Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Lana Semenić

Integracija podataka ekspresijskih mikročipova za istraživanje genskog potpisa upalnog raka dojke strojnim učenjem

Zagreb, 2024.

Ovaj rad izrađen je na Zavodu za molekularnu biologiju Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu pod vodstvom izv. prof. dr. sc. Rose Karlić i predan je na natječaj za dodjelu Rektorove nagrade u akademskoj godini 2023./2024.

Kratice

ACSM1	član obitelji acil-CoA sintetaza srednjih lanaca 1 (engl. *acyl-CoA synthetase medium chain family member 1*)
ACVRL1	receptor nalik na aktivin A tipa I (engl. *activin A receptor like type 1*)
ADGRF5	adhezijski receptor povezan s G proteinom F5 (engl. *adhesion G protein-coupled receptor F5*)
AJCC	Američka komisija za rak (engl. *American Joint Committee on Cancer*)
BCL2	protein leukemije/limfoma B-stanica 2 (engl. *B-cell leukemia/lymphoma 2 protein*)
BRINP3	neuralno specifični protein induciran koštanim morfogenetskim proteinom i retinoičnom kiselinom 3 (engl. *bone morphogenetic protein/retinoic acid inducible neural specific 3*)
cDNA	komplementarna DNA (engl. *copy DNA, complementary DNA*)
CENPF	centromerni protein F (engl. *centromere protein F*)
CDH5	kadherin 5 (engl. *cadherin* 5)
CROT	karnitin O-oktanoiltransferaza (engl. *carnitine O-octanoyltransferase*)
D903	skup 903 gena s najrazličitijom prosječnom ekspresijom između IBC i nIBC uzoraka
ECM	izvanstanični matriks (engl. *extracellular matrix*)
ER	estrogenski receptor (engl. *estrogen receptor*)
ERBB2	eritroblastični onkogen B 2 (engl. *erythroblastic oncogene B 2*)
FOS	Finkel-Biskis-Jinkins onkogen osteosarkoma (engl. *Finkel-Biskis-Jinkins osteosarcoma oncogene*)
G59	genski potpis upalnog raka dojke prema Zare, Postovit i Githaka (2021)
GEO	Omnibus ekspresije gena (engl. *Gene Expression Omnibus*)
GEP	profil ekspresije gena (engl. *gene expression profile*)
GO	Genska ontologija (engl. *Gene Ontology*)
GPR87	receptor povezan s G proteinom 87 (engl. *G protein-coupled receptor 87*)
HC115	skup 115 gena dobiven na temelju *XGBoost* modela pomoću hijerarhijskog klasteriranja
HER2	receptor ljudskog epidermalnog faktora rasta 2 (engl. *human epidermal growth factor receptor 2*)
IBC	upalni rak dojke (engl. *inflammatory breast cancer*)
IHC	imunohistokemija (engl. *immunohistochemistry*)
M912	skup 912 gena uključenih u *XGBoost* model
MMP	metaloproteinaza matriksa (engl. *matrix metalloproteinase*)
NBR2	gen susjedan *BRCA1* 2 (engl. neighbor of *BRCA1* 2)
nIBC	neupalni rak dojke (engl. *non-inflammatory breast cancer*)

NXN	nukleoredoksin (engl. *nucleoredoxin*)
ORA	analiza prezastupljenosti (engl. *overrepresentation analysis*)
ORC6L	podjedinica 6 kompleksa za prepoznavanje ishodišta replikacije (engl. *origin recognition complex subunit 6*)
PAM	prediktivna analiza mikročipa (engl. *prediction analysis of microarray*)
PCA	analiza glavnih komponenti (engl. *principal component analysis*)
PR	progesteronski receptor (engl. *progesterone receptor*)
RNA-seq	RNA sekvenciranje (engl. *RNA sequencing*)
ROSE	nasumično prekomjerno uzorkovanje (engl. *random over-sampling examples*)
S1PR1	receptor sfingozin-1-fosfata 1 (engl. *sphingosine-1-phosphate receptor 1*)
SEMA3E	semaforin 3E (engl. *semaphorin 3E*)
SERINC2	uključitelj serina 2 (engl. *serine incorporator 2*)
SLC12A1	član obitelji nositelja otopljene tvari 12 1 (engl. *solute carrier family 12 member 1*)
SLCO3A1	član obitelji prijenosnika otopljenih organskih aniona 3A1 (engl. *solute carrier organic anion transporter family member 3A1*)
TCIM	regulator transkripcije i imunosnog odgovora (engl. *transcriptional and immune response regulator*)
TDM	istreniran model poklapanja distribucija (engl. *training distribution matching*)
TGF	transformirajući faktor rasta (engl. *transforming growth factor*)
TNM	tumor, (limfni) čvor, metastaza (engl. *tumor, node, metastasis*)
TYMS	timidilat sintetaza (engl. *thymidylate synthetase*)

Oznake u matematičkim izrazima

A	točnost (engl. *accuracy*)
a	srednja udaljenost opservacije od ostalih opservacija u njenom klasteru
b	minimalna srednja udaljenost opservacije od opservacija iz drugih klastera
C	klaster (engl. *cluster*)
IQR	interkvartilni raspon (engl. *interquartile range*)
κ	Cohenova kapa (engl. *Cohen's kappa*)
Q_n	n-ti kvartil (engl. *nth quartile*)
SD	standardna devijacija (engl. *standard deviation*)

Sadržaj

1. Uvod

1.1 Rak dojke

1.1.1 Obilježja raka dojke i njegova klasifikacija

Rak dojke je pojam koji obuhvaća niz bolesti različitih histoloških obilježja, stadija, receptorskih statusa i ekspresije gena (Hortobagyi *et al.*, 2017). Prema podacima iz 2022., on je globalno vodeći tip raka u žena i po broju slučajeva i po broju preminulih (Bray *et al.*, 2024). Najvažnije skupine rizičnih čimbenika su reproduktivni i hormonalni čimbenici, životni stil i genetska predispozicija. Rijetko se pojavljuje i u muškaraca, koji čine oko 1% slučajeva (Hong i Xu, 2022). U novije vrijeme, došlo je do brojnih novih spoznaja o raku dojke zahvaljujući razvoju bioinformatike i genomike raka (Stover i Wagle, 2015).

U suvremenom liječenju raka dojke, mikroskopske karakteristike tkiva kao što su stupanj diferenciranosti i ekspresija receptora smatraju se jednako važnima kao makroskopske, anatomske klasifikacije. Na temelju tih kriterija može se definirati prognoza i odabrati optimalna terapija (Hortobagyi *et al.,* 2017). Primjerice, tumori pozitivni na estrogenski receptor (ER) obično se bolje liječe hormonskom terapijom, dok tumori pozitivni na HER2 (engl. *human epidermal growth factor receptor 2*) bolje odgovaraju na anti-HER2 terapiju (Cui *et al.*, 2005). Zanimljivo je da je ER bio jedan od prvih biomarkera iskorištenih za ciljanu terapiju raka, tamoksifen (Stover i Wagle, 2015). Od hormonskih receptora važan je i progesteronski (PR). Prema prisutnosti navedenih biomarkera definirana su tri podtipa raka dojke: luminalni (ER+ i/ili PR+, HER2-), HER2 receptorom obogaćeni (HER2+) i trostruko negativni podtip (ER-, PR-, HER2-). Ovdje će se oni nazivati imunohistokemijskim (IHC) podtipovima, prema metodi njihove detekcije. Također, tumori obogaćeni receptorom HER2 mogu se podijeliti s obzirom na prisutnost (ili nedostatak) hormonskih receptora, što je povezano s različitim odgovorom na terapiju. Osim prisutnosti HER2, u sklopu patološke procjene provjerava se i postoji li amplifikacija njegovog gena (Hortobagyi *et al.*, 2017).

Slično IHC podjeli, postoji podjela na temelju profila ekspresije gena (GEP, engl. *gene expression profile*), koja će biti korištena u ostatku rada. Prema njoj, AJCC (engl. *American Joint Committee on Cancer*) dijeli tumore dojke na luminalni A, luminalni B, HER2 i bazalni podtip (Hortobagyi *et al.*, 2017). U literaturi se ponekad spominje i peti, tzv. normalni podtip, koji nalikuje normalnim stanicama dojke (Sørlie *et al.*, 2003; Rouzier *et al.*, 2005; Smid *et al.*,

2008). Kao i IHC podtipovi, GEP podtipovi razlikuju se po prognozi i odgovoru na pojedine terapije. Prognoza za luminalni A tip je dobra, za luminalni B osrednja ili loša, a za HER2 i bazalni loša. Ne postoji striktna povezanost IHC i GEP klasifikacije, ali postoje neka poklapanja (Tablica 1). Primjerice, bazalni podtip obično će biti trostruko negativan, ali manji dio tumora uz markere bazalnog podtipa eksprimira i hormonske receptore (Dai *et al.*, 2016). Primjenu GEP klasifikacije koči nedovoljna dostupnost molekularno-dijagnostičkih metoda u zdravstvu (Hortobagyi *et al.*, 2017), ali je zbog svoje informativnosti pogodna u istraživanjima molekularne biologije raka.

Tablica 1. GEP i IHC podtipovi te njihovi tipični statusi biomarkera, prema Dai *et al.* (2016) i Hortobagyi *et al.* (2017). Znak "|" predstavlja uključivo "ili".

GEP podtip	**Status biomarkera**	**IHC podtip**
Luminalni A	[ER+|PR+]HER2-	Luminalni
Luminalni B	[ER+|PR+]HER2-	
	[ER+|PR+]HER2+	HER2-obogaćeni
HER2	ER-PR-HER2+	
Bazalni	ER-PR-HER2-	Trostruko negativni

1.1.2 Upalni rak dojke

Upalni rak dojke (IBC, engl. *inflammatory breast cancer*) je rijedak i agresivan tip raka dojke. Većina spada pod duktalne karcinome, tj. nastaje iz epitelnih stanica mliječnih kanalića. Unutar TNM sustava klasifikacije (engl. *tumor, node, metastasis*) pridružena mu je kategorija T4d. Ima brzu progresiju i velik metastatski potencijal. Simptomi su bol, osjetljivost na dodir, povećanje dojke te crvenilo, toplina i zadebljanje kože nalik narančinoj kori (fr. *peau d'orange*). Smatra se da bi uzrok promjenama u koži mogle biti tumorske embolije u dermalnim limfnim žilama, koje mogu, ali ne moraju biti očite u biopsiji kože. IBC se ne smatra zasebnim histološkim podtipom karcinoma dojke i još uvijek ne postoje čvrsti molekularni niti patološki dijagnostički kriteriji za njega. Uz to, simptomi na koži znaju biti dosta slični bakterijskim infekcijama poput mastitisa, što može dovesti do krive dijagnoze i odgode adekvatnog liječenja. Za dijagnozu je potrebno utvrditi postojanje invazivnog karcinoma u parenhimu ili dermalnim limfnim žilama te odrediti status biomarkera (hormonski receptori, HER2 i stupanj diferenciranosti). U istraživanjima učestalosti IHC podtipova IBC-a zabilježeno je 17-30% slučajeva trostruko

negativnog podtipa, dok opaženi udjeli luminalnog i HER2 podtipa značajno variraju (Robertson *et al.*, 2010; Van Uden *et al.*, 2015; Hortobagyi *et al.*, 2017; Mamouch *et al.*, 2018). Općenito, istraživanje IBC-a kasni u odnosu na neupalni rak dojke (nIBC, engl. *non-inflammatory breast cancer*). Dijagnostiku bi moglo olakšati definiranje profila ekspresije gena koji je specifičan za IBC.

1.2 Visokoprotočni eksperimenti i računalne metode u istraživanju raka

1.2.1 Genomika tumora i visokoprotočne metode

Ljudski genom sadrži oko 3×10^9 parova baza. Točna sekvenca (slijed) tih baza dobiva se primjenom metoda sekvenciranja. Osim što postoji referentni ljudski genom koji je prvotno složen (Lander *et al.*, 2001), a zatim ažuriran (Nurk *et al.*, 2022) u sklopu velikih genomskih projekata, danas je moguće primijeniti metode sekvenciranja na pojedince u dijagnostičke svrhe, što ima sve važniju ulogu u liječenju zloćudnih tumora (Berger i Mardis, 2018). Ne samo da postoje brojne kategorije po kojima se rak može klasificirati, kao što su TNM, IHC podtipovi i GEP podtipovi, nego i unutar tih kategorija postoje razlike između pojedinaca. Te razlike mogu se detektirati sekvenciranjem. Primjerice, tumori dojke razlikuju se po tzv. pogonskim (engl. *driver*) mutacijama koje su stanicama omogućile uspješnu proliferaciju. One se mogu nalaziti unutar različitih gena, pa je tako definirano čak 99 pogonskih gena za adenokarcinom dojke (Martínez-Jiménez *et al.*, 2020). Također, tumori se razlikuju po tome kojim mehanizmom su nastali. Svaka etiologija povezana je s određenim mutacijskim obrascima, npr. supstitucijama jedne baze. Tako je genom tumora nastalih zbog izlaganja ultraljubičastom zračenju bogat supstitucijama citozina timinom (C>T) (Alexandrov *et al.*, 2020). Za ovakav napredak u istraživanju raka velikim su dijelom zaslužne suvremene, visokoprotočne metode sekvenciranja (engl. *high-throughput sequencing*) (Berger i Mardis, 2018). Među njihovim ključnim svojstvima je paralelizacija, mogućnost istovremenog sekvenciranja velikog broja molekula. Paralelizacija se ne primjenjuje samo u sekvenciranju, već i u drugim visokoprotočnim metodama, kao što su DNA mikročipovi (Lesk, 2012).

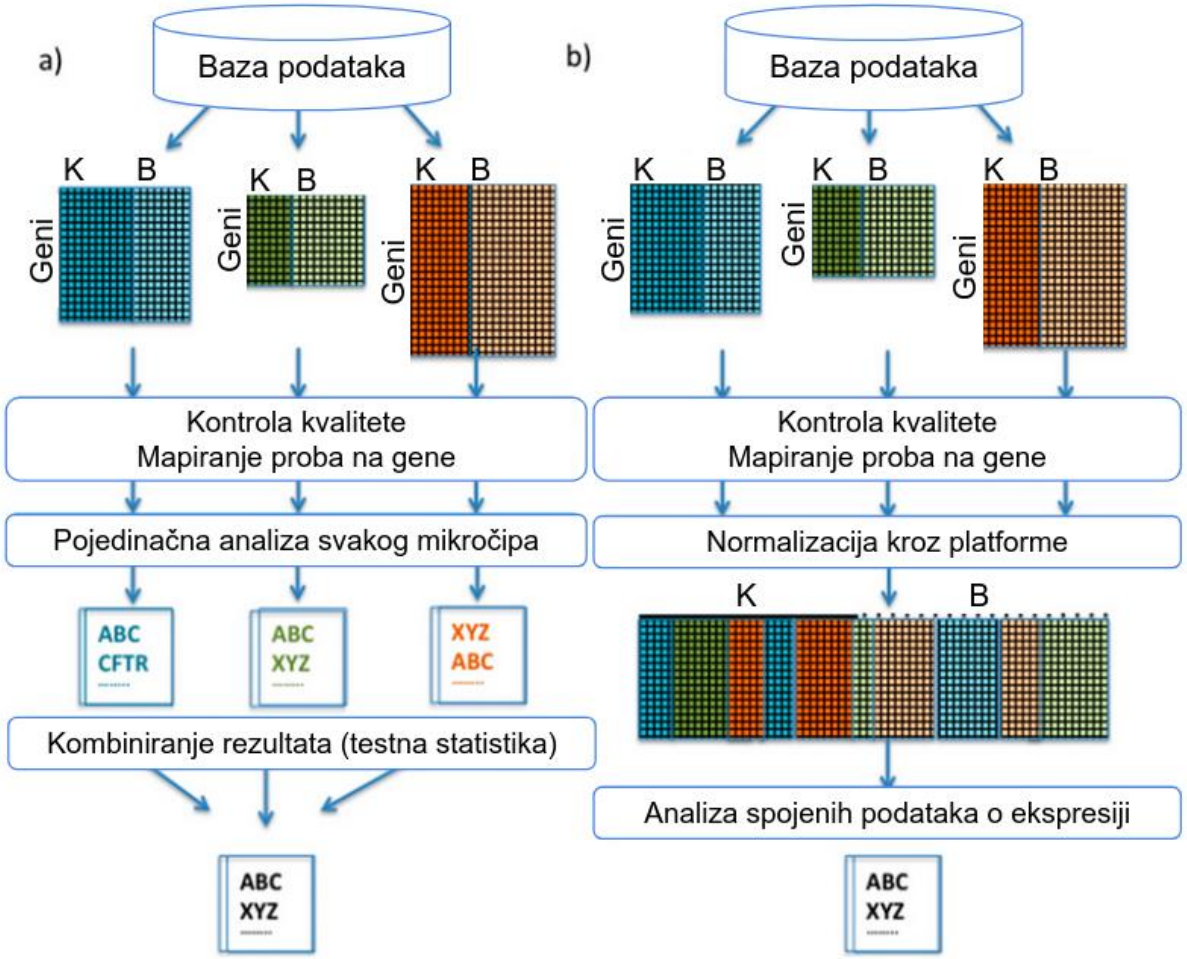
1.2.2 Mjerenje ekspresije gena DNA mikročipovima

DNA mikročip (engl. *microarray*) je pločica koja omogućava paralelnu detekciju velikog broja oligomera DNA. Površina im obično iznosi 2 cm^2 i sadrži od nekoliko tisuća do nekoliko stotina tisuća proba. Jednu probu čini mnogo kopija određenog oligonukleotida (duljine 20-80 bp) imobiliziranih na podlozi. Za analizu stanične mRNA koriste se ekspresijski čipovi. U tipičnom

eksperimentu mRNA se prvo izolira, a zatim koristi kao kalup za sintezu stabilnije cDNA (engl. *copy DNA*). Pritom se cDNA označava fluoroforima. Nakon toga se cDNA molekule nanose na mikročip u uvjetima koji potiču hibridizaciju s komplementarnim oligonukleotidnim probama. Mikročip se zatim ispire, čime se sa čipa uklanjaju cDNA koje nisu hibridizirale niti s jednom probom. Na probama gdje je došlo do hibridizacije moguće je detektirati i kvantificirati fluorescentni signal koji potječe od cDNA. Budući da je poznata sekvenca oligonukleotida koji čine svaku probu i zna se kojem genu odgovaraju, ovako se kvantificira ekspresija pojedinih gena na razini transkripcije. Ekspresijski mikročipovi mogu se primijeniti u dijagnostici, s obzirom na to da omogućavaju detekciju kompleksnih obrazaca u ekspresiji gena karakterističnih za određene bolesti, primjerice različite tipove leukemije (Lesk, 2012). Uz pomoć ekspresijskih mikročipova razvijena je i metoda klasifikacije raka dojke u GEP podtipove nazvana PAM50. Ime je dobila po PAM-u, prediktivnoj analizi mikročipova (engl. *predictive analysis of microarray*) i 50 gena na kojima se temelji (Parker *et al.*, 2009).

Podaci prikupljeni eksperimentima na DNA mikročipovima mogu se pohranjivati u javne baze podataka, gdje su dostupni istraživačima koji ih žele iskoristiti za nove analize. Ovo je posebno korisno kod proučavanja rijetkih bolesti, budući da je kod njih otežano prikupljanje uzoraka. Kada pojedinačni skupovi podataka nemaju dovoljan broj uzoraka, podaci iz različitih skupova mogu se analizirati zajedno. U slučaju mikročipova postoje dvije opcije: metaanaliza i normalizacija po platformama (engl. *cross-platform normalization, cross-normalization*) (Slika 1). Metaanaliza je kasna integracija podataka, što znači da se mikročipovi pretprocesiraju i analiziraju odvojeno, a zatim se rezultati tih analiza kombiniraju. Suprotno tome, normalizacija po platformama je rana integracija podataka jer se mikročipovi pretprocesiraju odvojeno, nakon čega se spajaju i zajedno analiziraju. Ovaj problem nije trivijalan, budući da postoje brojne razlike između mikročipova koje treba uzeti u obzir. Postoje skupovi podataka koje je lakše integrirati, primjerice podaci dobiveni na istim platformama mikročipova. Za njih se kaže da imaju nisku kompleksnost. S druge strane su skupovi podataka visoke kompleksnosti, npr. oni koji su dobiveni različitim platformama. Jedna velika razlika je u genima koje svaka platforma uključuje. Ovo je važno jer u analizu mogu biti uključeni samo geni čija se ekspresija mjerila na svakom mikročipu. Također se mogu razlikovati po osjetljivosti senzora za fluorescencijski signal, a istraživači koriste i različite protokole obrade uzoraka, npr. označavaju cDNA različitim fluoroforima. Zato treba obaviti opsežno pretprocesiranje gdje se podaci normaliziraju s obzirom na navedene razlike, pri čemu treba biti

oprezan da se zadrže realne, biološke razlike između uzoraka (Hardiman, 2004; Walsh *et al.*, 2015).



Slika 1. Metode integracije podataka dobivenih na mikročipovima: a) metaanaliza (kasna integracija), b) normalizacija po platformama (rana integracija). Slova K i B predstavljaju dvije skupine koje se istražuju, u ovom primjeru kontrolnu skupinu (K) i skupinu koja ima određenu bolest (B). Preuzeto i prilagođeno prema Walsh *et al.* (2015).

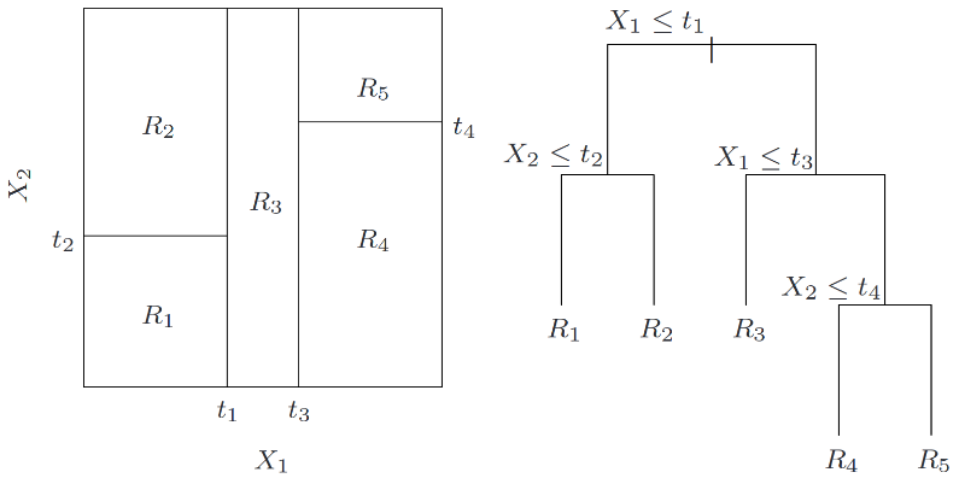
1.2.3 Primjena računalnih metoda u istraživanju tumora

Visokoprotočne metode omogućile su dobivanje velikih skupova podataka u kratkom vremenu. To je pratio razvoj računalnih metoda za njihovu obradu i interpretaciju (Lesk, 2012). Kod tumora se može istraživati niz bioloških svojstava, a ona zahtijevaju primjenu različitih algoritama. Kombinacija visokoprotočnih i računalnih metoda uvelike je doprinijela razumijevanju biološke podloge različitih tipova raka, razvoju ciljanih terapija te otkrivanju biomarkera i ekspresijskih profila (Berger i Mardis, 2018). Kad se nastoji pronaći ekspresijski profil neke bolesti, čest pristup je stvaranje klasifikatora koji na temelju izmjerene ekspresije gena svrstava uzorke u dijagnostičke kategorije. Pritom se pojavljuju određeni izazovi. Postoji velika količina ulaznih podataka, a relativno mali broj uzoraka, te je potrebno identificirati gene

koji najviše doprinose točnoj klasifikaciji. Za rješavanje ovakvih problema primjenjuje se strojno učenje. (Tibshirani *et al.*, 2002).

1.2.4 Strojno učenje

Strojno učenje obuhvaća niz alata za razumijevanje podataka koji se dijele na nadzirano i nenadzirano učenje. U nadziranom učenju postoje ulazne, nezavisne varijable (X) te izlazna, zavisna varijabla (Y). Nezavisne varijable još se nazivaju prediktori. Nadzirano učenje dobilo je svoj naziv po tome što opservacije s poznatim vrijednostima zavisne varijable služe kao vodilja u stvaranju modela. Od metoda nadziranog učenja često se koriste metode temeljene na stablima. Lijevo na Slici 2 prikazan je prostor kojeg čine prediktorske varijable X_1 i X_2. On je više puta prošao kroz binarne podjele. Prvo se podijelio na vrijednosti $X_1=t_1$. Tada je regija za koju vrijedi $X_1 \leq t_1$ podijeljena na $X_2=t_2$, što je stvorilo dvije nove regije, R_1 i R_2. Nakon još binarnih podjela nastalo je ukupno pet regija. Te se podjele mogu prikazati u obliku stabla (Slika 2, desno). Na vrhu stabla počinje se sa skupom svih opservacija, a u svakom se čvoru dijele na dva dijela prema određenom kriteriju. Regije se nalaze na terminalnim čvorovima koji se zovu i listovi. Na temelju regije u koju je opservacija svrstana model predviđa vrijednost varijable Y (Hastie, Tibshirani i Friedman, 2008).



Slika 2. Grafički prikaz principa na kojem se temelje metode bazirane na stablima. Lijevo je prikazan prostor koji čine prediktori X_1 i X_2. Taj prostor je binarno dijeljen, što je rezultiralo stvaranjem regija R_1, $R_2 \ldots R_5$. Desna slika prikazuje isto dijeljenje opservacija, ali u obliku dendrograma. Preuzeto i prilagođeno prema Hastie, Tibshirani i Friedman (2008).

Iako imaju ulazne, nezavisne varijable, modeli nenadziranog učenja ne predviđaju zavisnu varijablu, već im je cilj ukazati na određene pravilnosti u skupu podataka. Primjeri nenadziranog učenja su klasteriranje i analiza glavnih komponenti (PCA, engl. *principal*

component analysis). U klasteriranju je cilj razdijeliti opservacije u relativno distinktne grupe ili klastere. Za eksperimente na DNA mikročipovima često se primjenjuje hijerarhijsko klasteriranje gena i uzoraka koje se može prikazati dendrogramom sličnim onom na Slici 2. Glavne komponente skupa podataka daju niz njihovih najboljih linearnih aproksimacija. PCA omogućava smanjenje dimenzionalnosti, stoga se često koristi za interpretaciju višedimenzionalnih skupova podataka (Hastie, Tibshirani i Friedman, 2008).

1.2.5 Gradijentno pojačavanje i *XGBoost*

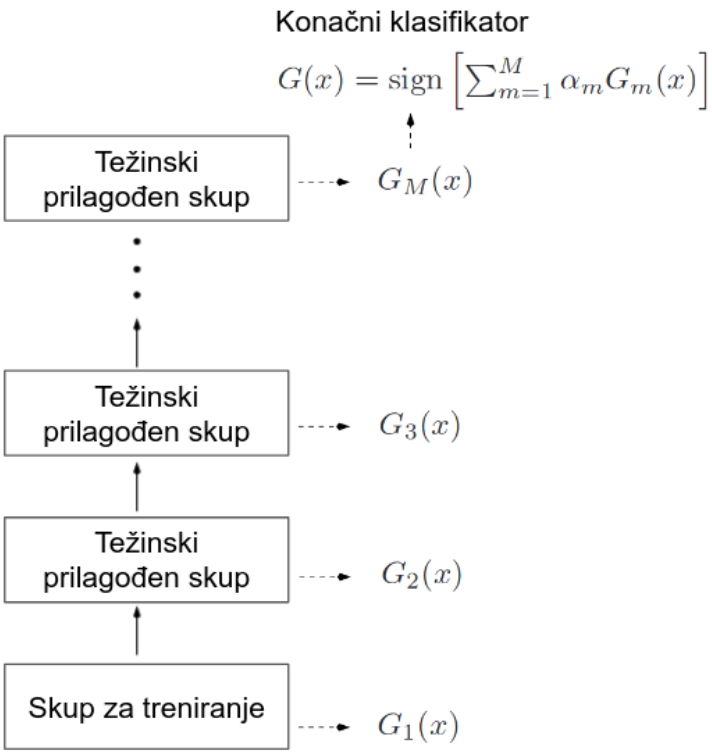
XGBoost (engl. *extreme gradient boosting*) je učinkovit i široko korišten model strojnog učenja koji implementira gradijentno pojačavanje (Chen i Guestrin, 2016).

Pojačavanje (engl. *boosting*) je metoda temeljena na stablima koja je prvotno osmišljena za klasifikaciju, ali je kasnije primijenjena i na regresiju. Prvi algoritam temeljen na pojačavanju nazvan je *AdaBoost* (engl. *adaptive boosting*) (Slika 3). Njegova osnovna ideja bila je kombiniranje slabih klasifikatora da bi se stvorio jak klasifikator. Klasifikator je slab ako nije bitno uspješniji od nasumičnog pogađanja. Pojačavanje uključuje niz koraka, a u svakom se koristi novi klasifikator i uvode se promjene u skup podataka za treniranje. Svaka opservacija u tom skupu podataka dobije novi težinski faktor (engl. *weight*). Kad bude točno klasificirana, težinski faktor joj se smanji, a u suprotnom poveća. Na taj način opservacije koje se teško klasificiraju postaju sve važnije i na njih se moraju usredotočiti klasifikatori koje slijede. Konačni klasifikator sastavljen je od pojedinačnih klasifikatora, a oni također imaju težinski faktor ovisan o uspješnosti koji se mijenjao u svakom koraku (Hastie, Tibshirani i Friedman, 2008; Schapire, 2013).

Na pojačavanje se može gledati kao na problem numeričke optimizacije gdje je cilj minimizirati funkciju pogreške (engl. *loss function*) dodavanjem slabih klasifikatora. Ovo se radi funkcionalnim gradijentnim spustom, procedurom koja podsjeća na gradijentni spust. U gradijentnom spustu se iterativno prilagođavaju neki parametri. Svaki korak je u smjeru koji lokalno rezultira najbržim smanjenjem funkcije greške, tj. u smjeru negativnog gradijenta. U metodi funkcionalnog gradijentnog spusta, na mjesto parametara dolaze klasifikatori: nakon što se izračuna pogreška, u model se treba dodati stablo koje ju smanjuje, tj. slijedi gradijent (Schapire i Freund, 2014; Brownlee, 2016).

Kombinacije pojačavanja i gradijentnog spusta mogu dovesti do prenaučenosti (engl. *overfitting*), gdje model točno predviđa klase iz podataka na kojima je treniran, ali ne i novih podataka s kojima se susreće prvi put. Ovo se može prevenirati na različite načine i zato

XGBoost ima velik broj parametara koji se moraju namjestiti, od kojih će biti navedeni samo neki. Brzina učenja (η) skalira utjecaj pojedinačnog stabla na konačan model. Ako se brzina učenja smanji, to znači da pojedinačna stabla nemaju veliki utjecaj na model, što ostavlja mjesta za veći broj budućih stabala koja bi ga trebala poboljšati. Zato smanjenje brzine učenja treba pratiti povećanje broja stabla. Općenito, dodavanje novih stabala treba prestati kad ono prestane poboljšavati model. Za stabla je poželjno i da budu kratka, tj. manje dubine, najčešće od četiri do osam razina. Još jedna opcija je primjena stohastičkog gradijentnog pojačavanja, gdje se u svakoj iteraciji koristi samo podskup podataka za treniranje. Primjerice, u jednoj iteraciji može se koristiti samo 50% nezavisnih varijabli iz početnog skupa podataka (Brownlee, 2016).



Slika 3. Shematski prikaz algoritma *AdaBoost*. *G* predstavlja klasifikator, *x* nezavisnu varijablu, a α težinski faktor (engl. *weight*) za klasifikatore. Preuzeto i prilagođeno prema Hastie, Tibshirani i Friedman (2008).

1.2.6 Genski potpis upalnog raka dojke

Postoji nekoliko pokušaja definiranja genskog potpisa karakterističnog za IBC pomoću podataka dobivenih DNA mikročipovima (Van Laere *et al*. 2005; Van Laere *et al.*, 2013; Woodward *et al.*, 2013; Bertucci *et al.*, 2014; Lerebours *et al.*, 2021; Zare, Postovit i Githaka, 2021). Rezultirali su skupovima gena koji se uglavnom ne preklapaju u značajnoj mjeri. U klasifikaciji uzoraka u kategorije IBC i nIBC zasad je najuspješniji model temeljen na genskom potpisu G59. Dobiven je jednom od metoda temeljenih na stablima koja se zove *bagging* (engl.

bootstrap aggregating) i ostvario je točnost od 97% na nezavisnom setu podataka. Pritom treba spomenuti da je funkcionirao samo kad su uzorci uzeti prije tretmana kemoterapijom (Zare, Postovit i Githaka, 2021). Iako se potpis G59 zasad pokazao iznimno uspješnim u klasifikaciji prije početka liječenja, trebao bi biti validiran na više uzoraka. Nedovoljan broj uzoraka čest je problem u istraživanjima IBC-a.

2. Opći i specifični ciljevi rada

IBC je tip raka dojke kojeg karakterizira vrlo brza progresija, zbog čega je ključno da se ubrza proces dijagnoze. Ovome bi uvelike pomoglo postavljanje dijagnostičkih kriterija temeljenih na ekspresiji gena. Međutim, IBC je slabo istražen, a radovi o ekspresiji gena daju različite ili čak kontradiktorne rezultate. Među mogućim razlozima su manjak sistematičnosti u podacima i nedovoljan broj pacijentica u istraživanjima. Danas postoje javne baze podataka gdje se pohranjuju podaci dobiveni visokoprotočnim metodama mjerenja ekspresije gena, kao što su DNA mikročipovi. Ti se podaci mogu iskoristiti za novu analizu u kojoj se drugačijim pristupom potencijalno može otkriti nešto što je u izvornom istraživanju ostalo neopaženo. U ovom radu cilj je riješiti problem malih uzoraka tako da se istovremeno analiziraju podaci iz tri različita istraživanja. Primjenom statističkih metoda podaci bi se trebali dovesti u oblik koji je ekvivalentan jednom istraživanju s većim brojem uzoraka. Ovo omogućava da se ti podaci zajedno koriste u metodama strojnog učenja koje se nastoje primijeniti za otkrivanje skupine gena odgovorne za razvoj IBC-a.

Specifični ciljevi rada su:

1. sistematizirati uzorke prema histološkim markerima (ER i HER2), vrsti raka dojke (IBC ili nIBC) i GEP podtipu (luminalni A, luminalni B, HER2 ili bazalni)

2. spojiti podatke iz tri različita eksperimenta na ekspresijskim mikročipovima primjenom metode normalizacije po platformama

3. provesti hijerarhijsko klasteriranje uzoraka prema ekspresiji gena

4. napraviti *XGBoost* modele strojnog učenja koji na temelju ekspresije gena razlikuju IBC i nIBC uzorke, pri čemu će jedan biti neovisan o GEP podtipu, a ostali specifični za određeni GEP podtip

3. Materijali i metode

3.1 Javno dostupni podaci

Za analizu su korišteni podaci dobiveni ekspresijskim mikročipovima iz baze podataka GEO (engl. *Gene Expression Omnibus*): mikročip 1 (GSE45581, Woodward *et al.*, 2013), mikročip 2 (GSE22597, Iwamoto *et al.*, 2011) i mikročip 3 (GSE17907, Sircoulomb *et al.*, 2010). Svi analizirani uzorci prikupljeni su prije bilo kakvog tretmana. Mikročip 1 (20 IBC, 20 nIBC) sadrži uzorke prikupljene biopsijom širokom iglom (IBC) ili kirurškim uklanjanjem (nIBC). Provedena je mikrodisekcija pri kojoj je netumorsko tkivo uklonjeno tako da udio tumorskih stanica bude >90%, a ekspresija gena mjerena je platformom *Whole Human Genome 4×44K* (*Agilent Technologies*). U slučaju mikročipa 2 (25 IBC, 57 nIBC), uzorci su prikupljeni aspiracijom tankom iglom, dok je ekspresija gena mjerena pomoću platforme *U133A* (*Affymetrix*). Uzorci u mikročipu 3 (21 IBC, 30 nIBC) imaju amplifikaciju lokusa *ERBB2* i prikupljeni su pri kirurškom uklanjanju tumora (nIBC) ili u sklopu dijagnostičke kirurške biopsije (IBC). Netumorsko tkivo uklonjeno je makrodisekcijom i svi uzorci imaju sadržaj tumorskih stanica >60%. Za mjerenje ekspresije korištena je platforma *HG-U133 plus 2.0* (*Affymetrix*).

3.2 Filtriranje gena

U analizu su uključeni geni za koje postoji barem jedna proba na svakom čipu i imaju svoju oznaku u sustavu *Entrez Gene ID*. Podaci u mikročipovima 1 i 2 preuzeti iz baze podataka GEO već su bili normalizirani, a mikročip 3 normaliziran je kvantil-normalizacijom pomoću funkcije *normalize.quantiles* iz *R* paketa *preprocessCore* (Bolstad, 2024).

Kod mikročipova velik broj gena nije predstavljen samo jednom probom, već više njih, pri čemu one ne moraju nužno pokazivati istu vrijednost ekspresije. Geni čije se probe previše razlikuju uklonjeni su iz analize jer se smatra da je došlo do greške u mjerenju, a za preostale gene odabrana je reprezentativna proba. Za svaki gen izračunata je standardna devijacija (*SD*). Iz analize su zatim izbačeni geni kojima pripadajuće probe ne zadovoljavaju uvjet:

$$SD \leq Q_3 + 1{,}5 \cdot IQR \quad (1)$$

gdje Q_3 predstavlja treći kvartil, a *IQR* interkvartilni raspon (engl. *interquartile range*). U nejednadžbu (1) uvrštene su vrijednosti koje odgovaraju mikročipu 1 ($SD \leq 5{,}34$), budući da se

tako izgubi manji broj gena nego uvrštavanjem podataka za druga dva mikročipa. Kao reprezentativna proba izabrana je ona s najvećom vrijednosti ekspresije, pri čemu se pretpostavlja da druge probe nisu dovoljno dobro vezale cDNA da se detektira stvarna vrijednost.

3.3 Priprema anotacijskih podataka

Svakom uzorku pridruženi su metapodaci iz GEO baze podataka koji uključuju ER status, HER2 status i bolest (IBC/nIBC). Ukoliko je uzorku nedostajao bilo koji od navedenih podataka, uklonjen je iz analize. Slijedila je standardizacija metapodataka. Kod mikročipa 1 autori su koristili numeričke vrijednosti za opisivanje receptorskih statusa, gdje veći broj predstavlja veći udio stanica u uzorku na kojima je prisutan receptor. Ove su vrijednosti pretvorene u binarne kategorije (+/-) prema Tablici 2, tako da odgovaraju anotacijama drugih mikročipova. Metapodacima su pridruženi GEP podtipovi pomoću funkcije *molecular.subtyping* iz *R* paketa *genefu* (Gendoo *et al.*, 2024).

Tablica 2. Tumačenje rezultata IHC testa za ER i HER2 receptor. Stupac *FISH* označava rezultate fluorescencijske hibridizacije *in situ* koja se provodila u slučaju dvosmislenog rezultata IHC testa. Prilagođeno prema Woodward *et al.* (2013).

Receptor	Vrijednost	FISH	Tumačenje
ER	0	/	ER-
	1-5	/	ER+
HER2	0	/	HER2-
	1+	/	HER2-
	2+	FISH-	HER2-
	2+	FISH+	HER2+
	3+	/	HER2+

3.4 Normalizacija po platformama

Vrijednosti ekspresije gena iz sva tri mikročipa spojene su u jednu matricu. Za uklanjanje nebioloških razlika (engl. *batch effect*) između uzoraka s različitih mikročipova korištena je funkcija *ComBat* iz *R* paketa *sva* (Leek *et al.*, 2024), pri čemu je primijenjen parametrijski

Bayesov model, a varijablama od interesa smatrali su se bolest (IBC ili nIBC), GEP podtip, ER status i HER2 status. Varijable od interesa mogući su uzroci varijabilnosti između podataka koja se želi zadržati (jer je ona predmet istraživanja), za razliku od nebiološke varijabilnosti koja se nastoji ukloniti.

3.5 Klasifikacija uzoraka metodom *XGBoost*

Napravljena su četiri modela strojnog učenja: jedan za sve GEP podtipove zajedno i jedan za svaki od sljedećih podtipova: luminalni A, HER2 i bazalni. Luminalni B podtip je isključen zbog premalog broja IBC uzoraka (njih samo devet). Cijeli postupak u ovom poglavlju ponovljen je za svaki od ta četiri slučaja.

Podaci su razdijeljeni u skup za treniranje i skup za testiranje modela jednakih veličina, uz zadržavanje prvotnog omjera klasa IBC i nIBC. U slučajevima kad omjer klasa nije bio podjednak (svi podtipovi zajedno i luminalni A) primijenjena je funkcija *ROSE* (engl. *random over-sampling examples*) implementirana u istoimenom *R* paketu (Lunardon, Menardi i Torelli, 2014) kojom su se generirali sintetički uzorci IBC-a za izjednačenje omjera klasa u setu za treniranje. Njome je umjetno povećan broj uzoraka IBC-a tako da bude jednak broju uzoraka nIBC-a, čime se sprječava da model bolje predviđa uzorke zastupljenije klase. Ta metoda nije primijenjena na setu za testiranje da sintetički uzorci ne bi dali pogrešnu predodžbu o uspješnosti modela. Ukupna točnost (*A*, engl. *accuracy*), tj. udio točnih predviđanja, nije uvijek prikladna za validaciju modela, budući da može maskirati činjenicu da model jednu klasu predviđa puno bolje nego drugu. U tome je bolja Cohenova kapa (κ), za koju vrijedi $\kappa \in [-1, 1]$ i model je bolji čim je vrijednost bliža 1. Iz vrijednosti u matrici zabune (engl. *confusion matrix*) računa se sljedećim izrazom:

$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)} \quad (2),$$

gdje *TP* predstavlja stvarne pozitive, *FP* lažne pozitive, *TN* stvarne negative, a *FN* lažne negative.

Izračunata je srednja vrijednost ekspresije za svaki gen, odvojeno za IBC uzorke i nIBC uzorke. Zatim je za svaki gen izračunat omjer njegove ekspresije u IBC i nIBC uzorcima. Za izradu modela strojnog učenja odabrano je 10% gena (903) kod kojih taj omjer najviše odstupa od 1. Korišten je *XGBoost* implementiran u istoimenom *R* paketu (Chen *et al.*, 2024), funkcija *xgb.train*. Vrijednost *eta* određivana je metodom pokušaja i pogreške. Na temelju odabrane

vrijednosti *eta*, parametar *nrounds* određen je peterostrukom unakrsnom validacijom (engl. *cross-validation*) preko funkcije *xgb.cv*, a parametri *booster*, *max_depth*, *min_child_weight*, *subsample* i *colsample_bytree* (Tablica 3) metodom nasumičnog pretraživanja (engl. *random search, grid search*) uz pomoć funkcije *tuneParams* iz *R* paketa *mlr* (Bischl *et al.*, 2016). Za klasifikaciju se mogu koristiti *XGBoost* modeli temeljeni na stablima (*gbtree*) i *XGBoost* linearni modeli (*gblinear*). Za razliku od modela *gblinear*, *gbtree* provodi selekciju najvažnijih prediktorskih gena, što je poželjno, ali ponekad ne uspijeva doseći točnost kakvu nudi *gblinear*.

Tablica 3. Parametri korišteni u izradi *XGBoost* modela.

<table>
<tr><th rowspan="2">Parametar</th><th colspan="4">Vrijednost</th></tr>
<tr><th>Miješani podtipovi</th><th>LumA</th><th>HER2</th><th>Bazalni</th></tr>
<tr><td>eta</td><td>0,1</td><td>0,3</td><td>0,1</td><td>0,3</td></tr>
<tr><td>n_rounds</td><td>197</td><td>22</td><td>8</td><td>14</td></tr>
<tr><td>booster</td><td>gbtree</td><td colspan="3">gblinear</td></tr>
<tr><td>max_depth</td><td>3</td><td colspan="3" rowspan="4">Parametar nije primjenjiv u odabranom boosteru</td></tr>
<tr><td>min_child_weight</td><td>1,35</td></tr>
<tr><td>subsample</td><td>0,522</td></tr>
<tr><td>colsample_bytree</td><td>0,894</td></tr>
</table>

3.6 Nenadzirano učenje

Analiza glavnih komponenti provedena je pomoću funkcije *prcomp*, a hijerarhijsko klasteriranje pomoću funkcije *hclust*, pri čemu su obje iz *R* paketa *stats* (R Core Team, 2024). U hijerarhijskom klasteriranju korištena je metoda potpune povezanosti (engl. *complete linkage*). Za određivanje optimalnog broja klastera korištena je funkcija *fviz_nblust* iz *R* paketa *factoextra* (Kassambara i Mundt, 2020). Pritom je računata prosječna širina siluete. Širina siluete *s* za opservaciju *i* iz klastera C_I računa se sljedećom formulom (Rousseeuw, 1987):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3),$$

gdje je $a(i)$ srednja udaljenost između opservacije i i ostalih opservacija u klasteru C_I, dok je $b(i)$ najmanja srednja udaljenost opservacije i od opservacija iz bilo kojeg klastera osim C_I. Ako je veličina klastera $|C_I|=1$, tada vrijedi $s(i)=0$. Pritom je računata euklidska udaljenost. Optimalan je onaj broj klastera za kojeg je prosječna širina siluete maksimalna.

3.7 Analiza prezastupljenosti

Provedena je analiza prezastupljenosti (ORA, engl. *overrepresentation analysis*) za odabrane skupove prediktorskih gena iz *XGBoost* modela za bazalni podtip. Prvi skup sadržavao je 10% gena kod kojih je omjer srednjih vrijednosti ekspresije u IBC i nIBC uzorcima najviše odstupao od 1. Drugi skup sadržavao je 115 gena na temelju kojih se IBC i nIBC uzorci najbolje odvajaju u hijerarhijskom klasteriranju. Proveden je hipergeometrijski test pomoću funkcije *enrichGO* iz *R* paketa *clusterProfiler* (Wu *et al.*, 2021). Pretraživane su ontologije za dijelove stanice (*Cellular Component*) i biološke procese (*Biological Process*). Pozadinski skup gena činili su svi geni na mikročipovima koji su prošli filtriranje opisano u potpoglavlju 3.2 (str. 11). Granična q-vrijednost iznosila je 0,05. q-vrijednost predstavlja udio lažnih pozitiva kad se rezultat u statističkom testu proglasi značajnim. Za razliku od nekorigirane p-vrijednosti, povoljna je za višestruko testiranje hipoteza.

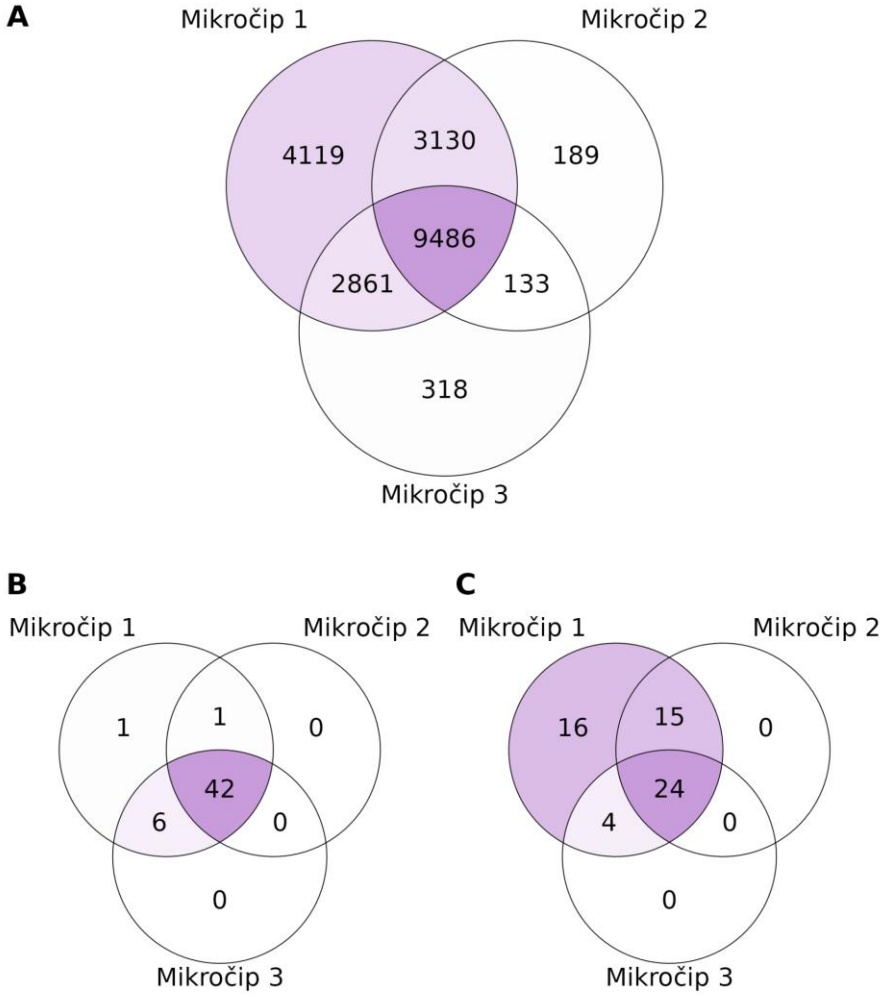
3.8 Statistička obrada i vizualizacija podataka

Cjelokupna statistička obrada podataka provedena je u programskm jeziku *R* (R Core Team, 2024). Za vizualizaciju podataka korišteni su *R* paketi *ComplexHeatmap* (Gu, Eils i Schlesner, 2016), *ggVennDiagram* (Gao i Dusa, 2024), *ggplot2* (Wickham, 2016) i *ggpubr* (Kassambara, 2023).

4. Rezultati

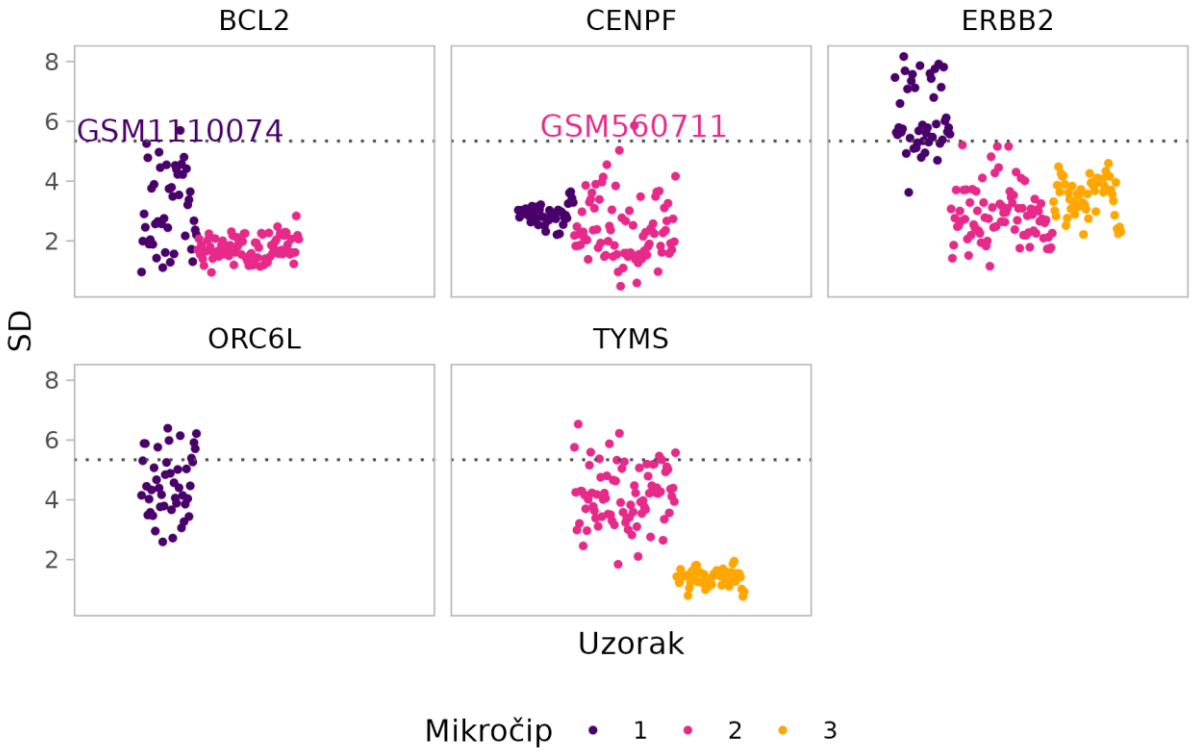
4.1 Sistematizacija i filtriranje podataka

Metoda normalizacije po mikročipovima zahtijevala je da se uklone svi geni čija se ekspresija nije mjerila na sva tri mikročipa. Korištene su tri različite platforme mikročipova koje uključuju različite gene, stoga je prilikom njihovog spajanja došlo do smanjenja broja gena (Slika 4). *Entrez* identifikacijska oznaka uspješno je pridružena uz 20236 gena, a njih 9486 (47%) nalazilo se na svakom mikročipu. Mikročip 1 bio je najbogatiji genima, a i jedini je sadržavao potpune skupove gena PAM50 i G59. Mikročip 2 i 3 sadržavali su 3130, odnosno 2861 gena koji su se preklapali s mikročipom 1, ali ne i međusobno.



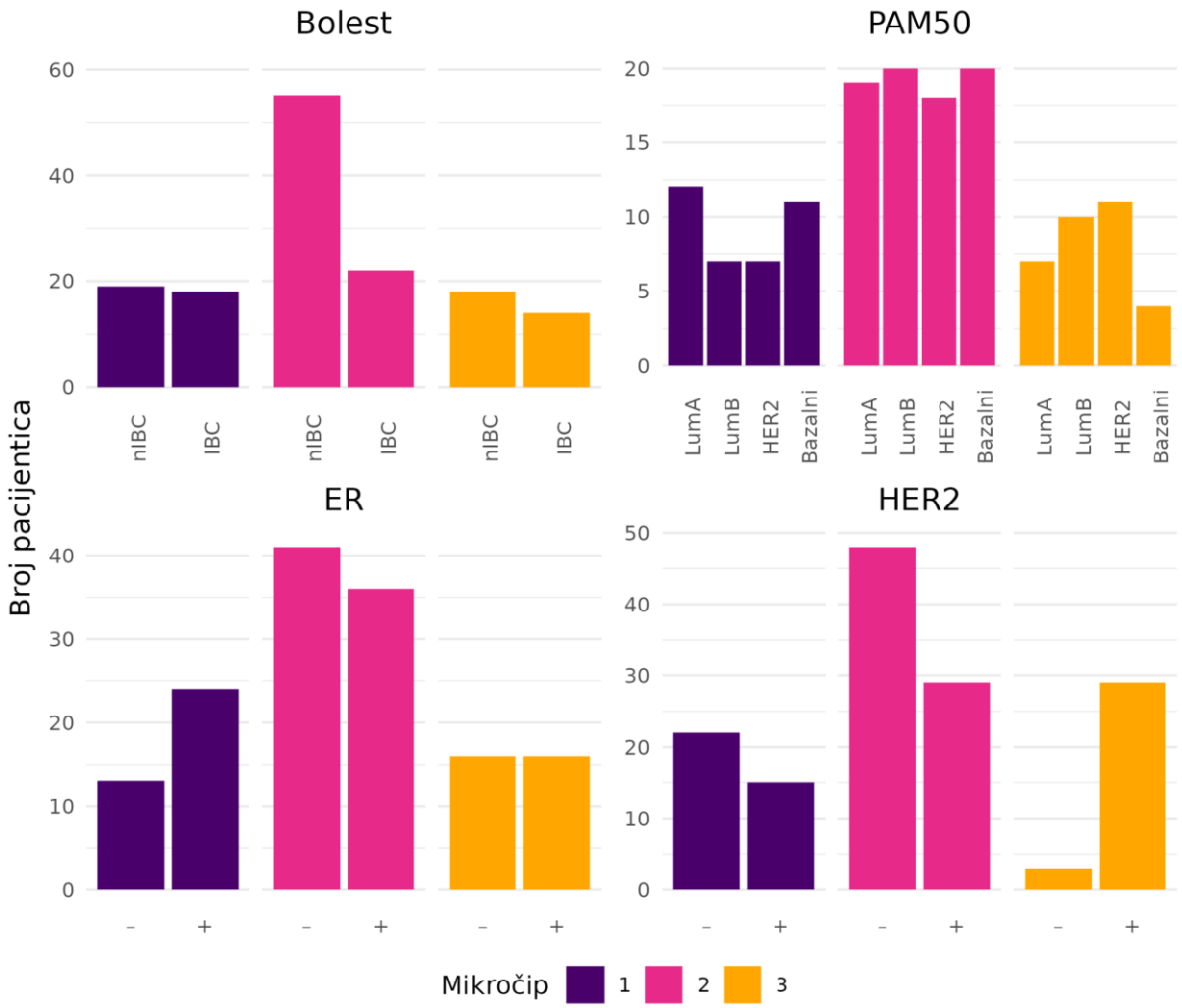
Slika 4. Vennovi dijagrami koji prikazuju broj gena na svakom mikročipu. (A) svi geni (ukupno 20236), (B) geni iz skupa PAM50, (C) geni iz skupa G59. Uključeni su samo geni kojima je uspješno pridružena *Entrez* identifikacijska oznaka.

Sljedeći korak bio je uklanjanje gena čije su probe imale previsoku standardnu devijaciju ekspresije. Pet gena s previsokom standardnom devijacijom spadalo je u skup PAM50: *BCL2*, *CENPF*, *ERBB2*, *ORC6L* i *TYMS*. U detaljnijoj analizi uočeno je da je *BCL2* zadovoljavao uvjet u svim uzorcima osim jednog, a isto je vrijedilo i za *CENPF* (Slika 5). Ova su dva gena zbog svoje važnosti u raku dojke zadržana, a uzorci izbačeni. Konačan broj gena iznosio je 9027 (39 PAM50, 20 G59). Osim navedena dva uzorka, izbačeno je 18 uzoraka s nedostajućim podacima o statusu ER i HER2 te devet uzoraka koji su svrstani pod normalni podtip u PAM50 klasifikaciji. Normalni podtip je izbačen jer broj uzoraka nije bio dovoljno velik za ozbiljniju analizu. Konačan broj uzoraka iznosio je 146 (54 IBC, 92 nIBC).



Slika 5. Standardne devijacije jačine fluorescentnog signala između proba koje pripadaju istom PAM50 genu. Prikazani su PAM50 geni za koje barem jedan uzorak prelazi graničnu SD (SD>5,34) označenu crtkanom linijom. Za svaki od gena *BCL2* i *CENPF* granicu prelazi samo jedan uzorak. Ako na grafičkom prikazu nema niti jednog uzorka iz nekog mikročipa, gen je na njemu predstavljen samo jednom probom. Tekst u boji prikazuje identifikacijsku oznaku uzorka u bazi podataka GEO.

Nakon filtriranja uzoraka analizirana je njihova raspodjela po sljedećim kategorijama: bolest (IBC ili nIBC), GEP podtip te status receptora ER i HER2 (Slika 6). U pojedinim kategorijama uočeni su veći disbalansi. Primjerice, mikročip 2 sadržavao je puno više nIBC uzoraka od IBC uzoraka, a mikročip 3 je većinom sadržavao HER2+ tumore, budući da su autori istraživali tumore s amplifikacijom gena *ERBB2*, koji kodira za receptor HER2.



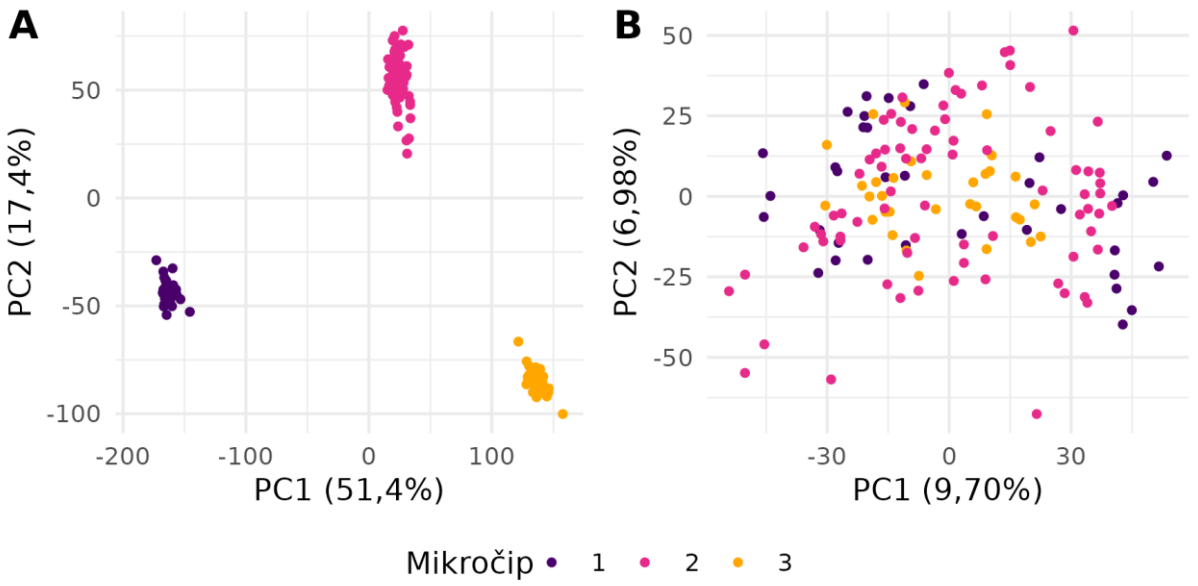
Slika 6. Broj pacijentica po bolesti (IBC ili nIBC), GEP (PAM50) podtipu raka dojke te statusu receptora ER i HER2.

4.2 Normalizacija po platformama

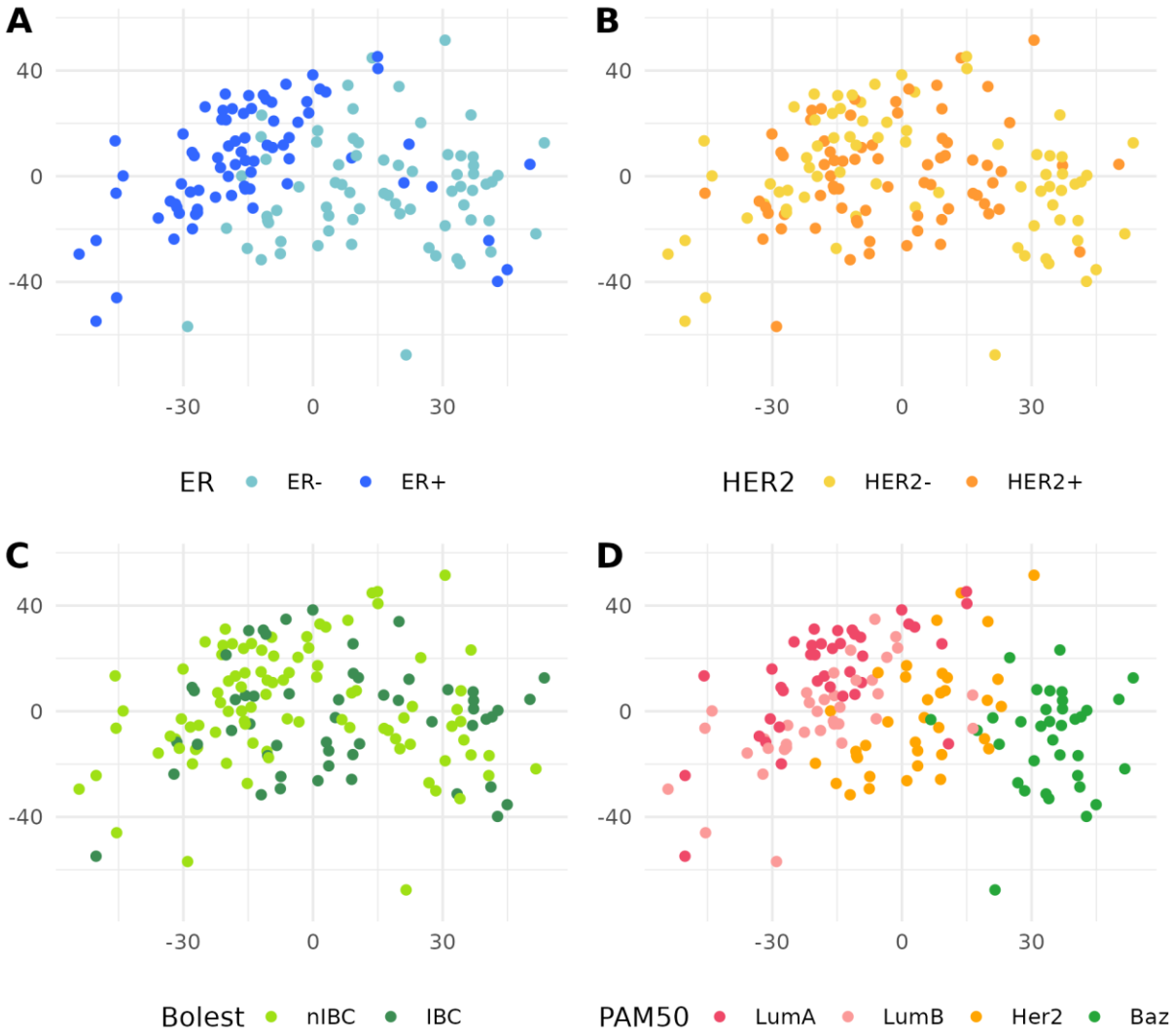
Vrijednosti ekspresije iz sva tri mikročipa bile su spojene u jednu matricu, a zatim normalizirane po platformama metodom *ComBat* s ciljem uklanjanja nebioloških razlika koje potječu od različitih platformi i eksperimentalnih protokola. U rezultatima PCA-a prije primjene metode *ComBat* uočena su tri jasna, nepreklapajuća klastera uzoraka koji potječu iz istih mikročipova (Slika 7A). Ovo više nije bio slučaj nakon normalizacije, gdje su uzorci iz različitih mikročipova bili raspršeni i međusobno izmiješani (Slika 7B).

Osim uklanjanja nebioloških razlika, zadatak metode *ComBat* bio je i zadržavanje bioloških razlika između uzoraka, konkretno onih uzrokovanih različitim bolestima (IBC i nIBC), GEP podtipovima te statusima receptora ER i HER2. Pomoću metoda nenadziranog učenja traženi su biološki uvjetovani obrasci. Prvo je proveden PCA na temelju ekspresije svih 9027 gena. Vizualizirane su prve dvije glavne komponente i uočeno je da su se uzorci relativno

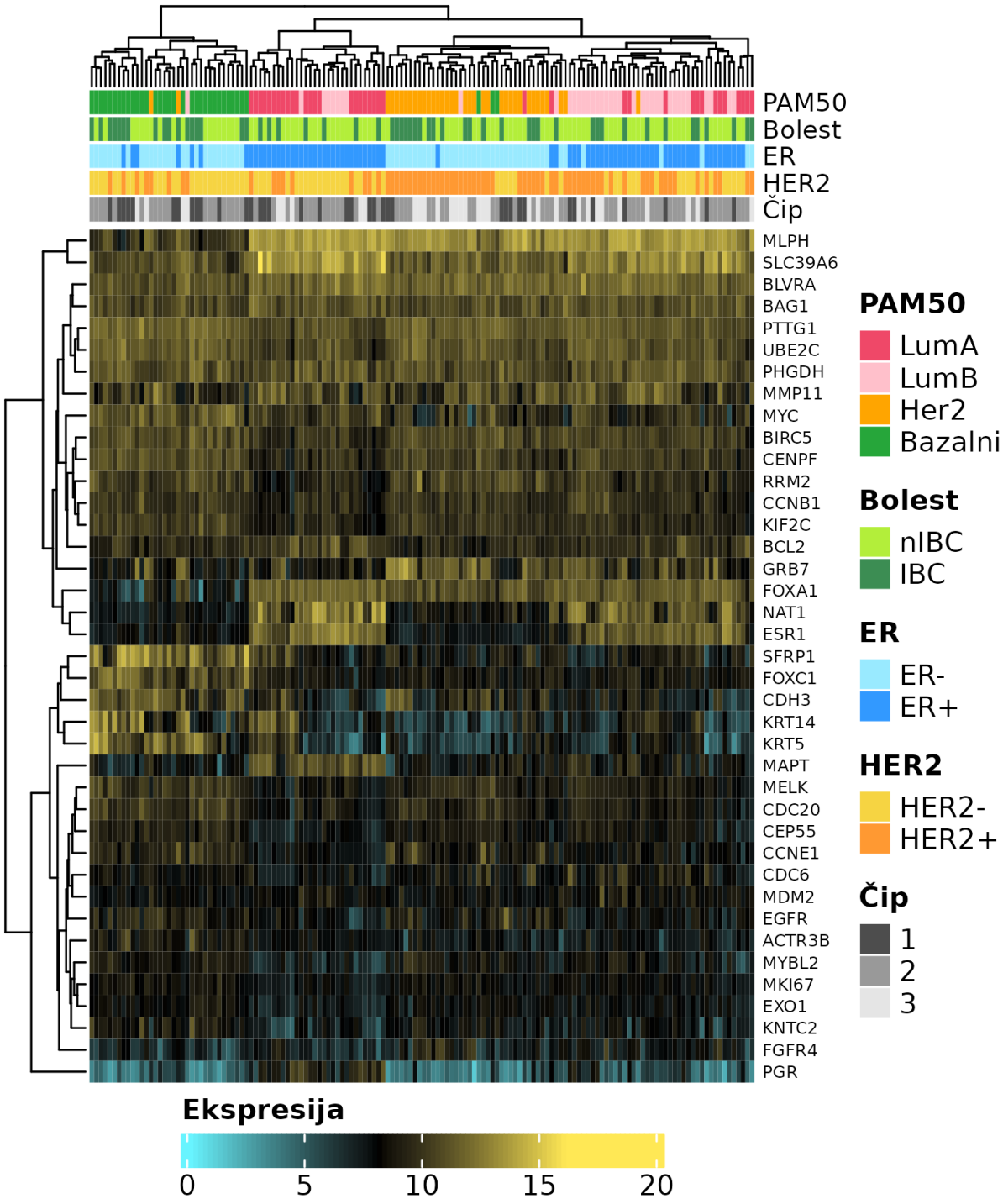
dobro odvajali s obzirom na status ER-a i GEP (PAM50) podtip, dok to nije bio slučaj za bolest niti HER2 status (Slika 8). Od GEP podtipova najviše su se miješali luminalni A i luminalni B, dva podtipa koje karakterizira status ER+. Zatim je vizualizirana ekspresija PAM50 gena (njih 39 koji su prošli filtriranje) kako bi se procijenilo je li normalizacija po platformama poremetila obrasce ekspresije karakteristične za GEP podtipove (Slika 9). Pritom su uzorci i geni hijerarhijski klasterirani. Odabran je ovaj skup gena jer je dobro istražen, što olakšava detekciju nepravilnosti. Podtipovi u anotaciji određeni su prije spajanja mikročipova i normalizacije po platformama. Uočeno je da su GEP podtipovi formirali svoje klastere, iako je bilo i miješanja. Međusobno su se najviše miješali luminalni A i luminalni B, a najčišći klaster stvarao je bazalni podtip. Hijerarhijsko klasteriranje uzoraka također je pokazalo da se uzorci nisu grupirali po mikročipovima iz kojih potječu. Na ovakvom je prikazu također vidljivo da GEP podtipovi u nekolicini slučajeva nisu imali očekivanu IHC klasu, pogotovo kod bazalnog podtipa koji se najčešće veže uz trostruko negativni status receptora, a ovdje je u nekoliko slučajeva bio ER+ i/ili HER2+.



Slika 7. Prve dvije glavne komponente u PCA-u provedenom na uzorcima prije (A) i poslije (B) normalizacije po platformama. Postotci u zagradama označavaju udio varijance koji objašnjava pojedina glavna komponenta.



Slika 8. Rezultati PCA-a provedenog na uzorcima nakon normalizacije po platformama, s označenim biološkim kategorijama. Koordinate točaka iste su kao na Slici 7, B, ali ovdje boje prikazuju ER status (A), HER2 status (B), bolest (C) i GEP podtip po PAM50 klasifikaciji (D). Kratice za GEP podtipove: LumA - luminalni A, LumB - luminalni B, Baz - bazalni podtip.



Slika 9. Toplinska mapa (engl. *heatmap*) koja prikazuje matricu ekspresije 39 PAM50 gena u 146 uzoraka nakon normalizacije po platformama metodom *ComBat*. Ekspresija se mjerila jačinom fluorescencijskog signala. Svaki uzorak ima svoj stupac, a svaki gen svoj redak. Uzorci i stupci su hijerarhijski klasterirani metodom potpune povezanosti (engl. *complete linkage*), što je prikazano dendrogramima. Ispod gornjeg dendrograma nalazi se anotacija pridružena svakom uzorku koja se odnosi na GEP podtip raka dojke (PAM50), bolest (IBC ili nIBC), statuse receptora ER i HER2 te čip kojem uzorak pripada.

4.3 Klasifikacija raka dojke u upalni i neupalni

Klasifikator raka dojke neovisan o GEP podtipu dobiven metodom *XGBoost* postigao je točnosti A_{nIBC}=74% i A_{IBC}=81% u skupu za testiranje, a Cohenova kapa (κ) iznosila je 0,53 (Slika 10). Model za HER2 podtip rezultirao je sličnom vrijednosti κ, ali većom razlikom u uspješnosti za nIBC i IBC (A_{nIBC}=67%, A_{IBC}=88%, κ=0,56). Nešto uspješniji bio je model za luminalni A podtip (A_{nIBC}=85%, A_{IBC}=83%, κ=0,65), a najboljim se pokazao model za bazalni podtip (A_{nIBC}=90%, A_{IBC}=86%, κ=0,76).

Svi podtipovi

Predviđena klasa \ Stvarna klasa	nIBC	IBC
nIBC	34	5
IBC	12	22

Luminalni A podtip

Predviđena klasa \ Stvarna klasa	nIBC	IBC
nIBC	10	1
IBC	3	6

HER2 podtip

Predviđena klasa \ Stvarna klasa	nIBC	IBC
nIBC	6	1
IBC	3	8

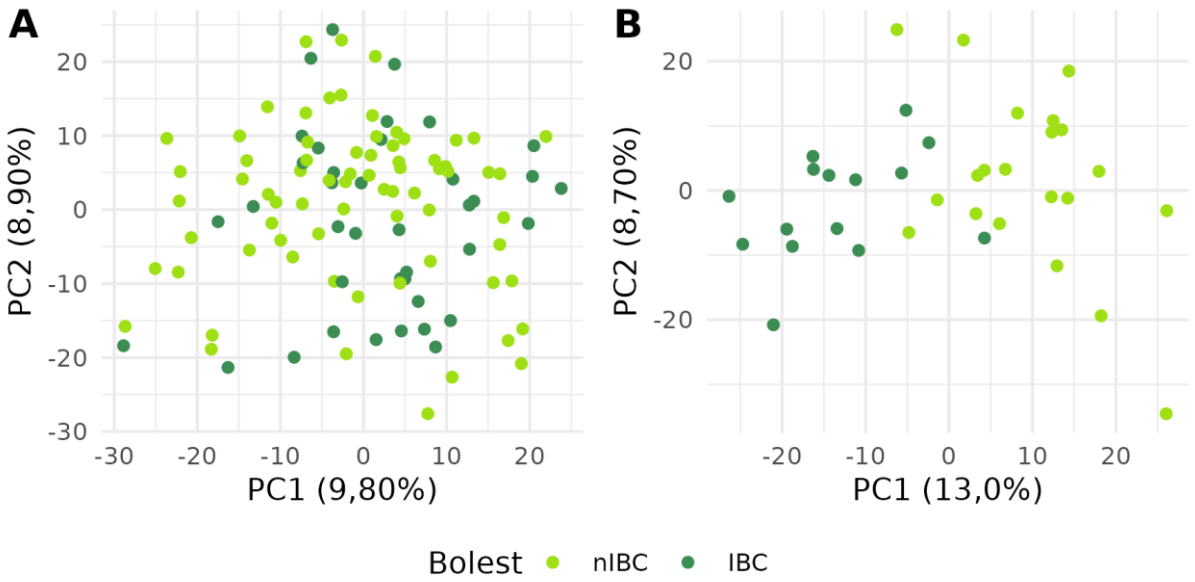
Bazalni podtip

Predviđena klasa \ Stvarna klasa	nIBC	IBC
nIBC	9	1
IBC	1	6

Slika 10. Matrice zabune (engl. *confusion matrices*) za klasifikator neovisan o GEP podtipu i klasifikatore specifične za GEP podtip (luminalni A, HER2 i bazalni). Prikazani su samo rezultati za uzorke u skupu za testiranje modela. Omjer uzoraka u skupovima za treniranje i skupovima za testiranje modela bio je 1:1.

Budući da je bio najuspješniji, za daljnje analize izabran je model specifičan za bazalni podtip. On uključuje 912 gena i njihov skup će se nazvati M912 (prema riječi *model*). Njega čine skup 10% gena kojima omjer srednje ekspresije u IBC i nIBC uzorcima najviše odstupa od 1 (n=903), nazvan D903 (prema engl. *differential*) i preostali G59 geni koji nisu zadovoljili

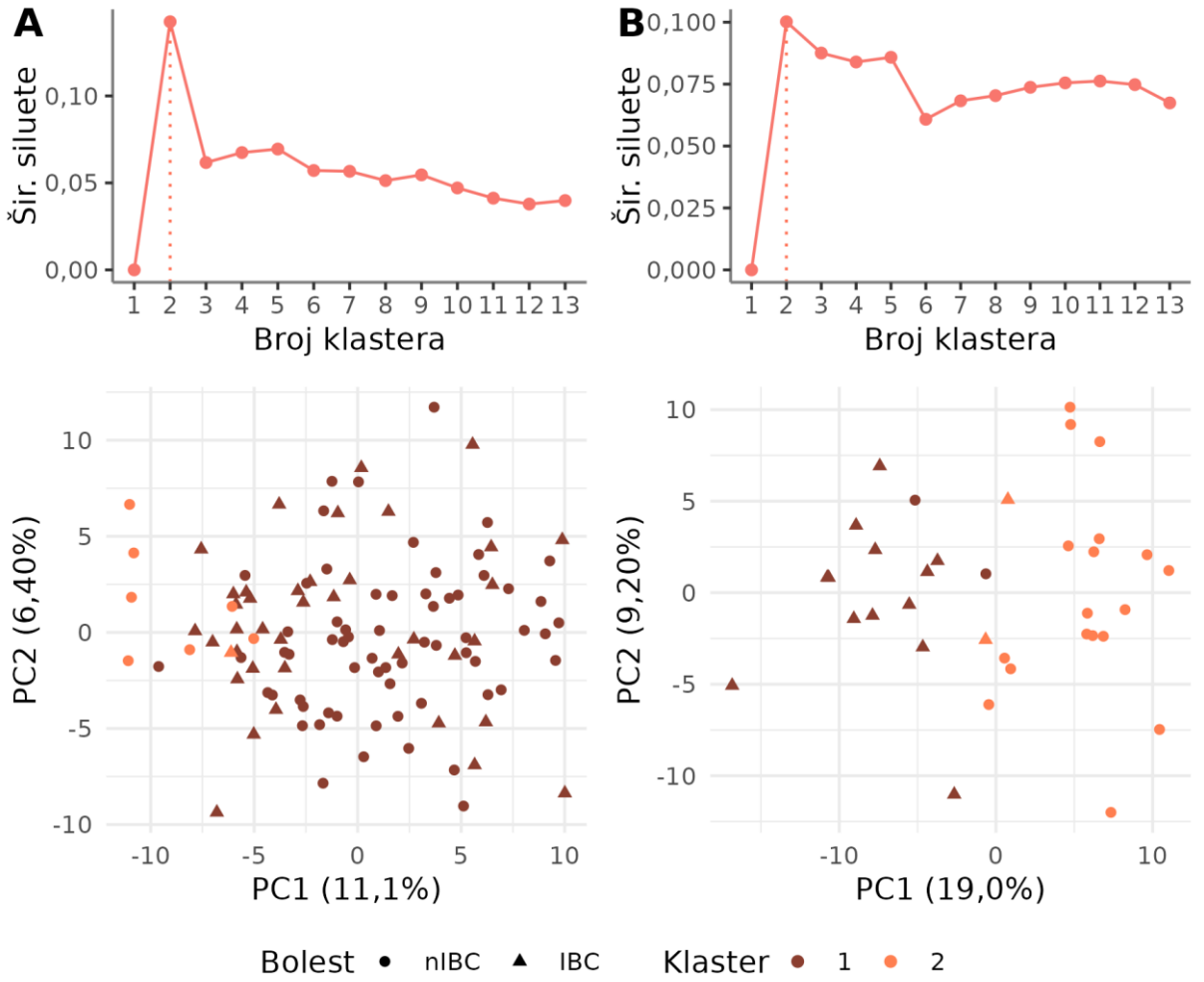
taj uvjet. Svi klasifikatori specifični za GEP podtip temeljili su se na linearnom *XGBoost* modelu jer se on u nasumičnom pretraživanju (potpoglavlje 3.5, str. 14) pokazao boljim od modela temeljenog na stablima. Zato su u tim modelima bili uključeni svi geni koji su bili prisutni u ulaznim podacima. Ovo je suprotno *XGBoost* modelima temeljenima na stablima, kakav je bio klasifikator neovisan o podtipu, gdje se odabire podskup gena koji najbolje klasificira uzorke. Za dodatan uvid u prediktivnu moć skupa M912 proveden je PCA. Prva analiza uključivala je uzorke svih podtipova osim bazalnog (Slika 11A). Ovdje se nije naziralo nikakvo razdvajanje IBC i nIBC uzoraka na temelju prve dvije glavne komponente. Druga analiza (Slika 11B) uključivala je samo uzorke bazalnog podtipa te su se IBC i nIBC prilično dobro odvajali. Zanimljivo je i da se skup 10% gena s najvećom razlikom u ekspresiji između IBC i nIBC uzoraka dosta razlikuje u slučaju bazalnog podtipa i miješanih podtipova. Poklapaju se u tek ~45% gena.



Slika 11. Rezultati PCA-a temeljenog na 912 gena uključenih u *XGBoost* model za klasifikaciju uzoraka bazalnog podtipa u IBC i nIBC (skup M912). Kad se analiza provede na uzorcima svih podtipova osim bazalnog (A), na grafu se ne uočava razdvajanje IBC i nIBC uzoraka. S druge strane, uzorci bazalnog podtipa razdvajaju se s obzirom na bolest (B).

Dva klastera uočena na PCA grafu pokušala su se replicirati drugom metodom nenadziranog učenja, hijerarhijskim klasteriranjem. Skup gena postepeno se sužavao na temelju njihovih koeficijenata u *XGBoost* linearnom modelu, pri čemu su se ponavljali PCA i hijerarhijsko klasteriranje. Za određivanje optimalnog broja klastera računata je prosječna širina siluete. Geni su poredani silazno prema apsolutnoj vrijednosti koeficijenata koji im je dodijeljen u modelu i prvo bi se uklanjali oni s dna liste. Kad bi se uključilo previše gena, hijerarhijsko

klasteriranje uglavnom je rezultiralo jednim vrlo velikim klasterom i jednim vrlo malim, gdje je manji sadržavao jedan ili tek nekoliko uzoraka. Premali broj gena rezultirao je klasterima s miješanim IBC i nIBC uzorcima. Razdvajanje IBC i nIBC uzoraka bilo je najuspješnije kad je uključeno 115 gena s najvećim apsolutnim vrijednostima koeficijenata. Ovaj skup gena nazvan je HC115 (po engl. *hierarchical clustering*). Optimalan broj klastera iznosio je dva i za nebazalne podtipove (Slika 12A) i za bazalni podtip (Slika 12B). U slučaju nebazalnih podtipova definirana su dva klastera. U klasteru 1 bila je većina IBC i nIBC uzoraka, a u klasteru 2 bilo je sedam preostalih nIBC uzoraka i jedan IBC uzorak. S druge strane, u slučaju bazalnog podtipa, klaster 1 sastojao se većinom od IBC uzoraka, a klaster 2 od nIBC uzoraka. Samo su se četiri uzorka našla u neodgovarajućem klasteru: dva IBC i dva nIBC. Unatoč smanjenju broja gena, IBC i nIBC uzorci bazalnog podtipa i dalje su se prilično dobro odvajali na PCA grafu.

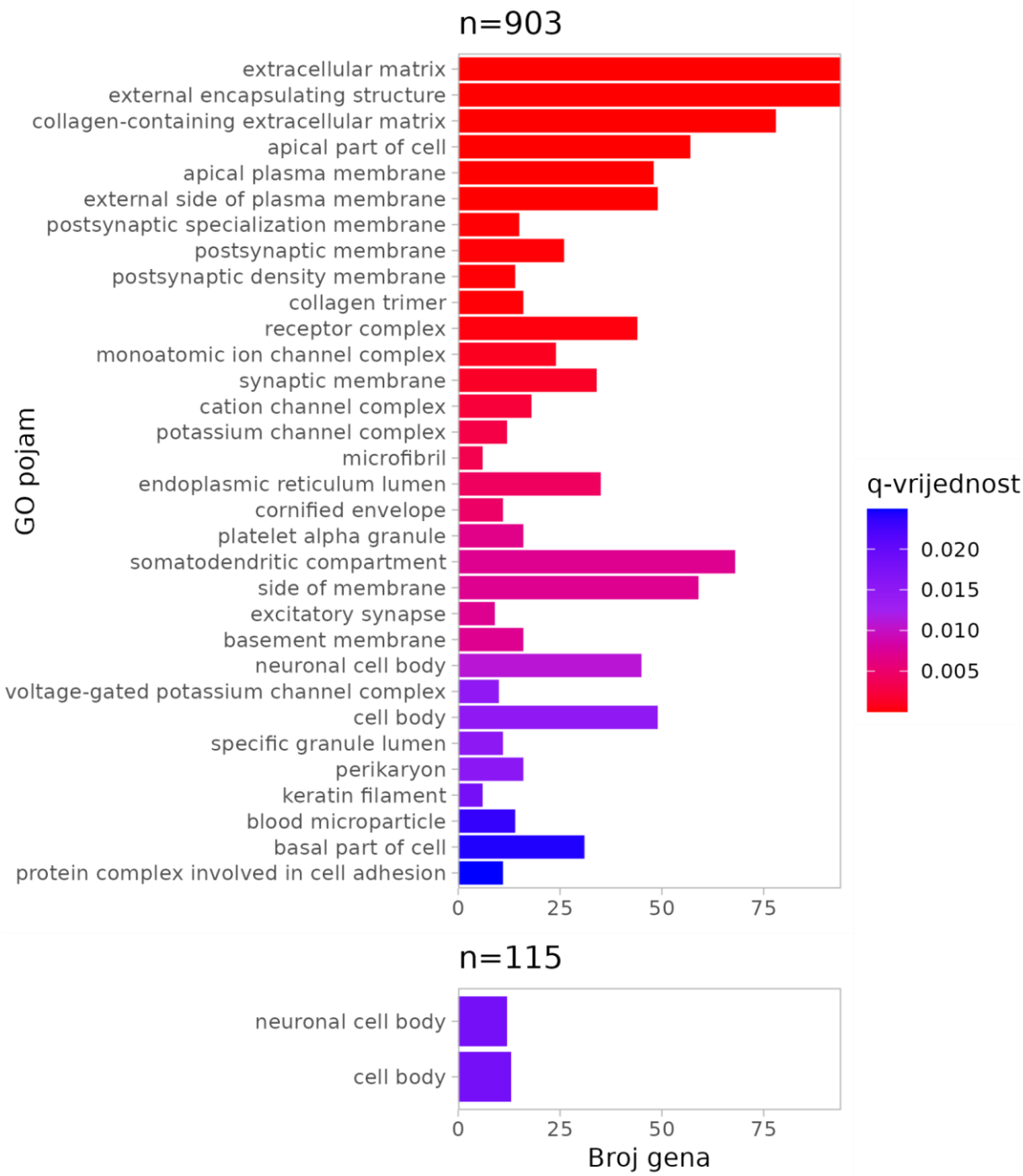


Slika 12. Hijerarhijsko klasteriranje uzoraka na temelju 115 gena s najvišim koeficijentima u *XGBoost* linearnom modelu. Analiza je provedena zasebno na uzorcima svih GEP podtipova osim bazalnog (A) i uzorcima bazalnog podtipa (B). Optimalan broj klastera određen je na temelju prosječne širine siluete te on u oba slučaja iznosi 2. Klasteri definirani hijerarhijskim klasteriranjem označeni su na PCA grafovima.

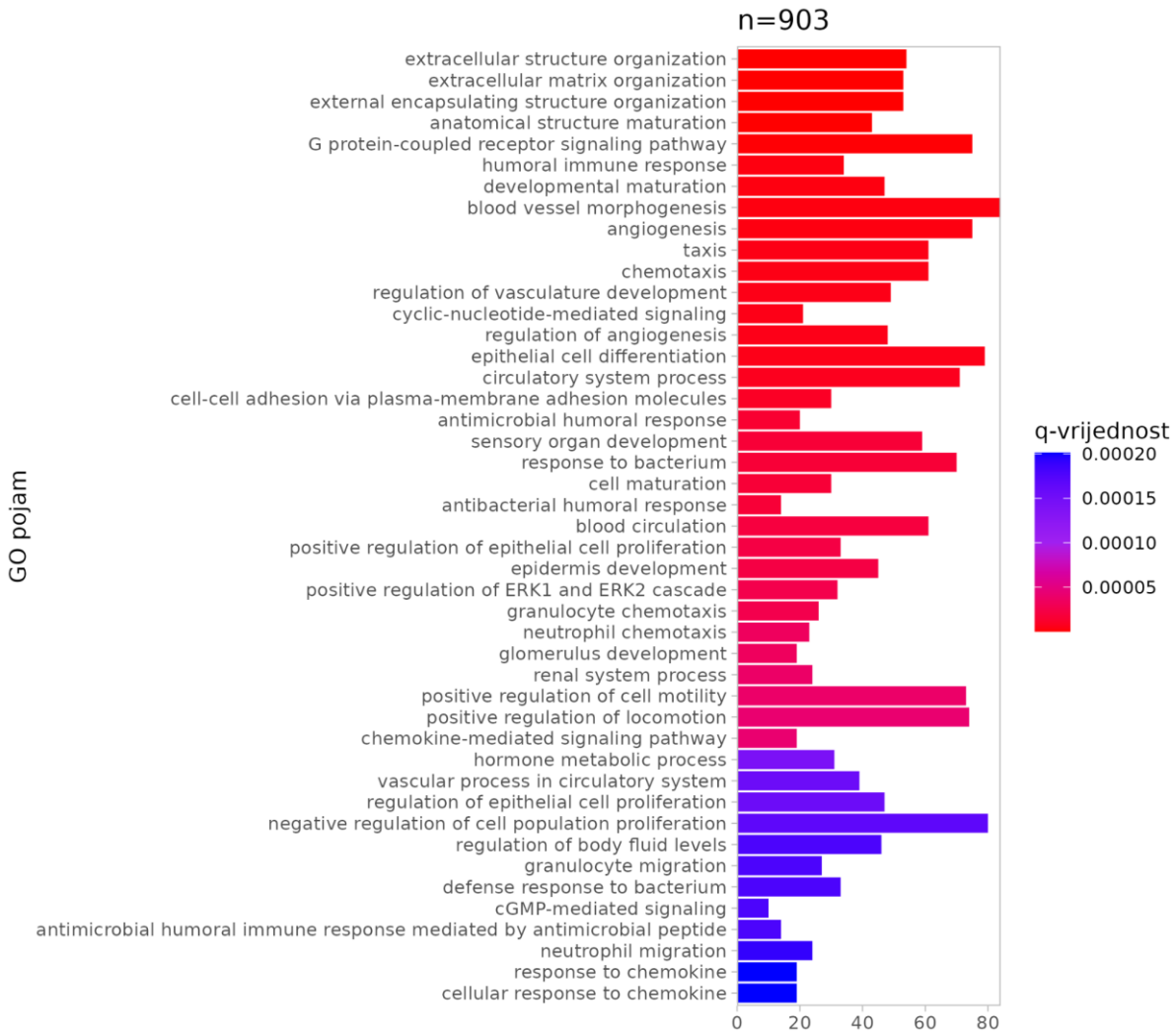
4.4 Geni prediktori u modelu za bazalni podtip

Za daljnju analizu gena modela korištena je metoda ORA. Prvo je provjereno jesu li u određenim staničnim odjeljcima prezastupljeni produkti gena uključenih u model strojnog učenja. To je učinjeno za D903 i HC115. Produkti gena iz D903 pokazali su se prezastupljenima u 32 dijela stanice (Slika 13, gore). Najzastupljeniji su bili pojmovi vezani za staničnu okolinu i apikalni dio stanice: *extracellular matrix* (q=1,54×10^{-24}), *external encapsulating structure* (q=1,54×10^{-24}), *collagen-containing extracellular matrix* (q=1,27×10^{-18}) i *apical part of the cell* (q=6,37×10^{-8}). U slučaju skupa HC115 bila su samo dijela stanice u kojima postoji prezastupljenost: stanično tijelo i stanično tijelo neurona (Slika 13, dolje).

Ista analiza provedena je za biološke procese. 278 bioloških procesa prezastupljeno je (q<0,05) za produkte gena iz D903 (Slika 14) i niti jedan za HC115. Statistička značajnost bila je najveća za pojmove vezane za razvoj te organizaciju izvanstaničnih struktura i izvanstaničnog matriksa (ECM, engl. *extracellular matrix*). Isticali su se i neki procesi vezani za imunosni sustav, među kojima su humoralni imunosni odgovor, odgovor na bakterije, kemotaksija imunosnih stanica (neutrofila i granulocita) i kemokinima posredovani signalni putevi. Također je bilo više pojmova vezanih za krvožilni sustav, poput angiogeneze i njezine regulacije, morfogeneze krvnih žila i krvne cirkulacije. S obzirom na rezultate GO analize skupa HC115 za stanične komponente, obraćena je pozornost i na pojmove vezane za živčani sustav (nisu prikazani na Slici 14): *maintenance of blood-brain barrier* (q=2,01×10^{-3}), *nervous system process* (q=5,06×10^{-3}), *regulation of neuron differentiation* (q=5,93×10^{-3}), *regulation of nervous system development* (q=1,36×10^{-2}) i *positive regulation of neuron differentiation* (q=1,37×10^{-2}).



Slika 13. Rezultati analize genske ontologije za stanične dijelove. GO pojmovi su poredani uzlazno po *q*-vrijednosti. Napomena: ovi pojmovi nisu prevedeni jer pripadaju ontologiji, strogo definiranoj nomenklaturi vezanoj za gene i genske produkte.



Slika 14. Rezultati analize genske ontologije za biološke procese. GO pojmovi su poredani uzlazno po *q*-vrijednosti. Bioloških procesa na popisu ima ukupno 278, a ovdje su prikazana samo prva 32. Napomena: ovi pojmovi nisu prevedeni jer pripadaju ontologiji, strogo definiranoj nomenklaturi vezanoj za gene i genske produkte.

5. Rasprava

5.1 Povećanje broja uzoraka i gubitak gena u pretprocesiranju

Spajanjem mikročipova dobiven je skup podataka s puno više uzoraka nego što je uobičajeno za transkriptomska istraživanja IBC-a, u kojima je tipičan broj uzoraka (IBC i nIBC zajedno) između 30 i 90 (Bertucci *et al.*, 2005; Nguyen *et al.*, 2006; Van Laere *et al.*, 2007; Zare, Postovit i Githaka, 2021). Pritom IBC uzoraka najčešće ima manje nego nIBC uzoraka, kao što je slučaj i u ovom radu. Nažalost, povećanje broja uzoraka bilo je popraćeno gubitkom gotovo polovice gena. Kad bi se mikročip 1 spajao samo s jednim, bilo kojim od preostala dva mikročipa, gubitak bi bio manji za oko 3000 gena, što i dalje nije idealno. Spajanje samo s mikročipom 2 bilo bi daleko povoljnije u pogledu očuvanja G59 gena, budući da ih on sadrži 55/59, duplo više od mikročipa 3. Jedna od potencijalnih primjena spajanja mikročipova je replikacija rezultata Zare, Postovit i Githaka (2021), gdje su G59 geni bili izrazito uspješni u razvrstavanju uzoraka u klase IBC i nIBC. Međutim, gubitak G59 gena je to u ovom skupu podataka spriječio. Zare, Postovit i Githaka (2021) su za treniranje i validaciju svog modela koristili podatke iz mikročipa 1, stoga bi se kao nezavisan skup podataka za provjeru njihovog rezultata ubuduće mogao iskoristiti mikročip 2. Što se tiče gena iz skupa PAM50, mikročip 3 ih sadrži šest više od mikročipa 2, zbog čega se klasifikacija uzoraka iz mikročipa 3 u PAM50 podtipove smatra pouzdanijom. Valja napomenuti da je gubitak do devet PAM50 gena i dalje prihvatljiv za klasifikaciju te se tolerirao u literaturi (Creighton, 2012).

Kako bi se ubuduće spriječio gubitak velikog broja gena, bilo bi preporučljivo spajati mikročipove istih platformi. Druga opcija je kombinirati podatke dobivene mikročipom s podacima dobivenim RNA sekvenciranjem (RNA-seq). U RNA-seq eksperimentima, cDNA se ne veže za komplementarne probe, već se sekvencira, te se kvantificira koliko se često određena sekvenca pojavljuje u rezultatima. Prednost ovog pristupa je mogućnost detekcije transkripata bilo kojeg gena, za razliku od mikročipova na kojima se mogu detektirati samo transkripti za koje postoje komplementarne probe na pločici (Lesk, 2012). Ovo je jedan od razloga zašto se danas rezultati RNA-seq eksperimenata sve češće viđaju u javnim bazama podataka (Clough *et al.*, 2023). Za kombiniranje mikročipova s RNA-seq podacima koriste se metode poput TDM-a (engl. *training distribution matching*), kvantil-normalizacije i neparanormalne normalizacije (Foltz, Greene i Taroni, 2023). U teoriji, spoj tri mikročipa dobiven u ovom radu spreman je za spajanje s RNA-seq podacima. Metoda TDM normalizira RNA-seq podatke prema mikročipu, dok mikročip ostaje netaknut (jer RNA-seq podaci imaju veći dinamički

raspon) (Thompson, Tan i Greene, 2016). Ovo je i poželjno jer je učinak uzastopnih normalizacija uglavnom nepoznat, stoga bi postojao rizik od iskrivljenja podataka (Walsh *et al.*, 2015). Valja napomenuti da je normalizacija pojedinačnih mikročipova prije normalizacije metodom *ComBat* opisana u potpoglavlju 3.4 (str. 12) u redu. Štoviše, *ComBat* podrazumijeva da je provedena (Johnson, Li i Rabinovic, 2007).

5.2 Uklanjanje nebioloških i zadržavanje bioloških razlika

Na PCA grafu vrlo se jasno vidjelo da su nebiološke razlike između mikročipova uspješno uklonjene. Teže je zaključiti jesu li biološke razlike zadržane jer ne postoji pouzdana metoda provjere je li nešto realna razlika između uzoraka ili artefakt. U ovom slučaju su za provjeru odabrane metode nenadziranog učenja. U slučaju kad je uključeno svih 9027 gena, PCA je pokazao da se uzorci uglavnom grupiraju po GEP podtipu i ER statusu koji je s njime povezan, a IBC i nIBC uzorci se miješaju. Pritom je otežano razdvajanje uzoraka luminalnog A i luminalnog B podtipa, možda zbog toga što su u mikročipu 2 i mikročipu 3 nedostajali PAM50 geni za precizniju klasifikaciju. Sve ovo je uočeno i u hijerarhijskom klasteriranju uzoraka na temelju 39 gena iz skupa PAM50. Woodward *et al.* (2013) su u hijerarhijskom klasteriranju na temelju 1000 gena dobili klaster bogat ER+ uzorcima i klaster gdje ima nešto više ER- uzoraka. Status ER-a je važan biomarker, ali opet je samo jedno od obilježja raka dojke, stoga je za očekivati da se uzorci po njemu odvajaju donekle, ali nesavršeno. Toplinska mapa (engl. *heatmap*) ekspresije PAM50 gena uspoređena je s radom Creighton (2012) i uočeno je da im se obrasci specifični za GEP podtipove poklapaju. Ovi rezultati idu u prilog tome da normalizacija po platformama nije uklonila biološke razlike.

U literaturi postoje primjeri kad su metode uklanjanja nebioloških razlika, uključujući *ComBat*, greškom uvele neku novu nebiološku razliku u uzorke, što je dovelo do lažnih rezultata (Nygaard, Rødland i Hovig, 2016; Price i Robinson, 2018). To se obično događa kad biološke grupe (koje su predmet istraživanja) nisu ravnomjerno raspoređene po nebiološkim grupama, u ovom slučaju mikročipovima. Neravnomjerna raspodjela uočena je kod mikročipa 2, gdje ima više nIBC uzoraka od IBC uzoraka, te kod mikročipa 3, gdje uvjerljivo dominiraju HER2+ uzorci. No, uklanjanje uzoraka iz većinskih klasa ne bi učinilo podatke imunima na ovaj problem (Zindler *et al.*, 2020). Ovaj rizik postoji kod svake metode normalizacije po platformama, zbog čega bi možda bilo mudrije koristiti metaanalizu kada je primjenjiva. Problem je što metaanaliza u potpunosti onemogućava stvaranje modela nadziranog učenja, budući da se podaci iz različitih mikročipova moraju spojiti u jedinstveni skup za treniranje.

Normalizaciju po platformama pritom nije moguće preskočiti, budući da su razlike između mikročipova veće od bilo kakvih razlika po biološkim kategorijama. Imajući ovo na umu, računalno dobiveni rezultati uvijek bi trebali biti provjereni eksperimentalno, u istom ili u različitom istraživanju, što je i uobičajena praksa. Stoga je bolje iskoristiti podatke u javnim bazama nego odustati od potencijalno vrlo korisnih metoda, pogotovo u slučaju rijetkih i smrtonosnih bolesti kao što je IBC.

5.3 Različiti pristupi klasifikaciji uzoraka upalnog i neupalnog raka dojke

Uspješnost klasifikacije uzoraka u IBC i nIBC varirala je ovisno o modelu. Spajanje mikročipova povećalo je broj uzoraka, što je otvorilo mogućnost izrade modela specifičnih za GEP podtipove. Klasifikacija neovisna o GEP podtipu nije bila uspješna. Vrijednost κ=0,53 je preniska da bi se model smatrao pouzdanim, a i točnost je niža od drugih modela koje su provjeravali Zare, Postovit i Githaka (2021) na dva skupa podataka gdje su uzorci uzeti prije liječenja kemoterapijom. Uz to, tijekom optimizacije tog modela primijećeno je da male promjene u parametrima dovode do vrlo različitog skupa prediktorskih gena, a opet vrlo slične uspješnosti u klasifikaciji. Međutim, i model temeljen na potpisu G59 ima svoje probleme. On nije mogao klasificirati uzorke pacijentica koje su liječene neoadjuvantnom kemoterapijom. Nije točno klasificirao niti jedan takav IBC uzorak iz skupa za testiranje modela. Autori navode da je uzrok tome utjecaj kemoterapije na ekspresiju gena, ali je svejedno neobično da je model u potpunosti izgubio svoju sposobnost klasifikacije.

G59 je neovisan o podtipu, što ne mora nužno biti prvotna namjera autora, budući da broj uzoraka pojedinih podtipova niti nije bio dovoljno velik za treniranje i validaciju modela specifičnih za podtip. Iako bi model neovisan o podtipu bio praktičniji za dijagnostiku, modeli specifični za podtip mogli bi se pokazati točnijima. Za razliku od modela za HER2 podtip, modeli za bazalni i luminalni A podtip vidno su bolji od modela neovisnog o podtipu, ali niti od njih nadmašio najbolji rezultat modela temeljenog na G59. Glavni razlog tome je najvjerojatnije gubitak gotovo polovice gena spajanjem mikročipova. Moguće je da postoje geni koji su važni prediktori za klasifikaciju, ali su izbačeni iz analize. Ako je gubitak gena zaista izvor problema u klasifikaciji, čini se da je njime najmanje pogođen bazalni podtip. Razdvajanje IBC i nIBC uzoraka u PCA-u i hijerarhijskom klasteriranju dodatno validira prediktivnu sposobnost modela specifičnog za bazalni podtip. Rezultati ukazuju na to da bi se vjerojatno isplatilo napraviti slične klasifikatore na kvalitetnijem skupu podataka. Također bi trebalo imati na umu da genski potpisi obično imaju manje od 100 gena (Walsh *et al.*, 2015),

što čini čak i najmanji skup prediktora definiran u ovom radu, HC115, većim nego što je poželjno.

5.4 Funkcija i lokalizacija produkata prediktorskih gena

U skupu gena D903 nalazi se 11 gena iz skupa G59: *ACVRL1*, *ADGRF5*, *CDH5*, *CROT*, *FOS*, *NXN*, *S1PR1*, *SEMA3E*, *SLCO3A1*, *TCIM* i *TGFA*, dok se u skupu HC115 od njih nalazi samo *FOS*. Uz to, mogu se naći određene sličnosti u lokalizaciji i funkciji produkata gena koji im pripadaju. Primjerice, G59 ima značajno poklapanje s pojmom *integral component of plasma membrane* (identifikacijska oznaka GO:0005887) (Zare, Postovit i Githaka, 2021), pod koji spadaju geni *ACVRL1*, *TGFA* i *SLCO3A1*, koji su zajednički skupovima D903 i G59. D903 je povezan s nizom pojmova vezanih za staničnu membranu, među kojima je i vanjski dio stanične membrane, uz kojeg postoji i veliko preklapanje s izvanstaničnim matriksom (ECM, engl. *extracellular matrix*) za kojeg su stanice pričvršćene membranskim komponentama. Glavni sastavni dio ECM-a su kolageni, na koje se također odnosi više pojmova u rezultatima GO analize. Akumulacija i križno povezivanje kolagenskih filamenata zaslužni su za skrućivanje tkiva koje je obilježje velikog broja zloćudnih tumora dojke (Najafi, Farhood i Mortezaee, 2019). Kao što je već spomenuto, krute kvržice uočavaju se rjeđe kod IBC-a nego kod nIBC-a, zbog čega je na prvi pogled smisleno da i ekspresija povezanih gena bude različita (Robertson *et al.*, 2010). Nadalje, preduvjet za uspješnu metastatsku kolonizaciju je prolazak stanica kroz ECM da bi došle do lumena krvne žile. Za razgradnju ECM-a koriste metaloproteinaze matriksa (MMP, engl. *matrix metalloproteinase*) koje su pojačano eksprimirane u gotovo svakom tumoru (Pecorino, 2012). Spadaju pod GO pojmove *external encapsulating structure* i *extracellular matrix*. U skupu D903 su četiri gena koji kodiraju za njih: *MMP1*, *MMP10*, *MMP12* i *MMP15*. Međutim, u slučaju gena vezanih za kolagen i metaloproteinaze mora se postaviti pitanje koje su stanice zapravo izvor tih signala. Iako je zabilježeno da stanice raka dojke u kulturi mogu proizvoditi kolagen (Minafra *et al.*, 1988), opće je prihvaćeno da se u raku ta uloga pripisuje fibroblastima. Tumorske stanice proizvode TGF-β (engl. *transforming growth factor β*) i ciklooksigenazu-2 da regrutiraju normalne fibroblaste. TGF-β ih tada pretvori u fibroblaste povezane s tumorom (engl. *cancer associated fibroblasts*) (Najafi, Farhood i Mortezaee, 2019). Osim što proizvode komponente izvanstaničnog matriksa, fibroblasti povezani s tumorom proizvode MMP-ove kao i tumorske stanice (Pecorino, 2012). Van Laere *et al.* (2013) su u IBC-u uočili smanjenje ekspresije gena uključenih u signalni put TGF-β, kao što je *TGFB2* (engl. *transforming growth factor beta 2*), u odnosu na nIBC. U skupu M912 nema gena za TGF-β, ali sadrži gene za njegove receptore: *ACVRL1* (Garrido-Martin *et al.*,

2010) i *TGFBR3* (engl. *transforming growth factor beta receptor 3*). Uloga *TGFBR3* u raku dojke je predmet rasprave jer sudjeluje u više signalnih puteva, a zasad se pokazalo su da je jače eksprimiran u stanicama raka dojke nego u fibroblastima (Jovanović *et al.* 2016).

U skupu D903 nalaze se razni geni vezani za angiogenezu, proces u kojem tumori stvaraju nove krvne žile iz već postojećih. Ovo omogućava opskrbu rastućeg broja tumorskih stanica hranjivim tvarima i kisikom te ulazak u krvotok i metastazu (Muz *et al.*, 2015). Prema Van der Auwera *et al.* (2004), IBC ima višu ekspresiju 10 gena vezanih za angiogenezu i limfangiogenezu (stvaranje limfnih žila) u odnosu na nIBC. Njih četiri se nalazi u D903: *FLT4*, *LYVE1*, *TEK* (*Tie-2*) i *TIE1*. Zaključili su da je IBC vrlo angiogeni i limfangiogeni tumor, što mu olakšava metastaziranje. Osim što je ciklooksigenaza-2 važna za privlačenje fibroblasta, ima proangiogeni učinak (Gately i Li, 2004), stoga je i ona bila uključena u istraživanje Van der Auwera *et al.* (2004), ali joj je povećanje ekspresije nije bilo statistički značajno. Unatoč tome, njezin gen (*PTGS2*) nalazi se u skupu D903. Uz povišenu ekspresiju angiogenetskih gena, u literaturi je zabilježena i veća gustoća krvnih mikrožila u IBC-u nego u nIBC-u (McCarthy *et al.* 2002).

Zanimljivo je da se u skupu D903 nalaze geni uključeni u imunosni odgovor na bakterije, a simptomi IBC-a znaju biti slični bakterijskoj infekciji (Robertson *et al.*, 2010). Ovdje je ponovno upitan izvor transkripata i postoji mogućnost da oni dolaze od stanica imunosnog sustava koje su infiltrirale tumor. Međutim, čak i skup G59 sadrži gene povezane s imunošću, a autori su radili na uzorcima s udjelom tumorskih stanica >90%. G59 sadrži neke gene iz signalnog puta interleukina-2 i kemokinskog signalnog puta, iako njihovo preklapanje nije statistički značajno (Zare, Postovit i Githaka, 2021). Među tim genima iz signalnog puta interleukina-2 su i dva iz skupa D903: *S1PR1* i *FOS* (koji pripada i skupu HC115). D903 sadrži gene vezane za kemokine, ali se ne preklapaju s onima iz G59.

Dobiven je i jedan naizgled neobičan rezultat: tijelo neurona bogato je produktima gena iz HC115. Stanice bazalnog podtipa su trostruko negativne po IHC klasifikaciji (uz neke iznimke), a kod trostruko negativnog raka dojke postoji povećan rizik od metastaziranja u središnji živčani sustav (Lin *et al.*, 2013). Stanice raka dojke čak mogu steći obilježja moždanih stanica, što se smatra prilagodbom na naseljavanje moždanog mikrookoliša (Neman *et al.*, 2014; Zeng *et al.*, 2019). Tumorska masa može unutar sebe sadržavati živčano tkivo (Bourgot *et al.*, 2020), ali s obzirom na to da se radi o bazalnom podtipu, čini se izglednije da se radi o adaptaciji tumorskih stanica.

5.5 Nečistoće u uzorcima tumora

Uzorci iz mikročipa 1 sadrže >90% tumorskih stanica (Woodward *et al.*, 2013), dok su u slučaju mikročipa 3 autori postavili puno niži kriterij: >60% tumorskih stanica (Sircoulomb *et al.*, 2010). Uz to, anotacija uzoraka iz mikročipa 3 ne sadrži izmjeren udio tumorskih stanica, stoga ih u ovom radu nije bilo moguće filtrirati. Mikročip 2 ne sadrži nikakvu informaciju vezanu za udio tumorskih stanica (Iwamoto *et al.*, 2011). Po ovom kriteriju, podaci iz mikročipa 1 mogu se smatrati kvalitetnima i zasigurno im se može pripisati dio dosadašnjeg uspjeha genskog profila G59. Kao što je već diskutirano, tumorska masa ne sadrži samo tumorske stanice, nego i druge stanice iz svog mikrookoliša, primjerice stromalne i imunosne (Bourgot *et al.*, 2020). One predstavljaju nečistoće koje otežavaju određivanje profila ekspresije gena. Ako se za to koristi statistički test, jedna od preporuka glasi da bi postotak tumorskih stanica u uzorku trebao biti >90%. Ovo bi trebalo učiniti da postotak lažno pozitivnih gena bude 5% ili manji. Prema literaturi, čini se da klasifikacijski modeli ipak bolje podnose nečistoće, ali je iz očitih razloga i dalje povoljnije da one budu čim manje (de Ridder *et al.*, 2005). Još jedna opcija je računalno procijeniti na temelju transkriptomskih podataka, što je i jedina opcija kada se radi na javno dostupnim podacima. Za to se koriste metode *ESTIMATE* (Yoshihara *et al.*, 2013), *PUREE* (Revkov *et al.*, 2023) i sl. Međutim, postoje dva razloga zašto to na ovom skupu podataka možda nije najbolja opcija. Prvo, model ne mora biti jednako dobar za sve tipove raka, čak i ako se zasad pokazao prilagodljivim. Moguće je da performanse ne bi bile dobre na rijetkom tipu raka kao što je IBC. Drugo, možda nedostaje previše gena koji bi bili ključni za preciznost procjene.

Na čistoću tumora mogle bi utjecati i metode kojima je tkivo secirano. Kod mikročipa 1 to je ručna mikrodisekcija, a kod mikročipa 3 makrodisekcija. Makrodisekcija je vjerojatno najčešća metoda prikupljanja uzoraka u kliničkoj onkologiji. U njoj se naprave presjeci tkiva, jedan tretiran histološkim bojama, drugi neobojan. Boja služi za razlikovanje tipova stanica. S neobojanog presjeka se skalpelom izrezuje regija od interesa, a obojani presjek služi kao vodilja. Ova metoda je dobra kad tumorska masa ima dobro definirane granice (što je slučaj kod veličine tumora) i nije heterogena (Walsh i Halushka, 2023). Kod IBC-a je problematično što u najvećem dijelu slučajeva ne postoji kvržica, već je tumor difuzne građe (Robertson *et al.*, 2010; Van Uden *et al.*, 2015). Ručna mikrodisekcija je slična makrodisekciji, samo što se u njoj tkivo gleda pod mikroskopom, a stanice iz regije od interesa prikupljaju pomoću igle, što je čini puno preciznijom (Walsh i Halushka, 2023). Prema ovome je mikrodisekcija povoljnija za IBC od makrodisekcije. Ručna mikrodisekcija može narušiti integritet tkiva, što bi smanjilo

preciznost, zbog čega bi još bolje bilo koristiti specijalizirani instrument (Eltoum, Siegal i Frost, 2002). Slične metode mogu se primijeniti i na citološkim uzorcima, primjerice onima dobivenima aspiracijom tankom iglom (Walsh i Halushka, 2023). Tako su prikupljeni uzorci u slučaju mikročipa 2, no autori pripadajućeg rada (Iwamoto *et al.*, 2011) nisu naveli ništa o razdvajanju tumorskih stanica od netumorskih.

6. Zaključci

Podaci iz tri različita eksperimenta na mikročipovima uspješno su spojeni i normalizacijom po platformama dovedeni do oblika spremnog za upotrebu u strojnom učenju. Nebiološke razlike između mikročipova su uklonjene, a u podacima nisu uočene nikakve nepravilnosti koje bi mogle biti uzrokovane normalizacijom. Ova metoda najpovoljnija je kad su dostupni podaci iz eksperimenata na istoj platformi mikročipa ili sve uključene platforme pokrivaju gotovo cijeli genom. To nažalost nije bio slučaj u ovom istraživanju, što je rezultiralo gubitkom gena i otežanom klasifikacijom uzoraka u upalni i neupalni rak dojke. Svejedno, veći broj uzoraka omogućio je stvaranje modela strojnog učenja specifičnih za PAM50 podtipove, a postoje i naznake da bi mogli biti uspješniji od modela neovisnih o podtipu. Normalizirani podaci smatraju se spremnima i za spajanje s RNA-seq podacima metodama poput TDM-a. Ukoliko se genski potpis G59 ne pokaže dovoljno dobrim za kliničku primjenu, razvoj modela ovisnih o podtipu bio bi smislen sljedeći korak. Definiranju genskog potpisa upalnog raka dojke uvelike bi pomogli kvalitetni eksperimentalni podaci, kojih nažalost nema dovoljno. Posebnu pozornost trebalo bi obratiti na udio tumorskih stanica u uzorku. Osim za istraživanje upalnog raka dojke, normalizacija po platformama u kombinaciji sa strojnim učenjem mogla bi se primijeniti za istraživanje ekspresije gena u drugim rijetkim bolestima.

7. Zahvale

Puno hvala izv. prof. dr. sc. Rosi Karlić na mentorstvu i Pauli Štancl, mag. biol. mol. na velikoj pomoći pri izradi ovog rada. Objema se zahvaljujem na njihovoj stručnosti, pristupačnosti i poticajnoj atmosferi. Također im se želim zahvaliti za sve što sam naučila na njihovim kolegijima koji su me zainteresirali za bioinformatiku i bili osnova za ovaj rad.

8. Popis literature

Alexandrov, L.B. et al. (2020) ‘The repertoire of mutational signatures in human cancer’, Nature, 578(7793), pp. 94–101. Dostupno na: https://doi.org/10.1038/s41586-020-1943-3.

Berger, M.F. i Mardis, E.R. (2018) ‘The emerging clinical relevance of genomics in cancer medicine’, *Nature Reviews Clinical Oncology*, 15(6), pp. 353–365. Dostupno na: https://doi.org/10.1038/s41571-018-0002-6.

Bertucci, F. *et al.* (2005) ‘Gene Expression Profiling Identifies Molecular Subtypes of Inflammatory Breast Cancer’, *Cancer Research*, 65(6), pp. 2170–2178. Dostupno na: https://doi.org/10.1158/0008-5472.CAN-04-4115.

Bertucci, F. *et al.* (2014) ‘Genomic profiling of inflammatory breast cancer: A review’, *The Breast*, 23(5), pp. 538–545. Dostupno na: https://doi.org/10.1016/j.breast.2014.06.008.

Bischl, B. *et al.* (2016) ‘mlr: Machine Learning in R’. Dostupno na: https://jmlr.org/papers/v17/15-066.html.

Bolstad, B. (2024) ‘preprocessCore: A collection of pre-processing functions’. Dostupno na: 10.18129/B9.bioc.preprocessCore.

Bourgot, I. *et al.* (2020) ‘Reciprocal Interplay Between Fibrillar Collagens and Collagen-Binding Integrins: Implications in Cancer Progression and Metastasis’, *Frontiers in Oncology*, 10. Dostupno na: https://www.frontiersin.org/articles/10.3389/fonc.2020.01488.

Bray, F. *et al.* (2024) ‘Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries’, *CA: A Cancer Journal for Clinicians*, 74(3), pp. 229–263. Dostupno na: https://doi.org/10.3322/caac.21834.

Brownlee, J. (2016) *XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn*. Machine Learning Mastery.

Chen, T. *et al.* (2024) ‘xgboost: Extreme Gradient Boosting’. Dostupno na: https://CRAN.R-project.org/package=xgboost.

Chen, T. i Guestrin, C. (2016) ‘XGBoost: A Scalable Tree Boosting System’, u *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. *KDD ’16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, pp. 785–794. Dostupno na: https://doi.org/10.1145/2939672.2939785.

Clough, E. *et al.* (2023) ‘NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update’, *Nucleic Acids Research*, 52(D1), pp. D138–D144. Dostupno na: https://doi.org/10.1093/nar/gkad965.

Creighton, C.J. (2012) 'The molecular profile of luminal B breast cancer', *Biologics: Targets and Therapy*, 6, pp. 289–297. Dostupno na: https://doi.org/10.2147/BTT.S29923.

Cui, X. *et al.* (2005) 'Biology of Progesterone Receptor Loss in Breast Cancer and Its Implications for Endocrine Therapy', *Journal of Clinical Oncology*, 23(30), pp. 7721–7735. Dostupno na: https://doi.org/10.1200/JCO.2005.09.004.

Dai, X. *et al.* (2016) 'Cancer Hallmarks, Biomarkers and Breast Cancer Molecular Subtypes', *Journal of Cancer*, 7(10), pp. 1281–1294. Dostupno na: https://doi.org/10.7150/jca.13141.

Eltoum, I.A., Siegal, G.P. i Frost, A.R. (2002) 'Microdissection of Histologic Sections: Past, Present, and Future':, *Advances in Anatomic Pathology*, 9(5), pp. 316–322. Dostupno na: https://doi.org/10.1097/00125480-200209000-00006.

Foltz, S.M., Greene, C.S. i Taroni, J.N. (2023) 'Cross-platform normalization enables machine learning model training on microarray and RNA-seq data simultaneously', *Communications Biology*, 6(1), pp. 1–10. Dostupno na: https://doi.org/10.1038/s42003-023-04588-6.

Gao, C.-H. i Dusa, A. (2024) 'ggVennDiagram: A "ggplot2" Implement of Venn Diagram'. Dostupno na: https://CRAN.R-project.org/package=ggVennDiagram.

Garrido-Martin, E.M. *et al.* (2010) 'Characterization of the human Activin-A receptor type II-like kinase 1 (ACVRL1) promoter and its regulation by Sp1', *BMC Molecular Biology*, 11, p. 51. Dostupno na: https://doi.org/10.1186/1471-2199-11-51.

Gately, S. i Li, W.W. (2004) 'Multiple roles of COX-2 in tumor angiogenesis: a target for antiangiogenic therapy', *Seminars in Oncology*, 31(2 Suppl 7), pp. 2–11. Dostupno na: https://doi.org/10.1053/j.seminoncol.2004.03.040.

Gendoo, D.M.A. *et al.* (2024) 'Computation of Gene Expression-Based Signatures in Breast Cancer'. Dostupno na: http://www.pmgenomics.ca/bhklab/software/genefu.

Gu, Z., Eils, R. i Schlesner, M. (2016) 'Complex heatmaps reveal patterns and correlations in multidimensional genomic data', *Bioinformatics*, 32(18), pp. 2847–2849. Dostupno na: https://doi.org/10.1093/bioinformatics/btw313.

Hardiman, G. (2004) 'Microarray Platforms – Comparisons and Contrasts', *Pharmacogenomics*, 5(5), pp. 487–502. Dostupno na: https://doi.org/10.1517/14622416.5.5.487.

Hastie, T., Tibshirani, S. i Friedman, H. (2008) *The Elements of Statistical Learning*. Drugo izdanje.

Hong, R. i Xu, B. (2022) 'Breast cancer: an up-to-date review and future perspectives', *Cancer Communications*, 42(10), pp. 913–936. Dostupno na: https://doi.org/10.1002/cac2.12358.

Hortobagyi, G. *et al.* (2017) 'AJCC Cancer Staging Manual'. Osmo izdanje. Springer International Publishing. Dostupno na: https://doi.org/10.1007/978-3-319-40618-3_48.

Iwamoto, T. *et al.* (2011) 'Different gene expressions are associated with the different molecular subtypes of inflammatory breast cancer', *Breast Cancer Research and Treatment*, 125(3), pp. 785–795. Dostupno na: https://doi.org/10.1007/s10549-010-1280-6.

Johnson, W.E., Li, C. i Rabinovic, A. (2007) 'Adjusting batch effects in microarray expression data using empirical Bayes methods', *Biostatistics*, 8(1), pp. 118–127. Dostupno na: https://doi.org/10.1093/biostatistics/kxj037.

Jovanović, B. *et al.* (2016) 'TβRIII Expression in Human Breast Cancer Stroma and the Role of Soluble TβRIII in Breast Cancer Associated Fibroblasts', *Cancers*, 8(11), p. 100. Dostupno na: https://doi.org/10.3390/cancers8110100.

Kassambara, A. (2023) 'ggpubr: "ggplot2" Based Publication Ready Plots'. Dostupno na: https://CRAN.R-project.org/package=ggpubr.

Kassambara, A. i Mundt, F. (2020) 'factoextra: Extract and Visualize the Results of Multivariate Data Analyses'. Dostupno na: https://CRAN.R-project.org/package=factoextra.

Lander, E.S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921. Dostupno na: https://doi.org/10.1038/35057062.

Leek, J.T. *et al.* (2024) 'sva: Surrogate Variable Analysis'.

Lerebours, F. *et al.* (2021) 'Hemoglobin overexpression and splice signature as new features of inflammatory breast cancer?', *Journal of Advanced Research*, 28, pp. 77–85. Dostupno na: https://doi.org/10.1016/j.jare.2020.08.009.

Lesk, A.M. (2012) *Introduction to genomics*. Drugo izdanje. Oxford ; New York: Oxford University Press.

Lin, N.U. *et al.* (2013) 'CNS Metastases in Breast Cancer: Old Challenge, New Frontiers', *Clinical Cancer Research*, 19(23), pp. 6404–6418. Dostupno na: https://doi.org/10.1158/1078-0432.CCR-13-0790.

Lunardon, N., Menardi, G. i Torelli, N. (2014) 'ROSE: a Package for Binary Imbalanced Learning'.

Mamouch, F. *et al.* (2018) 'Inflammatory Breast Cancer: A Literature Review', *World Journal of Oncology*, 9(5–6), pp. 129–135. Dostupno na: https://doi.org/10.14740/wjon1161.

Martínez-Jiménez, F. *et al.* (2020) 'A compendium of mutational cancer driver genes', *Nature Reviews Cancer*, 20(10), pp. 555–572. Dostupno na: https://doi.org/10.1038/s41568-020-0290-x.

McCarthy, N.J. *et al.* (2002) 'Microvessel density, expression of estrogen receptor alpha, MIB-1, p53, and c-erbB-2 in inflammatory breast cancer', *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 8(12), pp. 3857–3862.

Minafra, S. *et al.* (1988) 'Collagen biosynthesis by a breast carcinoma cell strain and biopsy fragments of the primary tumour', *Cell Biology International Reports*, 12(10), pp. 895–905. Dostupno na: https://doi.org/10.1016/0309-1651(88)90053-7.

Muz, B. *et al.* (2015) 'The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy', *Hypoxia*, p. 83. Dostupno na: https://doi.org/10.2147/HP.S93413.

Najafi, M., Farhood, B. i Mortezaee, K. (2019) 'Extracellular matrix (ECM) stiffness and degradation as cancer drivers', *Journal of Cellular Biochemistry*, 120(3), pp. 2782–2790. Dostupno na: https://doi.org/10.1002/jcb.27681.

Neman, J. *et al.* (2014) 'Human breast cancer metastases to the brain display GABAergic properties in the neural niche', *Proceedings of the National Academy of Sciences*, 111(3), pp. 984–989. Dostupno na: https://doi.org/10.1073/pnas.1322098111.

Nguyen, D.M. *et al.* (2006) 'Molecular Heterogeneity of Inflammatory Breast Cancer: A Hyperproliferative Phenotype', *Clinical Cancer Research*, 12(17), pp. 5047–5054. Dostupno na: https://doi.org/10.1158/1078-0432.CCR-05-2248.

Nurk, S. *et al.* (2022) 'The complete sequence of a human genome', *Science*, 376(6588), pp. 44–53. Dostupno na: https://doi.org/10.1126/science.abj6987.

Nygaard, V., Rødland, E.A. i Hovig, E. (2016) 'Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses', *Biostatistics*, 17(1), pp. 29–39. Dostupno na: https://doi.org/10.1093/biostatistics/kxv027.

Parker, J.S. *et al.* (2009) 'Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes', *Journal of Clinical Oncology*, 27(8), pp. 1160–1167. Dostupno na: https://doi.org/10.1200/JCO.2008.18.1370.

Pecorino, L. (2012) *Molecular biology of cancer: mechanisms, targets, and therapeutics*. Treće izdanje. Oxford: Oxford University Press.

Price, E.M. i Robinson, W.P. (2018) 'Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned', *Frontiers in Genetics*, 9. Dostupno na: https://doi.org/10.3389/fgene.2018.00083.

R Core Team (2024) 'R: A Language and Environment for Statistical Computing'. Vienna, Austria: R Foundation for Statistical Computing.

Revkov, E. *et al.* (2023) 'PUREE: accurate pan-cancer tumor purity estimation from gene expression data', *Communications Biology*, 6(1), pp. 1–10. Dostupno na: https://doi.org/10.1038/s42003-023-04764-8.

de Ridder, D. *et al.* (2005) 'Purity for clarity: the need for purification of tumor cells in DNA microarray studies', *Leukemia*, 19(4), pp. 618–627. Dostupno na: https://doi.org/10.1038/sj.leu.2403685.

Robertson, F.M. *et al.* (2010) 'Inflammatory Breast Cancer: The Disease, the Biology, the Treatment', *CA: A Cancer Journal for Clinicians*, 60(6), pp. 351–375. Dostupno na: https://doi.org/10.3322/caac.20082.

Rousseeuw, P.J. (1987) 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, 20, pp. 53–65. Dostupno na: https://doi.org/10.1016/0377-0427(87)90125-7.

Rouzier, R. *et al.* (2005) 'Breast Cancer Molecular Subtypes Respond Differently to Preoperative Chemotherapy', *Clinical Cancer Research*, 11(16), pp. 5678–5685. Dostupno na: https://doi.org/10.1158/1078-0432.CCR-04-2421.

Schapire, R.E. (2013) 'Explaining AdaBoost', u B. Schölkopf, Z. Luo, i V. Vovk (eds) *Empirical Inference*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 37–52. Dostupno na: https://doi.org/10.1007/978-3-642-41136-6_5.

Schapire, R.E. i Freund, Y. (2014) *Boosting: foundations and algorithms*. Cambridge, Mass.: MIT Press (Adaptive computation and machine learning).

Sircoulomb, F. *et al.* (2010) 'Genome profiling of ERBB2-amplified breast cancers', *BMC cancer*, 10, p. 539. Dostupno na: https://doi.org/10.1186/1471-2407-10-539.

Smid, M. *et al.* (2008) 'Subtypes of Breast Cancer Show Preferential Site of Relapse', *Cancer Research*, 68(9), pp. 3108–3114. Dostupno na: https://doi.org/10.1158/0008-5472.CAN-07-5644.

Sørlie, T. *et al.* (2003) 'Repeated observation of breast tumor subtypes in independent gene expression data sets', *Proceedings of the National Academy of Sciences*, 100(14), pp. 8418–8423. Dostupno na: https://doi.org/10.1073/pnas.0932692100.

Stover, D.G. i Wagle, N. (2015) 'Precision Medicine in Breast Cancer: Genes, Genomes, and the Future of Genomically-Driven Treatments', *Current oncology reports*, 17(4), p. 15. Dostupno na: https://doi.org/10.1007/s11912-015-0438-0.

Thompson, J.A., Tan, J. i Greene, C.S. (2016) 'Cross-platform normalization of microarray and RNA-seq data for machine learning applications', *PeerJ*, 4, p. e1621. Dostupno na: https://doi.org/10.7717/peerj.1621.

Tibshirani, R. *et al.* (2002) 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', *Proceedings of the National Academy of Sciences*, 99(10), pp. 6567–6572. Dostupno na: https://doi.org/10.1073/pnas.082099299.

Van der Auwera, I. *et al.* (2004) 'Increased Angiogenesis and Lymphangiogenesis in Inflammatory versus Noninflammatory Breast Cancer by Real-Time Reverse Transcriptase-PCR Gene Expression Quantification', *Clinical Cancer Research*, 10(23), pp. 7965–7971. Dostupno na: https://doi.org/10.1158/1078-0432.CCR-04-0063.

Van Laere, S. *et al.* (2005) 'Distinct Molecular Signature of Inflammatory Breast Cancer by cDNA Microarray Analysis', *Breast Cancer Research and Treatment*, 93(3), pp. 237–246. Dostupno na: https://doi.org/10.1007/s10549-005-5157-z.

Van Laere, S. *et al.* (2007) 'Distinct molecular phenotype of inflammatory breast cancer compared to non-inflammatory breast cancer using Affymetrix-based genome-wide gene-expression analysis', *British Journal of Cancer*, 97(8), pp. 1165–1174. Dostupno na: https://doi.org/10.1038/sj.bjc.6603967.

Van Laere, S.J. *et al.* (2013) 'Uncovering the Molecular Secrets of Inflammatory Breast Cancer Biology: An Integrated Analysis of Three Distinct Affymetrix Gene Expression Datasets', *Clinical Cancer Research*, 19(17), pp. 4685–4696. Dostupno na: https://doi.org/10.1158/1078-0432.CCR-12-2549.

Van Uden, D.J.P. *et al.* (2015) 'Inflammatory breast cancer: An overview', *Critical Reviews in Oncology/Hematology*, 93(2), pp. 116–126. Dostupno na: https://doi.org/10.1016/j.critrevonc.2014.09.003.

Walsh, C.J. *et al.* (2015) 'Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery', *Microarrays*, 4(3), pp. 389–406. Dostupno na: https://doi.org/10.3390/microarrays4030389.

Walsh, E.M. i Halushka, M.K. (2023) 'A Comparison of Tissue Dissection Techniques for Diagnostic, Prognostic, and Theragnostic Analysis of Human Disease', *Pathobiology*, 90(3), pp. 199–208. Dostupno na: https://doi.org/10.1159/000525979.

Wickham, H. (2016) 'ggplot2: Elegant Graphics for Data Analysis'. Dostupno na: https://ggplot2.tidyverse.org.

Woodward, W.A. *et al.* (2013) 'Genomic and expression analysis of microdissected inflammatory breast cancer', *Breast Cancer Research and Treatment*, 138(3), pp. 761–772. Dostupno na: https://doi.org/10.1007/s10549-013-2501-6.

Wu, T. *et al.* (2021) 'clusterProfiler 4.0: A universal enrichment tool for interpreting omics data', *The Innovation*, 2(3). Dostupno na: https://doi.org/10.1016/j.xinn.2021.100141.

Yoshihara, K. *et al.* (2013) 'Inferring tumour purity and stromal and immune cell admixture from expression data', *Nature Communications*, 4(1), p. 2612. Dostupno na: https://doi.org/10.1038/ncomms3612.

Zare, A., Postovit, L.-M. and Githaka, J.M. (2021) ‘Robust inflammatory breast cancer gene signature using nonparametric random forest analysis’, *Breast Cancer Research*, 23(1), p. 92. Dostupno na: https://doi.org/10.1186/s13058-021-01467-y.

Zeng, Q. *et al.* (2019) ‘Synaptic proximity enables NMDAR signalling to promote brain metastasis’, *Nature*, 573(7775), pp. 526–531. Dostupno na: https://doi.org/10.1038/s41586-019-1576-6.

Zindler, T. *et al.* (2020) ‘Simulating ComBat: how batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies’, *BMC Bioinformatics*, 21(1), p. 271. Dostupno na: https://doi.org/10.1186/s12859-020-03559-6.

9. Sažetak

Lana Semenić

Integracija podataka ekspresijskih mikročipova za istraživanje genskog potpisa upalnog raka dojke strojnim učenjem

Upalni rak dojke (IBC) je agresivan, rijedak i slabo istražen tip raka dojke. Jedan od glavnih izazova s kojim se istraživači suočavaju je pronalazak njegovog genskog potpisa, tj. skupa gena po čijoj se ekspresiji IBC razlikuje od neupalnog raka dojke (nIBC-a). Takav potpis mogao bi omogućiti bolju dijagnostiku ili čak ciljane terapije specifične za IBC. Jedna od glavnih prepreka u mnogim takvim istraživanja je nedovoljan broj uzoraka. Kako bi se riješio ovaj problem, tri javno dostupna skupa podataka dobivena ekspresijskim mikročipovima podvrgnuta su normalizaciji po platformama kako bi se spojili u jedinstveni skup podataka od 146 uzoraka (54 IBC i 92 nIBC). Uzorci iz ovog skupa podataka zatim su korišteni za treniranje i validaciju *XGBoost* modela nadziranog učenja koji je na temelju ekspresije gena određivao je li rak dojke upalni ili ne. Istrenirana su ukupno četiri modela – jedan neovisan o podtipu raka dojke i tri specifična za podtip. Među njima je najuspješniji bio model specifičan za bazalni podtip, točnosti od 88%. On je nešto manje točan od postojećeg modela temeljenog na G59 potpisu, koji je neovisan o podtipu. Međutim, treba istaknuti da je model za bazalni podtip bio uspješan na vrlo heterogenom skupu podataka, dok je G59 testiran na jednostavnijima, gdje uzorci potječu iz istog eksperimenta. Mane ovog pristupa bile su veliki gubitak gena tijekom pretprocesiranja, kao i upitna čistoća tumorskih uzoraka iz dva skupa podataka. Unatoč tome, ova analiza pokazuje da postoji temelj za istraživanje genskih potpisa ovisnih o podtipu normalizacijom po platformama, pri čemu se preporučuje korištenje skupova podataka veće kvalitete ili manje kompleksnosti (npr. ekspresijskih podataka s mikročipova iste platforme).

Ključne riječi: normalizacija po platformama, transkriptomika, *XGBoost*

10. Summary

Lana Semenić

Integrating Microarray Data for Machine Learning Analysis of Inflammatory Breast Cancer Gene Signature

Inflammatory breast cancer (IBC) is an aggressive, rare and poorly understood type of breast cancer. One of the main challenges researchers face is defining its gene signature, i.e., a set of genes that differentiates IBC from non-inflammatory breast cancer (nIBC). Such a signature could allow for more accurate diagnosis or even IBC-specific targeted therapies. A major obstacle in many of these studies is their small sample size. To address this issue, cross-platform normalization was carried out on three publicly available expression microarray datasets in order to integrate them into a single dataset comprising 146 samples (54 IBC and 92 nIBC). Samples from this dataset were then used to train and validate *XGBoost* supervised learning models to determine whether a cancer was inflammatory or not, based on gene expression. A total of four models were trained – one that was independent of breast cancer subtypes and three that were subtype-specific. Among them, the basal subtype-specific model was the most successful (88% accuracy), albeit less so than the existing G59 signature-based, subtype-independent model. However, it should be noted that the basal subtype-specific model performed well on a very heterogeneous dataset, unlike the single-experiment datasets the G59-based model was tested on. The main impediments to this approach were the large gene loss that occurred during preprocessing, as well as questionable tumor purity in samples derived from multiple data sources. Nevertheless, this analysis highlights that there are grounds for researching breast cancer subtype-specific signatures by performing cross-platform normalization, however, it is recommended to use datasets of higher quality and lower complexity (e.g., those that share a microarray platform).

Keywords: cross-platform normalization, transcriptomics, *XGBoost*

11. Životopis

Rođena sam 2.6.2000. u Zaboku. Završila sam XV. gimnaziju u Zagrebu (2019) i preddiplomski studij molekularne biologije na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu (2022). Na istom fakultetu pohađam diplomski studij molekularne biologije, modul računalna biologija, te u sklopu diplomskog rada istražujem molekulsku dinamiku bakterijskih manganovih metalosenzora. Laboratorijsku praksu odrađivala sam na Institutu Ruđer Bošković (2022) i Prirodoslovno-matematičkom fakultetu (2023, 2024). Tijekom studija sam sudjelovala u pripremama učenika XV. gimnazije za Državno natjecanje iz biologije te popularno-znanstvenim manifestacijama *Noć muzeja* i *Dan i noć na PMF-u*.