Extended abstract

Thesis: **"Machine learning methods in support of multiomics signature identification for breast cancer patient subpopulations"**

Author: **Joanna Tobiasz**

1 Motivation, aims, and thesis

Breast cancer is a highly heterogeneous disease with diverse clinical outcomes, manifesting various molecular and histological backgrounds. The clinical classification of breast cancer cases routinely used 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 for the identification of five intrinsic molecular subtypes of breast cancer in the early 2000s. Despite noteworthy inconsistencies with clinical classification, they are still referred to as the gold standard. 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 powerful tools to retrieve a more comprehensive insight into breast cancer stratification. Nonetheless, large data sets delivered by high-throughput analysis 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.

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 prognosis prediction. Examining disease subtypes can deliver clinically relevant information and 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.

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 choosing an appropriate machine learning approach for subpopulation detection. Moreover, the methods to assess the performance of tested methods were necessary.

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.

2 Background

Currently used, four 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).

The most common subtype is the hormone receptor-positive (HR+), defined with negative HER2 status (HER2-) and positive ER or PR statuses (ER+, PR+). HER2-positive (HER2+) breast cancer contains cases with both HER2+ status and HR- status. Triple-Negative Breast Cancer (TNBC) is defined as ER-, PR-, and HER2-. The last clinical subtype, defined as ER+, PR+, and HER2+, is called Triple Positive (TPBC) (Szymiczek, Lone, & Akbari, 2020).

Clinical subtypes demonstrate diversity in terms of therapy outcomes. Hence, new approaches to breast cancer classification were proposed with the advancements in high-throughput platforms. Gene expression profiling with hierarchical clustering allowed the identification of five molecular subtypes in Perou et al. (Perou, et al., 2000) and Sørlie et al. (Sørlie, et al., 2001). The first subtype, luminal A, is characterized by high expression levels of HRs and luminal epithelial genes and a low level of HER2. The luminal B subtype is also HR+, but its HR levels are low compared to luminal A. In some luminal B cases, HER2 levels are elevated. HER2-enriched subtype shows high levels of HER2 and low expression of luminal epithelial genes. 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 last subtype, normal-like, is not in use anymore as it was regarded as an artifact resulting from the contamination of tumor biospecimens with normal tissues (Parker, et al., 2009). 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. Hence, various machine learning approaches have been applied for cancer subtyping and further evaluating the obtained stratification.

50-gene Prediction Analysis of Microarray (PAM50) classifier is considered a gold standard for intrinsic molecular subtype prediction based on gene expression profiles. It was developed by (Parker, et al., 2009) using microarray data supported by the qRT-PCR results. Hence, this method is transcriptomic-based.

3 Materials

The data sets used for this study were collected from The Cancer Genome Atlas (TCGA Breast Invasive Carcinoma (BRCA) project. Only the primary tumor samples collected from the female patients were considered. The protein levels were measured with the Reverse Phase Protein Array (RPPA) platform. The mRNA gene expression levels were obtained with the Agilent custom 244K whole genome microarrays. Both data sets were downloaded from the Genomic Data Commons (GDC) Data Portal (Genomic Data Commons Data Portal, 2022) or Legacy Archive (Genomic Data Commons Legacy Archive, 2021) in the normalized form. Data sets were checked for the batch effect and adequately corrected if necessary. Data were discarded for those patients for whom a PAM50 classifier result was not available.

Moreover, TCGA Research Network provided demographic information concerning the patients, including age at the initial diagnosis, declared race, and ethnicity. Each patient was also annotated with the tissue source site (TSS), the medical center of the patient's initial diagnosis and sample collection. The clinical information provided per patient included 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. 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.

The relative proportions of 22 immune cell types in the tissue further characterized the tumor samples. 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.

4 Identification of patient subpopulations

Various combinations of clustering algorithms and feature engineering methods were tested for the subtyping based on the levels of 166 proteins. 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. DiviK algorithm consists of stepwise k-means clustering in a locally optimized feature domain, selected with the log2-scaled variances Gaussian Mixture Model (GMM) decomposition (Mrukwa & Polanska, 2022).

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 feature selection or extraction procedures were applied to prepare the data set for the clustering. Table 1 presents the summary and abbreviations of the variants, later used for referring to results.

Table 1 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*	×	✓	×	×	×	×

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 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 included the Principal Components Analysis (PCA) to select the top principal components (PC) explaining 90% of the variance in the data and Uniform Manifold Approximation and Projection (UMAP) performed on the PCA-reduced set (McInnes, Healy, & Melville, 2018).

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.

The results of different combinations of feature engineering and clustering algorithms are presented in Figure 1.

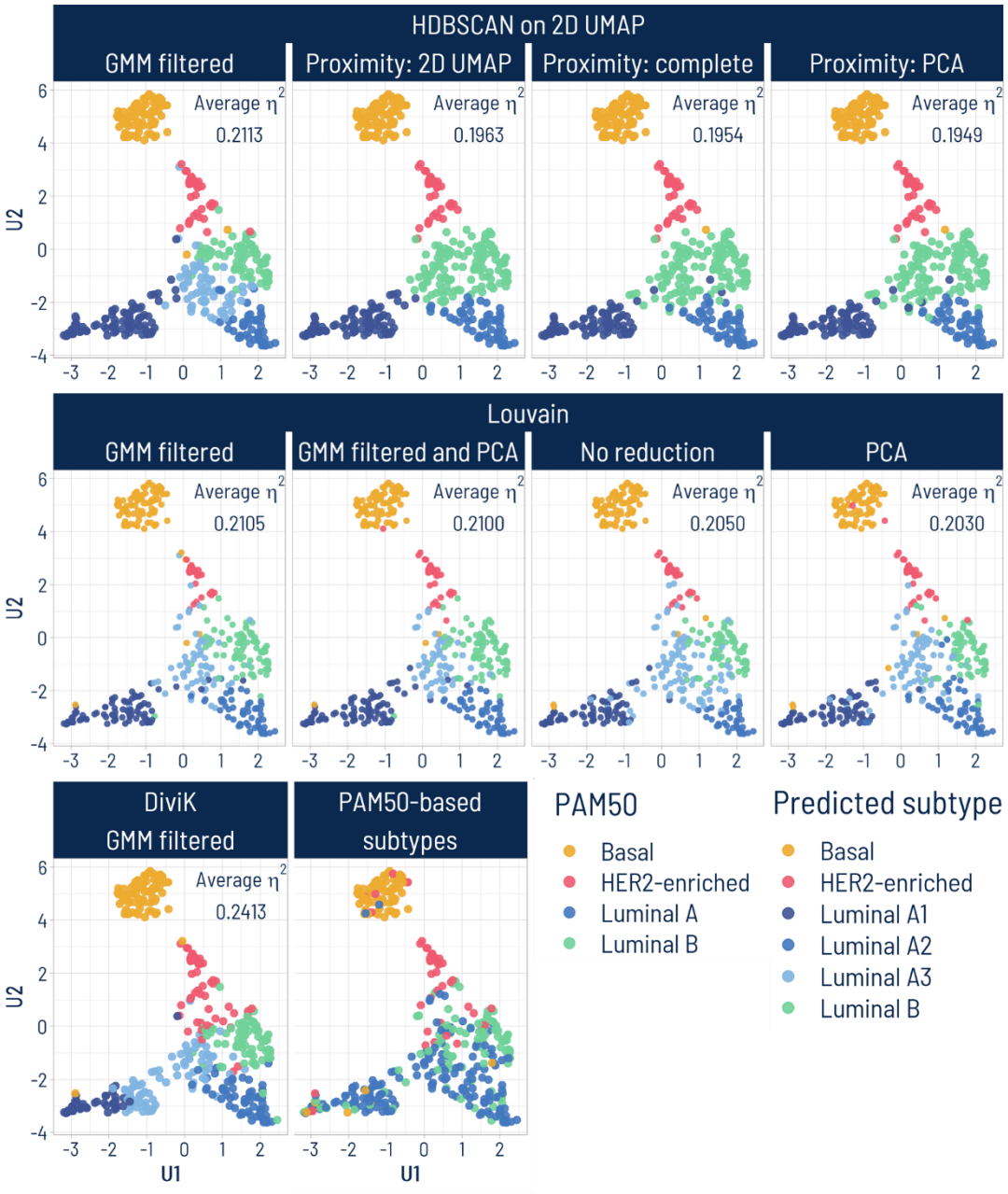

Figure 1. UMAP visualization with results of all clustering approaches and the original PAM50 subtype labels

Each figure corresponds to a 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)**.

Two effect-size-based metrics were proposed for selecting the most reliable clustering approach and, consequently, for defining breast cancer subtypes investigated in this work. Firstly, the levels of each protein were compared between the clusters with the η^2 effect size measure. The higher η^2 value, the higher the variance between the groups compared to the variances within the groups and the better the cluster separation. η^2 values for each protein 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.

However, all clusters are considered jointly, which is the limitation of η^2 metrics. Therefore, high η^2 values do not provide detailed information on whether all clusters are well-separated or just some are highly isolated. Thus, another metric 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. One hundred sixty-six *d* values were obtained per cluster for each evaluated clustering 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. Each cluster was annotated with the Q_3 of protein *d* absolute values. 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 2 (Tobiasz & Polanska, 2022). Furthermore, the 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.

Table 2 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 3.

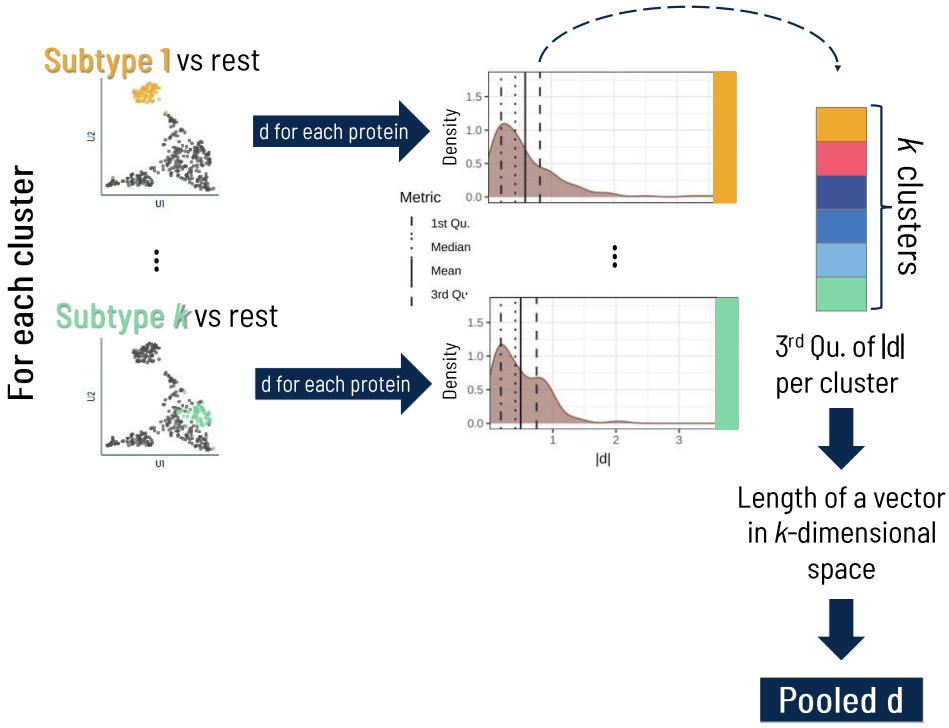

Figure 2. The procedure of pooled *d* calculation

Table 2. 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		
$H_{UMAP-C1}$	5	0.0764	0.1587	0.1963	0.3083	1.7053	0.7125
$H_{UMAP-C2}$	5	0.0749	0.1519	0.1954	0.3002	1.7204	0.7052
$H_{UMAP-C3}$	5	0.0785	0.1598	0.1949	0.3034	1.6847	0.7052
H_{UMAP-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-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-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

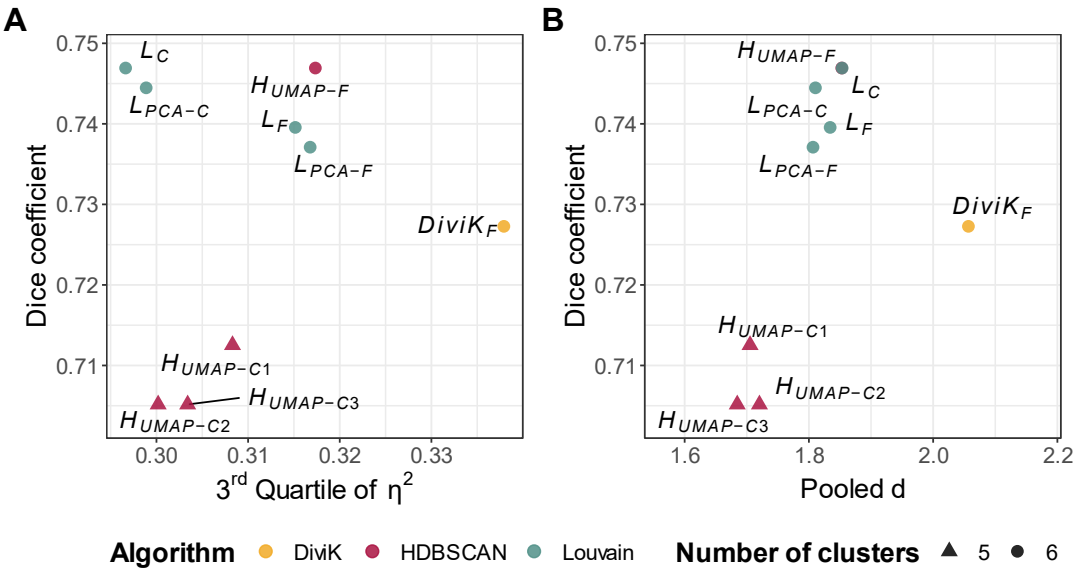

Figure 3. Comparison of η^2 and pooled d with Dice coefficient for tested clustering approaches

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. 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 regarding the number of cases in Table 3.

Table 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**

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. 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 the decision to divide luminal cases into four subgroups instead of only two luminal A and B, like the PAM50 predictor.

5.1 Survival analysis

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 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. The log-rank test was calculated for each comparison. It is the most common approach, in which the same importance is put on differences between the survival functions throughout the whole timespan of the study (Mantel, 1966; Peto & Peto, 1972; May, Hosmer, & Lemeshow, 2014). However, in the case of some comparisons, the differences in survival outcomes were mainly visible in the initial phases of the illness and therapy. Thus, the generalized Wilcoxon rank sum test, also called the Gehan-Wilcoxon test, was applied to compare the subpopulations. In this approach, the weights of differences between the survival outcomes are defined as the number of patients still at risk.

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). HR can be regarded as the effect size measure, interpreted analogously to the relative risk. The thresholds for HR interpretation were adjusted for the imbalance between the sizes of the compared groups. The survival analysis was performed for four endpoints: Overall Survival (OS), Disease-Specific Survival (DSS), Disease-Free Interval (DFI), and Progression

Free-Interval (PFI). The first two, however, are not recommended for the breast cancer cohort of TCGA, as the follow-up time is too short to observe a sufficient number of events.

The KM graphs for all four endpoints are shown in Figure 4 for luminal subpopulations identified with DiviK proteomic-based approach. 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.

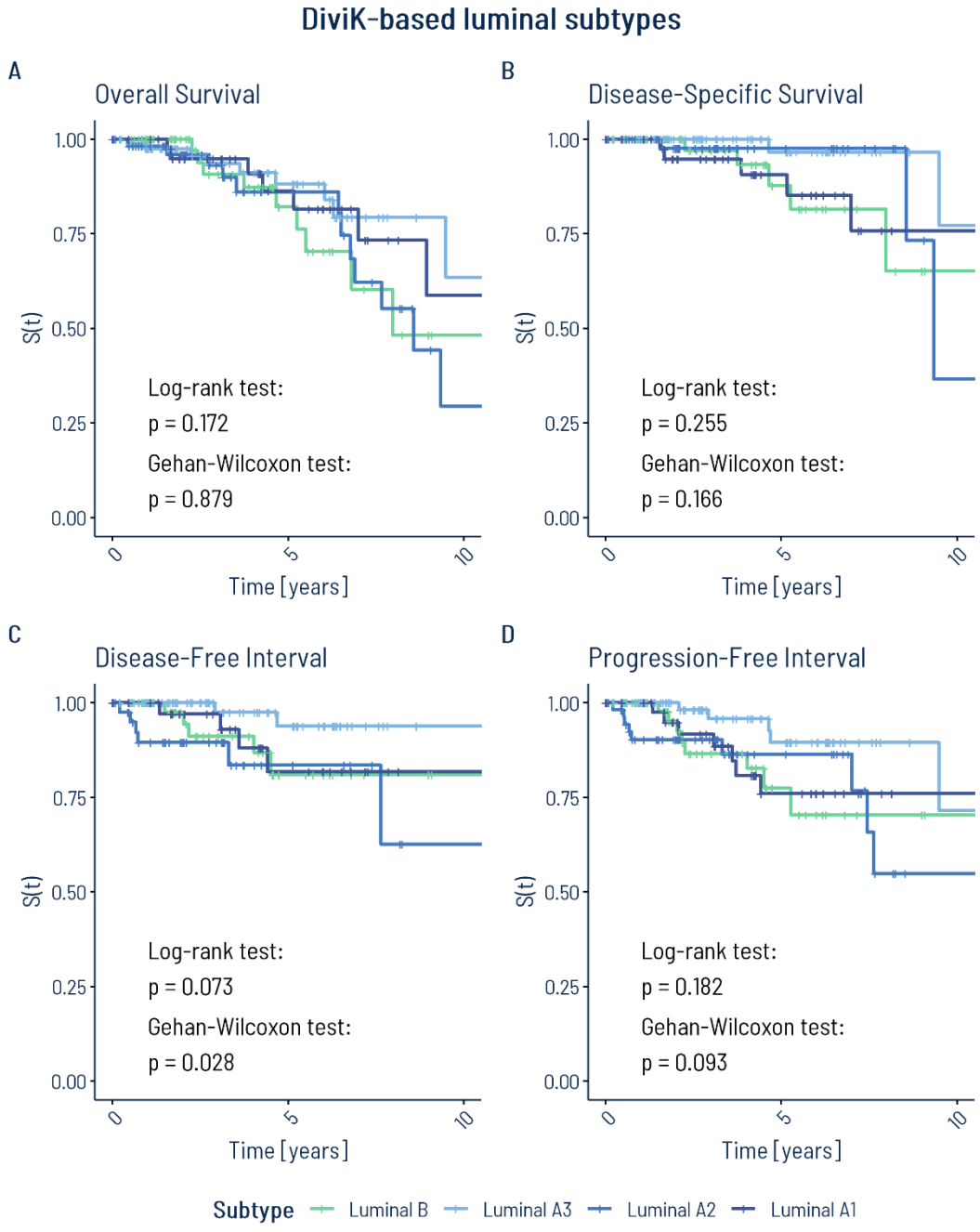

Figure 4. Kaplan-Meier survival curves of luminal subpopulations identified with DiviK

Moreover, the test statistics and p-values are presented in Table for DiviK-based and PAM50 luminal subtypes.

Table 4. 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

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. 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. The p-value is lower than 0.05 only for DFI. For DSS, two groups of similar curves can be noticed: one with luminal A2 and A3 subpopulations with a better prognosis and one consisting of luminal A1 and B subtypes with a worse outcome. 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.

5.2 Statistical analysis of demographic and clinical profiles

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

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. Results of the association analysis are shown in Table 5 for the subset of luminal subtypes. Cramér's V values are colored based on the effect size interpretation.

Table 5. 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 $\chi 2$ 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			

Feature	χ^2	p-value	Cramér's V	Cramér's V effect threshold		
				Small	Medium	Large
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			

The results indicate a 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.

Numerical variables used for the subtyping results evaluation included patient age at diagnosis and CIBERSORT immune cellular fraction estimates. They were compared between the subpopulations with tests selected according to the normality and variance homogeneity assumptions. Moreover, appropriate effect size measures supported the classical testing approach. Figure 5 summarizes the differentiation testing pipeline for comparing more than two subtypes.

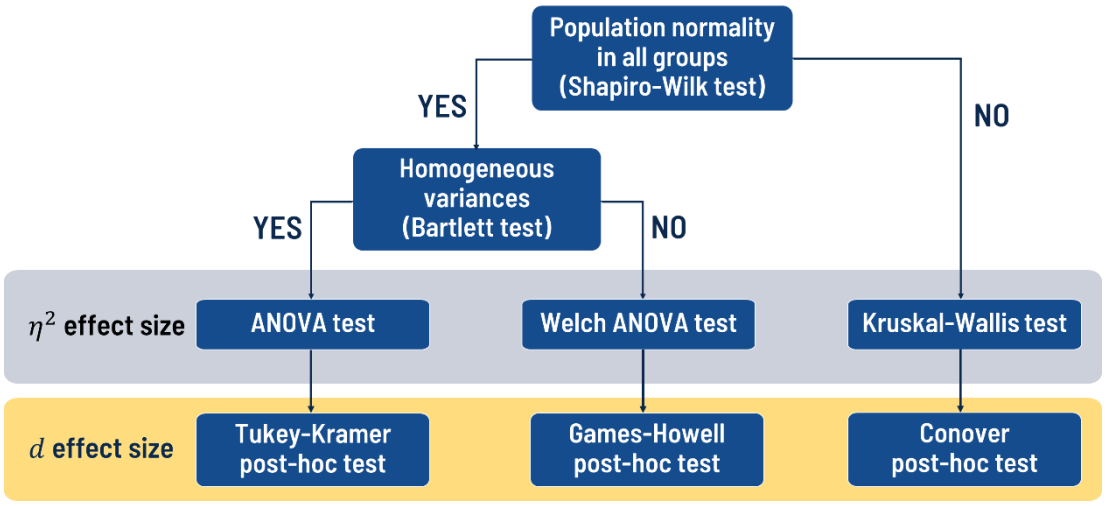

Figure 5. Differentiation testing pipeline for comparison of more than two groups

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 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. In those cases, outliers had a great impact on the test results.

The fractions 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.

6 Molecular signature of patient subpopulations

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. Also, 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 transcriptomic signatures, either specific for a single subtype or sufficient to differentiate the subtypes.

6.1 Subtype-specific marker identification

Subtype-specific markers were identified with the differentiation testing pipeline shown in Figure 5. 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). Their results were interpreted per omics to ensure that all measurements from the same platform were analyzed consistently.

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 appeared 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. Figure 6 presents the scheme of the subtype-specific marker identification process.

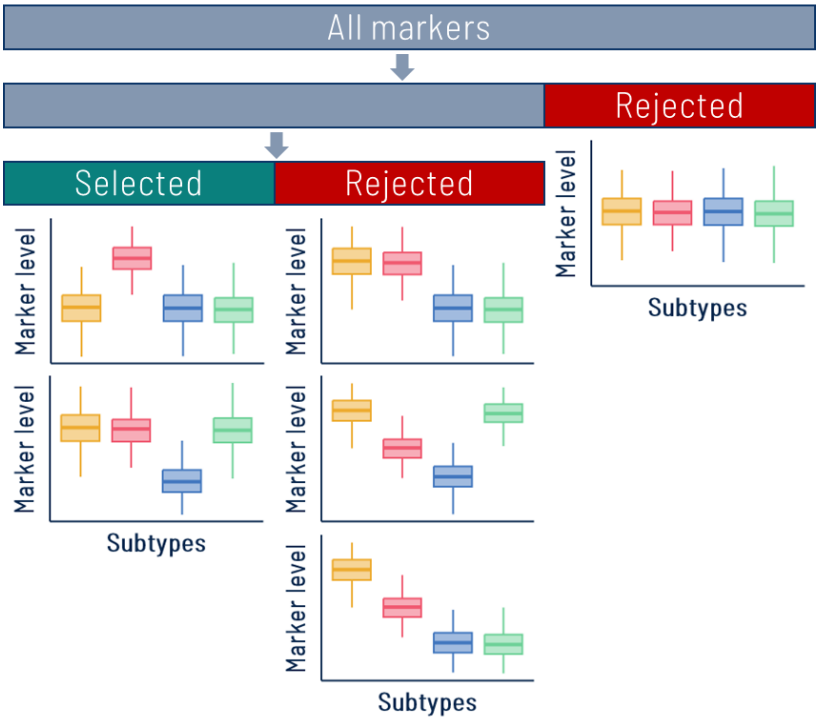

Figure 6. The subtype-specific marker identification process

Table 6 shows the numbers of subtype-specific markers identified with the effect-size-based approach. The number of transcriptomic markers is larger due to the bigger feature space.

Table 6. Number 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 the basal subpopulation, with many 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. Identified subtype-specific markers are listed regarding the direction of level changes compared to other subtypes in Figure 7 for the proteomic data set.

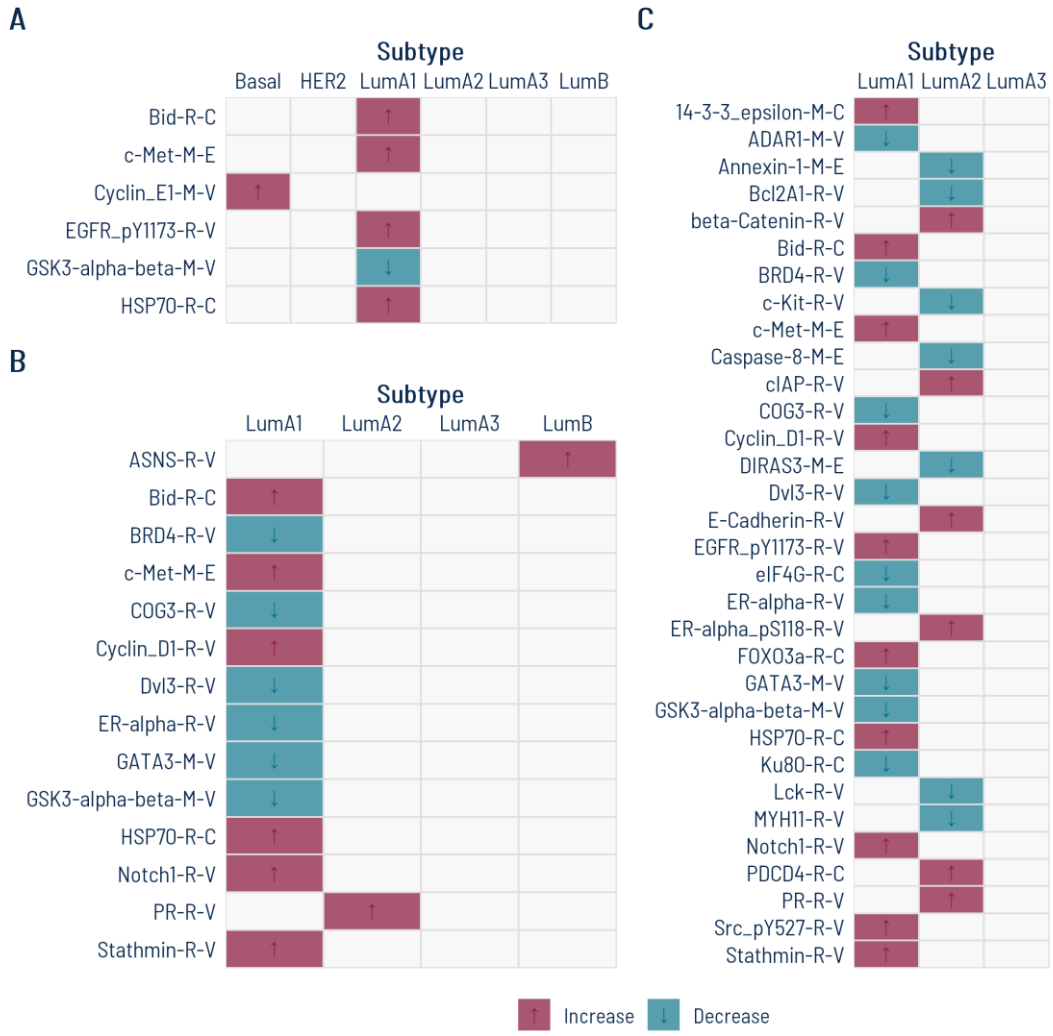

Figure 7. Subtype-specific markers identified based on the protein levels

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.

Over-Representation Analysis (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).

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. All pairwise comparisons provided significantly enriched KEGG pathways for the transcriptomic data set. Regardless of the comparison variant, the obtained pathways included those crucial for proper cell functioning and many involved in tumor biology.

6.2 Subtype differentiating signature

Another approach was proposed based on the multinomial logistic regression to identify the molecular signature, distinguishing all considered subtypes. Logistic regression is commonly applied as the classification method. However, it can also select meaningful features, such as the molecular signature of the identified subpopulations.

Multiple Random Cross-Validation (MRCV) procedure was used for model building with 100 iterations. 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 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 forward selection method. In each step, the model with the highest Bayes Factor (BF) was selected until BF dropped below ten or no more potential features were left. The performance of the resulting model was assessed based on the test set. Figure 8 presents the scheme of the MRCV procedure.

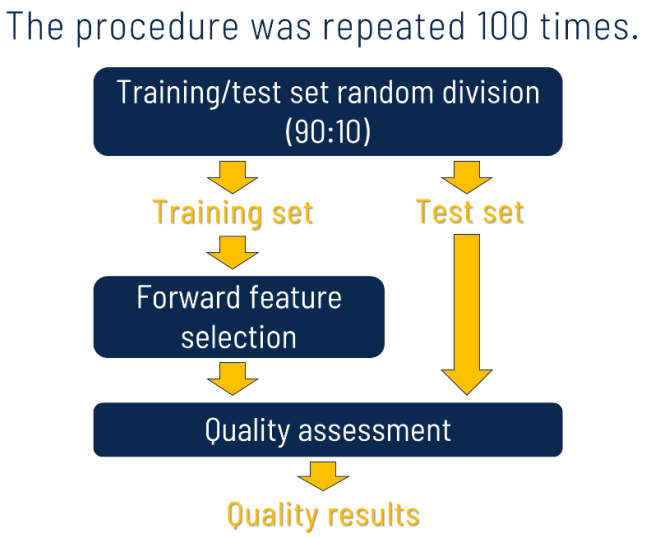

Figure 8. MRCV procedure for multinomial logistic regression model building

Outcomes of MRCV 100 repetitions served for the creation of the feature ranking. Features were sorted and assigned weights for each resulting model based on the selection order.

Each weight was multiplied by the model's overall balanced accuracy (BA) calculated on each test set. Products summed up among all 100 models gave an importance score for each feature. 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 author described a similar pipeline for the binary logistic regression (Henzel, et al., 2021) and (Kozielski, et al., 2021). Three variants of regression models were fit: for the proteomic data set, for the reduced transcriptomic data set, and for those two data sets combined.

Figure 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. The top 9 proteins were identified as the proteomic signature based on the elbow method.

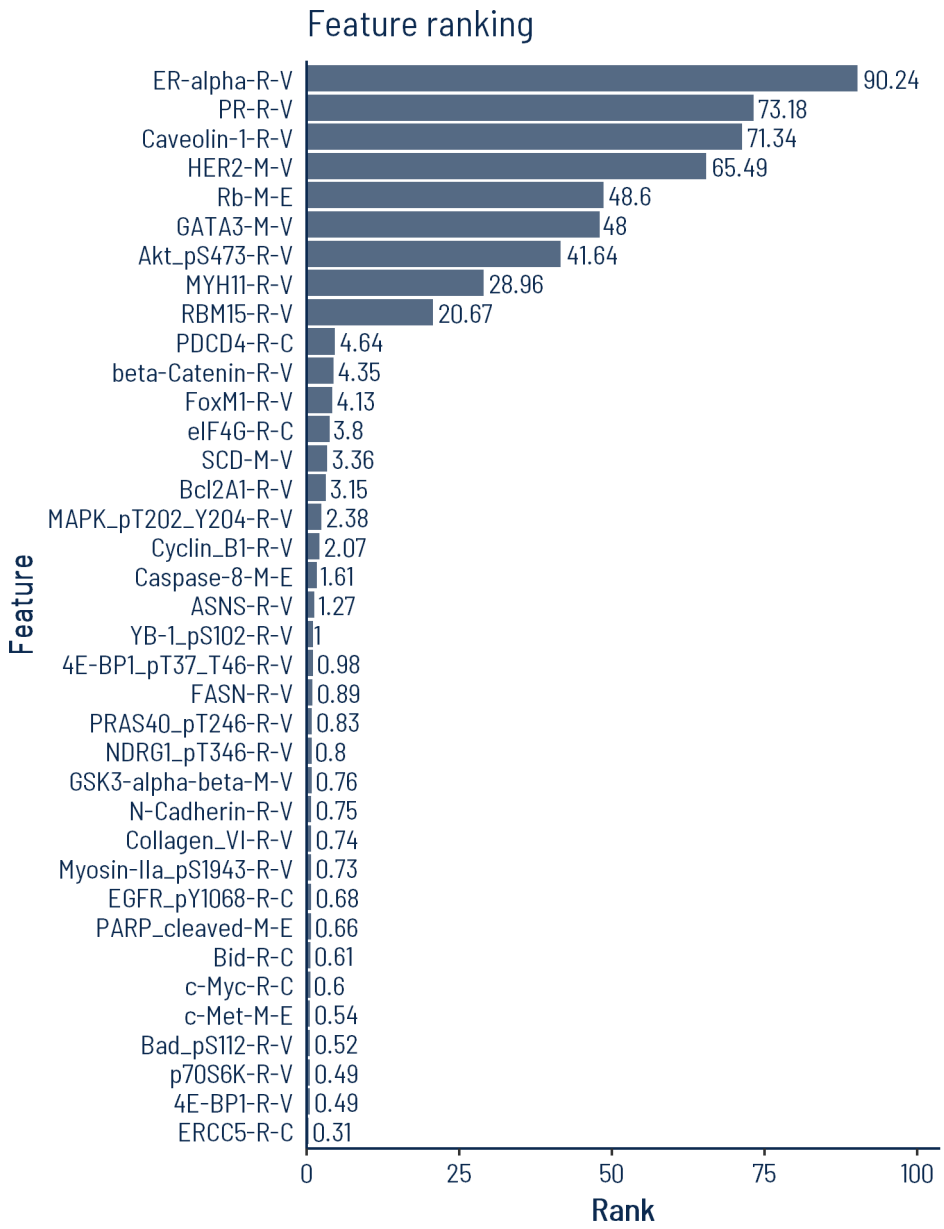

Figure 9. Feature ranking for the proteomic multimodal logistic regression model

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

Levels of proteins included in the subtype-differentiating proteomic signature are presented in Figure 10 for the top three proteins.

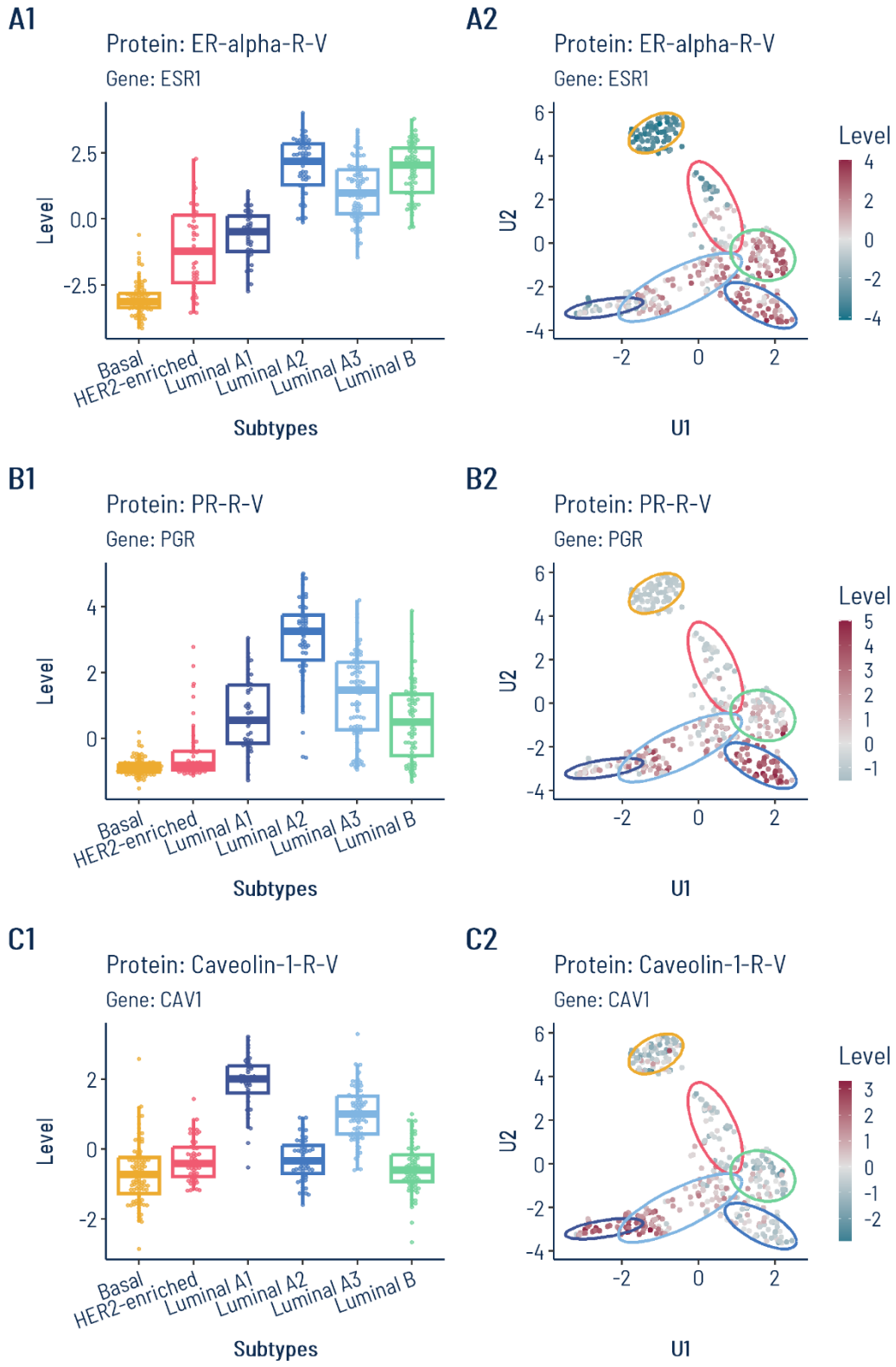

Figure 10. Levels of the top three proteins selected for the multinomial regression model concerning subpopulations identified with DiviK

Panels 1 show boxplots of protein levels per subtype. Panels 2 show the UMAP projection obtained based on the protein level data set, with the color of data points reflecting the protein level.

Figure 11 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.

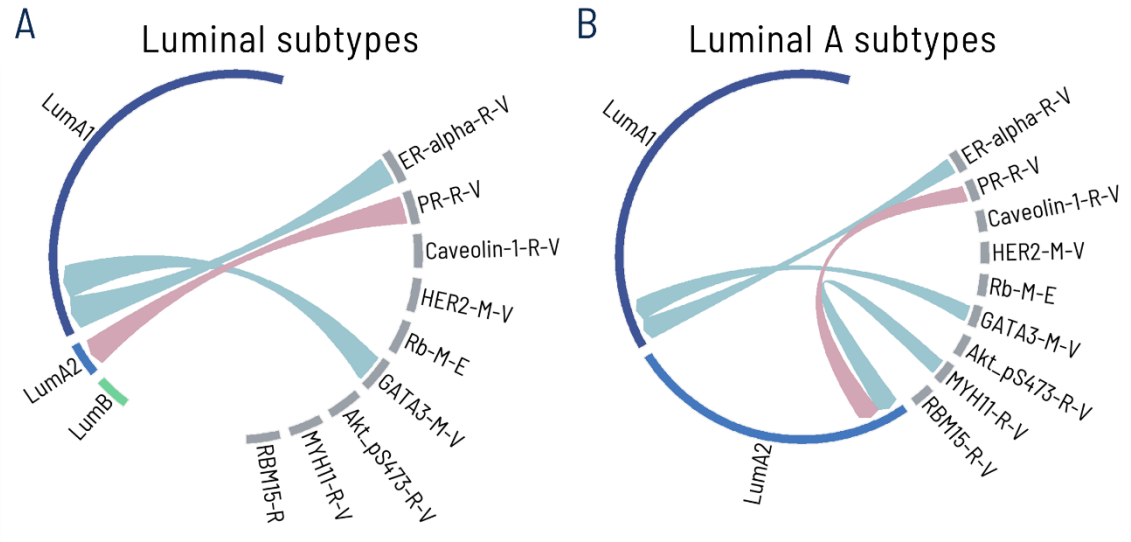

Figure 11. Comparison of proteomic model features and proteomic subtype-specific markers identified based on the effect size

A link means a particular protein was included in the model and 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).

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.

Feature ranking obtained in the MRCV procedure is shown in Figure 12. The maximal distance in the elbow plot was obtained for the sixth gene (*C7*). Hence, the top five genes formed the transcriptomic signature for subpopulations' differentiation. mRNA gene expression levels of those top 3 selected genes are presented in Figure 13.

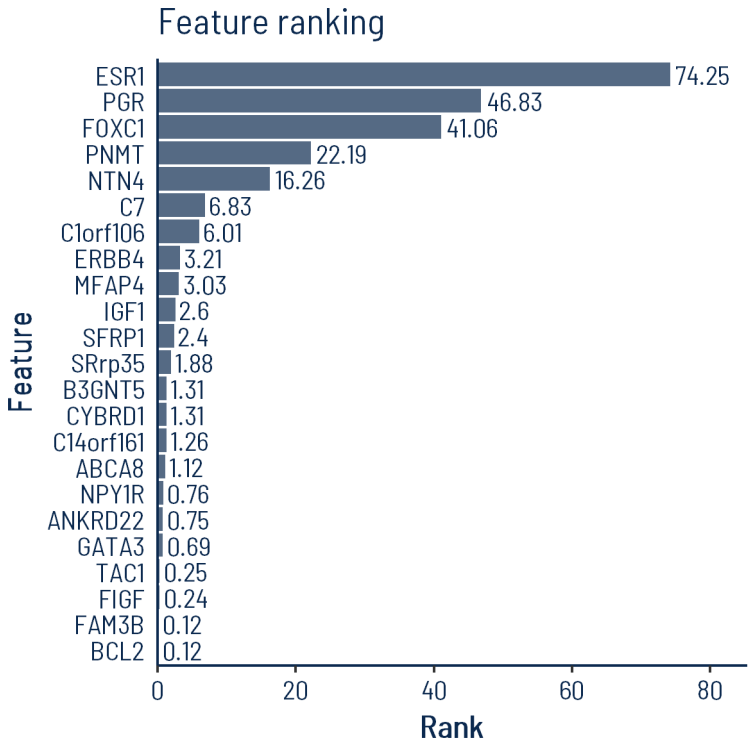

Figure 12. Feature ranking for the transcriptomic multimodal logistic regression model

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 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 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.

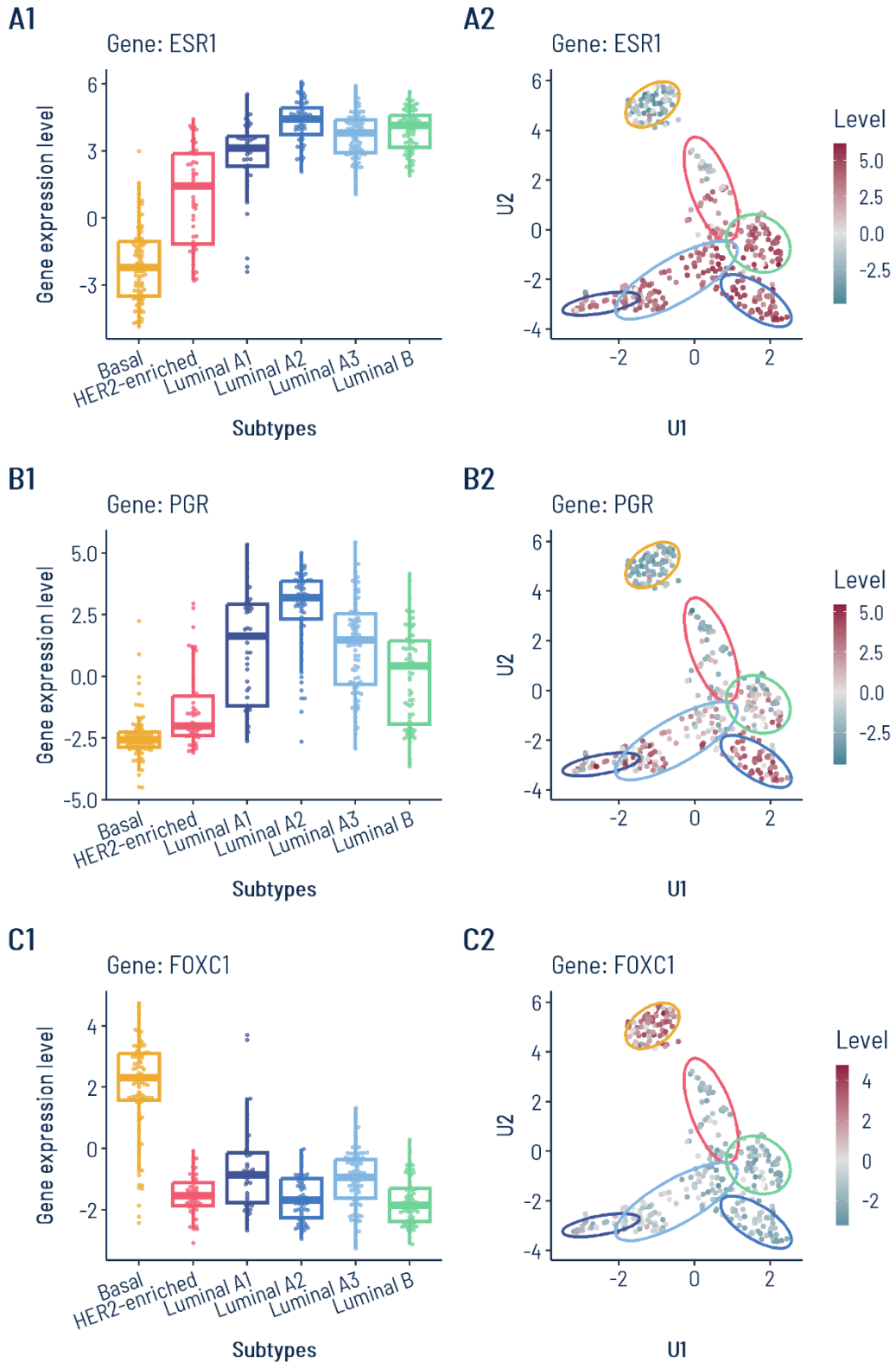

Figure 13. Levels of the top three transcripts selected for the multinomial regression model concerning subpopulations identified with DiviK

Panels 1 show boxplots of mRNA gene expression levels per subtype. Panels 2 show the UMAP projection obtained based on the protein level data set with the color of data points reflecting the mRNA gene expression level.

7 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. Thesis I was confirmed by the analysis outcomes shown in Chapter 4 (Identification of patient subpopulations). 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 (Clinical characteristics of patient subpopulations) and 6 (Molecular signature of patient subpopulations). The differences in survival experiences between the defined subpopulations were confirmed. HR+ and HR- subtypes were shown to vary in prognosis, and the newly revealed additional luminal subgroups were diverse in their survival outcome. 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. 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. 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 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. The pipeline proposed in this dissertation provided satisfactory results and dealt with the challenging data set.

All applied machine learning approaches proved 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. Given the insufficient follow-up time for cancer with a relatively good prognosis, 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 on the transcriptomic and proteomic levels, including the significant diversity within the luminal group. 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. 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, the luminal A1 subtype 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 References

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

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

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

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

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

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

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

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

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/*.

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

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

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

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

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

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

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

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*.

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

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

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

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

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

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

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

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., 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., 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

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

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

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/ $\less$i$\greater$neu$\less$/i$\greater$ 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

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

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

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

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

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

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

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 {\&}amp$\mathsemicolon$ Laboratory Medicine, 142*, 1364–1382. doi:10.5858/arpa.2018-0902-sa

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

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

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