in the clinical setting and, more importantly, in therapy response. Although several clinicopathological features have been used to discriminate between low- and high-risk patients, the identification of novel molecular features with prognostic value expands the disease overview. By recognising them, researchers and clinicians should be able to design more effective tailored therapies for patients at high risk; and reduce the prescription of high-dose chemotherapy to individuals at a low risk, thus reducing or minimising side effects.

"Is it possible to link cell line profiles with the breast cancer subtypes in order to provide consistent information for 'in vitro' drug tests?"

Cell lines are widely used to investigate the breast cancer clinical pathology and molecular heterogeneity. The stratification of tumour lineages according to intrinsic subtypes has also changed the functional management of laboratory models. In particular, cancer cell cultures are explored to test and validate molecular drivers for group targeted therapies. Accordingly, we classified the *in vitro* models by comparing the gene expression profile of cell lines and approximately 2000 primary breast tumours from the METABRIC data set (data not shown). The proposed method assessed the CM1 and PAM50 gene lists to classify 75 cell lines using an ensemble learning approach. Preliminary results showed an overall consensus on the cell lines classification; but a disagreement on the subtype labels attributed to widely used lineages, such as BT474, HCC1500, HCC1954, and SKBR3. These cell lines had different labelling across studies and platforms. This approach adds a new perspective to effective experimental models that are used to investigate intrinsic subtypes for improving the therapeutic decisions and the clinical outcomes. This information will be used to filter future lab experiments using the drug targets described in *Chapter 8*.

8.2 Future Directions

Microarray profiling has been crucial to the growth and maturation of bioinformatics techniques and has also laid a solid basis for the analysis of gene expression signatures. With the increasing popularity of microarray analysis, however, the perspective for understanding breast cancer disease profoundly change. Irrespective of the contribution of microarray studies, measurement techniques usually become obsolete over time. With the advent of new 'omics' sciences – including genomics, transcriptomics, epigenomics, proteomics, lipidomics, metabolomics, etc. – and technologies, gene expression microarrays are unlikely to escape this fate. Next generation sequencing is revolutionizing molecular research with the dissection of chromatin immunoprecipitation coupled to DNA microarrays (ChIP-chip), DNA (DNA-seq) and RNA sequencing (RNA-seq), and whole genome sequencing (WGS); in parallel with the analysis of protein chips and a range of other high-throughput measurements. These technologies aim at the

collective characterization and quantification of molecular features that may explain the structure, function and dynamics of cells, tissues and organisms. They complement microarray gene expression data and provide additional knowledge for uncovering the molecular imbalance in breast cancers, and within intrinsic subtypes.

In this thesis, the applied methodologies provide a feasible way to efficiently search the entire microarray gene expression space for candidate robust classification sets. A future goal is to develop a rank-based enrichment analysis method that compares the different 'omics' information according to their abilities in differentiating classes. These classes should, ultimately, be more homogeneous than the current intrinsic subtypes and able to compose the true taxonomy of breast cancer disease. In this context, the intrinsic pathways and molecular imbalances are expected to match the DAVID database, used throughout this thesis, but also other databases that remain unexplored. A number of publicly available bioinformatics tools – including, but not limited to, GoMiner, EaseGo, GOstat, Onto-express, GoToolBox, FatiGO and GOSSIP – reflect the biological processes most pertinent to revealing molecular phenomena and associated phenotypes. The power of many of these applications is in highlighting the most over-represented signatures, out of a list of hundreds or thousands features, and providing a systematic means of understanding data. However, with the increasing collection of complex data, actual integrated sources are required for data analysis and interpretation.

The emerging bioinformatics resources and tools will impose personal and social challenges for breast cancer researchers in the next decades. In this scenario, they will provide further hints about the relationship between genetic and environmental causes, leading to improvements in breast cancer detection and prevention. With regards to the treatment, patients will likely benefit from targeted therapies and avoid the risks of chemotherapy toxicity and adverse effects. Furthermore, microenvironment and peripheral system disorders in cancer will be also considered as primordial factors in the root nature of breast cancer disease. In the broad view, immune and inflammatory phenomena are critical, representing some of the most promising areas of research in the field. Ultimately, the understanding of individual features and major biological networks – within intrinsic subtypes – will allow advances in drug development over the next several years for truly tailored therapy in future clinical trials, leading toward effective personalised medicine.

8.3 Closing Note

The contribution of this thesis is to provide an overview of breast cancer research in context of applied bioinformatics. Additionally, a variety of powerful techniques for the comprehensive investigation of intrinsic subtypes, both at the genomic and transcriptomic level, is also employed and described. The implications of methods and techniques are discussed on their relevance and potential clinical impact; however, questions remain without answer due to the involvement of major, complex signalling pathways in the non-linear disease progression. We anticipate that the power of systemic approaches will increase as additional and complementary molecular features are defined, including multi-omics sources from the genome, transcriptome, epigenome, proteome, lipidome and metabolome. In sheer volume, this approach presents a qualitative and quantitative challenge for the future decades. The main goal is to convert this precious data into meaningful information, valuable knowledge and to broaden the horizons of clinical management.