SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO - MATEMATIČKI FAKULTET

BIOLOŠKI ODSJEK

Kristian Bodulić i Paula Štancl

RAČUNALNA ANALIZA SLJEDOVA OGULINSKE ŠPILJSKE SPUŽVICE (*EUNAPIUS SUBTERRANEUS* SKET & VELIKONJA, 1984) PRIKUPLJENIH TEHNOLOGIJOM SEKVENCIRANJA NANOPORAMA

Zagreb, 2019.

Ovaj rad izrađen je u Grupi za bioinformatiku na Zavodu za molekularnu biologiju Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu pod vodstvom prof. dr. sc. Kristiana Vlahovičeka. Rad je predan na natječaj za dodjelu Rektorove nagrade u akademskoj godini 2018./2019.

Popis kratica

IUCN International Union for Conservation of Nature

COI Cytochrome c oxidase subunit 1

ITS1 Innternal transcribed spacer 1

DNA Deoksiribonukleinska kiselina

pb Par baza

kb Kilobaza

PCR Polymerase chain reaction

SMRT Single molecule real time sequencing

ONT Oxford Nanopore Technologies

BWA Burrows-Wheeler Aligner

SAM Sequence Alignment Map

BAM Binary Alignment Map

VCF Variant Call Format

BCF Binary Variant Call Format

PCA Principal component analysis

DIAMOND Double index alignment of next-generation sequencing data

MEGAN Metagenome Analyzer

LCA Lowest common ancestor

SADRŽAJ

1. UVOD 1
1.1. Koljeno Porifera (spužve) 1
1.1.1. Građa tijela i osnovne funkcije.....1
1.1.2. Taksonomija i filogenetski odnosi.....3
1.1.3. Ogulinska špiljska spužvica (Eunapius subterraneus Sket & Velikona, 1984) 4
1.1.4. Genomi spužvi: sekvenciranje, prepreke i daljnja istraživanja 5
1.2. Sekvenciranje DNA 6
1.2.1. Metoda reverzibilnog zaustavljanja sinteze DNA (Illumina) 7
1.2.2. Sekvenciranje tehnologijom nanoporama.....8
2. OPĆI I SPECIFIČNI CILJEVI RADA.....12
3. MATERIJALI I METODE 13
3.1. Korištene knjižnice.....13
3.2. Statistička obrada podataka programskim jezikom R.....14
3.3. Sravnjenje sljedova iz knjižnice Ilumina na sljedove iz knjižnice ONT500 15
3.3.1. Program minimap2.....15
3.3.2. Program BWA.....16
3.3.3. Format SAM i program SAMBAMBA.....16
3.4. Ispravljanje sljedova upotrebom programa Pilon.....16
3.5. Profil pogrešaka kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT500 17
3.6. Analiza ispravljenosti sljedova upotrebom programa DIAMOND.....19
3.7. Taksonomska i funkcijska klasifikacija sljedova upotrebom programa MEGAN 20
4. REZULTATI 21
4.1. Statistička obrada kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT i ONT500.....21
4.2. Rezultati sravnjenje sljedova iz knjižnice Ilumina na sljedove iz knjižnice ONT500....23
4.3. Profil pogrešaka kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT500 24
4.3.1. Pogrešno očitane baze u kvalitetnim i nekvalitetnim sljedovima iz knjižnice ONT500.....26
4.3.2. Indeli u kvalitetnim i nekvalitetnim sljedovima iz knjižnice ONT500 27
4.4. Ispravljanje kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT500 28

4.4.1. Taksonomska klasifikacija sljedova iz knjižnice ONT500 31
4.4.2. Funkcijska klasifikacija proteina pronađenih u svim kvalitetnim i nekvalitetnim sljedovima knjižnice ONT500 34
4.4.3. Funkcijska klasifikacija proteina kvalitetnih i nekvalitetnih sljedova knjižnice ONT500 kojima je dodijeljena taksonomska kategorija Metazoa ili Porifera 39
5. RASPRAVA 44
6. ZAKLJUČCI 48
7. ZAHVALE 49
8. POPIS LITERATURE 50
9. SAŽETAK 54
10. SUMMARY 55
11. ŽIVOTOPISI 56
PRILOZI 57

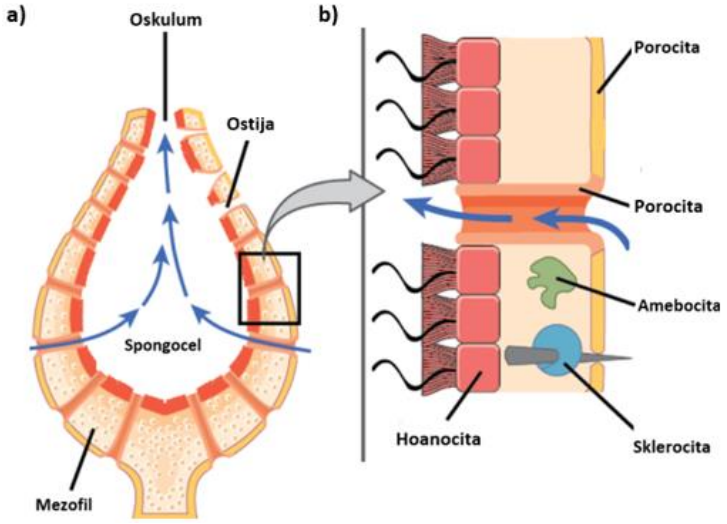
1. UVOD

1.1. Koljeno Porifera (spužve)

Koljeno spužvi predstavlja raznoliku skupinu višestaničnih organizama koja se među prvima odvojila od zajedničkog pretka životinja. To su široko rasprostranjeni organizmi koji žive u vodenom okolišu. Njihovo tijelo čini sustav pora koji omogućuje protok vode i izmjenu tvari s okolišem, što je u skladu sa sjedilačkim načinom života odraslih jedinki. Spužvama nedostaju pravi organski sustavi, njihovo tijelo građeno je od specijaliziranih stanica sa znatnim stupnjem nezavisnosti, odnosno relativno malim stupnjem koordinacije. Ova skupina organizama je relativno neistražena, brojna pitanja još uvijek su otvorena, a odgovor na njih mogao bi dati bolji uvid u biologiju spužvi, ali i u općenitiju problematiku poput filogenije životinja (Dunn i sur. 2015).

1.1.1. Građa tijela i osnovne funkcije

Oblik i veličina spužvi su različiti u pojedinih vrsta, jednostavnije vrste su u načelu radijalno simetrične, dok je veći broj vrsta asimetričan. Veličina tijela opisanih vrsta kreće se u rasponu od nekoliko milimetara do nekoliko metara u promjeru. Tijelo većine spužvi građeno je od mnoštva manjih pora (ostija) kroz koje ulazi voda, manjeg broja većih pora (oskuluma) kroz koje izlazi voda te središnje šupljine (spongocela) (Slika 1). Ovisno o složenosti spužve spongocel može biti jednostavna šupljina ili kompleksna mreža kanalića. Na histološkoj razini tijelo spužve sastoji se od tri sloja stanica što uključuje unutrašnji sloj, središnji sloj (mezohil) i površinski sloj (pinakoderma). Pinakodermu grade pinakocite, stanice sa zaštitnom ulogom među kojima su raspoređene ostije. Ostije su u jednostavnijih spužvi građene od porocita, cjevastih stanica koje kontroliraju ulazak vode u poru. Mezohil sadrži vlakna kolagena, različite vrste stanica i skelet. Jedna od osnovnih vrsta stanica mezohila su amebocite, pokretne stanice s brojnim ulogama poput stvaranja oocita ili diferencijacije u velik broj stanica što ih čini važnim za sposobnost regeneracije spužve. Jedna od vrsta stanica koje nastaju iz amebocita su sklerociti. Ove stanice proizvode spikule, sastavnice skeleta koje mogu biti građene od kalcijeva karbonata, silicijeva dioksida ili proteina spongina. Unutrašnji sloj spužve također grade pinakocite, ali i bičaste hoanocite. Uloge hoanocita su brojne pri čemu je važno izdvojiti pomaganje u strujanju vode kroz tijela spužve što omogućuje izmjenu tvari s okolinom. Također, ove stanice imaju sposobnost fagocitoze što omogućava probavu jednostaničnih organizama i manjih čestica. U probavi isto tako mogu sudjelovati i druge stanice spužve pri čemu pinakociti obično fagocitiraju čestice koje su prevelike da bi prošle kroz ostije. Strujanje vode važno je i za disanje te izlučivanje otpadnih tvari, pri čemu kisik difuzijom ulazi u stanice dok otpadne tvari, poput ugljikova dioksida i amonijaka, izlaze iz stanice u strujajuću vodu (Matoničkin 1990).



Slika 1. a) Plan građe tijela jednostavne spužve. b) Prikaz osnovnih stanica spužve. Smjer kretanja vode na oba prikaza označen je strelicama plave boje. Preuzeto i prilagođeno s https://courses.lumenlearning.com/wm-biology2/chapter/morphology-of-sponges/

Spužve se mogu razmnožavati nespolno i spolno. Najčešći oblici nespolnog razmnožavanja su pupanje i fragmentacija pri čemu se dio tijela spužve odvaja te nastavlja život kao kolonija ili kao nova jedinka. Rjeđi oblik nespolnog razmnožavanja uključuje stvaranje gemula, otpornih tjelešaca karakterističnih za slatkovodne spužve. Za gemule je karakterističan srednji sloj amebocita koji je okružen vanjskim slojem kolagena ili spongina u koji su uklopljene spikule. Ovakav raspored stanica omogućava preživljavanje gemule u nepovoljnim okolišnim uvjetima u stadiju dormancije pri čemu se pri ponovnom nastupanju povoljnih uvjeta iz gemule razvija nova spužva. S druge strane, spolno razmnožavanje spužve karakterizira hermafroditizam pri čemu se muške (spermiji) i ženske (oocite) gamete mogu otpuštati istovremeno ili u različitim dijelovima godine. Spermiji nastaju sazrijevanjem hoanocita pri čemu se otpuštaju kroz oskulum u okolnu vodu. Nasuprot tome, oocite nastaju diferencijacijom amebocita te se zadržavaju u spongocelu. Spermiji nošeni strujom vode mogu doprijeti u spongocel druge spužve pri čemu stapanjem s oocitom dolazi do oplodnje. Početni stadiji ličinke razvijaju se unutar spužve nakon čega se njezin pokretni stadij otpušta kroz oskulum u vodu. Nakon određenog vremena ličinka se smješta na podlogu gdje se razvija u odrasli oblik spužve (Matoničkin 1990).

1.1.2. Taksonomija i filogenetski odnosi

Koljeno Porifera podijeljeno je u četiri razreda u koje se svrstava između 8500 i 11000 opisanih vrsta od kojih je velika većina morska. Spužve razreda Hexactinellida (staklače) imaju spikule građene od silicijeva dioksida, a obično žive na morskom dnu. Za njih je karakteristična sincicijalna građa tkiva, odnosno prisustvo stanica s mnogojezgrenim citoplazmama. Nedostaju im pinakocite, a vanjski sloj čine amebocite kroz koje prolaze spikule. S druge strane, razred Calcarea (vapnenjače) karakteriziraju spikule građene od kalcijeva karbonata. To su isključivo morske spužve, obično male veličine sa staništem u plitkim vodama. U razred Demospongiae (kremenorožnjače) spada najveći broj dosad opisanih vrsta spužvi, uključujući sve dosad opisane slatkovodne spužve. Njihove spikule mogu biti građene od silicijeva dioksida, spongina ili kombinacije tih dvaju tvari. Raznolika su oblika i boje, a može ih se naći u svim dubinama. Skupina Homoscleromorpha donedavno se zbog morfoloških sličnosti smatrala podrazredom skupine Demospongiae, a novija filogenetska istraživanja svrstavaju ovu skupinu u zaseban razred. Spužve ovog razreda žive u moru, obično na tamnijim mjestima. Karakterizira ih prisutnost bazalne membrane, odnosno epitela sličnog ostalim životinjskim koljenima. Ako je prisutan, skelet grade male spikule od /silicijeva dioksida. Sistematika na nižim taksonomskim kategorijama otežana je morfologijom spužvi zbog čega su njezine promjene vrlo česte (van Soest i sur. 2012).

Filogenetski odnosi unutar skupine spužvi još uvijek nisu razriješeni. Smatra se da su se spužve od zajedničkog pretka svih životinja odvojile prije svih ostalih životinjskih koljena (Simion i sur. 2017). Zbog toga se životinje (Metazoa) dijele na spužve i skupinu Eumetazoa u koju se svrstavaju sve ostale životinje. Brojni autori smatraju da su spužve parafiletična skupina što temelje na povećanoj sličnosti razreda Homoscleromorpha sa skupinom Eumetazoa (Sperling i sur. 2007). Ipak, mnogi radovi objavljeni u zadnjem desetljeću dokazima molekularne filogenije pobijaju ovu teoriju i govore o spužvama kao o monofiletskoj skupini (primjerice Philippe i sur. 2009). Također, postoji neslaganje o položaju spužvi unutar filogenije bazalnih životinja. Iako brojni autori smatraju da su se spužve prve odvojile od zajedničkog pretka svih životinja, novija istraživanja na to mjesto stavljaju rebraše (Ctenophora) (Dunn i sur. 2008).

1.1.3. Ogulinska špiljska spužvica (*Eunapius subterraneus* Sket & Velikona, 1984*)*

Ogulinska špiljska spužvica (*Eunapius subterraneus* Sket & Velikona, 1984) (Slika 2) je endemska vrsta pronađena na području Velike Kapele i Ogulina. Prvi su je opisali Sket i Velikonja 1984. godine a 2004. godine uvrštena je u kategoriju ugroženih vrsta prema kriterijima organizacije IUCN (engl*. International Union for Conservation of Nature*). To je jedina poznata slatkovodna stigobiontska spužva na svijetu što je čini izuzetno važnom sastavnicom hrvatske faune. Karakterizira je rahlo osjetljivo tijelo bijele boje sa skeletom od vrlo velikih i lagano zakrivljenih spikula građenim od spongina. Ističe se vrlo veliki oskulum, a oblik tijela može biti jajolik ili tanjurast. Obično raste na stijenama birajući zaklonjena mjesta od jakog strujanja vode. Potpuno su depigmentirane te preferiraju rast u potpunoj tami, ali toleriraju i malu količinu svjetlosti (Bedek i sur. 2008).

Ogulinska špiljska spužvica svrstana je u razred Demospongiae, red Spongillida i porodicu Spongillidae *(World Porifera Database* 2018). U prilog svrstavanja ove spužve u rod *Eunapius* govore mnoga njezina morfološka obilježja. Ipak, u ove spužve postoje i obilježja koja je razlikuju od drugih pripadnika roda *Eunapius*, poput građe gemula ili pora. Ove razlike mogu biti objašnjene adaptacijom vrste na okoliš ili pogrešnom klasifikacijom ove vrste u rod *Eunapius*. S mogućnošću pogrešne klasifikacije slaže se i istraživanje provedeno na tri molekularna biljega (18S rDNA, COI (engl. *cytochrome c oxidase subunit* 1) i ITS2 (engl. *internal transcribed spacer*)), koje ovu vrstu smješta bliže vrstama slatkovodnih spužvi rodova *Ephydatia* i *Lubomirskia* nego drugim vrstama unutar roda *Eunapius* (Harcet i sur. 2010).

Slika 2. Ogulinska špiljska spužvica (*Eunapius subterraneus*) jajolika oblika. Preuzeto iz Bedek i sur. 2008)

1.1.4. Genomi spužvi: sekvenciranje, prepreke i daljnja istraživanja

Sekvenciranje genoma spužvi jedan je od ključnih koraka u opisivanju različitih aspekata njihove biologije, ali i općenitijih problema poput filogenetskih odnosa unutar životinja i, posebno, unutar koljena Porifera. Jedini sekvenciran spužvin genom do danas je genom vrste *Amphimedon queenslandica* (Srivastava i sur. 2010), kremenorožnjače pronađene na području Velikog koraljnog grebena. Komparativna analiza genoma spomenute vrste dala je uvid u porijeklo i evoluciju brojnih gena važnih za višestaničnost životinja. Smatra se da se broj takvih gena izrazito povećao prije odvajanja spužvi od ostalih životinja. Sukladno tome, u genomu spomenute vrste pronađene su genske obitelji koje imaju glavnu ulogu u komunikaciji stanica s okolinom. Te obitelji odgovorne su za svojstva poput kontrole staničnog ciklusa i regulacije rasta, programirane stanične smrti, međustanične adhezije, regulacije razvoja organizma, imunosnog i senzornog sustava te specijalizacije. Iako prisutnost članova navedenih genskih obitelji u genomu vrste *Amphimedon queenslandica* dokazuje pojavu tih obitelji prije odvajanja spužvi od zajedničkog pretka životinja, njihov broj i kompleksnost izuzetno su porasli tek tijekom razvoja skupine Eumetazoa. Također, u genomu ove vrste pronađeni su brojni geni odgovorni za biološke procese za koje se do tada pretpostavljalo da u spužvama nedostaju. Sve ovo upućuje na neočekivano veliku kompleksnost genoma zajedničkog pretka svih životinja, unatoč gubitku mnogih fizioloških i morfoloških karakteristika današnjih spužvi tijekom evolucije. Moguće je da su tak karakteristike u spužava prisutne u obliku koji još nije poznat pri čemu bi sekvenciranje genoma drugih vrsta spužvi moglo potvrditi ili odbaciti ovakvu pretpostavku (Srivastava i sur. 2010).

Sekvenciranje genoma drugih vrsta spužvi je aktualna tema današnje genomike, a sam proces otežan je brojnim okolišnim čimbenicima. Primjerice, utvrđena je prisutnost velikog broja organizama koji žive u simbiotskom odnosu sa spužvama (najčešće mutualizam i parazitizam). Spomenuti organizmi izuzetno su metabolički, morfološki i filogenetski raznoliki, s predstavnicima u sve tri domene života. Njihov mutualizam sa spužvama obično se temelji na lučenju brojnih sekundarnih metabolita čije su najčešće uloge obrana spužve od različitih predatora ili bakterija, a njihova primjena mogla bi imati značaj u medicini ili industriji. Isto tako, pokazana je i prisutnost fotosintetskih organizama, obično kod spužvi koje žive u vodi s malo otopljenih hranjivih tvari. Uzimajući u obzir činjenicu da mikrobne zajednice mogu sastavljati do 40% mase spužve, one su vrlo česti kontaminanti u eksperimentima sekvenciranja genoma spužvi (Taylor i sur. 2007). Ovakva onečišćenja teško je ukloniti eksperimentalnim ili *in silico* (računalnim) metodama, a jedna od često korištenih eksperimentalnih metoda je izolacija DNA iz primarne stanične kulture (primorfa) dobivene iz pročišćenog uzorka.

U zadnjih nekoliko godina došlo je do ubrzanog razvoja tehnologije sekvenciranja i pristupa koji se koriste u sastavljanju genoma. Kao što je već opisano, sekvenciranje i kvalitetnije sastavljanje genoma različitih vrsta spužvi moglo bi dati odgovore na mnoga pitanja

koja ne obuhvaćaju samo biologiju spužvi i njezinih simbionata, već i starija biološka pitanja poput filogenije i razvoja životinja te prijelaza organizama na višestaničnost.

1.2. Sekvenciranje DNA

Sekvenciranje DNA podrazumijeva određivanje redoslijeda njezinih baza što je omogućilo velik napredak u brojnim područjima molekularne biologije. Brz razvoj sekvenciranja DNA započinje sedamdesetih godina 20. stoljeća kada se javljaju metode sekvenciranja prve generacije: Maxam-Gilbertovo i Sangerovo sekvenciranje. Zbog veće jednostavnosti i mogućnosti automatizacije Sangerovo sekvenciranje ubrzo je postala vodeća metoda sekvenciranja, s primjenama u projektima poput sekvenciranja genoma mnogih vrsta, uključujući čovjeka. Tijekom sekvenciranja ljudskog genoma ova metoda je znatno poboljšana što je omogućilo njezinu automatizaciju i sekvenciranje fragmenata DNA do 900 pb (parova baza). Sangerovo sekvenciranje se danas nerijetko naziva „zlatnim standardom" sekvenciranja, posebno za primjene poput sekvenciranja manjih dijelova DNA u svrhu genotipizacije i kliničke prakse. Ipak, zbog male brzine i velike cijene, primjena ove metode nije pogodna za sekvenciranje većih dijelova DNA, poput kromosoma ili cijelog genoma (Stranneheim i sur. 2012).

Početkom 21. stoljeća dolazi do pojave novih metoda sekvenciranja nazvanih metodama druge generacije. Ovakve metode omogućuju paralelno sekvenciranje velikog broja fragmenata molekule DNA čime se brzina procesa višestruko povećava, uz drastično smanjenje cijene postupka. Postoji velik broj različitih metoda sekvenciranja druge generacije, a neke od njih su metoda reverzibilnog zaustavljanja sinteze DNA (Illumina), pirosekvenciranje, sekvenciranje ligacijom i sekvenciranje ionskim poluvodičima. Iako su biokemijski mehanizmi ovih metoda međusobno različiti, one dijele nekoliko sličnosti. Uzorci se prije sekvenciranja obično pocijepaju na manje dijelove. Na ovako nastale fragmente DNA veže se adapter koji u većini slučajeva služi pričvršćivanju uzorka za krutu podlogu (npr. kuglice ili pločice). Pričvršćeni fragmenti DNA umnažaju se tehnikom PCR (lančana reakcija polimerazom, engl. *polymerase chain reaction*) i grupiraju nakon čega dolazi do sekvenciranja. Ono se temelji na automatiziranom čitanju biokemijskih, fluorescencijskih ili električnih signala koji odgovaraju pojedinoj bazi ili bazama. Iako je sam proces čitanja baza relativno spor, pričvršćivanje fragmenata DNA na krutu podlogu omogućava paralelno sekvenciranje velikog broja molekula DNA, odnosno visoku efikasnost ovih metoda. Jedna od mogućnosti koju pružaju metode sekvenciranja druge generacije je sekvenciranje oba kraja fragmenta DNA, odnosno sekvenciranje uparenih krajeva (engl. *pair-end sequencing)* što se postiže korištenjem adaptera za oba kraja fragmenta DNA. Ova tehnika često je važna u smještanju repetitivnih sljedova pri sastavljanju genoma *de novo.* Glavni nedostatak ovih metoda je mogućnost očitanja razmjerno kratkih sljedova (50-700 pb) s većim udjelom pogrešaka u odnosu na metode prve generacije (Stranneheim i sur. 2012).

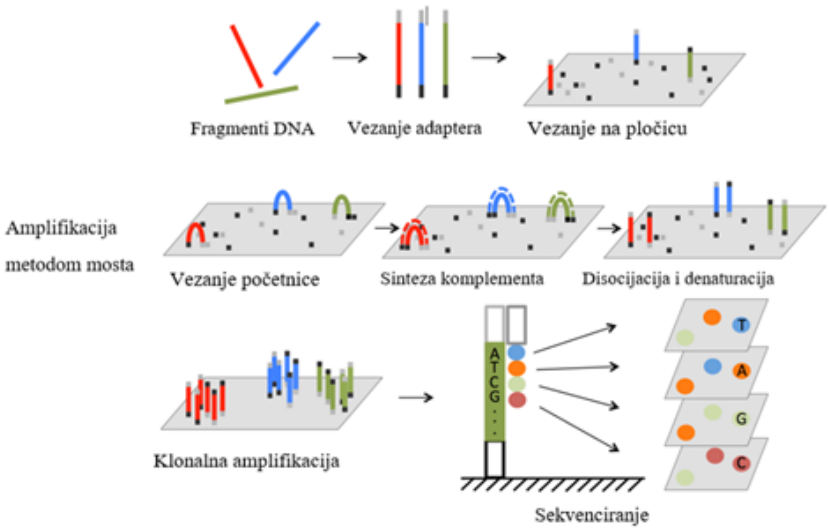
U posljednjih nekoliko godina javljaju se metode sekvenciranja treće generacije. Ove metode omogućavaju sekvenciranje u stvarnom vremenu (engl. *real time sequencing*) pojedinačnih molekula DNA, bez potrebe pripreme, odnosno umnažanja početnog uzorka, što sam proces čini još bržim i jeftinijim. Također, ove prednosti omogućavaju čitanje dugačkih sljedova (do više od 150 000 pb, a u nekim slučajevima i do 1Mb). Međutim, ove tehnologije još su u razvoju, a njihov glavni problem je relativno velik udio pogreške u čitanju baza. Najpoznatije tehnologije sekvenciranja treće generacije su SMRT (engl. s*ingle molecule real time sequencing*) razvijeno od tvrtke *Pacific Biosciences* te sekvenciranje nanoporama razvijeno od tvrtke *Oxford Nanopore Technologies* (ONT) (Stranneheim i sur. 2012).

1.2.1. Metoda reverzibilnog zaustavljanja sinteze DNA (Illumina)

Metoda reverzibilnog zaustavljanja sinteze DNA ubraja se u metode sekvenciranja druge generacije, što znači da je karakterizira paralelno sekvenciranje velikog broja fragmenata DNA i produkcija relativno kratkih sljedova. Kao i u ostalim metodama druge generacije, knjižnica DNA se na početku rascijepa na manje jednolančane dijelove nakon čega se na oba kraja svakog lanca DNA (u slučaju sekvenciranja uparenim krajevima) ligira fragment koji se sastoji od terminalne sekvence, adaptera, indeksa i mjesta vezanja početnice. Indeks je slijed od šest nukleotida koji omogućava razlikovanje uzoraka tijekom obrade podataka, a terminalna sekvenca služi vezanju fragmenta DNA za protočnu ćeliju (engl. *flowcell*). Protočna ćelija je staklena podloga na kojoj se paralelno odvija više reakcija sekvenciranja. Na njoj se nalaze dvije vrste oligonukleotida koje su međusobno komplementarne, a služe za vezanje terminalnih sekvenci adaptera. Nakon vezanja adaptera za jedan oligomer dolazi do amplifikacije metodom mosta (engl. *bridge amplification*) koja započinje vezanjem početnice za mjesto vezanja početnice adaptera. Nakon što DNA polimeraza sintetizira komplementarni lanac, originalni lanac se ispire tako da ostaje komplementarni lanac s adapterom koji se zatim veže za drugu vrstu oligonukleotida na ploči pri čemu nastaje struktura nalik mostu. Nakon vezanja početnice ponovno dolazi do sinteze DNA polimerazom pri čemu nastaje originalni lanac. Ovako nastala dvolančana DNA se denaturira te se svaki lanac ponovno veže za novi oligonukleotid na ploči. Ovaj proces ponavlja se mnogo puta (klonalna amplifikacija) čime nastaje velik broj klastera molekula DNA pričvršćenih za protočnu ćeliju (Kchouk i sur. 2017).

Nakon amplifikacije sljedova DNA slijedi sekvenciranje. Lanci komplementarni originalu nastali amplifikacijom se ispiru te se na slobodne adaptere originalnih lanaca vežu zaštitna skupina i početnice. U reakcijske smjese dodaju se četiri nukleotida obilježena različitim fluorescencijskim bojama te se pri ugradnji nukleotida zabilježava fluorescencijski signal za svaki klaster (lokalna nakupina umnoženih fragmenata na čvrstoj podlozi), pri čemu se odgovarajuće baze u svakom klasteru čitaju istovremeno. Nukleotidi koji se dodaju su modificirani s reverzibilnim terminatorom koji sprječava dodavanje sljedećeg nukleotida u

rastući lanac DNA. Nakon ugradnje nukleotida smjesa se ispire uz uklanjanje terminatora, a postupak se ponavlja do završetka sinteze novih lanaca svakog klastera (Slika 3). Zatim, sintetizirani lanci se ispiru, a uklanjanjem 3' zaštitne skupine dolazi do vezanja drugog adaptera originalnog lanca za drugu vrstu oligonukleotida na pločici pri čemu nastaje već spomenuta struktura mosta. Polimeraza ponovno sintetizira komplementarni lanac te se nastala dvolančana struktura denaturira. Originalni lanac se ispire, a komplementarni lanac se sekvencira na isti način kao i originalni lanac. Dobiveni podaci obično se računalno grupiraju prema indeksima nakon čega se pomoću preklapanja u sljedovima rade duži neprekinuti sljedovi (engl. *contings*) (Kchouk i sur. 2017).



Slika 3. Postupak sekvenciranja reverznim zaustavljanjem sinteze. Preuzeto i prilagođeno s https://www.intechopen.com/books/next-generation-sequencing-advances-applications-and-challenges/next-generation-sequencing-in-aquatic-models

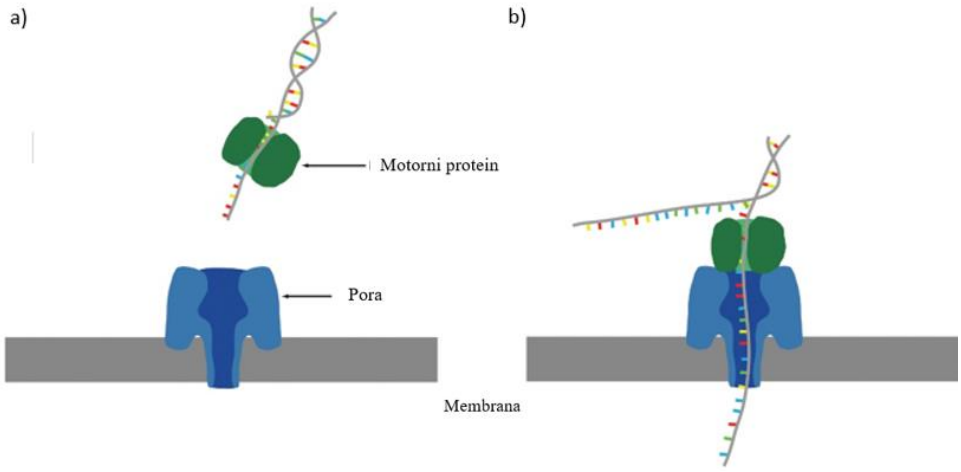
Relativno mala duljina sekvenciranih sljedova, čija se prosječna duljina kreće od 150 do 300 pb, već je istaknuta kao najveći problem ove metode, ali i ostalih metoda sekvenciranja druge generacije. Naime, ugradnja nukleotida nema 100%-tnu učinkovitost što može uzrokovati desinkronizaciju ugrađivanja nukleotida u različitim lancima. Posljedica ove pojave je pozadinski šum koji sprječava očitavanje daljnjih signala. Također, greška ove metode iznosi oko 0.1%, a može biti uzrokovana nepreciznošću polimeraze ili stvaranjem kimernih sekvenci. Ipak, danas je Illumina najčešća korištena metoda sekvenciranja, vjerojatno zbog svoje brzine, rasprostranjenosti i relativno niske cijene (Kchouk i sur. 2017).

1.2.2. Sekvenciranje tehnologijom nanoporama

Tehnologija sekvenciranja nanoporama ubraja se u metode sekvenciranja treće generacije što znači da se ovom tehnologijom sekvenciraju pojedinačne molekule DNA u stvarnom vremenu. Ova metoda zasniva se na prolasku molekule DNA kroz proteinsku poru u membrani na kojoj je prisutan električni potencijal (Slika 4). Prolaskom određenog

redoslijeda nukleotida kroz nanoporu dolazi do promjene električnog potencijala, odnosno do promjenjivog strujnog signala ovisnog o slijedu nukleotida, iz kojeg se može iščitati redoslijed tih nukleotida. Tvrtka ONT 2014. godine predstavlja MinION, iznimno mali uređaj za sekvenciranje koji se temelji na tehnologiji sekvenciranja nanoporama. Njegova protočna ćelija sastoji se od jednoslojne membrane visokog električnog otpora u koju su umetnute biološke pore. Prvobitna verzija ovog uređaja (R7) kao poru je koristila bakterijski porin MspA, dok najnovija verzija (R9) koristi bakterijski protein CsgG (de Lannoy i sur. 2017).

Za razliku od metoda druge generacije, priprema knjižnice pri sekvenciranju nanoporama ne zahtjeva umnažanje DNA te je samim time mnogo jednostavnija i brža. Također, sam protokol pripreme DNA za sekvenciranje ne zahtijeva ni cijepanje DNA na određenu duljinu fragmenta, pa je ovaj korak u pripremi moguće u potpunosti izostaviti ili uključiti, ovisno o krajnjem cilju eksperimenta. Na svaki se fragment DNA zatim vežu adapteri. Za jedan kraj molekule DNA veže se adapter oblika ukosnice (adapter HF), dok se za drugi kraj molekule veže adapter Y. Ovaj adapter navodi molekulu DNA prema nanopori i veže motorni protein koji nije aktivan sve do dolaska molekule DNA do nanopore. Motorni protein (obično helikaza) razmata molekulu DNA koja zatim prolazi kroz nanoporu brzinom koja omogućava pojedinačno čitanje baza. Nakon sekvenciranja prvog lanca adapter HF omogućuje uvlačenje komplementarnog lanca u nanoporu nakon čega dolazi do njegova sekvenciranja. Pristup sekvenciranja oba lanca DNA naziva se 2D sekvenciranjem, a omogućava višestruko povećavanje točnosti u odnosu na sekvenciranje samo jednog lanca DNA (1D sekvenciranje) (de Lannoy i sur. 2017). Glavni nedostatci 2D sekvenciranja su mogućnost vezanja ukosnice za neodgovarajući lanac čime nastaju kimerne molekule DNA te nepreciznije i sporije sekvenciranje komplementarnog lanca (White i sur. 2017).



Slika 4. Sekvenciranje tehnologijom nanopora. a) Vođenje molekule DNA motornim proteinom do pore. b) Prolazak molekule DNA kroz poru. Preuzeto i prilagođeno iz Leggett i Clark, 2017.

Prolaskom molekule DNA kroz nanoporu dolazi do promjene protoka iona što uzrokuje promjenu električnog potencijala na membrani. Važno je naglasiti da nukleotidi molekule DNA kroz poru prolaze pojedinačno pri čemu određeni nukleotid daje karakterističan električni signal ovisan o susjednim nukleotidima, odnosno o nukleotidima koji su u tom trenutku u kontaktu s nanoporom. Dobivenim električnim signalima pridružuju se događaji (engl. *events*) pri čemu svaki događaj predstavlja pomak molekule DNA u pori za jedan nukleotid. Metodama strojnog učenja iz niza događaja iščitava se slijed sekvenciranih baza (tzv. *basecalling*). Donedavno se smatralo da pentamer nukleotida u kontaktu s porom daje jednoznačni električni signal pri čemu bi se svakom iščitanom događaju prepisao takav pentamer od čijeg bi se neprekinutog slijeda formirao slijed DNA. Međutim, na električne signale utječu i brojne druge značajke kao što je sekundarna struktura ili specifični sljedovi DNA, poput homopolimera, zbog čega noviji programi događajima pripisuju promjenjiv broj nukleotida (de Lannoy i sur. 2017). Također, svakoj bazi određuje se vjerojatnost krivog čitanja iz koje se računa mjera kvalitete (engl. *quality score*). Dobiveni sljedovi svrstavaju se prema aritmetičkoj sredini njihove mjere kvalitete. Sljedovi s aritmetičkom sredinom mjere kvalitete većom od granične vrijednosti svrstavaju se u kvalitetne sljedove, dok se sljedovi s tom vrijednošću manjom od granične vrijednosti svrstavaju u nekvalitetne sljedove. Nekvalitetni sljedovi obično se ne koriste u postupcima poput sastavljanja genoma *de novo* (Tyler i sur. 2018).

Glavna prednost tehnologije sekvenciranja nanoporama leži u relativno brzom i jeftinom dobivanju dugih sljedova s prosječnom dužinom od nekoliko desetaka tisuća pb. Međutim, udio grešaka ove tehnologije procjenjuje se na 12%. Ovaj udio mnogo je veći od udjela grešaka metoda sekvenciranja druge generacije te često stvara probleme u primjenama ove tehnologije u različitim područjima (Kchouk i sur. 2017). Razvoj ove tehnologije kreće se u smjeru poboljšanja same kemije sekvenciranja i razvoja boljih algoritama za određivanje sljedova baza. Također, često se koriste i algoritmi za ispravljanje ovakvih sljedova pomoću točnijih sljedova dobivenih tehnologijom sekvenciranja druge generacije. Analiza grešaka sljedova dobivenih tehnologijom sekvenciranja nanoporama mogla bi pridonijeti razvoju ovakvih algoritama, a samim time i većoj primjeni te iskoristivosti takvih sljedova u rješavanju pitanja različitih područja znanosti. Primjerice, dugi sljedovi dobiveni ovom tehnologijom mogli bi imati brojne primjene u medicini, poput relativno brzog utvrđivanja polimorfizama određenog gena pacijenata ili određivanja uzročnika bolesti radi brzog i jeftinog postavljanja valjane dijagnoze, odnosno određivanja najprikladnije terapije.

Jedna od primjena tehnologije sekvenciranja nanoporama je i sastavljanje genoma *de novo*. Dugi sljedovi dobiveni ovom tehnologijom mogu olakšati sastavljanje različitih dijelova genoma, poput ponavljajućih sljedova ili sljedova niske kompleksnosti (engl. *low complexity regions*). Jedan od čestih pristupa sklapanja genoma je hibridno sklapanje, koje podrazumijeva korištenje sljedova dobivenih različitim tehnologijama sekvenciranja (Pop

2009). Genomi mnogih organizama sklapaju se korištenjem sljedova dobivenih tehnologijom sekvenciranja nanoporama i tehnologijom Illumina (primjerice Passera i sur. 2018). U ovom slučaju dugi sljedovi dobiveni sekvenciranjem nanoporama omogućavaju sklapanje dužih neprekinutih sljedova što podiže kvalitetu sastavljenog genoma. Spomenutim pristupom sklopljen je genom spužve *Eunapius subterraneus* (Glavaš 2018). Analiza grešaka i ispravljanje sljedova dobivenih tehnologijom sekvenciranja nanoporama pomoći će u povećavanju kvalitete sklopljenog genoma ove spužve. To nas dovodi korak bliže do odgovora na već postavljena pitanja unutar biologije spužvi, ali i filogenije te razvoja životinja općenito.

2. OPĆI I SPECIFIČNI CILJEVI RADA

Primjena tehnologije sekvenciranja nanoporama ima velik potencijal u različitim područjima biologije i medicine. Primjerice. dugi sljedovi DNA dobiveni ovom tehnologijom olakšavaju rješavanje brojnih znanstvenih pitanja moderne genomike, poput sekvenciranja genoma *de novo* ili određivanja strukturnih varijanti ili tandemskih polimorfizama sljedova DNA. Međutim, relativno velik udio grešaka u sljedovima dobivenih ovom tehnologijom ograničava i dovodi u pitanje njihovu primjenu. Cilj ovog istraživanja je kvantitativna analiza i ispravak grešaka sljedova DNA. Koristit ćemo podatke dobivene sekvenciranjem nanoporama ogulinske špiljske spužvice (*Eunapius subterraneus* Sket & Velikona, 1984) koji se mogu iskoristiti za poboljšanje kvalitete sklopljenog genoma.

Specifični ciljevi istraživanja su:

1. Usporediti kvalitetne i nekvalitetne sljedove genomske DNA ogulinske špiljske spužvice sekvencirane tehnologijom nanopora s naglaskom na odabrani uzorak od 500 najdužih kvalitetnih i nekvalitetnih sljedova.
2. Napraviti sravnjenje sljedova genomske DNA ogulinske špiljske spužvice dobivenih tehnologijom sekvenciranja Illumina na spomenutim uzorcima sljedova dobivenih tehnologijom sekvenciranja nanoporama.
3. Analizirati i ispraviti pogreške u spomenutim uzorcima sljedova sekvenciranih tehnologijom nanopora koristeći rezultate njihova sravnjenja sa sljedovima dobivenih tehnologijom sekvenciranja Illumina.
4. Provjeriti uspješnost ispravljanja kvalitetnih i nekvalitetnih sljedova sekvenciranih tehnologijom nanopora usporedbom utvrđenih pogrešaka prije i poslije ispravljanja.
5. Odrediti taksonomsku klasifikaciju i funkcijsku klasifikaciju identificiranih proteina pomoću Gene Ontology podataka kvalitetnih i nekvalitetnih sljedova prije i poslije ispravljanja.

3. MATERIJALI I METODE

3.1. Korištene knjižnice

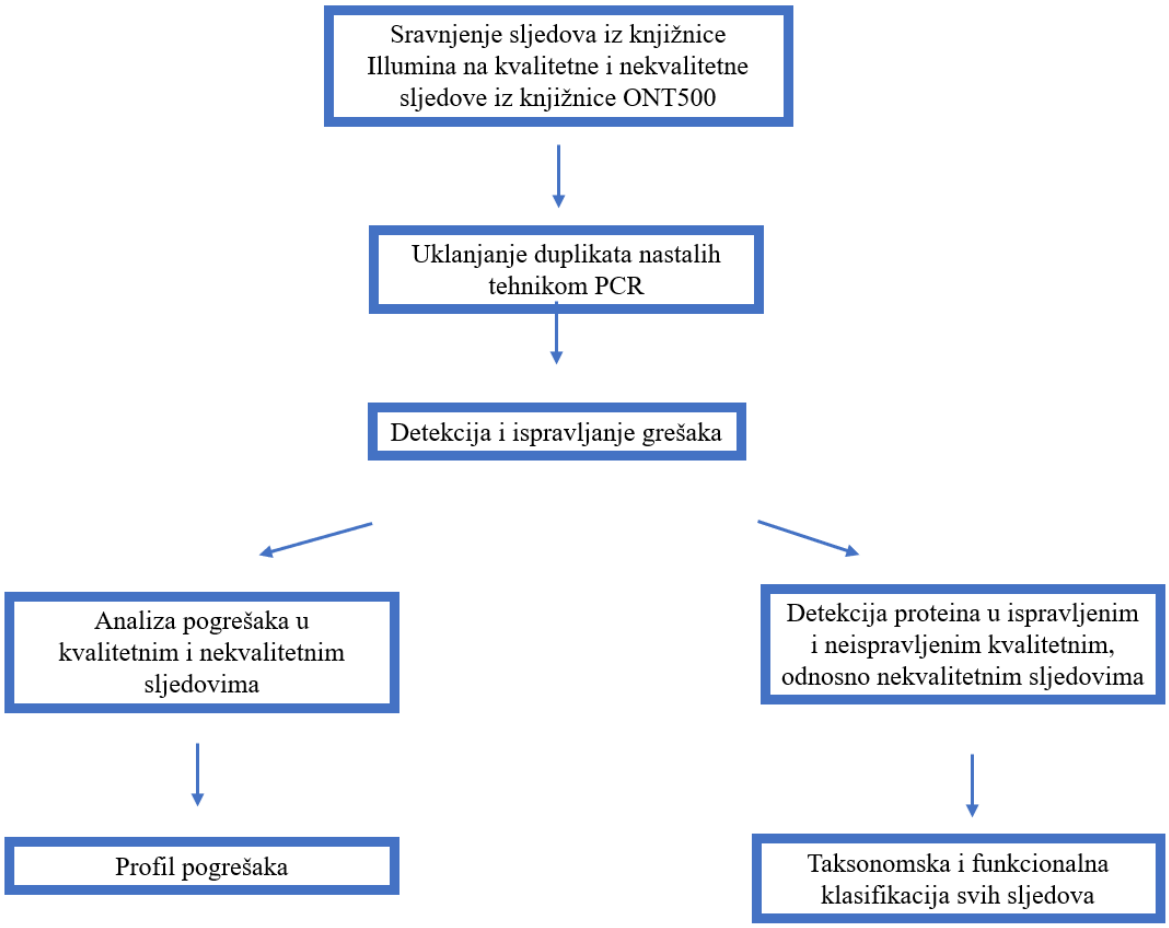
U ovom istraživanju koristili smo dvije otprije dostupne knjižnice. Prva knjižnica (Illumina) dobivena je uparenim sekvenciranjem krajeva na platformi Illumina HiSeq tvrtke Macrogen dok je druga knjižnica (ONT) dobivena 2D sekvenciranjem tehnologijom nanopora (uređaj minION R9, ONT). Oba sekvenciranja provedena su na istoj jedinci primorfa spužve *Eunapius subteranneus* u rujnu 2017. godine.

Sljedovima iz knjižnice Illumina uklonjeni su adapteri upotrebom programa Trimmomatic verzije 0.32 (Bolger i sur. 2014) koristeći sljedeće parametre: „leading:3 trailing:3 slidingwindow:4:15 minlen:50“. Normalizacija sljedova u knjižnici Illumina provedena je upotrebom programa BBnorm (Bushnell 2015). Normalizacijom se smanjuje količina podataka te se ispravljaju pogreške u samim sljedovima koje su nastale tokom sekvenciranja tehnologijom Illumina. Knjižnica ONT dobivena je prevoditeljem baza Albacore pri čemu su dobiveni sljedovi razvrstani na kvalitetne i nekvalitetne s granicom prosječne mjere kvalitete 7. Podaci o sljedovima u korištenim knjižnicama nalaze se u Tablici 1.

Tablica 1. Podaci o predobrađenim sljedovima dobivenih sekvenciranjem genomske DNA ogulinske špiljske spužvice

Knjižnica	Broj sljedova	Prosječna duljina slijeda / pb	Ukupna duljina sljedova / pb
Illumina	75 059 139	151	11 333 929 989
ONT	267 915	24 956	49 911 278

Velika količina podataka u knjižnici ONT zahtijeva velik broj resursa zbog čega smo analizirali i ispravili 500 najdužih kvalitetnih i 500 najdužih nekvalitetnih sljedova (knjižnica ONT500). Na Slici 5 prikazana je shema provedenih postupaka na knjižnici ONT500. Programska skripta pokrenuta u ljusc*i bash* koja provodi sve daljnje opisane postupke istraživanja nalazi se u Prilogu 1, dok se programski postupak za dobivanje slike opisanim postupcima nalazi u Prilogu 2.



Slika 5. Shema provedenih postupaka na knjižnici ONT500

3.2. Statistička obrada podataka programskim jezikom R

Napravili smo statistička analizu kvalitetnih i nekvalitetnih sljedova iz knjižnica ONT i ONT500 korištenjem programskog jezika R verzije 3.5.1. (R Core Team 2018). R je slobodno dostupno programsko okruženje pogodno za statističku obradu velikih količina podataka pri čemu posjeduje dobro razvijenu podršku za biološke podatke. Sadrži mnogo paketa s različitim funkcijama i bazama podataka koji znatno proširuju njegovu funkcionalnost. U analizi kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT koristili smo Anderson-Darlingov test normalnosti s razinom značajnosti postavljenom na 95% ($p<0,05$). Također, Wilcoxonovim testom usporedili smo mjere kvalitete i duljine spomenutih sljedova.

3.3. Sravnjenje sljedova iz knjižnice Ilumina na sljedove iz knjižnice ONT500

Pogreške koje se nalaze u kvalitetnim i nekvalitetnim sljedovima iz knjižnice ONT500 identificirali smo uspoređujući ih sa sljedovima iz knjižnice Illumina. Postupak uspoređivanja dvaju sljedova s ciljem pronalaska podsekvence većeg slijeda koja odgovara manjem slijedu s dopuštanjem određenog broja pogreški naziva se sravnjenje sljedova (engl. *sequence alignment*). Sravnjenje velikog broja sekvenci je računalno zahtjevan proces zbog velike količine podataka i grešaka u samim sljedovima koje se javljaju tijekom sekvenciranja. Postoje različiti programi koji u što kraćem vremenu pronalaze optimalno rješenje za sravnjenje velikog broja sljedova. Za sravnjenje kvalitetnih i nekvalitetnih sljedova knjižnice ONT500 sa sljedovima knjižnice Illumina koristili smo programe minimap2 i BWA (engl. *Burrows-Wheeler Aligner*). Rezultat sravnjenja su datoteke formata SAM (engl. *sequence alignment map*) koje smo dalje obrađivali pomoću programskog paketa SAMBAMBA.

3.3.1. Program minimap2

Program minimap2 slijedi tipičan postupak sravnjenja klica-lanac (engl. *seed-chain-align*) koji se koristi kod većine programa za sravnjenje genoma. Klice su kratke riječi određene fiksne duljine koje se traže u unesenim (engl. *query sequence*) i referentnim sljedovima (engl. *reference sequence*). Program minimap2 skuplja minimizatore, odnosno male dijelove klica (Roberts i sur. 2004) referentnih sljedova te ih indeksira u tzv. hash tablici (engl. *hash table*), Zatim, za svaki uneseni slijed program minimap2 uzima minimizatore unesenog slijeda kao klice, pronalazi točne podudarnosti, to jest sidra s referentnim slijedom pri čemu identificira skupove kolinearnih sidara kao lance. Ako se zahtijeva poravnanje na osnovnoj razini, program minimap2 primjenjuje dinamičko programiranje kako bi se proširio od krajeva lanaca i zatvorio područja između susjednih sidara u lancima (Li 2017).

Program minimap2 pokrenuli smo koristeći dvije skupine različitih parametara. U oba slučaja uneseni sljedovi predstavljaju sljedove iz knjižnice Illumina, dok su referentni sljedovi kvalitetni i nekvalitetni sljedovi iz knjižnice ONT500. U prvoj skupini, nazvanoj minimap2, parametri su sljedeći: “-x sr -k11 -w9 --end-bonus 50 -N 500 --secondary=no”. U drugoj skupini, nazvanoj minimap2-last, svi su parametri po zadanim vrijednostima, osim dodanih parametara “-x sr --no-long-join -r50 --end-bonus 50 --secondary=no”.

3.3.2. Program BWA

Burrows-Wheelerov alat za poravnanja je programski paket za sravnjenje unesenih sljedova na velike referentne sljedove. Program se temelji na Burrows-Wheelerovoj transformaciji pri čemu omogućava efikasna sravnjenja kratkih sljedova uz dopuštanje postojanja grešaka i praznina (Li i Durbin 2010). Koristili smo program BWA-MEM verzije 0.7.15. sa zadanim parametrima i parametrom „L 500“. Uneseni sljedovi ponovno predstavljaju sljedove iz knjižnice Illumina, dok su referentni sljedovi kvalitetni i nekvalitetni sljedovi iz knjižnice ONT500.

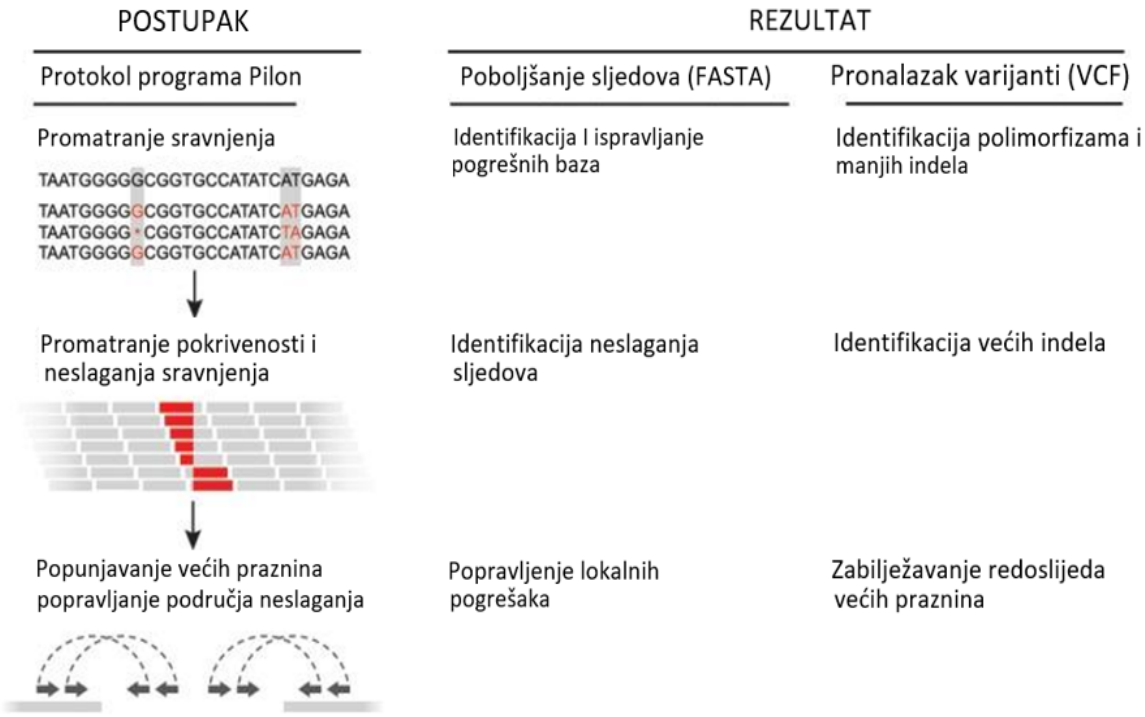
3.3.3. Format SAM i program SAMBAMBA

Format SAM je tekstualni format za spremanje podataka dobivenih sravnjenjem unesenih sljedova na referentne sljedove. Format SAM obično započinje sa zaglavljem čije se linije odnose na svako sravnjenje. Svaka linija počinje sa znakom „@“ nakon kojeg slijedi glavni dio linije koji se sastoji od 11 polja. Svako polje sadrži informacije o sravnjenju, kao što su kvaliteta sravnjenja slijeda, kvaliteta baze i položaj ulaznog slijeda na referentnom slijedu (Wysoker i sur. 2009). Program SAMBAMBA predstavlja skup alata koji omogućava manipulaciju datoteke formata SAM što uključuje filtriranje, sortiranje, indeksiranje i uklanjanje dupliciranih sljedova (Tarasov i sur. 2017).

Dobivene datoteke formata SAM pretvorili smo u format BAM (binarni oblik formata SAM, engl. *binary alignment map*). Nastale datoteke formata BAM smo zatim sortirali i indeksirali. Iz spomenutih datoteka također smo ukloonili duplicirane sljedove iz knjižnice Illumina nastale metodom PCR tijekom postupka sekvenciranja. Opisane obrade datoteka proveli smo programom SAMBAMBA.

3.4. Ispravljanje sljedova upotrebom programa Pilon

Radi detekcije i ispravka pogrešaka u kvalitetnim i nekvalitetnim sljedovima knjižnice ONT500, obrađenu datoteku formata BAM analizirali smo upotrebom programa Pilon. Program Pilon je skup alata koji istovremeno pronalazi i ispravlja varijante u referentnim sljedovima (Slika 6). Pri pronalasku i ispravljanju varijanti program Pilon koristi sravnjenja sljedova iz datoteke formata BAM za identificiranje specifičnih razlika između referentnog slijeda iz knjižnice ONT500 ii unesenih sljedova iz knjižnice Illumina. Referentni sljedovi se ispravljaju pri čemu nastaje datoteka formata VCF (engl. *variant call format*) u kojoj su zabilježene sve varijante referentnog slijeda, kao i datoteka formata FASTA koja sadrži ispravljene referentne sljedov (Walker i sur. 2014).



Slika 6. Pregled rada programa Pilon. U lijevom stupcu prikazani su konceptualni koraci programa Pilon, dok srednji i desni stupac opisuju svaki korak kojim program Pilon dovodi do poboljšanja referentnih sljedova, odnosno otkrivanja njihovih varijanti. Tijekom prvog koraka (gornji red) program Pilon skenira sravnjenja pri čemu utvrđuje mjesta neslaganja unesenih sljedova s referentnim sljedovima. Također, u prvom koraku ispravljaju se male pogreške i utvrđuju polimorfizmi te indeli referentnih sljedova. Tijekom drugog koraka (srednji red) program Pilon traži odstupanja pokrivenosti i usklađivanja kako bi identificirao potencijalna neslaganja unesenih i referentnih sljedova. U zadnjem koraku (donji red) program Pilon koristi sljedove i informaciju o uparenosti sljedova koji su usidreni na krajeve suprotnih područja praznine u referentnim sljedovia kako bi ispravio te sljedove. Rezultat je datoteka formata FASTA s ispravljenim referentnim sljedovima i/ili datoteka formata VCF koja sadrži varijante referentnih sljedova. Preuzeto i prilagođeno iz Walker i sur. 2014.

3.5. Profil pogrešaka kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT500

Pomoću dobivene datoteke formata VCF analizirali smo pogreške u kvalitetnim i nekvalitetnim sljedovima knjižnice ONT500. Format VCF je tekstualni format za spremanje podataka o polimorfizmu sljedova DNA poput SNP-ova (polimorfizmi jedne varijante, engl. *single nucleotide polymorhpism*), insercija, delecija i strukturnih varijanta. Sadrži meta informacijske linije, redak zaglavlja i odjeljak podataka u kojem svaka linija sadrži određene informacije o poziciji polimorfizma u sljedovima. Meta informacijske linije daju standardizirani opis oznaka i anotacija korištenih u odjeljku podataka. U odjeljku podataka nalazi se 8 stupaca koji sadrže imena referentnih sljedova, početak prve baze, pronađene varijante, referentnu bazu slijeda, promijenjenu bazu slijeda i informacije koje govore o dubini slijeda te kvaliteti

sravnjenja. Dubina slijeda (pokrivenost) predstavlja broj sljedova koji su sravnjeni s referentnim slijedom na određenoj poziciji pronađene varijante. Kvaliteta sravnjenja (engl. *mapping quality*) govori kolika je vjerojatnost da se određeni sljedovi nalaze na određenoj poziciji. Veće vrijednosti pokrivenosti i kvalitete sravnjenja podrazumijevaju veću pouzdanost pronađene varijante. Za prikaz navedenih informacija često se koristi i format BCF (engl. *binary variant call format*), birarni oblik formata VCF. Programi VCFtools i BCFtools koriste se za manipulaciju zapisa u spomenutim formatima, što uključuje filtriranje specifičnih varijanti, spajanje više datoteka u jednu i dobivanje sažete statistike o varijantama (Danecek i sur. 2011).

Koristeći programe VCFtools i BCFtools filtrirali smo dobivene datoteke formata VCF radi izdvajanja pogrešaka u referentnim sljedovima. Filtriranjem smo ostavili one varijante, odnosno pogreške, koje imaju kvalitetu sravnjenja veću od 20 i pokrivenost veću od 40. Također, uzeli smo samo one pogreške na čijoj se poziciji unutar 97% sljedova iz knjižnice Illumina pojavljuje ista baza. Sažetu statistiku o pogreškama dobili smo upotrebom programa BCFtools nakon čega smo rezultate statistički obradili upotrebom programskog jezika R. Statistička obrada obuhvaćala je analize stopa i udjela pogrešaka kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT500. Stope pogrešaka ovih dvaju skupina sljedova usporedili smo Wilcoxonovim testom s razinom značajnosti od 95% ($p<0.05$), dok smo udjele vrsta pogrešaka usporedili analizom PCA (engl. *Principal component analysis*). Također, posebno smo analizirali pogrešno očitane baze i indele navedenih sljedova pri čemu smo u analizi duljina indela proveli Anderson-Darlingov test normalnosti s razinom značajnosti postavljenom na 95% ($p<0,05$). Isto tako, statističke usporedbe indela u kvalitetnim i nekvalitetnim sljedovima proveli smo pomoću Wilcoxonova testa s razinom značajnosti postavljenom na 95% ($p<0,05$).

Radi usporedbe stopa i udjela pogrešaka ispravljenih i neispravljenih sljedova, programom Pilon napravljena je datoteka formata VCF za sravnjenje ispravljenih kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT500 te sljedova iz knjižnice Illumina. Pritom su korišteni isti parametri sravnjenja i filtriranja kao pri dobivanju datoteke VCF za sravnjenje neispravljenih sljedova. Usporedili smo duljine neispravljenih i ispravljenih sljedova. Također, iz dobivene datoteke formata VCF izračunali smo stope pogrešaka ispravljenih sljedova, kao i relativne promjene udjela parova krivo očitanih baza i njihovih točnih varijanti.

3.6. Analiza ispravljenosti sljedova upotrebom programa DIAMOND

Za provjeru učinkovitosti ispravljanja kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT500 usporedili smo proteine pronađene u navedenim sljedovima prije i poslije ispravljanja. Kvalitetni i nekvalitetni neispravljeni, odnosno ispravljeni sljedovi pretraženi su za proteine koji se nalaze u neredundantnoj proteinskoj bazi programom DIAMOND (engl. *double index alignment of next-generation sequencing data*).

DIAMOND je program za uspoređivanje nukleotidnih sekvenci s bazom kodirajućih proteinskih sljedova. Algoritam ovog programa pronalazi zajedničke klice između unesenog slijeda i sljedova u proteinskoj bazi nakon čega nastala sravnjena (pogotke) produljuje Smith-Waterman algoritmom lokalnog sravnjenja. Za razliku od većine ostalih programa za pretragu proteinskih sljedova, koji također koriste sravnjenje klica za pronalaženja pogotka, program DIAMOND ima nekoliko karakteristika koje ga čine bržim i osjetljivijim. Primjerice, ovaj program koristi modificiranu aminokiselinsku abecedu čiji je broj znakova smanjen s 20 na 11. One aminokiseline koje su slične po svojstvima označene su istim znakom (uglate zagrade označavaju jedan znak): [KREDQN] [C] [G] [H] [ILV] [M] [F] [Y] [W] [P] [STA] čime se postiže veća osjetljivost pretraživanja. Drugo poboljšanje programa je korištenje razmaknutih klica (engl. *spaced seeds*), odnosno duljih klica u kojima se koristi samo podskup pozicija (Slika 7). Broj i točan raspored tih pozicija nazivaju se težina i oblik razmaknute klice. Pojedinačna razmaknuta klica može imati bolje rezultate od susjedne neprekidne klice iste težine ako je njezin oblik prikladno odabra0n. Također, osjetljivost algoritma povećava se korištenjem dodatnih oblika klica. Prema zadanim postavkama, DIAMOND koristi skup od četiri oblika klica duljine 15-24 i težine 12 (Buchfink i sur. 2014).

Slika 7. a) Četiri oblika klice veličine 12 koje program DIAMOND koristi prema zadanim postavkama. Jedinice označavaju pozicije znakova koje se koriste, dok nule označavaju pozicije koje se zanemaruju. b) Ilustracija primjene razmaknute klice na slaganje slova između unesene i referentne sekvence. Preuzeto i prilagođeno iz Buchfink i sur. 2014.

Isto tako, DIAMOND koristi pristup dvostrukog indeksiranja pri čemu se indeksiraju unesene i referentne sekvence. Paralelnim prelaskom dva popisa indeksa identificiraju se sve

podudarne klice između unesene i referentne sekvence što dopušta računanje lokalnog poravnanja na odgovarajućim lokacijama klica.

Pokrenuli smo program DIAMOND za neispravljene i ispravljene kvalitetne, odnosno nekvalitetne sljedove knjižnice ONT500 koristeći sljedeće parametre: “--sensitive --evalue 10 -F 15 --range-culling”.

3.7. Taksonomska i funkcijska klasifikacija sljedova upotrebom programa MEGAN

Rezultate sravnjenja dobivene programom DIAMOND analizirali smo upotrebom programa MEGAN verzije 6.12.5. (Huson i sur. 2016). MEGAN je besplatan program koji upotrebom algoritma LCA (engl. *lowest common ancestor*) provodi taksonomsku i funkcijsku analizu sljedova. Algoritam LCA iz rezultata programa DIAMOND za svaki ulazni slijed odabire 10% najbolje sravnjenih referentnih sljedova iz proteinske baze. Zatim, pojedinom ulaznom slijedu dodjeljuje najvišu zajedničku taksonomsku kategoriju od svih kategorija odabranih referentnih sljedova iz proteinske baze.

Programom MEGAN odredili smo taksonomske kategorije kvalitetnim i nekvalitetnim neispravljenim, odnosno ispravljenim sljedovima iz knjižnice ONT500. Također, napravili smo funkcijsku analizu navedenih sljedova, tj. proteina pronađenih u njima koristeći klasifikaciju InterPro2GO unutar programa MEGAN pri čemu smo pronađenim proteinima odredili molekularnu funkciju, staničnu lokalizaciju i ulogu u biološkim procesima. InterPro je integrirani izvor proteinskih obitelji, domena i mjesta koji su dobiveni iz niza različitih proteinskih baza podataka. Svakom unosu u bazu InterPro pridodaje se pojam GO (engl. *Gene Ontology*) koji opisuje očuvanu funkciju ili lokalizaciju određenog proteina (Mulder i sur. 2003). Dakle, projekt GO omogućio je stvaranje seta hijerarhijski organiziranog vokabulara koji opisuje gene i genske produkte.

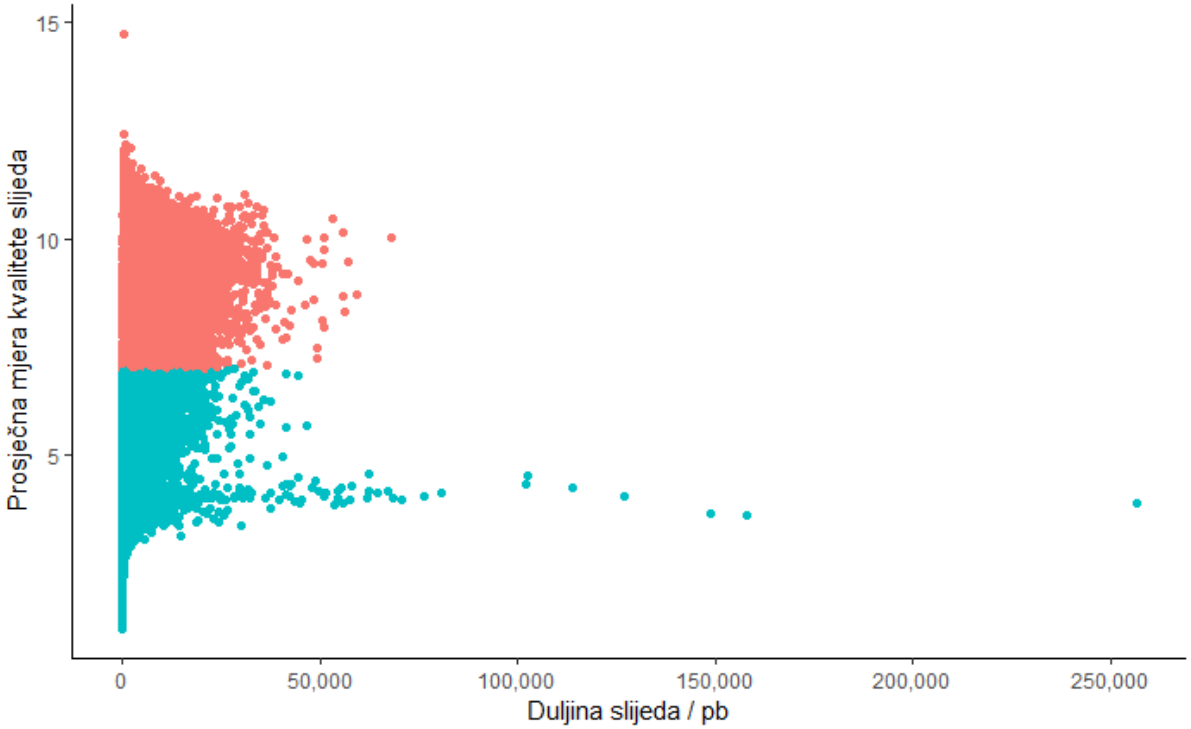
4. REZULTATI

4.1. Statistička obrada kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT i ONT500

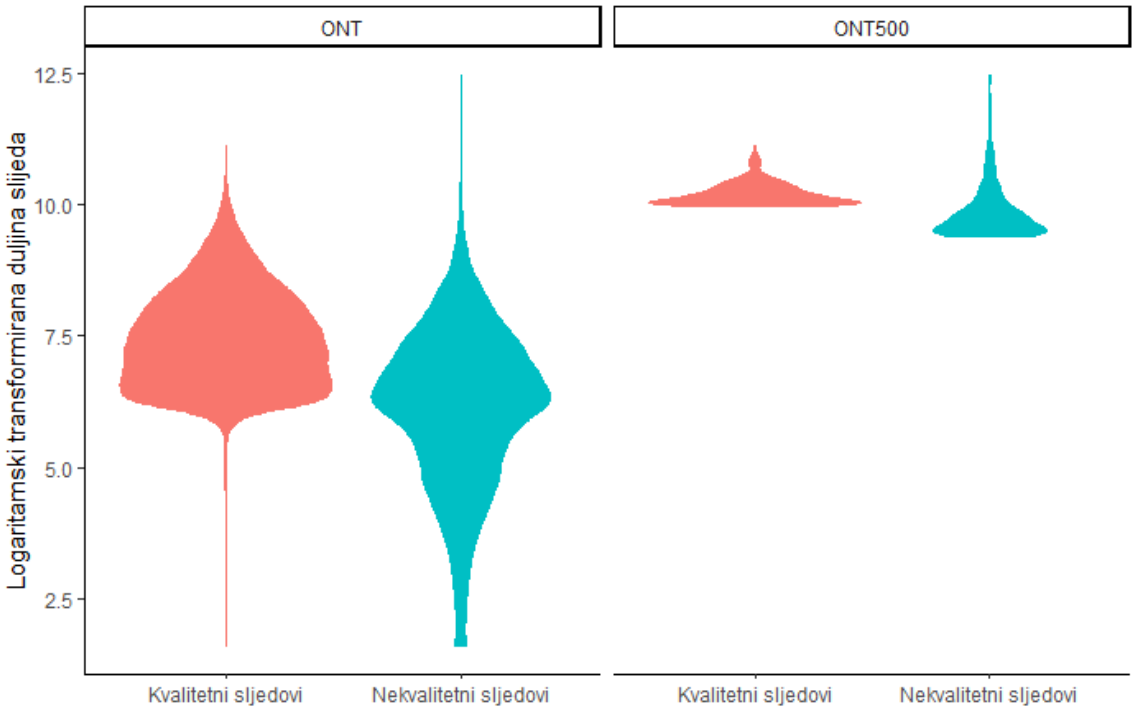
Statistički podaci analize kvalitetnih i nekvalitetnih sljedova prikazani su u Tablici 2. Rezultati Anderson-Darlingova testa normalnosti pokazali su da raspodjela duljina i kvalitete svih kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT nisu normalne ($p<0,05$). Također, rezultati Wilcoxonova testa pokazali su postojanje statistički značajne razlike između duljina i kvaliteta obje skupine sljedova ($p<0,05$). Vidljivo je da kvalitetni sljedovi imaju veću ukupnu duljinu, iako je njihov ukupan broj manji od ukupnog broja nekvalitetnih sljedova. Isto tako, najdulji nekvalitetni slijed veći je od najduljeg kvalitetnog slijeda, a medijan kvalitete kvalitetnih sljedova puno je veći od medijana kvalitete nekvalitetnih sljedova. Najdulji sljedovi iz knjižnice ONT spadaju u nekvalitetne sljedove, čija je prosječna mjera kvalitete manja od 7. Iako je ukupna duljina nekvalitetnih sljedova puno manja od ukupne duljine kvalitetnih sljedova, postoji velik broj nekvalitetnih sljedova relativno velike duljine s prosječnom ocjenom kvalitete blizu 7. Također, postoji nekoliko nekvalitetnih sljedova izuzetno velikih duljina, s maksimumom od 255 kb. (Slika 8). Isto tako, potrebno je istaknuti da je raspon duljine nekvalitetnih sljedova puno veći od raspona duljine kvalitetnih sljedova (Slika 9).

Tablica 2. Statistički podaci o kvaliteti i duljini kvalitetnih i nekvalitetnih sljedovima iz knjižnice ONT dobivenih sekvenciranjem genomske DNA ogulinske špiljske spužvice

Skup sljedova	Ukupna duljina / pb	Medijan duljina sljedova /pb	Maksimalna duljina slijeda / pb	Minimalna duljina slijeda / pb	Ukupan broj sljedova	Medijan mjere kvalitete sljedova
Kvalitetni sljedovi	268471694	1557	68106	5	97379	8,90
Nekvalitetni sljedovi	92889554	547	256645	5	78863	4,70



Slika 8. Ovisnost prosječne mjere kvalitete slijeda o duljini kvalitetnih (crveno) i nekvalitetnih (plavo) sljedova iz knjižnice ONT.



Slika 9. Raspodjela logaritamski transformiranih duljina kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT (lijevo) i knjižnice ONT500 (desno). Utvrdili smo značajnu statističku razliku između raspodjela duljina u kvalitetnim i nekvalitetnim sljedovima (Wilcoxonov test, $p<2{,}2*10^{-16}$).

4.2. Rezultati sravnjenje sljedova iz knjižnice Ilumina na sljedove iz knjižnice ONT500

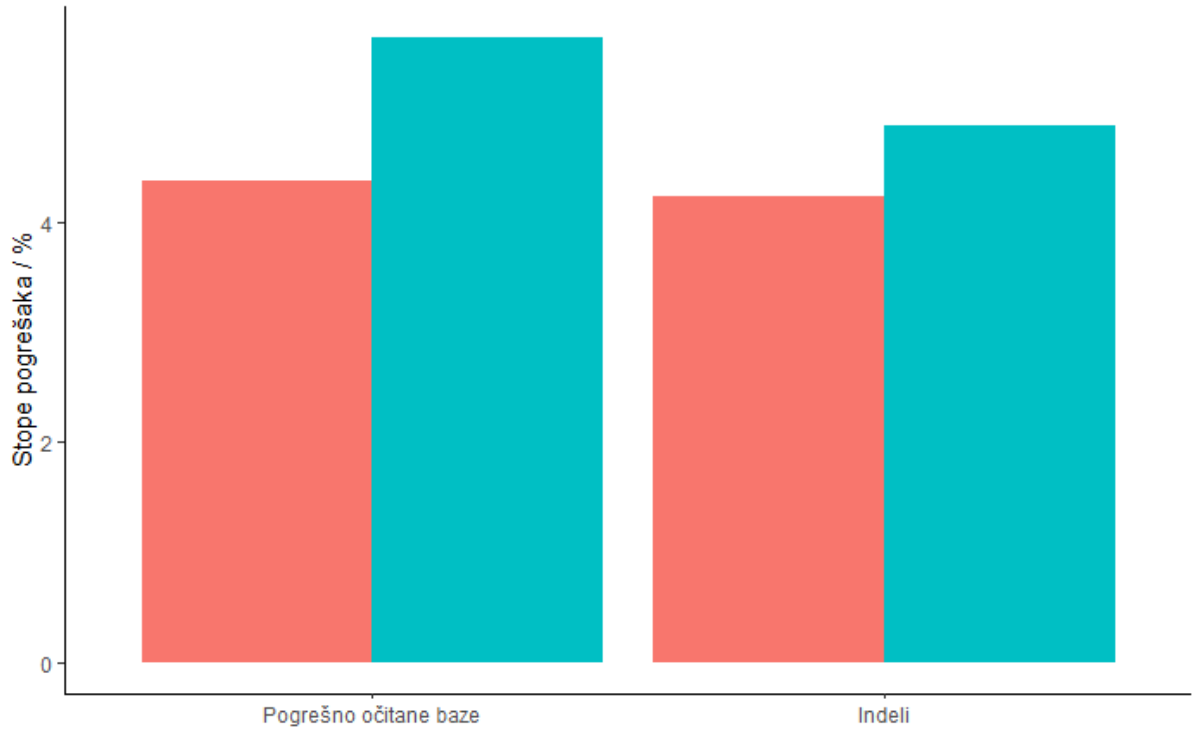
Napravili smo sravnjenje sljedova knjižnice Illumina na kvalitetne i nekvalitetne sljedove knjižnice ONT500 koristeći dvije skupine parametara programa minimap2 (minimap2 i minimap2-last) te program BWA. Statistički podaci o sravnjenjima prikazani su u Tablici 3. Vidljivo je da su za sve skupine parametara prosječna kvaliteta i prosječna pokrivenost sravnjenja, kao i udio sravnjenih sljedova iz knjižnice Illumina veći u sravnjenjima kvalitetnih sljedova. Najveća prosječna pokrivenost kvalitetnih i nekvalitetnih sljedova te najveći udio sravnjenih sljedova iz knjižnice Illumina dobiveni su skupinom parametara minimap2. Također, kvaliteta sravnjenja kvalitetnih sljedova najveća je upotrebom skupine minimap2, dok je kvaliteta sravnjena nekvalitetnih sljedova najveća upotrebom programa BWA. Korištenjem sve tri skupine parametara sravnjenja dobiveni udio GC sravnjenih dijelova kvalitetnih sljedova iznosi oko 43% dok je dobiveni udio GC sravnjenih dijelova nekvalitetnih sljedova veći. Za daljnje analize grešaka i ispravljanje kvalitetnih te nekvalitetnih sljedova iz knjižnice ONT500 koristili smo rezultate sravnjenja dobivene korištenjem skupine parametara minimap2.

Tablica 3. Statistički podaci o sravnjenjima sljedova iz knjižnice Illumina na kvalitetne i nekvalitetne sljedove iz knjižnice ONT500

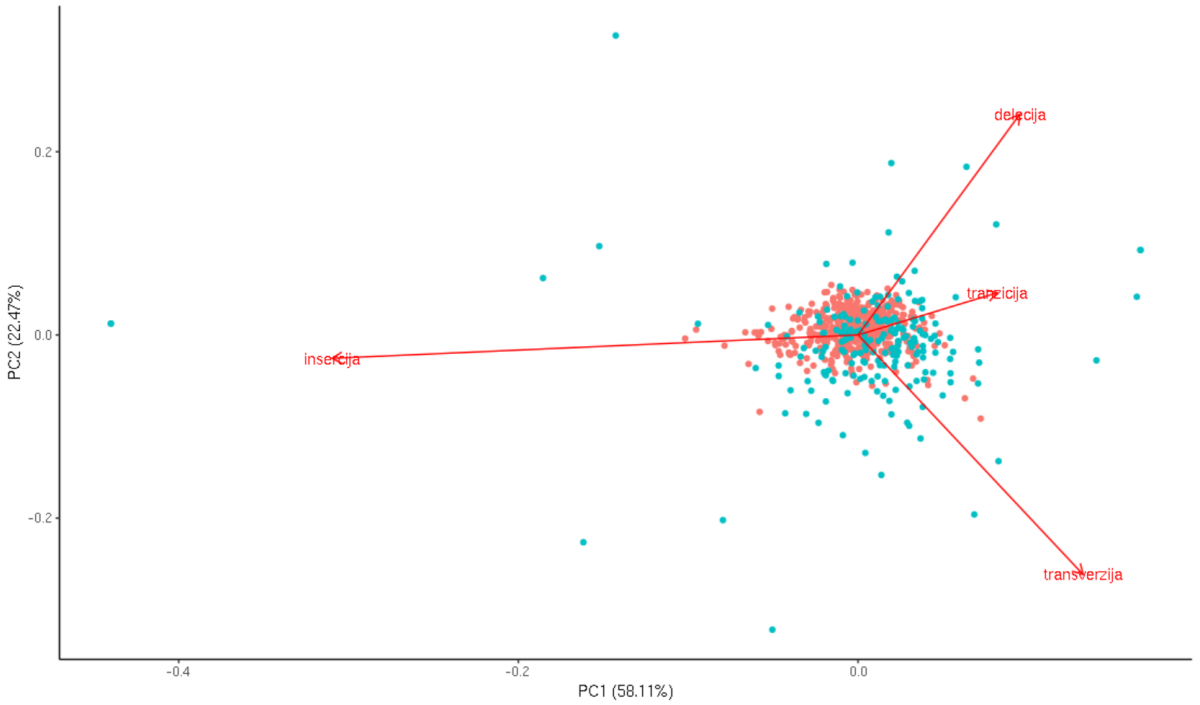
		Prosječna pokrivenost	Udio sravnjenih sljedova iz knjižnice Illumina / %	Udio GC / %	Prosječna kvaliteta sravnjenja
minimap2	Kvalitetni sljedovi	269,57	28,45	43,84	35,99
	Nekvalitetni sljedovi	95,14	10,68	48,70	12,25
minimap2-last	Kvalitetni sljedovi	124,52	15,99	42,99	23,26
	Nekvalitetni sljedovi	28,38	3,04	53,59	3,54
BWA	Kvalitetni sljedovi	210,22	20,17	43,39	35,92
	Nekvalitetni sljedovi	59,38	6,07	49,91	20,77

4.3. Profil pogrešaka kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT500

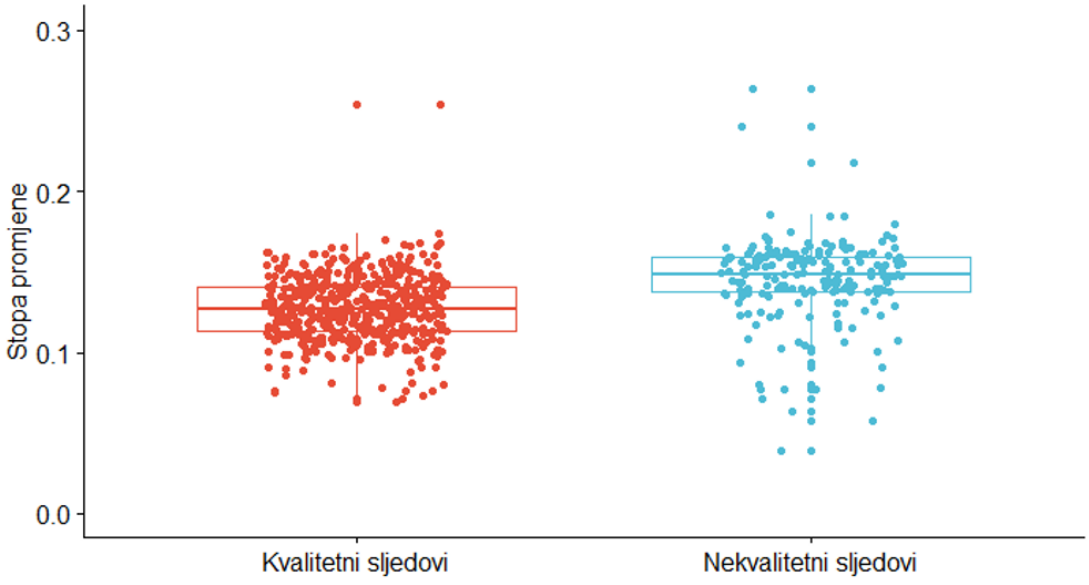
Prosječna stopa pogreške u kvalitetnim sljedovima iznosi 8,61%, a u nekvalitetnim sljedovima 10,55% pri čemu je stopa pogreške definirana kao broj grešaka normaliziran na duljinu slijeda. Na Slici 10 prikazane su stope pogrešno očitanih baza i indela u kvalitetnim i nekvalitetnim sljedovima. Stopa pogrešno očitanih baza i indela veća je u nekvalitetnim sljedovima, dok je u obje skupine sljedova stopa pogrešno očitanih baza veća od stopa indela. Pogreške kvalitetnih i nekvalitetnih sljedova također smo karakterizirali analizom glavnih komponenti, a rezultat analize prikazan je na Slici 11. Analizirana je i raspodjela stopa pogrešaka kvalitetnih i nekvalitetnih sljedova, a rezultat je prikazan na Slici 12. Proveli smo Wilcoxonov test čiji rezultat ukazuje na statistički značajnu razliku ($p<0,05$) između stopa promjena kvalitetnih i nekvalitetnih sljedova.



Slika 10. Stope pogrešno očitanih baza i indela u kvalitetnim (crveno) i nekvalitetnim (plavo) sljedovima iz knjižnice ONT500.



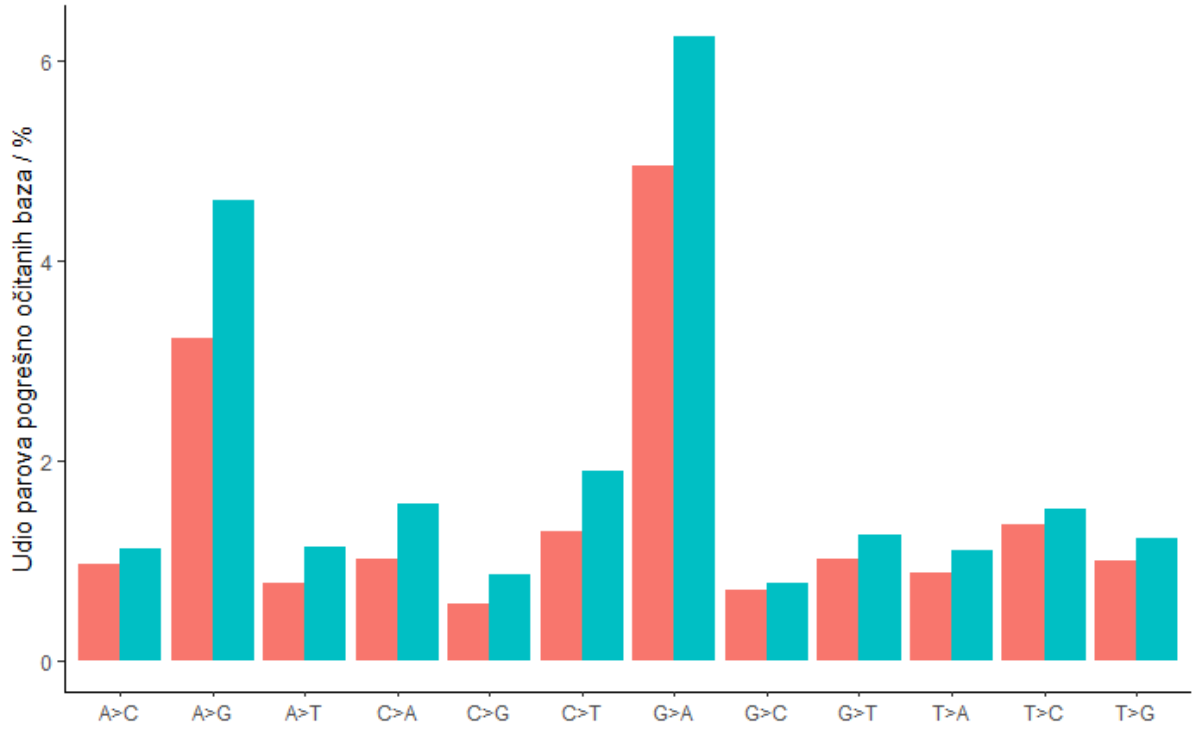
Slika 11. Analiza glavnih komponenti udjela insercija, delecija, tranzicija i transverzija po slijedu kod kvalitetnih (crveno) i nekvalitetnih (plavo) sljedova iz knjižnice ONT500. Vektori označeni crvenom bojom ukazuju na smjer i jačinu utjecaja pojedine varijable na raspodjelu sljedova. Transverzije i tranzicije definirane su kao parovi pogrešno očitanih baza i njihovih točnih varijanti pri čemu tranzicije uključuje parove purin - purin i pirimidin - pirimidin, dok transverzije uključuju parove purin - pirimidin te pirimidin - purin.



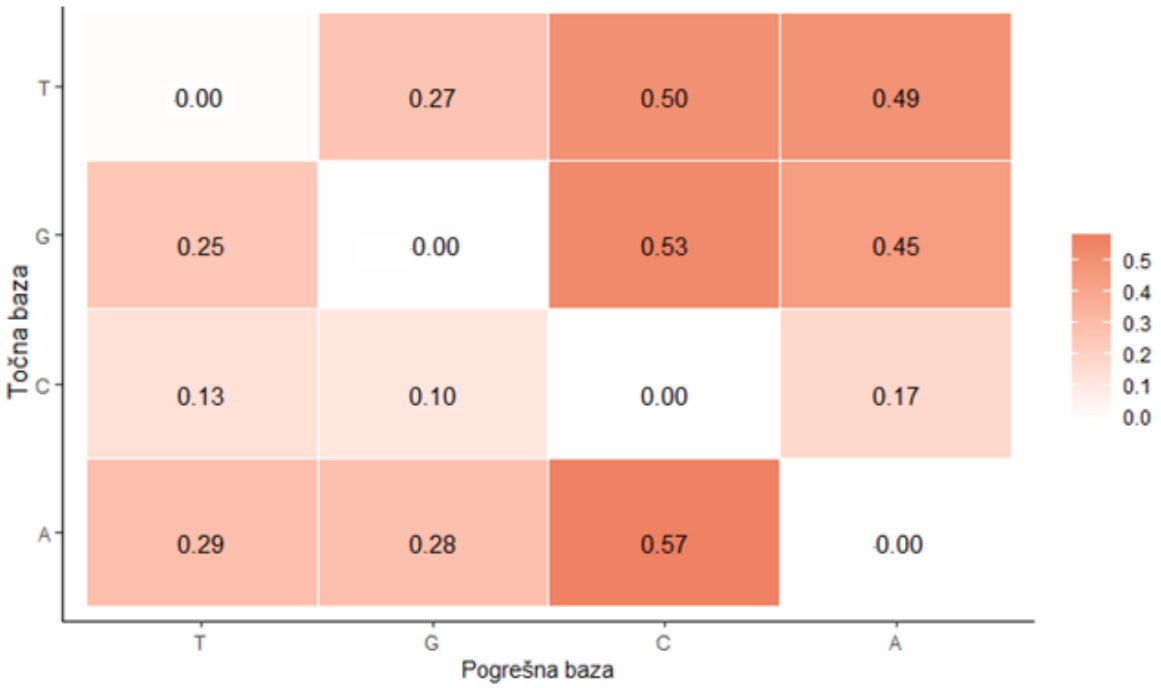
Slika 12. Kvantilni dijagram raspodjele stope promjena u kvalitetnim (crveno) i nekvalitetnim (plavo) sljedovima iz knjižnice ONT500. Stopa promjene računa se kao broj pogrešaka normaliziran na duljinu slijeda. Stope promjena u kvalitetnim i nekvalitetnim sljedovima statistički su značajne (Wilcoxonov test, $p<2,2*10^{-6}$).

4.3.1. Pogrešno očitane baze u kvalitetnim i nekvalitetnim sljedovima iz knjižnice ONT500

Na Slici 13 prikazani su udjeli parova pogrešno očitanih baza i njihovih točnih varijanti u kvalitetnim i nekvalitetnim sljedovima iz knjižnice ONT500. Radi usporedbe kvalitetnih i nekvalitetnih sljedova, udjele parova pojedinih baza i njihovih točnih varijanti u nekvalitetnim sljedovima podijelili smo s odgovarajućim udjelima parova pojedinih baza i njihovih točnih varijanti u kvalitetnim sljedovima te smo dobivene omjere oduzeli s 1. Dobivene vrijednosti prikazane su na Slici 14.



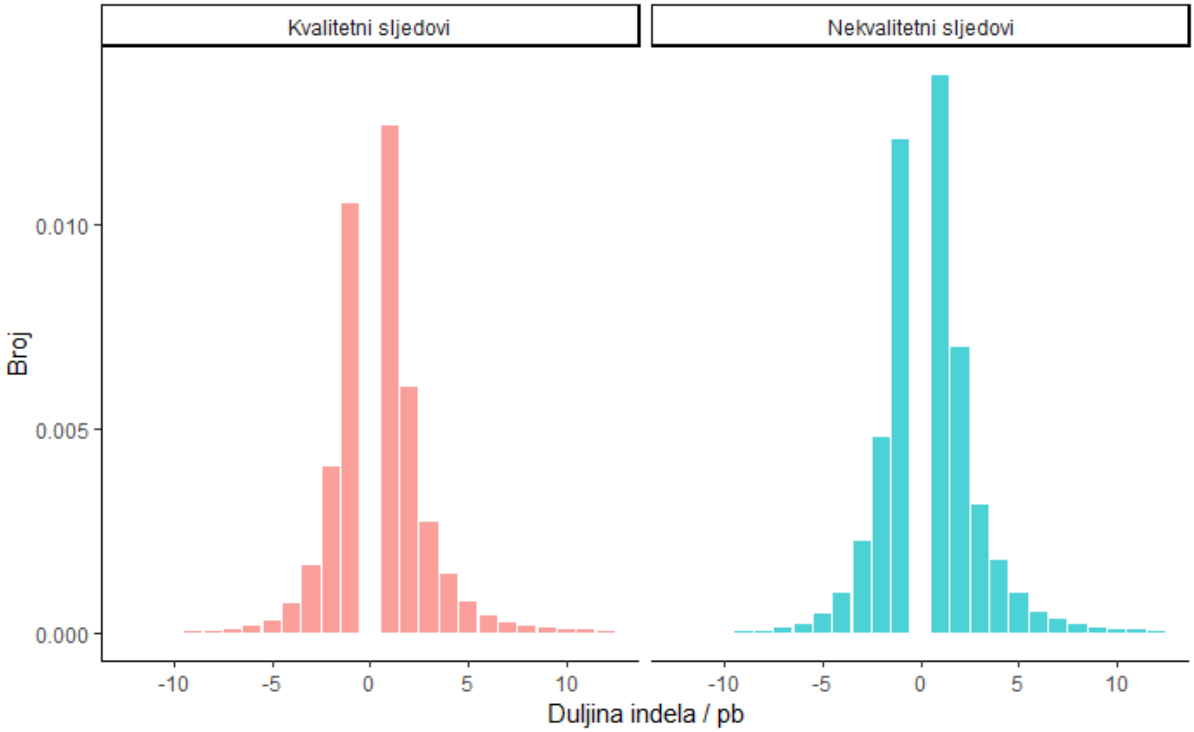
Slika 13. Udjeli parova pogrešno očitanih baza i njihovih točnih varijanti u kvalitetnim (crveno) i nekvalitetnim (plavo) sljedova iz knjižnice ONT500. Navedeni udio definiran je kao omjer broja parova pojedinih baza i njihovih točnih varijanti te broja pojedine vrste pogrešno očitane baze.



Slika 14. Omjeri udjela parova pogrešno očitanih baza i njihovih točnih varijanti u nekvalitetnim i kvalitetnim sljedovima iz knjižnice ONT500 oduzeti s 1. Raspon boja prikazan je logaritamskom skalom.

4.3.2. Indeli u kvalitetnim i nekvalitetnim sljedovima iz knjižnice ONT500

Pri analizi indela odredili smo broj i raspored duljina insercija te delecija u kvalitetnim, odnosno nekvalitetnim sljedovima. Broj insercija i delecija normalizirali smo s ukupnom duljinom analiziranih kvalitetnih i nekvalitetnih sljedova. Broj insercija i delecija veći je u nekvalitetnim sljedovima (Slika 15). Najdulja delecija u kvalitetnim sljedovima iznosi 2526 pb, dok najdulja delecija u nekvalitetnim sljedovima iznosi 93 pb. S druge strane, najdulja insercija u kvalitetnim sljedovima iznosi 167 pb, dok najdulja insercija u nekvalitetnim sljedovima iznosi 34 pb. Medijani duljina indela svih skupina sljedova iznose 1 pb. Rezultati Anderson-Darlingova testa normalnosti pokazali su da raspodjela indela kod kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT500 nije normalna ($p<0,05$). U skladu s time proveli smo Wilcoxonov test čiji su rezultati pokazali statistički značajnu razliku ($p<0,05$) između veličine indela u kvalitetnim i nekvalitetnim sljedovima.



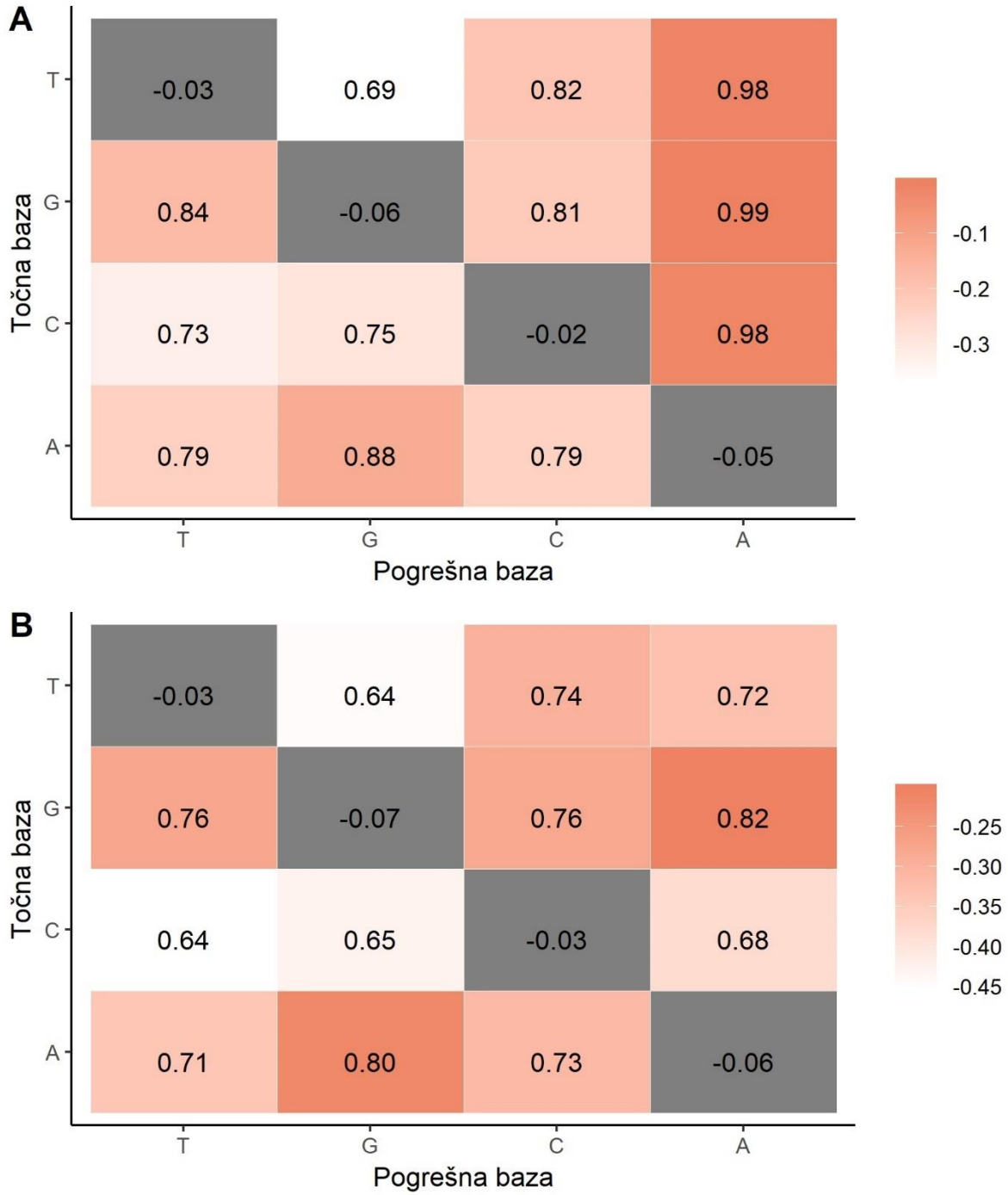
Slika 15. Raspodjela broja indela u rasponu duljine od -10 do 12 (N=XX, 99,3% uzorka) kod kvalitetnih (lijevo) i nekvalitetnih (desno) sljedova iz knjižnice ONT500. Pozitivne vrijednosti duljina indela označavaju insercije, dok negativne vrijednosti duljina označavaju delecije. Vrijednost duljine 0 označava nukleotidne supstitucije koje nisu prikazane. Brojevi indela normalizirani su s ukupnom duljinom analiziranih kvalitetnih i nekvalitetnih sljedova.

4.4. Ispravljanje kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT500

U Tablici 4. prikazani su statistički podaci o duljinama kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT500 prije i poslije ispravljanja programom Pilon. Vidljivo je da ispravljanjem programom Pilon nije došlo do značajnih promjena duljine sljedova. S druge strane, u ispravljenim sljedovima došlo je do značajnog smanjenja stope pogreške, koja nakon ispravljanja kod kvalitetnih sljedova iznosi 1,13%, a kod nekvalitetnih sljedova 2,33%. Također, analizirali smo relativne promjene učestalosti zamjena pojedinih baza u kvalitetnim i nekvalitetnim sljedovima, a rezultati su prikazani na Slici 16. Vidljivo je da je najmanja relativna promjena udjela parova krivo očitanih baza i njihovih točnih varijanti u nekvalitetnim sljedovima 69%, a najveća 99%. S druge strane, najmanja relativna promjena udjela parova krivo očitanih baza i njihovih točnih varijantu u kvalitetnim sljedovima iznosi 64%, a najveća 82%.

Tablica 4. Statistički podaci o duljinama kvalitetnin i nekvalitetnih sljedova iz knjižnice ONT500 prije i poslije ispravljanja programom Pilon

		Ukupna duljina sljedova /pb	Medijan duljina sljedova /pb	Standardna devijacija duljina sljedova	Maksimalna duljina slijeda / pb
Kvalitetni	Neispravljeni sljedovi	13 843 661	25 602	6 854	68 106
	Ispravljeni sljedovi	14 048 511	26 013	6 996	70 589
Nekvalitetni	Neispravljeni sljedovi	10 997 146	16 471	19 234	256 645
	Ispravljeni sljedovi	11 021 960	16 648	19 196	256 410

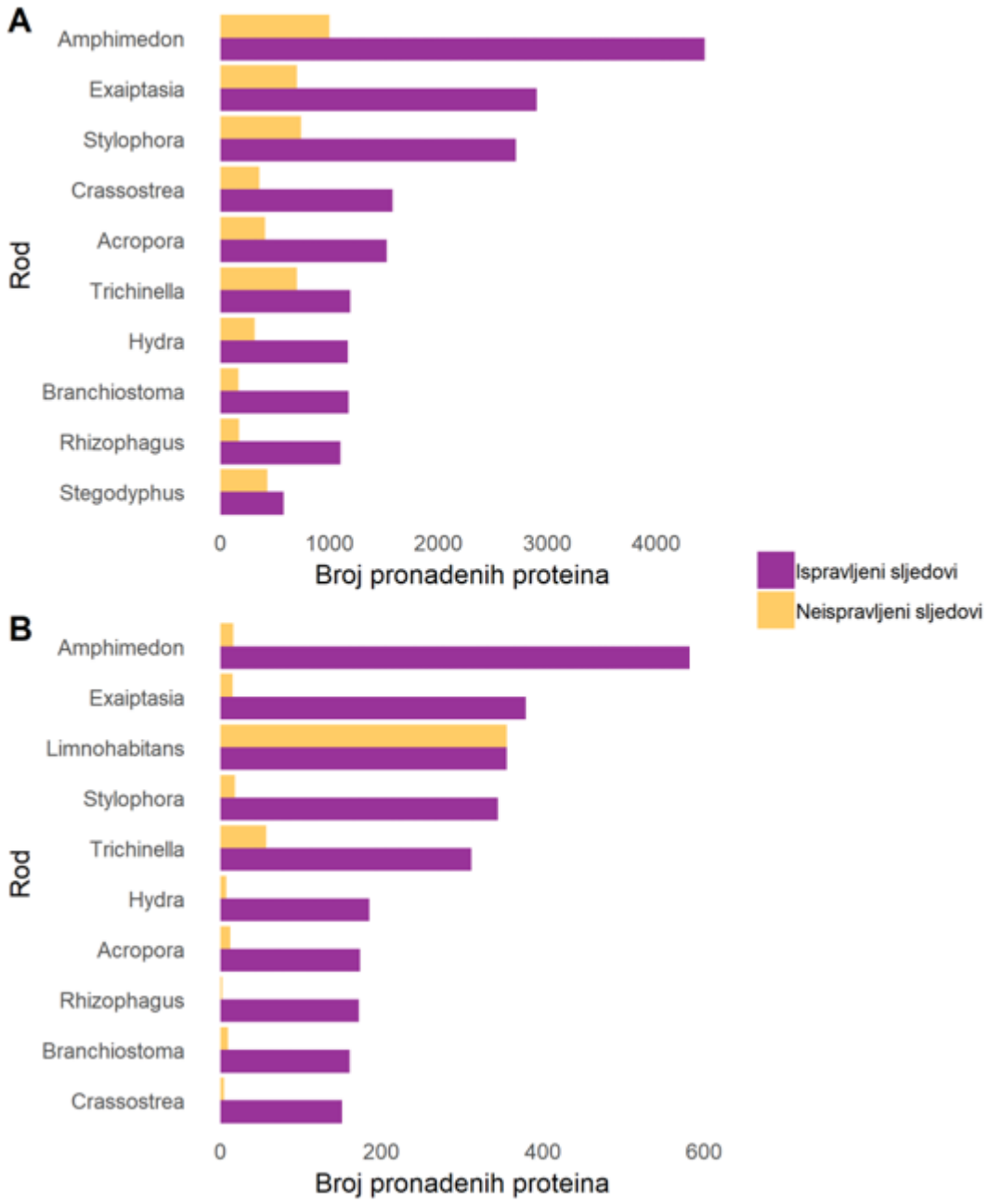


Slika 16. Prikaz relativne promjene udjela parova pogrešnih baza i njihovih točnih varijanti kod kvalitetnih (A) i nekvalitetnih (B) sljedova prije i poslije ispravljanja programom Pilon. Raspon boja prikazan je logaritamskom skalom.

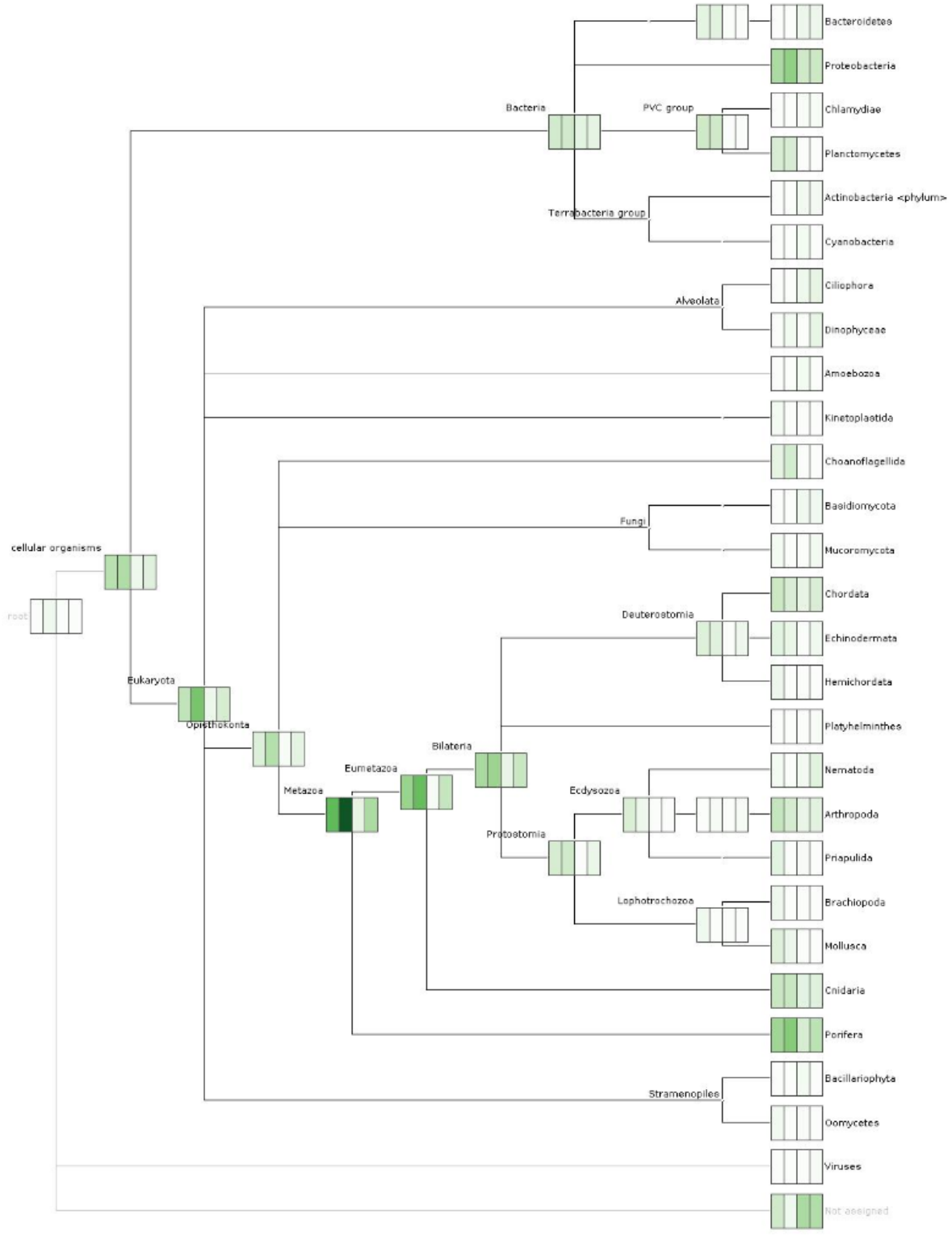
4.4.1. Taksonomska klasifikacija sljedova iz knjižnice ONT500

Analiza taksonomske klasifikacije pronađenih proteina u kvalitetnim i nekvalitetnim neispravljenim, odnosno ispravljenim sljedovima bio je drugi način provjere efikasnosti ispravljanja spomenutih sljedova. Programom DIAMOND napravili smo sravnjenja spomenutih skupina slijedova s neredundantnom bazom proteinskih sljedova. Ispravljanje kvalitetnih i nekvalitetnih sljedova povećalo je pronađeni broj proteina. Broj pronađenih proteina u kvalitetnim sljedovima prije ispravljanja iznosio je 33407, a nakon ispravljanja 98237. Isto tako, broj pronađenih proteina u nekvalitetnim sljedovima prije ispravljanja iznosio je 3532, a nakon ispravljanja 15837. Iz ovih podataka vidljiva je i prisutnost mnogo većeg broja pronađenih proteina u kvalitetnim sljedovima u odnosu na nekvalitetne sljedove. Na Slici 17 prikazano je deset najzastupljenijih rodova organizama čiji su proteini pronađeni u spomenutim sljedovima. Najzastupljeniji rod u kvalitetnim i u nekvalitetnim sljedovima nakon ispravljanja je rod *Amphimedon*. Većina pronađenih proteina u svim skupinama sljedova pripada eukariotskim organizmima, a jedini prokatiotski rod svrstan u deset najzastupljenijih rodova je rod *Limnohabitans*. Proteini koji pripadaju organizmima ovog roda pronađeni su u nekvalitetnim sljedovima, a broj takvih proteina nakon ispravljanja ostaje isti.

Sve skupine sljedova analizirali smo algoritmom LCA pomoću programa MEGAN. Ovim algoritmom svim skupinama sljedova pridružena je taksonomska kategorija do razine koljena. Postotak kvalitetnih slijedova kojima nije pridodana taksonomija iznosi 7,2% za neispravljene, odnosno 0,4% za ispravljene sljedove. S druge strane, broj nekvalitetnih sljedova kojima nije pridodana taksonomija puno je veći te iznosi 62,6% za neispravljene, odnosno 46% za ispravljene sljedove. Na Slici 18 prikazan je broj sljedova pridružen određenoj taksonomskoj kategoriji. Iako je većina sljedova svrstana u eukariotske organizme, relativno velik broj svih skupina sljedova svrstan je u prokariotsko koljeno Proteobacteria. Najveći broj sljedova pridružen je podcarstvu Metazoa i koljenu Porifera.



Slika 17. Prikaz broja pronađenih proteina svrstanih u deset najzastupljenijih rodova za kvalitetne (A) i nekvalitetne (B) sljedove iz knjižnice ONT500 prije i poslije ispravljanja. Proteini su pronađeni sravnjenjem s bazom neredundantnih proteinskih sljedova pomoću programa DIAMOND. Maksimalna e-vrijednost postavljena je na 10^{-5}.

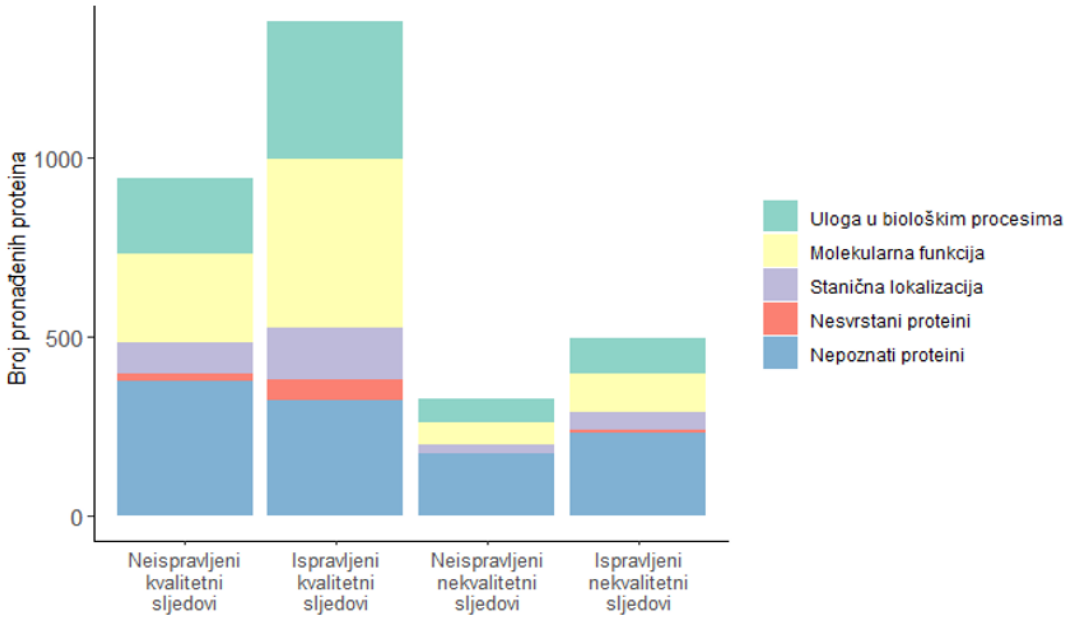


Slika 18. Kladogram taksonomije sljedova dobiven korištenjem algoritma LCA unutar programa MEGAN. Polja unutar pravokutnika označavaju sljedove iz knjižnice ONT500 sljedećim redoslijedom: neispravljeni kvalitetni, ispravljeni kvalitetni, neispravljeni nekvalitetni i ispravljeni kvalitetni sljedovi. Polja su prikazana toplinskom kartom gdje intenzivnije zeleno obojenje odgovara većem broju baza kod identificiranih proteina Raspon intenziteta boje je od 0 do 208 sljedova.

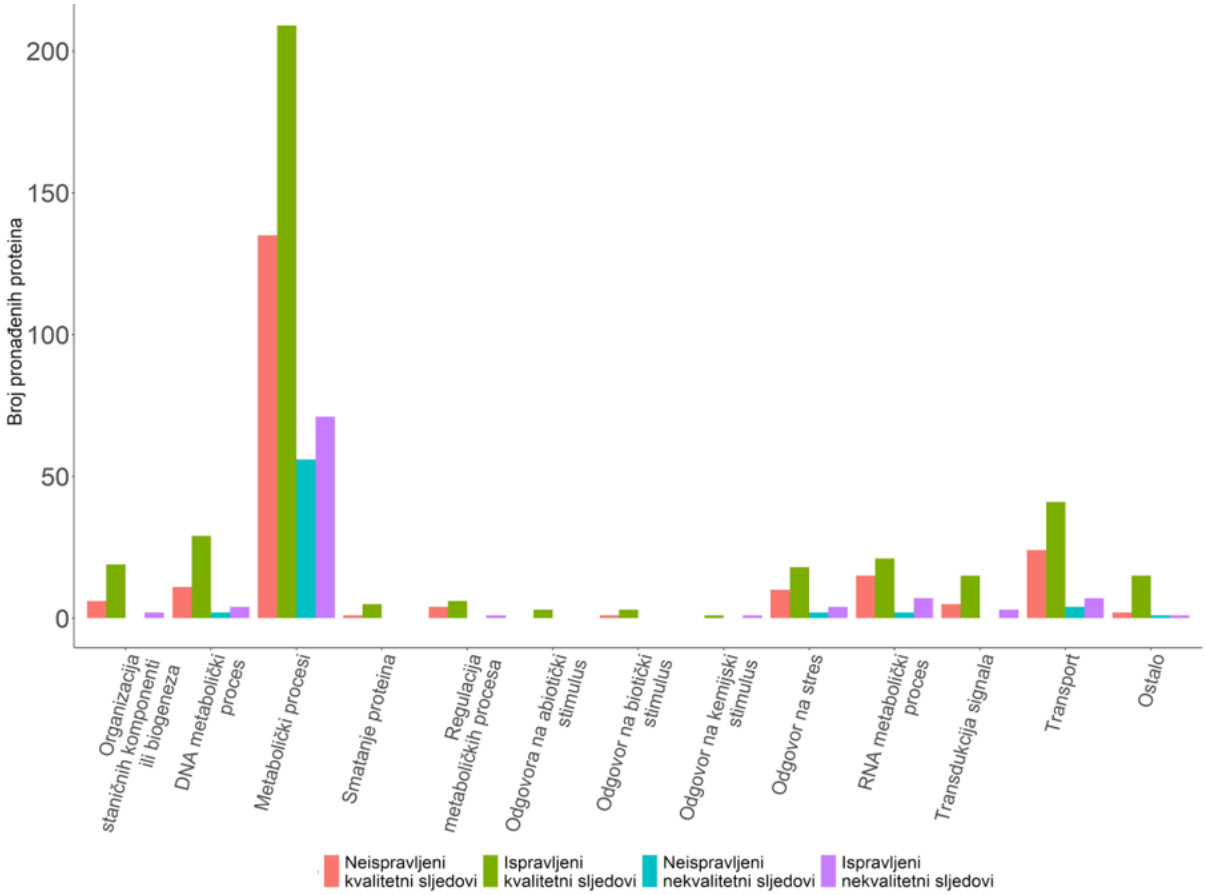
4.4.2. Funkcijska klasifikacija proteina pronađenih u svim kvalitetnim i nekvalitetnim sljedovima knjižnice ONT500

Nakon taksonomske klasifikacije kvalitetnih i nekvalitetnih neispravljenih, odnosno ispravljenih sljedova iz knjižnice ONT500 učinkovitost postupka ispravljanja provjerili smo i funkcijskom klasifikacijom proteina u navedenim sljedovima. Proteinima pronađenim programom DIAMOND odredili smo funkciju upotrebom klasifikacije InterPro2GO unutar programa MEGAN. Ispravljanjem kvalitetnih sljedova ukupan broj proteina kojima je ovom klasifikacijom dodijeljena funkcija povećao se s 944 na 1382. Primijetili smo porast proteina s dodijeljenom funkcijom kod nekvalitetnih sljedova, u kojima je prije ispravljanja funkcija određena 238 proteina, a nakon ispravljanja 469 proteina.

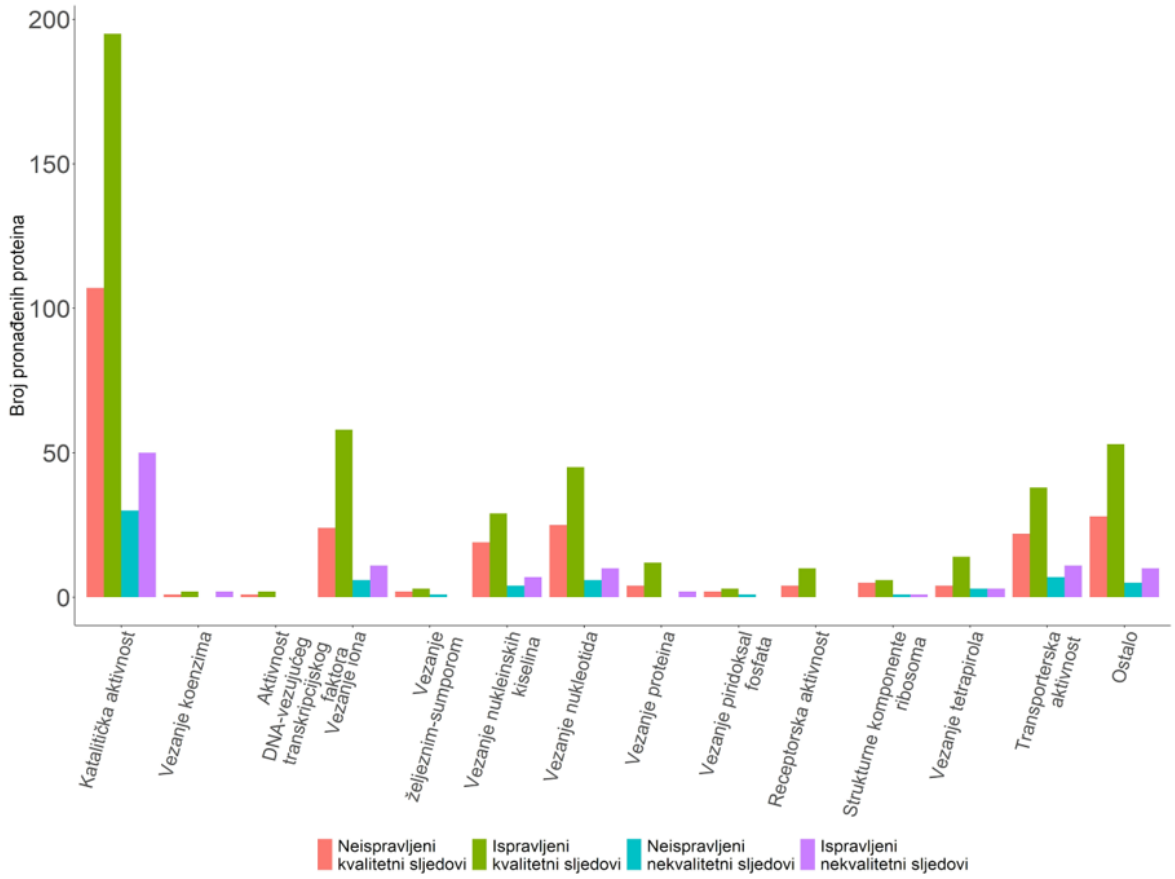
Ne računajući nepoznate proteine, proteinima pronađenima u svim skupinama sljedova najčešće je dodijeljena kategorija molekularne funkcije. Zatim slijede kategorija uloge u biološkim procesima te naposljetku kategorija lokalizacije proteina (Slika 19). Ispravljanjem kvalitetnih došlo je do smanjenja broja nepoznatih proteina, dok se ispravljanjem nekvalitetnih sljedova količina takvih proteina povećala. Najzastupljenije funkcije proteina pronađenih u svim skupinama sljedova unutar kategorije bioloških procesa su metabolički procesi (Slika 20). Najzastupljenije funkcije proteina pronađene u svim skupinama sljedova unutar kategorije molekularne funkcije su katalitička i druge vezne aktivnosti (Slika 21). Gledajući lokalizaciju proteina, proteini pronađeni u svim skupinama sljedova najčešće su dio stanične membrane (Slika 22). Općenito gledajući, količina pronađenih proteina svih funkcija veća je u kvalitetnim sljedovima te se ispravljanjem kvalitetnih i nekvalitetnih sljedova povećava.



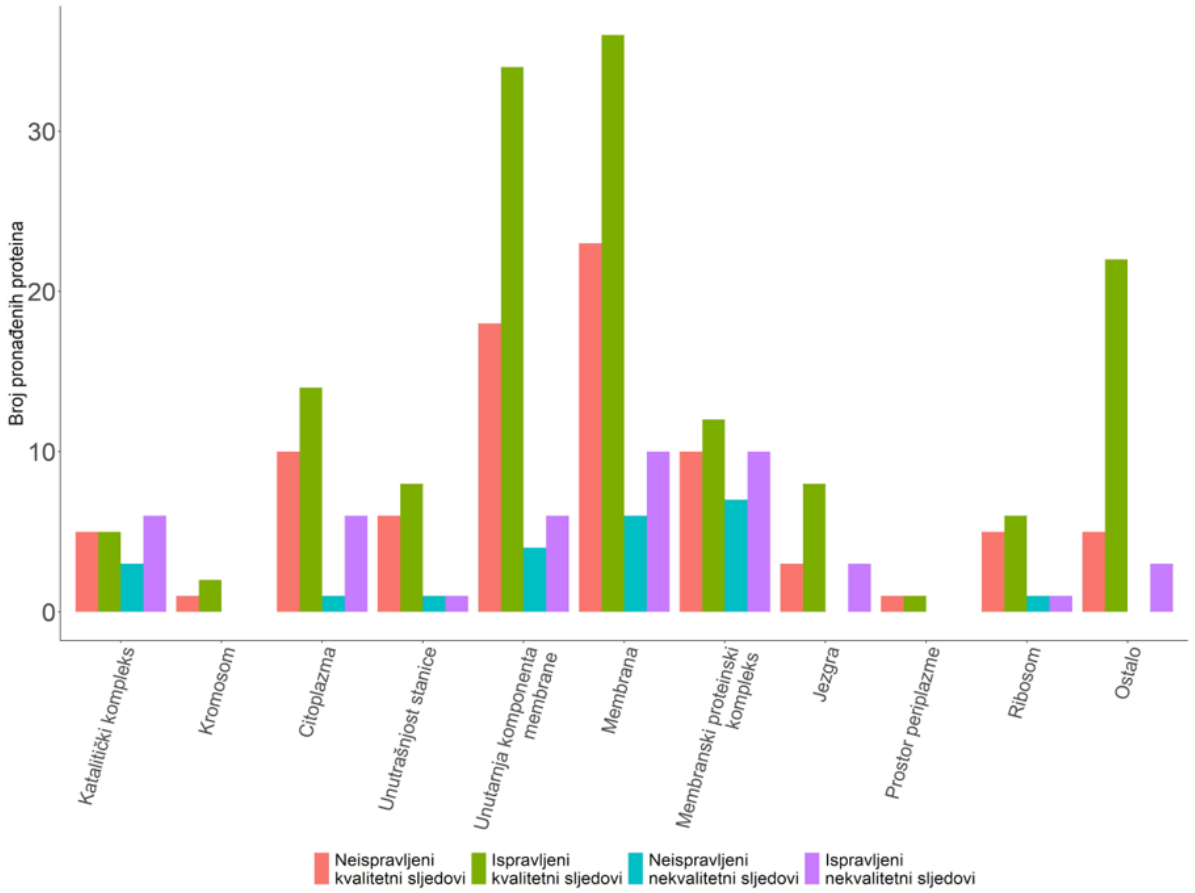
Slika 19. Prikaz zastupljenosti dodanih funkcijskih kategorija proteinima pronađenih u svim skupinama sljedova knjižnice ONT500. Funkcionalne kategorije određene su prema klasifikaciji InterPro2GO unutar programa MEGAN.



Slika 20. Prikaz broja proteina kojima je dodijeljena pojedina uloga u biološkim procesima prema klasifikaciji InterPro2GO unutar programa MEGAN. Analizirani su proteini pronađeni u svim skupinama sljedova iz knjižnice ONT500.



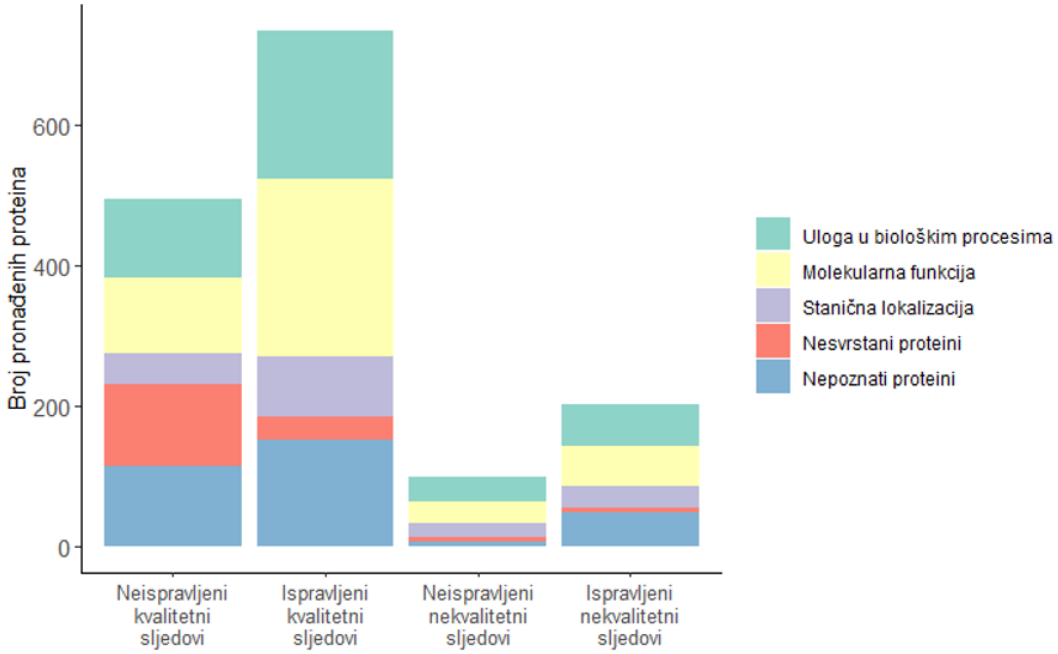
Slika 21. Prikaz broja proteina kojima je dodijeljena pojedina molekularna funkcija prema klasifikaciji InterPro2GO unutar programa MEGAN. Analizirani su proteini pronađeni u svim skupinama sljedova iz knjižnice ONT500.



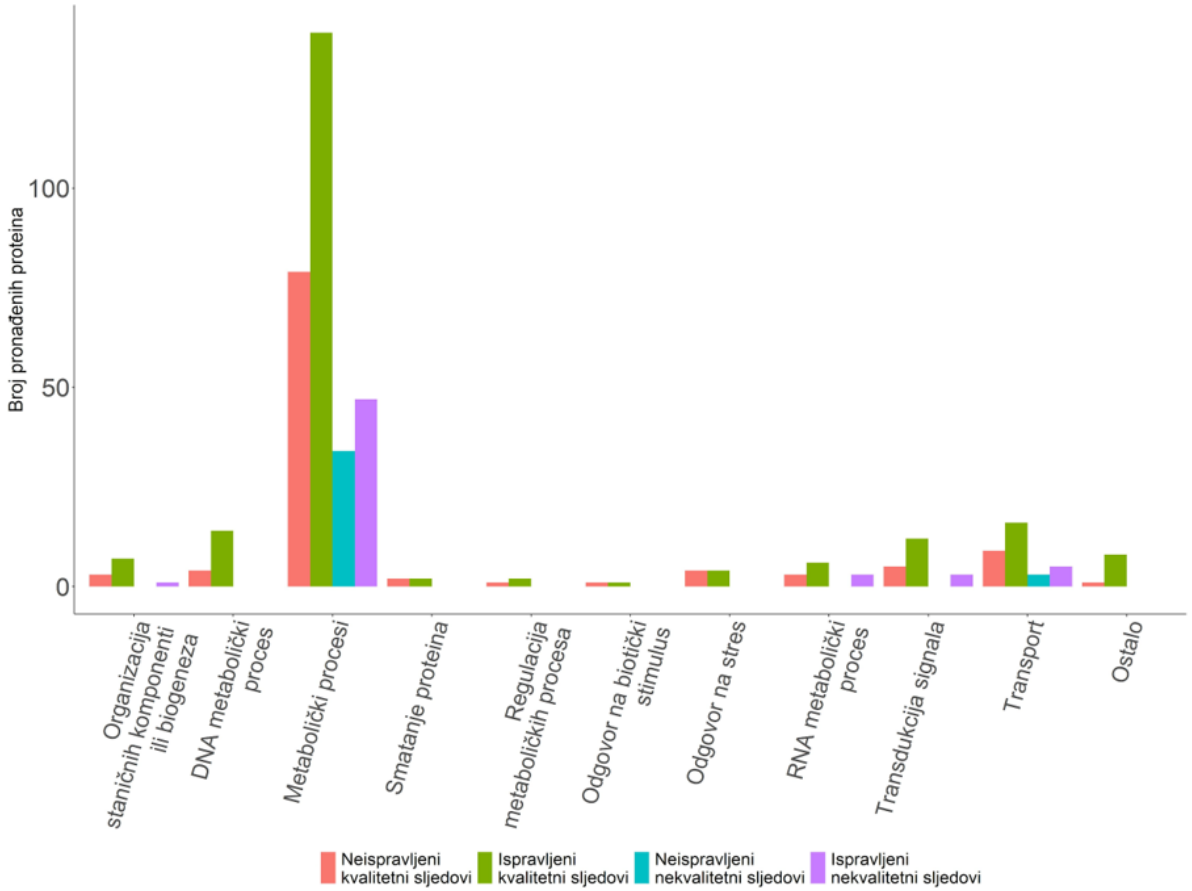
Slika 22. Prikaz broja proteina kojima je dodijeljena pojedina stanična lokalizacija prema klasifikaciji InterPro2GO unutar programa MEGAN. Analizirani su proteini pronađeni u svim skupinama sljedova iz knjižnice ONT500.

4.4.3. Funkcijska klasifikacija proteina kvalitetnih i nekvalitetnih sljedova knjižnice ONT500 kojima je dodijeljena taksonomska kategorija Metazoa ili Porifera

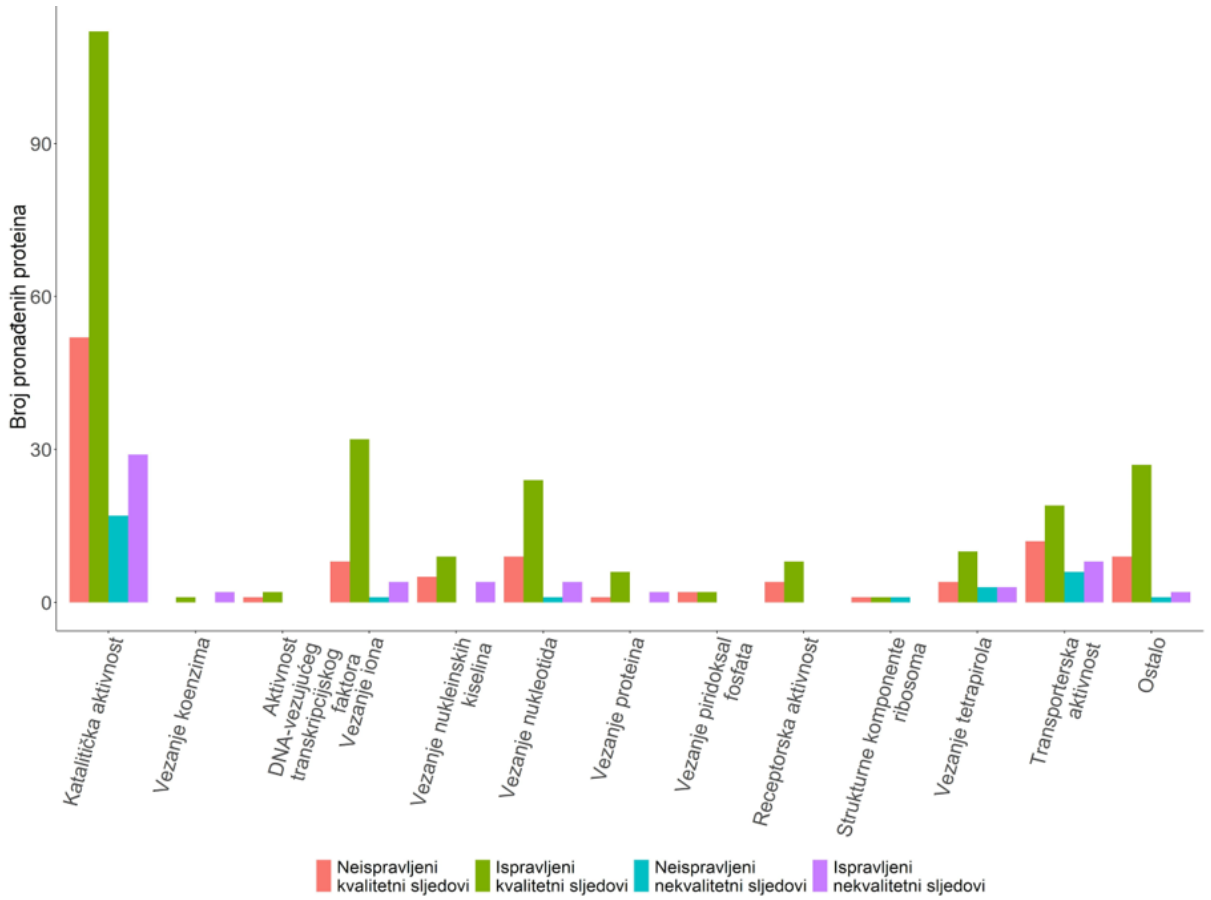
U sljedećem dijelu istraživanja napravili smo funkcijsku klasifikaciju InterPro2GO isključivo na sljedovima pridruženim taksonomskim kategorijama Metazoa i Porifera. Ispravljanjem kvalitetnih sljedova ukupan broj proteina u navedenim sljedovima kojima je ovom klasifikacijom dodijeljena funkcija povećao se s 493 na 734. Porast proteina s dodijeljenom funkcijom smo primijetili i kod nekvalitetnih sljedova skupine Metazoa i Porifera, u kojima je prije ispravljanja funkcija dodana 100 proteina, a nakon ispravljanja 202 proteina. Kao i u funkcijskoj analizi svih sljedova, proteinima pronađenima u svim skupinama sljedova pridruženim kategorijama Metazoa i Porifera najčešće je dodijeljena kategorija molekularne funkcije. Zatim, također slijede kategorija uloge u biološkim procesima te naposljetku kategorija lokalizacije proteina (Slika 23). Ispravljanjem kvalitetnih i nekvalitetnih sljedova broj pronađenih nepoznatih proteina se povećao. Najzastupljenije funkcije proteina pronađenih u svim skupinama spomenutih sljedova unutar kategorije bioloških procesa su metabolički procesi (Slika 24). Najzastupljenije funkcije proteina pronađene u svim skupinama spomenutih sljedova unutar kategorije molekularne funkcije su katalitička i druge vezne aktivnosti (Slika 25). Gledajući lokalizaciju proteina, proteini pronađeni u svim skupinama sljedova najčešće su dio stanične membrane (Slika 26). Kao i u analizi svih sljedova, količina pronađenih proteina svih funkcija u sljedovima kojima je pridružena taksonomska kategorija Metazoa ili Porifera veća je u kvalitetnim sljedovima te se ispravljanjem kvalitetnih i nekvalitetnih sljedova povećava.



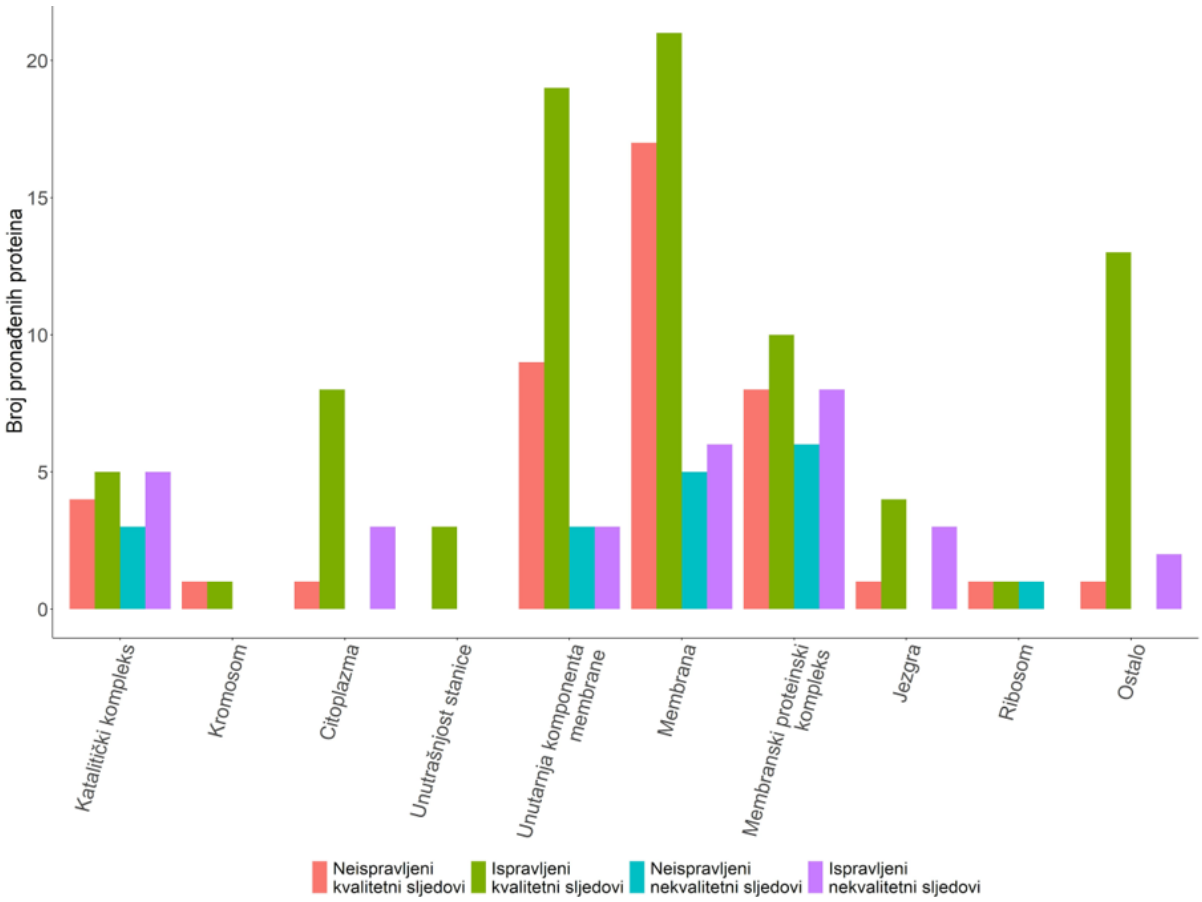
Slika 23. Prikaz zastupljenosti dodanih funkcijskih kategorija proteinima pronađenih kod neispravljenih i ispravljenih kvalitetnim, odnosno nekvalitetnim sljedovima knjižnice ONT500 pridruženih kategorijama Metazoa ili Porifera. Funkcijske kategorije dodane su prema klasifikaciji InterPro2GOunutar programa MEGAN.



Slika 24. Prikaz broja proteina kojima je dodijeljena pojedina uloga u biološkim procesima prema klasifikaciji InterPro2GO unutar programa MEGAN. Analizirani su proteini pronađeni u neispravljenim i ispravljenim kvalitetnim, odnosno nekvalitetnim sljedovima knjižnice ONT500 pridruženih kategorijama Metazoa ili Porifera.



Slika 25. Prikaz broja proteina kojima je dodijeljena pojedina molekularna funkcija prema klasifikaciji InterPro2GO unutar programa MEGAN. Analizirani su proteini pronađeni u neispravljenim i ispravljenim kvalitetnim, odnosno nekvalitetnim sljedovima knjižnice ONT500 pridruženih kategorijama Metazoa ili Porifera.



Slika 26. Prikaz broja proteina kojima je dodijeljena pojedina stanična lokalizacija prema klasifikaciji InterPro2GO unutar programa MEGAN. Analizirani su proteini pronađeni u neispravljenim i ispravljenim kvalitetnim, odnosno nekvalitetnim sljedovima knjižnice ONT500 pridruženih kategorijama Metazoa ili Porifera.

5. RASPRAVA

Istraživanje smo započeli analizom sljedova dobivenih tehnologijom sekvenciranja nanoporama (knjižnica ONT), pri čemu smo promatrali sljedove koji su prošli test kvalitete (kvalitetni sljedovi) i sljedove koji su pali test kvalitete (nekvalitetni sljedovi). Sljedovi su odvojeni programom Albacore na temelju prosječne ocjene kvalitete 7. Vrijednost ove granice je proizvoljna s relativno velikim rasponom korištenja od strane mnogih autora (Tyler i sur. 2018). Iz toga proizlazi da bi nekvalitetni sljedovi relativno velike duljine čija je prosječna mjera kvalitete blizu 7 mogli također biti iskorišteni za sastavljanje genoma ogulinske špiljske spužvice. Duljina slijeda ponajprije ovisi o korištenoj kemiji sekvenciranja te o prethodnoj obradi uzoraka, koja obično uključuje fragmentaciju DNA na željenu duljinu. Unatoč tome što kvalitetni i nekvalitetni sljedovi nisu odvojeni na temelju duljine, razlika raspodjela duljina ovih dvaju skupina je statistički značajna. U nekvalitetnim sljedovima primjećujemo veći raspon duljina nego u kvalitetnim sljedovima. Naime, velik udio nekvalitetnih sljedova zanemarivo male duljine vjerojatno je posljedica neuspješnog sekvenciranja takvih sljedova, ali i nepravilnog fragmentiranja u pripremi uzorka. Udio kvalitetnih sljedova kraćih duljina puno je manji, a postojanja takvih sljedova također se može objasniti nepravilnim fragmentiranjem uzorka tijekom njegove pripreme. Određen udio sljedova veće duljine vjerojatno je svrstan u nekvalitetne sljedove zbog stvaranja kimera koje značajno smanjuju prosječnu ocjenu kvalitete slijeda, što je karakteristično za sljedove dobivene 2D sekvenciranjem (White i sur. 2017). S druge strane, već spomenuti izrazito dugi sljedovi sadrže više pogrešaka od kraćih sljedova, vjerojatno zbog nepreciznosti tehnologije sekvenciranja samog uređaja, poput motornog proteina (Pontefract i sur. 2018). Budući da dugi sljedovi sadrže više informacija koje bi mogle koristiti u sastavljanju genoma ogulinske špiljske spužvice, za daljnju analizu odabrali smo 500 najduljih kvalitetnih i nekvalitetnih sljedova iz knjižnice ONT (knjižnica ONT500).

Sravnjenje ispravnijih sljedova dobivenih tehnologijom sekvenciranja Illumina (knjižnica Illumina) na referentne sljedove knjižnice ONT500 napravili smo trima skupinama parametara odabranih na temelju prethodnog iskustva. Nedopuštanjem sekundarnih sravnjenja i sprječavanjem opcije „soft clipping“ kod svih triju skupina parametara osigurali smo sravnjenje pojedinog slijeda iz knjižnice Illumina na najsličniji dio referentnih sljedova. Prilikom sravnjenja sljedova iz knjižnice Illumina na nekvalitetne referentne sljedove, manje vrijednosti prosječne kvalitete sravnjenja i pokrivenosti su očekivane. Naime, upravo zbog maje prosječne ocjene kvalitete takvih sljedova, u njima očekujemo više pogrešno očitanih nukleotida, a time i veće razlike u usporedbi s točnijom tehnologijom poput tehnologije Illumina. U svim provedenim sravnjenjima puno veći udio sljedova iz knjižnice Illumina uspješno se sravnio s kvalitetnim sljedovima knjižnice ONT500. Vrijednosti prosječne kvalitete sravnjenja sljedova knjižnice Illumina na sljedove knjižnice ONT500, kao i vrijednost pokrivenosti sljedova iz knjižnice ONT iz tog su razloga puno pouzdanije kod sravnjenja na kvalitetnim sljedovima.

Spomenute vrijednosti najveće su u sravnjenjima dobivenim korištenjem skupine parametara minimap2 zbog čega je spomenuto sravnjenje odabrano za korištenje u analizi pogrešaka i ispravljanju sljedova iz knjižnice ONT500.

Važno je spomenuti i udio broja gvanina i citozina (udio GC) sravnjenih sljedova iz knjižnice ONT500, čija vrijednost može upućivati na kontaminacije drugim organizmima. Jedini poznati podatak o udjelu GC u genomu spužvi je udio GC kod spužve *Amphimedon queenslandica*, koji iznosi 37,5% (NCBI Genomes 2019). Veći udio GC u sravnjenim sljedovima može ukazivati na biološku kontaminaciju u uzorcima sekvencirane DNA ogulinske špiljske spužvice. Najveći udio takvih kontaminacija vjerojatno čine prokarioti, odnosno prokariotski endosimbionti budući da je njihova zastupljenost u tijelu spužve najčešća. S druge strane, pokazano je da veća srodnost organizama ne ukazuje na sličan udio GC njihovih genoma (Li i sur. 2014). Iz tog razloga, moguća je i značajna razlika između udjela GC spužve *Amphimedon queenslandica* i ogulinske špiljske spužvice.

Filtriranjem varijanti sljedova knjižnice ONT500 izdvojili smo pogreške visoke pouzdanosti. Naime, pogreške pronađene pomoću sljedova s malom kvalitetom sravnjivanja, kao i pogreške pronađene na mjestima male pokrivenosti nisu pouzdane te mogu predstavljati lažno pozitivne rezultate. Granične vrijednosti pokrivenosti i kvalitete mapiranja sljedova knjižnice Illumina određene su na temelju prethodnog iskustva, ali i istraživanja drugih autora (poput Münz i sur. 2018). Smatramo da su odabrane granične vrijednosti omogućile minimalan broj lažno pozitivnih i lažno negativnih rezultata. Također, analizirali smo samo one pogreške za koje je u sravnjenju s knjižnicom Illumina bilo moguće jednoznačno utvrditi točni nukleotid (na sravnjenoj poziciji unutar 97% sljedova iz knjižnice Illumina pojavljuje se ista baza). Time smo iz analize izbacili varijante spužve uzrokovane njezinom heterozigotnošću dok smo istovremeno dozvolili i pogreške u sljedovima knjižnice Illumina koje su razmjerne dokumentiranoj stopi pogreške ove tehnologije.

Izračunata stopa pogrešaka veća je u nekvalitetnim sljedovima knjižnice ONT500 budući da ovi sljedovi imaju manju prosječnu ocjenu kvalitete. Analizom glavnih komponenti pokazali smo malu razliku u udjelu vrsti pogrešaka ovih dvaju skupina sljedova pri čemu su delecije češće kod kvalitetnih sljedova, a transverzije kod nekvalitetnih sljedova. Iz rezultata analize vidimo da je razlika u pogreškama ovih dvaju skupina sljedova relativno mala što predstavlja još jedan argument u korist mogućeg iskorištenja najtočnijih sljedova skupine nekvalitetnih sljedova u sastavljanju genoma ogulinske špiljske spužvice. Ipak, moguće je da su opažene razlike manje zbog relativno malog postotka sravnjenih sljedova iz knjižnice Illumina na nekvalitetne sljedove iz knjižnice ONT500.

Analizom pogrešaka također smo utvrdili veću učestalost pogrešno očitanih baza i indela u nekvalitetnim sljedovima knjižnice ONT500. Udjeli parova pogrešno očitanih baza i

njihovih točnih varijanti u obje skupine sljedova veći su za kemijski slične baze što upućuje na teže razlikovanje sličnijih baza pri prevođenju (*basecalling*) sljedova. Također, učestalost indela slična je učestalostima pogrešno pročitanih baza u obje skupine sljedova. Jedan od najčešćih razloga relativno velike učestalosti indela pri sekvenciranju tehnologijom nanopora je prisutnost homopolimera. Naime, sekvenciranje slijeda istih nukleotida ne dolazi do promjene električnog signala što često može uzrokovati preskakanje ili dodavanje baze čitanom slijedu (O'Donnell i sur. 2013). Osim učestalosti, analizirali smo i duljine indela pri čemu smo ustanovili statistički značajnu razliku duljine indela u kvalitetnim i nekvalitetnim sljedovima. Ipak, ova razlika mogla bi biti posljedica ekstremnih vrijednosti duljina indela, koje u nekvalitetnim sljedovima vjerojatno nisu pronađene zbog relativno malog postotka mapiranja sljedova knjižnice Illumina na takve sljedove.

Uspješnost ispravljanja programom Pilon provjerili smo usporedbom pogrešaka u ispravljenim i neispravljenim kvalitetnim, odnosno nekvalitetnim sljedovima. Usporedbom neispravljenih i ispravljenih sljedova nismo dobili statistički značajnu razliku duljina navedenih sljedova. Međutim, nakon ispravljanja sljedova vidljivo je značajno smanjenje stope pogrešaka u kvalitetnim i nekvalitetnim sljedovima. Isto tako, usporedbom udjela parova pogrešno očitanih baza i njihovih točnih varijanti u neispravljenim i ispravljenim sljedovima utvrdili smo smanjeni broj pogrešno očitanih baza u ispravljenim sljedovima. Preciznije, u obje skupine sljedova ispravljeno je više od 60 % svake krivo očitane baze. Navedenim rezultatima dokazali smo veću kvalitetu ispravljenih sljedova. Također, zanimljivo je vidjeti da je stopa pogreške ispravljenih kvalitetnih sljedova izuzetno niska zbog čega smatramo da korištenje takvih sljedova ima velik potencijal u sklapanju genoma ogulinske špiljske spužvice.

Kao provjeru uspješnosti postupka ispravljanja, odnosno povećanja kvalitete kvalitetnih i nekvalitetnih sljedova napravili smo taksonomsku klasifikaciju proteinskih sljedova pronađenih u svim skupinama sljedova. Kvalitetni sljedovi su u usporedbi s nekvalitetnim sljedovima sadržavali veći broj kodirajućih regija prije i poslije ispravljanja. Pokazali smo da ispravljanje sljedova svih skupina znatno povećava broj pronađenih proteinskih sljedova koji pripadaju rodu *Amphimedon*. Promjena broja spomenutih proteina najizraženija je u skupini nekvalitetnih sljedova što dokazuje veliko povećanje kvalitete takvih sljedova. Naime, proteini svrstani u rod *Amphimedon* taksonomski su najbliže ogulinskoj špiljskoj spužvici budući da njezin genom ne postoji u bazama bioloških podataka. No, to nije jedini dokaz poboljšanja kvalitete sljedova. Analizom sljedova koristeći algoritam LCA utvrdili smo da najveći broj sljedova svih skupina prije i poslije ispravljanja ima pridruženu taksonomsku kategoriju Metazoa, odnosno Porifera. Ti sljedovi najvjerojatnije čine dio genoma ogulinske špiljske spužvice. Međutim, relativno velik broj nekvalitetnih sljedova svrstan je u prokariotsko koljeno Protobacteria. Ovo opažanje možemo povezati s prvim dijelom taksonomske analize u kojoj je u nekvalitetnim sljedovima pronađen velik broj proteina koji pripada rodu *Limnohabitans*.

Rod *Limnohabitans* pripada koljenu Proteobacteria, a čine ga 4 vrste čija je jedna od glavnih karakteristika stanište slatkovodnih voda. Različiti sojevi spomenutih vrsta usko su povezani s velikim brojem nekultiviranih bakterija u brojnim slatkovodnim sustavima (Hahn i sur. 2010). Također, važno je naglasiti da je ovaj rod dokazani simbiont mnogih vrsta slatkovodnih spužvi (Jung i sur. 2014). Iz primjera spomenutog roda vidljiva je mogućnost primjene programa MEGAN u uklanjanju bioloških kontaminanata koje je teže ukloniti eksperimentalnim putem.

Osim taksonomske klasifikacije, pomoću podataka iz rječnika Gene Ontology odredili smo i funkcijsku ulogu pronađenih proteina svih kvalitetnih i nekvalitetnih sljedova prije i nakon njihova ispravljanja. Kategorija nepoznatih proteina označava proteine kojima je dodijeljena nepoznata funkcija u klasifikaciji InterPro2GO, dok u kategoriju nesvrstanih proteina spadaju proteini kojima nije dodijeljen pojam prema rječniku Gene Ontology. Ispravljanjem svih kvalitetnih i nekvalitetnih sljedova pronađeno je više proteina s određenom pridruženom funkcijom što predstavlja još jednu potvrdu uspješnog ispravljanja svih skupina sljedova. Za analizu funkcijske uloge sljedova, tj. njihovih proteina, koji pripadaju ogulinskoj špiljskoj spužvici odabrali smo samo sljedove taksonomskih kategorija Metazoa, odnosno Porifera. U ovom dijelu analize uočili smo povećan broj nepoznatih proteina pronađen nakon ispravljanja obiju skupina sljedova. Naime, jedan od glavnih problema prijašnjih istraživanja porijekla, evolucije i funkcije gene spužvi i ostalih bazalnih životinja je pristranost uzrokovana spoznajama o skupini Bilateria (Dunn i sur. 2015). Analizom genoma vrste *Amphimedon queenslandica* potvrđeni su brojni geni odgovorni za biološke procese za koje se smatralo da u spužvama nedostaju, poput imunosnog ili senzornog sustava (Srivastava i sur. 2010), a slični rezultati dobiveni su i analizom transkriptoma osam vrsta spužvi s predstavnicima u sva četiri razreda (Guzman i Conaco 2016). Pronalaskom većeg broja nepoznatih proteina najvjerojatnije smo pronašli gene koji još uvijek nisu karakterizirani u skupini spužvi pri čemu ti geni vjerojatno spadaju u njihovu skrivenu biologiju (engl. *hidden biology*). Skrivena biologija uključuje nove strukture i funkcije na staničnoj razini uzrokovane nepoznatim mehanizmima koji se ne nalaze u vrstama skupine Bilateralia. Sve ovo upućuje na veliku kompleksnost genoma ogulinske špiljske spužvice, kao i na veliku važnost njezina istraživanja. Ispravljanjem sljedova dobivenih tehnologijom sekvenciranja nanoporama došli smo korak bliže korištenju takvih sljedova u sastavljanju genoma ogulinske špiljske spužvice, koji bi u budućnosti mogao odgovoriti na brojna pitanja i rasvijetliti mnoge nove biološke spoznaje.

6. ZAKLJUČCI

Računalnom analizom sljedova ogulinske špiljske spužvice (*Eunapius subterraneus* Sket & Velikona, 1984) prikupljenih tehnologijom sekvenciranja nanoporama došli smo do sljedećih zaključaka:

1. Sljedovi koji su svrstani u skupinu kvalitetnih sljedova imaju veći medijan duljine i medijan mjere kvalitete od sljedova koji su svrstani u skupinu nekvalitetnih sljedova.
2. Optimirali smo parametre sravnjenja sljedova iz knjižnice Illumina na odabrani uzorak sljedova sekvenciranih tehnologijom nanoporama.
3. Analizom rezultata sravnjenja odredili smo da je stopa pogreške u kvalitetnim sljedovima 8,61%, a u nekvalitetnim sljedovima 10,55%.
4. Razlika broja i udjela pogrešaka kvalitetnih i nekvalitetnih sljedova je relativno mala što upućuje na mogućnost iskorištenja nekvalitetnih sljedova u sastavljanju genoma ogulinske špiljske spužvice.
5. Poboljšali smo kvalitetu obje skupine sljedova ispravljanjem pronađenih pogrešaka, što je dovelo do:
    - smanjenja stope pogrešaka u kvalitetnim sljedovima na 1,23%, odnosno smanjenja stope pogreške u nekvalitetnim sljedovima na 2,33%
    - ispravka više od 60% svake pogrešno očitane baze u objema skupinama sljedova
    - povećanja broja pronađenih proteina u ispravljenim kvalitetnim i nekvalitetnim sljedovima knjižnice ONT
    - povećanja broja kvalitetnih i nekvalitetnih sljedova svrstanih u taksonomske kategorije Metazoa i Porifera upotrebom algoritma LCA
    - povećanje broja funkcija pronađenih proteina u ispravljenim kvalitetnim i nekvalitetnim sljedovima prema rječniku Gene Ontology
    - povećanje broja nepoznatih proteina u kvalitetnim i nekvalitetnim sljedovima taksonomske kategorije Metazoa, odnosno Porifera

7. ZAHVALE

Zahvaljujemo mentoru prof. dr. sc. Kristianu Vlahovičeku na stručnom vodstvu, strpljenju, pruženom znanju te svom uloženom trudu i vremenu.

Zahvaljujemo asistenticama Maji Kuzman i Dunji Glavaš na svim pruženim savjetima, smjernicama i ugodnom vremenu provedenom u laboratoriju.

Posebno hvala našim obiteljima i prijateljima na pruženoj podršci.

8. POPIS LITERATURE

Bedek J., Bilandžija H., Jalžić B. (2008): Ogulinska špiljska spužvica Eunapius subterraneus Sket et Velikonja, 1984, rasprostranjenost i ekologija vrste i staništa. Modruški zbornik 2: 103-130.

Bolger A.M., Lohse M., Usadel B. (2014): Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114–2120.

Buchfink B., Xie C., Huson D.H. (2014): Fast and sensitive protein alignment using DIAMOND. Nat. Methods **12**: 59–60.

Bushnell B. (2015): BBMap - sourceforge.net/projects/bbmap/

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R. (2011): The variant call format and VCFtools. Bioinformatics **27**: 2156–2158.

Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sørensen M. V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. (2008): Broad phylogenomic sampling improves resolution of the animal tree of life. Nature **452**: 745–749.

Dunn C.W., Leys S.P., Haddock S.H.D. (2015): The hidden biology of sponges and ctenophores. Trends Ecol. Evol. **30**: 282–291.

Glavaš D. (2018): Sekvenciranje tehnologijom nanopora i sklapanje genoma ogulinske špiljske spužvice Eunapius subterraneus. Diplomski rad. Zagreb: Prirodoslovno–matematički fakultet

Guzman C., Conaco C. (2016): Comparative transcriptome analysis reveals insights into the streamlined genomes of haplosclerid demosponges. Sci. Rep. **6**: 1–10.

Hahn M.W., Kasalický V., Jezbera J., Brandt U., Jezberová J., Šimek K. (2010): Limnohabitans curvus gen. nov., sp. nov., a planktonic bacterium isolated from a freshwater lake. Int. J. Syst. Evol. Microbiol. **60**: 1358–1365.

Harcet M., Bilandžija H., Bruvo-Madarić B., Ćetković H. (2010): Taxonomic position of Eunapius subterraneus (Porifera, Spongillidae) inferred from molecular data - A revised classification needed? Mol. Phylogenet. Evol. **54**: 1021–1027.

Huson D.H., Beier S., Flade I., Górska A., El-Hadidi M., Mitra S., Ruscheweyh H.J., Tappu R. (2016): MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput. Biol. **12**: 1–12.

Jung D., Seo E.Y., Epstein S.S., Joung Y., Han J., Parfenova V. V., Belykh O.I., Gladkikh A.S., Ahn T.S. (2014): Application of a new cultivation technology, I-tip, for studying microbial diversity in freshwater sponges of Lake Baikal, Russia. FEMS Microbiol. Ecol. **90**: 417–423.

Kchouk M., Gibrat J.F., Elloumi M. (2017): Generations of Sequencing Technologies: From First to Next Generation. Biol. Med. **09**

de Lannoy C., Ridder D. de, Risse J. (2017): A sequencer coming of age: De novo genome assembly using MinION reads. F1000Research **6**: 1083.

Li H. (2017): Minimap2: pairwise alignment for nucleotide sequences. **34**: 3094–3100.

Li H., Durbin R. (2010): Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26**: 589–595.

Li X.Q., Du D. (2014): Variation, evolution, and correlation analysis of C+G content and genome or chromosome size in different kingdoms and phyla. PLoS One **9**: 1-18.

Matoničkin I. (1990): Beskralješnjaci: biologija nižih avertebrata. Školska knjiga: Zagreb

Mulder N.J., Apweiler R., Attwood T.K., Bairoch A., Barrell D., Bateman A., Binns D., Biswas M., Bradley P., Bork P., Bucher P., Copley R.R., Courcelle E., Das U., Durbin R., Falquet L., Fleischmann W., Griffiths-Jones S., Haft D., Harte N., Hulo N., Kahn D., Kanapin A., Krestyaninova M., Lopez R., Letunic I., Lonsdale D., Silventoinen V., Orchard S.E., Pagni M., Peyruc D., Ponting C.P., Selengut J.D., Servant F., Sigrist C.J.A., Vaughan R., Zdobnov E.M. (2003): The InterPro database, 2003 brings increased coverage and new features. Nucleic Acids Res. **31**: 315–318.

Münz M., Mahamdallie S., Yost S., Rimmer A., Poyastro-Pearson E., Strydom A., Seal S., Ruark E., Rahman N. (2018): CoverView: a sequence quality evaluation tool for next generation sequencing data. Wellcome Open Res. **3**

NCBI Genomes (2019): Amphimedon queenslandica. Dostupno na https://www.ncbi.nlm.nih.gov/genome/2698. Pristupljeno 26.4.2019.

O'Donnell C.R., Wang H., Dunbar W.B. (2013): Error analysis of idealized nanopore sequencing. Electrophoresis **34**: 2137–2144.

Passera A., Marcolungo L., Casati P., Brasca M., Quaglino F., Cantaloni C., Delledonne M. (2018): Hybrid genome assembly and annotation of Paenibacillus pasadenensis strain R16 reveals insights on endophytic life style and antifungal activity. PLoS One **13**: .

Philippe H., Derelle R., Lopez P., Pick K., Borchiellini C., Boury-Esnault N., Vacelet J., Renard E., Houliston E., Quéinnec E., Silva C. Da, Wincker P., Guyader H. Le, Leys S., Jackson

D.J., Schreiber F., Erpenbeck D., Morgenstern B., Wörheide G., Manuel M. (2009): Phylogenomics Revives Traditional Views on Deep Animal Relationships. Curr. Biol. **19**: 706–712.

Pontefract A., Hachey J., Zuber M.T., Ruvkun G., Carr C.E. (2018): Sequencing nothing: Exploring failure modes of nanopore sensing and implications for life detection. Life Sci. Sp. Res. **18**: 80–86.

Pop M. (2009): Genome assembly reborn: Recent computational challenges. Brief. Bioinform. **10**: 354–366.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Roberts M., Hayes W., Hunt B.R., Mount S.M., Yorke J.A. (2004): Reducing storage requirements for biological sequence comparison. Bioinformatics **20**: 3363–3369.

Simion P., Philippe H., Baurain D., Jager M., Richter D.J., Franco A. Di, Roure B., Satoh N., Quéinnec É., Ereskovsky A., Lapébie P., Corre E., Delsuc F., King N., Wörheide G., Manuel M. (2017): A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. Curr. Biol. **27**: 958–967.

van Soest R.W.M., Boury-Esnault N., Vacelet J., Dohrmann M., Erpenbeck D., Voogd N.J. de, Santodomingo N., Vanhoorne B., Kelly M., Hooper J.N.A. (2012): Global diversity of sponges (Porifera). PLoS One **7**: 1-23.

Sperling E.A., Pisani D., Peterson K.J. (2007): Poriferan paraphyly and its implications for Precambrian palaeobiology. Geol. Soc. London, Spec. Publ. **286**: 355–368.

Srivastava M., Simakov O., Chapman J., Fahey B., Gauthier M.E.A., Mitros T., Richards G.S., Conaco C., Dacre M., Hellsten U., Larroux C., Putnam N.H., Stanke M., Adamska M., Darling A., Degnan S.M., Oakley T.H., Plachetzki D.C., Zhai Y., Adamski M., Calcino A., Cummins S.F., Goodstein D.M., Harris C., Jackson D.J., Leys S.P., Shu S., Woodcroft B.J., Vervoort M., Kosik K.S., Manning G., Degnan B.M., Rokhsar D.S. (2010): The Amphimedon queenslandica genome and the evolution of animal complexity. Nature **466**: 720–726.

Stranneheim H., Lundeberg J. (2012): Stepping stones in DNA sequencing. Biotechnol. J. **7**: 1063–1073.

Tarasov A., Vilella A.J., Cuppen E., Nijman I.J., Prins P. (2017): Genome analysis Sambamba : fast processing of NGS alignment formats. Bioinformatics **31**: 2032–2034.

Taylor M.W., Radax R., Steger D., Wagner M. (2007): Sponge-Associated Microorganisms:

Evolution, Ecology, and Biotechnological Potential. Microbiol. Mol. Biol. Rev. **71**: 295–347.

Tyler A.D., Mataseje L., Urfano C.J., Schmidt L., Antonation K.S., Mulvey M.R., Corbett C.R. (2018): Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. Sci. Rep. **8**: 1–12.

Walker B.J., Abeel T., Shea T., Priest M., Abouelliel A., Sakthikumar S., Cuomo C.A., Zeng Q., Wortman J., Young S.K., Earl A.M. (2014): Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One **9**

White R., Pellefigues C., Ronchese F., Lamiable O., Eccles D. (2017): Investigation of chimeric reads using the MinION. F1000Research **6**: 631

World Porifera Database (2019): Species - Eunapius subterraneus Sket & Velikonja, 1984. Dostupno http://www.marinespecies.org/porifera/porifera.php?p=taxdetails&id=167168. Pristupljeno 8.4.2018.

Wysoker A., Fennell T., Marth G., Abecasis G., Ruan J., Li H., Durbin R., Homer N., Handsaker B. (2009): The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079.

9. SAŽETAK

RAČUNALNA ANALIZA SLJEDOVA OGULINSKE ŠPILJSKE SPUŽVICE (*EUNAPIUS SUBTERRANEUS* SKET & VELIKONJA, 1984) PRIKUPLJENIH TEHNOLOGIJOM SEKVENCIRANJA NANOPORAMA

Kristian Bodulić i Paula Štancl

Metode sekvenciranja sljedeće generacije stalno napreduju prema bržem i jeftinijem dobivanju dugih sljedova sa što manjom stopom pogreške. Primjena dugih sljedova dobivenih tehnologijom sekvenciranja nanoporama (Oxford Nanopore Technologies, ONT) ima velik potencijal u mnogim genomskim i biomedicinskim istraživanjima, a posebno u sastavljanju genoma *de novo*. Međutim, ova tehnologija još uvijek je ograničena velikim udjelom pogrešaka takvih sljedova, koji se na osnovu mjere kvalitete eliminiraju iz daljnjeg istraživanja. Cilj našeg istraživanja bio je identificirati i usporediti kvalitetne i nekvalitetne sljedove s obzirom na stopu i vrstu pogreške, a zatim i ispraviti pogreške u sljedovima prikupljenim sekvenciranjem nanoporama u svrhu dobivanja što točnijih sljedova. Koristili smo podatke iz pokusa sekvenciranja genomske DNA ogulinske špiljske spužvice (*Eunapius subteranneus* Sket i Velikonja, 1984) za koji su izrađene sekvence korištenjem dvije tehnologije; sekvenciranje metodom Illumina i nanoporama. Napravili smo sravnjenje točnijih i kraćih sljedova izknjižnice Illumina s odabranim uzorkom sljedova dobivenim tehnologijom sekvenciranja nanoporama. Dobiveno sravnjenje iskoristili smo za analizu i ispravljanje pogrešaka u uzorku sljedova dobivenih sekvenciranjem nanoporama. Odredili smo da je stopa pogreške u kvalitetnim sljedovima 8,61%, a u nekvalitetnim sljedovima 10,55%. Analizom glavnih komponenti ustanovili smo visok stupanj sličnosti navedenih skupina sljedova obzirom na udjele vrsta pogrešaka. U postupku ispravljanja, smanjili smo stopu pogreške u kvalitetnim sljedovima na 1,23%, odnosno u nekvalitetnim sljedovima na 2,33% pri čemu smo ispravili više od 60% svih pogrešno očitane baze u obje skupine sljedova. Uspješnost ispravljanja sljedova provjerili smo taksonomskom klasifikacijom sljedova, kao i funkcijskom klasifikacijom pronađenih proteina. Sravnjenjem analiziranog uzorka sljedova s bazom neredundantnih proteina programom DIAMOND pokazali smo višestruko povećanje broja pronađenih proteina uispravljenim sljedovima, uključujući proteine ogulinske špiljske spužvice. Također, upotrebom algoritma LCA (engl. *last common ancestor*) utvrdili smo povećanje broja sljedova pridruženih kategorijama Metazoa i Porifera nakon njihova ispravljanja. Isto tako, ispravljanjem sljedova pronašli smo veći broj proteina čija funkcija nije opisana kod ogulinske špiljske spužvice. Prikazani rezultati korak su prema poboljšanju kvalitete sljedova prikupljenih sekvenciranjem nanoporama. Upotrebom ispravljenih sljedova u sastavljanju genoma hrvatskog endema i jedine stigobiontske spužve na svijetu, ogulinske špiljske spužvice, zasigurno će se dobiti odgovori na brojna biološka pitanja, kao i uvidi u nove, za sada nepoznate značajke skrivene biologije spužvi.

Ključne riječi: tehnologije sekvenciranja sljedeće generacije, kvaliteta slijeda DNA, stopa pogreške sekvenciranja, taksonomska klasifikacija, funkcijska klasifikacija

10. SUMMARY

COMPUTATIONAL ANALYSIS OF DNA READS FROM ENDEMIC CAVE SPONGE *EUNAPIUS SUBTERRANEUS* OBTAINED BY NANOPORE SEQUENCING TECHNOLOGY

Kristian Bodulić i Paula Štancl

Next-generation sequencing technologies are constantly being improved towards a quicker generation of longer and less erroneous reads. Reads obtained by nanopore sequencing technology (Oxford Nanopore Technologies, ONT) present great potential in genomic and biomedical applications, especially in de novo genome assembly. However, reads generated with this technology still suffer from high sequencing error rates, resulting in their flagging as low quality and subsequent elimination from downstream processing. The goal of this study was to compare high-quality vs. low-quality long reads and correct identified errors in the genomic sequencing dataset obtained by both nanopore and Illumina sequencing of the endemic cave sponge (*Eunapius subterraneus* Sket & Velikonja, 1984). We aligned reads obtained by a less error-prone Illumina sequencing technology against a selected sample of nanopore reads in order to identify nanopore sequencing errors and correct them. We calculated the error rate of high-quality and low-quality nanopore reads to be 8,61% and10,55% respectively. Using principal component analysis we determined a high degree of similarity between both read groups regarding the ratio of error types. By correcting high-quality and low-quality reads we reduced their error rate to 1,23% and 2,33% respectively, while reducing the overall basecalling error rate by more than 60%. In order to estimate the correction efficiency, we also classified the reads by their taxonomy and function. By aligning the reads to the non-redundant protein database using DIAMOND, we demonstrated that their correction drastically increased the number of found proteins, including the proteins of the endemic cave sponge. Furthermore, using the LCA (last common ancestor) algorithm, we found that the correction increased the number of reads classified as Metazoa and Porifera. Finally, by examining the functional classification of the proteins found in the corrected reads, we identified an increased number of proteins whose function in the endemic cave sponge is still unknown. The demonstrated results represent a significant step in improving the nanopore sequencing quality of the endemic cave sponge genome. The use of corrected reads in assembling the genome of the endemic cave sponge, the only stygobitic sponge in the world, could provide answers to many biological questions and also shed a light on unknown features of the hidden biology of sponges.

Keywords: next-generation sequencing technologies, DNA sequence quality, sequencingerror rate, taxonomic classification, functional analysis

11. ŽIVOTOPISI

Kristian Bodulić

Kristian Bodulić rođen je 10.3.1997. godine u talijanskom gradu Erba (Como). U Zagrebu 2011. godine završava Osnovnu školu Pavleka Miškine, a 2015. godine Prirodoslovnu školu Vladimira Preloga. Tijekom svog srednjoškolskog obrazovanja sudjelovao je na mnogim natjecanjima, uključujući dva Državna natjecanja iz matematike i Državno natjecanje iz biologije na kojem je postigao treće mjesto. Nakon završetka srednjoškolskog obrazovanja upisao je Preddiplomski studij molekularne biologije na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu. Nakon završetka preddiplomskog studija s prosjekom ocjena 5,00, 2018. godine upisao je Diplomski studij molekularne biologije. Tijekom studija bio je demonstrator na kolegijima Osnove fizikalne kemije i Bioinformatika, a 2018. godine osvojio je Dekanovu nagradu za najboljeg studenta Preddiplomskog studija molekularne biologije. Također, od ožujka 2018. godine vrši ulogu suvoditelja bioinformatičke sekcije studentske udruge BIUS u sklopu koje organizira edukativna predavanja i radionice. Od travnja 2018. godine obavlja stručnu praksu u Grupi za bioinformatiku na PMF-u u Zagrebu gdje se bavi genomikom bazalnih životinja.

Paula Štancl

Paula Štancl rođena je 22. lipnja 1996. godine u Zagrebu. Završila je srednjoškolsko obrazovanje u XV. gimnaziji te je studentica prve godine diplomskog studija molekularne biologije na Prirodoslovno-matematičkom fakultetu u Zagrebu. Za vrijeme srednjoškolskog obrazovanja osvojila je prvo mjesto na državnom natjecanju iz biologije u izradi istraživačkih projekata pod naslovom „Od blata do mikro svjetla“. Dobitnica je brončane nagrade na međunarodnoj olimpijadi INESPO (*International Environment & Sustainability Project Olympiad)* u Amsterdamu 2015. i Stipendije grada Zagreba za izvrsnost 2018. godine. Tri godine je volontirala na Institutu Ruđer Bošković u Laboratoriju za evolucijsku genetiku u sklopu projekta Filostratigrafija nastanka i gubitka gena (PhyLoss). Također, od travnja 2018. godine volontira u Grupi za bioinformatiku na Zavodu za molekularnu biologiju Prirodoslovno-matematičkog fakulteta gdje se bavi genomikom bazalnih organizama. Suvoditeljica je sekcije Bioinformatike u studentskoj udruzi BIUS od 2018. godine. Organizirala je više edukativnih predavanja i radionica od kojih je vrijedno spomenuti radionicu na ZIMS (*Zagreb International Medical Summit for Medical Students and Young Doctors*) 2018. i *swapshop* radionicu „*Phylogenetic tree*“ na S3 (*Summer School of Science*) u Požegi 2016.

PRILOZI

Prilog 1. Programska skripta pokrenuta u ljusci *bash*

```
#!/bin/bash
##minimap2
MINIMAP2=path/to/minimap2
REF=ONT500.fasta
QUERY1=Illumina_paired_end1.fasta
QUERY2=Illumina_paired_end2.fasta
OUTPUT_MINIMAP=output.sam

$MINIMAP2 -x sr -a -k 11 -w 1 --end-bonus 50 -N 500 --secondary=no -
t 12 $REF $QUERY1 $QUERY2 > $OUTPUT_MINIMAP

##SAMBAMBA(sam-bam-sort_rmPCRdup-index)
SAMBAMBA=path/to/sambamba

$SAMBAMBA view -f bam -S -t 20 -o ${OUTPUT_MINIMAP}.bam $OUTPUT_MINI
MAP
$SAMBAMBA sort -t 20 -m 50G -o sorted_${OUTPUT_MINIMAP}.bam ${OUTPUT
_MINIMAP}.bam
$SAMBAMBA markdup -r -t 20 sorted_${OUTPUT_MINIMAP}.bam sorted_${OUT
PUT_MINIMAP}_noPCRdup.bam
$SAMBAMBA index sorted_${OUTPUT_MINIMAP}_noPCRdup.bam

##pilon
PILON=path/to/pilon
INPUT_PILON=$REF
OUTPUT_PILON_DIRPREFIX=pilon_directory_prefix
OUTPUT_PILON_FILE=pilon_file_prefix

java -Xmx350G -jar $PILON --genome $INPUT_PILON \
--frags sorted_${OUTPUT_MINIMAP}_noPCRdup.bam \
--output $OUTPUT_PILON_FILE \
--changes --vcf --tracks \
--threads 20 --verbose \
--outdir $OUTPUT_PILON_DIRPREFIX

##vcftools + bcftools
VCFTOOLS=path/to/vcftools
BCFTOOLS=path/to/bcftools
BCFTOOLS_INPUT=$OUTPUT_PILON_DIRPREFIX/${OUTPUT_PILON_FILE}.vcf
BCFTOOLS_OUTPUT=${OUTPUT_PILON_FILE}_dp40_mq20
$BCFTOOLS filter $BCFTOOLS_INPUT 'MIN(DP)>40 && MIN(MQ)>20' > $BCFTO
OLS_OUTPUT
VCFTOOLS_PREFIX=${BCFTOOLS_OUTPUT}_max0.03
$VCFTOOLS --vcf $BCFTOOLS_OUTPUT---max-maf 0.03 --out $VCFTOOLS_PREF
IX --recode
VCFTOOLS_OUTPUT=${VCFTOOLS_PREFIX}_recode.vcf
$BCFTOOLS stats -F $REF $VCFTOOLS_OUTPUT

##DIAMOND + MEGAN
DIAMOND=path/to/diamond
DATABASE=path/to/nrprotein/database
MEGANIZE=/path/to/meganizer
MEGAN_DATA=/path/to/megan_data
FORMAT_TAB='6 qseqid qlen sseqid slen staxids stitle qcovhsp length
```

```
pident mismatch qstart qend sstart send evalue bitscore'
FORMAT_DAA='100'

#Neispravljeni sljedovi
INPUT_DIAMOND_NEISPRAVLJENI=$REF
OUTPUT_DIAMOND_NEISPRAVLJENI_TAB=tabular_output.txt
OUTPUT_DIAMOND_NEISPRAVLJENI_DAA=output.daa
$DIAMOND blastx --sensitive --evalue 10 -F 15 --range-culling -d $DA
TABASE -q $INPUT_DIAMOND_NEISPRAVLJENI -o $OUTPUT_DIAMOND_NEISPRAVLJ
ENI_TAB-f $FORMAT_TAB -p 12
$DIAMOND blastx --sensitive --evalue 10 -F 15 --range-culling -d $DA
TABASE -q $INPUT_DIAMOND_NEISPRAVLJENI -o $OUTPUT_DIAMOND_NEISPRAVLJ
ENI_DAA -f $FORMAT_DAA -p 12
$MEGANIZE --in $OUTPUT_DIAMOND_NEISPRAVLJENI_DAA--longReads --lcaAlg
orithm longReads --acc2taxa $MEGAN_DATA/prot_acc2tax-Mar2018X1.abin
--acc2eggnog $MEGAN_DATA/acc2eggnog-Oct2016X.abin --acc2interpro2go
$MEGAN_DATA/acc2interpro-Nov2016XX.abin --acc2kegg $MEGAN_DATA/acc2k
egg-Dec2017X1-ue.abin --acc2seed $MEGAN_DATA/acc2seed-May2015XX.abin

##Ispravljeni sljedovi
INPUT_DIAMOND_ISPRAVLJENI=$OUTPUT_PILON_DIRPREFIX/${OUTPUT_PILON_FIL
E}.fasta
OUTPUT_DIAMOND_ISPRAVLJENI_TAB=tabular_output.txt
OUTPUT_DIAMOND_ISPRAVLJENI_DAA=output.daa
$DIAMOND blastx --sensitive --evalue 10 -F 15 --range-culling -d $DA
TABASE -q $INPUT_DIAMOND_ISPRAVLJENI -o $OUTPUT_DIAMOND_ISPRAVLJENI_
TAB -f $FORMAT_TAB -p 12
$DIAMOND blastx --sensitive --evalue 10 -F 15 --range-culling -d $DA
TABASE -q $INPUT_DIAMOND_ISPRAVLJENIf -o $OUTPUT_DIAMOND_ISPRAVLJENI
_DAA -f $FORMAT_DAA -p 12
$MEGANIZE --in $OUTPUT_DIAMOND_ISPRAVLJENI_DAA --longReads --lcaAlgo
rithm longReads --acc2taxa $MEGAN_DATA/prot_acc2tax-Mar2018X1.abin -
-acc2eggnog $MEGAN_DATA/acc2eggnog-Oct2016X.abin --acc2interpro2go $
MEGAN_DATA/acc2interpro-Nov2016XX.abin --acc2kegg $MEGAN_DATA/acc2ke
gg-Dec2017X1-ue.abin --acc2seed $MEGAN_DATA/acc2seed-May2015XX.abin
```

Prilog 2. Programski postupak za dobivanje slika napisan u programskom jeziku R.

```
#Korišteni paketi
library(ggplot2)
library(reshape2)
library(ggfortify)
library(ggpubr)
library(cowplot)

table <- readRDS("Podaci.Robj") #učitavanje podataka

#Slika 8.
ggplot(table[[1]], aes(sequence_length_template, mean_qscore_templat
e, color=skup_podataka)) +
  geom_point(show.legend = F) +
  theme_classic() +
  labs(x = "Duljina slijeda / pb",  y = "Prosječna mjera kvalitete s
lijeda") +
  scale_x_continuous(labels = scales::comma, breaks = seq(0, 257000,
50000))

#Slika 9.
ggplot(table[[1]], aes(x=skup_podataka, y=log(sequence_length_templa
te))) +
  geom_violin(aes(color=skup_podataka, fill=skup_podataka), show.leg
end=F) +
   labs(x = NULL,  y = "Logaritamski transformirana duljina slijeda"
) +
  theme_classic() +
  facet_grid(~filename)+
  scale_y_continuous(labels = scales::comma)

#Slika 10.
gr.m <- melt(table[[2]], "skup")
ggplot(gr.m, aes(x = factor(variable), y = value, fill = skup)) +
geom_col(position = "dodge", show.legend = F)+
  theme_classic() +
   labs(x = NULL, y="Stope pogrešaka / %")

#Slika 11.
autoplot(prcomp(table[[3]][, c(2,3,4,5)]), data = as.data.frame(pca)
, colour = 'skup',loadings = F, loadings.label = TRUE, loadings.labe
l.size = 5) + theme_classic() + theme(legend.position="none")

#Slika 12.
ggboxplot(table[[4]], x = "skup", y = "norm",
          color = "skup", palette = c("npg"), add="jitter") + stat_c
ompare_means() +  ylim(0,0.3) + labs(x=NULL, y="Stopa promjene")

#Slika 13.
ggplot(table[[5]], aes(x = factor(MODEL), y = value, fill = skup_pod
ataka)) +
geom_col(position = "dodge", show.legend = F)+
  theme_classic() +
   labs(x = NULL, y="Udio parova pogrešno očitanih baza / %")

#Slika 14.
ggplot(data =  table[[6]], aes(x = Var1, y = Var2)) +
```

```
  geom_tile(aes(fill = value), colour = "white") +
  geom_text(aes(label = sprintf("%1.2f", value)), vjust = 1) +
  scale_fill_gradient(low = "white", high = "salmon2") +
   labs(x = "Pogrešna baza",  y = "Točna baza")+
  theme_classic()

#Slika 15.
ggplot(table[[7]], aes(V3, V4, fill=skup)) +
  geom_col(alpha=0.7, show.legend = F) +
  theme_classic() +
  labs(x="Duljina indela / pb", y="Učestalost indela / %") +
  facet_wrap(~skup)

#Slika 16.
g1 <-ggplot(data =  table[[8]], aes(x = Var1, y = Var2)) +
  geom_tile(aes(fill = log(value)), colour = "white") +
  geom_text(aes(label = sprintf("%1.2f", value)), vjust = 1) +
  scale_fill_gradient(low = "white", high ="salmon2" ) +
   labs(x = "Pogrešna baza",  y = "Točna baza")+
  theme_classic()

g2 <-ggplot(data =  table[[9]], aes(x = Var1, y = Var2)) +
  geom_tile(aes(fill = log(value)), colour = "white") +
  geom_text(aes(label = sprintf("%1.2f", value)), vjust = 1) +
  scale_fill_gradient(low = "white", high = "salmon2") +
   labs(x = "Pogrešna baza",  y = "Točna baza")+
  theme_classic()
plot_grid(g1, g2, labels = "AUTO",nrow = 2, align = "v", rel_heights
=2, rel_widths=c(0.1,0.1))

#Slika 17.
jedan<-ggplot(table[[10]][1:20,],aes(x=rod,y=omjer2,fill=condition))
+
  labs(x = "Rod",y= "Udio pronađenih proteina")+
  scale_x_discrete(limits=c("Xenopus","Rhizophagus","Trichinella","H
ydra","Branchiostoma", "Acropora", "Crassostrea","Stylophora","Exaip
tasia","Amphimedon")) +
  scale_fill_manual(values=c("#993399","#FFCC66"))+
  geom_bar(position="dodge", stat="identity") +
  theme_minimal() +
  theme(legend.title = element_blank())+
  theme(panel.grid.major = element_blank(), panel.grid.minor = eleme
nt_blank())+
  coord_flip()

dva<-ggplot(table[[11]][1:20,],aes(x=rod,y=omjer2,fill=condition))+
  labs(x = "Rod",y= "Udio pronađenih proteina")+
  scale_x_discrete(limits=c("Rhizophagus","Hydra", "Crassostrea","Ac
ropora","Branchiostoma","Limnohabitans","Trichinella","Stylophora","
Exaiptasia","Amphimedon")) +
  scale_fill_manual(values=c("#993399","#FFCC66")) +
  geom_bar(position="dodge", stat="identity") +
  theme_minimal() +
  theme(legend.title = element_blank())+
  theme(panel.grid.major = element_blank(), panel.grid.minor = eleme
nt_blank()) +
  coord_flip()
```

```
plot_grid(jedan,dva, labels = c("A", "B"), align="v", ncol=1)

#Slike 19. i 23.
lapply(unique(table[[12]]$odvoji), function(x) ggplot(tt[odvoji==x],
aes(variable, value, fill=funkcija)) +
  geom_col(stats="identity", position = "stack", show.legend=T) +
  theme_classic() +
   labs(x = NULL, y="Broj pronađenih proteina") +
  scale_fill_brewer(palette = "Set3") +
  theme(axis.text.x = element_text(size=10), axis.text.y = element_t
ext(size=11), axis.title.y = element_text(size=11))+
  scale_x_discrete(labels = function(x) lapply(strwrap(x, width = 10
, simplify = FALSE), paste, collapse="\n")))

#Slike 20. i 24.
lapply(unique(table[[13]]$odvoji), function(x) ggplot(table[[13]][od
voji==x], aes(x = `Biološki procesi`, y = value, fill= variable)) +
geom_col(position = "dodge", legend.position="bottom") +
  theme_classic() +
   labs(x = NULL, y="Broj pronađenih proteina")    +
  theme(axis.text.x = element_text(size=21, angle=75, hjust=1), axis
.text.y = element_text(size=21), axis.title.y = element_text(size=21
), legend.text=element_text(size=19)) +
  scale_x_discrete(limits=unique(table[[13]][odvoji==x]$`Biološki pr
ocesi`), labels = function(x) lapply(strwrap(x, width = 22, simplify
= FALSE), paste, collapse="\n")) +
  scale_fill_discrete(labels = function(x) lapply(strwrap(x, width =
22, simplify = FALSE), paste, collapse="\n")) +
  theme(legend.position="bottom"))

#Slike 21. i 25.
lapply(unique(table[[14]]$odvoji), function(x) ggplot(table[[14]][od
voji==x], aes(x = `Molekularna funkcija`, y = value, fill= variable)
) +
geom_col(position = "dodge", legend.position="bottom") +
  theme_classic() +
   labs(x = NULL, y="Broj pronađenih proteina")    +
  theme(axis.text.x = element_text(size=21, angle=75, hjust=1), axis
.text.y = element_text(size=21), axis.title.y = element_text(size=21
), legend.text=element_text(size=19)) +
  scale_x_discrete(limits=unique(table[[14]][odvoji==x]$`Molekularna
funkcija`), labels = function(x) lapply(strwrap(x, width = 22, simpl
ify = FALSE), paste, collapse="\n")) +
  scale_fill_discrete(labels = function(x) lapply(strwrap(x, width =
22, simplify = FALSE), paste, collapse="\n")) +
  theme(legend.position="bottom"))

#Slike 22. i 26.
lapply(unique(table[[15]]$odvoji), function(x) ggplot(table[[15]][od
voji==x], aes(x = `Stanična lokalizacija`, y = value, fill= variable
)) +
geom_col(position = "dodge", show.legend=T) +
  theme_classic()  +
   labs(x = NULL, y="Broj pronađenih proteina")    +
  theme(axis.text.x = element_text(size=21, angle=75, hjust=1), axis
.text.y = element_text(size=21), axis.title.y = element_text(size=21
), legend.text=element_text(size=19)) +
```

```
  scale_x_discrete(limits=unique(table[[15]][odvoji==x]$`Stanična lo
kalizacija`), labels = function(x) lapply(strwrap(x, width = 22, sim
plify = FALSE), paste, collapse="\n")) +
  scale_fill_discrete(labels = function(x) lapply(strwrap(x, width =
22, simplify = FALSE), paste, collapse="\n")) +
  theme(legend.position="bottom"))
```