Streszczenie poszerzone

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

Autor: **mgr inż. Joanna Tobiasz**

1 Motywacja, cele i tezy

Rak piersi jest wysoce heterogeniczną chorobą o zróżnicowanych wynikach klinicznych, z różnymi podłożami molekularnym i histologicznym. Stosowana powszechnie klasyfikacja kliniczna raka piersi pozostaje niezmienna od lat i opiera się na ekspresji kilku genów i białek markerowych. Nie odzwierciedla więc ona idealnie portretu molekularnego chorych na raka piersi i ma wiele ograniczeń.

Profilowanie ekspresji genów umożliwiło na początku XXI wieku identyfikację pięciu podtypów molekularnych raka piersi. Pomimo pewnych niespójności z klasyfikacją kliniczną, nadal są one określane jako złoty standard. Wraz ze wzrostem wiedzy biologicznej i lepszym zrozumieniem podłoża molekularnego raka, klasyfikacja molekularna wydaje się jednak niewystarczająco odzwierciedlać złożony charakter raka piersi. Co więcej, różne mechanizmy wpływają na ekspresję genów pomiędzy warstwami transkryptomiczną a proteomiczną. Nie są one więc uwzględniane przez obecnie stosowane klasyfikacje raka piersi.

Wraz z postępem w technologiach wysokoprzepustowych do badania ekspresji poza poziomem transkryptomicznym oraz w metodach uczenia maszynowego do eksploracji wielkich danych biologicznych, pojawiły się narzędzia umożliwiające bardziej wszechstronny wgląd w strukturę raka piersi. Niemniej jednak, adekwatna i zaawansowana analiza statystyczna jest niezbędna, by odpowiednio scharakteryzować zmienność w danych i precyzyjnie wybrać cechy najlepiej wyjaśniające różnorodność i rozróżniające podtypy raka piersi. Dla wyciągnięcia poprawnych i istotnych biologicznie wniosków kluczowe jest opracowanie metodologii wykorzystującej dedykowane techniki uczenia statystycznego, w tym metody nienadzorowane, zapewniające podział na podtypy raka piersi niezależny od tego stosowanego obecnie.

Ponowna identyfikacja podtypów raka piersi może stanowić uzupełnienie istniejących podejść oraz odzwierciedlać wcześniej niewidoczne źródła różnorodności nowotworu. Dokładne określenie podtypu raka piersi jest kluczowe dla wyboru leczenia i pozwala na przewidywanie rokowań. Badania nad podtypami raka piersi mogą dostarczyć istotnych klinicznie informacji i pomóc odkryć nowe cele terapeutyczne. W przyszłości może to znaleźć zastosowanie w medycynie spersonalizowanej i zindywidualizować organizację terapii tak, by zapewnić każdemu pacjentowi możliwie optymalny plan leczenia i zmniejszyć efekty uboczne.

Niniejsza rozprawa doktorska miała na celu identyfikację i ocenę podpopulacji chorych na raka piersi. Istniejące podtypy molekularne tej choroby zostały opracowane za pomocą profilowania ekspresji genów, w tej pracy bazowano na profilach białkowych. Pierwszy etap badania wymagał zaproponowania odpowiedniego podejścia uczenia maszynowego do wykrywania podpopulacji. Ponadto, konieczne było zaproponowanie metod oceny jakości uzyskanych wyników.

Następnie, otrzymane za pomocą dobranego podejścia uczenia maszynowego podpopulacje chorych na raka piersi oceniono i scharakteryzowano pod kątem klinicznym. Praca również miała na celu dostarczenie narzędzi statystycznych i metod uczenia maszynowego do identyfikacji sygnatur molekularnych zidentyfikowanych podpopulacji. Na podstawie testowania statystycznego uzupełnionego odpowiednimi miarami wielkości efektu, otrzymano sygnaturę molekularną opisującą różnice proteomiczne i transkryptomiczne pomiędzy zidentyfikowanymi podpopulacjami. Została ona zweryfikowana za pomocą przeglądu literatury i dedykowanych metod analizy funkcjonalnej.

Mając na uwadze motywację i cel tej rozprawy doktorskiej, sformułowano następujące tezy:

I. Zastosowanie zaawansowanych metod uczenia maszynowego i modelowania matematycznego pozwala na identyfikację nowych molekularnych podpopulacji chorych na raka piersi.
II. W przypadku wysoce niezrównoważonych i zróżnicowanych pod względem wielkości prób, wszechstronne testowanie statystyczne poparte analizą wielkości efektu pozwala zdefiniować stabilne molekularne i kliniczne profile podtypów.

2 Wprowadzenie

Obecnie stosowane są cztery kliniczne podtypy raka piersi określane na podstawie obecności trzech kluczowych markerów: receptorów estrogenu (ER), receptorów progesteronu (PR), oraz receptora ludzkiego naskórkowego czynnika wzrostu (ang. human epidermal growth factor receptor 2 - HER2) (Jassem, Shan, & Buczek, 2020).

Najczęstszym podtypem jest hormonozależny (HR+) rak piersi, z ujemnym statusem HER2 (HER2-) i dodatnim statusem ER lub PR (ER+, PR+). Rak piersi HER2-dodatni (HER2+) obejmuje natomiast przypadki HER2+ i HR-. Potrójnie ujemny rak piersi (ang. Triple-Negative Breast Cancer - TNBC) nie posiada żadnego z trzech receptorów (ER-, PR- i HER2-). Ostatni podtyp kliniczny, zdefiniowany jako ER+, PR+ i HER2+, nazywany jest potrójnie pozytywnym rakiem piersi (Triple-Positive Breast Cancer - TPBC) (Szymiczek, Lone, & Akbari, 2020).

Podtypy kliniczne różnią się wynikami leczenia. Po pojawieniu się nowych metod badania genomu, zaproponowano nowe podejścia do klasyfikacji raka piersi. Profilowanie ekspresji genów z użyciem hierarchicznego klastrowania pozwoliło na identyfikację pięciu podtypów molekularnych w pracach Perou et al. (Perou, et al., 2000) oraz Sørlie et al. (Sørlie, et al., 2001). Pierwszy podtyp, luminalny A, charakteryzuje się wysokim poziomem ekspresji HR i genów nabłonka luminalnego oraz niskim poziomem HER2. Podtyp luminalny B jest również definiowany jako HR+, ale jego poziomy HR są niskie w porównaniu z luminalnym A. W niektórych przypadkach luminalnych B poziom HER2 są podwyższone. Podtyp HER2-wzbogacony wykazuje wysoki poziom HER2 i niską ekspresję genów nabłonka luminalnego. Najbardziej specyficznym podtypem jest podstawny, w którym geny luminalne, HR i HER2 mają niską ekspresję. Wysoką ekspresję wykazują natomiast geny charakterystyczne dla komórek podstawnych. Ostatni podtyp, „normalny", nie jest już używany, ponieważ został uznany za artefakt wynikający z zanieczyszczenia próbek nowotworowych sąsiadującą zdrową tkanką (Parker, et al., 2009). Początkowo podtypy kliniczne i molekularne uznawano za spójne. Nazwy HER2+ i HER2-wzbogacony, TNBC i podstawny oraz hormonozależny i luminalny często stosowane były zamiennie, przy czym podtypy luminalne A i B rozróżniane były na podstawie poziomu białka Ki67(Szymiczek, Lone, & Akbari, 2020; Sali, et al., 2020).

Niemniej jednak, wraz z rosnącą dostępnością nowoczesnych platform badawczych i rosnącą liczbą badań dotyczących profilowania guzów piersi, pojawiła się rozbieżność pomiędzy

klinicznymi i molekularnymi podtypami. W związku z tym zaczęto stosować różne podejścia uczenia maszynowego do podtypowania nowotworów i dalszej oceny uzyskanego podziału.

Klasyfikator PAM50 (ang. 50-gene Prediction Analysis of Microarray) jest standardowo stosowanym sposobem predykcji podtypu molekularnego w oparciu o profile ekspresji 50 wybranych genów. Klasyfikator ten został opracowany przez (Parker, et al., 2009) z wykorzystaniem danych mikromacierzowych wspartych wynikami ilościowej reakcji łańcuchowej polimerazy w czasie rzeczywistym qRT-PCR.

3 Materiały

Zestawy danych wykorzystane w tym badaniu pochodzą z projektu The Cancer Genome Atlas (TCGA Breast Invasive Carcinoma - BRCA). Uwzględniono tylko próbki guza pierwotnego pobrane od kobiet. Poziomy białek zostały zmierzone za pomocą platformy Reverse Phase Protein Array (RPPA). Poziomy ekspresji genów mRNA uzyskano za pomocą mikromacierzy Agilent custom 244K whole genome microarrays. Oba zestawy danych zostały pobrane z Genomic Data Commons (GDC) Data Portal (Genomic Data Commons Data Portal, 2022) lub Legacy Archive (Genomic Data Commons Legacy Archive, 2021) w formie znormalizowanej. Zestawy danych zostały sprawdzone pod kątem występowania efektu paczki i w razie potrzeby odpowiednio skorygowane. Odrzucono również dane dla tych pacjentek, dla których wynik klasyfikatora PAM50 nie był dostępny.

Ponadto TCGA Research Network udostępniło informacje demograficzne w tym wiek w momencie diagnozy, zadeklarowaną rasę i pochodzenie etniczne. Każda pacjentka została również opatrzony adnotacją TSS (ang tissue sourse site), czyli nazwą ośrodka, w którym została postawiona diagnoza i pobrano materiał biologiczny. Informacje kliniczne dostarczone dla każdej pacjentki obejmowały status przeżycia, czas od początkowej diagnozy do ostatniego kontaktu z pacjentką, a w przypadku śmierci - czas przeżyty od początkowej diagnozy. Zebrano również zapisy dotyczące badań kontrolnych po wyzdrowieniu, chociaż niestety te dane nie są spójne dla całego zbioru. Ponadto, dla każdej pacjentki dostępne są informacje dotyczące stadium choroby według AJCC (ang. American Joint Committee on Cancer) dotyczące wielkości guza T (ang. Tumor), zajętych węzłów chłonnych N (ang. Nodes), przerzutów M (ang. Metastasis) oraz stadium.

Próbki guza charakteryzowano również pod względem frakcji 22 typów komórek układu odpornościowego. Zostały one oszacowane w (Thorsson, et al., 2018) za pomocą metody CIBERSORT (Newman, et al., 2015) na podstawie danych z sekwencjonowania RNA (RNA-Seq).

4 Identyfikacja podpopulacji pacjentek

Różne kombinacje algorytmów klasteryzacji i metod inżynierii cech zostały zastosowane w celu identyfikacji podtypów raka piersi na podstawie pomiarów poziomów 166 białek. Wykorzystano algorytmy reprezentujące grupę metod opartych na gęstości, grafach i centroidach, odpowiednio: Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello, Moulavi, & Sander, 2013), metodę detekcji skupisk Louvain (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008), oraz niestandardową metodę DiviK - Divisive intelligent K-means (Mrukwa & Polanska, 2022). Algorytm DiviK polega na wielokrotnym, hierarchicznym klastrowaniu metodą k-średnich w lokalnie zoptymalizowanej przestrzeni cech, wybranej za pomocą dekompozycji mieszanin rozkładów Gaussowskich wariancji poziomów białka w skali logarytmicznej (Mrukwa & Polanska, 2022).

Zastosowane metody klastrowania w różnym stopniu radzą sobie z wysoką wymiarowością danych, dlatego w niektórych przypadkach wymagana była jej redukcja, a klastrowanie przeprowadzono albo na poziomach wszystkich białek, albo w obrębie zredukowanej przestrzeni cech. W zależności od algorytmu grupowania wykorzystano różne procedury selekcji lub ekstrakcji cech w celu przygotowania zbioru danych do klasteryzacji. W Tabeli 1 przedstawiono podsumowania i skróty wariantów metod, stosowane później do odwoływania się do wyników.

Tabela 1. Kombinacje algorytmów klastrowania i metod redukcji wymiarowości danych

Skróty dla każdej kombinacji zapisane są kursywą. DiviK jest oznaczony (*), aby wskazać, że filtracja oparta na GMM jest budowana w każdej iteracji algorytmu.

Tabela pochodzi z (Tobiasz & Polanska, 2022).

	Metoda inżynierii cech					
	Brak redukcji		**PCA**		**UMAP**	
Klastrowanie	**Pełna**	**Po filtracji GMM**	**Pełna**	**Po filtracji GMM**	**Pełna**	**Po filtracji GMM**
HDBSCAN	×	×	×	×	$H_{UMAP\text{-}C}$ ✓	$H_{UMAP\text{-}F}$ ✓
Louvain	L_C ✓	L_F ✓	$L_{PCA\text{-}C}$ ✓	$L_{PCA\text{-}F}$ ✓	×	×
DiviK*	×	✓	×	×	×	×

Do selekcji cech wykorzystano metodę dekompozycji GMM. Obliczono wariancje poszczególnych poziomów białka i przekształcono je do skali logarytmicznej. Następnie rozkład uzyskanych wartości został zdekomponowany zgodnie z opisem w (Marczyk, Jaksik, Polanski, & Polanska, 2019). Punkt przecięcia dwóch składowych odpowiadających największym wariancjom wyznaczał wartość progową dla filtracji: tylko białka o wyższej wariancji poziomów były brane pod uwagę w procedurze klasteryzacji.

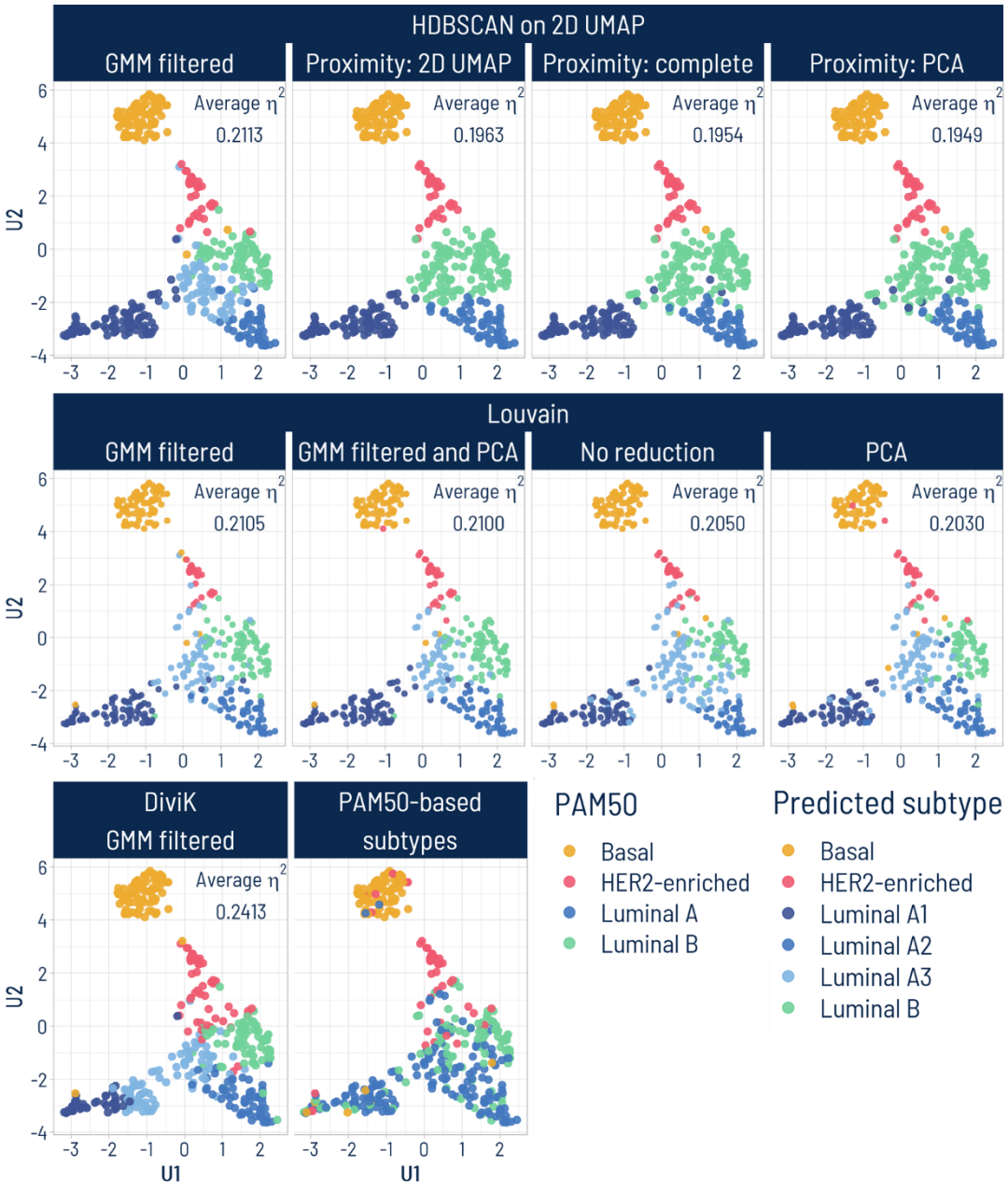
Metody ekstrakcji cech obejmowały analizę głównych składowych (PCA) w celu selekcji głównych składowych (PC) wyjaśniających 90% wariancji w danych oraz algorytm Uniform Manifold Approximation and Projection (UMAP) zastosowany na zbiorze zredukowanym przez PCA (McInnes, Healy, & Melville, 2018).

Ze względu na charakter metody HDBSCAN, po jej zastosowaniu niektóre pacjentki mogą pozostać nieprzypisane do żadnego wynikowego klastra. Jednak do dalszej analizy potrzebna jest nowa etykieta podtypu dla każdej pacjentki. Konieczne więc było przypisanie takich obserwacji do grup jak najbardziej do nich podobnych. Przetestowano następujące warianty predykcji przypisania do klastra, wszystkie oparte na odległości euklidesowej między punktem odpowiadającym danej pacjentce a centroidem klastra:

1. $H_{UMAP\text{-}C1}$: Bliskość w 2-wymiarowej przestrzeni UMAP;
2. $H_{UMAP\text{-}C2}$: Bliskość w pełnym zestawie danych (dla wszystkich białek);

3. $H_{UMAP\text{-}C3}$: Bliskość zestawie głównych składowych wyjaśniających 90% wariancji w danych.

Każdą pacjentkę przypisano zatem do klastra za pomocą dziewięciu kombinacji metod redukcji wymiarowości i klasteryzacji. Powstałe klastry uznane zostały za podpopulacje pacjentek i w dalszej części pracy będą tak nazywane lub będą opisane jako podtypy raka piersi. Wyniki różnych kombinacji algorytmów inżynierii cech i klastrowania przedstawiono na Rysunku 1.



Rysunek 1. Wizualizacja UMAP z wynikami wszystkich metod klasteryzacji i oryginalnymi etykietami podtypów PAM50

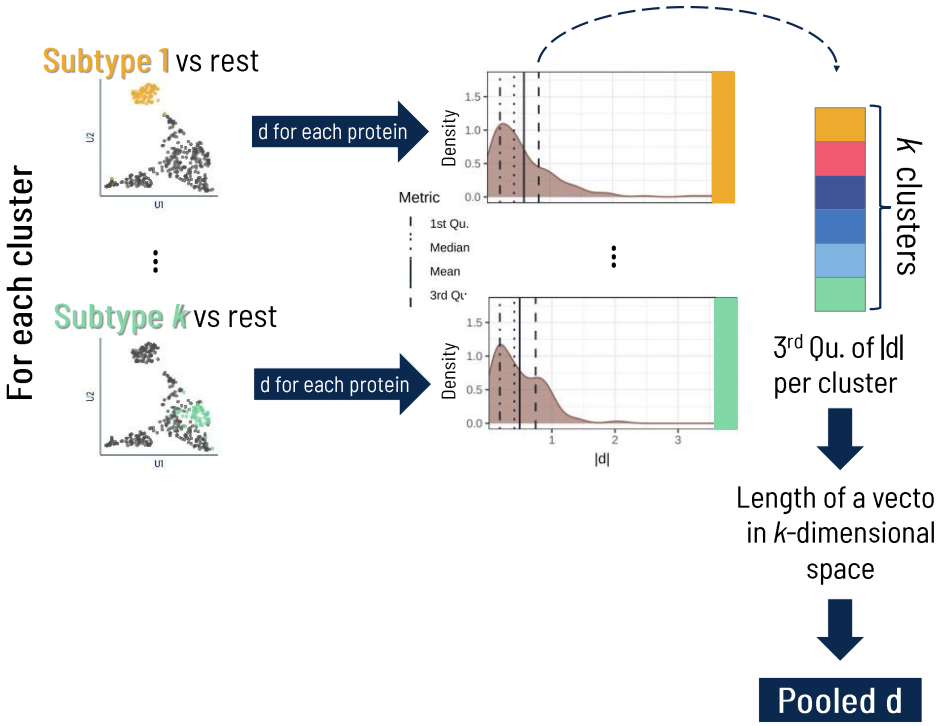
Każdy panel odpowiada innemu podejściu do grupowania połączonemu z różnymi procedurami przetwarzania przed grupowaniem lub po nim: redukcja wymiarowości danych przed klastrowaniem za pomocą selekcji i/lub ekstrakcji cech, lub w przypadku metody HDBSCAN, techniki przewidywania podtypu dla nieprzypisanych pacjentek. Kolor punktów oznacza podtyp: albo przewidywany w tym badaniu, albo uzyskany za pomocą klasyfikatora PAM50.

Rysunek został zaadaptowany z (Tobiasz & Polanska, 2022).

Do wyboru najbardziej rzetelnego podejścia do klastrowania, a w konsekwencji do zdefiniowania badanych w tej pracy podtypów raka piersi, zaproponowano dwie metryki oparte na wielkości efektu. Najpierw, poziomy każdego białka zostały porównane pomiędzy klastrami za pomocą miary wielkości efektu η^2. Im wyższa wartość η^2, tym większa wariancja między grupami w porównaniu z wariancjami wewnątrz grup i tym lepsza separacja klastrów. W obrębie każdego testowanego podejścia grupowania wartości η^2 obliczono dla każdego białka osobno. Aby zintegrować te wyniki, obliczono średnią, medianę, pierwszy kwartyl (Q_1) i trzeci kwartyl (Q_3) wartości η^2 białka.

Jednakże, ograniczeniem metryki η^2 jest rozpatrywanie wszystkich klastrów łącznie. Wysokie wartości η^2 nie dostarczają szczegółowych informacji o tym, czy wszystkie klastry są dobrze rozdzielone, czy tylko niektóre są silnie izolowane. Zaproponowano więc inną metrykę, modyfikując wielkość efektu *d* Cohena (Cohen, 2013). Koncepcja ta polegała na odniesieniu każdego uzyskanego klastra po kolei do wszystkich pozostałych klastrów rozpatrywanych łącznie. Efekt ten uzyskano poprzez porównanie średnich poziomów białka pomiędzy pacjentami przypisanymi i nieprzypisanymi do danej podpopulacji. Dla każdego ocenianego sposobu klastrowania uzyskano 166 wartości *d* na klaster. Aby łatwo porównać metody klastrowania, tylko jeden wynik powinien reprezentować każdą z nich. Listy wartości *d* dla każdej metody zostały więc zintegrowane w celu uzyskania jednego zbiorczego wyniku *d*. Każdy klaster został opisany Q_3 wartości bezwzględnych *d* dla białka. Wartości Q_3 zostały rzutowane jako punkt w przestrzeni *k*-wymiarowej, gdzie *k* było liczbą wykrytych podtypów. Na koniec obliczano zbiorczy wynik *d* jako odległość między utworzonym punktem a początkiem układu współrzędnych. Procedura uzyskiwania zintegrowanych wartości *d* została przedstawiona na Rysunku 2 (Tobiasz & Polanska, 2022). Ponadto obliczono indeks Dice'a (Dice, 1945) w celu oceny podobieństwa między podtypami wykrytymi za pomocą poszczególnych metod grupowania a podtypami uzyskanymi przez predyktor PAM50.

W Tabeli 2 przedstawiono wartości kwartyli i średniej η^2, zintegrowanej wartości d oraz wartości współczynnika Dice'a dla poszczególnych sposobów grupowania. Współczynniki Dice'a zostały porównane z wartościami zintegrowanego d i Q_3 η^2 na Rysunku 3.

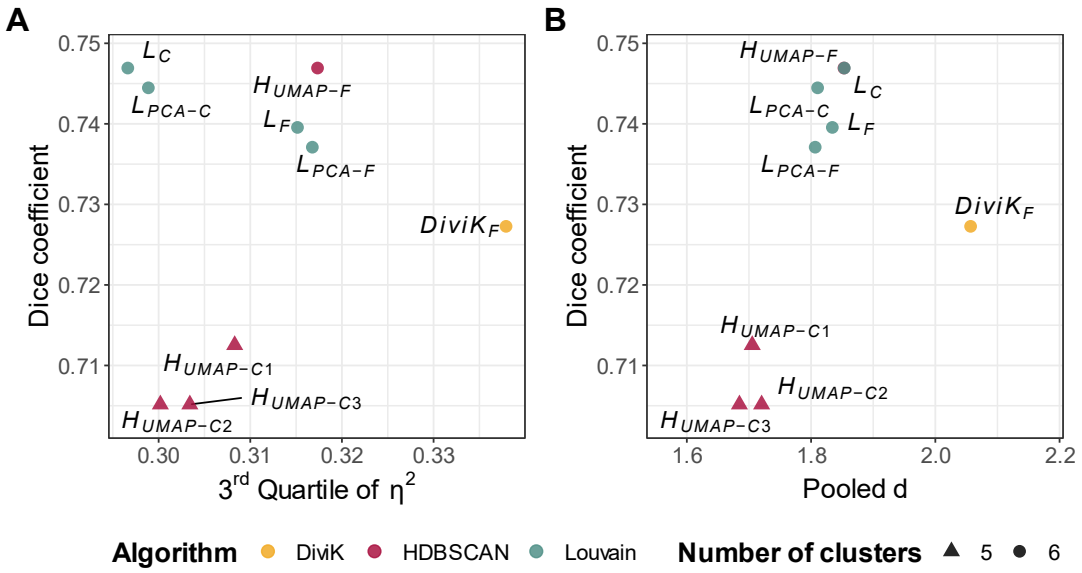


Rysunek 2. Procedura integracji wartości *d*

Tabela 2. Metryki uzyskane za pomocą różnych kombinacji metod inżynierii cech i algorytmów grupowania

Tabela pochodzi z (Tobiasz & Polanska, 2022).

Metoda	Liczba klastrów	η^2				Zintegrowane *d*	Współczynnik Dice'a
		Q_1	Mediana	Średnia	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



Rysunek 3. Porównanie η^2 i zintegrowanego d z indeksem Dice'a dla testowanych sposobów grupowania

Wartości współczynnika Dice'a w odniesieniu do trzeciego kwartyla η^2 (Panel A) oraz zintegrowanego d (Panel B).

Rysunek pochodzi z (Tobiasz & Polanska, 2022).

W przypadku porównania z etykietami podtypów PAM50 na podstawie współczynnika Dice'a wszystkie metody, które dawały sześć skupień, przewyższały te, które wykrywały tylko pięć podpopulacji. Najwyższy współczynnik Dice'a zaobserwowano dla algorytmu Louvain zastosowanego do całej przestrzeni cech oraz dla klastrowania HDBSCAN poprzedzonego selekcją cech opartą na GMM i ekstrakcją cech za pomocą UMAP. Ostatecznie jako najwłaściwszą metodę identyfikacji podpopulacji pacjentów wybrano podejście klastrowania DiviK. Wyniki klastrowania DiviK odniesiono do podtypów PAM50 w zakresie liczby przypadków w Tabeli 3.

Tabela 3. Liczba pacjentek w podtypach wykrytych za pomocą algorytmu DiviK w odniesieniu do podtypów PAM50

Podtyp PAM50	Podtyp wykryty za pomocą DiviK						SUMA
	Podstawny	HER2-wzbogacony	Luminalny				
			A1	A2	A3	B	
Podstawny	79	0	4	0	2	1	**86**
HER2-wzbogacony	8	34	2	0	2	4	**50**
Luminalny A	2	9	27	47	65	23	**173**
Luminalny B	0	11	11	14	18	44	**98**
SUMA	**89**	**54**	**44**	**61**	**87**	**72**	**407**

5 Charakterystyka kliniczna podpopulacji pacjentek

Zidentyfikowane podpopulacje chorych na raka piersi oceniono, badając profile kliniczne i demograficzne. Ta część analizy miała na celu przede wszystkim weryfikację, czy przeżywalność i wyniki kliniczne lub tło demograficzne niosą ze sobą jakiekolwiek znaczenie różnicujące i potwierdzają wyniki detekcji podpopulacji na podstawie profili białkowych. W szczególności część ta skupiała się na analizie porównawczej wykrytych podtypów luminalnych, które stanowiły główną modyfikację w stosunku do zestawu podtypów dostarczonych przez klasyfikator PAM50 oparty na transkryptomice. Celem było sprawdzenie, czy tło demograficzne, przeżycie i wyniki kliniczne wpływają na decyzję o podziale przypadków luminalnych na cztery podgrupy zamiast tylko dwóch luminalnych A i B, jak w przypadku predyktora PAM50.

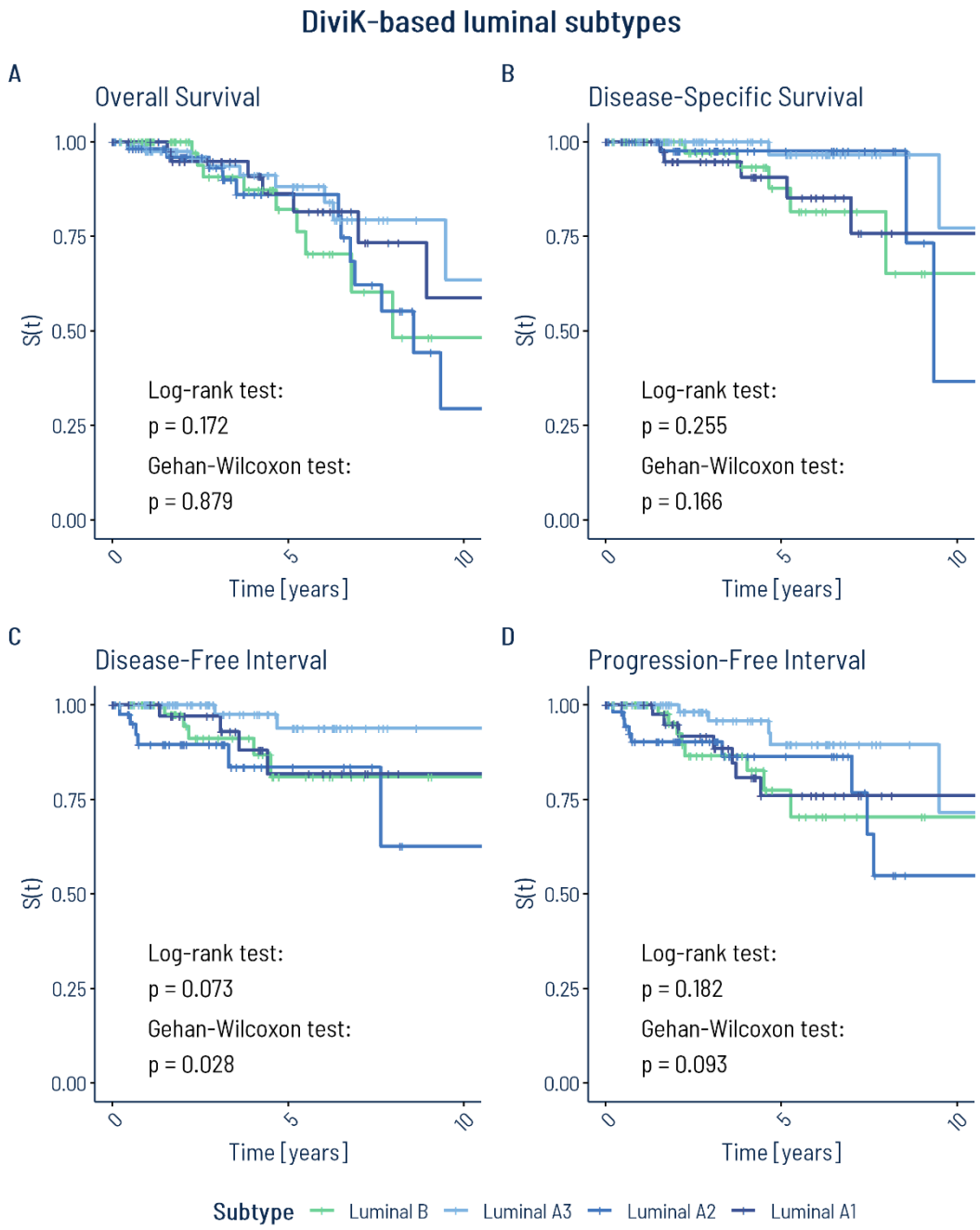
5.1 Analiza przeżywalności

Estymator funkcji przeżycia Kaplana-Meiera (KM) (Kaplan & Meier, 1992) został wykorzystany do wykreślenia krzywych przeżycia dla poszczególnych podpopulacji chorych na raka piersi. Przeżycia dla różnych podtypów zostało porównane wizualnie na podstawie wykresów KM. Przeprowadzono również odpowiednie testy statystyczne w celu ilościowego określenia różnic między grupami i sprawdzenia, czy były one istotne statystycznie. Dla każdego porównania wykonano test log-rank. Jest to najczęściej stosowane podejście, w którym taką samą wagę przykłada się do różnic pomiędzy funkcjami przeżycia w całym przedziale czasowym badania (Mantel, 1966; Peto & Peto, 1972; May, Hosmer, & Lemeshow, 2014). Jednak w przypadku niektórych porównań różnice w wynikach przeżycia były widoczne głównie w początkowych fazach choroby i terapii. Do porównania podpopulacji zastosowano więc uogólniony test Wilcoxona, zwany również testem Gehana-Wilcoxona. W tym podejściu wagi różnic między wynikami przeżycia określono jako liczbę pacjentów pozostających w badaniu.

Ponadto, wykorzystano model proporcjonalnego hazardu Coxa do oszacowania ilorazu hazardu (HR), odpowiadającego każdemu podtypowi w porównaniu z podtypem zdefiniowanym jako referencyjny (Cox, 1972). HR można uznać za miarę wielkości efektu, interpretowaną analogicznie do ryzyka względnego. Progi interpretacji HR zostały skorygowane ze względu na niezrównoważone wielkości porównywanych grup.

Analizę przeżycia przeprowadzono dla ogólnego przeżycia (Overall Survival - OS), przeżycia specyficznego dla raka piersi (Disease-Specific Survival - DSS), czasu bez choroby (Disease-Free Interval - DFI) oraz czasu bez progresji (Progression Free-Interval - PFI). Pierwsze dwie analizy nie są jednak zalecane dla projektu raka piersi w TCGA, ponieważ czas obserwacji jest zbyt krótki, aby zaobserwować wystarczającą liczbę zdarzeń.

Wykresy KM dla wszystkich czterech analiz przedstawiono na Rysunku 4 dla podpopulacji luminalnych zidentyfikowanych za pomocą DiviK. Porównanie podgrup luminalnych zostało tutaj wyróżnione w celu zbadania głównej różnicy pomiędzy wynikami podtypowania DiviK i PAM50. Podtypy HER2-wzbogacone i podstawny były wysoce zgodne pomiędzy oboma podejściami.



Rysunek 4. Krzywe przeżycia Kaplana-Meiera podpopulacji luminalnych zidentyfikowanych za pomocą DiviK

Ponadto w Tabeli 4 przedstawiono statystyki testowe i p-wartości dla podtypów luminalnych opartych na DiviK i PAM50.

Tabela 4. Wyniki testów log-rank i Gehana-Wilcoxona dla porównania funkcji przeżycia podtypów luminalnych zidentyfikowanych za pomocą DiviK lub na podstawie klasyfikatora PAM50

Punkt końcowy	χ^2		p-wartość	
	Test log-rank	Test Gehana-Wilcoxona	Test log-rank	Test Gehana-Wilcoxona
Podpopulacje wykryte za pomocą 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
Podtypy PAM50				
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

W przypadku porównywania podpopulacji luminalnych zidentyfikowanych za pomocą DiviK p-wartość była wyższa dla testu Gehana-Wilcoxona niż dla testu log-rank tylko dla OS, który jest najmniej wiarygodnym punktem końcowym spośród wszystkich tu rozpatrywanych. Nie można jednak dostrzec różnic w wynikach przeżywalności dla OS na podstawie wyników testu, ani krzywych KM. Kiedy położono większy nacisk na wczesne zmiany w przeżyciu w teście Gehana-Wilcoxona, p-wartość zmniejszyła się dla DSS, DFI i PFI. Wyniki te zostały również poparte wykresami KM, zwłaszcza dla DFI i PFI, gdzie można zaobserwować wyraźny spadek funkcji przeżycia dla przypadków luminalnych A2 w pierwszym roku obserwacji. P-wartość jest niższa od 0,05 tylko dla DFI. Dla DSS można zauważyć dwie grupy podobnych krzywych: jedną z podpopulacjami luminalnymi A2 i A3 o lepszym rokowaniu i jedną złożoną z podtypów luminalnych A1 i B o gorszym wyniku. Na podstawie wykresów KM można stwierdzić, że podtyp luminalny A3 może być powiązany z najlepszym rokowaniem dotyczącym nawrotów wśród wszystkich badanych podgrup chorych.

5.2 Analiza statystyczna demograficznego i klinicznego profilu

Kilka kategorycznych zmiennych odnoszących się do czynników demograficznych i klinicznych zostało przeanalizowanych, by zweryfikować ich związek z podpopulacjami

zidentyfikowanymi na podstawie danych RPPA. Oceniono również ich zależność od transkryptomicznych podtypów PAM50, aby porównać wyniki między tymi dwoma podejściami. W tym celu został przeprowadzony test niezależności χ^2 Pearsona. W przypadku tablic kontyngencyjnych 2x2, gdy testowano związek dwóch grup z dwiema kategoriami, zastosowano korektę Yatesa na nieciągłość. (Yates, 1934). Tabele kontyngencyjne powstałe dla różnych badanych kombinacji podtypów i zmiennych kategorycznych różniły się wymiarami. Utrudniało to porównanie wyników podtypowania za pomocą PAM50 i metody zaproponowaną w niniejszej rozprawie. P-wartość testu χ^2 Pearsona nie pozwala więc na zadowalającą charakterystykę zależności pomiędzy podtypami a czynnikami demograficznymi czy klinicznymi. W związku z tym, do oceny siły asocjacji obliczono wielkość efektu V Craméra. Wyniki analizy zależności przedstawiono w Tabeli 5 dla podzbioru podtypów luminalnych. Wartości V Craméra są pokolorowane na podstawie interpretacji wielkości efektu.

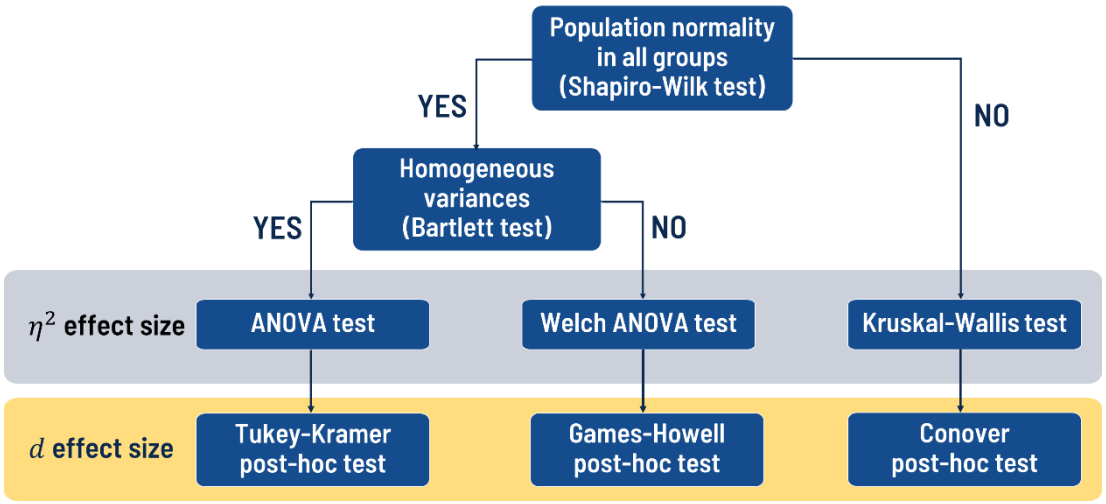
Tabela 5. Związek między kategorycznymi czynnikami demograficznymi i klinicznymi a podtypami luminalnymi zidentyfikowanymi za pomocą DiviK lub na podstawie klasyfikatora PAM50

Statystyki testowe i p-wartość z testu niezależności χ^2 Pearsona, wielkość efektu V Craméra oraz progi efektu dla małego, średniego i dużego efektu skorygowane ze względu liczbę kategorii lub grup.

Cecha	χ^2	p-wartość	V Craméra	Próg efektu V Craméra		
				Mały	Średni	Duży
Podpopulacje zidentyfikowane za pomocą DiviK						
Rasa	13.42	0.0368	0.1712	0.0707	0.2121	0.3536
Etniczność	0.23	0.9718	0.0346	0.1	0.3	0.5
AJCC Stadium	18.61	0.0287	0.1536	0.0577	0.1732	0.2887
AJCC Guz	19.34	0.0225	0.1566			
AJCC Węzły	13.23	0.1526	0.1292			
AJCC Guz (binarny)	13.86	0.0031	0.2295	0.1	0.3	0.5
AJCC Węzeł (binarny)	3.75	0.2900	0.1191			
AJCC przerzuty	2.23	0.5254	0.0922			
Podtypy PAM50						
Rasa	3.74	0.1543	0.1269	0.1	0.3	0.5
Etniczność	1.26	0.2610	0.0793			
AJCC Stadium	9.19	0.0269	0.1848			
AJCC Guz	14.40	0.0024	0.2309			
AJCC Węzły	0.91	0.8228	0.0580			
AJCC Guz (binarny)	13.25	0.0003	0.2215			
AJCC Węzeł (binarny)	0.67	0.4133	0.0497			
AJCC przerzuty	1.42	0.2335	0.0725			

Wyniki wskazują na niewielki, ale istotny statystycznie związek między podtypami opartymi na DiviK rozpatrywanymi łącznie a wszystkimi czynnikami kategorycznymi, poza pochodzeniem etnicznym i przerzutami, dla których efekt był pomijalny. Podobną zależność wykazano dla podtypów PAM50. Dla tego podejścia wykryto jednak niewielki związek z etnicznością, a nawet średni z rasą. Dla przypadków luminalnych efekt był również niewielki w odniesieniu do wszystkich czynników oprócz etniczności i przerzutów. Niemniej jednak dla zajętych węzłów chłonnych według AJCC nie wykazano istotnej zależności za pomocą testu Pearsona χ^2. Efekt był również nieistotny dla podtypów PAM50. Ponadto nie stwierdzono istotnej zależności pomiędzy czynnikami kategorycznymi a podpopulacjami luminalnymi A wykrytymi poprzez metodę DiviK za pomocą testu Pearsona χ^2. Wykryto natomiast niewielki efekt dla wszystkich czynników, poza pochodzeniem etnicznym i zbinaryzowaną wielkością guza.

Zmienne liczbowe wykorzystane do ewaluacji wyników podtypowania obejmowały wiek pacjenta w momencie diagnozy oraz oszacowania frakcji komórek układu odpornościowego CIBERSORT. Porównywano je między podpopulacjami za pomocą testów dobranych zgodnie z założeniami normalności i jednorodności wariancji. Ponadto, odpowiednie miary wielkości efektu wsparły klasyczne podejście do testowania. Na Rysunku 5 przedstawiono schemat testowania różnicowania dla porównania więcej niż dwóch podtypów.



Rysunek 5. Schemat testowania różnicowania dla porównania więcej niż dwóch grup

Podtypy we wszystkich badanych wariantach różniły się istotnie pod względem wieku. Niemniej jednak efekt ten był niewielki. Wyjątkiem było porównywanie przypadków PAM50

luminalnych A i B, dla których nie wykryto istotnych różnic, a wielkość efektu Cohena *d* sklasyfikowano jako bardzo małą.

Frakcje komórek układu odpornościowego różniły się istotnie pomiędzy wszystkimi podtypami dla 13 typów komórek: komórek B naiwnych i pamięci, komórek plazmatycznych, aktywowanych i spoczywających komórek T pamięci CD4, pęcherzykowych komórek T pomocniczych, monocytów, makrofagów M0, M1 i M2, spoczywających i aktywowanych komórek dendrytycznych oraz spoczywających komórek tucznych. Testy post hoc Conovera poparte wykresami wskazały na podwyższoną frakcję pęcherzykowych komórek T pomocniczych i brak spoczynkowych komórek tucznych w nowotworach podstawnych, znacząco odróżniając ten podtyp od innych. Wizualnie zweryfikowano, że dla komórek B pamięci, aktywowanych komórek T i komórek dendrytycznych występuje stosunkowo niewielka liczba niezerowych wyników.

Frakcje istotnie różniły się dla podzbioru podtypów luminalnych dla dziewięciu typów komórek odpornościowych: komórek B naiwnych i pamięci, komórek plazmatycznych, spoczynkowych komórek T CD4 pamięci, monocytów, makrofagów M0, M1 i M2 oraz spoczynkowych komórek dendrytycznych. Na podstawie testów post hoc Conovera, główne istotne różnice wykryto dla podtypu luminalnego A2 odniesionego do innych. Frakcja naiwnych komórek B była istotnie wyższa ze średnim efektem w podtypie luminalnym A2 w porównaniu z A1 i A3 oraz w podtypie luminalnym A3 w porównaniu z B. Największą liczbę niezerowych wartości zaobserwowano dla podtypu luminalnego A2. Ponadto frakcja komórek plazmatycznych była istotnie niższa w podtypie luminalnym A2 niż w luminalnym A3 i B, ze średnim efektem. W porównaniu z innymi podtypami luminalnymi, frakcje luminalne A2 makrofagów M1 były stosunkowo małe, a M2 były stosunkowo duże. Dla makrofagów M1 efekt był średni we wszystkich parach, natomiast dla M2 tylko w przypadku porównania przypadków luminalnych A1 i A3.

6 Sygnatury molekularne podpopulacji pacjentek

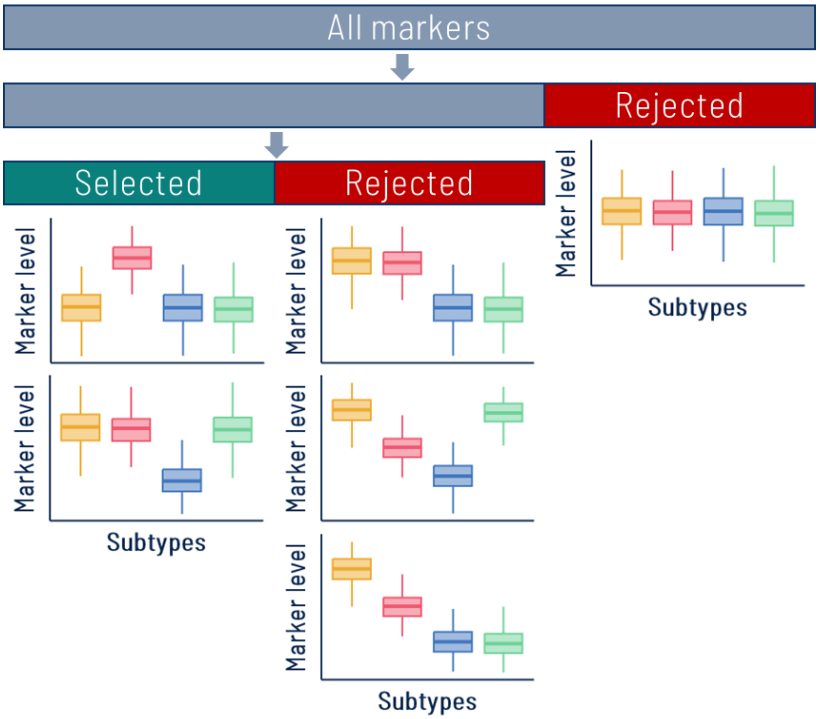
Podpopulacje chorych na raka piersi zostały wykryte za pomocą wybranej metody klasteryzacji zastosowanej na zbiorze danych RPPA. Spodziewano się zatem, że uzyskane podtypy będą różniły się poziomem białek. Niemniej jednak wymagana była dalsza analiza w celu identyfikacji profili białkowych charakterystycznych dla każdej z grup. Nie było również

wiadomo, czy podobne informacje można otrzymać na podstawie pomiarów ekspresji genów mRNA i czy sygnatury transkryptomiczne potwierdzą uzyskane podtypy. Ta część rozprawy ma zatem na celu scharakteryzowanie zidentyfikowanych podpopulacji chorych na raka piersi za pomocą sygnatur białkowych i transkryptomicznych, specyficznych dla danego podtypu lub pozwalających na rozróżnienie wszystkich podtypów.

6.1 Identyfikacja markerów specyficznych dla podtypu

Markery specyficzne dla podtypu zostały zidentyfikowane za pomocą procedury testowania różnicowania przedstawionej na Rysunku 5. Porównanie podtypów za pomocą wybranego podejścia przeprowadzono osobno dla każdego transkryptu lub białka. Testy na normalność i homogeniczność wariancji były również stosowane oddzielnie dla każdej cechy. Wyniki poddano następnie korekcie na wielokrotne testowanie Benjaminiego-Hochberga (Benjamini & Hochberg, 1995). Wyniki interpretowane były zbiorczo dla każdego zbioru danych, aby wszystkie pomiary z danej platformy były poddane tej samej ścieżce testowania.

Markery były identyfikowane na podstawie p-wartości bądź wielkości efektu. Biorąc pod uwagę dużą liczbę porównań i różniącą się wielkość podpopulacji, podejście wykorzystujące wielkość efektu okazało się być bardziej wiarygodnym rozwiązaniem. Markery specyficzne dla podtypu zostały zdefiniowane jako białka lub transkrypty o znacząco wyższym lub niższym poziomie tylko w jednym podtypie. Markery zostały zidentyfikowane w trzech przestrzeniach cech: danych proteomicznych, danych transkryptomicznych oraz danych transkryptomicznych ograniczonych do genów kodujących białka zmierzone za pomocą platformy RPPA. Na Rysunku 6 przedstawiono schemat procesu identyfikacji markerów specyficznych dla podtypu.



Rysunek 6. Proces identyfikacji markerów specyficznych dla podtypu

W Tabeli 6 przedstawiono liczby markerów specyficznych dla podtypu, zidentyfikowanych za pomocą podejścia opartego na wielkości efektu. Liczba markerów transkryptomicznych jest większa ze względu na większą przestrzeń cech.

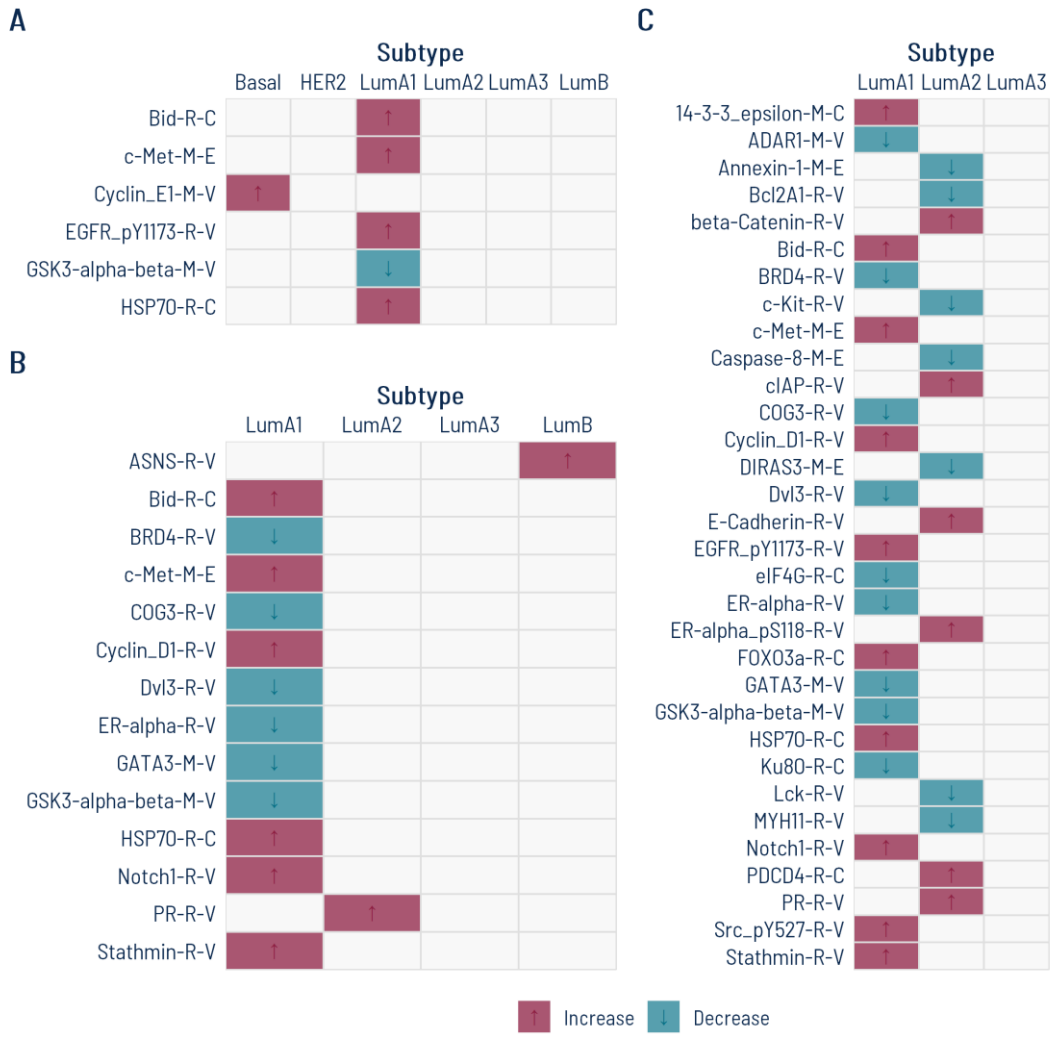
Tabela 6. Liczba markerów specyficznych dla podtypu wybranych na podstawie wielkości efektu

"P" oznacza zestaw danych dotyczących poziomów białek, "T" oznacza cały zestaw danych dotyczących poziomów ekspresji genów mRNA (transkryptomiczny), a "LT" oznacza zestaw danych transkryptomicznych ograniczony do genów kodujących białka zawarte w zestawie proteomicznym. (*) oznacza, że progi stosowane do interpretacji wielkości efektu η^2 i d Cohena zostały obniżone odpowiednio do średniej i dużej dla danych transkryptomicznych.

Podzbiór podtypów	**Wszystkie**			**Luminalne**			**Luminalne A**		
Przestrzeń cech	**P**	**T**	**LT**	**P**	**T**	**LT**	**P**	**T**	**LT**
Podstawny	1	1146	9	-	-	-	-	-	-
HER2-wzbogacony	0	21	0	-	-	-	-	-	-
Luminalny A1	5	0	0	12	0	0	19	1	0
Luminalny A2	0	2	0	1	13	1	13	45	1
Luminalny A3	0	0	0	0	0	0	0	0	0
Luminalny B	0	0	0	1	33	2	-	-	-
SUMA	**6**	**1169**	**9**	**14**	**46**	**3**	**32**	**46**	**1**

Na poziomie transkryptomicznym najbardziej znaczące różnice zaobserwowano dla podpopulacji podstawnej, z wieloma specyficznymi markerami. Ponadto, identyfikacja

markerów specyficznych dla podtypu HER2-wzbogaconego powiodła się jedynie na podstawie poziomu ekspresji genów mRNA. W celu selekcji markerów charakterystycznych dla podtypów luminalnych usunięto przypadki podstawne i HER2-wzbogacone. Następnie, największą liczbę specyficznych markerów stwierdzono dla podpopulacji luminalnych B i A2; ta ostatnia obserwacja została również wzmocniona przez późniejsze porównanie przypadków luminalnych A. Podejście oparte na wielkości efektu okazało się bardziej restrykcyjne. Zidentyfikowane markery specyficzne dla podtypu są wymienione w odniesieniu do kierunku zmian poziomu w porównaniu z innymi podtypami na Rysunku 7 dla proteomicznego zbioru danych.



Rysunek 7. Markery specyficzne dla podtypu zidentyfikowane na podstawie poziomu białek

Panele A, B i C zawierają markery wybrane przez porównanie odpowiednio wszystkich podtypów, podtypów luminalnych i podtypów luminalnych A. Fioletowe i turkusowe kolory wskazują, że poziom markera był odpowiednio wyższy lub niższy dla danego podtypu niż dla wszystkich pozostałych.

Analiza nadreprezentacji (Over-Representation Analysis - ORA) została przeprowadzona na zestawach otrzymanych specyficznych markerów, w oparciu o ścieżki sygnałowe oraz bazę Molecular Signatures Database (MSigDB) (Liberzon, et al., 2011; Liberzon, et al., 2015; Subramanian, et al., 2005).

Ze względu na stosunkowo małą liczbę zidentyfikowanych markerów dla obu zestawów danych i niewystarczającą ogólną liczbę białek zmierzonych metodą RPPA, ORA nie dała znaczących wyników dla ścieżek KEGG po zastosowaniu korekty Benjamini-Hochberga na wielokrotne testowanie (Benjamini & Hochberg, 1995). Jednakże, dla zbiorów MSigDB i przestrzeni cech transkryptomicznych wiele zestawów genów okazało się nadreprezentowanych w uzyskanych listach markerów specyficznych, szczególnie dla raków podstawnych i HER2-wzbogaconych. W przypadku transkryptów specyficznych dla podtypu podstawnego, wzbogacone zostały zestawy genów związanych z wczesną lub późną reakcją na estrogeny. Ponadto, ORA zastosowana na MSigDB ujawniła kilka powiązań z wcześniej opublikowanymi zestawami genów związanych z rakiem piersi, głównie w kontekście markerów specyficznych dla podtypów HER2-wzbogaconego i podstawnego (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).

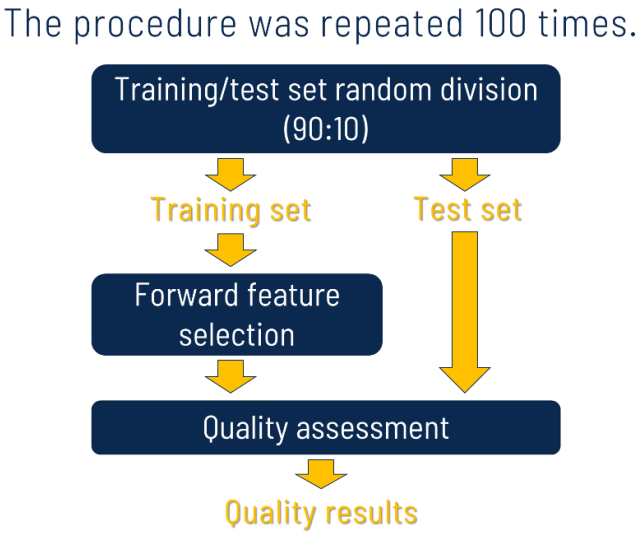
W celu rozwiązania problemu niewystarczającej wielkości zbioru dla ORA i dalszego zbadania różnic pomiędzy czterema ujawnionymi podpopulacjami luminalnymi, zastosowano test CERNO na bezwzględnych wartościach wielkości efektu d dla każdej pary podtypów luminalnych. Dla zestawu danych proteomicznych, po zastosowaniu korekty Benjaminiego-Hochberga (Benjamini & Hochberg, 1995), tylko zestawienie podtypów luminalnych A2 i B dostarczyło statystycznie istotnych wyników wzbogacenia. Dla zestawu danych transkryptomicznych wszystkie porównania parami dały istotnie wzbogacone ścieżki KEGG. Niezależnie od wariantu porównania, wśród uzyskanych ścieżek znalazły się te kluczowe dla prawidłowego funkcjonowania komórek oraz wiele związanych z biologią nowotworów.

6.2 Sygnatura różnicująca podtypy

Zaproponowano inne podejście oparte na wielomianowej regresji logistycznej do identyfikacji sygnatury molekularnej, wyróżniającej wszystkie rozważane podtypy. Regresja logistyczna jest powszechnie stosowana jako metoda klasyfikacji. Jednak może ona również posłużyć

do wyboru znaczących cech, takich jak sygnatura molekularna zidentyfikowanych podpopulacji.

Do budowy modelu zastosowano procedurę Multiple Random Cross-Validation (MRCV) ze 100 iteracjami. MRCV wybrano ze względu na ograniczoną liczbę pacjentek i wyraźny brak równowagi pomiędzy licznościami podpopulacji raka piersi. W każdej iteracji 10% pacjentek z każdego podtypu pozostawiano jako zbiór testowy, a pozostałe 90% służyło do dopasowania modelu. Na tym zbiorze budowano model wielomianowej regresji logistycznej metodą dołączania. W każdym kroku wybierano model o najwyższym czynniku Bayesa (Bayes Factor - BF), aż BF spadł poniżej 10 lub nie pozostało więcej potencjalnych cech do wyboru. Jakość powstałego modelu była oceniana na podstawie zbioru testowego. Na Rysunku 8 przedstawiono schemat procedury MRCV.



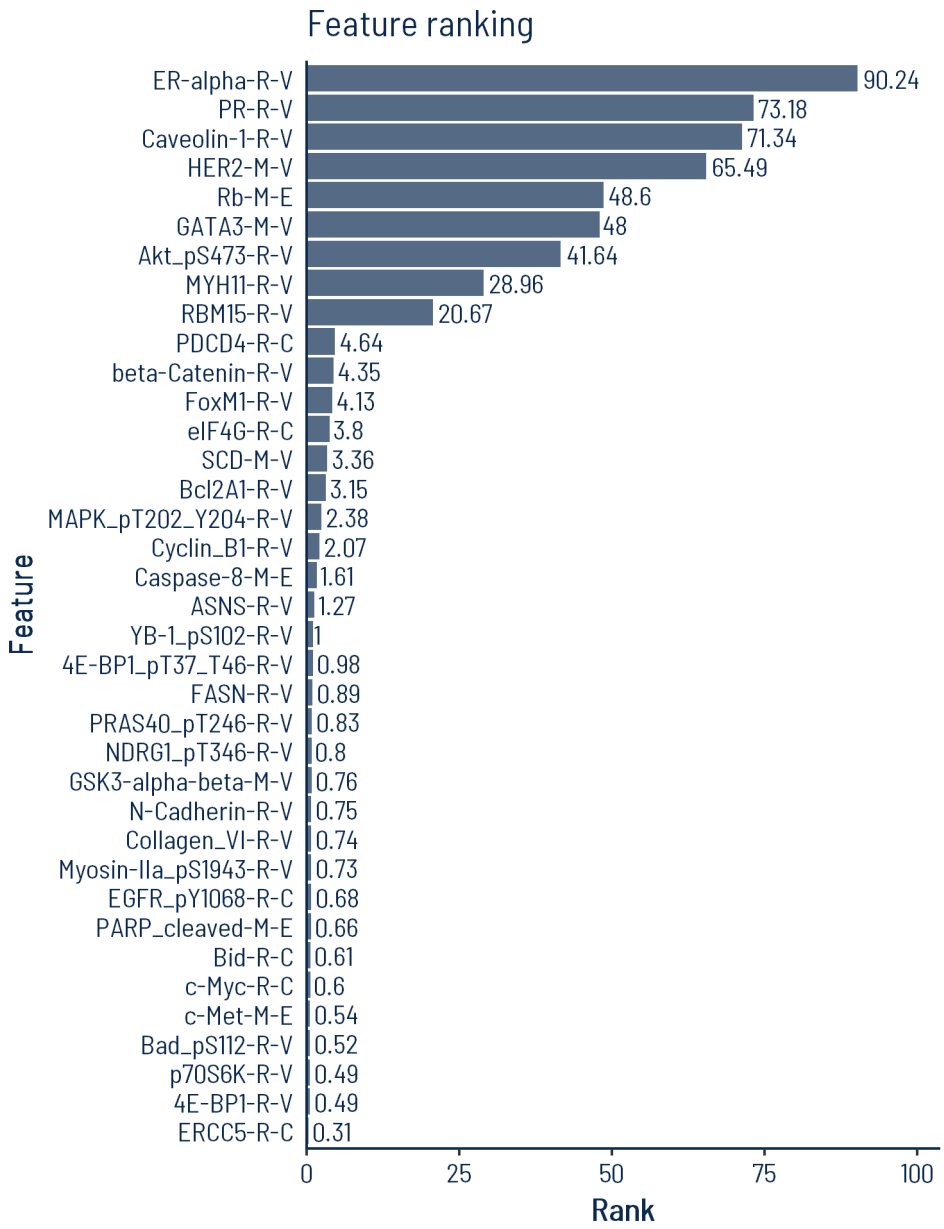
Rysunek 8. Procedura MRCV do budowy modelu wielomianowej regresji logistycznej

Wyniki 100 powtórzeń MRCV posłużyły do stworzenia rankingu cech. Cechy zostały posortowane i przypisano im wagi dla każdego modelu na podstawie kolejności wyboru.

Każda waga została pomnożona przez ogólną zbalansowaną dokładność modelu (BA) obliczoną na każdym zbiorze testowym. Zsumowane iloczyny dla wszystkich 100 modeli dały wynik rankingu dla każdej cechy. W ten sposób ranking cech łączy dwa podejścia oceny modeli: oparte na dobroci dopasowania (ang. goodness-of-fit-based), gdyż kolejność cech odpowiada BF, oraz oparte na jakości predykcji (prediction-quality-based), reprezentowane przez BA. Ranking cech posłużył do zidentyfikowania ostatecznej sygnatury molekularnej różnicującej wszystkie podtypy. Do wyboru progu odcięcia dla najlepszych cech zastosowano

metodę łokciową. Polegała ona na posortowaniu wyników rankingu cech, wykreśleniu ich i połączeniu linią najwyższych i najniższych wartości. Punktem przegięcia była wartość rankingu leżąca najdalej od linii. Wszystkie cechy z wynikami wyższymi niż punkt przegięcia były wybierane jako sygnatura modelu. Autorka opisała analogiczną metodę dla binarnej regresji logistycznej w (Henzel, et al., 2021) and (Kozielski, et al., 2021). Dopasowano trzy warianty modeli regresji: dla zestawu danych proteomicznych, dla zredukowanego zestawu danych transkryptomicznych oraz dla tych dwóch zestawów danych łącznie.

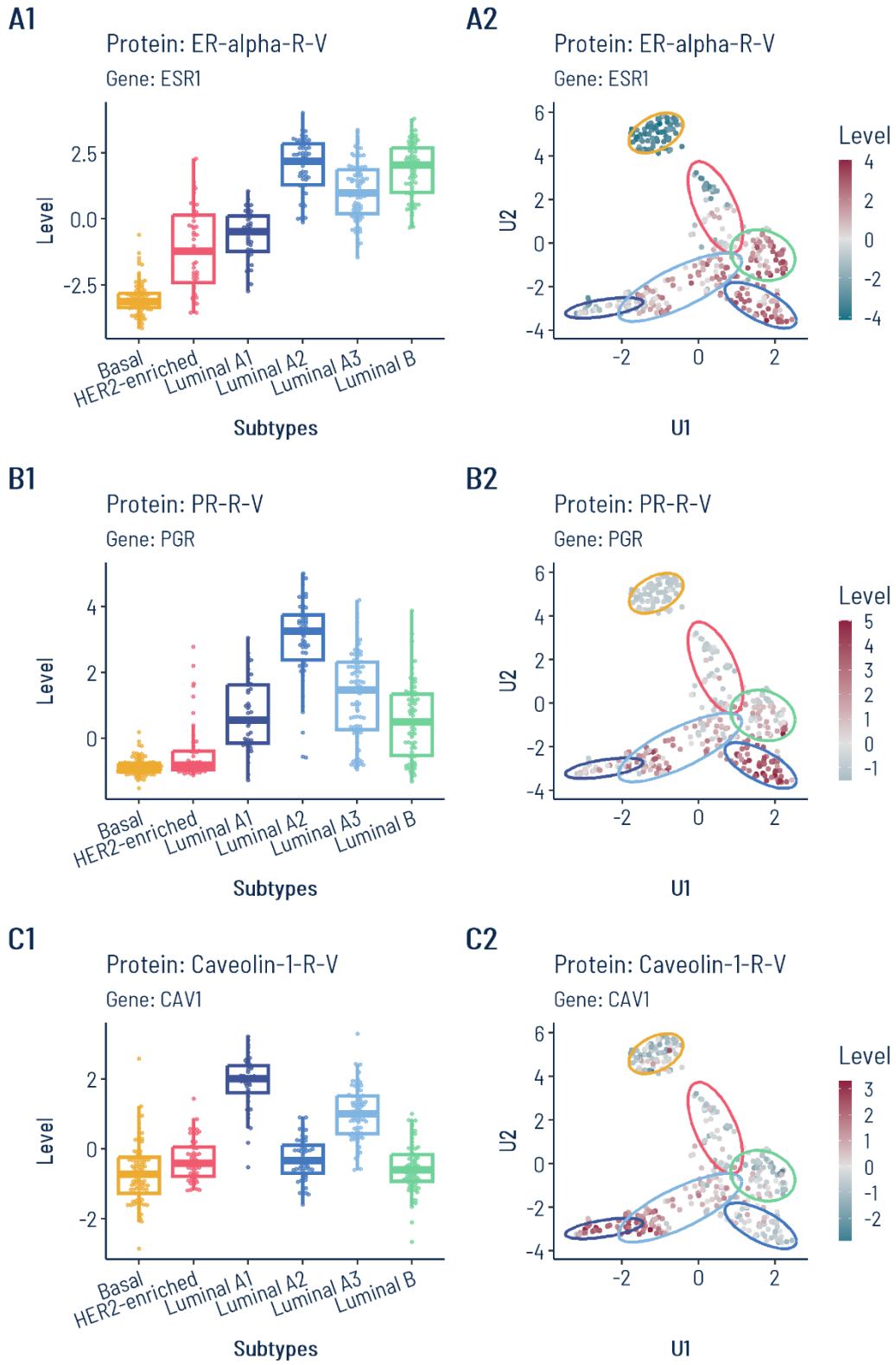
Na Rysunku 9 przedstawiono wyniki rankingu cech dla każdego białka uzyskane w procedurze MRCV. Dla przejrzystości wykres został obcięty, aby pokazać tylko najlepsze cechy, bez tych pojawiających się tylko w jednej iteracji MRCV. 9 najlepszych białek zostało zidentyfikowanych jako sygnatura proteomiczna w oparciu o metodę łokciową.



Rysunek 9. Ranking cech dla proteomicznego multimodalnego modelu regresji logistycznej

Dla przejrzystości wykres został obcięty, aby pokazać tylko cechy wybrane dla więcej niż jednego modelu w procedurze MRCV.

Poziomy białek wchodzących w skład sygnatury proteomicznej różnicującej podtypy przedstawiono na Rysunku 10 dla trzech najlepszych białek.



Rysunek 10. Poziomy trzech najlepszych białek wybranych do modelu regresji wielomianowej dotyczące podpopulacji zidentyfikowanych za pomocą DiviK

Po lewej stronie zaprezentowano wykresy pudełkowe poziomów białek dla każdego podtypu. Po prawej stronie zaprezentowano projekcję UMAP uzyskaną na podstawie zbiorze danych proteomicznych, przy czym kolor punktów odzwierciedla poziom białka.

Na Rysunku 11 porównano wybraną białkową sygnaturę modelu oraz zestawy specyficznych dla podtypu markerów zidentyfikowanych na podstawie wielkości efektu między wszystkimi podpopulacjami luminalnymi (panel A) lub między podpopulacjami luminalnymi A (panel B). Sygnatura modelu i luminalne specyficzne markery podtypu miały wspólne trzy białka.

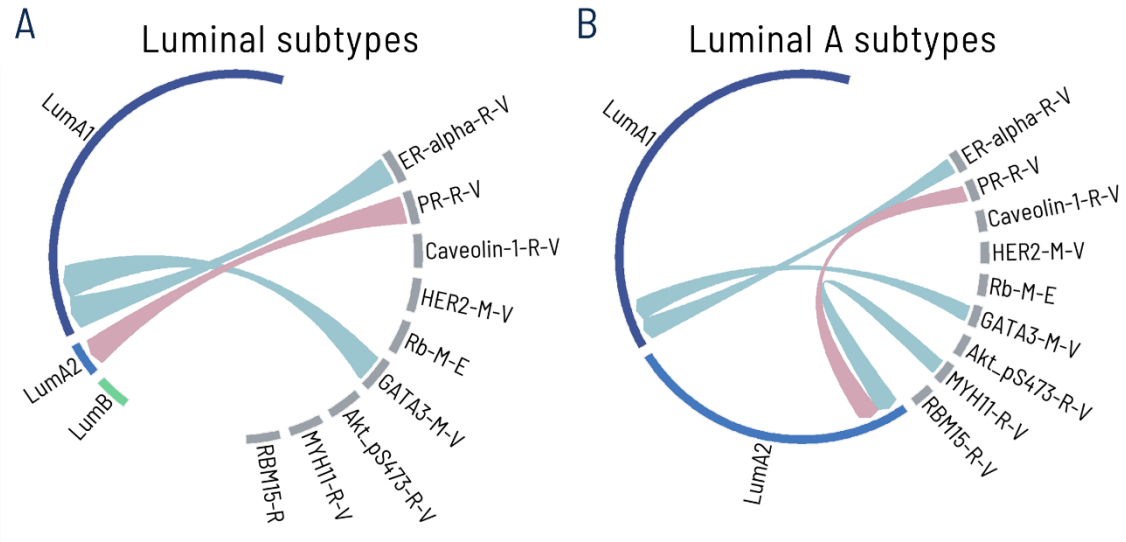
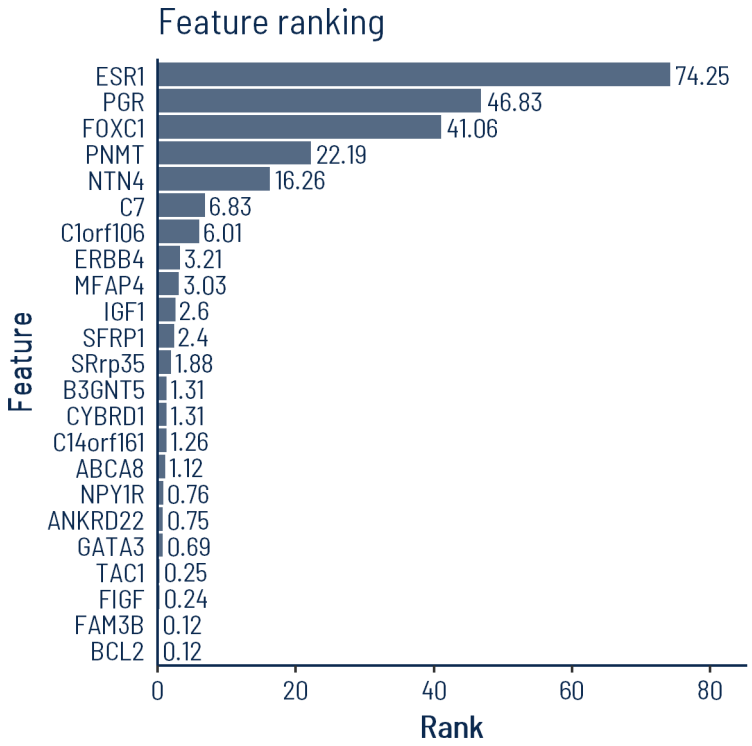


Figure 11. Porównanie cech modelu proteomicznego i markerów specyficznych dla podtypu proteomicznego, zidentyfikowanych na podstawie wielkości efektu

Połączenie oznacza, że określone białko zostało uwzględnione w modelu i zidentyfikowane jako marker specyficzny dla podtypu. Różowe i turkusowe kolory wskazują na wzrost lub spadek poziomu białka w porównaniu z innymi podtypami luminalnymi (panel A) lub innymi podtypami luminalnymi A (panel B).

Po usunięciu brakujących wartości, zbiór danych ekspresji genów na poziomie mRNA zawierał pomiary dla 17328 genów. Selekcja cech metodą dołączania byłaby niewystarczająca, dlatego zbiór danych ograniczono do jedynie 1124 genów o najwyższej wariancji. Próg wariancji został określony na podstawie dekompozycji GMM.

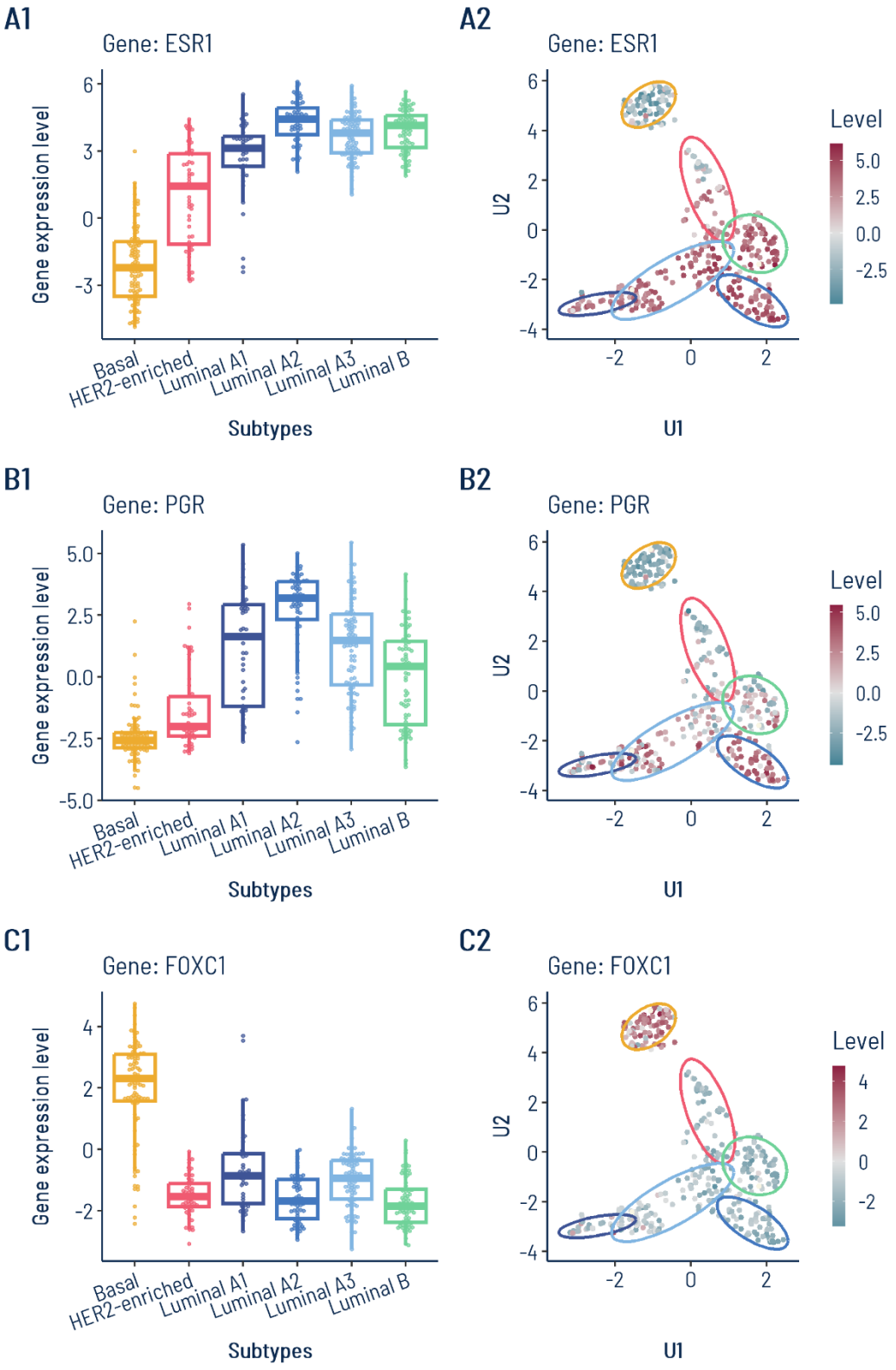
Ranking cech uzyskany w procedurze MRCV przedstawiono na Rysunku 12. Maksymalną odległość na wykresie łokciowym uzyskano dla szóstego genu (*C7*). W ten sposób pięć najlepszych genów utworzyło transkryptomiczną sygnaturę różnicującą podpopulacje. Poziomy ekspresji mRNA tych trzech najlepszych wybranych genów przedstawiono na Rysunku 13.



Rysunek 12. Ranking cech dla transkryptomicznego multimodalnego modelu regresji logistycznej

Dwa pierwsze geny wybrane do modelu (*ESR1* i *PGR*) kodują dwa najlepsze białka z sygnatury proteomicznej (receptory estrogenu i progesteronu). Niemniej jednak odpowiadające im geny i białka nie wykazywały tego samego wzorca, zwłaszcza w przypadku podpopulacji luminalnej A1.

Połączony zestaw pomiarów dla 166 białek i 1124 genów po filtracji opartej na GMM posłużył do stworzenia wspólnego modelu wielomianowej regresji logistycznej. Dziewięć najlepszych białek zostało zidentyfikowanych jako połączona sygnatura. Co ciekawe, wszystkie te cechy były proteomiczne, gdyż pierwszy poziom ekspresji genów mRNA zajął dopiero jedenastą pozycję w rankingu. Co więcej, kolejność tych najlepszych cech jest identyczna jak w przypadku modelu tylko proteomicznego.



Rysunek 13. Poziomy trzech najlepszych transkryptów wybranych do modelu regresji wielomianowej dotyczące podpopulacji zidentyfikowanych za pomocą DiviK

Po lewej stronie zaprezentowano wykresy pudełkowe poziomów białek dla każdego podtypu. Po prawej stronie zaprezentowano projekcję UMAP uzyskaną na podstawie zbiorze danych proteomicznych, przy czym kolor punktów odzwierciedla poziom białka.

7 Wnioski

Cele niniejszej rozprawy w zakresie identyfikacji podpopulacji chorych na raka piersi oraz ich oceny klinicznej i molekularnej zostały osiągnięte. Wyniki opisane w niniejszej rozprawie uzasadniają postawione tezy. Teza I została potwierdzona wynikami analizy przedstawionymi w rozdziale 4 (Identyfikacja podpopulacji pacjentek). Wykazano, że różne kombinacje algorytmów inżynierii cech i klasteryzacji umożliwiają wykrycie nowych podpopulacji chorych na raka piersi w oparciu o ich profile białkowe. Zaproponowane metryki porównania wyników klasteryzacji pozwoliły na wybór podejścia dającego najbardziej odrębne podpopulacje. Teza II została udowodniona w rozdziałach 5 (Charakterystyka kliniczna podpopulacji pacjentek) i 6 (Sygnatury molekularne podpopulacji pacjentek). Potwierdzono różnice w przeżywalności pomiędzy zdefiniowanymi podpopulacjami. Wykazano, że podtypy HR+ i HR- różnią się rokowaniem, a nowo ujawnione dodatkowe podgrupy luminalne były zróżnicowane pod względem wyników przeżywalności. Stwierdzono niewielki związek badanych podpopulacji z czynnikami demograficznymi lub klinicznymi, podobnie jak w przypadku podtypów opartych na PAM50. Wykryto również, że zidentyfikowane podpopulacje wykazują zróżnicowanie we frakcjach komórek układu odpornościowego, w tym również w podgrupach luminalnych. Metodologia testowania różnicowania oparta na klasycznych testach statystycznych i wielkości efektu pozwolił na określenie i funkcjonalną charakterystykę profili proteomicznych i transkryptomicznych większości ujawnionych podpopulacji. Wyłoniono sygnaturę białkową rozpoznającą wszystkie podtypy. Sygnatura transkryptomiczna pozwoliła głównie na rozdzielenie podtypów HR+ i HR-, ale słabo radziła sobie z rozróżnieniem wykrytych w tej pracy podtypów luminalnych.

Niniejsza rozprawa odpowiada na potrzeby ponownej identyfikacji ustalonej klasyfikacji raka piersi z wykorzystaniem uczenia maszynowego i podejść modelowania matematycznego. Najpierw, techniki uczenia maszynowego pozwoliły wydzielić podpopulacje chorych na raka piersi na podstawie poziomów białek. Następnie, uzyskane grupy zostały ocenione pod względem czynników demograficznych i klinicznych. Wreszcie, podtypy zostały scharakteryzowane molekularnie przy użyciu kompleksowych metod statystycznych i podejść uczenia statystycznego. Zaproponowana w tej pracy metodologia dostarczyła zadowalających wyników i poradziła sobie z wymagającym zbiorem danych.

Wszystkie zastosowane metody uczenia maszynowego wskazały, że podtyp luminalny A jest najbardziej zróżnicowany w zbiorze TCGA-BRCA i powinien być dalej podzielony na dwie lub trzy podgrupy. Etapy selekcji lub ekstrakcji cech przed klasteryzacją miały kluczowe znaczenie dla jakości wyników. Filtracja cech oparta na GMM poprawiła wykrywalność wysoce odrębnych klastrów, niezależnie od algorytmu klastrowania. Proponowane podejście oparte na centroidach z iteracyjnym grupowaniem k-średnich w lokalnie przefiltrowanej przestrzeni cech za pomocą GMM zapewniło najlepsze wyniki spośród wszystkich sprawdzonych podejść. Zidentyfikowano sześć podpopulacji chorych nazwanych na podstawie ich zgodności z etykietami PAM50 jako podtyp podstawny, HER2-wzbogacony, luminalny B oraz trzy podgrupy luminalne A: A1, A2 i A3.

Ocena demograficzna i kliniczna zidentyfikowanych podpopulacji podkreśliła znaczenie odpowiedniego podejścia do testowania statystycznego. Biorąc pod uwagę niewystarczający czas obserwacji dla nowotworów o stosunkowo dobrym rokowaniu, kluczowe było właściwe zdefiniowanie punktu końcowego odnoszącego się do czasu do nawrotu choroby, a nie do śmierci. Ponadto rozszerzenie klasycznego testu log-rank o ważone podejście Gehana-Wilcoxona umożliwiło wykrycie istotnych wczesnych zmian w przeżywalności między podpopulacjami. Oszacowanie wielkości efektu za pomocą HR interpretowanego z korektą dla niezrównoważonych grup częściowo rozwiązało problem zróżnicowanej wielkości prób badawczych i umożliwiło porównanie podpopulacji pomimo niewielkiej liczby zdarzeń uchwyconych podczas obserwacji. Wielkość efektu V Craméra pozwoliła na analizę związku między podpopulacjami a czynnikami demograficznymi lub klinicznymi w sposób dostosowany do różnej liczebności kategorii.

Wykazano większe zróżnicowanie w przeżywalności niż w przypadku powszechnie stosowanych podtypów PAM50. Co ciekawe, wykryte podtypy luminalne różniły się wynikami przeżywalności, zwłaszcza pod względem czasu do wystąpienia nowych zdarzeń nowotworowych. Podtyp luminalny A2 charakteryzował się rokowaniem porównywalnie złym do guzów HER2-wzbogaconych i podstawnych. Z drugiej strony, przypadki luminalne A3 wykazywały korzystne rokowania.

Podpopulacje ujawnione w tym badaniu na podstawie portretu białkowego wykazały niewielką zależność od czynników demograficznych i klinicznych, porównywalną z dobrze poznanymi podtypami PAM50. Cztery podtypy luminalne zidentyfikowane w tej rozprawie

wykazały niewielki związek z zajętymi węzłami chłonnymi, czego nie zaobserwowano w przypadku klasyfikacji PAM50 podtypów luminalnych A i B. Ponadto okazało się, że zaproponowane tu podpopulacje różnią się odpowiedzią immunologiczną zarówno wśród całej kohorty, jak i tylko w grupie luminalnej.

Klasyczne testy statystyczne i wielkość efektu zostały wykorzystane do identyfikacji niespecyficznych i specyficznych dla podtypu markerów zarówno w przestrzeni proteomicznej, jak i transkryptomicznej. Ze względu na dużą liczbę cech w porównaniu z wielkością próby, wielkość efektu przewyższała podejście klasyczne i dostarczyła bardziej rygorystycznej listy markerów specyficznych dla podtypów. Zróżnicowanie między podtypami w przestrzeni transkryptomicznej było mniejsze niż w proteomicznej.

Wybór metody był również kluczowy dla analizy funkcjonalnej. Ze względu na niewystarczające rozmiary listy markerów i małą liczbę wszystkich zmierzonych białek, metoda pierwszej generacji ORA nie dała zadowalających wyników. Niemniej jednak, test CERNO drugiej generacji przeprowadzony na oszacowaniach wielkości efektu dostarczył listy istotnie wzbogaconych ścieżek. Wyniki wskazują na wyraźne różnice pomiędzy zidentyfikowanymi podpopulacjami na poziomie transkryptomicznym i proteomicznym, w tym na znaczne zróżnicowanie w obrębie grupy luminalnej. Różnicujące geny i białka są zaangażowane w procesy mające znaczenie dla prawidłowego funkcjonowania komórek i rozwoju nowotworu.

Dedykowane podejście uczenia maszynowego pozwoliło zidentyfikować sygnaturę białkową odróżniającą wszystkie sześć ujawnionych podtypów. Podobnie uzyskano sygnaturę transkryptomiczną. Niektóre z sygnaturowych genów i białek mają dobrze poznaną rolę w raku piersi. Dla innych jednak związek z tą chorobą pozostaje nieznany.

Podtyp luminalny A1 wykazał wyraźne różnice w ekspresji genów i białek sygnaturowych w porównaniu do trzech pozostałych podgrup luminalnych. Wykazano pewne podobieństwa do nowotworów podstawnych i HER2-wzbogaconych, a także wyraźne różnice w porównaniu do wszystkich podtypów. Ponadto, zaobserwowano względny spadek ekspresji ER pomiędzy poziomem mRNA i białkowym. Sugeruje to, że przypadki luminalne A1 mogły zostać błędnie sklasyfikowane jako luminalne na podstawie profilowania genów i są bliższe guzom ER-, co nie może być odzwierciedlone w ich portretach transkryptomicznych.

Podsumowując, dane proteomiczne niosą informacje dotyczące podziału raka piersi, które pozostają ukryte na poziomie transkryptomicznym. Podtypowanie oparte na profilu białkowym uzupełnia klasyfikację molekularną raka piersi i dostarcza lepszych informacji na temat jego heterogeniczność, których nie odzwierciedla profilowanie ekspresji genów. W regulacji ekspresji pomiędzy warstwą mRNA a białkową biorą udział różne mechanizmy. Wyniki uzyskane w tej pracy sugerują, że procesy te wpływają na zachowanie nowotworu. Proteomiczne podpopulacje pacjentek wykazują różnice w wynikach klinicznych, których nie obserwowano w podtypach luminalnych PAM50. Profilowanie poziomu białek może zatem potencjalnie dostarczyć bardziej wszechstronnego wglądu w biologię guza i dostarczyć klinicznie istotnych informacji poza profilowaniem ekspresji genów. Zidentyfikowane markery mogą posłużyć do optymalizacji planowania terapii i przyczynić się do badań nad nowymi jej celami. Niemniej jednak, konieczna jest dalsza niezależna walidacja, aby potwierdzić potencjalne zastosowania prognostyczne i kliniczne oraz ocenić, czy obecne kliniczne i molekularne sposoby podtypowania mogą być uzupełnione o te wyniki i zastosowane w praktyce klinicznej.

8 Bibliografia

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