Sveučilište u Zagrebu
Prirodoslovno - matematički fakultet
Biološki odsjek

Dora Šribar

Sklapanje genoma ogulinske špiljske spužvice (*Eunapius subterraneus*) iz podataka dobivenih tehnologijom sekvenciranja nanoporama

Diplomski rad

Zagreb, 2016.

Ovaj rad, 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 na stručnoj podršci, obitelji na moralnoj podršci, a ponajviše mentoru prof. dr. sc. Kristianu Vlahovičeku.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Diplomski rad

SKLAPANJE GENOMA OGULINSKE ŠPILJSKE SPUŽVICE (*EUNAPIUS SUBTERRANEUS*) IZ PODATAKA DOBIVENIH TEHNOLOGIJOM SEKVENCIRANJA NANOPORAMA

Dora Šribar
Rooseveltov trg 6, 10000 Zagreb, Hrvatska

Koljeno Porifera iz carstva životinja (Metazoa) obuhvaća preko 8000 znanstveno opisanih vrsta spužvi. Zbog svoje jednostavne građe spužve imaju važnu ulogu u razumijevanju rane evolucije i odnosa između ostalih koljena iz carstva Metazoa. Spužve su i dobri modelni organizmi za razumijevanje bioloških procesa u višim životinjama. Uz morfološke podatke, sve važniju ulogu imaju i genomski podaci. Za složeniju funkcionalnu genomiku Porifera potrebni su nam genomi više predstavnika koljena, međutim, sekvenciranje i sklapanje čitavog genoma spužvi i dalje predstavlja izazov. Ogulinska špiljska spužvica *Eunapius subterraneus* zbog svojih osobitosti predstavlja zanimljiv modelni organizam. Pomoću dostupnih genomskih knjižnica sklopila sam, a potom i procijenila kvalitetu sklopljenog genoma ogulinske špiljske spužvice. Pritom sam koristila algoritme temeljene na de Bruijnovim grafovima i metodi preklapanje-raspored-konsenzus te primijenila hibridni pristup pri sklapanju genoma pomoću dugačkih sljedova dobivenih sekvenciranjem na uređaju The Oxford Nanopore Technologies MinION. Korištenjem hibridnog pristupa pomoću sljedova dobivenih na uređaju Oxford Nanopore Technologies MinION nije primijećeno poboljšanje u sklopljenom genomu. Moguće poteškoće u sastavljanju genoma predstavljaju onečišćenja, repetitivne sekvence i visoka heterozigotnost. Potrebna su daljnja istraživanja na području genomike Porifera.

(31 stranica, 9 slika, 6 tablica, 52 literaturnih navoda, jezik izvornika: hrvatski)

Rad je pohranjen u Središnjoj biološkoj knjižnici

Ključne riječi: hibridno sklapanje genoma, tehnologije sekvenciranja sljedeće generacije, preklapanje-raspored-konsenzus, de Bruijnov graf

Voditelj:	Prof. dr. sc. Kristian Vlahoviček
Ocjenitelji:	Prof. dr. sc. Kristian Vlahoviček Izv. prof. dr. sc. Sven Jelaska Doc. dr. sc. Damjan Franjević
Zamjena:	Izv. prof. dr. sc. Dunja Leljak-Levanić

Rad prihvaćen: 4.2.2016.

BASIC DOCUMENTATION CARD

University of Zagreb
Faculty of Science
Division of Biology

Graduation Thesis

GENOME ASSEMBLY OF ENDEMIC CAVE SPONGE (*EUNAPIUS SUBTERRANEUS*) USING DATA OBTAINED BY NANOPORE SEQUENCING TECHNOLOGY

Dora Šribar
Rooseveltov trg 6, 10000 Zagreb, Croatia

Porifera, a phylum within the Kingdom Animalia (Metazoa), consists of more than 8000 species of sponges. Owing to their simple morphology, sponges have an important role in understanding the early evolution and relationship between different phyla within the kingdom Animalia. Furthermore, sponges represent good models for biological processes in higher animals. Morphological features are becoming insufficient and we need more genomic data from different species within Porifera. However, the entire process of genome sequencing and assembly of Porifera still remains the challenge. Endemic cave sponge *Eunapius subterraneus* is an interesting model organism due to its unusual characteristics. Using various genomic libraries I conducted genome assembly and subsequent evaluation of the assembled genome of *Eunapius subterraneus*. I used various approaches based on both de Bruijn graphs and the Overlap-Layout-Consensus and conducted hybrid assembly using long reads obtained from the Oxford Nanopore Technologies MinION device. The hybrid assembly did not yield any improvement. Low assembly quality could be consequence of potential contamination, repetitive sequences and high heterozigosity rate. Further research of Porifera genomic is needed.

(31 pages, 9 figures, 6 tables, 52 references, original in: Croatian)

Thesis deposited in the Central Biological Library

Key words: hybrid assembly, Next generation sequencing, Overlap-Layout-Consensus, de Bruijn graphs

Supervisor:	Professor Kristian Vlahoviček, PhD
Reviewers:	Professor Kristian Vlahoviček, PhD Assoc. Prof Sven Jelaska, PhD Asst. Prof Damjan Franjević, PhD
Substitution:	Assoc. Prof Dunja Leljak-Levanić, PhD

Thesis accepted: February 4, 2016

Sadržaj

Kratice

BLAST	*engl.* Basic Local Alignment Search Tool
dNTP	Deoksiribonukleotid-trifosfat
Nk	Nukleotid
ONT	*engl.* The Oxford Nanopore Technologies
pb	Nukleotidni par
PCR	Lančana reakcija polimerazom
PRK	Preklapanje-raspored-konsenzus
SGA	*engl.* String Graph Assembler

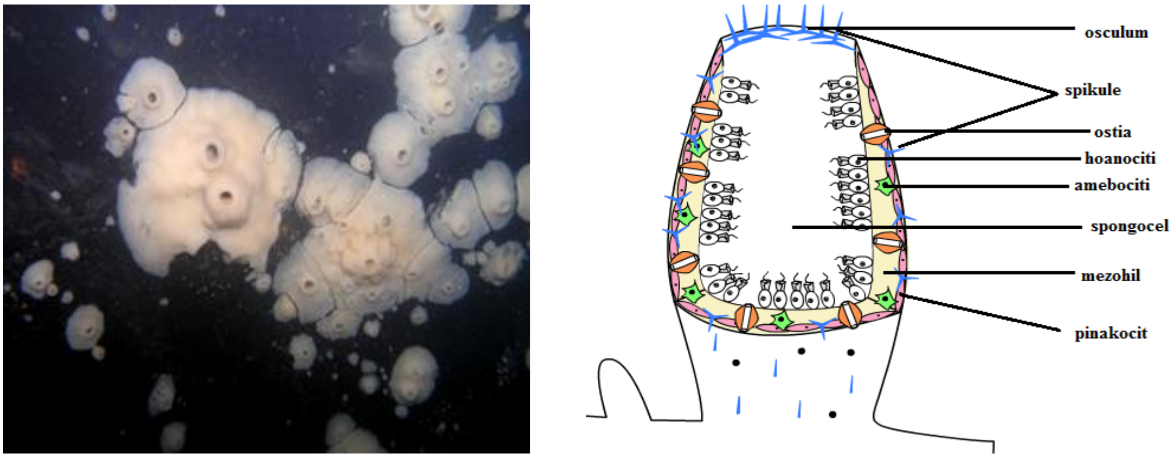
1 Uvod

1.1 Opće značajke spužava

Koljeno Porifera iz carstva životinja (Metazoa) obuhvaća preko 8000 znanstveno opisanih vrsta spužvi. Spužve su sesilni, vodeni, većinom morski organizmi. Smatraju se najjednostavnijim i najstarijim višestaničnim životinjama, s fosilima koji datiraju iz doba Prekambrija. Zbog svoje jednostavne građe spužve imaju važnu ulogu u razumijevanju rane evolucije i odnosa između ostalih koljena iz carstva Metazoa (Wörheide i sur. 2012).

1.1.1 Građa i životni ciklus spužava

Spužve su građene od visoko specijaliziranih i samostalnih stanica. Razlikuju se unutrašnji, srednji i vanjski sloj stanica (Slika 1.). Vanjski sloj sastoji se od spljoštenih i poligonalnih stanica pinakocita. Srednji sloj mezohil se sastoji od želatinoznog proteinskog i ugljikohidratnog matriksa, pokretnih stanica amebocita i kostura. Kostur spužve građen je od iglica ili spikula različitih oblika, veličine i sastava (silicijev dioksid i kalcijev karbonat) i/ili elastičnih proteinskih vlakana spongina. Unutrašnji sloj čine bičaste stanice hoanociti. Mnogobrojne male pore ostia prevode vodu u unutrašnje kanalne sustave ispresijecane hoanocitima. Hoanociti bičevima sinkroniziranim kretanjem proizvode vodenu struju koja služi za dovođenje kisika, hranjivih tvari i oslobađanje otpadnih tvari. Filtrirana voda i otpadni produkti se izbacuju kroz veliki otvor osculum.



Slika 1. Ogulinska špiljska spužvica *Eunapius subterraneus* (lijevo) i shema građe spužve (desno), preuzeto i prilagođeno s http://www.rufford.org/rsg/projects/jana_bedek_0 i http://www.biology.ualberta.ca/courses.hp/zool250/animations/Porifera.swf

Spužve se razmnožavaju spolno i nespolno. Jedna spužva izbacuje zrele spermatozoide te se oni vodom prenose do druge jedinke. Specijalizirani hoanociti dovode spermije do jajne stanice. Zigota se razvija u bičastu ličinku i ispušta u vodu gdje slobodno pluta prije nego što se pričvrsti za odgovarajuću podlogu i razvije u odraslu spužvu. Spužva se nespolno razmnožava pupanjem pomoću gemula. Gemule su izrazito otporne na nepovoljne uvjete kao što su nedostatak kisika i smrzavanje. Kod ovog oblika razmnožavanja odvaja se dio stanica iz kojeg se razvija novi organizam (Encyclopedia.com 2016).

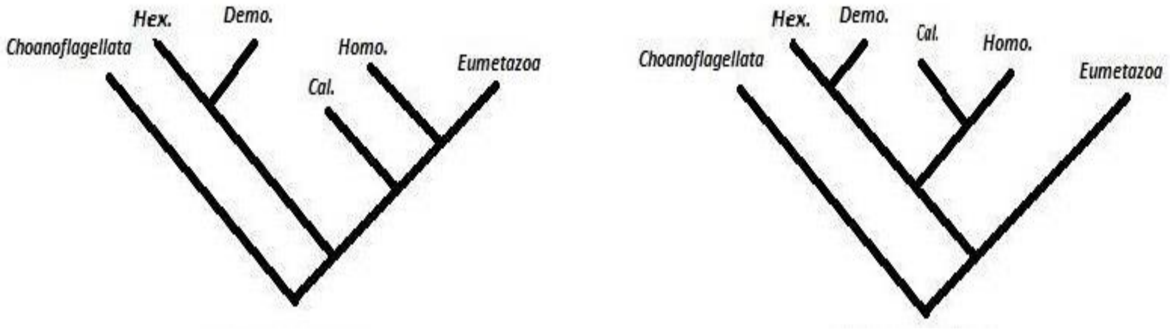
1.1.2 Sistematika i filogenija spužava

Koljeno Porifera dijeli se u 4 glavna razreda uglavnom temeljeno na građi skeletnih elemenata:

- **Hexactinellida** ili staklače naseljavaju morske dubine. Karakteriziraju ih triosne i šesterozrakaste silikatne spikule i nakupine stanica sinciciji (Janussen i sur. 2004; Leys i sur. 2007)
- **Calcispongiae** ili vapnenjače većinom naseljavaju morske plićake. Proizvode izvanstanične vapnenačke spikule (Cárdenas i sur. 2012)
- **Demospongiae** je najbrojniji razred slatkovodnih i morskih spužvi. Imaju monoosne i četveroosne silikatne spikule i spongioblaste koji izlučuju spongin (Boury-Esnault 2006).
- **Homoscleromorpha** su do nedavno svrstavane u razred Demospongia. Zanimljive su jer imaju bazalnu laminu ispod vanjskog i unutarnjeg sloja (Boute i sur. 1996; Gazave i sur. 2011).

Spužve se smatraju najprimitivnijim i najstarijim životinjama. Najstariji fosil spužve star je oko 750 milijuna godina (Reitner i Wörheide 2002). Karakteristike Metazoa npr. diploidna višestaničnost, oogeneza, spermatogeneza, mejoza i građa samog spermija te dostupni molekularni podaci dijele spužve i ostale više Metazoe od njihovih najbližih predaka okovratnih bičaša (Choanoflagellata) (Müller 1995; Rokas i sur. 2005).

U znanstvenim krugovima postoji debata o tome jesu li spužve monofiletska ili polifiletska skupina (Slika 2.). Monofiletska teorija, temeljena na morfologiji, smatra da su spužve monofiletska sestrinska skupina višim životinjama (Eumetazoa) (Zrzavy i sur. 1998). Parafiletska teorija, temeljena na velikom broju molekularnih i morfoloških dokaza, smatra da su razredi Calcarea i Homoscleromorpha bliži Eumatazoama nego ostali razredi iz koljena Porifera (Borchiellini i sur. 2001).



Slika 2. Parafiletsko (lijevo) i monofiletsko (desno) podrijetlo spužava, preuzeto i prilagođeno s https://u.osu.edu/eeob3320/2015/03/

1.1.3 Genomske značajke spužava

Uvidom u genom prve sekvencirane spužve *Amphimedon queenslandica* otkriveno je da ona sadrži brojne gene Eumetazoa, uključujući ključne gene za adheziju i staničnu signalizaciju, međustaničnu komunikaciju, imunosno prepoznavanje te transkripcijske faktore uključene u razvoj germinativne linije i spola. Unatoč nedostatku živčanog sustava, pronađeni su geni uključeni u razvoj živčanog sustava u Eumetazoama (Richards i sur. 2008). Međutim, *A. queenslandica* nedostaju pojedini geni koji su očuvani u drugim životinjama, upućujući na dodatno proširivanje genomskog repertoara nakon odvajanja od spužvi i/ili gubitak pojedinih gena u samoj *A. queenslandica* (Srivastava i sur. 2010; Riesgo i sur. 2014).

1.1.4 Značajke ogulinske špiljske spužvice

Ogulinska špiljska spužvica *Eunapius subterraneus* endemska je vrsta ogulinskog područja i jedina poznata podzemna slatkovodna spužva na svijetu (Slika 1.). Dosadašnja istraživanja su pokazala da je ona najvjerojatnije pravi stigobiont (vodeni organizam potpuno prilagođen na špiljske uvjete koji nikad ne živi u nadzemnim sustavima). Nedostatak pigmenta, usporen metabolizam, promijenjena fiziologija stanica i cijelog organizma, promijenjen način razmnožavanja najvjerojatnije su posebnosti ovog organizma u odnosu na nadzemne srodnike (Bedek i sur. 1984).

Dosadašnja genomska i filogenetska istraživanja analize mitohondrijske genomske DNA pokazala su da je, suprotno očekivanom, *Eunapius* bliži slatkovodnim vrstama *Ephydatia* *muelleri* i bajkalskoj *Lubomirska baikalenskiis* nego ostalim vrstama iz roda *Eunapius* (Harcet i sur. 2010).

1.2 Tehnologije sekvenciranja

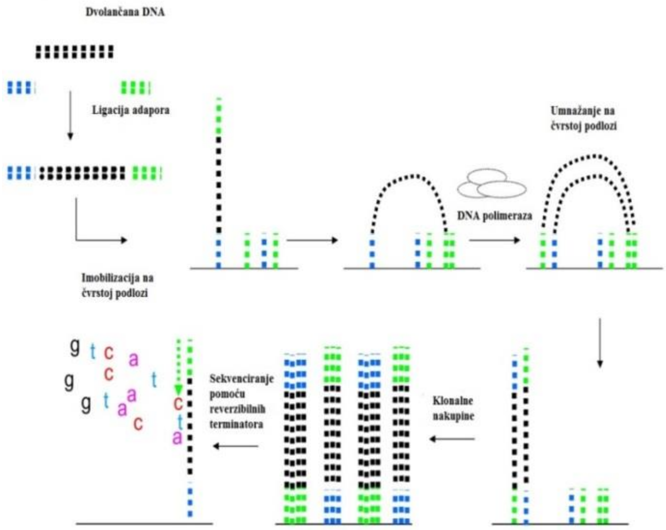
Sekvenciranje DNA obuhvaća skup postupaka i tehnologija kojima se određuje slijed dušičnih baza u molekuli DNA. Većina sekvenatora temelji se na metodi koju je razvio Sanger 1975. Metoda se temelji na umnažanju kratkog umnoženog uzorka DNA pomoću DNA polimeraze, odgovarajućih deoksinukleotida i specifičnih prekidajućih dideoksinukleotida koji dovode do zastoja u sintezi komplementarnog lanca. Sintetizirana DNA se razdvaja elektroforezom na temelju veličine te se očita specifičan slijed nukleotida (Sanger i sur. 1977). Danas su na tržištu dostupni različiti uređaji temeljeni na automatizaciji Sangerovog sekvenciranja koji koriste fluorescentno obilježene prekidajuće dideoksinukleotide, razdvajanje pomoću kapilarne elektroforeze i lasersku detekciju signala. Metoda temeljena na automatiziranom Sangerovom sekvenciranju još se naziva tehnologijom sekvenciranja prve generacije te je unatoč zahtjevnosti i danas najprimjenjivanija tehnologija sekvenciranja.

Tehnologije sekvenciranja sljedeće generacije postale su komercijalno dostupne od 2005. Pojam obuhvaća tehnologije koje se temelje se na sekvencijalnom dodavanju nukleotida imobiliziranim i umnoženim kalupima DNA. Međusobno se razlikuju primarno po načinu dobivanja kalupa DNA i načinu ispitivanja identiteta dodane baze (Linnarsson 2010). Prednost tehnologija sekvenciranja sljedeće generacije pred tehnologijama prve generacije je nedostatak potrebe za bakterijskim kloniranjem, mnogo više paralelnih reakcija sekvenciranja i nedostatak potrebe za elektroforezom što omogućuje veću količinu očitanih sljedova u kraćem vremenu. Glavni nedostaci su manja duljina očitanih sljedova (Quail i sur. 2011; Dijk i sur. 2014) i često manja pouzdanost točnosti svakog pročitanog nukleotida (Ansorge 2009).

Razlikuju se tehnologije kod kojih postoji potreba za umnažanjem DNA kalupa lančanom reakcijom polimerazom (*engl. Polymerase Chain Reaction*, PCR) i tehnologije kod kojih je moguće sekvenciranje pojedinačne molekule (*engl. Single molecule sequencing*). Umnažanje je potrebno kako bi se pojačao signal jer detektor nije u mogućnosti detektirati dodatak samo jedne baze na razini jedne molekule DNA. Tehnologije koje koriste umnažanje su Illumina, SOLiD i Roche 454 (Mitra 1999). Tijekom umnažanja DNA postoji mogućnost nejednolikog tj. pristranog umnažanja što naposljetku može dovesti do krivog signala i očitane baze (Dabney i Meyer 2012). S druge strane, tehnologije Pacific Biosciences sekvenciranje pojedinačne molekule u realnom vremenu (*engl. Single Molecule, Real-Time*) i Heliscope ne trebaju umnažanje jer mogu detektirati signal na razini jedne molekule DNA. S obzirom na potrebu za umnažanjem DNA kalupa, tehnologije sekvenciranja sljedeće generacije možemo podijeliti na tehnologije druge i treće generacije. Tehnologije treće generacije omogućavaju detekciju pojedinačne molekule i tzv. sekvenciranje u realnom vremenu (*engl. real time sequencing*) (Berglund i sur. 2011).

1.2.1 Metoda reverzibilnog zaustavljanja sinteze DNA (Illumina)

Illumina proizvodi veliki broj raznovrsnih platformi koje se temelje na tehnologiji koja se pojavila 2006. i brzo postala vrlo popularna među istraživačima zbog velikog broja sekvenciranih sljedova za nisku cijenu sekvenciranja (Hodkinson i Grice 2015). Svi koraci sekvenciranja i detekcije se odvijaju na protočnoj ćeliji (*engl. flow cell*) (Slika 3.). Illumina koristi umnažanje na čvrstoj podlozi (*engl. bridge amplifiaction*) i sekvenciranje pomoću sinteze. Priprema genomke knjižnice za sekvenciranje obuhvaća fragmentaciju, popravak krajeva, fosforilaciju 5' kraja, dodavanje A sljeda na 3' kraj kako bi se omogućila ligacija adaptera i naposljetku ligaciju adaptera (Head i sur. 2014). Na površini cijele protočne ćelije su pričvršćene početnice s 5' i 3' kraja (*engl.* „forward“ i „reverse“) komplementarne adapterima dodanim u prethodnim koracima pripreme knjižnice. Prvi korak je denaturacija dvolančane DNA u jednolančanu DNA. Pomoću adaptera jednolančana DNA hibridizira s početnicama na površini protočne ćelije. Nakon umnažanja početnih DNA kalupa oni se odstranjuju, a kopije pričvršćene za površinu umnažaju se pomoću početnica u nakupine identičnih DNA molekula (tzv. otoke ili klastere). Cijepa se jedan lanac DNA u dvolančanoj umnoženoj DNA pomoću specifičnih mjesta u oligo početnicama, te se blokira 3' slobodan kraj na pričvršćenom lancu kako bi se spriječila reakcija na slobodnom kraju. Tijekom sekvenciranja, nakupine umnoženih DNA molekula se čitaju po jedan nukleotid u svakom ponovljenom ciklusu. Deoksiribonukleotid-trifosfati (*engl. deoxynucleotide triphosphate,* dNTP) obilježeni različitom fluorescentnom bojom se ugrađuju u rastući DNA lanac.



Slika 3. Shema sekvenciranja pomoću Illumine, preuzeto i prilagođeno s http://www.gsejournal.org/content/44/1/21/figure/F2?highres=y

Uz flourecentnu boju dNTP sadrži i tzv. reverzibilni terminator koji sprečava ugradnju sljedećeg nukleotida. Nakon detekcije flourescentog signala, odcijepi se boja i deaktivira terminator te se omogućuje sljedeći ciklus dodavanja nukleotida. Nedostaci umnažanja su neefikasnost reakcije umnažanja, rehibridizija lanca kalupa DNA i pristranost DNA polimeraze prema pojedinim DNA kalupima (Bentley i sur. 2008).

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

Tehnologija se zasniva na prolasku nukleinske kiseline kroz molekularnu poru promjera nekoliko nanometara ugrađenu u membranu na koju se primjeni napon. Prolaskom molekule kroz poru, mijenja se ionska struja. Uređaj detektira promjene u struji na razini pentamera koji prolaze kroz membranu i „događaja“ kao što su vrijeme početka i trajanje prolaska pojedinačne molekule DNA kroz membranu te zatim prevodi dobiveni signal u odgovarajući slijed nukleotida u sekvenci pomoću ONT Metrichor „cloud“ servisa. Kad je propisno ligirana, dvolančana DNA sadrži adapter u obliku slova Y na jednom kraju i adapter u obliku ukosnice na drugom kraju. Dvolančana DNA se provlači kroz poru počevši na 5' kraju Y adaptera, slijedi lanac kalup, zatim idealno adaptor u obliku ukosnice i naposljetku komplementarni lanac. Pročitani sljedovi mogu uključivati podatke s jednog lanca (kalupa ili komplementarnog lanca) te ih nazivamo 1D lancima, dok informacija s 2 lanca dovodi do tzv. 2D sljedova veće kvalitete.

Prednosti su što je uređaj malen, prijenosan, visokoprotočan, može proizvesti sljedove dulje i od 100 kb, jeftin i brz (Eisenstein 2012; Bayley 2014).

1.3 Sklapanje genoma

Sklapanje genoma je rekonstrukcija izvorne DNA sekvence iz očitanih sljedova dobivenih sekvenciranjem. Sklapanje se temelji na nalaženju preklapanja i sklapanja preklapajućih sekvenci u neprekinute sljedove. Sklapanje genoma se može podijeliti u dvije osnovne skupine:

1. Sklapanje genoma *de novo* obuhvaća sklapanje sekvenci u neprekinute sljedove (*engl. contigs*) te točno uređivanje neprekinutih sljedova u prekinute sljedove (*engl. scaffolds*).
2. Mapiranje pomoću referentnog genoma obuhvaća sravnjenje očitanih sljedova s referentnim genomom (Bao i sur. 2011).

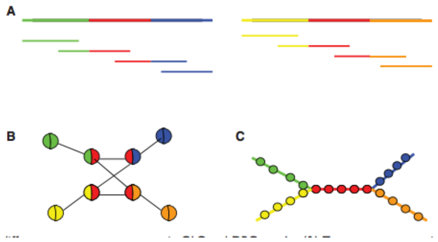
Kako u najvećem broju slučaju ne postoji referentni genom, potrebno je sklapanje genoma pomoću metode *de novo*. Pri sklapanju se mogu koristiti samostalni sljedovi ili upareni očitani sljedovi. Upareni sljedovi su jednake duljine kao samostalni sljedovi, ali kako znamo informaciju o udaljenosti pojedinih sljedova iz para, možemo složiti međusobno nepreklapajuće sljedove u preklapajuće sljedove. Sklapanje genoma možemo podijeliti na 4 osnovna koraka: predobradu podataka, konstrukciju preklapajućih sljedova, spajanje nepreklapajućih sljedova i popunjavanja praznina. Metode sklapanja genoma *de novo* možemo podijeliti u tri skupine s obzirom na algoritam koji koriste: metode temeljene na „pohlepnom“ algoritmu (*engl. greedy algorithm*), tzv. metode preklapanje-raspored-konsenzus (*engl. Overlap-layout-consensus* ,PRK) i metode po de Bruijnu (Miller i sur. 2010; Schatz i sur. 2010; Li i sur. 2012).

1.3.1 Metoda preklapanje-raspored-konsenzus

Kod metode PRK (preklapanje-raspored-konsenzus) traži se međusobno preklapanje između svih parova sljedova. Iz sljedova se izgradi graf u kojem čvorovi predstavljaju sljedove, a ukoliko postoji preklapanje između dva slijeda/čvora veće od granične vrijednosti, čvorovi se spoje rubom. Broj čvorova odgovara broju sljedova. Faza rasporeda obuhvaća pronalaženje puta u grafu kojem je svaki čvor posjećen točno jednom, što odgovara Hamiltonianovoj stazi i smatra se računalno NP kompletnim problemom. Za NP kompletne probleme još nije nađeno efikasno riješenje u realnom vremenu. Naposljetku se na temelju višestrukog sravnjenja nalazi konsenzus sekvenca (Pop 2009).

1.3.2 Metoda po de Bruijnu

Metoda po de Bruijnu uključuje razdvajanje sljedova na kraće k-mere, korištenja k-merova za izgradnju De Bruijnovog grafa i naposljetku očitanja odgovarajuće genomske sekvence iz grafa. Za razliku od PRK metode, pronalaženje puta u grafu predstavlja Eulerovu stazu, gdje svaki rub treba proći jednom, koju je mnogo lakše riješiti nego Hamiltonionovu stazu (Pevzner i sur 2001). Nadalje, ova metoda je mnogo manje računalno zahtjevna od PRK metode jer ne koristi prvi korak međusobnog sravnjenja sljedova (Slika 4.).



Slika 3. Usporedba PRK metode (B) i de Bruijnovih grafova (C) pri sklapanju odgovarajućih genomskih sekvenci (A), preuzeto sa http://bioinformaticsk-state.blogspot.hr/2012/08/introduction-to-two-common-sequence.html

2 Ciljevi rada

Razvojem tehnologija sekvenciranja, posebno tehnologija druge i treće generacije, moguće je dobiti veliku količinu podataka o genomskoj informaciji organizma u relativno kratkom vremenu. 2010. godine sekvenciran je, sklopljen i funkcionalno opisan prvi genom spužve *Amphimedon queenslandica* (Srivastava i sur. 2010). Uvidom u genom i transkriptom prve sekvencirane spužve pronađeno je iznenađujuće mnogo gena koji nalikuju onima morfološki kompleksnijih životinja. Za složeniju funkcionalnu genomiku Porifera potrebni su nam genomi više predstavnika koljena, međutim, sekvenciranje i sklapanje čitavog genoma spužvi i dalje predstavlja izazov.

Pomoću dostupnih genomskih knjižnica korištenjem bioinformatičkih metoda sklopit ću, a potom i procijeniti kvalitetu sklopljenog genoma ogulinske špiljske spužvice. Sklapanje kvalitetnog genoma omogućilo bi potpunije daljnje filogenetske analize, pomoglo bi u rasvjetljavanju prijelaza jednostaničnih organizama na višestanične te naposljetku detaljan pogled na događaje na razini DNA i RNA u životnom ciklusu ogulinske špiljske spužve.

3 Materijali i metode

3.1 Sekvencirane genomske knjižnice

Korištene su dvije knjižnice kratkih udaljenosti između parova očitanih fragmenata (*engl. paired end*) sekvencirane na uređajima Illumina MiSEQ. Na uređaju IIlumina sekvencirana je i jedna knjižnica uparenih očitanih fragmenata (*engl. mate-pair*) (Tablica 1.).

Tablica 1. Statistike podataka dobivenih nakon sekvenciranja genomske DNA spužve *Eunapius subterraneus* na Illumina uređajima

Knjižnica	Broj očitanih fragmenata	Prosječna duljina očitanih fragemenata (nk)	Ukupna duljina (nk)
Illumina MiSEQ 1	30 871 686	251	7 748 793 186
Illumina MiSEQ 2	31 456 208	251	7 895 508 208
Illumina LMP	42 651 284	101	4 307 779 684

Naposljetku, korišteni su i sljedovi DNA izolirani i sekvencirani prema odgovarajućem protokolu na uređaju The Oxford Nanopore Technologies MinION (ONT).

Program za detekciju baza očitanih na sljedovima ONT Metrichor izdvaja datoteke u dvije odvojene mape: "pass" ili uspješna mapa i "fail" ili neuspješna mapa. Uspješna mapa sadrži 2D sljedove sa srednjom kvalitetom slijeda Q>=9. Neuspješna mapa sadrži 2D sljedove Q<9, 1D sljedove kod kojih nije bilo moguća detekcija 2D slijeda i sve sljedove s neuspješnom detekcijom 1D slijeda nukleotida (Tablica 2.).

Tablica 2. Statistike podataka 2D sljedova dobivenih nakon sekvenciranja genomske DNA spužve *Eunapius subterraneus* na ONT uređaju

ONT MinIon sljedovi	Broj očitanih fragmenata	Prosječna duljina očitanih fragmenata (nk)	Ukupna duljina (nk)
Uspješni 2D sljedovi	1 492	4088	6 100 496
Neuspješni 2D sljedovi	3 472	3 767	13 079 097

3.2 Provjera kvalitete sekvenciranih sljedova

Prije i nakon predobrade sljedova nukleotida sekvencirane knjižnice analizirane su i vizualizirane pomoću programa FastQC za provjeru kvalitete sljedova dobivenih visokoprotočnim sekvenciranjem.

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

3.3 Predobrada sekvenciranih sljedova

3.3.1 Predobrada sljedova sekvenciranih na uređaju Illumina

Za predobradu i daljnju analizu sljedova sekvenciranih na uređaju Illumina koristila sam programe otvorenog koda iz programskog paketa BBMap/BBTools.

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

3.3.1.1 Uklanjanje adaptera i sljedova s nedovoljnom pouzdanošću očitanih nukleotida

BBDuk korišten je za uklanjanje adaptera i sljedova s nedovoljnom pouzdanošću očitanih nukleotida. Upotrijebljeni su standardni sljedovi adaptera koji su distribuirani zajedno s programskim paketom BBMap/BBTools.

Knjižnici Illumina MiSEQ 1 odstranjene su baze na krajevima sljedova kvalitete pouzdanosti očitanih nukleotida niže od 20 i sljedovi koji sadrže preklapanje s adapterskim sljedovima duljine 28 nt sa jednim dozvoljenim nepreklapanjem.

Knjižnici Illumina MiSEQ 2 odstranjene su baze na krajevima sljedova kvalitete pouzdanosti očitanih nukleotida niže od 18 i sljedovi koji sadrže preklapanje s adapterskim sljedovima duljine 28 nt s jednim dozvoljenim nepreklapanjem. Nadalje, primijenjeno je dodatno odstranjivanje adaptera na temelju preklapanja dvaju uparenih sljedova i jednakomjerno odstranjivanje adaptera.

Knjižnici Illumina LMP odstranjeni su sljedovi koji sadrže preklapanje s adapterskim sljedovima duljine 23 nt s jednim dozvoljenim nepreklapanjem.

3.3.1.2 *Spajanje parova očitanih fragmenata*

Korišten je BBMerge kako bi se na temelju preklapanja parovi očitanih fragmenata knjižnica Illumina MiSEQ 1 i Illumina MiSEQ 2 spojili u jedan slijed.

3.3.1.3 *Normalizacija i ispravljanje pogrešaka u sljedovima*

BBNorm se temelji na učestalosti zadanih k-mera. Kod normalizacija se pojedini sljedovi odbace s određenom vjerojatnošću temeljeno na omjeru ciljane pokrivenosti genoma i medijana učestalosti k-merova u slijedu. Za ispravljanje pogrešaka prvo se detektiraju susjedni k-merovi s vrlo različitom učestalošću. Zatim se promijeni sumnjiva baza u rijetkom k-meru ako se pritom dobije k-mer sa frekvencijom sličnoj frekvenciji susjednih k-merova.

Normalizacija dubine sljedova i ispravljanje pogrešaka na temelju učestalosti 31-mera izvršena je pomoću programa BBNorm.

3.3.2 Predobrada sljedova sekvenciranih na uređaju The Oxford Nanopore Technologies MinION

3.3.2.1 *Pretvorba FAST5 formata u FASTQ format*

Neobrađeni podaci dobiveni detekcijom baza nakon sekvenciranja na The Oxford Nanopore Technologies MinION uređaju nalaze se u tzv. formatu FAST5 što je jedna od inačica hijerarhijskog podatkovnog formata (*engl. Hierarchical Data Format*, HDF5). HDF5 format omogućava pohranjivanje i organizaciju velike količine podataka, ali dostupni programi iz područja računalne genomike ne podržavaju HDF5 format.

R je programski jezik otvorenog koda sa snažnom objektno orijentiranom podrškom. Zbog njegovih statističkih i grafičkih tehnika, kao i zbog sve većeg broja paketa koje kreiraju korisnici za specijalizirane zadaće, postaje sve popularniji, a širi se i broj znanstvenih područja gdje se primjenjuje. Paket poRe omogućava korisnicima manipulaciju, organizaciju, sažetak i vizualizaciju sljedova dobivenih sekvenciranjem na ONT uređaju (Watson i sur. 2014).

Korištenjem paketa poRe programskog jezika R iz FAST5 podataka dobivenih nakon detekcije baza izvukla sam odgovarajuće 2D FASTQ sljedove. (http://sourceforge.net/projects/rpore/)

Koristila sam posebno uspješne 2D sljedove te zajedničke uspješne 2D sljedove i neuspješne 2D sljedove dulje od 2000 nt.

3.3.2.2 Ispravljanje pogrešaka

Program Nanocorr koristi hibridni pristup, koristeći visoko kvalitetne kraće sljedove dobivene na Illumina MiSEQ uređaju kako bi ispravio dugačke, ali pune pogrešaka ONT sljedove. Algoritam sravnjuje kratke MiSEQ sljedove s dugačkim ONT sljedovima, pomoću dinamičkog programiranja odabire idealni skup MiSEQ sljedova koji premošćuju cijeli ONT slijed i naposljetku nalazi konsenzusni slijed. (Goodwin i sur. 2015)

Dostupne sljedove ispravila sam pomoću programa Nanocorr koristeći Illumina MiSEQ 1 knjižnicu (https://github.com/jgurtowski/nanocorr).

3.4 Sklapanje genoma

Koristila sam različite dostupne programe temeljene na de Bruijnovom grafu i metodi preklapanje-raspored-konsenzus (PRK).

3.4.1 SOAPdenovo

SOAPdenovo je program za sklapanje genoma temeljen na de Bruijnovom grafu specijaliziran za kratke sljedove sekvencirane na IlluminaGenome Analyzer uređaju (Luo i sur. 2012).

Koristeći program SOAPdenovo2 sklopila sam genom špiljske spužvice pomoću knjižnice uparenih sljedova Illumina MiSEQ 1 koristeći vrijednosti neparnih k-merova od 35-127. Isprobane su različite jačine spajanja sljedova tijekom faze sklapanja preklapajućih sljedova (http://soap.genomics.org.cn/soapdenovo.html).

3.4.2 SPAdes

SPAdes je program za sklapanje genoma također temeljen na de Bruijnovim grafovima. SPAdes može koristiti sljedove dobivene sekvenciranjem na Illumina ili IonTorrent uređaju i dodatne sljedove za hibridno sklapanje genoma sa PacBio, Oxford Nanopore i Sangerovim sljedovima što ga čini posebno korisnim. Program sadrži i dodatne korisne potprograme za ispravljanje pogrešaka u Illumina sljedovima BayesHammer i za sklapanje polimorfnih diploidnih genoma dipSPAdes (Bankevich i sur. 2012).

Koristeći SPAdes pomoću knjižnice Illumina MiSEQ 1 sklopila sam genom spužve. Pritom sam za ispravljanje pogrešaka u sljedovima koristila potprogram BayesHammer.

Koristeći SPAdes i knjižnice Illumina MiSEQ 1, Illumina MiSEQ 2 i Illumina LMP te sljedove dobivene na ONT uređaju, sklopila sam genom pomoću hibridnog sklapanja genoma. Pritom su korišteni 2D ONT sljedovi visoke kvalitete i svi pomoću Nanocorra ispravljeni 2D sljedovi (http://bioinf.spbau.ru/spades).

3.4.3 String Graph Assembler (SGA)

Slaganje genoma pomoću metode PRK primijenjeno je i pomoću računalnog programa String Graph Assembler (SGA). SGA koristi strukture podataka i algoritme koji značajno smanjuju potrebnu memoriju, te omogućuje da se i veće količine podataka slažu pomoću metode PRK (Simpson i Durbin 2012).

Računalnim programom SGA sklopila sam knjižnicu kratkih inserata Illumina MiSEQ 1. Programom SGA provedeno je dodatno ispravljanje očitanih fragmenata prema učestalosti k-mera u uzorku. Za pronalazak optimalne vrijednosti duljine preklapanja za stvaranje neprekidnih sljedova, programom SGA su isprobane duljine preklapanja između 55 nukleotida i 150 nukleotida s razmacima od 10 nukleotida (https://github.com/jts/sga).

3.4.4 CELERA

Celera je program za sklapanje genoma *de novo* korištenjem sljedova dobivenih „metodom sačmarice“ (*engl. de novo* whole-genome shotgun (WGS) DNA sequence assembler) te je temeljen na PRK metodi. Prihvaća sljedove dobivene različitim tehnologijama sekvenciranja uključujući Illuminu i ONT.

Koristeći CELERU sklopila sam genom pomoću Illumina MiSEQ 1 knjižnice (http://wgs-assembler.sourceforge.net/wiki/index.php?title=Main_Page).

3.5 Procjena kvalitete sklopljenog genoma

Uspješnost programa za sklapanje genoma se najčešće procjenjuje pomoću veličine i točnosti sklopljenih prekinutih i neprekinutih sljedova. Također se gleda i potrošnja resursa izraženo kao vrijeme procesiranja i zauzimanje memorije.

Najčešće se koriste sljedeće statističke mjere:

- N50 - duljina sklopljenog slijeda pri kojoj 50% ukupne duljine sljedova prisutno u sklopljenim sljedovima ove duljine ili dulje. Standardna mjera je za izražavanje cjelovitosti sklopljenog genoma
- Duljina najkraćeg neprekinutog slijeda
- Duljina najduljeg neprekinutog slijeda

Ukoliko nije drugačije navedeno, u statistiku ulaze sklopljeni sljedovi dulji od 500 pb.

3.5.1 BUSCO

BUSCO (*engl. Benchmarking Universal Single-Copy Orthologs*) je program za procjenu dovršenosti genoma koji koristi univerzalne ortologe u jednoj kopiji iz OrthoDB (www.orthodb.org) kako bi kvantitativno procijenio dovršenost sklopljenog genoma, anotiranih genskih setova i transkriptoma na temelju očekivanog sadržaja gena (Simão i sur. 2015).

Koristila sam BUSCO kako bih procijenila dovršenost sklopljenih genoma koristeći skup od 429 visoko očuvanih eukariotskih gena (http://busco.ezlab.org/).

3.5.2 BLAST *(engl. Basic Local Alignment Search Tool)*

BLAST (*engl. Basic Local Alignment Search Tool*) je algoritam za uspoređivanje nukleotidnih ili aminokiselinskih sljedova prema sličnosti. Omogućuje pretraživanje baze gena ili proteina, te kao rezultat daje sve sekvence koje s našom sekvencom od interesa imaju sličnost jednaku ili veću od određenog praga. Postoji nekoliko inačica programa ovisno o tipu sekvence koju uspoređujemo i tipu baze s kojom se uspoređuje (Altschul 1990).

Koristeći nukleotidni BLAST sklopljeni genomi sravnjeni su s nukleotidnom bazom sljedova te su sačuvani samo najbolji pogoci (*engl. hits*) sa sklopljenih sljedova koji su 100% identični sa sravnjenim slijedom u bazi (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Za izvlačenje odgovarajućih taksonomskih podataka korišten je programski paket za R taxize

(https://cran.r-project.org/web/packages/taxize/index.html).

4 Rezultati

4.1 Predobrada sekvenciranih sljedova

4.1.1 Predobrada sljedova dobivenih na Illumina uređaju

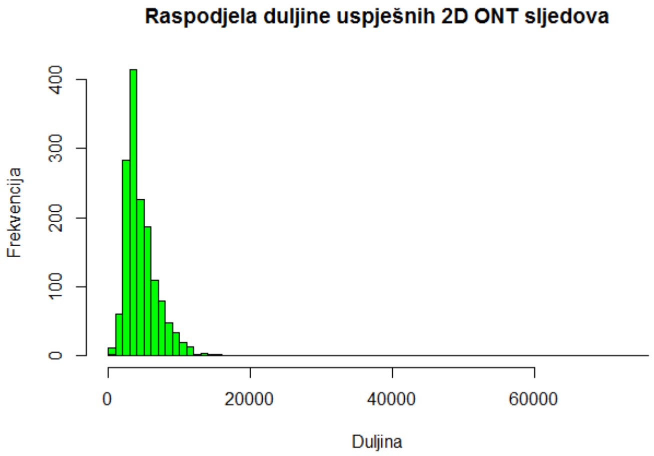
Analizom početnih sljedova dobivenih sekvenciranjem na Illumina uređajima, uočila sam pad kvalitete pouzdanosti očitanih fragmenata prema krajevima kod duljih MiSEQ sljedova te prisutnost adaptorskih sljedova. Nakon uklanjanja baza s nedovoljnom pouzdanošću očitanih nukleotida i adapterskih sljedova ukupna duljina pročišćenih sljedova iznosila je redom ~82%, ~86%, i ~99% ukupne duljine početnih sljedova dobivenih sekvenciranjem Illumina MiSEQ 1, Illumina MiSEQ 2 i Illumina LMP knjižnice (Prilog, Tablica 1.).

Kako bi se povećala duljina i kvaliteta sljedova, pročišćene uparene sljedove iz Illumina MiSEQ 1 i Illumina MiSEQ 2 spojila sam u samostalne sljedove na temelju nedvosmislenog preklapanja. Kako Illumina MiSEQ 2 sadrži značajan broj dvosmislenih sljedova (48%) koje je nemoguće spojiti, zadržana je informacija u obliku uparenih nespojenih sljedova (Prilog, Tablica 2.).

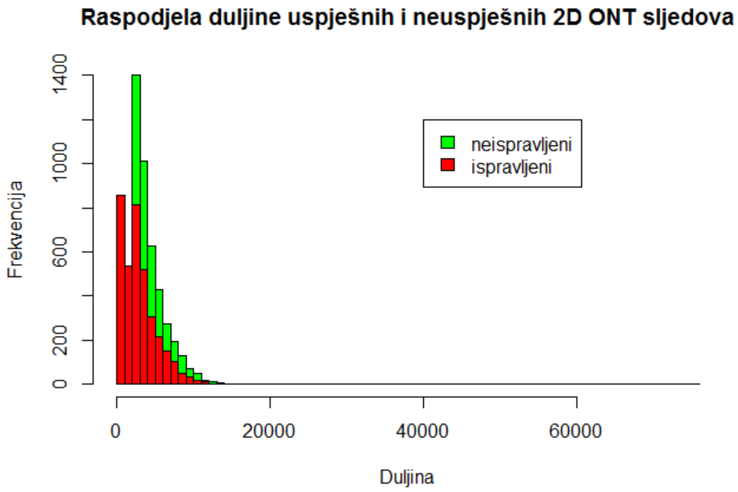
Hibridni pristup s ONT sljedovima zahtijeva veći broj knjižnica te se upotrebom potprograma za ispravljanje pogrešaka na temelju učestalosti k-mera BayesHammera nastoji smanjiti broj podataka i pojednostaviti graf za sljedeći korak sklapanja genoma. Međutim, pokazalo se da je taj korak za veći broj knjižnica zahtjevan i izrazito dugotrajan zato sam uparene nespojene knjižnice Illumina MiSEQ 2 i Illumina LMP umjesto pomoću programa BayesHammer dodatno ispravila pomoću programa BBNorm. Pritom je provedena i normalizacija kako bi se dodatno smanjilo računalno vrijeme u sljedećem koraku sklapanja genoma (Prilog, Tablica 3.).

4.1.2 Predobrada sljedova dobivenih na ONT uređaju

Sljedovi dobiveni sekvenciranjem na ONT uređaju pokazuju široki raspon duljina sljedova-prosječna duljina sljedova je oko 4 kb s najdužim 2D slijedom dugačkim oko 75 kb (Slika 5.). Međutim, problem predstavlja mali broj uspješnih 2D sljedova - njih samo 1492. Kako bi se dobilo više sljedova korišteni su i neuspješni 2D sljedovi dulji od 2000 baza. Zatim sam sljedove ispravila pomoću programa Nanocorr. Nakon ispravljanja sljedova ukupna duljina sljedova se smanjila te se povećao broj kraćih sljedova (Slika 6.). Ispravljeni sljedovi su ~85% identični sa neispravljenim sljedovima.



Slika 4. Raspodjela duljine uspješnih 2D ONT sljedova



Slika 5. Raspodjela duljine uspješnih i neuspješnih 2D ONT sljedova

4.2 Sklapanje genoma pomoću knjižnice Illumina MiSEQ 1

Kako bi se dobile grube smjernice za sljedeće korake koristeći pročišćenu knjižnicu Illumina MiSEQ 1 s uparenim sljedovima sklopljeni su genomi pomoću različitih programa. Kao primarna statistika za procjenu uspješnosti sklopljenog genoma uzeta je N50 preklapajućih i nepreklapajućih sljedova, zanemarujući moguće pogreške u sklapanju genoma. Prema statistikama o duljini sastavljenog genoma, odabrana je optimalna duljina preklapanja za SGA (Prilog, Tablica 4.) i duljina k-mera i jačina spajanja sljedova za SOAPdenovo (Prilog, Tablica 5.-7.). Za sklapanje sa SGA koristila sam statistike sklopljene s genomom koji je sklopljen s preklapanjem od 75 pb. Kod SOAPdenova koristila sam statistike dobivene sklapanjem genoma s 89-merom i najjačim spajanjem sličnih sljedova. Najbolje rezultate prema statistikama izlaznih preklapajućih i nepreklapajućih sljedova dao je program temeljen na PRK metodi Celera (N50=5336) s najmanjom fragmentacijom sklopljenih sljedova (Tablica 3.). Međutim, program je računalno zahtjevniji nego ostali programi. Pomoću SPAdesa je sklopljen genom s najduljom duljinom neprekinutih sljedova i relativno prihvatljivim N50. Međutim, genom sklopljen pomoću SPAdesa je mnogo fragmenitraniji u odnosu na genom sklopljen pomoću Celere. SGA i SOAPdenovo2, iako računalno najmanje zahtjevni, dali su podjednako slabe rezultate.

Tablica 3. Statistika podataka neprekinutih sljedova sklopljenih genoma pomoću programa CELERA, SGA, SOAPdenovo2 i SPAdes

Program	Celera	SGA	SOAPdenovo2	SPAdes
Broj neprekinutih sljedova	29 134	158 841	85 596	124 684
Ukupna duljina neprekinutih sljedova (pb)	120 214 202	187 246 968	105 088 749	219 341 702
Duljina najduljeg neprekinutog slijeda (pb)	126 345	91 854	52 927	164 764
N50 neprekinutih sljedova	5 336	1 366	1 437	2 789

Gledajući statistike složenih nepreklapajućih sljedova najbolje rezultate opet daje Celera (Tablica 4.). Međutim, sama statistika svih programa se nije značajno poboljšala u odnosu na preklapajuće sljedove što je očekivano zbog malog omjera duljine inserta i duljine uparenih sljedova.

Tablica 4. Statistika podataka prekinutih sljedova sklopljenih genoma pomoću programa CELERA, SGA, SOAPdenovo2 i SPAdes

Program	Celera	SGA	SOAPdenovo2	SPAdes
Broj prekinutih sljedova	25 172	157 124	75 441	123 225
Ukupna duljina	120 355 338	187 125 651	105 349 572	219 573 863

prekinutih sljedova (pb)				
Duljina najduljeg prekinutog slijeda (pb)	133 300	91 854	59 320	164 764
N50 prekinutih sljedova	6 584	1 398	1 874	2 913

4.3 Hibridno sklapanje genoma pomoću SPAdesa

Tablice 5. i 6. prikazuju statistiku podataka za sklapanje genoma pomoću programa koristeći knjižnice Illumina MiSEQ 1, Illumina MiSEQ 2 i Illumina LMP te podatke dobivene sekvenciranjem na ONT MinION uređaju. Statistika N50 pokazuje da kod korištenja ONT MinION sljedova dolazi do neznatnog poboljšanja i kod preklapajućih i nepreklapajućih sljedova. Što se tiče fragmentiranosti također nema značajnijeg poboljšanja.

Tablica 5. Statistika podataka neprekinutih sljedova hibridno sklopljenih genoma pomoću programa SPAdes

	Bez ONT sljedova	S uspješnim 2D ONT sljedovima	S neispravljenim uspješnim i neuspješnim sljedovima	S ispravljenim uspješnim i neuspješnim 2D sljedovima
Broj neprekinutih sljedova	202 025	202 020	202 025	201 965
Ukupna duljina neprekinutih sljedova (pb)	350 146 490	350 166 651	350 146 490	350 137 301
Duljina najduljeg neprekinutog slijeda (pb)	204 792	204 792	204 792	204 792
N50 neprekinutih sljedova	2 440	2 442	2 440	2 442

Zanimljivo je da postoji značajna razlika u duljini najduljeg prekinutog sljeda, s tim da je neočekivano najdulji slijed kod kojeg nisu korišteni ONT MinION sljedovi.

Tablica 6. Statistika podataka prekinutih sljedova hibridno sklopljenih genoma pomoću programa SPAdes

	Bez ONT slijedova	S uspješnim ONT sljedovima	S neispravljenim uspješnim i neuspješnim sljedovima	S ispravljenim uspješnim i neuspješnim 2D sljedovima
Broj prekinutih sljedova	200 721	200 717	200 721	200 641

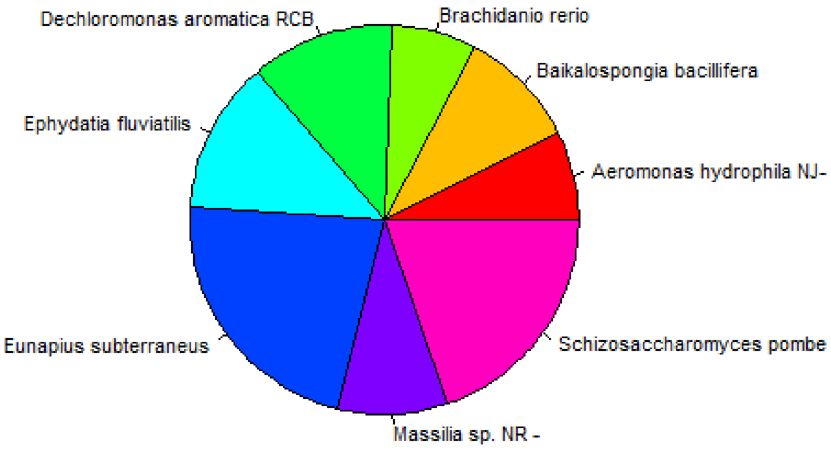
Ukupna duljina prekinutih sljedova (pb)	350 678 670	35 069 2850	350 678 670	350 675 439
Duljina najduljeg prekinutog slijeda (pb)	249 970	232 871	249 970	233 081
N50 prekinutih sljedova	2 459	2 460	2 459	2 461

4.3.1 Procjena sklopljenog genoma pomoću programa BUSCO-a

Pomoću programa BUSCO-a od 429 ključnih i visoko očuvanih eukariotskih gena u sklopljenom genomu nađeno je 40% potpuno dovršenih gena, oko 19% nepotpuno dovršenih gena dok 40 % gena nije pronađeno.

4.3.2 Sravnjenje sljedova sklopljenog genoma pomoću BLAST-a

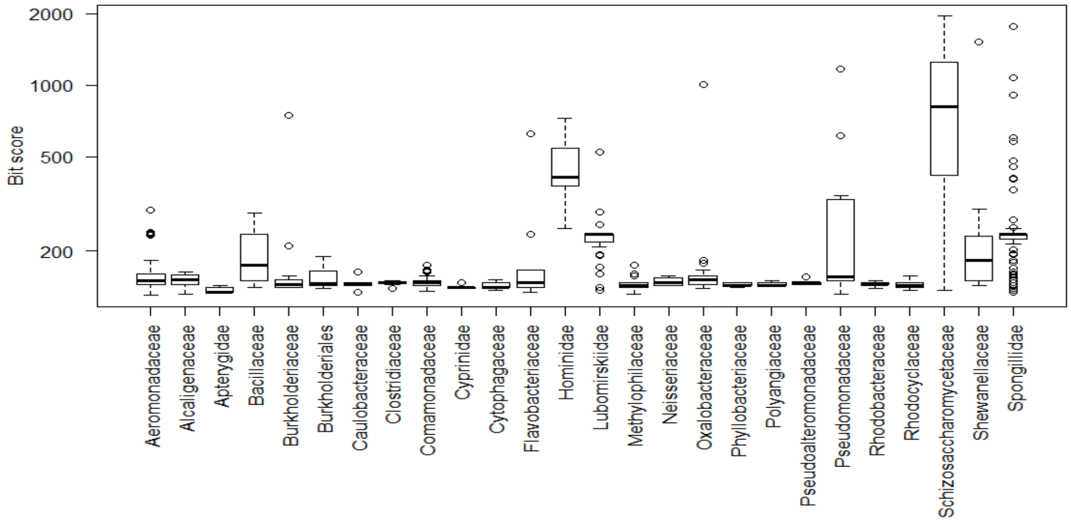
Rezultati sravnjenja pokazuju najveći udio organizama iz koljena Porifera, od kojih najveći broj pogodaka ima mitohondrijska DNA *Eunapius subterraneus* (slika 7.). Međutim, značajan broj pogodaka imaju i različite bakterije te kvasac *Schizosaccharomyces pombe*.



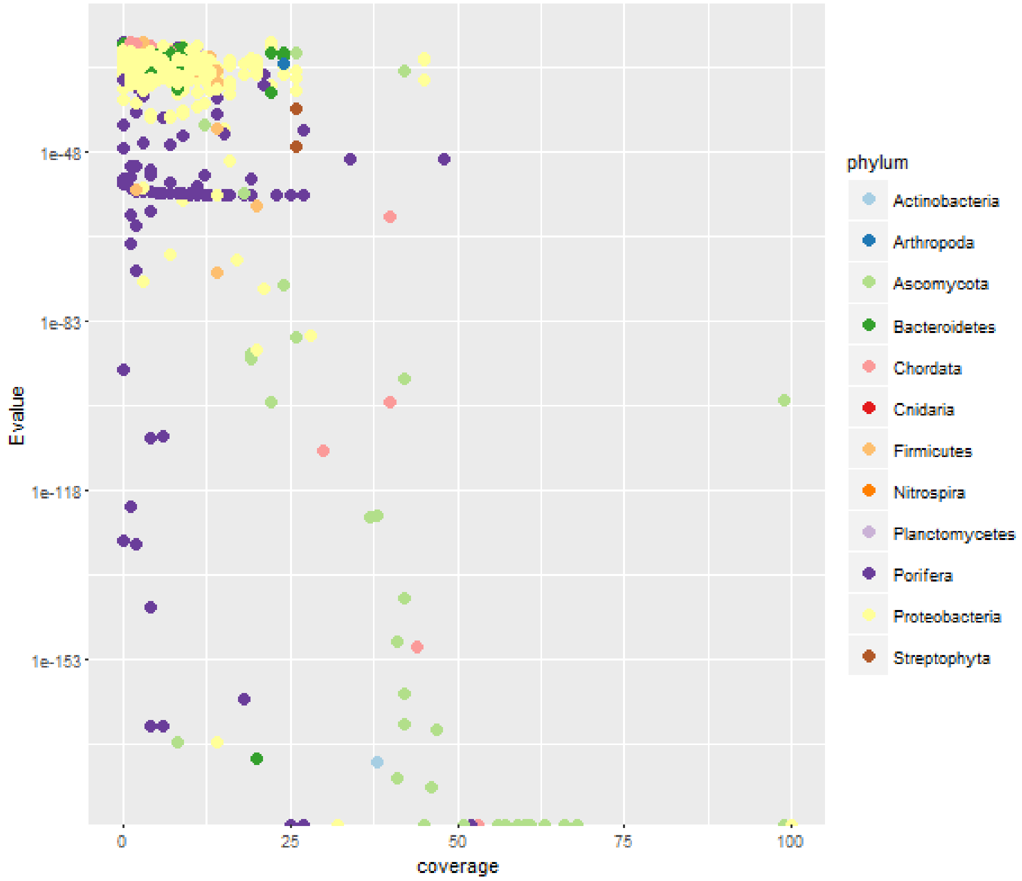
Slika 6. Udio najzastupljenijih organizama po broju pogodaka sravnjena s nukleotidnom bazom sljedova pomoću BLAST-a

Kako bi se vidjela značajnost dobivenih pogodaka, prikazala sam raspodjelu „Bit score“ za svaku porodicu pomoću „box plota“ (slika 8.) i ovisnost E-value i pokrivenosti (*engl. coverage*) po pojedinom koljenu (slika 9.). Pokrivenost nam govori koliki je postotak ispitivanog slijeda sravnjen s nađenim pogotkom u bazi. Najznačajniji pogotci su iz porodice

Schizosacharomycetacea, Hominidae i Spongillidae. Nadalje, viša pokrivenost i manji E-value sa Slike 9. isključuju mogućnost slučajnog pogotka. Iznenađujuće, najveću pokrivenost daju rezultati iz koljena Ascomycota.



Slika 7. „Box plotovi“ raspodjele Bit scorea po pojedinim porodicama.



Slika 8. Ovisnost E-value i pokrivenosti slijeda po pojedinim koljenima

5 Rasprava

U radu sam primarno koristila sljedove dobivene sekvenciranjem na uređaju Illumina. Unatoč optimizmu zbog uspješno sklopljenog genoma pande (Li i sur. 2010) koristeći isključivo sljedove Illumina s visokom pokrivenošću, sklapanje genoma pomoću kratkih sljedova Illumina predstavlja izazov. Koristila sam sljedove Illumina sa znatno nižom pokrivenošću od one koju su koristili Li i suradnici. Nadalje, Gnerre i suradnici (2010) su pokazali da su Sangerovim sekvenciranjem prve generacije dobiveni cjelovitiji sljedovi nego sljedovi dobiveni s kraćim sljedovima nove generacije. Postavlja se pitanje kolika je prednost jeftinijih visokoprotočnih Illumina sljedova s visokom pokrivenosti genoma nad slabijom pokrivenošću, ali duljim sljedovima klasičnog Sangerovog sekvenciranja.

Kako bi se pokušao prevladati problem kraćih sljedova, korišten je tzv. hibridni pristup u sklapanju genoma gdje su sljedovi dobiveni s Illumina tehnologijom sklopljeni koristeći i tehnologiju sekvenciranja nanoporama na uređaju ONT MinION. Prednost ovih sljedova je značajno veća duljina dobivenih sljedova, najdulji 2D slijed dugačak je čak 80 kb. Pojedina istraživanja potvrdila su 2D sljedove dugačke i do 147 kpb (Goodwin i sur. 2015). Problem u početnim podacima predstavlja ukupna duljina i broj 2D sljedova dobivenih na ONT uređaju. Ukupna duljina i broj sljedova i uspješnih visoko kvalitetnih 2D sljedova i 2D sljedova slabije kvalitete je znatno manja od prijavljenih duljina u dosadašnjim istraživanjima koja su također koristila ONT MinION uređaj (Loman i sur. 2015; Urban i sur. 2015).

Nadalje, problem predstavlja i prijavljeni niski postotak točnosti sljedova MinION (~85%) koji mogu dovesti do krivog spajanja sljedova u sklapanju. Pogreške sam pokušala ispraviti pomoću programa Nanocorr. Međutim, nije došlo do značajnijeg povećanja veličine i cjelovitosti isprekidanih sljedova, za razliku od rezultata koji su dobili Goodwin i suradnici (2015). Oni su uspjeli pomoću ispravljenih MinION sljedova zajedno s Illumina sljedovima sklopiti *de novo* genom kvasca povećavajući N50 sljedova s ~59 kpb na 479 kpb. Pri tom je pokazana visoka točnost sklapanja, 99% sklopljenog genoma mapiralo se na referentni genom te su uspjeli sklopiti genske kazete, rRNA, transpozone i ostale elemente koji su bili odsutni u genomima sklopljenima isključivo pomoću podataka Illumina. Koristeći slični princip Madoui i sur. (2015) također su uspjeli poboljšati sklapanje genoma. Smatram da je svakako potrebno dodatno sekvenciranje na uređaju ONT MinION kako bi se dobile dodatne informacije. Također bi bilo poželjno pokušati mijenjati različite parametre u protokolu izolacije i pripreme knjižnice za sekvenciranje ne bi li se dobili što dulji sljedovi.

Što se tiče samih programa, algoritam temeljen na metodi PRK je dao nadmoćno najbolje rezultate izlaznih sklopljenih sljedova. Međutim, problem predstavlja njihova računalna zahtjevnost te se danas za količinski nadmoćnije sljedove Illumina i dalje preporučuje korištenje de Bruijnovih grafova.

Preduvjet za dobru anotaciju genoma je kvalitetno sklopljen genom. Kako kod sekvenciranja *de novo* ne postoji referentni genom kako bi se procijenila kvaliteta sklopljenog genoma, glavnu mjeru za procjenu dovršenosti i cjelovitosti sklopljenog genoma predstavlja statistika sklopljenih neprekidajućih i prekidajućih sljedova, pogotovo N50. Jedno od pravila je da bi N50 prekinutih sljedova trebao biti dugačak kao prosječna duljina gena u genomu. Medijan duljine gena je otprilike proporcionalan ukupnoj duljini genoma, te se na taj način može približno pretpostaviti duljina gena od ~3000 pb (Yandell i Ence 2012) što se približno podudara s podacima dobivenim za srodnu vrstu *Ephydatia muelleri*. U slučaju *Eunapius subterraneus*, duljine N50 prekinutih sljedova dobivenih različitim programima kreću se oko 2000-3000 pb. Poželjnija bi bila viša N50 vrijednost za daljnju anotaciju genoma. Nadalje, treba uzeti u obzir mogućnost pogrešaka u sklapanju tj. neprikladnog spajanja dva slijeda što može uzrokovati veće, ali krivo spojene sljedove. Salzberg i sur. (2012) su pokazali da je točnost sklopljenog genoma varijabilna ovisno o programu te nema dobre korelacije sa samim statistikama.

Naposljetku, na samu kvalitetu sklopljenog genoma, uz navedenu kvalitetu sljedova i programe za sklapanje genoma, znatno utječu i same karakteristike genoma. Moguće je i kontaminacija genoma na što upućuje i sravnjenje sljedova s bazom nukleotida. Uz pogotke na spužvama prisutan je nezanemariv broj pogodaka iz mogućih kontaminanata kao što su ribe i različite bakterije. Pojedine porodice dobivenih pogodaka, prvenstveno bakterija i kvasaca, podudaraju se sa porodicama nađenih simbionata (Passarini i sur. 2014). Uz kontaminaciju, prepreku dobrom sklapanju genoma mogu predstavljati i repetitivne sekvence i visoka heterozigotnost. Sve navedeno utječe na povećanje kompleksnosti grafa, združivanje neprikladnih sljedova i onemogućavanja pronalaženje staze u grafu. Navedeni problemi bi se mogli riješiti korištenjem dugačkih sljedova npr. sljedova uređaja ONT koji su dovoljno dugi da premoste problematični slijed.

Potrebna su daljnja istraživanja genomike spužava. Unatoč nedostatku morfoloških karakteristika, spužve bi mogle imati mnogo značajnije uloge od dosad pretpostavljenog. Uz važnu ulogu u razumijevanju razvoja životinja, one bi mogle pomoći rasvijetliti mehanizme nastanka tumora. Nadalje, spužve su i komercijalno sve zanimljivije zbog potrebe za pronalaskom novih terapijskih i biotehnoloških izvora (Leal i sur. 2012).

6 Zaključci

- Najbolje rezultate daje sklapanje genoma pomoću programa temeljenog na metodi preklapanje-raspored-konsenzus. Međutim, značajnu prepreku i dalje predstavlja izrazita računalna zahtjevnost u odnosu na programe koji koriste de Bruijnove grafove
- Korištenjem hibridnog pristupa pomoću sljedova dobivenih na Oxford Nanopore Technologies MinION uređaju nije primijećeno poboljšanje u sklopljenom genomu. Najvjerojatnije je problem u malom broju kvalitetnih sljedova dobivenih sekvenciranjem
- Uz kvalitetu sekvenciranih ulaznih podataka i odabir različitih algoritama, na sklapanje genoma mogu utjecati same značajke genoma kao što su repetitivne sekvence i visoka heterozigotnost te moguća kontaminacija
- Zbog navedenih problema sklapanje genoma *de novo* i dalje predstavlja izazov, ali rješavanju problema moglo bi pomoći korištenje dugačkih ONT MinION sljedova
- Spužve su zanimljiv modelni organizam za daljnja istraživanja, pogotovo „metodu sačmarice“, najviše zbog njihove evolucijske važnosti, te mogućnosti korištenja kao modelnih organizama za razumijevanje procesa u višim životinjama.

7 Literatura

1. Altschul, S., 1990. Basic Local Alignment Search Tool. J. Mol. Biol. 215, 403–410.
2. Ansorge, W.J., 2009. Next-generation DNA sequencing techniques. N. Biotechnol. 25, 195–203.
3. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comp. Biol. 19, 455–477.
4. Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X., Song, Y.-Q., 2011. Evaluation of next-generation sequencing software in mapping and assembly. J. Hum. Genet. 56, 406–414.
5. Bayley, H., 2014. Nanopore Sequencing: From Imagination to Reality. Clin. Chem. 61, 25–31.
6. Bedek J, Bilandžija H.,Jalžić B., 2008. Ogulinska špiljska spužvica Eunapius subterraneus Sket et Velikonja, 1984, rasprostranjenost i ekologija vrste i staništa. Modruški zbornik 2, 103-130.
7. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Cheetham, R.K., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M.J., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M.D., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Catenazzi, M.C.E., Chang, S., Cooley, R.N., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fajardo, K.V.F., Furey, W.S., George, D., Gietzen, K.J., Goddard,

C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Jones, T.A.H., Kang, G.-D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., Mccauley, P.G., Mcnitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ng, B.L., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Pinkard, D.C., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Rodriguez, A.C., Roe, P.M., Rogers, J., Bacigalupo, M.C.R., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Sohna, J.E.S., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., Mccooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R., Smith, A.J., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59.

8. Berglund, E.C., Kiialainen, A., Syvänen, A.C., 2011. Next-generation sequencing technologies and applications for human genetic history and forensics. Invest. Genet. 2, 23.
9. Borchiellini, C., Manuel, M., Alivon, E., Boury-Esnault, N., Vacelet, J., Parco, Y.L., 2001. Sponge paraphyly and the origin of Metazoa. J. Evol. Biol. 14, 171–179.
10. Boury-Esnault, N., 2006. Systematics and evolution of Demospongiae. Can. J. Zool. 84, 205–224.
11. Boute, N., Exposito, J.-Y., Boury-Esnault, N., Vacelet, J., Noro, N., Miyazaki, K., Yoshizato, K., Garrone, R., 1996. Type IV collagen in sponges, the missing link in basement membrane ubiquity. Biol. Cell 88, 37–44.
12. Cárdenas, P., Pérez, T., Boury-Esnault, N., 2012. Sponge Systematics Facing New Challenges. Advances in Sponge Science: Phylogeny, Systematics, Ecology. Adv. Mar. Biol. 61, 79–209.

13. Dabney, J., Meyer, M., 2012. Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. Biotech. 52, 87-94.
14. Dijk, E.L.V., Auger, H., Jaszczyszyn, Y., Thermes, C., 2014. Ten years of next-generation sequencing technology. Trends Genet. 30, 418–426.
15. Eisenstein, M., 2012. The battle for sequencing supremacy. Nat. Biotechnol. 30, 1023–1026.
16. Encyclopedia.com, 2016. Porifera Facts, information, pictures | Encyclopedia.com articles about Porifera. Dostupno sa http://www.encyclopedia.com/topic/Porifera.aspx (pristupljeno 6. 1. 2016).
17. Gazave, E., Lapébie, P., Ereskovsky, A.V., Vacelet, J., Renard, E., Cárdenas, P., Borchiellini, C., 2011. No longer Demospongiae: Homoscleromorpha formal nomination as a fourth class of Porifera. Hydrobiologia 687, 3–10.
18. Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S., Jaffe, D.B., 2010. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA 108, 1513–1518.
19. Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M.C., Mccombie, W.R., 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Res. 25, 1750–1756.
20. Harcet, M., Bilandžija, H., Bruvo-Mađarić, B., Ćetković, H., 2010. Taxonomic position of Eunapius subterraneus (Porifera, Spongillidae) inferred from molecular data – A revised classification needed? Mol. Phylogenet. Evol. 54, 1021–1027.
21. Head, Steven R., H. Kiyomi Komori, Sarah A. LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R. Salomon, and Phillip Ordoukhanian. "Library Construction for Next-generation Sequencing: Overviews and Challenges." Biotechniques 56, February 1, 2014, 61.
22. Hodkinson, B.P., Grice, E.A., 2015. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. Adv. Skin. Wound Care 4, 50–58.
23. Janussen, D., Tabachnick, K.R., Tendal, O.S., 2004. Deep-sea Hexactinellida (Porifera) of the Weddell Sea. Deep Sea Res. Part II Top. Stud. Oceanogr. 51, 1857–1882.

24. Leal, M.C., Puga, J., Serôdio, J., Gomes, N.C.M., Calado, R., 2012. Trends in the Discovery of New Marine Natural Products from Invertebrates over the Last Two Decades – Where and What Are We Bioprospecting? PLoS ONE 7.
25. Leys, S., Mackie, G., Reiswig, H., 2007. The Biology of Glass Sponges. Adv. Mar. Biol. 52, 1–145.
26. Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., Fan, W., 2011. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief. Funct. Genomics 11, 25–37.
27. Linnarsson, S., 2010. Recent advances in DNA sequencing methods – general principles of sample preparation. Exp. Cell. Res. 316, 1339–1343.
28. Loman, N.J., Quick, J., Simpson, J.T., 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Meth. 12, 733–735.
29. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.-W., Wang, J., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Giga Sci. 1, 18.
30. Madoui, M.-A., Engelen, S., Cruaud, C., Belser, C., Bertrand, L., Alberti, A., Lemainque, A., Wincker, P., Aury, J.-M., 2015. Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genomics 16.
31. Miller, J.R., Koren, S., Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. Genomics 95, 315–327.
32. Mitra, R., 1999. In situ localized amplification and contact replication of many individual DNA molecules. Nucleic Acids Res. 27, 34-39.
33. Müller, W.E.G., 1995. Molecular Phylogeny of Metazoa (Animals): Monophyletic Origin. Naturwissenschaften 82, 321–329.
34. Passarini, M.R.Z., Miqueletto, P.B., Oliveira, V.M.D., Sette, L.D., 2014. Molecular diversity of fungal and bacterial communities in the marine sponge Dragmacidon reticulatum. J. Basic Microbiol. 55, 207–220.
35. Pevzner, P.A., Tang, H., Waterman, M.S., 2001. An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. USA 98, 9748–9753.
36. Pop, M., 2009. Genome assembly reborn: recent computational challenges. Brief. Bioinform. 10, 354–366.

37. Quail, M., Smith, M.E., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13, 341.
38. Reitner, J., Wörheide, G., 2002. Non-Lithistid Fossil Demospongiae — Origins of their Palaeobiodiversity and Highlights in History of Preservation,U: Hooper, J.N.A., Van Soest, R.W.M. (Ur.), Systema Porifera. Kluwer Academic / Plenum Publishers, New York, 52–68.
39. Richards, G.S., Simionato, E., Perron, M., Adamska, M., Vervoort, M., Degnan, B.M., 2008. Sponge Genes Provide New Insight into the Evolutionary Origin of the Neurogenic Circuit. Curr. Biol. 18, 1156–1161.
40. Riesgo, A., Farrar, N., Windsor, P.J., Giribet, G., Leys, S.P., 2014. The Analysis of Eight Transcriptomes from All Poriferan Classes Reveals Surprising Genetic Complexity in Sponges. Mol. Biol. Evol. 31, 1102–1120.
41. Rokas, A., 2005. Animal Evolution and the Molecular Signature of Radiations Compressed in Time. Science 310, 1933–1938.
42. Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M., Marcais, G., Pop, M., Yorke, J.A., 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 22, 557–567.
43. Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74, 5463–5467.
44. Schatz, M.C., Delcher, A.L., Salzberg, S.L., 2010. Assembly of large genomes using second-generation sequencing. Genome Res. 20, 1165–1173.
45. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212.
46. Simpson, J.T., Durbin, R., 2011. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 22, 549–556.
47. Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M.E.A., Mitros, T., Richards, G.S., Conaco, C., Dacre, M., Hellsten, U., Larroux, C., Putnam, N.H., Stanke, M., Adamska, M., Darling, A., Degnan, S.M., Oakley, T.H., Plachetzki, D.C., Zhai, Y., Adamski, M., Calcino, A., Cummins, S.F., Goodstein, D.M., Harris, C., Jackson, D.J., Leys, S.P., Shu, S., Woodcroft, B.J., Vervoort, M., Kosik, K.S.,

Manning, G., Degnan, B.M., Rokhsar, D.S., 2010. The Amphimedon queenslandica genome and the evolution of animal complexity. Nature 466, 720–726.

48. Urban, J.M., Bliss, J., Lawrence, C.E., Gerbi, S.A., 2015. Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. bioRxiv. DOI: 10.1101/019281
49. Watson, M., Thomson, M., Risse, J., Talbot, R., Santoyo-Lopez, J., Gharbi, K., Blaxter, M., 2014. poRe: an R package for the visualization and analysis of nanopore sequencing data. Bioinformatics 31, 114–115.
50. Wörheide, G., Dohrmann, M., Erpenbeck, D., Larroux, C., Maldonado, M., Voigt, O., Borchiellini, C., Lavrov, D., 2012. Deep Phylogeny and Evolution of Sponges (Phylum Porifera). Advances in Sponge Science: Phylogeny, Systematics, Ecology. Adv. Mar. Biol. 61, 1–78.
51. Yandell, M., Ence, D., 2012. A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. 13, 329–342.
52. Zrzavy, J., Mihulka, S., Kepka, P., Bezdek, A., Tietz, D., 1998. Phylogeny of the Metazoa Based on Morphological and 18S Ribosomal DNA Evidence. Cladistics 14, 249–285.

8 Prilozi

Tablica 1. Statistike podataka sljedova dobivenih nakon uklanjanje adaptera i sljedova s nedovoljnom pouzdanošću očitanih nukleotida sekvencirane genomske DNA spužve *Eunapius subterraneus*

Knjižnica	Broj očitanih fragmenata	Prosječna duljina očitanih fragemenata (nk)	Ukupna duljina (nk)
Illumina MiSEQ 1	30424088	210.4	6402230662
Illumina MiSEQ 2	31266376	217.4	6797649472
Illumina LMP	42476708	100.1	4249816037

Tablica 2. Statistike podataka sljedova dobivenih nakon spajanja parova očitanih fragmenata knjižnica Illumina MiSEQ 1 i Illumina MiSEQ 2

Knjižnica	Illumina MiSEQ 1	Illumina MiSEQ 2
Broj parova sljedova	15212044	15633188
Broj združenih sljedova	11535742 (75.83%)	8087201 (51.731%)
Broj dvosmislenih parova sljedova	3666173 (24.10%)	7522729 (48.120%)
Bez rješenja	10129 (0.07%)	23258 (0.15%)

Tablica 3. Statistike podataka sljedova dobivenih nakon normalizacije i ispravljanja pogrešaka u pročišćenim nespojenim uparenim sljedovima sekvenciranih knjižnica Illumina MiSEQ 1 Illumina LMP.

Knjižnica	Broj očitanih fragmenata	Prosječna duljina očitanih fragemenata (nk)	Ukupna duljina (nk)
Illumina MiSEQ 1	6260434	198.2	1241021620
Illumina LMP	18902454	100.1	1891635359

Tablica 4. Statistika podataka sklopljenih genoma pomoću programa SGA za različite duljine preklapanja za knjižnicu Illumina MiSEQ 1.

Preklapanje (nk)	55	65	75	85	95	105	115	125	135	145	155
Broj neprekinutih sljedova	150556	154805	158841	163270	166947	170240	173264	176156	178999	182169	185293
Ukupna duljina neprekinutih sljedova	174505881	181398376	187246968	192455766	196530224	199691451	202014343	203577247	204369306	204578170	203933228

(pb)											
Duljina najduljeg neprekinutog slijeda (pb)	123305	91857	91854	91857	91860	82733	79357	79360	72874	67365	62918
N50 neprekinutih sljedova	1324	1351	1366	1363	1356	1343	1325	1302	1273	1234	1189

Tablica 5. Statistika podataka prekinutih i neprekinutih sljedova sklopljenih genoma pomoću programa SOAPdenovo2 za Illumina MiSEQ 1 knjižnicu

k-mer	Broj neprekinutih sljedova	Duljina najduljeg neprekinutog slijeda (pb)	Ukupna duljina neprekinutih sljedova (pb)	N50 neprekinutih sljedova	Broj prekinutih sljedova	Duljina najduljeg prekinutog slijeda (pb)	Ukupna duljina prekinutih sljedova (pb)	N50 prekinutih sljedova
35	63461	26506	70357725	1218	52370	54530	71452332	1805
37	64877	26541	72529624	1232	53495	55002	73625370	1830
39	66191	26543	74661130	1249	54536	65528	75755871	1857
41	67356	35719	76644273	1264	55289	65528	77747295	1887
43	68614	34119	78530357	1279	56560	65528	79602994	1893
45	69665	42593	80292636	1292	57400	65528	81353918	1917
47	70549	42593	81915240	1305	58007	65528	82970513	1940
49	71431	42593	83467643	1319	58843	65528	84497139	1957
51	72128	42578	84800662	1334	59548	65528	85794790	1969
53	72827	34119	86145415	1346	60216	65627	87105478	1986
55	73499	43209	87361686	1355	60836	65695	88292430	1995
57	74221	34119	88552038	1364	61567	65528	89450206	1998
59	74846	35719	89737844	1375	62286	66776	90601671	1999
61	75460	34119	90778797	1383	62930	67108	91611253	2001
63	76062	43217	91904178	1393	63575	67047	92699119	2006
65	76699	43219	92912260	1399	64335	67013	93663745	2003
67	77362	43441	93975680	1404	65227	67026	94674748	1993
69	77917	38649	94920721	1410	65940	66922	95576027	1989
71	78524	50602	95889726	1415	66773	67242	96494852	1977
73	79149	50604	96895205	1421	67606	67244	97454283	1969
75	79772	50606	97839893	1425	68452	67210	98353142	1951
77	80454	50608	98789347	1430	69327	67232	99262122	1946
79	81122	51091	99724546	1433	70269	67125	100150946	1931
81	81886	50612	100688712	1434	71263	67112	101073645	1916
83	82628	51531	101678337	1436	72222	67116	102027651	1905

85	83383	51533	102598734	1438	73189	67120	102909796	1891
87	84141	52927	103572916	1442	74085	54021	103854476	1885
89	85024	52927	104534438	1440	75071	59320	104789775	1870
91	86021	52927	105528964	1435	76183	59127	105756698	1851
93	86861	52927	106455166	1432	77272	53348	106653805	1829
95	87765	52927	107473965	1429	78394	53348	107645908	1815
97	88788	52927	108450856	1427	79553	53348	108595165	1799
99	89831	53260	109238021	1419	80881	53260	109359123	1772
101	90291	46976	100288706	1236	82129	46976	100357056	1493
103	91511	46976	101720539	1235	83448	46976	101769287	1486
105	92885	46976	103178528	1234	84889	46976	103206861	1475
107	94261	46976	104633811	1232	86310	46976	104639392	1468
109	95663	46976	106077564	1228	87766	46976	106060182	1457
111	97244	53260	107573985	1223	89364	53260	107533717	1445
113	98807	53372	109024552	1216	90998	53372	108960483	1433
115	100472	53374	110498264	1208	92686	53374	110411314	1420
117	102084	53376	111948965	1203	94332	53376	111840696	1405
119	103753	53378	113406859	1196	96017	53378	113274760	1394
121	105275	45948	114768101	1193	97543	45948	114611235	1383
123	106882	45950	116162700	1185	99171	48457	115982650	1370
125	108560	45776	117546779	1179	100958	48457	117342477	1355
127	110546	45778	119118655	1170	102921	53796	118892824	1341

Tablica 6. Statistika podataka neprekinutih sljedova sklopljenih genoma s različitom jačinom preklapanja za 89-mer pomoću programa SOAPdenovo2 za knjižnicu Illumina MiSEQ 1

Snaga za spajanje	**Broj neprekinutih sljedova**	**Duljina najduljeg neprekinutog slijeda (pb)**	**Ukupna duljina neprekinutih sljedova (pb)**	**N50 neprekinutih sljedova**
0	83535	31448	91934925	1222
1	85024	52927	104534438	1440
2	85567	52927	105052230	1437
3	85596	52927	105088749	1437

Tablica 7. Statistika podataka prekinutih sljedova sklopljenih genoma s različitom jačinom preklapanja za 89-mer pomoću programa SOAPdenovo2 za knjižnicu Illumina MiSEQ 1

Snaga za spajanje	**Broj prekinutih sljedova**	**Duljina najduljeg prekinutog slijeda (pb)**	**Ukupna duljina prekinutih sljedova (pb)**	**N50 prekinutih sljedova**
0	74517	31448	92140812	1534
1	75071	59320	104789775	1870
2	75496	59320	105310756	1870
3	75441	59320	105349572	1874

Životopis

OSOBNE INFORMACIJE **Šribar Dora**

Hećimovićeva 3, 10000 Zagreb (Hrvatska)

+385981812590

dora.sribar@gmail.com
dora.sribar@zg.htnet.hr

Spol Žensko |

Datum rođenja 05/10/1989 |

Državljanstvo: hrvatsko

OBRAZOVANJE I OSPOSOBLJAVANJE

10/2013–02/2016 **Diplomski studij molekularne biologije**

Biološki odsjek, Prirodoslovno-matematički fakultet, Sveučilište u Zagrebu, Zagreb (Hrvatska)

Znanje i vještine vezane uz programiranje, statistiku i strojno učenje i analizu podataka

2008–2013 **Integrirani preddiplomski i diplomski studij Farmacije**

Farmaceutsko-biokemijski fakultet, Sveučilište u Zagrebu, Zagreb (Hrvatska)

Znanja i vještine vezane uz farmaceutsku profesiju

KONFERENCIJE I ŠKOLE

- 07 -24.08.2015. Research summer school in statistical omics, Split, Croatia
- 22-28.4.2013. 36. EPSA congress, Catania, Italy
- 2012. Somborac Bačura A, Rumora L, Šribar D, Popović-Grle S, Čepelak I, Žanić Grubišić TMMP-9 and TIMP-1 concentrations in plasma of patients with chronic obstructive pulmonary disease, Dubrovnik 2nd EFCC-UEMS Congress (congress poster)
- 2012. FARMEBS, Symposium of students of pharmacy and medical biochemistry

PRIZNANJA I NAGRADE

- Godišnja dekanova nagrada (2011/2012.) za rad "Arilesterazna aktivnost paraoksonaze 1 u bolesnika s kroničnom opstrukcijskom plućnom bolesti"
    - Mentor: Lada Rumora, dr.sc.
- Dobitnica stipendije za najuspješnije studente 2011-2013, Ministarstvo obrazovanja, znanosti i športa Republike Hrvatske
- 2005: 1.mjesto na državnom natjecanju iz kemije