REFINING MOLECULAR SUBTYPING IN BREAST AND COLON CANCERS USING GENE EXPRESSION AND PROTEOMICS DATA

ARCHITHA ELLAPPALAYAM

Propositions

1. Cancer diagnosis and treatment should be constantly updated to accommodate the newly discovered tumor biology.
(this thesis)

2. Precision medicine will reach its full potential when it effectively handles tumor heterogeneity.
(this thesis)

3. Positive impact of alternative medicine is substantially overrated.

4. A researcher should be measured by the scientific quality and the societal impact of their research.

5. A completely sustainable lifestyle cannot be afforded by all.

6. A four-day workweek should be embraced as the new norm.

Propositions belonging to the thesis, entitled

Refining Molecular Subtyping in Breast and Colon Cancers using Gene Expression and Proteomics Data

Architha Ellappalayam
Wageningen, September 12, 2023

Refining Molecular Subtyping in Breast and Colon Cancers using Gene Expression and Proteomics Data

Architha Ellappalayam

Thesis committee

Promotors
Prof. Dr Vitor A.P. Martins dos Santos
Personal chair, Bioprocess Engineering Group
Wageningen University & Research

Prof. Dr Maria Suarez Diez
Professor at the Laboratory of Systems and Synthetic Biology
Wageningen University & Research

Co-promotor
Dr Edoardo Saccenti
Assistant Professor at the Laboratory of Systems and Synthetic Biology
Wageningen University & Research

Other members
Prof. Dr Martien Groenen, Wageningen University, Wageningen
Dr Marleen Kok, Netherlands Cancer Institute, Amsterdam
Dr Arcangela Denicolo, Institute Oncologico Veneto, Italy
Dr Harmen van de Werken, Erasmus MC, Rotterdam

This research was conducted under the auspices of VLAG Graduate School (Biobased, Biomolecular, Chemical, Food and Nutrition Sciences)

Refining Molecular Subtyping in Breast and Colon Cancers using Gene Expression and Proteomics Data

Architha Ellappalayam

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 12 September 2023
at 11 a.m. in the Omnia Auditorium.

Architha Ellappalayam
Refining Molecular Subtyping in Breast and Colon Cancers using Gene Expression and Proteomics Data ,
170 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2023)
With references, with summary in English

ISBN: 978-94-6447-773-3
DOI: https://doi.org/10.18174/633837

Contents

1 Introduction 3

2 BluePrint Breast Cancer Molecular Subtyping Recognizes Single and Dual Subtype Tumors with Implications for Therapeutic Guidance 23

3 A twenty-nine HER2 biology Guided Gene Signature Improves Breast Cancer HER2-type Molecular Classification 47

4 Investigating the Concordance in Molecular Subtypes of Primary Colorectal Tumors and their Matched Synchronous Liver Metastasis 65

5 Molecular Subtyping of Triple-Negative Breast Cancer using Proteomics Data 93

6 General Discussion 117

7 Summary 129

References 133

8 List of Publications 157

9 Overview of completed training activities 161

10 About the author 163

11 Acknowledgements 165

Chapter 1

Introduction

1.1 Abstract

Advancements in research and technology have transformed cancer treatment, driving progress toward more effective therapies. In the continuing battle against cancer, this chapter provides a comprehensive exploration of breast and colorectal cancer, shedding light on their statistics, common causes, and the transformative advancements that drove progress in cancer treatment. The significance of moving beyond clinical subtypes was recognized, emphasizing the importance of molecular characterization in precision oncology and personalized treatment. By delving into the diagnosis workflow of breast cancer and unraveling the complexities of its molecular subtypes, a deeper understanding of tumor biology was gained, paving the way for improved therapeutic guidance. Additionally, the classification of colon cancer uncovered its clinical and molecular subtypes, revealing valuable insights into this complex disease. The benefits of precision oncology were highlighted, underscoring the need for precise diagnostics and the potential to maximize outcomes. The overarching aims of the thesis were discussed in detail. Chapter 2 focused on the BluePrint test's ability to identify dual-activated pathways in breast cancer patients. Chapter 3 expanded the BluePrint HER2 gene signature, while Chapter 4 investigated the concordance of molecular characteristics in primary and metastatic colon tumors. Chapter 5 explored subgroups within triple-negative breast cancer using proteogenomic datasets. Finally, Chapter 6 summarized the key findings, discussed their correspondence with existing subtyping techniques, and highlighted the research's limitations and prospects. My thesis aimed to increase knowledge to contribute to the understanding of breast and colorectal cancer subtypes, their molecular characteristics, and their implications for personalized treatment. Through these efforts, this thesis chapter aimed to transform cancer diagnostics, ultimately improving lives and strengthening communities.

1.2 The Continuing Battle Against Cancer: An Overview of Statistics and Common Causes

Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells in the body [1]. Cancer cells can invade nearby tissues and organs, and in some cases, they can spread to other parts of the body through the bloodstream or lymphatic system, a process known as metastasis [2]. Certain cancers like breast and prostate cancer are slow growing with little to no symptoms whereas cancers like pancreatic cancer are aggressive and quick to spread to other parts of the body.

According to estimates from the World Health Organization (WHO) in 2019, cancer is the first or second leading cause of death before the age of 70 years in 112 of 183 countries and ranks third or fourth in a further 23 countries [3]. It is also estimated to be responsible for about 10 million deaths annually. Additional statistics about the worldwide impact of cancer are shown in Figure 1.1. Among the most common types of cancer, breast, and colorectal rank among the top five prevalent types [4]. Both breast and colon cancer have a significant impact on patient quality of life, and healthcare costs, making them important research areas for cancer prevention, early detection, and treatment.

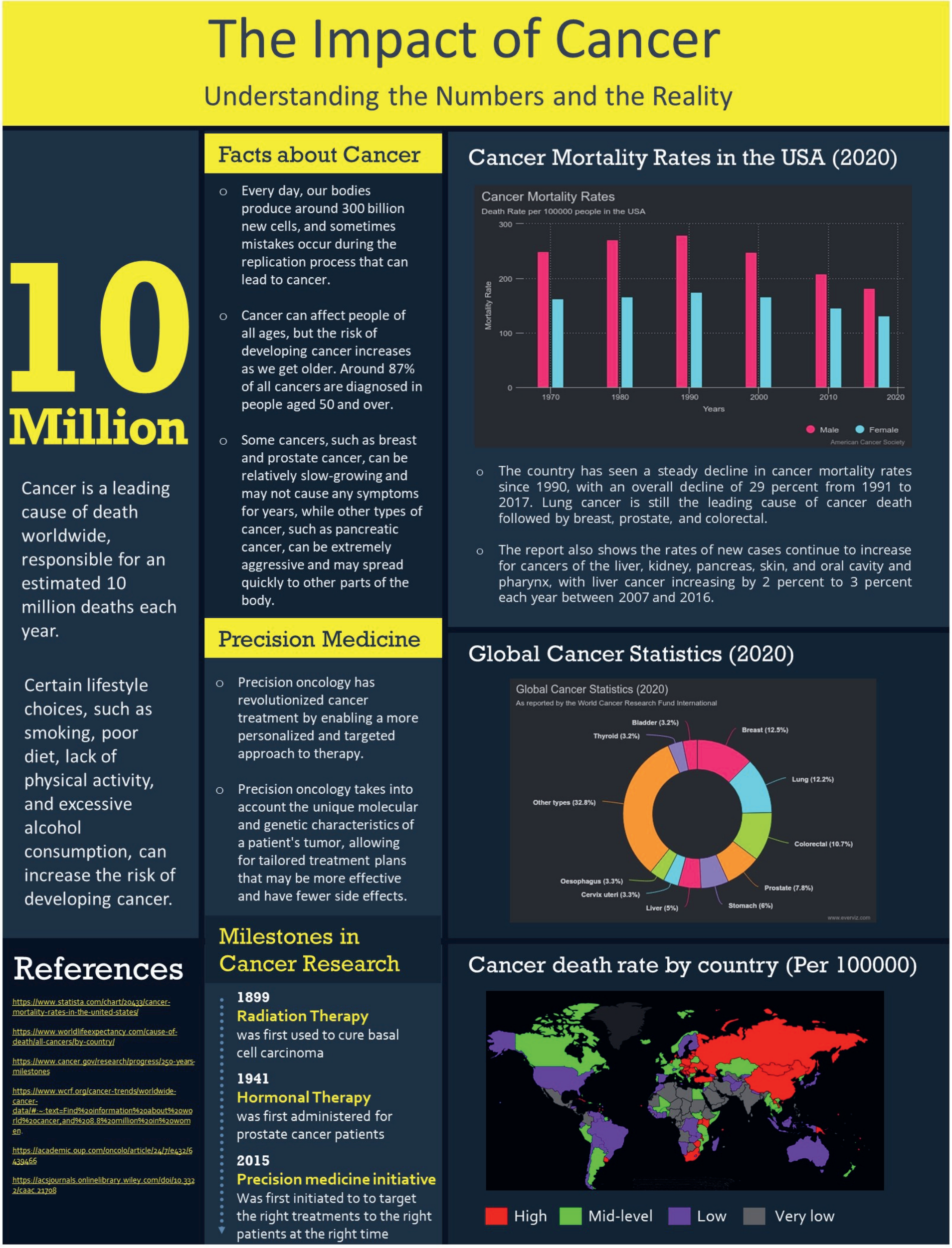
The causes of cancer are complex and can include genetic mutations, exposure to environmental toxins, and lifestyle factors such as tobacco use or diet [5]. Smoking and alcohol consumption are the leading risk factors worldwide associated with death from cancer [6]. In addition, some viruses, such as human papilloma (HPV) [7], hepatitis B and C, and human immunodeficiency (HIV) viruses, can increase the risk of developing certain types of cancer [8].

Cancer screening and detection methods are crucial in identifying cancer early when treatment is more effective [9]. Some common methods include biopsies, imaging and blood tests, physical exams, and genetic testing. All clinical screening tests contain information on key performance indicators like sensitivity, specificity, safety, cost, simplicity, and patient and clinician acceptability [10, 11] and evaluating these tests depends on several methodological, clinical, and ethical factors [12].

If the screening and detection methods confirm the presence of cancer, a healthcare team determines the specific type, stage, and extent of cancer, and a comprehensive treatment plan is developed. This plan is tailored to the individual's specific cancer type, stage, overall health, and personal preferences.

1.2.1 Breast Cancer

Breast cancer is a heterogeneous and multifaceted disease with a range of clinical, pathological, and molecular features. Breast cancer remains among the most prevalent types of tumor among women, where 1 in 8 women are diagnosed with breast cancer in their lifetime [13, 14]. However, breast cancer also has low mortality and



The Impact of Cancer

Understanding the Numbers and the Reality

10 Million

Cancer is a leading cause of death worldwide, responsible for an estimated 10 million deaths each year.

Certain lifestyle choices, such as smoking, poor diet, lack of physical activity, and excessive alcohol consumption, can increase the risk of developing cancer.

Facts about Cancer

- Every day, our bodies produce around 300 billion new cells, and sometimes mistakes occur during the replication process that can lead to cancer.
- Cancer can affect people of all ages, but the risk of developing cancer increases as we get older. Around 87% of all cancers are diagnosed in people aged 50 and over.
- Some cancers, such as breast and prostate cancer, can be relatively slow-growing and may not cause any symptoms for years, while other types of cancer, such as pancreatic cancer, can be extremely aggressive and may spread quickly to other parts of the body.

Precision Medicine

- Precision oncology has revolutionized cancer treatment by enabling a more personalized and targeted approach to therapy.
- Precision oncology takes into account the unique molecular and genetic characteristics of a patient's tumor, allowing for tailored treatment plans that may be more effective and have fewer side effects.

Milestones in Cancer Research

1899
Radiation Therapy
was first used to cure basal cell carcinoma

1941
Hormonal Therapy
was first administered for prostate cancer patients

2015
Precision medicine initiative
Was first initiated to to target the right treatments to the right patients at the right time

Cancer Mortality Rates in the USA (2020)

- The country has seen a steady decline in cancer mortality rates since 1990, with an overall decline of 29 percent from 1991 to 2017. Lung cancer is still the leading cause of cancer death followed by breast, prostate, and colorectal.
- The report also shows the rates of new cases continue to increase for cancers of the liver, kidney, pancreas, skin, and oral cavity and pharynx, with liver cancer increasing by 2 percent to 3 percent each year between 2007 and 2016.

Global Cancer Statistics (2020)

Cancer death rate by country (Per 100000)

🟥 High 🟩 Mid-level 🟪 Low ⬛ Very low

References

https://www.statista.com/chart/20433/cancer-mortality-rates-in-the-united-states/

https://www.worldlifeexpectancy.com/cause-of-death/all-cancers/by-country/

https://www.cancer.gov/research/progress/250-years-milestones

https://www.wcrf.org/cancer-trends/worldwide-cancer-data/#:~:text=Find%20information%20about%20world%20cancer,and%208.8%20million%20in%20women.

https://academic.oup.com/oncolo/article/24/7/e432/6439466

https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21708

Figure 1.1: A summary of cancer statistics and facts.

recurrence rates [15]. These low rates can be attributed to the early diagnosis and a plethora of treatment options post-diagnosis. Inherited genetic mutations such as BRCA1 and BRCA2 gene mutations can increase the risk of developing breast cancer [16]. Hormonal imbalances such as increased levels of estrogen and progesterone can also increase the risk. Several methods have been developed to classify breast cancer into multiple subtypes [17, 18], namely the clinical classification and molecular subtyping of breast tumors.

Clinical subtype classification of breast cancer is based on factors such as hormone receptor (HR) status, Human epidermal growth factor receptor (HER2) status, and tumor grade. The main clinical subtypes include hormone receptor-positive (HR+), HER2-positive, and triple-negative breast cancer (TNBC). Clinical subtyping in breast cancer relies on Immunohistochemistry (IHC) and Fluorescence In Situ Hybridization (FISH) tests to determine the status of receptors like Estrogen (ER), Progesterone (PR), and Human Epidermal Growth Factor Receptor (HER2) [19, 20]. IHC detects extra and intra-cellular proteins in the tissue samples using antibodies, which are labeled with a dye or radioactive material, to detect and visualize the location of the protein. FISH is a cytogenic method used to determine the presence of specific genetic alterations, such as amplifications or deletions, in genes associated with aggressive or treatment-resistant forms of the disease.

In addition, tumors are classified based on their molecular characteristics such as gene expression, mutations, and other biomarkers [21]. These subtypes have different prognoses and may require tailored treatment approaches. Molecular subtypes of breast cancer include luminal subtypes which are hormone receptor-positive and have a better prognosis, while HER2-enriched tumors show high HER2 expression and require targeted therapies. Triple-negative breast cancer lacks hormone receptors and HER2 expression, making it more aggressive. A detailed explanation of the molecular subtypes of breast cancer is discussed below in Section 1.3.1.

The diagnosis and treatment of breast cancer involve a well-defined workflow, shown in Figure 1.2. The process typically begins with screening methods such as mammography, which can detect early signs of cancerous growth. In cases where an abnormality is detected, further diagnostic tests, such as Computed Tomography (CT) scans and biopsies, are conducted to confirm the presence of cancer and determine its specific characteristics. Upon diagnosis, a multidisciplinary approach is employed to determine the most appropriate treatment strategy for each patient. Treatment options may include surgery to remove the tumor, followed by additional therapies such as radiation therapy to target any remaining cancer cells. Systemic treatments, including chemotherapy, hormonal therapy, and targeted therapy, may also be administered based on the individual's tumor subtype and specific molecular characteristics.

Hormone therapy is commonly used for hormone receptor-positive subtypes, while HER2-targeted therapies are effective for HER2-positive subtypes. Additionally, chemotherapy may be recommended for certain subtypes to target rapidly dividing cancer cells. The hormonal therapy includes selective estrogen receptor modulators (e.g., tamoxifen) [22], aromatase inhibitors (e.g., anastrozole, letrozole), and ovarian function suppression [23, 24]. These therapies aim to block the estrogen signaling pathway and reduce the growth of hormone-sensitive tumors. Targeted therapies such as trastuzumab [25], pertuzumab [26], and ado-trastuzumab emtansine (T-DM1) [27] specifically inhibit HER2 signaling and have significantly improved outcomes for HER2-positive breast cancer patients. These targeted therapies are often combined with chemotherapy to enhance their effectiveness. Treatment options for TNBC primarily involve chemotherapy, which may be given before surgery (neoadjuvant) or after surgery (adjuvant) [28]. Emerging therapies, such as immune checkpoint inhibitors (e.g., pembrolizumab), are showing promise in subsets of TNBC patients with high levels of immune cell infiltration [29].

1.2.2 Colorectal Cancer

Colorectal cancer (CRC) is the third most common cancer worldwide and the second leading cause of cancer death in the United States. It is estimated that about 1 in 20 people will develop colon cancer in their lifetime [30, 31]. The 5-year relative survival rate for people with localized colon cancer is approximately 90%. In women, colorectal cancers account for 13% of all new cancers and are the second most frequent tumors after those of the breast. Similar to breast cancer, the mortality rates of colon cancer have also declined rapidly which is thought to be a result of CRC prevention and earlier diagnosis through screening as well as the reduced prevalence of risk factors, and/or availability of improved treatment regimens [32, 33]. These risk factors that have reduced the prevalence of colon cancer include the decreased consumption of red and processed meats, increased awareness of the importance of physical activity and maintaining a healthy weight, smoking cessation efforts, and improved screening practices such as regular colonoscopies and stool-based tests [34].

Some of the regular screening tests such as colonoscopy, fecal occult blood test (FOBT), stool DNA test, flexible sigmoidoscopy, or virtual colonoscopy can detect colon cancer or precancerous polyps early before symptoms develop [35].

Once diagnosed, the next step is to determine the stage of the cancer, which helps guide treatment decisions. Staging involves assessing the size and extent of the tumor [36, 37], as well as the presence of any spread to nearby lymph nodes or distant organs [38, 39]. Treatment options for colon cancer depend on the stage and individual factors. The primary treatment modalities include surgery, chemotherapy, and radiation therapy. Surgery is often the first-line treatment, aiming to remove the tumor and nearby lymph nodes [40]. Chemotherapy may be given before or after surgery to shrink the tumor or to eliminate any remaining cancer cells. Radiation therapy may be used in certain cases to target specific areas [41]. Additionally, targeted therapies and immunotherapy may be utilized in specific situations [42, 43]. Often, a combination of treatments may be used to achieve the best results [44, 45].

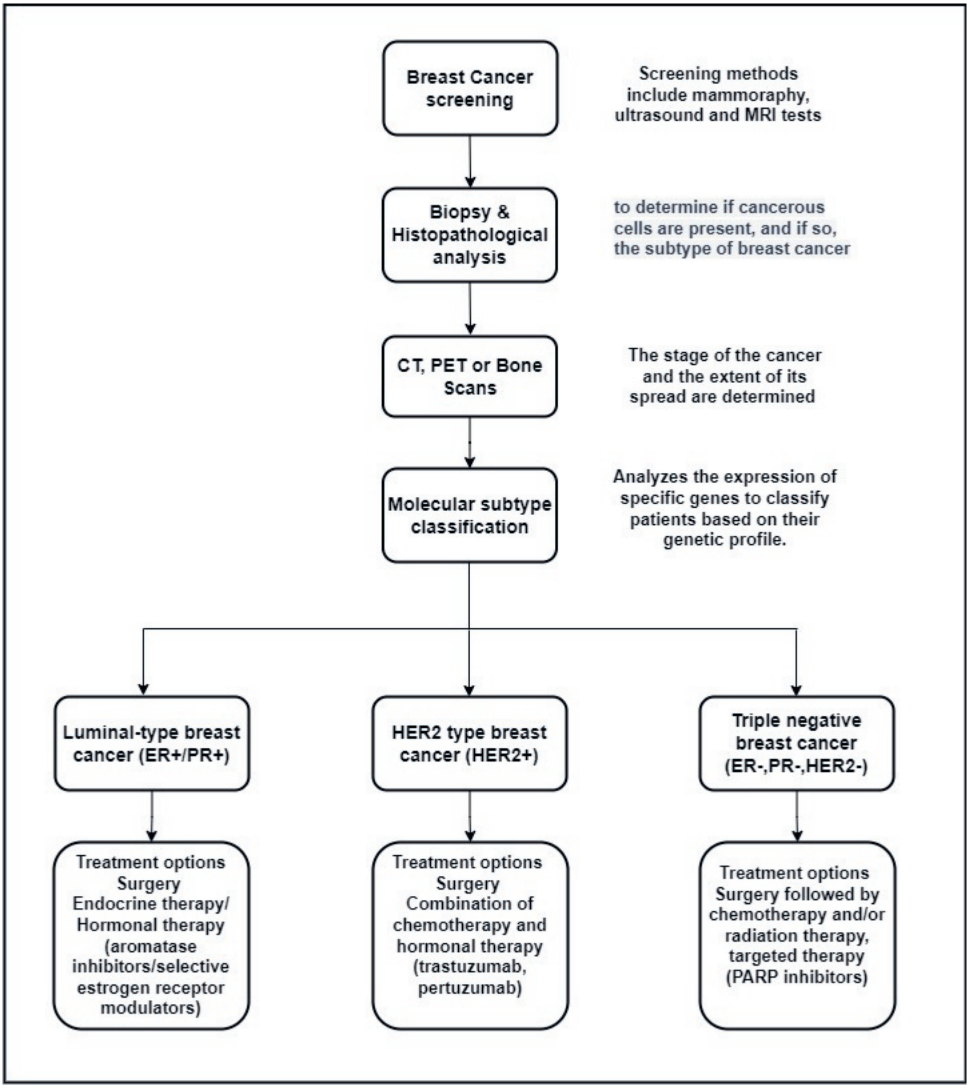

Breast Cancer screening
Screening methods include mammoraphy, ultrasound and MRI tests
Biopsy & Histopathological analysis
to determine if cancerous cells are present, and if so, the subtype of breast cancer
CT, PET or Bone Scans
The stage of the cancer and the extent of its spread are determined
Molecular subtype classification
Analyzes the expression of specific genes to classify patients based on their genetic profile.
Luminal-type breast cancer (ER+/PR+)
HER2 type breast cancer (HER2+)
Triple negative breast cancer (ER-,PR-,HER2-)
Treatment options Surgery Endocrine therapy/ Hormonal therapy (aromatase inhibitors/selective estrogen receptor modulators)
Treatment options Surgery Combination of chemotherapy and hormonal therapy (trastuzumab, pertuzumab)
Treatment options Surgery followed by chemotherapy and/or radiation therapy, targeted therapy (PARP inhibitors)

Figure 1.2: Workflow for Diagnosis and Treatment of Breast Cancer Based on Molecular Subtypes, CT - Computed Tomography, PET - Positron emission tomography, PARP - Poly (ADP-ribose) polymerases

Regular follow-up visits, imaging tests, and blood work are important for monitoring the response to treatment and detecting any signs of recurrence.

1.2.3 Transforming Cancer Treatment: Innovations in Research and Technology Driven Progress

In the last fifty years of clinical and medical research and trials, cancer treatment has come a long way, with significant advancements leading to new and improved treatments. Some of the improvements include:

- Early detection: Advances in technology and screening methods have made it possible to detect many types of cancer at an earlier stage when they are more treatable. Many studies have shown that a delay of more than three months between the discovery of symptoms and the start of cancer treatment is associated with an advanced clinical stage of the disease [46–48]. Earlier diagnosis and optimal treatment can lead to many thousands of patients being spared their life being cut short by cancer [49].

- Immunotherapy: This treatment harnesses the body's immune system to fight cancer, and has shown promising results in some types of cancer, including melanoma, leukemia, and lymphoma. Immunotherapy is emerging as an addition to conventional therapies [50]. Immune checkpoint blockade therapy, in particular, is one of the most impressive advancements in cancer therapeutics in recent years [51].

- Minimally invasive surgical techniques: Surgery techniques like laparoscopic or robotic surgery, result in a quicker recovery and less pain for patients. Minimally invasive surgery shows benefits when compared to open surgery, like better recurrence-free survival [52]. Alongside, techniques aiding minimally invasive surgery techniques such as three-dimensional vision, instruments' articulation, and greater ergonomics for the surgeon, offer a better therapeutic approach to the minimally invasive treatment [53].

- Personalized medicine: This is an innovative approach to tailoring disease prevention and treatment that takes into account differences in people's genes, environments, and lifestyles. The goal of precision medicine is to target the right treatments to the right patients at the right time. Rather than relying on a one-size-fits-all approach, precision medicine takes into account an individual's unique genetic profile, lifestyle factors, and disease characteristics to determine the most effective treatment options. This approach minimizes unnecessary treatments and reduces the risk of adverse side effects, maximizing the benefits for patients while minimizing the burden of ineffective interventions. Moreover, it has also shifted the focus of medicine from reactive to preventative measures [54, 55]. By identifying genetic predispositions or biomarkers

> associated with certain diseases, individuals at high risk can be identified and proactive measures can be taken to prevent or mitigate the development of the disease. Furthermore, precision medicine has a profound impact on research and drug development. By dissecting diseases at a molecular level and understanding the underlying mechanisms, researchers can discover novel targets for drug development and design therapies that are more precise and effective. This personalized approach to drug development holds immense promise for addressing complex diseases and unlocking innovative treatment options. In summary, precision medicine represents a paradigm shift in healthcare, offering personalized, targeted, and effective interventions. It holds the potential to transform patient care, optimize treatment outcomes, and pave the way for a future where healthcare is tailored to each individual's unique characteristics and needs.

Amidst the remarkable progress in cancer research, precision medicine emerges as a standout approach with immense potential for the future. This personalized approach holds the key to unlocking breakthroughs in cancer treatment, paving the way for more effective therapies and improved patient outcomes.

1.3 Precision matters: The Importance of Moving beyond Clinical subtypes to Molecular Characterization

The purpose of performing tumor subtype classification is to determine the optimal treatment plan based on tumor biology. While clinical subtypes can provide information about the underlying biology of a tumor, they have limited value in predicting patient outcomes and response to treatment. Another important question in the clinical management of tumors is how to avoid overtreatment or undertreatment of patients.

In addition, clinical subtyping relies on subjective interpretations of pathology and clinical features, leading to variability in classification among different clinicians and institutions. To overcome the limitations of clinical subtyping, a concerted effort has been to develop new classification models based on gene expression profiling reflecting the functionality of molecular pathways [56, 57].

Over the past few years, progress in cancer genomics research has established how genetic altercations can serve as effective biomarkers for early detection, monitoring, and prognosis of cancer. Through years of advancements in the field of bioinformatics and biotechnology, cancer gene profiling has expanded from a single gene to the whole genome through genome-wide association studies (GWAS). This has helped in defining the landscape of early-stage breast cancer, which is shown to be highly heterogeneous in breast tumors, driven by distinct molecular alterations

[58–60].

1.3.1 Precise Diagnostics for Personalised Treatment

The IHC and FISH tests that estimate the clinical subtypes of breast cancer may not always provide a complete picture of cancer's molecular subtype and genetic alterations, which can impact the choice of treatment. This is because they are limited by their ability to analyze only a few genes or proteins at a time. Hence, newer tools for determining breast cancer biology can help in individualizing the treatment for patients, where low-risk cancer patients can be approached with less-aggressive treatments, preventing unnecessary toxicity and high-risk cancers treated in the appropriate way.

The molecular subtypes of breast cancer are identified using gene expression profiling or molecular assays which analyze the activity of a set of genes in a tumor sample to determine which subtype the cancer belongs to. Table 1.1 presents an overview of the molecular subtypes of breast cancer, including their etiology, prevalence, and associated genetic alterations.

Table 1.1: Overview of the molecular subtypes of breast cancer, their causes, prevalence, genetic mutations, specific biomarkers, prognosis, and possible treatments. HER2 - Human epidermal growth factor receptor

Molecular subtype	Causes	Prevalence	Specific Biomarkers	Prognosis	Possible Treatments
Luminal	Hormonal imbalance, age, obesity	40%	Ki-67 Low, HER2 negative	Favourable	Hormonal Therapy, Chemotherapy, Targeted therapy
HER2	HER2 Gene Amplification	10-20%	HER2 Positive	Less Favourable	Chemotherapy, Targeted therapy
Triple-negative/Basal-like	Gene mutations (e.g. BRCA1)	10-20%	CK5/6 and/or EGFR Positive	Poor	Chemotherapy, Immunotherapy, PARP inhibitors

Luminal subtypes of tumors are associated with the luminal epithelial cells of the breast and are hormone receptor-positive, meaning they express estrogen receptor (ER) and/or progesterone receptor (PR) [61]. They are called "luminal" because they arise from the cells lining the milk ducts (lumina) of the breast. These tumors tend to grow slowly and are usually less aggressive than other types of breast cancer [62]. These tumors have the best prognosis [63]. These tumors can be treated with hormone therapies, which block the effects of estrogen and progesterone in the body. This includes selective ER modulators (SERMs), aromatase inhibitors, and ER downregulators. Among luminal subtypes, there are two subgroups, namely Luminal A tumors which are a subtype of Luminal tumors that have a lower proliferation rate and a better prognosis. Luminal B tumors, on the other hand, have a higher proliferation rate and are more aggressive. They may be less responsive to hormone therapy and more likely to recur [64].

HER2-positive breast cancer, also known as HER2-amplified, tests positive for

HER2 [65]. HER2 overexpression plays an important role in sustaining several pathways for tumor growth and they occur in about 10-20% of patients and is generally associated with a poor prognosis [66] HER2-positive tumors are typically more aggressive and fast-growing than other types of breast cancer. However, they can be effectively treated with targeted therapies such as trastuzumab (Herceptin), pertuzumab (Perjeta), and ado-trastuzumab emtansine (Kadcyla). HER2-positive breast cancer may also be treated with surgery, chemotherapy, and radiation therapy.

Triple-negative breast cancer (TNBC) is a subtype of breast cancer that lacks the expression of ER, PR, and HER2 [67]. It is also named "basal-like" because the tumor cells resemble the basal cells that line the mammary ducts. It is an aggressive form of breast cancer that tends to grow and spread more quickly than other subtypes [68]. It is also more likely to recur after treatment and has a poorer prognosis. Treatment options include chemotherapy, surgery, and radiation therapy [69]. Because TNBC does not express the three common receptors targeted by hormonal therapy or targeted therapy, such as HER2-targeted therapy, treatment options are limited [17].

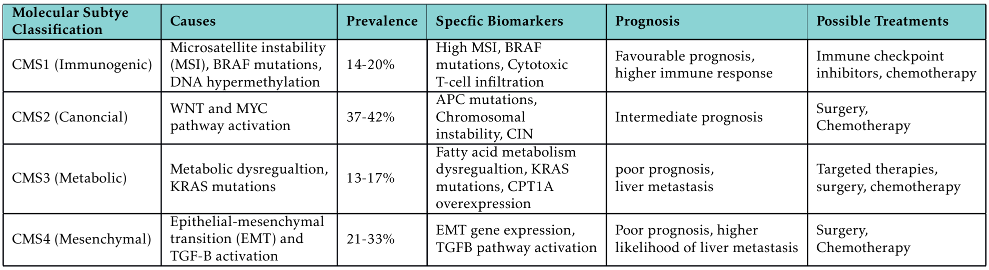
1.3.2 Classification of Colon Cancer: Unraveling Clinical and Molecular Subtypes

Very similar to breast cancer, colon cancer is also a lethal disease with heterogeneous outcomes and drug responses. Based on gene expression patterns, colorectal cancers are classified into subtypes based on the activation of different biological processes [70]. Colon cancer classification was primarily based on the histopathological appearance of the tumor tissue, which involves the examination of tissue samples under a microscope to identify the features of cancer cells, such as size, shape, growth pattern, glandular architecture, or degree of differentiation [71]. This method of classification is still used today and is an essential part of the diagnosis of colon cancer.

Colon cancer is classified into five **clinical subtypes**. Adenocarcinoma is the most common subtype, accounting for approximately 70-80% of colon cancer cases. Mucinous adenocarcinoma is characterized by the presence of extracellular mucin, secreted by the cancer cells [72]. Signet-ring cell carcinoma is a rare subtype characterized by the presence of cancer cells with a large, central mucin-filled vacuole that pushes the nucleus to the periphery [73]. Squamous cell carcinoma is a rare subtype that arises from the squamous cells that line the colon. Undifferentiated carcinoma is a subtype characterized by a lack of differentiation, making it difficult to identify the tissue of origin [74]. Each subtype has different clinical and pathological characteristics, which can impact treatment decisions and patient outcomes.

Colon cancers are classified into four types based on a classification system developed based on **molecular features**. This is known as the Consensus Molecular Subtype Classification (CMS) developed by the Consensus Molecular Subtype Consortium [75]. In addition, colorectal cancer also has an integrated molecular subtype classification system developed by The Cancer Genome Atlas, which classifies them into three major types, namely hypermutated, ultra-mutated, and Chromosomal Instability (CIN) [76]. An overview of the CMS molecular subtypes is shown in Table 1.2.

Table 1.2: Overview of Consensus Molecular Subtype Classification of Colon cancer, their causes, prevalence, specific biomarkers, prognosis, and possible treatments. CMS - Consensus Molecular Subtype, MSI - Microsatellite instability, CIN - Chromosomal instability



Molecular Subtye Classification	Causes	Prevalence	Specfic Biomarkers	Prognosis	Possible Treatments
CMS1 (Immunogenic)	Microsatellite instability (MSI), BRAF mutations, DNA hypermethylation	14-20%	High MSI, BRAF mutations, Cytotoxic T-cell infiltration	Favourable prognosis, higher immune response	Immune checkpoint inhibitors, chemotherapy
CMS2 (Canoncial)	WNT and MYC pathway activation	37-42%	APC mutations, Chromosomal instability, CIN	Intermediate prognosis	Surgery, Chemotherapy
CMS3 (Metabolic)	Metabolic dysregualtion, KRAS mutations	13-17%	Fatty acid metabolism dysregualtion, KRAS mutations, CPT1A overexpression	poor prognosis, liver metastasis	Targeted therapies, surgery, chemotherapy
CMS4 (Mesenchymal)	Epithelial-mesenchymal transition (EMT) and TGF-B activation	21-33%	EMT gene expression, TGFB pathway activation	Poor prognosis, higher likelihood of liver metastasis	Surgery, Chemotherapy

CMS1 tumors are characterized by high levels of immune activation and are associated with a good prognosis [77]. They are often referred to as "immune" tumors and are found in about 14-20% of colorectal tumors [78, 79]. It has a high frequency of BRAF mutations and microsatellite instability (MSI) [80]. Patients with CMS1 tumors may benefit from immunotherapy [81]. CMS2-type tumors have high copy number gains in oncogenes and copy number losses in tumor suppressor genes [76]. CMS2 tumors display epithelial differentiation and strong upregulation of WNT and MYC downstream targets, both of which have classically been implicated in CRC carcinogenesis [82, 83]. Approximately 39% of tumors of CMS2 tumors are stage III at the time of diagnosis, and standard adjuvant chemotherapy is recommended [75]. CMS3, also known as the metabolic subtype, has genomic features consistent with CIN but has relatively low somatic copy-number alterations (SCNAs) compared with CMS2 or 4. CMS3 also had more MSI than CMS2 and 4 (CIMP-low, intermediate hypermethylation). Approximately 30% CMS3 tumors are considered hypermutated (less common than CMS1 tumors, but more than CMS2 or 4 type tumors). Although KRAS mutants were present in every molecular subtype, they were more prevalent among CMS3 CRC (68%). [84]. CMS4 tumors exhibit low levels of hypermutation, MSS status, and very high SCNA counts. They display a mesenchymal phenotype as well [85]. CMS4 tumors are diagnosed at very advanced stages, with a very poor prognosis. These cancers have also been shown not to show any benefit from systemic adjuvant treatments [75].

1.4 The Benefits of Precision Oncology in Fighting Cancer

Traditional approaches to cancer treatment are not effective in all cases, as new information on tumor biology and treatment response becomes available it can be utilized to develop new strategies. In addition, some of these tumors are also increasingly becoming resistant to traditional treatments, which means new and more effective approaches are needed. Precision medicine considers the complexity of biological systems and the interactions between different factors. Precision medicine can help to develop more personalized and effective treatments that take into account the unique characteristics of each patient [86, 87].

Progress in sequencing technologies, which have evolved from single-gene sequencing to whole-exome, whole-genome, and whole-transcriptome sequencing, has produced a substantial amount of valuable information to boost cancer precision medicine [88]. Currently, several multigene assays have been developed to stratify patients according to their risk of relapse or to perform molecular subtype classification. This approach has led to the discovery and creation of targeted therapeutic agents in clinical trials.

Precision medicine continually evolves as new technologies and scientific discoveries emerge. This requires improving methods for analyzing genetic and other patient data to provide more accurate diagnoses and effective treatments. The development of targeted treatment drugs and therapies based on a patient's unique genetic and medical profiles is also constantly evolving. Therefore, precision medicine must identify and stratify patients well-suited to newly developed targeted treatment therapies and drugs. Continual refinement and advancement of precision medicine can lead to better patient health outcomes.

The most seminal work on breast cancer molecular subtype classification was performed by Perou *et al.* [89], who used DNA microarray-based gene expression profiling to classify five subtypes of breast cancer that are reproducible across patient populations and laboratories [89]. Sorlie *et.al.* built on this work and developed a 50-gene molecular classifier called the PAM50 [90]. The PAM50 test is used to classify breast cancer into the four subtypes shown in Table 1.1: Luminal A, Luminal B, HER2-enriched, and Basal-like. The PAM50 test is typically performed on a small amount of breast tumor tissue collected during a biopsy or surgery [91]. RNA is isolated from the tumor tissue and hybridized into a custom-designed microarray containing probes for the 50 genes. The Prediction Analysis of Microarrays (PAM) algorithm is a centroid-based classification method [92]. For each molecular subtype, the mean expression values for each gene across all samples are calculated in that subtype. These mean expression values are referred to as the centroids. The distance between each new sample and the centroid of each class is then determined. The sample is then assigned to the molecular subtype with the smallest distance. The PAM50 test has been shown to be a more accurate predictor of breast cancer prognosis than traditional clinical factors, such as tumor size, grade, and lymph node status [93]. In addition, the PAM50 test has been shown to predict the likelihood of response to specific types of chemotherapy and to guide treatment decisions in breast cancer patients [94].

OncotypeDX is a genomic test that analyzes the activity of certain genes within tumor tissue to help determine the risk of recurrence in early-stage breast cancer patients. The test measures the expression of 21 genes within the tumor tissue, including genes associated with cell proliferation, hormone receptors, and HER2 expression. [95]. The OncotypeDx assay was developed by selecting 250 candidate genes from the literature and quantifying their expression by reverse transcriptase polymerase chain reaction (RT-PCR) using mRNA extracted from formalin-fixed paraffin-embedded tumors of 447 patients from the NSABP B-20 study [96]. 16 genes were selected based on their statistical association with breast cancer recurrence. These 16 genes were combined with five reference "housekeeping" genes to produce the recurrence score (RS) which ranges from 0-100. The cutoff points for RS were prespecified into low-risk (<18), intermediate-risk (18-30), and high-risk (geq31) based on the NSABP B-20 study. Patients with a low recurrence score have a low risk of recurrence and may not benefit from chemotherapy, while patients with a high recurrence score have a higher risk of recurrence and may benefit from chemotherapy [95].

Another important test for breast cancer molecular subtyping is the BuePrint 80-gene assay [97]. The BluePrint test was developed using IHC-based clinical subtyping as a guide. The BluePrint 80-gene signature was developed using a cohort of 200 samples with concordant ER, PR, and HER2 status [97]. The study used a threefold cross-validation (CV) procedure to identify the genes that best discriminate between the three molecular subtypes. Within each CV iteration, they performed two-sample Welch t-tests on a randomly selected set of 133 of the 200 training samples to score all genes for their differential expression among the three classes. Genes were ranked according to their absolute t-statistics, and the threefold CV procedure was repeated a hundred times. Next, the 100 gene ranking scores were combined into a single ranking per gene, and the minimal number of genes with optimal performance was determined using a leave-one-out CV on all 200 training samples [98]. The optimal performance was achieved with a total of 80 unique genes. A centroid classification model was then built using the 80-gene profile [99]. The BluePrint test classifies breast cancer patients by measuring the similarity of the tumor to a Luminal-type (58 genes), Basal-type (28 genes), and HER2-type (4 genes) representative profile [100].

Precision medicine has revolutionized the treatment approaches to colon cancer as well, offering personalized strategies based on the molecular profile of the tumor. An example of one of the most notable molecular subtyping tests in colon cancer is the Oncotype DX Colon Cancer assay which is a genomic test used to assess the risk of recurrence in patients with stage II colon cancer [101]. This test is very similar to the Oncotype Dx breast cancer assay. This assay examines the expression levels of 12 specific genes involved in cancer progression, tumor invasion, and immune response. Based on the gene expression data, the Oncotype DX Colon Cancer assay calculates a Recurrence Score, which is a number between 0 and 100. The Recurrence Score is derived using a validated algorithm that combines the gene expression data with clinicopathological factors such as tumor stage and grade. The algorithm takes into account the relative contribution of each gene to the overall risk of recurrence. Patients with a low Recurrence Score may have a lower risk of recurrence and may not benefit significantly from adjuvant chemotherapy, allowing for potentially more personalized treatment plans. On the other hand, patients with a high Recurrence Score may have a higher risk of recurrence and may benefit from more aggressive treatment approaches, such as adjuvant chemotherapy [101].

Similarly, FoundationOne CDx is a comprehensive genomic profiling test that uses next-generation sequencing technology to analyze the DNA from a patient's tumor sample [102]. The test examines a broad panel of genes, including both well-known cancer-related genes and emerging targets. The test provides a detailed report that includes information on specific genetic alterations identified in the tumor sample. The report categorizes these alterations based on their clinical significance and relevance to available targeted therapies or clinical trials. The results may include information about potential treatment options, such as targeted therapies, immunotherapies, or clinical trials, which may be tailored to the molecular characteristics of the patient's tumor [103].

These commercial precision medicine tools and tests have significantly contributed to the individualized treatment approach in colon cancer, allowing oncologists to make more informed decisions based on the unique molecular characteristics of each patient's tumor.

In the upcoming sections, I will delve into the various objectives of our thesis, shedding light on the underlying motivations that propelled this research topic into focus.

1.4.1 Capturing the Complexities of Breast Cancer Molecular Subtyping and Dual Subtypes

Molecular subtyping in breast cancer is a critical aspect of cancer diagnosis and treatment, as it can help identify specific subtypes of cancer that have unique molecular characteristics and clinical behaviors. However, when a tumor possesses characteristics of more than one molecular subtype, it may open up the opportunity of providing a combination of treatment strategies to target both subtypes. Since 2010, there have been several studies published on the identification and features of dual molecular subtypes. In a study published in the Journal of the National Cancer Institute, researchers analyzed the molecular subtypes of 1492 breast cancer samples. They identified 26% of the samples as having multiple molecular subtypes.

They found that mixed molecular subtypes were associated with higher tumor grade, larger tumor size, and worse overall survival [17].

In another recent study published in the Journal of Clinical Oncology, researchers analyzed the genomic profiles of breast tumors to identify patients with mixed ER-positive and ER-negative subtypes [104]. They found that these patients had a worse prognosis compared to those with a single ER-positive or ER-negative subtype, suggesting that identifying mixed subtypes within ER-positive or ER-negative tumors is important for predicting patient outcomes and developing treatment strategies. While the existing subtypes of breast cancer (Luminal A/B, HER2-enriched, and Basal-like) have been useful in guiding treatment decisions, there has been a recognition that these subtypes are not always sufficient to capture the complexity of breast cancer. As a result, there have been efforts to refine these subtypes into single and dual subtypes.

In my thesis, I assessed the differences between molecular subtyping scores to help identify patients that belong to more than one molecular subtype which were previously not detected by standard molecular subtyping tests. I aimed to gain a deeper understanding of which pathways are activated in the single and the dual subtypes of breast tumors which may help to understand the specific biology of dual subtypes that distinguish them from the single subtypes. This study was performed on breast cancer microarray data as well as Proteomics data since microarray and proteomics data play crucial roles in precision medicine by providing comprehensive molecular profiles, identifying biomarkers, guiding treatment selection, predicting treatment response, and enhancing our understanding of disease heterogeneity.

1.4.2 HER2-Positive Breast Cancer: Unraveling the Complexity of Tumor Heterogeneity

One of the methods of improving precision medicine is the identification of new biomarkers and signature genes. HER2-positive breast cancer is known for its heterogeneity, which means that the tumor may have different genetic, molecular, and cellular features within the same patient. This heterogeneity can occur due to the acquisition of different genetic mutations and alterations during tumor growth and progression, leading to diverse subpopulations of tumor cells with different phenotypes, genotypes, and responses to treatment [105]. In HER2-positive breast cancer, this heterogeneity can have important clinical implications, as some tumoral cell subpopulations may be more resistant to certain treatments, while others may respond better. For example, some HER2-positive tumors may have high levels of a protein called PTEN, which is associated with better response to certain treatments, while others may have low levels of PTEN [106]. Heterogeneity can also affect the accuracy of HER2 testing, as different regions of the tumor may have different levels of HER2 protein expression or HER2 gene amplification, leading to discordant results between different tests or between the primary tumor and metastatic sites.

Therefore, understanding and addressing the heterogeneity in HER2-positive breast cancer is crucial for improving the precision and effectiveness of treatment and avoiding potential overtreatment or undertreatment. Hence, in my thesis, I focused on the identification of novel (potential) signature genes for HER2-positive patients which captures the latest tumor biology of the HER2 breast cancer.

1.4.3 Concordance between Primary and Metastatic Tumors

Primary and metastatic tumors may have different molecular and genetic features due to the evolutionary process of cancer. This heterogeneity can affect the treatment response and prognosis of the patient. Therefore, it is important to analyze both the primary and metastatic tumors to determine the most effective treatment approach for the patient. Concordance between primary and metastatic tumors is an important factor to consider in precision medicine. For example, if the primary tumor is HER2-negative but the metastatic tumor is HER2-positive, treatment with HER2-targeted therapy may be appropriate. This phenomenon has been observed in breast cancer. However, until now, no systematic studies have investigated the concordance of primary and metastatic tumors in colon cancer. Very few studies have reported some similarities between the two, however, the small sample size led to inconclusive results [107, 108].

In my thesis, samples were collected with the support of the IntraColor consortium and systematically analyzed the concordance between the primary and metastatic tumors in the case of colon cancer patients.

1.5 Transforming Cancer Diagnostics: Improving Lives, Strengthening Communities

When diagnosed early, patients have a better chance of full recovery and reduced risk of metastasis. In many cases in both breast and colon cancers, an early diagnosis can also result in a cure. This is very commonly seen in patients who are Luminal-type (ER+) in breast cancer and in CMS1-type in colon cancer patients [109, 110].

Another large societal impact of improving cancer diagnostics is the possibility of reducing the need for more expensive and invasive treatments. This can result in lower healthcare costs for patients and the healthcare system as a whole. In addition, such accurate diagnoses and effective treatments greatly improve the quality of life for patients with cancer. Constant improvement in cancer diagnostics leaves very little chance for misdiagnosis. With the appropriate treatment, these cancer diagnostics can help patients maintain their quality of life and reduce the physical, emotional, and financial burden of the disease.

Improving cancer diagnostics in turn improves the lives of cancer patients, which can help communities come together in support of those affected by the disease and raise awareness about the importance of early detection and effective treatment.

1.6 Aim and Outline of the Thesis

The overarching goal of this thesis is to **Refine Molecular Subtyping Diagnostics in Breast and Colon Cancers using Gene Expression and Proteomics Data**. The specific objectives are:

1. To identify and examine dual subtyping in breast cancer tumors, overall and within a particular subgroup to understand their tumor biology and possible implications to therapeutic guidance.
2. To identify an expanded HER2 gene signature to capture the full biological diversity of HER2+ tumors
3. To assess if molecular subtyping signatures are concordant in primary and metastatic tumors in colon cancer.

This work was performed using bioinformatics and systems biology approaches to analyze "omics" datasets in both breast and colon cancers. In this thesis, I hope to contribute to the advancement of oncology and improve patient outcomes through the application of precision medicine in cancer diagnosis.

In **Chapter 2**, I demonstrate how the BluePrint test identifies a proportion of breast cancer patients that have dual-activated pathways showing characteristics of more than one BluePrint subtype. A classification threshold was developed using bootstrapping and multi-modality detection and was evaluated on the Neoadjuvant Breast Registry Symphony Trial (NBRST) dataset.

In **Chapter 3**, I expanded the BluePrint HER2 gene signature by evaluating additional genes that may capture more heterogeneity within HER2+ tumors. I expanded the signature with genes that are upregulated in pathologically confirmed HER2+ tumors, thereby capturing the modern definition of HER2+ tumors while having excellent concordance with the current HER2 gene signature.

In **Chapter 4**, I investigated the concordance of molecular characteristics in primary colon tumors and their matched liver metastasis in metastatic colon cancer patients. I explored the gene expression profiles of the matched tissue pairs concerning several molecular subtyping signatures. In addition, I also explored the tumor microenvironment in these tumor tissues to identify whether they influenced the molecular subtype classification of the tumor.

In **Chapter 5**, I explored the subgroups within the triple-negative breast cancer subtype of breast cancer using public proteogenomic datasets. This was enabled using the similarity network fusion analysis. I also validate the findings of the study using additional validation proteogenomic datasets.

In **Chapter 6**, I discuss the key findings of the thesis and the correspondences and conflicts with the existing molecular subtyping techniques in cancer diagnostics. The limitations of the thesis are also elaborated in this chapter along with the implications of this research.

Chapter 2

BluePrint Breast Cancer Molecular Subtyping Recognizes Single and Dual Subtype Tumors with Implications for Therapeutic Guidance

Midas Kuilman*, Architha Ellappalayam*, Andrei Barcaru, Josien C. Haan, Rajith Bhaskaran, Diederik Wehkamp, Andrea R. Menicucci, William M. Audeh, Lorenza Mittempergher, and Annuska M. Glas

*Authors contributed equally to this work

Breast Cancer Research and Treatment. 2022; 195(3): 263–274

Abstract

Purpose

BluePrint (BP) is an 80-gene molecular subtyping test that classifies early-stage breast cancer (EBC) into Basal, Luminal, and HER2 subtypes. In most cases, breast tumors have one dominant subtype, representative of a single activated pathway. However, some tumors show a statistically equal representation of more than one subtype, referred to as dual subtype. This study aims to identify and examine dual subtype tumors by BP to understand their biology and possible implications for treatment guidance.

Methods

The BP scores of over 15,000 tumor samples from EBC patients were analyzed, and the differences between the highest and the lowest scoring subtypes were calculated. Based upon the distribution of the differences between BP scores, a threshold was determined for each subtype to identify dual versus single subtypes.

Results

Approximately 97% of samples had one single activated BluePrint molecular subtype, whereas approximately $\sim$ 3% of samples were classified as BP dual subtype. The most frequently occurring dual subtypes were the Luminal-Basal-type and Luminal-HER2-type. Luminal-Basal-type displays a distinct biology from the Luminal single type and Basal single type. Burstein's classification of the single and dual Basal samples showed that the Luminal-Basal-type is mostly classified as 'luminal androgen receptor' and 'mesenchymal' subtypes, supporting molecular evidence of AR activation in the Luminal-Basal-type tumors. Tumors classified as Luminal-HER2-type resemble features of both Luminal-single-type and HER2-single-type. However, patients with dual Luminal-HER2-type have a lower pathological complete response after receiving HER2-targeted therapies in addition to chemotherapy in comparison with patients with a HER2-single-type.

Conclusion

This study demonstrates that BP identifies tumors with two active functional pathways (dual subtype) with specific transcriptional characteristics and highlights the added value of distinguishing BP dual from single subtypes as evidenced by distinct treatment response rates.

2.1 Introduction

Breast cancer (BC) is a heterogeneous disease with respect to clinical, histopathological, and molecular features. Based on clinical behavior and genomic characteristics, multiple methods have been utilized to categorize BC into distinct subgroups, be it with clinical subtyping for hormone receptor (HR) protein status or more recently with molecular subtyping based on RNA assays [89, 90, 97, 111].

Clinical subtyping relies on well-established immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH) staining that determines estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status [19, 20]. The BluePrint (BP) 80-gene subtyping assay was developed to bridge clinical pathology and molecular subtyping, by using IHC-based receptor status and mRNA expression, resulting in a molecular diagnostic array with predictive value [97, 100]. Each of the three subtypes determined by BP (Basal-type, Luminal-type, and HER2-type) is scored according to their respective gene signatures (consisting of 28, 58, and 4 genes, respectively) reflecting specific functional pathways, with the highest score determining the subtype [97, 100]. In most cases, the highest score is significantly higher than the score of the other two subtypes, indicating a strong dominance of a single pathway activation in the tumor (so-called single subtype). However, in rare instances, the difference between the highest score and the second-highest score is statistically indiscernible, indicating that these tumors might be characterized by multiple activated pathways (dual subtype). Having a deeper understanding of which pathways are activated may help understanding the specific biology of BP dual subtypes that distinguish them from the single subtypes.

In addition to the standard BP subtypes (Basal-type, Luminal-type, and HER2-type), other studies have identified expression-based subtypes, which include normal-like, claudin-low, triple positive, and triple-negative [89, 112–115] types. Among others, Burstein and colleagues further classified the clinical triple-negative breast cancer (TNBC) subtype into basal-like immuno-activated (BLIA), basal-like immuno suppressed (BLIS), luminal androgen receptor (LAR), and mesenchymal-like (MES) [116]. Therefore, assessing further the differences between the BP scores may help identifying additional subtypes previously not detected by standard BP. Also, understanding the biological characteristics of BP dual subtypes may help in guiding more effective treatment plans.

2.2 Materials and Methods

2.2.1 Data

For this study, only data and no samples were collected, and all patient data were fully anonymized according to the 'General Data Protection Regulation' (GDPR) and

Table 2.1: (a) The BluePrint single and dual classification of the samples with full-genome data, which were used in differential expression analysis and their standard BluePrint classification (n = 7985) and (b) the BluePrint single and dual classification of the samples for which both the clinical information and the standard BluePrint classification are available

	Full genome	**Standard BluePrint**			**Total**
		Basal	**Luminal**	**HER2**	
(a) Single-dual subtype classification					
Basal-single	712	712	0	0	712
Luminal-single	6732	0	6732	0	6732
HER2-single	277	0	0	277	277
Luminal-Basal	122	51	71	0	122
Luminal-HER2	99	0	50	49	99
HER2-Basal	23	7	0	16	23
Luminal-HER2-Basal	20	5	11	4	20
Total	**7985**	775	6864	346	7985

	Full genome	**Targeted array**	**Standard BluePrint**			**Total**
			Basal	**Luminal**	**HER2**	
(b) Single-dual subtype classification						
Basal-single	150	440	590	0	0	590
Luminal-single	1727	6781	0	8508	0	8508
HER2-single	47	145	0	0	192	192
Luminal-Basal	32	124	58	98	0	156
Luminal-HER2	11	65	0	45	31	76
HER2-Basal	5	16	6	0	15	21
Luminal-HER2-Basal	6	24	10	13	7	30
Total	1978	7595	664	8664	245	9573

the ‘Health Insurance Portability and Accountability Act’ (HIPAA) and are in compliance with the ‘Data Protection Act’. This study was a retrospective analysis of (internal) studies between 2015 and 2020. These studies included those previously described in Beumer et al. [117], the FLEX registry trial (NCT03053193), the Neoadjuvant Breast Registry Symphony Trial (NBRST) (NCT01479101), and the Multi-Institutional Neo-adjuvant Therapy MammaPrint Project (MINT) trial (NCT01501487). Most samples comply with MammaPrint (MP) eligibility criteria [118, 119], stage I,II, or operable stage III breast cancer, tumor diameter less than 5cm, positive lymph nodes, with any ER/PR/HER2 status. Microarray processing was performed following standard procedure at Agendia [100] (Supplementary Methods 2.12.1). Agendia’s customized diagnostic arrays were either a targeted array or a full genome array, as previously described [100, 117, 120].

Of the 15,580 samples analyzed with BP, 7985 had full-genome expression data available of which 1978 with clinicopathological information (Table 2.1). All samples analyzed with the targeted array had clinicopathological information available (Table 2.1b).

The Neoadjuvant Breast Registry Symphony Trial (NBRST) [121–123] classified

BC patients according to MP and BP and compared it with conventional IHC/ FISH subtyping to predict treatment sensitivity. From the entire NBRST trial dataset (n = 1060), a subset that received HER2 targeted therapy (n = 289) was used to evaluate the association between the dual subtypes and response to HER2 targeted therapy.

The NBRST trial protocol was approved by Institutional Review Boards at all participating sites (ClinicalTrials.gov NCT01479101). All patients consented to participation in the study and clinical data collection. Part of the anonymized data (BP results and IHC) used in this study was generated from early-stage BC patients collected from standard diagnostic testing and was only used to identify potential dual subtypes and not for any gene expression analysis. The data from studies can be shared by the authors upon reasonable request.

2.2.2 BluePrint Single and Dual-subtype Classification

Standard BP scores of 15,580 samples were calculated followed by dual-subtype classification, which was based on bootstrap technique [124], and multi-modality detection. Details on the procedure can be found in the Supplementary methods 2.12.1 and Figure S2.6.

2.2.3 Conventional Subtype Classification

Clinicopathological information was available for 9573 of 15,580 samples, including IHC HR status for ER and PR, Ki-67, and IHC/FISH HER2 status (Supplementary Table2.2). Tumors with at least 1% positivity for either ER or PR were classified HR-positive (HR+), otherwise HR negative (HR-). Tumors with HER2 IHC 0, 1+ or 2+ (FISH non-amplified) score were considered HER2-negative (HER2-) while tumors with HER2 IHC 2+ (FISH amplified) and 3+ score were considered HER2 positive.

2.2.4 Burstein Classification

An algorithm published by Burstein et al., stratifies TNBCs into different subtypes by gene expression analyses of 80 signature genes. This algorithm was used to classify the Basal-single-type and Luminal-Basal-type samples into BLIA, BLIS, LAR, and MES [116].

2.2.5 Software and Statistics

Gene expression analysis was performed on full genome microarray data (n = 7985) using limma (v3.2) [125]. Hallmark and Oncogenic gene sets from the Molecular Signatures Database v7.3 were used for gene set enrichment analysis (GSEA) [126]. Genes were ranked based on the effect size ratio using the Cohen's D effect size [127]. Differentially expressed genes (DEG) were considered significant with a p-value ≤ 0.05 and a log2 fold change ≥ 1.

Computational analysis and visualization were performed using R (v3.6.1) [128]. Principal component analysis (PCA) was performed using the “prcomp” package (v3.6.2) [129] and visualized using “ggplot” (v3.3.2) [130]. Unpaired, two-sample t-tests were used to measure if the means of ER, PR, and Ki-67 positivity were significantly different between single and dual subtypes. Chi-square test of Independence was used to test for differences of categorical variables within the Burstein classification (BLIA, BLIS, LAR, and MES) and a multivariate logistic regression analysis for response to therapy (pathological complete response, pCR) between single and dual subtypes. Molecular subtype classification algorithms were used from the “Genefu” package [131].

2.3 Results

2.3.1 BluePrint Single and Dual Subtype Classification

Molecular subtyping of patient tumors (n = 15580) was performed at Agendia using the BP 80-gene assay as previously described [97, 100]. We applied the dual subtype classification method (see “Methods” for details) to assess the presence of multiple activated pathways.Most tumors were classified as single subtype (n = 15087, 96.8%) followed by 449 (2.9%) tumors classified as dual subtype, and 44 (0.3%) tumors as triple subtype (Table 2.1). The most common dual subtypes in this dataset were the Luminal-Basal-type and the Luminal-HER2-type. These had sufficient numbers for downstream analyses while HER2-Basal-type and Luminal-HER2-Basal-type were not sufficient in size [132] and not further analyzed (Table 2.1a).

To note, only 1.9% of Luminal-type tumors were identified as dual subtype, whereas this was the case for 9.6% of the Basal-type and 23.8% of the HER2- type tumors. Since our dataset was largely HR+HER2- (Table S1), in order to estimate the dual subtype prevalence in the overall BC clinical population, we iteratively created subsets representing expected distributions of clinical subtypes (https://seer.cancer.gov/statfacts/html/breast-subtypes.html) [133] (70% HR+HER2-, 13% HR+HER2+, 5% HR-/HER2+ and 12% HR-HER2-) and we detected 4.92% dual subtypes (95% CI 4.91-4.93) (Figure S2.7).

2.3.2 Principal Component Analyses using BluePrint Reveals Similarities between Subtypes

To understand the similarities between single and dual subtypes, we performed PCA based on the BP gene expression signatures (Figure 2.1(a-e)). We observed a clear distinction of single subtypes shown in the first two principal components (Figure 2.1a-c). Luminal-Basal-type cluster separately from both Basal-single-type (Figure 2.1d) and Luminal-single-type (Figure 2.1e), conversely, Luminal-HER2-type (Figure 2.1e, f) are more closely related with both Luminal-single-type (Figure 2.1b) and

HER2-single type (Figure 2.1c).

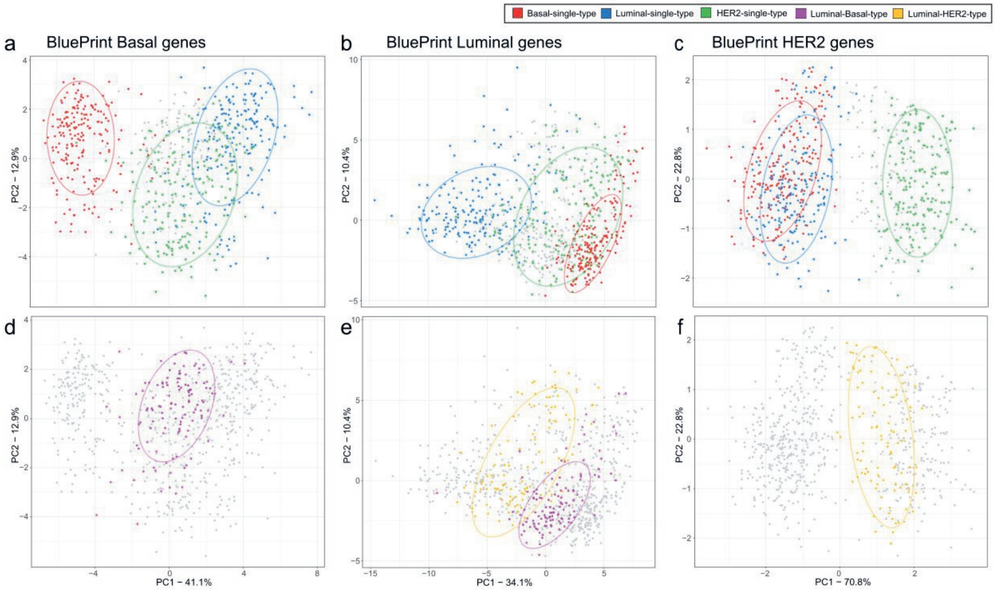


Figure 2.1: Principle component analysis using the three BluePrint signature gene sets (Basal-type, N = 28, panels **a** and **d**; Luminal-type, N = 58, panels **b** and **e**; HER2-type, N = 4, panels **c** and **f**). The x-axis shows variance explained for the first principle component (PC) and the y-axis show the variance explained for the second PC of the correspondent BluePrint signature gene set. **a–c** Clustering of Basal-single-type, Luminal-single-type, and HER2-single-type samples based on BluePrint signature genes. **d–f** Clustering of Luminal-Basal-type and Luminal-HER2-type based on BluePrint signature genes. **a–c** shows coloring of single subtype samples (blue, Luminal-single-type; green, HER2-single-type; red, Basal-single-type) whereas the dual subtype samples are colored grey. d–f shows this in reverse where the dual subtypes are colored (yellow, Luminal-HER2-type; pink, Luminal-Basal-type) and the single subtypes are shown in grey. The ellipses reported in each subfigure illustrate the 80% confidence intervals of the single and dual subtypes

2.3.3 Differential Gene Expression Analysis highlights Differences between BluePrint Dual and Single Subtypes

Differential expression analysis using full-genome data (n = 7985) was performed to evaluate global transcriptional differences between single and dual subtypes. As expected from the PCA, when compared with their corresponding single subtypes, more DEGs were found for the Luminal-Basal-type (446 DEGs) (Figure 2.2a, b) than for the Luminal-HER2-type (151 DEGs) (Figure 2.2e, f).

Among the up-regulated genes in Luminal-Basal-type (vs. both Basal-single-type and Luminal-single-type) were present *MUCL1*, a known tumor suppressor gene [134], and *CLCA2*, a negative regulator of cancer cell migration and invasion [135].

Among the most up-regulated genes in Luminal-HER2-type compared with Luminal-single-type tumors, we found *GRB7*, *TCAP*, and *ERBB2* which belong to the HER2 amplicon and are known to be overexpressed in pathologically confirmed HER2 tumors [136]. Indeed, these genes were also up-regulated in HER2-single-type tumors (Figure 2.2e). When comparing Luminal-HER2-type with HER2-single-type, *ESR1* was found to be upregulated, similarly as in Luminal-single-type tumors. Additionally, Luminal-HER2-type tumors were mainly classified as either Luminal B (n = 34/99, 34%) or HER2 enriched (n = 44/99, 44%) using the intrinsic subtype classified of the "Genefu" [91, 131]. Together, these data suggest that both ER and HER2 are activated in Luminal-HER2-type tumors.

2.3.4 Differences between BluePrint Single and Dual subtypes may Impact Therapy Response Pathways

A better understanding of the underlying biological characteristics of the dual subtypes may come from analyzing gene pathway regulation.

Comparison of Luminal-Basal-type with Basal-single-type revealed upregulation of two estrogen response (ESR) and one androgen response (AR)-related gene sets (Figure 2.2c). Same ESR gene sets were downregulated in Luminal-Basal-type versus Luminal-single-type, indicating that Luminal-Basal-type has intermediate ER levels. Conversely, AR was upregulated in Luminal-Basal-type, versus both the Basal-single-type and Luminal-single-type. G2M and E2F pathways [137, 138] were either downregulated or upregulated in Luminal-Basal-type compared with Basal-single-type and Luminal-single-type, respectively, indicating that Luminal-Basal-type are less proliferative than Basal-single-type, but more proliferative than Luminal-single-type tumors. Taken together, Luminal-Basal-type tumors show a distinct biology from their single counterparts with decreased proliferation than Basal-single-type and AR activation.

Compared with single HER2-single-type tumors, a Luminal-HER2 type shows downregulation of *MAPK* (MEK and RAF) signaling pathways and ER activation (Figure 2.2g). Clinical characteristics of the single and dual BP subtypes and their response to therapy may confirm these hypotheses and provide additional insights.

2.3.5 BluePrint Dual subtypes present clear Clinicopathological Differences from Single Subtypes

Standard BP Luminal-, HER2-, and Basal- type tumors were further stratified using the single-dual subtyping classification (Figure 2.3a). Additionally, conventional

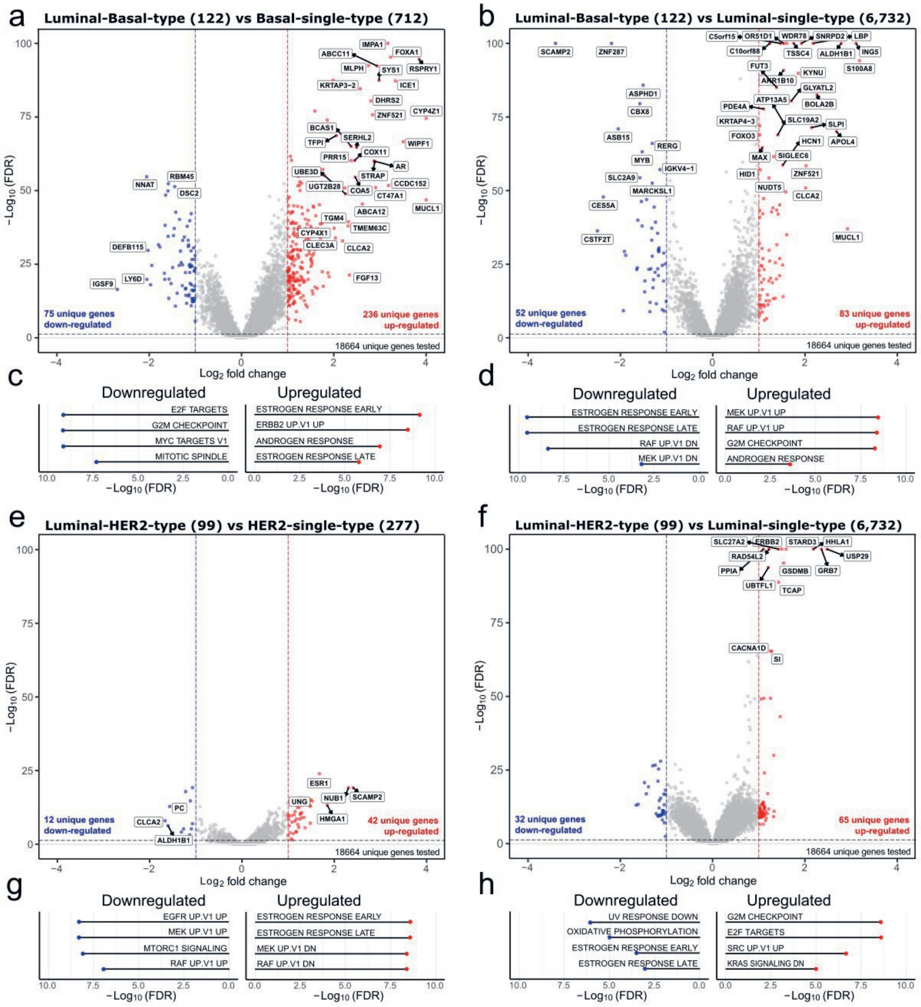


Figure 2.2: Differential gene expression analysis between BluePrint single and dual subtypes. The x-axis and y-axis report the Log2 fold change and the FDR-adjusted p-values (-Log10(FDR)), respectively. Number of tumor samples used for the analysis are shown in between brackets in titles. Significance thresholds of ≤ 0.05 FDR and a log2 fold change of ≥ 1 were used. Red and blue dots illustrate significant differentially expressed genes. The strongest differentially expressed genes are labeled (abs(logFC) ≥ 2 or -Log10 adj p-value ≥ 50). Differentially expressed genes are identified in the following comparisons: **a** Luminal-Basal-type versus Basal-single-type. **b** Luminal-Basal-type versus Luminal-single-type, **e** Luminal-HER2-type versus HER2-single-type, and **f** Luminal-HER2-type versus Luminal-single-type. Similarly, differentially expressed pathways are shown between **c** Luminal-Basal-type versus Basal-single-type. **d** Luminal-Basal-type versus Luminal-single-type, **g** Luminal-HER2-type versus HER2-single-type, and **h** Luminal-HER2-type versus Luminal-single-type. FDR= false discovery rate, UP = upregulated, DN = downregulated

clinical subtypes (based on IHC HR staining (ER, PR) and HER2 status) were further classified into BP single subtypes or dual subtypes (Figure 2.3b, c). Majority of HR-HER2- tumors were classified as Basal single-type (n = 150/176) (Figure 2.3b). while only 26 were dual subtypes of which 22 were Luminal- Basal type.

Most of the HR+ HER2- tumors were classified as Luminal-single-type (n = 4285/4548), but interestingly, 3% (n = 152/4548) was classified as Basal-single-type, which corresponds to more than half of all Basal-single-types identified by BP (n = 152/265) (Fig. 3b). Of the HR+ HER2- with a dual subtype, majority was Luminal-Basal-type (n = 57 / 86) (Figure 2.3c).

Most HR+ HER2+ tumors were classified as either Luminal-single-type (n = 165/272) or as HER2-single-type (n = 61/272) (Figure 2.3b) with the most frequent dual subtype being the Luminal-HER2-type (n = 30/34) (Figure 2.3c).

Luminal-single-type tumors had the highest IHC ER expression levels with the lowest levels observed in Basal-single-type tumors (Figure 2.3d). Dual subtypes showed intermediate ER expression, compared to their single counterparts (Figure 2.3d). ER low positive tumors (1 – 10% IHC) were mostly found in the Basal single-type (n = 62/147, 42%) and in the Luminal single-type (n = 58/147, 39%) (Figure 2.3d). However, considering the differences in sample size of the subtypes, a larger fraction of Basal-single-type (24%) was found to be ER low positive, compared with other subtypes. Proliferation measured by % Ki-67 positivity was significantly higher in Luminal-Basal-type and Luminal-HER2-type compared with Luminal-single-type, but significantly lower than Basal-single-type and HER2-single-type (Figure 2.3e). Indeed, there were significantly more Luminal-Basal-type (n = 34/47, 72.3%, p-value < 0.001) than Basal-single-type tumors (59/173, 34.1%) with Ki67 < 30%, threshold recently proposed for the so-called TNBC low proliferation (TNLP) tumors [139] (Figure 2.3e).

2.3.6 Burstein LAR and MES Subtypes are Identified using BluePrint Dual Subtype Classification

Since Luminal-Basal-type displays different transcriptional characteristics than Luminal-single-type and Basal-single-type, we classified them using the Burstein classifier to better understand their biology. Indeed, we found a significant association between BP single/dual subtypes and the Burstein BLIA, BLIS, LAR, and MES subtypes [116] (p-value < 0.001) with the Basal-single-type classified mostly as BLIA or BLIS, whereas the Luminal-Basal-type as LAR or MES (Figure 2.4a), irrespective of their standard BP subtype (Figure 2.4b).

When using the PAM50 [89, 91] intrinsic subtype classifier of the"Genefu" [131] package, the Luminal-Basal-type tumors were mostly classified as HER2 enriched (HER2-e) (Table S2.4a). When comparing "Luminal-Basal/HER2-e" against "Luminal-Basal/ non-HER2-e", common biomarkers for HER2 molecular classification were not differentially expressed (Table S2.4b). Indeed, approximately 98% of Luminal-

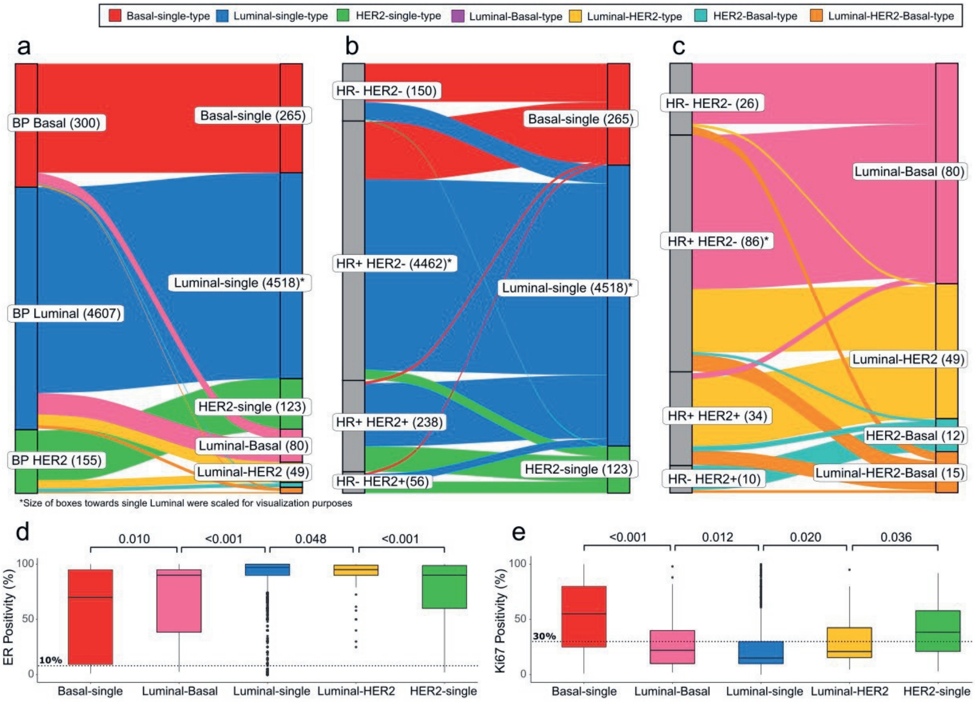


Figure 2.3: **a** Sankey plot showing the further stratification of the standard BluePrint (BP) Basal, Luminal, and HER2 subtypes with full-genome microarray data available, into the BP single and dual subtypes. **b** Sankey plot illustrating the re-classification of clinical-based subtypes (based on hormone receptors (HR) and human epidermal growth factor receptor 2 (HER2) status) to BP-based single-type molecular subtypes (Basal-single-type, Luminal-single-type, HER2-single-type). **c** Further stratification of the same clinical-based subtypes as in **(b)** to the BP-based dual subtypes (Luminal-HER2-type, Luminal-Basal-type, HER2-Basal-type, and Luminal-HER2-Basal-type). **d, e** Boxplots reporting for each single and dual subtype category (x-axis), the level and spread of estrogen receptor and Ki67 positivity based on Immunohistochemistry assessment (y-axis). Significant differential positivity between ER and Ki67 was assumed at a p-value < 0.05 determined with a t-test between subtype categories. To note, for 4511 of the 9573 tumor samples with clinical annotation, HR and HER2 status were not available (Table S2.1)

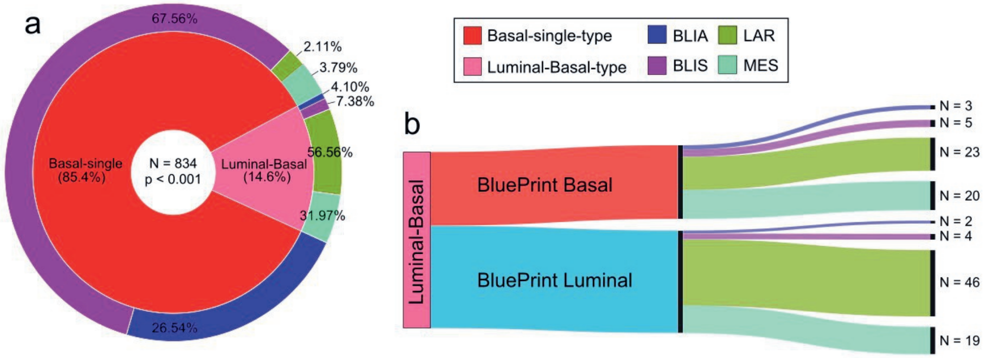


Figure 2.4: BluePrint (BP) dual subtype classification compared with Burstein's classification of triple-negative breast cancer tumors [116]. **a)**The inner circle contains percentages of the BP Basal-single-type and BP Luminal-Basal-type. The outer circle illustrates the correspondent Burstein classification into Basal-like immuno-activated (BLIA), Basal-like immuno-suppressed (BLIS), Luminal androgen receptor (LAR), or Mesenchymal (MES). **b)** Samples with the Luminal-Basal-type were split based on standard BluePrint classification to illustrate their distribution over BLIA, BLIS, LAR, and MES subtypes. Significant differential classification of Burstein subtypes was assumed at a p-value ≤ 0.05 determined with a Chi-square test of Independence between subtypes

Basal type tumors were clinically HER2- (Figure 2.3c).

2.3.7 BluePrint Dual Subtype Classification of the NBRST dataset shows Refined Prediction to Therapy

Our findings indicate that the Luminal-HER2-type shares clinical and genomic features with Luminal-single-type and HER2-single-type and previous studies suggest that HR and *HER2* co-expression is associated with endocrine and HER2-targeted therapy resistance [140, 141]. Therefore, to better understand how Luminal-HER2-type relates to HER2-targeted therapy response, we analyzed the NBRST dataset (see Methods for details) [122] and selected only pathologically confirmed HER2+ tumors (n = 289) with gene expression and HER2- targeted therapy response data available [either Trastuzumab (T) only or with Pertuzumab (P)]. Patient tumors were stratified using the BluePrint dual subtype classification (Figure 2.5).

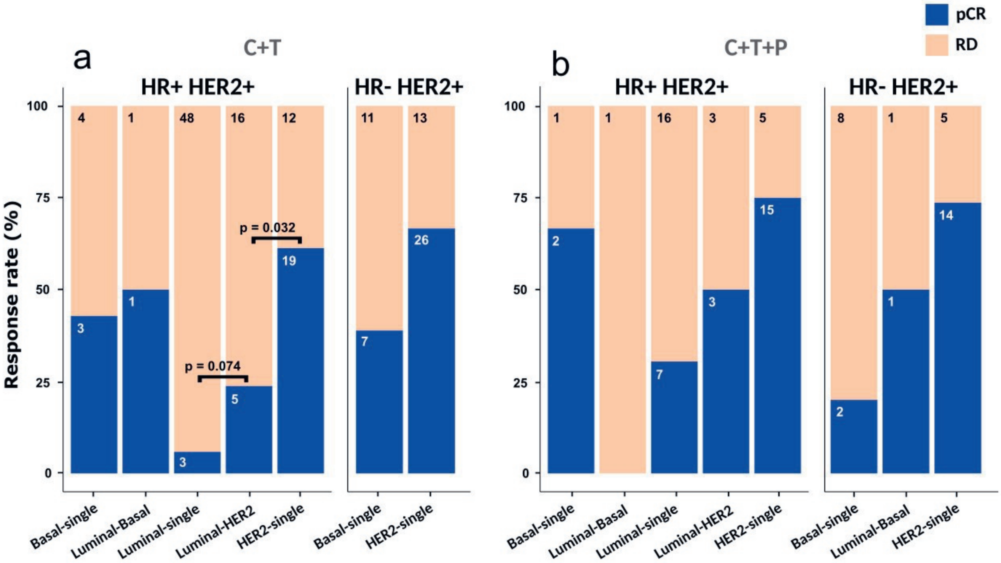


Figure 2.5: Distribution of pathologically confirmed HER2+ patients of the NBRST trial [121–123] based on the BluePrint single and dual subtype classification and their treatment response (N = 253). Patients are grouped based on their therapy regimen [chemotherapy (C) plus Trastuzumab (T) (panel **a**) or C + T and Pertuzumab (P) (panel **b**)], and their HR and HER2 status (HR+ HER2 + or HR- HER2+. The colored bars represents if a tumor did (pCR, blue) or did not [Residual Disease (RD), bisque] achieve pathological complete response (pCR). p-value determined with a chi-square test of independence between subtypes. Of the entire NBRST set (n = 289), 253 samples are showed due to low numerosity of HER2-Basal-type (n = 19) and Luminal-HER2-Basal (n = 17)

BP HER2-single-type showed higher pCR rate to chemotherapy (C) + T compared to Luminal-HER2-type (61.3% vs 23.8%, p = 0.032) (Figure 2.5a). Although not significant (also due to lower numerosity of Luminal-HER2-type), this trend remained for patients that received additional P (Figure 2.5b). Response rates of Luminal-HER2-type tumors was higher, but not significantly different than for Luminal-single-type. Instead, a significant higher response rate was observed for the HER2-single-type compared to the Luminal-HER2-type, after correcting for HR status, tumor stage, tumor grade, and therapy in a multivariate logistic regression analysis (p value = 0.006, Table S2.3).

2.4 Discussion

Molecular subtyping using the standard BP 80-gene assay enables to discern the tumor subtype by the underlying functional pathways and not merely by HR and HER2 status [97, 100]. In most cases, the assay identifies a single, dominant activated pathway distinctive of a Luminal-, Basal-, or HER2-type tumor. This information often confirms the pathologically defined subtype but in many cases further classifies tumors from their initial clinical subtype into a different molecular subtype. This phenomenon has clinical implications for the treatment of patients, perhaps most notably in the ER+/Basal and HER2+/Luminal subtypes which have been previously described [121, 122, 142, 143].

The vast majority of the breast cancer tumors analyzed in this study using the BP test show a single activated pathway (i.e., single BP subtype) (97%); however, less frequently, they exhibit multiple activated pathways (i.e., dual or triple subtype) (3%), as we showed in a preliminary analysis [144]. Notably, this dataset mostly reflects a HR+ population but upon sampling the data based on observed frequencies of clinical subtypes, such a percentage raises to approximately 5%. Importantly, the single and dual assessment performed on the NBRST dataset and also reported for the TRAIN2 [145] and APHINITY [146] patient cohorts show a higher number of dual subtypes, ranging from 11 to 30%, indicating that the dual subtype classification might have a greater clinical impact on a HER2+ population and that the potential clinical utility should be found in specific subgroups rather than in the entire EBC population. The analysis on the NBRST dataset was performed on limited numbers of dual subtypes (n = 32); however, the size was sufficient to generate statistically powerful results.

Overall, in this manuscript, we aimed to provide a better understanding of the biological diversity of EBC and these results should be taken with caution with respect to any immediate change in clinical management.

Next, by analyzing whole-transcriptomic data, we set out to understand if and how dual subtypes were distinct from single subtypes. For the analysis, we focused on the Luminal-Basal-type and Luminal-HER2-type tumors as the other dual subtypes were limited in size.

Neither the Basal nor the Luminal BP template genes were able to fully capture the biology of the Luminal-Basal-type tumors. The majority of tumors expressing typical Basal gene patterns are TNBC by pathology [147], and it is known that there is a large overlap between BP Basal subtypes and TNBCs. Therefore, we applied the TNBC Burstein classifier on the Basal-single-type and Luminal-Basal-type. Basal-single-type tumors were mostly classified as BLIA and BLIS while Luminal-Basal-type tumors were more likely to be either LAR or MES. Genes described by Burstein et al. to be up-regulated in the LAR subtype, such as *DHRS2*, *AGR2*, *FOXA1*, *AR*, and *MUCL1*, were indeed higher expressed in Luminal-Basal-type compared with the Basal-single-type samples. Since the majority of Luminal-Basal-type tumors were classified as LAR, and according to Burstein et al., those patients derive benefit from traditional anti-estrogen or anti-androgen therapy, we could speculate that Luminal-Basal-type cancers would benefit from such treatment as well. Furthermore, *ADH1B* and *FABP4* genes were up-regulated in Luminal-Basal-type samples compared with Basal-single-type samples. The upregulation of these genes is typical of the MES subtype, which is characterized by the dysregulation of cell cycle and DNA damage repair pathways. On the contrary, BLIS subtype-specific genes, *HORMAD1*, *SOX10*, *SERPINB5*, and *FOXC1*, were up-regulated in Basal-single-type samples compared with Luminal-Basal-type samples. Therefore, we could hypothesize that among the Basal-single-type samples, two subgroups are present which are indiscernible with the current dual subtype classification, but might have a different prognosis according to Burstein et al. and require additional analyses. Notably, majority of the Luminal-Basal-type showed a Ki67 positivity below 30% which might indicate that they share features with the TNLP tumors recently described by Bhargava and colleagues [139]. Additionally, no large agreement was found between any of the dual subtypes and the normal-like [89, 90] (Table S2.4). or claudin-low classifications [113, 131] (data not shown). Conversely, BluePrint Basal-, Luminal-, and HER2-single type classifications were largely concordant with the intrinsic subtypes (> 90%) (see Table S2.4). Interestingly, and perhaps unexpectedly, the Luminal-Basal-type tumors were mostly classified as HER2-e intrinsic subtype, possibly due to the absence of Luminal- and Basal-type biology in the BP Luminal-Basal-type.

Luminal-HER2-type samples consistently showed patterns of both ER and HER2 activation (by expression and IHC/FISH), which may suggest similarities to the clinically triple-positive tumors [114]. Expression of both ER and HER2 may lead to receptor crosstalk which has often been associated with resistance to both endocrine and HER2-targeted therapies [148]. However, down-regulation of the MAPK-related gene sets MEK and RAF may indicate no downstream activation of the HER2 pathway. Therefore, Luminal-HER2-type tumors are unlikely fueled through the HER2 pathway alone and HER2-targeted therapies might not be as effective as in the HER2-single-type tumors. This suggestion is strengthened by the observation in the NBRST data that Luminal-HER2-type tumors have a significantly lower pCR rate to neoadjuvant chemotherapy including HER2-targeted agents compared with HER2-single-type tumors (p-value < 0.032). This is supported by preliminary subanalysis of the TRAIN2 [145, 149] and APHINITY [146, 150, 151] trial datasets, suggesting that BluePrint HER2-single-type tumors derive the most benefit from HER2 dual-targeted treatment [146].

It has been suggested that clinically triple-positive tumors develop endocrine resistance as downstream-activated MAPK inhibits ER transcription and phosphorylates ER [141]; however, in this study, Luminal-HER2-type tumors may be only driven by the ER pathway, as MAPK is downregulated compared with HER2-single-type tumors and not significantly different from that of Luminal-single-type tumors. Further analysis on Luminal-HER2-type samples treated with endocrine therapy is required to investigate and confirm this hypothesis.

2.5 Conclusion

Our study showed that by further dissecting the BP scores, it is possible to identify a small proportion of EBCs that have dual-activated BP pathways. These dual subtypes display specific transcriptional and clinicopathological features supporting the idea that they represent a different biological subgroup than their single counterparts. Most dual BP subtypes are either Luminal-Basal-type or Luminal-HER2-type.

The Luminal-Basal-type shows lower proliferation levels compared with the Basal-single-type and AR activation. Interestingly, using the Burstein classification, Luminal-Basal tumors are mostly classified as LAR and MES subtypes.

The Luminal-HER2-type resembles features of both the Luminal-single-type and HER2-single-type. However, patients with Luminal-HER2-type tumors have a lower pCR rate after receiving HER2-targeted therapies in addition to chemotherapy compared with patients with a HER2-single-type.

Taken together, BP dual classification shows potential clinical utility in helping treatment decision for a limited, but still relevant, fraction of EBC patients with dual subtypes that may benefit from additional or alternative targeted therapies. Even though molecular subtyping is not yet standardly used in routine clinical diagnostics, increasing number of evidences are emerging indicating that molecular subtypes should become part of breast cancer management [152]. In this light, results presented here further support the need toward such transition and implementation.

Future work will be focused on further confirming and prospectively validating the findings described here in additional independent datasets.

2.6 Acknowledgemnts

We are indebted to Patricia Dauer and Jeff Fensterer for their thorough review of the paper.

2.7 Abbreviations

BP BluePrint
ER Estrogen receptor
PR Progesterone receptor
HER2 Human epidermal growth factor receptor
IHC Immunohistochemistry
FISH Fluorescence in-situ hybridization
EBC Early-stage breast cancer
MP MammaPrint
PCA Principal component analysis
DEG Differentially expressed gene
TNBC Triple-negative breast cancer
LAR Luminal androgen receptor
MES Mesenchymal
BLIA Basal-like immuno-activated
BLIS Basal-like immuno-suppressed
FDR False discovery rate
GSEA Gene set enrichment analysis

2.8 Author's Contributions

MMK analyzed the data, interpreted and visualized the results, and wrote the manuscript. AE collected and analyzed the data, interpreted the results, and wrote the manuscript. AB analyzed the data, interpreted the results, and reviewed the manuscript. JCH interpreted the results and reviewed the manuscript. RB collected the data and reviewed the manuscript. DW designed the research questions. ARM collected the data and reviewed the manuscript. WMA interpreted the results and reviewed the manuscript. LM designed and supervised the research study, interpreted the results, and wrote the manuscript. AG designed and supervised the research study, interpreted the results, and reviewed the manuscript.

2.9 Funding

This research did not receive any special grant from funding agencies in the public, commercial, or not-for-profit sectors. The NBRST trial was funded by Agendia Inc.

2.10 Conflicts of Interest

All authors (MMK, AE, AB, JCH, RB, DW, ACM, WMA, LM, and AMG) are non-commercial employees of Agendia, the commercial entity that markets the 80-gene signature as BluePrint. AMG is named inventor on the patent for the 80-gene signature used in this study. No writing assistance was utilized in the production of this manuscript.

2.11 Footnotes

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Midas M. Kuilman and Architha Ellappalayam have contributed equally as first authors.

2.12 Supplementary Information

2.12.1 Supplementary Methods

Microarray processing

Microarray processing was performed following standard procedure at Agendia [100]. Briefly, Total RNA was isolated from Formalin-Fixed-Paraffin-Embedded (FFPE) tissue with the RNeasy FFPE kit (Qiagen), DNase treated and amplified using a TransPLEX C-WTA kit (Rubicon Genomics, Ann Arbor, MI). Amplified cDNA was labeled using the Genomic DNA Enzymatic Labeling Kit (Agilent Technologies, Santa Clara, CA) and hybridized onto Agendia's diagnostic arrays (custom-designed, Agilent Technologies), according to the manufacturer's instructions.

BluePrint single and dual-subtype classification

The dual-subtype classification was performed as follows and it is visually summarized in Figure S1. First the standard BluePrint (BP) score was calculated from 15580 samples as previously described (Figure S1a-b) [100]. Briefly, for each tumor, three scores were generated, and the subtype with the highest score was the categorical subtype reported. Next, BP scores were scaled with a SoftMax function (Goodfellow IJ, et al (2016) 6.2.2.3 SoftMax Units for Multiple Output Distributions. Deep Learning. MIT Press. pp 180-184) (Fig S1c) to reduce variance and outlier impact, which allows for optimal threshold determination between single and dual subtypes. Using a bootstrap algorithm [124], samples were divided into 70% and 30% groups per BP subtype for 1000 iterations (FigS1d). For each iteration, the two highest scores were selected (Fig S1e) and the distance between them was calculated (Fig S1f). The distribution of the differences between BP scores was constructed (Fig S1g). If a bimodal distribution emerged (implying the presence of single and dual subtypes), the separation point (i.e., local minimum) between the two distributions was selected as a threshold candidate (Fig S1h). After 1000 bootstrap iterations, multiple threshold candidates were captured for each subtype. The maximum likelihood values of threshold distributions were taken as thresholds for the identification of dual subtypes (Fig S1i), which are reported in FigS1j. If the difference between the two highest BP SoftMax scores was lower or equal to the corresponding single-dual threshold, then the tumor was classified as a dual subtype comprised of the two highest molecular subtype scores. If tumors had similar scores for all three subtypes, they would be defined as triple subtypes.

Table 2.2: Clinical-pathological characteristics of the patients analyzed in this study.

	Standard BluePrint classification			
	Luminal (n=8664)	HER2 (n=245)	Basal (n=664)	All (n=9573)
Age at diagnosis (years, median, range)	62 (23-93)	61 (23-95)	58.5 (23-87)	62 (23-95)
Nodal Status				
0	1516	21	121	1658
1	1320	24	47	1391
2	335	2	6	343
3+	163	4	7	174
Missing	5330	194	483	6007
Grade				
1	1409	6	8	1423
2	2966	58	74	3098
3	869	66	309	1244
Missing	3420	155	273	3848
Clinical subtype based on receptor status				
HR+ HER2-	4343	33	172	4548
HR+ HER2+	180	83	9	272
HR- HER2+	20	37	9	66
HR- HER2-	64	2	110	176
missing	4057	90	364	4511
Ki67 Percentage				
Median positivity (1st - 3rd quantiles)	16 (10-30)	40 (21-60)	50 (20-80)	18 (10-32)

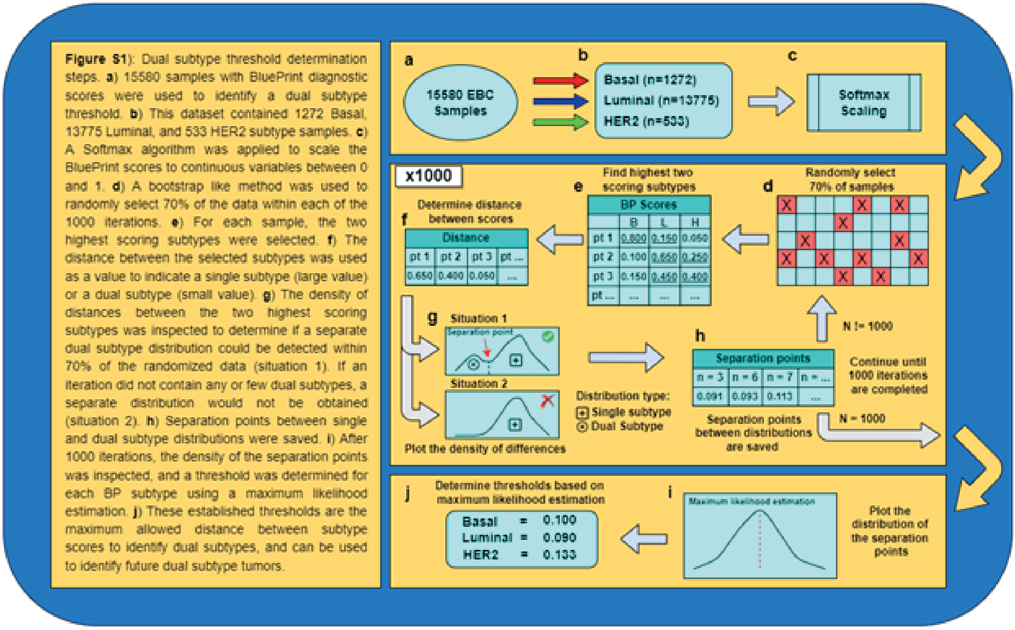


Figure S1): Dual subtype threshold determination steps. **a)** 15580 samples with BluePrint diagnostic scores were used to identify a dual subtype threshold. **b)** This dataset contained 1272 Basal, 13775 Luminal, and 533 HER2 subtype samples. **c)** A Softmax algorithm was applied to scale the BluePrint scores to continuous variables between 0 and 1. **d)** A bootstrap like method was used to randomly select 70% of the data within each of the 1000 iterations. **e)** For each sample, the two highest scoring subtypes were selected. **f)** The distance between the selected subtypes was used as a value to indicate a single subtype (large value) or a dual subtype (small value). **g)** The density of distances between the two highest scoring subtypes was inspected to determine if a separate dual subtype distribution could be detected within 70% of the randomized data (situation 1). If an iteration did not contain any or few dual subtypes, a separate distribution would not be obtained (situation 2). **h)** Separation points between single and dual subtype distributions were saved. **i)** After 1000 iterations, the density of the separation points was inspected, and a threshold was determined for each BP subtype using a maximum likelihood estimation. **j)** These established thresholds are the maximum allowed distance between subtype scores to identify dual subtypes, and can be used to identify future dual subtype tumors.

a
15580 EBC Samples
b
Basal (n=1272)
Luminal (n=13776)
HER2 (n=533)
c
Softmax Scaling
x1000
d
Randomly select 70% of samples
X
e
Find highest two scoring subtypes
f
Determine distance between scores
g
Situation 1
Separation point
Situation 2
Plot the density of differences
Distribution type:
⊞ Single subtype
⊙ Dual Subtype
h
Separation points between distributions are saved
N != 1000
Continue until 1000 iterations are completed
N = 1000
i
Maximum likelihood estimation
Plot the distribution of the separation points
j
Determine thresholds based on maximum likelihood estimation
Basal = 0.100
Luminal = 0.090
HER2 = 0.133

BP Scores			
	B	L	H
pt 1	0.800	0.150	0.050
pt 2	0.100	0.650	0.250
pt 3	0.150	0.450	0.400
pt ...	...	...	...

Distance			
pt 1	pt 2	pt 3	pt ...
0.650	0.400	0.050	...

Separation points			
n = 3	n = 6	n = 7	n = ...
0.091	0.093	0.113	...

Figure 2.6: Dual-subtype threshold determination steps

Table 2.3: Multivariate logistic expression to examine the subtype, HR status, tumor grade, tumor stage, and treatment variables to determine those that best predict response to HER2-targeted therapy. Only cases with complete clinical information were used.

				95% C.I. for EXP (B)			
			Exp (B)			Significance	
				Lower	Upper		N
Variables		Intercept	5.698	1.454	26.466	0.018	
	Subtype	Single HER2*	-	-	-	-	101
		Luminal-HER2	0.204	0.062	0.619	0.006	25
	HR status (IHC)	ER-negative*	-	-	-	-	56
		ER Positive	0.571	0.215	1.466	0.249	70
	Tumor grade	Grade II*	-	-	-	-	51
		Grade III	0.982	0.422	2.251	0.967	75
	Tumor stage	Stage I*	-	-	-	-	21
		Stage II	0.629	0.173	2.020	0.453	70
		Stage III	0.154	0.033	0.634	0.013	24
		Stage IV	0.295	0.052	1.658	0.160	11
	Treatment	C + T*	-	-	-	-	83
		C + T + P	1.844	0.768	4.632	0.179	43

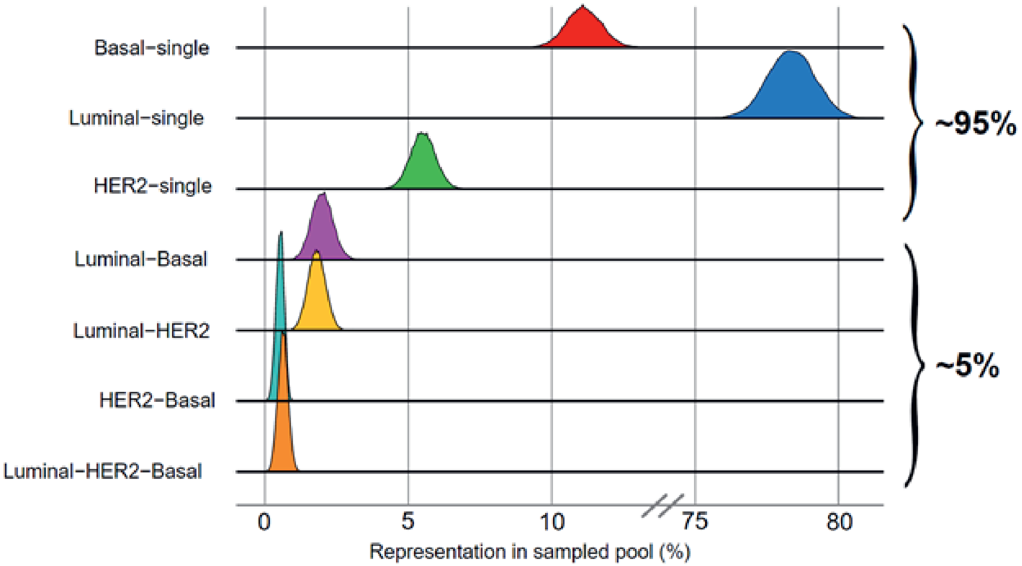


Figure 2.7: Distribution of the representation of single and dual subtypes in a sampled pool of tumor samples. The x-axis reports the proportion in percentage for each subtype (y-axis). Curly brackets on the right indicate the subtype prevalence obtained using the sampled pool. Sampled pool represents actual occurrences of clinical subtypes (70% HR+HER2-, 13% HR+HER2+, 5% HR-/HER2+ and 12% HR-HER2-) according to literature (https://seer.cancer.gov/statfacts/html/breast-subtypes.html) [29].

Table 2.4: Distribution of dual subtypes over molecular subtypes of Genefu.

		Genefu molecular subtyping classification				
		Luminal A	**Luminal B**	**HER2-e**	**Basal**	**Normal-like**
Single-dual subtype classification	**Basal-single-type**	0	2	42	660	8
	Luminal-single-type	3722	2303	163	109	435
	HER2-single-type	6	21	243	6	1
	Luminal-Basal-type	11	14	63	12	22
	Luminal-HER2-type	16	34	44	3	2
	HER2-Basal-type	0	0	14	8	1
	Luminal-HER2-Basal-type	0	13	13	3	1

Table 2.5: Log fold change and adjusted P value of 4 HER2 related genes by comparing Genefu HER2-e BP Luminal-Basal against Genefu non- HER2-e BP Luminal-Basal tumors.

Gene	Log fold change	Adjusted P value
ERBB2	0.346	0.138
GRB7	0.230	0.199
TCAP	-0.049	0.640
STARD3	0.199	0.159

Chapter 3

A twenty-nine HER2 biology Guided Gene Signature Improves Breast Cancer HER2-type Molecular Classification

Architha Ellappalayam*, Andrei Barcaru*, Josien C Haan, Midas M Kuilman, Rajith Bhaskaran, Laura J van ‘t Veer, Lorenza Mittempergher and Annuska M. Glas

*Authors contributed equally to this work

Abstract

Purpose

An 80-gene molecular subtyping test BluePrint (BP) classifies early-stage breast cancer (EBC) into Basal, Luminal, and HER2 subtypes by measuring the similarity to a 58-Luminal, 28-Basal and 4-HER2 gene signatures, respectively. In this study, we aimed to further explore HER2 biology to improve the precision of the HER2 signature.

Methods

Full genome microarray data of 1252 Formalin-Fixed Paraffin-Embedded (FFPE) EBC samples were used to develop an expanded HER2 gene signature. Differential expression analysis (DEA) was conducted to identify additional HER2-biology-relevant genes by comparing BluePrint HER2-type tumors with Basal- and Luminal-type tumors. Statistically significant differentially expressed genes were identified and a further filter based on the coefficient of variation (CV) was applied to select genes for inclusion in the expanded HER2 signature. Additionally, we compared the new signature with previously reported molecular subtyping signatures, using Principal Component Analysis (PCA).

Results

DEA followed by filtering based on CV resulted in the selection of 29 HER2-biology-associated genes. Among the 29 genes, 14 are part of the HER2 amplicon which is known to be upregulated in clinically confirmed HER2-positive (HER2+) tumors. These genes play a role in HER2, PI3K, and, AKT signaling pathways, important for cancer growth and proliferation. We observed that the expanded HER2 signature genes could discriminate the three molecular subtypes with a higher percentage of variance explained in the PCA analysis than the previously reported molecular subtype signatures.

Conclusion

The expanded 29-gene HER2-type signature includes known HER2 amplicon genes and others involved in several HER2-related oncogenic signaling pathways. By incorporating a broader range of genes and pathways relevant to the HER2-type signaling, the expanded HER2 gene signature has the potential to better predict response to HER2-targeted therapies and identify new therapeutic targets for patients with HER2-positive breast cancer.

3.1 Introduction

Breast cancer (BC) is a highly heterogeneous disease that encompasses biologically distinct entities with specific clinical and biological features [153]. There are several methods used to classify the heterogeneity into receptor subtypes either based on immunohistochemistry (IHC), Fluorescence In Situ Hybridization (FISH), or RNA-based assays [90, 112]. The subgroups of the former two are defined by the receptor protein status, whereas the subgroups of the latter are defined by the molecular, also called intrinsic subtypes. IHC is used to measure the presence of estrogen receptor (ER), progesterone receptor (PR), and IHC and/or FISH to measure the human epidermal growth factor receptor 2 (HER2) proteins. BC IHC subgroups are classified based on the presence or absence of these receptors and tumors lacking all these receptors are classified as triple-negative breast cancers (TNBC) [154].

IHC subtypes correlate to the molecular subtypes which were originally identified based on clustering patterns of gene expression by microarray [89, 90, 155]. There is a consensus over three distinct intrinsic molecular subtypes identified in literature namely: Luminal-type (which is predominantly Hormone receptor (HR) positive by IHC), HER2-type (which is predominantly HER2-positive by IHC and fluorescence in situ hybridization (FISH)) and Basal-type tumors, which are mostly the TNBC by IHC/FISH [100].

The BluePrint (BP) 80-gene subtyping assay is a test that classifies early-stage BC into three molecular subtypes, Luminal-type, Basal-type and, HER2-type by measuring the similarity to 58-Luminal, 28-Basal and, 4-HER2 gene signatures [97, 100]. BP was developed to bridge clinical pathology and molecular subtyping, by using IHC-based receptor status as the basis and provides a molecular diagnostic array with a high predictive value. The analytical validity of the BP test was confirmed and discussed previously, and it shows that BP is a precise and reproducible test, both using replicates and over time measurements of multiple control samples [100].

The general binary HER2 scoring system classifies breast cancers into 1) HER2-positive, when HER2 expression is scored either 3+ by immunohistochemistry (IHC) or 1+/2+ by IHC with gene amplification found by fluorescence in-situ hybridization (FISH); and 2) HER2-negative, when HER2 expression is scored either 0+ by IHC or scored 1+/2+ by IHC without FISH gene amplification [156]. Research studies have concluded HER2 low breast cancers to have low levels of HER2 expression, defined as +1 or 2+ by IHC and FISH-negative, and are not be considered positive according to current diagnostic guidelines [157]. However, some recent studies have shown that some HER2-low tumors may still respond to HER2-targeted therapies, particularly if they also have other features associated with HER2 positivity, such as high levels of HER2 heterogeneity [158]. The development of new HER2-targeted therapies and biomarkers that can better capture the heterogeneity of HER2 expression and amplification in breast cancer may also have implications for the treatment of HER2-low tumors in the future [159]. In this paper, we set out to identify an ex-

Table 3.1: Development and validation of the expanded HER2 gene signature: Distribution of the standard BP outcomes of the 1252 samples into training (n = 626) and test (n = 626) data sets

	Luminal-type	**Basal-type**	**HER2-type**	**Total**
Training set	208	209	209	626
Test set	208	209	209	626

panded HER2 signature that would capture the full biological diversity of HER2+ tumors.

3.2 Materials and Methods

3.2.1 Data

For this study, only data and no samples were collected, and all data were fully anonymized according to the 'General Data Protection Regulation' (GDPR), and the 'Health Insurance Portability and Accountability Act' (HIPAA) and are in compliance with the 'Data Protection Act'.

Full genome microarray data of 1552 invasive breast cancers were utilized for the identification and validation of the expanded HER2 signature. The samples were from patients meeting MammaPrint (MP) eligibility criteria [118, 119], stage I, II, or operable stage III breast cancer, tumor diameter ≤ 5 cm, and up to three positive lymph nodes, with any ER/PR/HER2 status.

Microarray sample data encompassing all three BP molecular subtypes (n = 1252 samples) were used to develop and validate the expanded HER2 signature. Comparison of the extended HER2 signature with previously reported molecular subtyping signatures was performed using microarray data of an additional 300 samples (100 Basal-type, 100 Luminal-type, and 100 HER2-type).

Microarray processing was performed previously following standard procedures at Agendia. Briefly, total RNA was isolated from Formalin Fixed Paraffin Embedded (FFPE) tissue with the RNeasy FFPE kit (Qiagen) following the manufacturer's instructions. Total RNA was DNase treated and amplified using a TransPLEX C-WTA whole transcriptome amplification kit (Rubicon Genomics, Ann Arbor, MI). Amplified cDNA was labeled using the Genomic DNA Enzymatic Labeling Kit (Agilent Technologies, Santa Clara, CA) and hybridized onto Agendia's custom-designed arrays, according to manufacturer's instructions [100, 117, 120].

3.2.2 Methods

BluePrint Test

Standard BP scores were calculated for the training and test set samples by comparing the expression of 80 BP genes with the three subtype gene signatures (Basal-type, Luminal-type, and HER2-type) [97, 100]. For each tumor, three scores were generated, and the subtype with the highest score was the categorical subtype reported. The samples and their BP molecular subtype classification are shown in Table 3.1.

3.2.3 Identification of Expanded HER2 Biology Genes

Differential expression analysis (DEA) of full genome microarray data was applied to compare the HER2-type against Basal-type, and HER2-type against Luminal-type samples from the sample set (n = 626 samples). A Benjamini-Hochberg adjusted p-value [160] of 0.05 and a log2 fold change (FC) of 1 and -1 were used as thresholds for selecting statistically significant up- and down-regulated probes, respectively. To be selected, the probes had to meet these requirements in both comparisons (i.e., HER2-type versus Basal-type and HER2-type versus Luminal-type).

Selected statistically significant up- and down-regulated probes matching the same gene were assessed using the Mann Whitney U test and were filtered using the effect size (ES) [127], i.e. only the probes with the largest ES were selected as candidate signature genes.

The final filtration step aimed to ensure the selection of a list of highly stable genes based on the coefficient of variation (CV).

3.2.4 Expanded HER2 Signature Genes Characterization

Enrichment analysis was performed on the highly stable genes using Reactome [161] and the significant biological processes were estimated using Gene Ontology [162].

The percentage of variance explained by the expanded HER2 signature was measured along with other publicly available molecular subtyping gene expression signatures using PCA on existing microarray data derived from 300 FFPE samples.

Previously reported signatures such as 28 signature genes developed by Desmedt *et.al.* [163], 19 signature genes from Lin & Hsu [164], 50 signature genes developed by Perou *et.al.* [89] based on the intrinsic subtypes [90] and 30 signature genes developed by Milioli [165] were used for the PCA analysis.

3.2.5 Software and Statistics

The gene expression analysis, statistical analysis, and the BP prediction using the expanded HER2 signature were built in Python 3.7.6 from Anaconda 3 distribution. The expanded HER2 signature was selected based on the effect size ratio using Cohen's D effect size [127]. Computational analysis and visualization were performed using R (v3.6.1) [128]. Principal component analysis (PCA) was performed using the "prcomp" package [129] (v3.6.2) and visualized using "ggplot" (v3.3.2) [130]. Gene ontology and pathway analysis were performed and visualized using the "pathfinder" package in R [166].

3.3 Results

3.3.1 Identification of the Expanded Set of HER2 Biology Genes

We set out to identify additional HER2 biology-related genes to improve the precision of our HER2 molecular signature. We used 626 samples as training data set to perform the differential expression analysis (DEA). The DEA of HER2-type samples versus Luminal-type and Basal-type revealed 44 unique candidate genes specific for the HER2-type which fulfilled the fold change and p-value selection criteria (see Material and Methods for details). It has been previously reported that clinical HER2-positive breast cancers are characterized by overexpression and amplification of the genes located on chromosome 17q12 [167, 168]. Therefore, the threshold for the maximum allowed CV was chosen on the HER2 amplicon genes' largest CV in order to include all HER2 amplicon genes that passed our DEA significance criteria as well as additional genes that are not on the HER2 amplicon. This CV of a maximum of 25% identifies the most stable genes. The 44 candidate genes were filtered using this threshold resulting in a final set of 29 genes. Details of the 29 genes are reported in Table 3.2.

3.3.2 Functional Annotation of the Expanded 29 HER2 Signature Genes

Among the newly identified 29 genes representing *HER2* biology, 14 out of the 29 genes belong to the *HER2* amplicon region of chromosome 17q12 and exhibit amplification in approximately 90% of HER2-positive tumors (*GSDMB, MED24, ORMDL3, FBXL20, MIEN1, STARD3, CDK12, GRB7, TCAP, ERBB2, PGAP3, RPL19, MED1* and *PSMD3*) [99, 168–173]. All of these genes have been implicated in the context of HER2-positive tumors. Several of these genes have been associated with metastasis and tumor progression. For instance, *MIEN1* has been linked to increased invasiveness and metastatic potential in breast cancer cells [174]. *ERBB2, MED24* and *GRB7* genes are present within a 280kb segment as the core of the amplicon [175] and are concurrently amplified in many HER2+ breast cancers [136, 176]. *ERBB2, GRB7* and *MED1* genes are involved in signal transduction pathways, which play a crucial role in transmitting signals from the extracellular environment to the nucleus, regulating cell growth, survival, and proliferation. *MED1, CDK12* and *FBXL20* genes are involved in transcriptional regulation and influence gene expression patterns [177]. *CDK12* is involved in cell cycle regulation, while *TCAP* plays

Table 3.2: Table showing the new expanded 29 signature genes for HER2-type, with their gene name, molecular function, ENSEMBL ID, and chromosome location.

Gene Symbol	Gene name	Ensembl ID	Chromosome location
PPIP5K1	Diphosphoinositol Pentakisphosphate Kinase 1	ENSG00000168781	chr15 43,533,462-43,590,272
STARD3	StAR Related Lipid Transfer Domain Containing 3	ENSG00000131748	chr17 39,637,090-39,664,201
FBXL20	F-Box And Leucine-Rich Repeat Protein 20	ENSG00000108306	chr17 39,252,663-39,402,556
FGFR2	Fibroblast Growth Factor Receptor 2	ENSG00000066468	chr10 121,478,330-121,598,458
GRB7	Growth Factor Receptor Bound Protein 7	ENSG00000141738	chr17 39,737,927-39,747,291
SIX1	SIX Homeobox 1	ENSG00000126778	chr14 60,643,421-60,658,259
TCAP	Titin-Cap	ENSG00000173991	chr17 39,665,349-39,666,554
ABCA12	ATP Binding Cassette Subfamily A Member 12	ENSG00000144452	chr2 214,931,542-215,138,626
GSDMB	Gasdermin B	ENSG00000073605	chr17 39,904,595-39,919,854
MED24	Mediator Complex Subunit 24	ENSG00000008838	chr17 40,019,097-40,061,215
PI15	Peptidase Inhibitor 15	ENSG00000137558	chr8 74,824,534-74,855,029
MIEN1	Migration And Invasion Enhancer 1	ENSG00000141741	chr17 39,728,496-39,730,532
ERBB2	Erb-B2 Receptor Tyrosine Kinase 2	ENSG00000141736	chr17 39,687,914-39,730,426
PGAP3	Post-GPI Attachment To Proteins Phospholipase 3	ENSG00000161395	chr17 39,671,122-39,696,797
MED1	Mediator Complex Subunit 1	ENSG00000125686	chr17 39,404,285-39,451,281
PRODH	Proline Dehydrogenase 1	ENSG00000100033	chr22 18,912,777-18,936,553
CDK12	Cyclin Dependent Kinase 12	ENSG00000167258	chr17 39,461,486-39,567,560
C2ORF72	Chromosome 2 Open Reading Frame 72	ENSG00000204128	chr2 231,037,523-231,049,719
SSFA2	ITPR Interacting Domain Containing 2	ENSG00000138434	chr2 181,891,730-181,930,738
PMAIP1	Phorbol-12-Myristate-13-Acetate-Induced Protein 1	ENSG00000141682	chr18 59,899,996-59,904,305
MFSD2A	Major Facilitator Superfamily Domain Containing 2A	ENSG00000168389	chr1 39,955,112-39,969,968
NANOS1	Nanos C2HC-Type Zinc Finger 1	ENSG00000188613	chr10 119,029,714-119,033,730
MNX1	Motor Neuron And Pancreas Homeobox 1	ENSG00000130675	chr7 156,994,051-157,010,663
CATSPERB	Cation Channel Sperm Associated Auxiliary Subunit Beta	ENSG00000133962	chr14 91,580,696-91,780,707
DST	Dystonin	ENSG00000151914	chr6 56,457,987-56,954,830
RPL19	Ribosomal Protein L19	ENSG00000108298	chr17 39,200,283-39,204,840
ORMDL3	ORMDL Sphingolipid Biosynthesis Regulator 3	ENSG00000172057	chr17 39,921,041-39,927,601
PSMD3	Proteasome 26S Subunit, Non-ATPase 3	ENSG0000010834	chr17 39,980,807-39,997,959
ITGB6	Integrin Subunit Beta 6	ENSG00000115221	chr2 160,099,667-160,200,313

a role in muscle cell development and growth [178]. *PSMD3* is involved in proteasomal degradation pathways [161], while *STARD3* is involved in intracellular lipid transport and membrane dynamics [179]. While these genes have individual functions and mechanisms, their collective involvement in HER2-positive breast cancer highlights the complex interplay of molecular pathways and networks underlying the disease.

3.3.3 Decoding HER2 Signature Genes by Reactome Pathway Analysis and Gene Ontology

Reactome pathway and Gene Ontology analyses were performed to identify the gene pathways captured by the 29 genes. Figure 3.1, shows 20 Reactome pathways identified, including expected enriched pathways like "Signaling by ERBB2" and other pathways that are associated with *ERBB2* overexpression such as PI3K/AKT signaling pathway. Dysregulation of the PI3K/Akt signaling pathway in HER2 breast tumors results in uncontrolled cell proliferation and endocrine resistance [180, 181]. The *RND1* GTPase cycle pathway was also significantly enriched. Common features among these pathways include their involvement in cell signaling, regulation of cell growth and survival, and interconnections with other signaling networks. Many of these pathways intersect and cross-talk with each other, forming a complex signaling network that controls various aspects of cellular behavior. Activation of these pathways promotes cancer cell proliferation, survival, and metastasis.

Figure 3.2 shows that the 23 most enriched GO pathways are "regulation of ERK1 and ERK2 cascade" and "positive regulation of transcription initiation from RNA polymerase II promoter" was found. Among the other significant biological processes are a cellular response to steroid hormone stimulus, epidermal growth factor signaling pathway, and peptidyl tyrosine phosphorylation. Some other more commonly found biological processes include cellular response to growth factor stimulus, translation, and positive regulation of cell growth.

3.3.4 Precision of the Expanded HER2 Signature Genes to Identify Molecular Subtypes

PCA analysis was performed to understand if the expanded HER2 signature genes increase the precision of identifying the three molecular subtypes by capturing an increased percentage of variance. The PCA plots of the expanded BP gene set, as well as PCA plots of previously reported molecular subtyping signatures on Agendia microarray gene expression data are shown in Figure 3.3. When the PCA of the expanded BP signature genes was applied on FFPE samples of all three BP subtypes, PC1 could capture 44.8% of the variance observed, while also showing three distinct clusters of the HER2-, Luminal- and Basal- molecular subtypes. Comparatively, previously reported molecular subtyping signature genes from Desmedt *et.al.*

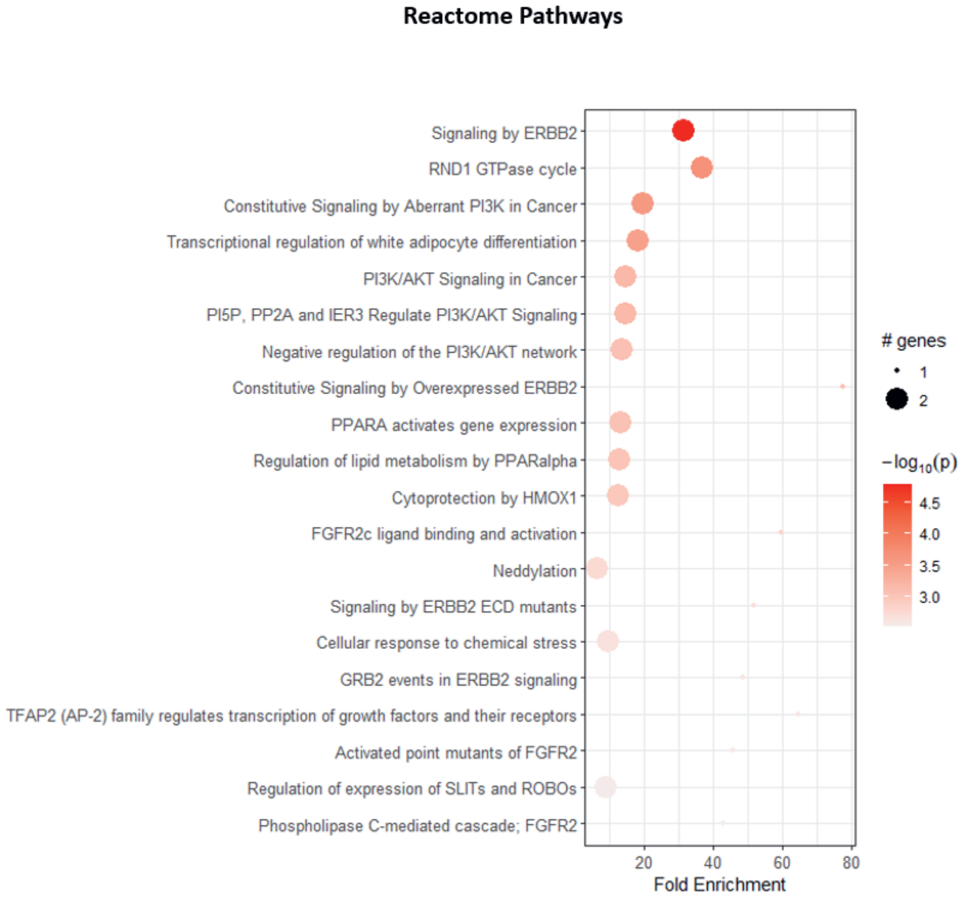


Figure 3.1: Overview of the Reactome Pathway Enrichment Analysis of the extended 29-gene HER2 signature. The x-axis represents the fold enrichment value (obtained by comparing the background frequency of total genes annotated to the GO-BP term to the sample frequency representing the number of genes inputted that fall under the same term) and the y-axis represents the significantly enriched pathways of the 29 genes. The size of the dots indicates the number of genes in the Reactome pathway and the color intensity indicates the -log10p-value.

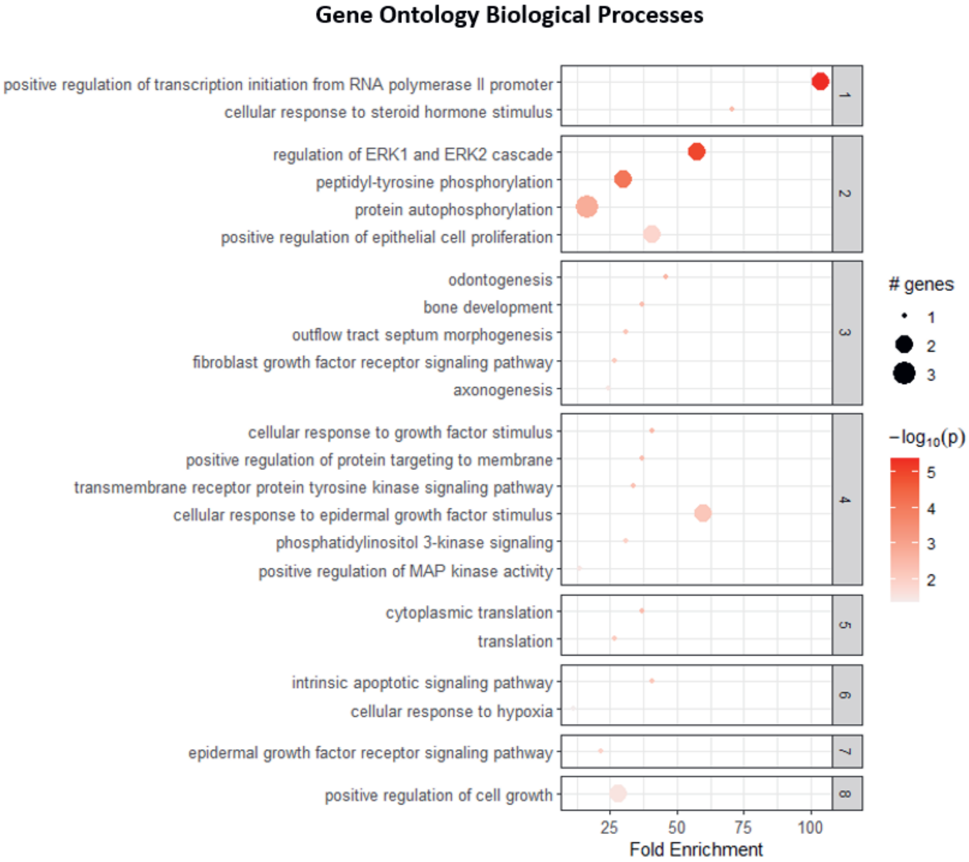


Figure 3.2: Overview of the Gene Ontology Biological Processes (GO-BP) identified in the extended 29-gene HER2 signature, clustered based on biologically relevant groups. The x-axis shows the Fold Enrichment value (obtained by comparing the background frequency of total genes annotated to the GO-BP term to the sample frequency representing the number of genes inputted that fall under the same term) and the y-axis shows the GO-BP enriched terms. The size of the dots indicates the number of the genes in the GO-BP and the color intensity indicates the -log10p-value.

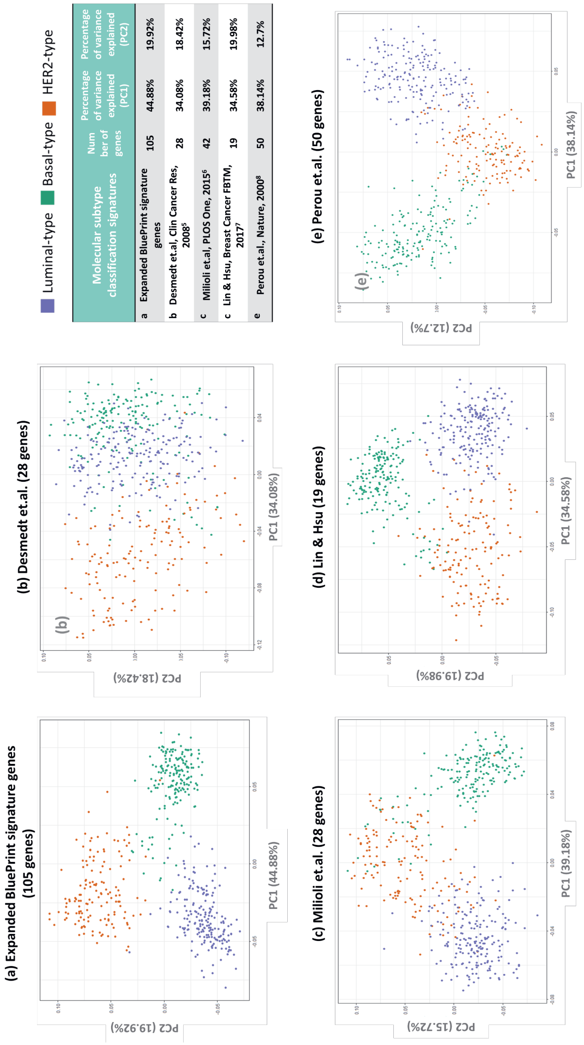
[163] and Perou *et.al.* [89] showed less distinct clusters between the subtypes, and a lower percentage of variance captured to distinguish the three types. The signature genes reported by Milioli *et.al.* [165] were able to show distinct clusters, however, many of the Luminal- and HER2-type samples were very closely clustered with each other, indicating that not much variance was captured among these subtypes. The Basal-like samples clustered separately from the other two molecular subtypes, without much clustering with the Luminal or HER2-type samples. The PCA with the signature genes of Lin & Hsu [164] indicated a strong separation for the basal-like samples, showing a lot more variance captured across the y-axis of the plots as compared to the x-axis. The Luminal and the HER2 samples were spread across the x-axis showing a lesser amount of variance captured. Overall, Figure 3.3 shows that the expanded Blueprint signature genes captured the maximum amount of variance compared to the other reported gene signatures.

3.4 Discussion

Around 16-18% of breast cancer cases are classified as HER2-positive based on the immunohistochemistry protein expression profile [182]. Heterogeneity in HER2-positive tumors can be attributed to genetic and epigenetic alterations and tumor microenvironment factors, but it is not always reflected in the clinical classification of breast cancer [183]. In this paper we set out to identify a HER2 signature that could discriminate with higher precision HER2-type tumors from Basal- and Luminal-type tumors not only based on the HER2 amplicon genes but on a larger set of genes, aiming to better capture the biology of HER2-type driven tumors. Analyses described in this study led to the discovery of an “expanded HER2 gene signature” that can be used in combination with the Basal- and Luminal-type BP signatures, to identify HER2-type tumors.

29 HER2-biology-associated genes passed our thresholds of fold change, p-value, and coefficient of variation. Many of these genes are located on chromosome 17q12 and are known to be co-amplified with *ERBB2* in human breast cancers [169, 170]. Pathway analysis of the 29 signature genes shows many major pathways like MAPK1/MAP3K signaling, *ERBB2* signaling, GRB7 events in ERBB2 signaling, and constitute signaling by aberrant PI3K in Cancer.

The PI3K (phosphoinositide 3-kinase) signaling pathway is one of the most frequently altered pathways in cancer, including HER2-positive breast cancer. In HER2-positive breast cancer, aberrant PI3K signaling is frequently observed due to the activation of the HER2 receptor, which can directly activate *PI3K* [184]. Several studies have demonstrated that constitutive PI3K signaling is associated with a poor prognosis in HER2-positive breast cancer patients [185]. Furthermore, preclinical studies have shown that inhibition of the PI3K pathway can enhance the response to *HER2* targeted therapies, suggesting that targeting this pathway may be a promising therapeutic strategy for HER2-positive breast cancer [186].



	Molecular subtype classification signatures	Number of genes	Percentage of variance explained (PC1)	Percentage of variance explained (PC2)
a	Expanded BluePrint signature genes	105	44.88%	19.92%
b	Desmedt et.al, Clin Cancer Res, 2008[5]	28	34.08%	18.42%
c	Milioli et.al, PLOS One, 2015[6]	42	39.18%	15.72%
c	Lin & Hsu, Breast Cancer FBTM, 2017[7]	19	34.58%	19.98%
e	Perou et.al., Nature, 2000[8]	50	38.14%	12.7%

Figure 3.3: Principal Component Analysis (PCA) plots depicting the expression profiles of the expanded BluePrint signature genes (n = 105) and genes of four previously reported signatures in a comprehensive set of 300 FFPE breast cancer samples (100 HER2-type, 100 Luminal-type, and 100 Basal-type samples). HER2-type samples are represented by the color green, Luminal samples by purple, and Basal samples by orange. (a) PCA of the 105 expanded BluePrint signature genes with a PC1 value of 44.8% and a PC2 value of 19.92% (b) PCA of the 28 signature genes by Desmedt et.al. with a PC1 value of 34.08% and a PC2 value of 18.42% (c) PCA of the 42 signature genes by Milioli et.al. with a PC1 value of 39.18% and a PC2 value of 15.72% (d) PCA of the 19 signature genes by Lin & Hsu with a PC1 value of 34.58% and a PC2 value of 19.98% (e) PCA of the 50 signature genes by Perou et.al. with a PC1 value of 38.14% and a PC2 value of 17.7% (f) A table with the PC1 and PC2 values of all the molecular subtype signatures genes

RND1 is a tumor suppressor that is frequently downregulated in cancer their expression correlates with the expression of the 70-gene poor prognosis signature, as previously mentioned in literature [99, 187]. Many biological processes related to receptor tyrosine kinases (RTK) are also found like the epidermal growth factor receptor (EGFR) pathway and peptidyl tyrosine phosphorylation. RTKs are known to regulate many signaling pathways like the MAPK and AKT signaling pathways which regulate cell proliferation, differentiation, inflammatory response, and apoptosis [188].

Fibroblast growth factor receptor 2 (*FGFR2*) and human epidermal growth factor receptor 2 (*HER2*) are both tyrosine kinase receptors that are frequently dysregulated in breast cancer. Studies have also shown that activating point mutations in *FGFR2* are associated with HER2-positive breast cancer, and may contribute to resistance to HER2-targeted therapies [189]. A study published in Clinical Cancer Research demonstrated that activating *FGFR2* mutations were associated with decreased response to trastuzumab and lapatinib, two HER2-targeted therapies commonly used in the treatment of HER2-positive breast cancer [190].

Finally, in HER2-positive breast cancer, SLIT/ROBO pathway is often dysregulated, leading to increased cell migration, invasion, and metastasis. One study published in Breast Cancer Research and Treatment found that the expression of *SLIT2* and *ROBO1* was significantly downregulated in HER2-positive breast cancer tissues compared to normal breast tissues. The study suggested that the loss of *SLIT2* and *ROBO1* expression may contribute to the aggressive phenotype of HER2-positive breast cancer [191].

Among the significant biological processes of gene ontology, *ERK1* and *ERK2* are kinase effectors of the *MAPK* cascades, whose pathways play a primary role in mediating cancer cell proliferation [192, 193]. More specifically, in BC both *ERK1* and *ERK2* show distinct patterns where higher expression of *ERK1* is associated with a good prognosis outcome and *ERK2* is associated with poor overall survival in patients [194]. Overexpression of *GRB7* is correlated with greater ER negativity, higher p53 immunopositivity, and other adverse parameters in breast cancers [195]. Similarly *MED1* is also overexpressed in breast cancer and promotes breast cancer cell proliferation and migration and may also serve as a novel target for therapy [196]. The *STARD3* gene is an important biomarker present in the HER2 amplicon [197, 198]. Indeed, several studies showed how the reduction of *STARD3* expression in HER2+ BC reduces their growth [199–201]. Recent studies also highlighted the prognostic and predictive value of *STARD3* protein expression in pathological complete response in HER2+ BC [202]. Notably, also other genes included in the 29 HER2 signature such as *MED24*, *FBXL20*, *MNX1* and *SIX1* have been shown to promote cancer cell proliferation and cell growth [203–205]. Interestingly, higher expression of *FGFR2* and *GSDMB* genes have been correlated to partial or no pathological complete response in early-stage HER2+BC patients [206, 207].

Peptidyl tyrosine phosphorylation is a key event in the activation of cellular signaling pathways that promote cell proliferation, differentiation, and survival. Protein autophosphorylation is often associated with tyrosine kinase receptors, including *HER2*, which are frequently overexpressed or amplified in breast cancer. Additionally, positive regulation of MAPK kinase activity has also been found to be increased in HER2-positive breast cancer [208]. Finally, positive regulation of transcription initiation from RNA polymerase II promoter has been found to be enriched in HER2-positive breast cancer, as HER2 signaling can activate transcription factors like NF-kB and AP-1, which regulate the expression of genes involved in cell growth and survival [209].

Cellular response to epidermal growth factor stimulus involves the binding of epidermal growth factor (EGF) ligands to the extracellular domain of the *EGFR*, which results in receptor dimerization and autophosphorylation of intracellular tyrosine residues [210]. Many studies have shown how inhibiting the EGFR signaling pathways through therapeutics like lapatinib and erlotinib increases tumor resistance in relapsing HER2-positive breast cancer [211]. Therefore, targeting this pathway could be a potential therapeutic strategy for HER2-positive breast cancer patients.

Additionally, when we explored the amount of variance captured in previously reported signature genes, we found that the expanded BP signature could capture the most amount of variance, which indicated that they could classify between the molecular subtypes, more specifically between the HER2 and non-HER2 molecular subtypes. The previously reported signature from Milioli *et.al.* was developed from the intrinsic subtype classifier developed by Perou *et.al.* [89]. This method used the concept of a Single Sample Predictor in order to classify the molecular subtypes into five types namely, Basal, Luminal A, Luminal B, HER2, and Normal-like subtypes. Additionally, we should also keep in mind that the variance was observed on Agendia's microarray gene expression samples, hence future studies can include calculating the variance observed on these samples on possibly more public datasets which includes all molecular subtypes.

Overall, the dysregulation or altered expression levels of these genes can influence key pathways and processes associated with tumor growth, metastasis, and response to therapy in HER2-positive breast cancer. These biological processes and pathways are closely related and interact with each other to drive HER2-positive breast cancer progression. Understanding the intricate interplay between these processes and pathways may provide new insights into the underlying mechanisms of HER2-positive breast cancer and offer novel targets for therapeutic intervention. For instance, targeting the ERK1/2 cascade or PI3K/AKT pathway may offer new strategies for HER2-positive breast cancer treatment. Additionally inhibiting *EGFR*, *PI3K*, or Rho GTPases may be effective in reducing tumor growth and metastasis.

Studies have shown that combining different therapies can be more effective than using them alone. For example, combining PPARA agonists with HER2-targeted therapies has been shown to enhance the efficacy of treatment in HER2-positive breast cancer. Future research could investigate other potential combinations of therapies that may improve outcomes for patients with HER2-positive breast cancer. Despite the success of HER2-targeted therapies, resistance remains a challenge in the treatment of HER2-positive breast cancer. Further research is needed to understand the mechanisms of resistance and to develop strategies to overcome it. For example, understanding the role of the PI3K/AKT pathway in mediating resistance to HER2-targeted therapies could help identify new targets for therapy.

In recent years, the definition of HER2-positivity has evolved as the clinical importance of heterogeneity of HER2-positive BC is increasingly emerging. HER2-positive breast cancer can vary in the extent and intensity of HER2 expression within the tumor, as well as in the degree of amplification. HER2 heterogeneity in breast cancer refers to the presence of different levels of HER2 expression or amplification within a single tumor. More recently, attention has been drawn to HER2-low tumors, which are those that have a low-level expression of the HER2 protein but above the threshold of HER2 negativity, and have no HER2 gene amplification [157, 158]. These tumors are HER2-negative by current diagnostic criteria, but recent studies have shown that they may have distinct clinical and biological characteristics that differentiate them from other fully HER2-negative tumors [159].

The presence of HER2 heterogeneity in tumors may result in diminished responses to conventional anti-HER2 therapies. Hence, validation of techniques that enable prospective identification of HER2 heterogeneity in tumors in order to tailor therapy appropriately is deemed necessary [212]. One of the studies previously by Schettini et.al. showed that HER2-low breast cancer tumors had a higher expression level than HER2-negative tumors [213].

With the evolving knowledge of the biology of the HER2-low tumors and their biomarkers, the clinical relevance of the HER2 classification system is also shifting to include HER2-low tumors as well, since they can also benefit from these HER2-targeted treatment therapies like trastuzumab deruxtecan [214, 215]. Some research is already underway through clinical trials, which showed that HER2-low metastatic patients had a partial response to trastuzumab-deruxtecan [216].

3.5 Conclusion

Our newly identified expanded 29-gene HER2 signature can further refine the molecular subtyping of HER2-positive tumors and eventually aid in improving treatment selection. This could involve exploring the role of specific genes or genetic alterations that are part of the expanded HER2 signature. Future studies can involve exploring potential therapeutic interventions targeting the identified pathways and genes. Additional studies can investigate on the efficacy of specific inhibitors or combination therapies in HER2-positive tumors, taking into account the molecular subtypes and genomic alterations associated with HER2-positive breast cancer.

3.6 Abbreviations

BP BluePrint
EBC Early stage breast cancer
DEA Differential expression analysis
CV Coefficient of variation
BC Breast cancer
IHC Immunohistochemistry
ER Estrogen receptor
PR Progesterone receptor
HER2 Human epidermal growth factor receptor 2
TNBC Triple-negative breast cancer
FF Fresh Frozen
FFPE Formalin-fixed paraffin-embedded
SD Standard deviation
QC Quality Control
WT-NGS Whole Transcriptome Next-generation Sequencing
GEA Gene expression analysis
FC Fold change
ES Effect size
GSEA Gene set enrichment analysis
GO Gene Ontology
IQR Interquartile Range

3.7 Author's Contributions

AE and AB collected, analyzed the data, interpreted, visualized the results, and wrote the manuscript. MMK and RB interpreted the results. JH and LJV reviewed the manuscript. LM and AG designed and supervised the research study and interpreted the results and reviewed the manuscript. All authors approved the manuscript.

3.8 Funding

This research did not receive any special grant from funding agencies in the public, commercial, or non-profit sectors.

3.9 Conflicts of Interest

The authors AE, AB, JH, MMK, RB, LV, LM, and AMG are employees of Agendia. Agendia is the commercial entity that markets the 80-gene signature as BluePrint.

AMG is a named inventor on the patent for the 80-gene signature used in this study. No writing assistance was utilized in the production of this manuscript.

Chapter 4

Investigating the Concordance in Molecular Subtypes of Primary Colorectal Tumors and their Matched Synchronous Liver Metastasis

A. Schlicker*, A. Ellappalayam*, I.J. Beumer*, H.J.M. Snel, L. Mittempergher, B. Diosdado, C. Dreezen, S. Tian, R. Salazar, F. Loupakis, F. Pietrantonio, C. Santos Vivas, M.M. Martinez-Villacampa, A. Villanueva, X. Sanjuán, M. Schirripa, M. Fassan, A. Martinetti, G. Fuca, S. Lonardi, U. Keilholz, A. M. Glas, R. Bernards, L. Vecchione

*Authors contributed equally to this work

International Journal of Cancer 2020 Oct 15;147(8):2303-2315

Abstract

To date, no systematic analyses are available assessing concordance of molecular classifications between primary tumors (PT) and matched liver metastases (LM) of metastatic colorectal cancer (mCRC). We investigated concordance between PT and LM for four clinically relevant CRC gene signatures. Twenty-seven fresh and 55 formalin-fixed paraffin-embedded pairs of PT and synchronous LM of untreated mCRC patients were retrospectively collected and classified according to the MSI-like, BRAF-like, TGFB activated-like and the Consensus Molecular Subtypes (CMS) classification. We investigated classification concordance between PT and LM and association of TGFBa-like and CMS classification with overall survival. Fifty one successfully profiled

matched pairs were used for analyses. PT and matched LM were highly concordant in terms of BRAF-like and MSI-like signatures, (90.2% and 98% concordance, respectively). In contrast, 40% to 70% of PT that were classified as mesenchymal-like, based on the CMS and the TGFBa-like signature, respectively, lost this phenotype in their matched LM (60.8% and 76.5% concordance, respectively). This molecular switch was independent of the microenvironment composition. In addition, the significant change in subtypes was observed also by using methods developed to detect cancer cell-intrinsic subtypes. More importantly, the molecular switch did not influence the survival. PT classified as mesenchymal had worse survival as compared to nonmesenchymal PT (CMS4 vs CMS2, hazard ratio [HR] = 5.2, 95% CI = 1.5-18.5, P = .0048; TGFBa-like vs TGFBi-like, HR = 2.5, 95% CI = 1.1-5.6, P = .028). The same was not true for LM. Our study highlights that the origin of the tissue may have major consequences for precision medicine in mCRC.

What's New?
No systematic analyses have assessed concordance of molecular classifications between primary tumors and matched liver metastases in metastatic colorectal cancer (mCRC). Here, the authors show that 40% to 70% of primary colon tumors cease to exhibit an epithelial-to- mesenchymal transition phenotype (EMT) at the transcription level in their matched liver metastasis (LM). While EMT-positive PT show worse outcome compared to EMT-negative PT, this is not true for LM. The data argue in favor of using the primary tumor for molecular analysis rather than distant metastases. Overall, this study highlights that tissue origin may have major consequences for precision medicine in mCRC.

4.1 Introduction

Colorectal cancer (CRC) is one of the most common cancers worldwide, with an estimated 1.2 million cases and over 600,000 deaths per year [217]. Due to its relatively asymptomatic progression, patients are frequently diagnosed with metastatic disease, which is associated with a five-year survival rate of around 10% [218]. Since biopsies and surgical tissue of metastatic lesions are difficult to obtain, treatment choice is mainly driven by the analysis of the archived primary tumor.

Coding mutations have been reported to be highly concordant between primary tumors (PT) and matched liver metastasis (LM) [219]. This is also the case for epigenetic and microbiome profiles [220–222]. In contrast, copy number profiles are discordant [223, 224] possibly pointing at larger genomic differences between PT and LM.

CRC can also be classified into different molecular subtypes based on gene expression patterns [75, 225–230]. The different molecular subtypes are characterized by the activation of different biological processes, such as microsatellite instability (MSI) and immune infiltration signaling, canonical epithelial signaling activation, metabolic dysregulation and mesenchymal characteristics. Although these subgroups have different prognosis, their predictive value, especially regarding the efficacy of targeted agents, remains under investigation. In this context, the MoTriColor consortium is currently exploring the efficacy of specific treatment strategies in molecularly defined CRC subgroups. Published and validated transcriptomic signatures were mainly developed in stage II and stage III disease, while metastatic CRC (mCRC) was not systematically investigated. Moreover, most of the data were generated from archival primary tumors.

Importantly, no systematic studies have investigated the concordance of classification of PT and LM according to different gene expression signatures. Few studies have reported about similarity in transcriptomic profiles between PT and matched LM, but they were inconclusive because of their small size and inclusion of synchronous and metachronous tumors [107, 108]. Here, we aimed to systematically study PT and their matched LM and to assess if gene expression signatures currently investigated in the MoTriColor consortium as well as the Consensus Molecular Subtype (CMS) classification [75] are concordant between matched pairs. Recently, Trumpi *et. al.*[231] reported that chemotherapy can affect the molecular classification of CRC. In addition, Isella *et. al.* [232] showed that the features of the mesenchymal subtype can be ascribed to its stromal component. Therefore, to avoid any bias that could be related to systemic treatment and different metastatic locations, we only included untreated primary CRC and their matched synchronous LM. Moreover, we investigated if the classification of some tumors, especially the ones classified as belonging to mesenchymal subgroups, could be influenced by the tumor microenvironment. Such data could help to understand if the transcriptomic molecular profiling of PT is sufficient to inform treatment choice or if molecular profiling of matching metastases is required to guide clinicians for individualized treatment recommendations in mCRC.

4.2 Materials and Methods

4.2.1 Patient Samples

We collected retrospectively samples of PT and matching synchronous LM from three different academic institutions: Catalan Institute of Oncology (ICO)-Bellvitge Biomedical Research Institute (IDIBELL - Barcelona), Istituto Nazionale dei Tumori (INT-Milan) and Istituto Oncologico Veneto (IOV-Padua). Samples were collected from treatment-naive cases with synchronous liver metastases at time of diagnosis and available clinical-pathological annotations. We restricted our study to these inclusion criteria to exclude potential effects of earlier treatments or different metastatic locations [231]. Clinico-pathological annotations included are reported in Table 4.1.

Based on these eligibility criteria, 82 matched mCRC pairs were collected. Of these, 24 fresh and 38 formalin-fixed paraffin-embedded (FFPE) tissue pairs were successfully processed and passed quality control. For 11 patients, we received both fresh and FFPE tissues of matched pairs from ICO-IDIBELL, which were used to investigate the influence of tissue preservation technique on gene expression (Figure 4.1).

Research was performed according to the principles of the Declaration of Helsinki. All patients were under clinical follow-up surveillance according to the Spanish or Italian National Guidelines. All patient samples and data were anonymized in accordance with national ethical guidelines [233] and study samples had Institutional Review Board approvals for the anonymized use of archival tissues. In particular, the Institutional Review Board of the INT approved the study (study number 117/15) and all alive patients signed a written informed consent. The Ethical Board (EB) of the IOV approved the study (study 2017/70) and the local EB of ICO-IDIBELL approved the study (PR030/17; study 2017/70). For ICO-IDIBELL, none of the patients signed a written informed consent form (ICF) because patients were dead or lost during the follow-up. The Spanish law allows using tumor samples collected before 2006 without an ICF if it is not possible to have it.

4.2.2 Microarray Processing and Quality Control

Total RNA was isolated from fresh-frozen and FFPE tissues with at least 30% of tumor cells. If possible, tissue enrichment was performed for samples that did not meet these criteria. RNA isolation and microarray processing were performed as described previously [225, 226, 228, 229]. For fresh tissue, RNA was isolated using the RNeasy micro kit (Qiagen, Hilden, Germany). Quality was assessed using an RNA 6000 Nano total RNA-Chip (Agilent Technologies, Santa Clara, California). Only

Table 4.1: Patients' characteristics for the successfully profiled matched pairs

		ICO		INT		IOV		p-value
		n	%	n	%	n	%	
Age at Diagnosis	median - range	59.3	53.4-73.3	64.4	39.5-78.5	70.0	35.6-76.8	0.015
Age at Surgery	median - range	62.6	52.6-75.1	64.4	39.5-78.5	73.7	35.6-76.8	0.022
Gender	Female	7	28.0	2	15.4	4	30.8	0.615
	Male	18	72.0	11	84.6	9	69.2	
Tumor Location	Right	7	28.0	3	23.1	4	30.8	0.055
	Left	7	28.0	5	38.5	9	69.2	
	Rectum	11	44.0	5	38.5	0	0.0	
Grade	G1	20	87.0	0	0.0	0	0.0	<0.001
	G2	0	0.0	6	46.2	10	76.9	
	G3	3	13.0	7	53.8	3	23.1	
Diameter	median - range	47.5	20-60	40	16-75	35	12-80	0.605
MSI status	MSS	2	8.0	12	92.3	9	69.2	<0.001
	MSI	0	0.0	1	7.7	1	7.7	
	Missing	23	92.0	0	0.0	3	23.1	
BRAF	Wild	0	0.0	10	76.9	12	92.3	<0.001
	Mutant	0	0.0	1	7.7	1	7.7	
	Missing	25	100.0	2	15.4	0	0.0	
KRAS	Wild	2	8.0	6	46.2	6	46.2	<0.001
	Mutant	0	0.0	6	46.2	7	53.8	
	Missing	23	92.0	1	7.7	0	0.0	

Note: Twenty-five matched pairs were provided by ICO (Catalan Institute of Oncology, Barcelona, Spain) while 13 matched pairs were provided both by INT (Istituto Nazionale dei Tumori di Milano, Italy) and IOV (Istituto Oncologico Veneto, Padua, Italy), respectively. The following clinical variables were considered: age at diagnosis, age at surgery, gender, tumor location (right side colon, left side colon, rectum), tumor grade (G1, G2, G3), tumor diameter, MSI-status, BRAF status, KRAS status. P values are referring to differences in distribution of clinical variables across the three different centers. Abbreviation: NA, not applicable.

Figure 1

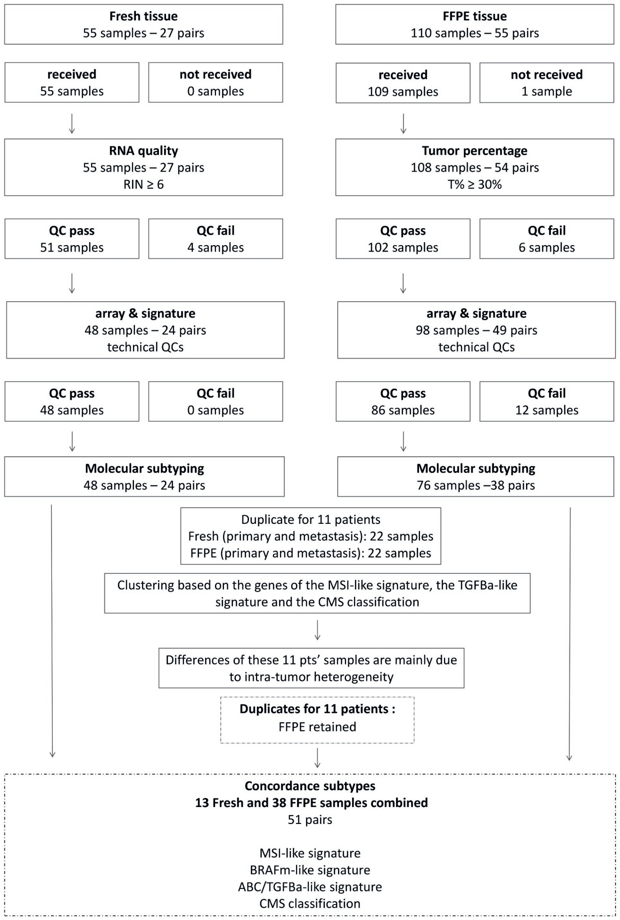


Figure 4.1: Workflow design of the study. A total of 55 fresh samples and 110 FFPE samples were collected, ending in 82 matched pairs. For 11 matched pairs, both fresh and FFPE tissue were collected. Upon RNA quality control and tumor content evaluation, samples that fulfill the criteria were processed on the array. Only successfully profiled PT with their corresponding matched LM were further analyzed for the different signatures. Further, clustering of these 11 pairs based on the genes belonging to the MSI-like and TGFBa-like signature as well as the CMS classification was performed. Because differences observed among those 11 pairs were mainly due to intratumor heterogeneity, the 11 FFPE matched pairs were retained for further analyses and combined with 13 fresh and 27 FFPE pairs. Therefore 51 matched pairs were finally molecularly classified based on the MSI-like signature, the BRAF-like signature, the ABC/TGFBa-like signature and the CMS classification

samples with RIN ≥ 6 were included in further analyses. Two hundred nanograms of total RNA were reverse transcribed, amplified and labeled with either Cy3 (sample) or Cy5 (reference sample) using the QuickAmp Labeling kit (Agilent Technologies), and subsequently purified using the Qiagen RNeasy mini kit. Cy3-labeled cDNA and Cy5-labeled cDNA were pooled (equimolar) and hybridized to the microarray.

For FFPE tissues, RNA was isolated using the RNeasy FFPE kit (Qiagen). Fifty nanograms of total RNA were reversed transcribed and amplified using the TransPLEX C-WTA whole-transcriptome amplification kit (RubicoFn, Ann Arbor, Michigan) and labeled with Cy3 using the Genomic DNA Enzymatic Labeling Kit (Agilent Technologies). For the microarray processing, cDNA was hybridized to custom full genome arrays (array design based on Agilent Catalog #G2514F) and washed according to the Agilent standard hybridization protocol (Agilent Oligo Microarray Kit, Agilent Technologies). Arrays were scanned with a dual laser scanner (Agilent Technologies).

Probes that showed nonuniformity of the signal as identified by the feature extraction software were omitted from further analyses. Image analysis of the scanned arrays was performed to quantify fluorescent intensities using Feature Extraction software version 9.5 and 11.5.1.1 (Agilent Technologies), for fresh and FFPE tissues, respectively. The feature extraction process included within-array normalization, which was performed using the default method for within-array normalization of Agilent microarrays (Lowess correction method using a linear polynomial [locally weighted linear least square regression]). Background correction was not applied. The final data sets contained expression values for 32,164 unique probes for our entire cohort. Expression values were calculated as sample/reference ratios using within-array normalized signals (log10[Cy3/Cy5]) for fresh tissue and represented the gMeanSignal intensities for FFPE tissue.

4.2.3 Data Analysis

Analyses and visualization of transcriptome data were performed in R [128] and RStudio. To investigate the contribution of the tissue preservation on gene expression levels, 11 unique patients' pairs for which both fresh and FFPE tissues were provided, underwent further analyses. In particular, 22 fresh samples (primary and metastasis) and 22 FFPE samples (primary and metastasis) were first median centered separately to remove probe specific bias and then combined together. Next, samples were clustered using the 64 genes of the MSI-like signature, the 277 genes of the TGFBa-like signature and the 266 unique genes of the CMS classification. For these 11 patients, we finally included the data from the FFPE tissue pairs, thus giving a total of 51 matched pairs. Hierarchical unsupervised clustering of these 51 matched pairs was performed on genome-wide transcriptome level. All clusterings (R version 3.1.3) were performed using the "ward.D2" method in the hclust function, and visualized using ggplot2 (version 3.0.01), dendextend (version 1.12.0), dplyr (version 0.8.3). Colors of the dendrogram bars were generated using ColorBrewer (version 2.0).

4.2.4 Signature and Classification Readout

Molecular subtyping was performed on the microarray transcriptome data of the 51 matched pairs, using five patented signatures for molecular subtype classification in colon cancer (MSI64-gene signature for fresh tissues, MSI-like FFPE signature, BRAF58-gene signature for fresh samples, BRAF mutant-like FFPE signature, ABC classification for fresh samples), one under patenting (TGFB activating-like signature) and the consensus classification (CMS classification). Additionally, we applied the CMScaller signatures, [234] which are based on cancer cell-intrinsic gene markers.

Proprietary Signatures

The MSI64-gene signature [229] was developed using fresh tissues to identify patients with a gene expression pattern similar to patients that were MSI-high by clinical tests, and categorizes tumors as microsatellite stable (MSS)-like or instable (MSI)-like. The BRAF58-gene signature [226, 228] was developed using fresh tissues to identify patients with a gene expression pattern similar to patients with BRAF V600E mutations, and categorizes tumors into BRAF wild-type (BRAFwt)-like and BRAF mutant (BRAFm)-like. Both signatures were also adapted for use in FFPE tissues. The ABC classification [225] was specifically developed for fresh tissues and identifies tumors as A-type (DNA mismatch repair-deficient epithelial subtype), B-type (proliferative epithelial subtype) or C-type (mesenchymal subtype). The TGFB activated (TGFBa)-like signature was developed specifically for FFPE tissue. The signature categorizes tumors into TGFB inactivated (TGFBi)-like and activated (TGFBa)-like, with the TGFBa-like group resembling the C-type and the TGFBi-like the AB-type of the ABC classification, respectively. The TGFBa-like signature, even if not yet published, is under investigation in the frame of the MoTriColor consortium. Classification results were generated using proprietary software based on MATLAB (MathWorks Inc, Natick, Massachusetts).

Publicly Available Classification

Probe sequences were aligned to the human transcriptome using NCBI-Blast to obtain the latest annotation information for generating the CMS calls. The CMS classification [75] was performed in R (version 3.1.1 [128]) and RStudio (version 0.98.994) using the CMS calls specific for the Agilent-platform. Additionally, the CMScaller signatures [234] were performed in R (version 0.99.1) which has as input a matrix with gene expression data and a CMS template.

4.2.5 Stroma Percentage and Microenvironment Assessment

The stroma percentage of the FFPE tissue slides were visually scored in a blinded manner. The scoring percentages of the hematoxylin and eosin (H&E) stained 5 µM thick sections were scanned on an Aperio ScanScope XT (Leica Biosystems, Wetzler, Germany) and uploaded to the Aperio eSlide Manager (Leica Biosystems). Preexisting healthy tissue, necrotic and mucinous areas were excluded from the scoring. 1X amplification was used to determine the relative percentages corresponding to the desmoplastic stroma. The tumor epithelium areas were determined within the tumor field. Stromal percentage (surrogate) was defined as 100 minus the tumor percentage. Moreover, to estimate the composition of the tumor microenvironment, we utilized the Microenvironment Cell Populations-counter (MCP-counter) method [235] which allows a robust quantification from transcriptomic data of both immune and stromal cell populations in heterogeneous tissue.

4.2.6 Statistical Analysis

Data were analyzed using SPSS 22.0 for Windows (SPSS Inc. Chicago, Illinois). For all statistical analyses, a two-sided P-value of .05 or less was considered statistically significant.

Sample population homogeneity and normality of distribution were tested using, respectively, a Pearson chi-square statistic for categorical variables and an Independent Samples Kruskal-Wallis (KW) test for the continuous variables.

The overall concordance of molecular profile between PT and matched LM was estimated using categorical classifications for all gene expression signatures. For BRAF-like, MSI-like and ABC/TGFBa-like signatures, the switch between tumor types was calculated using generalized estimating equations to fit a repeated measures logistic regression.

Sankey plots were generated using R (version 3.6.1) and the package networkD3 (version 0.4).

The relationship between stroma percentage (S%) and tissue type (PT/LM) or molecular subtypes was investigated using a paired t test for paired PT and LM. For the molecular subtypes either a Mann-Whitney (MW)-test (TGFBa-like signature) or Kruskal Wallis (KW)-test (CMS classifier) was used. Spearman's rank-order correlation (SpCorr) served to measure the correlation between the S% in PT and the S% in LM. ΔS% was defined as the difference in S% between matched tissue pairs and calculated as ΔS% = S% of LM - S% of PT. An independent t-test was used to investigate the relationship between ΔS% and a Boolean variable indicating a switch or not in molecular subtype of the TGFBa-like signature or the CMS classifier.

Survival analyses were performed using R (version 3.6.1) and the packages survival (version 2.44) and survminer (version 0.4). The Cox proportional hazards model was used to analyze the association of molecular subtypes with overall survival (OS), which was defined as the time from surgery until death from any cause.

Kaplan-Meier curves were used to compare the survival distributions of the molecular subtypes with OS.

4.3 Results

4.3.1 Study Population

To gain insights into the concordance of the transcriptomic profiles of PT and their matched LM, we collected 82 matched mCRC samples. As summarized in Figure 4.1, 48 (= 24 pairs) fresh tissue samples were processed and all passed quality control (QC). The FFPE tissue cohort contained 76 samples (= 38 pairs) that were available for molecular subtyping. When we compared the success rate of sample processing on the gene expression array, we did not observe statistically significant differences between the fresh (94.4%) and the FFPE (87.8%) cohorts (P = .259).

We next sought to investigate if tissue preservation could have an influence on the gene expression read-out. To this end, we looked at the expression of genes belonging to the MSI-signature, the TGFBa-like signature and the CMS classification in the 11 patients' pairs for which we received both fresh and FFPE tissues. Unsupervised clustering of these pairs showed that samples derived from the same patients were clustering together irrespective of tissue type (Figures 4.2A and S4.5A,B). This effect was most apparent when considering the MSI-like and TGFBa-like signatures (Figures 4.2A and S4.5A). Therefore, we concluded that gene expression differences between samples from the same patient were mainly due to intratumor heterogeneity rather than tissue preservation method, as previously reported for other solid malignancies [117].

4.3.2 Primary CRC and Matched Liver Metastasis differ at Gene Expression Level

We next aimed to investigate if transcriptomic profiles of the PT differed from those of their matched LM. Considering that the tissue preservation method did not influence the transcriptomic profiles, we combined the fresh and FFPE pairs. As reported in Figure 4.1, the final cohort of 51 successfully profiled matched pairs derived from 13 fresh pairs and 38 FFPE pairs, combined together. For patient characteristics, see Table 4.1. Overall, the distribution of the major clinical-pathological characteristics was similar between the three centers, except for tumor grading (P < .001), with samples from ICO being mainly characterized by well-differentiated tumors.

Furthermore, unsupervised clustering of the transcriptome profiles of these 51 matched pairs showed two major clusters without any obvious correlation with molecular subtyping calls or categorical clinical-pathological variables. As reported in Figure 4.2B, we neither observed a clear separation of PT from LM, since each cluster was characterized by both tissue types, nor a homogeneous clustering of the PT and

Figure 2

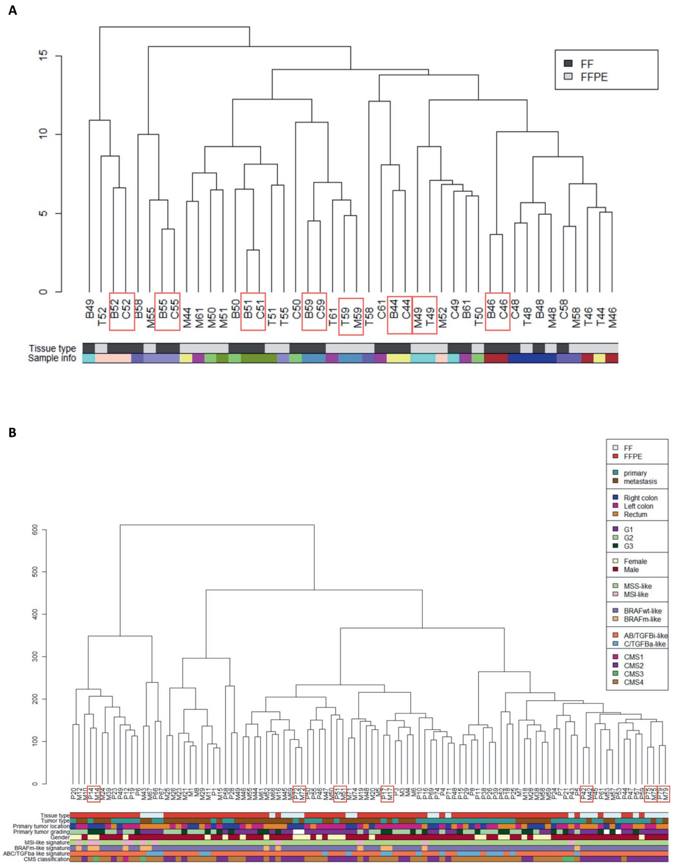


Figure 4.2: Clustering based on genes belonging to the MSI-like signature and transcriptome-wide gene expression. A, Clustering of the 11 matched pairs for which we received both fresh and FFPE tissue based on the genes belonging to the MSI-like signature (number genes = 64). Red rectangles highlight matched pairs that cluster together. T: primary tumor, FFPE tissue; M: matched liver metastasis, FFPE tissue; B: primary tumor, fresh tissue; C: matched liver metastasis, fresh tissue; Dendrogram bars: Tissue type: fresh tissue (dark gray), FFPE tissue (light gray); sample info: each color indicates samples belonging to the same patient. B, Unsupervised clustering based on genome wide transcriptomic profile of 51 matched pairs (13 fresh pairs and 38 FFPE pairs). Red rectangles highlight matched pairs that cluster together P: primary tumor; M: matched liver metastasis. Dendrogram bars: tissue type (fresh, FFPE); Tumor type (primary, metastasis); Primary tumor location (right colon, left colon, rectum); Primary tumor grading (G1, G2, G3); Gender (male, female); MSI-like signature (MSS-like, MSI-like); BRAFm-like signature (BRAFwt-like, BRAFm-like); ABC/TGFBa-like signature (AB/TGFBi-like, C/TGFBa-like); CMS classification (CMS1, CMS2, CMS3, CMS4)

their matching LM. Only 13% of the matched pairs (7 out of 51) clustered together indicating differences in the overall gene expression profiles. This exploratory analysis suggested that primary mCRC differ from their matched LM at the transcriptome-wide level.

4.3.3 The Mesenchymal Profile of Primary Tumors is not Always Retained in their Matched Liver Metastasis

Next, we aimed at comparing four established molecular gene signatures with potential clinical utility in PT and their matched LM. In particular, both PT and their matched LM were classified as MSI-like or MSS-like [229] and as BRAF m-like or BRAF wt-like [226–228]. Tumors were also classified as being TGFBa-like or TGFBi-like. For this purpose, we used the ABC classification [225] for the fresh tissue samples and the TGFBa-like signature for the FFPE tissue samples. It is important to note that the genes belonging to the C-group of the ABC classification and the TGFBa-like group are highly overlapping and they both identify tumors showing an epithelial-mesenchymal transition (EMT) phenotype. Therefore, we classified tumors as AB/TGFBi-like or C/TGBa-like to give a uniform nomenclature for fresh and FFPE samples. Finally, the CMS classification [75] was applied both to PT and their matched LM. A schematic overview of the concordance and changes of the different molecular subtypes between PT and their matched LM is reported in Figure 4.3 as well as in Table S4.5.

Overall, we observed high concordance for the BRAF-like and the MSI-like signatures between PT and LM, while lower concordance was observed for the TGFBa-like signature and the CMS classification. Four PT were classified as BRAF wt-like while their matched LM were classified as BRAF m-like. One PT was classified as BRAF m-like while its matched LM was classified as BRAF wt-like (Figure 4.3A). The overall concordance in terms of BRAF-like signature between PT and LM was 90.2%; the number of switches was not statistically significant (P = .177) (Table S4.1). Only one matched pair was not concordant in terms of MSI signature, with the PT classified as MSI-like and its matched LM as MSS-like (Figure 4.3A). The overall concordance of MSI-like signature between PT and LM was 98%; the number of switches was again not statistically significant (P = .313; Table S4.11).

Two pairs switched from AB/TGFBi-like in the PT to C/TGFBa-like in the LM (Figure 4.3A). More importantly, 10 out of 14 pairs (71%), whose PT were classified as C/TGFBa-like, were classified as AB/TGFBi-like in their matched LM showing an overall concordance of 76.5% (Table S4.1). This significant switch (P = .020) was also observed for the CMS4 classification (Figure 4.3B). Thirteen out of 32 pairs (40.6%), whose PT were classified as CMS4, were classified as CMS2 in their corresponding LM. One pair, whose PT was classified as CMS4, was classified as CMS3 in its matched LM. Furthermore, one pair switched from CMS1 in the PT to CMS4 in the LM and one pair switched from CMS3 in the PT to CMS2 in the LM. Within 16 PT

Figure 3

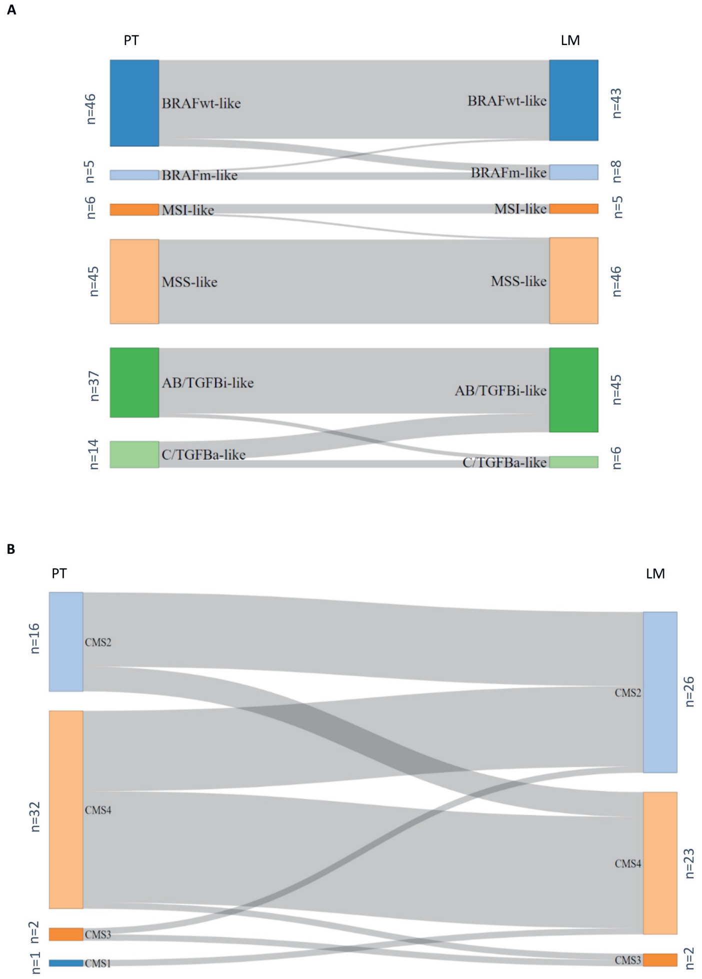


Figure 4.3: Molecular subtypes switch between PT and matched LM. A, Sankey plots showing the switch between PT and LM, in terms of MSI-like signature, BRAF-like signature, ABC/TGFBa-like signature. The molecular classification of the PT, with the corresponding number of pairs classified as such, is reported on the left of the Sankey plot. The molecular classification of the matched LM, with the corresponding number of pairs classified as such, is reported on the right of the Sankey plot. B, Sankey plots showing the switch between PT and LM in terms of CMS classification. The molecular classification of the PT, with the corresponding number of pairs classified as such, is reported on the left of the Sankey plot. The molecular classification of the matched LM, with the corresponding number of pairs classified as such, is reported on the right of the Sankey plot

that were classified as CMS2, four pairs switched to CMS4. These results indicated a 60.8% overall concordance for the CMS classification between PT and matched LM, with major significant switches (P = .050) regarding the mesenchymal subtype. Overall, the switches observed, both for the C/TGFBa-like signature and the CMS classification, indicate that the mesenchymal profile of 41% or 71%, depending on the classification used, is not retained in their matched LM.

4.3.4 The Loss of Mesenchymal Profile between Primary Tumors and their Matched Liver Metastases is Independent of the Tumor Microenvironment

Tumor microenvironment and in particular tumor stroma might play a role in determining the mesenchymal transcriptional profile of CRC [232]. We therefore investigated the composition of the microenvironment in order to better understand the switches of the mesenchymal profile observed between PT and their matched LM. To this end, we first quantified the stromal percentage in our samples. Because of a lack of further available tissue from the fresh pairs, we only considered the FFPE matched pairs for this analysis.

Overall, we observed similar population means of stroma percentages in PT and LM (P = .097). Also, no correlation was observed between stroma percentages in the matched pairs (rho =-.196, P =.252). We did not observe an association between stroma percentage (S%) and CMS classification (P = .127). However, we observed a significantly higher S% with the TGFBa-like subtype (P = .04). To understand if a difference in S% could be associated with a switch from PT to LM in CMS or TGFB classification, we analyzed if the difference in stroma percentage between PT and LM calculated as the S%LM - S%PT (ΔS%) was different between switchers and nonswitchers. We did not observe a significant difference in ΔS% between nonswitchers and switchers neither for the TGFBa-like signature (P = .607) nor for the CMS classification (P = .076), indicating that the difference in S% is not associated with a switch in molecular subtype classification. To confirm this observation, we further used MCP-counter in order to robustly quantify the absolute abundance of both immune and stromal cells using the transcriptomic data of our 51 matched samples. As reported in Figure S4.6, we did not observe a systematic difference in microenvironment composition between PT and LM. More importantly, no pattern was observed among the samples based on the CMS classification.

In addition, because a dedicated translation of the CMS classification to metastatic organs of CRC remains pending and by considering that gene expression signals might be strongly influenced by the organ of origin, we used the CMScaller classification to compare tumor classification between PT and LM. Overall, as reported in Figure S4.7, we observed a better distribution of the different subtypes both in the PT and LM, with more tumors classified as CMS3 and CMS1 as compared to the CMS classification. Nevertheless, the subtype assignments varied significantly

(Fisher test, P = .0033) between primary and metastasis.

In summary, these results indicate that the switch observed between PT and LM was not influenced by the stromal component, both evaluated as stromal percentage and by transcriptomic prediction, and that the classification of PT and matched LM are significantly different both by using the CMS classification and a classification that is based on cancer cell-intrinsic gene markers as the CMScaller.

4.3.5 The Molecular Profile of the Primary Tumor Determines the Outcome of mCRC patients Independent from the Molecular Profile of their Matched Liver Metastases

Next, we investigated if differences observed in CMS classification and TGFBa-like signature between PT and LM could affect patient overall survival (OS). Median OS (mOS) for PT was 165.9 months vs 37.3 months in CMS2 and CMS4, respectively (HR = 5.2, 95% CI = 1.5-18.5, P = .0048; Figure 4.4A) and 51.6 vs 24.0 months for TGFBi-like vs TGFBa-like, respectively (HR = 2.5, 95% CI = 1.1-5.6, P = .028; Figure S4.8A). These results confirmed that tumors classified as positive for a mesenchymal phenotype have a worse prognosis when compared to tumors classified as nonmesenchymal [75, 230]. In contrast, no mOS differences were observed among LM classified as CMS2 vs CMS4 (51.6 vs 42.1 months, respectively, HR = 1.5, 95% CI = 0.7-3.5, P = .28; Figure 4.8B) and TGFBa-like vs TGFBi-like (59.7 vs 45.4 months, respectively; Figure S4.8B). Finally, when we compared matched pairs that switched phenotype with the ones that did not switch phenotype, we did not observe major differences. Even if exploratory, these analyses confirmed previous observations [75, 230] that also report mesenchymal-like tumors to have a worse outcome compared to nonmesenchymal tumors. Interestingly, this effect was independent of the transcriptomic profile of their matched LM (Figures 4.4C and S4.8C).

4.4 Discussion

Currently, the treatment of mCRC is based on the molecular profile of the archived primary tissue and this is sufficient in most cases to identify mutations in genes that are predictive of response to conventional biological agents [219, 236]. Nevertheless, it has been shown that primary colon tumors and their matched metastases might differ in terms of copy number alterations [223, 224]. This raises the possibility that LM could have different actionable targets as compared to their matched PT. In addition, this could imply that the transcriptomic profile of PT and their matched LM might also differ. Different molecular classifications for CRC are currently under investigation for their predictive role in response to specific treatment strategies. It is therefore important to understand if the transcriptomic profile of archived primary tumor is sufficiently informative to predict the efficacy of certain treatment or the gene expression profile of their matched LM is required.

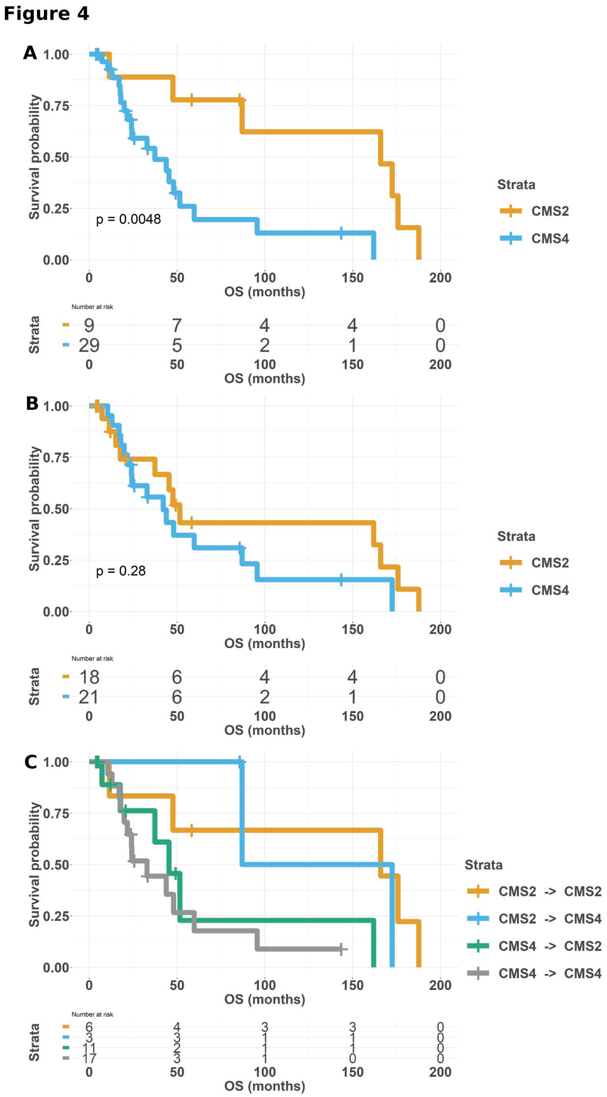


Figure 4.4: Estimate survival curves for the CMS classification. Kaplan-Meier plot of overall survival in months. A, Subjects were divided based on their primary tumor sample's CMS classification. CMS1 and CMS3 were excluded due to the small number of samples classified as such. The table below the survival plot contains the numbers of samples remaining in each group (strata) at each time point. B, Subjects were divided based on their metastatic sample's CMS classification; C, subjects were classified based on the CMS classification change from primary tumor samples to the matched metastatic sample. A two-sided P-value was not applied to C due to the small sample size of the groups

In this retrospective study, we analyzed the concordance of gene expression signatures with potential treatment implications in 51 matched samples of primary colon tumors and their matched synchronous LM. We observed that PT did not cluster together with their matched LM on transcriptome-wide gene expression level, indicating that the biology of PT might differ from the biology of their matched LM. When we looked at the concordance of different molecular subtypes, we found that both the BRAF-like and the MSI-like signatures were highly concordant between PT and matched LM. In contrast, major discordances were observed for the CMS classification and the TGFBa-like signature. Indeed, 41% of PT that were classified as CMS4 and 70% of PT that were classified as TGFBa-like lost their mesenchymal profile in the matched LM. These differences were statistically significant. Nevertheless, possibly due to the limited sample size, no major differences were observed for the other CMS subgroups. Both the TGFBa-like signature and CMS4 are characterized by high mesenchymal gene expression, which could be attributed to stromal cells as well as to cancer cells [232, 237–239]. When we looked at the differences in stroma percentage in our FFPE cohort, we did not observe a statistically significant difference in terms of stroma percentage in LM compared to their matched PT. We found that stroma percentage was statistically significantly associated with TGBa-like signature, but not with the CMS classification. Nevertheless, this was not associated with a change in the mesenchymal expression phenotype meaning that the switch observed between PT and LM was not influenced by the tumor stromal component. In addition, as also reported by Sandberg *et. al.* [240] there was no linear association between stroma percentage and CMS classification. Because a quantification of the stromal content represents a limited description of the tumor microenvironment, we additionally applied transcriptomic signatures to quantify the stromal contribution. The MCP-counter results showed no systematic differences in microenvironment composition between PT and LM, thus validating our findings of the visual stromal quantification of the FFPE samples. Importantly, CMScaller, which was designed to focus on expression of tumor cell-specific genes, also indicated that subtype assignments of many matched PT and LM were different as we have seen using the CMS classification. We are aware that a dedicated translation of CMS classifiers to colorectal tumors from different metastatic organs remains pending and that the CMScaller, as highlighted by Eide *et. al.*, [234] in its implementation is not recommended for use with samples with a different human stromal component than primary, like biopsies and metastatic tissue. Nevertheless, based on these results we can conclude that independent of the classification used, most of the PT classified as mesenchymal by gene expression lose this phenotype in their matched LM and this is independent of the tissue in which the tumor arises and its intrinsic microenvironment.

By looking at OS differences among molecular subgroups, we could confirm that PT classified as CMS2 and TGFBi-like have significantly longer mOS as compared to CMS4 and TGFBa-like PT tumors, respectively. Surprisingly, this effect was lost when the analyses were performed using LM as the basis for subgroup classification. Finally, no substantial differences were observed in terms of mOS between PT that switched their transcriptomic profile in the matched LM from epithelial to mesenchymal and from mesenchymal to epithelial, compared to tumors that did not change their expression profile. We are aware that the survival analyses need to be considered with caution because of the small sample size. No conclusions could be derived for other molecular subgroups due to low numbers of tumors classified as MSI-like and/or CMS1 and BRAF m-like and/or CMS3. In addition, survival estimates were not adjusted for relevant clinical variables, such as kind of treatment, radical resection of liver metastasis and the presence of other metastatic lesions. Due to our inclusion criteria, patient selection did not follow predefined criteria with respect to the treatment received. Moreover, 90% of patients received liver resection while 10% of patients received liver biopsies, thus implying a potential selection bias. Finally, with respect to the molecular classification, grouping our patients by considering other clinical variables would have led to even smaller subgroups and to inconclusive results. Despite these limitations, our cohort represents a unique series of synchronous mCRC where only LM were analyzed. By keeping in mind the limitations above reported, our data suggest that the transcriptomic profile of the PT is the driver of patient outcome rather than the profile of their matched LM. This may indicate that the PT has intrinsic properties that are constant despite changes induced by a different microenvironment. Our data argue in favor of using the PT rather than the distant metastases, for molecular analyses of mCRC.

4.5 Acknowledgements

We want to particularly acknowledge the patients and the Biobank HUB-ICO-IDIBELL (PT17/0015/0024) integrated in the Spanish National Biobanks Network for their collaboration and the patients of INT and IOV. This work was supported by the INTRACOLOR grant funded under the programme ERA-NET on Translational Cancer Research (TRANSCAN 2). Dr L. V. is a participant in the BIH-Charité Clinician Scientist Program funded by the Charité-Universitätsmedizin Berlin and the Berlin Institute of Health.

4.6 Conflicts of Interest

I. J. B., M. H. J. S., L. M, A. E., R. B. and A. M. G are employed by Agendia NV. S. T., A. M. G. and R. B. are one of the named inventors on patents for the gene signatures used in our study. R. B. is also shareholder in Agendia. C. D. is consultant by Agendia NV. A. E., L. M. and A. M. G reports grants from the European Union during the conduct of the study. A. S. is employee and shareholder of Bayer AG. L. V. is a member of the GI connect group, spouse is employee and shareholder of Bayer AG. U. K. reports grants and personal fees from Astra Zeneca, Bayer, BMS, Glycotope Merck Serono, MSD, Novartis, Pfizer, outside the submitted work. M. F. reports grants from Astellas and QED. M. F. is advisory board of Astellas and Tesaro. B. D., R. S., F. L., F. P., C. S.-V., M. M. V.-C., A. V., X. S., A. M., G. F., M. S. and S. L. declare no conflict of interest. Authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed. No writing assistance was utilized in the production of this manuscript.

4.7 Data Accessibility

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

4.8 Abbreviations

BRAFm-like	BRAF mutant like
BRAF	v-RAF murine sarcoma viral oncogene homolog B
BRAFwt-like	BRAF wild type like
CMS	consensus molecular subtype
CMS1	consensus molecular subtype 1
CMS2	consensus molecular subtype 2
CMS3	consensus molecular subtype 3
CMS4	consensus molecular subtype 4
CRC	colorectal cancer
EB	Ethical Board
EMT	epithelial to mesenchymal transition
FFPE	formalin-fixed paraffin-embedded
HR	hazard ratio
ICF	informed consent form
ICO-IDIBELL	Catalan Institute of Oncology - Bellvitge Biomedica Research Institute
INT	Instituto Nazionale dei Tumori
IOV	Instituto Oncologico Veneto
KRAS	Kirsten Rat Sarcoma Viral Oncogene Homolog
KW-test	Kruskal-Wallis test
LM	liver metastasis
mCRC	metastatic colorectal cancer
mOS	median overall survival
MoTriColor	Molecularly guided trials strategies in patients with advanced newly molecular defined subtypes of colorectal cancer
MSI	microsatellite instable
MW-test	Mann-Whitney
OS	overall survival
PT	primary tumor
QC	quality control
SpCorr	Spearman's rank-order correlation
TGFB	transforming growth factor-beta 1
TGFBa-like	transforming growth factor-beta 1 activating-like
TGFBi-like	transforming growth factor-beta 1 inactivating-like

4.9 Supplementary Information

(A)

Table 4.2: Concordance between PT and matched LM in terms of molecular profile
Binary response format describing the concordance of the BRAF-like signature (A), MSI-like signature (B) and the TGFBa-like signature (C) between PT and matched LM. Categorical response format describing the concordance of the CMS classification (D) between PT and matched LM. P-value has been generated by using Generalized Estimating Equations to fit a Repeated Measures Logistic Regression. $p<0.05$: differences between the tumor types.

		Metastasis			
		BRAFm	**BRAFwt**	**Total**	**Overall Concordance**
Primary	**BRAFm**	4	1	5	90.20% (p = 0.177)
	BRAFwt	4	42	46	
	Total	8	43	51	

(B)

		Metastasis			
		MSI	**MSS**	**Total**	**Overall Concordance**
Primary	**MSI**	5	1	6	98.00% (p = 0.313)
	MSS	0	45	45	
	Total	5	4	51	

(C)

		Metastasis			
		C/TGFBa	**AB/TGFBi**	**Total**	**Overall Concordance**
Primary	**C/TGFBa**	4	10	14	76.50% (p = 0.020)
	AB/TGFBi	2	35	37	
	Total	6	45	51	

(D)

		Metastasis					
		CMS1	**CMS2**	**CMS3**	**CMS4**	**Total**	**Overall Concordance**
Primary	**CMS1**	0	0	0	1	1	60.80% (p = 0.050)
	CMS2	0	12	0	4	16	
	CMS3	0	1	1	0	2	
	CMS4	0	13	1	18	32	
	Total	0	26	2	23	51	

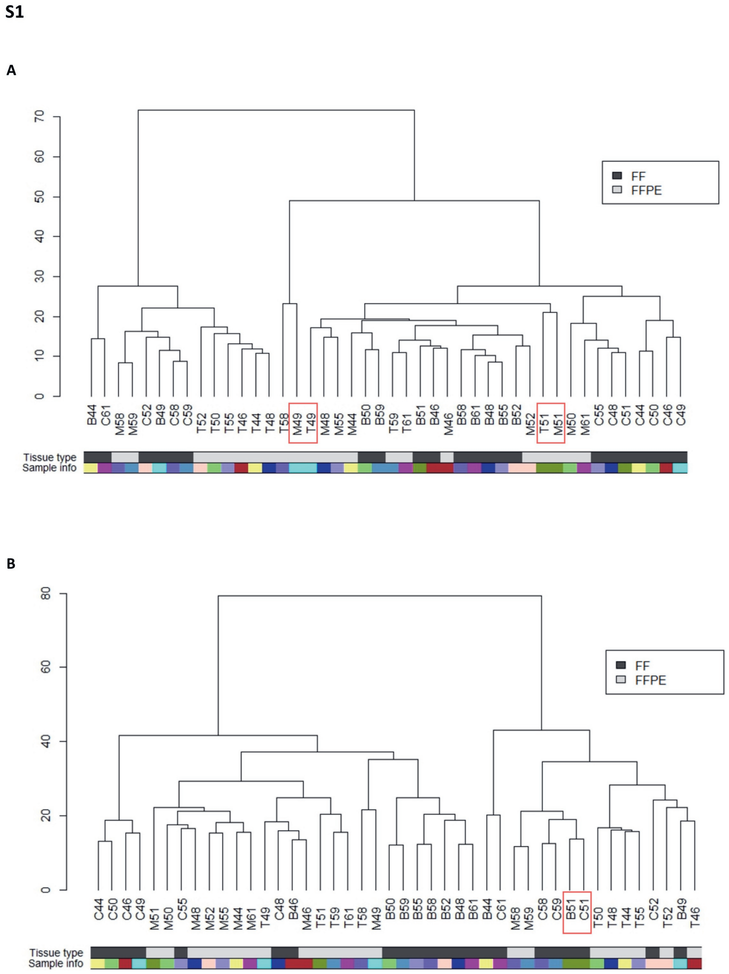


Figure 4.5: **Clustering based on genes belonging to the TGFBa-like and MSI-like signature** A. Clustering of the 11 matched pairs for which we received both fresh and FFPE tissue based on the genes belonging to the TGFBa-like signature (number genes=277). Red rectangle highlights matched pairs that cluster together. T: primary tumor, FFPE tissue; M: matched liver metastasis, FFPE tissue; B: primary tumor, fresh tissue; C: matched liver metastasis, fresh tissue; Dendrogram bars: Tissue type: fresh tissue (dark gray), FFPE tissue (light gray); sample info: each color indicates samples belonging to the same patient. B. Clustering of the 11 matched pairs for which we received both fresh and FFPE tissue based on the genes belonging to the CMS classification (number genes=266). Red rectangle highlights matched pairs that cluster together. T: primary tumor, FFPE tissue; M: matched liver metastasis, FFPE tissue; B: primary tumor, fresh tissue; C: matched liver metastasis, fresh tissue; Dendrogram bars: Tissue type: fresh tissue (dark gray), FFPE tissue (light gray); sample info: each color indicates samples belonging to the same patient.

S2

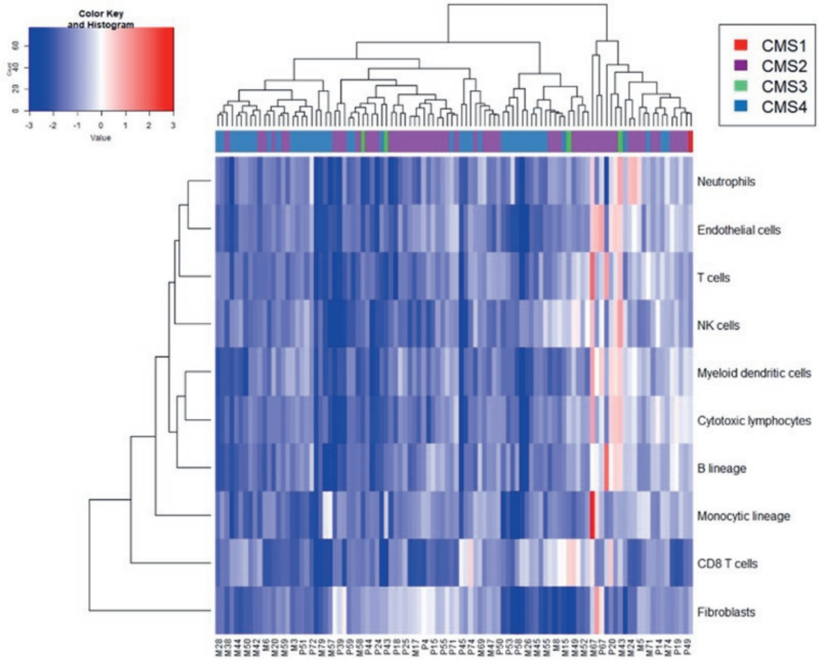


Figure 4.6: **MCP-counter scores for cell populations in colon samples and their association with the CMS signatures.**
Heatmap of MCP-counter scores for cell populations in colon samples (n = 102). Cell population values are in rows and samples in columns. Shades of red indicate high MCP-counter scores corresponding to a high abundance of the corresponding cell population; shades of blue correspond to low abundance of the corresponding cell population. The colored bar indicates the CMS signatures of the colon samples calculated using the CMS classifier package. The cell populations have an overall low abundance in the colon samples, with red highlighting high levels of microenvironment and blue low abundance levels.+ No pattern is observed among the samples based on the CMS subtypes and the samples with the higher expression in the microenvironment belong to CMS types CMS2, CMS3 and CMS4

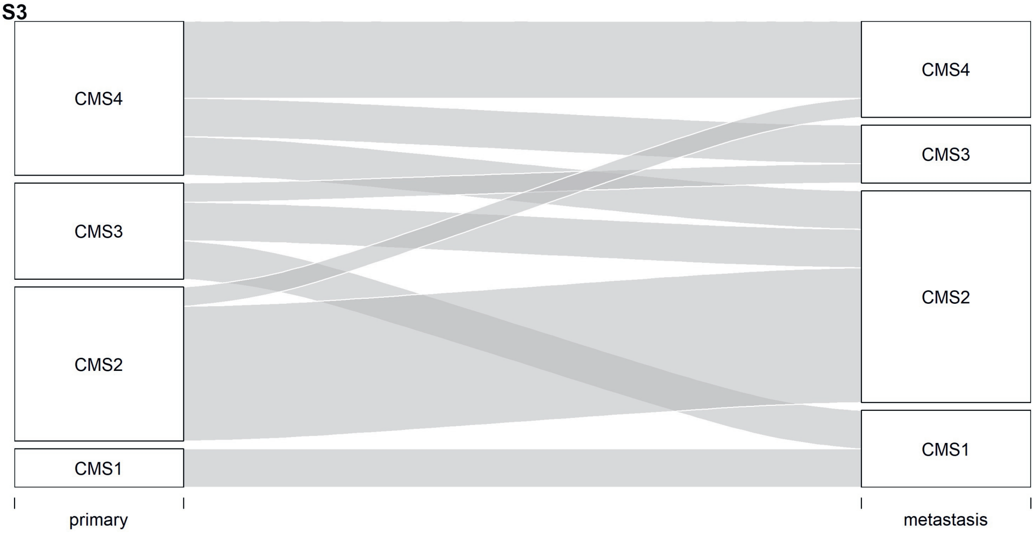


Figure 4.7: **CMScaller subtype assignments for PT and matched LM**
CMScaller subtype of PT and LM are indicated on the left and right side, respectively. Bars indicate the change of subtype assignments for different samples with bar width corresponding to the number of samples represented. As can be seen, the majority of PT samples were classified as CMS2 and CMS4. Subtype assignment for LM differs for many samples.

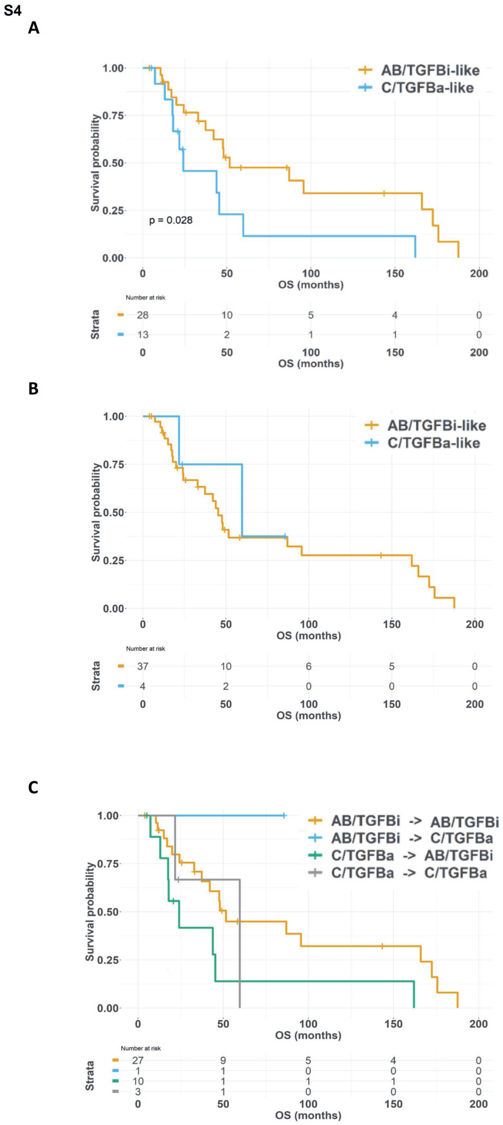


Figure 4.8: **Estimate survival curves for the TGFB-signature**
Kaplan-Meier plot of overall survival in months. A: Subjects were divided based on the TGFB-signature classification of their primary tumor samples. The table below the survival plot contains the numbers of samples remaining in each group (strata) at each time point. B: subjects were divided based on the TGFBsignature classification of their metastatic samples; C: subjects were classified based on the TGFB-signature classification change from primary tumor samples to the matched metastatic sample. A two-side p-value was not applied to Supplementary Figure 4B and 4C due to the small sample size of the groups.

Chapter 5

Molecular Subtyping of Triple-Negative Breast Cancer using Proteomics Data

Architha Ellappalayam, Cristina Furlan, Vitor A.P. Martins dos Santos, Maria Suarez Diez and Edoardo Saccenti.

5.1 Abstract

Triple-negative breast cancer (TNBC) is a highly aggressive and heterogeneous subtype of breast cancer with diverse molecular and genetic characteristics. Understanding the heterogeneity within TNBC is crucial for developing targeted therapies and improving patient outcomes. Proteogenomic approaches have gained importance in breast cancer research to understand how tumors develop and progress at the molecular level. By analyzing proteomics datasets, which provide a comprehensive view of protein expression, we aim to identify subgroups within TNBC and gain insights into their underlying biology. In this study, we utilized Similarity Network Fusion (SNF) analysis to identify subgroups within the triple-negative breast cancer (TNBC) subtype. The proteomics data from Anurag et al. and Krug et al. were used as the test and validation datasets, respectively. SNF combines multiple similarity networks into a single integrated network, capturing the underlying biological relationships. Our study identified two distinct subgroups within triple-negative breast cancer (TNBC), with one subgroup characterized by enriched pathways and processes specific to TNBC patients. Another subgroup exhibited pathways and processes associated with both HER2-type cancers and TNBC. The heterogeneity of TNBC and the importance of HER2 expression as a distinct subtype within this category are emphasized in our study. Further investigation of the HER2 subgroup within TNBC could provide valuable insights into its biology and potential therapeutic strategies.

5.2 Introduction

Breast cancer is a complex and heterogeneous disease with significant public health implications worldwide [241–244]. The heterogeneity in breast cancer poses challenges in diagnosis, treatment decisions, and predicting patient outcomes. Mapping and understanding the underlying mechanisms and causes of breast cancer heterogeneity is crucial for improved diagnosis, prognosis, and treatment strategies.

Table 5.1: Overview of the molecular subtypes of breast cancer and its prognosis, IHC status, and Treatment options. IHC - Immunohistochemistry, ER - Estrogen Receptor, PR - Progesterone Receptor, HER2 - Human epidermal growth factor receptor, PARP - Poly (ADP-ribose) polymerases

Molecular Subtype	**Prognosis**	**IHC Status**	**Treatment Options**
Luminal	Favorable	ER+/PR+, HER2-	Hormone therapy, targeted therapy
HER2-enriched	Less favourable	HER2+	HER2-targeted therapy, chemotherapy
TNBC	Poor	ER-, PR-, HER2-	Chemotherapy, targeted therapy, PARP inhibitors

The clinical subtypes of breast cancer are defined using Immunohistochemistry (IHC) which is a commonly used technique in breast cancer diagnostics, allowing for the detection and analysis of specific proteins in breast tissue samples. The molecular subtypes of breast cancer are identified using gene expression profiling of cancer biomarkers. The main molecular subtypes include Luminal, human epidermal growth factor receptor (HER2) enriched, and triple-negative breast cancer (TNBC). These subtypes have distinct molecular characteristics and varying prognoses. Luminal tumors are hormone receptor-positive with a favorable prognosis. HER2-enriched tumors express the HER2 protein and are aggressive. Triple-negative tumors lack hormone receptors and have poorer prognoses. An overview of the molecular subtypes of breast cancer is shown in Table 5.1.

TNBC is typically more aggressive than other subtypes of breast cancer and is also highly heterogeneous, with various TNBC subgroups identified based on molecular and genetic differences[90]. Some of these subgroups include mesenchymal-like, and claudin-low TNBC [113–115]. Each of these subgroups has distinct characteristics and clinical outcomes [112]. Previous studies have attempted to classify TNBCs into subgroups, such as the work by Burstein *et. al.* [116] and the Non-negative Matrix Factorization (NMF) classification of TNBC samples [245].

The heterogeneity of TNBC subtypes adds to the challenge of finding effective and specific targeted treatments for each molecular subtype. Failure to address the diverse subtypes within TNBC can impact the interpretation of clinical trial outcomes and limit the generalizability of results.

Proteogenomics is an emerging field that integrates genomic and proteomic data to enhance our understanding of cancer biology, identify potential therapeutic targets, and improve patient outcomes [246]. In breast cancer, proteogenomic approaches have become increasingly important as researchers seek to identify the molecular mechanisms driving tumor development and progression [247]. Proteogenomics also provides a more comprehensive view of the complex molecular landscape of breast cancer, which is characterized by significant heterogeneity [248, 249]. Exploring the heterogeneity at the proteomic level can enhance our understanding of TNBC's complexity and pave the way for personalized treatment strategies tailored to specific molecular subtypes

The challenge addressed here lies in the implementation of precision medicine strategies in breast cancer that consider the heterogeneity of TNBC subtypes and their associated treatment responses using proteomics data. Exploring the heterogeneity of TNBC across proteomics data is critical for gaining insights into its underlying biology, identifying potential therapeutic targets, and advancing personalized treatment strategies. Hence, in our study, we aimed to identify subgroups within TNBCs with the potential to inform treatment decisions. To identify distinct subgroups within the TNBC patient population, proteomics datasets were utilized.

We built upon publicly available data from two studies conducted under the auspices of the Clinical Proteomic Tumor Analysis Consortium (CPTAC). The CPTAC was launched to use proteomics technologies to improve the understanding of cancer biology and identify new targets for cancer therapy [250]. Anurag *et al.* identified proteogenomic markers associated with chemotherapy resistance and response in patients with TNBC. Krug *et al.* [251] is a comprehensive study that combined genomic and proteomic data to better understand the molecular mechanisms underlying breast cancer development and progression.

The results from this study can aid in improving the accuracy of breast cancer molecular subtype classification and provide insights into the underlying biology within TNBC subtypes, which can subsequently inform personalized treatment decisions.

5.3 Materials and Methods

5.3.1 Data

For the 71 samples from Anurag *et al.*, the peptides were labeled with 11-plex TMT reagents according to the manufacturer's instructions, and the acquired spectra were searched against the human proteome database resulting in the identification of 11063 proteins for the 71 samples. IHC and PAM50 assay along with Non-negative matrix factorization (NMF) were performed for molecular subtype classification. It is important to note that, in the PAM50 subtype classification, the Luminal A and the

Luminal B subtypes correspond to the Luminal molecular subtype, and the normal-like subtype corresponds to Luminal-type as well. Additional information is described in the Supplementary Data and Methods "Proteomic sample preparation" section of the Anurag *et al.* paper [252]. Since the study contained both pre- and on-treatment TNBC samples, only pre-treatment samples (n = 55 samples) were selected for this analysis to ensure consistency among all the study samples. The proteomics data can be retrieved via NCI Proteomics Data Commons with the accession identifiers PDC000408 (TNBC biopsies proteome raw files), PDC000409 (TNBC biopsies phosphoproteome raw files), and PDC000410 (TNBC PDX proteome raw files). In this study, we will refer to the dataset by Anurag et al. as the *Test* dataset.

The study by Krug *et al.* included 122 newly diagnosed breast cancer patients, out of which approximately 23% were classified as TNBC and therefore used in this study. The peptides were labeled with 10-plex TMT reagents according to the manufacturer's instructions and the acquired spectra were searched against the human proteome database resulting in the identification of 10107 proteins for the 122 samples. IHC and PAM50 classification was performed for molecular subtype classification. Additional information is described in the Supplementary Data and Methods "Proteomic sample preparation" section of the Krug *et al.* paper. Proteomics raw and characterized datasets can be retrieved through the CPTAC data portal (CPAC Data Portal) and at the Proteomic Data Commons. The accession number for the proteomic data at the CPTAC data portal is S060. The accession number for the proteomic data characterized by the Proteomic Data Commons is PDC: PDC000120. In this study, we will refer to the samples from Krug as the *Validation* dataset. Details of the patient data used in this study are detailed in Table 5.2. The test and validation datasets exhibit a significant overlap, with 9074 proteins overlapping between them.

Table 5.2: Overview of the Proteomics Dataset used in the studies by Anurag et.al. and Krug et.al. All samples analyzed in this table correspond to female individuals. IHC - Immunohistochemistry, TNBC - Triple-negative breast cancer, pCR - pathological complete response, PAM - Prediction Analysis of Microarray

Study	**Anurag et.al. Cancer Discovery (2022) (Test Dataset)**	**Krug et.al. Cell (2020)(Validation Dataset)**
Proteomic Data Commons ID	PDC000408	PDC000120
Total number of patients	59	122
Association to clinical trials	NCT02547987, NCT02124902	None
PAM50 - Basal like patients	34	29
IHC - TNBC patients	55	28
pCR Reponse data	Yes	No

5.3.2 Methods

Data Processing

The 55 proteome samples from the test dataset and the 29 samples from the validation dataset were selected based on Quality Control (QC) criterion. For each dataset, the samples were randomly divided into two equal subsets using R (version 3.6.1) [128]. Along with the proteome data, metadata information about the samples was also utilized in this study. The metadata included information like pathological complete response (pCR) to chemotherapy and residual cancer burden (RCB) which refers to the residual tumor burden in breast cancer patients after chemotherapy. The proteomics data were imported into the R programming environment for further analysis using the "readxl" package.

Similarity Network Fusion Analysis for subgrouping within TNBC subtype

Similarity Network Fusion (SNF) was applied to the proteomics data. SNF uses a machine learning algorithm to combine multiple similarity networks into a single integrated network that captures the underlying biological relationships. In our study, we applied SNF analysis to the Anurag (test dataset) and Krug *et. al.* (validation dataset) to find subgroups within the TNBC subtype. Pair-wise distances between samples are computed by considering the Euclidean distance between the protein profiles. A pair-wise distance matrix for each TNBC sample was computed which was then used to perform similarity network fusion. SNF was performed using SNFTool (version 2.3.1) [253] with the following parameter values: K (the number of neighbors) = 10, alpha (hyperparameter) = 0.5.

The results of the SNF analysis were hierarchically clustered using the "Complete Linkage" method using the stats packages and visualized using "ggplot2" and "gplots" packages [254]. Hierarchical clustering of the SNFTool results was performed using the "stats" package [128] and visualized with colored bars using the "dendextend" package [255]. The color scheme for the dendrograms was determined using the "randomcoloR" package.

Differential Abundance Analysis and Enrichment Analysis

Differences in protein abundances and their statistical significance were evaluated using the "limma" package [125]. Differences in protein abundance were considered significant when the adjusted p-value ≤ 0.05 and had a fold change value ≥ 1.5. Gene ontology enrichment analysis was performed and visualized using the "pathfindR" package [166]. The Reactome database [161] was utilized for pathway analysis on the differentially enriched proteins and visualized also using the "pathfindR" package.

The scripts used for the similarity network fusion analysis, the list of clustered samples by SNF analysis, differential abundance analysis, and visualization of the ontology and Reactome results are accessible in a GitHub repository .

5.4 Results

5.4.1 Subtyping of TNBC samples on the Test Dataset

Samples from the test dataset as clustered using SNF are shown in Figure 5.1

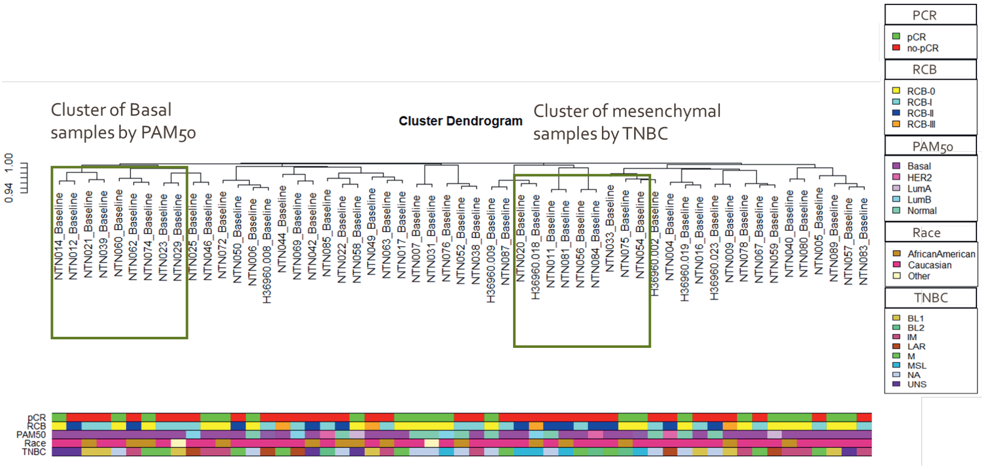


Figure 5.1: TNBC samples from the test dataset. The patient clusters are classified into Cluster 1, with clusters of basal samples, and Cluster 2, with clusters of mesenchymal samples. The color bars below the dendrogram include information on (from top to bottom): i) pathological Complete Response (pCR) ii) Residual cancer Burden (RCB) iii)PAM50 classification iv)individual race and v)TNBC status. Abbreviation (TNBC - Triple-negative breast cancer, BL - Basal-like, IM - Immunomodulatory, M- Mesenchymal, MSL - Mesenchymal stem-like, LAR - Luminal Androgen Receptor, UNS - Unspecified.)

Two clusters have been marked in the dendrogram. The first one labeled as "Cluster of basal samples by PAM50" contains nine samples in Cluster 1; the second one, labeled as "Cluster of mesenchymal samples by TNBC" also contains nine samples in Cluster 2.

To gain more insight into the clusters differential abundance analysis was performed between the two patient clusters. Up and down-regulated proteins of Cluster 1 are shown in Figure 5.2. the up-regulated proteins have higher in expression in Cluster 1 and the down-regulated proteins have lower expression in Cluster 1. Differential abundance analysis yielded 2082 proteins that were down-regulated and 2210 proteins that were up-regulated in Cluster 1. From the differentially abundant proteins, we performed Gene Ontology (GO) and Reactome pathway enrichment analysis. Among the GO terms we specifically considered the Biological Process (BP) ontology.

The enriched biological processes for the up-regulated proteins in Cluster 1 of

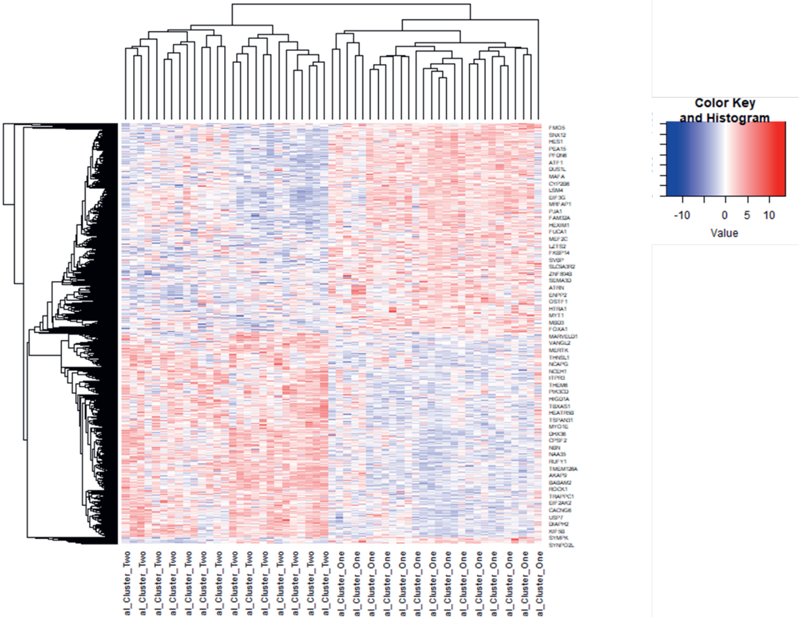


Figure 5.2: Biclustering Heatmap of the differentially enriched proteins between two clusters of samples of the test dataset. Values indicate log_2 foldchanges. Proteins with higher values in Cluster 1 are considered to be up-regulated and are shown in red; Proteins with lower values in Cluster 1 are considered to be down-regulated and shown in blue. Criteria for differential abundance: adjusted p-value < 0.05 and fold change ≥ 1.5 or $\leq 1/1.5$ ($log_2|FC| \geq log_2 1.5 \approx$ 0.58

the test dataset are shown in Figure 5.3. The enriched biological processes include *protein K-48 linked ubiquitination, protein ubiquitination, transcription by RNA polymerase II,* and *ubiquitin-dependent protein catabolic process.* Additional significant biological processes include mRNA splicing via spliceosome which plays a crucial role in gene abundance regulation, and its dysregulation has been associated with breast cancer [256].

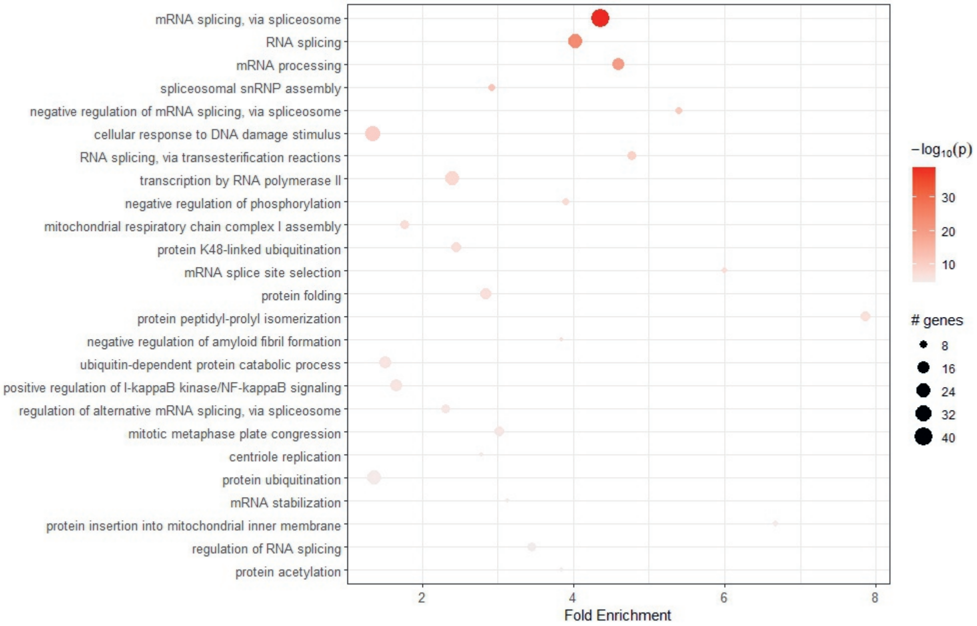


Figure 5.3: Biological Processes for the upregulated proteins of Cluster 1 in TNBC from the test dataset

Reactome pathways enriched among the up-regulated proteins of Cluster 1 of the test set are shown in Supplementary Figure 5.11. Some of the significant pathways are *mitochondrial elongation, initiation, translation, IL-12 signaling, mRNA Splicing, complex biogenesis,* and *mitochondrial biogenesis.* The above pathways represent a diverse set of cellular functions that are involved in various stages of cell cycle regulation, cellular stress response, and protein degradation pathways.

The biological processes for the down-regulated proteins of Cluster 1 of the test set are shown in Supplementary Figure 5.9. Among the down-regulated proteins the following biological processes are enriched: *the ubiquitin-dependent protein catabolic process and the regulation of small GTPases-mediated signal transduction* which is known to be associated with breast cancer, were found to be enriched. Dysregulation of these pathways can contribute also to the metastatic potential of TNBC breast

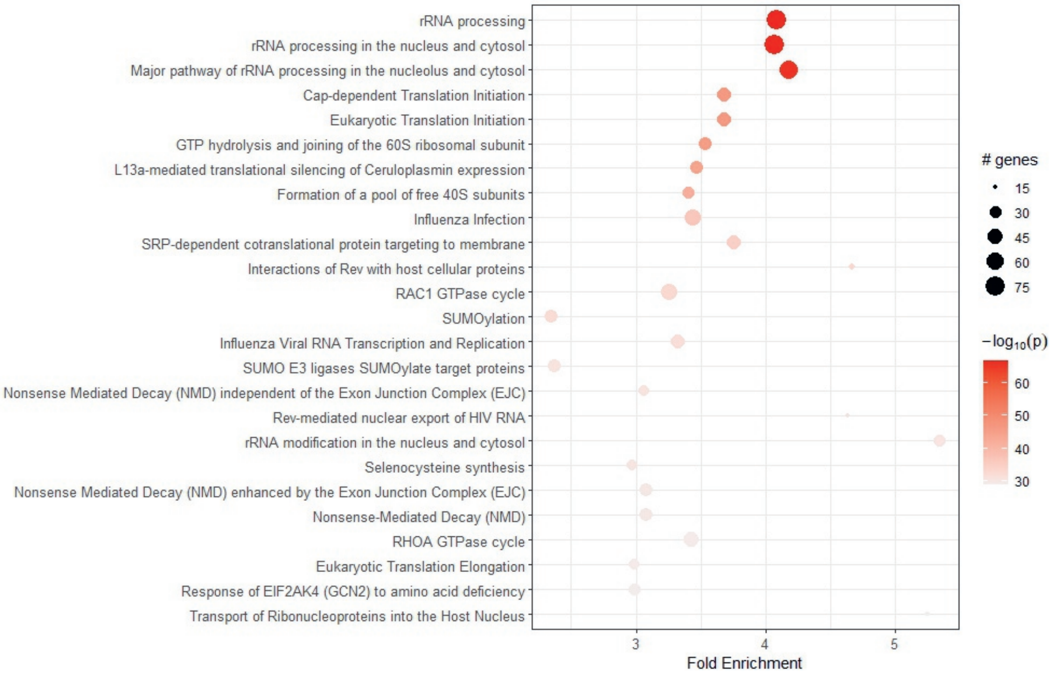


Figure 5.4: Reactome pathways for the downregulated proteins of Cluster 1 in TNBC from the test dataset

cancer cells [257].

The Reactome pathways enriched in the down-regulated proteins of Cluster 1 of the test set are shown in Figure 5.4. Some of the Reactome pathways include *Cap-dependent translation initiation, Eukaryotic translational initiation, Nonsense Mediated Decay (NMD) of the Exon Junction Complex, SRP-dependent cotranslational protein targeting to membrane, L-13 mediated translational silencing of Ceruloplasmin expression, SUMO E3 ligase SUMOylate target proteins, Influenza infection, formation of a pool of free 40S units, major pathway or rRNA processing in the nucleolus and cytosol* and *GTP hydrolysis and joining of the 60S ribosomal subunits*.

Based on the pathways listed, it appears that translation initiation, elongation, and peptide chain elongation were common processes in all types of breast cancer. Additionally, ribosomal biogenesis and RNA processing were also important pathways across all subtypes of breast cancer.

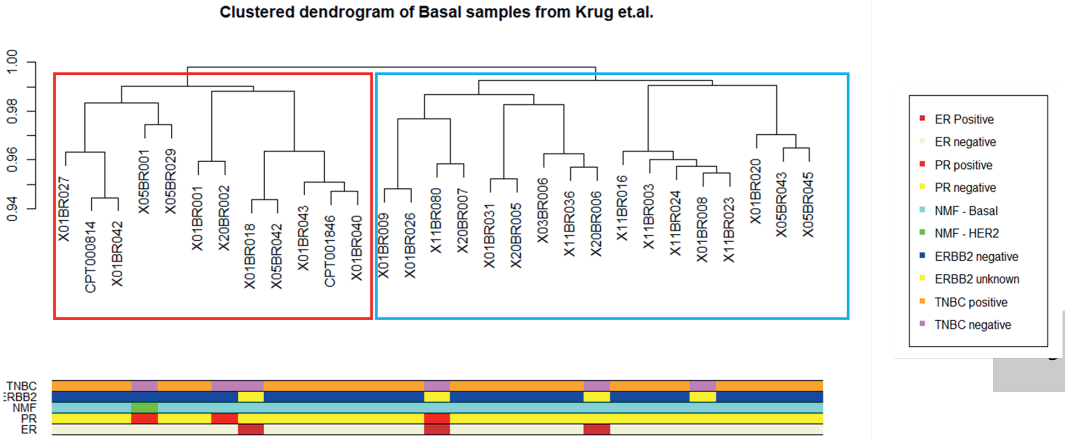


Figure 5.5: TNBC samples from the validation dataset. The color bars below the dendrogram include information on (from top to bottom): i) triple-negative breast cancer status ii) ERBB2 status iii) NMF classification iv) PR status v) ER status. Abbreviation (TNBC - Triple-negative breast cancer, NMF - Non-negative matrix factorization, ER - Estrogen, PR - Progesterone)

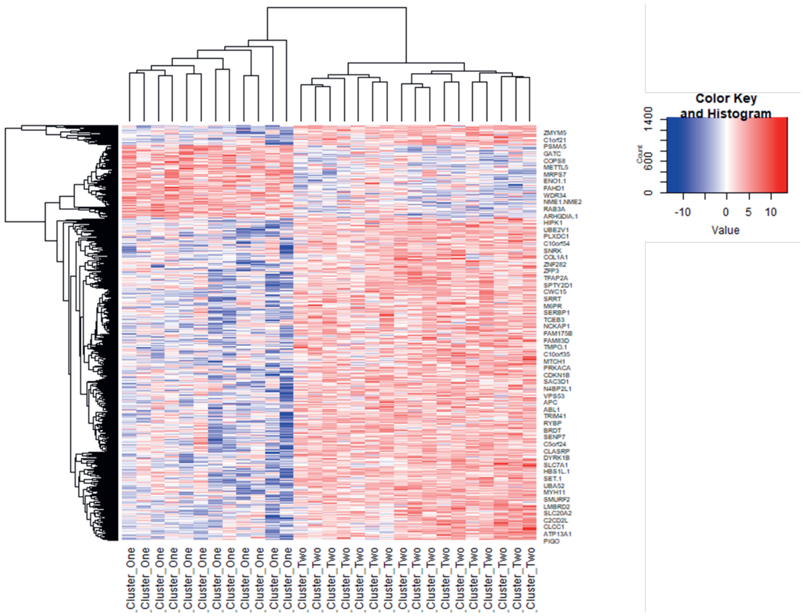


Figure 5.6: Biclustering Heatmap of the differentially enriched proteins between two TNBC clusters of samples of the validation dataset. Values indicate log_2 foldchanges. Proteins with higher values in Cluster 1 are considered to be up-regulated and are shown in red; Proteins with lower values in Cluster 1 are considered to be down-regulated and shown in blue. Criteria for differential abundance: adjusted p-value < 0.05 and fold change ≥ 1.5 or $\leq 1/1.5$ ($log_2|FC| \geq log_2 1.5 \approx 0.58$

5.4.2 Subtyping of the TNBC clusters on the Validation dataset

The Similarity Network Fusion classified the TNBC subset of the validation dataset into two clusters of 12 and 17 samples each. The clusters were visualized using dendrograms in Figure 5.5. From the colored bars, we could see that more ER and PR receptor-positive samples are present in Cluster 1, which has been marked in red in the figure. The NMF approach classified all the TNBC samples as Basal-like, except for one sample which was classified as HER2-type, which is also present in Cluster 1. Furthermore, the sample X05BR001 classified as HER2-type also had a TNBC negative and a PR positive status. For three of the samples in Cluster 2, which is marked in blue in the figure, the ERBB2 status was unknown. However, the TNBC status was positive for the three samples, indicating that these samples could belong to more than one molecular subtype and not just the TNBC one. Some distinct patterns were observed in the clustered dendrogram. To gain insights into these patterns, we conducted a differential abundance analysis between the two TNBC clusters, which is shown in Figure 5.6

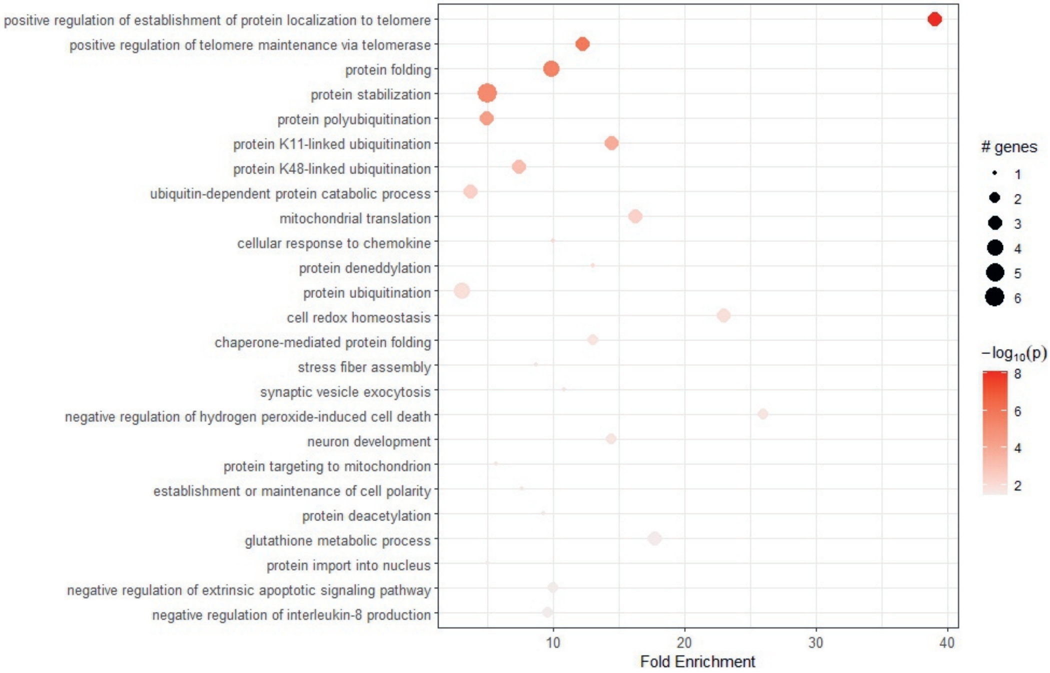


Figure 5.7: Biological Processes for the upregulated proteins of Cluster 1 in TNBC from the validation dataset

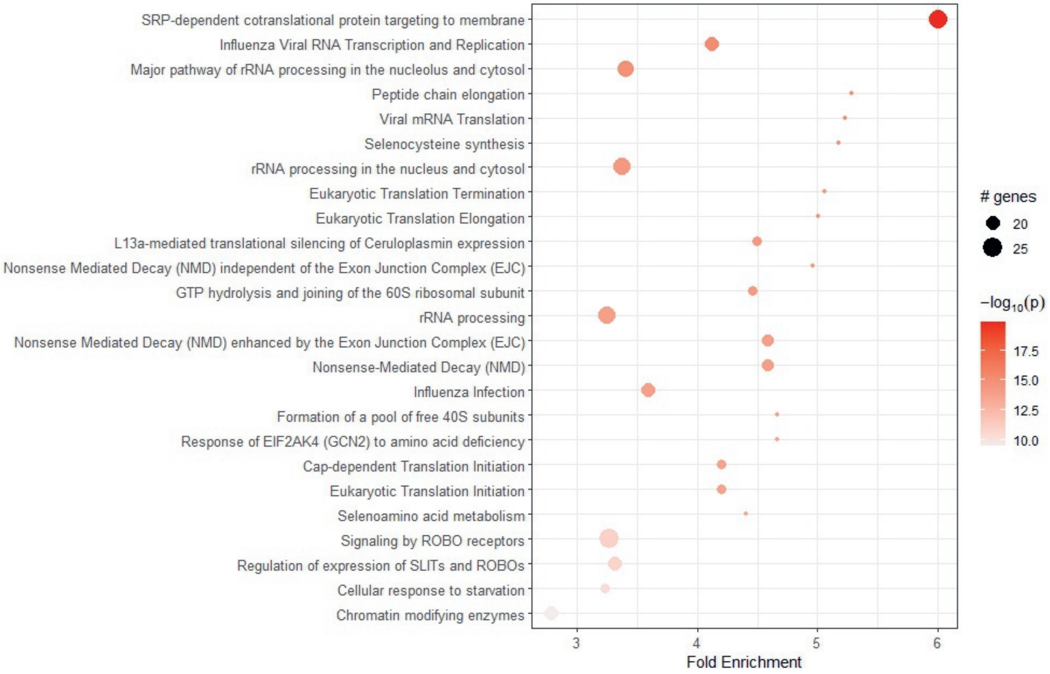


Figure 5.8: Reactome pathways for the down-regulated proteins of Cluster 1 in TNBC from the validation dataset

Differential abundance analysis yielded 181 proteins that were down-regulated and 839 proteins that were up-regulated in Cluster 1. Gene Ontology and pathway analysis were performed on both the up and down-regulated gene sets. The biological processes enriched in the up-regulated proteins of Cluster 1 in the validation dataset are shown in Figure 5.7.

The up-regulated proteins in both the test and validation datasets share many enriched biological processes like *protein ubiquitination, protein K-48 linked ubiquitination stabilization, and ubiquitin-dependent protein catabolic process,* which are all involved in regulating protein turnover and degradation. Additionally, significant biological processes like *positive regulation of telomere maintenance via telomerase* and *positive regulation of mRNA transcription by RNA polymerase II* were present. Many of the GO biological processes are associated with TNBC breast cancer and are also implicated in HER2 breast cancer [258, 259].

The Reactome pathways enriched in the up-regulated proteins of Cluster 1 in the validation dataset are shown in Supplementary Figure 5.12. The pathways listed here are more frequently associated with TNBC and HER2-positive breast cancer subtypes. The common Reactome pathways from the up-regulated proteins between these clusters are *mitochondrial elongation, initiation,* and *termination* pathways. Other enriched pathways enriched include *the regulation of ornithine decarboxylase* and *the metabolism of polyamines,* which are interrelated. *MAPK6* and *MAPK4* are members of the mitogen-activated protein kinase (MAPK) family, which regulate cell growth, differentiation, and survival. The above pathways represent a diverse set of cellular functions that are involved in various stages of cell cycle regulation, cellular stress response, and protein degradation pathways.

The biological processes for the down-regulated proteins of Cluster 1 in the validation dataset are shown in Supplementary Figure 5.10. Biological processes involved *Dysregulation of histone H2A monoubiquitination, midbody abscission, ubiquitin-dependent protein catabolic process* and *cellular response to DNA damage stimulus* which has been linked to the development and progression of breast cancer. The down-regulated proteins are shown to be associated very generally with breast cancer which can lead to genomic instability and the accumulation of mutations that contribute to tumorigenesis [260, 261].

The Reactome pathways enriched in the down-regulated proteins of Cluster 1 in the validation dataset are shown in Figure 5.8. The pathways are all related to different aspects of gene expression regulation, including transcription, mRNA processing, translation, and protein degradation. the other significant pathways include *SRP-dependent co-translation protein, Eukaryotic translation, L-13 mediated translational silencing,* and *Nonsense-mediated decay.* These pathways are associated more generally with breast cancer and not with any one molecular subtype of breast cancer.

5.5 Discussion

The enriched biological processes from the up-regulated proteins in both the test and validation datasets contain similar biological processes like *protein ubiquitination, protein K-48 linked stabilization, ubiquitin-dependent protein catabolic process,* and *transcription by RNA polymerase II*. The protein processes are all involved in protein turnover and degradation. Other significant processes include *positive regulation of telomere maintenance* and *mRNA splicing via splicosome*.

The biological processes from Gene Ontology have shown an association with both HER2-positive and TNBC breast tumors alike. HER2-positive breast cancers have been found to have higher telomerase activity, and targeting telomerase has been suggested as a potential therapeutic strategy for HER2-positive breast cancer [258, 259]. Dysregulation of protein polyubiquitination processes can contribute to the development of TNBC breast cancer. Protein folding and stabilization have been associated with both TNBC breast cancer and HER2-positive breast cancer. Aberrant protein folding and stabilization can lead to HER2 degradation or mislocalization, which can impair downstream signaling pathways. HER2-positive breast cancers have been found to have higher levels of chaperone proteins that facilitate protein folding and stabilization [262–264]. In HER2-positive breast cancer, HER2 overexpression can lead to upregulation of oncogenic proteins and cell proliferation. Biological processes for up-regulated proteins in both testing and validation TBNC samples show clear similarities, implying that our analysis identifies a common profile in this cluster.

The common features among the up-regulated Reactome pathways of the validation and test datasets include protein degradation, regulation of cellular processes, DNA damage response, and modulation of signaling pathways. Many of these pathways involve the degradation of specific proteins through ubiquitin-mediated proteolysis, which plays a crucial role in regulating protein levels and maintaining cellular homeostasis. Mapping to Reactome pathways showed enrichment on mRNA splicing, where one study identified a splice variant of HER2 called HER2Δ16 that is frequently overexpressed in HER2-positive breast cancer. In addition, dysregulation of "complex I assembly" may play a role in the development of HER2-positive breast cancer [265]. The Reactome pathways of the up-regulated proteins of the validation dataset, also show an association with both TNBC and HER2-type breast tumors, similar to the results that were observed from the test dataset. HER2-positive breast cancer cells have been found to have increased mitochondrial biogenesis and altered mitochondrial function, which may contribute to the aggressiveness of this type of cancer. [266]. IL-12 signaling has been associated with better prognosis in TNBC and HER2-positive breast cancers [267]. Studies have shown that inhibiting ODC sensitized TNBC cells to cytotoxic chemotherapy, suggesting that targeting the polyamine pathway could be a potential strategy to enhance the efficacy of chemotherapy for TNBC [268]. Regarding MAPK4 and MAPK6 signaling, a study showed evidence that MAPK4 plays a crucial role in the growth and survival of TNBC cells, and targeting MAPK4 may be a potential therapeutic strategy for TNBC [269]. From the above Reactome pathways, it is seen that not all the pathways are associated only with breast cancer, but with also HER-type cancer. The remaining significant pathways overlapped with the upregulated Reactome pathways from the test and validation datasets which show that there are similar patterns formed by the SNFTool in both these datasets, validating the findings.

Among the down-regulated biological processes in TNBC from the test dataset, many biological processes have an association with TNBC breast tumors. *Dysregulation of the Endoplasmic reticulum to the Golgi vesicle-mediated transport pathway* can contribute to the development of TNBC. Autophosphorylation of proteins, such as EGFR, can contribute to the development of TNBC by promoting cell growth and survival [270]. When observing the down-regulated biological processes of TNBC in the validation dataset, the biological processes are mostly involved among TNBC breast tumors, and not with the other molecular subtypes of breast tumors. TNBC breast cancer cells may be more sensitive to DNA-damaging agents leading to genomic instability and may have altered responses to DNA damage checkpoints [271]. Studies suggest that the dependency of triple-negative breast cancer cells on RNA splicing provides a potential therapeutic target for the treatment of TNBC [272].

The downregulated Reactome pathways from TNBC of the test and the validation dataset show pathways associated again with mostly TNBC breast tumors. SRP-dependent protein targeting could be a driver of breast cancer aggressiveness [273]. L13 ribosomal proteins bonded to UTRs that have been implicated in breast cancer, including those of the estrogen receptor alpha (ERα), cyclin D1, and *HER2* proteins. Dysregulation of these UTRs has been linked to the development and progression of breast cancer [274]. Dysregulation of the eukaryotic translation process has been implicated in breast cancer where high expression of some eukaryotic translation factors is associated with poor prognosis and may play important roles in breast cancer progression [275]. Nonsense-Mediated Decay (NMD) is dysregulated in breast cancer and contributes to tumor progression [276]. Altered mRNA splicing, including NMD, has been associated with breast cancer. The down-regulated pathways include protein synthesis and translation, RNA processing, viral infection and replication, chromatin organization and modification, cellular signaling, and cell cycle regulation. Many of these pathways involve the translation of mRNA into protein, including the initiation, elongation, and termination stages of translation.

From our study, two subgroups of TNBC breast cancer were identified, where one group encompassed mainly GO biological processes and significantly enriched pathways that belonged to TNBC breast cancer patients. Another subgroup of tumors that were observed in both the test and validation datasets was a cluster of patients that consisted of pathways and processes belonging to HER2-type cancers as well as TNBC cancers. Studies previously have identified a subset of TNBC patients characterized by low expression of HER2 (HER2-low), distinct from HER2-negative TNBC where they observed that HER2-low TNBC tumors had distinct molecular profiles and clinical characteristics, suggesting a potential role for HER2-low expression as a prognostic and predictive marker in TNBC [277, 278]. Overall, our study highlights the heterogeneity of TNBC and the significance of HER2 expression as a distinct subtype within this category. Further investigation of the HER2 subgroup within TNBC may provide valuable insights into the biology and potential therapeutic strategies for this specific subset of patients.

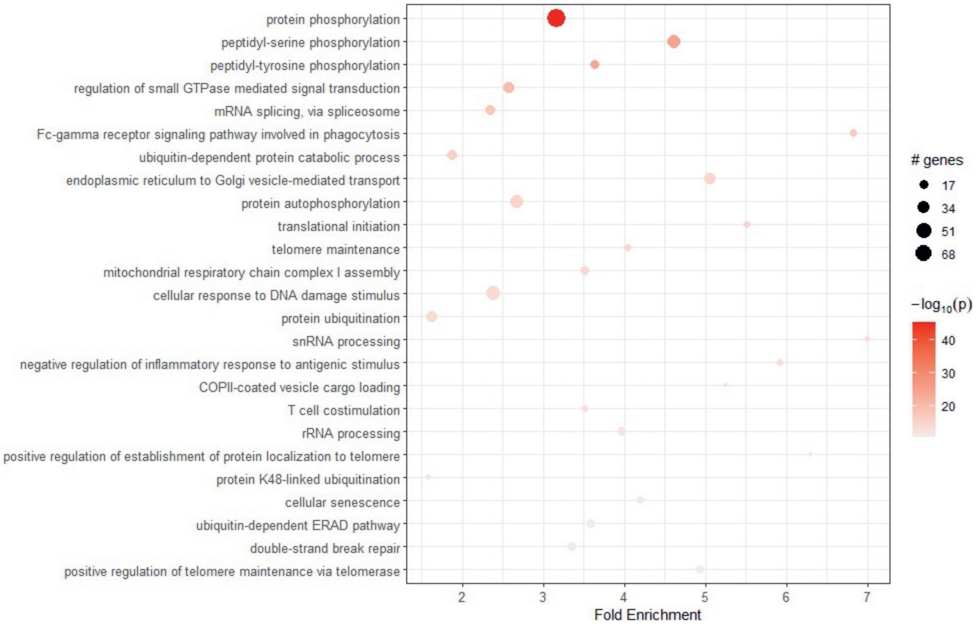
A commonality among all the previous studies is that the TNBC subtypes that had HER2 expression were HER2-low subgroups of patients. These HER2 2+ TNBC seems to have worse relapse-free survival, advocating for a dedicated clinical, biological, and therapeutic evaluation of this subgroup [279]. Traditional HER2-targeted therapies, such as trastuzumab (Herceptin), are not typically effective in HER2-low TNBC patients since the HER2 expression levels are low. However, there is ongoing research exploring alternative targeted therapies that may be effective in HER2-low subgroups, such as antibody-drug conjugates (ADCs) or other HER2-targeting agents.

5.6 Conclusions

Proteogenomics has the potential to transform precision medicine in breast cancer by enabling more precise diagnosis and treatment selection. From our analysis, we found that classifying subgroups within the TNBC molecular subtyping groups can aid in understanding the biology of these tumors, which will in turn improve the development of targeted therapies for these subgroups. It is imperative to emphasize the importance of incorporating diverse datasets and exploring innovative approaches to analyze them effectively. In our study, diverse datasets of TNBC were processed in different trials and used an SNF Analysis tool, which is a powerful method to integrate diverse datasets.

By using SNF analysis to proteomics data, we could classify TNBC subtypes into two groups based on their affiliation with TNBC and HER2 types. This could indicate that these patients may benefit from a combination of HER2-targeted and TNBC-targeted treatments. Future possibilities for this study include subgroup classification among the Luminal and the HER2 subtypes as well, to gain a better knowledge of subgroups within these molecular subtypes. This study also highlights the importance of having access to FAIR data to explore new data analysis since FAIR methods promote interoperability, enabling the combination of diverse datasets from different sources, which enhances the robustness and generalizability of the analysis results and advance the knowledge of breast cancer subtyping.

5.7 Supplementary Information



=

Figure 5.9: Biological processes for the downregulated proteins of Cluster 1 from the test dataset

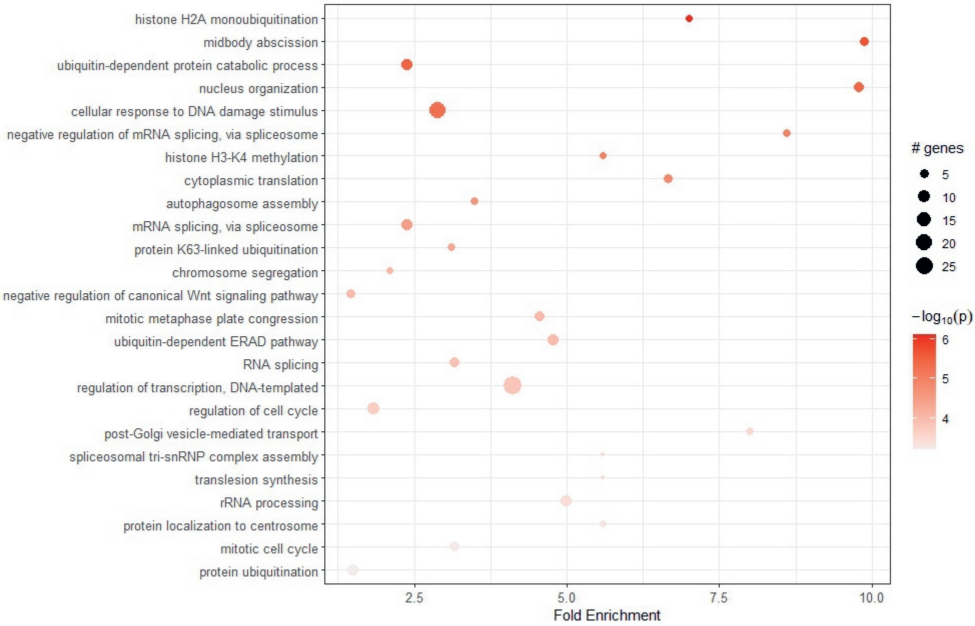


Figure 5.10: Biological processes for the downregulated proteins of Cluster 1 from the validation dataset

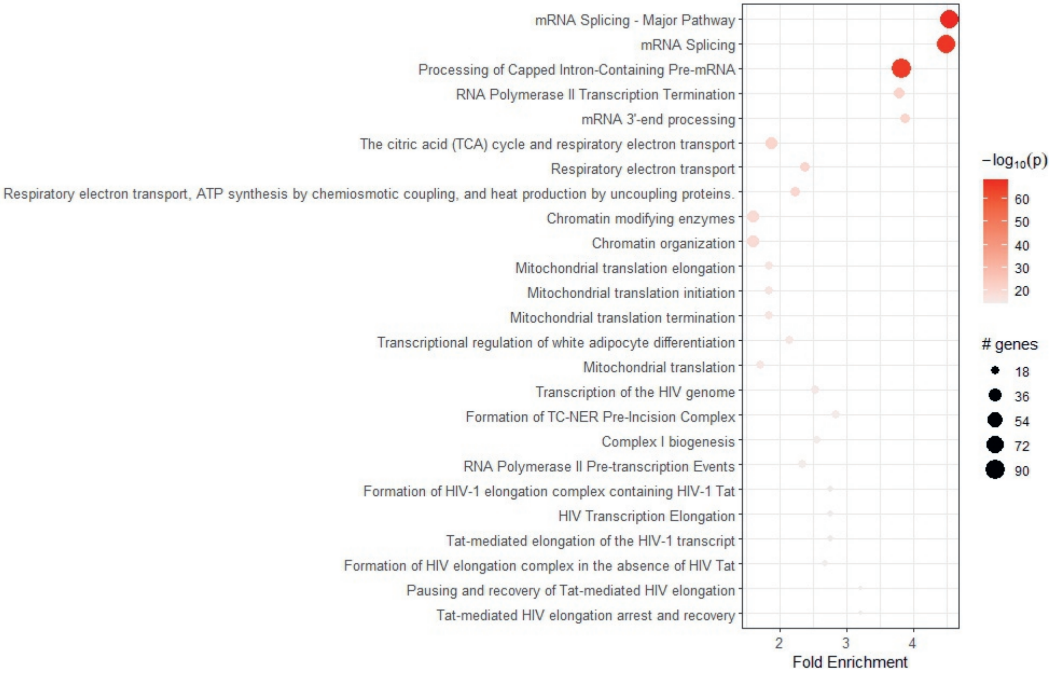


Figure 5.11: Reactome pathways for the upregulated proteins of Cluster 1 from the test dataset

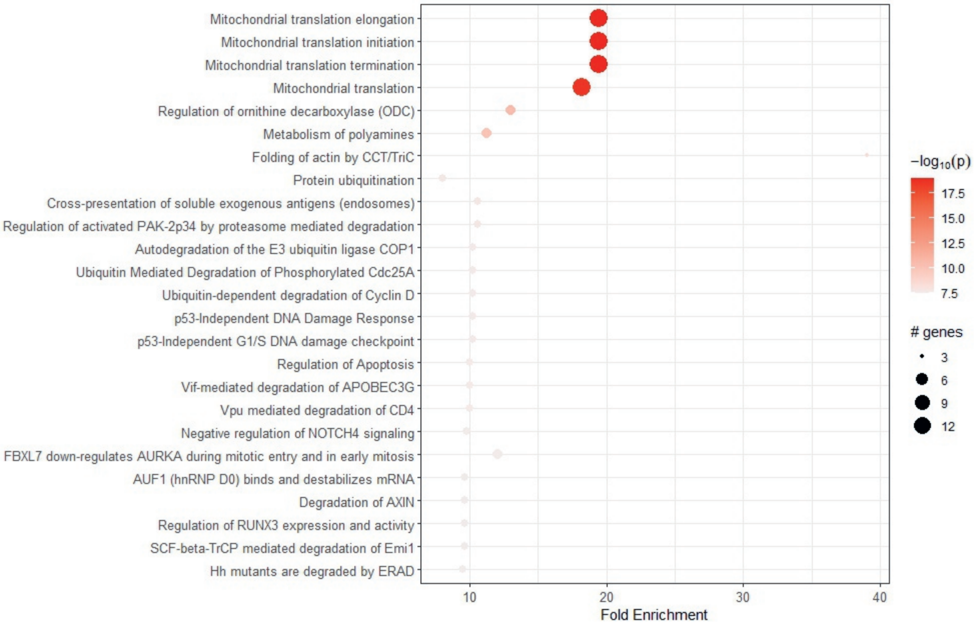


Figure 5.12: Reactome pathways for the upregulated proteins of Cluster 1 from the validation dataset

Chapter 6

General Discussion

Cancer, a multifaceted and enigmatic disease, continues to challenge the boundaries of medical understanding, driving the relentless pursuit of innovative solutions and personalized approaches in the battle against this formidable foe. In the realm of cancer research and treatment, molecular subtyping has witnessed remarkable advancements providing crucial insights into its underlying mechanisms and paving the way for personalized treatment strategies.

In this thesis, I address different research aims to **Refine Molecular Subtyping Diagnostics in Breast and Colon Cancers using Gene Expression and Proteomics Data**. The specific aims of this thesis were:

1. To identify and examine dual subtyping in breast cancer tumors, overall and within a particular subgroup to understand their tumor biology and possible implications to therapeutic guidance.
2. To identify an expanded HER2 gene signature in order to capture the full biological diversity of HER2+ tumors.
3. To assess if molecular subtyping signatures are concordant in primary and metastatic tumors in colon cancer.

The main findings associated with this thesis are shown in Table 6.1. First, I identified patients that belonged to more than one molecular subtype of breast cancer (BC) in **Chapters 2 and 5**. In **Chapter 2**, these dual subtypes of samples were shown to have different tumor biology than that of their respective single subtypes. In Chapter 5, the subgroups within triple-negative breast cancer (TNBC) consisted of a subtype with HER2+-associated pathways, suggesting treatment implications. Second, I expanded the HER2 gene signature to capture the ever-growing tumor biology behind HER2+ BC tumors in **Chapter 3**. The newly expanded gene signature captured the heterogeneity among the HER2+ tumors, which was incidentally also observed in **Chapter 5**, where the dual subtype samples of TNBC and HER2+ showed characteristics pertaining to HER2-low tumors. Finally, in **Chapter 4**, I investigated the concordance between the primary and liver metastasis of colon cancer tumors, which showed that the origin of the tumor tissue plays a major role in precision medicine in Colorectal Cancer.

Table 6.1: Overview of the studies and the main findings presented in this thesis.
Abbreviations: HER2 - Human epidermal growth factor receptor, CMS - Consensus Molecular Subtype, FFPE - Formalin Fixed Paraffin Embedded, PT - Primary tumor, LM - Liver metastasis, ICO-IDIBELL - Catalan Institute of Oncology - Bellvitge Biomedical Research Institute, INT - Istituto Nazionale dei Tumori di Milano, IOV - Istituto Oncologico Veneto, CPTAC - Clinical Proteomic Tumor Analysis Consortium, BC - Breast Cancer, DEA - Differential Expression Analysis, TNBC - Triple Negative Breast Cancer

	Chapter 2	**Chapter 3**	**Chapter 4**	**Chapter 5**
Aim of the study	To identify and examine dual subtype tumors to understand their biology and possible treatment implications	To identify an expanded HER2 signature that would capture the full biological diversity of HER2+ tumors.	To assess if gene expression signatures and CMS classifications are concordant in primary and metastatic tumors	To identify distinct molecular subgroups within the triple-negative breast cancer patient population
Cancer type	Breast Cancer	Breast Cancer	Colon Cancer	Breast Cancer
Data type	Full genome microarray data from FFPE breast tumor tissues	Full genome microarray data from FFPE breast tumor tissues	Full genome microarray data from fresh frozen and FFPE colon cancer tissues	Proteomic data from breast tissue specimens
Study Population	15,580 samples from internal studies	1552 samples from internal studies	51 matched pairs of PT and LM samples from ICO-IDIBELL, INT and IOV institutes	84 samples from the CPTAC for studies by Anurag et.al. and Krug et.al.
Main Findings	97% of samples had a single activated subtype, 3% were classified as BP dual subtype Most frequently occurring dual subtypes were Luminal-Basal -type and Luminal-HER2-type	29 genes were selected as a result of DEA and additional filtering These genes play a role in HER2, PI3K, and, AKT signaling pathways, important for cancer growth and proliferation	PT did not cluster together with their matched LM on gene expression level, indicating that biology of PT differs from LM Gene expression classifiers in PT in many cases did not reflect on LM, but did not affect the overall survival	Similarity network fusion analysis resulting in two clusters of samples Cluster 1 reflected pathways of TNBC and HER2+ enriched tumors whereas the Cluster 2 showed pathways of TNBC tumors alone

6.1 Unpacking the Complexities of Molecular Subtypes in Breast and Colon Cancers: Lessons from this thesis

6.1.1 Aim 1 - Exploring Dual Subtyping in Breast Cancer: Unveiling Tumor Biology and Therapeutic Implications

Chapter 2 found a novel method to classify the molecular subtype of tumors into single and dual subtypes. While previous research focused on the characteristics that defined the dual subtypes, our research also showed the clinical relevance of the dual subtypes samples. Rather than focusing on understanding the features of one single molecular subtype and its respective mixed subtype, our study devised a compelling methodology to classify the single and dual subtypes. In addition, some promising results on the treatment of HER2-single-type tumors with HER2 dual-targeted treatments were shown through the TRAIN2, Aphinity, and NBRST clinical trial datasets [122, 149, 151].

Similarly, in **Chapter 5** my research showed a robust way to find subgroups within a molecular subtype using similarity network fusion analysis implemented using the SNFTool package in R. Although this tool was intended to combine samples from two different datatypes, it is shown to be equally useful even when the same dataset is split into two and used for analysis. Hence, I used it on datasets from Krug and Anurag et.al, specifically on TNBC samples, from both the TNBC and HER2 types, which is one of the dual subtypes that was not explored in **Chapter 2** due to lack of samples. The results from our study showed the possible implications that the study of dual TNBC-HER2 subtypes could have on targeted treatment therapies since they could benefit from treatments like trastuzumab and pertuzumab.

6.1.2 Aim 2 - Unveiling the Full Spectrum: Expanding the HER2 Gene Signature to Capture Biological Diversity

Our study in **Chapter 3** introduced a comprehensive 29-gene HER2-type signature that incorporates known HER2 amplicon genes and others involved in important HER2-related oncogenic signaling pathways. An important feature of our expanded HER2 signature genes is that they accounted for a higher percentage of variance captured, highlighting their robust ability to differentiate between subtypes, compared to previously reported signature genes. This also could possibly indicate heterogeneity within HER2+ tumors could be captured by the expanded HER2 signature genes. By identifying the expanded HER2 signature genes encompassing new molecular characteristics and biological behaviors of different HER2-positive biomarkers, we can develop targeted therapies that are more tailored to the individual patient's tumor [280, 281].

6.1.3 Aim 3 - Evaluation of Molecular Subtyping Concordance Between Primary and Metastatic Colon Cancer Tumors

The study in **Chapter 4** highlights the discordance in gene expression profiles between primary colon tumors (PT) and matched metastases (LM) and the importance of considering the unique biology of each tumor site. Our findings challenge the conventional practice of solely relying on molecular profiling of archived primary tissue for treatment decisions in metastatic Colorectal Cancer (mCRC). Specifically, our results show that PT losing their mesenchymal phenotype in matched LM indicates that the tumor microenvironment and intrinsic properties of PT may influence their molecular characteristics. Overall, our results contribute to advancing the understanding of the complex molecular landscape of mCRC and provide insights into the potential clinical implications of PT and LM discordance.

6.2 Contributions to Methodological Improvements

The quest to improve molecular subtyping is a continuous effort, and my research has contributed to this goal in several ways.

In **Chapters 2, 3, and 4**, FFPE microarray expression data were used for the studies associated with dual subtyping, signature gene expansion in breast cancer, and concordance of Primary and metastatic tumors in colon cancer. In **Chapter 5**, proteomic data was used for subgroup identification within TNBC subtypes in breast cancer. I performed dual subtyping analysis in both the microarray data in **Chapter 2** as well as subgrouping analysis in the proteomics data in **Chapter 5**, although only the TNBC samples were analyzed in detail using the Proteomics dataset. From the microarray data, I was able to analyze the biology of the dual subtypes and compare it against their respective single subtypes. However, in the proteomics data, I could gain a better idea of protein-activated and degraded pathways in the TNBC-HER2 subtype, where the proteomics data shed more light on the protein-activated and degraded pathways. In my opinion, it could be beneficial to validate the results on both the proteomic and a whole-transcriptome level of the same samples, in order to better understand their underlying tumor biology.

In **Chapter 3**, I utilized an unsupervised clustering called Principal Component Analysis (PCA), which is a dimensionality reduction technique to transform high-dimensional data into a lower-dimensional space while preserving the most important patterns or variations in the data. I used PCA for the expanded BluePrint signature with 29 new HER2 signature genes and compared it with previously reported signature genes. However, in my opinion, these results should be evaluated with additional circumspection since the variance captured from the previously reported signature genes was estimated on FFPE microarray samples, which are developed from the same data processing methods as the samples that were used to identify the expanded gene signature. I believe that a more suitable approach would be to compare the variance captured amongst all these signatures in an unbiased microarray dataset, which does not pertain to any one of the signature gene sets alone.

In **Chapter 2, Chapter 3, and Chapter 4**, we utilized Formalin-Fixed Paraffin-Embedded (FFPE) samples as our primary tissue source. In **chapter 4**, in addition to the FFPE data, fresh frozen (FF) tissue samples were also present in the sample set. FF data is generally considered to be the gold standard since they provide higher-quality nucleic acids. FFPE however, stays stable for longer periods of time, which makes it useful for retrospective analysis. However, in **Chapter 4**, I clustered the FFPE and FF-matched pairs from 11 samples that showed that samples derived from the same patients were clustering together irrespective of tissue type. Although, generally the FF data is considered to hold more biological significance, I found the samples from both these data types to hold patterns based on intratumoral heterogeneity. Hence, I combined the FFPE and FF data and normalized them using median normalization in order to perform further analysis on the combined dataset. In the future, I believe we should approach the integration of datasets from different tissue types with an open mind since combining data can lead to larger dataset sizes, offering more robust analyses, and improving the generalizability of findings. Despite variations in tissue type, the inherent heterogeneity of tumors often transcends these distinctions.

In **Chapter 4** and **Chapter 5**, hierarchical clustering was employed to classify the samples and impose them with the molecular subtype classifications. In **chapter 4**, clustering of the fresh and the FFPE samples of both the primary and metastatic tumors based on the MSI signature was performed to show a non-homogeneous classification of primary and metastatic tumors. **Chapter 5** clustered the samples based on the results of the SNFTool. The additional colored bars added more information to understand the clustering patterns better. In my opinion, hierarchical clustering methods were better suited for interpreting patterns and variance among the dataset, than the PCA analysis. With PCA analysis, the additional variables that can be superimposed are far lesser than in hierarchical clustering analysis, where the colored bars could help us interpret the clustering patterns based on many more variables.

In **Chapter** 2 and **Chapter 3**, we made use of the BluePrint 80-gene molecular subtyping assay, in order to classify the breast tumor samples into Luminal-like, HER2-like, and Basal-like samples. In **Chapter 5**, the PAM50 test was used to classify the breast proteomics dataset into one of four molecular subtypes: Luminal A, Luminal B, HER2-enriched, and Basal-like. Although both the PAM50 and the BluePrint tests perform molecular subtype classification, for the purpose of studying any one subtype in detail, I found the BluePrint signature set to be better suited for this purpose. The PAM50 signature set classifies more molecular subtypes like Luminal-B and Normal-like. However, if I wanted to analyze the biomarkers of the Normal-like subtype, it would not be possible with PAM50, since they do not have a dedicated set of signature genes for this subtype. BluePrint, on the other hand, has 58 Luminal-type, 28 Basal-type, and 4 HER2-type signature genes, which enables a deeper understanding of any one molecular subtype.

In **Chapter** 2 and **Chapter 5**, we focused on identifying subgroups within the molecular subtype classification in breast cancer. **Chapter** 2 focused on Luminal-Basal and Luminal-HER2 samples and their tumor biology, whereas in **Chapter 5**, I focused on identifying subgroups within TNBC patients which resulted in samples that could resemble a HER2-Basal subtype within TNBC patients. I believe that the results from Chapter 5 could have given us more insight into the TNBC-HER2 type if there had been IHC and FISH information available. Previous studies have shown that samples with TNBC-HER2 characteristics mostly belonged to the HER2-low subtype, which shows low expressions of HER2 and is mostly misclassified as Basal or Luminal types. HER2-low samples are defined as +1 or 2+ by IHC and FISH-negative. Hence, if I had IHC and FISH information available for the Proteomics dataset, conclusions regarding the HER2 expression on the TNBC-HER2 samples could have been estimated.

In **Chapter** 2, we investigated the presence of dual subtypes in breast cancer tumors, which is an area that has been understudied in the past. We used approximately 15,600 samples to classify the single and dual subtypes. Predominantly, the samples that belonged to the Luminal and the HER2 types, were followed by the Basal type. It is aggressive and less common, accounting for only about 10-20% of all breast cancer cases. Therefore, the sample count for Basal subtypes is typically undercounted compared to other subtypes of breast cancer. In my opinion, another reason for the low count of these samples could be that the Basal-like samples are already detected as triple negative through laboratory tests like IHC and FISH. Once these patients are detected as triple-negative, they may be immediately provided with treatment options like chemotherapy, in order to increase their chances of recurrence-free survival. Hence, a very small proportion of patients may actually be sent for molecular subtype diagnostic tests like BluePrint or Oncotype, and more often directed immediately towards treatment options.

These factors can all contribute to the challenge of obtaining sufficient, high-quality samples for precision medicine research. In our work, the, smaller sample size associated with HER2-Basal and Luminal-HER2-Basal prevented us from drawing conclusions about these subtypes. Similarly, in **Chapter 4**, more concrete results could not be observed about the CMS1 and the CMS3 molecular subtypes of tumor tissues due to low numbers of tumors classified as MSI-like and/or CMS1 and BRAF m-like and/or CMS3.

6.3 Precision Medicine in Oncology: Limitations in Thesis

In the following sections, I identify areas of improvement and opportunities for further development within molecular subtyping diagnostics in the context of cancer.

This exploration aims to highlight the potential for enhancing the existing methodologies and approaches, paving the way for more thorough and precise diagnostic strategies in the future.

Transcriptomics: In my thesis, the microarray technique has been the basis of our gene expression analysis in **chapters 2, 3, and 4** for both breast and colon cancer analysis. However, with the advent of next-generation sequencing (NGS) technologies, **microarrays are rapidly becoming obsolete**. But still, the shift from microarray technologies to NGS has not happened as quickly as anticipated due to several reasons. NGS technologies were initially associated with complex workflows and data analysis challenges, requiring specialized bioinformatics skills. This limited the accessibility and ease of use for many researchers and clinicians. Moreover, microarray technologies had already established a strong presence in research and clinical settings, making it challenging to rapidly transition to a new technology. Over time, as NGS technologies have become more affordable, streamlined, and user-friendly, the adoption rate has increased. Today, NGS has gained prominence for its ability to provide comprehensive and high-throughput sequencing data, enabling researchers and clinicians to unravel complex genomic information and advance precision medicine approaches. I believe our focus should be shifted towards NGS technology which is likely to become a routine part of cancer diagnostics in the future [282].

Data accessibility: Having larger datasets is crucial for precision medicine [283]. A larger sample set can help improve the accuracy and generalizability of machine learning models used in precision medicine. However medical data has inherently been decentralized, to protect patient privacy, to limit the type and amount of data that can be shared, and who can access it, such as the European General Data Protection Regulation (GDPR) or the Health Insurance Portability and Accountability Act (HIPAA) in the USA. One practical solution that can be implemented is **data centralization**, which involves aggregating patient data from multiple healthcare providers, research institutions, and other sources into a single database for analysis and use in developing new treatments and diagnostics. However, while data centralization can provide benefits, such as enabling researchers to access larger and more diverse datasets, it also poses substantial challenges. This limits the development of precision medicine solutions [284, 285]. Centralizing data also raises questions about data ownership [286], especially when dealing with data from multiple sources, each with its own policies and regulations. Hence, alternative approaches should be developed to enable **data analysis without the need for it to be centralized in a location**.

Heterogeneous datasets: Precision medicine relies on a diverse range of patient data to be effective. Many genomic tests and treatment drugs are developed and tested on a homogeneous population, which does not guarantee the same treatment response across people from different ancestral and cultural backgrounds. This lack of diversity also limits the ability to identify new biomarkers and develop targeted therapies for underrepresented populations [287]. This highlights the need for **more racially diverse representation in breast and colon cancer research** to better understand the disease in different populations and improve precision medicine approaches.

Overlooked cohorts: Numerous research studies and clinical trials have focused on colon cancer in both males and females, actively recruiting and enrolling patients of both genders as there is a relatively similar incidence rate between males and females. Additionally, screening programs for colon cancer are recommended for both males and females starting at a certain age (usually around 50 years old), whereas breast cancer screenings are more aimed at women. These efforts have ensured that females are adequately screened and diagnosed at an early stage, leading to improved outcomes and survival rates for colon cancer. On the other hand, breast cancer in males is relatively rare, accounting for less than 1% of all breast cancer cases [288]. Due to its rarity, there has been a lack of research on breast cancer in males, as most studies on breast cancer have focused on female patients. This has resulted in a lack of knowledge and understanding about the disease in men, leading to delayed diagnoses, inadequate treatment, and worse outcomes. In my opinion, **this lack of awareness about breast cancer in men**, leads to the misconception that only women can get breast cancer. Additionally, men diagnosed with breast cancer may face social stigma and misconceptions, which can further complicate their experience. Hence, more research and awareness efforts are needed to address breast cancer in men.

6.4 Precision Medicine in Oncology - Future Opportunities

In this subsequent section, I will delve into potential solutions and avenues that hold promise for addressing the limitations identified earlier. By exploring innovative approaches, emerging technologies, and cutting-edge research, we aim to offer potential solutions that could contribute to overcoming the challenges and limitations faced in the field of molecular subtyping diagnostics.

Transcriptomics in Precision Medicine: With the slow decline of microarray technology, the rise of **NGS technology** has revolutionized the field of precision medicine by providing researchers and clinicians with a powerful tool to analyze genomic data at an unprecedented depth and scale. NGS has also facilitated the development of personalized cancer therapies and the identification of rare genetic diseases that were previously difficult to diagnose [289]. One of the most impactful discoveries with respect to NGS has been **liquid biopsies**, which is considered the future of cancer diagnostics. Liquid biopsies are a wonderful application of NGS technologies as they offer a non-invasive and real-time way of detecting cancer-related alterations in bodily fluids such as blood, urine, or saliva [290].

Decentralized Machine Learning: A promising solution for decentralized data in cancer research could be **Swarm Learning** [291], which is a **decentralized machine learning approach** that focuses on collaborative model training among a network of devices or institutions. Swarm learning has already been used to predict diseases like COVID and tuberculosis [291]. In breast cancer research, I discussed limitations in **Chapter 2** regarding the limited sample size of HER2-Basal and Luminal-HER2-Basal samples. Swarm Learning could potentially resolve this drawback, by using all Basal and HER2 samples from different institutions. Similarly, in **Chapter 5**, the presence of HER2-low samples in the TNBC-HER2 subgroup of samples could have been easily estimated through samples from the combined Swarm Learning datasets which could have enabled the presence of samples with IHC and FISH information.

Diversity in Precision Medicine: In Chapter 4, all the primary and metastatic colon tumor tissue samples were taken from the HUB-ICO-IDIBELL cohort is a set of breast cancer patients that has a higher representation of patients of European descent. This lack of diversity may limit the generalizability of precision medicine findings and may result in unequal access to precision medicine for certain populations. Precision medicine must continue to develop and refine with a focus on ensuring that the benefits are available to all patients **regardless of race, ethnicity, or socioeconomic status**. It is crucial to address this issue and increase diversity in cancer diagnostics by **improving inclusivity in genomic research and clinical trials**, incorporating data from diverse populations in databases, and enhancing training for healthcare professionals to improve their cultural competence. In addition to developing racially and ethnically diverse clinical trials and genomic research, **Swarm Learning** could also play a role here, in increasing the opportunity of **combining research data conducted in two different populations to create a diverse patient population** and ensure the generalizability of the research findings and assure equal access to precision medicine for all populations.

Breast Cancer in Males: Men with a family history of breast cancer in women have a higher risk of developing breast cancer [292]. **Male offspring of men with breast cancer also have an increased risk of breast cancer**, which is often attributed to BRCA2 mutations or other inherited or intrinsic features affecting androgen metabolism [293, 294]. To create an inclusive study population of both genders in breast cancer, study protocols, and recruitment strategies should be designed to include both male and female participants. This may involve collaborating with multiple research institutions, hospitals, and clinics to access a diverse patient population. This step can be enabled once again using Swarm Learning. Another possible strategy is to **conduct long-term studies that track patients over time**. This provides an opportunity to capture samples from both genders as patients progress through different stages of breast cancer, ensuring a more balanced representation. Finally, we should **raise awareness among healthcare providers, patients, and the general public** about the importance of including both genders in breast cancer research.

More data integration: In addition to the above future opportunities, with **data**

integration, researchers can easily combine a wealth of research data from different institutions, such as genomic, proteomic, and imaging data, and a more comprehensive understanding of cancer biology can be achieved. This results in developing more accurate and personalized cancer diagnoses and treatment plans. Integrating imaging data can also help to identify patients with specific tumor characteristics, such as high vascularization, that may benefit from certain treatments. A prime example of Data integration is the Cancer Genome Atlas (TCGA) project [295]. This is a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that aims to comprehensively characterize the genomic alterations in 33 different cancer types. By integrating data from multiple molecular profiling technologies and clinical information from thousands of patient samples, the TCGA project has helped identify new cancer driver genes and potential therapeutic targets. In my opinion, **data integration is very beneficial since it fosters collaboration and knowledge sharing among researchers and institutions**. By combining data from multiple sources, large-scale studies and research initiatives can be conducted to identify novel therapeutic targets, develop predictive models, and uncover new insights into cancer biology. This collaborative approach accelerates scientific discoveries and contributes to the collective knowledge base in precision oncology

Advanced computational approaches: Finally, advanced computational approaches, including machine learning and artificial intelligence (AI), can play a significant role in analyzing complex molecular datasets and identifying patterns that may be missed by conventional methods. Machine learning and AI have the potential to expedite this process by analyzing large datasets and identifying potential drug candidates with higher precision and efficiency. AI algorithms can also assist in virtual screening, target identification, lead optimization, and predicting drug toxicity, thereby accelerating the development of new therapies and reducing costs. AI-powered systems can assist in the early detection and accurate diagnosis of cancer, by quickly identifying patterns and anomalies that may be missed by human observers. The future holds the potential for even more precise and automated diagnostic systems that can assist healthcare professionals in making accurate and timely diagnoses.

Swarm learning: Swarm Learning allows for the distributed training of AI models based on standardized AI engines and allows for the merging of model parameters with equal rights for all members while protecting the machine learning models from attacks [291]. Swarm Learning can enable the development of more accurate and reliable AI models for precision medicine, without compromising patient privacy and confidentiality. It addresses issues such as privacy and data security by enabling collaborative analysis without sharing sensitive patient information. Additionally, swarm learning tackles data heterogeneity by integrating diverse datasets, including genomics, proteomics, and clinical data. It overcomes limitations posed by limited sample sizes through data pooling, enhances model generalizability by leveraging diverse datasets, and supports real-time learning for up-to-date knowledge. Swarm learning also ensures regulatory compliance by keeping data locally stored and processed. By tackling these challenges, swarm learning has the potential to advance accurate predictive models, identify novel biomarkers, optimize treatment strategies, and ultimately improve patient outcomes in precision medicine for cancer.

In summary, machine learning and artificial intelligence hold immense potential to transform healthcare by improving diagnostics, enabling personalized medicine, accelerating drug discovery, enhancing clinical decision-making, and facilitating remote patient monitoring.

6.5 Concluding Remarks

Our findings, together with previous studies shed light on the intricate landscape of molecular subtyping in breast and colon cancers. The identification of dual subtyping within breast cancer tumors emphasizes the need for personalized treatment approaches that consider the unique characteristics of each molecular subtype. The expansion of the HER2 gene signature reveals the heterogeneity of HER2-positive tumors, suggesting the potential for tailored treatment strategies based on the specific molecular features of individual tumors. Additionally, the observed concordance between primary and metastatic tumors in colon cancer for certain molecular subtypes highlights the relevance of tumor tissue origin in guiding targeted therapies. These insights underscore the importance of precise molecular subtyping in enhancing treatment decision-making and advancing the field of precision oncology.

Through molecular subtyping, we have gained valuable insights into the biological diversity of tumors, allowing for precise diagnosis and personalized treatment strategies. Despite recent advances, current precision medicine approaches are still limited by several factors, such as an incomplete understanding of disease mechanisms, inadequate biomarker identification, and lack of access to necessary resources, among others. Incorporating patient-specific factors, such as genetic background and environmental exposure, can improve cancer diagnosis and treatment by tailoring therapies to individual patients.

It may also be advantageous to make precision medicine more accessible and affordable to all patients. Efforts to reduce the cost of precision medicine could include increasing efficiency and streamlining processes, improving reimbursement models, and encouraging competition among providers. By leveraging the power of precision medicine, we are poised to transform cancer treatment, paving the way for more personalized and targeted therapies that can make a meaningful impact on patient lives.

Chapter 7

Summary

Cancer, a complex and multifaceted disease, represents a significant challenge in the field of medical research and healthcare. It encompasses a broad range of conditions characterized by the uncontrolled growth and division of abnormal cells within the body. Among the various types of cancer, breast and colon cancer continue to be the subject of intensive study and clinical focus. Molecular subtyping enables a deeper understanding of the disease and guides tailored treatment strategies, leading to improved outcomes. The molecular subtyping of breast and colon cancers is a vital avenue of research and clinical practice, aiding in the identification of patient-specific biomarkers, targeted therapy selection, and the development of precision medicine approaches. The goal of my thesis was to actively increase knowledge to contribute to refining molecular subtyping diagnostics in breast and colon cancers using gene expression and proteomics data.

In **Chapter 2** for breast cancer, I focused on developing a methodology to identify patients that may belong to more than one molecular subtype, indicating that they could benefit from more than one kind of targeted therapy. The study investigated the presence of dual subtypes in breast cancer tumors, which is an area that has been understudied in the past. The findings showed that the BluePrint assay was able to accurately classify breast cancer tumors into their respective subtypes, and identified a subset of tumors that showed characteristics of more than one molecular subtype. We developed a classification threshold using full genome microarray samples which separates the single and dual subtypes. We also showed that the classification of the subtypes on the NBRST dataset shows refined prediction to therapy. This dual subtype classification has important implications for therapeutic guidance, as it suggests that a combination of treatments targeted at both subtypes may be necessary for effective treatment. Overall, the study highlights the importance of molecular subtyping in guiding breast cancer treatment decisions and identifies a potential area for further research into the role of dual subtypes in breast cancer prognosis and treatment.

In **Chapter 3**, I expanded the HER2-type molecular subtype signature to capture more biology and heterogeneity in the HER2-type tumors. where the expanded HER2-type signature set can be combined into a molecular diagnostic test. I per-

7

formed differential expression analysis and filtered genes based on additional selecting criteria to select a set of 29 HER2 signature genes. I performed ontology and pathway analysis to understand the biology and the function of the new signature genes and also observed the amount of variance observed along with several other previously reported molecular subtype signatures. By understanding the unique molecular characteristics of a patient's tumor, clinicians can develop targeted therapies that are more effective and have fewer side effects than traditional treatments. The findings from the above two papers can be synergistically combined to develop a more comprehensive and accurate genomic testing approach.

Similarly, in **Chapter 4**, I investigated the concordance of primary and metastatic tumors for consensus molecular subtype classification in colon cancer, to show that the origin of the tissue has major consequences for the targeted therapies in metastatic tumors as well. This study aimed to investigate the concordance between the molecular subtypes of primary colorectal tumors and their corresponding synchronous liver metastases. We analyzed the gene expression profiles of 36 matched pairs of primary colorectal tumors and liver metastases using a molecular subtype classification system. The results showed a high degree of concordance between the molecular subtypes of the primary tumors and their matched liver metastases. The majority of the samples exhibited the same molecular subtype in both the primary and metastatic sites. However, some samples showed a change in molecular subtype from the primary tumor to the liver metastasis. This finding highlights the importance of monitoring the molecular subtypes of metastatic tumors, as they may differ from the primary tumor and impact treatment decisions. Overall, the study asserts that molecular subtype analysis can be a useful tool in guiding precision medicine approaches for the treatment of colorectal cancer metastasis.

In **Chapter 5** of the thesis, I revisited the topic of breast cancer and utilized a breast cancer proteomics dataset. The previous chapters on breast cancer primarily focused on identifying new subtypes and signatures using the BluePrint signature assay, which identified three molecular subtypes. In this chapter, my objective was to identify subgroups within the Triple negative molecular subtype of breast cancer from a proteomics dataset, using the method Similar Network Fusion. By doing this, I aimed to validate the proteomics dataset and evaluate its performance in identifying molecular subtypes. This chapter provided a new perspective on breast cancer subtyping and opened up possibilities for the development of novel subtyping assays.

In **Chapter 6** of my thesis, I summarize how the different chapters have contributed towards achieving the overall objective of my research, which was to develop and improve breast and colon cancer diagnostics using computational precision medicine methods. While discussing the outcomes of our studies, I also highlight the limitations and challenges that we encountered during the research process. Moreover, I also discuss the potential applications of our research in the future of precision medicine, particularly in the development of targeted therapies for breast and colon cancer patients. Overall, the insights gained from this thesis can pave the way for future research and clinical applications in the field of precision medicine.

References

[1] *American Cancer Society*. URL: https://training.seer.cancer.gov/disease/cancer/.

[2] Franks, L. M. and Knowles, M. A. "What is cancer". In: *Introduction to the cellular and molecular biology of cancer* 4 (1990), 4–9.

[3] (WHO), W. H. O. et al. "Global Health Estimates 2020: Deaths by cause, age, sex, by country and by region, 2000-2019: WHO; 2020". In: *URL: https://www.who. int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death [accessed 2022-11-29]* (2020).

[4] Bray, F. et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 68.6 (2018), 394–424.

[5] Gersten, O. and Wilmoth, J. R. "The cancer transition in Japan since 1951". In: *Demographic Research* 7 (2002), 271–306.

[6] Danaei, G. et al. "Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors". In: *The lancet* 366.9499 (2005), 1784–1793.

[7] Galani, E. and Christodoulou, C. "Human papilloma viruses and cancer in the post-vaccine era". In: *Clinical microbiology and infection* 15.11 (2009), 977–981.

[8] Perz, J. F. et al. "The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide". In: *Journal of hepatology* 45.4 (2006), 529–538.

[9] Atkins, H. et al. "Treatment of early breast cancer: a report after ten years of a clinical trial". In: *Br Med J* 2.5811 (1972), 423–429.

[10] Banks, E. "Hormone replacement therapy and the sensitivity and specificity of breast cancer screening: a review". In: *Journal of medical screening* 8.1 (2001), 29–35.

[11] Menon, U. et al. "Sensitivity and specificity of multimodal and ultrasound screening for ovarian cancer, and stage distribution of detected cancers: results of the prevalence screen of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS)". In: *The lancet oncology* 10.4 (2009), 327–340.

[12] Vahabi, M. "Breast cancer screening methods: a review of the evidence". In: *Health care for women international* 24.9 (2003), 773–793.

[13] DeSantis, C. et al. "Breast cancer statistics, 2013". In: *CA: a cancer journal for clinicians* 64.1 (2014), 52–62.

[14] Society, A. C. *Cancer facts & figures*. The Society, 2008.

[15] Quinn, M. and Allen, E. "Changes in incidence of and mortality from breast cancer in England and Wales since introduction of screening". In: *Bmj* 311.7017 (1995), 1391–1395.

[16] Yang, X. and Lippman, M. E. "BRCA1 and BRCA2 in breast cancer". In: *Breast cancer research and treatment* 54 (1999), 1–10.

[17] Carey, L. A. et al. "Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study". In: *Jama* 295.21 (2006), 2492–2502.

[18] Ellsworth, R. E. et al. "Molecular heterogeneity in breast cancer: State of the science and implications for patient care". In: *Seminars in cell & developmental biology*. Vol. 64. Elsevier. 2017, 65–72.

[19] Press, O. A. et al. "Characterization of HER2 status by fluorescence in situ hybridization (FISH) and immunohistochemistry (IHC)". In: *Histopathology: Methods and Protocols* (2014), 181–207.

[20] Badve, S. S. et al. "Estrogen-and progesterone-receptor status in ECOG 2197: comparison of immunohistochemistry by local and central laboratories and quantitative reverse transcription polymerase chain reaction by central laboratory". In: *Journal of Clinical Oncology* 26.15 (2008), 2473–2481.

[21] Moore, R. G. et al. "The use of multiple novel tumor biomarkers for the detection of ovarian carcinoma in patients with a pelvic mass". In: *Gynecologic oncology* 108.2 (2008), 402–408.

[22] Osborne, C. K. "Tamoxifen in the treatment of breast cancer". In: *New England Journal of Medicine* 339.22 (1998), 1609–1618.

[23] Cuzick, J. et al. "Effect of anastrozole and tamoxifen as adjuvant treatment for early-stage breast cancer: 10-year analysis of the ATAC trial". In: *The lancet oncology* 11.12 (2010), 1135–1141.

[24] Group, B. I. G. (1.-9. C. "A comparison of letrozole and tamoxifen in postmenopausal women with early breast cancer". In: *New England Journal of Medicine* 353.26 (2005), 2747–2757.

[25] Valabrega, G., Montemurro, F., and Aglietta, M. "Trastuzumab: mechanism of action, resistance and future perspectives in HER2-overexpressing breast cancer". In: *Annals of oncology* 18.6 (2007), 977–984.

[26] Capelan, M. et al. "Pertuzumab: new hope for patients with HER2-positive breast cancer". In: *Annals of oncology* 24.2 (2013), 273–282.

[27] Lambert, J. M. and Chari, R. V. *Ado-trastuzumab Emtansine (T-DM1): an antibody–drug conjugate (ADC) for HER2-positive breast cancer*. 2014.

[28] De Laurentiis, M. et al. "Treatment of triple negative breast cancer (TNBC): current options and future perspectives". In: *Cancer treatment reviews* 36 (2010), S80–S86.

[29] Schmid, P. et al. "Pembrolizumab for early triple-negative breast cancer". In: *New England Journal of Medicine* 382.9 (2020), 810–821.

[30] Torre, L. A. et al. "Global cancer statistics, 2012". In: *CA: a cancer journal for clinicians* 65.2 (2015), 87–108.

[31] Ferlay, J. et al. "Global cancer observatory: cancer today". In: *Lyon, France: international agency for research on cancer* 3.20 (2018), 2019.

[32] Favoriti, P. et al. "Worldwide burden of colorectal cancer: a review". In: *Updates in surgery* 68 (2016), 7–11.

[33] Burt, R. W. et al. "Colorectal cancer screening". In: *Journal of the National Comprehensive Cancer Network* 8.1 (2010), 8–61.

[34] Bretthauer, M. "Colorectal cancer screening". In: *Journal of internal medicine* 270.2 (2011), 87–98.

[35] Walsh, J. M. and Terdiman, J. P. "Colorectal cancer screening: scientific review". In: *Jama* 289.10 (2003), 1288–1296.

[36] Hamilton, W. "Cancer diagnosis in primary care". In: *British Journal of General Practice* 60.571 (2010), 121–128.

[37] Collins, L. G. et al. "Lung cancer: diagnosis and management". In: *American family physician* 75.1 (2007), 56–63.

[38] Moiel, D. and Thompson, J. "Early detection of colon cancer -the Kaiser Permanente Northwest 30-year history: how do we measure success? Is it the test, the number of tests, the stage, or the percentage of screen-detected patients?" In: *The Permanente Journal* 15.4 (2011), 30.

[39] Powers, B., Hoopes, P., and Ehrhart, E. "Tumor diagnosis, grading, and staging." In: *Seminars in veterinary medicine and surgery (small animal)*. Vol. 10. 3. 1995, 158–167.

[40] Hohenberger, W., Reingruber, B., and Merkel, S. "Surgery for colon cancer". In: *Scandinavian journal of surgery* 92.1 (2003), 45–52.

[41] Segal, N. H. and Saltz, L. B. "Evolving treatment of advanced colon cancer". In: *Annual review of medicine* 60 (2009), 207–219.

[42] Kaleta-Richter, M. et al. "The capability and potential of new forms of personalized colon cancer treatment: Immunotherapy and Photodynamic Therapy". In: *Photodiagnosis and photodynamic therapy* 25 (2019), 253–258.

[43] Arruebo, M. et al. “Assessment of the evolution of cancer treatment therapies”. In: *Cancers* 3.3 (2011), 3279–3330.

[44] Waks, A. G. and Winer, E. P. “Breast cancer treatment: a review”. In: *Jama* 321.3 (2019), 288–300.

[45] Litwin, M. S. and Tan, H.-J. “The diagnosis and treatment of prostate cancer: a review”. In: *Jama* 317.24 (2017), 2532–2542.

[46] Arndt, V. et al. “Patient delay and stage of diagnosis among breast cancer patients in Germany–a population based study”. In: *British journal of cancer* 86.7 (2002), 1034–1040.

[47] Afzelius, P. et al. “Patient’s and doctor’s delay in primary breast cancer: Prognostic implications”. In: *Acta Oncologica* 33.4 (1994), 345–351.

[48] Coates, A. S. “Breast cancer: delays, dilemmas, and delusions”. In: *The Lancet* 353.9159 (1999), 1112–1113.

[49] Hiom, S. “Diagnosing cancer earlier: reviewing the evidence for improving cancer survival.” In: *British journal of cancer* 112 (2015), S1–5.

[50] Farkona, S., Diamandis, E. P., and Blasutig, I. M. “Cancer immunotherapy: the beginning of the end of cancer?” In: *BMC medicine* 14.1 (2016), 1–18.

[51] Postow, M. A., Callahan, M. K., and Wolchok, J. D. “Immune checkpoint blockade in cancer therapy”. In: *Journal of clinical oncology* 33.17 (2015), 1974.

[52] Nakauchi, M. et al. “Prognostic factors of minimally invasive surgery for gastric cancer: Does robotic gastrectomy bring oncological benefit?” In: *World Journal of Gastroenterology* 27.39 (2021), 6659.

[53] Sibio, S., La Rovere, F., and Di Carlo, S. “Benefits of minimally invasive surgery in the treatment of gastric cancer”. In: *World Journal of Gastroenterology* 28.30 (2022), 4227–4230.

[54] Ciardiello, F. et al. “Delivering precision medicine in oncology today and in future -the promise and challenges of personalised cancer medicine: a position paper by the European Society for Medical Oncology (ESMO)”. In: *Annals of Oncology* 25.9 (2014), 1673–1678.

[55] Ashley, E. A. “Towards precision medicine”. In: *Nature Reviews Genetics* 17.9 (2016), 507–522.

[56] Ali, H. et al. “Association between CD8+ T-cell infiltration and breast cancer survival in 12 439 patients”. In: *Annals of oncology* 25.8 (2014), 1536–1543.

[57] Sorlie, T. et al. “et al.(2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications”. In: *Proc Natl Acad Sci U S A* 98 ().

[58] Polyak, K. et al. “Heterogeneity in breast cancer”. In: *The Journal of clinical investigation* 121.10 (2011), 3786–3788.

[59] Turashvili, G. and Brogi, E. “Tumor heterogeneity in breast cancer”. In: *Frontiers in medicine* 4 (2017), 227.

[60] Beca, F. and Polyak, K. “Intratumor heterogeneity in breast cancer”. In: *Novel biomarkers in the continuum of breast cancer* (2016), 169–189.

[61] Cho, N. “Molecular subtypes and imaging phenotypes of breast cancer”. In: *Ultrasonography* 35.4 (2016), 281.

[62] Erber, R. and Hartmann, A. “Histology of luminal breast cancer”. In: *Breast Care* 15.4 (2020), 327–336.

[63] Gao, J. J. and Swain, S. M. “Luminal a breast cancer and molecular assays: a review”. In: *The oncologist* 23.5 (2018), 556–565.

[64] Creighton, C. J. “The molecular profile of luminal B breast cancer”. In: *Biologics: Targets and Therapy* (2012), 289–297.

[65] Yarden, Y. “Biology of HER2 and its importance in breast cancer”. In: *Oncology* 61.Suppl. 2 (2001), 1–13.

[66] Iqbal, N. and Iqbal, N. “Human epidermal growth factor receptor 2 (HER2) in cancers: overexpression and therapeutic implications”. In: *Molecular biology international* 2014 (2014).

[67] Bertucci, F. et al. “How basal are triple-negative breast cancers?” In: *International journal of Cancer* 123.1 (2008), 236–240.

[68] Foulkes, W. D., Smith, I. E., and Reis-Filho, J. S. “Triple-negative breast cancer”. In: *New England journal of medicine* 363.20 (2010), 1938–1948.

[69] Cleator, S., Heller, W., and Coombes, R. C. “Triple-negative breast cancer: therapeutic options”. In: *The lancet oncology* 8.3 (2007), 235–244.

[70] Engstrom, P. F. et al. “Colon cancer”. In: *Journal of the National Comprehensive Cancer Network* 7.8 (2009), 778–831.

[71] Markowitz, S. D. et al. “Focus on colon cancer”. In: *Cancer cell* 1.3 (2002), 233–236.

[72] Young, J. I. et al. “Treatment and survival of small-bowel adenocarcinoma in the United States: a comparison with colon cancer”. In: *Diseases of the Colon & Rectum* 59.4 (2016), 306–315.

[73] Thota, R., Fang, X., and Subbiah, S. “Clinicopathological features and survival outcomes of primary signet ring cell and mucinous adenocarcinoma of colon: retrospective analysis of VACCR database”. In: *Journal of gastrointestinal oncology* 5.1 (2014), 18.

[74] Ozuner, G. et al. “Colorectal squamous cell carcinoma: a rare tumor with poor prognosis”. In: *International journal of colorectal disease* 30 (2015), 127–130.

[75] Guinney, J. et al. “The consensus molecular subtypes of colorectal cancer”. In: *Nature medicine* 21.11 (2015), 1350–1356.

[76] Network, C. G. A. et al. “Comprehensive molecular characterization of human colon and rectal cancer”. In: *Nature* 487.7407 (2012), 330.

[77] Müller, M. F., Ibrahim, A. E., and Arends, M. J. “Molecular pathological classification of colorectal cancer”. In: *Virchows Archiv* 469 (2016), 125–134.

[78] Halvorsen, T. and Seim, E. “Association between invasiveness, inflammatory reaction, desmoplasia and survival in colorectal cancer.” In: *Journal of clinical pathology* 42.2 (1989), 162–166.

[79] Koopman, M. et al. “Deficient mismatch repair system in patients with sporadic advanced colorectal cancer”. In: *British journal of cancer* 100.2 (2009), 266–273.

[80] Popat, S., Hubner, R., and Houlston, R. “Systematic review of microsatellite instability and colorectal cancer prognosis”. In: *Journal of clinical oncology* 23.3 (2005), 609–618.

[81] Boland, C. R. and Goel, A. “Microsatellite instability in colorectal cancer”. In: *Gastroenterology* 138.6 (2010), 2073–2087.

[82] Van der Flier, L. G. et al. “The intestinal Wnt/TCF signature”. In: *Gastroenterology* 132.2 (2007), 628–632.

[83] Zeller, K. I. et al. “An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets”. In: *Genome biology* 4.10 (2003), 1–10.

[84] Abdelkader, A., Hartley, C., and Hagen, C. “Tubulovillous adenomas with serrated features are precursors to KRAS mutant colorectal carcinoma”. In: *LABORATORY INVESTIGATION*. Vol. 97. NATURE PUBLISHING GROUP 75 VARICK ST, 9TH FLR, NEW YORK, NY 10013-1917 USA. 2017, 157A–157A.

[85] Fessler, E. et al. “TGFβ signaling directs serrated adenomas to the mesenchymal colorectal cancer subtype”. In: *EMBO molecular medicine* 8.7 (2016), 745–760.

[86] Ahn, A. C. et al. “The limits of reductionism in medicine: could systems biology offer an alternative?” In: *PLoS medicine* 3.6 (2006), e208.

[87] Tillmann, T. et al. “Systems medicine 2.0: potential benefits of combining electronic health care records with systems science models”. In: *Journal of medical Internet research* 17.3 (2015), e3082.

[88] Low, S.-K., Zembutsu, H., and Nakamura, Y. “Breast cancer: The translation of big genomic data to cancer precision medicine”. In: *Cancer science* 109.3 (2018), 497–506.

[89] Perou, C. M. et al. “Molecular portraits of human breast tumours”. In: *nature* 406.6797 (2000), 747–752.

[90] Sorlie, T. et al. “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications”. In: *Proc Natl Acad Sci USA* 98 (2001), 10869–74.

[91] Parker, J. S. et al. “Supervised risk predictor of breast cancer based on intrinsic subtypes”. In: *Journal of clinical oncology* 27.8 (2009), 1160.

[92] Tibshirani, R. et al. “Diagnosis of multiple cancer types by shrunken centroids of gene expression”. In: *Proceedings of the National Academy of Sciences* 99.10 (2002), 6567–6572.

[93] Pu, M. et al. “based PAM50 signature and long-term breast cancer survival”. In: *Breast cancer research and treatment* 179 (2020), 197–206.

[94] Dieci, M. et al. “Integrated evaluation of PAM50 subtypes and immune modulation of pCR in HER2-positive breast cancer patients treated with chemotherapy and HER2-targeted agents in the CherLOB trial”. In: *Annals of Oncology* 27.10 (2016), 1867–1873.

[95] Paik, S. et al. “A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer”. In: *New England Journal of Medicine* 351.27 (2004), 2817–2826.

[96] Fisher, B. et al. “Treatment of lymph-node-negative, oestrogen-receptor-positive breast cancer: long-term findings from National Surgical Adjuvant Breast and Bowel Project randomised clinical trials”. In: *The Lancet* 364.9437 (2004), 858–868.

[97] Krijgsman, O. et al. “A diagnostic gene profile for molecular subtyping of breast cancer associated with treatment response”. In: *Breast cancer research and treatment* 133 (2012), 37–47.

[98] Glas, A. M. et al. “Gene expression profiling in follicular lymphoma to assess clinical aggressiveness and to guide the choice of treatment”. In: *Blood* 105.1 (2005), 301–307.

[99] Van’t Veer, L. J. et al. “Gene expression profiling predicts clinical outcome of breast cancer”. In: *nature* 415.6871 (2002), 530–536.

[100] Mittempergher, L. et al. “Performance characteristics of the BluePrint® breast cancer diagnostic test”. In: *Translational oncology* 13.4 (2020), 100756.

[101] Clark-Langone, K. M. et al. “Translating tumor biology into personalized treatment planning: analytical performance characteristics of the Oncotype DX® Colon Cancer Assay”. In: *BMC cancer* 10.1 (2010), 1–11.

[102] Sunami, K. et al. “Feasibility and utility of a panel testing for 114 cancer-associated genes in a clinical setting: a hospital-based study”. In: *Cancer science* 110.4 (2019), 1480–1490.

[103] Takeda, M. et al. “Clinical application of the FoundationOne CDx assay to therapeutic decision-making for patients with advanced solid tumors”. In: *The Oncologist* 26.4 (2021), e588–e596.

[104] Hofvind, S. et al. “Mammographic morphology and distribution of calcifications in ductal carcinoma in situ diagnosed in organized screening”. In: *Acta radiologica* 52.5 (2011), 481–487.

[105] Hamilton, E. et al. “Targeting HER2 heterogeneity in breast cancer”. In: *Cancer Treatment Reviews* 100 (2021), 102286.

[106] Polyak, K. et al. “Heterogeneity in breast cancer”. In: *The Journal of clinical investigation* 121.10 (2011), 3786–3788.

[107] Pantaleo, M. et al. “Gene expression profiling of liver metastases from colorectal cancer as potential basis for treatment choice”. In: *British journal of cancer* 99.10 (2008), 1729–1734.

[108] Vignot, S. et al. “Comparative analysis of primary tumour and matched metastases in colorectal cancer patients: evaluation of concordance between genomic and transcriptional profiles”. In: *European Journal of Cancer* 51.7 (2015), 791–799.

[109] Ho, P. J. et al. “Impact of delayed treatment in women diagnosed with breast cancer: A population-based study”. In: *Cancer medicine* 9.7 (2020), 2435–2444.

[110] Grass, F. et al. “Impact of delay to surgery on survival in stage I-III colon cancer”. In: *European Journal of Surgical Oncology* 46.3 (2020), 455–461.

[111] Wallden, B. et al. “Development and verification of the PAM50-based Prosigna breast cancer gene signature assay”. In: *BMC medical genomics* 8.1 (2015), 1–14.

[112] Dai, X. et al. “Breast cancer intrinsic subtype classification, clinical use and future trends”. In: *American journal of cancer research* 5.10 (2015), 2929.

[113] Prat, A. et al. “Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer”. In: *Breast cancer research* 12.5 (2010), 1–18.

[114] Vici, P. et al. “Triple positive breast cancer: a distinct subtype?” In: *Cancer treatment reviews* 41.2 (2015), 69–76.

[115] Prat, A. et al. “Molecular characterization of basal-like and non-basal-like triple-negative breast cancer”. In: *The oncologist* 18.2 (2013), 123–133.

[116] Burstein, M. D. et al. “Comprehensive Genomic Analysis Identifies Novel Subtypes and Targets of Triple-Negative Breast CancerIdentification of Four Unique Subtypes of TNBCs”. In: *Clinical Cancer Research* 21.7 (2015), 1688–1698.

[117] Beumer, I. et al. "Equivalence of MammaPrint array types in clinical trials and diagnostics". In: *Breast cancer research and treatment* 156 (2016), 279–287.

[118] Cardoso, F. et al. "70-gene signature as an aid to treatment decisions in early-stage breast cancer". In: *New England Journal of Medicine* 375.8 (2016), 717–729.

[119] Piccart, M. et al. "70-gene signature as an aid for treatment decisions in early breast cancer: updated results of the phase 3 randomised MINDACT trial with an exploratory analysis by age". In: *The Lancet Oncology* 22.4 (2021), 476–488.

[120] Glas, A. M. et al. "Converting a breast cancer microarray signature into a high-throughput diagnostic test". In: *BMC genomics* 7.1 (2006), 1–10.

[121] Whitworth, P. et al. "Chemosensitivity and endocrine sensitivity in clinical luminal breast cancer patients in the prospective neoadjuvant breast registry symphony trial (NBRST) predicted by molecular subtyping". In: *Annals of surgical oncology* 24 (2017), 669–675.

[122] Whitworth, P. et al. "Chemosensitivity predicted by BluePrint 80-gene functional subtype and MammaPrint in the prospective neoadjuvant breast registry symphony trial (NBRST)". In: *Annals of surgical oncology* 21 (2014), 3261–3267.

[123] Whitworth, P. et al. "5-year outcomes in the NBRST trial: preoperative MammaPrint and BluePrint breast cancer subtype is associated with neoadjuvant treatment response and survival". In: *Proceedings of the 2020 San Antonio Breast Cancer Virtual Symposium, San Antonio, TX*. 2020.

[124] Efron, B. and Tibshirani, R. J. *An introduction to the bootstrap*. CRC press, 1994.

[125] Ritchie, M. E. et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic acids research* 43.7 (2015), e47–e47.

[126] Subramanian, A. et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (2005), 15545–15550.

[127] Cohen, J. "Statistical power analysis". In: *Current directions in psychological science* 1.3 (1992), 98–101.

[128] Team, R. C. et al. "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria". In: *http://www. R-project. org/* (2016).

[129] Sigg, C. D. and Buhmann, J. M. "Expectation-maximization for sparse and non-negative PCA". In: *Proceedings of the 25th international conference on Machine learning*. 2008, 960–967.

[130] Hadley, W. *Ggplot2: Elegrant graphics for data analysis*. Springer, 2016.

[131] Gendoo, D. M. et al. "Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer". In: *Bioinformatics* 32.7 (2016), 1097–1099.

[132] Wei, C., Li, J., and Bumgarner, R. E. "Sample size for detecting differentially expressed genes in microarray experiments". In: *BMC genomics* 5 (2004), 1–10.

[133] Giuliano, A. E. et al. "Breast cancer -major changes in the American Joint Committee on Cancer eighth edition cancer staging manual". In: *CA: a cancer journal for clinicians* 67.4 (2017), 290–303.

[134] Conley, S. et al. "HER2 drives Mucin-like 1 to control proliferation in breast cancer cells". In: *Oncogene* 35.32 (2016), 4225–4234.

[135] Sasaki, Y. et al. "CLCA2, a target of the p53 family, negatively regulates cancer cell migration and invasion". In: *Cancer biology & therapy* 13.14 (2012), 1512–1521.

[136] Nadler, Y. et al. "Growth factor receptor-bound protein-7 (Grb7) as a prognostic marker and therapeutic target in breast cancer". In: *Annals of oncology* 21.3 (2010), 466–473.

[137] Oshi, M. et al. "G2M cell cycle pathway score as a prognostic biomarker of metastasis in estrogen receptor (ER)-positive breast cancer". In: *International journal of molecular sciences* 21.8 (2020), 2921.

[138] Hollern, D. P. et al. "E2F1 drives breast cancer metastasis by regulating the target gene FGF13 and altering cell migration". In: *Scientific reports* 9.1 (2019), 1–13.

[139] Srivastava, P. et al. "Clinical-pathologic characteristics and response to neoadjuvant chemotherapy in triple-negative low Ki-67 proliferation (TNLP) breast cancers". In: *NPJ Breast Cancer* 8.1 (2022), 51.

[140] Lousberg, L., Collignon, J., and Jerusalem, G. "Resistance to therapy in estrogen receptor positive and human epidermal growth factor 2 positive breast cancers: progress with latest therapeutic strategies". In: *Therapeutic advances in medical oncology* 8.6 (2016), 429–449.

[141] Schiff, R. et al. "Cross-talk between estrogen receptor and growth factor pathways as a molecular target for overcoming endocrine resistance". In: *Clinical Cancer Research* 10.1 (2004), 331s–336s.

[142] Groenendijk, F. H. et al. “Estrogen receptor variants in ER-positive basal-type breast cancers responding to therapy like ER-negative breast cancers”. In: *NPJ Breast Cancer* 5.1 (2019), 15.

[143] Beitsch, P. et al. “Pertuzumab/trastuzumab/CT versus trastuzumab/CT therapy for HER2+ breast cancer: results from the prospective neoadjuvant breast registry symphony trial (NBRST)”. In: *Annals of Surgical Oncology* 24 (2017), 2539–2546.

[144] Kuilman, M. et al. “BluePrint molecular subtyping recognizes single and dual subtype tumors with consequences for therapeutic guidance”. In: *European Journal of Cancer* 138 (2020), S106–S107.

[145] Liefaard, M. et al. “Effect of pertuzumab plus neoadjuvant trastuzumab-based chemotherapy in early-stage HER2-positive breast cancer according to BluePrint molecularly defined breast cancer subtypes”. In: *chemotherapy* 2016 (2013).

[146] Krop, I. et al. “Abstract PD3-01: BluePrint performance in predicting pertuzumab benefit in genomically HER2-positive patients: a biomarker analysis of the APHINITY trial”. In: *Cancer Research* 81.4_Supplement (2021), PD3–01.

[147] Badve, S. et al. “Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists”. In: *Modern Pathology* 24.2 (2011), 157–167.

[148] Giuliano, M., Trivedi, M. V., and Schiff, R. “Bidirectional crosstalk between the estrogen receptor and human epidermal growth factor receptor 2 signaling pathways in breast cancer: molecular basis and clinical implications”. In: *Breast Care* 8.4 (2013), 256–262.

[149] Ramshorst, M. S. van et al. “Neoadjuvant chemotherapy with or without anthracyclines in the presence of dual HER2 blockade for HER2-positive breast cancer (TRAIN-2): a multicentre, open-label, randomised, phase 3 trial”. In: *The Lancet Oncology* 19.12 (2018), 1630–1640.

[150] Von Minckwitz, G. et al. “Adjuvant pertuzumab and trastuzumab in early HER2-positive breast cancer”. In: *New England Journal of Medicine* 377.2 (2017), 122–131.

[151] Piccart, M. et al. “Adjuvant pertuzumab and trastuzumab in early HER2-positive breast cancer in the APHINITY trial: 6 years’ follow-up”. In: *Journal of Clinical Oncology* 39.13 (2021), 1448–1457.

[152] Prat, A. et al. “Correlative biomarker analysis of intrinsic subtypes and efficacy across the MONALEESA phase III studies”. In: *Journal of Clinical Oncology* 39.13 (2021), 1458.

[153] Spitale, A. et al. “Breast cancer classification according to immunohistochemical markers: clinicopathologic features and short-term survival analysis in a population-based study from the South of Switzerland”. In: *Annals of oncology* 20.4 (2009), 628–635.

[154] Hon, J. D. C. et al. “Breast cancer molecular subtypes: from TNBC to QNBC”. In: *American journal of cancer research* 6.9 (2016), 1864.

[155] Goldhirsch, A. et al. “Strategies for subtypes-dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011”. In: *Annals of oncology* 22.8 (2011), 1736–1747.

[156] Esebua, M. et al. “Interobserver reproducibility for HER2/neu immunohistochemistry: A comparison of reproducibility for the HercepTest (TM) and the 4B5 antibody clone”. In: *Pathology, Research and Practice* 212 (2016), 190–195.

[157] Wolff, A. C. et al. “Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update”. In: *Archives of Pathology and Laboratory Medicine* 138.2 (2014), 241–256.

[158] Watanabe, S. et al. “Targeting of the HER2/HER3 signaling axis overcomes ligand-mediated resistance to trastuzumab in HER2-positive breast cancer”. In: *Cancer Medicine* 8.3 (2019), 1258–1268.

[159] Rimawi, M. F., Schiff, R., and Osborne, C. K. “Targeting HER2 for the treatment of breast cancer”. In: *Annual review of medicine* 66 (2015).

[160] Benjamini, Y. et al. “Controlling the false discovery rate in behavior genetics research”. In: *Behavioural brain research* 125.1-2 (2001), 279–284.

[161] Fabregat, A. et al. “The reactome pathway knowledgebase”. In: *Nucleic acids research* 46.D1 (2018), D649–D655.

[162] Consortium, G. O. “Gene ontology consortium: going forward”. In: *Nucleic acids research* 43.D1 (2015), D1049–D1056.

[163] Desmedt, C. et al. “Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes”. In: *Clinical cancer research* 14.16 (2008), 5158–5165.

[164] Lin, I.-H. and Hsu, M.-T. “Analysis of 10086 Microarray Gene Expression Data Uncovers Genes that Subclassify Breast Cancer Intrinsic Subtypes”. In: *Breast Cancer-From Biology to Medicine*. IntechOpen, 2017.

[165] Milioli, H. H. et al. “The discovery of novel biomarkers improves breast cancer intrinsic subtype prediction and reconciles the labels in the metabric data set”. In: *PLoS One* 10.7 (2015), e0129711.

[166] Ulgen, E., Ozisik, O., and Sezerman, O. U. “pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks”. In: *Frontiers in Genetics* 10 (2019), 858. URL: https://doi.org/10.3389/fgene.2019.00858.

[167] Arriola, E. et al. “Genomic analysis of the HER2/TOP2A amplicon in breast cancer and breast cancer cell lines”. In: *Laboratory investigation* 88.5 (2008), 491–503.

[168] Sahlberg, K. K. et al. “The HER2 amplicon includes several genes required for the growth and survival of HER2 positive breast cancer cells”. In: *Molecular oncology* 7.3 (2013), 392–401.

[169] Kauraniemi, P. and Kallioniemi, A. “Activation of multiple cancer-associated genes at the ERBB2 amplicon in breast cancer”. In: *Endocrine-related cancer* 13.1 (2006), 39–49.

[170] Staaf, J. et al. “High-resolution genomic and expression analyses of copy number alterations in HER2-amplified breast cancer”. In: *Breast Cancer Research* 12.3 (2010), 1–18.

[171] Luoh, S.-W. “Amplification and expression of genes from the 17q11 q12 amplicon in breast cancer cells”. In: *Cancer genetics and cytogenetics* 136.1 (2002), 43–47.

[172] Mertins, P. et al. “Proteogenomics connects somatic mutations to signalling in breast cancer”. In: *Nature* 534.7605 (2016), 55–62.

[173] Sircoulomb, F. et al. “Genome profiling of ERBB2-amplified breast cancers”. In: *BMC cancer* 10.1 (2010), 1–18.

[174] Kpetemey, M. et al. “MIEN1 drives breast tumor cell migration by regulating cytoskeletal-focal adhesion dynamics”. In: *Oncotarget* 7.34 (2016), 54913.

[175] Kauraniemi, P. et al. “Amplification of a 280-kilobase core region at the ERBB2 locus leads to activation of two hypothetical proteins in breast cancer”. In: *The American journal of pathology* 163.5 (2003), 1979–1984.

[176] Bai, T. and Luoh, S.-W. “GRB-7 facilitates HER-2/Neu-mediated signal transduction and tumor formation”. In: *Carcinogenesis* 29.3 (2008), 473–479.

[177] Yao, I. et al. “SCRAPPER-dependent ubiquitination of active zone protein RIM1 regulates synaptic vesicle release”. In: *Cell* 130.5 (2007), 943–957.

[178] Gregorio, C. C. et al. “The NH2 terminus of titin spans the Z-disc: its interaction with a novel 19-kD ligand (T-cap) is required for sarcomeric integrity”. In: *The Journal of cell biology* 143.4 (1998), 1013–1027.

[179] Wilhelm, L. P. et al. “STARD 3 mediates endoplasmic reticulum-to-endosome cholesterol transport at membrane contact sites”. In: *The EMBO journal* 36.10 (2017), 1412–1433.

[180] Nahta, R. “Pharmacological strategies to overcome HER2 cross-talk and Trastuzumab resistance”. In: *Current medicinal chemistry* 19.7 (2012), 1065–1075.

[181] Paplomata, E. and O’Regan, R. “New and emerging treatments for estrogen receptor-positive breast cancer: focus on everolimus”. In: *Therapeutics and clinical risk management* (2013), 27–36.

[182] Howlader, N. et al. “US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status”. In: *JNCI: Journal of the National Cancer Institute* 106.5 (2014).

[183] Escrivá-de-Romanıé, S. et al. “HER2-positive breast cancer: current and new therapeutic strategies”. In: *The Breast* 39 (2018), 80–88.

[184] Lee, J. J., Loh, K., and Yap, Y.-S. “PI3K/Akt/mTOR inhibitors in breast cancer”. In: *Cancer biology & medicine* 12.4 (2015), 342.

[185] Fujimoto, Y. et al. “Combination treatment with a PI3K/Akt/mTOR pathway inhibitor overcomes resistance to anti-HER2 therapy in PIK3CA-mutant HER2-positive breast cancer cells”. In: *Scientific reports* 10.1 (2020), 1–16.

[186] Chung, W.-P. et al. “PI3K inhibitors in trastuzumab-resistant HER2-positive breast cancer cells with PI3K pathway alterations”. In: *American Journal of Cancer Research* 12.7 (2022), 3067.

[187] Okada, T. et al. “The Rho GTPase Rnd1 suppresses mammary tumorigenesis and EMT by restraining Ras-MAPK signalling”. In: *Nature cell biology* 17.1 (2015), 81–94.

[188] Zhang, W. and Liu, H. T. “MAPK signal pathways in the regulation of cell proliferation in mammalian cells”. In: *Cell research* 12.1 (2002), 9–18.

[189] Azuma, K. et al. “Switching addictions between HER2 and FGFR2 in HER2-positive breast tumor cells: FGFR2 as a potential target for salvage after lapatinib failure”. In: *Biochemical and biophysical research communications* 407.1 (2011), 219–224.

[190] Fernández-Nogueira, P. et al. “Tumor-Associated Fibroblasts Promote HER2-Targeted Therapy Resistance through FGFR2 ActivationTAF and FGFR2 Activation in Breast Cancer Resistance”. In: *Clinical Cancer Research* 26.6 (2020), 1432–1448.

[191] Wang, B. et al. “Induction of tumor angiogenesis by Slit-Robo signaling and inhibition of cancer growth by blocking Robo activity”. In: *Cancer cell* 4.1 (2003), 19–29.

[192] Burotto, M. et al. “The MAPK pathway across different malignancies: a new perspective”. In: *Cancer* 120.22 (2014), 3446–3456.

[193] Sebolt-Leopold, J. S. and Herrera, R. “Targeting the mitogen-activated protein kinase cascade to treat cancer”. In: *Nature reviews cancer* 4.12 (2004), 937–947.

[194] Yu, S. et al. “ERK1 indicates good prognosis and inhibits breast cancer progression by suppressing YAP1 signaling”. In: *Aging (Albany NY)* 11.24 (2019), 12295.

[195] Bivin, W. W. et al. “GRB7 expression and correlation with HER2 amplification in invasive breast carcinoma”. In: *Applied Immunohistochemistry & Molecular Morphology* 25.8 (2017), 553–558.

[196] Nagpal, N. et al. “Essential role of MED1 in the transcriptional regulation of ER-dependent oncogenic miRNAs in breast cancer”. In: *Scientific Reports* 8.1 (2018), 1–14.

[197] Biéche, I. et al. “Two distinct amplified regions at 17q11–q21 involved in human primary breast cancer”. In: *Cancer research* 56.17 (1996), 3886–3890.

[198] Tomasetto, C. et al. “Identification of four novel human genes amplified and overexpressed in breast carcinoma and localized to the q11-q21. 3 region of chromosome 17”. In: *Genomics* 28.3 (1995), 367–376.

[199] Kao, J. and Pollack, J. R. “RNA interference-based functional dissection of the 17q12 amplicon in breast cancer reveals contribution of coamplified genes”. In: *Genes, Chromosomes and Cancer* 45.8 (2006), 761–769.

[200] Alpy, F. and Tomasetto, C. L. “STARD3: A lipid transfer protein in breast cancer and cholesterol trafficking”. In: *Cholesterol Transporters of the START Domain Protein Family in Health and Disease: START Proteins-Structure and Function* (2014), 119–138.

[201] Hongisto, V. et al. “The HER2 amplicon includes several genes required for the growth and survival of HER2 positive breast cancer cells -A data description”. In: *Genomics data* 2 (2014), 249–253.

[202] Lodi, M. et al. “STARD3: A New Biomarker in HER2-Positive Breast Cancer”. In: *Cancers* 15.2 (2023), 362.

[203] Manne, R. K. et al. “FBXL20 promotes breast cancer malignancy by inhibiting apoptosis through degradation of PUMA and BAX”. In: *Journal of Biological Chemistry* 297.4 (2021).

[204] Zhu, L. et al. “Increased SIX-1 expression promotes breast cancer metastasis by regulating lncATB-miR-200s-ZEB1 axis”. In: *Journal of Cellular and Molecular Medicine* 24.9 (2020), 5290–5303.

[205] Hasegawa, N. et al. “Mediator subunits MED1 and MED24 cooperatively contribute to pubertal mammary gland development and growth of breast carcinoma cells”. In: *Molecular and cellular biology* 32.8 (2012), 1483–1495.

[206] Hanker, A. B. et al. “HER2-Overexpressing Breast Cancers Amplify FGFR Signaling upon Acquisition of Resistance to Dual Therapeutic Blockade of HER2FGFR Signaling Promotes Resistance to Dual HER2 Blockade”. In: *Clinical Cancer Research* 23.15 (2017), 4323–4334.

[207] Hergueta-Redondo, M. et al. “Gasdermin B expression predicts poor clinical outcome in HER2-positive breast cancer”. In: *Oncotarget* 7.35 (2016), 56295.

[208] Esteva, F. J. et al. “Clinical utility of serum HER2/neu in monitoring and prediction of progression-free survival in metastatic breast cancer patients treated with trastuzumab-based therapies”. In: *Breast Cancer Research* 7.4 (2005), 1–8.

[209] Ji, Z. et al. “Inflammatory regulatory network mediated by the joint action of NF-kB, STAT3, and AP-1 factors is involved in many human cancers”. In: *Proceedings of the National Academy of Sciences* 116.19 (2019), 9453–9462.

[210] Milanezi, F., Carvalho, S., and Schmitt, F. C. “EGFR/HER2 in breast cancer: a biological approach for molecular diagnosis and therapy”. In: *Expert review of molecular diagnostics* 8.4 (2008), 417–434.

[211] Zanini, E. et al. “The Tumor-Suppressor Protein OPCML Potentiates Anti–EGFR-and Anti–HER2-Targeted Therapy in HER2-Positive Ovarian and Breast Cancer”. In: *Molecular Cancer Therapeutics* 16.10 (2017), 2246–2256.

[212] Hamilton, E. et al. “Targeting HER2 heterogeneity in breast cancer”. In: *Cancer Treatment Reviews* 100 (2021), 102286.

[213] Schettini, F. et al. “Clinical, pathological, and PAM50 gene expression features of HER2-low breast cancer”. In: *NPJ breast cancer* 7.1 (2021), 1.

[214] Tarantino, P. et al. “HER2-low breast cancer: pathological and clinical landscape”. In: *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* 38.17 (2020), 1951–1962.

[215] Rinnerthaler, G., Gampenrieder, S. P., and Greil, R. “HER2 directed antibody-drug-conjugates beyond T-DM1 in breast cancer”. In: *International journal of molecular sciences* 20.5 (2019), 1115.

[216] Modi, S. et al. “Antitumor activity and safety of trastuzumab deruxtecan in patients with HER2-low–expressing advanced breast cancer: results from a phase Ib study”. In: *Journal of Clinical Oncology* 38.17 (2020), 1887.

[217] Jemal, A. et al. “Global cancer statistics”. In: *CA: a cancer journal for clinicians* 61.2 (2011), 69–90.

[218] Cidón, E. U. “The challenge of metastatic colorectal cancer”. In: *Clinical Medicine Insights: Oncology* 4 (2010), CMO–S5214.

[219] Vakiani, E. et al. “Comparative genomic analysis of primary versus metastatic colorectal carcinomas”. In: *Journal of clinical oncology* 30.24 (2012), 2956–2962.

[220] Udali, S. et al. “DNA methylation and hydroxymethylation in primary colon cancer and synchronous hepatic metastasis”. In: *Frontiers in genetics* 8 (2018), 229.

[221] Konishi, K. et al. “DNA methylation profiles of primary colorectal carcinoma and matched liver metastasis”. In: *PloS one* 6.11 (2011), e27889.

[222] Bullman, S. et al. “Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer”. In: *Science* 358.6369 (2017), 1443–1448.

[223] Mamlouk, S. et al. “DNA copy number changes define spatial patterns of heterogeneity in colorectal cancer”. In: *Nature communications* 8.1 (2017), 14093.

[224] Kawamata, F. et al. “Copy number profiles of paired primary and metastatic colorectal cancers”. In: *Oncotarget* 9.3 (2018), 3394.

[225] Roepman, P. et al. “Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition”. In: *International journal of cancer* 134.3 (2014), 552–562.

[226] Tian, S. et al. “A combined oncogenic pathway signature of BRAF, KRAS and PI3KCA mutation improves colorectal cancer classification and cetuximab treatment prediction”. In: *Gut* 62.4 (2013), 540–549.

[227] Popovici, V. et al. “Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer”. In: *J Clin Oncol* 30.12 (2012), 1288–1295.

[228] In’t Veld, S. G. et al. “A computational workflow translates a 58-gene signature to a formalin-fixed, paraffin-embedded sample-based companion diagnostic for personalized treatment of the BRAF-mutation-like subtype of colorectal cancers”. In: *High-throughput* 6.4 (2017), 16.

[229] Tian, S. et al. “A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency”. In: *The Journal of pathology* 228.4 (2012), 586–595.

[230] Huang, S. et al. “MED12 controls the response to multiple cancer drugs through regulation of TGF-β receptor signaling”. In: *Cell* 151.5 (2012), 937–950.

[231] Trumpi, K. et al. “Neoadjuvant chemotherapy affects molecular classification of colorectal tumors”. In: *Oncogenesis* 6.7 (2017), e357–e357.

[232] Isella, C. et al. “Stromal contribution to the colorectal cancer transcriptome”. In: *Nature genetics* 47.4 (2015), 312–319.

[233] Schouten, P. C. et al. “Ovarian Cancer-Specific BRCA-like Copy-Number Aberration Classifiers Detect Mutations Associated with Homologous Recombination Deficiency in the AGO-TR1 Trial”. In: *Clin. Cancer Res* 21.1673 (2021), 1078–0432.

[234] Eide, P. et al. *CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. Sci. Rep. 7, 16618*. 2017.

[235] Becht, E. et al. “Immune and Stromal Classification of Colorectal Cancer Is Associated with Molecular Subtypes and Relevant for Precision ImmunotherapyDistinct Immune Phenotypes of Colorectal Cancer Molecular Subtypes”. In: *Clinical cancer research* 22.16 (2016), 4057–4066.

[236] Van Cutsem, E. et al. “ESMO consensus guidelines for the management of patients with metastatic colorectal cancer”. In: *Annals of Oncology* 27.8 (2016), 1386–1422.

[237] Vellinga, T. et al. “Collagen-rich stroma in aggressive colon tumors induces mesenchymal gene expression and tumor cell invasion”. In: *Oncogene* 35.40 (2016), 5263–5271.

[238] Calon, A. et al. “Stromal gene expression defines poor-prognosis subtypes in colorectal cancer”. In: *Nature genetics* 47.4 (2015), 320–329.

[239] Becht, E. et al. “de Reynié s A (2016b). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression”. In: *Genome Biology* 17.1 (), 218.

[240] Sandberg, T. P. et al. “Molecular profiling of colorectal tumors stratified by the histological tumor-stroma ratio-Increased expression of galectin-1 in tumors with high stromal content”. In: *Oncotarget* 9.59 (2018), 31502.

[241] Elzamly, S. et al. “Epithelial-mesenchymal transition markers in breast cancer and pathological responseafter neoadjuvant chemotherapy”. In: *Breast cancer: basic and clinical research* 12 (2018), 1178223418788074.

[242] Siegel, R. L. et al. “Cancer statistics, 2022”. In: *CA: a cancer journal for clinicians* 72.1 (2022), 7–33.

[243] Chen, W. et al. “Cancer statistics in China, 2015”. In: *CA: a cancer journal for clinicians* 66.2 (2016), 115–132.

[244] Lee, B. L. et al. “Breast cancer in Brazil: present status and future goals”. In: *The lancet oncology* 13.3 (2012), e95–e102.

[245] Devarajan, K. “Nonnegative matrix factorization: an analytical and interpretive tool in computational biology”. In: *PLoS computational biology* 4.7 (2008), e1000029.

[246] Srinivas, P. R. et al. “Proteomics for cancer biomarker discovery”. In: *Clinical chemistry* 48.8 (2002), 1160–1169.

[247] Hondermarck, H. et al. “Proteomics of breast cancer for marker discovery and signal pathway profiling”. In: *PROTEOMICS: International Edition* 1.10 (2001), 1216–1232.

[248] Bertucci, F., Birnbaum, D., and Goncalves, A. “Proteomics of breast cancer: principles and potential clinical applications”. In: *Molecular & Cellular Proteomics* 5.10 (2006), 1772–1786.

[249] Shukla, H. D. “Comprehensive analysis of cancer-proteogenome to identify biomarkers for the early diagnosis and prognosis of cancer”. In: *Proteomes* 5.4 (2017), 28.

[250] Thangudu, R. R. et al. “Abstract LB-242: Proteomic Data Commons: A resource for proteogenomic analysis”. In: *Cancer Research* 80.16_Supplement (2020), LB–242.

[251] Krug, K. et al. “Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy”. In: *Cell* 183.5 (2020), 1436–1456.

[252] Anurag, M. et al. “Proteogenomic markers of chemotherapy resistance and response in triple-negative breast cancer”. In: *Cancer Discovery* 12.11 (2022), 2586–2605.

[253] Wang, B. et al. “Similarity network fusion for aggregating data types on a genomic scale”. In: *Nature methods* 11.3 (2014), 333–337.

[254] Wickham, H. and Wickham, H. “Data analysis”. In: *ggplot2: elegant graphics for data analysis* (2016), 189–201.

[255] Galili, T. “dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering”. In: *Bioinformatics* (2015). DOI: 10.1093/bioinformatics/btv428. URL: https://academic.oup.com/bioinformatics/article/31/22/3718/240978/dendextend-an-R-package-for-visualizing-adjusting.

[256] An, J. et al. “Identification of spliceosome components pivotal to breast cancer survival”. In: *RNA biology* 18.6 (2021), 833–842.

[257] Humphries, B., Wang, Z., and Yang, C. “Rho GTPases: big players in breast cancer initiation, metastasis and therapeutic responses”. In: *Cells* 9.10 (2020), 2167.

[258] Alhareeri, A. A. et al. “Telomere lengths in women treated for breast cancer show associations with chemotherapy, pain symptoms, and cognitive domain measures: a longitudinal study”. In: *Breast Cancer Research* 22.1 (2020), 1–18.

[259] Nawaz, S. et al. “Telomerase expression in human breast cancer with and without lymph node metastases”. In: *American journal of clinical pathology* 107.5 (1997), 542–547.

[260] Bartkova, J. et al. “DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis”. In: *Nature* 434.7035 (2005), 864–870.

[261] Ali, R. et al. “DNA damage repair in breast cancer and its therapeutic implications”. In: *Pathology* 49.2 (2017), 156–165.

[262] Dastsooz, H. et al. “A comprehensive bioinformatics analysis of UBE2C in cancers”. In: *International journal of molecular sciences* 20.9 (2019), 2228.

[263] Bromberg, K. D. et al. “Increased expression of the E3 ubiquitin ligase RNF5 is associated with decreased survival in breast cancer”. In: *Cancer research* 67.17 (2007), 8172–8179.

[264] Mani, A. and Gelmann, E. P. “The ubiquitin-proteasome pathway and its role in cancer”. In: *Journal of clinical oncology* 23.21 (2005), 4776–4789.

[265] Desmurs, M. et al. “C11orf83, a mitochondrial cardiolipin-binding protein involved in bc 1 complex assembly and supercomplex stabilization”. In: *Molecular and cellular biology* 35.7 (2015), 1139–1156.

[266] Porporato, P. E. et al. “Mitochondrial metabolism and cancer”. In: *Cell research* 28.3 (2018), 265–280.

[267] Parihar, R. et al. “A phase I study of interleukin 12 with trastuzumab in patients with human epidermal growth factor receptor-2-overexpressing malignancies: analysis of sustained interferon γ production in a subset of patients”. In: *Clinical Cancer Research* 10.15 (2004), 5027–5037.

[268] Geck, R. C. et al. “Inhibition of the polyamine synthesis enzyme ornithine decarboxylase sensitizes triple-negative breast cancer cells to cytotoxic chemotherapy”. In: *Journal of Biological Chemistry* 295.19 (2020), 6263–6277.

[269] Wang, W. et al. “MAPK4 promotes triple negative breast cancer growth and reduces tumor sensitivity to PI3K blockade”. In: *Nature communications* 13.1 (2022), 245.

[270] Hsu, J. L. and Hung, M.-C. “The role of HER2, EGFR, and other receptor tyrosine kinases in breast cancer”. In: *Cancer and Metastasis Reviews* 35 (2016), 575–588.

[271] Weigman, V. J. et al. “Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival”. In: *Breast cancer research and treatment* 133 (2012), 865–880.

[272] Chan, S. et al. “Basal-A Triple-Negative Breast Cancer Cells Selectively Rely on RNA Splicing for SurvivalRNA Splicing Dependencies in TNBC”. In: *Molecular cancer therapeutics* 16.12 (2017), 2849–2861.

[273] Erdoğan, G. et al. “Investigation of SRP9 protein expression in breast cancer”. In: *Molecular Biology Reports* (2022), 1–7.

[274] Schuster, S. L. and Hsieh, A. C. “The untranslated regions of mRNAs in cancer”. In: *Trends in cancer* 5.4 (2019), 245–262.

[275] Hassan, M. K. et al. “The expression profile and prognostic significance of eukaryotic translation elongation factors in different cancers”. In: *PloS one* 13.1 (2018), e0191377.

[276] Otani, Y. et al. “The Exon Junction Complex Core Represses Cancer-Specific Mature mRNA Re-splicing: A Potential Key Role in Terminating Splicing”. In: *International journal of molecular sciences* 22.12 (2021), 6519.

[277] Hu, X. et al. “Clinical and biological heterogeneities in triple-negative breast cancer reveals a non-negligible role of HER2-low”. In: *Breast Cancer Research* 25.1 (2023), 1–22.

[278] Dehghani, M. et al. “The effects of low HER2/neu expression on the clinicopathological characteristics of triple-negative breast cancer patients”. In: *Asian Pacific Journal of Cancer Prevention: APJCP* 21.10 (2020), 3027.

[279] Jacot, W. et al. “Prognostic value of HER2-low expression in non-metastatic triple-negative breast cancer and correlation with other biomarkers”. In: *Cancers* 13.23 (2021), 6059.

[280] Perez, E. A. et al. “HER2 testing: current status and future directions”. In: *Cancer treatment reviews* 40.2 (2014), 276–284.

[281] Loibl, S. and Gianni, L. “HER2-positive breast cancer”. In: *The Lancet* 389.10087 (2017), 2415–2429.

[282] Morganti, S. et al. “Complexity of genome sequencing and reporting: next generation sequencing (NGS) technologies and implementation of precision medicine in real life”. In: *Critical reviews in oncology/hematology* 133 (2019), 171–182.

[283] Kaissis, G. A. et al. “Secure, privacy-preserving and federated machine learning in medical imaging”. In: *Nature Machine Intelligence* 2.6 (2020), 305–311.

[284] Rajkomar, A., Dean, J., and Kohane, I. “Machine learning in medicine”. In: *New England Journal of Medicine* 380.14 (2019), 1347–1358.

[285] Savage, N. “Calculating disease”. In: *Nature* 550.7676 (2017), S115–S117.

[286] Froelicher, D. et al. “Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption”. In: *Nature communications* 12.1 (2021), 5910.

[287] Clark, L. T. et al. “Increasing diversity in clinical trials: overcoming critical barriers”. In: *Current problems in cardiology* 44.5 (2019), 148–172.

[288] Siegel, R. L., Miller, K. D., and Jemal, A. “Cancer statistics, 2018”. In: *CA: a cancer journal for clinicians* 68.1 (2018), 7–30.

[289] Meldrum, C., Doyle, M. A., and Tothill, R. W. “Next-generation sequencing for cancer diagnostics: a practical perspective”. In: *The Clinical Biochemist Reviews* 32.4 (2011), 177.

[290] De Rubis, G., Krishnan, S. R., and Bebawy, M. “Liquid biopsies in cancer diagnosis, monitoring, and prognosis”. In: *Trends in pharmacological sciences* 40.3 (2019), 172–186.

[291] Warnat-Herresthal, S. et al. “Swarm learning for decentralized and confidential clinical machine learning”. In: *Nature* 594.7862 (2021), 265–270.

[292] D’avanzo, B. and La Vecchia, C. “Risk factors for male breast cancer”. In: *British journal of cancer* 71.6 (1995), 1359–1362.

[293] Couch, F. J. et al. “BRCA2 germline mutations in male breast cancer cases and breast cancer families”. In: *Nature genetics* 13.1 (1996), 123–125.

[294] Haraldsson, K. et al. “BRCA2 germ-line mutations are frequent in male breast cancer patients without a family history of the disease”. In: *Cancer research* 58.7 (1998), 1367–1371.

[295] *The Cancer Genome Atlas*. URL: `https://www.cancer.gov/tcga`.

Chapter 8

List of Publications

Andreas Schlicker*, **Architha Ellappalayam***, Ines J. Beumer*, Mireille HJ Snel, Lorenza Mittempergher, Begona Diosdado, Christa Dreezen et al. "Investigating the concordance in molecular subtypes of primary colorectal tumors and their matched synchronous liver metastasis." International Journal of Cancer 147, no. 8 (2020): 2303-2315.

Midas M Kuilman*, **Architha Ellappalayam***, Andrei Barcaru, Josien C. Haan, Rajith Bhaskaran, Diederik Wehkamp, Andrea R. Menicucci, William M. Audeh, Lorenza Mittempergher, and Annuska M. Glas. "BluePrint breast cancer molecular subtyping recognizes single and dual subtype tumors with implications for therapeutic guidance." Breast Cancer Research and Treatment 195, no. 3 (2022): 263-274.

Haan, Josien C., Rajith Bhaskaran, **Architha Ellappalayam**, Yannick Bijl, Christian J. Griffioen, Ersan Lujinovic, William M. Audeh, Penault-Llorca, Lorenza Mittempergher, and Annuska M. Glas. "MammaPrint and BluePrint comprehensively capture the cancer hallmarks in early-stage breast cancer patients." Genes, Chromosomes and Cancer 61, no. 3 (2022): 148-160.

Architha Ellappalayam, Cristina Furlan, Vitor A.P. Martins dos Santos, Maria Suarez Diez and Edoardo Saccenti. "Molecular Subtyping of Triple-Negative Breast Cancer using Proteomics Data". *Manuscript prepared for submission.*

Architha Ellappalayam*, Andrei Barcaru*, Josien C Haan, Midas M Kuilman, Rajith Bhaskaran, Laura J van ‘t Veer, Lorenza Mittempergher and Annuska M. Glas. "A twenty-nine HER2 biology Guided Gene Signature Improves Breast Cancer HER2-type Molecular Classification". *Manuscript prepared for submission.*

Habibi, Mehran, Danijela Jelovac, Rima Couzi, Cesar Augusto Santa-Maria, Catherine Klein, Marissa White, Nivali Naik, **Ellappalayam Architha** et al. "Abstract P5-09-02: Impact of neoadjuvant endocrine therapy on tumor transcriptome in patients with early-stage breast cancer from the FLEX trial." Cancer Research 83, no. 5Supplement (2023): P5-09.

Kuilman, Midas M., **Architha Ellappalayam**, Lorenza Mittempergher, Diederik Wehkamp, Bob Chan, Rajith Bhaskaran, and Annuska M. Glas. "Corrigendum to “518 Poster-BluePrint molecular subtyping recognizes single and dual subtype tumors with consequences for therapeutic guidance”[Eur J Cancer 138 (Supplement 1)(October 2020) S106–S107]." European Journal of Cancer 174 (2022): 328.

Ellappalayam, A., A. Barcaru, M. M. Kuilman, R. Bhaskaran, W. M. Audeh, L. Mittempergher, and A. M. Glas. "An expanded 29-gene HER2 signature robustly identifies genomic HER2-Type early-stage breast tumors." European Journal of Cancer 175 (2022): S78-S79.
*Authors contributed equally

Chapter 9

Overview of completed training activities

Discipline Specific Activities	**Organizer**
MATLAB workshop - Programming Techniques (2019)	MathWorks
DCS MasterClass on Data Visualisation (2019)	Dutch Chemometrics Society
Data Visualisation Workshop (2022)	RSG Netherlands
BioSB Machine Learning course (2020)	BioSB
BioSB conference + Ph.D. retreat (2020)	BioSB
BioSB conference + Ph.D. retreat (2021)	BioSB
BioSB conference + Ph.D. retreat (2022)	BioSB
Bioinformatics Virtual Symposia (2021)	RSG Belgium
StrengthFinder Workshop + 3 follow-up sessions (2019-2021)	GallupAccess

General courses	
VLAG Ph.D. week (2019)	**VLAG**
StrengthFinder Workshop + 3 follow-up sessions (2019-2021)	GallupAccess
Scientific Writing (2019-2020)	Wur I'nto languages
Objectives and Key Results training (2019-2021)	OKR Workboard
Mindfulness and Meditation (2020)	Meditatiestation
Ph.D. competence assessment (2021)	WGS

Other activities	
Preparation of research proposal (2019)	VLAG
Research & Development meetings - Irvine & Amsterdam (2019-2022)	Agendia
Research & Development meetings - AMS R & D (2020-2022)	Agendia
BioSB Program Committee (2021)	BioSB
BioSB Educational Committee (2020-2022)	BioSB
Young Computational Biologists group (2019-2022)	RSG Netherlands
BioSB Organisational Committee (2022)	BioSB
Ph.D. trip (2022)	SSB/MIB

Chapter 10

About the author

Architha Ellappalayam

Architha Ellappalayam was born on July 4, 1994, in Chennai, India. After completing her secondary education, she pursued a bachelor's degree in Computer Science in the same city. In 2014, Architha embarked on an internship for her bachelor's at Rotterdam in the Netherlands. It was here that she discovered a love for the country and decided to make it her permanent home. Architha continued her academic journey at Wageningen University, pursuing a master's degree in Bioinformatics starting in 2015. Driven by her family's experience with her grandfather's cancer, Architha was drawn to the field of cancer research. For her master's internship, she worked at Erasmus MC, investigating the methylation patterns of testicular germ cell tumors. In March 2018, Architha began her career as an Applied Scientist at Agendia, a global biotech company. Impressed by the research conducted at Agendia, she decided to pursue a collaborative Ph.D. program with Wageningen University. This dream was realized in April 2019, when Architha became an external Ph.D. candidate in the Department of Systems and Synthetic Biology. During her Ph.D., she worked on several projects spanning from technical adjustments and investigations of Agendia's current tests to developing new biomarkers for early-stage breast cancer samples. In June 2023, Architha transitioned to the role of Business Analyst at the Hyve in Utrecht. Outside of her academic and professional pursuits, Architha is an avid traveler who enjoys building miniature houses, painting, and learning to play the violin.

Chapter 11

Acknowledgements

First and foremost, I'd like to express my deepest gratitude to my Promoter **Vitor Martins dos Santos** for welcoming my idea of a collaborative Ph.D. with such enthusiasm and most importantly, for staying by my side through everything that has happened and more. I had the pleasure of meeting my supervisors **Maria Suarez Diez** and **Edoardo Saccenti** at the start of 2016 during the Molecular Systems Biology course, which quickly became my favorite course during my Masters. Their guidance and support inspired me to pursue my Master's thesis, internship, and eventually my Ph.D. with them. Throughout my academic journey, they have been my unwavering source of guidance, always willing to lend a helping hand and offer words of encouragement. Despite any challenges, they have always helped me see the positive in every situation. Your enthusiasm for my collaborative Ph.D. idea gave me the motivation and confidence to pursue this project to completion. This would not have been possible without you.

Second, I'd like to thank **Annuska Glas** from Agendia for her support and guidance, and for enabling my PhD till the end. A very special thanks to my first manager, **Lorenza Mittempergher**, in addition to your supervision, you also took the time to mold me to face the professional world with your constructive feedback and provided never ending support during every step of this Ph.D.

During the final few months of PhD, I have a special few people to thank in addition to my supervisors who were extremely supportive and encouraging to me. During the final two months of my Ph.D., **Cristina Furlan** graciously agreed to guide me on a project and offered me plenty of support and guidance. Your help was crucial in completing my proteomics project, and I am immensely grateful for her unwavering support throughout. Additionally, I would like to extend my gratitude again to **Edoardo Saccenti** for his important contributions to the project.

I am grateful to **Laura van 't Veer** and **Josien Haan** for their invaluable assistance in completing my manuscript at the last minute. Despite the tight deadline, they generously volunteered their time and expertise to help me finish the work. Their dedication and support were instrumental in helping me meet the deadline, and I am deeply appreciative of their efforts. **Sonia Katz's** kindness and generosity came to my aid at a time when I needed it most. When I was at a crucial point in my thesis, Sonia immediately stepped in and offered several options for me to work on. I was overwhelmed and grateful for her help, and I am incredibly thankful for her timely and selfless assistance. To **Marco Anteghini, Sara Moreno Paz, Sara Bennito, and Sanjeevan**, I had not been around for long at our department to get to know you all much more personally, but none of you ever even thought twice before very quickly making me feel comfortable and helping me around, be it at Wageningen, or during our wonderful PhD trip. Thank you so much to all of you!

To **Midas Kuilman**, work never really felt like work thanks to all the fun of being in a team with you. You have been a joy to work with, and I shall miss our coffee break conversations a lot. To **Andrei Barcaru**, our Signatures and Biomarkers team was short-lived yet we had a fantastic period together while it lasted. You were the coolest manager Midas and I could hope for, with you, I definitely miss our very strange lunch conversation topics :). Stay awesome, both of you!

Without the support of my family, I would not be here. First is to my **mom, Uma**. As I sit down to express my gratitude towards you, I am filled with an overwhelming sense of love and admiration for everything you have done for me. You are the epitome of a strong and selfless woman who has made countless sacrifices to give me the life I have today. You have instilled in me the values of hard work, determination, and perseverance, and have always encouraged me to pursue my dreams no matter how big they may seem. Your passion and fierce nature have inspired me to become the best version of myself, and I am grateful for having you as my role model. Words will never be enough to express how much I appreciate and love you. You are the reason I am where I am today, and I want to dedicate my Ph.D. to you as a small token of my gratitude.

To dear **Noud**, as I reflect on my journey to completing my Ph.D., I cannot help but feel grateful for your unwavering support and encouragement. Your presence in my life has been a constant source of comfort and joy, and I am lucky to have you by my side. Through the ups and downs of this journey, you have been my rock, always there to lift me up when I needed it most. I appreciate the way you keep me grounded and calm, even in the most hectic of times. And let's not forget the fun we've had along the way, from playing board games to exploring new places together. I am excited to see where our journey takes us next, and I am grateful to have you as my partner in life. Thank you for everything, Noud.

To my **Amatha**, my dear grandmother. You have been my backbone, my constant support, an ever-present force in my life. You have raised me to be a wonderful, level-headed, and hardworking woman. I am sure no other Amatha in the world has supervised their granddaughter in their daily progress of the PhD, without skipping a single day ever. No other Amatha has managed to play the role of a mother, father, and grandmother all at the same time. Because there is really no one like you.

To my **chithi**, my whole childhood was filled with so many wonderful memories of all the days I have spent with you, Archu, Sanju, and Chithappa. You have been immensely supportive in every phase of my life and I am forever grateful for your support. My two amazing sisters, **Archana** and **Sanjana Senthilkumar**, who have always tried to keep me in high spirits during the highs and lows of the Ph.D. I would like to thank you from the bottom of my heart for everything that you have done, and I know that this will never be enough said. In addition, my very beautiful Ph.D. cover was designed by Archana, with constructive feedback courtesy of Sanjana. So thank you again, girls.

My deep appreciation and my heartfelt gratitude to **Willeke** and **Ger**, you have become some of the most precious people in my life. From the moment I met you, you have been incredibly welcoming, kind, and generous in spirit. Your love and guidance have been invaluable to me, and I am so grateful for all the help they have given me up until now. Thank you both from the bottom of my heart.

I would like to take this opportunity to express my deepest gratitude to my three wonderful friends, **Bhagyashree, Priyanka** and **Ram**. I have known you all since I was just 16 years old, and over the past 13 years, we have built a beautiful friendship that I truly cherish. Each of you holds a very special place in my heart, and I cannot thank you enough for the impact you have had. Thank you for being such an integral part of my life.

To **Rahul Anna, Vidhya** and **Avy**, my wonderful friends but actually more like family. Thank you, Rahul Anna, for constantly having my back and occasionally giving me a reality check urging me to not slack off in my PhD! Thank you, Vidhya, my lovely friend, my happy place for your beautiful friendship and for always making me feel at home with you.

To **Martine** and **Gaby de Kok**, thank you for being a constant source of support and love in my life. From the very first day, I moved to the Netherlands, you welcomed me with open arms and treated me like a daughter. You have been there for me through thick and thin, and I cannot thank you enough for all that you have done for me. Your kindness, generosity, and unwavering love have made a profound impact on my life, and I am forever grateful.

Thank you **Ashley Gallagher** for being the most wonderful friend I could ever ask for. I am glad the day we were scooped up to sit at the same office at Amstel Science Park, although I have more memories of us at the cafeteria or walking around the campus than I do of that office. Our friendship has since transcended to laughing at our miseries together, endless gossiping, and wishing for the next vacation to come sooner. Thanks again and I hope this was not too sappy for you!

To **Marielle** and **Lim**, I am always thankful for the day I decided to sit next to Lim on the first day at Wageningen University and Marielle joined us on the seat behind. What started on that day has continued on to this day, our lives changed in ways we dint imagine, yet our friendship was constant through it all and deepened with time. To many more years of our ever-growing lives and never-ending friendship!

To **Marjolein Sprangers**, it's funny when I look back to see that one accidental text about battery recommendations has bloomed into this wonderful friend of ours. I cherish the moments we've shared over coffee, lunches, and dinners. You're one of the few people with whom I can be completely honest, and I'm grateful for your presence in my life. Thank you, Marjolein, for your kindness and friendship. I am also deeply grateful for your willingness to serve as my paranymph for my PhD defense.

I am thankful to have had the opportunity to be a member of the **YoungCB/RSG Netherlands** during my PhD journey, and later to lead the group as its President for a few years. This experience has introduced me to a wonderful set of individuals in the field of bioinformatics and computational biology. The group has given me the chance to develop and enhance my management and leadership skills, and I am thankful to all the board members for welcoming me into the team. I already miss our wonderful lunch meetings and the joy of organizing successful events together.

The research described in this thesis was financially suuported by Agendia B.V, Amsterdam.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by Archana Senthilkumar