SILESIAN UNIVERSITY OF TECHNOLOGY
Faculty of Automatic Control, Electronics
and Computer Science

Machine learning methods in support of multiomics signature identification for breast cancer patient subpopulations

Doctoral Dissertation

Joanna Tobiasz

Supervisor:
Professor Joanna Polańska, PhD, DSc

Co-supervisor:
Christos Hatzis, PhD

2023
Gliwice, POLAND

In the preface of this work, I would like to express my sincere thanks:

To Prof. Joanna Polańska for her trust in me, invaluable guidance, help, and the opportunity she gave me.

To Dr. Christos Hatzis for the opportunity to cooperate, his expertise, and his devoted time.

To my dear friends in the Department of Data Science and Engineering for their advice, help, smile, and for being by my side through all the ups and downs.

To my beloved Parents for their love, support, patience, and wisdom.

Abstract

Breast cancer is a highly heterogeneous disease with a diverse molecular portrait. The commonly used clinical classification of breast cancer subtypes relies on levels of several protein markers, while molecular classification is defined based on gene expression profiling. Both those divisions remain unchanged for years and do not sufficiently reflect the complex structure of the disease and observed clinical experience diversity. With the rapid progress in molecular biology, more accurate characterization of breast cancer subtypes may support the search for new therapeutic targets. This dissertation aims to develop machine learning-based methods to identify novel subpopulations of breast cancer patients and examine their unique molecular and clinical characteristics.

The tested combinations of feature engineering and clustering approaches and proposed comparison methods allow the division of patients based on their proteomic profiles. Six subpopulations were identified. They were evaluated demographically, clinically, and molecularly based on their protein and transcriptomic profiles. Suitable classic statistical analysis methods supported by effect size estimates and machine learning algorithms allowed for dealing with the comparison groups' different, sometimes insufficient, sizes.

Three of the six subpopulations derived from the proteomic profile were highly consistent with commonly used transcriptomic-based subtypes: basal, HER2-enriched, and luminal B. Nevertheless, the transcriptomic-based luminal A subtype was highly heterogeneous and divided into three subgroups in this work. Revealed subpopulations vary in survival experience and proteomic and transcriptomic profiles. Novel luminal subtypes are less differentiated at the transcriptomic level than in proteomic space. The sets of markers specific for certain subpopulations and the signature enabling distinction between all subtypes were obtained.

The obtained profiles of revealed subpopulations, especially the proteomic one, may potentially complement the used classifications of breast cancer and support the search for novel targeted therapies with the development of personalized medicine. Nevertheless, the independent validation of those findings is required to assess clinical applications further.

Streszczenie

Nowotwór piersi jest chorobą mocno zróżnicowaną o wysoce heterogenicznym obrazie molekularnym. Stosowana powszechnie klasyfikacja kliniczna oparta jest na poziomie kilku białek markerowych, natomiast klasyfikacja molekularna powstała na podstawie profili ekspresji genów. Oba te podziały pozostają niezmienne od lat i nie odzwierciedlają dostatecznie złożonej struktury tej choroby oraz zróżnicowania zachowań klinicznych. Określenie podtypu raka piersi jest kluczowe przy wyborze terapii. W obliczu szybkiego rozwoju biologii molekularnej, dokładniejsze scharakteryzowanie podtypów nowotworu piersi może wesprzeć poszukiwanie nowych celi terapeutycznych. Celem tej pracy było zastosowanie metod uczenia maszynowego do identyfikacji i klinicznego oraz molekularnego scharakteryzowania podpopulacji pacjentek z rakiem piersi.

Przetestowane kombinacje metod inżynierii cech i klastrowania oraz zaproponowane sposoby ich porównania pozwoliły na pogrupowanie pacjentek w oparciu o profil białkowy. Zidentyfikowano sześć podpopulacji pacjentek, które oceniono pod kątem demograficznym, klinicznym i molekularnym na podstawie profili białkowych i transkryptomicznych. Dobrane metody klasycznej analizy statystycznej wspartej miarą wielkości efektu i uczeniem maszynowym pozwoliły zmierzyć się z problemem różnych, czasem niedostatecznych, rozmiarów grup.

Trzy z uzyskanych na podstawie profilu białkowego podpopulacji wykazały dużą zgodność ze stosowanymi podtypami opartymi o poziom mRNA: podstawnym, HER2-wzbogaconym oraz luminalnym B. Transkryptomiczny podtyp luminalny A okazał się wysoce zróżnicowany i został podzielony w tej pracy na trzy podgrupy. Otrzymane podpopulacje różnią się przeżywalnością oraz profilami białkowymi i transkryptomicznymi. Uzyskano zestawy markerów specyficznych dla podpopulacji oraz sygnaturę pozwalającą na rozróżnienie wszystkich podtypów. Nowe podtypy luminalne są mniej zróżnicowane w przestrzeni transkryptomicznej niż w białkowej.

Profile wykrytych podpopulacji, zwłaszcza białkowy, mogą uzupełnić stosowane klasyfikacje raka piersi i wesprzeć poszukiwania nowych terapii celowanych w rozwoju medycyny spersonalizowanej. Niemniej jednak, do dalszej oceny przydatności klinicznej otrzymanych wyników niezbędna jest ich niezależna walidacja.

Contents

1 Introduction

1.1 Motivation

Breast cancer is a highly heterogeneous disease with diverse clinical outcomes, manifesting various molecular and histological backgrounds (Szymiczek, Lone, & Akbari, 2020). The routinely used clinical classification of breast cancer cases remains unmodified over several decades, based on expressions of several marker genes and proteins. Hence, it does not perfectly reflect the molecular portraits of breast cancer patients and has several limitations.

Gene expression profiling allowed the identification of five intrinsic molecular subtypes of breast cancer in the early 2000s. They are still referred to as the gold standard, despite noteworthy inconsistencies with clinical classification, even though one of those subtypes is already widely regarded as an artifact and rarely used. With the increased biological knowledge and a better understanding of tumor molecular background, the intrinsic classification appears to insufficiently reflect the complex character of breast cancer and the diversity of tumor behaviors. Moreover, various mechanisms affect the gene expression between transcriptomic and proteomic layers, which remain unrepresented by currently used breast cancer classifications.

Advances in high-throughput technologies for expression investigation beyond the transcriptomic level and in machine learning approaches for biological big data mining now provide the possibility to retrieve a more comprehensive insight into breast cancer stratification. Nonetheless, large data sets delivered by high-throughput analytical techniques require thoughtful and statistically advanced analysis to appropriately assess the variability in the data and accurately select the most informative features explaining the diversity and distinguishing breast cancer subtypes. Therefore, providing a pipeline with dedicated statistical learning techniques, including unsupervised methods to deliver stratification uninfluenced by well-established breast cancer subtyping, is worthwhile and crucial for drawing biologically relevant conclusions.

The re-identifying breast cancer subtypes may complement the existing subtyping approaches and reflect previously hidden sources of tumor diversity. Accurate breast cancer subtype determination is crucial for treatment choice and allows for better prognosis prediction. Besides, a broad examination of disease subtypes can deliver clinically relevant information that could be used to discover new candidate therapeutic targets. This may find applications in personalized medicine and improve therapy tailoring, which now aims to provide each patient with a possibly optimized and individualized treatment plan to reduce side effects.

1.2 Aim of the work

This dissertation aimed to identify and evaluate breast cancer patient subpopulations. As the already existing and well-established intrinsic molecular subtypes were developed with gene expression profiling, the re-identification in this work relies on the proteomic profiles. The first step of the investigation required the development of machine learning-based approaches for subpopulation detection and the methods to assess the performance of tested algorithms.

Subsequently, the breast cancer subpopulations proposed with the appropriate machine learning pipeline must be evaluated and characterized. The purpose was to investigate the revealed subtypes regarding their clinical experience. The final goal was to provide statistical tools and machine learning methods for identifying molecular signatures of revealed subpopulations. Based on the statistical test supported by the corresponding effect size measures, the molecular signature describing the proteomic and transcriptomic differences between identified patient subpopulations was delivered and investigated with a literature review and dedicated functional analysis methods.

Based on the motivation and the aim of this dissertation, the following theses have been formulated:

I. The application of advanced machine learning and mathematical modeling methods allows the identification of novel molecularly different subpopulations of breast cancer patients.
II. In the case of highly imbalanced and varying-in-size samples, comprehensive statistical testing supported by effect size analysis allows the definition of robust molecular and clinical subtype profiles.

1.3 Chapter contents

The second chapter, "Background", contains the biological background of the dissertation. Clinical and intrinsic molecular classifications are described, along with their discordance, methods required for their development and application, and limitations. This chapter also provides information concerning various approaches to cancer subtyping.

The third chapter, "Materials and methods", presents the data sets used in this project. A description of the pipeline applied for the analysis is provided. Firstly, the batch effect identification and correction methods are explained. Subsequently, various machine learning attempts to identify patient subpopulations are described, including feature engineering methods and clustering algorithms. Moreover, the proposed metrics for the comparison of clustering outcomes are presented. Next, procedures for evaluation of the obtained patient subpopulations are described. The analysis of survival outcomes and a comparison of demographic and clinical profiles of each subpopulation, involving both numerical and categorical variables, are presented. Finally, the procedure for molecular characterization of identified subpopulation signatures is explained.

The fourth chapter, "Identification of patient subpopulations", presents the clustering results obtained with various machine learning approaches, their comparison, and the chosen solution outcome serving as a proposed breast cancer subpopulations further referred to also as subtypes.

The fifth chapter, "Clinical characteristics of patient subpopulations", contains the evaluation of the proposed subpopulations with their clinical

and demographic characteristics, referred to the breast cancer intrinsic subtypes. The most essential survival analysis results and comparison of available clinical and demographical data are presented.

The sixth chapter, “Molecular signature of patient subpopulations”, presents the most important results of various approaches to identifying molecular signatures of revealed subpopulations. Proteomic and transcriptomic differences between subtypes are investigated, and a list of potential biomarkers is proposed.

The seventh chapter, “Summary and conclusions”, summarizes the most important achievements of this work.

The eight chapter provides supplementary materials.

2 Background

2.1 Breast cancer

Breast cancer is the most commonly diagnosed cancer in women and the primary reason for female cancer death (Bray, et al., 2018). According to World Health Organization (WHO), only in 2020, breast cancer was the cause of 685 000 deaths and was diagnosed in 2.3 million women globally. Furthermore, at the end of 2020, 7.8 females alive were diagnosed with breast cancer in the previous five years, which means breast cancer is now the most widespread cancer worldwide (World Health Organization, 2023).

Breast cancer is also a highly heterogeneous disease, with large diversity at pathological, molecular, and clinical levels. Various attempts at the task of breast cancer stratification have been made worldwide. Prognosis, aggressiveness, and therapy response vary among breast tumors due to their molecular background, specific tumor biology, and sensitivity to treatment options. Hence, accurate classification and proper diagnosis are crucial for the optimal treatment choice and planning (Jassem, Shan, & Buczek, 2020; Szymiczek, Lone, & Akbari, 2020; Norum, Andersen, & Sørlie, 2014).

.

2.2 Clinical classification of breast cancer

Currently used clinical breast cancer subtypes are determined based on the presence of three key markers: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) (Jassem, Shan, & Buczek, 2020). Moreover, some approaches also use the cellular proliferation marker Ki67 for patient stratification. Both ER and PR are nuclear hormone receptors (HRs) that serve as transcriptional regulators of many genes' expression. Their status (positive or negative, according to the verified thresholds) is often consistent, as PR synthesis depends on estrogen (Osborne, Yochmowitz, Knight, & McGuire, 1980; Mueller, Haymond, Davis, Williams, & Espina, 2018). HER2 participates in proliferation pathways activation. Its gene *ERBB2* is regarded as oncogenic, and its amplification is associated with higher cancer invasiveness and worse prognosis (Szymiczek, Lone, & Akbari, 2020; Slamon, et al., 1987; Fragomeni, Sciallis, & Jeruss, 2018)

The routinely used breast cancer clinical classification involves four subtypes. The most common one, accounting for up to 70% of cases, is the hormone receptor-positive (HR+). It is defined with negative HER2 status (HER2-) and positive ER or PR statuses (ER+, PR+). HR+ tumors are associated with better prognosis and relatively low aggressiveness. They are also sensitive to HR-targeted endocrine therapy, frequently allowing for successful treatment with good clinical outcomes (Jassem, Shan, & Buczek, 2020; Szymiczek, Lone, & Akbari, 2020; Cho, 2016).

HER2-positive (HER2+) breast cancer contains cases with enriched HER2 receptor but low ER and PR hormone levels. This cancer subtype is associated with poor prognosis and high invasiveness (Szymiczek, Lone, & Akbari, 2020). It weakly responds to endocrine treatment or chemotherapy (Mueller, Haymond, Davis, Williams, & Espina, 2018). However, it is sensitive to the targeted anti-HER2 therapy based on the humanized monoclonal antibody (Jassem, Shan, & Buczek, 2020). Overexpression of HER2 is observed in approximately 20%-25% of breast cancer patients (Garrett & Arteaga, 2011).

The worst prognosis and the most limited treatment options are associated with Triple-Negative Breast Cancer (TNBC), defined as ER-, PR-, and HER2-. Insensitive to dedicated treatments like endocrine or anti-HER2 therapies, TNBC is a subject of research and exploration in search of potential therapeutic targets. Currently, TNBC is treated mainly with chemotherapy with a response outperforming other clinical subtypes (Dai, et al., 2015; Norum, Andersen, & Sørlie, 2014; Szymiczek, Lone, & Akbari, 2020).

The last clinical subtype, defined as ER+, PR+, and HER2+, is called Triple Positive (TPBC). A mix of endocrine, chemo-, and anti-HER2 therapies can be applied for this subtype. Moderate prognosis is associated with those cases (Szymiczek, Lone, & Akbari, 2020).

In clinical practice, ER, PR, and HER2 statuses are determined using immunohistochemistry (IHC), a simple, cost-effective, and thus widely available technique (Jassem, Shan, & Buczek, 2020; Zaha, 2014). IHC involves detecting the protein of interest by the specific primary antibody, to which the so-called secondary antibody is later attached with a reporter molecule. Following the antibody-antibody binding, another substrate is added, which reacts with the receptor molecule, generating the color complex visible under the microscope in specific locations (The Human Protein Atlas, 2023). In equivocal cases of HER2 status, IHC is supported by fluorescent in situ hybridization (FISH)– a technique using fluorescent oligonucleotides complementary to the DNA fragment of interest, in this case, the amplified *ERBB2* gene. Fluorescent DNA probes bind to the studied sequence and produce a colored signal, detectable with the fluorescent microscope (Wolff, et al., 2018; Hu, et al., 2014).

Therefore, both IHC and FISH are based on the visual inspection, requiring manual counting of the detected complexes or automatic image analysis. In that manner, both methods allow for visual assessment of tissue morphology and tumor heterogeneity. On the other hand, considerable limitations have been reported for both IHC and FISH. Those mainly result from technical factors and scoring subjectivity, despite the effort made by the American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP) to standardize both

approaches. IHC results can be biased by sample type, preparation procedure, and choice of equipment and subtracts. Technical variations may also influence FISH outcomes. Moreover, interpreting obtained results depends on the scoring system and selected thresholds (Szymiczek, Lone, & Akbari, 2020). About 6% of all breast cancers are considered borderline tumors in terms of ER, meaning that the fraction of tumor cells demonstrating the nuclear staining is 1%-9% (Allred, et al., 2009). According to ASCO/CAP guidelines, a percentage below 1% indicates negative ER status. HER2 scoring systems classify samples according to a metric determined as the sum of the staining intensity and positive cell fraction. Categories 0 and 1+ indicate HER2-, while 3+ mean HER2+. Category 2+ contains equivocal cases, which include approximately 4%-12% of all breast cancer patients (Wolff, et al., 2018). Thus, a notable proportion of breast cancer cases fails to be unambiguously determined regarding all three markers. Furthermore, the scoring is frequently biased by differences in expertise or judgment made by laboratory technicians who carry it out. Various factors can influence the experts' assessment (Szymiczek, Lone, & Akbari, 2020).

2.3 Intrinsic molecular classification of breast cancer

Currently used clinical subtypes demonstrate too high diversity in terms of therapy outcomes and response to applied treatment. Hence, new approaches to breast cancer classification were proposed with the advancements in high-throughput platform development and a better understanding of molecular biology of cancer. Initially, the within-subtype diversity in clinical outcomes was assumed toe reflected in gene expression pattern variation (Szymiczek, Lone, & Akbari, 2020). In 2000, (Perou, et al., 2000) examined the gene expression of 65 breast cancer specimens with complementary DNA (cDNA) microarrays. Hierarchical clustering of that data revealed four intrinsic subtypes of breast tumors, varying in their molecular portraits. Those clusters were denoted as ER+/Luminal-Like, Erb-B2+ (HER2), Basal-Like, and Normal Breast-Like.

In 2001, (Sørlie, et al., 2001) published an extension to work by Perou et al., also based on cDNA microarray data clustering. As a result, the previous ER+/Luminal-Like subtype was further divided into two or even three novel

subgroups: Luminal A, Luminal B, and eventually Luminal C. Additional Luminal C subtype did not progress to further applications. Normal Breast-Like subtype, very rare in all cohorts, is now only occasionally in use, as it was suspected to be an artifact resulting from the sample contamination with normal epithelial or stromal cells (Parker, et al., 2009; Weigelt, et al., 2010). Finally, four intrinsic molecular subtypes are well-established.

The first subtype, Luminal A, is characterized by high expression levels of HRs and luminal epithelial genes and a low level of HER2. Hence, those tumors can be successfully treated with endocrine therapy and are associated with a good prognosis and survival outcome. Chemotherapy is recommended for treatment only in high-risk cases (Jassem, Shan, & Buczek, 2020). Luminal A cases are also low grade and have a low proliferation. They were reported to show frequent non-silent mutations in PIK3CA, TP53, GATA3, and MAP3K1 pathways, as well as cyclin D1 amplification and high *RB1* expression (Dai, et al., 2015; Norum, Andersen, & Sørlie, 2014; Fragomeni, Sciallis, & Jeruss, 2018; The Cancer Genome Atlas Network, 2012).

Compared to luminal A, luminal subtype B has low HR levels, high grade, and beneficial survival outcomes. In some luminal B cases, HER2 levels are elevated, worsening the prognosis, but also being a target for anti-HER2 therapy (Norum, Andersen, & Sørlie, 2014; Guiu, et al., 2012). Generally, the prognosis in luminal B cases is moderate. This subtype is sensitive to endocrine therapy, but chemotherapy is also often applied. Luminal B subtype shows mutations mainly in TP53, PIK3CA, and GATA3 pathways (Dai, et al., 2015; Norum, Andersen, & Sørlie, 2014; Fragomeni, Sciallis, & Jeruss, 2018; The Cancer Genome Atlas Network, 2012).

HER2-enriched subtype shows high levels of HER2 and low expression of luminal epithelial genes. Moreover, it has a high proliferation rate and poor prognosis, but it is sensitive to anti-HER2 targeted therapy, which significantly improves the clinical outcome (Dai, et al., 2015; Szymiczek, Lone, & Akbari, 2020). Chemotherapy also is used for treatment (Jassem, Shan, & Buczek, 2020). However, contrary to the corresponding clinical subtype, the HER2-enriched intrinsic subtype is not determined by only HER2 and ER statuses. It is instead characterized

by the entire EGFR/HER2 signaling pathway, which also involves EGFR, HER3, and HER4 (Godoy-Ortiz, et al., 2019; Moasser, 2007). Furthermore, HER2-enriched cases frequently show TP53, PIK3CA, and PIK3RI pathway mutation(The Cancer Genome Atlas Network, 2012).

The last and the most specific intrinsic subtype is Basal-Like, in which luminal genes, HR, and HER2 are not expressed. However, genes characteristic for basal cells are highly expressed. The prognosis for this subtype is the worst, as Basal-Like tumors manifest high grade, invasiveness, aggressiveness, and progression rate (Dai, et al., 2015; Norum, Andersen, & Sørlie, 2014). Chemotherapy is a man treatment option due to the lack of targeted therapy options against this subtype (Jassem, Shan, & Buczek, 2020; Szymiczek, Lone, & Akbari, 2020). Basal-Like tumors were reported to show mutations in the TP53 pathway and *BRCA1* genes, overexpression of EGFR, and elevated WNT pathway activation (The Cancer Genome Atlas Network, 2012).

Initially, clinical and intrinsic subtypes were regarded as consistent. HER2+ and HER2-enriched, TNBC and Basal-Like, and HR+ and Luminal subtypes were assumed interchangeable, with Luminal A and B being distinguishable based on Ki67 protein levels (Szymiczek, Lone, & Akbari, 2020; Sali, et al., 2020).

Nevertheless, with the growing availability of high-throughput platforms and the increasing number of studies concerning breast tumor profiling, a noteworthy discrepancy between clinical and intrinsic subtypes has been suggested. Approximately one-third of tumors subtype-labeled both clinically and molecularly have been reported to demonstrate discordant outcomes (Prat, et al., 2015). For instance, about half of HER2-enriched cancers are negative for *ERBB2* amplification but still manifest significant similarities to HER2-amplified tumors in their molecular profiles (Daemen & Manning, 2018). Thus, HER2 amplification is not requisite for the HER2-enriched subtype but serves as only one of the factors (Szymiczek, Lone, & Akbari, 2020). Furthermore, only about 70% of TNBCs are Basal-Like. In the Basal-Like group, 23% of cases are clinically labeled with subtypes other than TNBC. That is an important observation, as tumors sensitive to targeted hormone or HER2 therapy also appear in the Basal-Like group

(Bertucci, et al., 2008). The highest concordance was observed for HR+ and Luminal subtypes, as only 5% of HR+ cases were categorized as HER-enriched and even less as Basal-Like (Szymiczek, Lone, & Akbari, 2020).

2.4 Gene expression profiling for intrinsic subtyping

Gene expression profiling methods aim to quantitatively characterize each examined gene with its expression level. In developing multigene tools for breast cancer prognosis prediction based on the molecular profile, four leading technologies are mainly used: quantitative Real-Time Polymerase Chain Reaction (qRT-PCR), DNA microarrays, RNA-sequencing (RNA-Seq), and the most recent one - NanoString nCounter® (Szymiczek, Lone, & Akbari, 2020).

qRT-PCR is based on cDNA amplification with probes modified to produce a fluorescent signal. Gene expression level is characterized as the number of amplification cycles required to achieve the set threshold (Szymiczek, Lone, & Akbari, 2020). However, PCR is target-specific, meaning the primers complementary to the appropriate sequence determine the cDNA fragment to amplify. Thus, the number of genes examined in that manner must be limited (National Center for Biotechnology Information, 2023).

DNA microarrays are the essential technology for both this dissertation and determining the "gold standard" intrinsic breast cancer subtypes. They were developed as a primary high-throughput technology, allowing for simultaneous measurements of thousands of genes. DNA microarrays are based on the hybridization of fluorescently labeled sample genetic material to the oligonucleotide probes at the microarray. The expression levels of genes corresponding to probes are estimated based on the intensity of the generated fluorescent signal (Szymiczek, Lone, & Akbari, 2020).

RNA-Seq is based on the Next Generation Sequencing (NGS) of the library prepared with cDNA produced from the sample material and ligated with adapters. The obtained sequences are subsequently mapped to the transcriptome, counted for particular locations, and pre-processed in a dedicated manner. RNA-Seq provides a more comprehensive insight into the transcriptomic profile and is now

widely used for gene expression characterization compared to DNA microarrays. However, in terms of gene expression of the breast cancer markers ER, PR, and HER2, both technologies provide comparable results (Fumagalli, et al., 2014). The high consistency of gene expression profiling outcome between DNA microarrays and RNA-Seq for the TCGA cohort was also demonstrated (Guo, et al., 2013).

NanoString's nCounter® technology is a state-of-the-art method that uses two probes complementary to the target sequence. One, called a capture probe, is responsible for immobilization and purification, the other - a reporter probe – carries a unique fluorescent barcode. However, this technology allows only up to 800 genes to be measured in a single run (Szymiczek, Lone, & Akbari, 2020; BioXpedia, 2023).

Gene expression profiling methods also have certain limitations, apart from varying numbers of genes to be measured simultaneously. Issues connected with poor reproducibility and batch effect were stated. The analysis pipeline was reported to influence the obtained results greatly. Hence, proper standardization, normalization, and batch effect removal are crucial for stability and reproducibility (Szymiczek, Lone, & Akbari, 2020; Larsen, Thomassen, Tan, Sørensen, & Kruse, 2014). Furthermore, the bulk gene expression profiling risks experiencing bias from intra-tumoral heterogeneity. The tumor sample can be substantially contaminated with neighboring histologically benign tissue. The genetic material from tumor and non-tumor cells is mixed afterward, and gene expression is measured jointly. Hence, the normal cell expression level may affect the measurements to the extent depending on the proportion of non-tumor cells in the specimen (Elloumi, et al., 2011).

Despite the abovementioned concerns, several commercial multigene tests were developed and are currently in use for breast cancer prediction and screening patients for chemotherapy. However, those signatures mainly include ER-related and proliferation genes and thus are limited to ER+ cases (Prat, et al., 2012). Oncotype DX® (Exact Sciences), Breast Cancer Index SM (BCI) (bioTheranostics), EndoPredict (Myriad Genetics, Inc.), and MammaTyper® (Cerca Biotech) tests apply

qRT-PCR. MammaPrint® (Agendia) and BluePrint® (Agendia) assays are based on microarray technology. 50-gene Prediction Analysis of Microarray (PAM50) Prosigna Risk of Recurrence assay (Nanostring Technologies) uses the nCounter® system (Jassem, Shan, & Buczek, 2020; Szymiczek, Lone, & Akbari, 2020; Vieira & Schmitt, 2018; Győrffy, et al., 2015; Wallden, et al., 2015).

The PAM50 classifier mentioned above was first published by (Parker, et al., 2009). It aimed to predict the chemotherapy benefit and breast cancer prognosis. Moreover, it allows the intrinsic subtype diagnosis for four molecular subtypes: basal-like, HER2-enriched, luminal A, and luminal B.

The classifier was developed using microarray data supported by the qRT-PCR results. Firstly, 1906 candidate genes for the analysis were selected based on the literature review. Expression levels of those genes measured for 189 breast tumor samples were median centered and hierarchically clustered with average linkage and Pearson correlation as the distance metrics. Results provided the set of prototypical genes and 122 tumor samples, for which significant clusters representing five intrinsic subtypes (four mentioned above and normal-like) were detected. Subsequently, the qRT-PCR and several minimalization procedures performed on the prototypic samples delivered the 50 genes distinguishing the subtypes. Finally, the reproducibility and robustness of that gene signature were assessed with three approaches of classification based on the nearest of the five centroids: Prediction Analysis of Microarray (PAM) (Tibshirani, Hastie, Narasimhan, & Chu, 2002), simple nearest centroid method (Hu, et al., 2006), and Classification of Nearest Centroid (Dabney, 2005). PAM outperformed the other methods regarding subtype prediction reproducibility; hence, this transcriptomics-based predictor is called PAM50. Normal-like subtype was later recognized as the artifact resulting from the contamination of tumor samples with normal breast tissue and was no longer considered.

Since it was first proposed, the PAM50 classifier has become a “gold standard” in the molecular classification of breast cancer. The outcomes of this predictor for the microarray data will be used in this dissertation as the reference.

PAM50 has also been further modified. The commercially applied Prosigna test relies on the nCounter® adaptation of PAM50 (Wallden, et al., 2015).

2.5 Breast cancer subtyping approaches

Various machine learning approaches have been applied for cancer subtyping and further evaluating the obtained stratification. Hierarchical clustering is the most common method. It was initially used to propose intrinsic molecular subtypes based on gene expression profiling. In (Perou, et al., 2000), it delivered four clusters, and later in work by (Sørlie, et al., 2001), this division was extended to five or even six groups. Hierarchical clustering of gene expression levels also served for developing the PAM50 predictor, as described above (Parker, et al., 2009). For evaluating subtypes obtained in those works, mainly classical overall or relapse-free survival analysis served complemented with clinical information regarding the therapy used, the tumor size, or the number of lymph nodes affected.

The Cancer Genome Atlas (TCGA) Network published the breast cancer patients cohort analysis results concerning six platforms, incorporating Agilent mRNA expression microarrays, DNA methylation and single nucleotide polymorphism arrays, miRNA and exome sequencing, and protein levels. Apart from exome sequencing, each platform was used separately for subtyping with a reliable method depending on the data type. The PAM50 predictor and hierarchical clustering were applied for mRNA expression microarray data. Furthermore, non-negative matrix factorization served for subtype identification on miRNA sequencing data and protein levels. It was later complemented with visual inspection and semi-supervised hierarchical clustering with Pearson correlation as a distance metric. Given the beta-distribution of DNA methylation data, the recursively partitioned mixture model served for this analysis. The results were compared to PAM50 results, clinical classification based on HER2 and HR statuses, tumor size, node status, and selected gene mutations. Moreover, the coordinated analysis and comparison were performed for all subtyping results, which revealed high concordance with PAM50 labels (The Cancer Genome Atlas Network, 2012; Brunet, Tamayo, Golub, & Mesirov, 2004).

In (Sotiriou, et al., 2003), hierarchical agglomerative clustering served for breast cancer stratification based on the gene expression profiles. Both Euclidean and one minus Pearson correlation distance metrics were tested. Two general patient groups were obtained, representing ER+ and ER- cases. Both were further split into three smaller subgroups. Gene expression profiles were assessed for association with ER status and clinical factors, including tumor size, node status, and menopausal status. Moreover, relapse-free survival was investigated with Cox proportional hazard model for additional evaluation of results.

In (Hu, et al., 2006), hierarchical clustering again served for subtyping the combined data set coming from (Sørlie, et al., 2001), (Sørlie, et al., 2003), (van't Veer, et al., 2002), and (Sotiriou, et al., 2003). The data were grouped into five already known intrinsic subtypes, and a new group was also revealed with overexpressed Interferon (IFN)-regulated genes. Moreover, an updated gene signature differentiating these subpopulations was obtained. The results were evaluated with relapse-free survival analysis based on a log-rank test and hazard ratio estimation.

In (El-Rehim, et al., 2005), tissue microarray technology combined with IHC served to analyze protein levels of breast cancer specimens. Five groups differentiating in proteomic profiles were revealed with hierarchical clustering. Biomarkers characteristic for each subtype were identified with a neural network approach. The groups were evaluated with the classic log-rank test of overall and disease-free survival. Furthermore, tumor grade, size, and histologic type were examined regarding the identified groups.

In (Jönsson, et al., 2010)., hierarchical clustering on DNA copy-number data provided six clusters. They were referred to known PAM50 labels, ER statuses, and mutations in the *BRCA1* gene. Moreover, the overall survival of the revealed subtypes was compared with the classic log-rank approach.

In (Lehmann, et al., 2011), k-means and consensus clustering approaches were used to divide the TNBC gene expression data set. The optimal number of clusters was found with the area under the curve of the consensus distribution function and was later visually inspected with Principal Components Analysis (PCA)

results (Hotelling, 1933). Each revealed subtype was referred to all remaining ones and tested for gene enrichment using the GSE-A method (Subramanian, et al., 2005). Subtype gene signatures were obtained with the Kruskal-Wallis test, followed by Bonferroni correction for multiple testing. Moreover, pairwise comparisons evaluated the subtypes based on the survival outcome compared with the log-rank test and Cox proportional hazard model.

In (Guedj, et al., 2011), hierarchical clustering, Gaussian mixture models, and k-means clustering were parallelly applied to the mRNA expression microarray measurements. A set of samples that were assigned to the same cluster in all three approaches served for analysis of variance. In that manner, the genes with the highest intragroup homogeneity and intergroup heterogeneity were selected. Hierarchical clustering conducted on this reduced data set provided six clusters and gene patterns characteristic for each cluster. The revealed subtypes were compared based on disease-free survival, age, metastatic sites, copy-number alterations, and differential activation of signaling pathways.

With the recent emergence of advanced machine learning approaches for multimodal data integration, state-of-the-art methods for cancer subtyping have started to appear lately. The two kinds of multiomics clustering approaches have been developed so far. The first one involves the separate analysis of each modality and then combining the results. The advantage of this approach is simplicity and better control over technical factors like batch effect per modality. The second type of methods involves combining the different data types before the joint model creation (Liu, Cheng, Jin, & Hu, 2022). Some methods are based on integrating data by estimating per-modality similarities; others incorporate dimensionality reduction methods or probabilistic modeling (Rappoport & Shamir, 2018). All approaches are rapidly developing, despite several challenges and limitations. Those involve various patterns of missing values, batch effects, heterogeneity of the data, including various ranges, scales, and distributions, a growing number of features with few observations, and complex correlation structure (Argelaguet, Cuomo, Stegle, & Marioni, 2021; Stuart & Satija, 2019).

Nevertheless, several approaches for multimodal subtyping have been proposed recently. MODEC approach uses manifold optimization to transform the multimodal data into a low-dimensional latent subspace, which then serves for a deep-learning-based clustering module. The obtained subtypes are evaluated based on survival analysis and comparison of clinical features (Zhang & Kiryu, 2022). In (Liu, Cheng, Jin, & Hu, 2022), another approach was proposed, which uses Bayesian tensor factorization for multimodal integration and consensus clustering with k-means for subtype identification. Obtained six subtypes are evaluated with survival analysis, differential analysis of gene expressions, and gene set enrichment analysis. Moreover, demographical and clinical factors serve for subtype characterization. The results are also compared with PAM50 labels. In (Wei, et al., 2022), a multi-kernel learning approach was proposed. It firstly optimizes the Gaussian kernel parameters per omics, then combines them into one fused kernel, which finally serves for k-means clustering to identify the cancer subtypes. The outcomes are evaluated with a classic comparison of survival and analysis of pathway activity. Significant features are moreover identified with the Kruskal-Wallis test. In (Sienkiewicz, et al., 2022), an approach called SUMO was proposed, which uses non-negative matrix factorization of patient-similarity networks and consensus clustering to detect molecular cancer subtypes. Subtypes are evaluated based on the log-rank comparison of survival.

The methods described above represent only a selection of recently published approaches, to the author's knowledge too recent to be addressed already in several methods reviews (Rappoport & Shamir, 2018; Duan, et al., 2021). In summary, multimodal clustering can provide a more comprehensive insight into tumor biology and increase the understanding of cancer behavior, as they do not rely on a single level in a gene expression process. However, due to the high complexity, incorporating multimodal approaches into daily clinical practice might be challenging.

3 Materials and methods

3.1 Data sets

The data sets used for this study were collected as a part of the TCGA Breast Invasive Carcinoma (BRCA) project. Only the primary tumor samples collected from the female patients were considered. The data files were acquired from the Genomic Data Commons (GDC) Data Portal (Genomic Data Commons Data Portal, 2022) or Legacy Archive (Genomic Data Commons Legacy Archive, 2021), depending on the file type. GDC Data Transfer Tool (GDC Data Transfer Tool, 2020) served for downloading files from the repository. The metadata, additional clinical and demographic information, and sample, patient, and file annotations were gathered with the GenomicDataCommons R package (Morgan & Davis, 2021).

3.1.1 Proteomic data

In the TCGA-BRCA project, the protein levels were measured with the Reverse Phase Protein Array (RPPA) platform. A single RPPA slide is stained with the specific antibody, which allows for measuring the levels of only one protein per array. Each RPPA slide is constructed as an array of 48 grids containing 11x11 spots formed in 2 columns of 5 spots for five 2-fold serial sample dilutions and one spot for the appropriate control lysate. Hence, one RPPA slide can measure a single protein level for up to 1056 samples. The sample preparation process for the RPPA methodology has been described in (Akbani, et al., 2014; The Cancer Genome Atlas Network, 2012).

TCGA Research Network provides three levels of RPPA data. Level 1 consists of the raw data with spot signal intensities. Level 2 contains the results of pre-processing with the SuperCurve non-parametric model built for each slide. This step involves adjusting raw spot intensities for the spatial bias correction, fitting the monotone-increasing B-spline model between log2-scaled protein concentration and signal intensities, and quality assessment (Hennessy, et al., 2007; Hu, et al., 2007; Tibes, et al., 2006; Coombes, 2012). Level 3 data is the final set of protein measurements following the correction for loadings and median-centering across antibodies (Hu, et al., 2007; Gonzalez-Angulo, et al., 2011;

Coombes, 2012). The level 3 data for 876 female patients were used in the analysis described further.

3.1.2 mRNA gene expression data

In the TCGA-BRCA project, the mRNA gene expression levels were obtained with the Agilent custom 244K whole genome microarrays. The sample preparation, hybridization, and processing procedures were described in (The Cancer Genome Atlas Network, 2011). Three levels of mRNA gene expression profiling data are available from TCGA Research Network: raw, probe-level, and gene-level (The Cancer Genome Atlas Network, 2012). The final gene-level data set consists of gene expression values after lowess normalization and log2-transformation of the Cy5 and Cy3 channels ratio representing the sample and the reference, respectively (The Cancer Genome Atlas Network, 2011). Moreover, the subtype labels obtained for the 50-gene PAM50 predictor (Parker, et al., 2009) were also provided in (The Cancer Genome Atlas Network, 2012). In total, the gathered set of gene expression levels represented 521 primary tumor female samples labeled with the PAM50 subtype.

3.1.3 Biospecimen and clinical data set

TCGA Research Network provides demographic information concerning the patients, including age at the initial diagnosis, declared race, and ethnicity. Each patient is also annotated with the tissue source site (TSS), which is the medical center of the patient's initial diagnosis and sample collection.

The clinical information provided per patient includes the vital status, time from the initial diagnosis to the last contact with a patient, and, in the case of a patient’s death, the time survived from the initial diagnosis. The follow-up records were also collected, although, unfortunately, follow-up intervals and collected details are not consistent for the whole cohort. However, if the patient was examined after the initial treatment, the current disease status was provided, the time between initial diagnosis and follow-up examination, and in the case of recurrence, the time until the new tumor was diagnosed. Moreover, the American Joint Committee on Cancer (AJCC) cancer staging fields of tumor T, regional nodes N, metastases M, and stage are available per patient. Those, however, must

be treated with caution as different AJJC Cancer Staging Manual editions were used throughout the timespan of the TCGA-BRCA project.

Following the inspection of the reported survival time since the diagnosis and follow-up data, several inconsistencies were found, mainly connected with death from recurring cancer. Those problems were also reported in (Huo, et al., 2017) and (Liu, et al., 2018). Hence, for the survival analysis, TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) dataset was used, which was curated, standardized, and described in (Liu, et al., 2018).

3.1.4 Immune cellular fraction estimates

The relative proportions of 22 immune cell types in the tissue served to characterize the tumor samples further. The immune cellular fractions for the TCGA-BRCA cohort were estimated in (Thorsson, et al., 2018) with the CIBERSORT method (Newman, et al., 2015) for cell composition identification based on RNA-Seq data.

3.2 Batch effect identification and correction

The batch effect remains a common problem in high-throughput data analysis. It is unavoidable that in the large-scale studies, the samples must be grouped into batches to be processed together. Not all samples can be processed identically in one laboratory on the same day. Hence, experimental conditions changing over time and technical factors associated with the location or sample grouping may cause additional bias in the data and the risk of hidden biological background.

To avoid technical bias in the data, TCGA Research Network has put extensive effort into properly designing all experiments and appropriately organizing the sample acquisition and further processing. The biospecimens were collected and preserved at various TSSs. However, for quality control, anonymization, and analyte isolation, the samples were transported overnight to Biospecimen Core Resources laboratories, where they were grouped into batches of a fixed set of patients. The batches of analytes were later shipped to Genome Characterization Centers on various dates for further experiments to generate the measurements. The analytes were processed on various plates annotated

with the unique Plate Identifier (ID) (MD Anderson Cancer Center, 2020; National Cancer Institute, 2020).

In 2012, TCGA Research Network published the TCGA-BRCA project summary, including the batch effect identification results on the set of mRNA expression, the part of which was used for this dissertation (The Cancer Genome Atlas Network, 2012). The batch effect verification was based on the visual assessment of plots generated using hierarchical clustering and PCA. The average linkage algorithm and one minus the Pearson correlation coefficient dissimilarity measure were used for the hierarchical clustering. The results were presented in the form of a dendrogram with samples colored with regard to the batch ID or TSS. For the PCA, the first four components were plotted with batch centroids marked. Neither of those visualization manners led to the detection of a batch effect in the data.

However, as only the subsets of the TCGA-BRCA cohort served for this dissertation, the additional batch effect detection was also conducted to ensure no technical bias.

3.2.1 Data dimensionality reduction methods

As in the TCGA Research Network approach, data visualization served to verify whether there is a batch effect corresponding to categorical technical factors like TSS, plate ID, or, in the case of RPPA data, the experiment design. Two combined dimensionality reduction methods were used to plot the data set in the two-dimensional (2D) space: PCA and Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, & Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2018).

PCA is a linear transformation method that, in the basic 2D example, aims to fit the line that minimizes the sum of distances to the data points. It involves centering the data set, scaling it if necessary, and multiplying it by the rotation matrix. This matrix is constructed from the eigenvectors of the data covariance matrix, sorted by decreasing eigenvalues. This operation provides the new data matrix with the set of principal component (PC) values per observation. The PCs

are sorted by the decreasing proportion of variance in the data explained by the particular PC. PCA is thus a simple and reproducible method, insensitive to any additional parameters, that allows reduction to any number of dimensions not larger than the original data. Moreover, PCA explains how the given PC correlates with each original variable (Hotelling, 1933).

UMAP is a non-linear data dimensionality reduction method based on modeling the manifold with a fuzzy topological structure. This procedure aims to provide the low dimensionality space with the topological structure reflecting the high dimensional topology as closely as possible. Unlike PCA, UMAP is a graph-based method, requires several parameters to be considered, and does not provide information on how the final extracted features correspond to the original variables. However, it was proven to generate high-quality embeddings of various large data sets. It is commonly used for high-throughput data analysis, especially for visualization purposes (McInnes, Healy, & Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2018).

For the data dimensionality reduction for both the protein and mRNA expression data sets, the UMAP algorithm was applied to the PCA-reduced data sets, consisting of top PCs explaining 90% of the variance in the data. The Euclidean distance served as the similarity measure, providing satisfactory results compared to the correlation-based distances or cosine similarity. In this manner, the 2D data set was obtained and visualized as the scatterplot with points corresponding to the primary tumor samples, one per patient. If the technical bias exists in the data, the samples included in the same technical factor category (for instance, collected at the same TSS or assigned to the same plate with identical plate ID) tend to group at the UMAP embedding. This should be easily visually assessed based on the scatterplot. PCA and UMAP do not deal with the missing data, so the features with incomplete records were removed.

3.2.2 Batch effect identification and correction methods

In the case of categorical technical factors, like TSS or plate ID, the structure of the potential batch effect is known. However, when the experimental conditions

changing over the study's timespan are the potential technical bias source, the batch categories are difficult to predict. Consequently, batch effect correction becomes challenging as many efficient tools for this task require prior batch ID assignment.

BatchI R package (Papiez, Marczyk, Polanska, & Polanski, 2018), dedicated to high-throughput data, provides the method to identify the subseries of experiments for the data sorted on the timescale. The algorithm uses the dynamic programming approach (Bellman, 1961; Jackson, et al., 2005) to partition the samples into estimated batches to maximize the inter-batch dispersion with possibly small intra-batch dispersion. As this approach requires the number of subseries into which the samples should be split, the BatchI tool also provides the method to select the optimal number of batches from a chosen range. The method relies on calculating δ guided PCA statistics (Reese, et al., 2013) and using a permutation test to verify whether the statistics is larger than expected to be obtained by chance. Hence, the p-value can be used to select the optimal number of batches and verify whether the batch effect exists in the data (Papiez, Marczyk, Polanska, & Polanski, 2018).

In the case of the RPPA data, the exact date of the experiment was unknown. However, for the mRNA expression, the scan date was extracted from level 1 raw files, and the samples were sorted accordingly. The BatchI algorithm was applied with the average intensity among all features as the quality score, as recommended in the package documentation for the microarray data (Papiez, Marczyk, Polanska, & Polanski, 2018). The optimal number of batches was chosen from the range of 2 to the number of unique scan dates.

If the batch effect was detected in the data and its structure was known, the measurements were corrected with the ComBat algorithm included in the “sva” R package (Leek, et al., 2017). The method is dedicated primarily to the microarray data and adjusts the data for batch effects using the parametric empirical Bayes frameworks. The algorithm provides the expression data corrected for the batch effect. However, it requires batch labels, so the batches must be identified before adjustment (Johnson, Li, & Rabinovic, 2006).

3.3 Identification of patient subpopulations

Various machine learning approaches were applied to the protein level measurements obtained using the RPPA platform to detect subpopulations of breast cancer patients and explore the data set composition. In total, levels of 281 proteins were measured and gathered from GDC Data Portal. However, the set of investigated proteins was inconsistent in the cohort, and many proteins were omitted for some samples. Hence, many missing values were detected in the data set. Consequently, for the sake of further described analysis methods, proteins with missing records were removed from the data set. As the aim was also to refer the results to molecular subtype established based on gene expression with PAM50 predictor (Parker, et al., 2009), the RPPA data set was limited to samples with available PAM50 subtype label, as explained in Chapter 3.1.2. Thus, the final data set used for this investigation consisted of measurements of 166 proteins for 407 patients.

The summary of cases included in the subpopulation identification step is presented in Table 3.1 regarding their PAM50 subtype etiquette. The cohort was highly imbalanced in terms of the PAM50 subtype. This, however, is the limitation observed for the TCGA-BRCA project as a whole, where HR+ statuses were reported for the majority of cases (The Cancer Genome Atlas Network, 2012; Tobiasz, Hatzis, & Polanska, Breast Cancer Heterogeneity Investigation: Multiple k-Means Clustering Approach, 2019). Hence, the imbalance does not result from reducing the data set to PAM50-labeled patients.

Table 3.1 Summary of cases considered for the subpopulation identification regarding their PAM50 subtype label

The table is taken from (Tobiasz & Polanska, 2022).

PAM50 subtype	No. patients	Percentage of patients [%]
Basal	86	21.13
HER2-enriched	50	12.28
Luminal A	173	42.51
Luminal B	98	24.08
TOTAL	**407**	**100.00**

Both the 281 proteins measured and 166 selected for this study due to missing records are a small fraction of the whole human protein universe. To initially investigate the function space covered by 166 proteins used for further steps of this work, the Reactome pathway Over-Representation Analysis (ORA) was performed on the set of genes annotated to those proteins. ORA is a first-generation enrichment analysis method based on the hypergeometric test. It aims to verify whether the set of measured proteins included more representatives of the particular pathway than it would be expected to occur by chance (Fabregat, et al., 2017; Fabregat, et al., 2015). Hence, ORA provided the list of Reactome pathways enriched among proteins used for this study. The results were corrected for False Discovery Rate (FDR) with the Benjamini-Hochberg method (Benjamini & Hochberg, 1995).

3.3.1 Clustering algorithms

Various combinations of clustering algorithms and feature engineering methods were tested for the subtyping. Representative methods of density-based, graph-based, and centroid-based approaches to data grouping were used: Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello, Moulavi, & Sander, 2013), Louvain community detection (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008), and custom Divisive intelligent K-means (DiviK) (Mrukwa & Polanska, 2022), respectively.

3.3.1.1 Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)

HDBSCAN is a hierarchical extension of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. The method involves the construction of the simplified tree of significant clusters from a clustering hierarchy standing for all possible DBSCAN solutions. The algorithm then uses the tree to select the optimal cuts and produce the clustering outcome based on the cluster stability (Campello, Moulavi, & Sander, 2013).

The main drawbacks of this approach, especially challenging for the relatively small RPPA data set used in this study, include the need to specify several parameters on which the clustering outcome strongly depends. Not only

the differences in the assignment of patients to the clusters were observed, but also the number of clusters detected varied greatly. For this study, it was assumed that each cluster should consist of at least 30 observations to avoid obtaining multiple little subgroups of patients. Also, the clusters close to each other were merged. Moreover, at least three observations were required in the neighborhood of each core point. The leaf cluster selection method was chosen to obtain the homogenous clusters, which means that clusters were selected from the leaves of the condensed tree. The Python HDBSCAN implementation was used for the calculations (McInnes & Healy, Accelerated Hierarchical Density Based Clustering, 2017; McInnes, Healy, & Astels, hdbscan: Hierarchical density based clustering, 2017).

Another challenge of using the HDSCAN algorithm is that it leaves so-called "noisy points" unassigned to any resulting cluster. Hence, in this study, some patients were not included in any of the detected subpopulations, which imposed the postprocessing of the results to predict the cluster assignment for them.

3.3.1.2 Louvain community detection

The graph-based Louvain community detection algorithm is based on the two-phase modularity optimization and community aggregation process. In the first phase, randomly ordered nodes are replaced sequentially between the communities until no further improvement in the modularity can be achieved. In the second phase, the updated network is constructed, in which communities resulting from the previous step serve as new nodes. This procedure is iteratively repeated until the obtained network is stable with the maximized modularity. High modularity indicates dense intra-community connections and sparse inter-community ones (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008).

Similarly to HDBSCAN, the Louvain algorithm requires several parameters. In this study, the Jaccard similarity index between the neighbors of two nodes served as the weight between them, and the five nearest neighbors were considered for graph construction to ensure the desired subpopulation sizes and hence to avoid obtaining many too-small clusters or too few large groups, providing too general information about the data set structure. The "bluster" R package was used for calculations (Lun, 2021).

3.3.1.3 Divisive intelligent K-means

DiviK algorithm consists of stepwise k-means clustering in a locally optimized feature domain. The method consists of three phases repeated iteratively til improvement in sample grouping is observed (Mrukwa & Polanska, 2022).

The first phase of each iteration is reducing the number of features, proteins in this case, that should be considered for grouping. The feature engineering procedure is performed independently at each step. In (Mrukwa & Polanska, 2022), the optimization of feature space involves Gaussian Mixture Models (GMM) decomposition of the distributions of feature averages and variances. The crossing points of components serve as thresholds for removing variables with the lowest average considered noise and keeping variables with the highest variance. In this dissertation, as the original number of proteins used for clustering was relatively small and no noise was observed in the data, the average-based filtration step was omitted, and variance-based filtration performed satisfactorily. Thus, log2-scaled variances of protein levels were GMM decomposed for the number of components from 1 to 5 chosen based on the Bayesian Information Criterion (BIC) (Schwarz, 1978), as described in (Marczyk, Jaksik, Polanski, & Polanska, 2019). To avoid obtaining very low and wide Gaussian components due to the outliers, only the values inside the interval $(\bar{x} \pm s)$ were considered, where $\bar{x}$ is the mean log2-scaled variance and s is its standard deviation. For the threshold determination, only components with the standard deviation higher than 0.01 were taken into account to reduce the impact of extreme and thin peaks observed in some cases.

The second phase of the DiviK algorithm is the evaluation of the cluster diversity, which provides information on whether the data should be further split and the optimal number of clusters. For this procedure, the GAP statistics (Tibshirani, Walther, & Hastie, 2001) is used, which is applicable for comparing a null model including only one cluster versus a multi-cluster partition. The calculations were carried out using MATLAB implementation. GAP statistics refers a total within-cluster dispersion to the expected one estimated through clustering 100 reference sets generated from the uniform distribution. Then, the optimal number of clusters is chosen as the smallest number of clusters

satisfying Equation 3.1 (Mrukwa & Polanska, 2022; Tibshirani, Walther, & Hastie, 2001; Tobiasz, Hatzis, & Polanska, 2019).

$$\boldsymbol{Gap(k) \geq Gap(k+1) - s_{k+1}} \quad \textbf{3.1}$$

Where:

$Gap(k)$ denotes the GAP statistics for k clusters,

s_{k+1} denotes the standard error for clustering into $k+1$ clusters.

The third phase of the DiviK algorithm is the centroid-based k-means clustering into the number of clusters selected with the GAP statistics. This study used the k-means MATLAB implementation with the squared Euclidean distance measure and k-means++ algorithm to determine the centroid seeds (Arthur & Vassilvitskii, 2006).

All three phases of the DiviK approach described above are performed independently for each cluster resulting from the previous iteration. The patient subgroup is not further partitioned when the GAP-statistics-based stop criterion is fulfilled or the subgroup consists of 10 patients or fewer. The advantage of this approach is that the feature selection method is built-in every iteration and reflects the variability of the considered subgroup only.

3.3.2 Feature engineering

The used clustering methods deal with the high dimensionality of data to a different extent, so data dimensionality reduction was required in some cases, and the clustering was applied either to the levels of all proteins or the reduced feature space. Depending on the grouping algorithm, various combinations of feature selection or extraction procedures were applied to prepare the data set for the clustering.

For the feature selection, the GMM decomposition approach was used. The variances of each protein levels were calculated and transformed to the logarithmic scale. Then, the distribution of resulting values was decomposed as described in (Marczyk, Jaksik, Polanski, & Polanska, 2019). The optimal number of Gaussian components was selected from 2 to 10 using BIC (Schwarz, 1978).

The intersection point of the two components corresponding to the highest variances determined the threshold value for filtration: only the proteins with a higher variance of levels were considered in the clustering procedure.

The feature extraction methods applied to prepare the data for clustering included the PCA to select the top PC explaining 90% of the variance in the data and UMAP performed on the PCA-reduced set, as described in Chapter 3.2.1.

3.3.3 Methods combinations

Various combinations of clustering algorithms and data dimensionality reduction methods were applied to the protein level measurements. Table 3.2 presents the summary and abbreviations of the variants, which will be later used for referring to results.

Table 3.2 Combinations of clustering algorithms and data dimensionality reduction methods

Abbreviations for each combination are written in italics. DiviK is marked with (*) to indicate that the GMM-based filtration is built in each algorithm iteration.

The table is taken from (Tobiasz & Polanska, 2022).

	Feature engineering					
	No reduction		**PCA**		**UMAP**	
Clustering	**Complete**	**GMM filtered**	**Complete**	**GMM filtered**	**Complete**	**GMM filtered**
HDBSCAN	×	×	×	×	$H_{UMAP\text{-}C}$ ✓	$H_{UMAP\text{-}F}$ ✓
Louvain	L_C ✓	L_F ✓	$L_{PCA\text{-}C}$ ✓	$L_{PCA\text{-}F}$ ✓	×	×
DiviK*	×	✓	×	×	×	×

As mentioned in Chapter 3.3.1.1, following the HDBSCAN algorithm, some patients may be left unassigned to any resulting cluster. However, for further analysis, a new subtype label is required for each patient. Hence, merging the left cases with the groups as similar as possible was necessary. The following variants of the cluster assignment prediction were tested, all based on the Euclidean distance between the data point and the cluster centroid:

1. $H_{UMAP\text{-}C1}$: Proximity in 2-dimensional UMAP;
2. $H_{UMAP\text{-}C2}$: Proximity in the dataset with all protein levels (complete);

3. $H_{UMAP\text{-}C3}$: Proximity in the set of top PCs explaining 90% of the variance.

Finally, the set of cluster assignments was obtained per patient for each of the nine combinations of data dimension reduction and clustering. The resulting clusters are considered patient subpopulations and will be described as such or as breast cancer subtypes. They were referred to the PAM50 subtype labels and named based on them. For instance, the cluster named "Basal" contains mainly samples labeled as "Basals" by the PAM50 predictor.

The UMAP embedding created, as explained in Chapter 3.2.1, served to visualize the data in the 2D space as a scatterplot. Data points were colored by the subtype predicted in this work or by the PAM50 predictor.

3.4 Metrics for outcome comparison

As described in Chapter 3.3, various combinations of clustering algorithms and feature engineering methods were tested to identify patient subpopulations. Moreover, in the case of the HDBSCAN algorithm, there was a need to predict the cluster assignments for patients left out as noise, and three different approaches to that task were applied. Results of new subtypes identification based on the protein levels differed between the variants of approaches tested in terms of both the patient's assignment to clusters and the final number of subpopulations detected. Moreover, the outcomes strongly depended on the parameters determined for the clustering algorithms. Hence, there was a need to define a reliable method for clustering outcomes comparison that would serve to select the appropriate machine learning approach for subpopulation identification.

Comparing various clustering approaches was challenging, as the proposed method should deal with the number of issues related to the problem of breast cancer subtyping. Firstly, even though it was possible to estimate the expected range of a possible number of clusters based on the literature review (The Cancer Genome Atlas Network, 2012), visualizations in the UMAP embedding, or other similar studies (Tobiasz, Hatzis, & Polanska, Breast Cancer Heterogeneity Investigation: Multiple k-Means Clustering Approach, 2019), the correct exact number remained unknown. Moreover, the resulting subpopulation sizes strongly

varied, which was also expected as the cohort was highly imbalanced. The number of features used for clustering in most approaches was large but not constant since different variants of data dimension reduction techniques were applied. Finally, various dissimilarity degrees between clusters were observed and expected due to the biological background and breast cancer's heterogeneous character. For instance, the basal subtype was assumed to be far more isolated from other tumors, while the luminal cases would instead group together and possibly tend to further split into less numerous subgroups.

Two approaches were therefore tested for the task of clustering outcome evaluation. Both rely on the effect size measures, which should address the problem of various cluster sizes. As a result, a new metrics was proposed that should satisfactorily handle the challenges mentioned above. It served for the selection of the most reliable clustering approach and subpopulation detection outcome and, consequently, for the definition of breast cancer subtypes investigated in this work.

3.4.1 η^2 effect size

Firstly, the levels of each protein were compared between the clusters with the η^2 effect size measure, given by Equation 3.2 (Cohen, 2013).

$$\boldsymbol{\eta^2 = \frac{SS_{among}}{SS_{total}}} \qquad \textbf{3.2}$$

Where:

SS_{among} denotes the inter-cluster sum of squares,

SS_{total} denotes the total sum of squares defined as the sum of inter- and intra-cluster sums of squares.

Equation 3.2 indicates that the higher η^2 value, the higher the variance between the groups compared to the variances within the groups. Hence, the higher η^2 value, the better the cluster separation. However, all clusters are considered jointly, which is the limitation of η^2 metrics. Therefore, high η^2 value do not provide detailed information on whether all clusters are well-separated or just some are highly isolated. Consequently, η^2 may be a less reliable measure of clustering

outcome when some distinctly outlying groups are expected, just like it might occur for breast cancer subpopulations (Schwarz, 1978; Tobiasz & Polanska, 2022).

Nevertheless, η^2 was calculated for each protein to evaluate the clustering outcomes. Hence, 166 η^2 values were obtained per clustering approach. To integrate those scores per method, mean, median 1st quartile (Q_1), and 3rd quartile (Q_3) of protein η^2 values were computed.

3.4.2 Pooled *d* metrics

As η^2 measure considers all clusters together, another metrics was proposed by modifying Cohen's *d* effect size (Cohen, 2013). The concept relied on referring each obtained cluster one by one to all remaining clusters considered jointly. This effect has been achieved by comparing the average protein levels between patients assigned and unassigned to a given subpopulation, as in Equation 3.3 (Tobiasz & Polanska, 2022).

$$\boldsymbol{d = \frac{\bar{x}_{subtype} - \bar{x}_{remaining}}{\sqrt{MS_{within}}}} \qquad 3.3$$

Where:

$\bar{x}_{subtype}$ denotes the mean protein level for patients assigned to the particular breast cancer subtype,

$\bar{x}_{remaining}$ denotes the mean protein level for all remaining patients meaning those assigned to other subtypes,

MS_{within} denotes mean intra-subtype sums of squares.

According to Equation 3.3, *d* values are positive when the protein level is increased in the given subtype compared to others and negative otherwise. The higher the absolute value of *d*, the bigger the difference between the investigated subpopulation and the remaining patients.

Therefore, 166 *d* values were obtained per cluster for each evaluated clustering approach. The number of *d* values for each protein equaled the number of subpopulations detected with the given approach. To easily compare the clustering approaches, one score should represent each. Therefore, several lists

of *d* scores per method were integrated to obtain one pooled *d* score. Hence, each cluster was annotated with the Q_3 of protein *d* absolute values.

Consequently, several vectors of 166 *d* values were reduced to just one vector with Q_3 per cluster. Those Q_3 values were projected as a point in the *k*-dimensional space, where *k* was the number of subtypes detected. Finally, the pooled *d* score was calculated as the distance between the created point and the beginning of the coordinate system. The procedure for obtaining pooled *d* values per clustering approach is presented in Figure 3.1 (Tobiasz & Polanska, 2022).

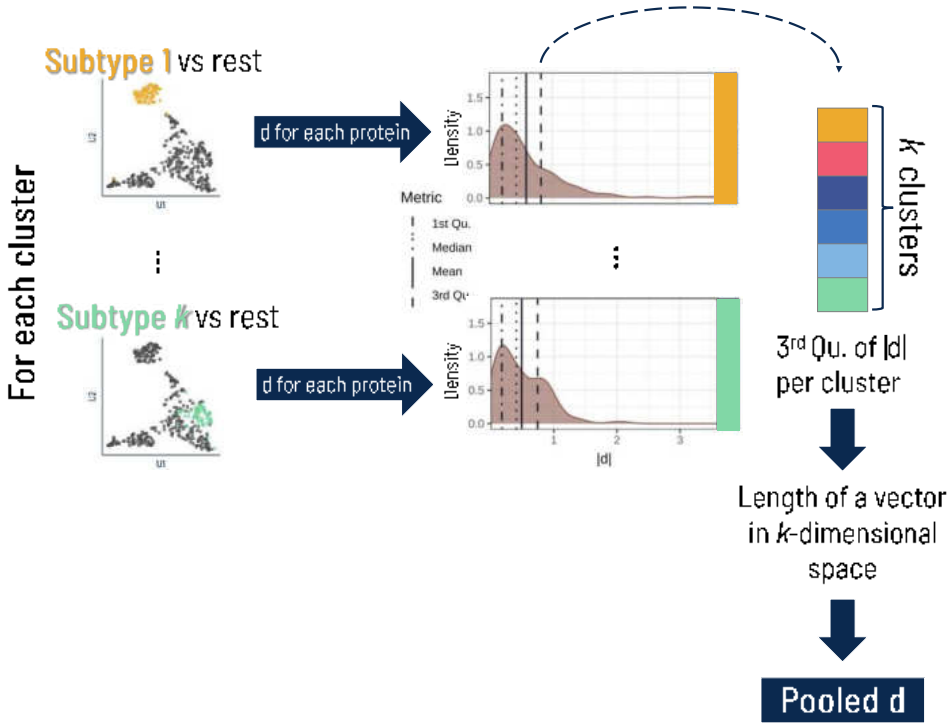


Figure 3.1 Procedure of pooled *d* calculation

3.4.3 Metrics evaluation

Dice coefficient (Dice, 1945) was calculated to assess the similarity between the subtypes detected with each clustering approach and those given by the PAM50 predictor. This coefficient measures the agreement between two categorical variables. However, it also assumes that they both consist of the same categories. In this work, however, a single PAM50 subtype may correspond to more than one subtype detected with the clustering approaches described above. A patient assignment to any cluster corresponding to their PAM50 subtype was considered

a match in that situation. The resulting Dice coefficient values were referred to the pooled *d* scores and Q_3 of η^2 effect size.

To investigate the differences in results obtained with various clustering approaches and to verify how the proposed pooled *d* metrics reflects them, the corresponding clusters from the best and worst method according to pooled *d* score were compared. Their per cluster *d* values (before the integration) were plotted against each other.

Moreover, the clustering approaches with the lowest and the highest pooled *d* values were compared regarding the number of characteristic proteins and their biological functions. The proteins with significantly increased or decreased levels in the given subpopulation were identified based on the absolute *d* values. The thresholds for at least large ($|d| \geqslant 0.8$) or very large ($|d| \geqslant 1.2$) Cohen's *d* effect served for this selection (Cohen, 2013). The resulting lists of proteins were matched to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa, Furumichi, Tanabe, Sato, & Morishima, 2016) to obtain the number of pathways in which the proteins are involved.

3.5 Clinical characteristics of patient subpopulations

The identified subpopulations of breast cancer patients were evaluated by investigating individuals' clinical and demographic profiles in different subtypes. This part of the analysis mainly aimed to verify whether the survival and clinical experiences or the demographic background carry any differentiating significance and support the protein-based detection of subpopulations.

3.5.1 Survival analysis

3.5.1.1 Survival outcome endpoints

Four clinical survival outcome endpoints were investigated in this work, as defined in the study by (Liu, et al., 2018), which not only introduced the standardized TCGA-CDR data set but also provided recommendations for the survival analysis of each TCGA cancer type, including breast tumors. The summary of the endpoint types is shown in Table 3.3.

Table 3.3 Summary of the survival outcome endpoint types

The table is adapted from (Liu, et al., 2018). (*) indicates that a particular endpoint type is not recommended for The Cancer Genome Atlas Breast Invasive Carcinoma project due to a short-term follow-up interval.

Endpoint type	Event of interest	Censored observation	Time to event	Time to censoring
Overall Survival (OS)*	Death from any cause	Alive	Date of death	Date of last contact
Disease-Specific Survival (DSS)*	Death from diagnosed cancer type	Alive, dead without tumor	Date of death with the diagnosed cancer	Date of last contact or death without tumor
Disease-Free Interval (DFI)	New tumor progression event following the disease-free period after the initial treatment (*locoregional recurrence, distant metastasis, new primary tumor in the same organ, death from advancing the same tumor*)	Dead, tumor-free alive, alive with new primary tumor in a different organ	Date of the first occurrence of new tumor progression event after the disease-free period	Date of last contact or death
Progression-Free Interval (PFI)	New tumor event (*disease progression, locoregional recurrence, distant metastasis, new primary tumor, death with tumor*)	Dead without tumor, alive without new tumor events	Date of the first occurrence of new tumor event	Date of last contact or death without tumor

The first endpoint, Overall Survival (OS), was defined as the period from the initial cancer diagnosis until death from any cause. The censored time was determined by the date of the last contact with a patient. OS is the most used survival outcome as it is unequivocal and straightforward to gather from GDC Data Portal. However, it is also biased as it does not distinguish cancer and non-cancer deaths, consequently not reflecting the tumor aggressiveness and responsiveness to therapy. Hence, OS may weaken the clinical study, especially in the relatively old patient cohort (Liu, et al., 2018).

Disease-Specific Survival (DSS) was the second endpoint considered. It was determined as the time between the initial diagnosis and the death from the specific cancer type, which in the case of this study was breast cancer. The censored time was thus a period from initial diagnosis until the last contact with the patient or until the patient's death from a cause different than cancer. Hence, DSS reflects tumor biology better than OS. On the other hand, DSS may be biased for TCGA as clinical data include only the tumor status during death, which might not always be identical to the cause of death. Thus, a patient with a tumor who died for another reason, even utterly independent of cancer, cannot be distinguished from a patient who died due to cancer (Liu, et al., 2018).

Disease-Free Interval (DFI) was another survival outcome endpoint used. It is the period between the diagnoses of the initial tumor and the new tumor event (NTE) if the patient had been considered disease-free after the first diagnosis and treatment. NTE was defined as one of the following: locoregional recurrence, distant metastasis, new primary tumor in the same organ, or death due to the same tumor. Thus, the censored time was a period between the initial diagnosis and the last contact or death. Patients with a new primary tumor in another organ, tumor-free, or dead, were censored. This endpoint is the most ambiguous for the TCGA cohort for several reasons. Firstly, it can be equivocal whether the patient had ever been determined disease-free following the initial diagnosis and treatment. Hence, sometimes it is unclear if the follow-up record reported the new cancer occurrence or the first tumor that had not been successfully treated yet. Moreover, technically, the time zero for DFI should be regarded as the date when a patient was determined to be disease-free following the first treatment. However, this information was not provided. Nevertheless, DFI reflects the tumor biology better than OS and is more informative than DSS when the follow-up period is not long enough to capture many deaths in the cohort, which is the case for less aggressive cancer types with a relatively good prognosis (Liu, et al., 2018).

The last used endpoint was Progression-Free Interval (PFI), the period from the initial diagnosis until the first occurrence of NTE, defined as the disease progression, locoregional recurrence, distant metastasis, new primary tumor,

or cancer death. Hence, the censored time is the time to either the last contact with a patient or death without a tumor. This endpoint is not biased with non-cancer deaths like OS and requires a shorter follow-up time than death-dependent endpoints. Moreover, contrary to DFI, PFI is easier to derive from the TCGA cohort, as it does not demand the information if the patient had ever achieved a tumor-free status. Hence, it is unnecessary to distinguish between ongoing initial tumors and new ones (Liu, et al., 2018).

The study (Liu, et al., 2018) mentions several TCGA-BRCA cohort survival analysis limitations. The first problem is a relatively short-term follow-up period, often insufficient to observe the event of interest. This is especially the case of less aggressive tumor types like breast cancer, for which it is unlikely to capture enough events during the study interval to produce reliable and statistically significant outcome determinations. In (Liu, et al., 2018), the assumption was made that a sufficient follow-up interval is indicated by the median censored time longer than the median event time. This condition was not fulfilled in the breast cancer cohort for OS and DSS. For PFI, both median times for PFI were very close, so the authors assumed the follow-up to be long enough to support a reliable analysis. However, in this dissertation, the median event and censoring times were also compared to assess the quality of results, as only a subset of the TCGA-BRCA cohort served for the survival analysis. The minimum follow-up period depends on the endpoint used and is shorter for PFI and DFI than for OS and DSS. OS and DSS require a patient's death, which occurs after tumor progression or recurrence and happens less frequently. Hence, in (Liu, et al., 2018) for breast cancer, it is recommended to use PFI and DFI while treating OS or DSS with caution due to the need for a longer follow-up.

This work considers not only DFI and PFI but also OS and DSS to provide a more comprehensive insight into the potential differences between the identified subpopulations. However, the results are discussed carefully and with an awareness of the problem.

3.5.1.2 Survival analysis methods

The survival function's Kaplan-Meier (KM) estimator (Kaplan & Meier, 1992) was used to plot the survival curves for the breast cancer patients' subpopulations. The advantage of using KM instead of the regular survival function defined as the fraction of patients alive at a specific time point is that even though it considers the times for all available cases, it distinguishes the uncensored and censored observations. Hence, it reduces the bias resulting from patients leaving the study for reasons other than investigated, like missing follow-up, and it also allows survival investigation even if the event of interest (like death) has not occurred yet for the whole cohort (Kaplan & Meier, 1992; May, Hosmer, & Lemeshow, 2014).

The comparison of survival experiences for different subtypes was visually examined based on the KM graphs. The appropriate statistical testing was also performed to quantify the differences between the groups and verify if they were statistically significant. As the survival data are usually right-skewed, the classic rank-based non-parametric approaches for statistical comparison might have been used, provided the complete survival times for the whole cohort (May, Hosmer, & Lemeshow, 2014). However, there were many censored observations in the patient group used for this study, so the tests dedicated to survival records were applied for the comparative analysis.

Each test is based on the contingency table of a group by vital status generated for each observed survival time. Generally, the test statistics to compare survival outcomes is defined as in Equation 3.4 (May, Hosmer, & Lemeshow, 2014).

$$\boldsymbol{\chi^2} = \frac{\left[\sum_{i=1}^{m}\left(w_i(d_{1i}-\hat{e}_{1i})\right)\right]^2}{\sum_{i=1}^{m}\left(w_i^2\hat{v}_{1i}\right)} \sim \boldsymbol{\chi^2_{k-1}} \qquad 3.4$$

Where:

m denotes the number of timepoints observed,

k denotes the number of groups compared,

d_{1i} denotes the number of events (e.g., deaths) in the first group for the i-th timepoint,

$\hat{e}_{1i}$ denotes the expected number of events in the first group for the i-th timepoint, defined as the product of the group size and total number of events in the cohort, and divided by the cohort size,

$\hat{v}_{1i}$ denotes the estimator of d_{1i} variance from the hypergeometric distribution,

w_i denotes the weight for the i-th timepoint, depending on the test.

The log-rank test was calculated for each comparison. It is the most common approach, in which all weights w_i are equal to one, which means that the same importance in put on differences between the survival functions throughout the whole timespan of the study (Mantel, 1966; Peto & Peto, 1972; May, Hosmer, & Lemeshow, 2014). However, it was observed that in the case of some comparisons, the differences in survival outcomes are mainly visible in the initial phases of the illness and therapy, while with time, the survival curves become more similar to each other, and the impact of censored variables increases as many patients has not experienced the event of interest yet. Thus, the generalized Wilcoxon rank sum test, also called the Gehan-Wilcoxon test, was applied to compare the subpopulations. In this approach, the weights are defined as the number of patients still at risk for each survival timepoint, as in Equation 3.5 (Gehan, 1965; Breslow, 1970; May, Hosmer, & Lemeshow, 2014).

$$\boldsymbol{w_i = n_i} \tag{3.5}$$

Where:

w_i denotes the weight for the i-th timepoint for the Gehan-Wilcoxon test,

n_i denotes the number of patients at risk for the i-th timepoint.

Therefore, the Gehan-Wilcoxon test emphasizes the differences between survival functions for the smaller time values. Hence, the Gehan-Wilcoxon test is more likely to detect early differences in survival experience than the log-rank test (May, Hosmer, & Lemeshow, 2014).

Moreover, the Cox proportional hazard model was fitted to estimate the hazard ratio (HR) corresponding to each subtype compared to the one defined as the reference (Cox, 1972). As explained in (Olivier, May, & Bell, 2017), the hazard

ratio can be regarded as the effect size measure, interpreted analogously to the relative risk (RR). The thresholds for RR or HR interpretation were adjusted for the imbalance between the sizes of the compared groups, represented by so-called allocation probability. Equation 3.6 served for this correction for positive HR (Olivier, May, & Bell, 2017). For negative HR, the threshold reciprocal was used. Consequently, HRs provided by the same Cox proportional hazard model were interpreted separately for each subpopulation based on a threshold depending on the balance between the sizes of the given subpopulation and the reference one.

$$\boldsymbol{HR_{\alpha} = 1 + \frac{\alpha}{(1-\alpha)\pi}} \qquad 3.6$$

Where:

π denotes the allocation probability as defined in Equation 3.7

α denotes the threshold for the correlation coefficient, equal to 0.1, 0.3, and 0.5 for small, medium, and large effect sizes, respectively (Cohen, 2013).

$$\boldsymbol{\pi = \frac{N_{group}}{N_{group}+N_{reference}}} \qquad 3.7$$

Where:

N_{group} denotes the size of a group for which HR is calculated,

$N_{reference}$ denotes the size of the reference group.

The KM estimator, log-rank test, and Cox proportional hazard model were calculated using the R package "survival" (Therneau, 2021). The survival curves were generated with the R package "survminer" (Kassambara, Kosinski, & Biecek, 2021). The Gehan-Wilcoxon test was performed using the R package "PHInfiniteEstimates" (Kolassa & Zhang, 2023). The analyses were performed for patient subpopulations detected as described in Chapters 3.3 and 3.4, and for subtypes obtained with the PAM50 classifier.

Confounding factors may be present in the cohort, and there is a risk of bias resulting from features like cancer stage, race, or age. Unfortunately, the small number of events captured for the subset of patients serving for this work did not allow for adjusting for confounding factors during the survival analysis. However,

as described in the next chapter, the obtained subpopulations were also compared regarding demographic and clinical background to estimate the impact those factors may have on the subtyping.

3.5.2 Statistical analysis of demographic and clinical profiles

3.5.2.1 Categorical variable analysis

Several categorical variables related to demographic and clinical factors were considered to verify their association with subpopulations identified on RPPA data. Moreover, the relationship with transcriptomic-based PAM50 subtypes was also evaluated to compare the outcomes between those two subtyping approaches.

Two categorical demographic factors were considered: race and ethnicity. The examined clinical categories were connected with AJCC Cancer Staging, involving the following pathologic stage fields: Tumor (T) describing the size of the tumor and its spread to the neighboring tissues, Nodes (N) denoting cancer spread to nearby lymph nodes, Metastasis (M) defining whether cancer passed on to other parts of the body, and the stage itself (National Cancer Institute, 2022). Several issues were reported in (The Cancer Genome Atlas Network, 2012) concerning the usage of AJCC data for the TCGA-BRCA cohort. The main problem was that various AJCC Cancer Staging Manual editions were used throughout sample collection and patient diagnosis. As a result, the TCGA-BRCA cohort was staged based on the mix of standards, mainly containing the 6th edition released in 2002 or the 7th edition released in 2010. For some patients, the edition was not reported and was impossible to determine based on the results. Hence, in (The Cancer Genome Atlas Network, 2012), the authors attempted to convert all older versions to the 7th edition of the staging manual, which was unsuccessful in some cases. Thus, in this work, the converted stage records provided as Supplementary Materials in (The Cancer Genome Atlas Network, 2012) were used when available. The original cancer stage provided in TCGA-BRCA was accepted for the unconverted cases.

Furthermore, to limit the number of categories for the association analysis, the stages were joined into the following classes: Stage I, Stage II, Stage III,

and Stage IV. More detailed staging information (for instance, the division into stages IIIA, IIB, and IIIC) was ignored due to insufficient sample size per subtype. Moreover, as recommended in (The Cancer Genome Atlas Network, 2012), pathologic fields connected with tumor size (T) and spread to lymph nodes (N) were binarized. T was coded as T1 and "other" to split tumors smaller than 2 cm and larger, respectively. N was coded as negative when no spread was observed and positive otherwise. The third pathologic field, M, was originally binary, corresponding to metastasis or lack of it.

Pearson χ^2 test of independence was conducted to check for the association between each of the demographic or clinical categorical factors and analyzed subtypes. For the 2-by-2 contingency table case, when two groups were tested for association with two categories, Yates's correction for continuity was applied (Yates, 1934).

Notably, contingency tables generated for different tested combinations of subtypes and categorical variables differed in dimensions. This impeded the comparison of subtyping outcomes provided by PAM50 and the method proposed in this dissertation. Pearson χ^2 test p-value, therefore, fails to provide a good characterization of dependency between the subtypes and demographic or clinical factors. Consequently, Cramér's V effect size was calculated to assess the strength of the association, as in Equation 3.8 (Cohen, 2013; Cramér, 1999).

$$\boldsymbol{V = \sqrt{\frac{\chi^2}{n \cdot min(c-1, r-1)}}} \qquad \textbf{3.8}$$

Where:

$\boldsymbol{\chi^2}$ denotes the test statistics from Pearson χ^2 test of independence,

n denotes the total number of observations,

c denotes the number of columns in the contingency table (e.g., the number of groups),

r denotes the number of rows in the contingency table (e.g., the number of categories).

The interpretation of Cramér's V depends on the smaller number of dimensions of the contingency table. The thresholds are adapted from Cohen's *w* cut-off values, as in Equation 3.9 (Cohen, 2013).

$$V_{thr} = \frac{w}{\sqrt{min\,(c-1, r-1)}} \quad 3.9$$

Where:

w denotes the threshold for Cohen's *w* effect size, equal to 0.1, 0.3, and 0.5 for small, medium, and large effect sizes, respectively (Cohen, 2013),

c denotes the number of columns in the contingency table (e.g., the number of groups),

r denotes the number of rows in the contingency table (e.g., the number of categories).

3.5.2.2 Numerical variable analysis

Numerical variables used for the subtyping results evaluation included patient age at diagnosis and CIBERSORT immune cellular fraction estimates. Numerical variables were compared between the subpopulations with appropriate tests selected according to the normality and variance homogeneity assumptions.

Shapiro-Wilk test (Shapiro & Wilk, 1965) per subtype was conducted for normality verification, followed by the Benjamini-Hochberg correction for multiple testing (Benjamini & Hochberg, 1995) applied for each group separately, when multiple variables were simultaneously tested, like for CIBERSORT data. In the case of confirmed normality, the Bartlett test of homogeneity of variances (Bartlett, 1937) was performed. One-way Analysis of Variance (ANOVA) dedicated to comparing more than two groups was used, supposed population normality and variance homogeneity assumption fulfillment. If the normality assumption was not satisfied, the non-parametric Kruskal-Wallis test replaced ANOVA. ANOVA or Kruskal-Wallis tests were also followed by the Benjamini-Hochberg multiple testing correction for CIBERSORT data (Benjamini & Hochberg, 1995). For variables varying significantly between the subtypes according to ANOVA or Kruskal-Wallis, respectively Tukey-Kramer

or Conover post hoc (Conover & Iman, 1979) tests were applied to assess the differences between every combination of two breast cancer subtypes. It is worth mentioning that the Tukey-Kramer test is based on the studentized range distribution, and thus it provides protection for multiple pairwise comparisons. Conover test, however, uses t-Student distribution and hence requires additional correction for multiple comparisons, which in the case of this work was obtained with the Bonferroni method (Conover & Iman, 1979; Haynes, 2013).

All tests for group comparisons were extended with calculations of the effect size measures. For the one-way ANOVA test, Equation 3.2 (Cohen, 2013) gave the η^2 effect size measure in Chapter 3.4.1. For the Kruskal-Wallis test, the non-parametric modification was used as in Equation 3.10 (Tomczak & Tomczak, 2014). This η^2 effect size modification represents the regular ANOVA η^2 effect size measure for ANOVA performed on the ranks instead of the original values.

$$\boldsymbol{\eta^2 = \frac{H-k-1}{n-k}} \qquad \textbf{3.10}$$

Where:

H denotes Kruskal-Wallis test statistics,

n denotes the total number of observations,

k denotes the number of groups.

For parametric post hoc comparisons, the effect size interpreted analogously to Cohen's *d* is based on the group means and ANOVA mean square value, according to Equation 3.11.

$$\boldsymbol{d_{ij} = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MS_{within}}}} \qquad \textbf{3.11}$$

Where:

$\bar{x}_i$ and $\bar{x}_j$ denote mean value in *i*-th and *j*-th groups, respectively,

MS_{within} denotes the ANOVA mean square reflecting the mean intra-group sums of squares.

Similarly, the non-parametric equivalent is defined based on the ANOVA test performed on ranks instead of the original values, as is given in Equation 3.12.

$$\boldsymbol{d = \frac{\overline{R_i} - \overline{R_j}}{\sqrt{MS_{within}}} = \frac{\overline{R_i} - \overline{R_j}}{\sqrt{s^2 \frac{n-1-H}{n-k}}}} \quad \textbf{3.12}$$

Where:

$\overline{R_i}$ and $\overline{R_j}$ denote mean ranks in *i*-th and *j*-th groups, respectively,

MS_{within} denotes the rank-based ANOVA mean square,

n denotes the total number of observations (in all considered groups from the Kruskal-Wallis test),

k denotes the total number of groups,

H denotes the Kruskal-Wallis test statistics,

s^2 is given by Equations 3.13 and 3.14.

$$\boldsymbol{s^2 = \frac{1}{n-1}\left[R - \frac{n(n+1)^2}{4}\right]} \quad \textbf{3.13}$$

$$\boldsymbol{R = \sum_{j=1}^{k} \sum_{i=1}^{n_j} r_{ij}^2} \quad \textbf{3.14}$$

Where:

r_{ij} denotes the rank of the *i*-th element in the *j*-th group.

Table 3.4 contains thresholds for η^2 and Cohen's d effect size interpretation.

Table 3.4 Thresholds for η^2 and Cohen's d effect size interpretation

The table is adapted from (Cohen, 2013; Sawilowsky, 2009).

Effect size interpretation	η^2	d
Very small	-	0.01
Small	0.01	0.2
Medium	0.06	0.5
Large	0.14	0.8
Very large	-	1.2
Huge	-	2

Figure 3.2 summarizes the differentiation testing pipeline for comparing more than two subtypes.

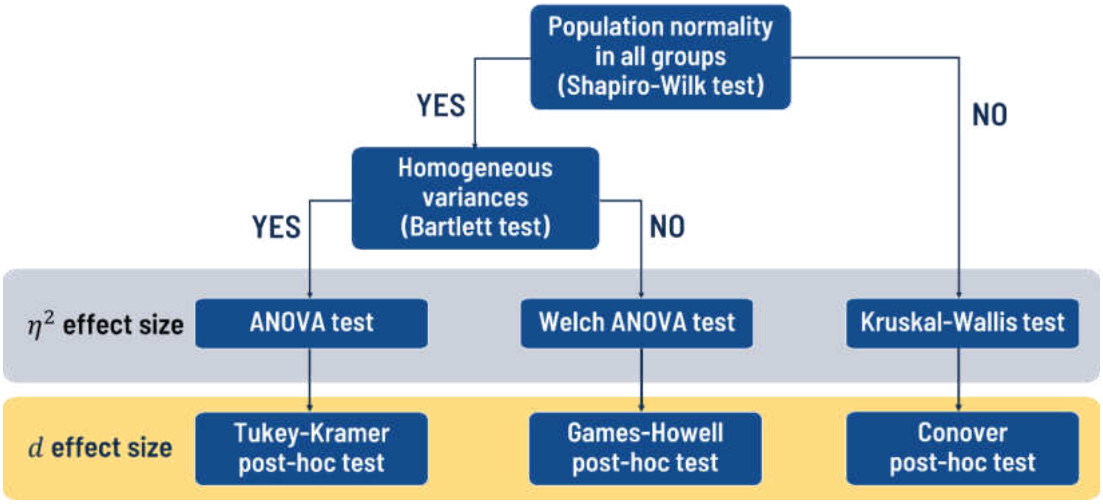


Figure 3.2 Differentiation testing pipeline for comparison of more than two groups

Provided population normality with heterogenous variances, the Welch ANOVA and Games-Howell post hoc tests would have been conducted, which, however, did not happen for any variable investigated in this work. Comparison of two groups only occurred in the rare cases of testing the subset of luminal PAM50 subtypes as the reference to subtyping proposed in this dissertation. For this situation, the differentiation testing pipeline included two-sample t-tests, modified Welch's test, or U-Mann-Whitney test, depending on the population normality and homogeneity of variances assumptions.

3.6 Molecular signature of patient subpopulations

As described in Chapter 3.3, considered breast cancer patient subpopulations were detected by the chosen clustering method applied to the RPPA data set. Hence, the obtained subtypes were expected to differ in their protein levels. Nevertheless, further analysis was required to identify proteomic profiles characteristic of each group. Furthermore, it remained unclear whether similar information can be gathered from mRNA gene expression measurements and if transcriptomic signatures support the obtained subtyping. Therefore, this part aims to characterize the identified breast cancer subpopulations with proteomic and, if possible, transcriptomic signatures, which are either specific for a single subtype or sufficient to differentiate the subtypes.

3.6.1 Subtype-specific marker identification

Subtype-specific markers were identified using the differentiation testing pipeline described in Chapter 3.5.2.2 and shown in Figure 3.2. The subtype comparison with the selected testing approach was performed separately for each transcript or protein. Tests for normality and variance homogeneity assumption verification were also applied feature-wisely with Benjamini-Hochberg correction for multiple testing (Benjamini & Hochberg, 1995). However, their results were interpreted per omics to ensure that all measurements from the same platform were analyzed consistently. Consequently, if the normality assumption was not fulfilled for most investigated proteins or transcripts, the entire data set was compared with a non-parametric approach.

The markers were identified based on either p-values or effect sizes. Considering large numbers of comparisons and subpopulations varying in size, the effect-size-based approach appears to be a more reliable solution. Subtype-specific markers were defined as proteins or transcripts with a significantly higher or lower level in only one subtype. The markers were identified in three feature spaces: proteomic data, transcriptomic data, and transcriptomic data limited to genes coding the proteins measured by the RPPA platform.

The subtype-specific marker identification process started with selecting features with significantly varying levels among the subtypes. Those markers were defined as non-specific. Secondly, post hoc test results were analyzed to filter the features with levels significantly different in all pairwise comparisons for a given subtype and non-differential in comparisons for the remaining subtypes. Lastly, the direction of changes per subtype was verified to be the same in all comparisons to ensure that the marker level is either higher in a given subtype referred to others or lower. The approach based on p-values, also called the test-based approach, used ANOVA or Kruskal-Wallis test for the first phase and an appropriate post hoc test for the second phase. For the effect-size-based method, η^2 and d values served, respectively, calculated as described in Chapter 3.5.2.2. Figure 3.3 presents the scheme of the subtype-specific marker identification process.

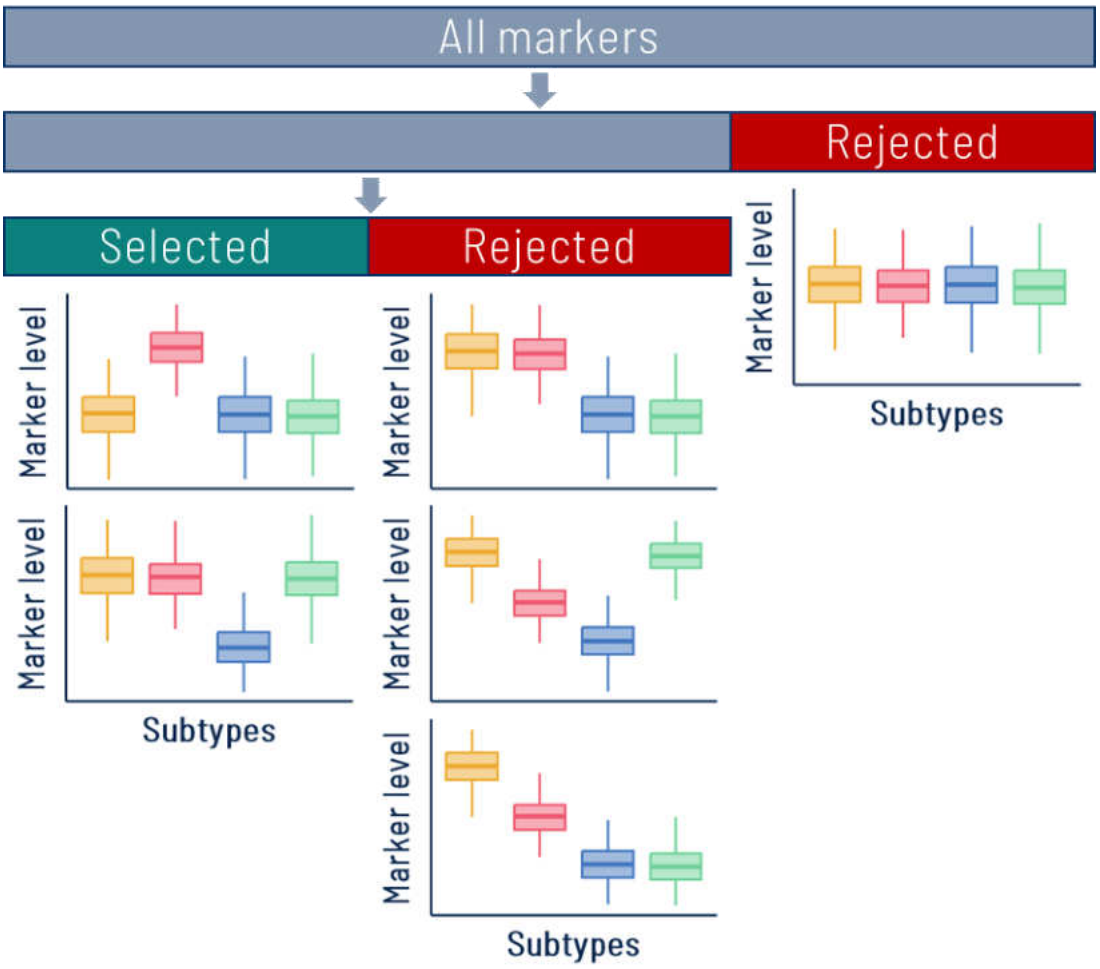


Figure 3.3 Subtype-specific marker identification process

ORA was performed on the sets of selected subtype-specific markers, including KEGG signaling pathways and Molecular Signatures Database (MSigDB) terms (Liberzon, et al., 2011; Liberzon, et al., 2015; Subramanian, et al., 2005). Since the size of the investigated protein universe and, in many cases, the size of the selected marker lists were insufficient for ORA, the second generation of enrichment analysis was also performed for the most interesting pairwise subtype comparisons. The CERNO test for KEGG pathways was conducted for genes ordered with the absolute values of Cohen's *d* effect size. Using χ^2 distribution, the test aims to verify if genes in a particular gene set are more likely to appear at the top of the ordered gene list, in this case, ordered by the size of differentiation between subtypes. The "tmod" R package was used for the CERNO test calculations (Weiner, 2022; Zyla, et al., 2019). For the enrichment analysis, proteins measured with RPPA were annotated with analogous gene names.

3.6.2 Subtype differentiating signature

As described in Chapter 3.6.1, the sets of proteomic and transcriptomic subtype-specific markers were identified, as well as the lists of markers with levels varying across subtypes determined as non-specific. Nonetheless, non-specific markers defined in that manner cover various scenarios: all subtype-specific markers are included, as well as features with significant differences in only one, several, or even all pairwise post hoc comparisons. Another approach was proposed based on the logistic regression to identify the molecular signature allowing for distinguishing all considered subtypes.

3.6.2.1 Multinomial logistic regression

Logistic regression is the method to estimate the probability of a binary response as a function of independent variables. It assumes that the dependent variable for each observation is described as Bernoulli-distributed data with an unknown probability of success. The logits of those probabilities are expressed as a linear function of independent variables in Equation 3.15. Consequently, the probabilities are estimated as the logistic function of all covariates, given by Equation 3.16. The unknown estimates β_i are chosen to maximize the likelihood function (James, Witten, Hastie, & Tibshirani, 2021).

$$\boldsymbol{logit(p) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j} \quad \textbf{3.15}$$

$$\boldsymbol{p = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_j x_j}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_j x_j}}} \quad \textbf{3.16}$$

Where:

p denotes the probability of belonging to a particular category,

β_j denotes the regression coefficient for the j-th independent variable,

x_j denotes the value of the j-th independent variable.

When the dependent variable has more than two categories, this approach can be extended to multinomial logistic regression, in which one of the classes serves as a baseline. Then, the probability of falling to each of the remaining categories can be estimated from Equation 3.17, while the probability

for the baseline category is given by Equation 3.18 (James, Witten, Hastie, & Tibshirani, 2021).

$$\boldsymbol{P(Y = k|x)} = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kj}x_j}}{\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lj}x_j}} \quad \textbf{3.17}$$

$$\boldsymbol{P(Y = 1|x)} = \frac{1}{\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lj}x_j}} \quad \textbf{3.18}$$

Where:

Y denotes the dependent variable,

k denotes a particular category of Y,

x_j denotes the value of the j-th independent variable,

β_{kj} and β_{lj} denote the regression coefficients for the j-th independent variable for k-th and l-th categories, respectively,

K denotes the number of Y categories.

Hence, the multinomial logistic regression provides the probability of belonging to each class, apart from the reference category. In this study, the subtype was the dependent variable, while protein or mRNA gene expression levels were potential independent variables.

3.6.2.2 Feature selection

Logistic regression is commonly applicable as the classification method. However, it can also be used to select meaningful features, such as the molecular signature of the identified breast cancer patient subpopulations.

Many models may be fitted for the given categories, varying in the sets of independent variables included. Those models can be assessed with the goodness-of-fit criteria like Akaike Information Criterion (AIC) (Akaike, 1974) or BIC (Schwarz, 1978), and with the ability to predict the actual category. The satisfactory model should accurately estimate the probability per category while remaining relatively simple. According to the parsimony rule, selecting as few relevant features as possible is crucial, i.e. features that provide high-quality prediction and are not redundant. Hence, selecting predictors for a high-quality

multinomial logistic regression model should be a signature distinguishing the proposed subtypes.

3.6.2.2.1 Forward selection method

There are many approaches to the problem of selecting features for the model, including forward and backward stepwise selection, genetic algorithm, or lasso shrinkage method (James, Witten, Hastie, & Tibshirani, 2021). In this work, the forward selection was used, as contrary to the backward method, it does not require creating a full model consisting of all possible covariates. This would be problematic, as the number of features in both sets is relatively large compared to the sample size. Moreover, the forward selection method provides the order in which features were chosen as the predictors, which gives an insight into their informative value. The forward method involves an iterative extending the model with a single feature, starting from the model with no independent variables. The new model is created and assessed with the quality index for each of the remaining features. Here, Bayes Factor (BF) served as the quality index, which can be derived based on the difference in BIC values between the two models, according to Equations 3.19 and 3.20 (Jeffreys, 1998; Schwarz, 1978; Wagenmakers, 2007). Thresholds for BF interpretation are shown in Table 3.5.

$$\boldsymbol{BF_{01} = \exp\left(\frac{\Delta BIC_{10}}{2}\right)} \quad \textbf{3.19}$$

$$\boldsymbol{\Delta BIC_{10} = BIC(H_1) - BIC(H_0)} \quad \textbf{3.20}$$

Where:

H_1 denotes the model with more features,

H_0 denotes the model with fewer features,

BIC denotes Bayesian Information Criterion as defined in (Schwarz, 1978).

The model with the highest BF was selected in each step of the forward selection method. The model was extended until BF dropped below ten or no more potential features were left. Moreover, the constraint was that the model's number of features should not be higher than the total number of patients used for training divided by 20.

Table 3.5 Thresholds for Bayes Factor interpretation

The table is adapted from (Jeffreys, 1998).

BF	Strength of evidence
$< 10^0$	Negative (supports M_2)
10^0 to $10^{1/2}$	Barely worth mentioning
$10^{1/2}$ to 10^1	Substantial
10^1 to $10^{3/2}$	Strong
$10^{3/2}$ to 10^2	Very strong
$> 10^2$	Decisive

3.6.2.2.2 Multiple random cross-validation

Multiple Random Cross-Validation (MRCV) procedure was used for model building with the number of iterations equal to 100. MRCV was chosen due to the limited number of patients and the high imbalance between the breast cancer subtypes. In each iteration, 10% of patients from each group (subtype) were left as the test set, and the remaining 90% served for training. The multinomial logistic regression model was built on this set using the abovementioned forward method. The performance of the resulting model was assessed based on the test set. Figure 3.4 presents the scheme of the MRCV procedure.

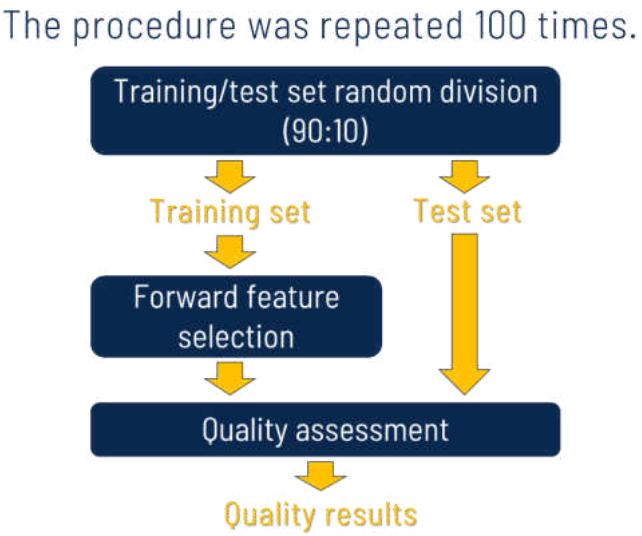


Figure 3.4 Multiple Random Cross-Validation procedure for multinomial logistic regression model building

3.6.2.2.3 Feature ranking

Outcomes of MRCV 100 repetitions served for the creation of the feature ranking. For each resulting model, features were sorted by the selection order and assigned a weight given by Equation 3.21.

$$w = 1 - \frac{k-1}{m} \quad 3.21$$

Where:

k denotes the feature order from firstly to lastly selected,

m denotes the maximal number of features in all 100 models.

Each weight was then multiplied by the corresponding model quality represented by the overall balanced accuracy (BA) calculated on each test set. Products summed up among all 100 models gave an importance score for each feature. Overall BA was defined as the weighted mean of balanced accuracies per category, with weights corresponding to the group sizes. The category balanced accuracy is a mean of sensitivity and specificity also calculated per category, according to Equations 3.22 and 3.23, respectively.

$$Sensitivity = \frac{TP}{P} \quad 3.22$$

$$Specificity = \frac{TN}{N} \quad 3.23$$

Where:

P denotes the number of observations in the given category,

N denotes the number of observations in the remaining categories,

TP denotes the number of correctly predicted observations in the given category,

TN denotes the number of observations from the remaining categories which are not predicted as the given category.

Hence, feature ranking merges two approaches of model assessment: goodness-of-fit-based, as the order of features corresponds to BF, and prediction-quality-based, represented by BA. Feature ranking served to identify the final molecular signature differentiating all subtypes. The elbow method was used

to select the cut-off for top features. It involved the feature ranking scores sorting, plotting, and connecting the highest and lowest values by line. The inflection point was the score with the maximal distance to the resulting line. All features with scores higher than the inflection point were selected as the model signature.

The analogous ranking-based feature selection method for binomial logistic regression was described in (Kozielski, et al., 2021; Henzel, et al., 2021).

Subtype differentiating signatures with the procedure described above were identified in three feature spaces: RPPA-derived protein levels, mRNA gene expression levels, and those two data sets combined. As the multinomial logistic regression does not deal with missing values, the features with incomplete records were removed for this part. Moreover, since the number of potential independent variables in the transcriptomic data set was huge, the model creation was preceded by GMM-based feature selection, as explained in Chapter 3.3.2.

4 Identification of patient subpopulations

4.1 Functional space of measured proteins

Reactome pathway ORA was performed to investigate the functional space covered by the set of proteins, which levels served for breast patent subpopulations' identification. The results are presented in Figure 4.1, in Voronoi diagram produced by the Reactome analysis tool. The whole figure space represents the entire Reactome pathway database, partitioned into smaller regions representing levels in the pathway hierarchy. The highlighted pathways were over-represented and colored based on the ORA hypergeometrical test p-values. The functional space of investigated proteins included mainly apoptosis, signaling, gene expression, immunological functions, and cellular response to stress.

Figure 4.1 Voronoi diagram generated with Reactome pathway Over-Representation Analysis (ORA) for the set of proteins used in this study

Color denotes over-represented pathways with ORA hypergeometrical test p-values lower than 0.05.

4.2 Batch effect

The dimensions of the original data set of 166 protein measurements for 876 female primary tumor samples were reduced, as previously described in Chapter 3.2.1. The resulting UMAP embedding was used to project the data set in the 2D space. Each data point corresponds to one sample and is colored by plate ID, as presented in Figure 4.2A, or by design referring to the locations of samples on slides, as presented in Figure 4.2C. As can be noticed in Figure 4.2A and Figure 4.2C, some samples with identical plate IDs or design labels tend to group. This is especially visible for data points annotated with design "5" and plate ID "A43F", situated at the bottom of the plots. Hence, the ComBat correction of batch effect was conducted, with the design labels as the batch etiquettes. The dimensions of the corrected data set were again reduced with PCA followed by UMAP, and analogous 2D visualizations were produced, in which colors denoted plate IDs or designs, as presented in Figure 4.2B and Figure 4.2D, respectively.

As shown in Figure 4.2, ComBat correction of batch effect performed satisfactorily, as distinct patterns of a specific plate ID or design are no longer visible. Data points annotated with design "5" and plate ID "A43F", which were grouped in the original data set, are now uniformly scattered. No communities of samples with the same plate ID can be detected based on the visual assessment of UMAP projections. Hence, the protein level data set corrected for batch effect served for further analysis steps.

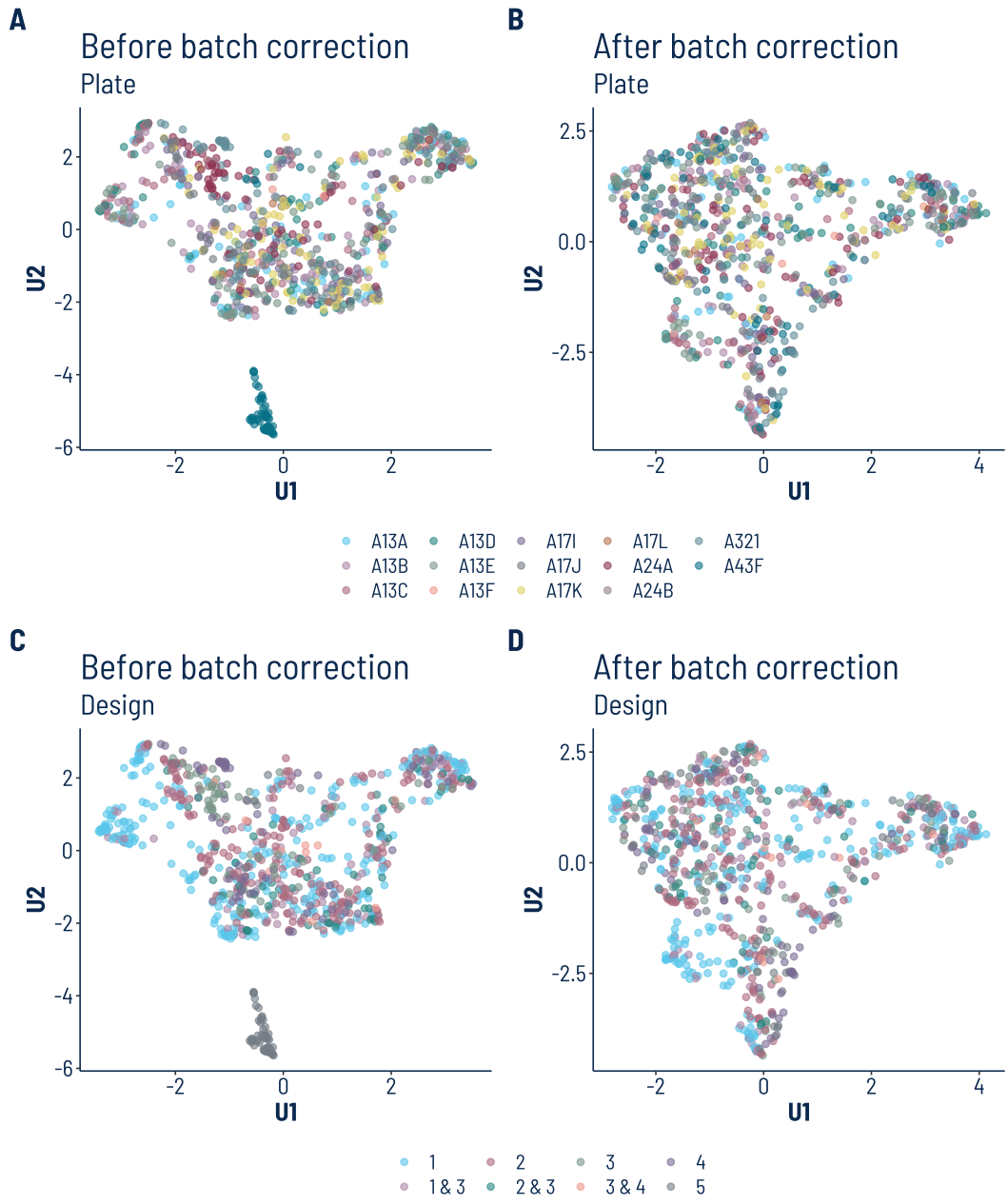


Figure 4.2 UMAP visualizations of protein level data set before and after batch effect correction

Data point color denotes plate ID in Panels A and B, and design in Panels C and D. Panels A and C show UMAP embedding of uncorrected data, while Panels B and D of data corrected for batch effect.

4.3 Clustering algorithms

Firstly, various combinations of clustering algorithms and, if necessary, feature selection and extraction techniques were applied to the data set of 166 protein levels measured for 407 patients. The summary of all tested variants of machine learning approaches is presented in Table 3.2 in Chapter 3.3.3. Figure 4.3 shows

the UMAP 2D visualization of the clustering results and subtype labels obtained with PAM50 transcriptomics-based predictor for all considered approaches.

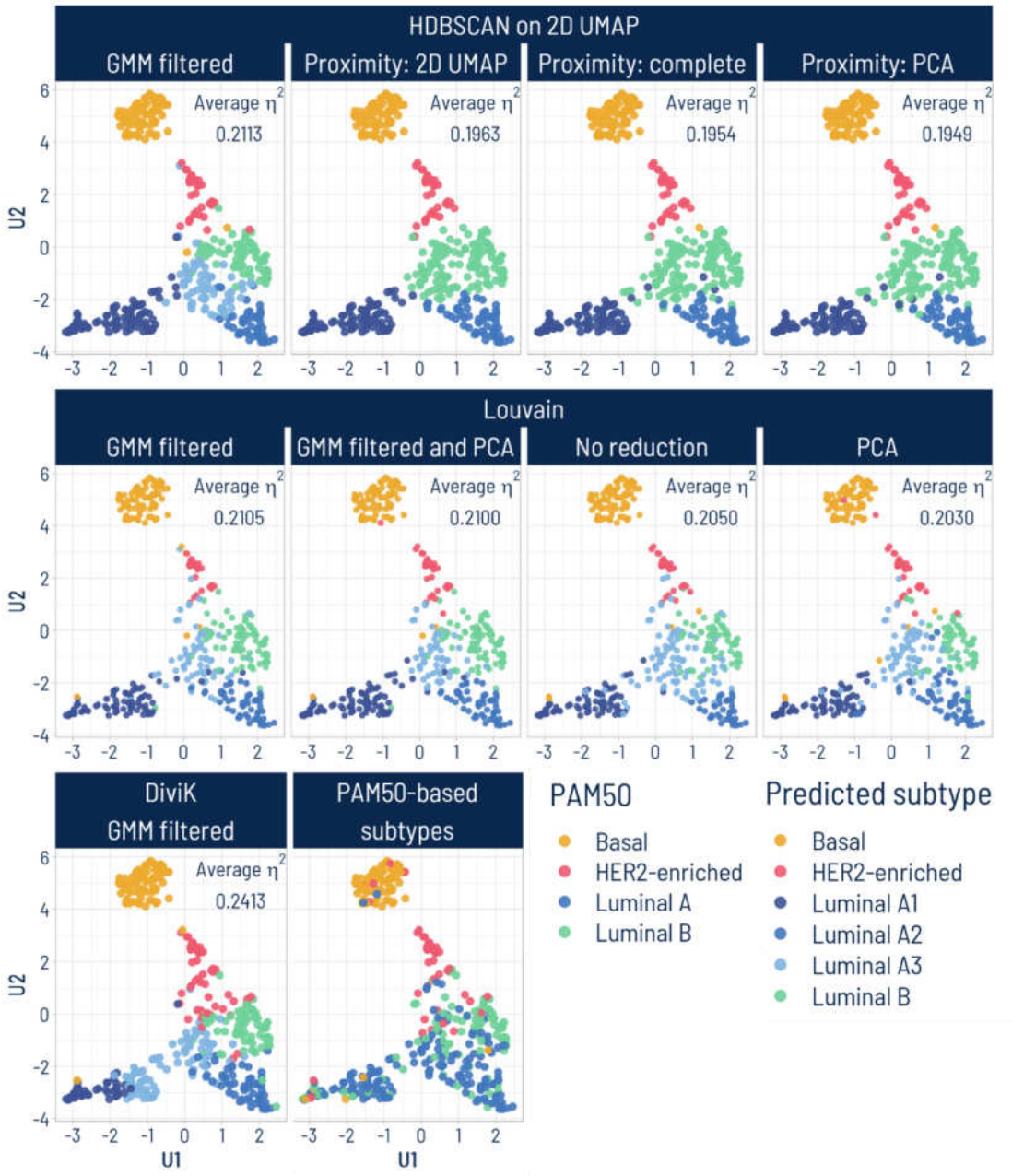


Figure 4.3 UMAP visualization with results of all clustering approaches and the original PAM50 subtype labels

Each figure corresponds to different clustering approach combined with various pre- or postprocessing procedures: data dimension reduction prior to clustering with feature selection and/or extraction, or in the case of the HDBSCAN method, the techniques to predict the subtype for unassigned patients. The data point color marks subtype: either predicted in this study, or obtained with the PAM50 predictor.

The figure is adapted from (Tobiasz & Polanska, 2022).

Interestingly, all results suggest that the luminal A subtype, the most numerous in the TCGA-BRCA cohort, is highly diverse and consists of several subpopulations. In all approaches, the final number of clusters detected was higher than the number of possible PAM50 etiquettes.

For the data set used in this study, the HDBSCAN algorithm only worked when applied to the 2D feature space, so feature extraction with PCA followed by UMAP was required. HDBSCAN approaches without GMM-based feature selection provided five clusters representing basal, HER2-enriched, luminal A, and luminal B subtypes. Moreover, luminal A cases were split into two subgroups.

All remaining combinations of machine learning approaches (HDBSCAN preceded by GMM-based selection, Louvain algorithm, and DiviK method) detected six clusters: one per basal, HER2-enriched, and luminal B subtypes, and three corresponding to luminal A subtypes. Interestingly, as can be noticed in Figure 4.3, the third luminal A cluster contains many cases labeled as luminal B by the PAM50 predictor. Moreover, the luminal A subpopulation located at the left-hand side of the UMAP embedding (for small values of both UMAP components) is a mixture of samples highly heterogeneous in terms of the PAM50 labels: not only luminal A and B subtypes can be noticed there, but also representatives of basal or HER2-enriched cases. The basal subpopulation forms a distinctly isolated community, while the HER2-enriched subtype is located between basal and luminal cases, slightly overlapping with the latter ones.

4.4 Clustering outcome comparison

The obtained clustering outcomes were evaluated with the η^2 and pooled d metrics, as explained in Chapter 3.4, and detected subpopulations were compared with PAM50 subtypes using Dice coefficients.

Figure 4.4A shows the distributions of η^2 values per clustering approach. Furthermore, in Figure 4.4B, the per cluster distributions of absolute d values are presented in the example of the DiviK method with a variance-based filtration step built-in.

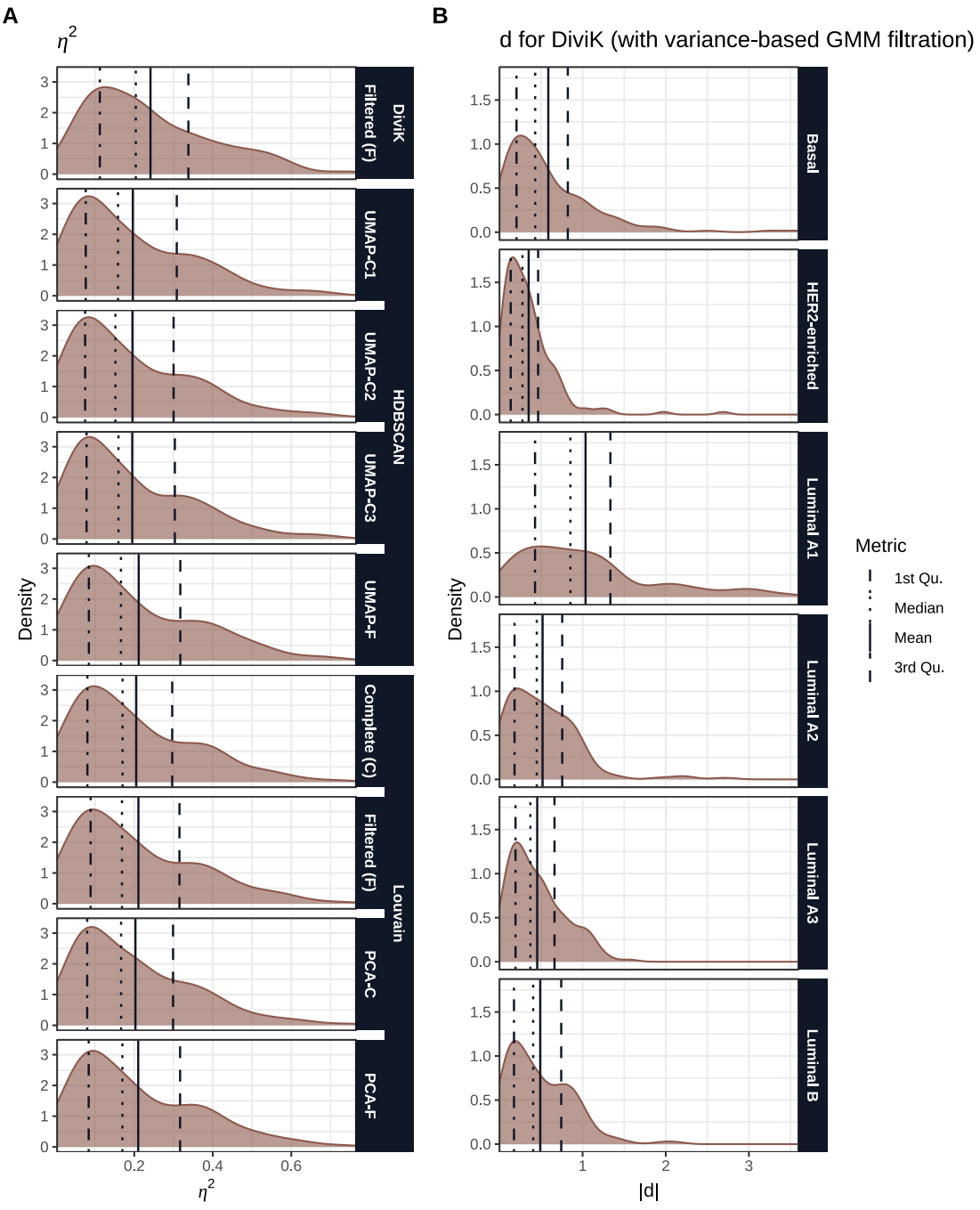


Figure 4.4 The distributions of metrics values

The metrics quartiles, median, and mean values are marked with vertical lines. Panel A density plots shows distributions of η^2 values per method. Panel B density plots show distributions of absolute d values per subtype for the DiviK method with variance-based Gaussian Mixture Model feature filtration.

The figure is taken from (Tobiasz & Polanska, 2022).

The distributions shown in Figure 4.4 indicate Q_3 to be an appropriate representation of both metrics' values for all proteins. Remaining relatively resistant to the outliers' influence, Q_3 still sufficiently reflects the impact

of proteins with levels significantly varying between the subtypes. As can be seen in Figure 4.4, three clusters detected by DiviK corresponding to the luminal A subtype were sorted by the decreasing Q_3 of the absolute d values and numbered accordingly (A1, A2, and A3, respectively).

Table 4.1 shows the values of η^2 quartiles and mean, pooled d scores, and Dice coefficient values per clustering approach. The Dice coefficients were compared with pooled d and Q_3 of η^2 in Figure 4.5.

Table 4.1 Metrics obtained with various combinations of feature engineering methods and clustering algorithms

The table is taken from (Tobiasz & Polanska, 2022).

Method	No. clusters	η^2				Pooled d	Dice Coeff.
		Q_1	Median	Mean	Q_3		
$H_{UMAP\text{-}C1}$	5	0.0764	0.1587	0.1963	0.3083	1.7053	0.7125
$H_{UMAP\text{-}C2}$	5	0.0749	0.1519	0.1954	0.3002	1.7204	0.7052
$H_{UMAP\text{-}C3}$	5	0.0785	0.1598	0.1949	0.3034	1.6847	0.7052
$H_{UMAP\text{-}F}$	6	0.0844	0.1661	0.2113	0.3173	1.8529	0.7469
L_C	6	0.0806	0.1702	0.2050	0.2966	1.8534	0.7469
$L_{PCA\text{-}C}$	6	0.0800	0.1665	0.2030	0.2989	1.8105	0.7445
L_F	6	0.0889	0.1687	0.2105	0.3151	1.8342	0.7396
$L_{PCA\text{-}F}$	6	0.0839	0.1698	0.2100	0.3168	1.8066	0.7371
DiviK	6	0.1123	0.2040	0.2413	0.3379	2.0568	0.7273

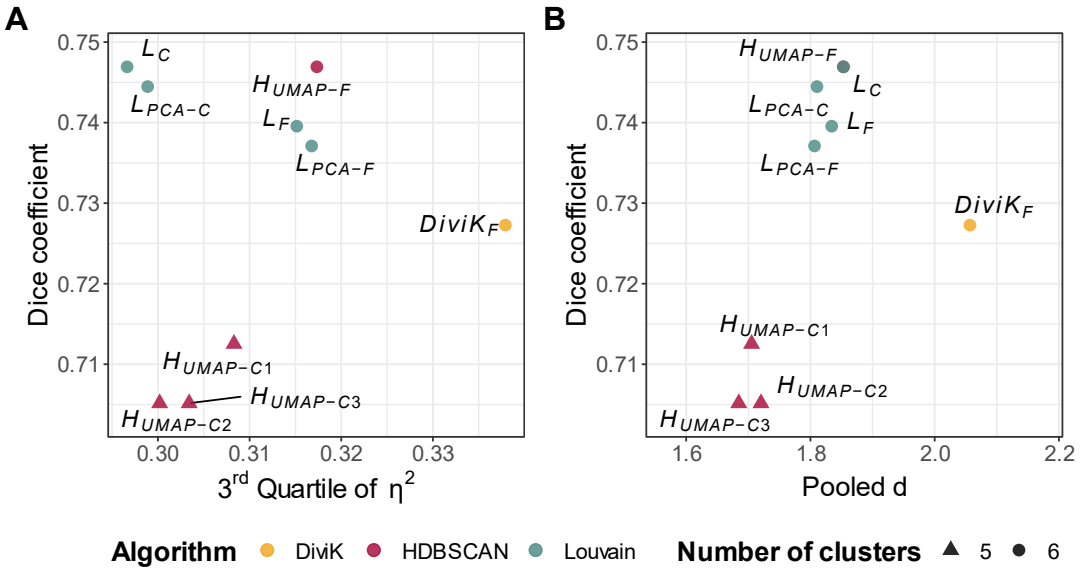


Figure 4.5 Comparison of η^2 and pooled d with Dice coefficient for tested clustering approaches

Panel A shows the 3rd quartile of η^2 versus Dice coefficient values plotted versus the 3rd quartile of η^2 (Panel A) and pooled d (Panel B).

The figure is taken from (Tobiasz & Polanska, 2022).

As for the comparison to PAM50 subtype labels based on the Dice coefficient, all methods which gave six clusters outperformed those which detected just five subpopulations. The highest Dice coefficient was observed for the Louvain algorithm applied to the whole feature space and for HDBSCAN clustering preceded by GMM-based feature selection and feature extraction with UMAP.

As it can be concluded from Table 4.1 and Figure 4.5, the DiviK method gave the maximal values of both effect-size-based metrics: η^2 and pooled d. The worst metrics values were obtained for HDBSCAN approaches with no GMM-based feature filtration. Pooled d scores were lowest for the variant in which a subtype of unassigned patients was predicted based on the proximity in top PCs, explaining 90% of the variance in the data. The main difference between those two approaches contrasting in their evaluation outcome is that the DiviK algorithm detected an additional luminal A3 cluster, taking over some cases included in the HDBSCAN approach, mainly in luminal B and luminal A1 subtypes.

Those two contrasting procedures were further analyzed by comparing the protein d values for corresponding luminal clusters: A1 versus A1, A2 versus A2,

B versus B, and DiviK luminal A3 versus HDBSCAN luminal B. The scatterplots for those comparisons are shown in Figure 4.6. Moreover, for those two clustering approaches contrasting in the evaluation, Table 4.2 contains the total numbers of proteins with significantly increased or decreased levels for the particular subtype and the numbers of KEGG pathways those proteins participate in. The significantly higher or lower-level proteins were selected with the thresholds for at least large and very large Cohen's *d* effect size (Cohen, 2013).

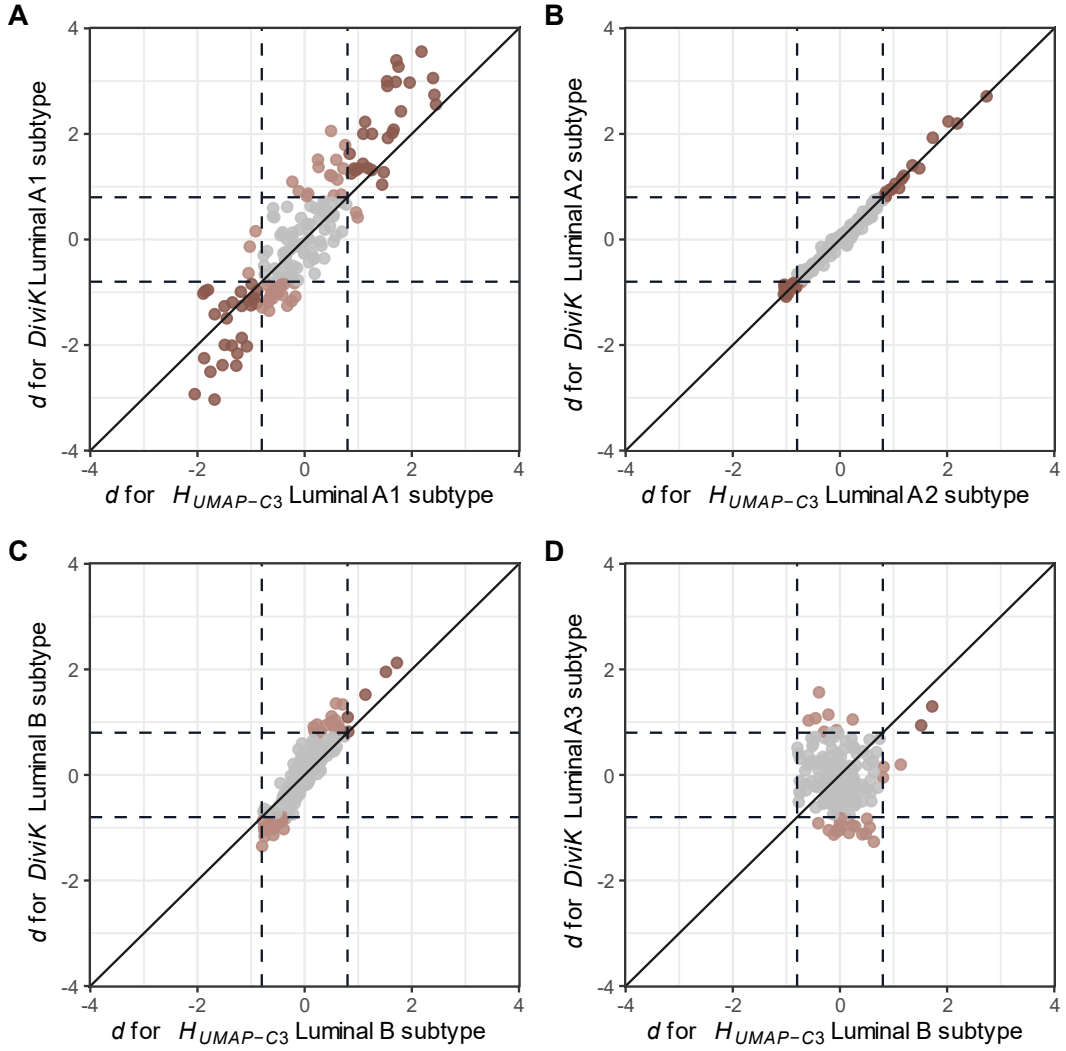


Figure 4.6 Protein *d* values for the best and the worst approach according to pooled *d* metrics

Y-axis refers to the DiviK algorithm with variance-based GMM filtration, while the X-axis to the HDBSCAN algorithm with the proximity in the set of top principal components explaining 90% of the variance for prediction, preceded by UMAP dimension reduction ($H_{UMAP\text{-}C3}$). *d* values are compared for the corresponding: luminal A1 subtypes (Panel A), luminal A2 subtypes (Panel B), luminal B subtypes (Panel C), and DiviK luminal A3 versus $H_{UMAP\text{-}C3}$ luminal B subtypes (Panel D). Dashed lines mark the threshold values for the large

effect size equal to -0.8 and 0.8 (Cohen, 2013). Values for proteins with small or medium effects in both compared approaches are marked in grey.

The figure is taken from (Tobiasz & Polanska, 2022).

Table 4.2 Total numbers of differentiating proteins and corresponding KEGG pathways for the best and worst approach according to pooled *d* metrics

The best approach is the DiviK algorithm with variance-based GMM filtration, while the worst one is the HDBSCAN algorithm with the proximity in the set of top principal components explaining 90% of the variance for prediction, preceded by UMAP dimension reduction ($H_{UMAP\text{-}C3}$). Differentiating proteins per subtype were selected based on the thresholds for large and very large effects (Cohen, 2013).

The table was taken from (Tobiasz & Polanska, 2022).

Subtype	At least large |d|				At least very large |d|			
	No. proteins		No. KEGG pathways		No. proteins		No. KEGG pathways	
	$H_{UMAP\text{-}C3}$	DiviK	$H_{UMAP\text{-}C3}$	DiviK	$H_{UMAP\text{-}C3}$	DiviK	$H_{UMAP\text{-}C3}$	DiviK
Basal	41	44	60	61	16	19	31	42
HER2-enriched	12	9	47	31	5	4	27	23
Luminal A1	59	89	83	86	34	54	76	80
Luminal A2	37	38	65	64	6	7	4	4
Luminal A3	-	28	-	36	-	3	-	4
Luminal B	5	39	2	79	2	6	0	10

Barely any differences in *d* values can be noticed for luminal A2 subtypes, as the data points form an almost straight line. Comparing DiviK luminal A3 and HDBSCAN luminal B subtypes does not favor any clustering outcome. Nevertheless, an increase in protein absolute *d* values can be noticed for the DiviK approach for the luminal B subtype, and especially for the luminal A1 subtype. Those observations are also supported by the results presented in Table 4.2, where the number of proteins is distinctly higher for the DiviK approach regarding luminal A1 and B subtypes. However, a similar effect on the number of KEGG pathways can be noticed only for the luminal B subtype.

Finally, the DiviK clustering approach was selected as the most appropriate method of patient subpopulation identification. DiviK clustering results are referred to the PAM50 subtypes in terms of the number of cases in Table 4.3, fractions of DiviK clusters per PAM50 subtype in Table 4.4, and fractions of PAM50 subtype per DiviK cluster in Table 4.5.

Table 4.3 Number of patients in DiviK-based clusters referred to PAM50 subtypes

PAM50 subtype	DiviK-based predicted subtype						TOTAL
	Basal	HER2-enriched	Luminal				
			A1	A2	A3	B	
Basal	79	0	4	0	2	1	**86**
HER2-enriched	8	34	2	0	2	4	**50**
Luminal A	2	9	27	47	65	23	**173**
Luminal B	0	11	11	14	18	44	**98**
TOTAL	**89**	**54**	**44**	**61**	**87**	**72**	**407**

Table 4.4 Percentage of patients of each PAM50 subtype in DiviK-based clusters

PAM50 subtype	DiviK-based predicted subtype						TOTAL
	Basal	HER2-enriched	Luminal				
			A1	A2	A3	B	
Basal	91.86	0.00	4.65	0.00	2.33	1.16	**100.00**
HER2-enriched	16.00	68.00	4.00	0.00	4.00	8.00	**100.00**
Luminal A	1.16	5.20	15.61	27.17	37.57	13.29	**100.00**
Luminal B	0.00	11.22	11.22	14.29	18.37	44.90	**100.00**

Table 4.5 Percentage of patients of each DiviK-based cluster per PAM50 subtypes

PAM50 subtype	DiviK-based predicted subtype					
	Basal	HER2-enriched	Luminal			
			A1	A2	A3	B
Basal	88.76	0.00	9.09	0.00	2.30	1.39
HER2-enriched	8.99	62.96	4.55	0.00	2.30	5.56
Luminal A	2.25	16.67	61.36	77.05	74.71	31.94
Luminal B	0.00	20.37	25.00	22.95	20.69	61.11
TOTAL	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

4.5 Conclusions and discussion

Obtained results suggest that the RPPA data set should be divided into five or six subgroups, with two or three clusters for the luminal A subtype and one cluster per remaining subtype. Based on the η^2 and pooled d metrics, the clustering

outcome given by the DiviK algorithm was selected as the one providing the most distinct six subpopulations. Hence, three subgroups corresponding to luminal A cases were obtained. Some other methods performed better in terms of the agreement with the transcriptomics-based PAM50 predictor, as indicated by the Dice coefficient values. However, the aim was not to maximize the similarity to the original PAM50-based subtypes but to select the approach which would provide possibly distant clusters. Hence, the purpose was to maximize the η^2 and pooled d scores, with the Dice coefficient as the additional information.

Both effect-size-based metrics were higher when six clusters were obtained instead of five. Interestingly, GMM-based feature selection improved the Q_3 of η^2 for HDBSCAN and Louvain algorithms and increased the pooled d score in the case of HDBSCAN. It is especially visible in Figure 4.5A, where data points corresponding to Louvain approach with and without the filtration step are distinctly separated. Therefore, referring the pooled d scores to other criteria, like the Dice coefficient, seems beneficial as it provides a more comprehensive insight.

The clustering approaches with the best and worst performance according to pooled d score differ mainly in luminal group splitting. The worst approach gave only two luminal A subtypes and one bigger luminal B subtype. On the other hand, the DiviK algorithm selected as the best approach distinguished one more luminal A cluster, which covered some luminal A1 and luminal B cases. Furthermore, the size HER2-enriched subpopulation was greater for the DiviK algorithm, as some patients assigned as luminal B cases in the HDBSCAN approach were included there.

The selection of additional luminal A3 subtype increased the absolute d values for luminal A1 and B subpopulations and, consequently, the number of proteins with large and very large effects. Luminal A2 clusters do not differ much between the contrasting approaches. This indicates that the patients with moderate protein levels were possibly assigned to the additional luminal A3 by the DiviK algorithm. Consequently, it caused the luminal A1 and B subpopulations to be less numerous and more extreme in their protein levels, leading to increased d absolute values. Interestingly, the number of differentiating proteins and involved KEGG pathways decreased for the HER2-enriched subtype for the DiviK algorithm.

The obtained results are relatively concordant with the outcomes of similar investigations. Applying the DiviK algorithm on RNA-Seq TCGA-BRCA data provided five clusters, although the cohort size was over two times bigger than in this study. However, two homogeneous subtypes were detected, highly consistent with PAM50 basal and HER2-enriched groups. The remaining three clusters corresponded to luminal cases, one almost equally balanced between luminal A and B cases (45.0% and 52.2%, respectively) and two containing mainly luminal A tumors (Tobiasz, Hatzis, & Polanska, 2019). In (The Cancer Genome Atlas Network, 2012), RPPA measurements for 403 TCGA breast cancer samples were hierarchically clustered, which gave seven subgroups. One, however, was very small. The clusters again showed great agreement with PAM50 labels, mainly in basal and HER2-enriched subtypes. The luminal RPPA-defined clusters included one mainly luminal A, one composed of both luminal A and B cases, one mainly luminal A with several luminal B and HER2-enriched cases, and one cluster highly heterogeneous. Those results for luminal tumors demonstrate similarities with DiviK outcome, as distinguishing between luminal A and B tumors seems challenging, some HER2-enriched cases show similarity with luminal subtypes, and there is one small and highly heterogeneous subgroup. The luminal group was reported to be diverse or even a continuum (Szymiczek, Lone, & Akbari, 2020).

To conclude, the applied effect-size-based methods of clustering outcome evaluation performed satisfactorily. The d score representing the differences between the given cluster and all remaining ones was proposed. All metrics' results were consistent regarding the best machine learning approach for breast cancer subpopulation identification - the custom DiviK-approach with GMM-based feature selection and stepwise k-means clustering outperformed other methods. Moreover, the GMM-based feature selection before clustering improved the cluster separability.

Finally, six breast cancer patient subpopulations were identified and named based on the concordance with the PAM50 subtype etiquettes: basal, HER2-enriched, luminal B, and three luminal A subtypes (A1, A2, A3, sorted from the one most distinct to remaining subpopulations to the one most similar).

5 Clinical characteristics of patient subpopulations

The subtyping outcomes obtained with the DiviK algorithm were evaluated by investigating patients' clinical and demographic profiles in different subpopulations. In particular, this part was focused on a comparative analysis of the detected luminal subtypes, which were the main modification compared to the set of subtypes provided by the PAM50 transcriptomic-based classifier. The purpose was to verify whether demographic background, survival, and clinical outcomes have any differentiating significance and if they support the decision to divide luminal cases into four subgroups instead of only two luminal A and B, like in the PAM50 predictor.

5.1 Survival analysis

Median event and censored times for each subtype were compared to assess the quality of survival outcome data provided by TCGA. The median values are presented in Table 5.1 regarding all subpopulations identified using DiviK clustering on protein levels in Chapter 4 and Table 5.2 for the subtypes based on the transcriptomic PAM50 classifier. In some cases, e.g., for DFI of all investigated subtypes, the median event time was impossible to estimate because the survival function did not drop to 50% throughout the entire timespan of the follow-up. For this reason, calculations of 95% confidence interval (CI) of median event time also failed for most of the subtypes and endpoints, and hence those results were not shown in Table 5.1 and Table 5.2.

Table 5.1 Median event and censored times per endpoint type for the identified subpopulations

Subtype	Overall Survival		Disease-Specific Survival		Disease-Free Interval		Progression-Free Interval	
	Event	Cens.	Event	Cens.	Event	Cens.	Event	Cens.
Basal	20.42	2.97	-	2.94	-	2.97	-	2.94
HER2-enriched	12.21	3.48	12.21	3.13	-	2.88	12.21	3.13
Luminal A1	11.69	4.43	-	4.23	-	4.21	-	4.28
Luminal A2	8.56	3.15	9.34	3.10	-	2.96	-	3.10
Luminal A3	-	3.17	-	2.86	-	2.87	-	2.87
Luminal B	7.98	2.10	-	2.10	-	2.10	-	2.10

Table 5.2 Median event and censored times per endpoint type for the PAM50 subtypes

Subtype	Overall Survival		Disease-Specific Survival		Disease-Free Interval		Progression-Free Interval	
	Event	Cens.	Event	Cens.	Event	Cens.	Event	Cens.
Basal	20.42	3.30	-	3.12	-	3.12	-	3.12
HER2-enriched	17.69	3.41	-	2.88	-	2.73	-	2.88
Luminal A	10.81	3.19	-	3.13	-	3.15	-	3.15
Luminal B	8.94	2.74	10.80	2.55	-	2.58	-	2.22

The median censored times were less than the median event times in each case when the estimation of both and comparison was possible. This confirms the limitation of the survival outcome data for the TCGA-BRCA cohort, mentioned in (Liu, et al., 2018) and Chapter 3.5.1. The follow-up was not long enough to observe a sufficient number of events. As indicated in (Liu, et al., 2018), this is a common issue for survival analysis of less aggressive tumors like breast cancer. Hence, the following results of the survival experience comparison must be considered carefully.

The KM graphs of survival functions for all four considered endpoints are shown in Figure 5.1 for luminal subpopulations identified with DiviK proteomic-based approach and in Figure 5.2 for luminal PAM50 subtypes. A comparison of luminal subgroups is highlighted here to investigate the main difference between DiviK- and PAM50-based subtyping approaches. HER2-enriched and basal subtypes were highly concordant for both proteomic and transcriptomic subtyping. However, the KM curves for all subpopulations, including HER2-enriched and basal cases, are shown in Supplementary Figure 8.1 and Supplementary Figure 8.2 for DiviK-based and PAM50 assignments, respectively. Moreover, KM graphs limited to only three luminal A subgroups detected with DiviK are presented in Supplementary Figure 8.3. All plots were truncated at a 10-year follow-up time for clarity, as suggested in (Liu, et al., 2018), since hardly any events of interest could have been observed after that time for the TCGA-BRCA cohort. Each KM graph is accompanied by the p-values resulting from log-rank and Gehan-Wilcoxon tests conducted to verify whether the survival functions were the same for all investigated subtypes. Moreover, the test statistics and p-values are presented in Table 5.3 for DiviK-based

and PAM50 luminal subtypes, Supplementary Table 8.1 for all subtypes, and Supplementary Table 8.4 for the comparison of luminal A subpopulations detected with DiviK.

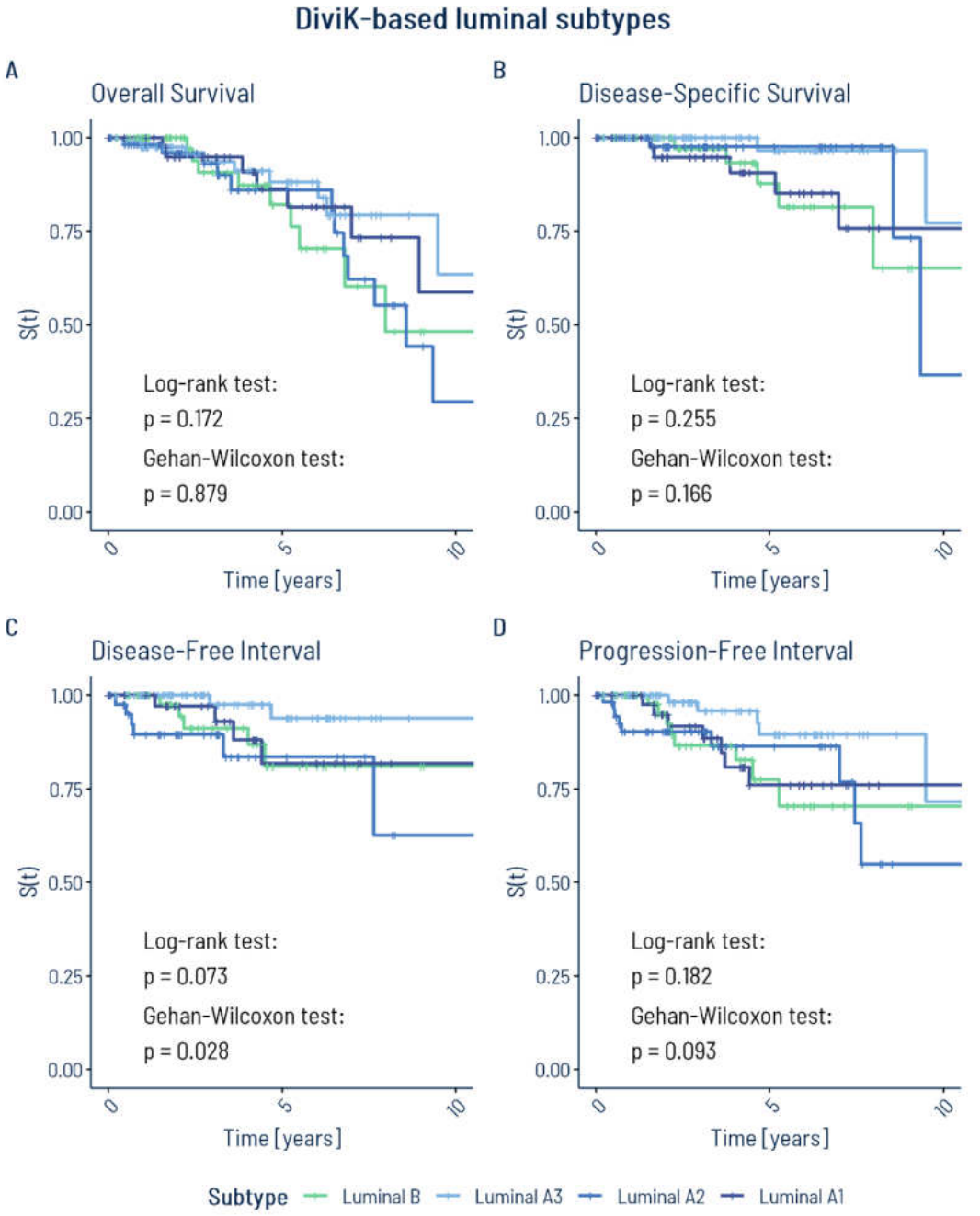


Figure 5.1 Kaplan-Meier survival curves of luminal subpopulations identified with DiviK

Interestingly, for comparing luminal subpopulations identified with DiviK, the p-value was higher for the Gehan-Wilcoxon test than for the log-rank test only for OS, which is the most biased endpoint among all considered here. However,

no differences in survival outcomes can be spotted for OS based on both test results and KM curves (Figure 5.1A). When the emphasis was placed more on the early changes in the survival experience in the Gehan-Wilcoxon test, the p-value decreased for DSS, DFI, and PFI. Those results were also supported by the KM graphs, especially for DFI and PFI, where the distinct drop in the survival function of luminal A2 cases can be observed during the first year of follow-up (Figure 5.1C and D). The p-value is lower than 0.05 only for DFI. For DSS, two groups of similar curves can be noticed (Figure 5.1B): one with luminal A2 and A3 subpopulations with a better prognosis and one consisting of luminal A1 and B subtypes with a worse outcome. That is a rather interesting result, especially considering the UMAP visualization created from the protein levels, which showed luminal A1 and B clusters located far from each other (Figure 4.3). Based on the KM graphs, it can be concluded that the luminal A3 subtype generally can be associated with the best prognosis regarding recurrence among all investigated patient subgroups. Supplementary Figure 8.3 and Supplementary Table 8.4 provide similar results concerning comparing luminal A subtypes only, without luminal B. For that subset, both log-rank and Gehan-Wilcoxon p-values for DFI were lower than 0.05.

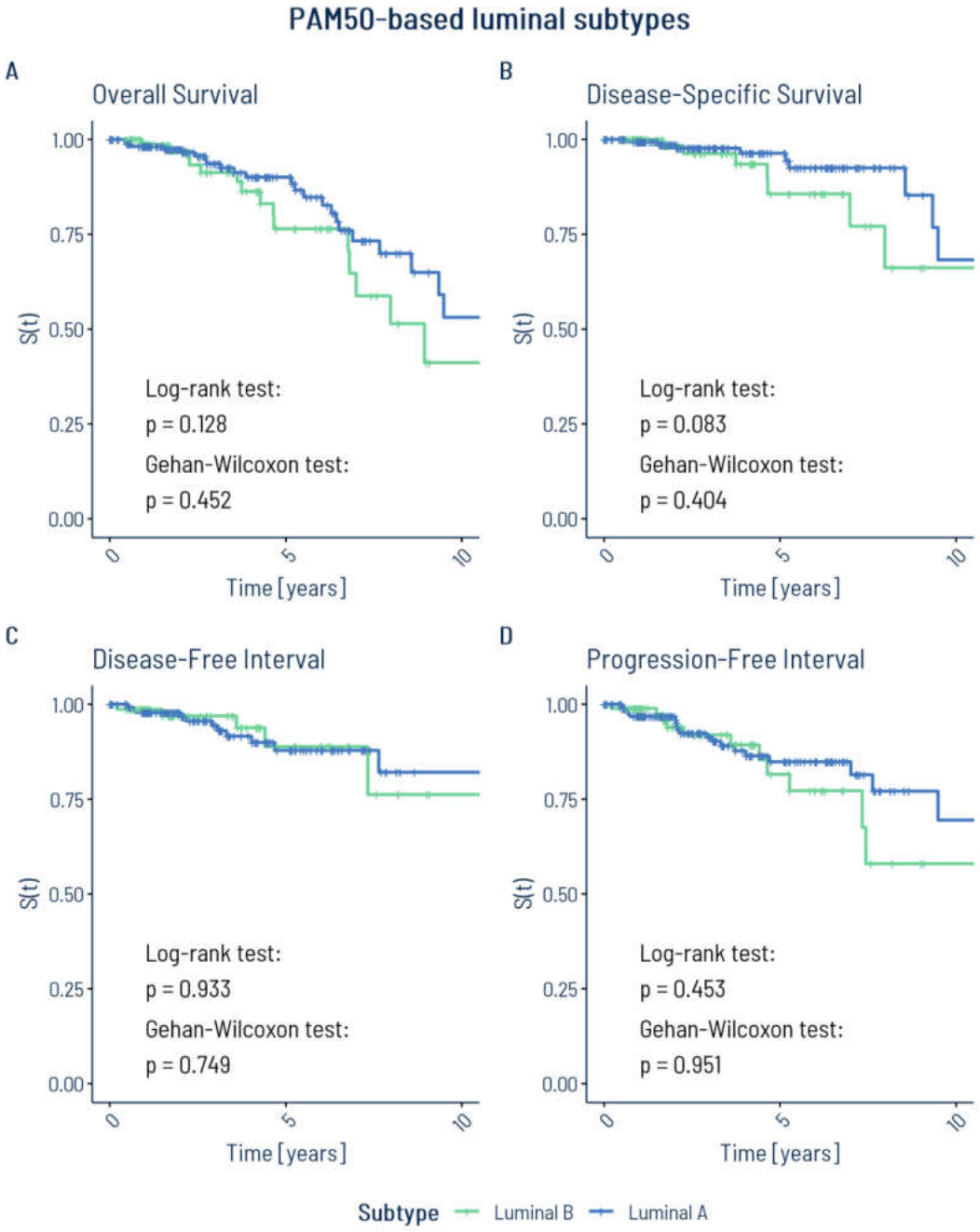


Figure 5.2 Kaplan-Meier survival curves of luminal PAM50 subtypes

Contrary to results for DiviK-based luminal subpopulations, for PAM50 subtypes, the log-rank p-values were lower than for the Gehan-Wilcoxon test for all endpoints apart from DFI. This suggests higher differences between the survival outcomes for patients with longer follow-up time. As observed in KM plots (Figure 5.2), the luminal A subtype shows a slightly better prognosis than luminal B in the late phases following the initial diagnosis, although no significant differences were detected.

Table 5.3 Results of log-rank and Gehan-Wilcoxon tests for comparison of survival functions of luminal subtypes identified with DiviK or based on PAM50 classifier

Endpoint type	χ^2		p-value	
	Log-rank test	Gehan-Wilcoxon test	Log-rank test	Gehan-Wilcoxon test
Subpopulations identified with DiviK				
Overall Survival	4.99	0.68	0.1724	0.8788
Disease-Specific Survival	4.06	5.08	0.2552	0.1661
Disease-Free Interval	6.97	9.12	0.0730	0.0277
Progression-Free Interval	4.87	6.41	0.1818	0.0932
PAM50-based subtypes				
Overall Survival	2.32	0.57	0.1280	0.4521
Disease-Specific Survival	3.01	0.70	0.0828	0.4043
Disease-Free Interval	0.01	0.10	0.9333	0.7488
Progression-Free Interval	0.56	0.003	0.4530	0.9512

Table 5.4 and Table 5.5 show the Cox proportional hazard analysis results for luminal subtypes based on DiviK clustering on protein levels and PAM50 predictor, respectively. The luminal B group served as the reference. Analogues results are presented for all subtypes, including basal and HER2-enriched, in Supplementary Table 8.2 and Supplementary Table 8.3. For those analyses, basal subtypes were used as a baseline. Moreover, Supplementary Table 8.5 shows the Cox regression analysis outcomes for only the luminal A subpopulations identified with DiviK. In this subset, the luminal A3 subtype was chosen as a reference. As previously mentioned, a low number of events was captured due to the short follow-up span. Nevertheless, some conclusions can be carefully drawn based on the HR treated as the effect size measure given appropriate thresholds adjustment for imbalanced classes as described in Chapter 3.5.1.2.

For Cox regression analysis on DiviK-based luminal subpopulations, only small or neglectable effect was observed for luminal A1 and A2 subtypes, referred to luminal B. However, depending on the endpoint, the medium and large effect was detected in favor of luminal A3 tumors. This supports the observations made based on KM graphs (Figure 5.1) that the luminal A3 subpopulation is associated

with a better prognosis in terms of death and recurrence. Moreover, the luminal A2 subtype showed a slightly worse survival outcome than luminal B, differentiating this subgroup from the other two luminal A subpopulations. The exemption here is DSS, for which all luminal A subtypes showed a lower risk than luminal B.

Table 5.4 Cox proportional hazard analysis of identified luminal subpopulations

Subtype	N	N_e	N_c	HR	HR effect	HR allocation probability adjusted critical value		
						α = 0.1	α = 0.3	α = 0.5
Overall Survival								
Luminal A1	44	8	36	0.657	Small	HR < 0.773; HR > 1.293	HR < 0.47; HR > 2.13	HR < 0.275; HR > 3.636
Luminal A2	61	14	47	1.275	Small	HR < 0.805; HR > 1.242	HR < 0.517; HR > 1.934	HR < 0.314; HR > 3.18
Luminal A3	87	9	78	0.533	Medium	HR < 0.831; HR > 1.203	HR < 0.561; HR > 1.783	HR < 0.354; HR > 2.828
Luminal B	72	9	63	Reference				
Disease-Specific Survival								
Luminal A1	43	5	38	0.759	Small	HR < 0.771; HR > 1.297	HR < 0.466; HR > 2.146	HR < 0.272; HR > 3.674
Luminal A2	58	4	54	0.776	Small	HR < 0.801; HR > 1.249	HR < 0.51; HR > 1.961	HR < 0.309; HR > 3.241
Luminal A3	85	2	83	0.213	Large	HR < 0.83; HR > 1.205	HR < 0.558; HR > 1.792	HR < 0.351; HR > 2.847
Luminal B	72	5	67	Reference				
Disease-Free Interval								
Luminal A1	38	4	34	0.927	No effect	HR < 0.77; HR > 1.298	HR < 0.465; HR > 2.15	HR < 0.271; HR > 3.684
Luminal A2	46	6	40	1.748	Small	HR < 0.79; HR > 1.266	HR < 0.494; HR > 2.025	HR < 0.295; HR > 3.391
Luminal A3	79	2	77	0.25	Large	HR < 0.833; HR > 1.201	HR < 0.563; HR > 1.776	HR < 0.356; HR > 2.81
Luminal B	64	5	59	Reference				
Progression-Free Interval								
Luminal A1	44	7	37	0.839	No effect	HR < 0.773; HR > 1.293	HR < 0.47; HR > 2.13	HR < 0.275; HR > 3.636
Luminal A2	61	9	52	1.101	No effect	HR < 0.805; HR > 1.242	HR < 0.517; HR > 1.934	HR < 0.314; HR > 3.18
Luminal A3	87	5	82	0.359	Medium	HR < 0.831; HR > 1.203	HR < 0.561; HR > 1.783	HR < 0.354; HR > 2.828
Luminal B	72	8	64	Reference				

Table 5.5 Cox proportional hazard analysis of luminal PAM50 subtypes

"LumA" denotes luminal A subtype. "LumB" denotes luminal B subtype.

Subtype	N	N_e	N_c	HR	HR effect	HR allocation probability adjusted critical value		
						α = 0.1	α = 0.3	α = 0.5
Overall Survival								
LumA	173	26	147	0.612	Small	HR < 0.852; HR > 1.174	HR < 0.598; HR > 1.671	HR < 0.39; HR > 2.566
LumB	98	16	82	Reference				
Disease-Specific Survival								
LumA	169	10	159	0.437	Medium	HR < 0.852; HR > 1.174	HR < 0.598; HR > 1.672	HR < 0.389; HR > 2.568
LumB	96	8	88	Reference				
Disease-Free Interval								
LumA	147	11	136	1.046	No effect	HR < 0.852; HR > 1.173	HR < 0.6; HR > 1.668	HR < 0.391; HR > 2.558
LumB	82	5	77	Reference				
Progression-Free Interval								
LumA	173	20	153	0.752	Small	HR < 0.852; HR > 1.174	HR < 0.598; HR > 1.671	HR < 0.39; HR > 2.566
LumB	98	11	87	Reference				

Referring PAM50 luminal A subtype to luminal B showed only a small effect in terms of OS and PFI and a medium effect for DSS. For all endpoints, the hazard was lower for the luminal A group.

When HER2-enriched and basal subpopulations were also considered in the comparison, the p-value of the Gehan-Wilcoxon test remained lower than that of the log-rank test for all considered endpoints but OS (Supplementary Table 8.1). Statistically significant differences between the survival functions of all subtypes were detected based on only the Gehan-Wilcoxon test for DSS and PFI. In KM graphs (Supplementary Figure 8.1A and B), it can be noticed that overall and disease-specific survival is distinctly worse for HER2-enriched and basal patients than luminal cases. However, for PHI and DFI recommended for breast cancer, luminal A subpopulations showed even poorer survival outcomes than HER2-enriched and basal subtypes, especially in the early phases following the initial diagnosis (Supplementary Figure 8.1C and D).

In Cox regression analysis performed on all DiviK-based subtypes, only small or even neglectable effects were shown for all subtypes referred to basals, apart from

the luminal A3 (Supplementary Table 8.2). For this subpopulation, the effect size was large per each endpoint but OS, indicating improved prognosis. Moreover, for PFI, the hazard decrease in luminal A3 subpopulation compared to basals was statistically significant. Generally, HR values suggest an increased risk for HER2-enriched subpopulation in reference to basals, regardless of the endpoint type. The increased risk was also observed for the luminal A2 subtype regarding DFI. Interestingly, in terms of OS, only the luminal A3 subtype showed a lower risk than basal, which is a rather unexpected result, given the medical knowledge according to which basal tumors are considered aggressive and associated with poor clinical outcomes. This seems to be another reason to doubt OS reliability and treat this endpoint with little confidence.

For comparison of all PAM50 subtypes, Gehan-Wilcoxon p-values were lower than their log-rank counterparts for all considered endpoints. This indicated the early changes in survival outcomes, which were proved to be statistically significant for OS, DSS, and PFI (Supplementary Table 8.1). The OS survival function was distinctly lower for the HER2-enriched subtype than for all remaining ones (Supplementary Figure 8.2A). KM graphs showed a worse prognosis for HER2-enriched and basal subtypes than for both luminal ones (Supplementary Figure 8.2). Moreover, the luminal A subtype seems to have the best survival experience, especially regarding OS and DSS (Supplementary Figure 8.2A and B). This, however, needs to be regarded with some uncertainty as OS and DSS are not recommended for breast cancer due to the short follow-up period (Liu, et al., 2018). Cox proportional hazard analysis for all PAM50 subtypes showed only small or neglectable effects for DSS, DFI, and PFI. The medium effect was detected only in terms of OS for HER2-enriched and luminal B subtypes, indicating increased risk compared to basal tumors (Supplementary Table 8.3).

5.2 Subpopulation demographic and clinical profile

5.2.1 Categorical variable analysis

Table 5.6 summarizes categorical demographic and clinical factors which association with breast cancer subtypes was tested. The subsets of subtypes

considered for the dependency analysis included: all subtypes identified with DiviK or PAM50, luminal subtypes coming from those two approaches, and three luminal A subpopulations identified with DiviK. Dimensions of contingency tables generated for statistical testing depended on the subtype list and categorical variables.

Table 5.6 Summary of demographic and clinical categorical data

Feature	Number of patients with records	Percentage of available records	Number of categories
Race	352	86.49%	4/3*
Ethnicity	307	75.43%	2
AJCC Cancer Stage	403	99.02%	4
AJCC T	405	99.51%	4
AJCC T binarized	405	99.51%	2
AJCC N	406	99.75%	4
AJCC N binarized	406	99.75%	2
AJCC M	405	99.51%	2

* The number of race categories was equal to 4 for all subtypes considered and 3 for luminal or luminal A subpopulations only.

Results of the association analysis are presented in Table 5.7 for all subtypes, in Table 5.8 for the subset of luminal subtypes, and for three luminal A subpopulations detected with the DiviK approach in Table 5.9. The tables show test statistics and p-values of Pearson χ^2 test of independence and Cramér's V effect with thresholds for interpretation adjusted for the size of corresponding contingency tables. Table cells with Cramér's V values are colored based on the effect size interpretation. Cramér's V results for binarized AJCC pathologic fields are referred to the thresholds in Figure 5.3, with the color and shape of data points corresponding to the subset of subtypes and method of subtype identification, respectively. Furthermore, Figure 5.4 shows Cramér's V values for the remaining AJCC staging features, also with regard to the subtyping approach and subset of subtypes tested. It is worth noting that although all those variables have four categories, the thresholds for Cramér's V interpretation vary as in some cases (e.g., for PAM50 luminal subtypes) the number of groups is smaller than the number of categories.

Table 5.7 Association between categorical demographic and clinical factors and all subtypes identified with DiviK or based on PAM50 classifier

Test statistics and p-value from Pearson's χ^2 test of independence, Cramér's V effect size of the association, and small, medium, and large effect thresholds adjusted for the number of categories.

Feature	χ^2	p-value	Cramér's V	Cramér's V effect threshold		
				Small	Medium	Large
Subpopulations identified with DiviK						
Race	29.13	0.0155	0.1661	0.0577	0.1732	0.2887
Ethnicity	1.14	0.9508	0.0609	0.1	0.3	0.5
AJCC Stage	29.32	0.0146	0.1557	0.0577	0.1732	0.2887
AJCC Tumor	29.33	0.0146	0.1554			
AJCC Node	28.79	0.0171	0.1538			
AJCC Tumor Binarized	18.72	0.0022	0.2150	0.1	0.3	0.5
AJCC Node Binarized	13.09	0.0225	0.1796			
AJCC Metastasis	2.58	0.7649	0.0798			
PAM50-based subtypes						
Race	34.68	0.0001	0.1812	0.0577	0.1732	0.2887
Ethnicity	4.09	0.2514	0.1155	0.1	0.3	0.5
AJCC Stage	20.13	0.0172	0.1290	0.0577	0.1732	0.2887
AJCC Tumor	24.28	0.0039	0.1414			
AJCC Node	13.42	0.1447	0.1049			
AJCC Tumor Binarized	19.58	0.0002	0.2199	0.1	0.3	0.5
AJCC Node Binarized	8.28	0.0406	0.1428			
AJCC Metastasis	2.78	0.4263	0.0829			

Table 5.8 Association between categorical demographic and clinical factors and luminal subtypes identified with DiviK or based on PAM50 classifier

Test statistics and p-value from Pearson's χ^2 test of independence, Cramér's V effect size of the association, and small, medium, and large effect thresholds adjusted for the number of categories.

Feature	χ^2	p-value	Cramér's V	Cramér's V effect threshold		
				Small	Medium	Large
Subpopulations identified with DiviK						
Race	13.42	0.0368	0.1712	0.0707	0.2121	0.3536
Ethnicity	0.23	0.9718	0.0346	0.1	0.3	0.5
AJCC Stage	18.61	0.0287	0.1536	0.0577	0.1732	0.2887
AJCC Tumor	19.34	0.0225	0.1566			
AJCC Node	13.23	0.1526	0.1292			
AJCC Tumor Binarized	13.86	0.0031	0.2295	0.1	0.3	0.5
AJCC Node Binarized	3.75	0.2900	0.1191			
AJCC Metastasis	2.23	0.5254	0.0922			
PAM50-based subtypes						
Race	3.74	0.1543	0.1269	0.1	0.3	0.5
Ethnicity	1.26	0.2610	0.0793			
AJCC Stage	9.19	0.0269	0.1848			
AJCC Tumor	14.40	0.0024	0.2309			
AJCC Node	0.91	0.8228	0.0580			
AJCC Tumor Binarized	13.25	0.0003	0.2215			
AJCC Node Binarized	0.67	0.4133	0.0497			
AJCC Metastasis	1.42	0.2335	0.0725			

Table 5.9 Association between categorical demographic and clinical factors and luminal A subtypes identified with DiviK

Test statistics and p-value from Pearson's χ^2 test of independence, Cramér's V effect size of the association, and small, medium, and large effect thresholds adjusted for the number of categories.

Feature	$\chi 2$	p-value	Cramér's V	Cramér's V effect threshold		
				Small	Medium	Large
Race	6.16	0.1879	0.1358	0.0707	0.2121	0.3536
Ethnicity	0.22	0.8959	0.0392	0.1	0.3	0.5
AJCC Stage	10.09	0.1211	0.1625	0.0707	0.2121	0.3536
AJCC Tumor	2.41	0.8779	0.0795			
AJCC Node	10.77	0.0957	0.1675			
AJCC Tumor Binarized	0.79	0.6731	0.0644	0.1	0.3	0.5
AJCC Node Binarized	2.36	0.3074	0.1109			
AJCC Metastasis	2.09	0.3513	0.1047			

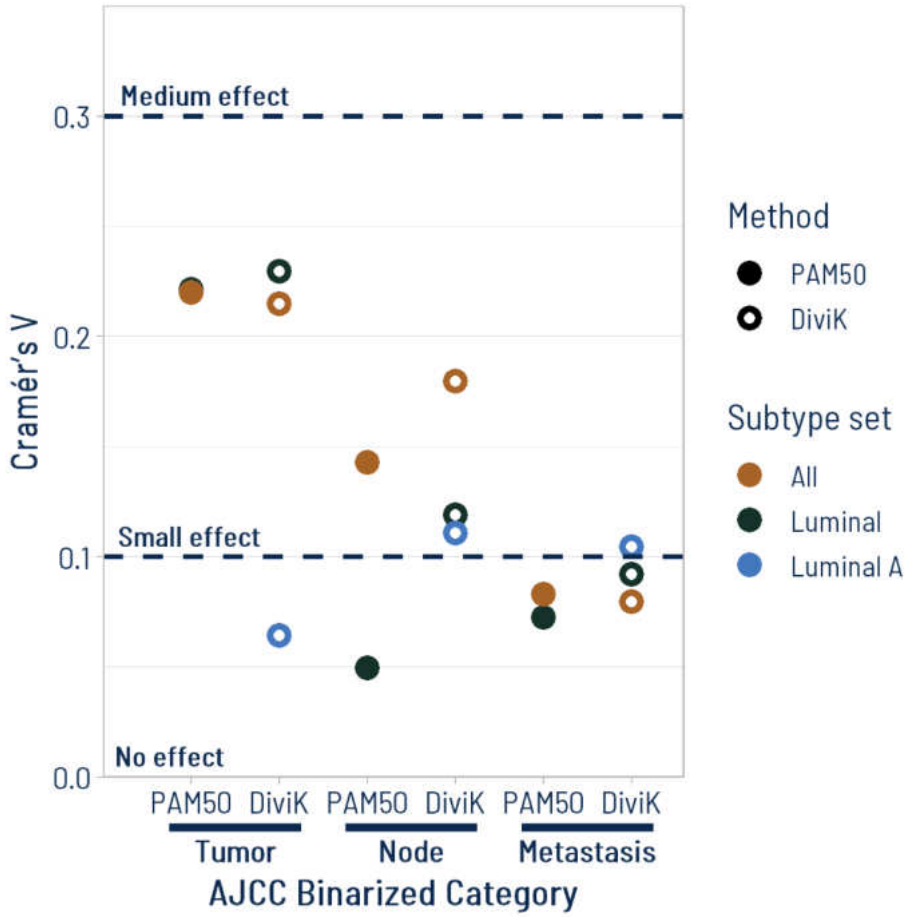


Figure 5.3 Cramér's V results for binarized AJCC pathologic fields

The color of data points corresponds to the subset of subtypes, while the shape marks the subtyping approach.

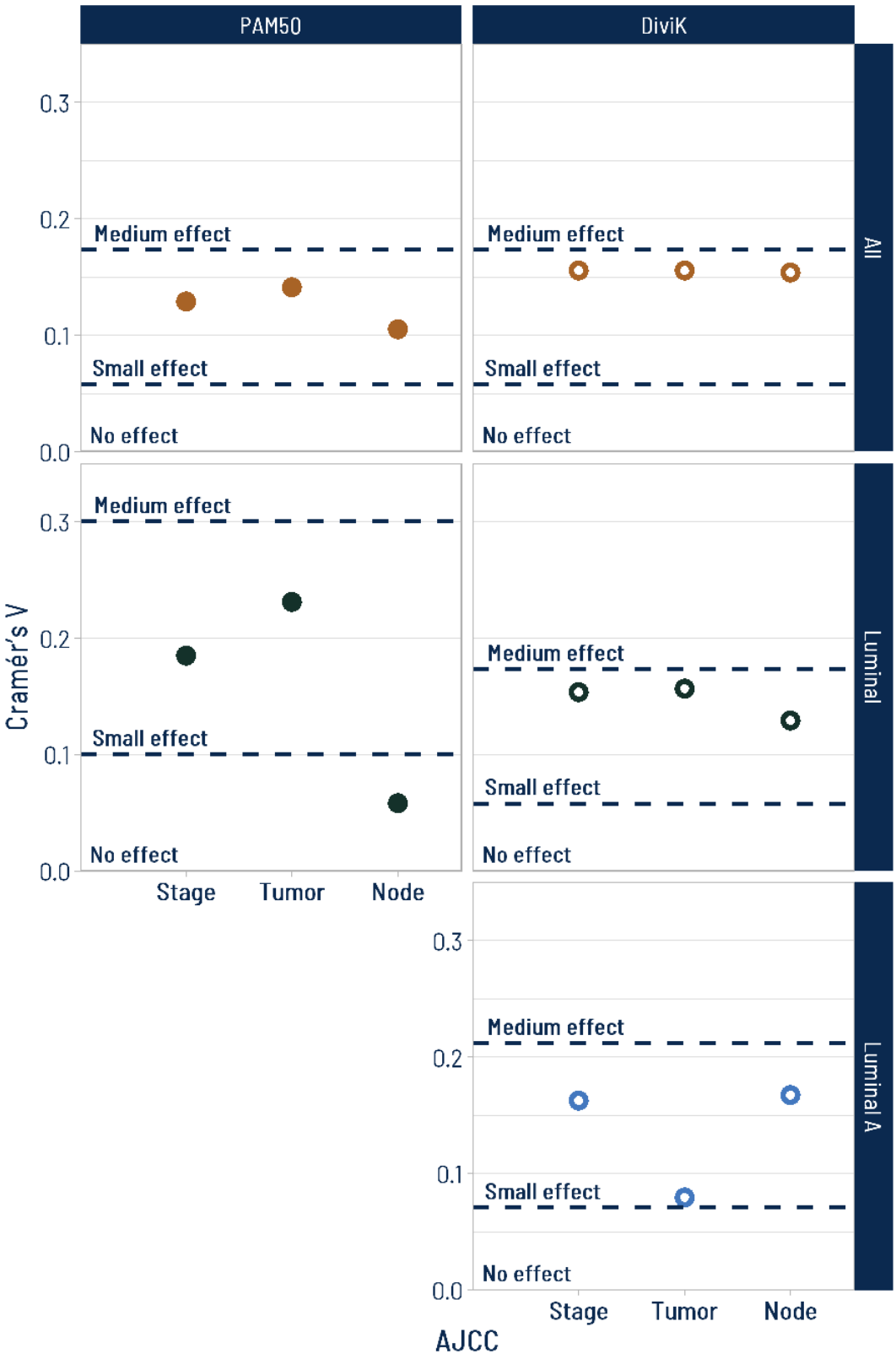


Figure 5.4 Cramér's V results for non-binary AJCC pathologic fields

The color of data points corresponds to the subset of subtypes, while the shape marks the subtyping approach. Thresholds for Cramér's V interpretation depend on the contingency table dimensions.

The results indicate small but statistically significant association between DiviK-based subtypes considered together and all categorical factors, apart from ethnicity and metastasis, for which the effect was negligible. A similar dependency was shown for PAM50 subtypes. However, a small association with ethnicity and even moderate with race was detected for this approach. For luminal cases, the effect was also small regarding all factors but ethnicity and metastasis. Nonetheless, for the AJCC node fields, no significant dependency was shown by the Pearson χ^2 test. The effect was also negligible for PAM50 subtypes. Furthermore, no significant dependency between categorical factors and luminal A subpopulations identified with DiviK was found with the Pearson χ^2 test. However, a small association effect was observed for all factors, apart from ethnicity and binarized tumor size.

5.2.2 Numerical variable analysis

The ANOVA test was used to compare the patient's age at the time of diagnosis, as the assumptions for population normality and homogeneity of variances were fulfilled. There were only two luminal PAM50 categories, so in that case, t-test and Cohen's d replaced ANOVA and η^2. Table 5.10 contains results of testing for age differences between the subtypes identified with DiviK or the PAM50 predictor.

Table 5.10 Comparison of age at diagnosis between the subtypes

(*) indicates that since there were only two luminal PAM50 subtypes, age was compared with the t-test and Cohen's d effect size. The number of groups was bigger for the remaining comparisons, so ANOVA and η^2 effect size were used.

Subtype subset	DiviK-based subtypes		PAM50-based subtypes	
	p-value	Effect size	p-value	Effect size
All	0.0011	0.050	0.0146	0.026
Luminal	0.0231	0.037	0.6334*	0.001 0.061*
Luminal A	0.0096	0.049	-	-

Based on the ANOVA p-values, the subtypes in all tested variants differed significantly in age. Nonetheless, the effect was small. The exemption was comparing PAM50 luminal A and B cases, for which no significant differences were

detected, and Cohen's *d* effect size was classified as very small. The boxplots of age at diagnosis in subpopulations identified with DiviK are shown in Figure 5.5.

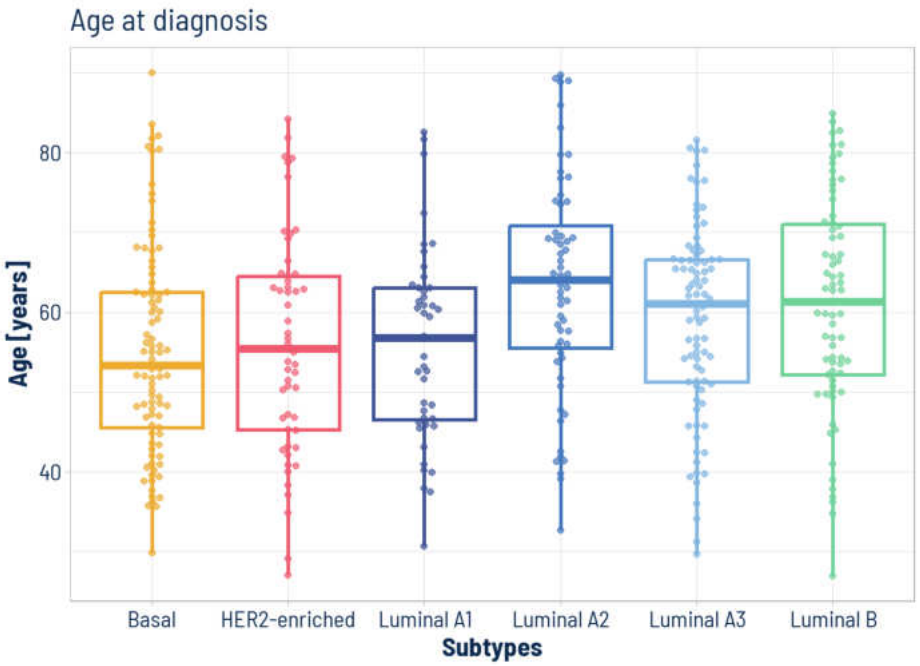


Figure 5.5 Age at diagnosis boxplots for patient subpopulations identified with DiviK

The median age at diagnosis for the luminal A2 subpopulation was the highest. Conover post hoc tests indicated that patients in this subpopulation were significantly older than those with luminal A1, basal, and HER2-enriched tumors. The effect was medium for those comparisons, while for the remaining ones was small or negligible. For PAM50 subtypes, patients with basal tumors were significantly younger during diagnosis referred to luminal A or B cases. However, no other significant differences were identified, and all effect sizes were small or negligible.

CIBERSORT immune cellular fraction estimates for 20 of 22 cell types were tested for differentiation between all subpopulations identified with DiviK or only luminal subtypes. Two cell types (eosinophils and naïve CD4 T cells) were rejected since their estimated fractions equaled 0 for almost all samples. Since the normality assumption was not fulfilled for most cell types, a non-parametric approach served for testing with FDR calculated using the Benjamini-Hochberg method (Benjamini & Hochberg, 1995). Table 5.11 contains the results of Kruskal-Wallis tests for comparison of all subtypes and only luminal ones.

Table 5.11 Kruskal-Wallis test and η^2 results for immune cellular fractions

Cell type	All subtypes			Luminal subtypes		
	p	FDR	η^2	p	FDR	η^2
B cells naive	$<10^{-4}$	0.0001	0.052	$<10^{-4}$	0.0005	0.073
B cells memory	0.0006	0.0013	0.037	0.0002	0.0018	0.057
Plasma cells	0.0014	0.0028	0.032	0.0030	0.0143	0.034
T cells CD8	0.7495	0.7890	-0.011	0.6981	0.7348	-0.014
T cells CD4 memory resting	0.0053	0.0089	0.024	0.0197	0.0439	0.019
T cells CD4 memory activated	$<10^{-4}$	0.0001	0.052	0.2472	0.3167	-0.003
T cells follicular helper	$<10^{-4}$	$<10^{-4}$	0.118	0.0428	0.0856	0.012
T cells regulatory Tregs	0.0700	0.0842	0.008	0.0825	0.1270	0.006
T cells gamma delta	0.5728	0.6364	-0.008	0.5351	0.6296	-0.011
NK cells resting	0.0584	0.0834	0.009	0.9149	0.9149	-0.017
NK cells activated	0.8224	0.8224	-0.012	0.6156	0.6840	-0.012
Monocytes	0.0017	0.0031	0.031	0.0029	0.0143	0.035
Macrophages M0	$<10^{-4}$	$<10^{-4}$	0.071	0.0036	0.0143	0.033
Macrophages M1	$<10^{-4}$	$<10^{-4}$	0.094	0.0119	0.0298	0.023
Macrophages M2	$<10^{-4}$	$<10^{-4}$	0.073	0.0099	0.0298	0.025
Dendritic cells resting	0.0292	0.0449	0.014	0.0116	0.0298	0.023
Dendritic cells activated	0.0002	0.0006	0.042	0.2533	0.3167	-0.004
Mast cells resting	$<10^{-4}$	$<10^{-4}$	0.303	0.0825	0.1270	0.006
Mast cells activated	0.0648	0.0842	0.008	0.1366	0.1951	0.002
Neutrophils	0.0716	0.0842	0.008	0.0598	0.1086	0.009

The fractions varied significantly among all subtypes for 13 immune cell types: naïve and memory B cells, plasma cells, activated and resting memory CD4 T cells, follicular helper T cells, monocytes, macrophages M0, M1, and M2, resting and activated dendritic cells, and resting mast cells. Conover post hoc tests supported by plots indicated an elevated fraction of follicular helper T cells and lack of resting mast cells in basal tumors, significantly distinguishing this subtype from others. Visual inspection revealed a relatively small number of non-zero records for memory B cells, activated T cells, and dendritic cells.

The fraction significantly varied for the subset of luminal subtypes for nine immune cell types: naïve and memory B cells, plasma cells, resting memory

CD4 T cells, monocytes, macrophages M0, M1, and M2, and resting dendritic cells. According to the Conover post hoc test results, the main significant differences were detected for the luminal A2 subtype referred to others. The fraction of naïve B cells was significantly higher with medium effect in luminal A2 compared to A1 and A3 and in luminal A3 compared to B. The highest number of non-zero records was observed for the luminal A2 subtype. Moreover, plasma cell fraction was significantly lower in the luminal A2 subtype than in luminal A3 and B, with a medium effect. Interestingly, compared to other luminal subtypes, luminal A2 fractions of macrophages M1 and M2 were relatively small and big, respectively. For macrophages M1, the effect was medium in all those pairs, while for M2, only if luminals A1 and A3 were compared.

The selection of differentiating cellular fractions is visualized in Figure 5.6 (resting memory CD4+ T cells in Panels A1-2 and follicular helper T cells in Panels B1-2), Figure 5.7 (macrophages M1 in Panels A1-2 and M2 in Panels B1-2), and Figure 5.8 (resting mast cells in Panels A1-2 and naïve B cells in Panels B1-2).

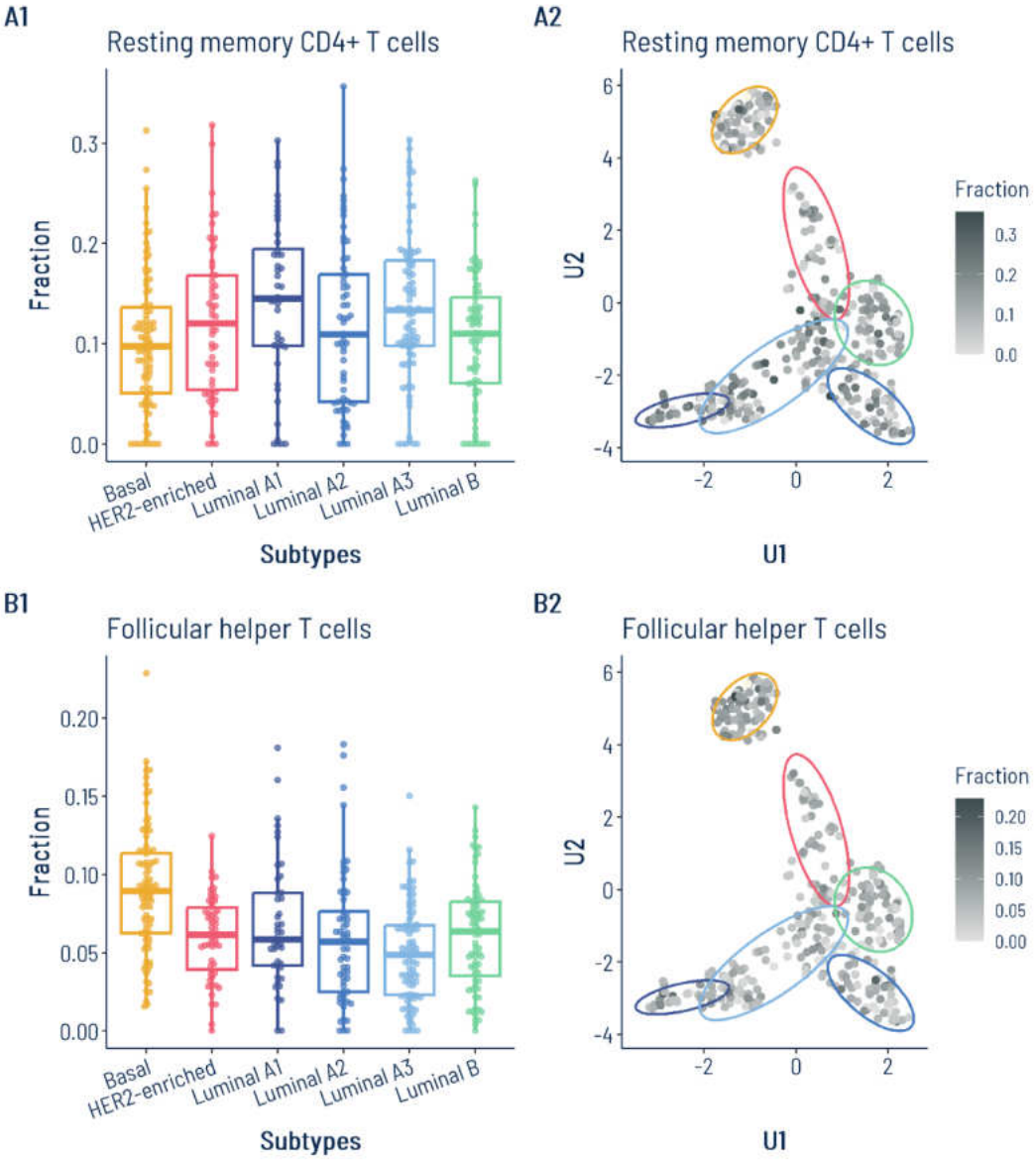


Figure 5.6 Fractions of resting memory CD4+ and follicular helper T cells with regard to subpopulations identified with DiviK

Panels A1 and B1 show boxplots of cellular fractions per subtype for resting memory CD4+ T cells and follicular helper T cells, respectively. Panels A2 and B2 show the UMAP projection obtained in Chapter 4 based on the protein level data set with the color of data points reflecting the fraction of resting memory CD4+ T cells and follicular helper T cells, respectively.

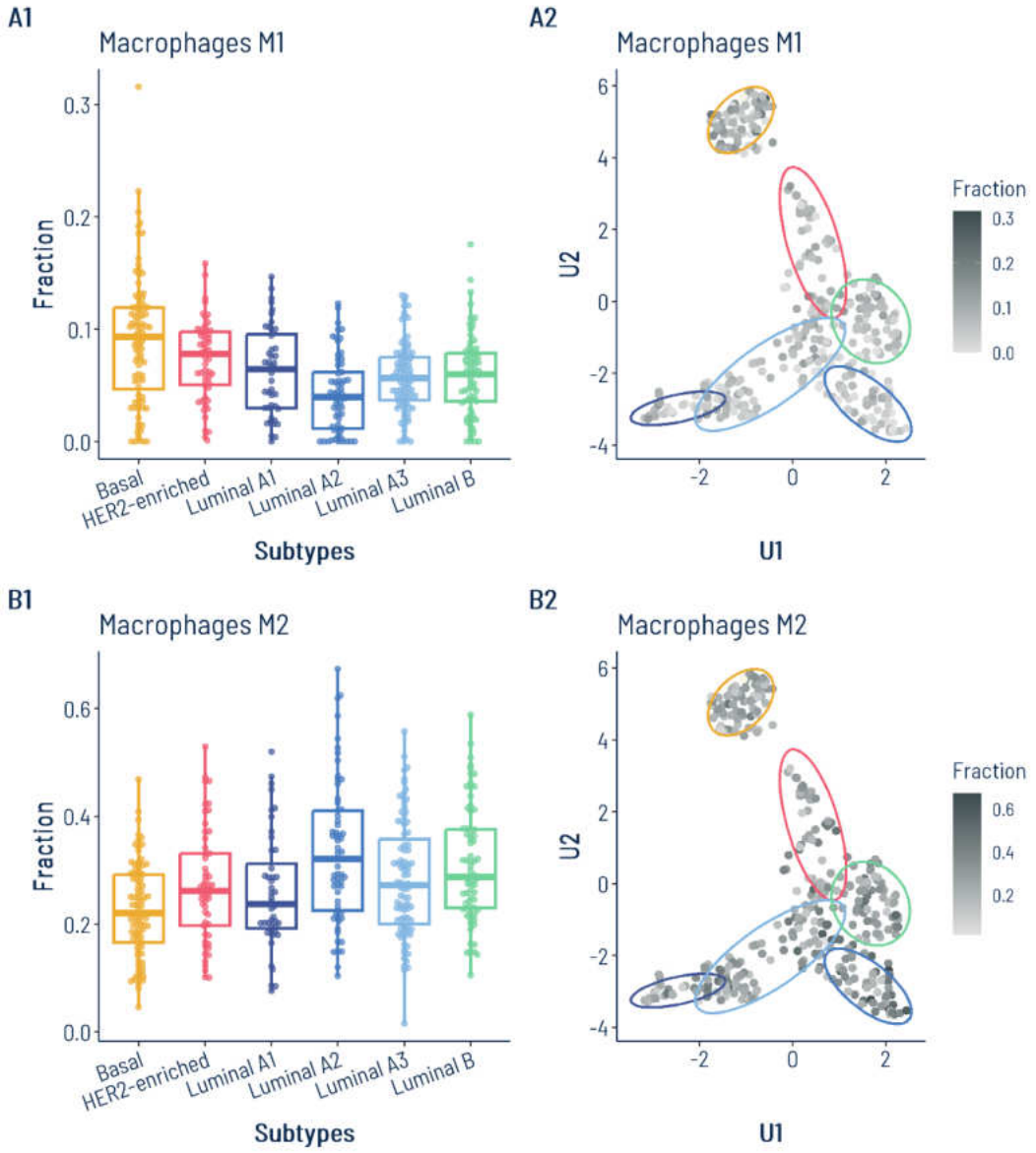


Figure 5.7 Fractions of macrophages M1 and M2 with regard to subpopulations identified with DiviK

Panels A1 and B1 show boxplots of cellular fractions per subtype for macrophages M1 and M2, respectively. Panels A2 and B2 show the UMAP projection obtained in Chapter 4 based on the protein level data set with the color of data points reflecting the fraction of macrophages M1 and M2, respectively.

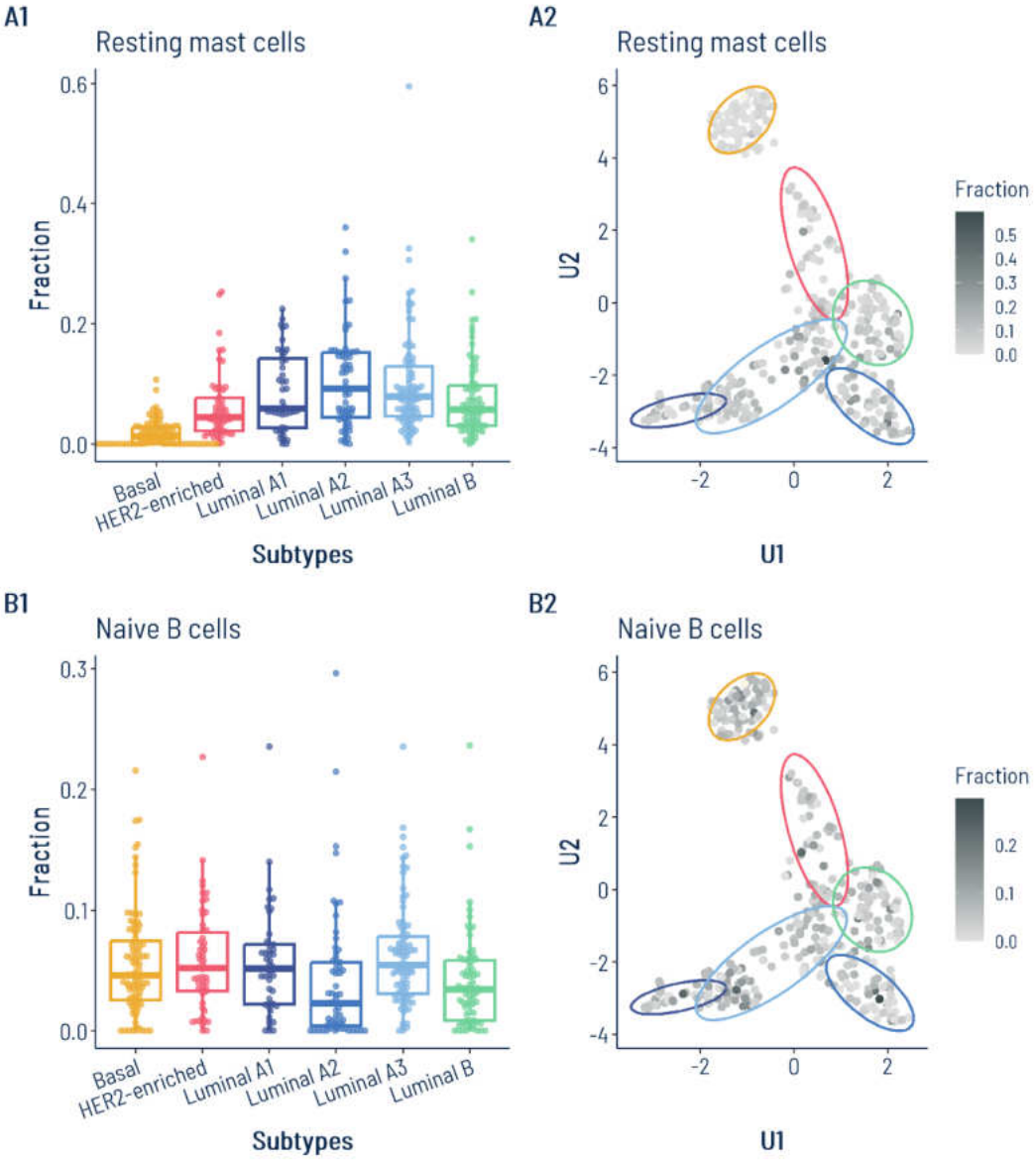


Figure 5.8 Fractions of resting mast cells and naïve B cells with regard to subpopulations identified with DiviK

Panels A1 and B1 show boxplots of cellular fractions per subtype for resting mast cells and naïve B cells, respectively. Panels A2 and B2 show the UMAP projection obtained in Chapter 4 based on the protein level data set with the color of data points reflecting the fraction of resting mast cells and naïve B cells, respectively.

5.3 Conclusions and discussion

Insufficient follow-up timespan remains a problem in survival analysis of the TCGA-BRCA cohort. As expected for less aggressive tumors like breast cancer, too few events were captured to provide statistical strength in the survival study, especially regarding the Cox proportional hazard regression. As indicated in (Liu, et al., 2018), too short follow-up period may be especially insufficient

for ER+ breast tumors, like luminal ones, which are known to have a better clinical outcome than ER- ones. ER+ cases are most of the TCGA-BRCA cohort and, consequently, the majority of samples used in this work. Nevertheless, the proposed statistical testing pipeline allowed for comparing survival experiences between the identified subpopulations. Those methods included: visual inspection of KM graphs, the Gehan-Wilcoxon test emphasizing early differences in survival outcomes, and HR effect size with threshold adjusted for the group imbalance. Moreover, as recommended in (Liu, et al., 2018), PFI and DFI were regarded as more trustworthy during the analysis and result interpretation.

The four detected luminal subpopulations differed in terms of prognosis and survival outcome. This was not the case when comparing transcriptomics-based PAM50 subtypes luminal A and B. Hence, the improvement was observed in reference to the PAM50 predictor, for which luminal subtypes did not show any significant variety, even though they outperformed more aggressive HER2-enriched and basal tumors in terms of survival outcome.

The identified luminal A3 subpopulation achieved the best prognosis, especially regarding progression- and disease-free intervals. Luminal A1 and B subtypes were the most similar in their survival experience, which is somewhat unexpected as those two luminal subtypes were the most distant from each other in the UMAP visualization. However, the most intriguing results were obtained for the luminal A2 subpopulation, which showed the worst prognosis among all luminal subtypes. Moreover, in terms of DFI and PFI, the early survival outcome for this subpopulation was even poorer than for HER2-enriched and basal cases, which are known to be more aggressive and more likely to develop relapse in the initial years following the diagnosis and treatment. However, the DSS experience for the luminal A2 subpopulation was much better than DFI and PFI – during the first few years following diagnosis, it was as good as for the luminal A3 subtype.

The results indicate only a small association between investigated subtypes and demographic or clinical categorical factors, regardless of the subset of subtypes considered. The relationship between examined factors and breast cancer clusters

identified here based on the proteomic data remained similar to that observed concerning PAM50 subtypes. Therefore, it cannot be concluded that subpopulations proposed in this work reflect the cancer stage or patient demographic background to more extend than the original PAM50 groups. Hence, the obtained subtyping is expected to result from the molecular background and tumor biology rather than the patient's age or tumor stage during the diagnosis.

Significant differences in patient age at diagnosis were detected between all subtypes and the luminal subset. The effect size, however, remained small. The differentiation for subpopulations identified based on proteomic data in this study was similar to the one between PAM50 subtypes. In the DiviK-based subtype, the median age at diagnosis was the highest for the luminal A2 subpopulation and the lowest for the basal subtype. DiviK-based luminal A2 subpopulation was the oldest at the time of diagnosis, while for PAM50, the highest age was observed in the luminal B group. Nevertheless, regardless of the subtyping approach, the lowest median age at diagnosis was observed for basal tumors.

Results of CIBERSORT data indicated significant differences in specific immune cellular fractions for both all DiviK subpopulations and only luminal ones. Thus, additional luminal subtypes appear to vary in their immune response, which supports the decision to divide the luminal A subtype into subgroups. The luminal A2 subpopulation showed the most considerable differences in immune cell proportions referred to other luminal tumors. Interestingly, luminal A2 revealed a distinctly lower M1/M2 macrophage ratio than the remaining subpopulations, while basal subtypes the highest. M1/M2 ratio serves as an indicator of the immune response. M2 macrophages downregulate inflammation by enhancing cell proliferation, angiogenesis, and tissue repair. Contrary, M1 macrophages upregulate inflammation by inhibiting cell proliferation and inducing tissue damage (Fujiwara & Kobayashi, 2005; Wynn, Chawla, & Pollard, 2013; Wang, Liang, & Zen, 2014; Boutilier & Elsawa, 2021; Liu, Geng, Hou, & Wu, 2021).

To summarize, the subtypes show slight but considerable differences in survival outcomes and still appear not greatly affected or biased by demographic factors. The strong impact of clinical factors like cancer stage on subpopulation

composition was not detected. However, subpopulations, including four groups of luminal tumors, differ in immune cellular proportions.

6 Molecular signature of patient subpopulations

After evaluating revealed breast cancer patient subpopulations with their demographic and clinical profiles, the molecular differences between the obtained subtypes were investigated based on the protein and mRNA gene expression levels. The sets of subtype-specific markers and the signature of features distinguishing between the subtypes, but not specifically for only one of them, were identified.

6.1 Batch effect

The batch effect verification and correction in the protein level data set were described in Chapter 4.2. To reduce the dimensionality of the mRNA gene expression data set using PCA and UMAP as described in Chapter 3.2.1, missing values had to be removed. Hence, genes for which the records were not fully complete for the whole cohort were rejected. Consequently, the number of features in the data set was limited to 17328 genes from the initial 17814 genes. The UMAP projection presented in Figure 6.1 allowed for visually verifying whether samples of the same TSS, plate, or similar scan date group together.

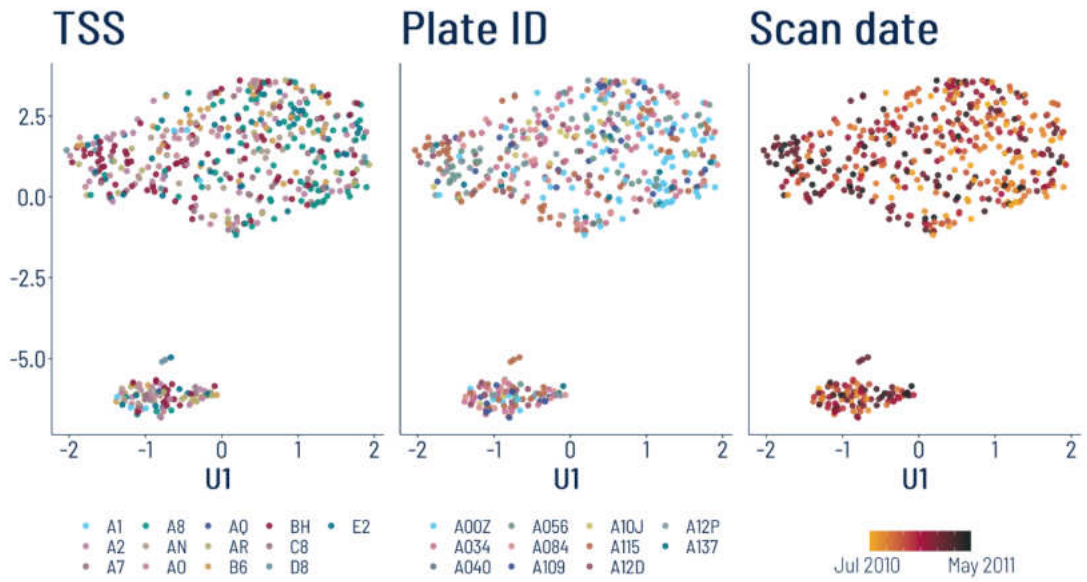


Figure 6.1 UMAP visualizations of the mRNA gene expression data set for the batch effect identification

Data points are colored according to the technical factors: tissue source site (TSS), plate identifier (ID), and scan date.

In the UMAP visualizations in Figure 6.1, no color pattern is visible. Samples do not group according to TSS, plate ID, or date of the experiment, indicating no batch effect resulting from those technical factors.

Moreover, the BatchI algorithm (Papiez, Marczyk, Polanska, & Polanski, 2018) was applied to the data sorted with the scan date to detect bias resulting from the experimental conditions changing over time. The division into two batches was selected as optimal for the tested number of batches ranging from 2 to 43, which is the number of unique experiment dates. Thus, the smallest possible solution was chosen. The resulting sample division into subranges on the timescale is presented in Figure 6.2 versus the mean signal intensity per sample, used at the quality index in the BatchI algorithm. However, the p-value for this split was equal to 0.14. This indicates that no significant batch effect was detected, which is also supported by Figure 6.2. Several gaps marking longer periods between the experiments can be noticed. Even though one of those separates the detected batches, no distinct change in the mean intensity is visible in Figure 6.2.

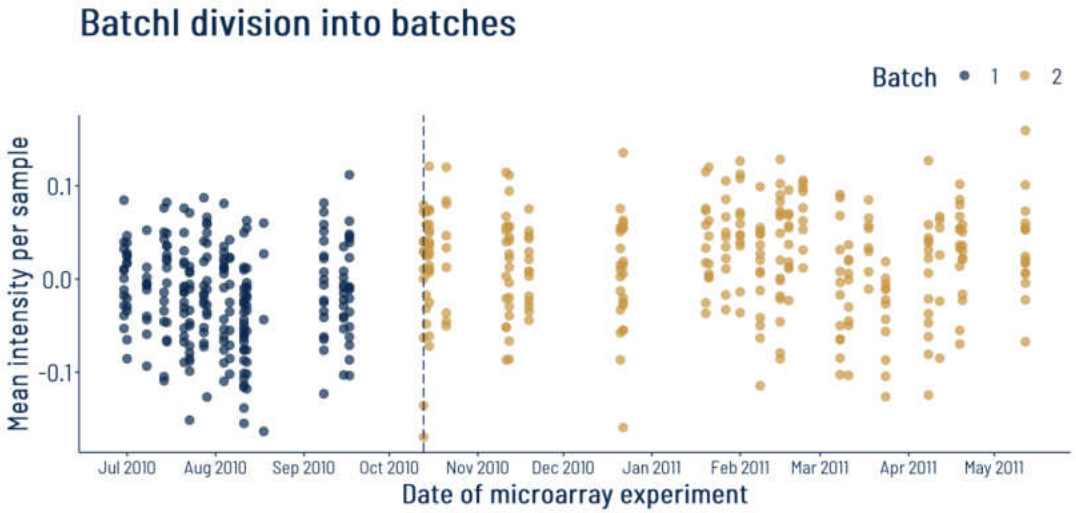


Figure 6.2 BatchI mRNA gene expression sample division into batches on the timescale

The data point color denotes the batches obtained with the BatchI algorithm (Papiez, Marczyk, Polanska, & Polanski, 2018).

Nevertheless, the proportions of samples in each batch per TSS were showed in Figure 6.3. Interestingly, the obtained batches were imbalanced in terms of TSS. This indicates that samples from certain sources arrived earlier and were examined earlier without mixing with other centers.

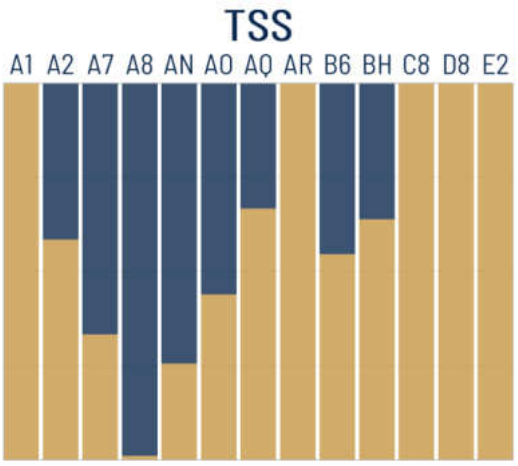


Figure 6.3 Proportion of tissue source sites (TSSs) in the batches detected with the BatchI algorithm

Color denotes the batches obtained with the BatchI algorithm (Papiez, Marczyk, Polanska, & Polanski, 2018).

To conclude, no significant batch effect was detected for the mRNA gene expression data set; thus, no correction was needed. However, the results suggest the experiment was poorly designed in terms of mixing samples from different TSSs.

6.2 Subtype-specific marker identification

Subtype-specific markers were identified in three feature spaces (proteomic, transcriptomic, or transcriptomic limited to genes coding investigated proteins), as described in Chapters 3.5.2.2 and 3.6.1. A non-parametric approach to differentiation testing was applied, as shown in Figure 3.2. since the normality assumption was not fulfilled in all compared groups for most proteins and transcripts.

As expected, smaller differentiation was observed on the transcriptomic level. Hence, for the differentiation pipeline shown in Figure 3.2, the restrictive thresholds corresponding to large and very large effect sizes were used for η^2 and Cohen's *d*, respectively, in the case of proteomics data. However, those cut-offs

were lowered to medium and large effects for mRNA gene expression levels, respectively. The threshold values are shown in Table 3.4.

Table 6.1 contains the numbers and percentages of markers non-specific for subtypes, with regard to the feature space, set of subtypes compared, and approach to differentiation verification: either based on p-values or effect sizes. For the p-values-based selection, the significance level equaled 0.05.

Table 6.1 Numbers and percentages of non-specific markers with regard to subtype set, feature space, and metrics used as a measure of differentiation

(*) denotes that the thresholds used for η^2 and Cohen's *d* effect size interpretation were lowered to medium and large, respectively, for the transcriptomic data.

Subtype set	Feature space	p-value-based	Effect-size-based	No. of all features
All	Proteomic	166(100.00%)	103 (62.05%)	166
	Transcriptomic	13870 (77.86%)	7994 (44.87%)*	17814
	Limited transcriptomic	117 (87.31%)	83 (61.94%)*	134
Luminal	Proteomic	162 (97.59%)	65 (39.16%)	166
	Transcriptomic	6975 (39.15%)	997 (5.60%)*	17814
	Limited transcriptomic	78 (58.21%)	19 (14.18%)*	134
Luminal A	Proteomic	146 (87.95%)	47 (28.31%)	166
	Transcriptomic	2805 (15.75%)	90 (0.51%)*	17814
	Limited transcriptomic	38 (28.36%)	5 (3.73%)*	134

A higher proportion of features was selected as non-specific markers in the proteomic data set than in the full transcriptomic one. According to the Kruskal-Wallis test, all proteins' levels differed significantly between all six DiviK-based subpopulations. Reducing the transcriptomic data set to RPPA-measured protein-coding genes increased the fraction of non-specific markers within the set. Nevertheless, the proportion remained much lower than in the case of protein levels. This implies that at the transcriptomic level, the great part of the differentiating effects between the revealed subtypes is not noticeable yet, contrary to the proteomic space.

Table 6.2 and Table 6.3 show the numbers of identified subtype-specific markers with the p-value-based or effect-size-based approach, respectively.

Obviously, the number of transcriptomic markers is larger due to the much bigger feature space.

Table 6.2 Numbers of subtype-specific markers selected based on p-values

"P" denotes the protein levels data set, "T" denotes the whole mRNA gene expression levels (transcriptomic) data set, and "LT" denotes the transcriptomic data set limited to genes coding the proteins included in the protein levels data set. (*) indicates the thresholds used for η^2 and Cohen's d effect size interpretation were lowered to medium and large, respectively, for the transcriptomic data.

Subtype set	All subtypes			Luminal			Luminal A		
Feature space	**P**	**T***	**LT***	**P**	**T***	**LT***	**P**	**T***	**LT***
Basal	1	1499	6	-	-	-	-	-	-
HER2-enriched	0	51	1	-	-	-	-	-	-
Luminal A1	6	0	0	19	25	1	30	140	1
Luminal A2	2	54	1	5	459	6	18	1812	25
Luminal A3	1	2	0	4	13	0	8	74	0
Luminal B	1	21	1	5	923	10	-	-	-
TOTAL	**11**	**1627**	**9**	**33**	**1420**	**17**	**56**	**2026**	**26**

Table 6.3 Numbers of subtype-specific markers selected based on effect sizes

"P" denotes the protein levels data set, "T" denotes the whole mRNA gene expression levels (transcriptomic) data set, and "LT" denotes the transcriptomic data set limited to genes coding the proteins included in the protein levels data set. (*) indicates the thresholds used for η^2 and Cohen's d effect size interpretation were lowered to medium and large, respectively, for the transcriptomic data.

Subtype set	All subtypes			Luminal			Luminal A		
Feature space	**P**	**T**	**LT**	**P**	**T**	**LT**	**P**	**T**	**LT**
Basal	1	1146	9	-	-	-	-	-	-
HER2-enriched	0	21	0	-	-	-	-	-	-
Luminal A1	5	0	0	12	0	0	19	1	0
Luminal A2	0	2	0	1	13	1	13	45	1
Luminal A3	0	0	0	0	0	0	0	0	0
Luminal B	0	0	0	1	33	2	-	-	-
TOTAL	**6**	**1169**	**9**	**14**	**46**	**3**	**32**	**46**	**1**

On the transcriptomic level, the most considerable differences were revealed for basal subpopulation, with a huge number of specific markers. Moreover, identification of HER2-enriched-specific markers was achievable only based

on mRNA gene expression levels. To select markers characteristic for luminal subtypes, basal and HER2-enriched cases were removed. Subsequently, the highest number of specific markers was found for luminal B and A2 subpopulations; the latter observation was also reinforced in the luminal A cases comparison. As can be concluded based on those results, the effect-size-based approach occurred more restrictive. The visual inspection of marker-level boxplots per subtype and the distribution of marker levels in UMAP embedding also supported this observation. Those figures are not shown here due to their large size and number. Therefore, the further analysis of results was focused on the lists of markers identified based on the effect sizes.

Identified subtype-specific markers are listed regarding the direction of level changes compared to other subtypes in Figure 6.4 for the proteomic data set, Figure 6.5 for the entire transcriptomic data set, and Figure 6.6 for the transcriptomic set reduced to luminal subpopulations.

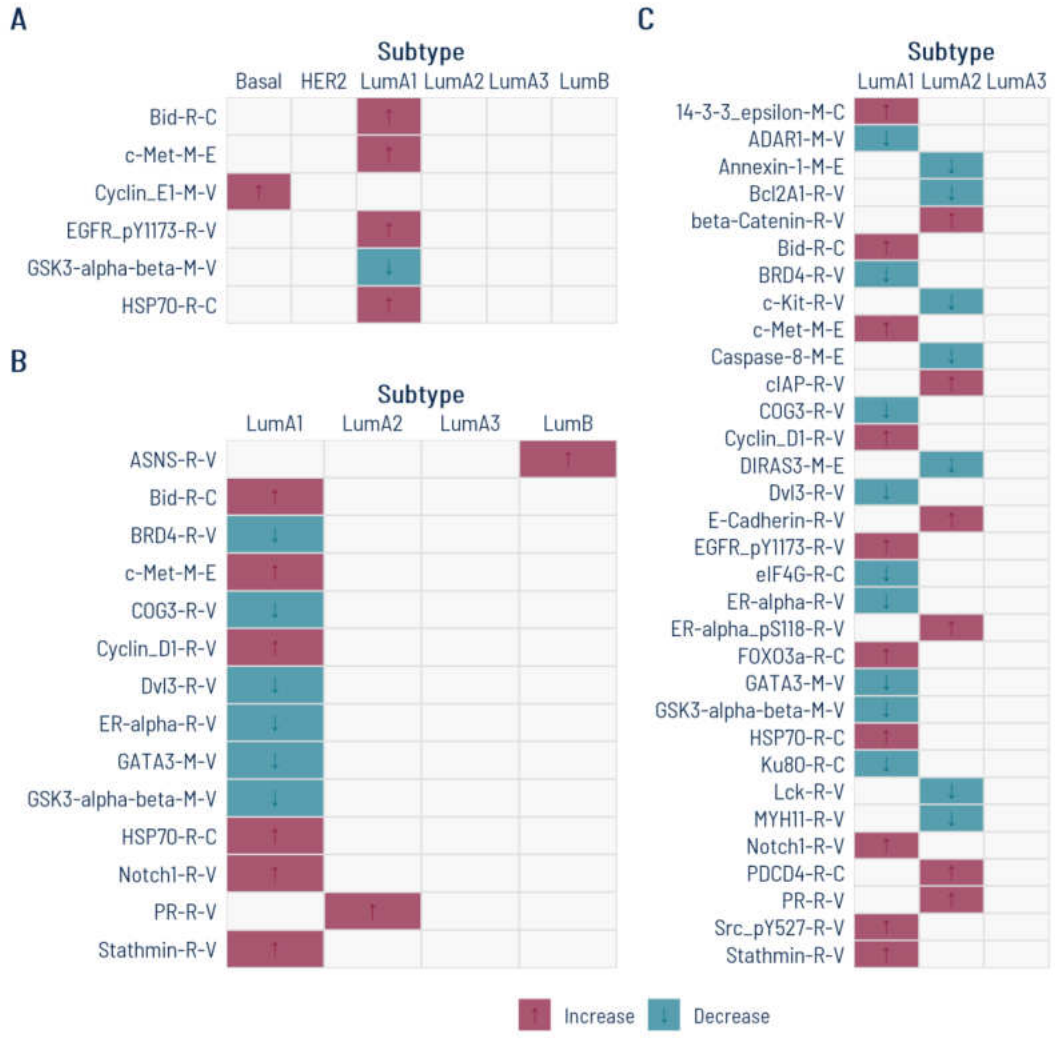


Figure 6.4 Subtype-specific markers identified based on the protein levels

Panels A, B, and C show markers selected by comparing all subtypes, luminal subtypes, and luminal A subtypes, respectively. Purple and turquoise colors indicate the marker level was respectively higher or lower for a given subtype than for all remaining ones.

When only luminal or luminal A groups were compared, the selected markers were balanced in the direction of changes. Nonetheless, when all subpopulations were considered, only one protein specific to the luminal A1 subtype revealed a lower level than in the remaining cases. Interestingly, PR were overexpressed in the luminal A2 subpopulation compared to other luminal cases.

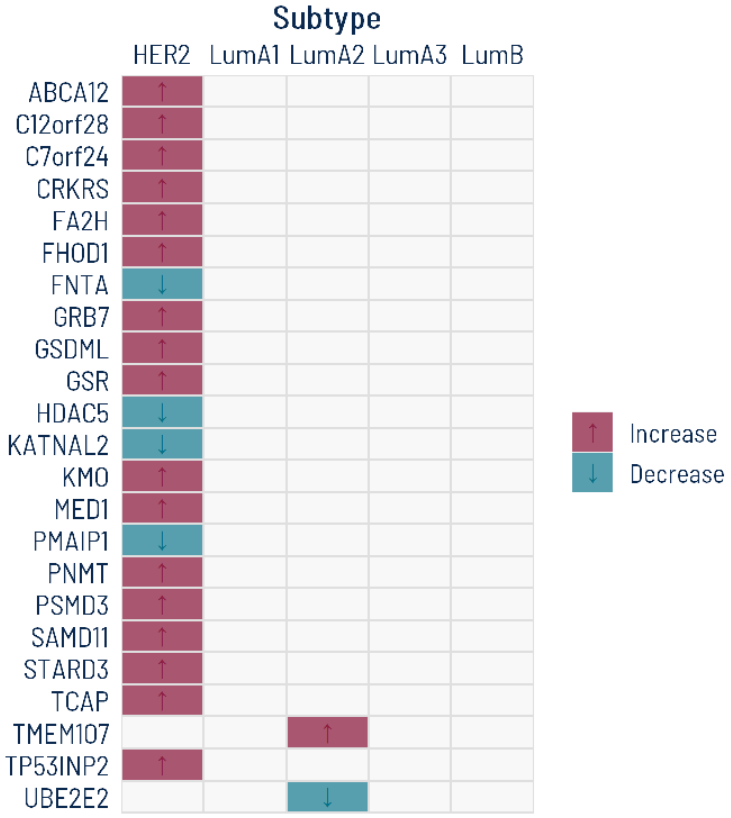


Figure 6.5 Subtype-specific markers identified based on the mRNA gene expression levels

Purple and turquoise colors indicate the marker level was respectively higher or lower for a given subtype than for all remaining ones. For clarity, basal-specific markers were not included due to their huge number.

Almost all HER2-enriched markers appeared to be overexpressed when compared to other subpopulations. Only two genes were found to be characteristically over- and under-expressed in the luminal A2 subtype; the second one (*UBE2E2*) is involved in the ubiquitin-mediated proteolysis pathway. Only markers specific for luminal A2 and B tumors were identified when all luminal subpopulations were considered. After luminal B group rejection, all but one marker were characteristic for the luminal A2 group, most under-expressed compared to other luminal A cases.

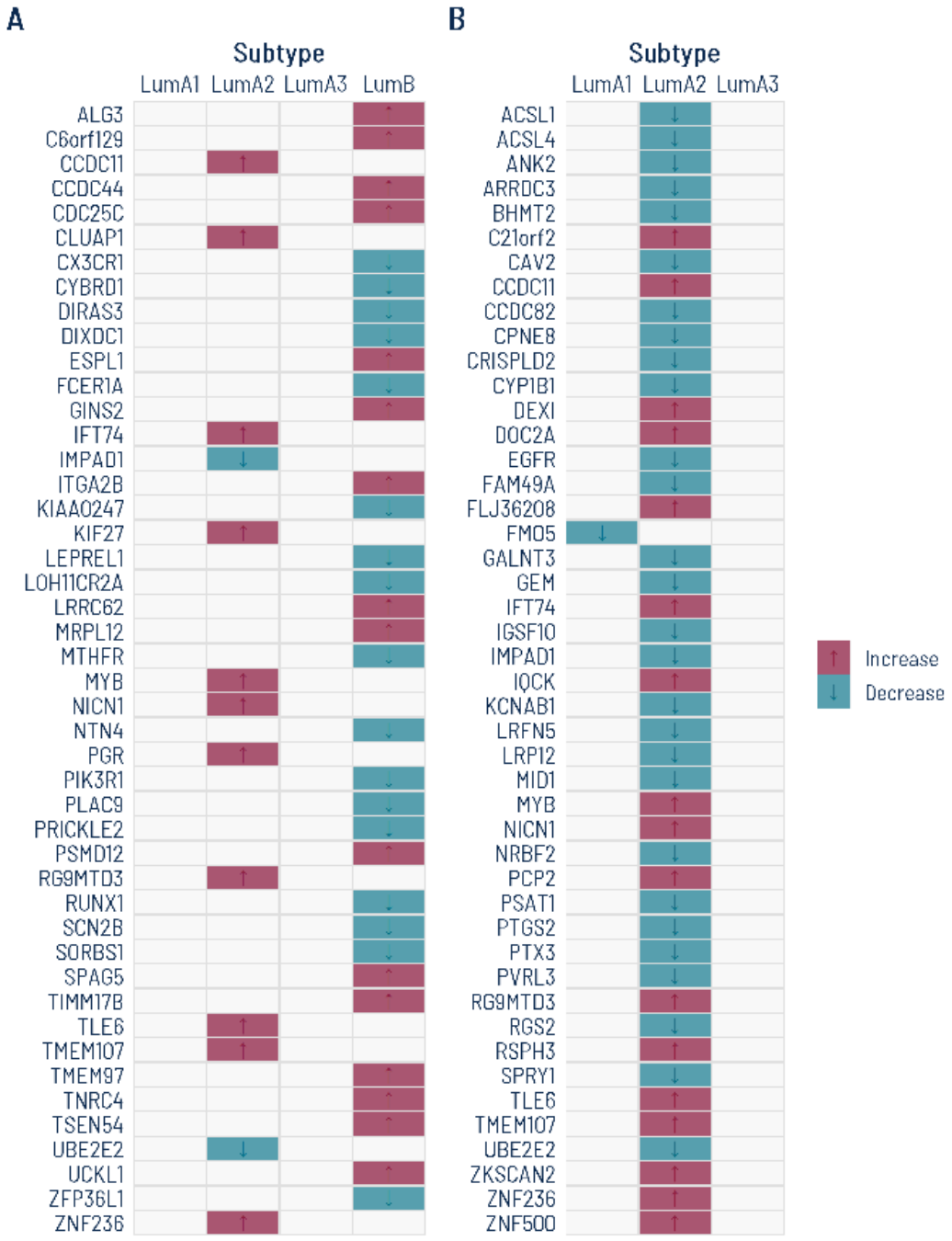


Figure 6.6 Luminal subtype-specific markers identified based on the mRNA gene expression levels

Panels A, B, and C show markers selected by comparing all subtypes, luminal subtypes, and luminal A subtypes, respectively. Purple and turquoise colors indicate the marker level was respectively higher or lower for a given subtype than for all remaining ones.

Due to the relatively small number of identified markers for both data sets and the insufficient RPPA-measured protein universe, ORA did not produce significant results for KEGG pathways following the Benjamini—Hochberg correction for multiple testing (Benjamini & Hochberg, 1995). However, for MSigDB

collections and transcriptomic feature space, many gene sets were overrepresented in the obtained lists of subtype-specific markers, especially for basal and HER2-enriched tumors. For the basal-specific transcripts, hallmark gene sets related to an early or late response to estrogen were enriched. Moreover, ORA applied on MSigDB revealed several overlaps with previously published breast cancer-related gene sets, mainly in the context of markers specific for HER2-enriched and basal subtypes (Doane, et al., 2006; Charafe-Jauffret, et al., 2005; Farmer, et al., 2005; Yang, et al., 2005; Smid, et al., 2008; van't Veer, et al., 2002).

To solve the problem of insufficient set size for ORA and further investigate the differences between four revealed luminal subpopulations, the CERNO test was applied on absolute values of *d* effect size per each luminal subtype pairwise comparison. For the proteomic data set, following the Benjamini-Hochberg correction for multiple testing (Benjamini & Hochberg, 1995), only the comparison of luminal A2 versus B subtypes provided statistically significant enrichment results. Significantly enriched KEGG pathways for that comparison are shown in Figure 6.7, referred to log-scaled FDR, and Figure 6.8, referred to both log-scaled FDR and Area Under Curve (AUC), serving as the effect measure in the CERNO test. In the obtained list of enriched pathways, several directly related to cancer biology and those involving hormone receptors can be found.

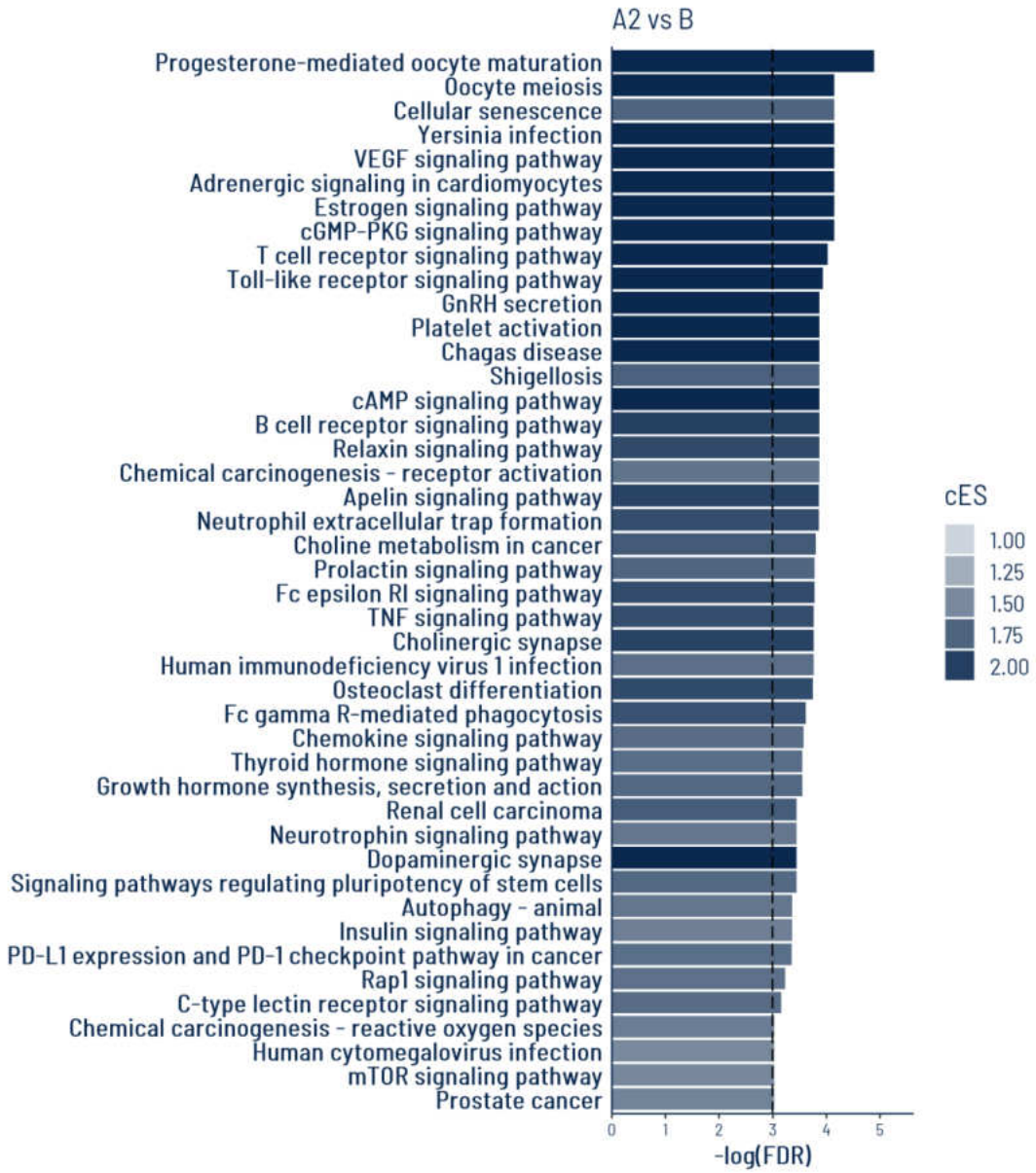


Figure 6.7 Significantly enriched KEGG pathways in comparison of luminal A2 and B subpopulations based on protein levels

cES denotes the CERNO test statistics divided by the number of genes in the module multiplied by 2. The black broken line marks FDR equal to 0.05.

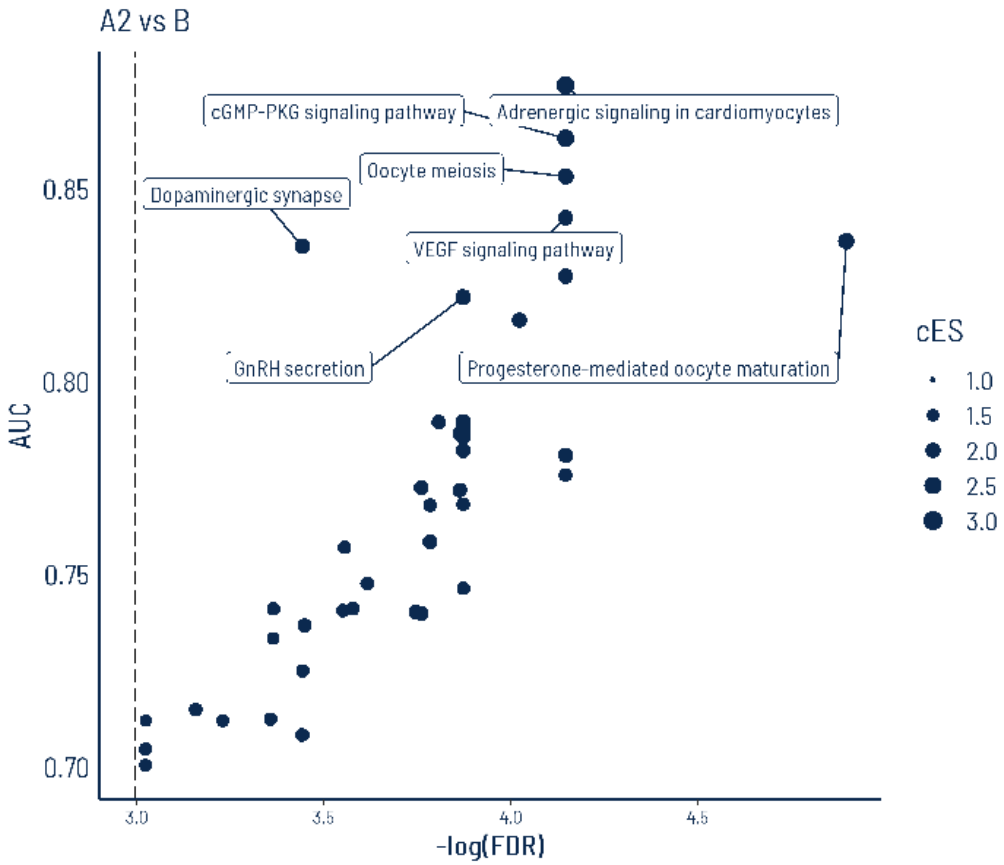


Figure 6.8 FDR and AUC values for significantly enriched KEGG pathways in comparison of luminal A2 and B subpopulations based on protein levels

cES denotes the CERNO test statistics divided by the number of genes in the module multiplied by 2. The black broken line marks FDR equal to 0.05.

All pairwise comparisons provided significantly enriched KEGG pathways for the transcriptomic data set. They are shown with their log-scaled FDRs in Figure 6.9 for luminal A2 versus other luminals and Figure 6.10 for the remaining pairs. Figure 6.11 presents those results as AUC versus FDR. As can be noticed in the figures, a smaller number of significantly enriched KEGG pathways was detected in pairwise comparisons of luminal A2 subpopulation versus other luminal tumors. The list of enriched pathways was longer when the luminal B subtype was referred to luminal A1 or A3 groups. Regardless of the comparison variant, the obtained pathways included those crucial for the proper cell functioning and many involved in tumor biology.

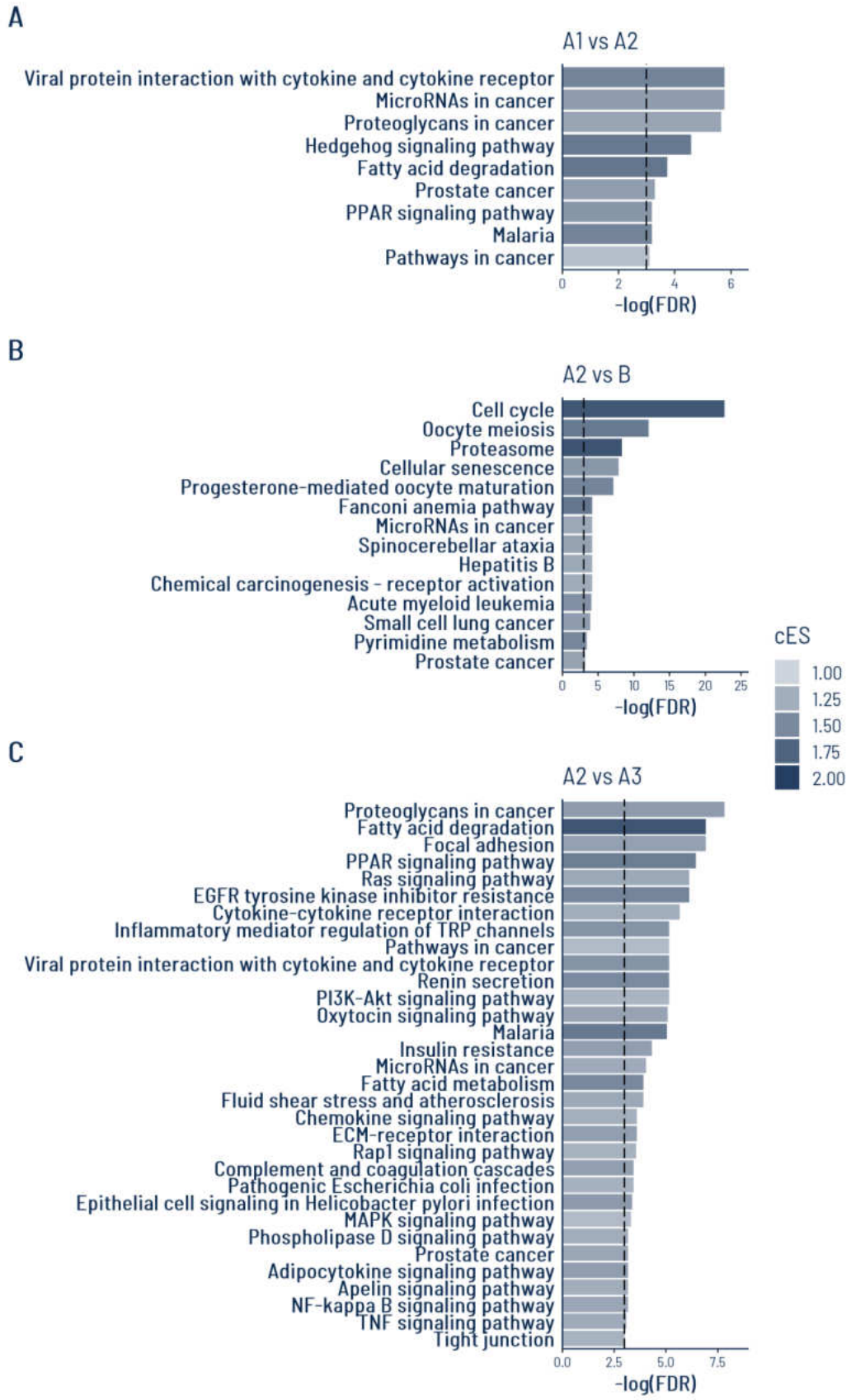


Figure 6.9 Significantly enriched KEGG pathways in comparison of luminal A2 versus other luminal subpopulations based on mRNA gene expression levels

Panels A, B, and C present results of the following comparisons of luminal subpopulations: A1 versus A2, A2 versus B, and A2 versus A3, respectively. cES denotes the CERNO test statistics divided by the number of genes in the module multiplied by 2. The black broken line marks FDR equal to 0.05.

Figure 6.10 Significantly enriched KEGG pathways in comparison of luminal B versus A1 and A3 subpopulations based on mRNA gene expression levels

Panels A and B present results of the following comparisons of luminal subpopulations: A1 versus B and A3 versus B, respectively. cES denotes the CERNO test statistics divided by the number of genes in the module multiplied by 2. The black broken line marks FDR equal to 0.05.

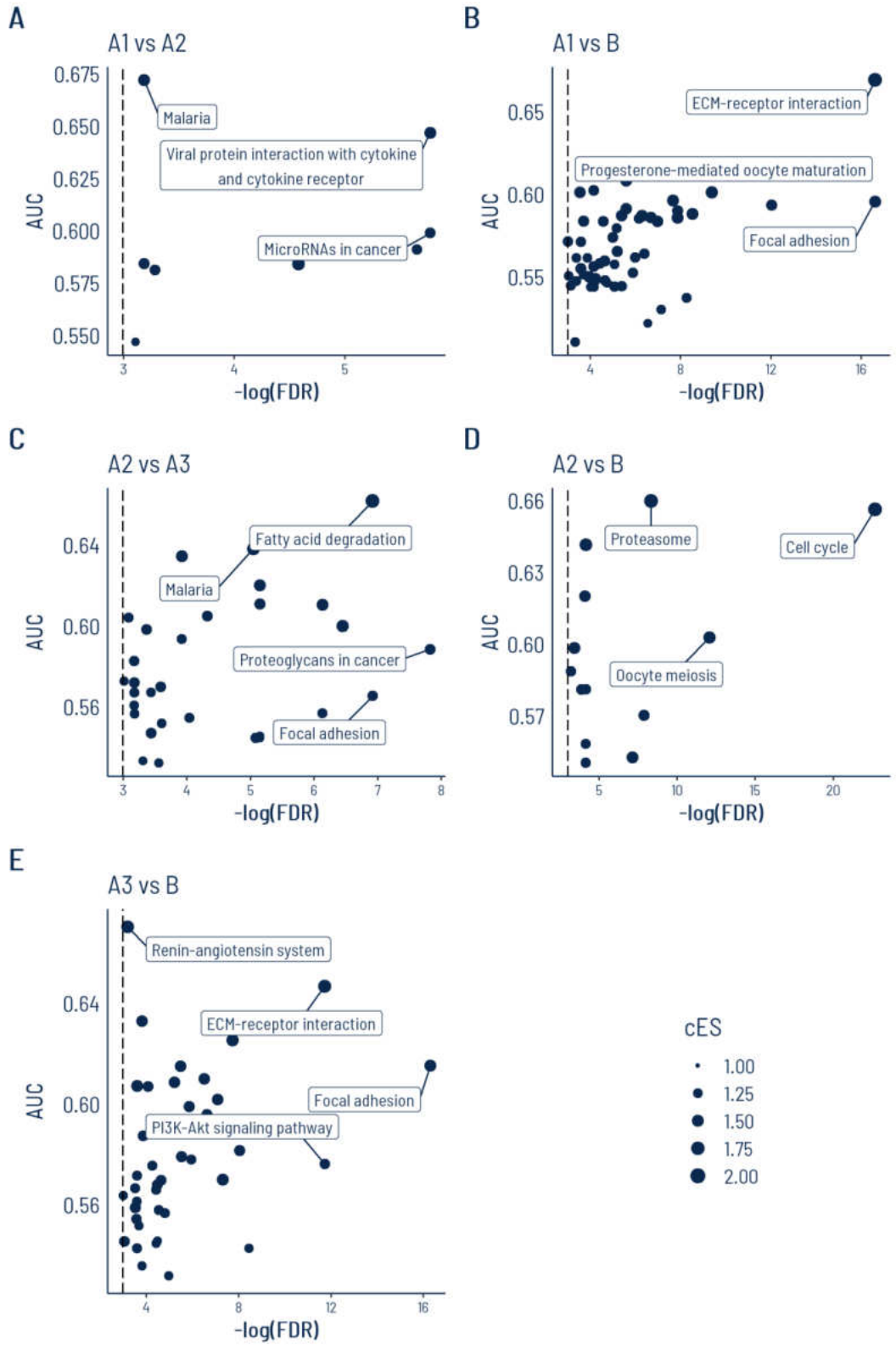


Figure 6.11 FDR and AUC values for significantly enriched KEGG pathways in pairwise comparisons of luminal subpopulations based on mRNA gene expression levels

Panels A, B, C, D, and E present results of the following comparisons of luminal subpopulations: A1 versus A2, A1 versus B, A2 versus A3, A2 versus B, and A3 versus B,

respectively. cES denotes the CERNO test statistics divided by the number of genes in the module multiplied by 2. The black broken line marks FDR equal to 0.05.

6.3 Subtype differentiating signature

A feature selection procedure for a multinomial logistic regression model was used to identify molecular signatures differentiating the subtypes. Three variants of regression models were fitted: for the proteomic data set, for the reduced transcriptomic data set, and for those two data sets combined.

6.3.1 Proteomic signature

Figure 6.12 illustrates the feature ranking scores per protein obtained in the MRCV procedure. For clarity, the plot was truncated to show only top features, without those appearing in only one out of 100 MRCV iterations. Figure 6.13 presents the elbow plot for the entire, untruncated feature ranking in Panel A and per feature shortest distances to the line in Panel B. As can be seen in the plot, the maximal distance was obtained for the tenth feature (PDCD4-R-C), so the top nine proteins were identified as the proteomic signature.

The multinomial logistic regression coefficients from a model fitted only with selected independent variables are presented in Figure 6.14. Levels of proteins included in the subtype-differentiating proteomic signature are presented in Figure 6.15 (top three proteins), Figure 6.16 (proteins 4-6), and Figure 6.17 (proteins 7-9). As seen in the plots, the selected proteins distinctly vary between identified subpopulations, not only in the whole set of subtypes but also among the luminal ones. The signature appears to be highly informative in distinguishing the subtypes based on the proteomic profile.

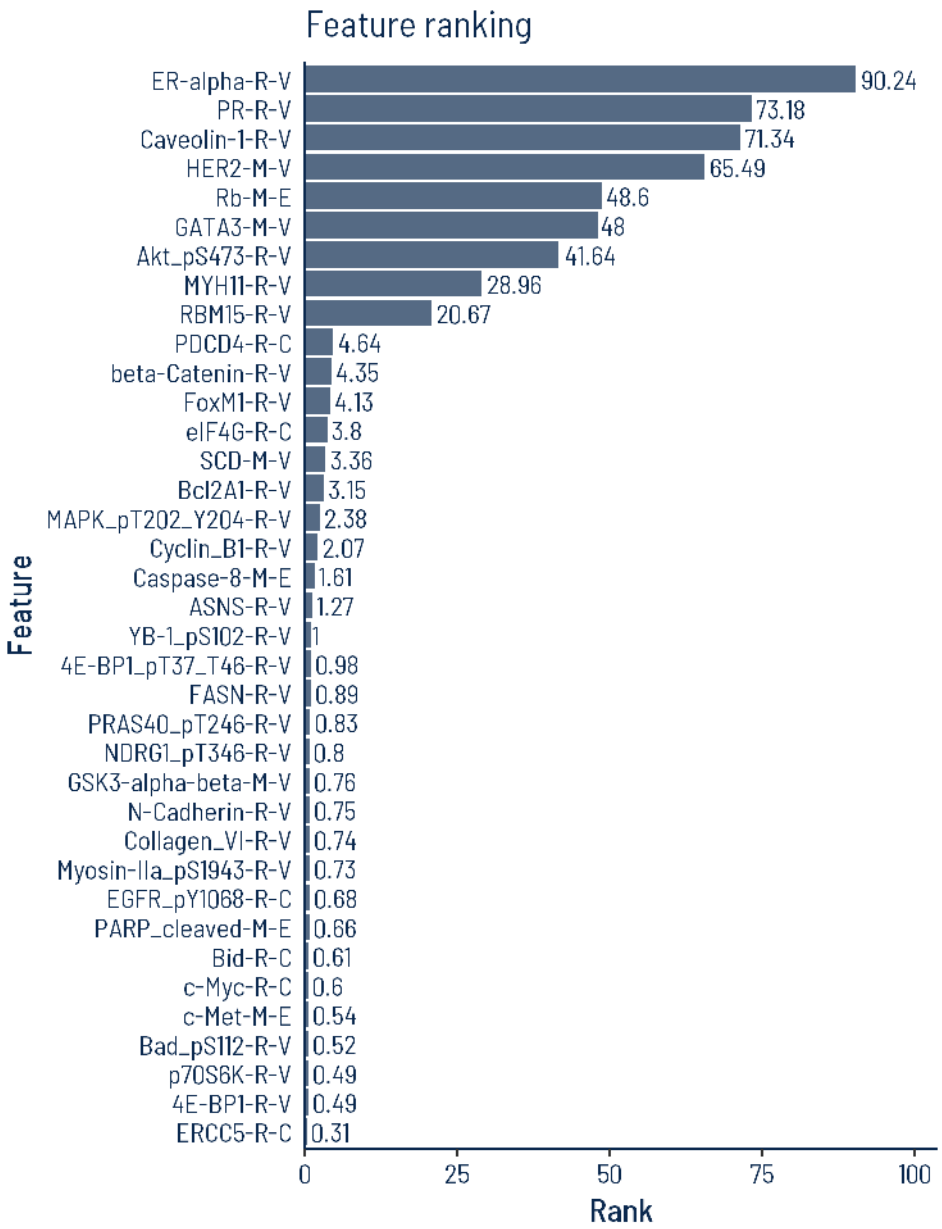


Figure 6.12 Feature ranking for the proteomic multimodal logistic regression model

For clarity, the plot was truncated to show only features selected for more than one model in the MRCV procedure.

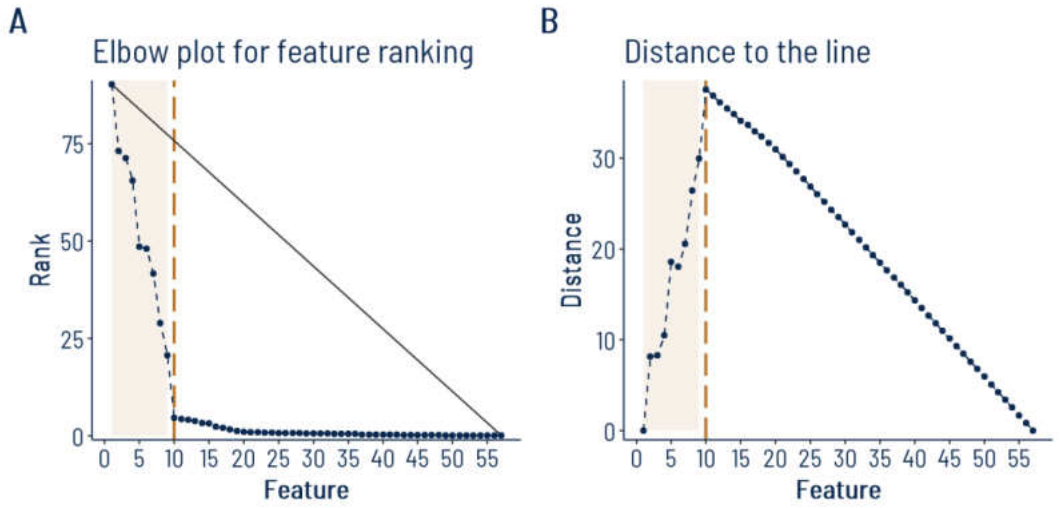


Figure 6.13 Proteomic signature identification with the elbow method

Panel A shows the elbow plot for all features selected in the MRCV procedure. Panel B shows the shortest distance between each data point and the black line joining the data points with the highest and lowest ranking score in Panel A. Brown broken line marks the feature with the highest distance serving as the cut-off. Data points representing features identified as the proteomic signature are highlighted with a light brown background.

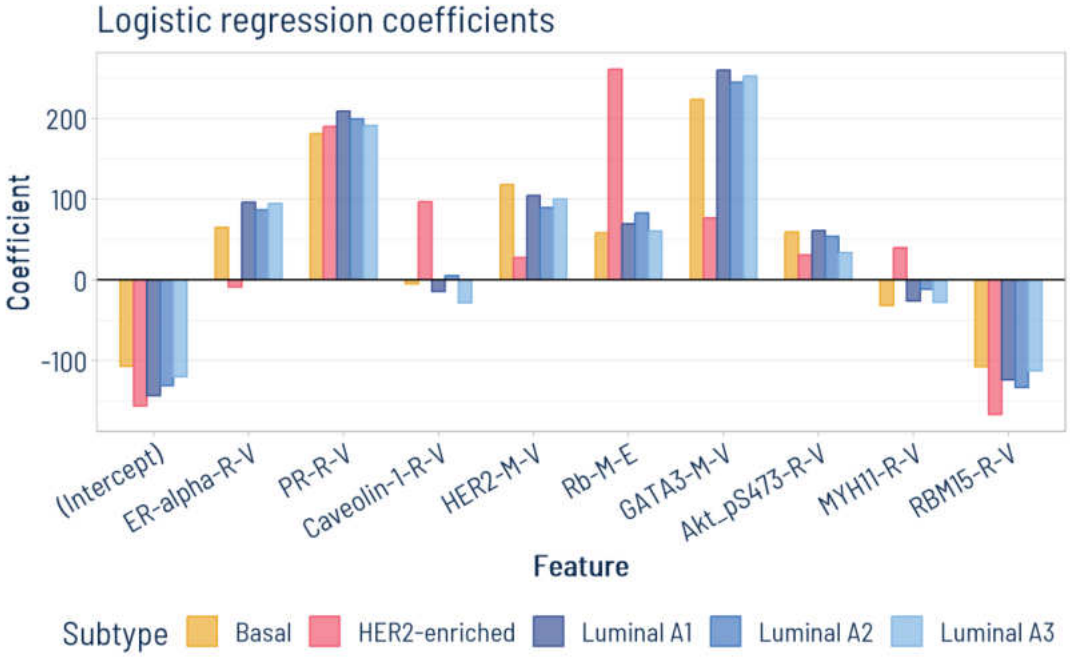


Figure 6.14 Multinomial logistic regression coefficients for the model fitted using the selected proteomic signature

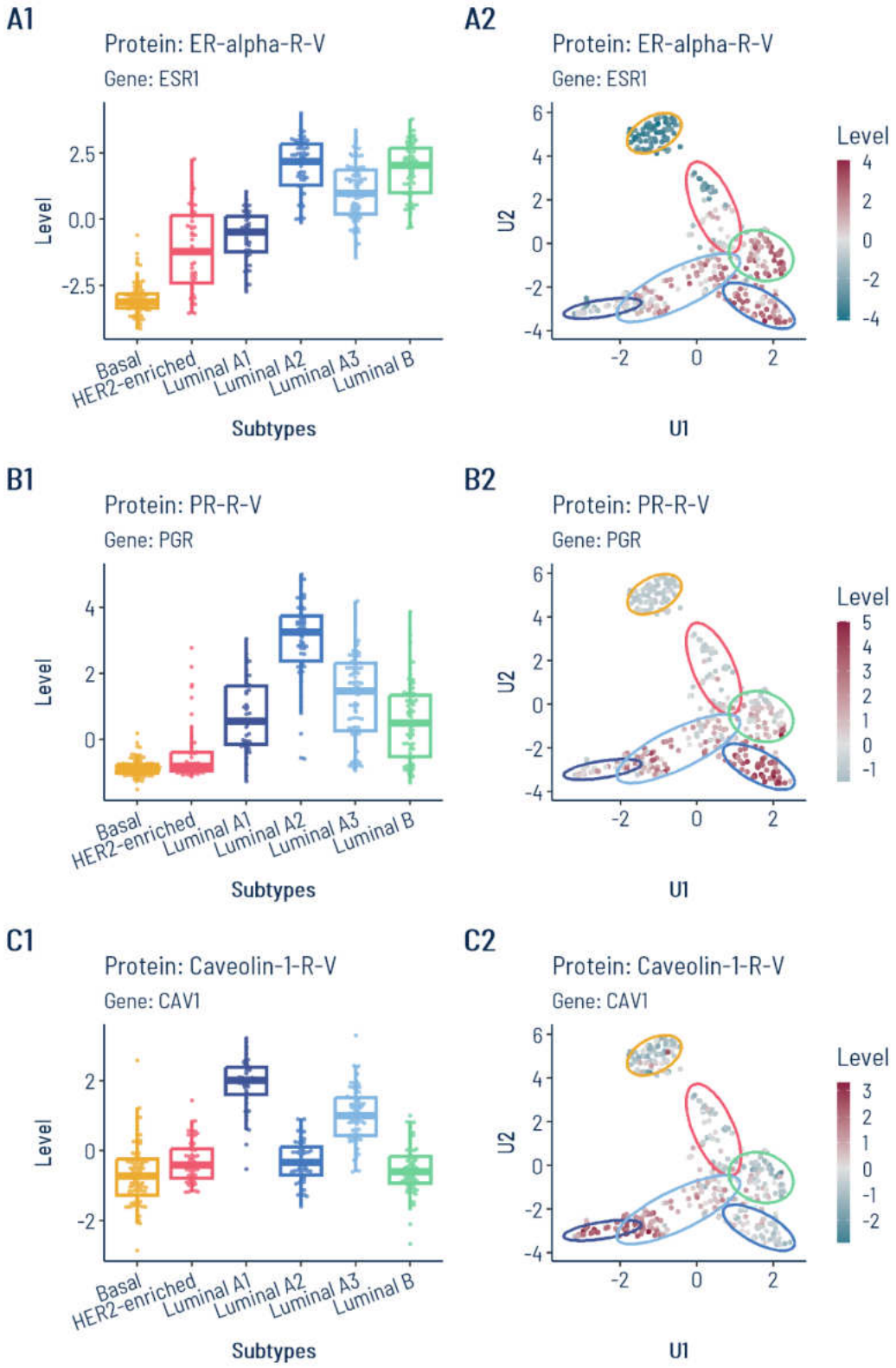


Figure 6.15 Levels of the top three proteins selected for the multinomial regression model with regard to subpopulations identified with DiviK

Panels A show boxplots of protein levels per subtype. Panels B show the UMAP projection obtained in Chapter 4 based on the protein level data set, with the color of data points reflecting the protein level.

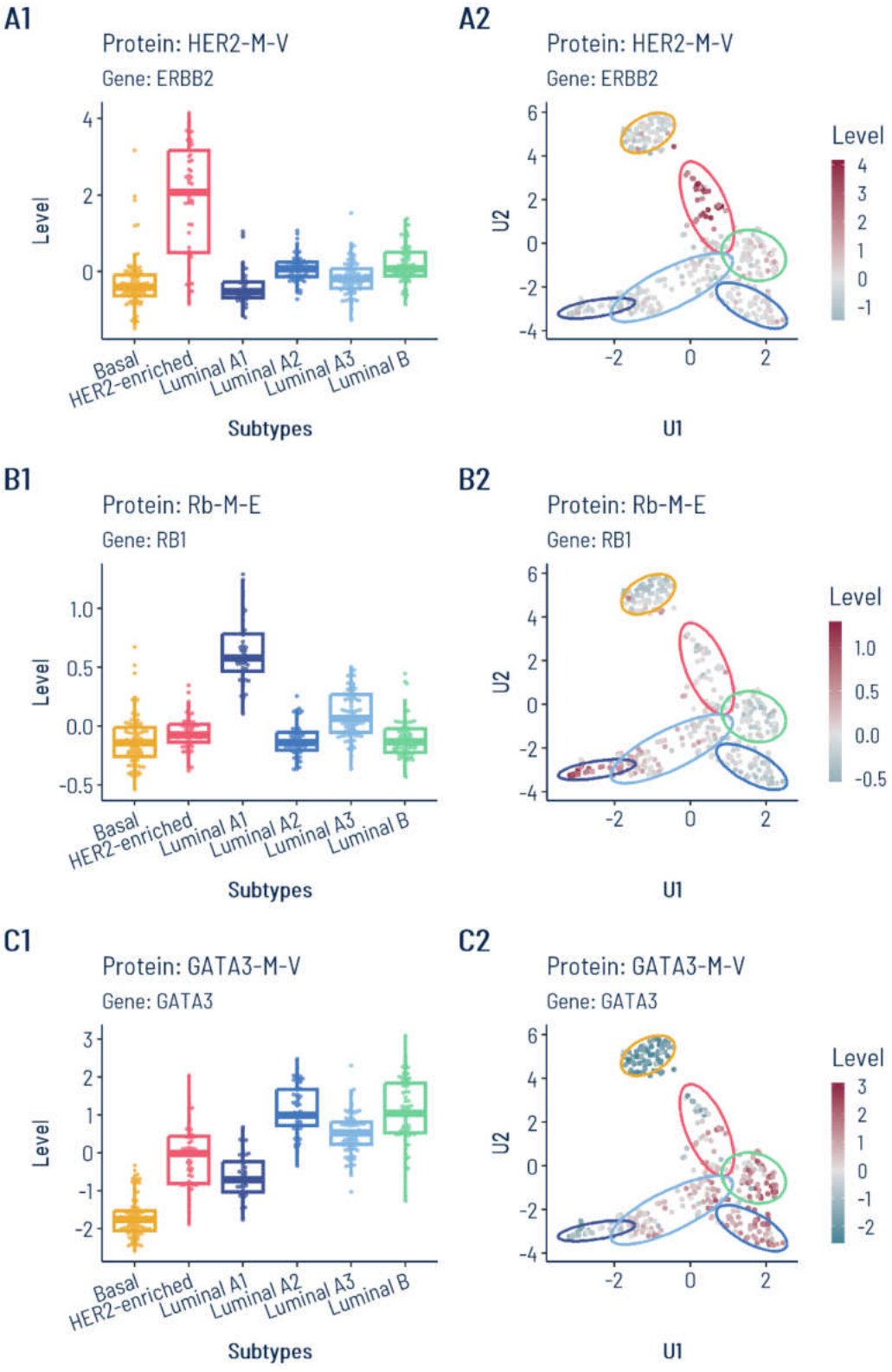


Figure 6.16 Levels of the proteins 4-6 out of 9 selected for the multinomial regression model with regard to subpopulations identified with DiviK

Panels A show boxplots of per subtype protein levels. Panels B show the UMAP projection obtained in Chapter 4 based on the protein level data set with the color of data points reflecting the protein level.

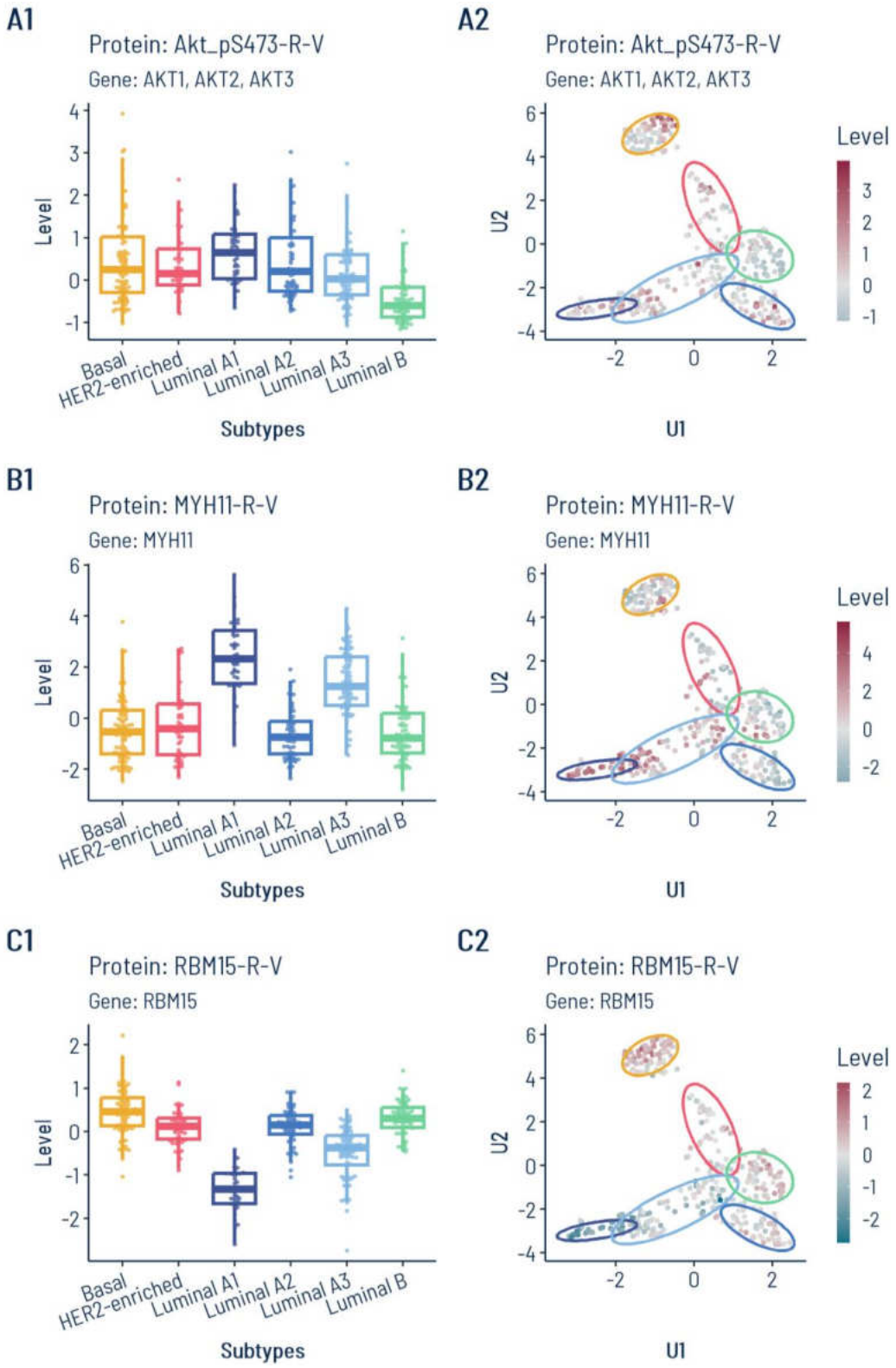


Figure 6.17 Levels of the proteins 7-9 out of 9 selected for the multinomial regression model with regard to subpopulations identified with DiviK

Panels A show boxplots of per subtype protein levels. Panels B show the UMAP projection obtained in Chapter 4 based on the protein level data set with the color of data points reflecting the protein level.

Figure 6.18 compares the selected model-based protein signature and the sets of subtype-specific markers selected based on the effect sizes between all luminal subpopulations (Panel A) or between luminal A subpopulations (Panel B). The model-based signature and luminal-wise subtype-specific markers shared three proteins. The first one is the first feature selected for the model signature - estrogen receptor (ER-alpha-R-V), which low levels were specific for the luminal A1 subtype. The second feature added to the mode – progesterone receptor (PR-R-V) was identified to have characteristically higher levels in the luminal A2 subpopulation. GATA3 transcription factor (GATA3-M-V), added to the model as the sixth independent variable, also had specifically low levels in luminal A1 cases.

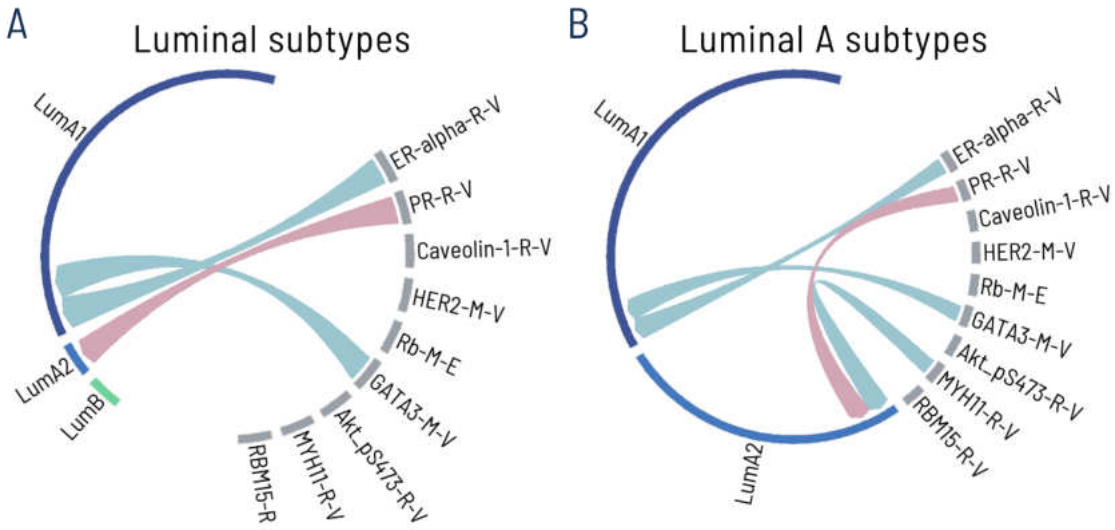


Figure 6.18 Comparison of proteomic model features and proteomic subtype-specific markers identified based on the effect size

A link means that a particular protein was included in the model and was also identified as the subtype-specific marker. Pink and turquoise colors indicate the increase or decrease in protein level compared to other luminal subtypes (Panel A) or other luminal A subtypes (Panel B).

Spearman rank correlation between the protein levels included in the proteomic model signature and all subtype-specific markers, luminal subtype-specific markers, or luminal A subtype-specific markers are visualized in Supplementary Figure 8.4, Supplementary Figure 8.5, and Supplementary Figure 8.6, respectively. For clarity, a negative and positive association is presented separately in Panels A and B, respectively. Moreover, only correlation coefficients

with absolute values higher than 0.3 are shown. Interestingly, GATA3 transcription factor and ER levels are strongly correlated.

6.3.2 Transcriptomic signature

After removing the missing records, the mRNA gene expression data set included measurements for 17328 genes. Feature selection with the forward method would be insufficient, so the data set was limited to only 1124 genes with the highest variance within the cohort. The variance threshold was identified based on the GMM decomposition, presented in Figure 6.19.

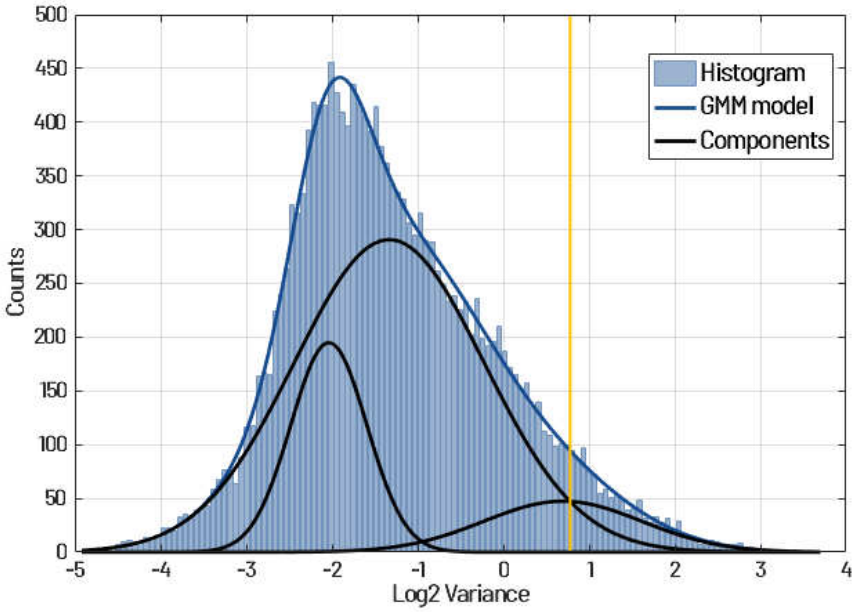


Figure 6.19 Gaussian Mixture Model decomposition of log-2 scaled mRNA gene expression levels' variances for feature selection

The yellow line marks the cut-off value.

Feature ranking obtained in the MRCV procedure is shown in Figure 6.20 and as the elbow plot in Figure 6.21A. Furthermore, Figure 6.21B shows the shortest distances to the line per feature. The maximal distance was obtained for the sixth gene (*C7*). Hence, the top 5 genes formed the transcriptomic signature for subpopulations' differentiation.

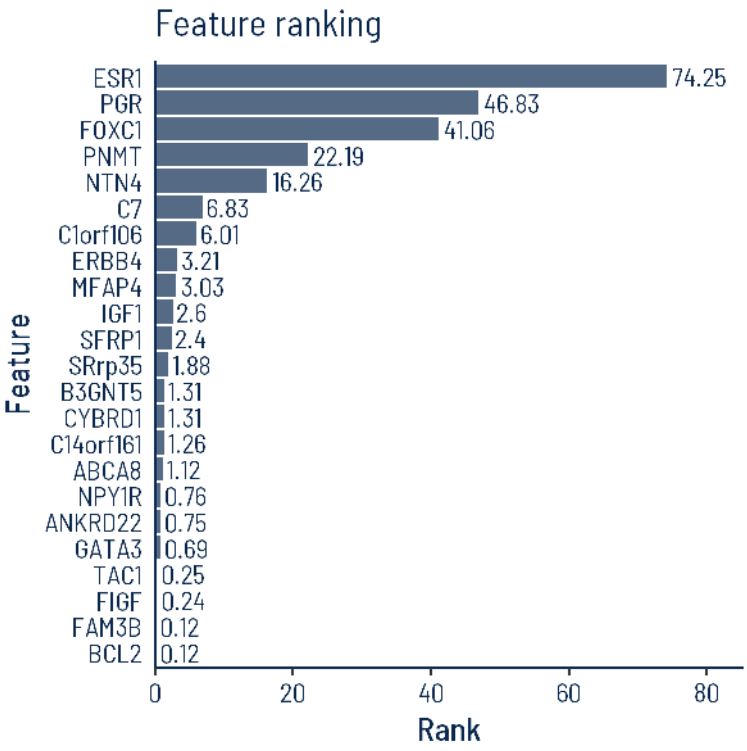


Figure 6.20 Feature ranking for the transcriptomic multimodal logistic regression model

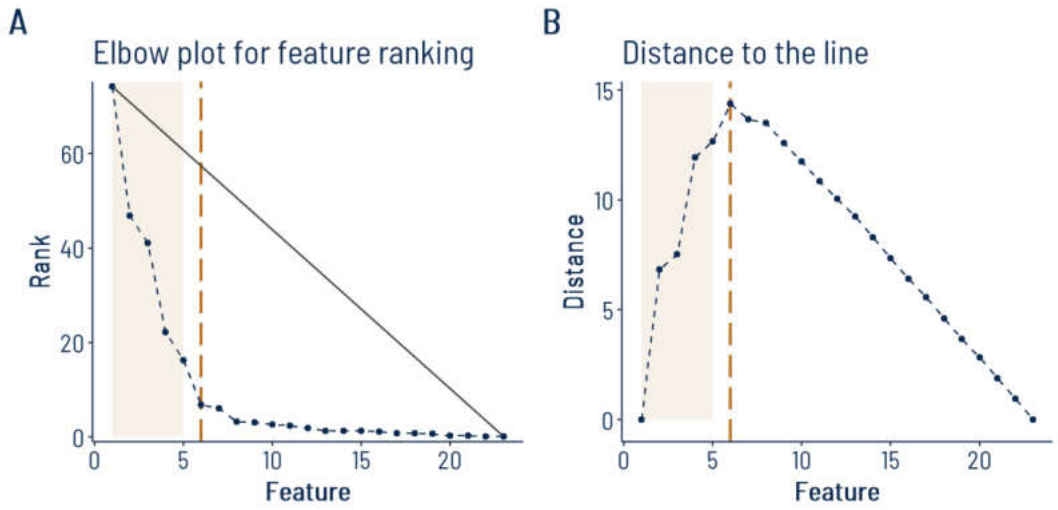


Figure 6.21 Transcriptomic signature identification with the elbow method

Panel A shows the elbow plot for all features selected in the MRCV procedure. Panel B shows the shortest distance between each data point and the black line joining the data points with the highest and lowest ranking score in Panel A. Brown broken line marks the feature with the highest distance serving as the cut-off. Data points representing features identified as the transcriptomic signature are highlighted with a light brown background.

Figure 6.22 shows the multinomial logistic regression coefficients for a model in which the identified transcriptomic signature served

as the independent variables. mRNA gene expression levels of those five selected genes are presented in Figure 6.23 (top three genes) and Figure 6.24 (last two genes). As can be noticed in those figures, the selected genes cannot differentiate the luminal subtypes as well as proteins. However, more distinct differences in expression levels can be observed for HER2-enriched and basal tumors compared to luminals.

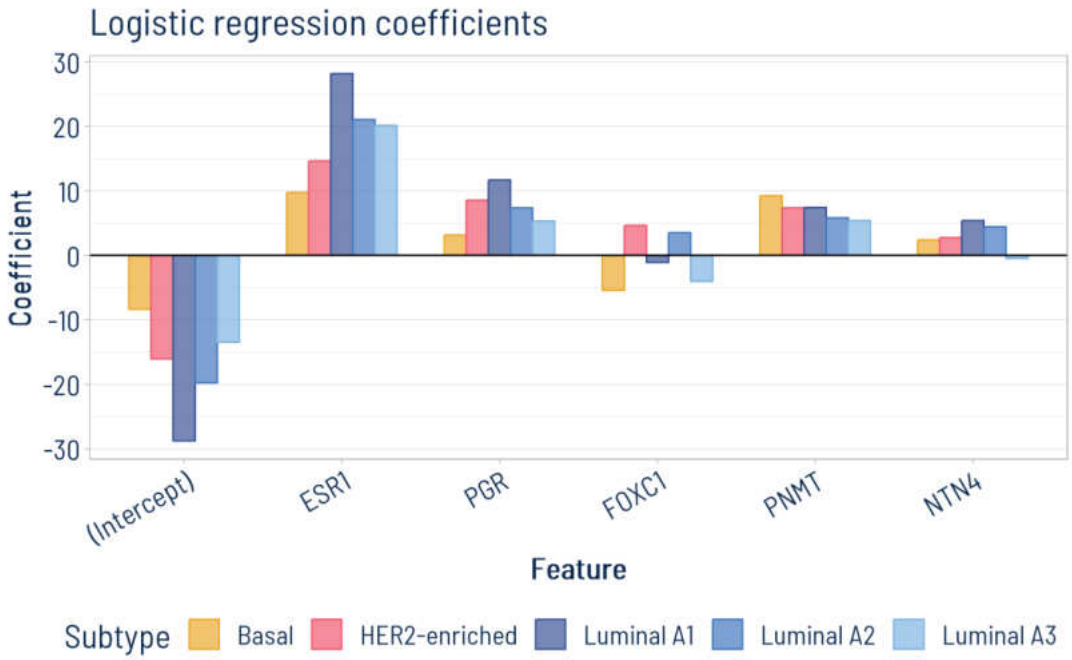


Figure 6.22 Multinomial logistic regression coefficients for the model fitted using the selected transcriptomic signature

Interestingly, the first two genes selected for the model (*ESR1* and *PGR*) code the top two proteins from the proteomic signature (estrogen and progesterone receptors). Nonetheless, the corresponding genes and proteins did not show the same pattern, especially in the case of the luminal A1 subpopulation. The differences between median gene expression levels for the luminal subgroups were smaller than in the case of protein levels. Luminal A1 and HER2-enriched tumors revealed almost equal median levels of estrogen receptors and distinctly lower than the other luminal subpopulations. Proteins coded by the remaining genes included in the signature (*FOXC1*, *PNMT*, *NTN4*) were not measured by the RPPA platform, so the mRNA gene expression and protein levels cannot be compared here.

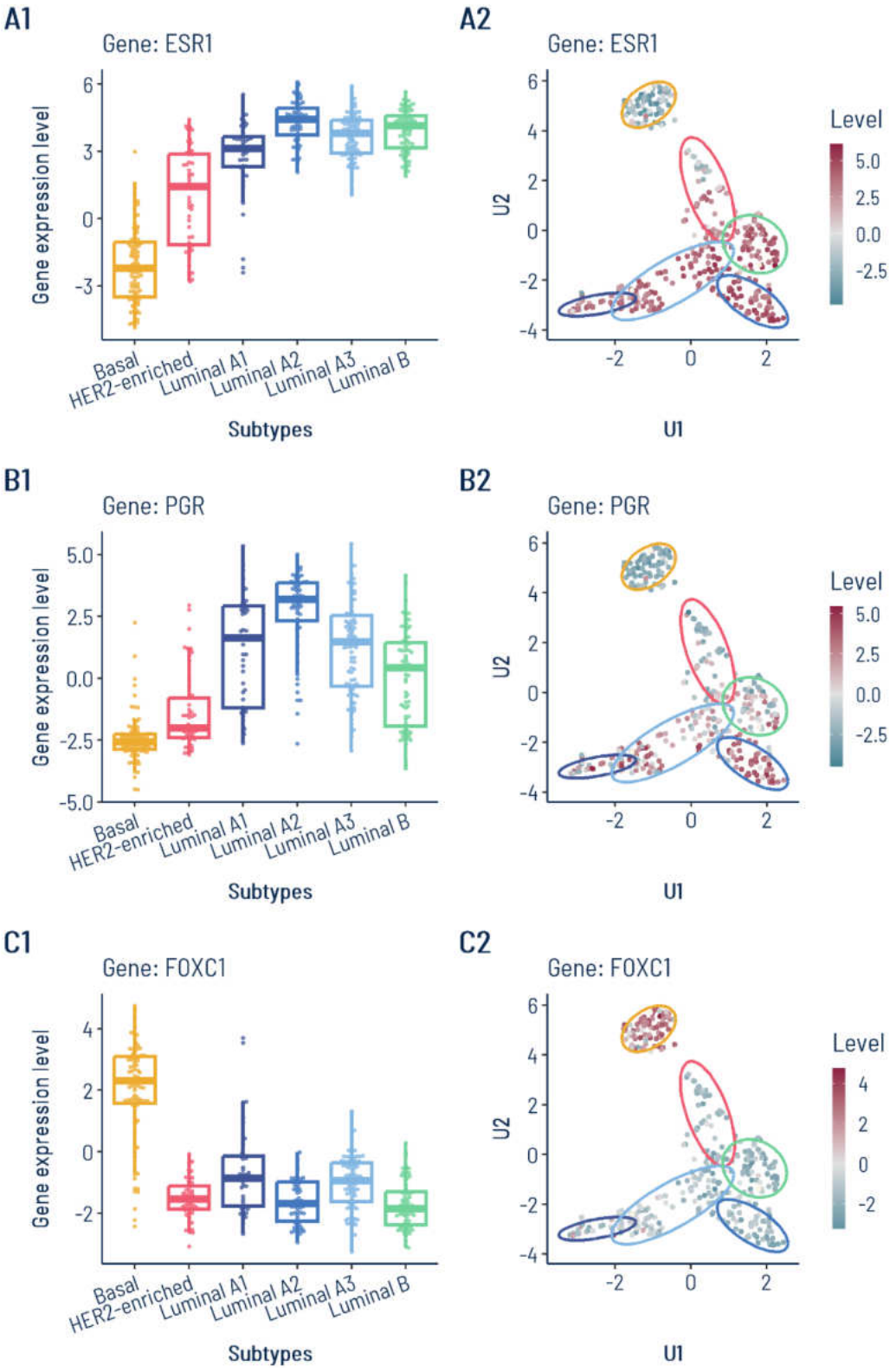


Figure 6.23 Levels of the top three transcripts selected for the multinomial regression model with regard to subpopulations identified with DiviK

Panels A show boxplots of mRNA gene expression levels per subtype. Panels B show the UMAP projection obtained in Chapter 4 based on the protein level data set with the color of data points reflecting the mRNA gene expression level.

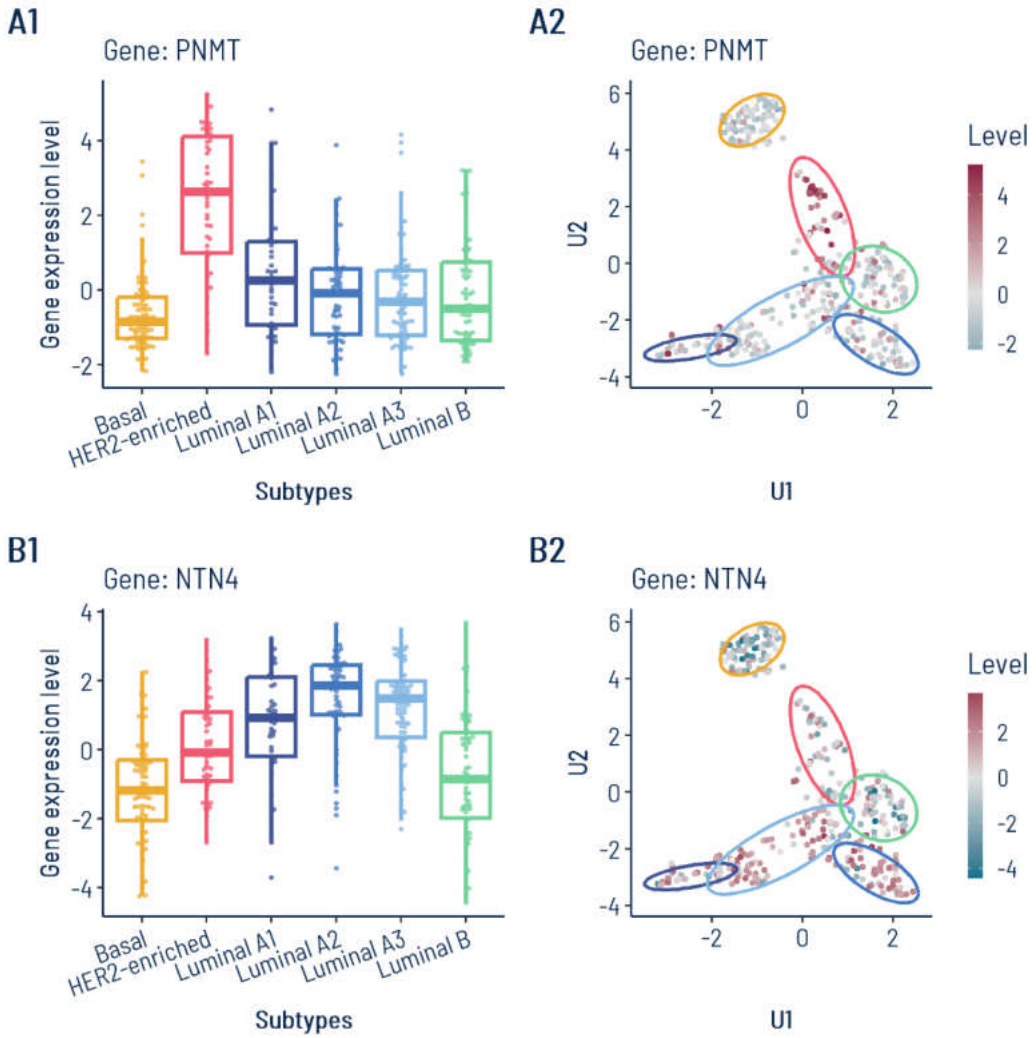


Figure 6.24 Levels of the last two transcripts selected for the multinomial regression model with regard to subpopulations identified with DiviK

Panels A show boxplots of mRNA gene expression levels per subtype. Panels B show the UMAP projection obtained in Chapter 4 based on the protein level data set with the color of data points reflecting the mRNA gene expression level.

6.3.3 Combined signature

The combined set of measurements for 166 proteins and 1124 genes following the GMM-based filtration served the creation of the joint multinomial logistic regression model. The feature ranking resulting from the MRCV procedure is shown in Figure 6.25, which for clarity, was truncated to features that at least twice appeared in the models through MRCV 100 iterations. The full ranking is presented as an elbow plot in Figure 6.26A. Figure 6.26B shows the shortest distances to the line from each data point. As can be seen in the plot, the maximal distance was obtained for the tenth feature (PDCD4-R-C), and the top nine proteins were identified as the combined signature. Interestingly, all those features were

proteomic, as the first mRNA gene expression level has the eleventh position in the ranking. Furthermore, the order of those top features is identical as in the case of the proteomic-only model.

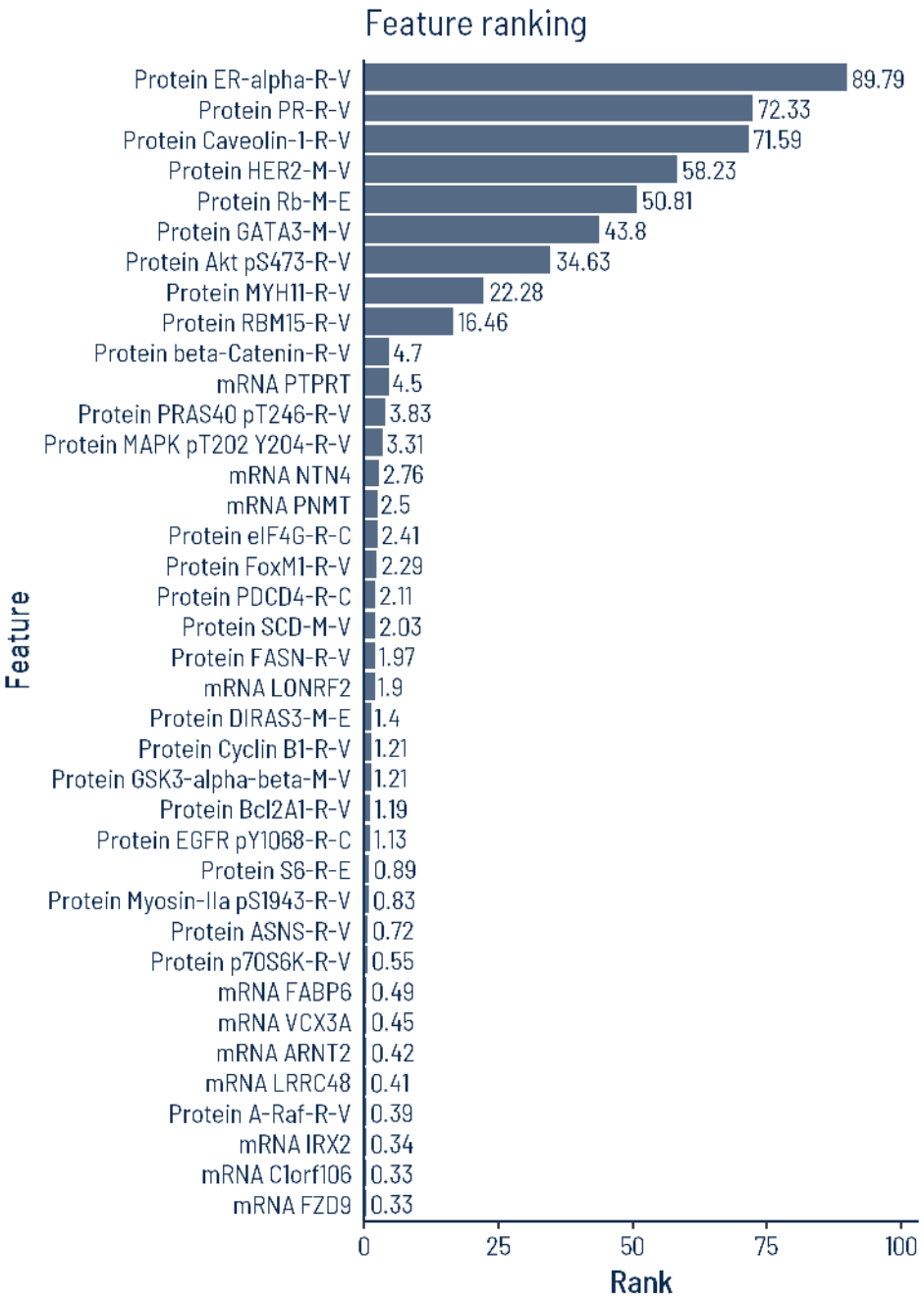


Figure 6.25 Feature ranking for the combined proteomic and transcriptomic multimodal logistic regression model

For clarity, the plot was truncated to show only features selected for more than one model in the MRCV procedure.

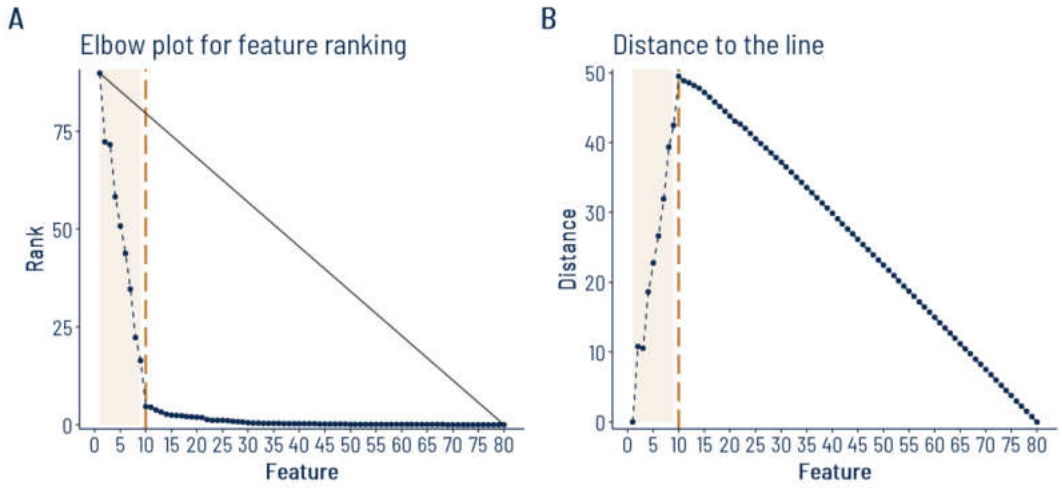


Figure 6.26 Combined proteomic and transcriptomic signature identification with the elbow method

Panel A shows the elbow plot for all features selected in the MRCV procedure. Panel B shows the shortest distance between each data point and the black line joining the data points with the highest and lowest ranking score in Panel A. Brown broken line marks the feature with the highest distance serving as the cut-off. Data points representing features identified as the combined proteomic and transcriptomic signature are highlighted with a light brown background.

The quality of subtype prediction for each of the three applied approaches was assessed with balanced accuracy, sensitivity, and specificity. The average results obtained in MRCV procedures per training and testing sets, accompanied by standard deviations as error bars, are shown in Figure 6.27 and Supplementary Table 8.6. The same features were selected based on proteomic and combined data sets, so the performance for those two models is highly similar. All models efficiently deal with basal subtype prediction, while the quality for the remaining subtypes is worse. Especially for the test sets, sensitivity is lower than specificity, which is understandable for the multiclass task with the imbalance in group sizes. The performance of transcriptomics-based models is, however, interesting. It was expected to be worse than the protein-based one, as the subpopulations were identified in the protein-level space. However, the prediction quality, especially for luminal subpopulations, is inferior. The model seems unable to distinguish between luminal subtypes, with a slightly better result for luminal B cases. Luminal A1 cases were rarely detected, indicating that this subpopulation is indefinite on the transcriptomic level while being highly distinguishable in its proteomic profile with many specific markers.

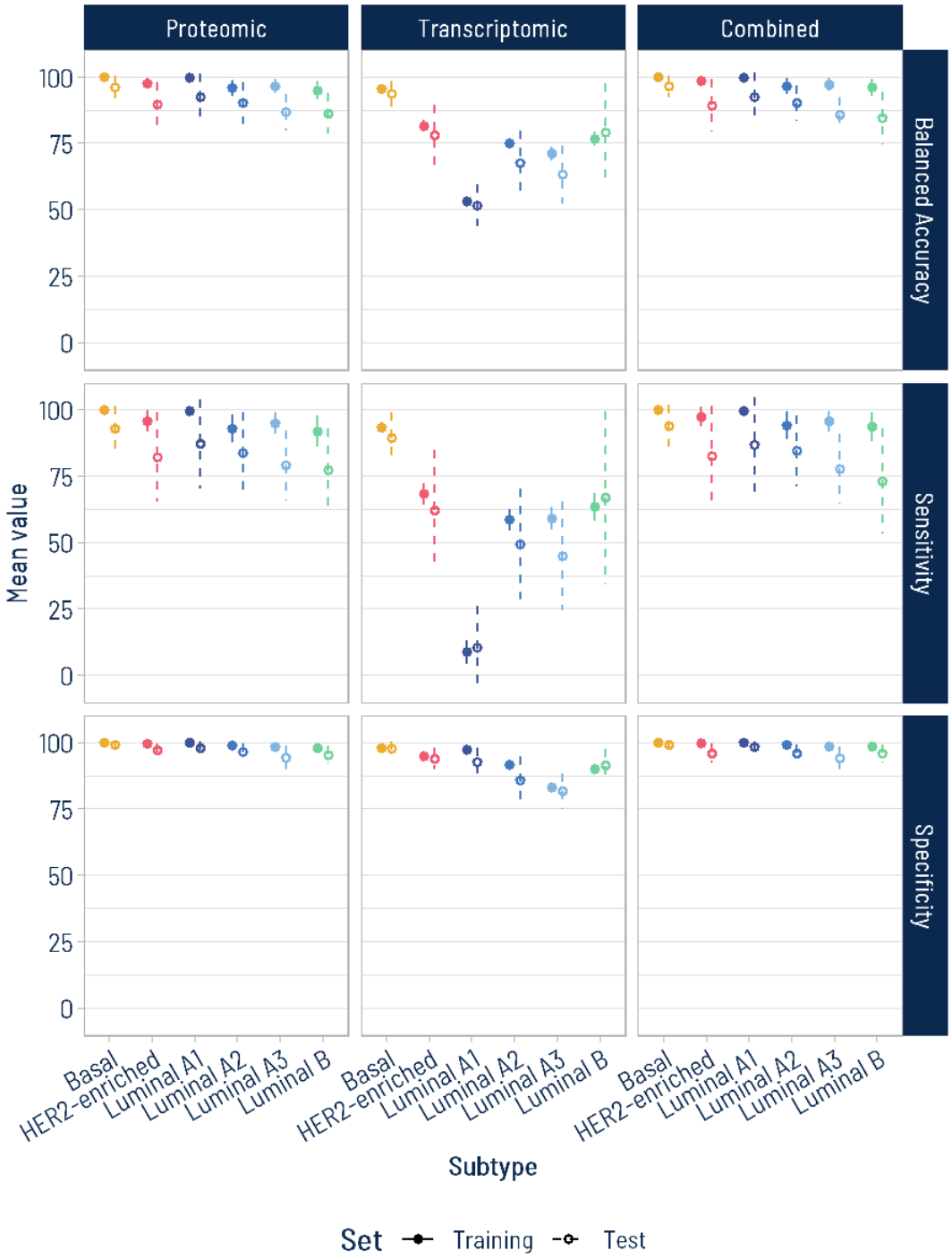


Figure 6.27 Mean balanced accuracy, sensitivity, and specificity per subtype for all considered models with regard to training and test sets of MRCV procedure

Error bars represent the standard deviation.

6.4 Conclusions and discussion

Molecular differences in protein and mRNA gene expression levels were revealed between the breast cancer subpopulations determined with the DiviK approach

based on protein levels. Transcriptomic or proteomic subtype-specific markers were identified for all subtypes. However, the number of those markers varied between the approaches and subtypes. After rejecting HER2-enriched and basal tumors, DiviK-detected luminal subgroups were also successfully characterized with proteomic and transcriptomic signatures. As could have been expected, proteomic-based differentiation was more significant since the subtypes were determined based on RPPA measurements. Nevertheless, smaller but considerable variability was also observed in mRNA gene expression levels. Hence, the identified subpopulations can be characterized with both proteomic and, with slightly worse quality, transcriptomic profiles.

Even though ORA occurred challenging due to insufficient set sizes, it can be concluded that selected marker lists include proteins and genes which role in tumor biology, breast cancer development, cell reaction to stress, cell proliferation, or response to therapy has been either well-established already or reported.

Cyclin E1 protein, identified as the basal-specific marker with an increased level compared to other subtypes, was reported to be overexpressed in TNBC in several studies (Aziz, et al., 2022; Milioli, Alexandrou, Lim, & Caldon, 2020; Llobet, et al., 2020). It participates in several crucial KEGG pathways, including p53 signaling, cell cycle, microRNAs in cancer, cellular senescence, PI3K-Akt signaling pathway, and pathways in cancer. Overexpression of asparagine synthetase (ASNS), which high levels were revealed to be characteristic of the luminal B subtype, has been associated with poor prognosis in breast cancer (Qin, Yang, & Zhan, 2020). However, it was reported to be overexpressed mainly in TNBC (Lin, et al., 2018). Overexpression of the *FMO5* gene, revealed here for the luminal A1 subpopulation, was indicated to be associated with better survival and identified as ER-responsive, participating in breast cancer drug metabolism (Bièche, Girault, Urbain, Tozlu, & Lidereau, 2004).

Signatures differentiating all subtypes were also identified with the multinomial logistic regression approach. Distinguishing between all subpopulations based on transcriptomic data set occurred to be challenging.

The obtained model did not deal with discriminating the luminal subgroups. Adding the transcripts to the proteomic model also did not improve its performance significantly. Nonetheless, the selected transcriptomic and proteomic signatures provided valuable and comprehensive insight into the differences between subpopulations, especially luminal ones, and the changes between the corresponding gene and protein levels.

The first two features added to the models were ER and PR or their genes. Their role in breast cancer is well-known and established. Interestingly, the revealed subpopulations differ regarding PR and ER levels not only between basal, HER2-enriched, and luminal tumors but also within the luminal group. The luminal A2 subpopulation appeared to have higher ER and PR levels, while luminal A1 had the lowest. Moreover, ER levels in that group were more decreased and similar to the HER2-enriched subtype than the *ESR1* gene. The role of the third proteomic feature - caveolin - in breast cancer is equivocal as it has been reported to both suppress and promote breast cancer progression (Qian, et al., 2019; Ren, et al., 2021; Savage, et al., 2007). In this dissertation, caveolin was observed to be overexpressed in luminal A1 and A3 subtypes. HER2 protein is obviously overexpressed in the HER2-enriched subtype. Interestingly, the RB1 level was distinctly higher in the luminal A1 subtype. This protein has been related to therapy response, but mainly in TNBCs (Robinson, et al., 2013), and the loss of heterozygosity has been observed at its locus in basal and luminal B tumors (Herschkowitz, He, Fan, & Perou, 2008). The role of the GATA3 transcription factor in breast cancer is also well-established, as it serves as a diagnostic marker for luminal A and B subtypes, creating a transcription factors' network with ER and FOXA1 (Perou & Borresen-Dale, 2010; Martin, Orlando, Yokobori, & Wade, 2021; Takaku, Grimm, & Wade, 2015). Hence, the decreased GATA3 levels in the luminal A1 subtype seemed interesting. AKT protein, also included in the proteomic signature with lower levels in luminal B tumors, plays a crucial role in the PI3K-AKT signaling pathway in cell metabolism, growth, proliferation, apoptosis, and angiogenesis. Its connections to carcinogenesis were broadly studied (Miricescu, et al., 2020; Risso, Blaustein, Pozzi, Mammi, & Srebrow, 2015; Zhang,

et al., 2020). MYH11 levels were found to be relatively high for luminal A1 and A3 subpopulations. This protein is involved in contraction production. However, its potential role in breast cancer was not revealed yet. The last protein included in the signature, RBM15, has been reported to be regulated by BARX2 and ER, hence affecting cell growth and invasion in breast cancer samples (Zheng, et al., 2021; Stevens & Meech, 2006). *FOXC1* gene, upregulated in DiviK-based basal cluster, is a well-known marker and therapeutic target for basal breast cancers (Elian, Yan, & Walter, 2017; Mott, Su, & Pack, 2018). *PNMT* gene, a fourth feature in the transcriptomic signature, was overexpressed in the HER2-enriched subtype. This gene has been observed before to be co-expressed with the HER2-coding gene *ERBB2* (Dressman, et al., 2003). The last gene of the transcriptomic signature, *NTN4*, had slightly higher expression levels in luminal A subgroups. Similar findings were reported for luminal samples in (Yi, et al., 2022), where *NTN4* was concluded to serve as the breast cancer prognostic marker and immune infiltration hallmark.

7 Summary and conclusions

The goals of this thesis in the identification of breast cancer patient subpopulations and their clinical and molecular evaluation have been achieved. The results described in this dissertation justify the thesis formulated in Chapter 1.2. Thesis I was confirmed by the analysis outcomes shown in Chapter 4. It was demonstrated that various tested combinations of feature engineering and clustering algorithms reveal novel subpopulations of breast cancer patients based on their proteomic profiles. The proposed metrics for clustering outcome comparison allowed the selection of the approach producing the most distinct subpopulations. Thesis II was proved in Chapters 5 and 6. In Chapter 5.1, the differences in survival experiences between the defined subpopulations were confirmed. Not only HR+ and HR- subtypes were shown to vary in prognosis, but also the newly revealed additional luminal subgroups were diverse in their survival outcome. In Chapter 5.2, a small association between investigated subpopulations and demographic or clinical factors was found, similar to PAM50-based subtypes. It was also detected that identified subpopulations demonstrate diversity in immune cell fractions,

including the luminal subgroups. In Chapter 6.2, the differentiation testing pipeline relying on classical statistical testing and effect size estimation allowed the definition and functional characterization of proteomic and transcriptomic profiles of the majority revealed subpopulations. Furthermore, in Chapter 6.3, proteomic signature distinguishing between all subtypes was selected. The transcriptomic signature allowed mainly HR+ and HR- subtype recognition but performed poorly in distinguishing between revealed luminal subtypes.

This dissertation addressed the need for the re-identification of established breast cancer classification with the use of machine learning and mathematical modeling approaches. Firstly, machine learning techniques recognized breast cancer patient subpopulations in the set of protein levels. Subsequently, the obtained clusters were evaluated regarding demographic and clinical factors. Finally, the subtypes were characterized molecularly with comprehensive statistical methods and statistical learning approaches.

All applied machine learning approaches delivered evidence that the luminal A intrinsic subtype is the most heterogeneous in the TCGA-BRCA cohort and should be further divided into two or three subgroups. Feature selection or extraction steps before clustering were crucial for the outcome quality. GMM-based feature filtration improved the detection of highly distinct clusters, regardless of the clustering algorithm. The proposed centroid-based approach with iterative k-means clustering in locally GMM-filtered feature space provided the best results among all tested approaches. It identified six patient subpopulations named according to their consistency with PAM50 labels as basal, HER2-enriched, luminal B, and three luminal A subgroups: A1, A2, and A3.

The demographic and clinical evaluation of identified subpopulations highlighted the importance of an appropriate statistical testing approach, especially for such a challenging data set. Given the insufficient follow-up time for cancer with a relatively good prognosis, such as breast cancer, it was crucial to properly define an endpoint relating to time to relapse rather than death. Furthermore, extending the classical log-rank test with a weighted Gehan-Wilcoxon approach enabled the detection of significant early changes in survival between

subpopulations. Estimating the effect size using HR interpreted with adjustment for unbalanced groups partially resolved the problem of varying study sample sizes and allowed subpopulations to be compared despite the small number of events of interest captured during follow-up. Cramér's V effect size allowed analysis of the association between subpopulations and demographic or clinical factors in a manner adjusted to varying category numbers.

Greater diversity in survival experience was shown than in the case of well-established PAM50-based subtypes. Interestingly, the revealed luminal subtypes varied in their survival outcome, especially regarding the time to new cancer events. The luminal A2 subtype was associated with a prognosis comparably poor to HER2-enriched and basal tumors. On the other hand, luminal A3 cases showed a favorable prognosis.

Subpopulations revealed in this study based on the proteomic portrait demonstrated a slight dependency on demographic and clinical factors, comparable to well-established PAM50-based subtypes. Four luminal subtypes identified in this dissertation demonstrated a small association with lymph nodes affected, which was not observed for the PAM50 classification of luminal A and B subtypes. Moreover, the subpopulations proposed here were suggested to vary in their immune response among both the whole cohort and only the luminal group.

Classical statistical tests and effect size were used to select non-specific and subtype-specific markers in both proteomic and transcriptomic spaces. Due to the large number of features compared to sample sizes, effect size outperformed the classic approach and provided a more rigorous list of markers specific to subtypes. Transcriptomic differentiation between subtypes was smaller than proteomic one.

The method choice was also crucial for the functional analysis. Due to insufficient marker lists and protein universe sizes, the first-generation method ORA did not perform satisfactorily. Nevertheless, the second-generation CERNO test conducted on effect size estimates delivered the lists of significantly enriched pathways. The results indicate distinct differences between identified subpopulations, including the significant diversity within the luminal group, both

on the transcriptomic and proteomic levels. The differentiating genes and proteins are involved in various processes meaningful for proper cell functioning and cancer development.

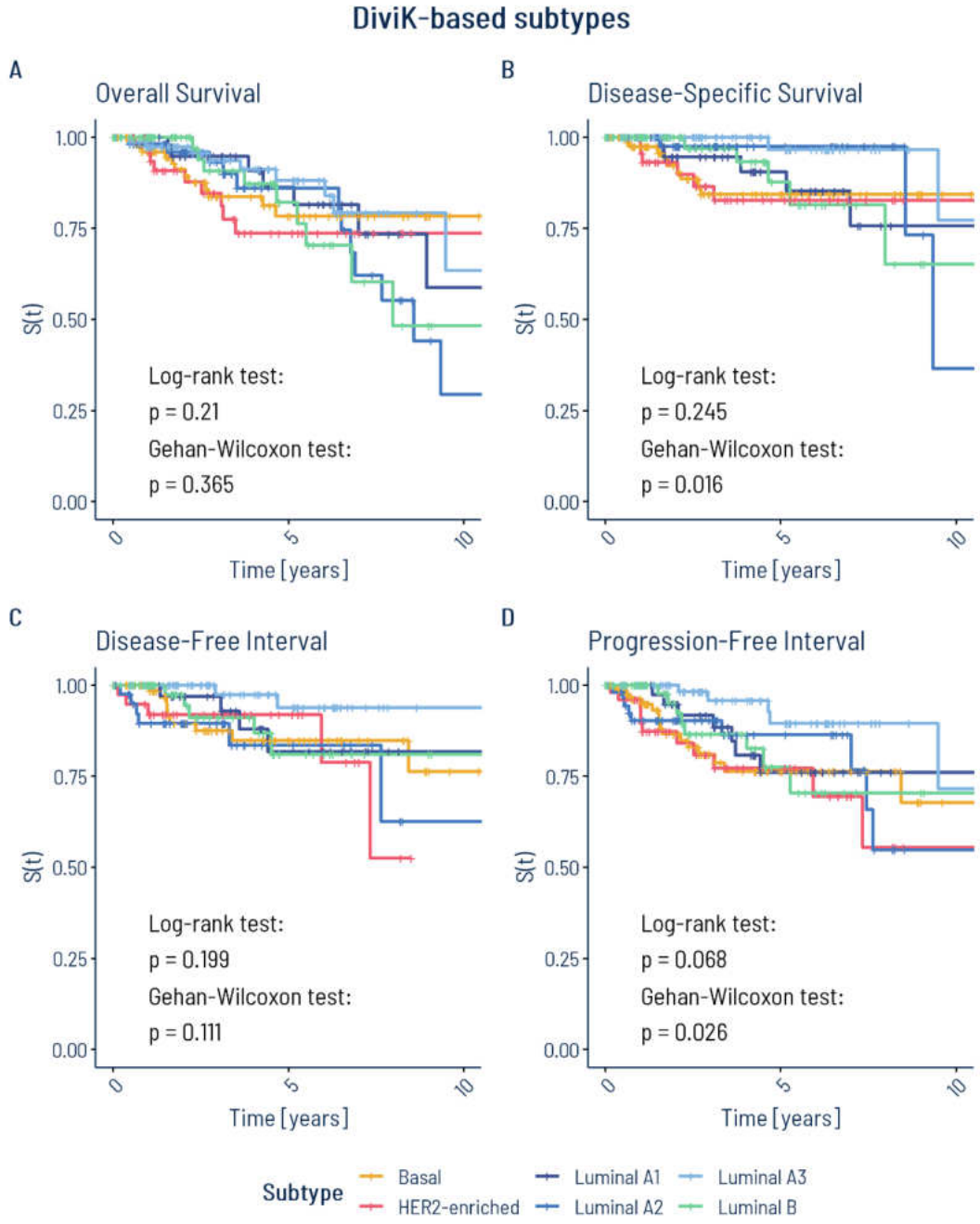
Finally, the dedicated machine learning approach identified the protein signature distinguishing all six revealed subtypes. Similarly, the transcriptomic signature was obtained. However, it was insufficient for differentiating between luminal subgroups. Some of the signature genes and proteins are well-established in their role in breast cancer. For some, however, the association with this disease remains unknown.

Interestingly, one of the revealed luminal A subtypes – luminal A1 - demonstrated distinct differences in the expression of signature genes and proteins compared to the three remaining luminal subgroups. Some similarities to basal and HER2-enriched tumors were demonstrated, as well as distinct differences compared to all subtypes. Moreover, a relative drop in ER expression was observed between mRNA and protein levels. This suggests that luminal A1 cases might have been misclassified as luminal based on gene profiling and are closer to ER- tumors, which cannot be reflected in their transcriptomic portraits.

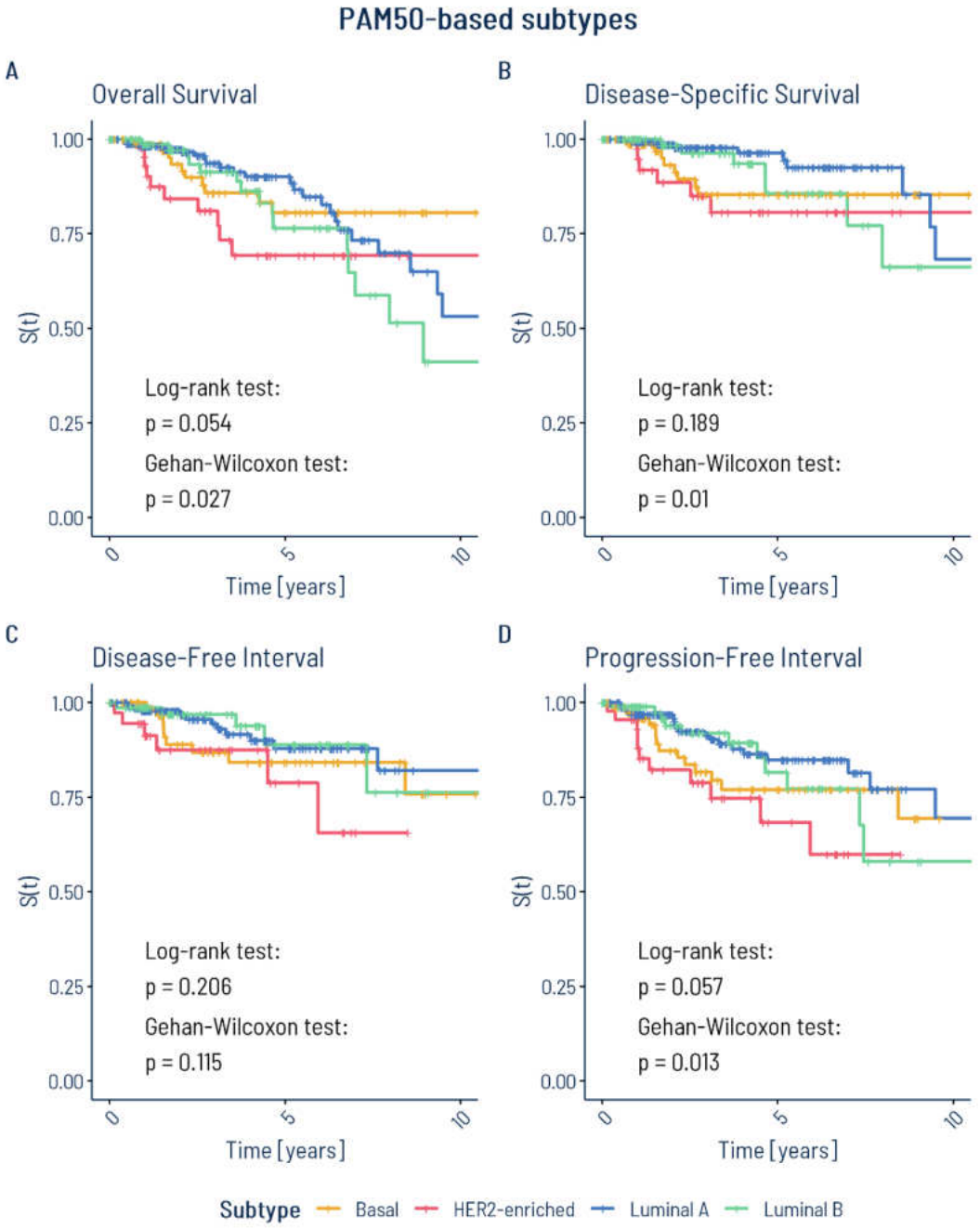
To conclude, proteomic data carry information concerning breast cancer stratification, which remains hidden at the transcriptomic level. Subtyping based on the proteomic profile complements the intrinsic molecular classification of breast cancer and provides superior information on breast cancer heterogeneity not reflected by gene expression profiling. Various mechanisms participate in expression regulation between the mRNA and protein layer. Therefore, the results obtained in this dissertation suggest that those processes impact tumor behavior. Proteomic-based patient subpopulations demonstrate differences in clinical outcome, which were not observed in PAM50 luminal subtypes. Hence, profiling of protein levels can potentially deliver a more comprehensive insight into tumor biology and provide clinically relevant information beyond gene expression profiling. Identified markers can possibly serve for the optimization of therapy planning and contribute to new targeting options research. Nonetheless, further independent validation is required to gain evidence supporting the potential

prognostic or clinical applications and assess whether the current clinical and intrinsic subtyping approaches can be complemented with those findings and applied in the clinical routine.

8 Supplementary materials



Supplementary Figure 8.1 Kaplan-Meier survival curves of all subpopulations identified with DiviK



Supplementary Figure 8.2 Kaplan-Meier survival curves of all PAM50 subtypes

Supplementary Table 8.1 Results of log-rank and Gehan-Wilcoxon tests for comparison of survival functions of all subtypes identified with DiviK or based on PAM50 classifier

Endpoint type	χ^2		p-value	
	Log-rank test	Gehan-Wilcoxon test	Log-rank test	Gehan-Wilcoxon test
Subpopulations identified with DiviK				
Overall Survival	7.15	5.44	0.2099	0.3648
Disease-Specific Survival	6.69	13.94	0.2447	0.0160
Disease-Free Interval	7.31	8.96	0.1986	0.1108
Progression-Free Interval	10.28	12.70	0.0677	0.0264
PAM50-based subtypes				
Overall Survival	7.66	9.18	0.0536	0.0270
Disease-Specific Survival	4.78	11.27	0.1888	0.0103
Disease-Free Interval	4.57	5.93	0.2060	0.1153
Progression-Free Interval	7.52	10.72	0.0571	0.0133

Supplementary Table 8.2 Cox proportional hazard analysis of all identified subpopulations

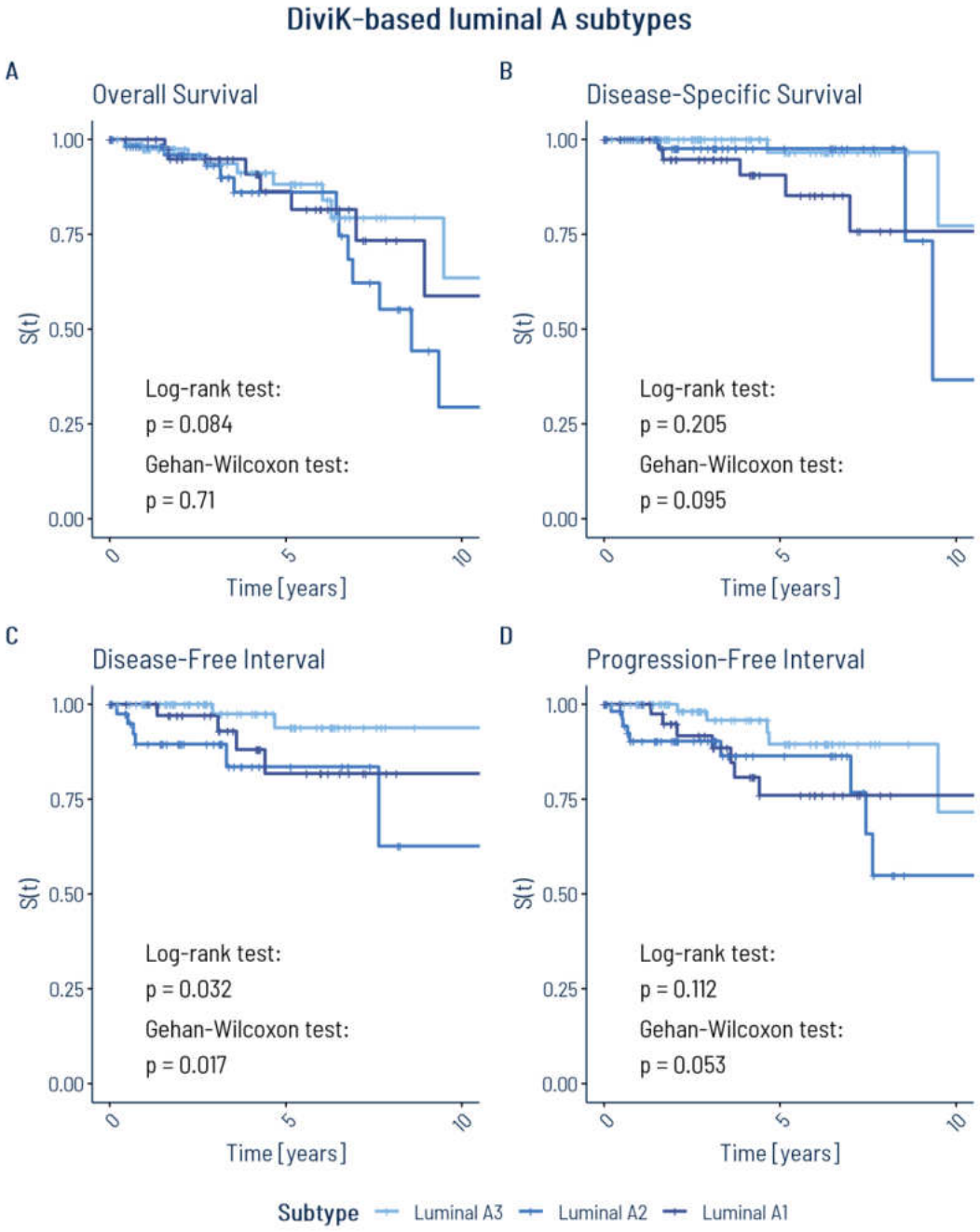
Subtype	N	N_e	N_c	HR	HR effect	HR allocation probability adjusted critical value		
						α = 0.1	α = 0.3	α = 0.5
Overall Survival								
Basal	89	13	76	Reference				
HER2-enriched	54	11	43	1.836	Small	HR < 0.773; HR > 1.294	HR < 0.468; HR > 2.135	HR < 0.274; HR > 3.648
Luminal A1	44	8	36	1.05	No effect	HR < 0.749; HR > 1.336	HR < 0.436; HR > 2.295	HR < 0.249; HR > 4.023
Luminal A2	61	14	47	2.016	Small	HR < 0.785; HR > 1.273	HR < 0.487; HR > 2.054	HR < 0.289; HR > 3.459
Luminal A3	87	9	78	0.837	No effect	HR < 0.816; HR > 1.225	HR < 0.536; HR > 1.867	HR < 0.331; HR > 3.023
Luminal B	72	9	63	1.445	Small	HR < 0.801; HR > 1.248	HR < 0.511; HR > 1.958	HR < 0.309; HR > 3.236
Disease-Specific Survival								
Basal	87	9	78	Reference				
HER2-enriched	52	7	45	1.579	Small	HR < 0.771; HR > 1.297	HR < 0.466; HR > 2.146	HR < 0.272; HR > 3.673
Luminal A1	43	5	38	0.957	No effect	HR < 0.749; HR > 1.336	HR < 0.436; HR > 2.296	HR < 0.249; HR > 4.023

Subtype	N	N_e	N_c	HR	HR effect	HR allocation probability adjusted critical value		
						α = 0.1	α = 0.3	α = 0.5
Luminal A2	58	4	54	0.817	No effect	HR < 0.783; HR > 1.278	HR < 0.483; HR > 2.071	HR < 0.286; HR > 3.5
Luminal A3	85	2	83	0.241	Large	HR < 0.816; HR > 1.225	HR < 0.536; HR > 1.867	HR < 0.331; HR > 3.024
Luminal B	72	5	67	1.002	No effect	HR < 0.803; HR > 1.245	HR < 0.514; HR > 1.946	HR < 0.312; HR > 3.208
Disease-Free Interval								
Basal	79	9	70	Reference				
HER2-enriched	41	5	36	1.315	No effect	HR < 0.755; HR > 1.325	HR < 0.444; HR > 2.254	HR < 0.255; HR > 3.927
Luminal A1	38	4	34	0.819	No effect	HR < 0.745; HR > 1.342	HR < 0.431; HR > 2.32	HR < 0.245; HR > 4.079
Luminal A2	46	6	40	1.516	Small	HR < 0.768; HR > 1.302	HR < 0.462; HR > 2.165	HR < 0.269; HR > 3.717
Luminal A3	79	2	77	0.225	Large	HR < 0.818; HR > 1.222	HR < 0.538; HR > 1.857	HR < 0.333; HR > 3
Luminal B	64	5	59	0.884	No effect	HR < 0.801; HR > 1.248	HR < 0.511; HR > 1.958	HR < 0.309; HR > 3.234
Progression-Free Interval								
Basal	89	15	74	Reference				
HER2-enriched	54	12	42	1.483	Small	HR < 0.773; HR > 1.294	HR < 0.468; HR > 2.135	HR < 0.274; HR > 3.648
Luminal A1	44	7	37	0.738	Small	HR < 0.749; HR > 1.336	HR < 0.436; HR > 2.295	HR < 0.249; HR > 4.023
Luminal A2	61	9	52	0.965	No effect	HR < 0.785; HR > 1.273	HR < 0.487; HR > 2.054	HR < 0.289; HR > 3.459
Luminal A3	87	5	82	0.308	Large	HR < 0.816; HR > 1.225	HR < 0.536; HR > 1.867	HR < 0.331; HR > 3.023
Luminal B	72	8	64	0.846	No effect	HR < 0.801; HR > 1.248	HR < 0.511; HR > 1.958	HR < 0.309; HR > 3.236

Supplementary Table 8.3 Cox proportional hazard analysis of all PAM50 subtypes

"HER2" denotes HER2-enriched subtype, "LumA" luminal A subtype, and "LumB" luminal B subtype.

Subtype	N	N_e	N_c	HR	HR effect	HR allocation probability adjusted critical value		
						α = 0.1	α = 0.3	α = 0.5
Overall Survival								
Basal	86	11	75	Reference				
HER2	50	11	39	2.83	Medium	HR < 0.768; HR > 1.302	HR < 0.462; HR > 2.166	HR < 0.269; HR > 3.72
LumA	173	26	147	1.408	Small	HR < 0.857; HR > 1.166	HR < 0.609; HR > 1.642	HR < 0.4; HR > 2.497
LumB	98	16	82	2.137	Medium	HR < 0.827; HR > 1.209	HR < 0.554; HR > 1.805	HR < 0.348; HR > 2.878
Disease-Specific Survival								
Basal	84	8	76	Reference				
HER2	48	6	42	1.977	Small	HR < 0.766; HR > 1.306	HR < 0.459; HR > 2.179	HR < 0.267; HR > 3.75
LumA	169	10	159	0.691	Small	HR < 0.857; HR > 1.166	HR < 0.609; HR > 1.642	HR < 0.4; HR > 2.497
LumB	96	8	88	1.326	Small	HR < 0.828; HR > 1.208	HR < 0.554; HR > 1.804	HR < 0.348; HR > 2.875
Disease-Free Interval								
Basal	76	9	67	Reference				
HER2	42	6	36	1.723	Small	HR < 0.762; HR > 1.312	HR < 0.454; HR > 2.204	HR < 0.262; HR > 3.81
LumA	147	11	136	0.664	Small	HR < 0.856; HR > 1.169	HR < 0.606; HR > 1.65	HR < 0.397; HR > 2.517
LumB	82	5	77	0.624	Small	HR < 0.824; HR > 1.214	HR < 0.548; HR > 1.826	HR < 0.342; HR > 2.927
Progression-Free Interval								
Basal	86	14	72	Reference				
HER2	50	11	39	1.84	Small	HR < 0.768; HR > 1.302	HR < 0.462; HR > 2.166	HR < 0.269; HR > 3.72
LumA	173	20	153	0.679	Small	HR < 0.857; HR > 1.166	HR < 0.609; HR > 1.642	HR < 0.4; HR > 2.497
LumB	98	11	87	0.858	No effect	HR < 0.827; HR > 1.209	HR < 0.554; HR > 1.805	HR < 0.348; HR > 2.878



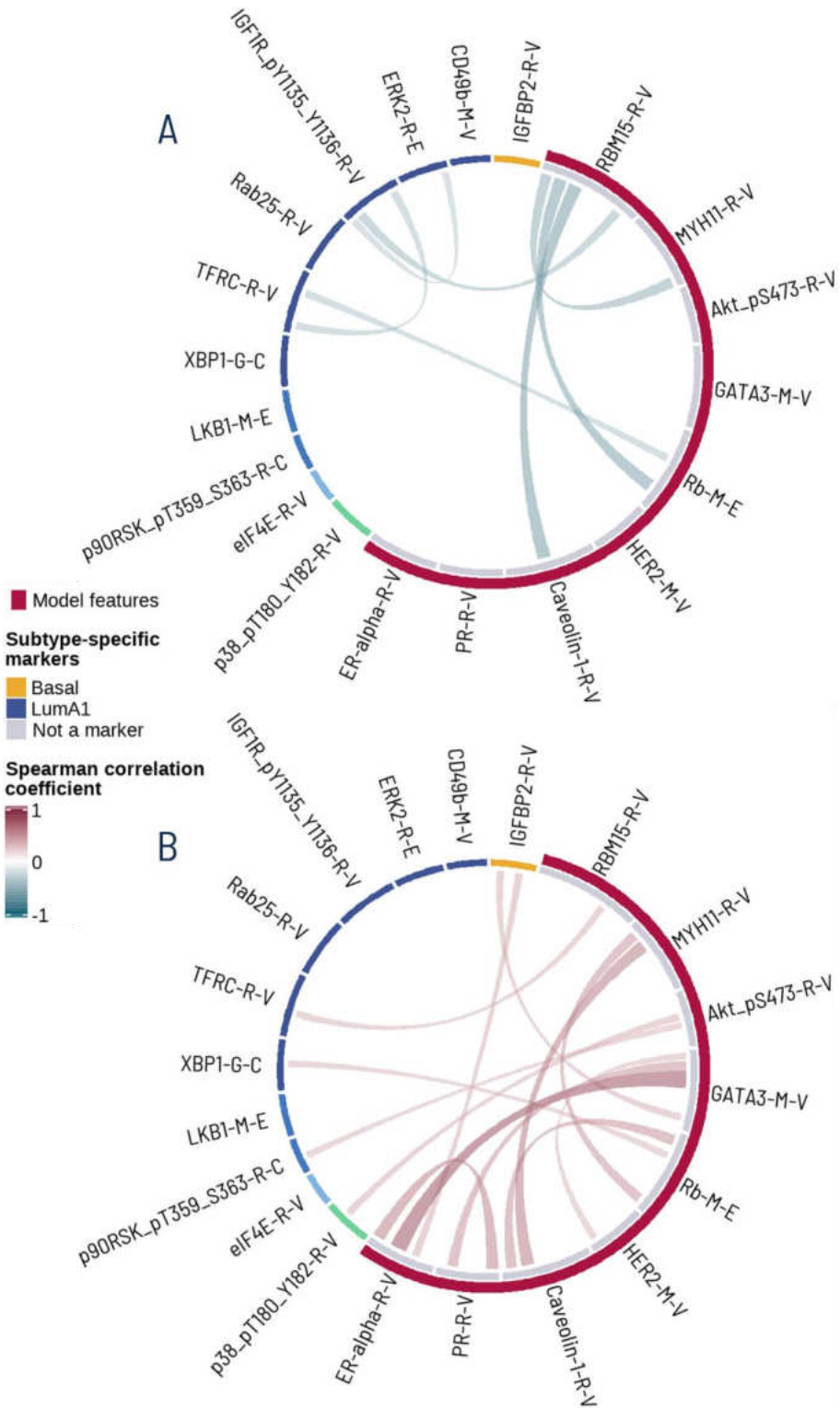
Supplementary Figure 8.3 Kaplan-Meier survival curves of luminal A subpopulations identified with DiviK

Supplementary Table 8.4 Results of log-rank and Gehan-Wilcoxon tests for comparison of survival functions of luminal A subtypes identified with DiviK

Endpoint type	χ^2		p-value	
	Log-rank test	Gehan-Wilcoxon test	Log-rank test	Gehan-Wilcoxon test
Subpopulations identified with DiviK				
Overall Survival	4.96	0.68	0.0837	0.7101
Disease-Specific Survival	3.17	4.71	0.2054	0.0947
Disease-Free Interval	6.87	8.12	0.0322	0.0173
Progression-Free Interval	4.38	5.86	0.1121	0.0533

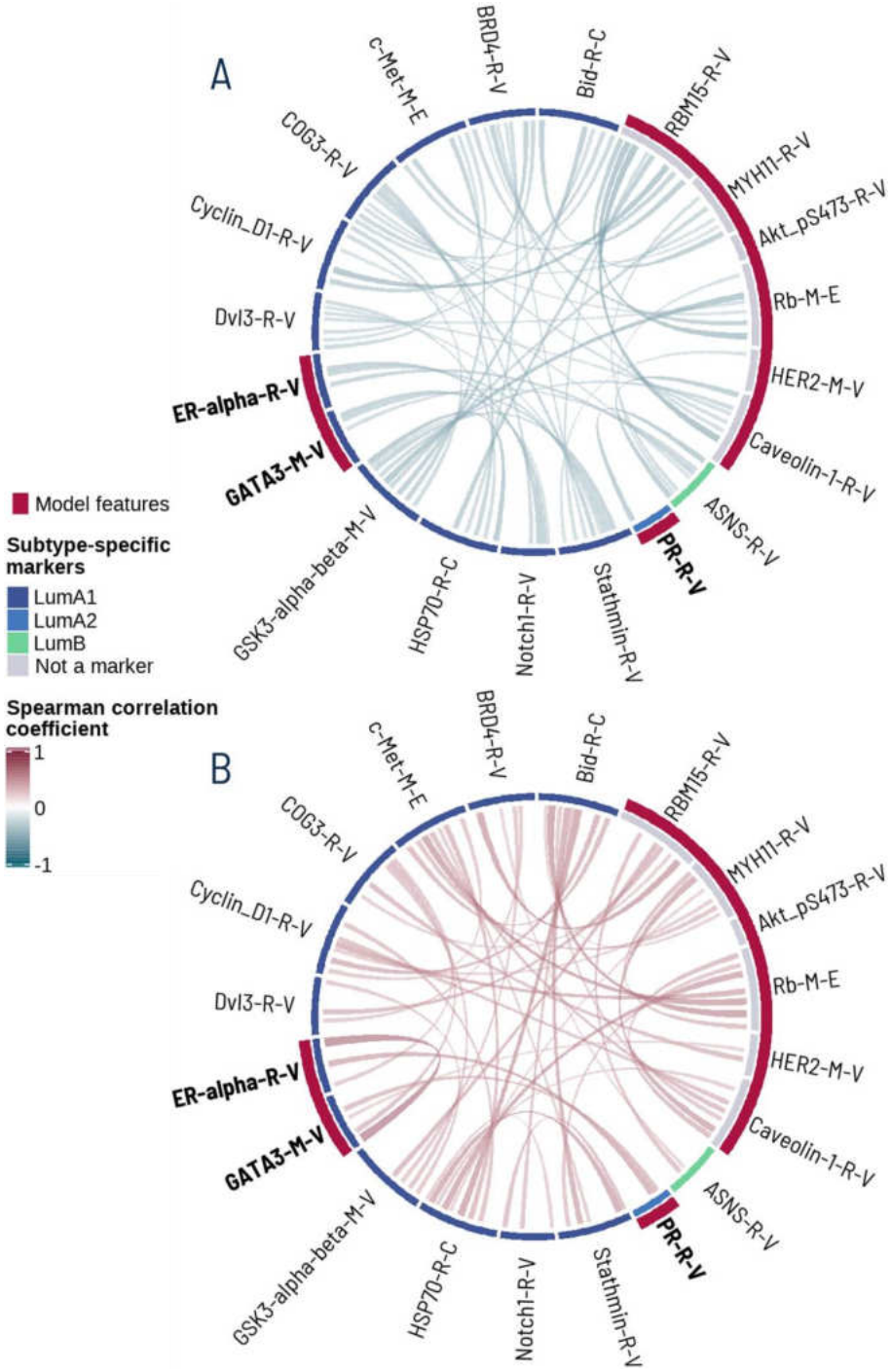
Supplementary Table 8.5 Cox proportional hazard analysis of identified luminal A subpopulations

Subtype	N	N_e	N_c	HR	HR effect	HR allocation probability adjusted critical value		
						α = 0.1	α = 0.3	α = 0.5
Overall Survival								
Luminal A1	44	8	36	1.237	No effect	HR < 0.751; HR > 1.331	HR < 0.439; HR > 2.276	HR < 0.251; HR > 3.977
Luminal A2	61	14	47	2.426	Medium	HR < 0.788; HR > 1.27	HR < 0.49; HR > 2.04	HR < 0.292; HR > 3.426
Luminal A3	87	9	78	Reference				
Disease-Specific Survival								
Luminal A1	43	5	38	3.518	Medium	HR < 0.751; HR > 1.331	HR < 0.439; HR > 2.276	HR < 0.251; HR > 3.977
Luminal A2	58	4	54	3.845	Large	HR < 0.785; HR > 1.274	HR < 0.486; HR > 2.057	HR < 0.289; HR > 3.466
Luminal A3	85	2	83	Reference				
Disease-Free Interval								
Luminal A1	38	4	34	3.686	Medium	HR < 0.745; HR > 1.342	HR < 0.431; HR > 2.32	HR < 0.245; HR > 4.079
Luminal A2	46	6	40	6.76	Large	HR < 0.768; HR > 1.302	HR < 0.462; HR > 2.165	HR < 0.269; HR > 3.717
Luminal A3	79	2	77	Reference				
Progression-Free Interval								
Luminal A1	44	7	37	2.325	Medium	HR < 0.751; HR > 1.331	HR < 0.439; HR > 2.276	HR < 0.251; HR > 3.977
Luminal A2	61	9	52	3.041	Medium	HR < 0.788; HR > 1.27	HR < 0.49; HR > 2.04	HR < 0.292; HR > 3.426
Luminal A3	87	5	82	Reference				



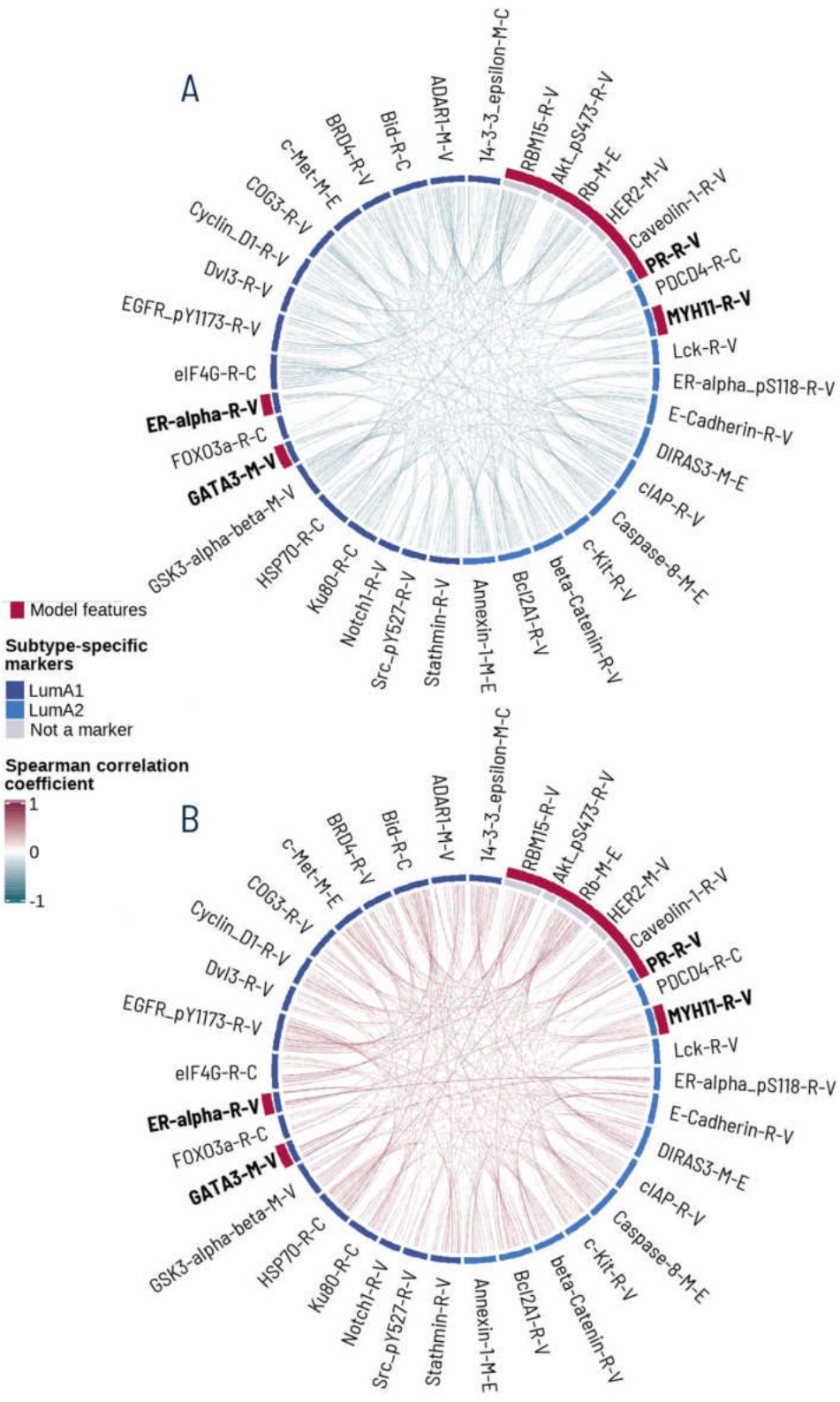
Supplementary Figure 8.4 Spearman rank correlation between protein levels for model-based proteomic signature and effect-size-based proteomic subtype-specific markers

Panel A shows links corresponding to Spearman rank correlation coefficient < -0.3. Panel B shows links corresponding to Spearman rank correlation coefficient > 0.3. "LumA1" denotes the luminal A1 subpopulation.



Supplementary Figure 8.5 Spearman rank correlation between protein levels for model-based proteomic signature and effect-size-based proteomic subtype-specific markers identified for the subset of luminal subpopulations

Panel A shows links corresponding to Spearman rank correlation coefficient < -0.3. Panel B shows links corresponding to Spearman rank correlation coefficient > 0.3. “LumA1”, “LumA2”, and “LumB” denote luminal A1, A2, and B subpopulations, respectively.



Supplementary Figure 8.6 Spearman rank correlation between protein levels for model-based proteomic signature and effect-size-based proteomic subtype-specific markers identified for the subset of luminal A subpopulations

Panel A shows links corresponding to Spearman rank correlation coefficient < -0.3. Panel B shows links corresponding to Spearman rank correlation coefficient > 0.3. "LumA1" and "LumA2" denote luminal A1 and A2 subpopulations, respectively.

Supplementary Table 8.6 Subtype prediction quality for all considered models with regard to training and test sets of the MRCV procedure

Metrics	Basal	HER2-enriched	Luminal A1	Luminal A2	Luminal A3	Luminal B
Proteomic model						
Training set						
Balanced Accuracy	100 ± 0	97.68 ± 2.19	99.74 ± 0.78	95.97 ± 2.96	96.65 ± 2.64	95 ± 3.5
Sensitivity	100 ± 0	95.8 ± 3.96	99.5 ± 1.51	92.91 ± 5.27	94.97 ± 3.99	91.97 ± 5.7
Specificity	100 ± 0	99.56 ± 0.45	99.98 ± 0.08	99.03 ± 0.72	98.33 ± 1.39	98.02 ± 1.37
Test set						
Balanced Accuracy	96.13 ± 4.31	89.64 ± 8.64	92.57 ± 8.69	90.2 ± 7.72	86.75 ± 6.96	86.29 ± 7.79
Sensitivity	93 ± 8.45	82.2 ± 16.79	87.25 ± 16.85	83.83 ± 15.44	79.11 ± 13.24	77.29 ± 15.74
Specificity	99.26 ± 1.58	97.09 ± 2.66	97.89 ± 2.5	96.56 ± 3.24	94.39 ± 4.34	95.3 ± 3.48
Transcriptomic model						
Training set						
Balanced Accuracy	95.65 ± 0.96	81.64 ± 2.25	53.16 ± 2.13	75.14 ± 2.05	71.2 ± 2.64	76.77 ± 2.73
Sensitivity	93.39 ± 1.85	68.43 ± 4.04	8.93 ± 4.43	58.65 ± 3.96	59.17 ± 4.29	63.53 ± 5.18
Specificity	97.9 ± 0.47	94.86 ± 0.76	97.4 ± 0.85	91.62 ± 0.78	83.22 ± 1.53	90.01 ± 1.05
Test set						
Balanced Accuracy	93.65 ± 4.8	78.04 ± 11.36	51.64 ± 7.78	67.59 ± 12.16	63.34 ± 10.95	79.24 ± 18.28
Sensitivity	89.56 ± 9.45	62.2 ± 22.54	10.5 ± 15.56	49.33 ± 20.91	45 ± 20.43	67 ± 32.66
Specificity	97.74 ± 2.81	93.89 ± 4.02	92.78 ± 5.38	85.85 ± 8.81	81.68 ± 6.78	91.48 ± 6.19
Combined model						
Training set						
Balanced Accuracy	100 ± 0	98.54 ± 1.97	99.76 ± 1.06	96.65 ± 2.96	97.17 ± 2.49	96.09 ± 3.39
Sensitivity	100 ± 0	97.41 ± 3.55	99.58 ± 1.92	94.15 ± 5.19	95.71 ± 3.82	93.62 ± 5.59
Specificity	100 ± 0	99.67 ± 0.44	99.95 ± 0.23	99.16 ± 0.78	98.63 ± 1.24	98.55 ± 1.25
Test set						
Balanced Accuracy	96.48 ± 4.14	89.3 ± 9.85	92.57 ± 9.03	90.24 ± 6.54	85.93 ± 6.58	84.56 ± 9.98
Sensitivity	93.89 ± 7.96	82.6 ± 19	86.75 ± 17.93	84.5 ± 13.24	77.67 ± 13.17	73.14 ± 19.82
Specificity	99.06 ± 1.91	96 ± 3.74	98.39 ± 2.1	95.97 ± 3.15	94.19 ± 4.23	95.97 ± 3.45

Acknowledgments

The results shown here are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

References

Akaike, H. (1974, 12). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723. doi:10.1109/tac.1974.1100705

Akbani, R., Ng, P. K., Werner, H. M., Shahmoradgoli, M., Zhang, F., Ju, Z., . . . Mills, G. B. (2014, 5). A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nature Communications, 5*. doi:10.1038/ncomms4887

Allison, K. H., Hammond, M. E., Dowsett, M., McKernin, S. E., Carey, L. A., Fitzgibbons, P. L., . . . Wolff, A. C. (2020, 4). Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update. *Journal of Clinical Oncology, 38*, 1346–1366. doi:10.1200/jco.19.02309

Allred, D. C., Carlson, R. W., Berry, D. A., Burstein, H. J., Edge, S. B., Goldstein, L. J., . . . Wolff, A. C. (2009, 9). NCCN Task Force Report: Estrogen Receptor and Progesterone Receptor Testing in Breast Cancer by Immunohistochemistry. *Journal of the National Comprehensive Cancer Network, 7*, S–1–S–21. doi:10.6004/jnccn.2009.0079

Argelaguet, R., Cuomo, A. S., Stegle, O., & Marioni, J. C. (2021, 5). Computational principles and challenges in single-cell data integration. *Nature Biotechnology, 39*, 1202–1215. doi:10.1038/s41587-021-00895-7

Arthur, D., & Vassilvitskii, S. (2006, 6). *k-means++: The Advantages of Careful Seeding.* Stanford InfoLab. Stanford. Retrieved from http://ilpubs.stanford.edu:8090/778/

Aziz, D., Lee, C., Chin, V., Fernandez, K. J., Phan, Z., Waring, P., & Caldon, C. E. (2022, 4). High cyclin E1 protein, but not gene amplification, is prognostic for basal-like breast cancer. *The Journal of Pathology: Clinical Research, 8*, 355–370. doi:10.1002/cjp2.269

Bartlett, M. S. (1937, 5). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences, 160*, 268–282. doi:10.1098/rspa.1937.0109

Bellman, R. (1961, 6). On the approximation of curves by line segments using dynamic programming. *Communications of the ACM, 4*, 284. doi:10.1145/366573.366611

Benjamini, Y., & Hochberg, Y. (1995, 1). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bertucci, F., Finetti, P., Cervera, N., Esterni, B., Hermitte, F., Viens, P., & Birnbaum, D. (2008). How basal are triple-negative breast cancers? *International Journal of Cancer, 123*, 236–240. doi:10.1002/ijc.23518

Bièche, I., Girault, I., Urbain, E., Tozlu, S., & Lidereau, R. (2004, 3). Relationship between intratumoral expression of genes coding for xenobiotic-metabolizing enzymes and

benefit from adjuvant tamoxifen in estrogen receptor alpha-positive postmenopausal breast carcinoma. *Breast Cancer Research, 6*. doi:10.1186/bcr784

BioXpedia. (2023). *What is NanoString nCounter*. Retrieved from https://www.bioxpedia.com/nanostring-ncounter-technology/

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008, 10). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*, P10008. doi:10.1088/1742-5468/2008/10/p10008

Boutilier, A. J., & Elsawa, S. F. (2021, 6). Macrophage Polarization States in the Tumor Microenvironment. *International Journal of Molecular Sciences, 22*, 6995. doi:10.3390/ijms22136995

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018, 9). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 68*, 394–424. doi:10.3322/caac.21492

Breslow, N. (1970). A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika, 57*, 579–594. doi:10.1093/biomet/57.3.579

Brunet, J.-P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004, 3). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences, 101*, 4164–4169. doi:10.1073/pnas.0308531101

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining* (pp. 160–172). Springer Berlin Heidelberg. doi:10.1007/978-3-642-37456-2_14

Charafe-Jauffret, E., Ginestier, C., Monville, F., Finetti, P., Adélaïde, J., Cervera, N., . . . Bertucci, F. (2005, 11). Gene expression profiling of breast cell lines identifies potential new basal markers. *Oncogene, 25*, 2273–2284. doi:10.1038/sj.onc.1209254

Cho, N. (2016, 10). Molecular subtypes and imaging phenotypes of breast cancer. *Ultrasonography, 35*, 281–288. doi:10.14366/usg.16030

Cohen, J. (2013, 5). *Statistical Power Analysis for the Behavioral Sciences.* Routledge. doi:10.4324/9780203771587

Conover, W., & Iman, R. (1979, 2). *Multiple-comparisons procedures. Informal report.* Office of Scientific and Technical Information (OSTI). doi:10.2172/6057803

Coombes, K. R. (2012). *Classes and methods for "class discovery" with microarrays or proteomics*. Retrieved from R package version 2.13.4.

Cox, D. R. (1972, 1). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*, 187–202. doi:10.1111/j.2517-6161.1972.tb00899.x

Cramér, H. (1999, 4). *Mathematical Methods of Statistics (PMS-9), Volume 9.* Princeton University Press. Retrieved from https://www.ebook.de/de/product/3646411/harald_cramer_mathematical_methods_of_statistics_pms_9_volume_9.html

Dabney, A. R. (2005, 9). Classification of microarrays to nearest centroids. *Bioinformatics, 21*, 4148–4154. doi:10.1093/bioinformatics/bti681

Daemen, A., & Manning, G. (2018, 1). HER2 is not a cancer subtype but rather a pan-cancer event and is highly enriched in AR-driven breast tumors. *Breast Cancer Research, 20*. doi:10.1186/s13058-018-0933-y

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research, 5*(10), 2929–2943.

Dice, L. R. (1945, 7). Measures of the Amount of Ecologic Association Between Species. *Ecology, 26*, 297–302. doi:10.2307/1932409

Doane, A. S., Danso, M., Lal, P., Donaton, M., Zhang, L., Hudis, C., & Gerald, W. L. (2006, 2). An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. *Oncogene, 25*, 3994–4008. doi:10.1038/sj.onc.1209415

Dressman, M. A., Baras, A., Malinowski, R., Alvis, L. B., Kwon, I., Walz, T. M., & Polymeropoulos, M. H. (2003, 5). Gene expression profiling detects gene amplification and differentiates tumor types in breast cancer. *Cancer research, 63*(9), 2194–2199.

Duan, R., Gao, L., Gao, Y., Hu, Y., Xu, H., Huang, M., . . . Jia, S. (2021, 8). Evaluation and comparison of multi-omics data integration methods for cancer subtyping. (E. Wang, Ed.) *PLOS Computational Biology, 17*, e1009224. doi:10.1371/journal.pcbi.1009224

Elian, F. A., Yan, E., & Walter, M. A. (2017, 11). FOXC1, the new player in the cancer sandbox. *Oncotarget, 9*, 8165–8178. doi:10.18632/oncotarget.22742

Elloumi, F., Hu, Z., Li, Y., Parker, J. S., Gulley, M. L., Amos, K. D., & Troester, M. A. (2011, 6). Systematic Bias in Genomic Classification Due to Contaminating Non-neoplastic Tissue in Breast Tumor Samples. *BMC Medical Genomics, 4*. doi:10.1186/1755-8794-4-54

El-Rehim, D. M., Ball, G., Pinder, S. E., Rakha, E., Paish, C., Robertson, J. F., . . . Ellis, I. O. (2005). High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *International Journal of Cancer, 116*, 340–350. doi:10.1002/ijc.21004

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., . . . D\textquotesingleEustachio, P. (2015, 12). The Reactome pathway Knowledgebase. *Nucleic Acids Research, 44*, D481–D487. doi:10.1093/nar/gkv1351

Fabregat, A., Sidiropoulos, K., Viteri, G., Forner, O., Marin-Garcia, P., Arnau, V., . . . Hermjakob, H. (2017, 3). Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics, 18*. doi:10.1186/s12859-017-1559-2

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., . . . Iggo, R. (2005, 5). Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene, 24*, 4660–4671. doi:10.1038/sj.onc.1208561

Fragomeni, S. M., Sciallis, A., & Jeruss, J. S. (2018, 1). Molecular Subtypes and Local-Regional Control of Breast Cancer. *Surgical Oncology Clinics of North America, 27*, 95–120. doi:10.1016/j.soc.2017.08.005

Fujiwara, N., & Kobayashi, K. (2005, 6). Macrophages in Inflammation. *Current Drug Target -Inflammation & Allergy, 4*, 281–286. doi:10.2174/1568010054022024

Fumagalli, D., Blanchet-Cohen, A., Brown, D., Desmedt, C., Gacquer, D., Michiels, S., . . . Haibe-Kains, B. (2014, 11). Transfer of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-Sequencing technology. *BMC Genomics, 15*. doi:10.1186/1471-2164-15-1008

Garrett, J. T., & Arteaga, C. L. (2011, 5). Resistance to HER2-directed antibodies and tyrosine kinase inhibitors. *Cancer Biology & Therapy, 11*, 793–800. doi:10.4161/cbt.11.9.15045

GDC Data Transfer Tool. (2020). *https://docs.gdc.cancer.gov/Data_Transfer_Tool/Users_Guide/Getting_Started/*.

Gehan, E. A. (1965, 6). A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. *Biometrika, 52*, 203. doi:10.2307/2333825

Genomic Data Commons Data Portal. (2022). *https://portal.gdc.cancer.gov/*. Retrieved from https://portal.gdc.cancer.gov/

Genomic Data Commons Legacy Archive. (2021). *https://portal.gdc.cancer.gov/legacy-archive*.

Godoy-Ortiz, A., Sanchez-Muñoz, A., Parrado, M. R., Álvarez, M., Ribelles, N., Dominguez, A. R., & Alba, E. (2019, 10). Deciphering HER2 Breast Cancer Disease: Biological and Clinical Implications. *Frontiers in Oncology, 9*. doi:10.3389/fonc.2019.01124

Gonzalez-Angulo, A. M., Hennessy, B. T., Meric-Bernstam, F., Sahin, A., Liu, W., Ju, Z., . . . Mills, G. B. (2011, 7). Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. *Clinical Proteomics, 8*. doi:10.1186/1559-0275-8-11

Guedj, M., Marisa, L., de Reynies, A., Orsetti, B., Schiappa, R., Bibeau, F., . . . Theillet, C. (2011, 7). A refined molecular taxonomy of breast cancer. *Oncogene, 31*, 1196–1206. doi:10.1038/onc.2011.301

Guiu, S., Michiels, S., André, F., Cortes, J., Denkert, C., Leo, A. D., . . . Reis-Filho, J. S. (2012, 12). Molecular subclasses of breast cancer: how do we define them? The IMPAKT 2012 Working Group Statement. *Annals of Oncology, 23*, 2997–3006. doi:10.1093/annonc/mds586

Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D. C., & Shyr, Y. (2013, 8). Large Scale Comparison of Gene Expression Levels by Microarrays and RNAseq Using TCGA Data. (P. Provero, Ed.) *PLoS ONE, 8*, e71462. doi:10.1371/journal.pone.0071462

Győrffy, B., Hatzis, C., Sanft, T., Hofstatter, E., Aktas, B., & Pusztai, L. (2015, 1). Multigene prognostic tests in breast cancer: past, present, future. *Breast Cancer Research, 17*. doi:10.1186/s13058-015-0514-2

Haynes, W. (2013). Bonferroni Correction. In *Encyclopedia of Systems Biology* (pp. 154–154). Springer New York. doi:10.1007/978-1-4419-9863-7_1213

Hennessy, B. T., Lu, Y., Poradosu, E., Yu, Q., Yu, S., Hall, H., . . . Mills, G. B. (2007, 12). Pharmacodynamic Markers of Perifosine Efficacy. *Clinical Cancer Research, 13*, 7421–7431. doi:10.1158/1078-0432.ccr-07-0760

Henzel, J., Tobiasz, J., Kozielski, M., Bach, M., Foszner, P., Gruca, A., . . . Sikora, M. (2021, 11). Screening Support System Based on Patient Survey Data—Case Study on Classification of Initial, Locally Collected COVID-19 Data. *Applied Sciences, 11*, 10790. doi:10.3390/app112210790

Herschkowitz, J. I., He, X., Fan, C., & Perou, C. M. (2008, 9). The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas. *Breast Cancer Research, 10*. doi:10.1186/bcr2142

Hotelling, H. (1933, 10). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*, 498–520. doi:10.1037/h0070888

Hu, J., He, X., Baggerly, K. A., Coombes, K. R., Hennessy, B. T., & Mills, G. B. (2007, 6). Non-parametric quantification of protein lysate arrays. *Bioinformatics, 23*, 1986–1994. doi:10.1093/bioinformatics/btm283

Hu, L., Ru, K., Zhang, L., Huang, Y., Zhu, X., Liu, H., . . . Miao, W. (2014, 2). Fluorescence in situ hybridization (FISH): an increasingly demanded tool for biomarker research and personalized medicine. *Biomarker Research, 2*. doi:10.1186/2050-7771-2-3

Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., . . . Perou, C. M. (2006, 4). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics, 7*. doi:10.1186/1471-2164-7-96

Huang, S., Chaudhary, K., & Garmire, L. X. (2017, 6). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in Genetics, 8*. doi:10.3389/fgene.2017.00084

Huo, D., Hu, H., Rhie, S. K., Gamazon, E. R., Cherniack, A. D., Liu, J., . . . Olopade, O. I. (2017, 12). Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncology, 3*, 1654. doi:10.1001/jamaoncol.2017.0595

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., . . . Tsai, T. T. (2005, 2). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters, 12*, 105–108. doi:10.1109/lsp.2001.838216

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021, 7). *An Introduction to Statistical Learning.* Springer-Verlag GmbH. Retrieved from https://www.ebook.de/de/product/40099270/gareth_james_daniela_witten_trevor_hastie_robert_tibshirani_an_introduction_to_statistical_learning.html

Jassem, J., Shan, A., & Buczek, D. (2020, 12). Changing paradigms in breast cancer treatment. *European Journal of Translational and Clinical Medicine, 3*, 53–63. doi:10.31373/ejtcm/130486

Jeffreys, H. (1998, 8). *Theory of Probability.* OUP Oxford. Retrieved from https://www.ebook.de/de/product/3605842/harold_jeffreys_theory_of_probability.html

Johnson, W. E., Li, C., & Rabinovic, A. (2006, 4). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics, 8*, 118–127. doi:10.1093/biostatistics/kxj037

Jönsson, G., Staaf, J., Vallon-Christersson, J., Ringnér, M., Holm, K., Hegardt, C., . . . Borg, Å. (2010, 6). Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Research, 12*. doi:10.1186/bcr2596

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2016, 11). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research, 45*, D353–D361. doi:10.1093/nar/gkw1092

Kaplan, E. L., & Meier, P. (1992). Nonparametric Estimation from Incomplete Observations. In *Springer Series in Statistics* (pp. 319–337). Springer New York. doi:10.1007/978-1-4612-4380-9_25

Kassambara, A., Kosinski, M., & Biecek, P. (2021). *survminer: Drawing Survival Curves using 'ggplot2'.* Retrieved from https://CRAN.R-project.org/package=survminer

Kolassa, J. E., & Zhang, J. (2023). *PHInfiniteEstimates: Tools for Inference in the Presence of a Monotone Likelihood.* Retrieved from https://CRAN.R-project.org/package=PHInfiniteEstimates

Kozielski, M., Henzel, J., Tobiasz, J., Gruca, A., Foszner, P., Zyla, J., . . . others. (2021). Enhancement of COVID-19 symptom-based screening with quality-based classifier optimisation. *Bulletin of the Polish Academy of Sciences. Technical Sciences, 69*.

Larsen, M. J., Thomassen, M., Tan, Q., Sørensen, K. P., & Kruse, T. A. (2014). Microarray-Based RNA Profiling of Breast Cancer: Batch Effect Removal Improves Cross-Platform Consistency. *BioMed Research International, 2014*, 1–11. doi:10.1155/2014/651751

Leek, J., Johnson, W., Parker, H., Fertig, E., Jaffe, A., Zhang, Y., . . . Torres, L. (2017). sva. *sva*. Bioconductor. doi:10.18129/B9.BIOC.SVA

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., & Pietenpol, J. A. (2011, 7). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation, 121*, 2750–2767. doi:10.1172/jci45014

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015, 12). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems, 1*, 417–425. doi:10.1016/j.cels.2015.12.004

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011, 5). Molecular signatures database (MSigDB) 3.0. *Bioinformatics, 27*, 1739–1740. doi:10.1093/bioinformatics/btr260

Lin, H. H., Chung, Y., Cheng, C.-T., Ouyang, C., Fu, Y., Kuo, C.-Y., . . . Ann, D. K. (2018, 8). Autophagic reliance promotes metabolic reprogramming in oncogenic KRAS-driven tumorigenesis. *Autophagy, 14*, 1481–1498. doi:10.1080/15548627.2018.1450708

Liu, J., Geng, X., Hou, J., & Wu, G. (2021, 7). New insights into M1/M2 macrophages: key modulators in cancer progression. *Cancer Cell International, 21*. doi:10.1186/s12935-021-02089-2

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., . . . Mariamidze, A. (2018, 4). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell, 173*, 400–416.e11. doi:10.1016/j.cell.2018.02.052

Liu, Q., Cheng, B., Jin, Y., & Hu, P. (2022, 1). Bayesian tensor factorization-drive breast cancer subtyping by integrating multi-omics data. *Journal of Biomedical Informatics, 125*, 103958. doi:10.1016/j.jbi.2021.103958

Llobet, S. G., van der Vegt, B., Jongeneel, E., Bense, R. D., Zwager, M. C., Schröder, C. P., . . . van Vugt, M. A. (2020, 9). Cyclin E expression is associated with high levels of replication stress in triple-negative breast cancer. *npj Breast Cancer, 6*. doi:10.1038/s41523-020-00181-w

Lun, A. (2021). *bluster: Clustering Algorithms for Bioconductor*. Retrieved from R package version 1.8.0.

Mantel, N. (1966, 3). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports, 50*(3), 163–170.

Marczyk, M., Jaksik, R., Polanski, A., & Polanska, J. (2019). GaMRed – adaptive filtering of high-throughput biological data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1–1. doi:10.1109/tcbb.2018.2858825

Martin, E. M., Orlando, K. A., Yokobori, K., & Wade, P. A. (2021, 12). The estrogen receptor/GATA3/FOXA1 transcriptional network: lessons learned from breast cancer. *Current Opinion in Structural Biology, 71*, 65–70. doi:10.1016/j.sbi.2021.05.015

May, S., Hosmer, D. W., & Lemeshow, S. (2014, 3). *Applied Survival Analysis.* John Wiley & Sons. Retrieved from https://www.ebook.de/de/product/7746386/susanne_may_david_w_jr_hosmer_stanley_lemeshow_applied_survival_analysis.html

McInnes, L., & Healy, J. (2017). Accelerated Hierarchical Density Based Clustering. *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*, (pp. 33–42).

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software, 2*, 205.

McInnes, L., Healy, J., & Melville, J. (2018, 2). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

MD Anderson Cancer Center. (2020). *TCGA Batch Effects Viewer*. Retrieved from https://bioinformatics.mdanderson.org/public-software/tcga-batch-effects/

Milioli, H. H., Alexandrou, S., Lim, E., & Caldon, C. E. (2020, 5). Cyclin E1 and cyclin E2 in ER+ breast cancer: prospects as biomarkers and therapeutic targets. *Endocrine-Related Cancer, 27*, R93–R112. doi:10.1530/erc-19-0501

Miricescu, D., Totan, A., Stanescu-Spinu, I.-I., Badoiu, S. C., Stefani, C., & Greabu, M. (2020, 12). PI3K/AKT/mTOR Signaling Pathway in Breast Cancer: From Molecular Landscape to Clinical Aspects. *International Journal of Molecular Sciences, 22*, 173. doi:10.3390/ijms22010173

Moasser, M. M. (2007, 4). The oncogene HER2: its signaling and transforming functions and its role in human cancer pathogenesis. *Oncogene, 26*, 6469–6487. doi:10.1038/sj.onc.1210477

Morgan, M., & Davis, S. (2021). *GenomicDataCommons: NIH / NCI Genomic Data Commons Access.* Retrieved from https://bioconductor.org/packages/GenomicDataCommons, http://github.com/Bioconductor/GenomicDataCommons

Mott, L., Su, K., & Pack, D. W. (2018, 3). Evaluation of FOXC1 as a therapeutic target for basal-like breast cancer. *Cancer Gene Therapy, 25*, 84–91. doi:10.1038/s41417-018-0010-9

Mrukwa, G., & Polanska, J. (2022, 12). DiviK: divisive intelligent K-means for hands-free unsupervised clustering in big biological data. *BMC Bioinformatics, 23*. doi:10.1186/s12859-022-05093-z

Mueller, C., Haymond, A., Davis, J. B., Williams, A., & Espina, V. (2018, 1). Protein biomarkers for subtyping breast cancer and implications for future research. *Expert Review of Proteomics, 15*, 131–152. doi:10.1080/14789450.2018.1421071

Nassar, A., Khoor, A., Radhakrishnan, R., Radhakrishnan, A., & Cohen, C. (2014). Correlation of HER2 overexpression with gene amplification and its relation to chromosome 17 aneuploidy: a 5-year experience with invasive ductal and lobular carcinomas. *International journal of clinical and experimental pathology, 7*(9), 6254–6261.

National Cancer Institute. (2020). *Center for Cancer Genomics*. Retrieved from NCI's Genome Characterization Pipeline: https://www.cancer.gov/about-nci/organization/ccg/research/genomic-pipeline

National Cancer Institute. (2022). *Dictionary of Cancer Terms*. Retrieved from AJCC staging system: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/ajcc-staging-system

National Center for Biotechnology Information. (2023). *National Library of Medicine*. Retrieved from Real-Time qRT-PCR: https://www.ncbi.nlm.nih.gov/probe/docs/techqpcr/

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., . . . Alizadeh, A. A. (2015, 3). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods, 12*, 453–457. doi:10.1038/nmeth.3337

Norum, J. H., Andersen, K., & Sørlie, T. (2014, 5). Lessons learned from the intrinsic subtypes of breast cancer in the quest for precision therapy. *British Journal of Surgery, 101*, 925–938. doi:10.1002/bjs.9562

Olivier, J., May, W. L., & Bell, M. L. (2017, 3). Relative effect sizes for measures of risk. *Communications in Statistics - Theory and Methods, 46*, 6774–6781. doi:10.1080/03610926.2015.1134575

Osborne, C. K., Yochmowitz, M. G., Knight, W. A., & McGuire, W. L. (1980, 12). The value of estrogen and progesterone receptors in the treatment of breast cancer. *Cancer, 46*, 2884–2888. doi:10.1002/1097-0142(19801215)46:12+<2884::aid-cncr2820461429>3.0.co;2-u

Papiez, A., Marczyk, M., Polanska, J., & Polanski, A. (2018, 10). BatchI: Batch effect Identification in high-throughput screening data using a dynamic programming algorithm. (B. Berger, Ed.) *Bioinformatics, 35*, 1885–1892. doi:10.1093/bioinformatics/bty900

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., . . . Bernard, P. S. (2009, 3). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology, 27*, 1160–1167. doi:10.1200/jco.2008.18.1370

Perou, C. M., & Borresen-Dale, A.-L. (2010, 11). Systems Biology and Genomics of Breast Cancer. *Cold Spring Harbor Perspectives in Biology, 3*, a003293–a003293. doi:10.1101/cshperspect.a003293

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., . . . Botstein, D. (2000, 8). Molecular portraits of human breast tumours. *Nature, 406*, 747–752. doi:10.1038/35021093

Peto, R., & Peto, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society. Series A (General), 135*, 185. doi:10.2307/2344317

Prat, A., Parker, J. S., Fan, C., Cheang, M. C., Miller, L. D., Bergh, J., . . . Perou, C. M. (2012, 11). Concordance among gene expression-based predictors for ER-positive breast cancer treated with adjuvant tamoxifen. *Annals of Oncology, 23*, 2866–2873. doi:10.1093/annonc/mds080

Prat, A., Pineda, E., Adamo, B., Galván, P., Fernández, A., Gaba, L., . . . Muñoz, M. (2015, 11). Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast, 24*, S26–S35. doi:10.1016/j.breast.2015.07.008

Qian, X.-L., Pan, Y.-H., Huang, Q.-Y., Shi, Y.-B., Huang, Q.-Y., Hu, Z.-Z., & Xiong, L.-X. (2019, 2). Caveolin-1: a multifaceted driver of breast cancer progression and its application in clinical treatment. *OncoTargets and Therapy, Volume 12*, 1539–1552. doi:10.2147/ott.s191317

Qin, C., Yang, X., & Zhan, Z. (2020, 10). High Expression of Asparagine Synthetase Is Associated with Poor Prognosis of Breast Cancer in Chinese Population. *Cancer Biotherapy and Radiopharmaceuticals, 35*, 581–585. doi:10.1089/cbr.2019.3295

Rappoport, N., & Shamir, R. (2018, 10). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research, 46*, 10546–10562. doi:10.1093/nar/gky889

Reese, S. E., Archer, K. J., Therneau, T. M., Atkinson, E. J., Vachon, C. M., de Andrade, M., . . . Eckel-Passow, J. E. (2013, 8). A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics, 29*, 2877–2883. doi:10.1093/bioinformatics/btt480

Ren, L., Zhou, P., Wu, H., Liang, Y., Xu, R., Lu, H., & Chen, Q. (2021, 8). Caveolin-1 is a prognostic marker and suppresses the proliferation of breast cancer. *Translational Cancer Research, 10*, 3797–3810. doi:10.21037/tcr-21-1139

Risso, G., Blaustein, M., Pozzi, B., Mammi, P., & Srebrow, A. (2015, 5). Akt/PKB: one kinase, many modifications. *Biochemical Journal, 468*, 203–214. doi:10.1042/bj20150041

Robinson, T. J., Liu, J. C., Vizeacoumar, F., Sun, T., Maclean, N., Egan, S. E., . . . Zacksenhaus, E. (2013, 11). RB1 Status in Triple Negative Breast Cancer Cells Dictates Response to Radiation Treatment and Selective Therapeutic Drugs. (A. Ahmad, Ed.) *PLoS ONE, 8*, e78641. doi:10.1371/journal.pone.0078641

Sali, A. P., Sharma, N., Verma, A., Beke, A., Shet, T., Patil, A., . . . Desai, S. B. (2020, 10). Identification of Luminal Subtypes of Breast Carcinoma Using Surrogate Immunohistochemical Markers and Ascertaining Their Prognostic Relevance. *Clinical Breast Cancer, 20*, 382–389. doi:10.1016/j.clbc.2020.03.012

Savage, K., Lambros, M. B., Robertson, D., Jones, R. L., Jones, C., Mackay, A., . . . Reis-Filho, J. S. (2007, 1). Caveolin 1 Is Overexpressed and Amplified in a Subset of Basal-like and Metaplastic Breast Carcinomas: A Morphologic, Ultrastructural, Immunohistochemical, and in situ Hybridization Analysis. *Clinical Cancer Research, 13*, 90–101. doi:10.1158/1078-0432.ccr-06-1371

Sawilowsky, S. S. (2009, 11). New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods, 8*, 597–599. doi:10.22237/jmasm/1257035100

Schwarz, G. (1978, 3). Estimating the Dimension of a Model. *The Annals of Statistics, 6*. doi:10.1214/aos/1176344136

Shapiro, S. S., & Wilk, M. B. (1965, 12). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika, 52*, 591. doi:10.2307/2333709

Sienkiewicz, K., Chen, J., Chatrath, A., Lawson, J. T., Sheffield, N. C., Zhang, L., & Ratan, A. (2022, 1). Detecting molecular subtypes from multi-omics datasets using SUMO. *Cell Reports Methods, 2*, 100152. doi:10.1016/j.crmeth.2021.100152

Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A., & McGuire, W. L. (1987, 1). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science, 235*, 177–182. doi:10.1126/science.3798106

Smid, M., Wang, Y., Zhang, Y., Sieuwerts, A. M., Yu, J., Klijn, J. G., . . . Martens, J. W. (2008, 5). Subtypes of Breast Cancer Show Preferential Site of Relapse. *Cancer Research, 68*, 3108–3114. doi:10.1158/0008-5472.can-07-5644

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., . . . Børresen-Dale, A.-L. (2001, 9). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences, 98*, 10869–10874. doi:10.1073/pnas.191367098

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., . . . Botstein, D. (2003, 6). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences, 100*, 8418–8423. doi:10.1073/pnas.0932692100

Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., . . . Liu, E. T. (2003, 8). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences, 100*, 10393–10398. doi:10.1073/pnas.1732912100

Stevens, T. A., & Meech, R. (2006, 4). {BARX2 and estrogen receptor-alpha (ESR1) coordinately regulate the production of alternatively spliced ESR1 isoforms and control breast cancer cell growth and invasion. *Oncogene, 25*, 5426–5435. doi:10.1038/sj.onc.1209529

Stuart, T., & Satija, R. (2019, 1). Integrative single-cell analysis. *Nature Reviews Genetics, 20*, 257–272. doi:10.1038/s41576-019-0093-7

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005, 9). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences, 102*, 15545–15550. doi:10.1073/pnas.0506580102

Szymiczek, A., Lone, A., & Akbari, M. R. (2020, 12). Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review. *Clinical Genetics, 99*, 613–637. doi:10.1111/cge.13900

Takaku, M., Grimm, S. A., & Wade, P. A. (2015, 12). GATA3 in Breast Cancer: Tumor Suppressor or Oncogene? *Gene Expression, 16*, 163–168. doi:10.3727/105221615x14399878166113

The Cancer Genome Atlas Network. (2011, 6). Integrated genomic analyses of ovarian carcinoma. *Nature, 474*, 609–615. doi:10.1038/nature10166

The Cancer Genome Atlas Network. (2012, 9). Comprehensive molecular portraits of human breast tumours. *Nature, 490*, 61–70. doi:10.1038/nature11412

The Human Protein Atlas. (2023). *Immunohistochemistry*. Retrieved from The Human Protein Atlas: https://www.proteinatlas.org/learn/method/immunohistochemistry

Therneau, T. M. (2021). *A Package for Survival Analysis in R.* Retrieved from https://CRAN.R-project.org/package=survival

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Yang, T.-H. O., . . . Mariamidze, A. (2018, 4). The Immune Landscape of Cancer. *Immunity, 48*, 812–830.e14. doi:10.1016/j.immuni.2018.03.023

Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G. B., & Kornblau, S. M. (2006, 10). Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular Cancer Therapeutics, 5*, 2512–2521. doi:10.1158/1535-7163.mct-06-0334

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002, 5). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences, 99*, 6567–6572. doi:10.1073/pnas.082099299

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*, 411–423. doi:10.1111/1467-9868.00293

Tobiasz, J., & Polanska, J. (2022). How to Compare Various Clustering Outcomes? Metrices to Investigate Breast Cancer Patient Subpopulations Based on Proteomic Profiles. In *Bioinformatics and Biomedical Engineering* (pp. 309–318). Springer International Publishing. doi:10.1007/978-3-031-07802-6_26

Tobiasz, J., Hatzis, C., & Polanska, J. (2019, 10). Breast Cancer Heterogeneity Investigation: Multiple k-Means Clustering Approach. *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE).* IEEE. doi:10.1109/bibe.2019.00080

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in sport sciences, 21*.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., . . . Friend, S. H. (2002, 1). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*, 530–536. doi:10.1038/415530a

Vieira, A. F., & Schmitt, F. (2018, 9). An Update on Breast Cancer Multigene Prognostic Tests—Emergent Clinical Biomarkers. *Frontiers in Medicine, 5*. doi:10.3389/fmed.2018.00248

Wagenmakers, E.-J. (2007, 10). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review, 14*, 779–804. doi:10.3758/bf03194105

Wallden, B., Storhoff, J., Nielsen, T., Dowidar, N., Schaper, C., Ferree, S., . . . Parker, J. S. (2015, 8). Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Medical Genomics, 8*. doi:10.1186/s12920-015-0129-6

Wang, N., Liang, H., & Zen, K. (2014, 11). Molecular Mechanisms That Influence the Macrophage M1-M2 Polarization Balance. *Frontiers in Immunology, 5*. doi:10.3389/fimmu.2014.00614

Wei, Y., Li, L., Zhao, X., Yang, H., Sa, J., Cao, H., & Cui, Y. (2022, 11). Cancer subtyping with heterogeneous multi-omics data via hierarchical multi-kernel learning. *Briefings in Bioinformatics, 24*. doi:10.1093/bib/bbac488

Weigelt, B., Mackay, A., A\textquotesinglehern, R., Natrajan, R., Tan, D. S., Dowsett, M., . . . Reis-Filho, J. S. (2010, 4). Breast cancer molecular profiling with single sample

predictors: a retrospective analysis. *The Lancet Oncology, 11*, 339–349. doi:10.1016/s1470-2045(10)70008-5

Weiner, J. (2022). *tmod: Feature Set Enrichment Analysis for Metabolomics and Transcriptomics.* Retrieved from https://CRAN.R-project.org/package=tmod

Wolff, A. C., Hammond, M. E., Allison, K. H., Harvey, B. E., Mangu, P. B., Bartlett, J. M., . . . Dowsett, M. (2018, 5). Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *Archives of Pathology & Laboratory Medicine, 142*, 1364–1382. doi:10.5858/arpa.2018-0902-sa

World Health Organization. (2023). *Breast cancer*. Retrieved from World Health Organization: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Wynn, T. A., Chawla, A., & Pollard, J. W. (2013, 4). Macrophage biology in development, homeostasis and disease. *Nature, 496*, 445–455. doi:10.1038/nature12034

Yang, F., Foekens, J. A., Yu, J., Sieuwerts, A. M., Timmermans, M., Klijn, J. G., . . . Jiang, Y. (2005, 10). Laser microdissection and microarray analysis of breast tumors reveal ER-alpha related genes and pathways. *Oncogene, 25*, 1413–1419. doi:10.1038/sj.onc.1209165

Yates, F. (1934). Contingency Tables Involving Small Numbers and the $\upchi$ 2 Test. *Supplement to the Journal of the Royal Statistical Society, 1*, 217. doi:10.2307/2983604

Yi, L., Lei, Y., Yuan, F., Tian, C., Chai, J., & Gu, M. (2022, 6). NTN4 as a prognostic marker and a hallmark for immune infiltration in breast cancer. *Scientific Reports, 12*. doi:10.1038/s41598-022-14575-2

Zaha, D. C. (2014). Significance of immunohistochemistry in breast cancer. *World Journal of Clinical Oncology, 5*, 382. doi:10.5306/wjco.v5.i3.382

Zhang, L.-Y., Zhang, Y.-Q., Zeng, Y.-Z., Zhu, J.-L., Chen, H., Wei, X.-L., & Liu, L.-J. (2020, 5). TRPC1 inhibits the proliferation and migration of estrogen receptor-positive Breast cancer and gives a better prognosis by inhibiting the PI3K/AKT pathway. *Breast Cancer Research and Treatment, 182*, 21–33. doi:10.1007/s10549-020-05673-8

Zhang, Y., & Kiryu, H. (2022, 9). MODEC: an unsupervised clustering method integrating omics data for identifying cancer subtypes. *Briefings in Bioinformatics, 23*. doi:10.1093/bib/bbac372

Zheng, F., Du, F., Zhao, J., Wang, X., Si, Y., Jin, P., . . . Yuan, P. (2021, 5). The emerging role of RNA N6-methyladenosine methylation in breast cancer. *Biomarker Research, 9*. doi:10.1186/s40364-021-00295-8

Zyla, J., Marczyk, M., Domaszewska, T., Kaufmann, S. H., Polanska, J., & Weiner, J. (2019, 6). Gene set enrichment for reproducible science: comparison of CERNO and eight other algorithms. (J. Wren, Ed.) *Bioinformatics, 35*, 5146–5154. doi:10.1093/bioinformatics/btz447

Tables

Figures

Abbreviations

2D
Two-dimensional,

AIC
Akaike Information Criterion,

AJCC
American Joint Committee on Cancer,

ANOVA
Analysis of Variance,

ASCO/CAP
American Society of Clinical Oncology/College of American Pathologists,

AUC
Area Under Curve,

BA
Balanced Accuracy,

BF
Bayes Factor,

BIC
Bayesian Information Criterion,

BRCA
Breast Invasive Carcinoma,

cDNA
Complementary DNA,

CI
Confidence Interval,

Abbreviations

DBSCAN

Density-Based Spatial Clustering of Applications with Noise,

DFI

Disease-Free Interval,

DiviK

Divisive intelligent K-means,

DSS

Disease-Specific Survival,

ER

Estrogen receptor,

FDR

False Discovery Rate,

FISH

Fluorescent in situ hybridization,

GDC

Genomic Data Commons,

GMM

Gaussian Mixture Models,

HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise,

HER2

Human epidermal growth factor receptor 2,

HR

Hazard Ratio,

Hormone receptor,

ID

Identifier,

IHC

Immunohistochemistry,

KEGG

Kyoto Encyclopedia of Genes and Genomes,

KM

Kaplan-Meier,

M

AJCC pathologic field Metastasis,

MRCV

Multiple Random Cross-Validation,

MSigDB

Molecular Signatures Database,

N

AJCC pathologic field Nodes,

NGS

Next Generation Sequencing,

NTE

New Tumor Event,

ORA

Over-Representation Analysis,

OS

Overall Survival,

PAM

Prediction Analysis of Microarray,

PAM50

50-gene Prediction Analysis of Microarray,

PC

Principal component,

PCA

Principal Components Analysis,

PFI

Progression-Free Interval,

PR

Progesterone receptor,

Q_1

First quartile,

Q_3

Third quartile,

qRT-PCR

quantitative Real-Time Polymerase Chain Reaction,

RNA-Seq

RNA-sequencing,

RPPA

Reverse Phase Protein Array,

RR

Relative Risk,

T

AJCC pathologic field Tumor,

TCGA

The Cancer Genome Atlas,

TCGA-CDR

TCGA Pan-Cancer Clinical Data Resource,

TNBC

Triple-Negative Breast Cancer,

TPBC

Triple Positive Breast Cancer,

TSS

Tissue Source Site,

UMAP

Uniform Manifold Approximation and Projection,

WHO

World Health Organization,

Funding

This work was funded partially by:

- European Social Fund grant AIDA POWR.03.02.00-00-I029
- SUT's grants for Support and Development of Research Potential in years 2018-2023