Računalna analiza knjižnica dobivenih metodama sekvenciranja sintezom prema kalupu i sekvenciranja nanoporama

Habulin, Dunja

Master's thesis / Diplomski rad

2016

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

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:078287

Sveučilište u Zagrebu
Prirodoslovno – matematički fakultet
Biološki odsjek

Dunja Habulin

Računalna analiza knjižnica dobivenih metodama sekvenciranja sintezom prema kalupu i sekvenciranja nanoporama

Diplomski rad

Zagreb, 2016.

Ovaj rad je izrađen pri Zavodu za molekularnu biologiju, pod vodstvom prof. dr. sc. Kristiana Vlahovičeka, predan je na ocjenu Biološkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu radi stjecanja zvanja magistra molekularne biologije.

Zahvaljujem se Maji Fabijanić, Filipu Horvatu i Dori Šribar na stručnoj podršci, obitelji, prijateljima i Martinu na moralnoj podršci te mentoru prof. dr. sc. Kristianu Vlahovičeku.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Diplomski rad

Računalna analiza knjižnica dobivenih metodama sekvenciranja sintezom prema kalupu i sekvenciranja nanoporama

Dunja Habulin

Rooseveltov trg 6, 10000 Zagreb, Hrvatska

Danas je sve popularnije za analizu i sklapanje genoma *de novo* koristiti hibridne podatke, odnosno sljedove očitane korištenjem dviju ili više različitih metoda sekvenciranja. Metoda sekvenciranja sintezom prema kalupu (Illumina) i metoda sekvenciranja nanoporama (Oxford Nanopore) vrlo su dobra kombinacija zbog velike količine i preciznosti očitanih sljedova dobivenih metodom Illumina te duljine očitanih sljedova dobivenih metodom Oxford Nanopore. U ovom radu analizirala sam knjižnice dobivene hibridnim sekvenciranjem genomske DNA spužve *Ephydatia mülleri*. Spužve pripadaju koljenu Porifera i moguća su prekretnica u ranoj evoluciji carstva Metazoa. Cilj je sklopiti što više genoma ove skupine kako bi se razjasnili brojni evolucijski događaji i odnosi između ostalih koljena carstva Metazoa. S obzirom da su spužve obligatni simbionti s mnogim bakterijskim vrstama, kao i da ih nije moguće uzgojiti u sterilnim laboratorijskim uvjetima, pripremljene knjižnice za sekvenciranje sadrže i određenu razinu kontaminirajuće DNA. Pomoću metoda strojnog učenja klasterirala sam dobivene sljedove obiju tehnologija sekvenciranja kako bih analizirala zastupljenost kontaminantnih sljedova koji ne pripadaju genomu spužve. Sljedove unutar pojedinih klastera sastavila sam u neprekinute nizove. Usporedbom rezultata klasteriranja i učinkovitosti pronalaska kontaminantnih sljedova, prednost imaju sljedovi dobiveni metodom Oxford Nanopore. Bez obzira na ovaj metodološki pristup, i dalje postoje problemi u pouzdanoj identifikaciji kontaminanata, stoga je neophodno ovaj postupak unaprijediti dodatnim računalnim metodama.

(45 stranica, 20 slika, 7 tablica, 45 literaturnih navoda, jezik izvornika: hrvatski)

Rad je pohranjen u Središnjoj biološkoj knjižnici

Ključne riječi: tehnologije sekvenciranja sljedeće generacije, nenadzirano strojno učenje, sklapanje neprekinutih sljedova

Voditelj: Prof. dr. sc. Kristian Vlahoviček

Ocjenitelji: Prof. dr. sc. Kristian Vlahoviček
Prof. dr. sc. Zlatko Liber
Doc. dr. sc. Tomislav Ivanković

Zamjena: Izv. prof. dr. sc. Dunja Leljak-Levanić

Rad prihvaćen: 4. srpnja 2016.

BASIC DOCUMENTARY CARD

University of Zagreb
Faculty of Science
Division of Biology

Graduation thesis

Computational analysis of short read Illumina and nanopore sequencing libraries

Dunja Habulin

Rooseveltov trg 6, 10000 Zagreb, Croatia

It is increasingly popular to use hybrid data, i.e. reads from two or more different sequencing libraries, for analysis and *de novo* genome assembly. The Illumina and the Oxford Nanopore sequencing methods are a very good combination used for this, since the Illumina produces extensive amounts of precise data and the Oxford Nanopore has shown to produce long reads. I analysed genomic DNA libraries of *Ephydatia mülleri* sponge, which were a result of hybrid sequencing. Sponges belong to the Porifera phylum and might be a milestone in the early evolution of Metazoa kingdom. Therefore it is very important to assemble as many genomes as possible in this particular group in order to clarify various events in the metazoan evolution and relations between other groups within this kingdom. Considering the fact that sponges are indeed obligatory symbionts with many bacterial species and the fact that they are impossible to grow in sterile laboratory conditions, prepared sequencing libraries also contain a certain level of DNA contamination. I clustered reads belonging to both sequencing technologies using machine learning methods, in order to analyse the abundance of contaminant reads. Reads within particular cluster were used to make contigs. After the comparison of both clustering results and contaminant detection efficiency, reads from the Oxford Nanopore sequencing technologies have shown better results. Regardless of the applied methodology, there are still problems with certain contaminant identification and it is therefore essential to further enhance the protocol by using additional computational methods.

(45 pages, 20 figures, 7 tables. 45 references, original in: Croatian)

Thesis deposited in the Central Biological Library.

Key words: next generation sequencing technologies, unsupervised machine learning, contig assembly

Supervisor: Professor Kristian Vlahoviček, PhD

Reviewers: Professor Kristian Vlahoviček, PhD
Professor Zlatko Liber, PhD
Asst. Tomislav Ivanković, PhD

Substitution: Assoc. Dunja Leljak-Levanić, PhD

Thesis accepted: July 4, 2016

Sadržaj

Kratice

BLAST	*engl*. Basic Local Alignment Search Tool
kb	kilobaza
nk	Nukleotid
ONT	*engl*. The Oxford Nanopore Technologies
pb	Nukleotidni par
PCR	*engl*. lančana reakcija polimerazom
PRK	Preklapanje-raspored-konsenzus
t-SNE	Stoihastičko pridruživanje susjedima temeljeno na t-distribuciji

1 Uvod

1.1 Tehnologije sekvenciranja

Sekvenciranje danas vjerojatno spada u jednu od najbrže razvijajućih metoda u molekularnoj biologiji, a podrazumijeva čitanje slijeda dušičnih baza u molekuli DNA. Prvu metodu sekvenciranja razvio je Sanger 1975. godine. Tadašnja tehnologija podrazumijevala je umnažanje kratkog umnoženog uzorka DNA pomoću enzima DNA polimeraze. Uz enzim, potrebni su deoksinukleotidi i ono što je za ovu metodu specifično i neophodno, a to su dideoksinukleotidi koji onemogućavaju nastavak sinteze komplementarnog lanca. Nakon umnožavanja i sinteze, uzorci se razdvajaju na elektroforetskom gelu te se očita slijed dušičnih baza u analiziranoj molekuli (Sanger *i sur.*, 1977). Ova metoda pripada metodama sekvenciranja prve generacije, a danas su prisutni uređaji koji se temelje na Sangerovoj metodi, ali je metoda automatizirana, koriste se fluorescentno obilježeni prekidajući nukleotidi, uzorci se razdvajaju koristeći kapilarnu elektroforezu, a postoji i laserska detekcija signala. Metoda je i danas prilično zastupljena, ali je tehnologija ipak omogućila razvoj metoda koje su preciznije, omogućavaju duže očitane sljedove i ono najvažnije, veliku količinu podataka u kratkome vremenu.

Sangerova metoda i inačice danas se zovu tehnologije sekvenciranja prve generacije. Početkom 21. stoljeća pojavila se druga skupina tehnologija, tzv. tehnologije sekvenciranja druge generacije. Razlike u odnosu na Sangerovu metodu su da nije potrebno bakterijsko kloniranje, moguće je izvoditi više reakcija sekvenciranja paralelno, a smanjeni su i gubitci te je povećana brzina dobivanja podataka jer elektroforeza više nije potrebna. Prednost ovih tehnologija u odnosu na tehnologije prve generacije je količina dobivenih podataka i brzina kojom se sekvenciranje odvija, a smanjeni su i gubitci jer više ne postoji potreba za elektroforezom. Glavni nedostaci su manja duljina sljedova (Quail *i sur.*, 2012) i manja preciznost i točnost dobivenih sljedova (van Dijk *i sur.*, 2014).

Tehnologije sekvenciranja druge generacije umnažaju DNA molekule lančanom reakcijom polimerazom (*engl. Polymerase Chain, Reaction*) ili PCR-om, a zahvaljujući posebnim vrstama umnažanja poput umnažanja na čvrstoj podlozi (*engl. bridge amplification*) ili

umnažanja u emulziji (*engl. emulsion PCR*) (Metzker, 2010), omogućena je veća količina podataka i veća brzina. Umnažanje je i dalje potrebno kako bi signal bio jači, jer detektor ne može uočiti i zapisati signal koji potječe od ugradnje samo jedne fluorescentne baze. Tehnologije druge generacije su Illumina, SOLiD i Roche 454 (Mitra i Church, 1999).

Treća generacija metoda napredovala je u kontekstu da više nema potrebe za umnažanjem, već se sekvencira jedna molekula DNA (*engl. single molecule sequencing*). Tehnologije koje ne koriste umnažanje imaju bolje razvijene detektore i komorice u kojima se sekvencira jedna jedina molekula DNA (*engl. zero-mode waveguide detectors*) (Metzker, 2010). Na taj se način izbjegava mogućnost pristranosti u našim podacima jer nema nejednolikog umnožavanja, odnosno krivo očitane baze. Tehnologije Pacific Biosciences omogućuju sekvenciranje jedne molekule u realnom vremenu (*engl. Single Molecule, Real-Time*), a uz Heliscope tehnologije ne koriste umnažanje.

1.1.1 Metoda sekvenciranja sintezom prema kalupu DNA (Illumina)

Nakon dolaska na tržište 2005. godine, Illumina je nove platforme razvijala prilično brzo, tako da ih do danas ima čak pet. Popularnost Illumine vrlo je brzo narasla zbog velikog broja sljedova koji se dobiju odjednom te niske cijene sekvenciranja (Hodkinson i Grice, 2015). Protočna ćelija (*engl. flow cell*) omogućava simultane reakcije, čime se omogućuje veći broj očitanih sljedova u kraćem vremenu. Illumina koristi umnažanje na čvrstoj podlozi, a prvi korak je priprema knjižnice. Priprema knjižnice obuhvaća fragmentiranje DNA, popravak krajeva, fosforilaciju 5' kraja te dodavanje slijeda A na 3' kraj. Nakon toga dodaju se adapteri na krajeve DNA pripremljene u prethodnom koraku. Protočna ćelija cijelom svojom površinom ima pričvršćene početnice s 5' i 3' kraja (*engl.* „forward“ i „reverse“) koje su komplementarne adapterima koji se nalaze na krajevima fragmenata. Fragmenti DNA dvolančani su, a kako bi se omogućile reakcije sinteze, molekule DNA denaturiraju se u jednolančane fragmente. Tada je omogućena hibridizacija fragmenata s početnicama na podlozi. Kada hibridiziraju, odvija se umnažanje početnih kalupa. Tako nastaju kopije početnih lanaca, nakon čega se kalupi odstranjuju. Slijedi umnažanje kopija pomoću početnica, te nastaju nakupine identičnih molekula DNA (*engl. clonal cluster*). Jedan lanac u sada dvolančanoj DNA cijepa se na specifičnim mjestima u oligonukleotidnim početnicama, te se blokira 3' slobodan kraj na pričvršćenom lancu kako bi se spriječila reakcija na slobodnom kraju. Princip ovog sekvenciranja je da se čita jedan po jedan nukleotid u svakom

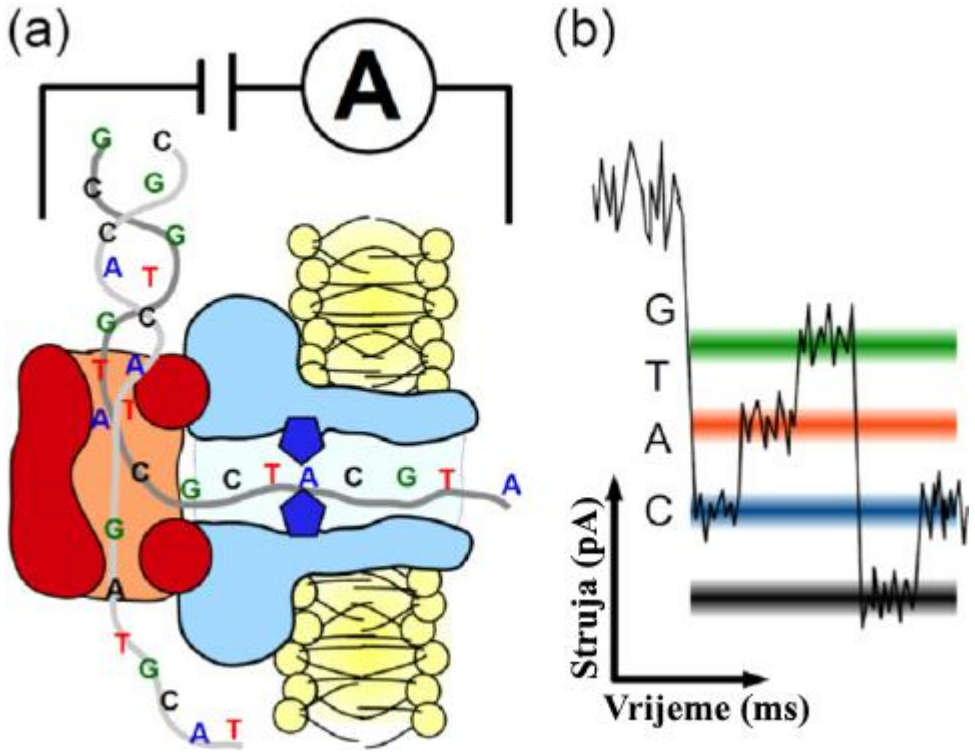
ponovljenom ciklusu pomoću signala kojeg odašilju ugrađeni fluorescentno obilježeni deoksiribinukleotidi.

1.1.1.1 Vrste knjižnica prema udaljenosti očitanih sljedova

Ovisno o duljini umetnutog dijela postoje tri vrste knjižnica. Knjižnice kratkih udaljenosti između parova očitanih fragmenata (*engl. paired end*), knjižnice uparenih očitanih sljedova (*engl. mate-pair*) i knjižnice dugih uparenih očitanih sljedova (*engl. long mate-pair*).
Knjižnice kratkih udaljenosti između parova očitanih fragmenata koriste se kada je dio oko kojeg sekvenciramo dugačak između 200 i 800 parova baza. Dobijemo dva očitana slijeda između kojih se nalazi dio kojemu se ne zna slijed, ali se (u pravilu) zna dužina.
Knjižnica uparenih očitanih razlikuje se od knjižnice kratkih udaljenosti po tome što se omogućuje sekvenciranje dijelova DNA koji su udaljeniji i do 12 kilobaza, pa je vrlo korisna u koraku sastavljanja prekinutih sljedova (*engl. scaffolds).*

1.1.2 Metoda sekvenciranja nanoporama (The Oxford Nanopore Technologies MinION, ONT)

Metoda sekvenciranja nanoporama odvija se na uređaju The Oxford Nanopore Technologies MinION. Površina tog uređaja prekrivena je nanoporama koje leže na membrani na kojoj se nalazi mreža električnog potencijala. Kroz nanoporu prolazi molekula DNA, a uređaj detektira promjene u struji prolaskom molekule. Očitava se slijed pentamera i „događaja“ kao što su vrijeme početka i trajanje prolaska molekule DNA kroz poru (Slika 1.). Da bi molekula DNA mogla biti sekvencirana ovim uređajem, potrebna je priprema knjižnice. Protokol obuhvaća zatvaranje dvolančane DNA ukosnicom na jednom kraju, a na drugom kraju dodaje joj se adapter na kojem se nalazi protein koji "vodi" DNA do pore. Zadaća proteina također je i odmotavanje molekule DNA pri ulasku u poru. Protein se nalazi na 5' kraju molekule i taj kraj prvi ulazi kroz poru. Dobiveni signal prevodi se u slijed nukleotida pomoću ONT Metrichor „cloud“ servisa. Pri idealnim uvjetima redom ulaze kalup, ukosnica, a na kraju komplementarni lanac i tako nastaje informacija s dva lanca, tzv. 2D slijed. No, često dolazi do komplikacija pa nastaju i 1D sljedovi koji nose informaciju sa samo jednog lanca. 2D sljedovi veće su kvalitete i nose točniju informaciju.



Slika 1 Prikaz promjene električnog polja prilikom prolaska određenog nukleotida kroz poru koja se nalazi na sintetskoj membrani pločice uređaja MinION. Preuzeto i prilagođeno s *https://www.researchgate.net/figure/271772842_fig3_Figure-3-Scheme-depicting-the-sequencing-technique-of-Oxford-Nanopore-Technologies.*

Uređaj je malen i jeftin, prednjači u duljini očitanih sljedova koji mogu biti duljine i preko 100 kb, uz to je visokoprotočan i brz (Eisenstein, 2012; Goodwin *i sur.*, 2015).

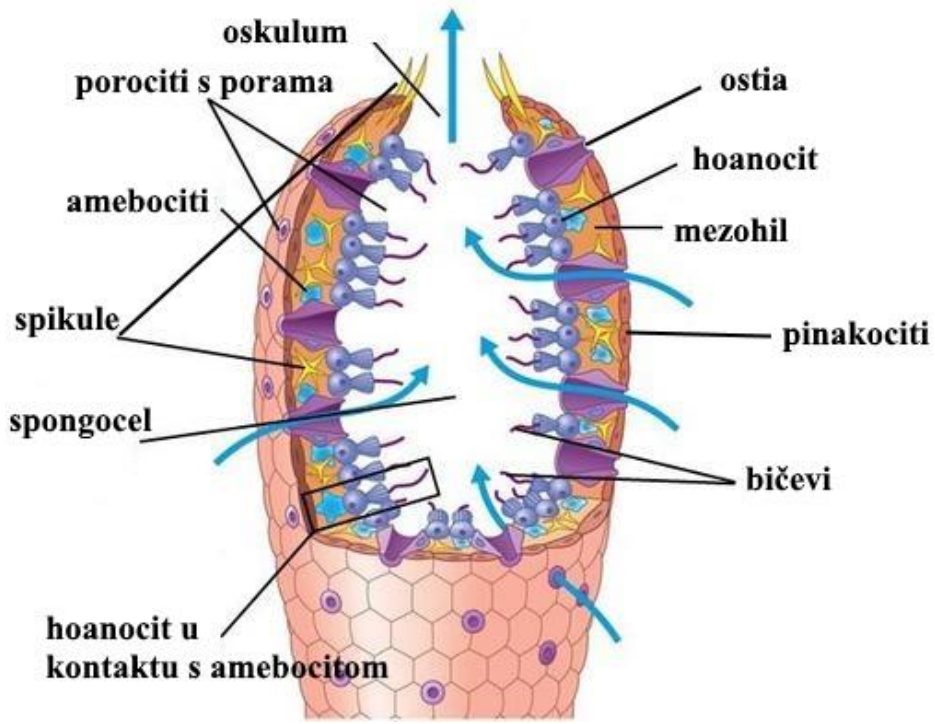
1.2 Općenito o skupini spužve

Spužve spadaju u carstvo životinja (Metazoa) i koljeno Porifera. One su taksonomski vrlo raznolike, u koljenu Porifera postoji preko 8000 znanstveno opisanih vrsta spužvi. Sesilni su i većinom morski organizmi, ali postoje i vrste koje žive u slatkim vodama. Smatraju se skupinom Metazoa koja se najranije počela granati, te najjednostavnijim i najstarijim višestaničnim životinjama. Upravo iz tih razloga važne su za razumijevanje početaka razvitka carstva Metazoa i odnosa unutar toga carstva (Wörheide *i sur.*, 2012). Za razliku od morfološki kompleksnijih životinja (Eumetazoa), spužve se zbog nedostatka pravih tkiva, organa i jednostavnog embrionalnog razvoja svrstavaju u skupinu životinja Parazoa (Ereskovsky i Dondua, 2006).

1.2.1 Životni ciklus i građa spužava

Tijelo spužava grade tri tipa stanica i tri sloja: vanjski, srednji i unutarnji sloj, od kojih je svaki visokospecijaliziran. Na Slici 2. možemo vidjeti na koji su način stanice raspoređene u tijelu spužve. Skelet spužve građen je od iglica (spikula) koje su različitih oblika i različitog mineralnog sastava (kombinacije silicijevog dioksida i kalcijevog karbonata) i/ili proteina spongina. Stanice vanjskoj sloja poligonalne su i spljoštene, a zovu se pinakocite. Srednji sloj naziva se mezohil i sastoji se od proteinskog želatinoznog i ugljikohidratnog matriksa u koji su uronjene pokretne stanice amebocite, skleroblase (stanice koje tvore skelet) i spolne stanice (Müller, 2006). U unutrašnjem sloju nalaze se bičaste stanice hoanociti. Važnost ovog sloja je u provođenju vode u unutrašnje kanalne sustave ispresjecane hoanocitima. Oni su zaslužni za strujanje vode kroz tijelo spužve jer bičevima proizvode vodenu struju koja dovodi kisik i hranjive tvari do stanica te odvodi otpadne tvari. Oskulum je veliki otvor koji spužvama služi za izbacivanje otpadnih produkata i filtrirane vode.

Razmnožavanje spužava dijeli se na spolno i nespolno. Spolno razmnožavanje podrazumijeva izbacivanje zrelih spermatozoida koji se vodom prenose do druge jedinke, a oplodnja je unutrašnja. Prilikom oplodnje važnu ulogu imaju specijalizirani hoanociti koji dovode spermije do jajne stanice, a oplodnjom nastaje zigota koja se razvija u bičastu ličinku i ispušta u vodu. Neko vrijeme pluta, a zatim se učvrsti za podlogu i tada počinje razvoj u odraslu spužvu. Pri nespolnom razmnožavanju nastaju rasplodna tijela gemule (Müller, 2006), oblik u kojem spužva preživljava nepovoljne uvjete, čak i nedostatak kisika ili smrzavanje. Dio stanica gemule odvoji se i nastaje novi organizam.



Slika 2 Građa spužve. Preuzeto i prilagođeno s *http://mrsdmarine.weebly.com/porifera.html.*

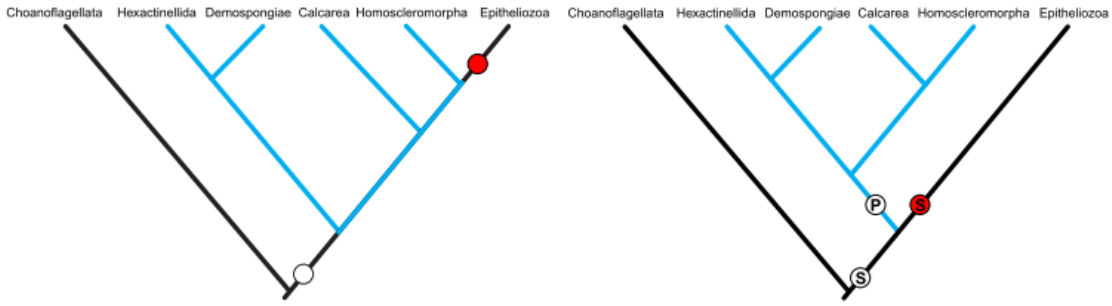
1.2.2 Filogenija koljena Porifera

U koljenu Porifera postoji ukupno preko 26 000 različitih vrsta (Hooper *i sur.*, 2013), a dijeli se na četiri glavna razreda ovisno o građi skeletnih elemenata:

1. **Demospongiae** je razred koji pokazuje najveću raznolikost i broji najveći broj vrsta. Spužve u ovoj skupini imaju silikatne iglice i/ili kostur od organskih vlakana ili kolagena.
2. **Hexactinellida**, drugi naziv im je staklače. Te su spužve isključivo morske. Građene su od silikatnih iglica triosne simetrije, a meko tkivo tvore nakupine stanica sinciciji.
3. **Homoscleromorpha** je mala grupa morskih spužava koja ima jedinstvena svojstva poput bazalne lamine ispod vanjskog i unutarnjeg sloja te viviparne ličinke.
4. **Calcarea** ili **Calcispongiae** prema starijoj literaturi, kod nas zvane vapnenjače. Ove su spužve većinom male i žive u plićacima. Morfološa značajka su vapnenačke spikule. (Van Soest *i sur.*, 2012)

Porijeklo spužvi još uvijek nije u potpunosti usuglašeno među znanstvenicima, na pitanje jesu li spužve monofiletskog ili parafiletskog porijekla još uvijek se traži odgovor.

Istraživanja 18S rRNA uočila su da su neke skupine spužava sličinije višim životinjama nego međusobno, te opisuju spužve kao parafiletsku skupinu (Borchiellini *i sur.*, 2001). Novija monofiletska istraživanja koriste podatke cDNA od 128 protein kodirajućih gena, fokusiraju se na jedinstvenost građe tijela spužve i kanalnog sustava, te zaključuju da je cijela skupina spužvi monofiletska sestrinska skupina višim životinjama (Eumetazoa) (Philippe *i sur.*, 2009).



Slika 3 Dvije teorije podrijetla spužava. Na lijevoj slici vidimo prikaz parafiletskog porijekla, prema kojem je spužvasti oblik tijela (bijeli kružić) nastao u zadnjem zajedničkom pretku Porifera i Epitheliozoa, a zatim se postepeno izgubio (crveni kružić). Desna slika prikazuje shemu monofiletske teorije, prema kojoj je spužvasti oblik tijela (bijeli kružić) nastao ili u liniji Porifera (P) ili u matičnoj grupi Metazoa (S) (Wörheide *i sur.*, 2012).

Skupina spužvi ima značajke prema kojima ih možemo smatrati prvim životinjama, ali su prilično primitivne građe i funkcionalnosti u odnosu na razvijenije životinje. Smatraju se najstarijim životinjama, a prva pojava ostataka spikula zabilježena je na kraju prekambija (kasni proterozoik) (Reitner i Worheide, 2002). Ne posjeduju organe ni živčani sustav, već u spužvama vidimo preteču organizacije stanica u tkiva. Spužve su građene od stanica koje su prilično labavo vezane u skupine i imaju sposobnost diferencijacije (Thacker *i sur.*, 2014). Ipak, postoje važne karakteristike koje dijele spužve od najbližih predaka okovratnih bičaša (Choanoflagellata), a svrstavaju ih u carstvo životinja. Višestaničnost, spermatogeneza, oogeneza, građa spermija, mejoza te ostali dostupni podaci glavni su razlozi svrstavanja spužava u carstvo Metazoa (Müller, 1995; Rokas *i sur.*, 2005).

1.2.3 Genetika spužava

Prvi genom spužve sekvenciran je u svibnju 2010. godine. Spužva *Amphimedon queenslandica* jedina je spužva za koju postoje opisi funkcionalnosti genoma. Prilično iznenađujuće je koliko se svojstava viših životinja nalazi u genomu spužava, obzirom na

jednostavnost građe. Spužve nemaju razvijeni živčani sustav, no pronađeni su geni uključeni u razvoj ovog sustava. Uz to, postoje i geni za adheziju, međustaničnu komunikaciju, kontrolu stanične smrti, signalizaciju te čak i imunosno prepoznavanje (Srivastava *i sur.*, 2010). Kasnija istraživanja na osam vrsta spužava došla su do sličnih zaključaka proučavajući transkriptome te pokušavajući razjasniti priču o životinjskoj kompleksnosti. Rezultati ovih istraživanja pokazuju da je genetička kompleksnost nastala vrlo rano u evoluciji te da spužve ili imaju kriptičnu fiziološku i morfološku kompleksnost i/ili su izgubile neke gene (Riesgo *i sur.*, 2014).

1.2.3.1 *Značajke spužve* Ephydatia mülleri

Ova je spužva slatkovodna i spada u skupinu *Demospongiae*. Prilično je rasprostranjena svugdje u svijetu, a uz još nekoliko spužvi predstavlja izvrstan modelni organizam za proučavanje evolucije, genske regulacije, razvoja i fiziologije ove skupine. Jedna od glavnih značajki ove spužve je da ona tijekom zimskog razdoblja stvara tisuće gemula koje sadržavaju matične stanice (arheocite). Te se stanice skupljaju i omogućavaju laku diferencijaciju u odrasle jedinke (Rivera *i sur.*, 2011).

1.3 Strojno učenje i klasteriranje

Podaci dobiveni nakon sekvenciranja nazivaju se sirovima i često ih nije pametno koristiti za sklapanje genoma, što je u slučaju spužava neophodno obzirom da još uvijek nedostaje mnogo informacija o ovoj skupini. Problem kod genoma spužava je visoka razina kontaminacije zbog simbioze s brojnim bakterijama (Thacker i Freeman, 2012), pa je prilično teško razdvojiti moguće sljedove spužava i kontaminirajuće sljedove. Metode poput pretraživanja baza metodom lokalnog sravnjivanja (*engl*. *Basic Local Alignment Search Tool, BLAST*) rješavanju ovog problema još uvijek nisu urodile plodom zbog toga jer su sljedovi premalo slični odnosno previše divergentni (Chan *i sur.*, 2013), te zbog velike količine podataka koju današnje tehnologije sekvenciranja omogućuju (Liu *i sur.*, 2008).

Tehnike strojnog učenja dijele se na nadzirane i nenadzirane. U slučaju nadziranih tehnika podrazumijeva se da za svaku opaženu varijablu i = 1,....,*n* imamo zavisnu varijablu x_i te pripadajuću nezavisnu varijablu y_i jer želimo model koji povezuje odgovor (nezavisnu varijablu) sa zavisnim varijablama kako bismo točno predvidjeli odgovor za buduće opažene varijable, odnosno zavisne varijable ili kako bismo bolje razumjeli odnos između zavisne i

nezavisne varijable. No postoje slučajevi kada za opažene varijable imamo vektor zavisnih varijabli, ali nemamo pripadajuće nezavisne varijable. Tehnike kojima ćemo analizirati ovakve skupove podataka zovu se tehnike nenadziranog strojnog učenja, jer nemamo nezavisnu varijablu koja će "nadzirati" našu analizu. Jedan od pristupa unutar tehnika nenadziranog strojnog učenja je klasteriranje. Klasteriranje je vrlo pogodno za skupove s mnogo podataka koji su preveliki i čiji su članovi međusobno prilično divergentni da bi se u njima na neki drugačiji način našle razlike (James *i sur.,* 2000).

1.3.1 Smanjenje dimenzionalnosti podataka

Kod velikih skupova podataka veliki je problem vizualizirati višedimenzionalne podatke i prikazati ih u dvije dimenzije. Postoji više pristupa, ali metode klasteriranja najprije će promijeniti (*engl. transform*) naše podatke kako bi se oni sveli na nižu dimenziju čime bi se omogućila vizualizacija (James *i sur.,* 2000).

1.3.1.1 Stoihastičko pridruživanje susjedima temeljeno na t-distribuciji

Stoihastičko pridruživanje susjedima temeljeno na t-distribuciji (*engl. t-distributed stochastic neighbor embedding*) ili t-SNE vrsta je klasteriranja koje smanjuje dimenzionalnost podataka. Najveći problem prilikom smanjenja dimenzionalnosti naših podataka je održati lokalnu, ali i globalnu strukturu naših podataka, gdje se pod globalnom strukturom podrazumijeva (prirodna) prisutnost klastera na određenim skalama. Algoritam t-SNE prvo radi analize parova u višedimenzionalnom objektu, primjerice u matrici, i pridružuje distribuciju vjerojatnosti svakom paru na način da točke koje su međusobno slične imaju veliku vjerojatnost da će međusobno biti susjedi u klasteru, dok manje slične točke imaju vrlo nisku vjerojatnost da će se naći u istom klasteru. Lokacije točaka na mapi određuju se minimizacijom udaljenosti. Rezultat je mapa (ili matrica) koja pokazuje sličnosti u višedimenzionalnom skupu podataka (Van Der Maaten i Hinton, 2008).

1.3.2 Metode koje ne uključuju sravnjenje temeljene na učestalosti ponavljanja riječi

Najveća prednost metoda koje ne uključuju sravnjenje (*engl. alignment-free methods*) je što koristeći strojno učenje skraćujemo vrijeme potrebno za obradu podataka te omogućujemo klasteriranje unutar velikog skupa podataka koji su međusobno prilično divergentni. Unutar

ovih metoda nalaze se metode temeljene na učestalosti ponavljanja riječi u slijedu (*engl. methods based on k-mer/word frequency*).

1.3.2.1 Divergencija po Jensenu i Shannonu

Uočeno je da postoji uzorak oligonukleotida koji je specifičan za vrstu, odnosno neka vrsta potpisa koja omogućuje razlikovanje prokariota od eukariota (većinom je to $k = 4$, tetranukleotidi) (Gori *i sur.*, 2011). Računanjem učestalosti ponavljanja svih mogućih tetramera kroz svaki od naših sljedova, te pravilnim klasteriranjem, moguće je uočiti specifične klastere. Matrice udaljenosti dobivaju se, između ostalog, kao rezultat uspoređivanja dvije distribucije vjerojatnosti. Ovakav način računanja matrica udaljenosti naziva se divergencija po Jensenu i Shannonu (*D*). Divergencija po Jensenu i Shannonu kvantificira razliku između dviju ili više distribucija vjerojatnosti, što se koristi za usporedbu sastava različitih sljedova. Postoje tri razloga zašto je baš ova distribucija prikladna kao mjera razlike divergencija između distribucija vjerojatnosti: (i) *D* je povezana s ostalim funkcionalima teorije informacija, kao što je to entropija po Shannonu ili divergencija po Kullbacku (na kojima se divergencija po Jensenu i Shannonu temelji), s kojima dijeli matematička svojstva kao i intuitivnu interpretabilnost, (ii) *D* se može poopćiti tako da se može koristiti i za usporedbu više od dvije distribucije, te (iii) omogućuje tzv. težinsku usporedbu distribucija (*engl. weighted distribution compare*), gdje svakom podslijedu (*engl. subsequence*) pridaje određeni član (težina) kako bi se ujednačila njihova raspodjela duljina, odnosno umanjio utjecaj razlike u duljinama iz kojih se računa distribucija vjerojatnosti (Grosse *i sur.*, 2002).

Prema Formuli 1.

$$D\left(p^{(1)}, p^{(2)}\right) = H\left[\pi^{(1)} \times p^{(1)} + \pi^{(2)} \times p^{(2)}\right] - \left(\pi^{(1)} \times H\left[p^{(1)}\right] + \pi^{(1)} \times H\left[p^{(2)}\right]\right)$$

definirana je divergencija po Jensenu i Shannonu, gdje su $p^{(1)}$ i $p^{(2)}$ dvije distribucije, $\pi^{(1)}$ i $\pi^{(2)}$ težine za svaku od distribucija, a *H* je entropija po Shannonu.

Kao što je već navedeno, divergencija po Jensenu i Shannonu temelji se na entropiji po Shannonu koja govori o informativnosti nekog događaja, gdje je informacija negativni logaritam distribucije vjerojatnosti. Općenito, entropija je mjera nepredvidljivosti sadržaja informacije, a ako je neki događaj vjerojatniji od drugog tada on nosi manje informacije. Tako su događaji koji su rjeđi više informativni. Mjerna jedinica informativnosti je *shannon* ili

uobičajenije - bit. Potrebno nam je $log_2(n)$ bita kako bismo opisali varijablu koja poprima jednu od *n* vrijednosti ako je *n* potencija broja 2.

$$H = -\sum p \times log_2\left(\frac{1}{p}\right)$$

Formula 2. prikazuje entropiju po Shannonu, gdje je p je distribucija vjerojatnosti.

Divergencija po Jensenu i Shannonu ima veliku primjenu, primjerice za mjerenje udaljenosti nasumičnih grafova, za testiranje točkovne procjene određenih statistika ili za kvantificiranje kompleksne heterogenosti sljedova DNA (Grosse *i sur.*, 2002).

1.4 Metode sklapanja genoma

Metode sklapanja genoma koriste se za rekonstrukciju genoma korištenjem kraćih očitanih sljedova DNA dobivenih sekvenciranjem. Kraći očitani sljedovi pokušavaju se preklopiti, a ti preklopljeni dijelovi složiti u neprekinuti slijed ili neprekinuti niz (*engl. contig*).

Osnovne skupine metoda sklapanja genoma su sklapanje genoma *de novo* i sklapanje na temelju referentnog genoma. Metode sklapanja genoma *de novo* kompleksnije su samim time što nemamo referentni genom, a započinje se slaganjem preklapajućih očitanih sljedova DNA u neprekinute nizove (*engl. contigs*), a zatim se neprekinuti nizovi točno uređuju u prekinute sljedove (*engl. scaffolds*). Metode koje koriste referentni genom u suštini su mapiranje dobivenih očitanih sljedova na referentni genom.

Bez obzira na veliku količinu podataka i raznolikost genoma koji se mogu pronaći današnjim bazama podataka, i dalje postoji velik broj organizama za koje ne postoji referentni genom tako da se vrlo često pribjegava metodama *de novo.* Obzirom na kompleksnost zadatka, traže se očitani sljedovi što veće točnosti i duljine. Tako postoje samostalni sljedovi (*engl. single reads*) i upareni očitani sljedovi (*engl. paired end reads*). Upareni očitani sljedovi danas se sve više koriste zbog toga što smanjuju mogućnost greške te podižu kvalitetu sklapanja – ako znamo udaljenost između uparenih sljedova, olakšavamo slaganje nepreklapajućih sljedova u preklapajuće.

Postoje tri skupine metoda sklapanja genoma *de novo* obzirom na algoritam koji koriste: metode temeljene na „pohlepnom“ algoritmu (*engl. greedy algorithm*), zatim metoda

preklapanje-raspored-konzenzus (*engl. overlap layout consensus*) te metode po De Brujinu odnosno De Brujinov graf (Miller *i sur.*, 2010; Schatz *i sur.*, 2010; Li *i sur.*, 2012).

1.4.1 Metoda preklapanje-raspored-konsenzus

Metoda preklapanje-raspored-konzenzus (PRK) uzima naše očitane sljedove i traži preklapanja među njima na način da se koristi metoda parnih sravnjenja (*engl. pairwise alignment*). Na temelju preklapanja izgradi se graf na način da je slijed prikazan kao čvor (*engl. node*) na grafu, a ukoliko postoji preklapanje između dva slijeda, njihovi se čvorovi spoje stazom (*engl. path*), s time da broj čvorova uvijek odgovara broju sljedova. Bit ovog grafa je da je svaki čvor posjećen samo jednom, a takva staza se zove Hamiltonianova staza. Hamiltonianova staza smatra se NP teškim računalnim problemom, a za takve probleme još uvijek nije pronađeno rješenje. Zadnji korak je na temelju višestrukih sravnjenja pronaći konsenzusne sekvence. Algoritmi koji koriste ovu metodu često imaju duže vrijeme izvršavanja, ali su točniji (Pop, 2009).

1.4.2 Metoda po de Brujinu

Ova se metoda, za razliku od metode preklapanje-raspored-konzenzus, temelji na Eulerovoj stazi u kojoj je put kroz graf izgrađen tako da se svakom stazom prođe točno jednom. Također, čvorovi na grafu nisu očitani sljedovi, već se očitani sljedovi dijele na manje komadiće određene veličine, k-mere. Obzirom da u ovom slučaju nemamo NP težak problem, već Eulerovu stazu, metoda je komputacijski manje intenzivna (Pevzner *i sur.*, 2001).

2 Ciljevi rada

Iako su metode pripreme uzoraka veoma napredovale, kao i stupnjevi sterilnosti današnjih laboratorija, i dalje se susrećemo s problemom kontaminacija u uzorcima izoliranih nukleinskih kiselina. Tome ne pomaže ni nezanemariv postotak grešaka nastalih prilikom sekvenciranja i/ili prekratki sljedovi dobiveni modernim tehnologijama kao što su metoda sekvenciranja sintezom prema kalupu i metoda sekvenciranja nanoporama. Proučavajući skupinu Porifera uočeno je da bile one mogle biti prekretnica u evoluciji Metazoa te su vrlo zanimljive svijetu znanstvenika. No, s obzirom na samo jedan sklopljeni genom, ovo je područje koje se tek istražuje i još je uvijek prerano za konstruktivne zaključke.

Sklapanje genoma vrlo je kompleksan i osjetljiv zadatak, a kako bismo ga odradili što točnije, potrebna je kvalitetna predobrada podataka. Koristeći računalne metode statistike i strojnog učenja pokušat ću što efikasnije detektirati kontaminirajuće sljedove koje su se našle u uzorku DNA spužava prilikom izolacije DNA. Time bi se napravio veliki korak prema boljem i kvalitetnijem sklapanju genome te boljoj interpretaciji rezultata.

3 Metode i materijali

3.1 Pregled sekvenciranih knjižnica

Koristila sam dvije knjižnice sekvencirane na uređajima Illumina MiSeq, a obje su knjižnice kratkih udaljenosti (*engl. paired end*).

Tablica 1 Statistike podataka dobivenih sekvenciranjem genomske DNA spužve *Ephydatia mülleri* dobivenih na uređajima Illumina MiSeq.

Knjižnica	Broj očitanih fragmenata	Prosječna duljina očitanih fragmenata (nk)	Ukupna duljina (nk)
Illumina MiSeq 1	30 674 946	251	7 699 411 446
Illumina MiSeq 2	45 239 016	251	11 354 993 016

Također, koristila sam i tri knjižnice dobivene sekvenciranjem na uređaju The Oxford Nanopore Technologies MinION (ONT) (Tablica 2.).

ONT Metrichor je program koji detektira baze očitane na uređaju ONT (*engl. base calling*), a očitane sljedove dijeli u dvije mape: uspješna (*engl. pass*) i neuspješna (*engl. fail*). 2D sljedovi sa srednjom kvalitetom većom ili jednakom od 9 pripadaju uspješnoj mapi, a 2D sljedovi sa srednjom kvalitetom manjom od 9, 1D sljedovi kod kojih nije detektiran 2D slijed te svi sljedovi s neuspješnom detekcijom 1D sljeda neuspješnoj mapi.

Tablica 2 Statistike podataka dobivenih sekvenciranjem genomske DNA spužve *Ephydatia mülleri* dobivenih na uređaju The Oxford Nanopore Technologies MinION.

Knjižnica	Vrsta slijeda	Broj očitanih fragmenata	Prosječna duljina očitanih fragmenata (nk)	Ukupna duljina (nk)
Oxford Nanopore 1	Uspješno pročitani 2D sljedovi	7 507	744.22	5 586 849
	Neuspješno pročitani 2D sljedovi	9 996	417.87	4 177 000
Oxford Nanopore 2	Uspješno pročitani 2D sljedovi	11 471	939.19	10 773 453
	Neuspješno pročitani 2D sljedovi	57 615	483.60	27 862 872
Oxford Nanopore 3	Uspješno pročitani 2D sljedovi	3 676	740.65	2 722 635
	Neuspješno pročitani 2D sljedovi	22 382	453.43	10 148 596

3.2 Predobrada sljedova dobivenih sekvenciranjem

3.2.1 Provjera kvalitete sljedova dobivenih sekvenciranjem

Prije nego što se uopće krene u predobradu sljedova, potrebno je svakoj knjižnici provjeriti kvalitetu kako bismo znali na što se trebamo usredotočiti u prvim koracima predobrade, a i nakon predobrade. U tu svrhu koristila sam FastQC program za vizualizaciju kvalitete sljedova dobivenih metodama Illumina i Oxford Nanopore.

(www.bioinformatics.babraham.ac.uk/projects/fastqc/)

3.2.2 Predobrada sljedova dobivenih sekvenciranjem metodom Illumina

Koristila sam programe otvorenog koda iz programskog paketa BBMap/BBTools za prvotnu predobradu i analizu sljedova sekvenciranih metodom Illumina.

(https://sourceforge.net/projects/bbmap/)

Za uklanjanje adaptera i sljedova s nedovoljnom pouzdanošću koristila sam BBDuk. Kao referenca pri uklanjanju adaptera korišten je skup adaptera koji se nalazi u programskom paketu BBMap/BBTools.

Knjižnici Illumina MiSeq 1 uklonjeni su sljedovi koji su imali preklapanje s referentnim adapterima duljine 23 nukleotida, te je primijenjeno dodatno odstranjivanje adaptera na temelju preklapanja dvaju uparenih sljedova i jednakomjerno odstranjivanje adaptera.

Knjižnici Illumina MiSeq 2 također su uklonjeni sljedovi koji su imali preklapanje s referentnim adapterima duljine 23 nukleotida, a potrebno je bilo i odstraniti skup baza na krajevima sljedova čija je ukupna kvaliteta bila manja od 10 po Phredovoj skali. Također je primijenjeno i dodatno odstranjivanje adaptera na temelju preklapanja dvaju uparenih sljedova i jednakomjerno odstranjivanje adaptera.

Korišten je BBMerge kako bi se preklapajući parovi očitanih sljedova Illumina MiSeq 1 i 2 knjižnica spojili u jedan slijed.

Na kraju svakog koraka sljedovi su vizualizirani FastQC programom.

3.2.3 Predobrada sljedova dobivenih sekvenciranjem metodom Oxford Nanopore

Očitane sljedove nakon sekvenciranja uređajem The Oxford Nanopore Technologies MinION dobivamo u datoteci oblika FAST5 koja je nepregledna i ostali programi iz područja računalne genomike ne podržavaju ovaj format. FAST5 je jedna od inačica hijerarhijskog podatkovnog formata (*engl*. *Hierarchical Data Format, HDF5*) i omogućuje pohranjivanje i organizaciju velike količine podataka.

Kako bih promijenila oblik u kojem su pohranjeni sljedovi, koristila sam programski jezik R koji je otvorenog koda. R je funkcionalan jezik i sve se više koristi upravo zbog mnogo dostupnih paketa sa širokom primjenom u svijetu statistike, strojnog učenja i računalne

genomike, a broj područja primjenjivosti sve je veći. Paket poRe namijenjen je vizualizaciji, manipulaciji, sažimanju i organizaciji sljedova dobivenih sekvenciranjem na uređaju ONT (Watson i *sur.*, 2015). Ovaj paket koristila sam za dobivanje 2D sljedova u obliku FASTQ iz postojećeg oblika FAST5. Oblik FASTQ uobičajen je oblik za manipulaciju i obradu sljedova dobivenih sekvenciranjem, posebice zato što sadrži kvalitetu za svaki očitani nukleotid.

Koristila sam Oxford Nanopore 2D sljedove koji prelaze prag kvalitete koji iznosi 9, odnosno uspješno pročitane 2D sljedove te 2D sljedove koji ne prelaze prag zadani kvalitete, odnosno neuspješno pročitane 2D sljedove, no samo one koji su duži od 2 000 nukleotida.

S obzirom na razinu greške koja je prilikom sekvenciranja na ONT uređaju još uvijek previsoka - oko 85%, potrebno je ove sljedove ispraviti (Jain *i sur.*, 2015). Za ispravljanje sljedova koristila sam LoRDEC, program koji je prvotno namijenjen za korekciju sljedova dobivenih tehnologijama sekvenciranja Pacific Biosciences, koje također proizvode duge sljedove, no prilagođen je i za sljedove dobivene sekvenciranjem na uređaju ONT. Brz je, zauzima manje memorije nego slični programi te omogućuje ispravljanje pogrešaka s efikasnošću i do 99 % (Salmela i Rivals, 2014).

LoRDEC ima hibridni pristup, koristi kratke očitane sljedove visoke kvalitete, pa su tako knjižnice dobivene sekvenciranjem metodom Illumina MiSeq odlične upravo za tu svrhu. Sljedove svih triju knjižnica dobivenih sekvenciranjem na uređaju ONT ispravila sam koristeći sljedove knjižnica Illumina MiSeq 1 i 2.

(http://atgc.lirmm.fr/lordec/)

Sljedovi su na kraju svakog koraka vizualizirani FastQC programom.

3.3 Tehnike nenadziranog strojnog učenja

3.3.1 Računanje učestalosti tetranukleotida

Za računanje učestalosti tetranukleotida koristila sam paket *Biostrings* dostupan unutar skupa programa Bioconductor u programskom jeziku R. U njemu se nalaze funkcije koje omogućuju računanje učestalosti svakog od 256 mogućih tetranukleotida u svakom očitanom DNA sljedu. Ovaj sam paket koristila i za samo učitavanje sljedova DNA unutar sučelja programskog jezika R kao i za većinu manipulacija datotekama FASTA koje su bile potrebne. Zbog velike količine podataka dobivenih metodom sekvenciranja sintezom prema kalupu koji

zahtijevaju mnogo memorije i ostalih resursa, uključujući i vrijeme, sve sam analize radila na podskupu dostupnih podataka koji je bio nasumično izabran koristeći alate iz paketa BBMap/BBTools.

(https://bioconductor.org/packages/release/bioc/html/Biostrings.html)

3.3.2 Računanje matrice udaljenosti

Koristila sam divergenciju po Jensenu i Shannonu kako bih dobila matricu udaljenosti koja predstavlja različitost sljedova u odnosu na distribuciju svakog od tetramera. Matrica udaljenosti koju dobijemo kada na svoje podatke primijenimo divergenciju po Jensenu i Shannonu prikazuje kolika je sličnost između svakog para sljedova unutar naših podataka. Funkcija prolazi kroz redove matrice koji predstavljaju naše sljedove i uspoređuje jedan po jedan par.

U programskom jeziku R implementirala sam divergenciju po Jensenu i Shannonu služeći se Formulom 1.

Gledajući matricu u kojoj su redovi naši sljedovi, a stupci svaki od 256 mogućih tetramera, ova divergencija gleda distribuciju svakog slijeda u odnosu na učestalost svakog tetramera. Tako imamo onoliko različitih distribucija koliko različitih sljedova, čiju sličnost gledamo u odnosu na učestalost tetramera. Divergencija po Jensenu i Shannonu temelji se na entropiji po Shannonu kojom se opisuje informativnost određenih podataka, u našem slučaju, sljedova. Što je entropija veća, slijed je informativniji, odnosno kompleksniji.

3.3.3 Smanjenje dimenzionalnosti matrice udaljenosti

Kako bih omogućila vizualizaciju klastera na dvodimenzionalnom raspršenom grafikonu (*engl. scatter plot*), koristila sam paket *Rtsne* unutar programskog jezika R koji vrši stoihastičko pridruživanje susjedima temeljeno na t-distribuciji. Istoimena funkcija unutar ovog paketa kao argument uzima matricu udaljenosti dobivenu korištenjem koda za divergenciju po Jensenu i Shannonu i vraća matricu dimenzija n × 2, gdje je *n* broj sljedova koji su dani funkciji za divergenciju po Jensenu i Shannonu. Tako svaki slijed ima svoje dvije koordinate *x* i *y* koje su onda prikazane na raspršenom grafikonu.

3.4 Pretraživanje baze podataka

Očitane sljedove potrebno je identificirati, odnosno pretražiti baze podataka i pronaći potencijalni organizam iz kojeg određeni slijed potiče. Koristila sam algoritam BLAST za uspoređivanje nukleotidnih ili aminokiselinskih sljedova prema (lokalnoj) sličnosti. Preko njega se pretražuju baze gena ili proteina, a kao rezultat dobivamo sljedove koji s našim slijedom imaju određenu sličnost. Ovisno o tome čime želimo pretraživati bazu, postoji više inačica ovog alata (Altschul *i sur.,* 1990).

S obzirom na prirodu mojih podataka odabrala sam nukleotidni BLAST ili BLASTN, sačuvavši samo najsličniji slijed u bazi za svaki očitani slijed. Pretraživala sam bazu bez ponavljajućih sljedova (*engl. nonredundant database*) koja ima informacije o taksonomiji pronađenog slijeda.

3.5 Sklapanje neprekinutih sljedova

Algoritmi za sklapanje genoma koriste očitane sljedove kako bi stvorili neprekinute sljedove, što je prvo korak u sklapanju genoma. U neprekinuti niz sklapaju one očitane sljedove koji su međusobno najsličniji. Koristila sam program za sklapanje genoma Tadpole koji je dio paketa BBMap/BBTools. Svrha sklapanja neprekinutih sljedova programom Tadpole bila je da potvrdim ili opovrgnem sličnost sljedova u pojedinom klasteru. Izvukla sam sve sljedove ovisno o klasteru te u svakoj skupini pokušala sklopiti neprekinute nizove. Tadpole radi na način da cijepa sljedove na k-mere određene duljine čija je zadana vrijednost je $k = 31$ nk, maksimalna $k = 62$, minimalna $k = 1$, zatim im traži preklapanja i slaže k-mere u dužu nit – neprekinuti slijed. Veće vrijednosti k uzorkuju nastanak kraćih k-mera.

4 Rezultati

4.1 Predobrada sljedova dobivenih sekvenciranjem

Knjižnicama dobivenima metodama sekvenciranja sintezom prema kalupu i sekvenciranja nanoporama za početak sam provjerila kvalitetu sljedova te ih obradila na način da im se sveukupna kvaliteta poboljša kako bi bile spremne za daljnje analize.

4.1.1 Predobrada sljedova dobivenih metodom sekvenciranja sintezom prema kalupu

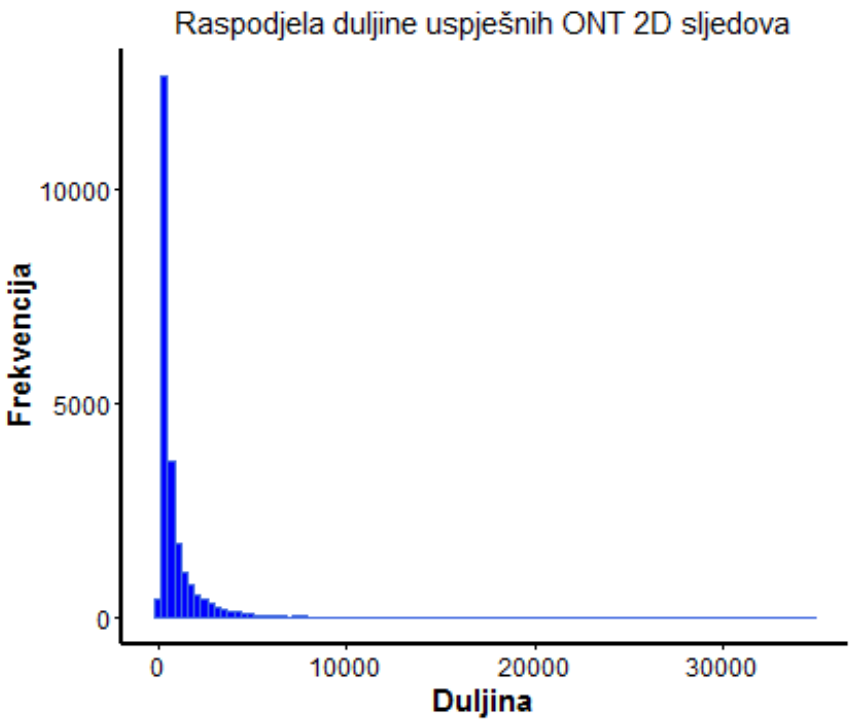
Prvotnom vizualizacijom kvalitete sljedova knjižnice Illumina MiSeq 1 uočila sam prisutnost previše učestalih sljedova koji su odgovarali sljedovima adaptera Illumina. Jedna skupina uparenih sljedova imala je odličnu kvalitetu, dok je kod druge skupine prisutan pad u kvaliteti sljedova prema krajevima. Tako su prvoj skupini uklonjeni samo adapterski sljedovi, dok je drugoj bilo potrebno ukloniti i baze s nedovoljnom pouzdanošću (Prilog, Tablica 1.).

Nakon pročišćavanja spojila sam samostalne sljedove knjižnica Illumina MiSeq 1 i 2 na temelju nedvosmislenog preklapanja kako bi se povećala duljina i kvaliteta sljedova koristeći BBMerge. Illumina MiSeq 1 ima ~30% dvosmislenih sljedova, što nije zanemarivo, i ta je informacija zadržana. Iako Illumina MiSeq ima nešto manje dvosmislenih sljedova (~20%), i ti su sljedovi zadržani (Prilog, Tablica 2.).

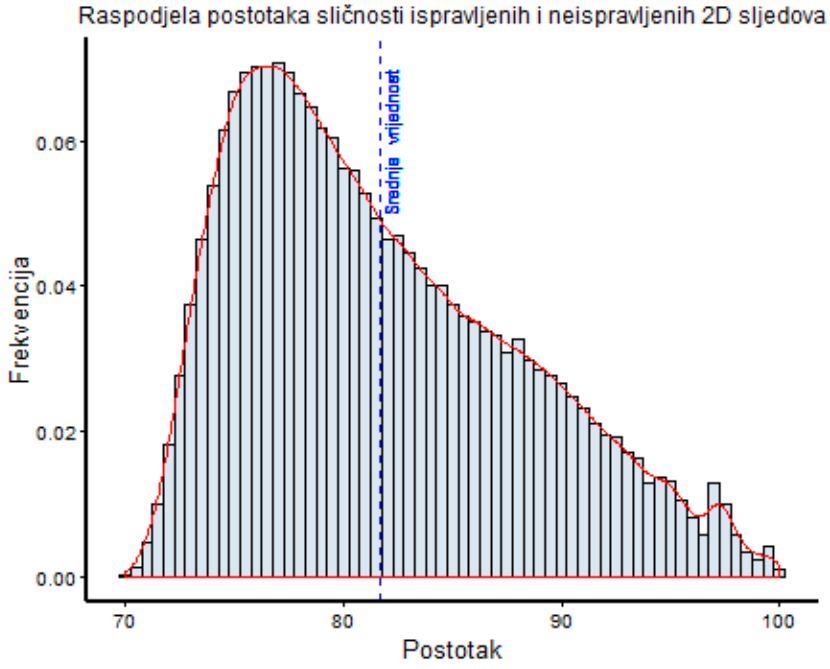
4.1.2 Predobrada sljedova dobivenih metodom sekvenciranja nanoporama

Sukladno onome što metoda Oxford Nanopore obećava, a to su dugački očitani sljedovi, najduži uspješno pročitani 2D sljedovi ONT knjižnica imaju redom duljinu oko 34 kb, 13 kb te oko 14 kb, a srednje vrijednosti duljina mogu se vidjeti u Tablici 2. Problem kod uspješno pročitanih 2D sljedova jest njihov broj, koji je dosta manji u odnosu na broj 2D sljedova koji ne prelaze zadani prag kvalitete (Tablica 2.), pa sam koristila i neuspješno pročitane 2D sljedove, ali samo one duže od 2 kb. Očitane sljedove svih triju Oxford Nanopore knjižnica spremila sam u jednu datoteku FASTA, a ukupan raspon duljina može se vidjeti na Slici 4.

Sljedovi su zatim ispravljeni programom LoRDEC koristeći obje knjižnice Illumina MiSeq za ispravljanje. Nakon ispravljanja ukupna duljina smanjila se za oko 1 %, a sličnost sljedova prije i nakon ispravljanja je oko 81 % (Slika 5.)



Slika 4 Raspodjela duljine uspješno očitanih 2D ONT sljedova prije ispravljanja. Najveća duljina je 34 kb.



Slika 5 Raspodjela postotaka sličnosti nakon ispravljanja ONT sljedova s prikazom krivulje gustoće te srednje vrijednosti postotka sličnosti između ispravljenih i neispravljenih sljedova. Srednja vrijednost sličnosti sljedova je oko 81 %.

4.2 Primjena metoda nenadziranog strojnog učenja

4.2.1 Učestalost tetranukleotida

S obzirom na veliku količinu podataka unutar knjižnica Illumina MiSeq 1 i 2, svakoj sam knjižnici u programskom jeziku R napravila nasumičan podskup od oko 20 000 sljedova (Tablica 3.).

S obzirom da sljedova Oxford Nanopore knjižnica nije bilo mnogo, uzela sam sve 2D sljedove koji prelaze zadani prag kvalitete i 2D sljedove koji ne prelaze zadani prag kvalitete, a dulji su od 2 kb. Neuspješno pročitanih 2D sljedova dužih od 2 kb bilo je sve zajedno iz sve tri knjižnice svega ~ 2.6 % (Tablica 3.). Na kraju sam radila s podskupom od ~ 25 000 sljedova sekvenciranih metodom nanoporama, iz tri različite tehnologije.

Tablica 3 Prikaz veličina nasumičnih podskupova i postotaka u odnosu na početne veličine knjižnica Illumina MiSeq 1 i 2 te Oxford Nanopore 1, 2 i 3.

Knjižnica	Postotak sljedova		Veličina podskupa (sljedovi)
Illumina MiSeq 1	0.1 %		**20 008**
Illumina MiSeq 2	0.09 %		**24 285**
Oxford Nanopore 1, 2 i 3	Uspješno pročitani 2D sljedovi	100 %	22 654
	Neuspješno pročitani 2D sljedovi veći od 2 kb	2.64 %	2 373
	Ukupno		**25 027**

Pomoću paketa *Biostrings* dostupnog unutar skupa programa Bioconductor izračunala sam učestalosti tetranukleotida u svakom očitanom sljedu u pojedinom podskupu. Rezultat je matrica s *n* redova, gdje je *n* veličina podskupa, i 256 stupaca koji odgovaraju svakom od

mogućih tetramera. Gledajući tu matricu, distribuciju tetramera predstavlja pojedini stupac, odnosno učestalost pojedinog tetramera kroz sljedove.

4.2.2 Divergencija po Jensenu i Shannonu

Implementiranu funkciju za divergenciju po Jensenu i Shannonu koristila sam kako bih dobila novu matricu u kojoj su zabilježene udaljenosti među sljedovima na temelju razlika u distribuciji tetramera. Funkcija prolazi redovima matrice i na temelju razlike u učestalostima pojedinog tetramera bilježi udaljenosti između očitanih sljedova određenog podskupa. Tako dobivamo gornjetrokutastu matricu dimenzija n x n, gdje je *n* veličina podskupa, u kojoj su zabilježene sve međusobne udaljenosti između sljedova, s dijagonalom 0 (udaljenosti između istovjetnih sljedova). Takva je matrica argument koji predajemo funkciji za stoihastičko pridruživanje susjedima temeljeno na t-distribuciji, odnosno smanjenje dimenzionalnosti naših podataka kako bismo ih mogli vizualizirati.

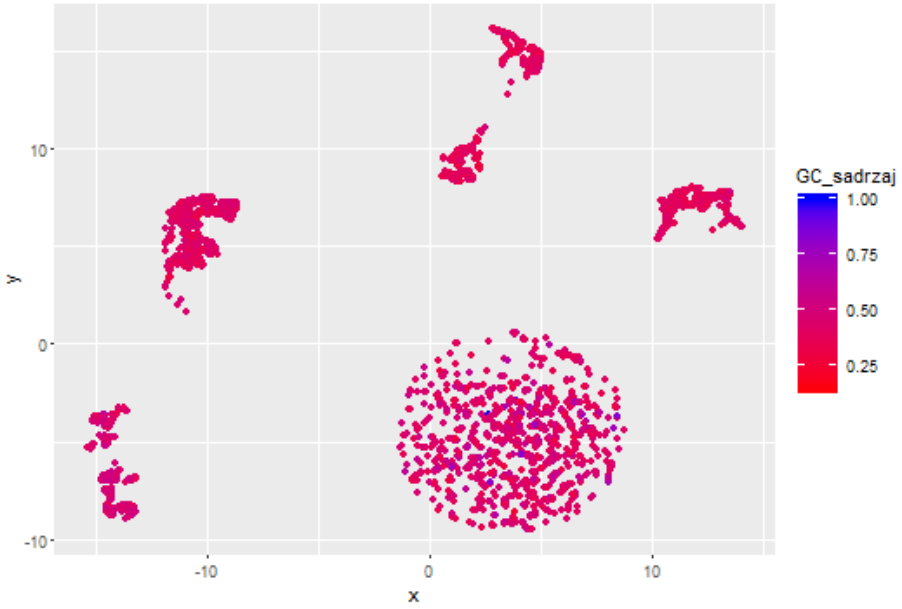
4.2.3 Klasteriranje i analiza rezultata

4.2.3.1 Stoihastičko pridruživanje susjedima temeljeno na t-distribuciji

Korištenjem funkcije *Rtsne* dostupne u istoimenom paketu u programskom jeziku R, smanjila sam dimenzionalnost svojih podataka s n x n, gdje je *n* veličina podskupa, na n x 2. Dva stupca prestavljaju podatke koji se mogu iskoristiti za prikaz klastera na raspršenom grafikonu. Sadržaj sam izračunala koristeći funkcije iz paketa *Biostrings*, a raspršeni grafikoni napravljeni su također pomoću programskog jezika R.

U analizu svake knjižnice uključeno je filtriranje reverznih komplemenata u tetramerima. Naime, tetrameri koji su međusobno reverzni komplementi funkcijom koja računa frekvencije broje se kao dvije frekvencije, iako je to zapravo jedan tetramer, pa nam treba i jedna frekvencija. Zato sam koristeći programski jezik R pronašla tetramere koji su međusobno reverzno komplementarni te njihove frekvencije prepolovila i zbrojila kako bih dobila frekvenciju za taj tetramer.

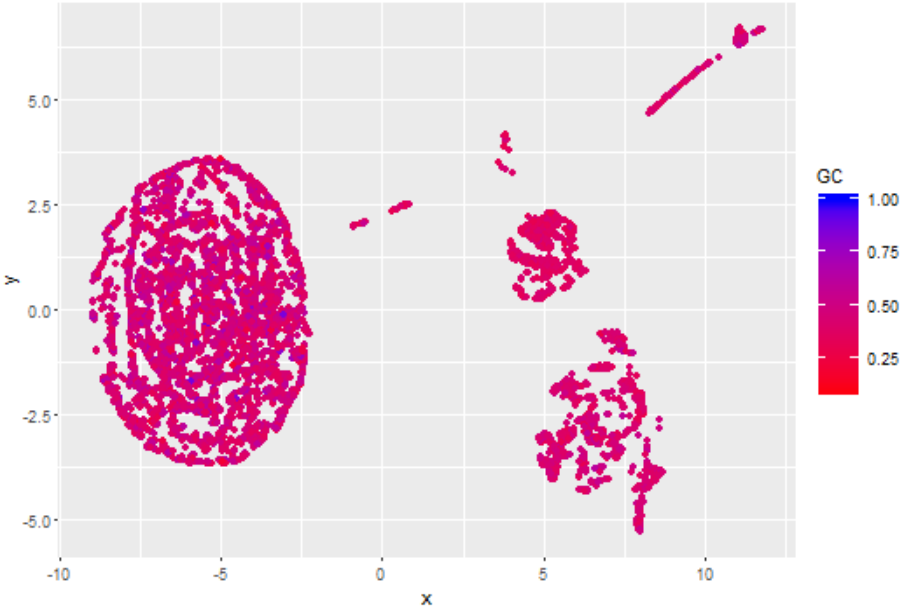
Na Slici 6. može se vidjeti nekoliko prilično oštro podijeljenih klastera knjižnice Illumina MiSeq 1. Kako bih bolje vizualizirala razlike između potencijalnih klastera, točke sam obojila prema sadržaju dinukleotida GC, i to na način da plava boja pokazuje sljedove s visokim postotkom sadržaja GC, a crvena s niskim postotkom. Raspon sadržaja prilično je širok, od manje od 25 % do blizu 100 %, tako da među sljedovima i u tom pogledu ima razlika.



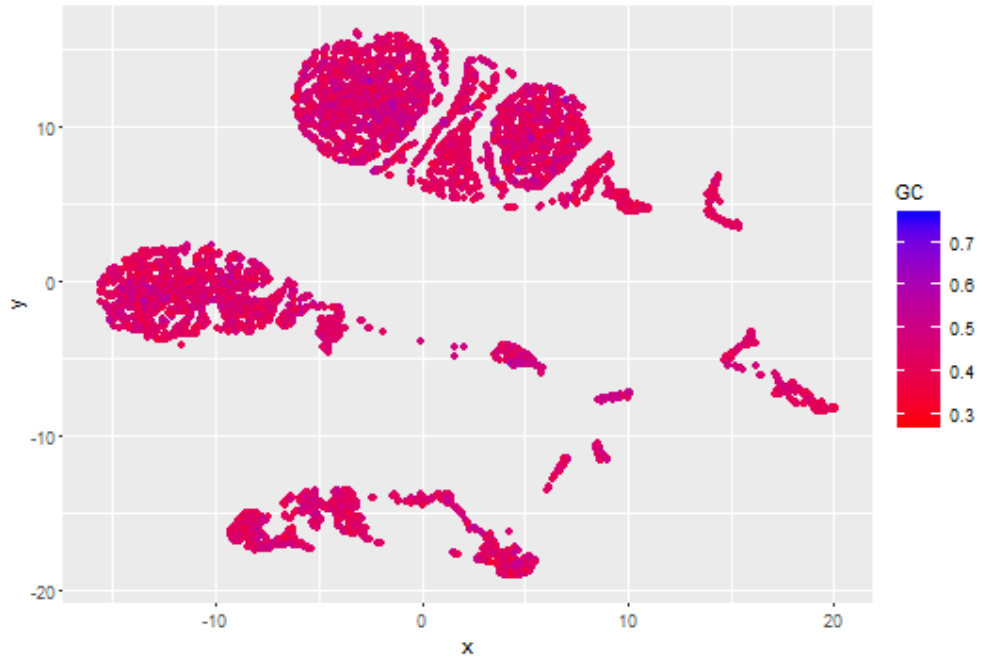
Slika 6 Prikaz klastera sljedova iz podskupa knjižnice Illumina MiSeq 1, raspršeni grafikon rezultata stoihastičkog pridruživanja susjedima temeljenog na t-distribuciji, obojano prema sadržaju dinukleotida GC.

Sljedovi iz podskupa knjižnice Illumina MiSeq 2 vizualno su se podijelili u nekoliko klastera (Slika 7.). Na raspršenom grafikonu sljedovi su obojani prema sadržaju dinukleotida GC, kao i na prethodnoj slici.

Ista analiza napravljena je i na Oxford Nanopore sljedovima, a rezultat smanjenja dimenzija prikazan je na Slici 8.



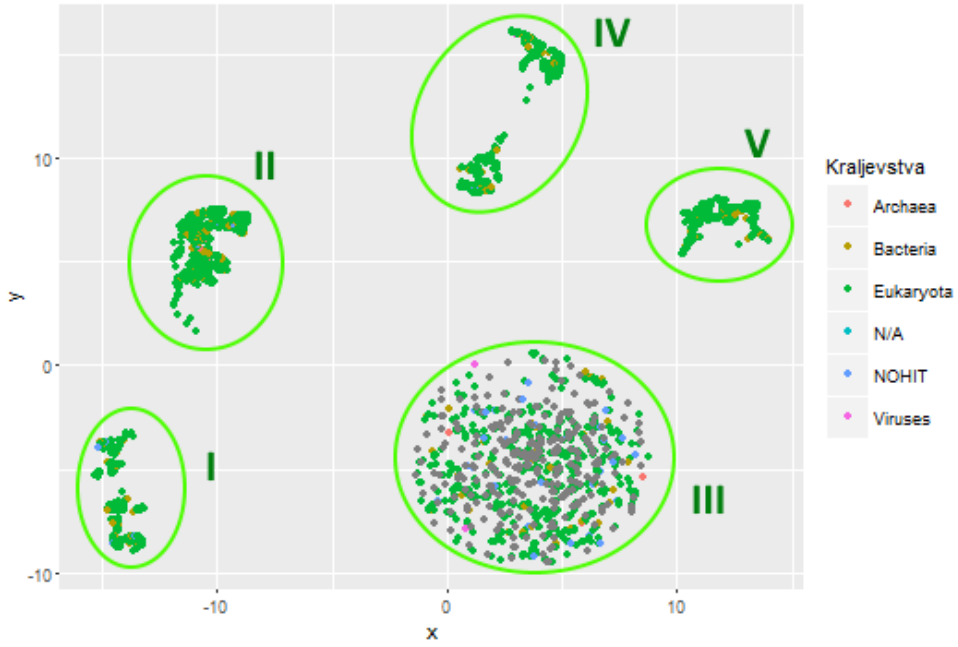
Slika 7 Prikaz klastera sljedova iz podskupa knjižnice Illumina MiSeq 2, raspršeni grafikon rezultata stoihastičkog pridruživanja susjedima temeljenog na t-distribuciji, obojano prema sadržaju dinukleotida GC.



Slika 8 Prikaz klastera sljedova iz skupa svih triju knjižnica Oxford Nanopore, raspršeni grafikon rezultata stoihastičkog pridruživanja susjedima temeljenog na t-distribuciji, obojano prema sadržaju dinukleotida GC.

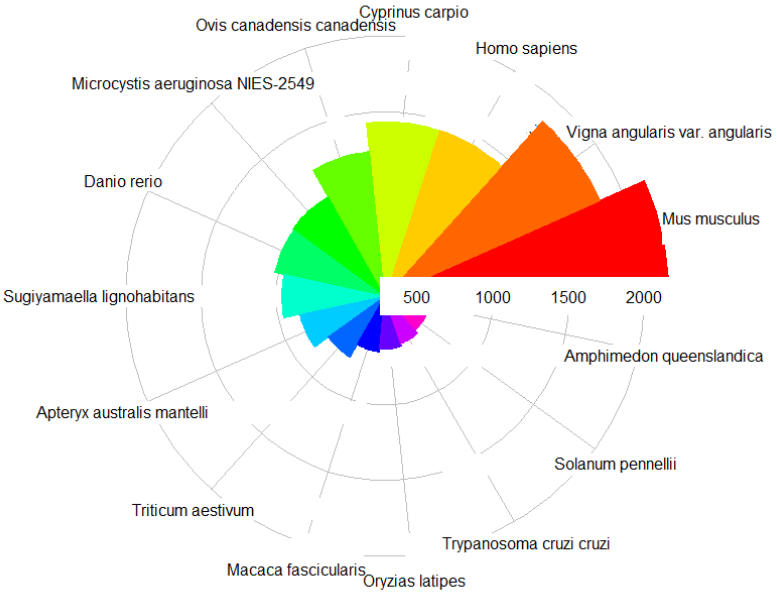
4.2.3.2 Pretraživanje baze podataka i sklapanje neprekinutih sljedova knjižnice Illumina MiSeq 1

Svaki rezultat klasteriranja, Illumina MiSeq 1, Illumina MiSeq 2 i Oxford Nanopore, dalje sam analizirala pretražujući baze podataka. Sljedovima svake knjižnice pretražila sam baze bez ponavljajućih sljedova kako bih pronašla kojim vrstama ovi sljedovi pripadaju, odnosno koji sljedovi imaju pogodak u bazi. Prvo sam analizirala knjižnicu Illumina MiSeq 1.

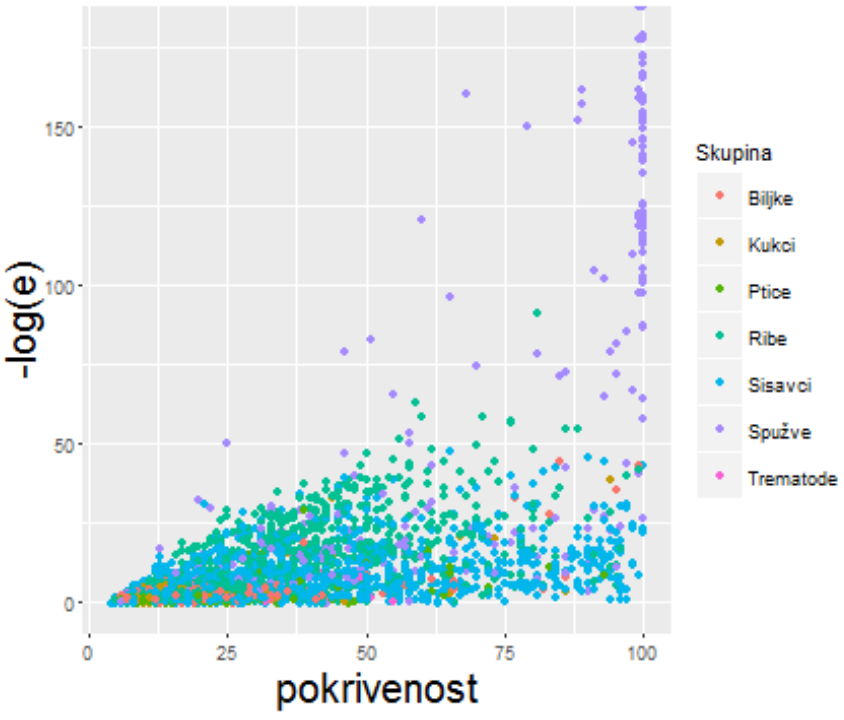


Slika 9 Prikaz raspršenog grafikona podskupa sljedova iz knjižnice Illumina MiSeq 1, rezultati stoihastičkog pridruživanja susjedima temeljenog na t-distribuciji te oznake klastera. Sljedovi su obojani prema tome kojem kraljevstvu pripadaju. "N/A" oznaku imaju sljedovi koji imaju pogodak u bazi, ali on pripada sintetskom organizmu, plazmidu, nekultiviranim bakterijama i sl. te stoga nije opisan u bazi. "NOHIT" oznaku imaju sljedovi koji nemaju nikakav pogodak u bazi. Zaokruženo je 5 klastera temeljeno na njihovoj udaljenosti. Siva boja točkica nastala je kao kombinacija boja točkica koje su zbog nedovoljne razlučivosti jedne na drugoj. Klasteri su podijeljeni na temelju njihove međusobne udaljenosti.

Na Slici 9. prikazani su klasteri podskupa sljedova knjižnice Illumina MiSeq 1, podijeljeni na temelju njihove međusobne udaljenosti obzirom da drugog uzorka za podjelu nije bilo. Obojeni su prema tome pripadaju li eukariotima, bakterijama, arhejama, virusima ili nemaju opisanog pogotka. Najveći dio sljedova pripada eukariotskim organizmima (Slika 10.). Oko 11 % pogodaka pripada mišu i čovjeku, a 1% pogodaka pripada spužvi *A. queenslandica*. Na Slici 11. prikazan je odnos negativnog logaritma e vrijednosti i kvalitete pokrivenosti pogotka u bazi nakon BLAST-a. Što je negativni logaritam e vrijednosti veći, to je sama e vrijednost manja. Niske e vrijednosti znače da je statistički manje vjerojatno da će se neki pogodak dogoditi slučajno što čini rezulat značajnijim, a više vrijednosti kvalitete prekrivanja znače bolji, kvalitetniji pogodak. Pogotci sa skupinom spužve su među najzastupljenijima, s najnižim e vrijednostima i prilično visokom kvalitetom pokrivenosti. Dosta dobru pokrivenost pokazuju i sljedovi iz skupine sisavaca, no e vrijednosti su nešto veće.



Slika 10 15 najzastupljenijih rodova knjižnice Illumina MiSeq 1 prilikom sravnjenja s nukleotidnom bazom sljedova pomoću BLAST-a. U prvih 15 najzastupljenijih nalazi se i rod *Amphimedon*, točnije spužva *Amphimedon queenslandica*.



Slika 11 Ovisnost negativnog logaritma e vrijednosti o pokrivenosti slijeda knjižnice Illumina MiSeq 1 u ovisnosti o skupini kojoj sljedovi pripradaju.

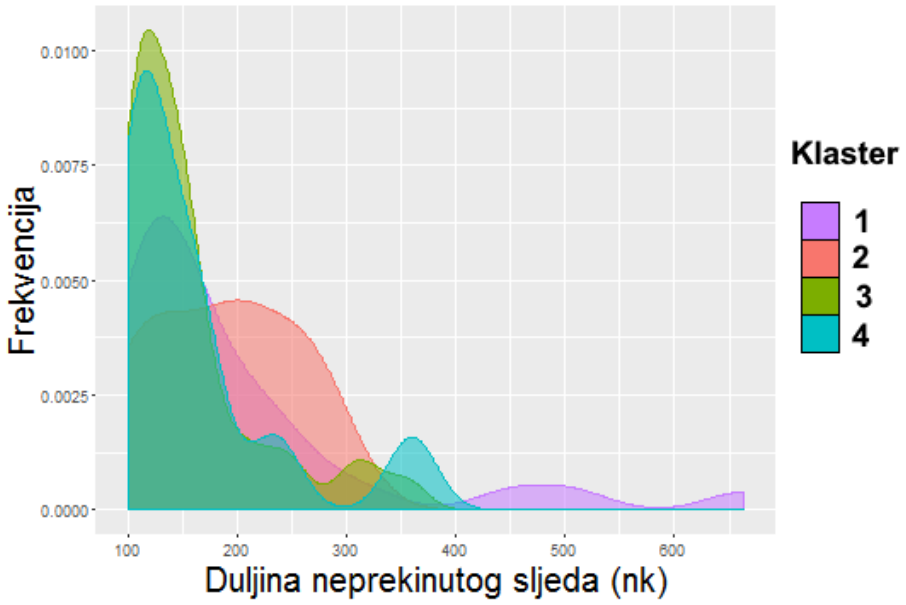
S obzirom da klasteriranje podskupa sljedova knjižnice Illumina MiSeq 1 nije bilo potpuno precizno, odnosno ne postoji gotovo nijedan klaster koji nema sljedove i eukariotskog i prokariotskog porijekla, bile su potrebne daljnje analize. Pomoću BLAST-a provjerila sam međusobnu sličnost unutar sljedova jednog klastera, s graničnom e vrijednosti 10 i granicom postotka sličnosti 0 %.

Kroz klastere nema većih odstupanja od srednje vrijednosti postotka sličnosti, dok su i e vrijednosti prilično niske (Tablica 4.).

Tablica 4 Prikaz ukupnog broja sljedova u svakom klasteru knjižnice Illumina MiSeq 1, srednjih vrijednosti postotka identičnosti (PI), srednjih vrijednosti e vrijednosti, srednjih vrijednosti *bitscorea* i srednjih vrijednosti kvaliteta prekrivenosti nakon BLAST-a.

Klaster	**Broj sljedova u klasteru**	**Srednja vrijednost PI (%)**	**Srednja vrijednost e vrijednosti**	**Srednja vrijednost *bitscorea***	**Srednja vrijednost kvalitete pokrivenosti**
1	2663	91.12	3.70×10^{-14}	296.30	78.48
2	2611	92.20	1.40×10^{-13}	481.10	72.84
3	10095	90.52	4.05×10^{-5}	206.80	35.32
4	2926	91.99	8.93×10^{-14}	444.50	51.42
5	1713	93.72	1.92×10^{-13}	614.50	50.00

Zadnji korak analize klastera bilo je sklapanje neprekinutih sljedova pomoću programa koji se inače koristi za sklapanje genoma, Tadpole. Sklopila sam neprekinute nizove uzimajući sljedove pojedinog klastera kako bih potvrdila ili opovrgnula sličnost sljedova po klasteru (Slika 12.). U programu Tadpole koristila sam zadanu duljinu k-mera, k = 31. S tom vrijednošću k u klasteru 1 sklopljeno je 30 neprekinutih sljedova, a u klasteru 3 čak 37. U klasteru 2 sklopljeno je 10 neprekinutih sljedova, a u klasteru 4 11. U klasteru 5 nije sklopljen nijedan neprekinuti slijed koristeći zadanu vrijednost k, no koristeći manju vrijednost k = 22 u tome je klasteru sklopljen jedan kratak neprekinuti niz. Detaljniji podaci o duljini neprekinutih sljedova nalaze se u Prilogu (Tablica 3.).



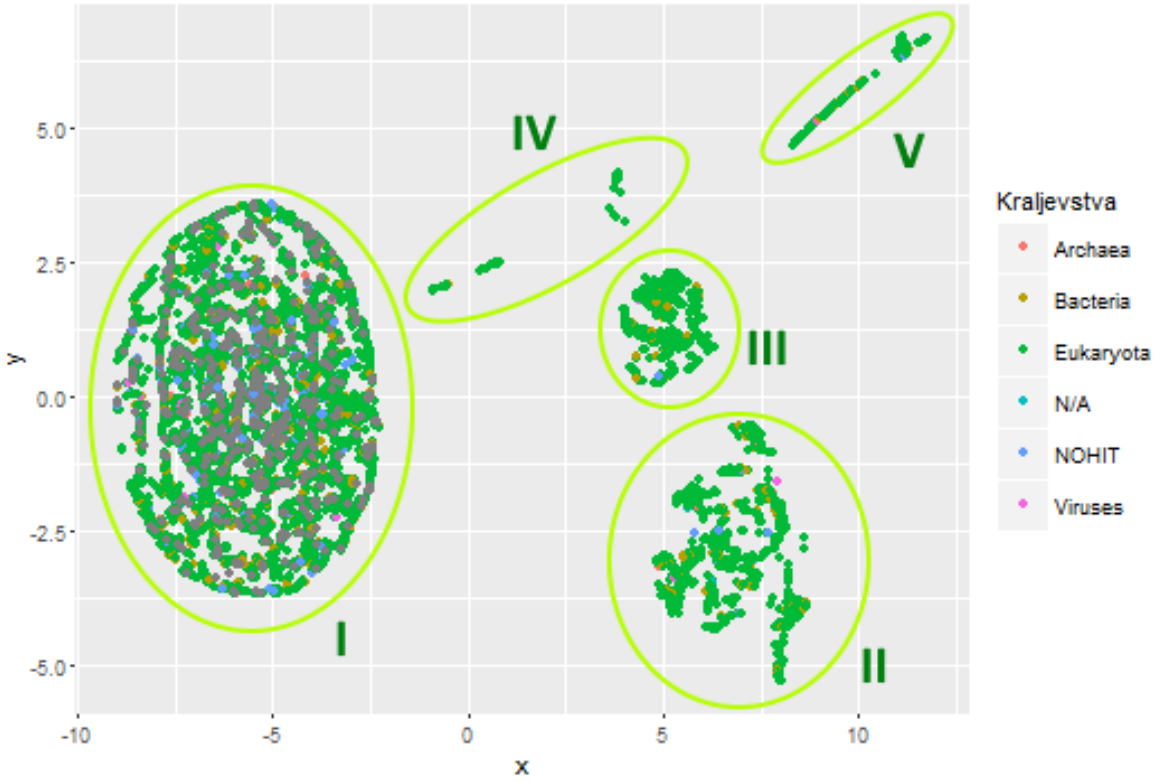
Slika 12 Raspodjela duljina neprekinutih sljedova klastera knjižnice Illumina MiSeq 1 nakon sklapanja programom Tadpole koristeći zadanu vrijednost *k* = 31. Nije prikazan klaster 5 jer u njemu nije sklopljen nijedan neprekinuti slijed koristeći zadanu vrijednost *k*. Klaster 1 ima najviše sklopljenih sljedova (88).

4.2.3.3 Pretraživanje baze podataka i sklapanje neprekinutih sljedova knjižnice Illumina MiSeq 2

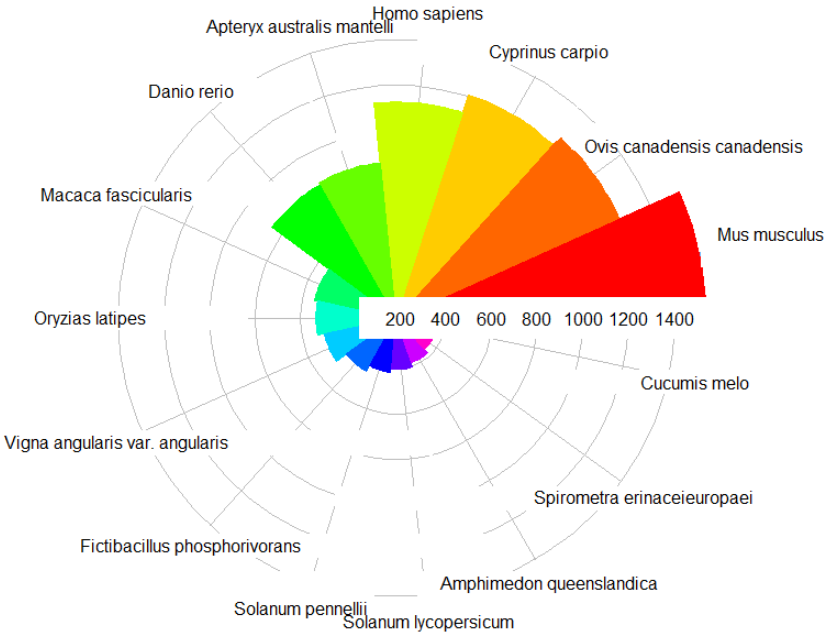
Na knjižnici Illumina MiSeq 2 ponovljene su analize napravljene na prethodnoj knjižnici. Za početak sam pretražila bazu BLAST-om u potrazi za organizmima kojima bi neki od sljedova mogli odgovarati, te napravila raspršeni grafikon t-SNE na kojem su sljedovi obojani prema kraljevstvu kojem pripadaju (Slika 13.). U ovoj knjižnici također nema posebnog uzorka na temelju kojeg bi se klasteri međusobno razlikovali pa sam ih podijelila na temelju njihove međusobne udaljenosti na grafikonu. Većina sljedova pripada eukariotima (Slika 14.), a skupine toga carstva imaju najveću pokrivenost te najnižu e vrijednost (Slika 15.). Blizu 10% ukupnih pogodaka pripada rodovima *Mus* i *Homo,* a oko 1% spužvi *A. queenslandica.* Slično kao i kod prethodne knjižnice, najkvalitetniji pogotci su s organizmima iz skupine spužvi, tamo su prisutne najveće vrijednosti negativnog logaritma e vrijednosti i najbolja kvaliteta prekrivenosti (Slika 15.).

Zatim sam izdvojila klastere te BLAST-om tražila sličnosti između sljedova u pojedinom klasteru i prikazala podatke o sličnosti ovisno o klasteru (Tablica 5.). Na kraju sam pokušala sklopiti neprekinute sljedove unutar klastera, koristeći zadanu vrijednost *k* = 31. Podaci o sklopljenim neprekinutim sljedovima nalaze se u Tablici 4. u Prilogu, a vizualan odnos

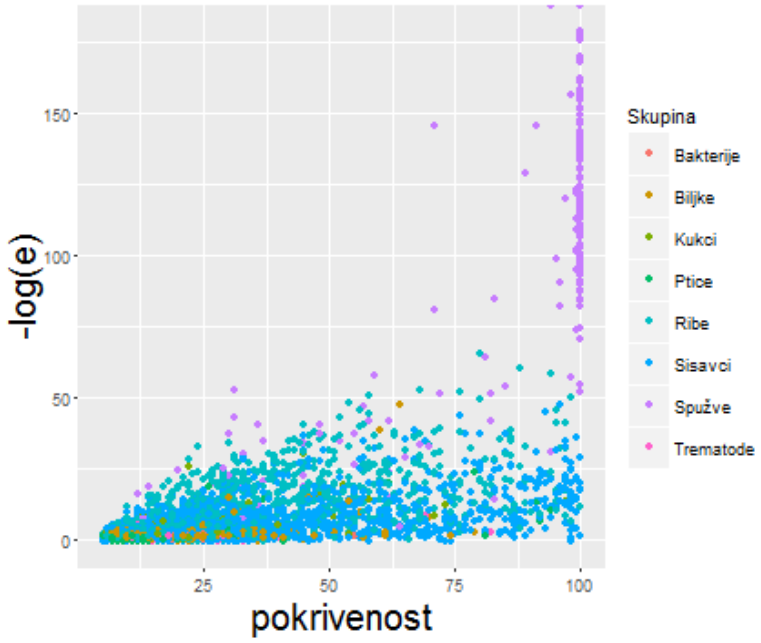
duljine neprekinutog slijeda i frekvencije prikazan je na histogramu Slike 16. Klaster 1 imao je čak 88 sklopljenih sljedova, što je ne čudi obzirom na prilično visoku kvalitetu prekrivenosti i niske e vrijednosti (Tablica 4.), kao i kod klastera 2, gdje je sklopljeno 50 neprekinutih sljedova. U klasterima 3 i 4 sklopljena su samo dva kraća neprekinuta slijeda, uzimajući u obzir niže kvalitete pokrivenosti. U klasteru 5 sklopljena su tri neprekinuta slijeda. Mijenjajući vrijednost *k* u programu Tadpole, odnosno duljinu k-mera, jedina bitna razlika uočena je kod klastera 5 gdje je sklopljeno čak 10 klastera s vrijednošću *k* većom od zadane (*k* = 60) (Prilog, Tablica 4.).



imaju sljedovi koji imaju pogodak u bazi, ali on pripada sintetskom organizmu, plazmidu, nekultiviranim bakterijama i sl. te stoga nije opisan u bazi. "NOHIT" oznaku imaju sljedovi koji nemaju nikakav pogodak u bazi. Siva boja točkica nastala je kao kombinacija boja točkica koje su zbog nedovoljne razlučivosti jedne na drugoj.



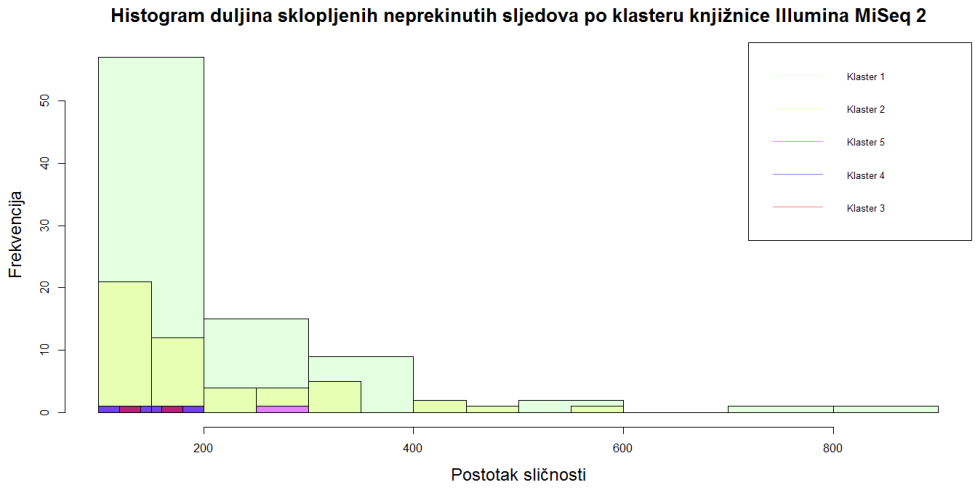
Slika 14 15 najzastupljenijih rodova knjižnice Illumina MiSeq prilikom sravnjenja s nukleotidnom bazom sljedova pomoću BLAST-a .



Slika 15 Ovisnost negativnog logaritma e vrijednosti o pokrivenosti slijeda knjižnice Illumina MiSeq 2 prema skupini kojoj sljedovi pripadaju.

Tablica 5 Prikaz ukupnog broja sljedova u svakom klasteru knjižnice Illumina MiSeq 2, srednjih vrijednosti postotaka identičnosti (PI), srednjih vrijednosti e vrijednosti (e), srednjih vrijednosti *bitscorea* te srednjih vrijednosti kvalitete prekrivenosti nakon BLAST-a.

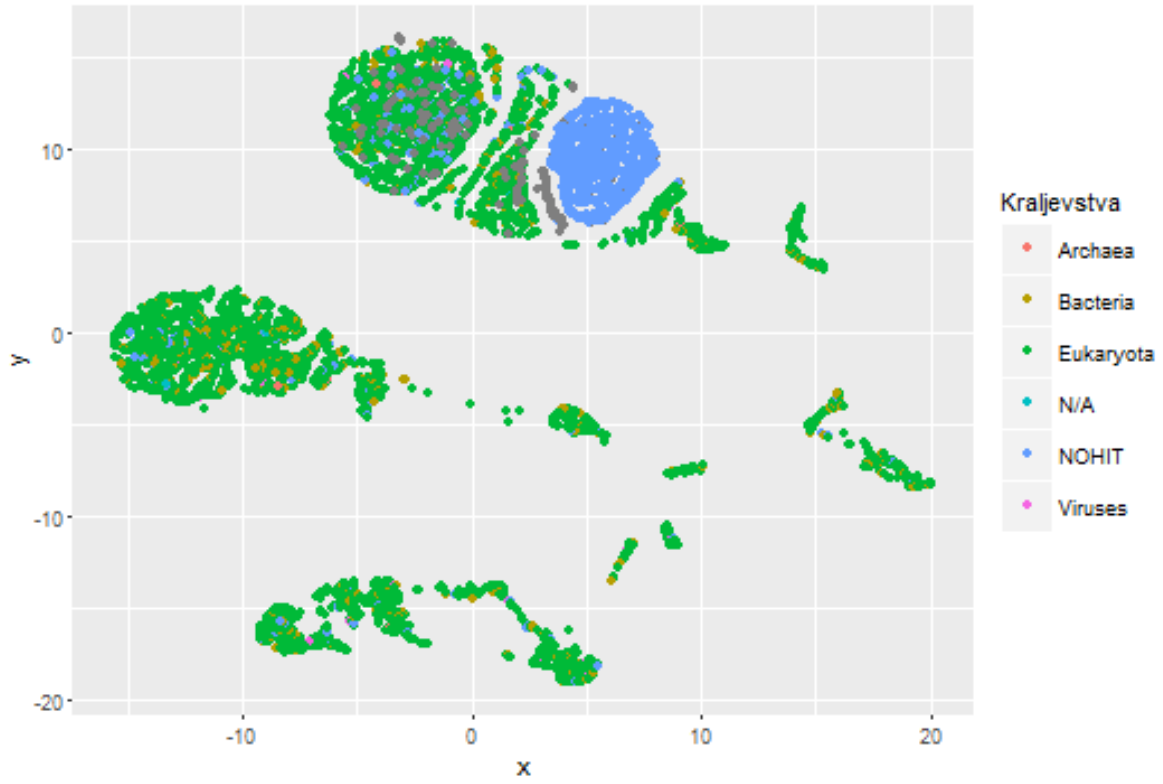
Klaster	Broj sljedova u klasteru	Srednja vrijednost PI (%)	Srednja vrijednost e	Srednja vrijednost *bitscorea*	Srednja vrijednost kvalitete prekrivenosti
1	13068	90.610	6.141×10^{-7}	198.42	85.04
2	5963	91.240	2.338×10^{-8}	195.85	81.93
3	2360	93.033	9.050×10^{-14}	205.10	50.76
4	1278	93.301	1.008×10^{-14}	202.57	52.98
5	1643	91.343	3.477×10^{-14}	211.00	86.13



Slika 16 Raspodjela duljina neprekinutih sljedova klastera knjižnice Illumina MiSeq 2 nakon sklapanja programom Tadpole koristeći zadanu duljinu k-mera, k = 31. Najviše neprekinutih sljedova sklopljeno je u klasteru 1, gdje su sklopljeni i najduži neprekinuti sljedovi.

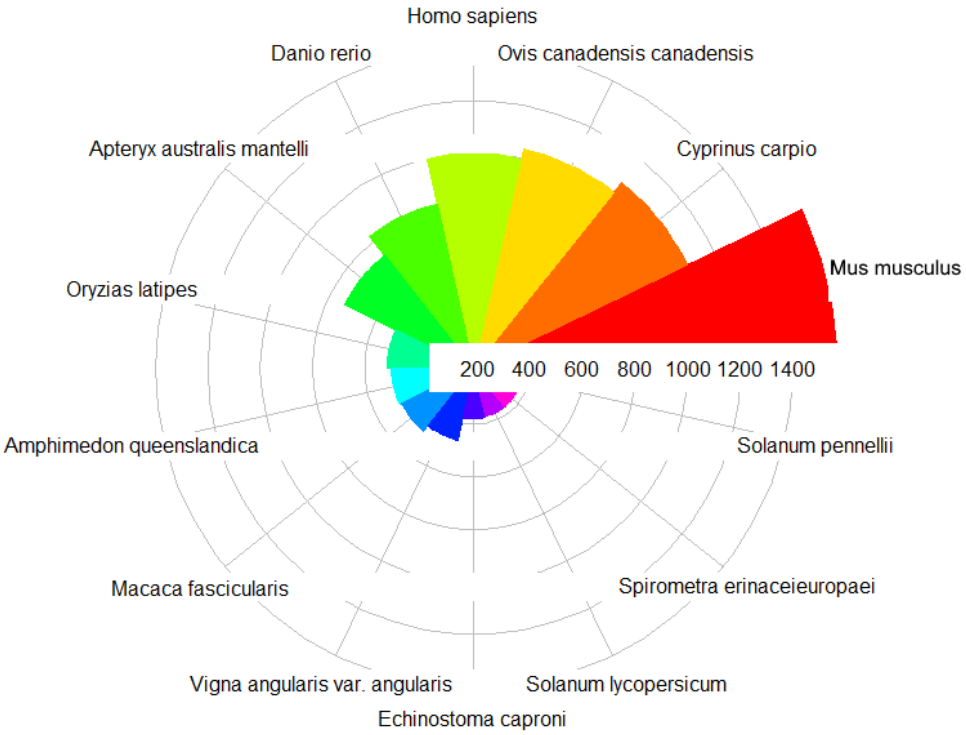
4.2.3.4 Pretraživanje baze podataka i sklapanje neprekinutih sljedova knjižnice Oxford Nanopore

Zadnja knjižnica čijom analizom sam se bavila jest Oxford Nanopore knjižnica. Prikaz klastera nalazi se na Slici 17., a rezultat je klasteriranja t-SNE gdje su sljedovi obojani ovisno o ishodu BLAST-a. U ovom slučaju nisu svi klasteri toliko jasno odvojeni kao u prethodnim knjižnicama, ali vidi se klaster koji je prilično jednolično obojan plavom bojom koja označava „NO HIT“ sljedove, odnosno sljedove za koje nije pronađen pogodak u bazi.

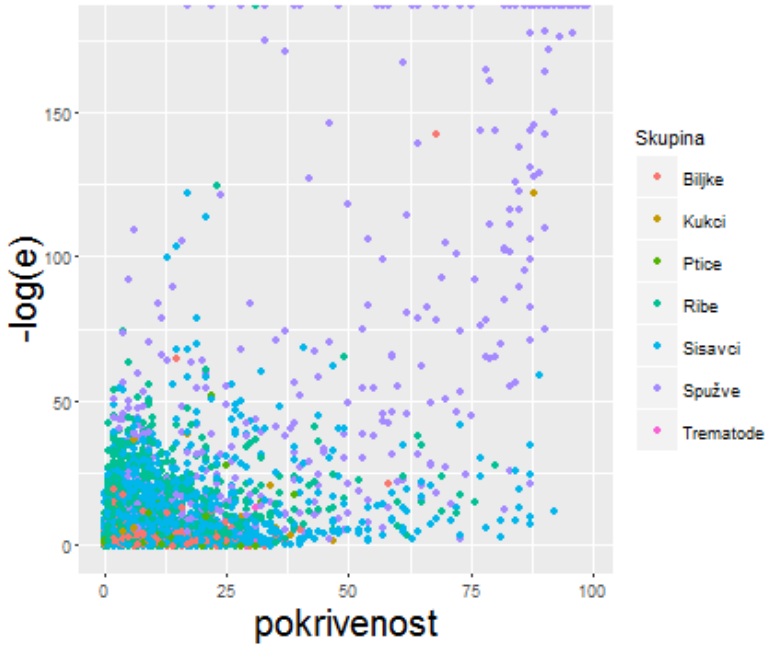


Slika 17 Prikaz sljedova knjižnice Oxford Nanopore nakon klasteriranja t-SNE, obojano prema kraljevstvu kojem pripadaju, ako su BLAST-om pronađeni pogotci za određeni slijed. "N/A" oznaku imaju sljedovi koji imaju pogodak u bazi, ali on pripada sintetskom organizmu, plazmidu, nekultiviranim bakterijama i sl. te stoga nije opisan u bazi. "NOHIT" oznaku imaju sljedovi koji nemaju nikakav pogodak u bazi. Vidi se područje pretežno obojeno plavom bojom, za razliku od prethodnih rezultata klasteriranja. Siva boja točkica nastala je kao kombinacija boja točkica koje su zbog nedovoljne razlučivosti jedne na drugoj.

I ovoj knjižnici među pogotcima u bazi prevladavaju eukariotski organizmi (Slika 18.). Grafikon odnosa negativnog logaritma e vrijednosti i kvaliete pokrivenosti prikazan je na Slici 19.



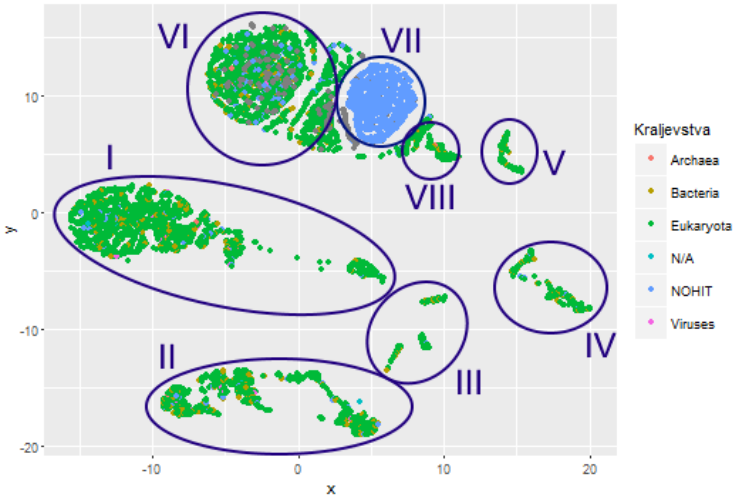
Slika 18 15 najzastupljenijih rodova prilikom sravnjenja s nukleotidnom bazom sljedova pomoću BLAST-a knjižnice Oxford Nanopore.



Slika 19 Ovisnost negativnog logaritma e vrijednosti o pokrivenosti slijeda po najzastupljenijim rodovima nakon BLAST-a knjižnice Oxford Nanopore. Primjećuje se veća raspršenost sljedova zbog odnosa negativnog logaritma e vrijednosti i pokrivenosti.

Nakon općenite analize knjižnice analizirala sam klastere zasebno. Klasteri ove knjižnice bili su raspršeniji u odnosu na prethodne no ipak je postojao dio sljedova koji su se grupirali kao „NOHIT“ i bili obojani pretežno plavom bojom (Slika 17.).

Klastere sam i ovaj put odabrala na temelju njihove međusobne udaljenosti, s tim da sam plavo obojeni klaster definirala kao zaseban (Slika 20.).



Slika 20 Prikaz podijele na klastere nakon analize t-SNE knjižnice Oxford Nanopore. Na slici vidimo osam klastera, a najzanimljiviji je klaster 7 zbog priličnog homogenog sastava sljedova („NOHIT“). "N/A" oznaku imaju sljedovi koji imaju pogodak u bazi, ali on pripada sintetskom organizmu, plazmidu, nekultiviranim bakterijama i sl. te stoga nije opisan u bazi. "NOHIT" oznaku imaju sljedovi koji nemaju nikakav pogodak u bazi.

Klasterima sam zatim tražila međusobnu sličnost BLAST-om (Tablica 6.)

Tablica 6 Prikaz ukupnog broja sljedova u svakom klasteru knjižnice Oxford Nanopore, srednjih vrijednosti postotaka identičnosti (PI), srednjih vrijednosti e vrijednosti i srednjih vrijednosti *bitscorea* te srednih vrijednosti kvalitete prekrivenosti nakon BLAST-a.

Klaster	**Broj sljedova u klasteru**	**Srednja vrijednost PI (%)**	**Srednja vrijednost e vrijednosti**	**Srednja vrijednost *bitscorea***	**Srednja vrijednost kvalitete prekrivenosti**
1	6714	86.95	3.98×10^{-6}	374.22	81.35
2	2346	78.76	4.18×10^{-12}	302.07	66.50
3	578	82.90	1.83×10^{-12}	285.08	85.44
4	1564	87.03	1.51×10^{-12}	350.85	68.77
5	782	86.65	2.44×10^{-12}	386.06	71.31
6	6317	84.51	9.84×10^{-7}	382.01	68.92
7	2959	85.77	6.87×10^{-12}	351.27	86.41
8	702	84.66	1.43×10^{-6}	454.27	72.77

Neprekinutih sljedova koje sam dobila programom Tadpole bilo je premalo da bih ih prikazala histogramom pa sam podatke o sljedovima prikazala tablicom (Tablica 7.).

Tablica 7 Podaci o neprekinutim sljedovima sklopljenim programom Tadpole koristeći sljedove iz klastera knjižnice Oxford Nanopore podijeljenih na način prikazan na Slici 20. Korištena je zadana duljina k-mera (*k* = 31).

Klaster	Najkraći sklopljeni neprekinuti slijed (nk)	Srednja duljina sklopljenog neprekinutog slijeda (nk)	Najkraći sklopljeni neprekinuti slijed (nk)	Broj sklopljenih neprekinutih sljedova
1	102.0	124.4	192.0	5
2	108.0	118.5	129.0	2
3	-	-	-	0
4	108.0	108.0	108.0	1
5	129.0	129.0	129.0	1
6	101.0	111.8	126.0	5
7	111.0	111.0	111.0	1
8	124.0	124.0	124.0	1

Iako su sljedovi u spornom klasteru 7 prilično slični i imaju visoku kvalitetu prekrivenosti ipak su dali samo jedan relativno kratak neprekinuti slijed. Prilikom sklapanja neprekinutih sljedova programom Tadpole u klasteru 7 promijenila sam vrijednost *k*, smanjila je i povećala, no nije bilo razlike. K-meri veličine od 23 do 26 nisu dali nijedan neprekinuti slijed, a duljine od 26 do 42 jedan, i to iste dužine kao i kad se koristila zadana vrijednost (*k* = 31). Mijenjajući duljinu korištenih k-mera kod ostalih klastera uočena su određena poboljšanja, iako nijedno nije znatno. U klasteru 1 povećanjem vrijednosti *k* na 34 i više dobiveno je 7 neprekinutih sljedova, što su dva neprekinuta slijeda više nego koristeći zadanu vrijednost *k* = 31. U klasteru 6 povećanjem vrijednosti *k* na 40 i više dobiveno je 7 neprekinutih sljedova, u odnosu na njih 5 dobivenih koristeći zadanu vrijednost *k.* Povećavanjem duljine k-mera na 46 i više u klasteru 8 dobila sam 2 neprekinuta slijeda. U ostalim klasterima promjena duljine k-mera nije utjecala na ishod.

5 Rasprava

U ovom sam radu koristila sljedove sekvencirane na uređaju Illumina i sljedove sekvencirane na uređaju ONT MinION kako bih usporedila koji će sljedovi dati bolje rezultate klasteriranja i omogućiti precizniji pronalazak kontaminanata. Nakon pretraživanja baza BLAST-om dobiveni su rezultati iz kojih je bilo teško iščitati koji bi sljedovi mogli biti kontaminacije zbog različith e vrijednosti i kvaliteta pokrivenosti unutar istih skupina organizama, ali su se ipak u rezultatima BLAST-a primijetile zanimljivosti. Najzastupljeniji rodovi nakon BLAST-a svih triju knjižnica jesu sisavci, ribe i ptice. BLAST podskupa sljedova iz knjižnice Illumina MiSeq 1 pokazao je da najveće negativne logaritme e vrijednosti i najveću pokrivenost pokazuju sljedovi koji pripadaju skupini spužve. Ista situacija primjećuje se i kod knjižnice Illumina MiSeq 2. Sljedovi koji prema BLAST-u pripadaju skupini sisavaca, a nalaze se u knjižnicama Illumina MiSeq 1 i 2, pokazuju niže e vrijednosti (višu statističku značajnost), ali dobre kvalitete prekrivenosti. Ti bi sljedovi mogli predstavljati jako divergentne sljedove ili kontaminacije. Također, zanimljivo je da se u 15 najzastupljenijih rodova nakon BLAST-a knjižnice Illumina MiSeq 2 nalazi rod *Fictibacillus.* Obzirom na visoke e vrijednosti i nisku prekrivenost moguće je da je porijeklo ovih sljedova od bakterije simbionta ili se radi o sljedovima u bakterijskom genomu koji su jako divergentni u odnosu na bakterijske sljedove.

U knjižnici Oxford Nanopore situacija nakon BLAST-a malo je drugačija. Sljedovi koji pripadaju skupini spužve ipak su raspršeniji u smislu e vrijednost i kvaliteta prekrivenosti, iako su to i dalje statistički najznačajniji i najkvalitetniji pogotci. Nakon BLAST-a ove knjižnice primjećuje se i puno sljedova s visokim e vrijednostima, odnosno niskim negativnim logaritmom e vrijednosti, a s malom prekrivenošću. Ovakva situacija posljedica je veće dužine sljedova dobivenih sekvenciranjem nanoporama, jer su sljedovi dobiveni metodom sekvenciranja sintezom prema kalupu prekratki da bi se pronašli pogotci u bazi s ovakvim karakteristikama. Ti bi sljedovi mogli predstavljati kontaminacije ili evolucijski očuvane domene u spužvama.

Alternativa metodama koje koriste sravnjenje može biti i klasteriranje koje bi trebalo biti pogodnije za pronalazak kontaminacija u uzorku podataka koji su jako divergentni. No, klasteriranje nije dalo željene rezultate. Mogući problem kod klasteriranja t-SNE koristeći sljedove knjižnica Illumina MiSeq 1 i 2 jest duljina tih sljedova. Laczny i suradnici (2014) koristili su sljedove dugačke između 1000 i 2000 nukleotida, za razliku od sljedova koji su

korišteni u ovom radu, čija je srednja vrijednost duljine oko 250 nk, te su uspjeli dobro odvojiti različite skupine organizama. Distribucija uzoraka korištenih u navedenom radu i u mojem radu bila je ista - nejednolika, tako da problem vjerojatno leži u duljini sljedova. U mojem slučaju organizmi eukariotskog i prokariotskog porijekla prilično su loše odijeljeni.

Gledajući knjižnice Illumina MiSeq 1 i 2, obje imaju slične probleme. Unutar klasterska sličnost ponekad nije imala veze s količinom i duljinom sklopljenih neprekinutih nizova. Postojali su klasteri s visokim postotkom sličnosti, niskim e vrijednostima, dobrom kvalitetom prekrivenosti, a s tek dva kratka sklopljena neprekinuta niza. Isto tako bilo je i obratnih situacija, slabija međusobna sličnost sljedova i slabija kvaliteta prekrivenosti, a veći broj neprekinutih sljedova. No, neke su skupine ipak opravdale visoku i statistički značajnu sličnost koji im je pripisao BLAST, dajući puno neprekinutih sljedova od kojih su neki imali duljinu do 800 pb. Mogući problem je sama tehnologija Illumine koja obećava kraće sljedove s visokom pokrivenošću u usporedbi sa Sangerovim sekvenciranjem prve generacije koje daje sljedove slabije pokrivenosti, ali ipak veće duljine. Određena istraživanja Sangerovim sekvenciranjem dobila su cjelovitije sljedove od sljedova dobivenih koristeći metode novih generacija sekvenciranja (Gnerre *i sur.*, 2010).

Divergencija po Jensenu i Shannonu koristi se na biološkim uzorcima (Merkin *i sur.*, 2012), pa smatram da je najveći problem kod mojih uzoraka bila njihova količina. Svega ~ 20 000 sljedova dobivenih metodom sekvenciranja sintezom prema kalupu očito nije bilo dovoljno kako bi se napravila kvalitetna klasifikacija. Istraživanja koja su koristila metode klasteriranja na biološkim podacima koristeći Illuminu imala su mnogo veći broj sljedova (Brawand *i sur.*, 2011; Merkin *i sur.*, 2012), a samim time i bolji ishod, odnosno razlika između skupina unutar podataka bila je bolje vidljiva i s oštrijim granicama.

Sljedovi sekvencirani na uređaju ONT MinION korišteni su kako bi se pokušao rješiti problem prekratkih sljedova dobivenih metodom sekvenciranja sintezom prema kalupu. Ovi sljedovi ipak su nešto duži, najdulji 2D slijed dugačak je 34 kb. To je odlično u usporedbi s duljinom sljedova dobivenima metodom sekvenciranja sintezom prema kalupu (250 nk), no u odnosu na duljinu od čak 147 kb dobivenu u drugim istraživanjima (Goodwin *i sur.*, 2015), maksimalna veličina od 34 kb zapravo je trebala biti veća. Međutim, postoje osvrti da je rezultat od 147 kb rjetkost i da se zapravo vrlo često dobivaju sljedovi duljine od oko 10 kb (Urban *i sur.*, 2015). Problem je očito u tehnologiji sekvenciranja koja daje sljedove kraće nego što je očekivano. Još jedan problem je i ukupna duljina i broj 2D sljedova koji prelaze

zadani prag kvalitete dobivenih metodom sekvenciranja nanoporama. Ukupni broj sljedova možda i ne bi bio toliki problem s obzirom da je s istim brojem sljedova u potpunosti sklopljen bakterijski genom (Loman *i sur.,* 2015), no broj uspješno pročitanih 2D sljedova u odnosu na neuspješno pročitane 2D sljedove razočarava. Također, ukupna duljina knjižnica dobivenih ovom metodom razlikuje se od onog što je objavljeno u dosadašnim istraživanjima koja su opisala podatke dobivene nakon sekvenciranja nanoporama (Loman *i sur.,* 2015; Urban *i sur.,* 2015).

Sljedeći problem koji je mogao utjecati na ishod klasteriranja i sklapanja neprekinutih sljedova jest točnost sljedova dobivenih na uređaju ONT MinION koja je oko 85 %. Greške u sljedovima mogle su presuditi u klasifikaciji sljedova i u sklapanju neprekinutih sljedova. Iako sam sljedove pokušala ispraviti koristeći sljedove knjižnica Illumina MiSeq 1 i 2, nije došlo do velikog poboljšanja što se tiče dužine i cjelovitosti ovih sljedova, za razliku od nekih drugih rezultata (Goodwin *i sur.,* 2015; Madoui *i sur.,* 2015).

Iako postoje problemi s duljinom uspješno pročitanih 2D sljedova, njihovim brojem i ukupnom duljinom sljedova, sljedovi dobiveni ovom metodom ipak su uspjeli nakon klasteriranja dati barem jedan klaster koji je pretežito jednoličan. Zanimljivo je da sljedovi koji pripadaju ovom klasteru pripadaju i skupini koja je označena kao "NO HIT" skupina, odnosno tim sljedovima BLAST nije pronašao nijedan pogodak u bazi. Ti su sljedovi najzanimljiviji jer je za njih najveća mogućnost da pripadaju upravo istraživanoj spužvi *Ephydatia mülleri,* s obzirom da njen genom nije anotiran u bazama podataka. No program Tadpole ni u ovom klasteru nije uspio sklopiti veći broj neprekinutih sljedova, sklopio je tek jedan neprekinuti slijed dužine (samo) 111 nk, bez obzira na promjenu vrijednosti *k.*

Potrebno je skupiti više sljedova dobivenih na uređaju ONT MinION, pokušati ih ispraviti s nekim drugim programom te promijeniti korake u protokolu izolacije i pripreme knjižnice za sekvenciranje kako bi se omogućilo da su dobiveni sljedovi duži i točniji.

Problem s malo sklopljenih sljedova u ostalim klasterima u sve tri knjižnice (izuzev klastera pretežno jednoličnog klastera knjižnice Oxford Nanopore) je to što su sljedovi ili ipak prerazličiti, usprkos podacima dobivenim nakon BLAST-a ili u programu koji se koristio za sklapanje neprekinutih sljedova. Još jedna prepreka u klasteriranju je velik broj ponavljajućih sljedova i visoka heterozigotnost. Sve navedene značajke podataka korištenih u ovom radu uvelike utječu na klasteriranje pomoću učestalosti tetramera i izdvajanje kontaminacija koje se zbog navedenih karakteristika genoma spužava teže izdvajaju. Problemi bi se mogli rješiti

dugačkim i točnim ONT MinION sljedovima koji će premostiti problematične dijelove, ali i upotrebom dodatnih računalnih metoda u sferi strojnog učenja koje bi omogućile bolje klasteriranje i izdvajanje sljedova koji predstavljaju kontaminacije uprkos kompleksnosti genoma spužava.

Ova skupina i dalje je vrlo zanimljiva jer je već mnogo istraživača pokušalo raznim metodama dati odgovore na pitanja vezana uz genom spužava. Jedan sklopljeni genom nije dovoljan kako bismo razumijeli razvoj viših životinja ili same mehanizme unutar organizma spužve. Upravo zbog toga što se spužve smatraju mogućom prekretnicom u razvoju Eumetazoa potrebno je nestaviti s istraživanjima na ovoj skupini i pronaći dodatne metodološke pristupe analizi kontaminacija jer su one još uvijek veliki problem te onemogućavaju kvalitetno sklapanje genoma.

6 Zaključci

Sklapanje genoma određenog organizma može biti jako zahtjevan posao, pogotovo ako ne postoji referentni genom pa je potrebno genom sklopiti *de novo*. Takav je slučaj s većinom organizama iz skupine spužvi jer postoji samo jedan sklopljeni genom. *Ephydatia mülleri* modelni je organizam za proučavanje skupine Porifera i predstavlja zanimljiv izvor podataka za daljnje proučavanje genetike ove skupine. U ovom radu korišteni su očitani sljedovi dobiveni dvjema metodama sekvenciranja, metodom sekvenciranja sintezom prema kalupu (Illumina) i metodom sekvenciranja nanoporama (Oxford Nanopore). Glavni cilj ovog rada bila je predobrada sljedova kojoj slijedi sklapanje genoma. Veliki problem kvalitetnom i preciznom sklapanju genoma ove spužve predstavljaju kontaminacijski sljedovi. Koristeći metode nenadziranog strojnog učenja pokušala sam pronaći kontaminacijske sljedove u tri korištene knjižnice: Illumina MiSeq 1, Illumina MiSeq 2 i Oxford Nanopore. Zbog velike količine podataka koji traže dugotrajnu obradu koristila sam nasumične podskupove od ~ 20 000 očitanih sljedova iz knjižnica dobivenih metodom sinteze prema kalupu. Nakon što sam napravila matricu udaljenosti između svakog para sljedova u ovisnosti o učestalosti tetramera koristeći divergenciju po Jensenu i Shannonu te metodom stoihastičkog pridruživanja susjedima temeljenom na t-distribuciji pokušala pravilno klasterirati očitane sljedove, mogu zaključiti sljedeće:

- najbolje rezultate klasteriranja dali su sljedovi dobiveni metodom sekvenciranja nanoporama. No i dalje je sam rezultat prilično neprecizan, pa je potrebno skupiti još podataka i primijeniti dodatne računalne metode,
- knjižnice dobivene metodom sekvenciranja sintezom prema kalupu nisu dale poželjne rezultate, rezultati klasteriranja bili su prilično loši. Problem je vjerojatno duljina sljedova, ali i količina podataka. Valja imati na umu da veća količina podataka traži duže komputacijsko vrijeme te je računalna obrada izrazito zahtjevna,
- mali broj neprekinutih sljedova koji prati ove podatke vjerojatno je rezultat kontaminacija u klasterima obzirom da metode klasteriranja nisu precizno podijelile podatke, ali je lako moguće da je problem i u samom programu Tadpole, iako su uočena određena poboljšanja koristeći vrijednost *k* različitu od zadane. Potrebno je isprobati i druge programe koji su komputacijski zahtjevniji i proces traje puno duže,

- obzirom na probleme u ovom radu, smatram da bi se postigao veliki napredak razvojem Oxford Nanopore tehnologije u smislu dužih sljedova koji imaju manje grešaka.

No, bez obzira na prepreke istraživanja skupine spužvi treba nastaviti jer svi ovi problemi s kojima se istraživanja na ovom području susreću govore o kompleksnosti i zamršenosti genoma skupine Porifera, što bi moglo ići u prilog tome da su spužve važan korak u evoluciji viših životinja. Istraživanja treba usmjeriti prema pronalasku novih računalnih metoda i metode primijeniti na podatke dobivene sekvenciranjem genomskih DNA ove skupine. Važno je upotrijebiti dodatne metode strojnog učenja kako bi se što bolje izdvojili kontaminacijski sljedovi jer je moguće da su upravo kontaminacije ono što priječi kvalitetno sklapanje genoma skupine spužava, a time i razjašnjavanje evolucijskih odnosa unutar viših životinja.

7 Literatura

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–10.

Borchiellini C, Manuel M, Alivon E, Boury-Esnault N, Vacelet J, Parco Y Le (2001). Sponge paraphyly and the origin of Metazoa. *J Evol Biol* **14**: 171–179.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, *i sur.* (2011). The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Chan CX, Ragan MA, Chan C, Beiko R, Darling A, Ragan M, *i sur.* (2013). Next-generation phylogenomics. *Biol Direct* **8**: 3.

Dijk EL Van, Auger H, Jaszczyszyn Y, Thermes C (2014). Ten years of next-generation sequencing technology. *Trends Genet* **30**: 418–26.

Eisenstein M (2012). The battle for sequencing supremacy. *Nat Biotechnol* **30**: 1023–1026.

Ereskovsky, Alexander V, Dondua AK (2006). The problem of germ layers in sponges (Porifera) and some issues concerning early metazoan evolution. *Zool Anzeiger - A J Comp Zool* **245**: 65–76.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, *i sur.* (2010). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* **108**: 1513–8.

Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz M, McCombie WR (2015a). Oxford Nanopore Sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res* **25**: 1750–56.

Gori F, Mavroedis D, Jetten MSM, Marchiori E (2011). Genomic signatures for metagenomic data analysis: Exploiting the reverse complementarity of tetranucleotides. *IEEE Int Conf Syst Biol ISB* 149–54

Grosse I, Bernaola-Galvan P, Carpena P, Roman-Roldan R, Oliver J, Stanley HE (2002). Analysis of symbolic sequences using the Jensen-Shannon divergence. *Phys Rev E - Stat Nonlinear, Soft Matter Phys* **65**: 1–16.

Hodkinson BP, Grice EA (2015). Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. *Adv wound care* **4**: 50–8.

Hooper JNA, Hall KA, Ekins M, Erpenbeck D, Wörheide G, Jolley-Rogers G (2013). Managing and sharing the escalating number of sponge „unknowns“: the SpongeMaps project. *Integr Comp Biol* **53**: 473–81.

Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M (2015). Improved data analysis for the MinION nanopore sequencer. *Nat Methods* **12**: 351–56.

James G, Witten D, Hastie T, Tibshirani R (2000). An introduction to Statistical Learning. *Springer-Verlag New York*

Laczny CC, Pinel N, Vlassis N, Wilmes P, Konstantinidis KT, Braff J, *i sur.* (2014). Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. *Sci Rep* **4**: 5345–55.

Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, *i sur.* (2012). Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics* **11**: 25–37.

Liu Z, Meng J, Sun X (2008). A novel feature-based method for whole genome phylogenetic analysis without alignment: Application to HEV genotyping and subtyping. *Biochem Biophys Res Commun* **368**:

Loman NJ, Quick J, Simpson JT (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* **12**: 733–35.

Maaten L Van Der, Hinton G (2008). Visualizing Data using t-SNE. *J Mach Learn Res* **9**: 2579–605.

Madoui M-A, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, *i sur.* (2015). Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* **16**: 327.

Merkin J, Russell C, Chen P, Burge CB, Stamm S, Lareau LF, *i sur.* (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**: 1593–9.

Metzker ML (2010). Sequencing technologies - the next generation. *Nat Rev Genet* **11**: 31–

46.

Miller JR, Koren S, Sutton G (2010). Assembly algorithms for next-generation sequencing data. *Genomics* **95**: 315–27.

Mitra RD, Church GM (1999). In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res* **27**: 34-9.

Müller WEG (1995). Molecular phylogeny of Metazoa (animals): monophyletic origin. *Naturwissenschaften* **82**: 321–9.

Müller WEG (2006). The stem cell concept in sponges (Porifera): Metazoan traits. *Semin Cell Dev Biol* **17**: 481–91.

Pevzner PA, Tang H, Waterman MS (2001). An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA* **98**: 9748–53.

Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, *i sur.* (2009). Phylogenomics Revives Traditional Views on Deep Animal Relationships. *Curr Biol* **19**: 706–12.

Pop M (2009). Genome assembly reborn: Recent computational challenges. *Brief Bioinform* **10**: 354–66.

Quail M, Smith ME, Coupland P, Otto TD, Harris SR, Connor TR, *i sur.* (2012). A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* **13**: 341.

Reitner J, Worheide G (2002). Non-Lithistid Fossil Demospongiae – Origins of their Palaeobidiversity and Highlights in History of Preservation. U: Hooper, JNA, Van Soest, RWM (Ur), *Syst Porifera. Kluwer Academic / Plenum Publishers, New York* 52–68.

Riesgo A, Farrar N, Windsor PJ, Giribet G, Leys SP (2014). The analysis of eight transcriptomes from all poriferan classes reveals surprising genetic complexity in sponges. *Mol Biol Evol* **31**: 1102–20.

Rivera AS, Hammel JU, Haen KM, Danka ES, Cieniewicz B, Winters IP, *i sur.* (2011). RNA interference in marine and freshwater sponges: actin knockdown in Tethya wilhelma and Ephydatia muelleri by ingested dsRNA expressing bacteria. *BMC Biotechnol* **11**: 67.

Rokas A, Kruger D, Carroll SB (2005). Animal Evolution and the Molecular Signature of Radiations Compressed in Time. *Science* **310**: 1933–8.

Salmela L, Rivals E (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics* **30**: 3506–14.

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**: 5463–7.

Schatz MC, Delcher AL, Salzberg SL (2010). Assembly of large genomes using second-generation sequencing. *Genome Res* **20**: 1165–73.

Soest RWM Van, Boury-Esnault N, Vacelet J, Dohrmann M, Erpenbeck D, Voogd NJ De, *i sur.* (2012). Global diversity of sponges (Porifera). *PLoS One* **7**: e35105.

Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier ME a, Mitros T, *i sur.* (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* **466**: 720–6.

Thacker RW, Díaz MC, Kerner A, Vignes-Lebbe R, Segerdell E, Haendel MA, *i sur.* (2014). The Porifera Ontology (PORO): enhancing sponge systematics with an anatomy ontology. *J Biomed Semantics* **5**: 39.

Thacker RW, Freeman CJ (2012). Sponge-microbe symbioses: recent advances and new directions. *Adv Mar Biol* **62**: 57–111.

Urban JM, Bliss J, Lawrence CE, Gerbi SA (2015). Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. *bioRxiv* doi:10.1101/019281.

Watson M, Thomson M, Risse J, Talbot R, Santoyo-Lopez J, Gharbi K, *i sur.* (2015). poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics* **31**: 114–5.

Wörheide G, Dohrmann M, Erpenbeck D, Larroux C, Maldonado M, Voigt O, *i sur.* (2012). *Deep Phylogeny and Evolution of Sponges (Phylum Porifera). Adv Mar Biol* **61**:1-78

8 Prilozi

Tablica 1. Statistička obrada podataka sljedova knjižnica Illumina MiSeq 1 i 2, nakon uklanjanja adaptera i niskokvalitetnih baza očitanih sljedova.

Knjižnica	Broj očitanih fragmenata	Prosječna duljina očitanih fragmenata (nk)	Ukupna duljina fragmenata (nk)
AA miseq1	30 674 946	245.61	7 534 184 258
A4 (miseq2)	45 239 016	235.89	10 671 229 334

Tablica 2. Statistička obrada podataka o sljedovima knjižnica Illumina MiSeq 1 i 2, nakon spajanja na temelju nedvosmislenog preklapanja.

Knjižnica	Broj parova sljedova	Broj združenih sljedova	Broj dvosmislenih parova sljedova	Bez rješenja
Illumina MiSeq 1	15337473	10643296 (69.39 4%)	4694170 (30.606 %)	0.000 %
Illumina MiSeq 2	22619508	18265077 (80.749 %)	4354425 (19.251 %)	0.000 %

Tablica 3. Sažetak podataka o duljini neprekinutih sljedova knjižnice Illumina MiSeq 1 nakon sklapanja programom Tadpole. U petom klasteru koristeći zadanu vrijednost k = 31 nije sklopljen niti jedan neprekinuti slijed, ali je koristeći manju vrijednost k ipak sklopljen jedan neprekinuti niz.

Klaster	Broj sklopljenih neprekinutih sljedova	Najmanja dužina neprekinutog slijeda (nk)	Srednja vrijednost dužine neprekinutog slijeda (nk)	Najveća dužina neprekinutog slijeda (nk)
1	30	100.0	153.5	664.0
2	10	100.0	188.1	279.0
3	37	100.0	155.9	356.0
4	11	103.0	160.1	360.0
5	0	-	-	-
5 ***k <= 29***	1	100.0	100.0	100.0

Tablica 4. Sažetak podataka o duljini neprekinutih sljedova knjižnice Illumina MiSeq 2 nakon sklapanja programom Tadpole koristeći zadanu vrijednost *k* = 31. Povećavajući vrijednost *k*, povećao se broj neprekinutih sljedova sklopljenih u klasteru 5.

Klaster	Broj sklopljenih neprekinutih sljedova	Najmanja duljina neprekinutog sljeda	Srednja duljina neprekinutog sljeda	Najveća duljina neprekinutog sljeda
1	86	100.0	203.3	870.0
2	50	101.0	202.7	575.0
3	2	128.0	145.5	163.0
4	2	106.0	135.0	164.0
5	3	132.0	194.0	296.0
5 *k* = 60	10	100.0	141.3	249.0

ŽIVOTOPIS

OSOBNI PODATCI

Ime:	Dunja Habulin
Adresa:	Augusta Šenoe 3, Zabok (Hrvatska)
Broj mobilnog telefona:	099/6938-716
e-mail adresa:	dunja.habulin@gmail.com
Datum rođenja	9. siječnja 1993.
Državljanstvo	hrvatsko

OBRAZOVANJE

2014. – 2016.	Diplomski studij Molekularne biologije na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu (Računalni modul: listopad 2015. – travanj 2016.)
2011. – 2014.	Preddiplomski studij Molekularne biologije na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu
2007. – 2011.	Opća gimnazija Antuna Gustava Matoša u Zaboku

LABORATORIJSKA PRAKSA I KONFERENCIJE

veljača 2016. - sad	Projekt sklapanja genoma spužava u Bioinformatičkom laboratoriju na Prirodoslovno-matematičkom fakultetu u Zagrebu, pod vodstvom prof. dr. sc. Kristiana Vlahovičeka
listopad 2014. – lipanj 2015.	Istraživanje poremećaja S-adenozilhomocistein hidrolaze u ljudi, u laboratoriju za Translacijsku medicinu na Institutu Ruđer Bošković u Zagrebu, pod vodstvom dr. sc. Olivera Vugreka
studeni 2014.	Konferencija Hrvatskog društva za istraživanje raka (HDIR) "From Bench to Clinic"
prosinac 2013. – lipanj 2014.	Istraživanja vezana uz izoleucil-t-RNA sintetazu na Zavodu za biokemiju na Prirodoslovno-matematičkom fakultetu u Zagrebu, pod vodstvom prof. dr. sc. Ite Gruić
listopad 2013. – prosinac 2013.	Laboratorijska praksa u Laboratoriju za epigenetiku na Prirodoslovno-matematičkom fakultetu u Zagrebu, pod vodstvom prof. dr. sc. Vlatke Zoldoš