

Meta-analiza transkriptomskih podataka i razvoj modela strojnoga učenja za detekciju i klasifikaciju karcinoma štitnjače

Anadolac, Alisa

Master's thesis / Diplomski rad (sveučilišni)

2026

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Split, Faculty of Science / Sveučilište u Splitu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: <https://um.nsk.hr/um:nbn:hr:166:107787>

Rights / Prava: [Zaštićeno autorskim pravom.](#) / [In copyright.](#)

Download date / Datum preuzimanja: **2026-09-07**



Repository / Repozitorij:

[Digital Repository of Faculty of Science in Split / Digitalni repozitorij Prirodoslovno-matematičkog fakulteta u Splitu](#)



Dabar

Sveučilište u Splitu
Prirodoslovno-matematički fakultet
Odjel za biologiju

Alisa Anadolac

**META-ANALIZA TRANSKRIPTOMSKIH
PODATAKA I RAZVOJ MODELA
STROJNOGA UČENJA ZA DETEKCIJU I
KLASIFIKACIJU KARCINOMA ŠTITNJAČE**

Diplomski rad

Split, 2026.

Sveučilište u Splitu
Prirodoslovno-matematički fakultet
Odjel za biologiju

Alisa Anadolac

**META-ANALIZA TRANSKRIPTOMSKIH
PODATAKA I RAZVOJ MODELA
STROJNOGA UČENJA ZA DETEKCIJU I
KLASIFIKACIJU KARCINOMA ŠTITNJAČE**

Diplomski rad

Split, 2026.

Ovaj rad, izrađen u Splitu 2026. godine pod mentorstvom nasl. doc. dr. sc. Andree Gelemanović i komentorstvom prof. dr. sc. Ane Maravić, u sklopu Obzor Europa projekta „InnoThyroGen – Innovating Thyroid Health with Genomics and Predictive Algorithms through Collaborative Excellence“ (šifra projekta 101187880), predan je na ocjenu Odjelu za biologiju Prirodoslovno-matematičkog fakulteta Sveučilišta u Splitu radi stjecanja zvanja magistre molekularne biologije (mag. biol. mol.)

IZJAVA

kojom izjavljujem s punom materijalnom i moralnom odgovornošću da sam diplomski rad s naslovom „Meta-analiza transkriptomskih podataka i razvoj modela strojnoga učenja za detekciju i klasifikaciju karcinoma štitnjače“ izradila samostalno pod mentorstvom nasl. doc. dr. sc. Andree Gelemanović i komentorstvom prof. dr. sc. Ane Maravić. U radu je primijenjena metodologija znanstveno istraživačkog rada i korištena je literatura koja je navedena na kraju diplomskog rada. Tuđe spoznaje, stavove, zaključke, teorije i zakonitosti koje su izravno ili parafrazirajući navedeni u završnom radu na uobičajen, standardan način citirani su i povezani s korištenim bibliografskim jedinicama. Rad je pisan u duhu hrvatskog jezika.

Studentica

Alisa Anadolac

Hvala voljenim roditeljima, posebice mojoj majci jer je vjerovala i ustrajala uz mene sve ovo vrijeme.

Hvala bratu, mojoj zvijezdi vodilji, za sugestije i savjete za sve aktivnosti mimo fakulteta koje su me oblikovale danas kao osobu.

Neizmjereno hvala divnoj mentorici koja je uvijek imala riječi ohrabrenja i podrške u trenucima obostranog kaosa.

Hvala predivnim prijateljima, bivšim i sadašnjim, mozaik sam satkan od vaše ljubavi.

Hvala sebi na hrabrosti.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Splitu

Diplomski rad

Prirodoslovno-matematički fakultet

Odjel za biologiju

Ruđera Boškovića 33, 21000 Split, Hrvatska

Meta-analiza transkriptomskih podataka i razvoj modela strojnoga učenja za detekciju i klasifikaciju karcinoma štitnjače

Alisa Anadolac

U ovom je radu integrativnom bioinformatičkom analizom javno dostupnih RNA-seq podataka istražen transkriptomski profil tumorskog (T), zdravog (H) i tumoru susjednog normalnog tkiva (NAT) štitnjače primjenom pojedinačnih analiza diferencijalne ekspresije, meta-analize, rane integracije i strojnog učenja. Usporedbe T prema H i T prema NAT pokazale su da se NAT transkriptomski razlikuje od neovisnog zdravog tkiva te da izbor kontrole utječe na broj i sastav identificiranih gena. Meta-analizom je izdvojen zajednički tumorski potpis, a funkcionalna analiza istaknula je procese povezane s izvanstaničnim matriksom, staničnom adhezijom, kolagenom i epitelnom diferencijacijom. Razvijeni modeli uspješno su razlikovali tumorske od kontrolnih uzoraka. Dobiveni rezultati omogućili su izdvajanje ponovljivih transkriptomskih značajki papilarnog karcinoma štitnjače i pokazali važnost vrste kontrolnog tkiva za njihovu interpretaciju.

Ključne riječi: papilarni karcinom štitnjače, transkriptomika, strojno učenje

Rad sadrži: 71 stranicu, 36 slika, 1 tablicu, 60 literaturnih navoda, originalni jezik: hrvatski

Mentor: nasl. doc. dr. sc. Andrea Gelemanović

Komentor: prof. dr. sc. Ana Maravić

Povjerenstvo: nasl. doc. dr. sc. Andrea Gelemanović

prof. dr. sc. Ana Maravić

doc. dr. sc. Ivica Šamanić

Datum obrane: 26. kolovoza 2026.

Rad je napisan na hrvatskom jeziku i pohranjen u repozitoriju Prirodoslovno-matematičkog fakulteta u Splitu, koji je dio nacionalnog repozitorijskog sustava Dabar.

BASIC DOCUMENTATION CARD

University of Split

Master Thesis

Faculty of Science

Department of Biology

Ruđera Boškovića 33, 21000 Split, Croatia

Meta-analysis of transcriptomic data and development of a machine learning model for thyroid cancer detection and classification

Alisa Anadolac

In this study, an integrative bioinformatic analysis of publicly available RNA-seq data was used to investigate the transcriptomic profiles of tumor (T), healthy (H), and normal adjacent-to-tumor (NAT) thyroid tissues through individual differential expression analyses, meta-analysis, early data integration, and machine learning. Comparisons of T vs. H and T vs. NAT showed that NAT is transcriptionally distinct from independent healthy tissue and that the choice of control tissue affects the number and composition of identified genes. Meta-analysis revealed a shared tumor-associated transcriptomic signature, while functional enrichment analysis highlighted processes related to the extracellular matrix, cell adhesion, collagen, and epithelial differentiation. The developed models successfully distinguished tumor from control samples. These results enabled the identification of reproducible transcriptomic features of papillary thyroid carcinoma and demonstrated the importance of control tissue selection for their interpretation.

Key words: papillary thyroid carcinoma, transcriptomics, machine learning

Thesis consists of: 71 pages, 36 figures, 1 table, 60 references; original language: Croatian

Thesis Supervisor: Assist. Prof. Andrea Gelemanović, PhD

Thesis Cosupervisor: Prof. Ana Maravić, PhD

Thesis Evaluation and Defense Committee: Assist. Prof. Andrea Gelemanović, PhD

Prof. Ana Maravić, PhD

Assist. Prof. Ivica Šamanić, PhD

Date of defence: 26th August 2026

The thesis is written in Croatian and deposited in the repository of the University of Split, Faculty of Science, which is part of the national repository system, Dabar.

SADRŽAJ

1.	UVOD	1
1.1	Papilarni karcinom štitnjače	2
1.2	Transkriptomika kao alat za molekularnu karakterizaciju karcinoma	3
1.3	Važnost odabira kontrolnog tkiva	6
1.4	Analiza diferencijalne ekspresije gena	7
1.5	Javno dostupni repozitoriji	9
1.5.1	Gene Expression Omnibus (GEO)	9
1.5.2	The Cancer Genome Atlas (TCGA)	10
1.5.3	Genotype-Tissue Expression (GTEx)	11
2.	CILJ ISTRAŽIVANJA	13
3.	MATERIJALI I METODE	14
3.1	Izvori i odabir podataka	14
3.2	Računalno okruženje i programski paketi	15
3.3	Obrada i usklađivanje podataka	16
3.4	Kontrola kvalitete podataka	16
3.5	Analiza diferencijalne ekspresije gena	18
3.6	Grafička interpretacija diferencijalne ekspresije gena	18
3.7	Meta-analiza pojedinačnih RNA-seq skupova podataka	19
3.8	Funkcionalna analiza obogaćenja	20
3.9	Rana integracija i objedinjavanje pojedinačnih RNA-seq podataka	21
3.10	Izrada modela strojnog učenja	22
4.	REZULTATI	23
4.1	Skup podataka GSE165724	23
4.2	Skup podataka GSE213647	28
4.3	TCGA/GTEx	30
4.4	Meta-analiza i preklapanje rezultata pojedinačnih studija	36
4.5	Funkcionalna analiza obogaćenja	39
4.5.1	Usporedba T naspram H	40
4.5.2	Usporedba T naspram NAT	42
4.6	Rana integracija RNA-seq skupova podataka	45
4.6.1	Rana integracija za usporedbu tumorskog i zdravog tkiva	46

4.6.2	Rana integracija za usporedbu tumorskog i NAT tkiva	47
4.7	Modeli strojnog učenja	49
4.8	Usporedba rezultata meta-analize i strojnog učenja	52
5.	RASPRAVA	57
6.	ZAKLJUČAK	65
7.	LITERATURA	66

1. UVOD

Štitnjača je endokrina žlijezda smještena u prednjem dijelu vrata, čija je osnovna uloga proizvodnja hormona tiroksina (T4) i trijodtironina (T3). Ti hormoni sudjeluju u regulaciji metabolizma, potrošnje energije, termogeneze, rasta i razvoja te utječu na funkciju gotovo svih organskih sustava. Tkivo štitnjače većinski je građeno od folikularnih stanica organiziranih u folikule, iz kojih potječe većina benignih i malignih novotvorina žlijezde [1]. Čvorovi štitnjače vrlo su česti u općoj populaciji, osobito u žena i starijih osoba, a većina ih je benigna. Njihova procjena obuhvaća klinički pregled, određivanje koncentracije tireostimulirajućeg hormona, ultrazvučni pregled te, kada je indicirano, punkciju tankom iglom. Ultrazvukom se procjenjuju obilježja povezana s rizikom od zloćudnosti, dok citološka analiza aspiriranog materijala omogućuje daljnju klasifikaciju čvora i odabir odgovarajućega načina liječenja ili praćenja [2].

Karcinom štitnjače danas je najčešći maligni tumor endokrinog sustava i njegova učestalost u populaciji stalno raste. Iako ovakvi pacijenti uglavnom imaju dobru prognozu i uspješno se liječe operativnim zahvatom i radioaktivnim jodom, u kliničkoj dijagnostici i dalje postoje ozbiljni izazovi. Glavni problem u svakodnevnoj praksi jest pouzdano razlikovanje dobroćudnih čvorova štitnjače od malignih tumora. Kada citološki nalaz dobiven punkcijom tankom iglom ostane neodređen, liječnici su često primorani poslati pacijenta na operaciju kako bi bili sigurni u dijagnostiku. To nerijetko dovodi do nepotrebnih kirurških zahvata i može odgoditi pravovremeno liječenje pacijenata koji doista imaju tumor.

Razvoj tehnologija sekvenciranja sljedeće generacije (engl. *next-generation sequencing*, NGS) poput sekvenciranja RNA (RNA-seq), ali i starije tehnologije genskih čipova (engl. *microarrays*), promijenilo je način na koji proučavamo tumore. Za razliku od sekvenciranja genoma koje pruža relativno statičan uvid u genomski profil pojedinca, transkriptomika nam omogućuje da u realnom vremenu pratimo koji su geni i u kojoj mjeri aktivni u stanici. Budući da takva istraživanja stvaraju ogromnu količinu podataka, suvremena se biomedicina sve više oslanja na bioinformatiku i metode strojnog učenja kako bi se prepoznali uzorci u transkriptomskim podacima te pronašli novi dijagnostički biomarkeri.

Glavni cilj ovog diplomskog rada jest razviti i evaluirati modele strojnog učenja za klasifikaciju karcinoma štitnjače u odnosu na kontrolne uzorke na temelju isključivo transkriptomskih podataka te identificirati potencijalne biomarkere. Kako bi se pojačala snaga studije i dobili što pouzdaniji rezultati, u ovom radu će se provesti rana i kasna (u obliku meta-analize) integracija nekoliko neovisnih skupova podataka iz javno dostupnih repozitorija.

Poseban naglasak stavljen je na metodološko pitanje o kojem se u literaturi često ne vodi dovoljno računa, a to je odabir kontrolnog tkiva. Stoga će se u radu usporediti rezultati dobiveni korištenjem tkiva uzetog neposredno uz tumor (engl. *normal adjacent to tumor*, NAT) s rezultatima dobivenim korištenjem tkiva potpuno zdravih osoba, kako bi se utvrdilo koliko taj odabir utječe na same performanse modela i konačni izbor najvažnijih gena.

1.1 Papilarni karcinom štitnjače

Papilarni karcinom štitnjače (eng. *papillary thyroid carcinoma*, PTC) je najčešći podtip karcinoma štitnjače s učestalošću od 80 %. Karcinom nastaje iz folikularnih epitelnih stanica te pripada skupini diferenciranih tumora. Osim papilarnog karcinoma štitnjače opisano je nekoliko drugih varijanti: obzirom na veličinu (papilarni), prirodu granica tumora (enkapsulirani), arhitekturu (folikularni), stanične karakteristike (hobnail), dodatne komponente tumora i stromalne značajke (papilarni karcinom sa stromom sličnom fasciitisu), te zasebna varijanta koju karakteriziraju visoko nediferencirane stanice (anaplastični karcinom) [3], [4]. Većina varijanti, kao što su npr. folikularni ili papilarni karcinom štitnjače, povezani su s pozitivnom stopom preživljavanja, no izuzetno rijetka varijata, anaplastični karcinom štitnjače, gotovo uvijek završava sa smrtnim slučajem [5]. Jedna od ključnih značajki papilarnog karcinoma štitnjače je njegova sposobnost prodiranja u susjedne strukture poput limfnih čvorova što povećava incidenciju recidiva, no većina oboljelih ima pozitivnu prognozu [6]. PTC se može pojaviti u bilo kojoj dobi, no generalno se otkriva u trećem do petom desetljeću života pacijenata, s prosječnom dobi od 40 godina. Incidencija PTC-a raste s godinama, a žene su češće pogođena skupina nego muškarci, u omjerima od 2:1 do 4:1 [7]. Glavni čimbenik rizika papilarnog karcinoma štitnjače je ionizirajuće zračenje, naročito kod djece, no na razvoj bolesti također utječu nasljedni čimbenici poput prethodne obiteljske anamneze i nasljednih sindroma, ali i razni okolišni čimbenici [7].

S molekularne strane, papilarni karcinom štitnjače nastaje zbog manjeg broja genetičkih promjena koje dovode do poremećaja u signalnim putevima. Takve genetičke promjene aktiviraju MAPK, mitogenom aktiviranu protein kinazu, ključni signalni put za regulaciju proliferacije, diferencijacije te preživljavanje stanica. Najčešća mutacija je *BRAF* V600E, koja se ujedno koristi i kao glavni biomarker za determinaciju bolesti, dok se u značajnom dijelu tumora pojavljuju mutacije gena *NRAS*, *HRAS* i *KRAS*, kao i genske fuzije koje uključuju *RET* i *NTRK*. Na osnovu molekularnih obilježja, papilarni karcinom štitnjače

dijelimo u dvije glavne skupine: *BRAF* slični tumori i *RAS* slični tumori. *BRAF* slični tumori su tumori koji obuhvaćaju prethodno spomenutu *BRAF* V600E mutaciju, a očituju se u izraženijoj aktivaciji MAPK signalnog puta. Posljedično, dolazi do smanjenja ekspresije gena zaduženih za normalnu funkciju štitnjače, sintezu hormona štitnjače i unos joda. *RAS* slični tumori s druge strane zadržavaju viši stupanj diferencijacije, odnosno tumorske stanice u većoj mjeri zadržavaju molekularna obilježja normalnih folikularnih stanica štitnjače. Kod *RAS* sličnih tumora očuvana je ekspresija gena povezanih s normalnom funkcijom štitnjače, primjerice gena uključenih u unos joda i sintezu hormona štitnjače. Zaključno, karcinom štitnjače ne može se promatrati kao potpuno homogena bolest, nego kao skupina tumora koji dijele osnovne morfološke značajke, ali se razlikuju prema histološkim, molekularnim i prognostičkim obilježjima [8].

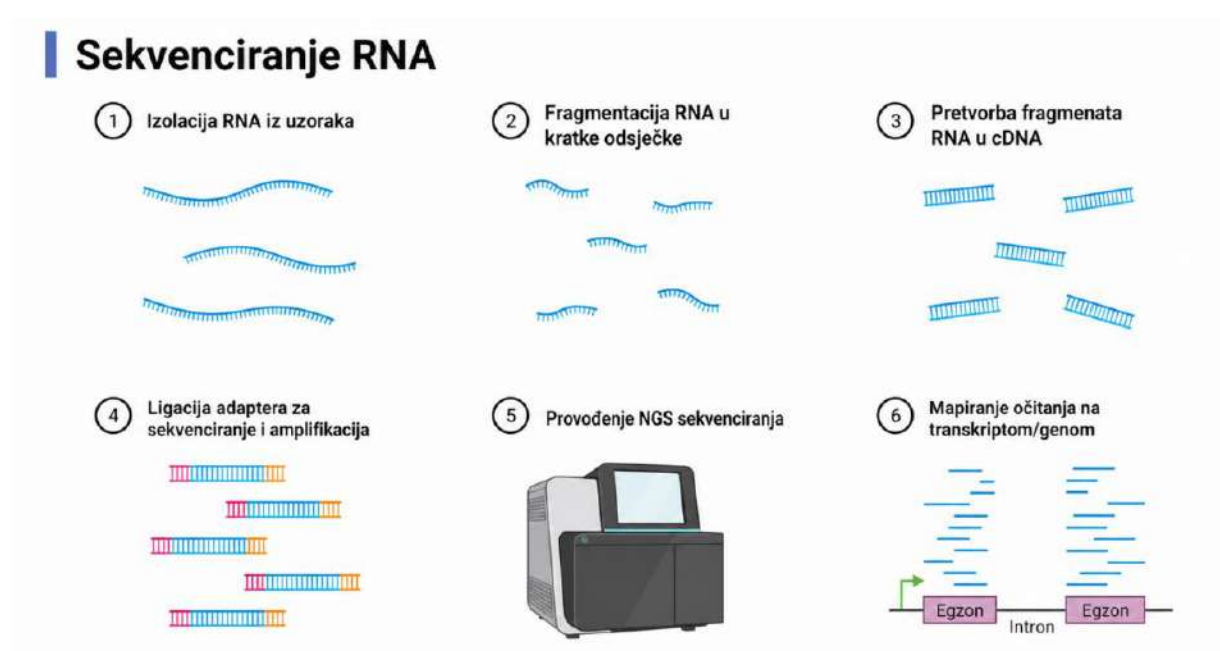
1.2 Transkriptomika kao alat za molekularnu karakterizaciju karcinoma

Transkriptom je skup svih RNA molekula koje su u određenom trenutku prisutne u stanici ili tkivu. Za razliku od genoma, koji je relativno stabilan, transkriptom se mijenja ovisno o vrsti stanice, razvojnem i fiziološkom stanju, okolišnim čimbenicima i prisutnosti bolesti. Transkriptomska analiza stoga omogućuje određivanje aktivnih gena i prepoznavanje promjena u njihovoj ekspresiji između različitih bioloških skupina, primjerice tumorskog i netumorskog tkiva [9].

Sekvenciranje svih RNA molekula, odnosno RNA-seq, visokoprotlačna je metoda kojom se kvalitativno i kvantitativno analizira RNA prisutna u uzorku [10], [11]. Postupak najčešće započinje izolacijom ukupne RNA, nakon čega se provodi izdvajanje mRNA ili uklanjanje ribosomske RNA. RNA se zatim prevodi u komplementarnu DNA, odnosno cDNA, koja se priprema u obliku knjižnice i sekvencira. Dobivena sekvencijska očitavanja prolaze kontrolu kvalitete, poravnavaju se s referentnim genomom ili transkriptomom te se pridružuju odgovarajućim genima. Broj očitavanja pridružen pojedinom genu predstavlja osnovu za procjenu njegove ekspresije (**Slika 1.**). Osim kvantifikacije poznatih gena, RNA-seq može omogućiti otkrivanje novih transkripata, alternativnih izoformi i fuzijskih transkripata [12].

U istraživanju tumora transkriptomski podaci nadopunjuju rezultate genomske analize jer omogućuju povezivanje promjena na razini DNA s njihovim transkripcijskim posljedicama. Dok se analizom DNA otkrivaju somatske mutacije, promjene broja kopija i strukturne

varijante, RNA-seq omogućuje kvantifikaciju ekspresije gena te otkrivanje promijenjenog prekrajanja RNA, fuzijskih transkripata i alelno specifične ekspresije [13], [14]. Analizom različito izraženih gena i obogaćenosti bioloških puteva mogu se istražiti molekularni procesi povezani s proliferacijom, izbjegavanjem stanične smrti, promijenjenim metabolizmom, invazijom, diferencijacijom i odgovorom tumora na terapiju [13], [15]. Transkriptomsko profiliranje može se primijeniti i za molekularnu klasifikaciju tumora, otkrivanje dijagnostičkih i prognostičkih biomarkera, predviđanje terapijskog odgovora te pronalaženje potencijalnih terapijskih meta [13], [16], [17].



Slika 1. Shematski prikaz postupka RNA sekvenciranja [18].

Konvencionalno sekvenciranje RNA, odnosno tzv. *bulk* RNA-seq, provodi se na ukupnoj RNA izoliranoj iz velikog broja stanica te daje agregirani profil ekspresije svih staničnih populacija prisutnih u analiziranom uzorku [11], [19]. Izmjereni signal stoga ovisi o udjelu pojedinih staničnih populacija, ali i o količini RNA koju te populacije sadržavaju pa ne predstavlja ekspresiju prosječne pojedinačne stanice [20], [21]. U usporedbi sa sekvenciranjem RNA na razini pojedinačnih stanica, *bulk* RNA-seq tehnički je jednostavniji, financijski pristupačniji i prikladniji za obradu većeg broja uzoraka. Dodatne su prednosti dobro uspostavljeni laboratorijski protokoli, razvijeni bioinformatički postupci i pouzdana kvantifikacija ekspresije gena na razini cijelog uzorka [11], [19], [22]. Jedan od najvažnijih postupaka u obradi *bulk* RNA-seq podataka jest analiza diferencijalne ekspresije. Njome se

određuju geni čija se razina ekspresije statistički značajno razlikuje između ispitivanih skupina. Analiza obuhvaća uklanjanje izrazito slabo eksprimiranih gena, normalizaciju broja očitavanja, procjenu biološke varijabilnosti i statističko testiranje hipoteze postoje li razlike između ispitivanih skupina uzoraka. Među najčešće korištenim programskim alatima nalaze se R paketi *limma*, *edgeR* i *DESeq2*. *Limma* se temelji na linearnim modelima, dok *edgeR* i *DESeq2* podatke o broju očitavanja modeliraju negativnom binomnom distribucijom. Odabir odgovarajućeg statističkog postupka, pravilna normalizacija i uključivanje relevantnih bioloških i tehničkih čimbenika važni su za pouzdano određivanje diferencijalno eksprimiranih gena [10].

Razvoj sekvenciranja RNA pojedinačnih stanica (engl. *single-cell RNA sequencing*, scRNA-seq) omogućio je određivanje transkriptomskih profila pojedinačnih stanica unutar heterogenog tkiva [22]. Za razliku od *bulk* RNA-seq-a, kojim se dobiva agregirani signal svih stanica u uzorku, scRNA-seq omogućuje prepoznavanje rijetkih staničnih populacija, razlikovanje funkcionalnih stanja tumorskih stanica te detaljnije istraživanje heterogenosti tumorskog mikrookoliša [22], [23]. Na taj način razlikuju se maligne, stromalne, endotelne i imunološke stanice te se određuje razlika genske ekspresije između različitih staničnih populacija [23].

Rezultati scRNA-seq analize podložni su tehničkoj i biološkoj varijabilnosti. Na njih mogu utjecati način prikupljanja i disocijacije tkiva, kvaliteta i vijabilnost stanica, dubina sekvenciranja, učinci serije, rijetko bilježenje transkripata te izbor postupaka filtriranja, normalizacije, redukcije dimenzionalnosti i grupiranja stanica [11], [24], a sama bioinformatička analiza je kompleksnija i zahtjevnija. Budući da se stanične populacije najčešće određuju računalnim grupiranjem prema sličnosti ekspresijskih profila, dobiveni klasteri ne moraju nužno odgovarati biološki zasebnim staničnim tipovima. Zbog toga je njihovu stabilnost i ponovljivost potrebno procijeniti u biološkim replikatima, neovisnim uzorcima ili vanjskim skupovima podataka te ih, kada je moguće, potvrditi poznatim biljezima ili drugim eksperimentalnim metodama. Upravo je identifikacija i anotacija staničnih populacija najizazovniji korak u scRNA-seq analizi jer zahtijeva izvrsno poznavanje specifičnih molekularnih markera i biološkog konteksta proučavanog tkiva [23], [25].

1.3 Važnost odabira kontrolnog tkiva

U transkriptomskim istraživanjima tumora izbor odgovarajućega kontrolnog tkiva ključan je za pravilno određivanje promjena u ekspresiji gena. Ekspresija gena tumorskog tkiva najčešće se uspoređuje sa zdravim tkivom osoba bez dijagnosticirane maligne bolesti ili s histološki normalnim tkivom uzetim u neposrednoj blizini tumora. Iako se obje vrste uzoraka u literaturi klasificiraju kao normalno tkivo, one nisu biološki istovjetne [26]. U kontekstu ovog rada potpuno zdravo tkivo označeno je kraticom H (engl. *healthy*), dok se kratica NAT (engl. *normal adjacent to tumor*) odnosi na histološki normalno tkivo neposredno u blizini tumora.

Prednost NAT tkiva kao kontrolnog uzorka jest to što potječe od istoga bolesnika kao i uzorak tumorskog tkiva. Takvim pristupom mogu se smanjiti razlike povezane s genetskom pozadinom, dobi, spolom i drugim osobinama pojedinca za koje se zna da utječu na profil genske ekspresije. Tumorsko i susjedno tkivo često se prikupljaju tijekom istog kirurškog zahvata te prolaze slične postupke obrade i pohrane. Međutim, histološki normalan izgled ne znači nužno i molekularno normalno stanje. Tkivo koje okružuje tumor može biti izloženo signalnim molekulama koje izlučuju tumorske, stromalne i imunosne stanice, lokalnoj upali, hipoksiji, promjenama izvanstaničnog matriksa i drugim učincima tumorskoga mikrookoliša. U njemu također mogu biti prisutne rane molekularne promjene koje još nisu dovele do vidljivih histoloških obilježja tumora [27].

Dokaz o različitoj molekularnoj osnovi NAT-a opisali su Aran i suradnici koji su integrirali transkriptomске podatke 6 506 uzoraka iz projekata GTEx (engl. *The Genotype-Tissue Expression*) i TCGA (engl. *The Cancer Genome Atlas*), obuhvaćajući zdravo, tumorsko i tumoru susjedno normalno tkivo iz osam organskih sustava. Rezultati su pokazali da NAT ima zaseban ekspresijski profil koji se često nalazi između profila zdravog i tumorskog tkiva. U NAT-u je utvrđena aktivacija gena povezanih s upalnim odgovorom i stimulacijom endotelnih stanica, zbog čega su autori zaključili da tumor može utjecati na molekularna obilježja okolnog tkiva. NAT se stoga ne može smatrati međusobnom zamjenom za tkivo osoba bez tumora [26]. Usporedbe T nasuprot H, T nasuprot NAT i NAT nasuprot H pružaju različite, ali komplementarne biološke informacije. Njihovom zajedničkom analizom moguće je razlikovati promjene karakteristične za tumor od promjena prisutnih u njegovu neposrednom okolišu te preciznije procijeniti kako izbor kontrolnoga tkiva utječe na interpretaciju transkriptomskih obilježja papilarnog karcinoma štitnjače.

1.4 Analiza diferencijalne ekspresije gena

Analiza diferencijalne ekspresije gena statistički je postupak kojim se utvrđuje razlikuje li se razina ekspresije pojedinih gena između unaprijed definiranih bioloških skupina, primjerice tumorskog i kontrolnog tkiva [11], [28]. U istraživanju tumora takva analiza omogućuje prepoznavanje gena i molekularnih procesa povezanih s nastankom, napredovanjem i biološkim obilježjima tumora [11], [15]. Cilj analize nije samo izdvojiti statistički značajne gene, nego procijeniti smjer i veličinu promjene njihove ekspresije uz uvažavanje biološke i tehničke varijabilnosti uzoraka [29], [30].

Polaznu točku analize RNA-seq podataka čini matrica ekspresije u kojoj retci predstavljaju gene, stupci uzorke, a vrijednosti brojeve očitavanja pridružene pojedinom genu [29], [31]. Brojevi očitavanja mogu se dobiti poravnanjem očitavanja na referentni genom i njihovim pridruživanjem anotiranim genima ili kvantifikacijom na razini transkripata i naknadnim objedinjavanjem na razinu gena [32], [33], [34]. Odabir referentnog genoma, anotacije i postupka kvantifikacije može utjecati na dobivene vrijednosti ekspresije i konačan popis diferencijalno eksprimiranih gena [19]. Prije statističkog testiranja potrebno je provesti kontrolu kvalitete očitavanja i dobivene matrice ekspresije [19]. Ona obuhvaća provjeru kvalitete baza, uspješnosti poravnanja ili kvantifikacije, veličine knjižnica, raspodjele ekspresijskih vrijednosti i međusobne sličnosti uzoraka [35]. Analiza glavnih komponenti (engl. *principal component analysis*, PCA) često se primjenjuje za otkrivanje nedostatnih uzoraka, učinaka serije i povezanosti glavnih izvora varijabilnosti s ispitivanim biološkim skupinama [29]. Uzorke koji odstupaju ne treba automatski isključiti jer njihovo izdvajanje može biti posljedica tehničkoga problema, pogrešnih metapodataka ili stvarne biološke heterogenosti [19].

Geni s vrlo malim brojem očitavanja najčešće se uklanjaju prije statističke analize jer njihova varijabilnost i veličina promjene mogu narušiti rezultate analize [29]. Uključivanje velikoga broja slabo eksprimiranih gena povećava broj statističkih testova i može smanjiti statističku snagu za otkrivanje stvarnih promjena. Neovisno filtriranje na temelju ukupne razine ekspresije može povećati broj pouzdano identificiranih gena [36].

Sirovi brojevi očitavanja nisu neposredno usporedivi između uzoraka jer ovise o dubini sekvenciranja i sastavu RNA knjižnice. Normalizacijom se korigiraju tehničke razlike uz nastojanje da se očuvaju stvarne biološke promjene. Pritom *DESeq2* koristi faktore veličine procijenjene metodom medijana omjera, dok *edgeR* najčešće primjenjuje metodu TMM, odnosno skraćeni prosjek logaritama omjera ekspresije. Brojevi očitavanja diskretne su

vrijednosti kod kojih je varijanca često veća od srednje vrijednosti, zbog čega se najčešće modeliraju negativnom binomnom distribucijom. *DESeq2* za svaki gen procjenjuje srednju ekspresiju i disperziju koja opisuje varijabilnost među biološkim uzorcima [37]. Informacije se pritom dijele između gena kako bi procjene disperzije i promjene ekspresije bile stabilnije u istraživanjima s ograničenim brojem uzoraka. Statistički model treba oblikovati u skladu s eksperimentalnim dizajnom i dostupnim metapodacima, dakle osim vrste tkiva, u model se mogu uključiti dob, spol, laboratorijska serija i drugi čimbenici koji mogu utjecati na ekspresiju gena [29]. Ako tumorsko i tumoru susjedno normalno tkivo potječu od istoga bolesnika, identitet bolesnika može se uključiti u upareni model kako bi se odvojila varijabilnost među pojedincima od učinka vrste tkiva. Usporedba tumorskoga ili NAT tkiva s neovisnim zdravim tkivom ne može se upariti jer analizirani uzorci ne potječu od istih osoba [26].

Veličina promjene ekspresije najčešće se izražava kao logaritam omjera ekspresije s bazom 2, odnosno $\log_2 FC$ (engl. *log₂ fold change*). Vrijednost $\log_2 FC = 1$ odgovara približno dvostruko višoj ekspresiji u ispitivanoj skupini u odnosu na kontrolu, dok apsolutna vrijednost $\log_2 FC = 1,5$ odgovara promjeni od približno 2,83 puta višoj ekspresiji. Predznak vrijednosti ovisi o redoslijedu skupina koje se uspoređuju pa pozitivna vrijednost za usporedbu T prema H označava višu ekspresiju u tumorskom tkivu, dok bi negativna vrijednost označavala nižu ekspresiju u tumorskom tkivu u odnosu za zdravo [29]. Smjer usporedbe stoga treba jasno navesti za usporedbe T prema H, odnosno T prema NAT [26].

Za svaki se gen testira nulta hipoteza prema kojoj između uspoređivanih skupina nema razlike u ekspresiji. Budući da se istodobno testiraju tisuće gena, p-vrijednosti se najčešće prilagođavaju Benjamini-Hochberg postupkom radi kontrole stope lažnih otkrića. Prilagođena p-vrijednost opisuje statističku potporu rezultatu, dok $\log_2 FC$ opisuje smjer i veličinu promjene ekspresije [29]. Statističku i biološku značajnost zato je potrebno razmatrati zajedno jer mala promjena može biti statistički značajna u velikom skupu uzoraka [38]. Ne postoji univerzalna kombinacija pragova koja je obvezna za sva RNA-seq istraživanja jer njihov odabir ovisi o cilju analize i obilježjima podataka [19]. Kriteriji $padj \leq 0,01$ (engl. *adjusted p-value*, prilagođena p-vrijednost), $|\log_2 FC| \geq 1,5$ (apsolutna promjena $\log_2 FC$ kako bi se uzeli u obzir i povišeno i sniženo eksprimirani geni) i $baseMean \geq 20$ (prosječan broj normaliziranih očitavanja određenoga gena u svim uzorcima uključenim u *DESeq2* analizu) predstavljaju strogu analitičku odluku kojom se rezultati ograničavaju na statistički značajne, izraženije i dovoljno zastupljene promjene [29]. Stroži kriteriji povećavaju specifičnost eksprimiranih gena, ali mogu isključiti biološki važne gene s manjim ili varijabilnijim promjenama [19].

Rezultati analize diferencijalno eksprimiranih gena se najčešće prikazuju vulkanskim dijagramom (engl. *volcano plot*), na kojemu se uspoređuju veličina promjene i statistička značajnost svih analiziranih gena. PCA dijagram (engl. *principal component analysis*, analiza glavnih komponenti) prikazuje sličnosti među uzorcima, dok vulkanski dijagram prikazuje koji geni istodobno imaju izraženu promjenu i visoku statističku značajnost. Međutim, vulkanski dijagram ne prikazuje raspodjelu ekspresije pojedinog gena među uzorcima pa ga je potrebno tumačiti zajedno s normaliziranim vrijednostima i drugim grafičkim prikazima [39].

Na rezultate analize diferencijalno eksprimirane analize mogu utjecati veličina uzorka, broj bioloških replika, heterogenost tkiva, dubina sekvenciranja i tehnički učinci serija sekvenciranja. Statistički značajna diferencijalna ekspresija zato ne dokazuje uzročnu ulogu gena u nastanku tumora, nego predstavlja osnovu za funkcionalno tumačenje i potvrdu u neovisnim skupovima podataka ili drugim eksperimentalnim metodama [19].

1.5 Javno dostupni repozitoriji

Razvoj visokoprotocnog sekvenciranja doveo je do stvaranja velikih količina molekularnih podataka čija vrijednost nadilazi ciljeve pojedinačnih istraživanja [19]. Pohranjivanje podataka u javno dostupne repozitorije omogućuje njihovu ponovnu analizu, provjeru reproducibilnosti rezultata i usporedbu neovisnih skupina ispitanika [40]. Javno dostupni podatci osobito su važni u istraživanju tumora jer omogućuju analizu većeg broja uzoraka nego što bi ga najčešće bilo moguće prikupiti u okviru jednoga istraživanja. Među najvažnijim izvorima transkriptomskih podataka nalaze se repozitoriji *Gene Expression Omnibus* (GEO), *The Cancer Genome Atlas* (TCGA) i *Genotype-Tissue Expression* (GTEx).

1.5.1 Gene Expression Omnibus (GEO)

Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) međunarodni je javni repozitorij funkcionalnih genomskih podataka koji održava National Center for Biotechnology Information (NCBI), sastavnica američkoga National Library of Medicine. GEO pohranjuje podatke dobivene mikročipovima, sekvenciranjem RNA, ali i drugim metodama funkcionalne genomike i epigenomike. Prema podacima objavljenima 2024. godine, GEO je sadržavao više od 200 000 istraživanja i 6,5 milijuna pojedinačnih uzoraka [41].

Zapisi u GEO-u sadržavaju obrađene podatke, opis eksperimentalnog dizajna, karakteristike uzoraka te informacije o korištenoj platformi i postupcima obrade [41], [42].

Neobrađena očitavanja iz eksperimenata sekvenciranja najčešće su povezana sa zapisima u bazi *Sequence Read Archive* (<https://www.ncbi.nlm.nih.gov/sra>, SRA), dok se obrađene matrice ekspresije mogu preuzeti izravno iz GEO [41]. Dostupnost podataka i metapodataka omogućuje drugim istraživačima ponavljanje izvorne analize ili postavljanje novih istraživačkih pitanja koja nisu bila obuhvaćena izvornim radom.

Podatci su u GEO-u organizirani kao međusobno povezani zapisi koji opisuju istraživanje, uzorke i korištenu platformu [42]. Oznaka GSE (*GEO Series*) odnosi se na cjelokupno istraživanje, oznaka GSM (*GEO Sample*) na pojedinačni uzorak, a oznaka GPL (*GEO Platform*) na korištenu eksperimentalnu platformu [41], [42]. Takva organizacija omogućuje povezivanje ekspresijskih vrijednosti s pripadajućim biološkim obilježjima uzoraka i opisom eksperimentalnoga postupka [42]

GEO također sadržava internetski alat GEO2R, kojim se može provesti osnovna analiza diferencijalne ekspresije između korisnički definiranih skupina uzoraka. Za velik broj ljudskih *bulk* RNA-seq istraživanja dostupne su standardizirano izračunate matrice brojeva očitavanja, čime se olakšava ponovna analiza podataka programima kao što je *DESeq2* [41]. Ipak, prije ponovne analize potrebno je provjeriti značenje pojedinih skupina, dostupnost metapodataka i način na koji su uzorci prikupljeni i obrađeni [19], [41]. Velika raznolikost bolesti, tkiva, organizama i eksperimentalnih tehnologija glavna je prednost GEO-a, ali istodobno predstavlja važno analitičko ograničenje [41], [42]. Istraživanja se mogu razlikovati prema načinu prikupljanja tkiva, pripremi knjižnica, platformi, dubini sekvenciranja, referentnom genomu i anotaciji gena [19]. Izravno spajanje ekspresijskih matrica iz različitih istraživanja može zato dovesti do tehnički uvjetovanih razlika koje se pogrešno tumače kao biološki signal [43].

1.5.2 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (<https://www.cancer.gov/ccg/research/genome-sequencing/tcga>) bio je zajednički program ustanova National Cancer Institute i National Human Genome Research Institute usmjeren na sveobuhvatnu molekularnu karakterizaciju ljudskih tumora [44]. Programom je analizirano više od 20 000 tumorskih i odgovarajućih netumorskih uzoraka koji obuhvaćaju 33 skupine zloćudnih bolesti [45]. Uz podatke o ekspresiji gena, TCGA sadržava podatke o somatskim mutacijama, promjenama broja kopija, metilaciji DNA, mikroRNA i kliničkim obilježjima bolesnika [44].

Podatci TCGA danas su dostupni preko platforme Genomic Data Commons (<https://portal.gdc.cancer.gov/>, GDC), koja omogućuje pretraživanje, standardiziranu obradu i

preuzimanje genomskih i kliničkih podataka [46]. GDC obrađuje sekvencijske podatke standardiziranim bioinformatičkim postupcima i poravnava ih na zajedničku verziju referentnoga genoma, čime se povećava usporedivost podataka iz različitih projekata [46], [45]. Obradeni podatci i većina osnovnih metapodataka dostupni su bez posebne autorizacije, dok pristup osjetljivim individualnim genomskim podacima zahtijeva odgovarajuće odobrenje preko baze *The database of Genotypes and Phenotypes* (<https://dbgap.ncbi.nlm.nih.gov/home/>, dbGaP) [45], [46].

Za istraživanje papilarnoga karcinoma štitnjače posebno je važan projekt TCGA-THCA, koji sadržava molekularne i kliničke podatke o karcinomima štitnjače. Uz tumorsko tkivo, TCGA sadržava uzorke označene kao *Solid Tissue Normal*, koji najčešće predstavljaju histološki normalno tkivo prikupljeno u blizini tumora. Takvo tumoru susjedno normalno tkivo označava se kraticom NAT i ne treba ga automatski izjednačiti sa zdravim tkivom osoba koje nisu oboljele od tumora [26].

1.5.3 Genotype-Tissue Expression (GTEx)

Projekt *Genotype-Tissue Expression* (<https://gtexportal.org/home/>) razvijen je radi proučavanja ekspresije gena i njezine genetske regulacije u različitim netumorskim ljudskim tkivima. GTEx sadržava RNA-seq, genomske i pripadajuće metapodatke iz velikoga broja tkiva prikupljenih od gotovo 1 000 zdravih donora. Projekt omogućuje proučavanje tkivno specifične ekspresije, varijabilnosti među pojedincima i povezanosti genetskih varijanti s ekspresijom i prekrajanjem RNA [47].

GTEx je u istraživanju tumora koristan kao izvor netumorskih kontrolnih tkiva obzirom da potpuno zdrava tkiva nisu zastupljena u TCGA-u [26]. Otvoreni dostupni podatci uključuju obrađene ekspresijske matrice i dio anonimiziranih metapodataka, dok su neobrađena sekvencijska očitavanja i potpuni podatci o donorima zaštićeni [48]. Uzorci zdravoga tkiva iz GTEx-a su u ovom radu označeni kao skupina H (engl. *healthy*), dok netumorski uzorci iz TCGA-a predstavljaju skupinu NAT, čime se omogućuje zasebna usporedba zdravog, tumoru susjednog i tumorskog tkiva [26].

Podatci TCGA i GTEx međusobno se nadopunjuju, ali se ne bi trebali izravno uspoređivati bez odgovarajućeg usklađivanja bioinformatičke obrade podataka i primjene metoda za uklanjanje tehničkog šuma (tzv. *batch effect*) [26]. Razlike u izvoru uzoraka, vremenu do stabilizacije tkiva, pripremi knjižnica, platformama i izvornim postupcima kvantifikacije mogu uzrokovati učinke serije [43], [47]. Zbog toga je za zajedničku analizu

potrebno koristiti istu verziju referentnoga genoma, istu anotaciju gena i usporediv postupak kvantifikacije ekspresije [49].

Projekt *Toil* (<https://toil.ucsc-cgl.org/>) omogućio je ujednačenu obradu velikoga broja RNA-seq uzoraka iz TCGA-a i GTEx-a te stvaranje međusobno usporedivih ekspresijskih matrica [49]. Bioinformatička harmonizacija smanjuje tehničke razlike, ali ne uklanja biološke razlike povezane s donorima, načinom prikupljanja tkiva i sastavom uzoraka [26], [47]. Analiza zdravog, tumoru susjednog i tumorskog tkiva pokazala je da NAT često ima ekspresijski profil između zdravog i tumorskog tkiva, ovisno o organskom sustavu. Taj nalaz opravdava odvojene usporedbe T prema H te T prema NAT pri proučavanju transkriptomskih promjena u papilarnom karcinomu štitnjače [26].

2. CILJ ISTRAŽIVANJA

Glavni cilj rada je bioinformatičkom analizom transkriptomskih podataka i primjenom metoda strojnog učenja identificirati ključne transkriptomske biomarkere tumora štitnjače i procijeniti utjecaj izbora kontrolne skupine na rezultate diferencijalne ekspresije i klasifikacijsku preciznost modela. U radu su analizirani skupovi podataka koji sadrže tumorske uzorke štitnjače, zdravo tkivo štitnjače i/ili histološki normalno tkivo uz tumor. Time se omogućava usporedba rezultata dobivenih unutar pojedinačnih skupova podataka, ali i procjena njihove robusnosti kroz integracijske pristupe.

Specifični ciljevi ovog rada su:

- I. Identificirati diferencijalno eksprimirane gene u usporedbama tumorskih uzoraka s dvije različite kontrolne skupine: T prema H, odnosno tumor u odnosu na zdravo tkivo, te T prema NAT, odnosno tumor u odnosu na netumorsko, histološki normalno tkivo smješteno uz tumor.
- II. Provesti meta-analizu, odnosno kasnu integraciju rezultata diferencijalne ekspresije iz više skupova podataka čime se izdvajaju robusni geni koji se konzistentno pojavljuju kao diferencijalno eksprimirani u više neovisnih skupova podataka. Fokus je usmjeren na usporedbe gena dobivenih analizama T prema H i T prema NAT kako bi se utvrdilo dovodi li izbor kontrolne skupine do bitnih različitih bioloških zaključaka.
- III. Izraditi integrirane skupove podataka, odnosno provesti ranu integraciju transkriptomskih podataka dobivenih iz više skupova podataka, te na njima razviti i evaluirati modele strojnog učenja za klasifikaciju tumorskih i netumorskih uzoraka.

3. MATERIJALI I METODE

3.1 Izvori i odabir podataka

U ovom istraživanju analizirani su javno dostupni transkriptomski podaci tumora štitnjače i pripadajućih netumorskih uzoraka. Skupovi podataka odabrani su iz repozitorija *Gene Expression Omnibus* (<https://www.ncbi.nlm.nih.gov/geo/>, GEO), te iz integriranog TCGA-GTEx skupa dostupnog putem platforme UCSC Xena (<https://xena.ucsc.edu/public>). Pretraga GEO repozitorija bila je usmjerena na RNA-seq skupove povezane s papilarnim karcinomom štitnjače. U obzir su uzeti samo podaci dobiveni na isključivo ljudskim uzorcima te generirani visokoprotočnim sekvenciranjem RNA (RNA-seq).

Pri odabiru skupova podataka važni su bili dostupnost ekspresijske matrice, jasno opisani metapodaci i mogućnost pouzdanog razvrstavanja uzoraka u biološki relevantne skupine. Posebno su traženi skupovi koji su osim uzoraka papilarnog tumora štitnjače imali odgovarajuću jednu ili više netumorskih kontrolnih skupina. Kontrolne skupine u ovom radu obuhvaćaju zdravo tkivo štitnjače (engl. *healthy*, H) i histološki netumorsko tkivo smješteno uz tumor (engl. *normal adjacent to tumor*, NAT). Skupovi podataka koji nisu sadržavali dovoljno informacija za određivanje statusa uzorka, nisu bili povezani s tkivom štitnjače, potjecali su iz drugih organizama ili su bili dobiveni mikročipovima odnosno scRNA-seq metodama nisu uključeni u analizu.

Odabrani su sljedeći RNA-seq skupovi podataka koji su zadovoljavali sve kriterije: GSE213647 i GSE165724. Skup GSE213647 korišten je kao jedan od glavnih skupova za analizu papilarnog karcinoma štitnjače. U njemu su obrađeni tumorski i netumorski uzorci za koje su bili dostupni podaci o tipu tkiva i relevantnim obilježjima uzoraka, uključujući dob, spol, godinu dijagnoze i laboratorijski kit. Ti su metapodaci korišteni za filtriranje uzoraka, definiranje analiziranih skupina i, prema potrebi, procjenu mogućih izvora tehničke ili biološke varijabilnosti. Ukupno je bilo dostupno 632 uzorka, od kojih je 609 ušlo u završnu analizu: 347 uzorka 347 uzoraka tumorskog tkiva (T), 262 uzorka tumoru susjednog tkiva (NAT), 21 uzorak drugih podtipova carcinoma (5 PDTC i 16 ATC) te jedan uzorak PTC koji nije prošao filtriranje.

Skup GSE165724 također je korišten za analizu tumorskih i netumorskih uzoraka štitnjače. Uzorci su prema metapodacima razvrstani u tumorske uzorke, zdrave kontrolne uzorke i NAT uzorke. Time je ovaj skup bio koristan za usporedbu tumorskog tkiva s različitim

vrstama kontrola te za procjenu koliko izbor kontrolne skupine utječe na dobivene rezultate diferencijalne ekspresije. Ukupno je bilo dostupno 16 uzoraka tumorskog tkiva (T), 46 uzoraka tumorskog susjednog tkiva (NAT) i 12 uzoraka zdravog tkiva (H).

Uz GEO podatke, u radu je korišten i integrirani TCGA-GTEx skup preuzet putem UCSC Xena platforme. TCGA je služio kao izvor tumorskih uzoraka štitnjače i uzoraka tkiva uz tumor, dok je GTEx korišten kao izvor zdravog tkiva štitnjače. Iz podataka su uklonjeni metastatski uzorci, uzorci koji nisu pripadali tkivu štitnjače i uzorci bez jasno određene oznake tipa uzorka. Na taj su način definirane tri skupine: tumorsko tkivo, zdravo tkivo i NAT uzorci.

Ovisno o strukturi pojedinog skupa, ekspresijski podaci obrađeni su kao zajednička matrica ili su pojedinačne datoteke po uzorku spojene u jedinstvenu matricu ekspresije. Prije statističke analize provedeno je usklađivanje matrice ekspresije i metapodataka kako bi svaki stupac ekspresijske matrice odgovarao točnom uzorku u tablici metapodataka. Uzorci s nepotpunim ili neprimjenjivim metapodacima isključeni su iz daljnje analize.

Pripremljeni podaci korišteni su za identifikaciju diferencijalno eksprimiranih gena u pojedinačnim skupovima, kroz usporedbu skupina T prema H te T prema NAT, kasnu integraciju rezultata u obliku meta-analize te ranu integraciju ekspresijskih matrica za potrebe strojnog učenja. Takav pristup omogućio je procjenu stabilnosti potencijalnih biomarkera i ispitivanje utjecaja odabira kontrolne skupine na rezultate diferencijalne ekspresije i klasifikaciju tumorskih uzoraka.

3.2 Računalno okruženje i programski paketi

Cjelokupna bioinformatička obrada podataka provedena je u statističko-programskom okruženju R, verzija 4.6.0. Za dohvat, organizaciju i obradu podataka korišteni su sljedeći R paketi: *GEOquery* 2.80.0, *tidyverse* 2.0.0, *dplyr* 1.2.1, *purrr* 1.2.2, *data.table* 1.18.4, *readxl* 1.4.5, *openxlsx* 4.2.8.1 i *tibble* 3.3.1. Analiza diferencijalne ekspresije provedena je paketima *DESeq2* 1.52.0, *apeglm* 1.34.0, *limma* 3.68.2 i *metaRNASeq* 1.0.8, dok su za anotaciju gena i funkcijsku analizu korišteni *AnnotationDbi* 1.74.0, *org.Hs.eg.db* 3.23.1, *biomaRt* 2.68.0, *clusterProfiler* 4.20.0, *ReactomePA* 1.56.0 i *enrichplot* 1.32.0. Za vizualizaciju rezultata korišteni su *ggplot2* 4.0.3, *ggpubr* 0.6.3, *ggrepel* 0.9.8, *EnhancedVolcano* 1.30.0, *pheatmap* 1.0.13, *RColorBrewer* 1.1.3 i *VennDiagram* 1.8.2. Modeli strojnog učenja i evaluacija njihove uspješnosti izrađeni su pomoću paketa *caret* 7.0.1, *Boruta* 10.0.0, *randomForest* 4.7.1.2 i *pROC* 1.19.0.1.

3.3 Obrada i usklađivanje podataka

Ulazni podaci za analizu bili su neobrađeni brojevi očitavanja po genu (engl. *raw read counts*) dobiveni metodom RNA-seq. Za skup GSE213647 očitavanja su bila dostupna kao zasebne datoteke za svaki uzorak pojedinačno, dok je za skup GSE165724 korištena već objedinjena ekspresijska matrica dostupna u GEO repozitoriju. Podaci iz GEO repozitorija preuzeti su pomoću paketa *GEOquery*, dok su pripadajuće tablične datoteke i metapodaci učitani funkcijama iz paketa *data.table*, *readxl* i *openxlsx*. Pojedinačne datoteke spojene su prema identifikatorima gena u zajedničku matricu ekspresije, pri čemu su stupci predstavljali uzorke, a retci gene. Iz identifikatora gena uklonjene su tehničke oznake, uključujući verzije Ensembl identifikacijskih oznaka, a ponovljeni (duplicirani) identifikatori objedinjeni su zbrajanjem pripadajućih očitavanja funkcijom *rowsum()*. Dobivene matrice ekspresije dodatno su provjerene kako bi sadržavale isključivo numeričke vrijednosti, a nazivi uzoraka usklađeni su s odgovarajućim metapodacima. Sređivanje metapodataka provedeno je funkcijama iz paketa *dplyr* (*transmute()*, *filter()*, *rename()*, *mutate()*, *case_when()*), pri čemu su iz izvornih metapodataka izdvojene varijable relevantne za analizu: GEO identifikator uzorka, tip tkiva, dob, spol, godina dijagnoze, korišteni kit za pripremu sekvencijske knjižnice i pripadnost eksperimentalnoj skupini. Na temelju oznake tkiva uzorci su razvrstani u tumorsku i kontrolnu skupinu, dok su iz analize isključeni uzorci s nepotpunim ili neusklađenim metapodacima, kao i oni koji nisu pripadali promatranim skupinama. Usklađivanje matrica očitavanja i pripadajućih metapodataka provedeno je prema GEO identifikatorima uzoraka pomoću funkcija *match()* i *intersect()*, čime je osigurano da redoslijed uzoraka u matrici ekspresije u potpunosti odgovara redoslijedu u tablici metapodataka. Nakon filtriranja uzoraka postupak usklađivanja ponovljen je kako bi se potvrdila podudarnost dviju tablica.

Pri objedinjavanju više skupova podataka (rana integracija) zadržane su samo varijable koje su zajedničke među skupovima, a pripadnost izvornom skupu podataka uvedena je kao dodatna (tzv. *batch*) varijabla.

3.4 Kontrola kvalitete podataka

Nakon usklađivanja ekspresijskih matrica i metapodataka provedena je kontrola kvalitete podataka radi procjene strukture uzoraka i mogućih izvora tehničke varijabilnosti. Za svaki analizirani skup podataka izrađen je objekt klase *DESeqDataSet* funkcijom

DESeqDataSetFromMatrix() iz paketa *DESeq2*, nakon čega su uklonjeni geni s vrlo niskim brojem očitavanja, jer takvi geni mogu povećati statistički šum i smanjiti pouzdanost rezultata. U većini analiza zadržani su geni koji su imali najmanje 10 očitavanja u barem 10 % uključenih uzoraka. Zatim je provedena normalizacija procjenom faktora veličine knjižnice (engl. *size factors*), čime su korigirane razlike u ukupnom broju očitavanja među uzorcima. U slučajevima u kojima je uočena tehnička varijabilnost, u statistički model uključene su odgovarajuće kovarijable, primjerice kit za pripremu sekvencijske knjižnice, godina dijagnoze ili izvorni skup podataka (*batch*).

Za procjenu strukture podataka korištena je transformacija stabilizacije varijance (engl. *variance stabilizing transformation*, VST), funkcijom *vst()* iz paketa *DESeq2*. VST transformirani podaci korišteni su isključivo za kontrolu kvalitete i vizualizaciju, a ne kao ulazni podaci za samu analizu diferencijalne ekspresije. Na temelju transformiranih vrijednosti izrađena je analiza glavnih komponenti (engl. *principal component analysis*, PCA) pomoću funkcije *plotPCA()*, čime je omogućena provjera grupiranja uzoraka prema biološkoj skupini i uočavanje mogućih tehničkih učinaka. U okviru skupa GSE213647, PCA dijagram služio je u svrhu jasnoće razdvajanja tumorskih od kontrolnih uzoraka papilarnog karcinoma štitnjače. Obzirom da je početni prikaz ukazivao na prisutnost tehničke varijabilnosti, kao mogući izvori te varijabilnosti dodatno su analizirani kit za pripremu sekvencijske knjižnice i godina postavljanja dijagnoze. Radi jasnijeg vizualnog prikaza, utjecaj navedenih varijabli uklonjen je primjenom funkcije *removeBatchEffect()* iz paketa *limma*. Prilikom objedinjavanja skupova GSE213647 i GSE223765 dodatno je uvedena varijabla serije (izvornog skupa podataka) radi kontrole mogućeg *batch* efekta. Korekcija *batch* efekta primijenjena je pri izradi PCA prikaza i procjeni grupiranja uzoraka, dok je sama statistička analiza diferencijalne ekspresije provedena modelom koji je uključivao sve relevantne biološke i tehničke čimbenike.

Za skup TCGA/GTEX, kontrola kvalitete obuhvatila je selekciju tumorskih uzoraka štitnjače i normalnih netumorskih uzoraka štitnjače iz baze TCGA te zdravih uzoraka tkiva štitnjače iz baze GTEX, dok su metastatski uzorci isključeni iz analize. Nakon filtriranja nisko eksprimiranih gena provedena je VST transformacija, a PCA analiza izrađena je u dvije verzije, prije i nakon korekcije kovarijabli. S obzirom na to da su spol i dobna skupina mogli predstavljati dodatne izvore varijabilnosti, njihov je utjecaj uklonjen isključivo za potrebe PCA vizualizacije (ponovno pomoću *removeBatchEffect()* funkcije), dok je razlika između tumorskih i kontrolnih uzoraka i dalje ostala vidljiva.

3.5 Analiza diferencijalne ekspresije gena

Analiza diferencijalne ekspresije provedena je radi identifikacije gena čija se razina ekspresije značajno razlikuje između tumorskih i kontrolnih uzoraka tkiva štitnjače. Za cjelokupnu statističku obradu korišten je paket *DESeq2*.

Uzorci su razvrstani u dvije glavne skupine: tumorske uzorke i kontrolne uzorke, pri čemu su rađene dvije paralelne usporedbe ovisno o dostupnosti kontrolnih uzoraka u pojedinačnom skupu podataka – usporedba skupine tumorskih uzoraka štitnjače (T) prema potpuno zdravim uzorcima štitnjače (H) te usporedba skupine tumorskih uzoraka štitnjače (T) prema histološki normalnim netumorskim uzorcima štitnjače (NAT). Nakon uklanjanja gena s vrlo niskom ekspresijom te normalizacijom podataka, procijenjene su razlike u ekspresiji gena između tumorskih i kontrolnih uzoraka. Za pouzdaniju procjenu promjene ekspresije korištena je metoda smanjivanja $\log_2$ promjene ekspresije pomoću paketa *apecglm*. Time su ublažene ekstremne vrijednosti kod gena s niskim brojem očitavanja, što omogućuje stabilniju interpretaciju rezultata.

Značajno diferencijalno eksprimirani geni određeni su prema prilagođenoj p-vrijednosti (*padj*), veličini promjene ekspresije ($\log_2FC$) i srednjoj razini ekspresije (*baseMean*). Kao kriteriji značajnosti korišteni su prilagođena p-vrijednost manja ili jednaka 0,01, apsolutna vrijednost $\log_2$ promjene ekspresije veća ili jednaka 1,5 te srednja razina ekspresije veća ili jednaka 20. Geni s pozitivnom $\log_2$ promjenom ekspresije smatrani su pojačano eksprimiranim u tumorskim uzorcima, dok su geni s negativnom vrijednošću smatrani sniženo eksprimiranim u tumorskim uzorcima u odnosu na kontrolne uzorke.

Dobiveni rezultati spremljeni su u tablice sa svim analiziranim genima te u zasebne tablice sa značajno diferencijalno eksprimiranim genima. Ti su rezultati korišteni za daljnju vizualizaciju, usporedbu skupova podataka i funkcionalnu interpretaciju.

3.6 Grafička interpretacija diferencijalne ekspresije gena

Rezultati analize diferencijalne ekspresije prikazani su pomoću više vrsta grafičkih prikaza kako bi se omogućila jasnija interpretacija promjena u ekspresiji gena između tumorskih i kontrolnih uzoraka.

Za prikaz ukupne raspodjele diferencijalno eksprimiranih gena korišteni su vulkanski dijagrami. U koordinatnom sustavu na osi x prikazana je vrijednost $\log_2$ promjene ekspresije,

dok je na osi y prikazan negativni logaritam p-vrijednosti s bazom 10, a svaki gen prikazan je kao pojedinačna točka, pri čemu položaj gena ovisi o veličini promjene ekspresije i statističkoj značajnosti. Geni s izraženom pozitivnom $\log_2$ promjenom ekspresije predstavljaju gene pojačano eksprimirane u tumorskim uzorcima, dok geni s negativnom $\log_2$ promjenom predstavljaju gene sniženo eksprimirane u tumorskim uzorcima u odnosu na kontrolu. Najznačajniji geni dodatno su označeni nazivima radi lakšeg prepoznavanja potencijalno važnih biomarkera.

Za usporedni prikaz smjera i veličine promjena ekspresije zajedničkih gena generirana je toplinska mapa (engl. *heatmap*). Prikazano je 30 gena zajedničkih meta-analizama tumora prema zdravom tkivu i tumora prema NAT-u, odnosno 15 gena s najizraženijom povećanom ekspresijom gena i 15 gena s najizraženijom smanjenom ekspresijom gena u tumorskim uzorcima. Stupci predstavljaju rezultate pojedinačnih skupova, a retci eksprimirane gene. Boja svakog polja odgovara vrijednosti $\log_2$ promjene ekspresije gena. Ružičasta označava povećanu, plava sniženu ekspresiju, a vrijednosti približne 0 odgovaraju bijeloj boji.

Svi grafički prikazi korišteni su kao dopuna tabličnim rezultatima analize diferencijalne ekspresije. Njihova je svrha bila olakšati prepoznavanje najizraženijih promjena u genskoj ekspresiji, usporedbu tumorskih i kontrolnih uzoraka te odabir gena za daljnju biološku interpretaciju.

3.7 Meta-analiza pojedinačnih RNA-seq skupova podataka

Meta-analiza provedena je zasebno za usporedbe tumorskog tkiva sa zdravim tkivom (T prema H) i tumorskog tkiva s tkivom susjednim tumoru (T prema NAT). U analizu T prema H uključeni su skupovi GSE165724 i TCGA/GTEx, dok su u analizu T prema NAT-u uključeni GSE213647, GSE165724 i TCGA. Skup GSE213647 nije uključen u usporedbu T prema H jer njegovi netumorski uzorci predstavljaju NAT, a ne tkivo zdravih osoba.

Ulazni podaci za meta-analizu bile su već generirane tablice rezultata diferencijalne ekspresije gena dobivene analizom svakog skupa podataka ponaosob. Oznake gena usklađene su pretvaranjem genskih simbola u Ensembl identifikatore, nakon čega su za svaki kontrast zadržani samo oni geni koji su prisutni u svim uključenim skupovima podataka. Geni bez pripadajuće p-vrijednosti su uklonjeni, a p-vrijednosti jednake 0 zamijenjene su vrijednošću deset puta manjom od najmanje pozitivne p-vrijednosti.

P-vrijednosti pojedinačnih analiza objedinjene su paketom *metaRNASeq* primjenom Fisherove metode i metode inverzne normalne transformacije. U metodi inverzne normalne

transformacije korišten je broj uzoraka pojedinačnih skupova kao težinski faktor. Za svaki je gen zatim određen smjer promjene ekspresije prema predznaku vrijednosti $\log_2FC$. Gen je smatran značajnim u meta-analizi ako je imao prilagođenu p-vrijednost $\leq 0,05$ prema objema metodama i jednak smjer promjene ekspresije u svim uključenim skupovima.

Dodatno su primijenjeni kriteriji $p_{adj} \leq 0,01$, $|\log_2FC| \geq 1,5$ i $baseMean \geq 20$ iz pojedinačnih analiza. U završni popis za usporedbu T prema H uključeni su geni koji su zadovoljavali navedene kriterije u oba skupa podataka. Za usporedbu T prema NAT-u zadržani su geni koji su kriterije zadovoljavali u najmanje dva od tri uključena skupa. Na kraju su završni popisi dviju meta-analiza međusobno uspoređeni kako bi se odredili zajednički i jedinstveni geni te provjerila podudarnost smjera njihove promjene ekspresije. Za razliku od rane integracije, u meta-analizi nisu spajane ekspresijske matrice ni pojedinačni uzorci, nego rezultati prethodno provedenih analiza diferencijalne ekspresije, stoga se ovakav pristup naziva kasna integracija.

3.8 Funkcionalna analiza obogaćenja

Funkcionalna analiza obogaćenja provedena je kako bi se značajno diferencijalno eksprimirani geni povezali s biološkim procesima i signalnim putevima relevantnima za papilarni karcinom štitnjače. Analiza je provedena u programskom jeziku R pomoću paketa *clusterProfiler*, *ReactomePA*, *org.Hs.eg.db* i *biomaRt*.

U analizu su uključeni geni koji su zadovoljili kriterije značajnosti definirane u analizi diferencijalne ekspresije. Kao značajno diferencijalno eksprimirani geni smatrani su oni s prilagođenom p-vrijednošću manjom ili jednakom 0,01, apsolutnom vrijednošću $\log_2$ promjene ekspresije većom ili jednakom 1,5 te srednjom razinom ekspresije većom ili jednakom 20. Geni su prema smjeru promjene ekspresije analizirani kao povišeno i sniženo eksprimirani u tumorskim uzorcima u odnosu na kontrolne uzorke.

Za potrebe funkcionalne analize simboli gena povezani su s odgovarajućim anotacijskim identifikatorima pomoću paketa *org.Hs.eg.db* i *biomaRt*. Obogaćenje je ispitano za biološke procese koristeći *Gene Ontology* (<https://geneontology.org/>) bazu podataka i signalne puteve koristeći KEGG (<https://www.genome.jp/kegg/>) i Reactome (<https://reactome.org/>) bazu podataka kako bi se utvrdilo jesu li među diferencijalno eksprimiranim genima značajno zastupljene određene funkcionalne kategorije. Posebna pažnja posvećena je procesima povezanim s proliferacijom stanica, regulacijom staničnog ciklusa,

imunostnim odgovorom, upalom, remodeliranjem izvanstaničnog matriksa i signalnim putevima uključenima u tumorsku biologiju.

Statistička značajnost procijenjena je na temelju Benjamini-Hochbergove prilagođene p-vrijednosti. Kao značajno obogaćenima smatrani su pojmovi i putevi s prilagođenom p-vrijednošću manjom od 0,05. Rezultati funkcionalne analize prikazani su tablično i grafički, pri čemu su grafički prikazi korišteni za isticanje najbolje rangiranih bioloških procesa i signalnih puteva.

3.9 Rana integracija i objedinjavanje pojedinačnih RNA-seq podataka

Rana integracija provedena je zasebno za analize T prema H i T prema NAT. Analiza T prema H obuhvaćala je skupove podataka GSE165724 i TCGA/GTEX, odnosno ukupno 765 uzoraka. U analizi T prema NAT obuhvaćeni su skupovi podataka GSE165724, GSE213647 i TCGA što čini 1 181 uzorak. Prije spajanja matrica genskih očitavanja iz različitih RNA-seq skupova podataka usklađene su oznake gena. Iz Ensembl identifikatora uklonjene su oznake verzija, a genski simboli iz podataka TCGA/GTEX prevedeni su u Ensembl identifikatore pomoću paketa *org.Hs.eg.db* i *AnnotationDbi*. Zadržani su samo oni geni koji su prisutni u svim skupovima uključenima u pojedinu usporedbu. Matrice očitavanja spojene su prema uzorcima, a metapodacima dodana je oznaka izvornog skupa podataka radi kontrole učinka (*batch*). Prije analize zadržani su geni koji su imali najmanje deset očitavanja u najmanje 10 % uzoraka.

Diferencijalna ekspresija određena je paketom *DESeq2* primjenom modela $\sim \text{batch} + \text{tkivo}$, kojim je istodobno uzet u obzir izvorni skup podataka i vrsta analiziranoga tkiva. Vrijednosti $\log_2FC$ stabilizirane su metodom *apecglm*. Značajnima su smatrani geni koji su zadovoljili kriterije $padj \leq 0,01$, $|\log_2FC| \geq 1,5$ i $baseMean \geq 20$.

Za PCA prikaz provedena je transformacija stabilizacije varijance, nakon čega je učinak studije uklonjen funkcijom *removeBatchEffect()*, uz očuvanje učinka vrste tkiva. Ova korekcija korištena je samo za vizualizaciju, dok je diferencijalna ekspresija izračunata iz izvornih brojeva očitavanja modelom *DESeq2* koji je uključivao varijablu *batch*.

3.10 Izrada modela strojnog učenja

Model strojnog učenja izrađen je pomoću paketa *caret* i *Boruta* u programskom jeziku R, koristeći algoritam *Boruta* za izbor najvažnijih značajki (gena) te model nasumičnih šuma (engl. *random forest*) kako bi se klasificirali tumorski od kontrolnih uzoraka. Cilj analize bio je identificirati gene koji najviše doprinose razlikovanju tumorskih i kontrolnih uzoraka papilarnog karcinoma štitnjače.

Prije izrade modela, normalizirana objedinjena ekspresijska matrica nakon rane integracije svih pojedinačnih RNA-seq skupova podataka, je podijeljena u dvije skupine, na trening i testni skup podataka u omjeru 70 : 30, pri čemu je korištena stratificirana podjela kako bi se očuvao omjer tumorskih i kontrolnih uzoraka u oba skupa. Na trening skupu najprije je provedena analiza diferencijalne ekspresije, a u daljnju analizu uključeni su geni koji su zadovoljili prethodno definirane kriterije značajnosti. Za odabir najinformativnijih gena korišten je *Boruta* algoritam, koji se temelji na metodi nasumičnih šuma. *Boruta* uspoređuje važnost stvarnih varijabli s važnosti nasumično permutiranih varijabli, tzv. sjena značajki, te zadržava samo one gene koji pokazuju veću važnost od slučajno generiranih varijabli. Na taj su način izdvojeni geni kandidati koji najviše doprinose razlikovanju tumorskih i kontrolnih uzoraka. Odabrani geni korišteni su za izradu klasifikacijskog modela koristeći algoritam nasumičnih šuma.

Model je najprije treniran na trening skupu, a njegova učinkovitost procijenjena je na neovisnom testnom skupu. Uspješnost modela procijenjena je pomoću matrice zabune te pokazatelja točnosti, osjetljivosti i specifičnosti. Dodatno je provedena ROC analiza (engl. *Receiver Operating Characteristic*) i izračunata površina ispod ROC krivulje, odnosno AUC vrijednost (engl. *Area Under the Curve*). Vrijednosti rezultata kreću se od 0 do 1, gdje vrijednost $AUC = 1$ predstavlja savršeno odvajanje uzoraka (odnosno savršenu klasifikaciju uzoraka), vrijednost $AUC = 0,5$ ukazuje na nasumičnu klasifikaciju (odnosno da svaki novi uzorak ima 50 % šanse da bude klasificiran kao tumorski ili kao kontrolni), dok vrijednost $AUC = 0$ znači da je model pogrešan te da nema nikakvu prediktivnu moć.

Rezultati ove analize korišteni su za izdvajanje gena s najvećim potencijalom za razlikovanje tumorskih i kontrolnih uzoraka te za dodatnu validaciju kandidata dobivenih analizom diferencijalne ekspresije.

4. REZULTATI

4.1 Skup podataka GSE165724

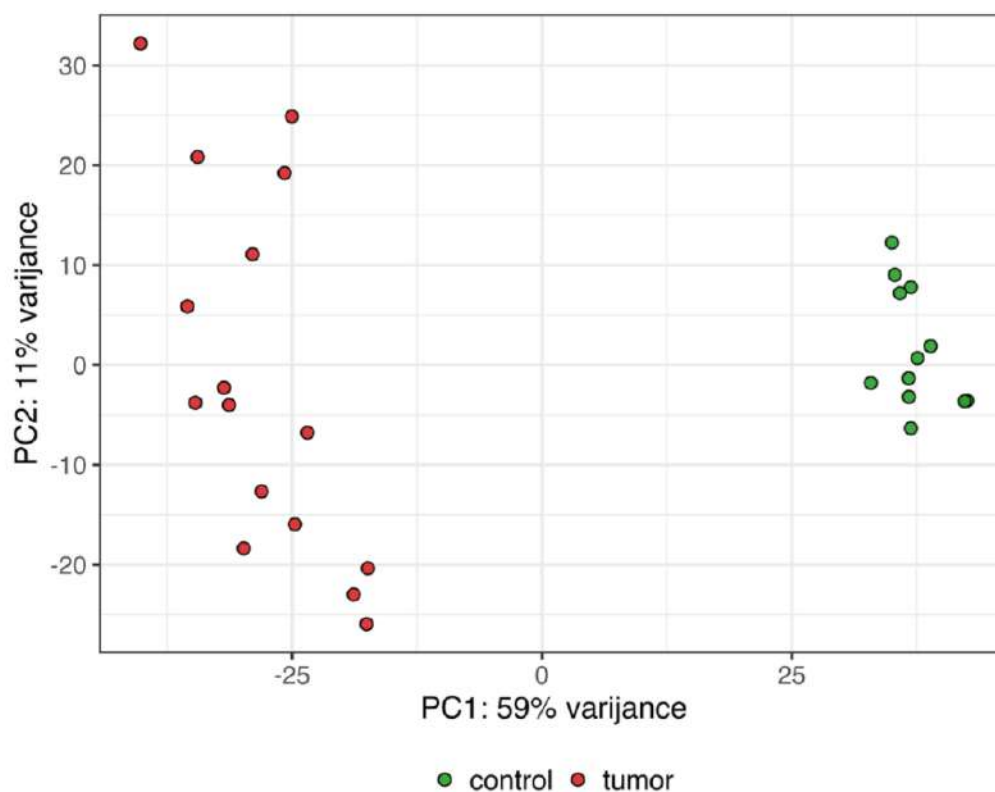
Skup podataka GSE165724 sadrži 74 uzorka tkiva štitnjače i ekspresijsku matricu s 60 675 genskih zapisa. Uzorci su prema pripadajućim metapodacima razvrstani u tri skupine: 16 uzoraka tumorskog tkiva (T), 46 uzoraka tumoru susjednog normalnog tkiva (NAT) i 12 uzoraka zdravog kontrolnog tkiva (H).

Provedene su tri analize: T prema H, T prema NAT te NAT prema H. Prije analize uklonjeni su geni s niskim brojem očitavanja, a diferencijalna ekspresija određena je pomoću paketa *DESeq2*. Značajnima su smatrani geni koji su zadovoljili kriterije $padj \leq 0,01$, $|\log_2FC| \geq 1,5$ i $baseMean \geq 20$.

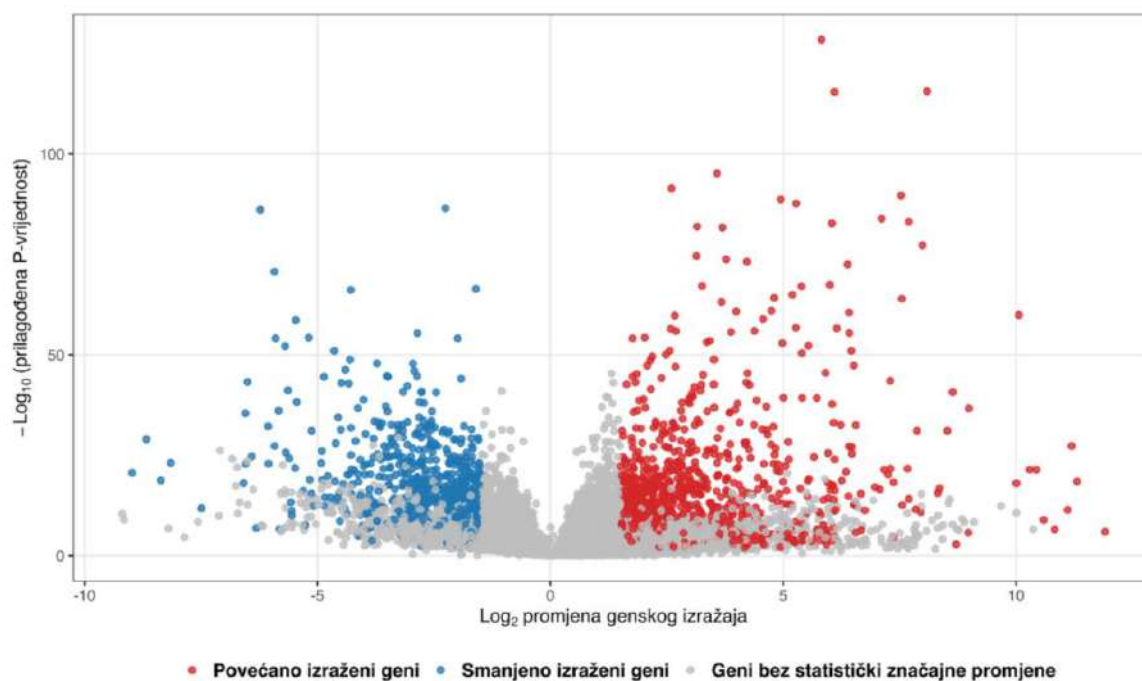
Analizom glavnih komponenti (PCA) utvrđeno je jasno razdvajanje tumorskih i zdravih kontrolnih uzoraka (**Slika 2.**). Prva glavna komponenta, PC1, objašnjava 59 % ukupne varijance, dok je druga glavna komponenta, PC2, objašnjava dodatnih 11 %. Prve dvije komponente zajedno su stoga objašnjavale 70 % varijance među uzorcima. Tumorski uzorci bili su smješteni na negativnoj strani osi PC1, a zdravi kontrolni uzorci na pozitivnoj strani, bez izraženog preklapanja između skupina. Takav raspored pokazuje da je tip tkiva glavni izvor varijabilnosti u analiziranim ekspresijskim podacima te upućuje na izražene razlike između transkriptomskih profila tumorskog i zdravog tkiva.

Analizom diferencijalne ekspresije identificirana su ukupno 2 092 statistički značajno diferencijalno izražena gena gdje je 1 211 gena bilo povišeno, a 881 gen sniženo izražen u tumorskom tkivu, u odnosu na zdravo. Raspodjela diferencijalno izraženih gena prikazana je vulkanskim dijagramom (**Slika 3.**). Pozitivne vrijednosti $\log_2FC$ označavaju gene čija je ekspresija povišena u tumorskom tkivu, dok negativne vrijednosti označavaju gene čija je ekspresija snižena u tumoru u odnosu na zdravo tkivo. Veća vrijednost na osi y odgovara manjoj prilagođenoj p-vrijednosti, odnosno većoj statističkoj značajnosti promjene.

Među istaknutim povišeno izraženim genima nalazili su se *DPP4*, *LIPH*, *LAMB3*, *TMC6*, *RAB27A*, *PRR15*, *TENM1* i *CDH3*, dok su među sniženo izraženim genima bili *ACACB*, *BMP8A*, *CYTH3*, *GPM6A*, *DLG2* i *PTPRQ*. Prisutnost velikoga broja gena s izraženim i statistički značajnim promjenama potvrđuje veliku razliku između molekularnih profila papilarnog karcinoma štitnjače i zdravoga tkiva, što je već uočeno i na PCA prikazu.

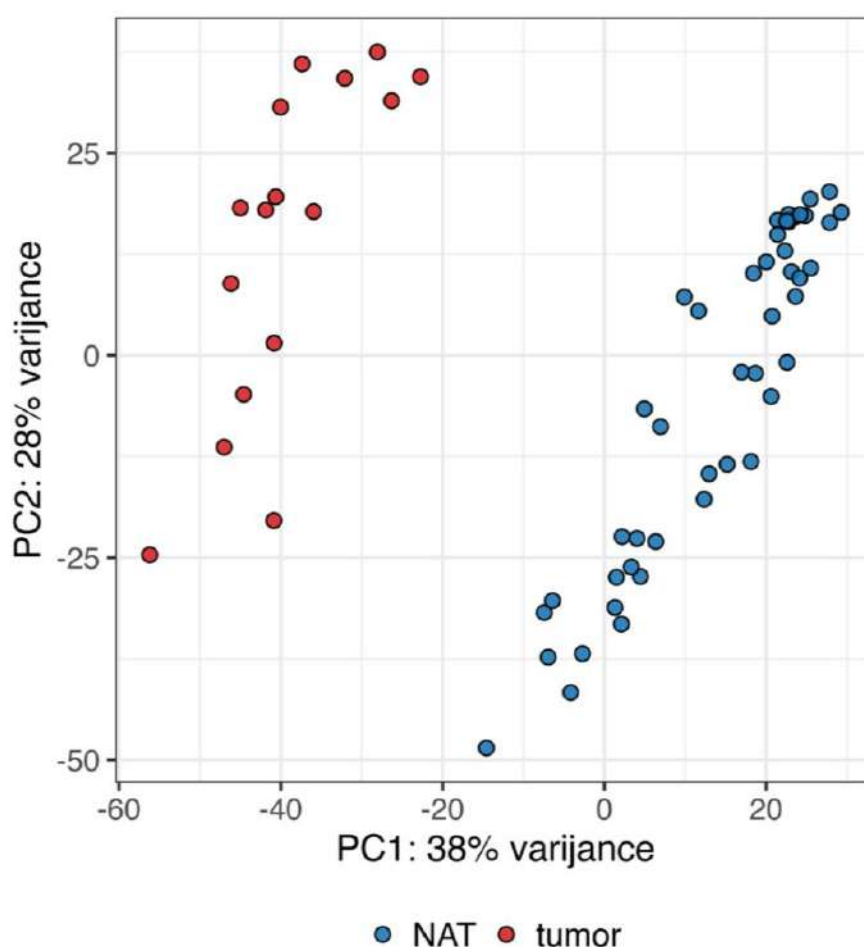


Slika 2. PCA prikaz tumorskog i zdravog tkiva u skupu GSE165724. Crveno je označeno tumorsko (T), a zeleno zdravo tkivo (H). PC1 objašnjava 59 %, a PC2 11 % ukupne varijance.

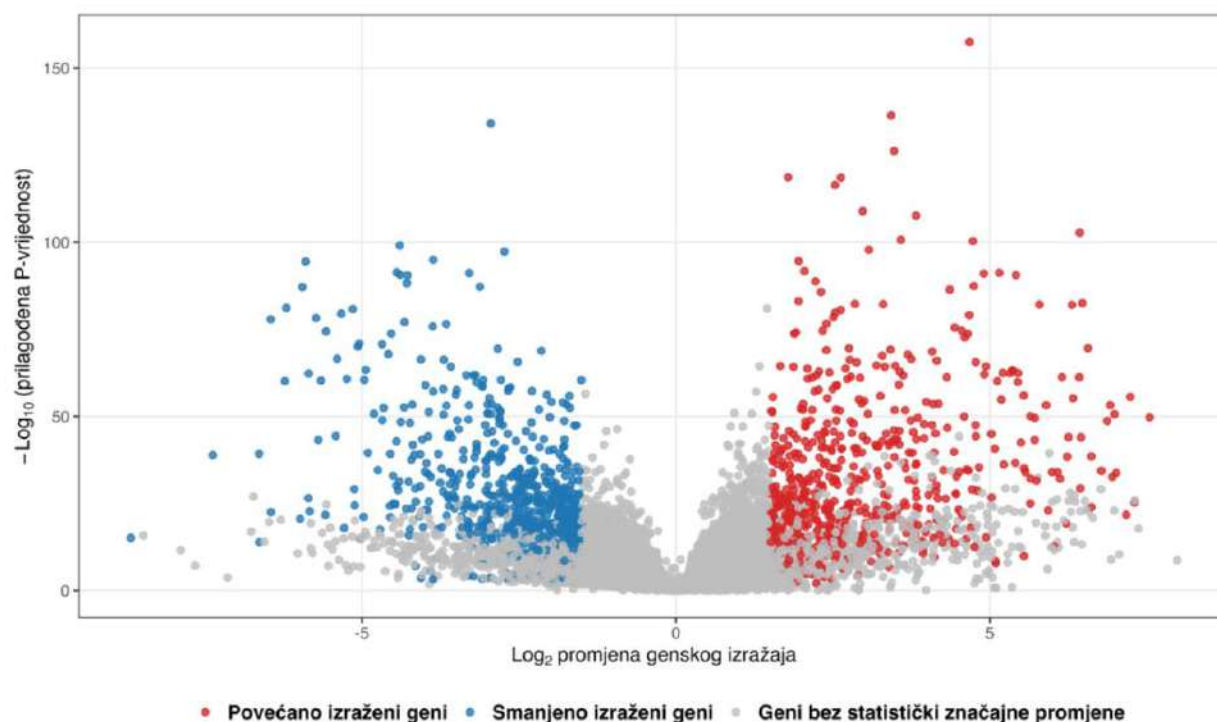


Slika 3. Vulkanski dijagram diferencijalne ekspresije između tumorskog i zdravog tkiva u skupu GSE165724. Crveno su označeni geni povišene, plavo geni snižene ekspresije u tumoru, a sivo geni bez značajne promjene. Identificirana su 2 092 statistički značajno diferencijalno izražena gena ($\text{padj} \leq 0,01$, $|\log_2\text{FC}| \geq 1,5$, $\text{baseMean} \geq 20$).

U usporedbi tumorskih i NAT uzoraka analizirano je 25 084 gena, a identificirana su 1 573 značajno diferencijalno eksprimirana gena. Utvrđeno je da je 722 gena bilo povišeno, a 851 gen sniženo eksprimiran u tumorskim uzorcima u odnosu na NAT. PCA prikaz (**Slika 4.**) pokazao je da PC1 objašnjava 38 %, a PC2 28 % varijance, što zajedno čini 66 % ukupne varijance, dok vulkanski dijagram prikazuje smjer i značajnost promjena (**Slika 5.**). Među najznačajnije povišenim genima u tumoru bili su *TENM1*, *PROS1*, *GALE*, *DOCK9*, *SHROOM4*, *GALNT7*, *MET* i *FAXC*, dok su među sniženima bili *DLG2*, *GPM6A*, *ARHGAP24*, *SLC4A4*, *FABP4*, *ASXL3*, *MATN2* i *MPPED2*.

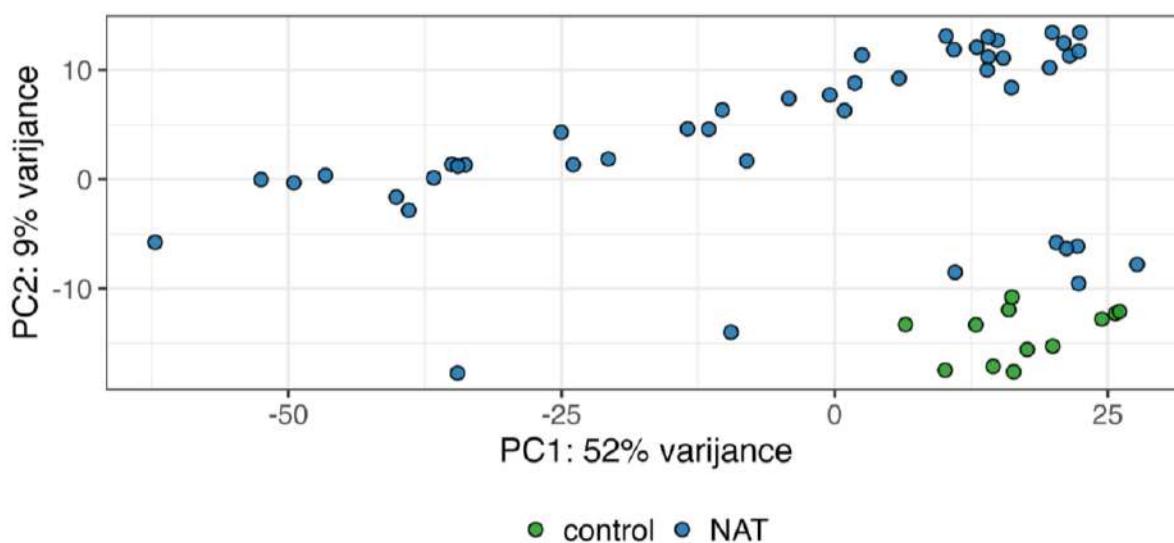


Slika 4. PCA prikaz tumorskog i tumoru susjednog normalnog tkiva (NAT) u skupu GSE165724. Crveno je označeno tumorsko tkivo (T), a plavo NAT. PC1 objašnjava 38 %, a PC2 28 % ukupne varijance.

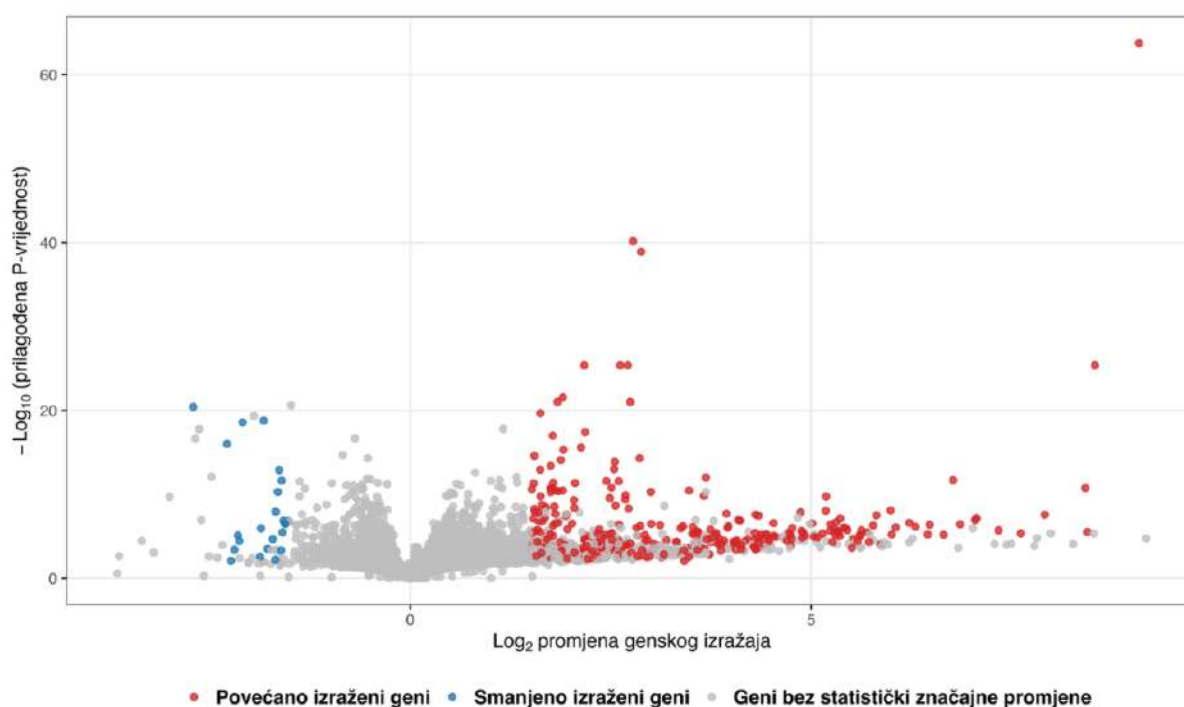


Slika 5. Vulkanski dijagram diferencijalne ekspresije između tumorskog i NAT tkiva u skupu GSE165724. Crveno su označeni geni povišene, plavo geni snižene ekspresije u tumoru, a sivo geni bez značajne promjene. Identificirana su 1 573 statistički značajno diferencijalno izražena gena ($\text{padj} \leq 0,01$, $|\log_2\text{FC}| \geq 1,5$, $\text{baseMean} \geq 20$).

Usporedba NAT uzoraka s potpuno zdravim tkivom pokazala je manji broj promjena nego dvije usporedbe koje uključuju tumor. Analizirano je 24 600 gena, a pronađeno je 426 značajno diferencijalno eksprimiranih gena. Većina ih je bila povišeno eksprimirana u NAT uzorcima, ukupno 403 gena, dok su 23 gena bila sniženo eksprimirana (**Slika 7.**). PCA analiza (**Slika 6.**) pokazala je da PC1 objašnjava 52 %, a PC2 9 % varijance. Iako se NAT uzorci ne predstavljaju kao tumorsko tkivo, njihovo odvajanje od kontrolnih uzoraka upućuje na to da normalno tkivo uz tumor može već imati promijenjen transkriptomski profil u odnosu na neovisno kontrolno tkivo. Među najznačajnijim promjenama nalazili su se *XIST*, *RPS18P5*, *SCARNA9/SCARNA9L* i drugi transkripti, dok je među sniženim genima bio i gen *MTIG*.



Slika 6. PCA prikaz NAT i zdravog tkiva (H) u skupu GSE165724. Plavo je označeno NAT, a zeleno zdravo tkivo. PC1 objašnjava 52 %, a PC2 9 % ukupne varijance. Različita raspodjela skupina pokazuje da NAT nije transkriptomski istovjetan neovisnom zdravom tkivu.



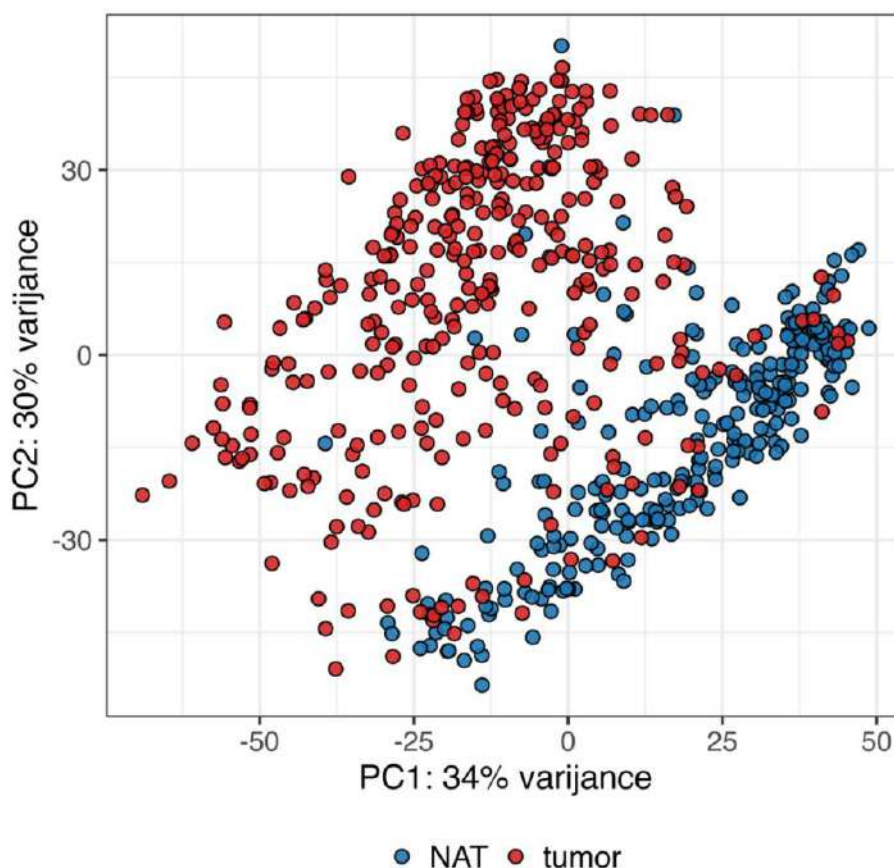
Slika 7. Vulkanski dijagram diferencijalne ekspresije između NAT i zdravog tkiva u skupu GSE165724. Crveno su označeni geni povišene, plavo geni snižene ekspresije u NAT, a sivo geni bez značajne promjene. Identificirano je 426 statistički značajno diferencijalno izraženih gena ($p_{adj} \leq 0,01$, $|\log_2FC| \geq 1,5$, $baseMean \geq 20$).

Najizraženije razlike u GSE165724 skupu podataka uočene su u usporedbi tumorskog i potpuno zdravog tkiva štitnjače, gdje je identificirano više od dvije tisuće značajno promijenjenih gena. Usporedba tumora i NAT uzoraka također je pokazala velik broj diferencijalno eksprimiranih gena, ali nešto manji nego usporedba tumora sa zdravim kontrolnim tkivom. Razlike između NAT i zdravog tkiva usmjeravaju na biološke razlike ovih dviju skupina.

4.2 Skup podataka GSE213647

Skup GSE213647 obuhvaća RNA-seq podatke dobivene iz uzoraka papilarnog karcinoma štitnjače (PTC) i pripadajućih normalnih uzoraka tkiva štitnjače. Nakon učitavanja sirovih ekspresijskih podataka i usklađivanja s metapodacima, u konačnu analizu uključeno je 609 uzoraka od početnih 632: 347 tumorskih uzorka i 262 kontrolna uzorka normalnog tkiva. Od probranih uzoraka 457 ih je pripadalo ženskim, a 152 muškim ispitanicima. Obzirom je u metapodacima uočena prisutnost tehničkih razlika povezanih s vrstom kita za pripremu sekvencijske knjižnice i godinom dijagnoze, navedene su varijable uključene u *DESeq2* model zajedno s biološkim uvjetom, prema dizajnu $\sim \text{kit} + \text{godina dijagnoze} + \text{stanje (tumor ili NAT)}$. 23 uzorka nisu bila dostatna za analizu jer su imali nepotpune podatke ili su bili histološki drugačiji tip karcinoma koji nije u fokusu ovog istraživanja.

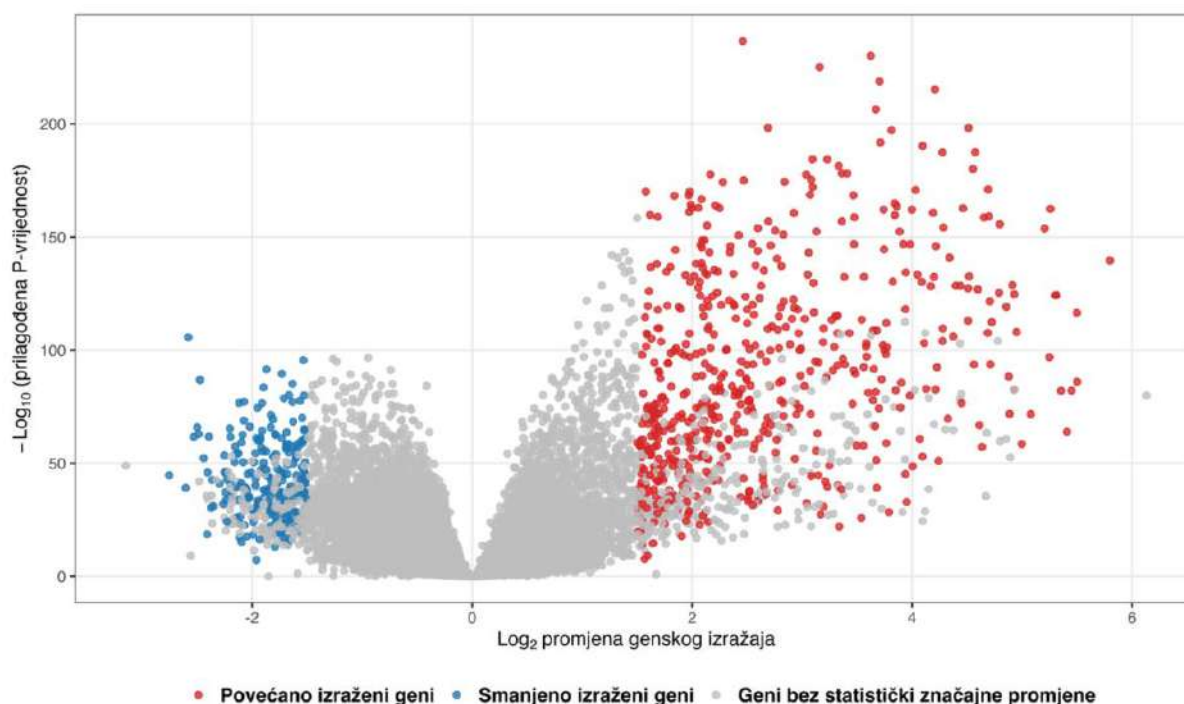
Nakon uklanjanja gena s niskim brojem očitavanja analizirano je 23 046 gena. PCA analiza provedena je na VST-transformiranim podacima te PC1 prikazuje 34 % varijance, a PC2 prikazuje 30 % varijance što daje 64 % ukupne razlike među uzorcima (**Slika 8.**). Kontrolni uzorci grupirani su zajedno na donjoj desnoj strani, dok su tumorski uzorci grupirani gore lijevo što upućuje na jasnu razliku u transkriptomskom profilu između tumorskog tkiva PTC i zdravog tkiva.



Slika 8. PCA prikaz tumorskog (T) i NAT tkiva u skupu GSE213647. Crveno je označeno tumorsko tkivo, a plavo uzorci označeni kao susjedno normalno tkivo (NAT). PC1 objašnjava 34 %, a PC2 30 % ukupne varijance.

Analizom diferencijalne ekspresije između tumorskih i kontrolnih uzoraka identificirano je 847 značajno diferencijalno eksprimiranih gena uz prag $p_{adj} \leq 0,01$, $|\log_2FC| \geq 1,5$ i $baseMean \geq 20$ (**Slika 9.**). U tumorskim uzorcima 588 gena imalo je povišenu ekspresiju, dok je 259 gena imalo sniženu ekspresiju u odnosu na kontrolno tkivo. Veći broj povišeno eksprimiranih gena upućuje na aktivaciju specifičnih signalnih i metaboličkih puteva karakterističnih za tumorsko tkivo.

Među najznačajnije povišeno eksprimiranim genima u PTC uzorcima izdvojili su se *GALE*, *KCNQ3*, *SPTBN2*, *SERPINA1*, *FN1*, *NECTIN4*, *PROS1* i *CLDN16*. Posebno su izražene promjene zabilježene za gene *FN1* i *CLDN16*, koji su imali $\log_2FC$ veći od 4. Među najznačajnije sniženo eksprimiranim genima nalazili su se *LYVE1*, *CITED2*, *MATN2*, *C8orf88*, *MMRNI*, *CPAP*, *TPPP* i *ANGPTL1*. Snižena ekspresija gena poput *LYVE1*, *MATN2* i *MMRNI* može upućivati na promjene u stromalnim, vaskularnim i izvanstaničnim komponentama tkiva tijekom tumorske preobrazbe.



Slika 9. Vulkanski dijagram diferencijalne ekspresije između tumorskog i NAT tkiva u skupu GSE213647. Crveno su označeni geni povišene, plavo geni snižene ekspresije u tumoru, a sivo geni bez značajne promjene. Identificirano je 847 statistički značajno diferencijalno izraženih gena ($p_{adj} \leq 0,01$, $|\log_2FC| \geq 1,5$, $baseMean \geq 20$).

Sveukupno, analiza skupa GSE213647 pokazala je izražene razlike u genskoj ekspresiji između papilarnog karcinoma štitnjače i normalnog tkiva. Dobivenih 847 značajno promijenjenih gena predstavlja osnovu za daljnju usporedbu s drugim skupovima podataka i za izdvajanje zajedničkih molekularnih obrazaca povezanih s PTC-om.

4.3 TCGA/GTEx

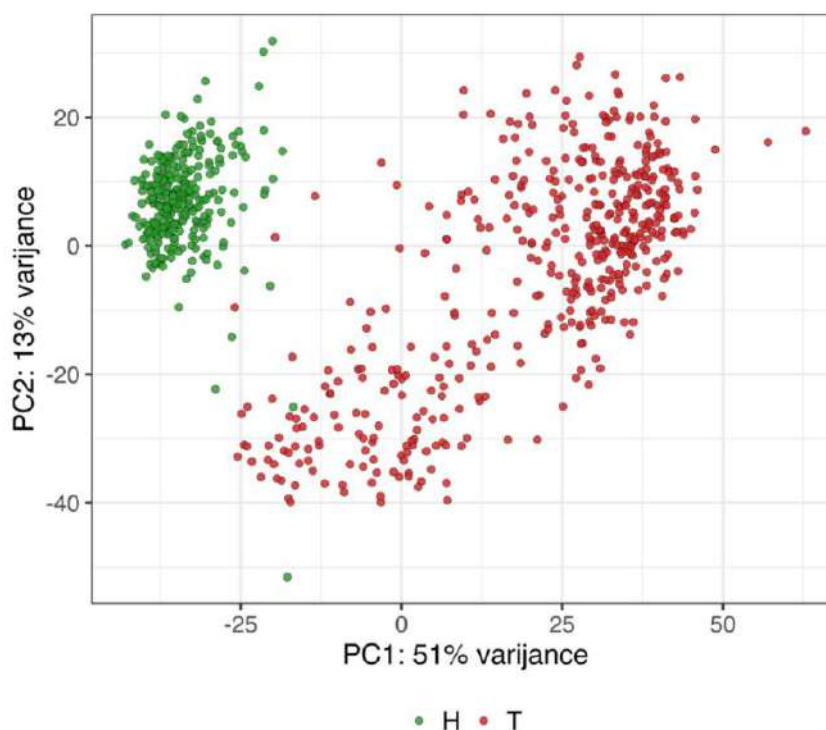
Objedinjeni TCGA/GTEx skup podataka korišten je za usporedbu tumorskog, NAT i zdravog tkiva štitnjače. Podaci su preuzeti u obliku već pripremljene ekspresijske matrice, pri čemu su za analizu izdvojeni uzorci tkiva štitnjače. U početnom pregledu podataka identificirani su zdravi uzorci štitnjače iz GTEx baze te tumorski i NAT uzorci iz TCGA baze. Metastatski uzorci nisu uključeni u analizu.

Nakon filtriranja u analizu je uključeno 789 uzoraka tkiva štitnjače: 279 zdravih kontrolnih uzoraka (H), 458 tumorskih uzoraka (T) i 52 histološki normalna netumorska uzorka (NAT). U *DESeq2* model uključene su i kovarijable dobne skupine i spola, kako bi se smanjio

utjecaj tih čimbenika na rezultate diferencijalne ekspresije, odnosno da glavna biološka razlika proistječe na temelju razlike između tumorskih i kontrolnih skupina.

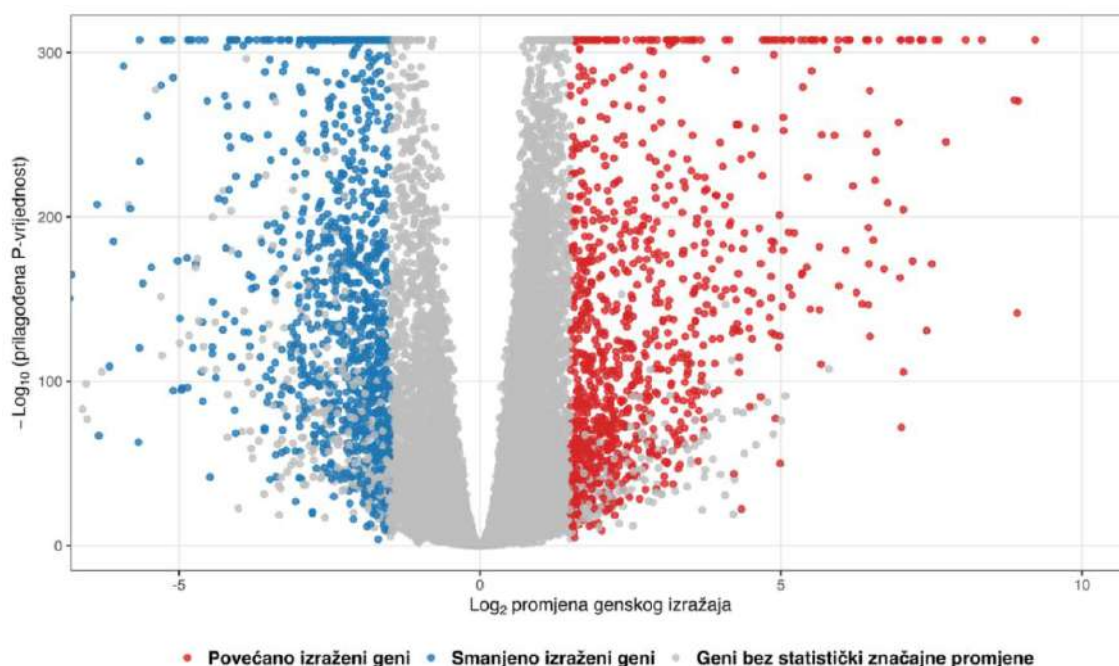
Analiza glavnih komponenti pokazala je razdvajanje uzoraka prema tipu tkiva: tumorski uzorci odvajali su se od zdravih GTEx kontrola, dok su NAT uzorci zauzimali međupoložaj između zdravog i tumorskog tkiva. Takav raspored upućuje na izražene transkriptomске razlike između tumorskog i zdravog tkiva te opet na činjenicu da NAT uzorci nisu potpuno jednaki zdravim kontrolama. Diferencijalna ekspresija gena analizirana je za dvije usporedbe: tumorsko tkivo naspram zdravog tkiva te tumorsko tkivo naspram NAT tkiva. Značajno diferencijalno eksprimirani geni definirani su prema kriterijima $padj \leq 0,01$, $|\log_2FC| \geq 1,5$ i $baseMean \geq 20$.

Prilikom usporedbe tumorskog tkiva u odnosu na zdrave kontrole, PCA je pokazala jasno razdvajanje skupina; PC1 objašnjava 51 %, a PC2 13 % varijance (**Slika 10.**). U usporedbi tumorskog tkiva i zdravih kontrola identificirano je ukupno 2 150 značajno diferencijalno eksprimiranih gena. Od toga je 1 037 gena bilo povećano, a 1 113 gena smanjeno izraženo u tumorskom tkivu u odnosu na zdravo tkivo što potvrđuje razliku između transkriptomskog profila papilarnog karcinoma štitnjače i zdravog tkiva štitnjače (**Slika 11.**).



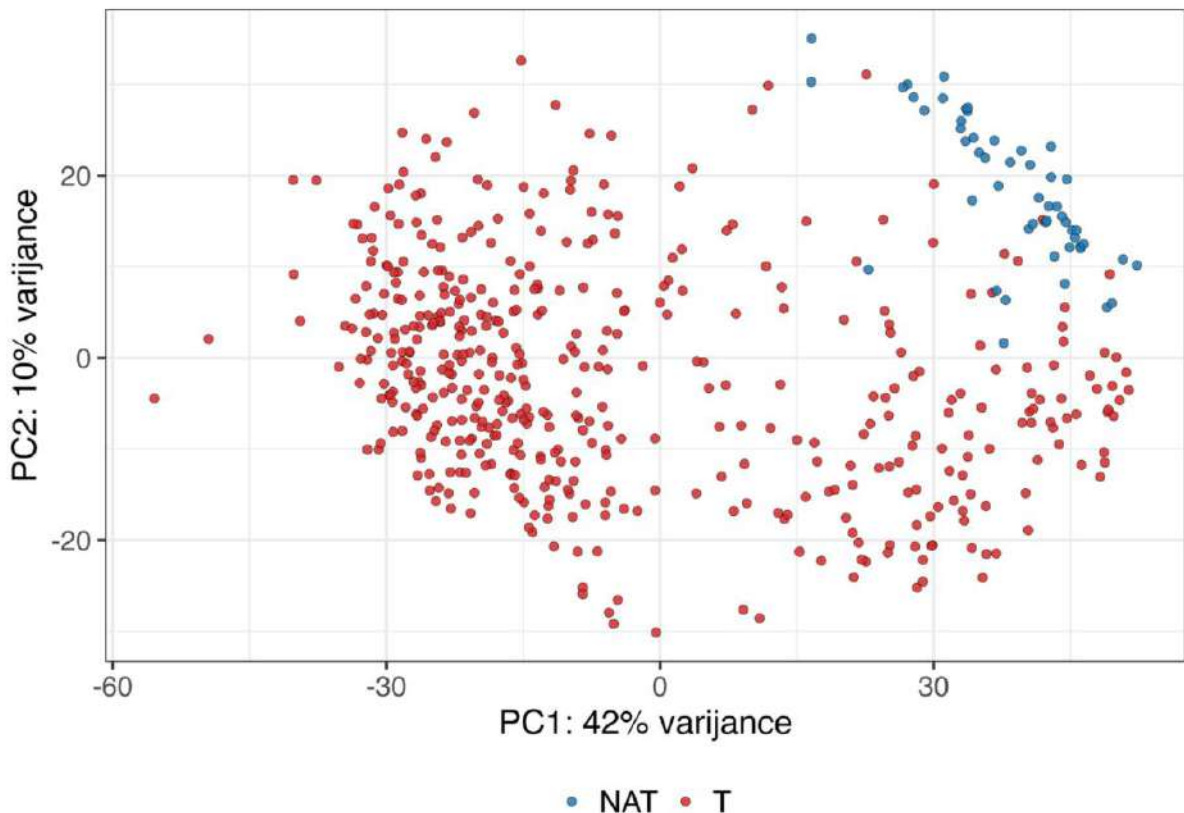
Slika 10. PCA prikaz tumorskog i zdravog tkiva u objedinjenom skupu TCGA/GTEx. Crveno je označeno tumorsko tkivo iz TCGA, a zeleno zdravo tkivo štitnjače iz GTEx. PC1 objašnjava 51 %, a PC2 13 % ukupne varijance.

Među najznačajnije povišeno eksprimiranim genima u tumorskim uzorcima izdvojili su se *NIPAL3*, *LAMC2*, *CDH3*, *NGEF*, *TRPC5*, *APLP2*, *NRCAM* i *MET*. Posebno izražene promjene zabilježene su za gene *CDH3*, *NGEF*, *TRPC5* i *CLDN16*, čije su vrijednosti $\log_2FC$ bile veće od 4. Najznačajniji sniženo eksprimirani geni bili su *TFPI*, *RBM6*, *ARHGAP33*, *PKD1*, *ABCC2*, *PLEKHH1*, *SEC31B* i *CHRD*. Promjene ekspresije gena povezanih sa staničnom adhezijom, signalizacijom, proliferacijom i izvanstaničnim matriksom upućuju na opsežno transkriptomsko preoblikovanje tumorskoga tkiva u odnosu na zdravo tkivo štitnjače.



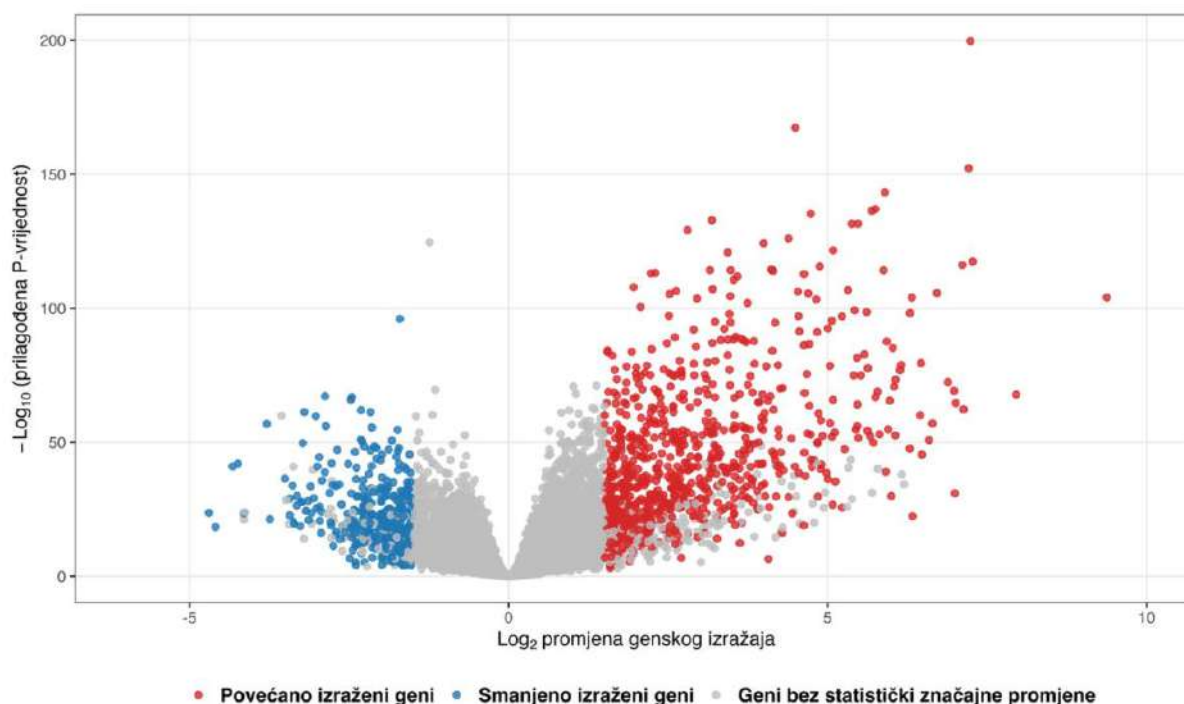
Slika 11. Vulkanski dijagram diferencijalne ekspresije između tumorskog TCGA i zdravog GTEx tkiva. Crveno su prikazani geni povišene, plavo geni snižene ekspresije u tumoru, a sivo geni bez značajne promjene. Identificirano je 2 150 statistički značajno diferencijalno izraženih gena ($padj \leq 0,01$, $|\log_2FC| \geq 1,5$, $baseMean \geq 20$).

U usporedbi tumorskog i NAT tkiva, PC1 objašnjava 42 %, a PC2 10 % varijance, uz djelomično razdvajanje tumorskih i NAT uzoraka (**Slika 12.**). Identificirana su 1 204 značajno diferencijalno eksprimirana gena, od čega je 855 gena bilo povećano, a 349 gena smanjeno izraženo u tumorskim uzorcima u odnosu na NAT. Manji broj diferencijalno eksprimiranih gena u usporedbi T prema NAT u odnosu na T prema H pokazuje da je NAT tkivo molekularno sličnije tumorskom tkivu nego zdravom GTEx tkivu, ali se od tumora ipak jasno razlikuje (**Slika 13.**).



Slika 12. PCA prikaz tumorskog i NAT tkiva iz skupa TCGA. Crveno je označeno tumorsko tkivo, a plavo NAT. PC1 objašnjava 42 %, a PC2 10 % ukupne varijance.

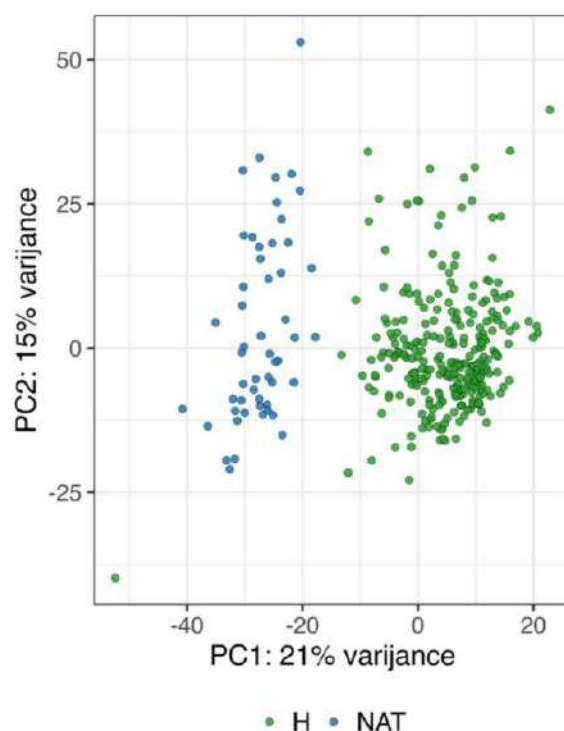
U usporedbi tumorskoga tkiva s NAT identificirana su ukupno 1 204 statistički značajno diferencijalno izražena gena, od kojih je 855 bilo povećano, a 349 smanjeno izraženo u tumoru. Među najznačajnije povećano izraženim genima izdvojili su se *GABRB2*, *LRP4*, *SLC22A31*, *PRR15*, *HMGA2*, *LIPH*, *TENM1* i *GALE*. Posebno izražene promjene zabilježene su za gene *GABRB2*, *SLC22A31*, *PRR15*, *HMGA2* i *LIPH*, čije su vrijednosti $\log_2FC$ bile veće od 5. Među najznačajnije smanjeno izraženim genima nalazili su se *RPS6KA5*, *PLA2R1*, *PAPSS2*, *RASSF9*, *CDHR3*, *MMRN1*, *SBSPON* i *LYVE1*. Smanjena ekspresija gena *MMRN1*, *LYVE1* i *MATN2* može upućivati na promjene u vaskularnim, stromalnim i izvanstaničnim komponentama tumorskoga tkiva u odnosu na NAT.



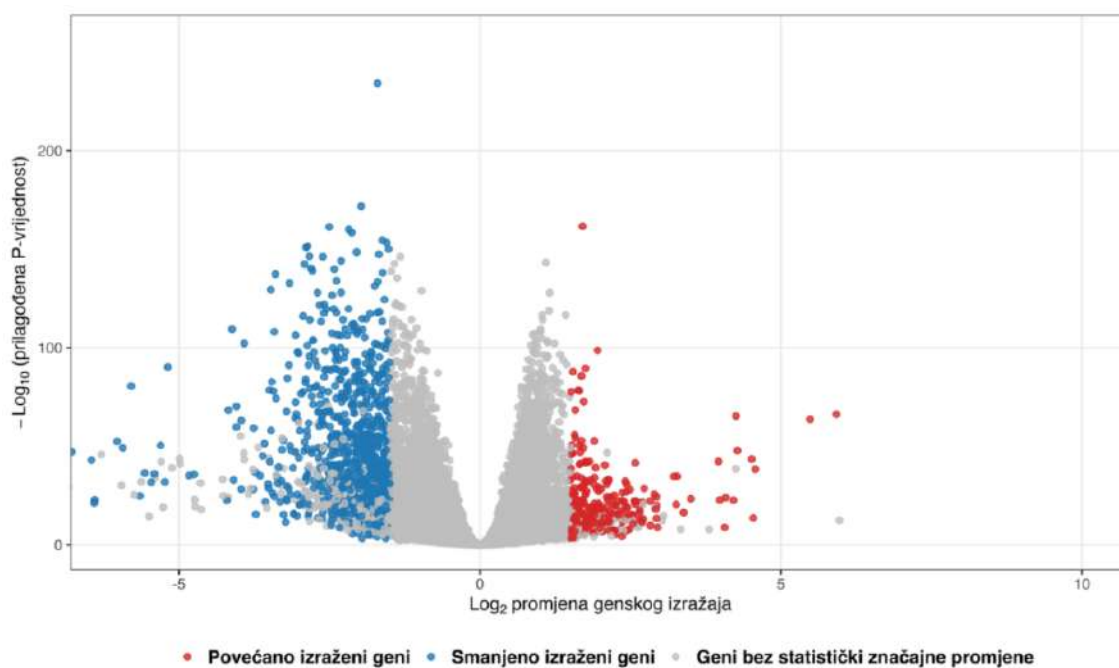
Slika 13. Vulkanski dijagram diferencijalne ekspresije između tumorskog i NAT tkiva iz TCGA. Crveno su označeni geni povišene, plavo geni snižene ekspresije u tumoru, a sivo geni bez značajne promjene. Identificirana su ukupno 1 204 statistički značajno diferencijalno izražena gena, od kojih je 855 bilo povećano, a 349 smanjeno izraženo u tumorskom tkivu u odnosu na NAT ($padj \leq 0,01$, $|\log_2FC| \geq 1,5$, $baseMean \geq 20$).

U usporedbi NAT tkiva iz TCGA sa zdravim tkivom štitnjače iz GTEx-a uzorci su se razdvajali ponajprije duž PC1, koja objašnjava 21 % varijance, dok PC2 objašnjava 15 % (Slika 14.). Vulkanski dijagram pokazao je značajne promjene u oba smjera, što potvrđuje da se ekspresijski profil NAT razlikuje od profila neovisnoga zdravog tkiva (Slika 15.).

U usporedbi NAT prema H identificirano je ukupno 1 247 statistički značajno diferencijalno izraženih gena, od kojih je 280 bilo povećano, a 967 smanjeno izraženo u NAT-u. Od povećano izraženih gena izdvojili su se *CDH1* i *CXADR*, koji sudjeluju u međustaničnoj adheziji, organizaciji epitelnih spojeva i održavanju polariteta stanica. Visoke pozitivne vrijednosti $\log_2FC$ zabilježene su za *FOSB* i *EGR2*, transkripcijske regulatore uključene u odgovor na izvanstanične podražaje te regulaciju proliferacije i diferencijacije, kao i za *FCAMR* i *LGALS2*, gene povezane s imunološkim odgovorom i sastavom tumorskoga mikrookoliša. Dobiveni rezultati pokazuju da se ekspresijski profil NAT tkiva razlikuje od profila neovisnoga zdravog tkiva štitnjače. Geni su izdvojeni primjenom kriterija $padj \leq 0,01$, $|\log_2FC| \geq 1,5$ i $baseMean \geq 20$.



Slika 14. PCA prikaz NAT tkiva iz TCGA i zdravog tkiva štitnjače iz GTEx. Plavo je označeno NAT, a zeleno zdravo tkivo. PC1 objašnjava 21 %, a PC2 15 % ukupne varijance.

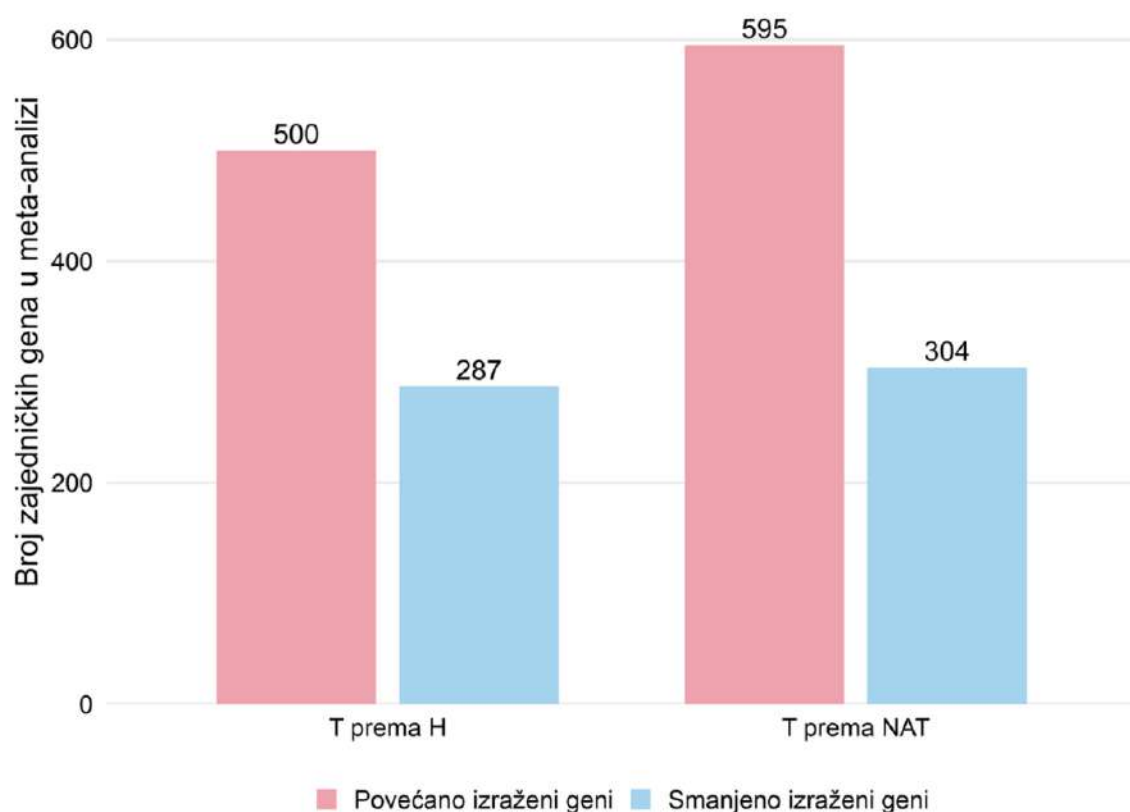


Slika 15. Vulkanski dijagram diferencijalne ekspresije između NAT tkiva iz TCGA i zdravog tkiva štitnjače iz GTEx. Crveno su prikazani geni povišene, plavo geni snižene ekspresije u NAT, a sivo geni bez značajne promjene. Identificirano je 1 247 statistički značajno diferencijalno izraženih gena ($\text{padj} \leq 0,01$, $|\log_2\text{FC}| \geq 1,5$, $\text{baseMean} \geq 20$), od kojih je 280 bilo povećano, a 967 smanjeno izraženo u NAT-u u odnosu na zdravo tkivo.

Rezultati TCGA/GTEX objedinjenog skupa potvrđuju da je najveća razlika prisutna između tumorskog i zdravog tkiva. Uključivanje NAT uzoraka dodatno pokazuje da se susjedno normalno tkivo ne može u potpunosti smatrati ekvivalentom zdravog tkiva, što je važno pri interpretaciji rezultata i usporedbi s drugim skupovima podataka.

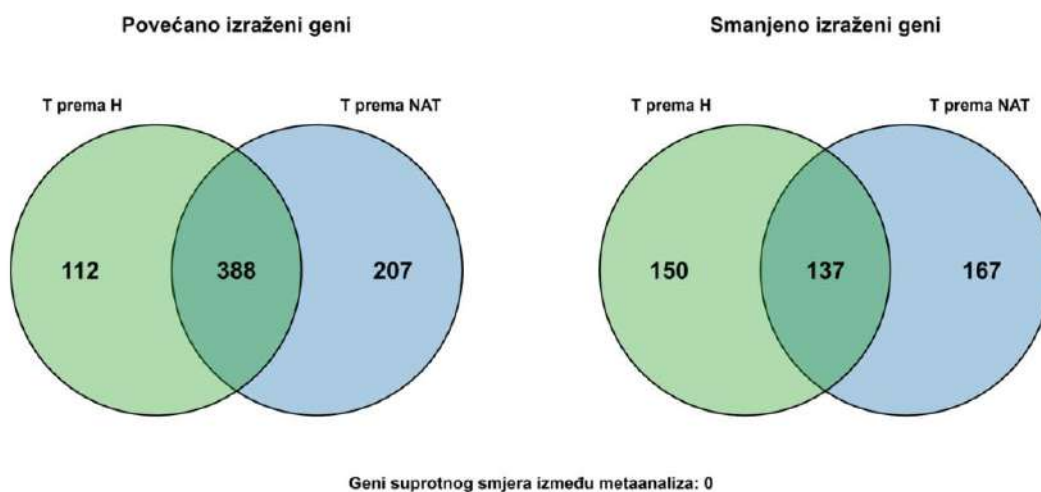
4.4 Meta-analiza i preklapanje rezultata pojedinačnih studija

Kasna integracija rezultata diferencijalne ekspresije pojedinačnih skupova podataka provedena je odvojeno za usporedbe T prema H i T prema NAT, kako bi se izdvojili robustni rezultati. U meta-analizu T prema H uključeni su rezultati skupova GSE165724 i TCGA/GTEX, dok su u meta-analizu T prema NAT uključeni rezultati skupova GSE165724, GSE213647 i TCGA. Rezultati meta-analize obuhvatili su gene sa značajnim kombiniranim rezultatima (prilagođena p-vrijednost manja ili jednaka od 0,05) i podudarnim smjerom promjene ekspresije među uključenim skupovima. U usporedbi T prema H izdvojeno je 787 gena, od kojih je 500 imalo povišenu, a 287 sniženu ekspresiju u tumorskom tkivu u odnosu na zdravu kontrolu. U usporedbi T prema NAT izdvojeno je 899 gena, uključujući 595 gena povišene i 304 gena snižene ekspresije u tumorskom tkivu u odnosu na NAT kontrolu (**Slika 16.**).



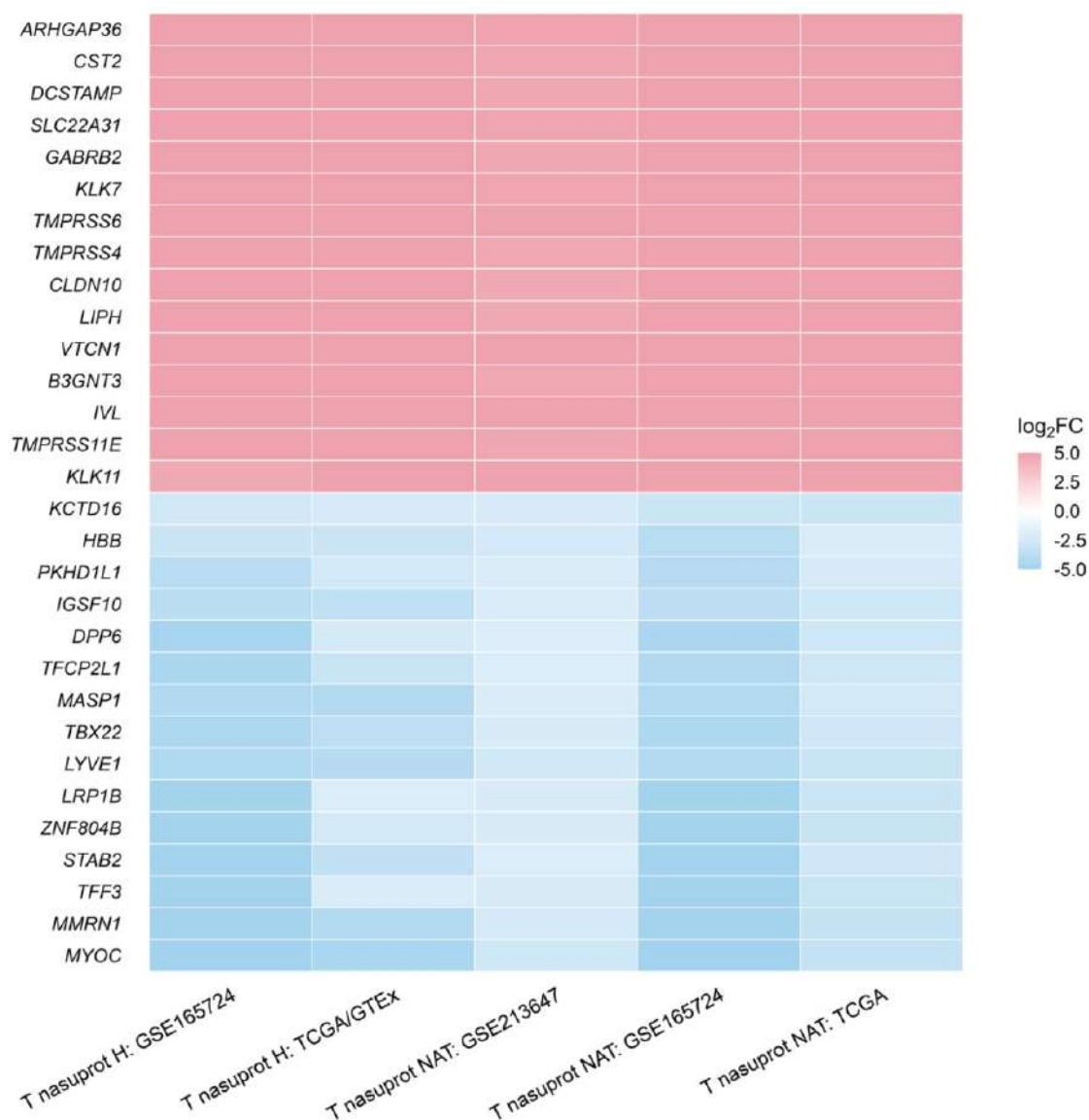
***Slika 16.** Broj zajedničkih gena identificiranih meta-analizom u usporedbama tumorskog sa zdravim tkivom (T prema H) i tumorskog s NAT tkivom (T prema NAT).*

Preklapanjem popisa zajedničkih gena identificiranih u ove dvije meta-analize (T prema H te T prema NAT), pronađeno je 525 zajedničkih gena u obje meta-analize. Dodatnih 262 gena bila su svojstvena usporedbi T prema H, a 374 usporedbi T prema NAT (**Slika 17.**). Velik zajednički dio pokazuje postojanje stabilnoga tumorskog transkriptomskog signala, dok geni svojstveni pojedinoj usporedbi upućuju na utjecaj vrste kontrolnog tkiva.



Slika 17. Preklapanje zajedničkih gena dobivenih u dvije meta-analize između usporedbi *T* prema *H* i *T* prema *NAT*. Zajedničko je 525 gena, dok su 262 gena svojstvena usporedbi *T* prema *H*, a 374 usporedbi *T* prema *NAT*.

Za 30 zajedničkih gena s najvećim apsolutnim promjenama ekspresije uspoređene su vrijednosti $\log_2FC$ u pet uključenih analiza (**Slika 18.**). Među genima dosljedno povišene ekspresije u tumorskom tkivu isticali su se *ARHGAP36*, *CST2*, *GABRB2*, *LIPH*, *TMPRSS4* i *CLDN10*, dok su među dosljedno sniženima bili *LYVE1*, *MMRN1*, *STAB2*, *TFF3*, *DPP6* i *MYOC*.



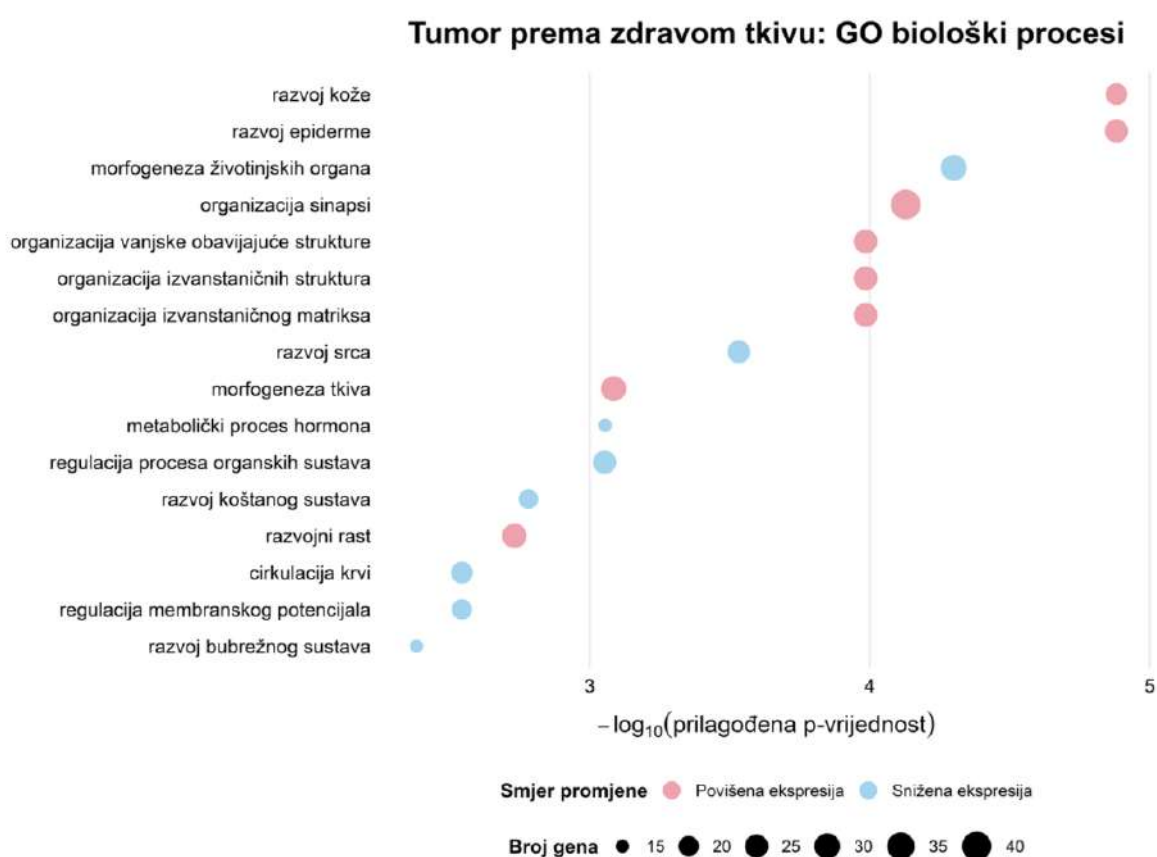
Slika 18. Toplinska mapa 30 zajedničkih gena s najvećim apsolutnim promjenama ekspresije u analiziranim skupovima. Crveno označava povišenu, a plavo sniženu ekspresiju u papilarnom karcinomu štitnjače u odnosu na kontrolne uzorke štitnjače; intenzitet boje odgovara vrijednosti $\log_2FC$.

4.5 Funkcionalna analiza obogaćenja

Funkcionalna analiza provedena je zasebno za zajedničke gene dobivene u meta-analizama T prema H i T prema NAT. Geni povišene i snižene ekspresije analizirani su odvojeno, a značajnima su smatrani pojmovi s Benjamini-Hochbergovom prilagođenom p-vrijednošću manjom ili jednakom 0,05.

4.5.1 Usporedba T naspram H

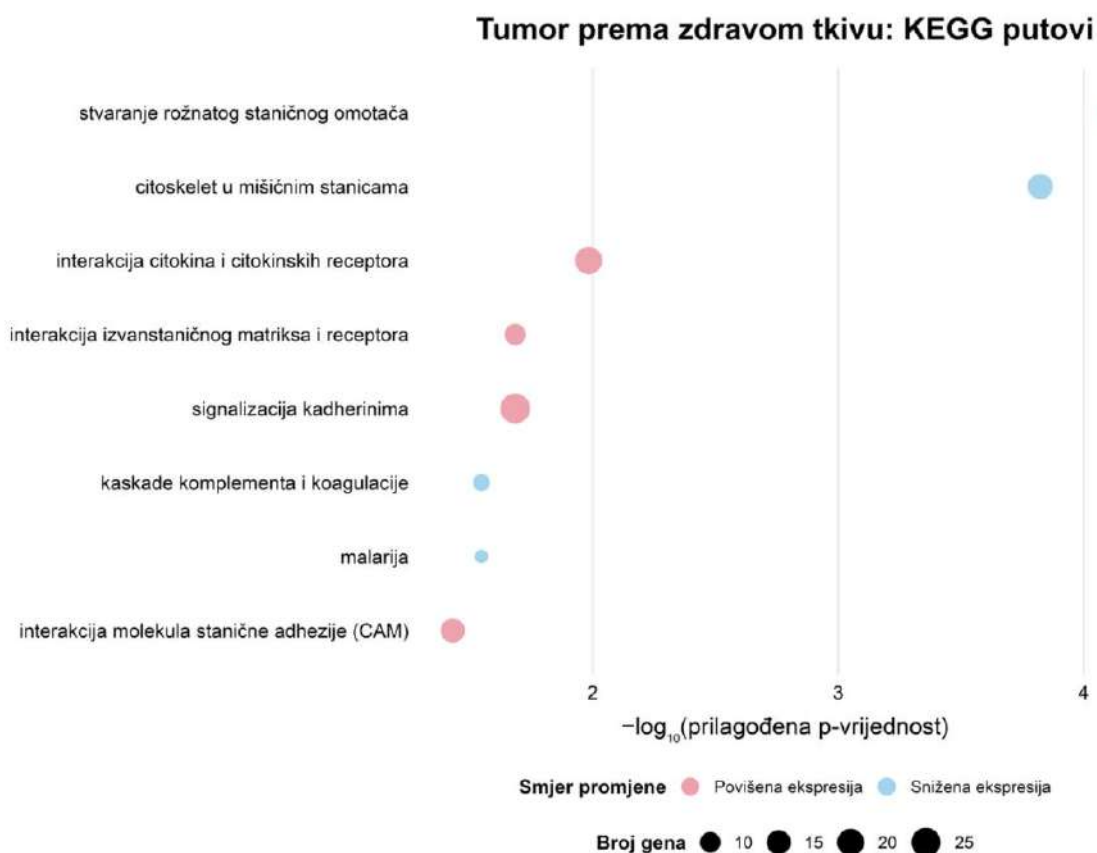
U analizu T prema H uključeno je 787 zajedničkih gena (500 povišene i 287 snižene ekspresije u tumorskom tkivu), uz pozadinski skup od 14 690 mapiranih gena. Analizom *Gene Ontology* (GO) identificirano je 198 značajnih pojmova bioloških procesa, 40 staničnih komponenti i 36 molekularnih funkcija. Među povišeno izraženim genima prevladavali su biološki putevi povezani uz razvoj epidermisa, organizacijom izvanstaničnog matriksa i morfogeneze tkiva, dok su sniženo izraženi geni bili povezani s razvojem organa, metabolizmom hormona i regulacijom membranskoga potencijala (Slika 19.).



Slika 19. Obogaćenje GO bioloških procesa za zajedničke gene dobivene u meta-analizi usporedbe T prema H. Boja označava smjer promjene ekspresije (crveno su prikazani procesi obogaćeni među genima povišene, a svijetloplavo među genima snižene ekspresije u tumoru), položaj na osi x statističku značajnost, a veličina točke broj gena pridruženih pojedinom putu.

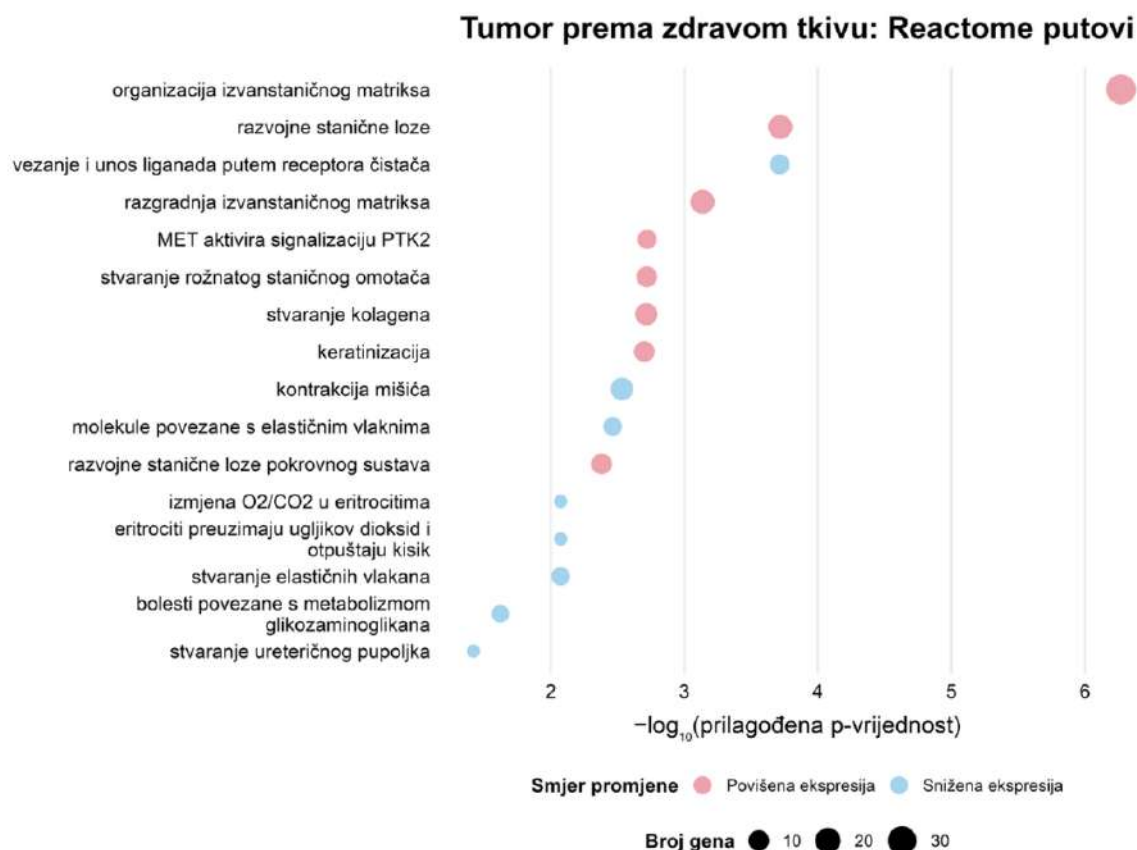
KEGG analizom identificirano je osam značajno obogaćenih puteva. Među genima povišene ekspresije izdvojeni su interakcija citokina i citokinskih receptora, interakcija izvanstaničnoga matriksa i receptora, signalizacija kadherinima i stanična adhezija, dok su

među sniženima bili putovi citoskeleta u mišićnim stanicama te komplementa i koagulacije (Slika 20.).



Slika 20. KEGG analiza obogaćenja za zajedničke gene dobivene u meta-analizi usporedbe T prema H.. Boja označava smjer promjene ekspresije (crveno su prikazani procesi obogaćeni među genima povišene, a svijetloplavo među genima snižene ekspresije u tumoru), položaj na osi x statističku značajnost, a veličina točke broj gena pridruženih pojedinom putu.

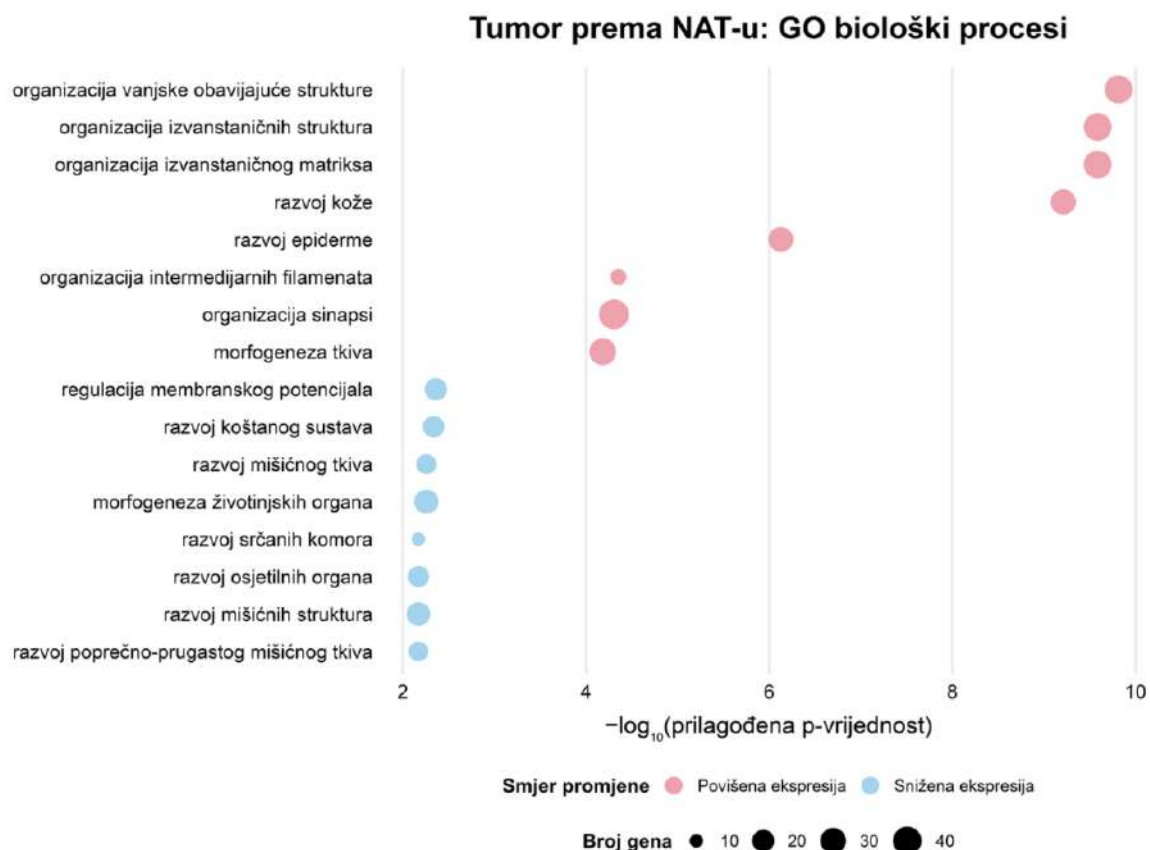
Analizom signalnih puteva iz Reactome baze podataka pronađeno je 29 značajno obogaćenih puteva. Najizraženiji rezultati među genima povišene ekspresije odnosili su se na organizaciju i razgradnju izvanstaničnog matriksa, stvaranje kolagena, keratinizaciju i aktivaciju PTK2 signalizacije posredovanu receptorom MET. Među genima snižene ekspresije isticali su se putevi kontrakcije mišića i elastičnih vlakana (Slika 21.).



Slika 21. Reactome analiza obogaćenja za zajedničke gene dobivene u meta-analizi usporedbe T prema H. Prikazani su najbolje rangirani putovi među genima povišene i snižene ekspresije u tumorskom tkivu. Boja označava smjer promjene ekspresije (crveno su prikazani procesi obogaćeni među genima povišene, a svijetloplavo među genima snižene ekspresije u tumoru), položaj na osi x statističku značajnost, a veličina točke broj gena pridruženih pojedinom putu.

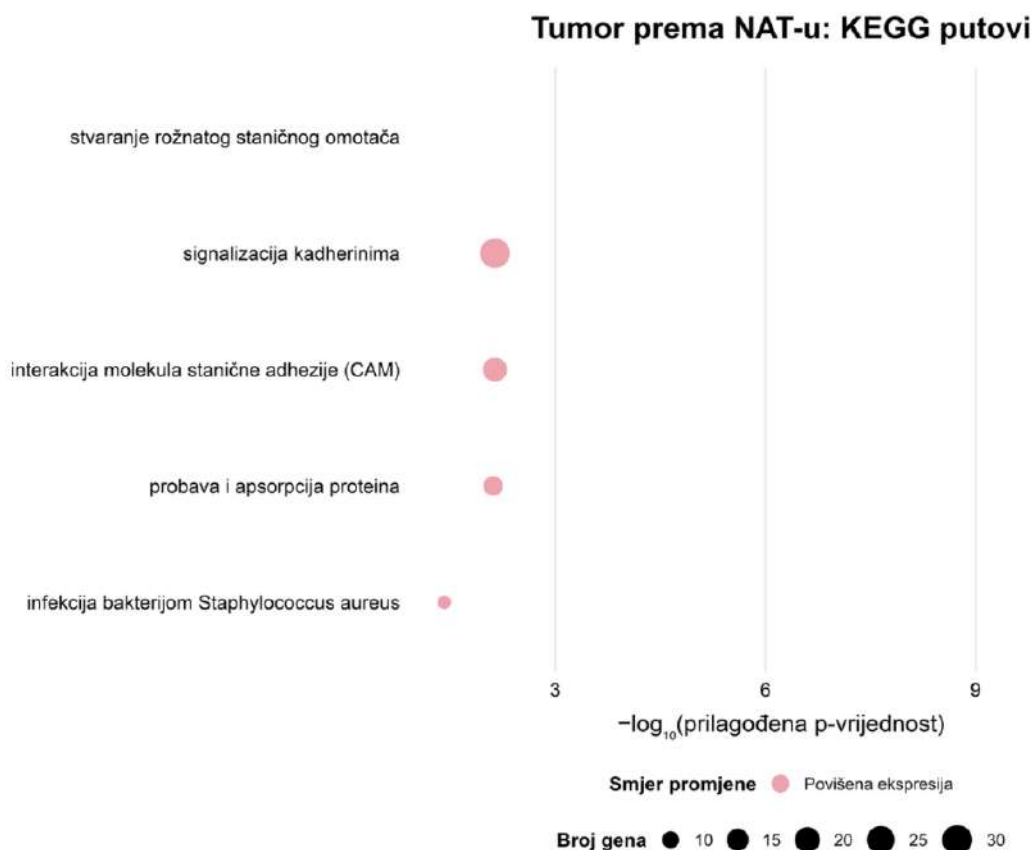
4.5.2 Usporedba T naspram NAT

U analizu T prema NAT uključeno je 899 zajedničkih gena nakon provedene meta-analize (595 povišene i 304 snižene ekspresije), uz pozadinski skup od 14 572 mapirana gena. Identificirana su 102 značajna *Gene Ontology* (GO) pojma bioloških procesa, 47 staničnih komponenti i 40 molekularnih funkcija. Najizraženije obogaćenje odnosilo se na organizaciju izvanstaničnog matriksa, razvoj epidermisa, intermedijarne filamente i morfogenezu tkiva (**Slika 22.**).



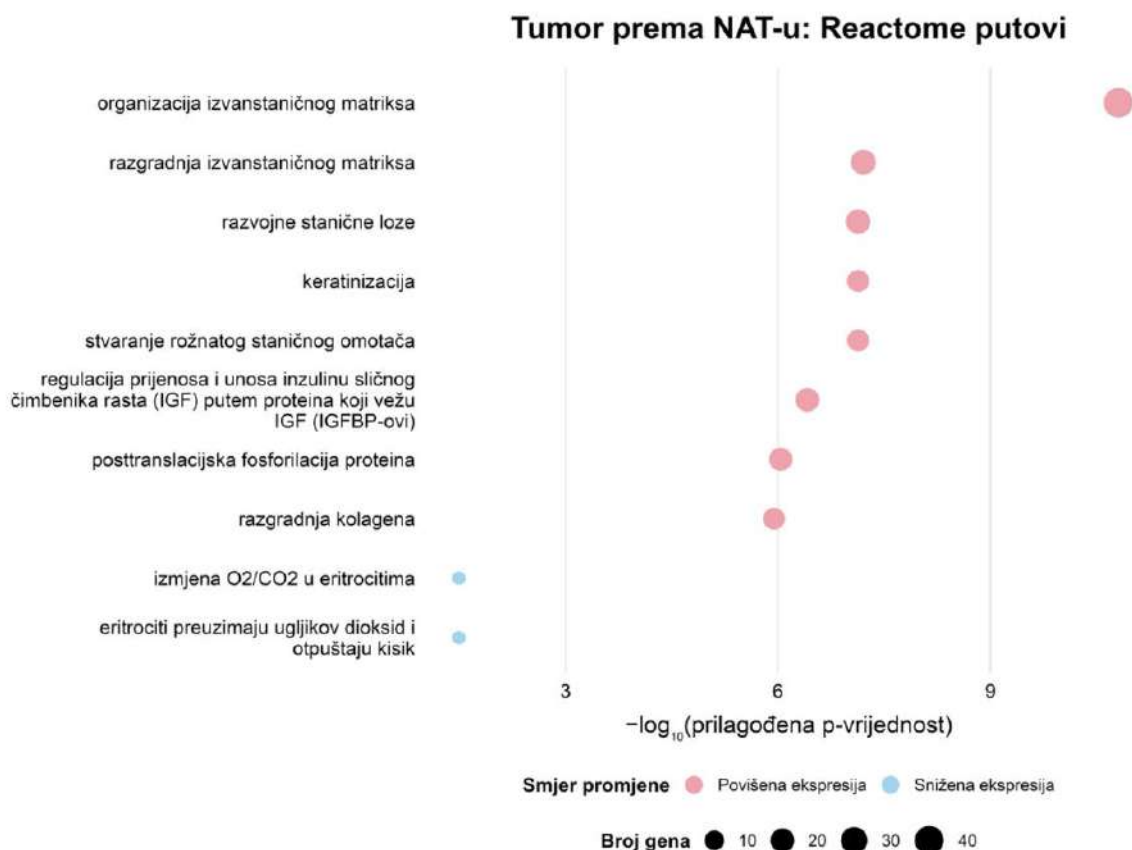
Slika 22. Obogaćenje GO bioloških procesa za zajedničke gene dobivene u meta-analizi usporedbe T prema NAT. Boja označava smjer promjene ekspresije (crveno su prikazani procesi obogaćeni među genima povišene, a svijetloplavo među genima snižene ekspresije u tumoru), položaj na osi x statističku značajnost, a veličina točke broj gena pridruženih pojedinom putu.

KEGG analizom pronađeno je pet značajno obogaćenih puteva i svi su pripadali genima povišene ekspresije. Izdvojeni su stvaranje orožnjelog omotača, signalizacija kadherinima, interakcija molekula stanične adhezije te probava i apsorpcija proteina (**Slika 23.**).



Slika 23. KEGG analiza obogaćenja za zajedničke gene dobivene u meta-analizi usporedbe T prema NAT. Svi prikazani značajni putovi obogaćeni su među genima povišene ekspresije u tumorskom tkivu. Boja označava smjer promjene ekspresije (crveno su prikazani procesi obogaćeni među genima povišene, a svijetloplavo među genima snižene ekspresije u tumoru), položaj na osi x statističku značajnost, a veličina točke broj gena pridruženih pojedinom putu.

Analizom signalnih puteva iz Reactome baze podataka identificirana su 33 značajno obogaćena puta. Među najbolje rangiranim bili su organizacija i razgradnja izvanstaničnog matriksa, razgradnja kolagena, keratinizacija, regulacija dostupnosti inzulinu sličnoga čimbenika rasta i posttranslacijska fosforilacija proteina (**Slika 24.**).



Slika 24. Ractome analiza obogaćenja za zajedničke gene dobivene u meta-analizi usporedbe T prema NAT. Boja označava smjer promjene ekspresije (crveno su prikazani procesi obogaćeni među genima povišene, a svijetloplavo među genima snižene ekspresije u tumoru), položaj na osi x statističku značajnost, a veličina točke broj gena pridruženih pojedinom putu.

Obje usporedbe pokazale su obogaćenje procesa povezanih s izvanstaničnim matriksom, staničnom adhezijom, razvojnim procesima i keratinizacijom. Usporedba T prema NAT bila je izraženije usmjerena na remodeliranje matriksa, dok je T prema H obuhvatila širi raspon promjena, uključujući citokinsku signalizaciju, cirkulaciju i hormonske procese.

4.6 Rana integracija RNA-seq skupova podataka

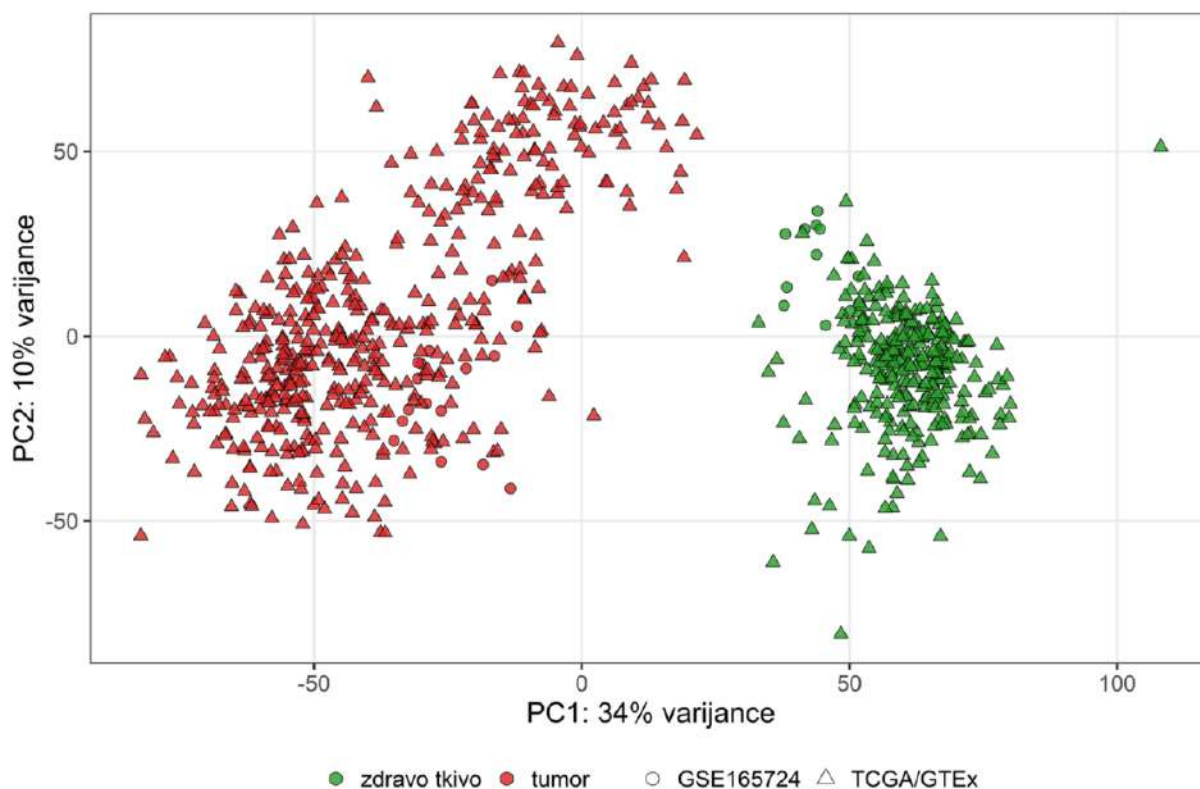
Rana integracija provedena je spajanjem matrica brojeva očitavanja pojedinačnih RNA-seq skupova prema zajedničkim Ensembl identifikatorima gena. Diferencijalna ekspresija analizirana je paketom *DESeq2* na cjelokupnoj objedinjenoj matrici. Svakom je uzorku pridružena oznaka izvornoga skupa podataka, koja je u modelu predstavljala učinak serije. Primijenjen je dizajn $\sim \text{batch} + \text{tkivo}$ (T, H ili NAT), čime je pri procjeni razlika između vrsta

tkiva uzet u obzir učinak studije iz koje uzorak potječe. Analize su zasebno provedene za usporedbe tumorskog i zdravog tkiva te tumorskog i NAT tkiva.

4.6.1 Rana integracija za usporedbu tumorskog i zdravog tkiva

U ranu integraciju za usporedbu T prema H uključeni su skupovi GSE165724 i TCGA/GTEX. Nakon usklađivanja bilo je dostupno 18 395 zajedničkih gena, od kojih je 15 715 zadržano nakon filtriranja nisko izraženih gena. Objedinjeni skup sadržavao je 765 uzoraka, odnosno 474 tumorska i 291 uzorak zdravog tkiva.

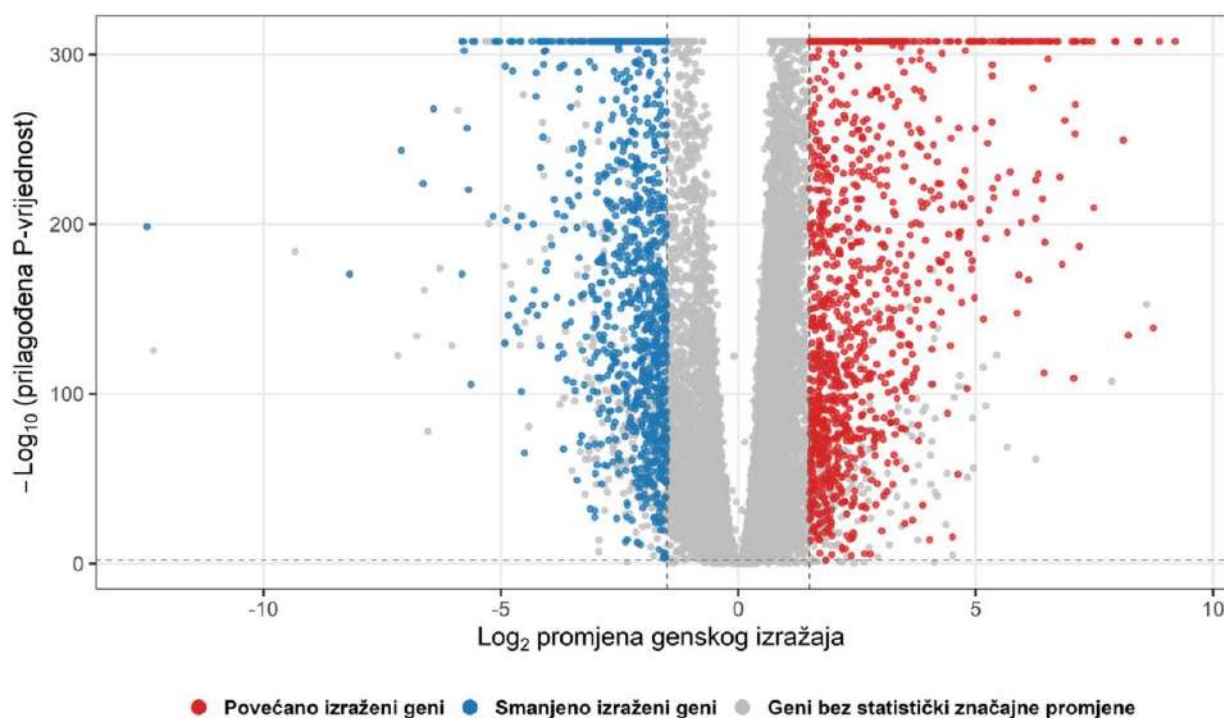
Nakon umanjivanja tehničkih razlika između uključenih skupova podataka, PCA prikaz (Slika 25.) jasno pokazuje razdvajanje tumorskih i zdravih uzoraka. Duž prve glavne komponente, PC1 objašnjava 34 % varijance, a PC2 objašnjava dodatnih 10 % varijance. Takvo grupiranje pokazuje da je vrsta tkiva bila jedan od glavnih izvora razlika u ekspresijskim profilima objedinjenih uzoraka.



Slika 25. PCA objedinjene analize tumorskoga i zdravoga tkiva. Bojama su označene vrste tkiva, a oblicima točaka izvorni skupovi podataka. PC1 objašnjava 34 %, a PC2 10 % ukupne varijance.

DESeq2 analizom identificirano je 2 049 statistički značajno diferencijalno izraženih gena, od kojih je 1 065 bilo povećano, a 984 smanjeno izraženo u tumoru (Slika 26.). Među

genima s najvećim povećanjem ekspresije nalazili su se *DCSTAMP*, *ADCY8*, *SLC18A3*, *TRPC5*, *ARHGAP36*, *CST1*, *PNPLA5* i *IGFL2*, dok su među genima s najvećim smanjenjem ekspresije bili *PRL*, *CHGA*, *DEFA1*, *PGC*, *SOX10*, *GP2*, *ADH1B* i *FCN2*. Geni su izdvojeni prema kriterijima $padj \leq 0,01$, $|\log_2 FC| \geq 1,5$ i $baseMean \geq 20$.

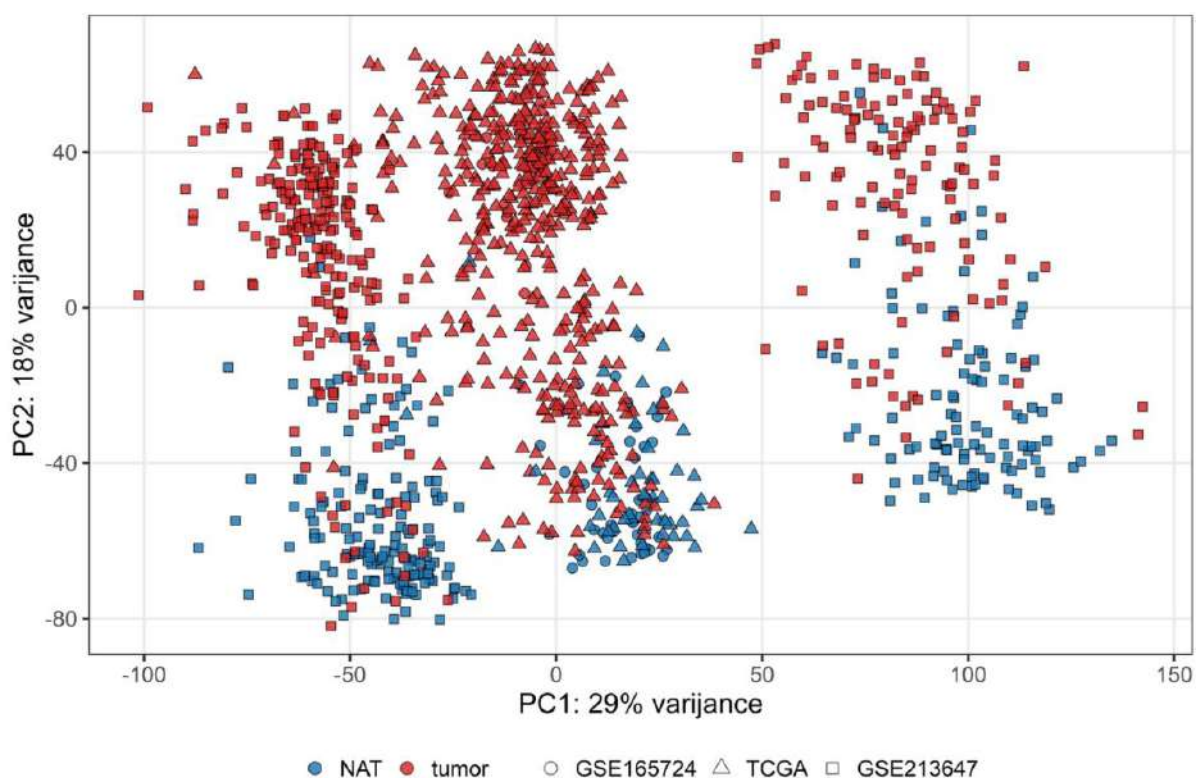


Slika 26. Vulkanski dijagram diferencijalne ekspresije između tumorskog i zdravog tkiva nakon rane integracije skupova GSE165724 i TCGA/GTEX. Crveno su prikazani povećano, plavo smanjeno izraženi geni, a sivo geni bez statistički značajne promjene.

4.6.2 Rana integracija za usporedbu tumorskog i NAT tkiva

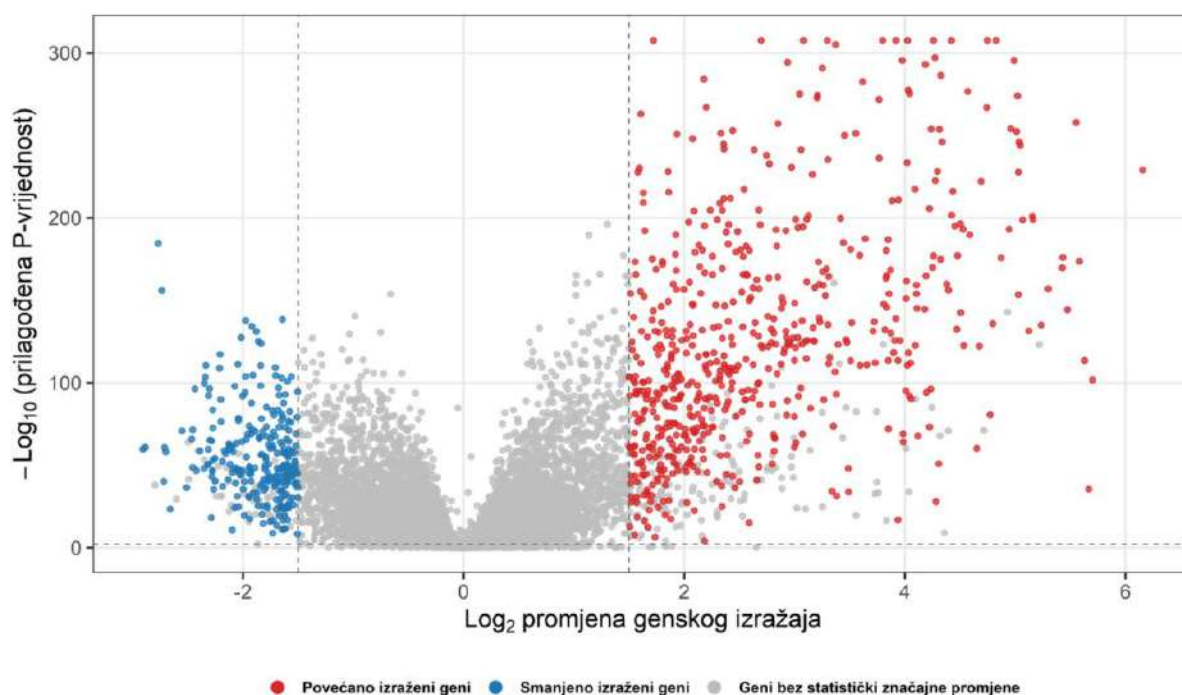
U ranu integraciju za usporedbu T prema NAT-u uključeni su skupovi GSE213647, GSE165724 i TCGA. Nakon usklađivanja identificirana su 18 393 zajednička gena, od kojih su 15 653 zadržana nakon filtriranja. Objedinjeni skup sadržavao je 1 181 uzorak, uključujući 821 tumorski i 360 NAT uzoraka.

Prva glavna komponenta PCA (**Slika 27.**) objašnjavala je 29 %, a druga 18 % ukupne varijance. Tumorski i NAT uzorci pokazivali su djelomično razdvajanje, ali uz veće međusobno preklapanje nego u prethodnoj usporedbi tumorskog i zdravog tkiva. To pokazuje da je ekspresijski profil NAT-a sličniji tumorskom profilu nego profil neovisnoga zdravog tkiva, uz prisutnu heterogenost među uzorcima i skupovima podataka.



Slika 27. PCA prikaz objedinjene analize tumorskog i NAT tkiva. Bojama su označene vrste tkiva, a oblicima točaka skupovi GSE165724, TCGA i GSE213647. PC1 objašnjava 29 %, a PC2 18 % ukupne varijance.

DESeq2 analizom identificiran je 861 statistički značajno diferencijalno izražen gen, od kojih je 600 bilo povećano, a 261 smanjeno izraženo u tumoru (**Slika 28.**). Među genima s najvećim povećanjem ekspresije nalazili su se *PRSS2*, *SLC18A3*, *KRT6B*, *KRT6A*, *IVL*, *TMPRSS6*, *ARHGAP36* i *VTCN1*, dok su među genima s najvećim smanjenjem bili *MYOC*, *SEC14L3*, *LYVE1*, *MMRN1*, *TFF2*, *TFF3*, *KCNA1* i *AGR3*. Geni su izdvojeni prema kriterijima $p_{adj} \leq 0,01$, $|\log_2FC| \geq 1,5$ i $baseMean \geq 20$.



Slika 28. Vulkanski dijagram diferencijalne ekspresije između tumorskog i NAT tkiva nakon rane integracije skupova GSE165724, TCGA i GSE213647. Identificiran je 861 značajno diferencijalno izražen gen, od kojih je 600 bilo povećano, a 261 smanjeno izraženo u tumoru.

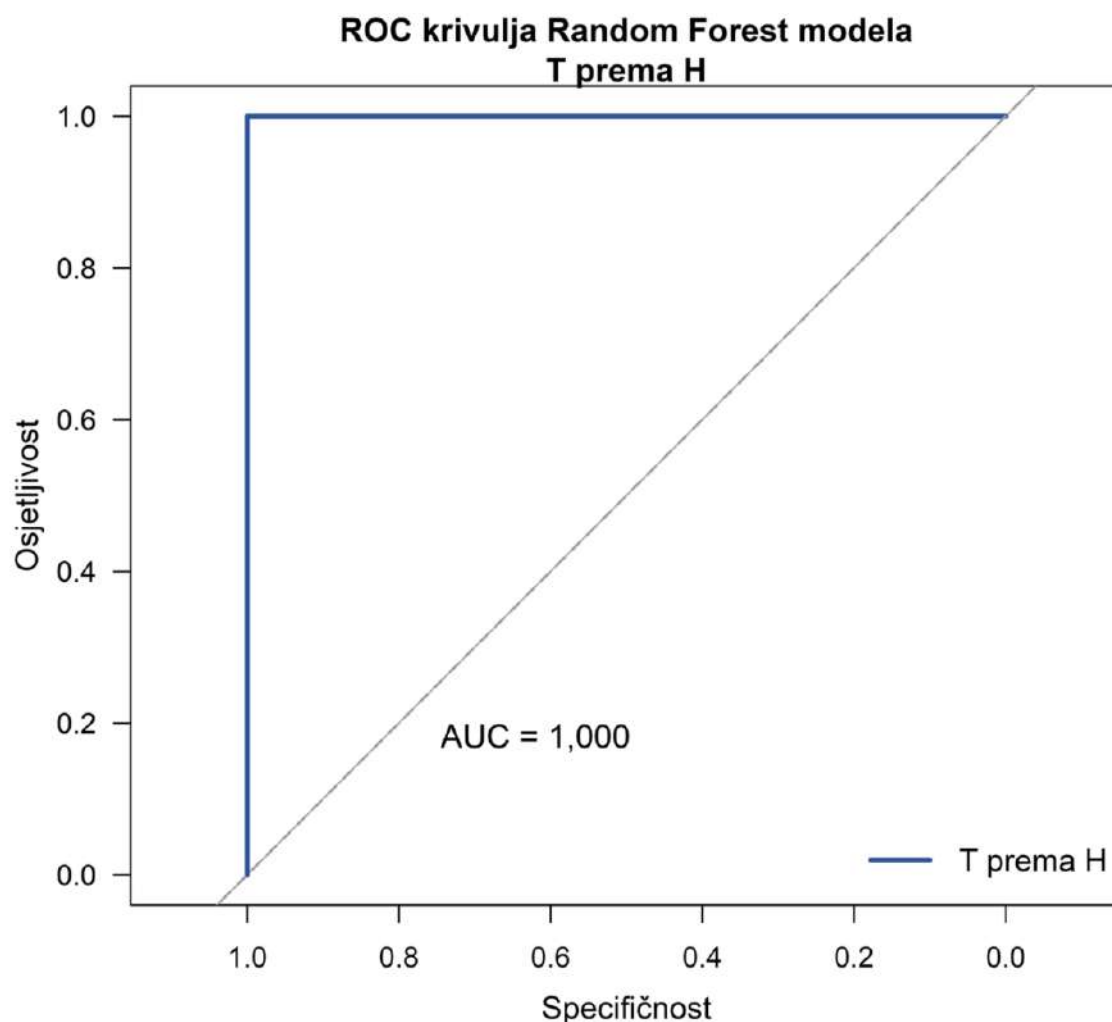
Rezultati rane integracije stoga pokazuju izražene promjene u genskoj ekspresiji između tumorskog i NAT tkiva, osobito u procesima epitelne diferencijacije, proteolize i organizacije tumorskog mikrookoliša. Međutim, nepotpuno razdvajanje uzoraka na PCA prikazu pokazuje da uz vrstu tkiva postoje i dodatni izvori varijabilnosti, što predstavlja ograničenje ovoga pristupa integraciji.

4.7 Modeli strojnog učenja

Modeli strojnog učenja su izrađeni odvojeno za usporedbe T prema H i T prema NAT ranom integracijom ekspresijskih matrica. U analizu T prema H uključeni su GSE165724 i TCGA/GTEx, a u analizu T prema NAT GSE165724, GSE213647 i TCGA. Podaci su stratificirano podijeljeni na skup za treniranje (70 %) i neovisni testni skup (30 %), pri čemu testni uzorci nisu korišteni za odabir gena ni treniranje modela.

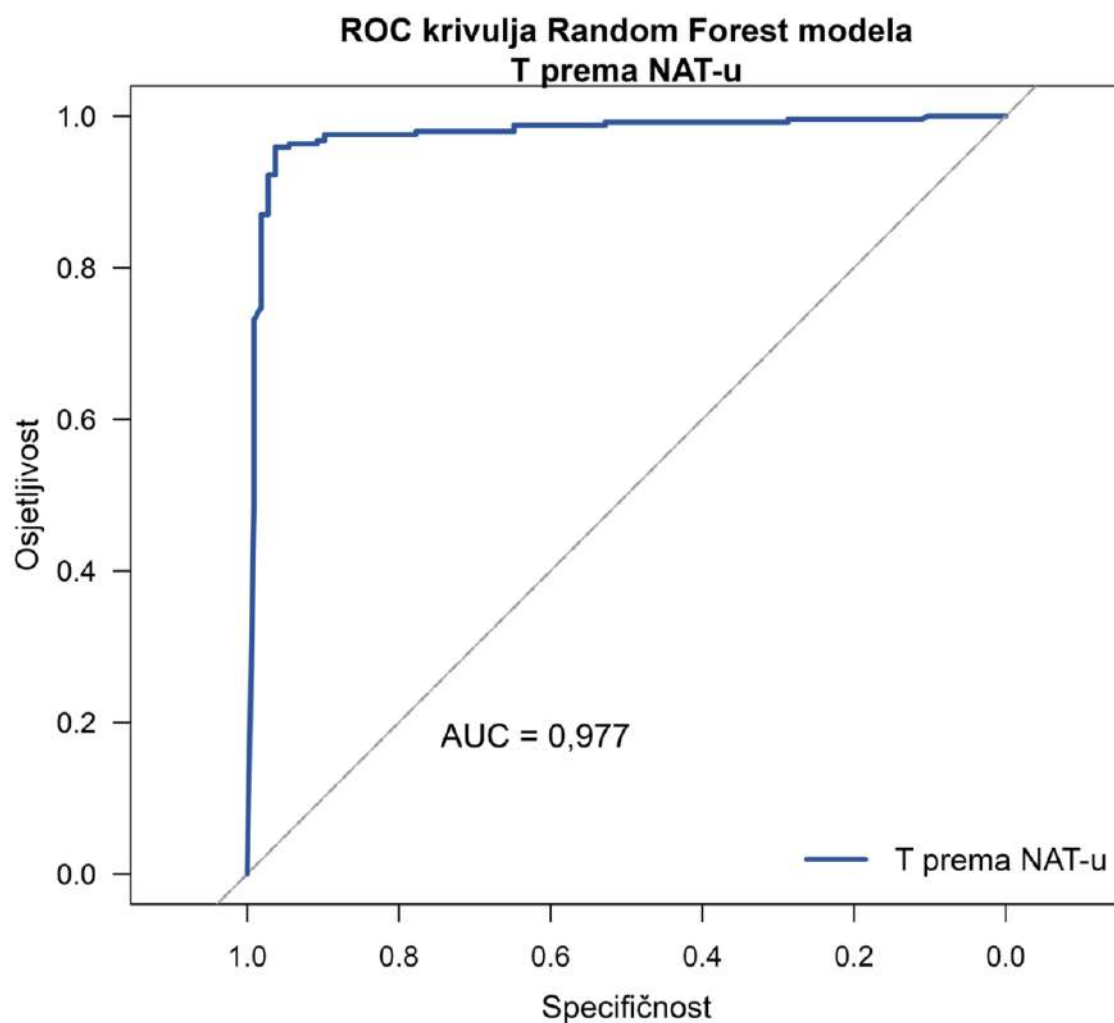
Za T prema H skup za treniranje sadržavao je 536 uzoraka (332 tumorska i 204 zdrava), a testni skup 229 uzoraka (142 tumorska i 87 zdravih). Analiza *DESeq2* provedena samo na

trening skupu izdvojila je 2 021 značajan gen; 300 najbolje rangiranih gena uključeno je u *Boruta* odabir najvažnijih gena, kojim je potvrđeno 229 značajki (gena). Model nasumičnih šuma ostvario je na testnom skupu točnost, osjetljivost, specifičnost i AUC od 1,000 (**Slika 29.**).



Slika 29. ROC krivulja modela nasumičnih šuma za razlikovanje tumorskog i zdravog tkiva na izdvojenom testnom skupu. Površina ispod krivulje iznosila je 1,000.

Za T prema NAT skup za treniranje sadržavao je 827 uzoraka (575 tumorskih i 252 NAT), a testni skup 354 uzorka (246 tumorskih i 108 NAT). U trening skupu identificirano je 848 značajnih gena; nakon *Boruta* analize od 300 kandidatnih gena zadržano je 109 značajki (gena). Model je na testnom skupu postigao točnost 0,955, osjetljivost 0,963, specifičnost 0,935 i AUC 0,977 (**Slika 30.**).



Slika 30. ROC krivulja modela nasumičnih šuma za razlikovanje tumorskoga i NAT tkiva na izdvojenom testnom skupu. Površina ispod krivulje iznosila je 0,977.

Prema relativnoj važnosti varijabli, model T prema H najviše se oslanjao na gene *CCND1*, *AP1S1*, *DRAXIN*, *CLDN7* i *CDH1*, dok su u modelu T prema NAT najvažniji bili *GABRB2*, *CDH3*, *KLHDC8A*, *LIPH* i *LRP4*. Deset najbolje rangiranih gena za svaki model prikazano je u **Tablici 1**.

Tablica 1. Deset gena s najvećom relativnom važnošću u modelima nasumičnih šuma.

T naspram H		T naspram NAT	
Gen	Važnost	Gen	Važnost
<i>CCND1</i>	100,00	<i>GABRB2</i>	100,00
<i>AP1S1</i>	77,40	<i>CDH3</i>	70,44
<i>DRAXIN</i>	66,76	<i>KLHDC8A</i>	40,03
<i>CLDN7</i>	65,52	<i>LIPH</i>	34,43
<i>CDH1</i>	64,36	<i>LRP4</i>	32,75
<i>FUCA1</i>	61,17	<i>GABRD</i>	31,47
<i>CORO2A</i>	60,67	<i>KCNQ3</i>	25,53
<i>TGFA</i>	59,82	<i>CDKN2B</i>	18,74
<i>PSD3</i>	58,06	<i>HMGA2</i>	17,59
<i>SDC4</i>	57,98	<i>LYVE1</i>	14,70

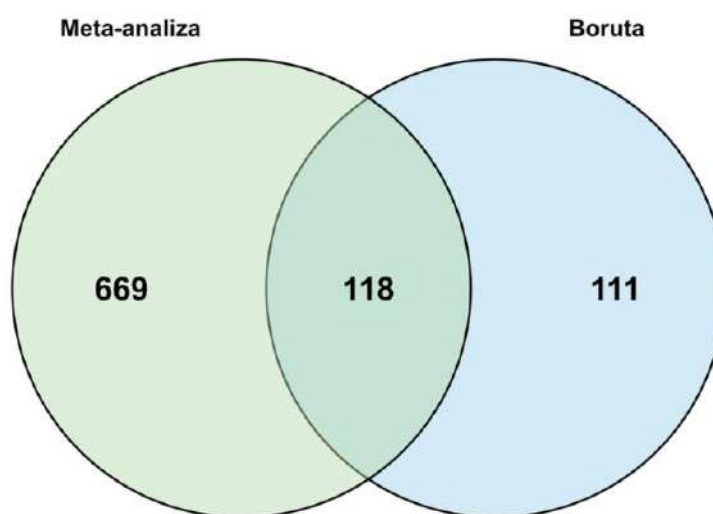
Oba su modela pokazala visoku sposobnost razlikovanja tumorskog od kontrolnog tkiva. Savršena klasifikacija u usporedbi T prema H odnosi se na izdvojeni testni skup ove analize, dok je model T prema NAT također ostvario vrlo visoku, ali ne savršenu klasifikacijsku uspješnost.

4.8 Usporedba rezultata meta-analize i strojnog učenja

Kako bi se procijenila podudarnost dvaju analitičkih pristupa, završni geni dobiveni meta-analizom uspoređeni su s genima odabranima metodom *Boruta*. Preklapanje je najprije analizirano za ukupan skup gena, a zatim zasebno za gene povećano i smanjeno izražene u tumorskom tkivu. Time je provjereno jesu li geni koji pokazuju konzistentnu diferencijalnu ekspresiju kroz više skupova podataka ujedno važni za klasifikaciju tumorskih i kontrolnih uzoraka.

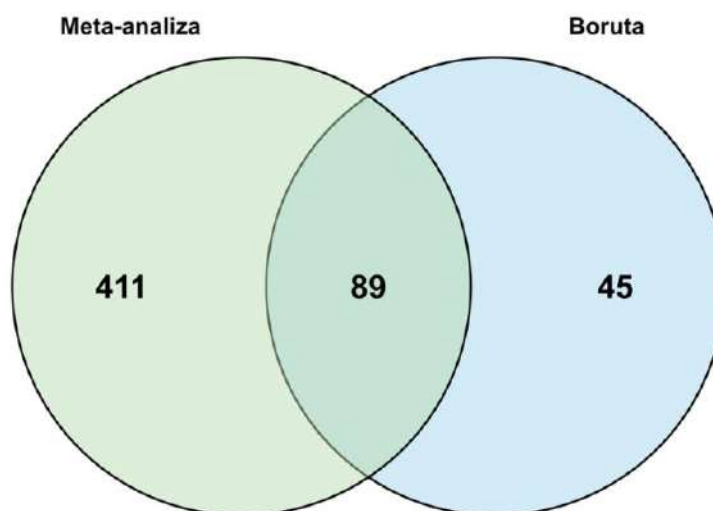
Za usporedbu T prema H pronađeno je ukupno 118 zajedničkih gena (**Slika 31.**). Analizom prema smjeru promjene utvrđeno je da je 89 zajedničkih gena bilo povećano izraženo (**Slika 32.**), dok je 29 gena bilo smanjeno izraženo u tumorskom tkivu (**Slika 33.**). Nijedan zajednički gen nije pokazao suprotan smjer promjene između meta-analize i *Boruta* analize.

T prema H: ukupno



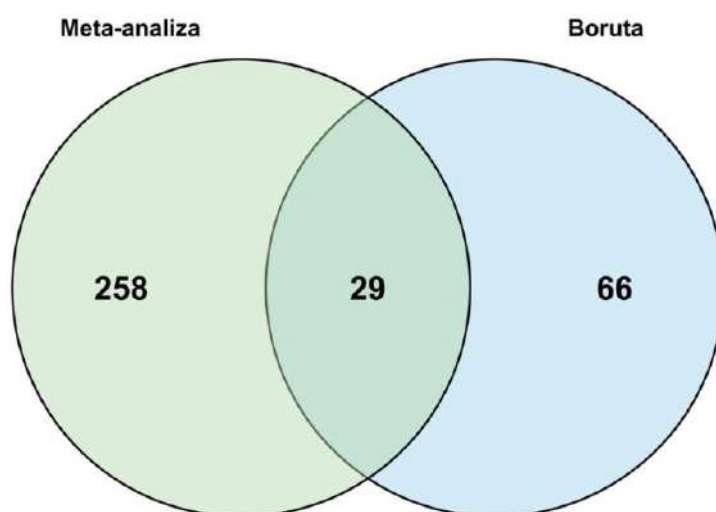
Slika 31. Vennov dijagram preklapanja završnih gena meta-analize i gena odabranih metodom Boruta za usporedbu T prema H. Pronađeno je ukupno 118 zajedničkih gen.

T prema H: povećano izraženi geni



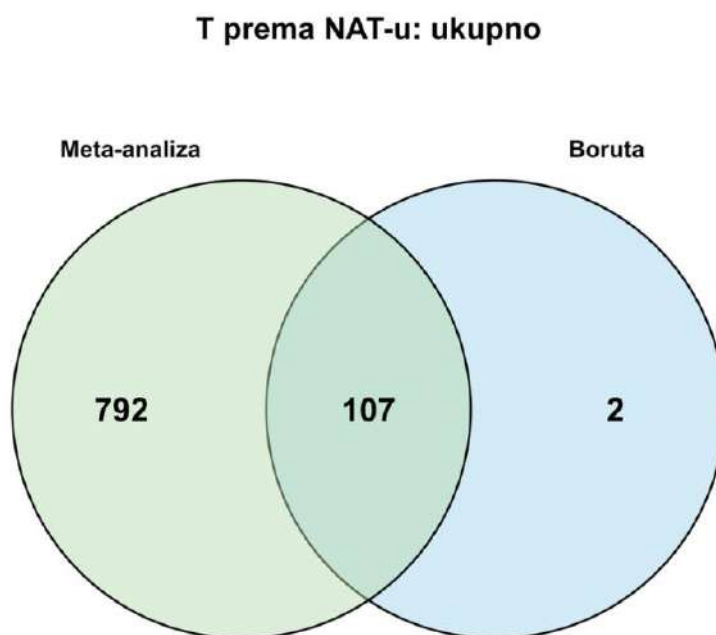
Slika 32. Vennov dijagram preklapanja gena povećano izraženih u tumorskom tkivu između meta-analize i Boruta analize za usporedbu T prema H. Pronađeno je 89 zajedničkih gena.

T prema H: smanjeno izraženi geni

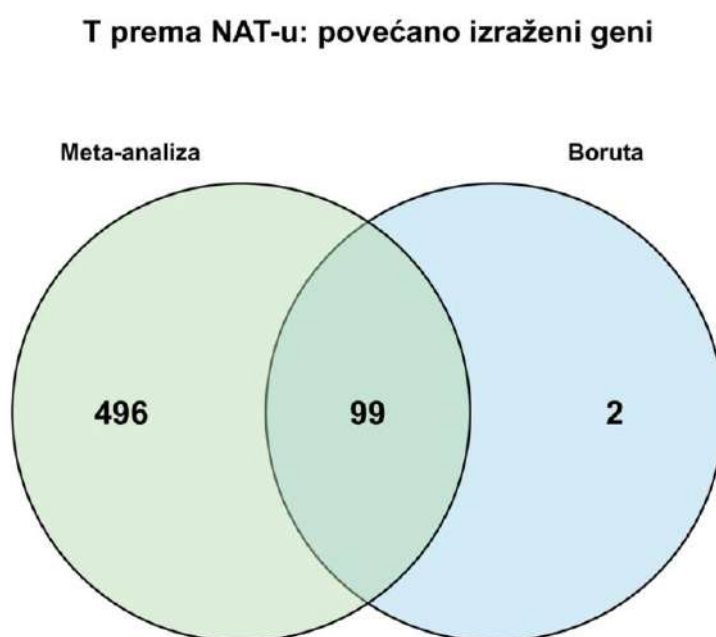


Slika 33. Vennov dijagram preklapanja gena smanjeno izraženih u tumorskom tkivu između meta-analize i Boruta analize za usporedbu T prema H. Pronađeno je 29 zajedničkih gena.

U usporedbi T prema NAT-u pronađeno je ukupno 107 gena zajedničkih meta-analizi i Boruta analizi (**Slika 34.**). Od toga je 99 gena bilo povećano izraženo u tumorskom tkivu (**Slika 35.**), dok je osam gena bilo smanjeno izraženo (**Slika 36.**). Ni u ovoj usporedbi nisu pronađeni geni sa suprotnim smjerom promjene između dvaju analitičkih pristupa.

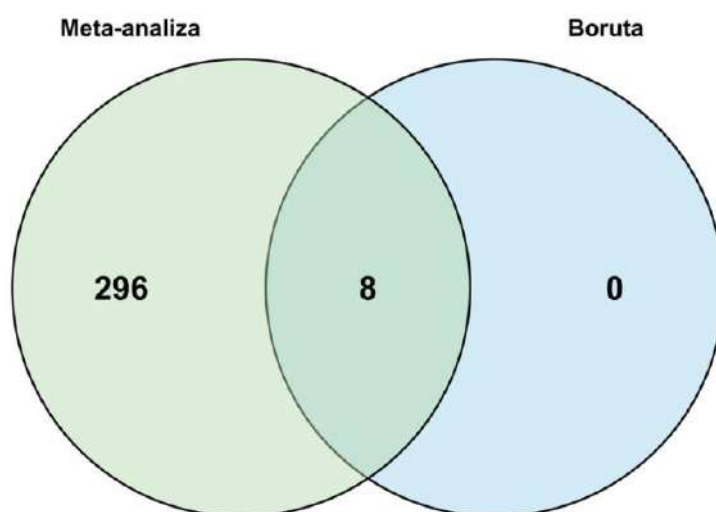


Slika 34. Vennov dijagram preklapanja završnih gena meta-analize i gena odabranih metodom Boruta za usporedbu T prema NAT-u. Pronađeno je ukupno 107 zajedničkih gena.



Slika 35. Vennov dijagram preklapanja gena povećano izraženih u tumorskom tkivu između meta-analize i Boruta analize za usporedbu T prema NAT-u. Pronađeno je 99 zajedničkih gena.

T prema NAT-u: smanjeno izraženi geni



Slika 36. Vennov dijagram preklapanja gena smanjeno izraženih u tumorskom tkivu između meta-analize i Boruta analize za usporedbu T prema NAT-u. Pronađeno je osam zajedničkih gena.

Dobivena preklapanja pokazuju da dio gena koji ima konzistentnu diferencijalnu ekspresiju kroz više skupova podataka ujedno pripada genima važnima za klasifikaciju tumorskih i kontrolnih uzoraka. Veći udio povećano izraženih gena u objema usporedbama upućuje na dominantnu zastupljenost gena aktiviranih u tumorskom tkivu među zajedničkim rezultatima meta-analize i strojnog učenja. Takvi geni predstavljaju posebno relevantne kandidate za daljnju procjenu mogućih biomarkera papilarnoga karcinoma štitnjače.

5. RASPRAVA

U ovom su radu analizirani javno dostupni RNA-seq skupovi podataka kako bi se odredile transkriptomске razlike između papilarnog karcinoma štitnjače i kontrolnog tkiva te identificirali potencijalni transkriptomski biomarkeri. Analiza je obuhvatila usporedbu tumorskog tkiva sa zdravim tkivom štitnjače (T prema H) i tumoru susjednim histološki normalnim tkivom (T prema NAT). Primjenom pojedinačnih analiza diferencijalne ekspresije, rane integracije, meta-analize i strojnog učenja bilo je moguće procijeniti stabilnost dobivenih rezultata te odrediti u kojoj mjeri izbor kontrolnog tkiva utječe na konačan popis gena i uspješnost klasifikacije.

Analiza glavnih komponenti (PCA) pokazala je da se ekspresijski profili tumorskih i kontrolnih uzoraka razlikuju, ali i da opseg razdvajanja ovisi o analiziranom skupu podataka i vrsti kontrole. U skupu GSE165724 tumorski i zdravi uzorci jasno su se razdvojili duž prve glavne komponente, koja je objašnjavala 59 % varijance, dok je druga komponenta objašnjavala dodatnih 11 %. Takav raspored pokazuje da je razlika između tumorskog i zdravog tkiva bila glavni izvor varijabilnosti u tom skupu.

Razdvajanje tumorskog i NAT tkiva u skupu GSE165724 također je bilo vidljivo, pri čemu su prve dvije glavne komponente zajedno objašnjavale 66 % varijance. Usporedba NAT-a i zdravog tkiva pokazala je da se i te dvije netumorske skupine međusobno razlikuju. Taj nalaz upućuje na to da histološki normalno tkivo uz tumor nije transkriptomski istovjetno tkivu štitnjače zdravih osoba.

U skupu GSE213647 razdvajanje tumorskih i kontrolnih uzoraka bilo je manje potpuno. Prve dvije glavne komponente objašnjavale su 64 % ukupne varijance, ali je između skupina postojalo djelomično preklapanje. To pokazuje da uz vrstu tkiva postoje i drugi izvori varijabilnosti. Oni mogu uključivati biološku heterogenost tumora, različite udjele stromalnih, endotelnih i imunosnih stanica, kliničke razlike među uzorcima i tehničke čimbenike. Budući da su kontrolni uzorci u tom skupu tretirani kao NAT, njihova djelomična sličnost s tumorskim uzorcima podudara se s rezultatima drugih analiza u ovom radu.

Zajednički PCA prikaz TCGA/GTEX podataka dodatno je pokazao razliku između triju vrsta tkiva. Zdravi uzorci grupirali su se pretežno na jednoj, a tumorski na drugoj strani prve glavne komponente, koja je objašnjavala 50 % varijance. Većina NAT uzoraka nalazila se između tih skupina ili bliže zdravom tkivu, ali uz pomak prema tumorskom profilu. Takav raspored može se opisati kao prijelazan transkriptomski položaj NAT-a između zdravog i

tumorskog tkiva. Ipak, NAT uzorci nisu tvorili potpuno homogenu skupinu, zbog čega se njihov profil ne može smatrati jednostavnom sredinom između H i T [26].

U ranoj integraciji objedinjene su matrice brojeva očitavanja iz pojedinačnih studija. Usporedba T prema H uključivala je skupove GSE165724 i TCGA/GTEx, dok su u usporedbu T prema NAT-u uključeni GSE165724, TCGA i GSE213647. Kako bi se u statističkoj analizi uzeo u obzir izvor podataka, u *DESeq2* model uključeni su vrsta tkiva i učinak studije. Time je omogućeno određivanje razlika povezanih s tumorskim statusom uz istodobno kontroliranje dijela varijabilnosti povezane s izvornim skupom podataka.

PCA prikaz rane integracije T prema NAT-u ipak nije pokazao potpuno razdvajanje tumorskih i NAT uzoraka. Osim boje, koja je označavala vrstu tkiva, oblik točaka pokazivao je izvorni skup podataka. Dio uzoraka grupirao se prema studiji, što znači da su nakon korekcije ostale razlike koje se nisu mogle potpuno opisati varijablom učinka studije. To nije razlog za odbacivanje analize, nego ograničenje koje treba uzeti u obzir pri tumačenju rezultata rane integracije.

Utjecaj vrste kontrole jasno se vidi u broju diferencijalno izraženih gena. U analizi TCGA/GTEx identificirano je 2 150 značajno diferencijalno izraženih gena između tumorskog i zdravog tkiva, dok su između tumorskog i NAT tkiva iz TCGA identificirana 1 204 gena. Sličan odnos dobiven je ranom integracijom. U usporedbi T prema H identificirano je 2 049 gena, od kojih je 1 065 bilo povećano, a 984 smanjeno izraženo u tumoru. U usporedbi T prema NAT-u identificiran je 861 gen, od kojih je 600 bilo povećano, a 261 smanjeno izraženo.

Veći broj diferencijalno izraženih gena u usporedbi T prema H pokazuje da je ekspresijski profil tumora udaljeniji od profila neovisnog zdravog tkiva nego od profila NAT-a. Tumor i NAT potječu od istoga bolesnika, nalaze se u istom anatomskom području i dijele dio bioloških i tehničkih obilježja. Zbog toga su njihove međusobne razlike manje od razlika između tumora i zdravog tkiva iz neovisnih izvora. Ipak, taj se brojčani odnos nije izravno ponovio u meta-analizi. Za T prema H dobiveno je 787 završnih gena, od kojih je 500 bilo povećano, a 287 smanjeno izraženo u tumoru. Za T prema NAT-u identificirano je 899 gena, od kojih je 595 bilo povećano, a 304 smanjeno izraženo. Razlika se može objasniti načinom provođenja meta-analize. Meta-analiza T prema NAT-u uključivala je tri skupa podataka, dok su u analizu T prema H uključena dva. Broj završnih gena zato ne ovisi samo o veličini razlike između tkiva, nego i o broju studija, njihovoj veličini i dosljednosti promjena kroz pojedinačne analize.

Izravnom usporedbom NAT tkiva iz TCGA sa zdravim tkivom štitnjače iz GTEx-a identificirano je 1 247 diferencijalno izraženih gena, od kojih je 280 bilo povećano, a 967

zmanjeno izraženo u NAT-u. Ovaj rezultat potvrđuje da NAT nije transkriptomski jednak zdravom tkivu. Budući da je NAT uzet iz neposredne blizine tumora, njegov molekularni profil može biti promijenjen djelovanjem lokalnog mikrokoliša. Parakrini signali, lokalna upala, promjene izvanstaničnog matriksa, aktivacija stromalnih stanica i promijenjena zastupljenost različitih staničnih populacija mogu pridonijeti razlikama između NAT-a i zdravog tkiva. Među genima povećano izraženima u NAT-u izdvojili su se *CDH1* i *CXADR*, povezani s međustaničnim vezama i organizacijom epitela, te transkripcijski regulatori *FOSB* i *EGR2*. Promjene gena *FCAMR* i *LGALS2* mogu upućivati i na razlike u imunološkoj komponenti tkiva. Ti rezultati pokazuju da se razlike između NAT-a i zdravog tkiva ne odnose samo na pojedinačne gene, nego uključuju organizaciju epitela, regulaciju transkripcije i lokalni stanični sastav.

Prednost NAT-a kao kontrole proizlazi iz činjenice da tumor i NAT potječu od istoga bolesnika. Time se smanjuje varijabilnost povezana s konstitucijskom genetskom pozadinom, dobi, spolom i drugim obilježjima pojedinca. Uzorci se također najčešće prikupljaju i obrađuju u jednakim uvjetima. Međutim, zbog promjena u tkivu koje okružuje tumor NAT ne bi trebalo biti jedina kontrola. Ako su određene rane molekularne promjene prisutne i u NAT-u, one se u usporedbi T prema NAT-u mogu smanjiti ili potpuno izgubiti.

Zdravo tkivo daje širu fiziološku referencu i omogućuje prepoznavanje većeg broja promjena u odnosu na normalnu funkciju štitnjače. Međutim, ono uvodi dodatnu varijabilnost među pojedincima i moguće razlike povezane s izvorom podataka. NAT i H zato odgovaraju na različita pitanja. Usporedba T prema H pokazuje širi odmak tumora od fiziološkog transkriptoma, dok usporedba T prema NAT-u naglašava promjene specifične za tumor u odnosu na njegovo neposredno okruženje. Uključivanje obiju vrsta kontrole omogućilo je potpunije tumačenje tumorskog ekspresijskog profila.

Funkcionalna analiza završnih meta-analitičkih gena pokazala je da su u obje usporedbe obogaćeni procesi povezani s organizacijom i razgradnjom izvanstaničnog matriksa, staničnom adhezijom, kolagenom, morfogeneza tkiva, intermedijarnim filamentima i keratinizacijom. To pokazuje da su uz promjene ekspresije pojedinačnih gena u PTC-u prisutne opsežnije promjene u organizaciji tumorskog tkiva i njegova mikrokoliša. Izvanstanični matriks čini strukturno i signalno okruženje stanica. Njegovo remodeliranje može promijeniti mehanička svojstva tkiva, veze između stanica i njihovih receptora te uvjete u kojima se odvijaju migracija i proliferacija. Istodobno obogaćenje organizacije i razgradnje matriksa pokazuje da se tijekom tumorske preobrazbe ne mijenja samo količina njegovih sastavnica, nego i njihova struktura i

dinamika. Obogaćenje razgradnje kolagena dodatno upućuje na promjene koje mogu olakšati preoblikovanje tkiva i kretanje stanica kroz okolni prostor [50].

Obogaćenje procesa stanične adhezije može se povezati s promjenama veza među tumorskim stanicama i njihova odnosa s matriksom. Organizirani stanični spojevi važni su za održavanje strukture i polariteta folikularnog epitela. Promjene gena povezanih s kadherinima, adhezijskim molekulama i interakcijama matriksa i receptora zato mogu odražavati gubitak dijela normalne organizacije epitelnih stanica i uspostavljanje drukčijih međustaničnih odnosa u tumoru[51]. Procesi označeni kao keratinizacija i stvaranje rožnatog staničnog omotača ne znače da u tumorskom tkivu nastaje doslovno histološko orožnjavanje. Ti funkcionalni pojmovi uključuju gene keratina, intermedijarnih filamenata i epitelne diferencijacije. Njihovo obogaćenje u ovoj analizi može se tumačiti kao promjena citoskeleta, epitelne organizacije i diferencijacije tumorskih stanica[52].

U usporedbi T prema H dobiven je širi raspon funkcionalnih promjena. Uz izvanstanični matriks, kolagen i adheziju, obogaćeni su bili procesi povezani s citokinskom signalizacijom, hormonima, cirkulacijom, komplementom, koagulacijom i funkcijama diferenciranoga tkiva. Takav rezultat odgovara većem broju identificiranih gena i većoj transkriptomskoj udaljenosti tumora od neovisnog zdravog tkiva.

U usporedbi T prema NAT-u dominantni su bili lokalni procesi organizacije i razgradnje izvanstaničnog matriksa, razgradnje kolagena, stanične adhezije, keratinizacije te regulacije dostupnosti inzulinu sličnoga čimbenika rasta. Tumor i NAT dijele dio lokalnog mikrookoliša pa su u ovoj usporedbi izraženije promjene neposredno povezane s tumorom i lokalnim remodeliranjem tkiva. Usporedba T prema NAT-u nije ograničena isključivo na izvanstanični matriks, ali je na njega više usmjerena od usporedbe T prema H.

Usporedbom završnih rezultata meta-analiza identificirano je 525 zajedničkih gena. Svih je 525 gena zadržalo jednak smjer promjene u objema analizama: 388 gena bilo je povećano, a 137 smanjeno izraženo u tumorskom tkivu. Nije pronađen nijedan gen koji bi u jednoj usporedbi bio povećano, a u drugoj smanjeno izražen. Taj rezultat pokazuje da odabir kontrole utječe na ukupan broj i statističku izraženost promjena, ali ne mijenja smjer zajedničkog tumorskog signala. Zajedničkih 525 gena može se opisati kao transkriptomska jezgra PTC-a koja je prepoznata neovisno o tome je li kao kontrola korišteno zdravo ili NAT tkivo. To ne znači da su svi ti geni specifični isključivo za PTC niti da svaki ima izravnu ulogu u nastanku tumora. Dio promjena može odražavati proliferaciju, gubitak diferencijacije, upalu, remodeliranje matriksa ili promijenjeni stanični sastav. Njihova vrijednost proizlazi iz dosljednosti smjera promjene u dvjema različitim vrstama usporedbe.

Rezultati meta-analize uspoređeni su i s genima izdvojenima *Boruta* analizom, koja na temelju modela nasumičnih šuma identificira sve relevantne gene za klasifikaciju, izdvajajući one čija je važnost statistički značajno veća od nasumičnog šuma. U usporedbi T prema H preklapalo se 118 gena, od kojih je 89 bilo povećano, a 29 smanjeno izraženo u tumoru. U usporedbi T prema NAT-u preklapalo se 107 gena, od kojih je 99 bilo povećano, a 8 smanjeno izraženo. Ni u jednoj usporedbi nisu pronađeni geni sa suprotnim smjerom promjene. Takva podudarnost pokazuje da je dio gena istodobno dosljedno promijenjen kroz studije i informativan za razlikovanje tumorskog od kontrolnog tkiva.

Meta-analiza i *Boruta* pritom imaju različite uloge. Meta-analiza traži gene koji pojedinačno pokazuju statistički značajnu i dosljednu promjenu ekspresije među studijama [53]. *Boruta* procjenjuje važnost gena za klasifikaciju kada se promatraju zajedno s ostalim značajkama. Ona ne traži nužno najmanji mogući skup gena, nego sve značajke koje nose relevantnu klasifikacijsku informaciju. Zbog korelacije ekspresije među genima, gen s najvećom promjenom ekspresije ne mora nužno imati i najveći dodatni doprinos klasifikaciji [54].

Kada se usporedba ograničila na deset vodećih značajki modela nasumičnih šuma i 30 vodećih gena meta-analize, za T prema H nije bilo preklapanja, dok su se za T prema NAT-u preklapili geni *GABRB2* i *LYVE1*. Odsutnost preklapanja na samom vrhu ljestvice T prema H ne znači da se dva pristupa ne slažu jer je na razini šireg popisa pronađeno 118 zajedničkih gena. Ona pokazuje da statistička značajnost promjene i doprinos multivarijantnoj klasifikaciji nisu jednaka svojstva. Nasuprot tomu, istodobno visoko rangiranje *GABRB2* i *LYVE1* u usporedbi T prema NAT-u daje dodatnu težinu njihovom odabiru.

Najvažnija značajka modela T prema H bio je *CCND1*, poznat onkogen koji kodira ciklin D1, regulator prijelaza stanice iz faze G1 u fazu S staničnoga ciklusa. Njegovo izdvajanje odgovara izraženom proliferacijskom signalu tumorskoga tkiva. Budući da je regulacija staničnog ciklusa povezana s MAPK signalizacijom, visoka važnost *CCND1* može odražavati pojačanu aktivaciju proliferacijskih programa u tumorskim uzorcima [55]. Međutim, iz ove analize nije moguće zaključiti da je promjena njegove ekspresije izravno uzrokovana određenom mutacijom ili signalnim putem.

Među važnim značajkama modela T prema H nalazio se i *DUSP6*, regulator ERK/MAPK signalizacije. Njegovo izdvajanje dodatno povezuje rezultate modela s promjenama proliferacijskih signalnih puteva. *CLDN7* i *CDH1* sudjeluju u organizaciji epitelnih spojeva, dok *SDC4* povezuje stanice s izvanstaničnim matriksom. Važnost tih gena podudara se s funkcionalnim obogaćenjem procesa stanične adhezije i organizacije matriksa.

Vodeća značajka modela T prema NAT-u bio je *GABRB2*. Njegovo izdvajanje u modelu nasumičnih šuma pokazuje da ekspresija toga gena snažno pridonosi razlikovanju tumorskog i NAT tkiva. Budući da se *GABRB2* pojavio i među visoko rangiranim kandidatima meta-analize T prema NAT-u, njegov signal nije bio ograničen na samo jedan analitički pristup. U ovom se radu zato može opisati kao robustan kandidat za razlikovanje tumorskog tkiva od njegova neposrednog okruženja, a već je od prije poznat u literaturi kao potencijalno važan biomarker upravo za PTC [56], [57]. Među vodećim genima modela T prema NAT-u nalazio se i *HMGA2*[58] , regulator kromatinske strukture i genske transkripcije. Njegova povećana važnost može se povezati s promjenama proliferacije, diferencijacije i staničnoga identiteta. *CDH3* [59] kodira adhezijsku molekulu P-kadherin, dok je *CDKN2B* [60]regulator staničnoga ciklusa. Njihovo zajedničko izdvajanje ponovno povezuje klasifikacijski signal s promjenama proliferacije i epitelne organizacije. Gen *LYVE1* , koji je bio smanjeno izražen u tumoru, povezan je s limfnim endotelom. Njegova promjena može odražavati reorganizaciju limfnih struktura ili promijenjenu zastupljenost endotelnih stanica u tumorskom tkivu. Budući da je analiziran *bulk* RNA-seq, nije moguće odrediti potječe li promjena iz tumorskih stanica ili iz razlike u staničnom sastavu. Ipak, preklapanje *LYVE1* između vodećih meta-analitičkih i klasifikacijskih rezultata pokazuje da se radi o dosljednom signalu u ovoj analizi.

Strojno učenje pokazalo je vrlo visoku uspješnost klasifikacije. Model T prema H na testnom je skupu postigao točnost, osjetljivost, specifičnost i AUC od 1,00. Ipak, rezultat od 1,00 ovdje valja tumačiti s oprezom te se rezultat treba validirati u neovisnim skupovima podataka. Model T prema NAT-u postigao je točnost od 0,955, osjetljivost od 0,963, specifičnost od 0,935 i AUC od 0,977. Niža, ali još uvijek vrlo visoka uspješnost modela T prema NAT-u očekivana je jer NAT dijeli dio molekularnih obilježja s tumorskim tkivom. Razlikovanje tumora od NAT-a zato predstavlja zahtjevniji klasifikacijski problem od razlikovanja tumora od zdravog tkiva.

Važna metodološka prednost rada jest to što je podjela uzoraka na skup za učenje i testni skup provedena prije odabira gena. *DESeq2* analiza, odabir 300 vodećih gena, *Boruta* selekcija, standardizacija i učenje modela nasumičnih šuma provedeni su isključivo na 70 % uzoraka namijenjenih učenju. Preostalih 30 % uzoraka nije korišteno u odabiru značajki ni izgradnji modela, nego samo za završnu evaluaciju. Time je spriječen prijenos informacija iz testnog skupa u postupak odabira značajki i smanjen rizik umjetnog povećanja procijenjene uspješnosti modela.

Visoku klasifikacijsku uspješnost modela T prema H potrebno je sagledati u kontekstu načina njegove evaluacije. Iako testni uzorci nisu sudjelovali u odabiru značajki ni treniranju

modela, izdvojeni su iz istih studija kao i uzorci za učenje. Provedena evaluacija stoga predstavlja internu provjeru modela, dok bi za potvrdu njegove primjenjivosti bila potrebna validacija na potpuno neovisnoj skupini uzoraka. Dodatno ograničenje proizlazi iz djelomične povezanosti vrste tkiva s izvorom podataka, osobito u usporedbi tumorskih uzoraka iz projekta TCGA sa zdravim uzorcima iz projekta GTEx. Zbog toga se ne može isključiti mogućnost da model, osim bioloških razlika između tumorskog i zdravog tkiva, djelomično prepoznaje i tehničke razlike povezane s izvornim projektima.

Dodatno ograničenje predstavlja nedostatak ujednačenih kliničkih podataka. Nije bilo moguće sustavno procijeniti utjecaj stadija, veličine tumora, histološke varijante, mutacijskog statusa, metastaza, recidiva i odgovora na terapiju. Zbog toga se ne može zaključiti jesu li izdvojeni geni povezani samo s prisutnošću tumora ili i s njegovom agresivnošću. Također nije bilo moguće ispitati bi li uključivanje kliničkih parametara dodatno poboljšalo klasifikacijsku moć modela.

Ograničenje predstavlja i uporaba *bulk* RNA-seq podataka. Ekspresijski profil pojedinog uzorka predstavlja prosjek svih stanica prisutnih u tkivu. Povećana ili smanjena ekspresija zato može nastati zbog promjene aktivnosti gena unutar tumorskih stanica, ali i zbog različitog udjela fibroblasta, imunskih i endotelinih stanica. To je posebno važno pri tumačenju gena povezanih s izvanstaničnim matriksom, upalom i krvnim ili limfnim žilama.

Za procjenu stvarne dijagnostičke vrijednosti modela potrebna je potvrda na potpuno neovisnim uzorcima. Posebno bi bilo važno ispitati zadržavaju li izdvojeni geni klasifikacijsku vrijednost kada se PTC uspoređuje s benignim čvorovima i citološki neodređenim uzorcima. Razlikovanje tumorskog od potpuno zdravog tkiva jednostavniji je zadatak od razlikovanja malignih i benignih proliferativnih promjena.

Unatoč ograničenjima, kombiniranje pojedinačnih analiza, rane integracije, meta-analize i strojnog učenja predstavlja važnu snagu ovog rada. Rana integracija omogućila je izgradnju zajedničkih klasifikacijskih modela na većem broju uzoraka, dok je meta-analiza omogućila izdvajanje promjena koje su bile dosljedne među neovisnim studijama. Preklapanje meta-analitičkih gena i značajki koje je izdvojila *Boruta* dodatno je suzilo popis kandidata na gene koji su istodobno statistički konzistentni i informativni za klasifikaciju.

Sveukupno, rezultati pokazuju da papilarni karcinom štitnjače ima snažan i ponovljiv transkriptomski profil koji obuhvaća promjene proliferacijskih programa, epitelne organizacije, stanične adhezije i izvanstaničnog matriksa. Istodobno je pokazano da izbor kontrolnog tkiva bitno utječe na veličinu i funkcionalno značenje dobivenog signala. NAT

smanjuje varijabilnost među pojedincima, ali zbog promjena povezanih s lokalnim tumorskim mikrookolišem ne predstavlja potpuno zdravo tkivo.

Zajedničkih 525 gena iz dviju meta-analiza predstavlja skup dosljedno promijenjenih gena koji nije ovisio o vrsti kontrole. Preklapanje 118 gena između meta-analize i *Borute* za T prema H te 107 gena za T prema NAT-u dodatno je izdvojilo kandidate koji su istodobno statistički i prediktivno relevantni. Među njima se posebno ističu *CCND1*, *DUSP6*, *GABRB2*, *HMGA2* i *LYVE1*. Njihovo izdvajanje u više analiza pokazuje da dobiveni klasifikacijski rezultati nisu temeljeni na nasumičnim značajkama, nego na genima povezanim sa staničnim ciklusom, signalizacijom, diferencijacijom, adhezijom i tumorskim mikrookolišem. Iako je njihovu stvarnu biološku i dijagnostičku vrijednost potrebno potvrditi na neovisnim i klinički dobro opisanim uzorcima, identificirani transkriptomski profil predstavlja vrijedno polazište za buduća translacijska istraživanja te potencijalni temelj za razvoj pouzdanijih molekularnih biomarkera u dijagnostici papilarnog karcinoma štitnjače.

6. ZAKLJUČAK

Integrativnom analizom javno dostupnih RNA-seq podataka istražene su transkriptomске značajke papilarnog karcinoma štitnjače te utjecaj odabira kontrolnog tkiva na rezultate diferencijalne ekspresije i strojnog učenja. Zasebnom usporedbom tumorskog tkiva sa zdravim tkivom i tumoru susjednim normalnim tkivom (NAT) utvrđene su jasne razlike u genskoj ekspresiji, ali i značajan utjecaj vrste kontrole na broj i sastav izdvojenih gena. NAT se pritom nije pokazao transkriptomski jednakim neovisnom zdravom tkivu, zbog čega se te dvije vrste kontrolnih uzoraka ne mogu smatrati međusobno zamjenjivima.

Meta-analizom neovisnih skupova podataka izdvojeni su geni s ponovljivim promjenama ekspresije, čime je utvrđen stabilan transkriptomski signal papilarnog karcinoma štitnjače. Funkcionalna analiza povezala je te promjene s organizacijom i razgradnjom izvanstaničnoga matriksa, staničnom adhezijom, kolagenom, epitelnom diferencijacijom i keratinizacijom. Usporedba sa zdravim tkivom obuhvatila je širi raspon bioloških promjena, dok je usporedba s NAT-om snažnije istaknula procese lokalnog remodeliranja tumorskog tkiva.

Ranom integracijom podataka razvijeni su modeli strojnog učenja koji su uspješno razlikovali tumorske od kontrolnih uzoraka. Podudarnost dijela gena izdvojenih meta-analizom i metodom nasumičnih šuma omogućila je prepoznavanje kandidata koji istodobno pokazuju konzistentnu diferencijalnu ekspresiju i važnost za klasifikaciju uzoraka. Dobiveni rezultati pokazuju da povezivanje meta-analize, rane integracije i strojnog učenja omogućuje pouzdanije izdvajanje potencijalnih biomarkera, ali je njihovu dijagnostičku vrijednost potrebno potvrditi u neovisnim skupinama bolesnika i eksperimentalnim istraživanjima.

7. LITERATURA

- [1] M. A. Shahid, M. A. Ashraf, i S. Sharma, „Physiology, Thyroid Hormone“, *StatPearls*, lip. 2023, Pristupljeno: 23. kolovoz 2026. [Na internetu]. Dostupno na: <https://www.ncbi.nlm.nih.gov/books/NBK500006/>
- [2] P. Vignali, E. Macerola, A. M. Poma, R. Sparavelli, i F. Basolo, „Indeterminate Thyroid Nodules: From Cytology to Molecular Testing“, *Diagnostics (Basel)*, sv. 13, izd. 18, ruj. 2023, doi: 10.3390/DIAGNOSTICS13183008.
- [3] E. D. Yulian, Sugiarto, R. Z. Martiana, Y. Wahyono, M. F. Ham, i F. M. R. Aqil, „Anaplastic transformation of metastatic papillary thyroid carcinoma lesion histologically mimicking a squamous cell carcinoma: A case report“, *Int. J. Surg. Case Rep.*, sv. 124, str. 110396, stu. 2024, doi: 10.1016/J.IJSCR.2024.110396.
- [4] S. Dizbay Sak, „Variants of papillary thyroid carcinoma: Multiple faces of a familiar tumor“, *Turk Patoloji Derg.*, sv. 31, str. 34–47, kol. 2015, doi: 10.5146/TJPATH.2015.01313.
- [5] R. V. Lloyd, D. Buehler, i E. Khanafshar, „Papillary Thyroid Carcinoma Variants“, *Head and Neck Pathology 2010 5:1*, sv. 5, izd. 1, str. 51–56, sij. 2011, doi: 10.1007/S12105-010-0236-9.
- [6] F. Limaïem, A. Rehman, i T. Mazzoni, „Papillary Thyroid Carcinoma“, *StatPearls*, ožu. 2024, Pristupljeno: 14. kolovoz 2026. [Na internetu]. Dostupno na: <https://www.ncbi.nlm.nih.gov/books/NBK536943/>
- [7] M. I. Abdullah, S. M. Junit, K. L. Ng, J. J. Jayapalan, B. Karikalan, i O. H. Hashim, „Papillary Thyroid Cancer: Genetic Alterations and Molecular Biomarker Investigations“, *Int. J. Med. Sci.*, sv. 16, izd. 3, str. 450, 2019, doi: 10.7150/IJMS.29935.
- [8] N. Agrawal i ostali, „Integrated Genomic Characterization of Papillary Thyroid Carcinoma“, *Cell*, sv. 159, izd. 3, str. 676–690, lis. 2014, doi: 10.1016/j.cell.2014.09.050.
- [9] „Transcriptome Fact Sheet“. Pristupljeno: 17. kolovoz 2026. [Na internetu]. Dostupno na: <https://www.genome.gov/about-genomics/fact-sheets/Transcriptome-Fact-Sheet>
- [10] Z. Wang, M. Gerstein, i M. Snyder, „RNA-Seq: a revolutionary tool for transcriptomics“, *Nature Reviews Genetics 2008 10:1*, sv. 10, izd. 1, str. 57–63, sij. 2009, doi: 10.1038/nrg2484.

- [11] R. Stark, M. Grzelak, i J. Hadfield, „RNA sequencing: the teenage years“, *Nature Reviews Genetics* 2019 20:11, sv. 20, izd. 11, str. 631–656, srp. 2019, doi: 10.1038/s41576-019-0150-2.
- [12] K. R. Kukurba i S. B. Montgomery, „RNA Sequencing and Analysis“, *Cold Spring Harb. Protoc.*, sv. 2015, izd. 11, str. pdb.top084970, stu. 2015, doi: 10.1101/PDB.TOP084970.
- [13] M. Cieřlik i A. M. Chinnaiyan, „Cancer transcriptome profiling at the juncture of clinical translation“, *Nature Reviews Genetics* 2017 19:2, sv. 19, izd. 2, str. 93–109, pros. 2017, doi: 10.1038/nrg.2017.96.
- [14] C. Calabrese i ostali, „Genomic basis for RNA alterations in cancer“, *Nature* 2020 578:7793, sv. 578, izd. 7793, str. 129–136, velj. 2020, doi: 10.1038/s41586-020-1970-0.
- [15] D. Hanahan, „Hallmarks of Cancer: New Dimensions“, *Cancer Discov.*, sv. 12, izd. 1, str. 31–46, sij. 2022, doi: 10.1158/2159-8290.CD-21-1059.
- [16] L. Zhao, V. H. F. Lee, M. K. Ng, H. Yan, i M. F. Bijlsma, „Molecular subtyping of cancer: current status and moving toward clinical applications“, *Brief. Bioinform.*, sv. 20, izd. 2, str. 572–584, ožu. 2019, doi: 10.1093/BIB/BBY026.
- [17] V. Jobanputra i ostali, „Clinical interpretation of whole-genome and whole-transcriptome sequencing for precision oncology“, *Semin. Cancer Biol.*, sv. 84, str. 23–31, ruj. 2022, doi: 10.1016/J.SEMCANCER.2021.07.003.
- [18] C. Ziegenhain i ostali, „Comparative Analysis of Single-Cell RNA Sequencing Methods“, *Mol. Cell*, sv. 65, izd. 4, str. 631-643.e4, velj. 2017, doi: 10.1016/j.molcel.2017.01.023.
- [19] A. Conesa i ostali, „A survey of best practices for RNA-seq data analysis“, *Genome Biology* 2016 17:1, sv. 17, izd. 1, str. 13-, sij. 2016, doi: 10.1186/S13059-016-0881-8.
- [20] D. Aran, M. Sirota, i A. J. Butte, „Systematic pan-cancer analysis of tumour purity“, *Nature Communications* 2015 6:1, sv. 6, izd. 1, str. 8971-, pros. 2015, doi: 10.1038/ncomms9971.
- [21] M. Bortolomeazzi, M. R. Keddar, F. D. Ciccarelli, i L. Benedetti, „Identification of non-cancer cells from cancer transcriptomic data“, *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, sv. 1863, izd. 6, str. 194445, lip. 2020, doi: 10.1016/J.BBAGRM.2019.194445.

- [22] B. Hwang, J. H. Lee, i D. Bang, „Single-cell RNA sequencing technologies and bioinformatics pipelines“, *Experimental & Molecular Medicine* 2018 50:8, sv. 50, izd. 8, str. 1–14, kol. 2018, doi: 10.1038/s12276-018-0071-8.
- [23] M. Crow i J. Gillis, „Single cell RNA-sequencing: replicability of cell types“, *Curr. Opin. Neurobiol.*, sv. 56, str. 69–77, lip. 2019, doi: 10.1016/J.CONB.2018.12.002.
- [24] M. D. Luecken i F. J. Theis, „Current best practices in single-cell RNA-seq analysis: a tutorial“, *Molecular Systems Biology* 2019 15:6, sv. 15, izd. 6, str. MSB188746-, lip. 2019, doi: 10.15252/MSB.20188746.
- [25] M. Crow, A. Paul, S. Ballouz, Z. J. Huang, i J. Gillis, „Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor“, *Nature Communications* 2018 9:1, sv. 9, izd. 1, str. 884-, velj. 2018, doi: 10.1038/s41467-018-03282-0.
- [26] D. Aran i ostali, „Comprehensive analysis of normal adjacent to tumor transcriptomes“, *Nature Communications* 2017 8:1, sv. 8, izd. 1, str. 1077-, lis. 2017, doi: 10.1038/s41467-017-01027-z.
- [27] H. He i ostali, „Transcriptome analysis discloses dysregulated genes in normal appearing tumor-adjacent thyroid tissues from patients with papillary thyroid carcinoma“, *Scientific Reports* 2021 11:1, sv. 11, izd. 1, str. 14126-, srp. 2021, doi: 10.1038/s41598-021-93526-9.
- [28] A. Conesa i ostali, „A survey of best practices for RNA-seq data analysis“, *Genome Biology* 2016 17:1, sv. 17, izd. 1, str. 13-, sij. 2016, doi: 10.1186/S13059-016-0881-8.
- [29] M. I. Love, W. Huber, i S. Anders, „Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2“, *Genome Biology* 2014 15:12, sv. 15, izd. 12, str. 550-, pros. 2014, doi: 10.1186/S13059-014-0550-8.
- [30] J. Costa-Silva, D. Domingues, i F. M. Lopes, „RNA-Seq differential expression analysis: An extended review and a software tool“, *PLoS One*, sv. 12, izd. 12, str. e0190152, pros. 2017, doi: 10.1371/JOURNAL.PONE.0190152.
- [31] S. Anders i W. Huber, „Differential expression analysis for sequence count data“, *Genome Biology* 2010 11:10, sv. 11, izd. 10, str. R106-, lis. 2010, doi: 10.1186/GB-2010-11-10-R106.
- [32] A. Dobin i ostali, „STAR: ultrafast universal RNA-seq aligner“, *Bioinformatics*, sv. 29, izd. 1, str. 15–21, sij. 2013, doi: 10.1093/BIOINFORMATICS/BTS635.

- [33] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, i C. Kingsford, „Salmon provides fast and bias-aware quantification of transcript expression“, *Nature Methods* 2017 14:4, sv. 14, izd. 4, str. 417–419, ožu. 2017, doi: 10.1038/nmeth.4197.
- [34] Y. Liao, G. K. Smyth, i W. Shi, „featureCounts: an efficient general purpose program for assigning sequence reads to genomic features“, *Bioinformatics*, sv. 30, izd. 7, str. 923–930, tra. 2014, doi: 10.1093/BIOINFORMATICS/BTT656.
- [35] P. Ewels, M. Magnusson, S. Lundin, i M. Käller, „MultiQC: summarize analysis results for multiple tools and samples in a single report“, *Bioinformatics*, sv. 32, izd. 19, str. 3047–3048, lis. 2016, doi: 10.1093/BIOINFORMATICS/BTW354.
- [36] R. Bourgon, R. Gentleman, i W. Huber, „Independent filtering increases detection power for high-throughput experiments“, *Proc. Natl. Acad. Sci. U. S. A.*, sv. 107, izd. 21, str. 9546–9551, svi. 2010, doi: 10.1073/PNAS.0914005107;WEBSITE:WEBSITE:PNAS-SITE;WGROUPE:STRING:PUBLICATION.
- [37] M. D. Robinson i A. Oshlack, „A scaling normalization method for differential expression analysis of RNA-seq data“, *Genome Biology* 2010 11:3, sv. 11, izd. 3, str. R25-, ožu. 2010, doi: 10.1186/GB-2010-11-3-R25.
- [38] A. Zhu, J. G. Ibrahim, i M. I. Love, „Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences“, *Bioinformatics*, sv. 35, izd. 12, str. 2084–2092, lip. 2019, doi: 10.1093/BIOINFORMATICS/BTY895.
- [39] A. McDermaid, B. Monier, J. Zhao, B. Liu, i Q. Ma, „Interpretation of differential gene expression results of RNA-seq data: review and integration“, *Brief. Bioinform.*, sv. 20, izd. 6, str. 2044, stu. 2018, doi: 10.1093/BIB/BBY067.
- [40] M. D. Wilkinson i ostali, „The FAIR Guiding Principles for scientific data management and stewardship“, *Scientific Data* 2016 3:1, sv. 3, izd. 1, str. 160018-, ožu. 2016, doi: 10.1038/sdata.2016.18.
- [41] E. Clough i ostali, „NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update“, *Nucleic Acids Res.*, sv. 52, izd. D1, str. D138–D144, sij. 2024, doi: 10.1093/NAR/GKAD965.
- [42] T. Barrett i ostali, „NCBI GEO: archive for functional genomics data sets—update“, *Nucleic Acids Res.*, sv. 41, izd. D1, str. D991–D995, sij. 2013, doi: 10.1093/NAR/GKS1193.
- [43] J. T. Leek i ostali, „Tackling the widespread and critical impact of batch effects in high-throughput data“, *Nature Reviews Genetics* 2010 11:10, sv. 11, izd. 10, str. 733–739, ruj. 2010, doi: 10.1038/nrg2825.

- [44] C. Hutter i J. C. Zenklusen, „The Cancer Genome Atlas: Creating Lasting Value beyond Its Data“, *Cell*, sv. 173, izd. 2, str. 283–285, tra. 2018, doi: 10.1016/j.cell.2018.03.042.
- [45] „TCGA Resources | NCI Genomic Data Commons“. Pristupljeno: 18. kolovoz 2026. [Na internetu]. Dostupno na: <https://gdc.cancer.gov/about-data/gdc-data-processing/resources-tcga-users>
- [46] A. P. Heath i ostali, „The NCI Genomic Data Commons“, *Nature Genetics* 2021 53:3, sv. 53, izd. 3, str. 257–262, velj. 2021, doi: 10.1038/s41588-021-00791-5.
- [47] F. Aguet i ostali, „The GTEx Consortium atlas of genetic regulatory effects across human tissues“, *Science (1979)*., sv. 369, izd. 6509, str. 1318–1330, ruj. 2020, doi: 10.1126/SCIENCE.AAZ1776.
- [48] „GTEx Portal“. Pristupljeno: 18. kolovoz 2026. [Na internetu]. Dostupno na: <https://gtexportal.org/home/downloads/adult-gtex/overview>
- [49] J. Vivian i ostali, „Toil enables reproducible, open source, big biomedical data analyses“, *Nature Biotechnology* 2017 35:4, sv. 35, izd. 4, str. 314–316, tra. 2017, doi: 10.1038/nbt.3772.
- [50] X. Wang i ostali, „Cell-state transitions and microenvironmental remodeling in thyroid cancer progression revealed by single-cell and spatial transcriptomics“, *Front. Immunol.*, sv. 17, str. 1904196, srp. 2026, doi: 10.3389/FIMMU.2026.1904196/FULL.
- [51] X. Zhu, X. Wang, Y. Gong, i J. Deng, „E-cadherin on epithelial–mesenchymal transition in thyroid cancer“, *Cancer Cell International* 2021 21:1, sv. 21, izd. 1, str. 695–, pros. 2021, doi: 10.1186/S12935-021-02344-6.
- [52] V. Karantza, „Keratins in health and cancer: more than mere epithelial cell markers“, *Oncogene*, sv. 30, izd. 2, str. 127–138, sij. 2011, doi: 10.1038/ONC.2010.456.
- [53] A. Rau, G. Marot, i F. Jaffrézic, „Differential meta-analysis of RNA-seq data from multiple studies“, *BMC Bioinformatics*, sv. 15, izd. 1, ožu. 2014, doi: 10.1186/1471-2105-15-91.
- [54] M. B. Kurska i W. R. Rudnicki, „Feature Selection with the Boruta Package“, *J. Stat. Softw.*, sv. 36, izd. 11, str. 1–13, ruj. 2010, doi: 10.18637/JSS.V036.I11.
- [55] E. A. Musgrove, C. E. Caldon, J. Barraclough, A. Stone, i R. L. Sutherland, „Cyclin D as a therapeutic target in cancer“, *Nature Reviews Cancer* 2011 11:8, sv. 11, izd. 8, str. 558–572, srp. 2011, doi: 10.1038/nrc3090.
- [56] Y. Jin i ostali, „GABRB2 plays an important role in the lymph node metastasis of papillary thyroid cancer“, *Biochem. Biophys. Res. Commun.*, sv. 492, izd. 3, str. 323–330, lis. 2017, doi: 10.1016/J.BBRC.2017.08.114.

- [57] I. Petrini, R. L. Cecchini, M. Mascaró, I. Ponzoni, i J. A. Carballido, „Papillary Thyroid Carcinoma: A thorough Bioinformatic Analysis of Gene Expression and Clinical Data“, *Genes (Basel)*., sv. 14, izd. 6, str. 1250, lip. 2023, doi: 10.3390/GENES14061250.
- [58] C. Van Branteghem, N. Henry, L. Craciun, i C. Maenhaut, „HMGA2 Overexpression in Papillary Thyroid Cancer Promotes Thyroid Cell Dedifferentiation and Invasion, and These Effects Are Counteracted by Suramin“, *Int. J. Mol. Sci.*, sv. 26, izd. 4, velj. 2025, doi: 10.3390/IJMS26041643.
- [59] J. Jonklaas i ostali, „Novel biomarker SYT12 may contribute to predicting papillary thyroid cancer outcomes“, *Future Sci. OA*, sv. 4, izd. 1, str. FSO249, sij. 2017, doi: 10.4155/FSOA-2017-0087.
- [60] P. Liu i X. Yu, „Lymphatic metastasis of papillary thyroid carcinoma: mechanism and clinicopathological physiology“, *Front. Endocrinol. (Lausanne)*., sv. 16, str. 1725077, 2026, doi: 10.3389/FENDO.2025.1725077.