Računalna optimizacija sljedova genomske DNA spužve Eunapius subterraneus sekvenciranih tehnologijom nanopora

Jelić Matošević, Zoe

Sveučilište u Zagrebu

Prirodoslovno – matematički fakultet

Biološki odsjek

Zoe Jelić Matošević

Računalna optimizacija sljedova genomske DNA spužve *Eunapius subterraneus* sekvenciranih tehnologijom nanopora

Diplomski rad

Zagreb, 2018.

Ovaj rad je izrađen pri Zavodu za molekularnu biologiju, Biološki odsjek, PMF i Laboratoriju za molekularnu genetiku Zavoda za molekularnu biologiju, Institut Ruđer Bošković, pod vodstvom prof. dr. sc. Kristiana Vlahovičeka i dr. sc. Helene Ćetković, predan je na ocjenu Biološkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu radi stjecanja zvanja magistra molekularne biologije.

Zahvaljujem prof. dr. sc. Kristianu Vlahovičeku i dr.sc. Heleni Ćetković na mentorstvu, trudu i vremenu uloženom u ovaj diplomski rad.

Zahvaljujem obitelji na podršci i motivaciji da završim ovaj studij, Maji na beskrajnom entuzijazmu i savjetovanju i Kikiju na infuziji motivacije u ključnom trenutku.

Posebno hvala Mikiju jer je uvijek bio tu.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Biološki odsjek Diplomski rad

RAČUNALNA OPTIMIZACIJA SLJEDOVA GENOMSKE DNA SPUŽVE *EUNAPIUS SUBTERRANEUS* SEKVENCIRANIH TEHNOLOGIJOM NANOPORA

Zoe Jelić Matošević

Rooseveltov trg 6, 10000 Zagreb, Hrvatska

Ogulinska špiljska spužvica, *Eunapius subterraneus* Sket & Velikonja, 1984 (Porifera, Demospongia), jedina je stigobiontna slatkovodna spužva na svijetu. Spužve (Porifera) su najstarije živuće koljeno životinja. Karakterizira ih jednostavna tjelesna građa te nedostatak pravih tkiva i organa. Zbog filogenetski bazalnog položaja u odnosu na ostale životinje, spužve su idealni modelni organizmi za proučavanje evolucije gena i genoma te razvoja višestaničnosti unutar životinja (Metazoa). Kako bi to bilo moguće potrebno je sekvencirati i sklopiti genome spužvi koje želimo proučavati. Metodom sekvenciranja tehnologijom nanopora u kombinaciji s metodom reverzibilnog zaustavljanja sinteze (Illumina) mogu se sklopiti čak i repetitivni genomi viših eukariota, ali često je potrebno ispraviti sljedove iz nanopora prije samog sklapanja genoma. Dodatan problem u radu sa spužvama predstavlja kontaminacija sekvenciranih knjižnica. U ovom radu izolirala sam visokomolekularnu genomsku DNA *Eunapius subterraneus* iz primorfa te sam pokazala kako ovaj pristup doprinosi smanjenju kontaminacije u uzorcima. Sljedove iz nanopora ispravila sam koristeći program NaS te sam dokazala njihovu točnost pretragom baze neredundantnih proteinskih sljedova programom DIAMOND. Za iskorištavanje punog potencijala tehnologije sekvenciranja nanoporama u budućnosti potrebno je razvijati algoritme prilagođene specifičnom profilu greške ove tehnologije.

(41 stranica, 15 slika, 10 tablica, 34 literaturnih navoda, jezik izvornika: hrvatski)

Rad je pohranjen u središnjoj biološkoj knjižnici

Ključne riječi: *Eunapius subterraneus*, tehnologija sekvenciranja nanoporama, sklapanje genoma

Voditelji: prof. dr. sc. Kristian Vlahoviček
dr. sc. Helena Ćetković

Ocjenitelji: prof. dr. sc. Kristian Vlahoviček
izv. prof. dr. sc. Damjan Franjević
doc. dr. sc. Tomislav Ivanković

Zamjena: doc. dr. sc. Rosa Karlić

Rad prihvaćen: 19. rujna, 2018.

BASIC DOCUMENTATION CARD

University of Zagreb

Faculty of Science

Division of Biology

Graduation thesis

COMPUTER OPTIMIZATION OF GENOMIC DNA NANOPORE SEQUENCES FROM THE SPONGE *EUNAPIUS SUBTERRANEUS*

Zoe Jelić Matošević

Rooseveltov trg 6, 10000 Zagreb, Croatia

The Ogulin cave sponge, *Eunapius subterraneus* Sket & Velikonja, 1984 (Porifera, Demospongia), is the only freshwater stygobiont sponge in the world. Sponges (Porifera) are the oldest living animal phylum. Their bodies have a simple structure and lack true tissues and organs. Because of their basal position among the Metazoa they are useful for researching gene and genome evolution as well as the evolution of multicellularity in the phylum Metazoa. To do that we need to sequence and assemble their genomes. The nanopore sequencing technology in combination with Illumina's reversible-terminator-based sequencing by synthesis can help to produce good assemblies even of repetitive higher-eukaryote genomes. However, a correction step for the nanopore reads is often necessary before assembly. When working with sponges sample and library contamination present an additional problem. I have isolated high-molecular-weight genomic DNA from sponge primmorphs of *Eunapius subterraneus* and shown that DNA isolation from primmorphs can drastically reduce sample contamination. I have corrected nanopore reads using the programme NaS and proven the accuracy of the corrected reads by comparing them to the non-redundant protein database using DIAMOND. In the future for nanopore sequencing to live up to its potential, algorithms that can deal with the specific error profile of nanopore reads will have to be developed.

(41 pages, 15 figures, 10 tables, 34 references, original in: croatian)

Thesis deposited in the Central Biological Library

Key words: *Eunapius subterraneus*, nanopore sequencing, genome assembly

Supervisors: prof. dr. sc. Kristian Vlahoviček
dr. sc. Helena Ćetković

Reviewers: Professor Kristian Vlahoviček
Assoc. Prof. Damjan Franjević
Asst. Tomislav Ivanković

Substitution: Asst. Rosa Karlić

Thesis accepted: September 19, 2018.

Kratice

PCR	Lančana reakcija polimerazom
EDTA	Etilendiamintrioctena kiselina
NFW	Voda bez nukleaze (eng. ***n**uclease **f**ree **w**ater*)
ITS2	eng. *internal transcribed spacer 2*
BLAST	eng. *Basic Local Alignment Search Tool*
nt	Nukleotid
TAE	Tris-acetatni EDTA pufer
ddNTP	Dideoksi nukleotid
ddATP	Dideoksi adenin
ddTTP	Dideoksi timin
ddGTP	Dideoksi gvanin
ddCTP	Dideoksi citozin

Sadržaj

1 Uvod

1.1 Modelni organizam – ogulinska špiljska spužvica *Eunapius subterraneus*

1.1.1 Opća obilježja spužvi

Spužve (koljeno Porifera) se smatraju jednima od najjednostavnijih višestaničnih životinja koje su se odvojila od drugih Metazoa prije više od 800 milijuna godina (Love *i sur.*, 2009.). Tijelo spužve je građeno kao sustav kanala kroz koji se filtrira voda kako bi se hvatale hranjive čestice. U odraslom stanju spužve su sesilni organizmi bez pravih tkiva i organa kao i prepoznatljivih osjetilnih ili živčanih struktura. Unatoč jednostavnoj građi, spužve posjeduju velik broj gena ključnih za razvoj i funkcioniranje višestaničnih životinja (Srivastava *i sur.*, 2010.). Iz ovog razloga spužve su zanimljiv model za istraživanje nastanka višestaničnosti i stanične diferencijacije kod životinja.

Tijelo spužve građeno je od tri sloja stanica. Unutrašnjost tijela obložena je hoanocitama, bičastim stanicama s okovratnikom zaslužnima za strujanje vode kroz tijelo spužve. Vanjski sloj je obložen poligonalnim stanicama pinakocitama, a središnji sloj, mezohil, sadrži kolagen, elemente skeleta i nekoliko tipova pokretnih stanica (Leys i Hill, 2012.). Skeleti spužvi građeni su od kombinacija silicijevog dioksida, kalcijevog karbonata i/ili spongina. Oblik spikula – struktura koje čine skelet – najvažnije je morfološko svojstvo za klasifikaciju spužvi.

Koljeno Porifera taksonomski je podijeljeno u četiri razreda (Wörheide *i sur.*, 2012.);

1. Calcarea (vapnenjače) imaju skelet građen od kalcijevog karbonata i uglavnom nastanjuju topla i plitka morska staništa.
2. Hexactinellida (staklače) su spužve čiji je skelet građen od silicijevog dioksida. Većina predstavnika ove skupine nastanjuje duboka mora. Specifične su po sincitijskom tkivu, koje čini vanjski sloj stanica.
3. Demospongia (kremenorožnjače) su najveća skupina spužava i nastanjuju brojna staništa, uključujući polarne vode i slatkovodna staništa. Skelet im je građen od proteina spongina ili silicijevog dioksida, ili oboje.
4. Homoscleromorpha su se donedavno svrstavale među kremenorožnjače. Ova mala skupina spužvi je također i jedina čiji pripadnici imaju bazalnu membranu u odraslom stadiju.

Dodatno, spužve se mogu po morfološkim značajkama podijeliti na tri osnovna tipa građe. Spužve s askonskim tipom građe su radijalno simetrične i tijela im se sastoje od spongocela,

šupljine koja je s unutarnje strane obložena hoanocitama, te otvora oskuluma. Sikonoidne spužve su također radijalno simetrične, no unutar spongocela postoje radijalni kanalići, također obloženi hoanocitima. Leukon je najsloženiji tip građe, a ovakve spužve imaju velik broj kanala koji povezuju površinu tijela s komoricama obloženim hoanocitama u unutrašnjosti tijela. Spužve ovakvog tipa građe nisu radijalno simetrične.

Morfološka podjela spužvi nije povezana s taksonomskom već s načinom prehrane određene spužve, te primjerice kod vapnenjača nalazimo sva tri oblika (Leys i Hill, 2012.).

1.1.2 Spužve kao modelni organizmi

Početkom prošlog stoljeća Wilson je prvi uspio rekonstruirati čitavu spužvu iz stanica, i pritom je pokazao da je agregacija spužvinih stanica vrsno specifična (Wilson, 1910.). U međuvremenu je pokazano da spužve posjeduju i auto- i alorekogniciju. Autorekognicija se odnosi na sposobnost organizma da prepoznaje vlastite stanice, a alorekognicija je sposobnost prepoznavanja stranih stanica. Eskperimenti na spužvama *Suberites domuncula* i *Geodia cyonium* pokazali su da spužve posjeduju autorekogniciju i alorekogniciju na razini jedinke (Müller i Müller, 2003.).

Agregati spužvinih stanica su koristan model za proučavanje de-, re- i transdiferencijacije spužvinih stanica (Lavrov i Kosevich, 2016.). Mogućnost proučavanje ponašanja spužvinih stanica i agregata u kontroliranim laboratorijskim uvjetima je od velikog značaja zbog njihovog bazalnog položaja naspram drugih skupina višestaničnih životinja. Upravo bazalne grupe su ključne za datiranje nastanka svojstava specifičnih za čitave skupine, kao i stvaranje slike o posljednjem zajedničkom pretku vrsti u tim skupinama.

Sekvenciranje spužve *Amphimedon queenslandica* po prvi put je omogućilo uvid u genom pripadnika skupine morfološki najjednostavnijih životinja, te je pokazano da je u sadržaju i strukturi vrlo blizak genomima složenijih životinja (Srivastava *i sur.*, 2010.).

Amphimedon posjeduje genetsku podlogu nužnu za uspostavu šest ključnih obilježja višestaničnosti kod životinja; regulaciju staničnog ciklusa i rasta, programiranu staničnu smrt, međustaničnu adheziju i adheziju na izvanstanični matriks, razvojne signalne puteve i alorekogniciju.

Dijelovi genetske mašinerije nužne za uspostavu ovih šest obilježja prisutni su i u jednostavnijim eukariotima, no određeni geni su specifični upravo za životinje.

Razlučivanje trenutka u kojem se pojavilo neko obilježje ili gen koristeći pritom samo jednog predstavnika skupine nije veoma pouzdano. Primjerice, skupine transkripcijskih faktora T-box i Runx smatralo se specifičnima za višestanične životinje jer nisu pronađene u genomu *Monosiga brevicolis*, pripadnika skupine Choanoflagellata –. najbližih jednostaničnih srodnika životinja. Analiza genoma *Capsaspora owczarzaki* iz skupine Filasterea (također bliskih jednostaničnih srodnika životinja) otkrila je da su te skupine gena nastale puno ranije (Sebé-Pedrós *i sur.*, 2011.). Iz navedenoga je očito je da će razlučivanje događaja koji su doveli do nastanka višestaničnosti i rane evolucije Metazoa zahtijevati analizu većeg broja dosad neistraženih genoma.

1.1.3 Eunapius subterraneus

Ogulinsku spiljsku spužvicu *Eunapius subterraneus* prikazanu na slici 1 prvi put su opisali Sket i Velikonja 1984. (Sket i Velikonja, 1986.) *E. subterraneus* je jedini do danas poznati predstavnik spužvi među stigobiointima i prema kriterijima IUCN uvrštena je na popis ugroženih vrsti. Dosad je pronađena na šest lokaliteta u špiljama u blizini Ogulina (Hrvatska): Tounjčica špilja, Mikašinovića špilja, Rudnica špilja VI, Mandelaja, Crnačka špilja i Izvor špilja Gojak (Bilandžija *i sur.*, 2007.).

Slika 1 Ogulinska spiljska spužvica *Eunapius subterraneus*. Preuzeto sa https://www.hbsd.hr/SkupineZ_spuzvica.html

Spada u razred Demospongiae, podred Spongillina. Bijele je boje i krhke građe, te ne posjeduje mikrosklere (spikule nevidljive golom oku). Silikatne markosklere (spikule vidljive golim okom) su igličastog, blago zaobljenog oblika s urezima. Iako je izvorno po morfološkim obilježjima smještena u rod Eunapius, molekularne analize 18S rDNA, ITS2 i mitohondrijske citokrom oksidaze I, smještaju *E. subterraneus* odvojeno od ostalih pripadnika roda *Eunapius*

u skupinu s pripadnicima roda *Ephydatia* i nekoliko drugih vrsta slatkovodnih spužvi (Harcet *i sur.*, 2010.).

Jedan od uzroka nerazriješene klasifikacije ove i srodnih spužvi je plastičnost fenotipa kod spužvi koji često varira ovisno o okolišnim uvjetima, te stoga nije začuđujuće da se i izgled jedinki *E. subterraneus* razlikuje među lokalitetima (Bilandžija *i sur.*, 2007.). Upravo kod ovakvih slučajeva molekularne analize su ključne za ispravnu taksonomsku klasifikaciju vrsta.

1.2 Metode sekvenciranja

Otkriće da je DNA nasljedna molekula te Watsonovo i Crickovo razotkrivanje strukture DNA pred znanstvenu zajednicu postavilo je novi izazov: određivanje sljedova nukleotida u molekulama DNA. Prvi veliki proboj u sekvenciranju DNA nastupio je s pojavom Sangerove dideoksi metode sekvenciranja 1977., koju danas svrstavamo pod metode sekvenciranja prve generacije (Heather i Chain, 2016.).

U nastojanju da se poveća broj molekula koje se mogu istovremeno sekvencirati i time ubrza proces sekvenciranja osmišljene su metode sekvenciranja druge generacije, od kojih će ovdje biti opisana metoda reverzibilnog zaustavljanja sinteze DNA (Illumina). Obilježja metoda sekvenciranja druge generacije su masivno paralelno sekvenciranje velikog broja molekula DNA umnoženih lančanom reakcijom polimeraze, ali i mala duljina sekvenciranih sljedova u usporedbi sa Sangerovom metodom.

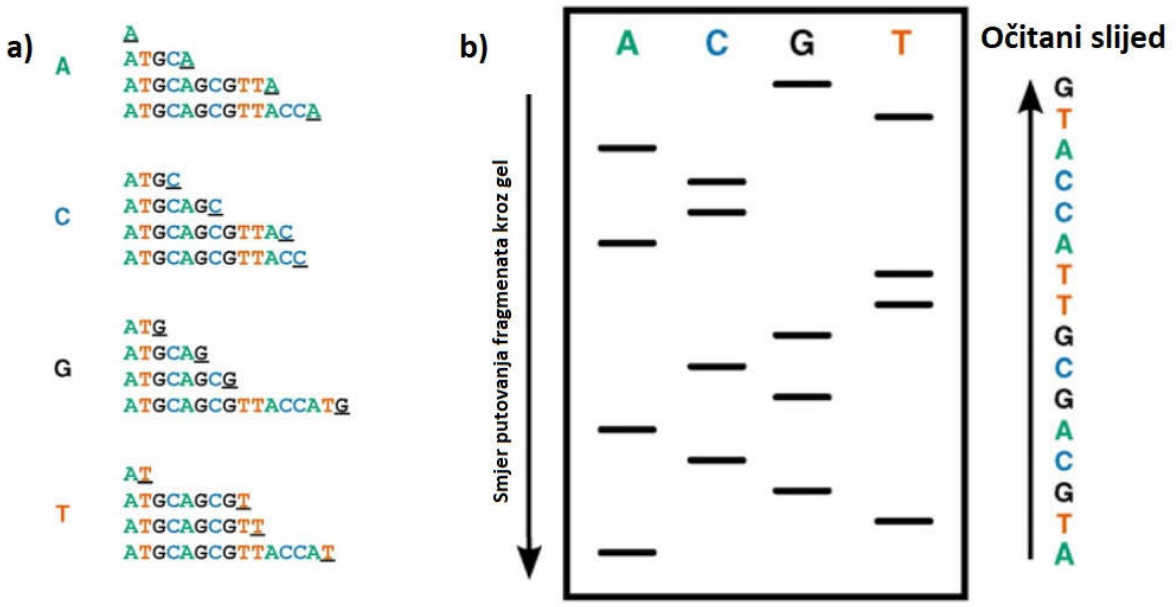
U najnovije vrijeme razvijaju se metode koje sekvenciraju pojedinačne molekule DNA te postižu velike duljine sekvenciranih sljedova, no uz manju točnost. U ovom radu bit će opisano sekvenciranje tehnologijom nanopora na uređaju ONT MinION tvrtke Oxford Nanopore Technologies.

1.2.1 Metode sekvenciranja prve generacije (Sangerova dideoksi metoda)

Sangerova dideoksi metoda sekvenciranja temelji se na ugradnji dideoksinukleotida koji ne posjeduju 3'-OH skupinu te na taj način onemogućavaju daljnju ugradnju nukleotida u rastući lanac. Omjer dNTP-ova i ddNTP-ova u reakciji je ugođen tako da se ddNTP-ovi ugrađuju s puno manjom učestalošću. U reakciji tako nastaju molekule DNA različitih duljina.

U svom izvornom obliku, sekvenciranje po Sangeru provodi se u četiri odvojene reakcije. Svaka od reakcija sadržava DNA polimerazu, jednu početnicu, kalup, dNTP-ove te jedan od četiri ddNTP-ova. Na taj način osigurano je da u svakoj reakciji svaki fragment završava prisutnim ddNTP-om. Dobiveni fragmenti razdvajaju se poliakrilamidnom gel elektroforezom

te se iz gela može očitati slijed DNA. Shematski prikaz gela prikazan je na slici 2. Vizualizacija gela postiže se autoradiografijom, tako da su prije početka reakcije početnice, dNTP-ovi ili ddNTP-ovi radioaktivno obilježeni.



Slika 2 Sekvenciranje po Sangeru, pod a) su prikazani fragmenti prisutni u svakoj od četiri reakcije za molekulu DNA čiji je slijed očitan iz gela pod b). Preuzeto i prilagođeno iz (Heather i Chain, 2016.)

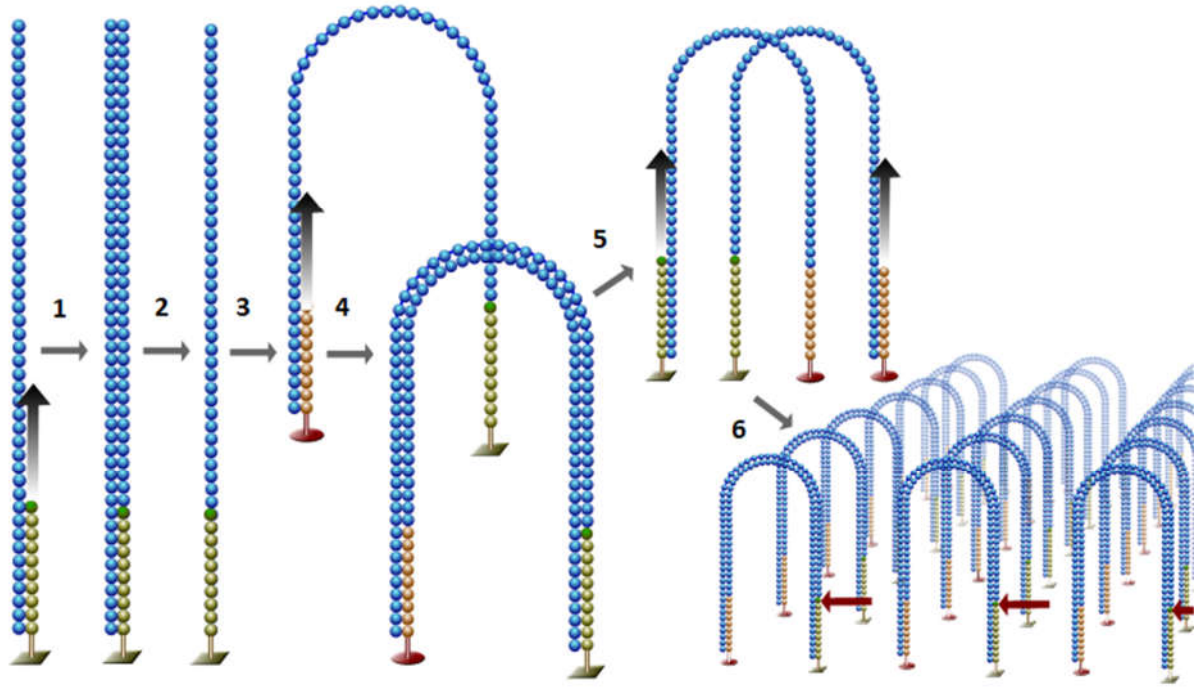
Sangerova dideoksi metoda poboljšana je uvođenjem fluorescentno a ne radioaktivno obilježenih nukleotida, što je omogućilo provođenje sekvenciranja u jednoj reakcijskoj smjesi koristeći četiri različita fluorescentna biljega za četiri tipa ddNTP-ova. Dodatno je optimirana korištenjem kapilarne elektroforeze, koja je omogućila veću razlučivost fragmenata, a time i sekvenciranje fragmenata duljine i do 1000 nt (Heather i Chain, 2016.).

1.2.2 Metoda reverzibilnog zaustavljanja sinteze DNA (Illumina)

Prvi uređaj za sekvenciranje tvrtke Illumina pojavio se na tržištu 2006., a danas je sekvenciranje metodom Illumina najraširenija metoda sekvenciranja druge generacije. Temelji se na umnažanju fragmenata DNA na čvrstoj podlozi (eng. *bridge amplification*) i reverzibilnom zaustavljanju sinteze DNA (Reuter *i sur.*, 2015.).

Sekvenciranje se odvija na staklenim pločicama na kojima su učvršćena dva tipa međusobno komplementarnih oligonukleotida. Prvi korak u pripremi uzoraka za sekvenciranje je fragmentacija DNA, nakon čega slijedi ligacija adaptera na dobivene fragmente. Nanošenjem ovako pripremljenih fragmenata na pločicu u pažljivo baždarenoj koncentraciji osigurava se razlučivost različitih fragmenata pri sekvenciranju (Goodwin *i sur.*, 2016.).

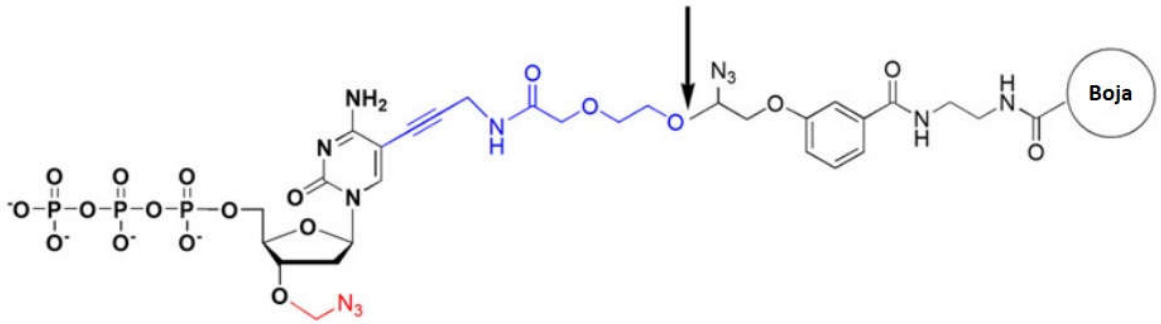
Fragmenti se pomoću adaptera komplementarno vežu na jedan od dva tipa učvršćenih oligonukleotida. Nakon prvog ciklusa umnažanja, originalni fragmenti se ispiru i nastaje struktura nalik mostu (vidi sliku 3) koja omogućava da se za drugi ciklus umnažanja kao početnica koristi drugi od dva tipa učvršćenih nukleotida. Naizmjeničnim stvaranjem struktura nalik mostu i umnažanjem s oba tipa oligonukleotida nastaju nakupine identičnih dvolančanih molekula DNA.



Slika 3 Umnažanje DNA fragmenta na Illumina pločici. 1. Fragment se pomoću adaptera veže na pločicu. U prvom ciklusu umnažanja zeleno označeni oligonukleotid se koristi kao početnica. 2. Originalni DNA fragment se ispire. 3. i 4. Nastaje struktura nalik mostu kako bi se u drugom ciklusu umnažanja kao početnica koristio narančasto označeni tip oligonukleotida. 5. i 6. Naizmjeničnim umnažanjem s narančastog i zeleno označenog tipa oligonukleotida nastaje lokalizirana nakupina identičnih dvolančanih DNA fragmenata. Preuzeto i prilagođeno sa https://en.wikipedia.org/wiki/File:DNA_Sequencing_Bridge_Amplification.png

Nakon amplifikacije cijepa se jedan od lanaca kako bi na pločici ostale samo jednolančane molekule DNA. Na fragmente se veže početnica za sekvenciranje i potom se kroz broj ciklusa ovisan o specifičnoj platformi ponavljaju koraci: dodaju se reaktanti (fluorescentno obilježeni deoksinukleotidi, DNA polimeraza i kofaktori), nevezani nukleotidi se ispiru, detektira se fluorescentni signal i uklanjaju se skupine koje blokiraju daljnju ugradnju nukleotida (Goodwin *i sur.*, 2016.). Cijeli proces istovremeno se odvija na milijunima fragmenata DNA.

Na slici 4 prikazan je modificirani nukleotid kakav se koristi za sekvenciranje metodom reverzibilnog zaustavljanja sinteze. Nakon ugradnje modificiranog nukleotida azidna skupina ($-N_3$) sprječava daljnju ugradnju nukleotida.



Slika 4 3'-*O*-azidometil-dNTP. Strelicom je označeno mjesto cijepanja fluorescentne skupine. Preuzeto i prilagođeni iz (Chen *i sur.*, 2013.).

Nakon pobuđivanja i detekcije fluorescentnog signala, azidna skupina se uklanja ostavljajući slobodnu 3'-OH skupinu. Uklanja se i boja kako ne bi interferirala sa signalom u idućem ciklusu. Na slici 4 je strelicom označeno mjesto cijepanja fluorescentne skupine, a plavo označeni dio molekule predstavlja „molekularni ožiljak" koji nije prisutan u prirodnim dNTP-ovima.

Po završetku predviđenog broja ciklusa sekvenciranje se ponavlja s drugog kraja molekule, te se tako dobivaju knjižnice sparenih sljedova (eng. *paired end*). Ukoliko su fragmenti više nego duplo dulji od broja nukleotida koje platforma može očitati dobiva se informacija o međusobnoj poziciji i udaljenosti parova sljedova.

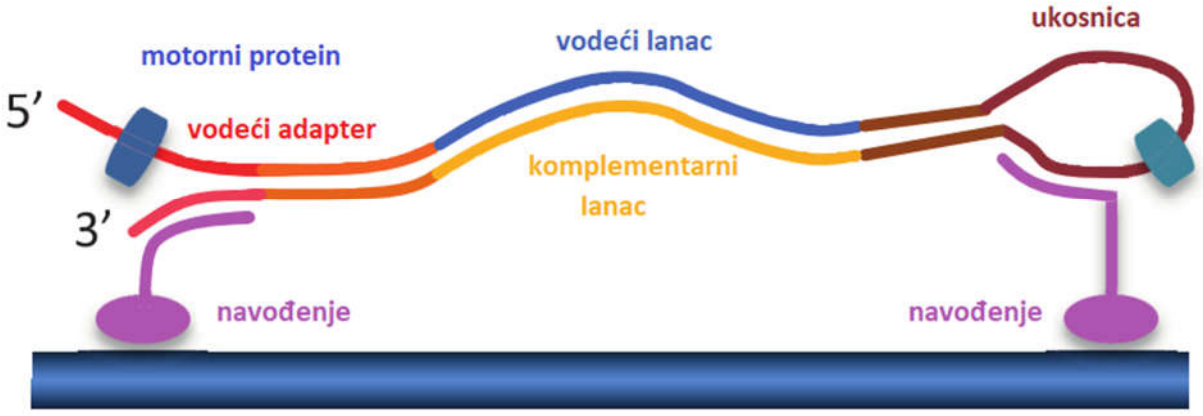
Sekvenatori tvrtke Illumina (MiSeq, HiSeq) mogu sekvencirati sljedove duljine do 300 nt u manje od 3 dana uz grešku od samo 0.1%, pritom očitavajući, ovisno o platformi, između 10^{10} i 10^{12} baznih parova (Goodwin *i sur.*, 2016.).

1.2.3 Metoda sekvenciranja tehnologijom nanopora (The Oxford Nanopore Technologies MinION, ONT)

Sekvenciranje tehnologijom nanopora svrstava se u metode sekvenciranja treće generacije jer se sekvenciraju pojedinačne molekule velike duljine. Ono što ju izdvaja od svih ostalih metoda je što ne koristi komplementarno sparivanje nukleotidnih baza za sekvenciranje. Umjesto toga, mjeri se strujni signal koji nastaje prolaskom jednolančanog polimera (DNA, RNA ili čak proteina) kroz biološku poru učvršćenu u električno nabijenoj membrani.

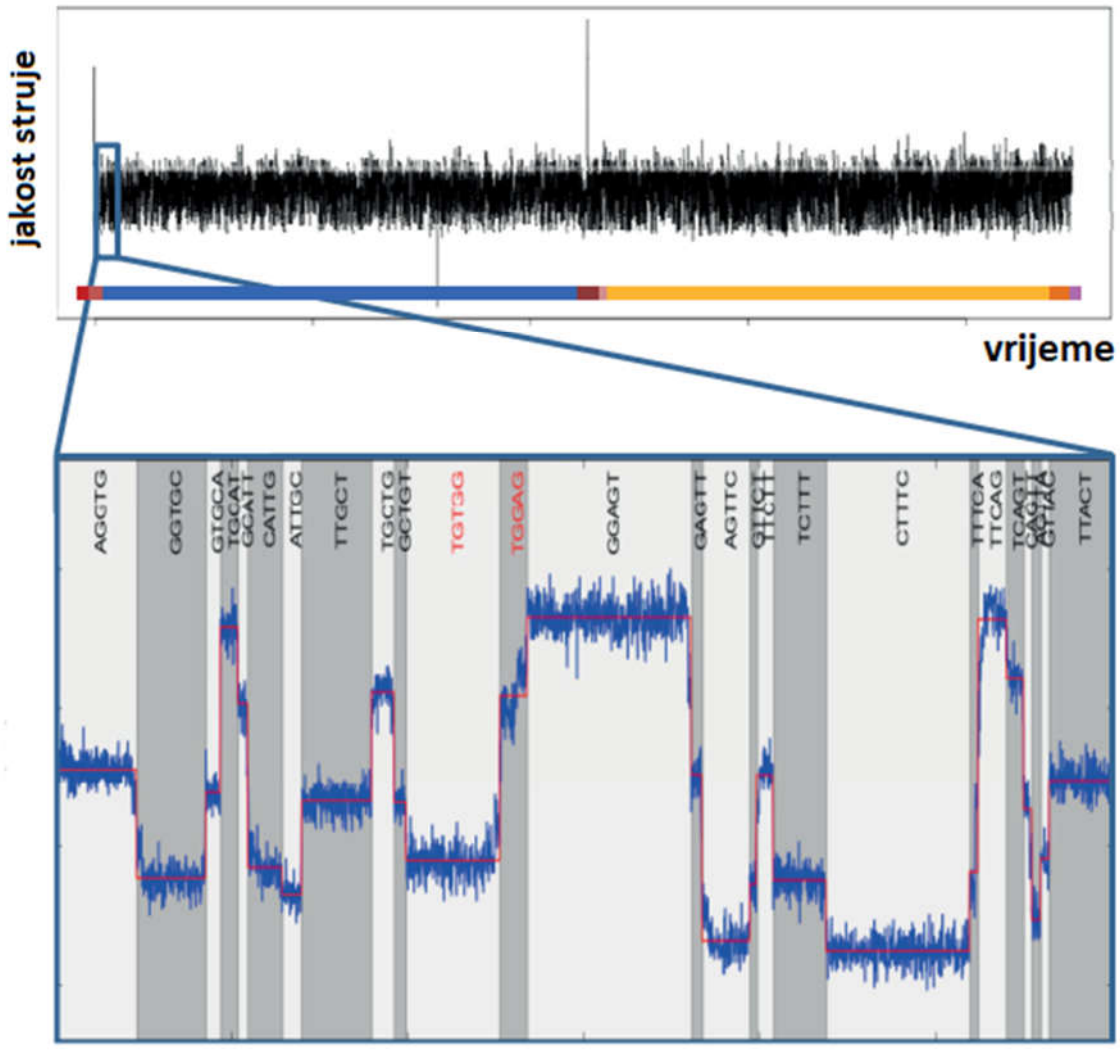
Ova metoda sekvenciranja je tehnologija u razvoju te se protokoli za pripremu uzoraka, kemija za sekvenciranje i program za očitavanje slijeda nukleotida iz signala stalno unaprjeđuju. Najveći nedostatak sekvenciranja tehnologijom nanopora je razmjerno niska razina točnosti dobivenih sljedova i velik udio sljedova odbačenih zbog niske kvalitete, no sa svakim novim izdanjem tehnologije vidljivi su napredci u navedenim mjerilima (Ip *i sur.*, 2015.; Jain *i sur.*, 2017.; Lu *i sur.*, 2016.).

Priprema knjižnica za sekvenciranje tehnologijom nanopora zahtijeva ligaciju dva tipa adaptera na dvolančanu DNA. Vodeći adapter je oblika slova Y te jedan njegov dio služi za navođenje DNA u blizinu pore, a drugi dio se veže na motorni protein, koji odmotava dvolančanu DNA (Slika 5). Drugi adaptor je u obliku ukosnice. Nakon što vodeći lanac DNA prođe kroz poru, ukosnica osigurava da će nakon njega kroz poru proći komplementarni lanac.



Slika 5 Prikaz DNA fragmenta na protočnoj ćeliji s vezanim adapterima i proteinima. Preuzeto i prilagođeno iz (Ip *i sur.*, 2015.)

Očitavanje slijeda nukleotida iz signala događa se u stvarnom vremenu (no može se i naknadno ponoviti). Promjene u izmjerenoj struji dijele se na događaje – svaki događaj predstavlja pomak molekule u pori za jedan nukleotid. Iz niza događaja i pridruženih vrijednosti struje pomoću metoda strojnog učenja se iščitava slijed nukleotida, pritom pretpostavljajući da na protok struje kroz poru u svakom trenutku utječe 5 ili 6 nukleotida koji se u njoj nalaze (Ip *i sur.*, 2015.). Primjer očitanog signala prikazan je na slici 6. Očitani sljedovi za vodeći i komplementarni lanac se, ukoliko je to moguće, kombiniraju u tzv. dvodimenzionalni (2D) slijed. Ukoliko to nije moguće dobivaju se jednodimenzionalni (1D) sljedovi.



Slika 6 Interpretacija promjena u protoku struje korištenjem skrivenih Markovljevih modela. Preuzeto i prilagođeno iz (Ip *i sur.*, 2015.)

U najnovijoj verziji tehnologije (R9), za očitavanje slijeda nukleotida koristi se rekurzivna neuralna mreža (Jain *i sur.*, 2017.).

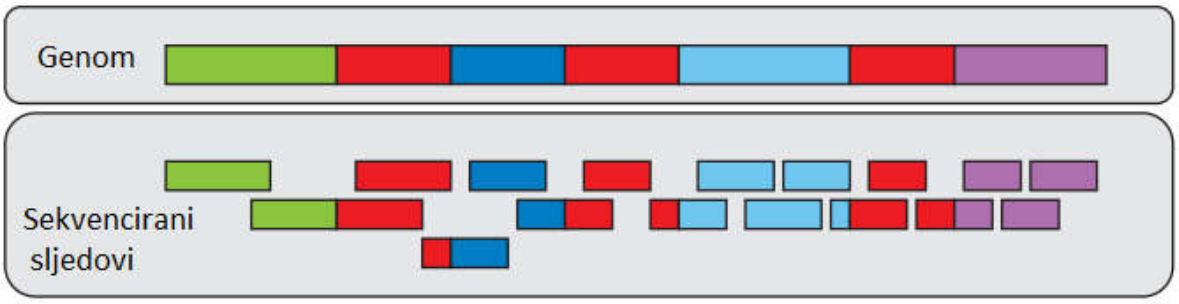
Početkom 2017. uveden je i novi protokol za pripremu knjižnica za sekvenciranje, koji ne uključuje kovalentno povezivanje komplementarnih lanaca ukosnicom. Umjesto toga oslanja se na to da će nakon prolaska vodećeg lanca u oko 60% slučajeva komplementarni lanac slijediti ulaskom u poru. Tako dobiveni sljedovi nazivaju se $1D^2$ sljedovi (Leggett i Clark, 2017.).

Veza između očitanog signala i DNA slijeda koji producira taj signal je kod tehnologije sekvenciranja nanoporama puno složenija nego kod tehnologija druge generacije kao što je metoda reverzbilinog zaustavljanja sinteze (Illumina). Glavni uzrok tome je što jedan signal

predstavlja više nukleotida, i tek se istodobnom analizom više susjednih signala može očitati baza na pojedinoj poziciji. Kod metode reverzibilnog zaustavljanja sinteze (Illumina) taj je odnos pak puno izravniji – svaki fluorescentni signal predstavlja jedan nukleotid. Brzina prolaska DNA kroz poru, koja nije uvijek jednaka, predstavlja dodatan problem pri očitavanju nukleotida kod metode sekvenciranja nanoporama. Kako bi se ovi problemi minimizirali, potrebno je poboljšanje kemije sekvenciranja i programa za očitavanje sljedova (Jain *i sur.*, 2017.).

1.3 Metode sklapanja genoma

Rekonstrukcija izvornog slijeda, odnosno genoma, postiže se sklapanjem sekvenciranih fragmenata (Slika 7). Dva su osnovna tipa sklapanja genoma: mapiranje sljedova na referentni genom (ukoliko takav postoji) i *de novo* sklapanje genoma. U ovom radu biti će opisano *de novo* sklapanje genoma.



Slika 7 Primjer genoma s repetitivnim sljedovima (crveno) i DNA fragmenata dobivenih sekvenciranjem. Preuzeto i prilagođeno iz (Simpson i Pop, 2015.)

Sklapanje genoma odvija se u dva koraka. U prvom koraku nastaju neprekinuti sljedovi (eng. *contigs*) koji se dobivaju preklapanjem sekvenciranih sljedova, a u drugom koraku se pronalazi redoslijed i međusobna orijentacija neprekinutih sljedova te se na taj način dobivaju veći, prekinuti sljedovi (eng. *scaffolds*) (Nagarajan i Pop, 2013.).

Dva najčešća pristupa sklapanju genoma su preklapanje-raspored-konsenzus i metoda po de Bruijnu.

1.3.1 Metoda preklapanje-raspored-konsenzus (eng. *Overlap-Layout-Consensus*)

Kod ovog pristupa sklapanja genoma prvi korak je pronalaženje preklapanja između sljedova. Umjesto da se pomoću dinamičkog programiranja svi sljedovi međusobno poravnaju, izrađuje se indeks k-mera (svih podnizova sljedova određene duljine) te se sljedovi sa odgovarajućim brojem zajedničkih k-mera međusobno poravnavaju. Postoje i napredniji

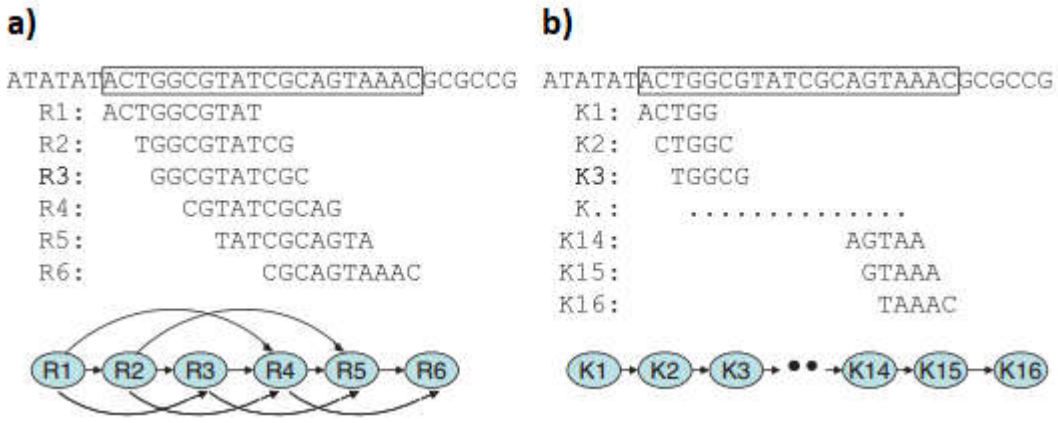
algoritmi za ubrzavanje ovog vremenski najzahtjevnijeg koraka. U drugom koraku (raspored) konstruira se usmjereni graf u kojemu čvorovi predstavljaju sljedove a bridovi poravnanja među sljedovima (Simpson i Pop, 2015.). U posljednjem koraku (konsenzus) se traženjem Hamiltonovog puta kroz graf dobiva konsenzus-sekvenca (Li *i sur.*, 2012.).

Pronalaženje preklapanja među sljedovima i pronalaženje Hamiltonovog puta u grafu su računalno izuzetno zahtjevni problemi. Programi poput Celera (Myers *i sur.*, 2000.), koji koriste metodu preklapanje-raspored-konsenzus, uspješno su se koristili za sklapanje genoma iz fragmenata dobivenih Sangerovim sekvenciranjem, no pojavom metoda sekvenciranja druge generacije koje generiraju velike količine kratkih sljedova nastala je potreba za razvojem drugačijih metoda za sklapanje genoma (Nagarajan i Pop, 2013.).

1.3.2 Metoda po de Bruijnu

Kod sklapanja genoma metodom po de Bruijnu sljedovi se prvo razdijele u k-mere. Konstruira se usmjereni graf u kojem čvorovi predstavljaju k-1-mere (podnizove k-mera duljine k – 1), odnosno preklapanje među k-merima, a bridovi k-mere. Ulazni čvor nekog brida predstavlja lijevi k-1-mer odgovarajućeg k-mera, a izlazni čvor predstavlja desni. Traženjem Eulerovog puta kroz graf rekonstruira se sekvenca genoma (Slika 8).

Razdiobom sljedova u k-mere gubi se dio informacije sadržan u sljedovima, no većina programa za sklapanje genoma koji koriste metodu po de Bruijnu koriste cjelovite sljedove za završno uređivanje grafa (Nagarajan i Pop, 2013.).



Slika 8 Sklapanje genoma. A) Prikazani su sljedovi dobiveni sekvenciranjem i graf konstruiran metodom preklapanje-raspored-konsenzus. Čvorovi u grafu predstavljaju sljedove, a bridovi preklapanje među sljedovima. B) Prikazani su k-meri konstruirani iz sljedova i graf konstruiran metodom po de Bruijnu. Čvorovi u grafu predstavljaju k-1-mere, a bridovi k-mere. Preuzeto i prilagođeno iz (Li *i sur.*, 2012.)

Sklapanje genoma metodom po de Bruijnu je memorijski i vremenski puno efikasnije od metode preklapanje-raspored-konsenzus. Na slici 8 prikazan je primjer sklapanja kratkog genomskog slijeda metodom preklapanje-raspored-konsenzus i metodom po de Bruijnu. Iako je broj čvorova u grafu konstruiranom metodom po de Bruijnu veći, postignuta je ušteda memorije jer je svaka genomska pozicija u grafu predstavljena samo jednom (Li *i sur.*, 2012.).

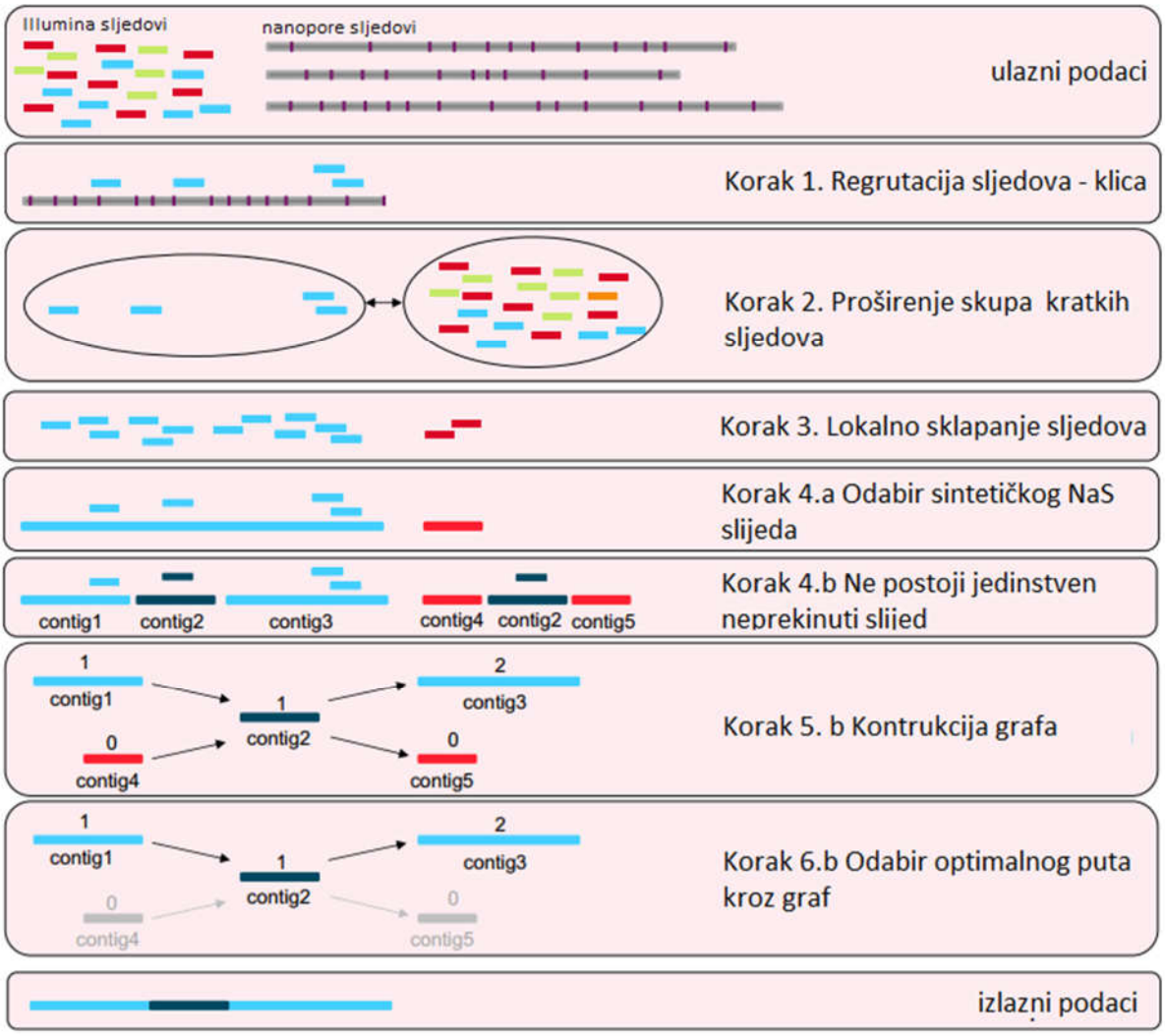
Nužan preduvjet za sklapanje genoma metodom po de Bruijnu je visoka točnost sekvenciranih sljedova jer je nužno da su k-1-meri susjednih k-mera identični. Takav pristup nije pogodan za korištenje dugačkih sljedova niske točnosti kakve proizvodi uređaj ONT MinION. Razdioba sljedova u k-mere također potkopava prednosti koje donosi velika duljina sljedova iz nanopora za razrješenje repetitivnih regija u genoma. Za uspješno korištenje dugačkih sljedova niže točnosti za sklapanje genoma su stoga razvijeni novi pristupi koji nastoje iskoristiti visoku točnost sljedova druge generacije i veliku duljinu sljedova treće generacije za sklapanje genoma visoke kvalitete, a nazivaju se hibridnim pristupima (Simpson i Pop, 2015.).

1.4 Metoda ispravljanja sljedova iz nanopora

Pristupi ispravljanju sljedova iz nanopora dijele se na hibridne, u kojima se za ispravljanje nanopora koriste kratki sljedovi velike točnosti (primjerice Illumina), i *de novo*, u kojima se preklapanjem velikog broja sljedova iz nanopora dobiva točna sekvenca (Bouri i Lavenier, 2017.). Većina hibridnih pristupa temelji se na pronalasku sličnih regija između dugačkih i kratkih sljedova i ispravljanju dugih sljedova koristeći kratke. Ovi pristupi mogu biti vrlo uspješni u nekim slučajevima, no ovise o tome da je cijela dužina dugog slijeda dovoljno kvalitetna da bi se pronašla sličnost sa kratkim sljedovima. Ukoliko to nije slučaj dobivaju se „mozaični“ sljedovi s naizmjenično visoko- i niskokvalitetnim regijama. Pristup korišten u ovom radu, u daljnjem tekstu nazivan NaS, zaobilazi taj problem lokalnim sklapanjem kratkih sljedova za koje je utvrđeno da dolaze iz iste genomske regije kao i slijed iz nanopora (Madoui *i sur.*, 2015.).

Na slici 9 nalazi se shematski prikaz algoritma NaS. U prvom koraku Illumina sljedovi se mapiraju na sljedove iz nanopora pomoću programa LAST (Kielbasa *i sur.*, 2011.) (za osjetljivo pretraživanje) ili BLAT (Kent, 2002.) (za brzo pretraživanje) te se izabiru poravnanja visoke kvalitete. Potom se odabrani sljedovi koriste kao klice za proširenje skupa sljedova. Skup se proširuje sljedovima koji sadrže određen broj zajedničkih k-mera sa sljedovima već prisutnima u skupu, pomoću programa Compareads (Maillet *i sur.*, 2012.). Ovaj korak se iterativno ponavlja do maksimalnog broja ponavljanja ili do postizanja određenog broja

sljedova u skupu. Sljedovi se potom sklapaju koristeći Newbler (Chaisson i Pevzner, 2008.), program za sklapanje genoma temeljen na metodi preklapanje-raspored-konsenzus. U ovom kontekstu metoda preklapanje-raspored-konsenzus nije prespora zbog toga što se u svakom skupu nalazi relativno mali broj Illumina sljedova.



Slika 9 Shematski prikaz metode ispravljanja sljedova iz nanopora NaS. U prvom koraku se regrutiraju sljedovi-klice, čime nastaje početni skup sljedova mapiranih na slijed iz nanopora. U drugom koraku se ovaj skup iterativno povećava traženjem preklapanja Illumina sljedova sa sljedovima već prisutnima u skupu. U koraku 3 se svi sljedovi u skupu sklapaju metodom preklapanje-raspored-konsenzus. Ukoliko postoji slijed koji nedvosmisleno predstavlja ispravljeni slijed iz nanopora algoritam završava (4.a), no ukoliko to nije slučaj (4.b) međusobna orijentacija neprekinutih sljedova (eng. *contig*) određuje se pomoću sparenih sljedova (5.b), a najbolji put kroz graf bira se ovisno o pokrivenosti neprekinutih sljedova sljedovima-klicama (6.b). Sekvenca dobivena razrješenjem grafa predstavlja ispravljeni slijed iz nanopora. Preuzeto i prilagođeno iz (Madoui *i sur.*, 2015.)

Ukoliko nakon sklapanja postoji neprekinuti slijed za koji se nedvosmisleno može reći da odgovara ispravljenom slijedu iz nanopora, ovim korakom algoritam završava. Uvjet nedvosmislenosti provjerava se uzimajući u obzir duljinu dobivenih neprekinutih sljedova i

pokrivenost sljedovima-klicama regrutiranima u prvom koraku. Problem nastaje u repetitivnim regijama gdje može doći do stvaranja više neprekinutih sljedova, a proširenjem skupa sljedova pomoću programa Compareads i sljedovi iz drugih genomskih lokacija mogu ući u skup. U tim slučajevima se međusobna orijentacija neprekinutih sljedova razrješava koristeći sparene sljedove, a optimalan put kroz graf se bira tako da prednost imaju neprekinuti sljedovi s najvećom pokrivenošću sljedovima-klicama (Madoui *i sur.*, 2015.).

Program NaS stvara sljedove visoke kvalitete, no s obzirom na to da koristi lokalizirano sklapanje genoma za očekivati je da će biti podložan istim greškama kao i uobičajeno sklapanje genoma - urušavanju ponavljanja i nemogućnosti razrješenja ponavljajućih regija. Ipak, ovi bi problemi mogli biti slabije izraženi zbog manje skale i lokalne prirode postupka.

2 Cilj istraživanja

Pri *de novo* sklapanju složenih genoma viših eukariota velike probleme stvaraju repetitivni elementi. Metode sekvenciranja sljedeće generacije su najpogodnije za dobivanje velikog broja sljedova iz nepoznatih genoma, no takvi sljedovi ne mogu razriješiti repetitivne regije pri sklapanju genoma. Dugi sljedovi kakve proizvodi uređaj ONT MinION su potencijalno rješenje ovog problema, no u dosadašnjim pokušajima sklapanja genoma *Eunapius subterran*eus nisu uspjeli doprinijeti povećanju kvalitete genoma zbog vlastite preniske kvalitete. Dodatna komplikacija pri sklapanju genoma ove spužve je visoka razina kontaminacije sekvenciranih knjižnica.

Cilj ovog istraživanja je smanjiti razinu kontaminacija u uzorcima izolacijom genomske DNA iz primorfa, stanične kulture spužve, te ispraviti sljedova iz nanopora i validirati njihovu točnost kako bi se olakšalo sklapanje genoma *E. subterraneus*.

3 Materijali i metode

3.1 Uzgoj primorfa

Prema protokolu prilagođenom iz Chernogor *i sur.*, 2011. sam uzgojila primorfe iz jedinke spužve *Eunapius subterraneus* sakupljene speleoronilačkim tehnikama s lokaliteta špilje Tounjčica (Tounj, Ogulin).

Isprala sam spužvu 4 puta sa po 30 mL sterilne izvorske vode. Potom sam ju stavila u 10 mL sterilne izvorske vode i usitnila pomoću sterilne špatule.

Uzorak sam potom filtrirala kroz filtere veličine 200 µm, 100 µm i 40 µm, tim redoslijedom u sterilnu epruvetu. Filtrat sam inkubirala 30 minuta na 8°C.

Nakon inkubacije u epruveti je bilo moguće uočiti želatinozni sloj između taloga i supernatanta. Uklonila sam supernatant ne dodirujući želatinozni sloj, dodala 20 mL sterilne izvorske vode i uzorak ponovno inkubirala 30 minuta na 8°C. Postupak sam ponovila 2 puta, odnosno dok supernatant nije postao bistar. Nakon inkubacije sam uklonila supernatant do konačnog volumena od 5 mL. Odredila sam broj spužvinih stanica pomoću Burker-Turk komorice za brojanje i svjetlosnog mikroskopa. Uzorak sam razdijelila u dvije Erlenmeyerove tikvice i razrijedila sterilnom izvorskom vodom kako bih dobila broj stanica $1x10^7$ st/ml. Inkubirala sam uzorke 2 sata na 8°C, pri čemu sam ih svakih 30 minuta ispirala tako da sam polovicu volumena uklonila i zamijenila sterilnom vodom. Pri posljednjem ispiranju dodala sam sterilnu vodu sa gentamicinom u koncentraciji 50 mg/L i ostavila uzorke na 8°C kako bi nastali primorfi.

3.2 Priprema uzoraka za sekvenciranje

Izolirala sam genomsku DNA iz spužvinih stanica (primorfa) te sam ju vizualizirala pomoću agarozne gel-elektroforeze. Potom sam potvrdila vrstu umnažanjem jezgrinog ITS2 PCR-om. Umnoženi fragment sam vizualizirala agaroznom gel-elektroforezom i sekvencirala Sangerovom dideoski metodom.

3.2.1 Izolacija genomske DNA iz primorfa

Iz uzgojenih primorfa izolirana sam genomsku DNA koristeći komplet QIAGEN Blood & Cell Culture DNA koji sadrži kolonice QIAGEN Genomic-tip 20/G i pufere QIAGEN Genomic DNA Buffer Set prema uputama proizvođača za izolaciju genomske DNA iz kulture stanica.

U ovoj metodi izolacije DNA nakon lize stanica te degradacije proteina i RNA, uzorak se nanosi na silikatnu kolonicu na koju se DNA veže u uvjetima visoke ionske jakosti. Uz odgovarajuće pufere ispiru se sve nepoželjne molekule, nakon čega se DNA eluira s kolonice u puferu niske ionske jakosti.

Koncentraciju DNA izmjerila sam na uređaju BioSpec-nano Micro-volume UV-VIS Spectrophotometer.

3.2.2 Agarozna gel-elektroforeza

Gel-elektroforeza se temelji na putovanju molekula kroz gel u električnom polju. Naboj nukleinskih kiselina je negativan i ujednačen po dužini, stoga se pod utjecajem električnog polja razdvajaju s obzirom na razlike u masi. Za razdvajanje molekula DNA nužno je koristiti gel s većim porama, te je u gelu agaroze moguće razdvojiti fragmente veličine 50 pb-50 kb. Agaroza je linearni polimer u kojem se izmjenjuju ostatci D-galaktoze i 3,6-anhidro-L-galaktoze.

Gel sam priredila otapanjem agaroze u TAE puferu s etidij-bromidom, zagrijavanjem do vrenja, te izlijevanjem u odgovarajući kalup s češljem za formiranje jažica. Polimerne molekule hlađenjem tvore gel međusobno se povezujući vodikovim vezama. Veličina pora, a time i raspon molekulskih masa molekula koje se mogu uspješno odijeliti, kontrolira se mijenjanjem koncentracije agaroze u gelu. 4.5 µL izolirane genomske DNA pomiješala sam sa 0.5 µL boje za praćenje elektroforeze (0.25% bromfenol plavilo, 0,25% ksilencijanola FF i 30% glicerola u vodi) te nanijela na 0.8%tni agarozni gel u TAE puferu. Boja služi za lakše nanošenje uzorka u jažicu i praćenje putovanja uzorka kroz gel te sadrži glicerol koji pospješuje taloženje, odnosno povećava gustoću uzorka. Kao standard veličine koristila sam DNA-biljeg, GeneRuler (Fermentas) (6 µL).

Provela sam elektroforezu u trajanju od ~45 minuta pri ~55 V. Vizualizacija DNA postiže se pomoću etidij-bromida, uključenog u pufer za izradu agaroznog matriksa. Etidij-bromid se interkalira u DNA i pobuđen UV svjetlom fluorescira u narančastom dijelu spektra. Vrpce DNA vizualizirane su promatranjem gela na transluminatoru pri 312 nm.

Snimila sam gelove pomoću uređaja GBOX EF Gel Documentation System (Syngene).

3.2.3 Genetička karakterizacija

3.2.3.1 Lančana reakcija polimerazom

Lančana reakcija polimerazom (PCR) je metoda umnažanja određenog fragmenta DNA u *in vitro* uvjetima. Reakcija se odvija u tri osnovna koraka – denaturacija molekule DNA, specifično vezanje oligonukleotidnih početnica te sinteza DNA pomoću enzima DNA polimeraze. Ovi koraci se ponavljaju kroz 30-40 ciklusa. Broj molekula željenog fragmenta DNA se eksponencijalno povećava svakim ciklusom, no broj ciklusa je ograničen potrošnjom i degradacijom reaktanata. Prije prvog ciklusa provodi se početna denaturacija produljenog trajanja, a nakon posljednjeg ciklusa provodi se završno produživanje lanaca (sinteza DNA). Svaka smjesa za PCR reakciju mora sadržavati kalup (plazmid, DNA-fragment ili genomsku DNA), uzvodnu i nizvodnu početnicu, dNTP-ove, termostabilnu DNA-polimerazu (najčešće *Taq*-DNA-polimerazu), te pripadajući pufer za odabranu polimerazu (koji sadrži Mg^{2+} ione kao kofaktor DNA-polimeraze te za stabilizaciju dNTP-ova). Temperatura denaturacije ovisi o duljini kalupa i njegovom sastavu – za dulje fragmente sa velikim udjelom GC parova potrebna je viša temperatura denaturacije. Temperatura sljepljivanja ovisi o samim početnicama, tako da se kao T_a (engl. *annealing temperature* - temperatura sljepljivanja) obično uzima T_m (temperatura mekšanja)[˚C] – 5[˚C]. Temperatura elongacije najčešće iznosi oko 72 ˚C, dok vrijeme same sinteze DNA ovisi o duljini i koncentraciji ciljnog slijeda, te temperaturi elongacije.

U svrhu potvrde vrste jedinke iz koje je izolirana DNA provela sam lančanu reakciju polimerazom kojom je umnožen jezgrin ITS2 (eng. *internal transcribed spacer*), molekularni marker za određivanje nižih taksonomskih kategorija. ITS2 je dio ribosomskih rRNA gena koji se izrezuje prije sastavljanja ribosoma (Slika 10). Za razliku od visoko očuvanih strukturnih ribosomskih gena koji ga okružuju, ITS2 ima visoku razinu varijabilnosti među vrstama i stoga je pogodan za određivanje nižih taksonomskih kategorija.

Slika 10 Struktura ribosomskih gena kod eukariota

Za umnažanje ITS2 lančanom reakcijom polimerazom koristila sam početnice iz (Harcet *i sur.*, 2010.), prikazane u tablici 1.

Tablica 1 Oligonukleotidne početnice za umnažanje jezgrinog ITS2

ITS-5.8F	5'CGGCTCGTGCGTCGATGAAGAAC3'
ITS-28R	5'CGCCGTTACTGGGGGAATCCCTGTTG3'

U reakcijsku smjesu sam dodala:

9,5 µL NFW,

1 µL razrijeđene genomske DNA (20 ng/µL),

1 µL početnice ITS-5.8F (10pM),

1 µL početnice ITS-28R (10pM),

12,5 µL *Taq PCR Reaction Mix With MgCl2* (Sigma),

za ukupan volumen reakcijske smjese 25 µL. Lančanu reakcija polimerazom provela sam u uvjetima opisanima u tablici 2.

Tablica 2 PCR uvjeti za umnažanje biljega ITS2

temperatura	**trajanje**	**broj ponavljanja**
95˚C	1 min	1x
95˚C	30 s	30x
60˚C	30 s	
72˚C	30 s	
72˚C	5 min	1x
8˚C	~	1x

3.2.3.2 Agarozna gel-elektroforeza i izolacija PCR fragmenata iz gela

Umnoženi fragmenti ITS2 razdvojila sam i vizualizirala pomoću agarozne gel-elektroforeze.

Za razdvajanje PCR fragmenata koristila sam 0,8%-tni agarozni gel u TAE puferu, a fragmente sam vizualizirala pomoću etidij bromida. Nanijela sam čitav volumen reakcijske smjese na gel. Kao standard veličine koristila sam GeneRuler DNA-biljeg (Fermentas). Gel sam slikala pomoću uređaja GBOX EF Gel Documentation System (Syngene).

Očekivana veličina PCR fragmenta je 410 bp (ITS2 je dugačak 361 nt (Harcet *i sur.*, 2010.), a početnice su dugačke 23 i 26 nt).

PCR fragment izrezala sam iz gela promatrajući ga kroz transiluminator. Za pročišćavanje fragmenata DNA, veličine 70 pb do 10 kb, iz agaroznog gela, koristila sam *QIAquick Gel Exstraction Kit* (QIAGEN). Kombinacijom centrifugiranja i selektivnog vezanja na silikagel matriks može se pročistiti i do 10 μg DNA.

Pufer za vezanje otapa agarozu i osigurava točnu koncentraciju soli i pH kod kojih dolazi do adsorpcije DNA na silikagel membranu. Do 95% DNA veže se na matriks pri visokoj koncentraciji soli, i optimalnom pH (pH ≤7,5), dok se neželjene nečistoće kao što su soli, enzimi, agaroza, etidij-bromid, te deterdženti ispiru s kolone. Elucija je također ovisna o koncentraciji soli i pH elucijskog pufera. Suprotno adsorpciji, uspješnoj eluciji pogoduju niska koncentracija soli i bazični pH.

3.2.3.3 Određivanje slijeda nukleotida Sangerovom dideoksi metoom

Kako bih odredila slijed nukleotida izoliranog PCR fragmenata (ITS2) koristila sam Sangerovu dideoksi metoda sekvenciranja molekule DNA detaljno opisanu u poglavlju 1.2.1. Ona se zasniva na zaustavljanju enzimatske sinteze lanca DNA ugradnjom dideoksiribonukleozid-trifosfata.

Slijed nukleotida fragmenta ITS2 odredila sam pomoću sekvenatora *ABI PRISM 3100 - Avant DNA Genetic Analyser* (Applied Biosystems) te kompleta *ABI PRISM BigDye Terminator v3.1 Ready Reaction Cycle Sequencing Kit*. Ovaj uređaj koristi kapilarnu elektroforezu (uzorci se nakon završene reakcije razdvajaju putujući kroz vrlo dugu i tanku kapilaru ispunjenu polimerom) i "četverobojnu biokemiju" (svaki ddNTP obilježen je drugom fluorescencijskom bojom) tako da se reakcija provodi u istoj mikroepruveti. CCD kamera pretvara informaciju o fluorescenciji u elektronički signal koji se onda prenosi na računalo i obrađuje ABI-PRISM-*Avant Genetic Analyzer Data Collection Software Version 2.0* i prikazuje u obliku elektroferograma. Na y-osi prikazuje se relativna koncentracija boje, a na x-osi vrijeme. Svaki pik na elektroferogramu predstavlja po jedan fragment DNA (čija je sinteza zaustavljena ugradnjom modificiranog nukleotida) i koristi se u određivanju nukleotidne sekvence. Rezultati se korisniku prikazuju pomoću programa ABI PRISM *DNA Sequencing Analysis Software Version 5.11*. Za sekvenciranje Sangerovom dideoksi metodom koristila sam iste početnice kao i za PCR reakciju u prethodnom poglavlju.

Pretragom baze neredundantnih nukleotidnih sekvenci dostupne na www.ncbi.nlm.nih.gov pomoću alata BLAST koristeći zadane parametre potvrdila sam vrstu *Eunapius subterraneus* analizom slijeda nukleotida fragmenta ITS2.

3.3 Računalna analiza sljedova

Koristeći program BLAST usporedila sam dvije knjižnice kratkih sljedova Illumina s bazom bakterijskih genoma (Camacho *i sur.*, 2009.) kako bih ustvrdila pridonosi li izolacija genomske DNA iz primorfa smanjenju kontaminacije uzorka u usporedbi s izolacijom iz cijele spužve.

Pomoću programa NaS i knjižnice kratkih sljedova Illumina ispravila sam sljedove iz nanopora te sam validirala ispravljene sljedove pretragom baze neredundantnih proteinskih sljedova programom DIAMOND (Buchfink *i sur.*, 2014.).

3.3.1 Procjena razine kontaminacije

Kako bih ustvrdila je li izolacija genomske DNA iz primorfa doprinijela smanjenju kontaminacije u uzorku analizirala sam 2 knjižnice sljedova dobivenih sekvenciranjem tehnologijom Illumina. Pritom je jedna knjižnica (u daljnjem tekstu Miseq2015) dobivena sekvenciranjem genomske DNA izolirane iz tkiva cijele spužve, dok je druga knjižnica (u daljnjem tekstu Macrogen) dobivena sekvenciranjem genomske DNA izolirane iz primorfa.

Koristeći alat BLAST usporedila sam knjižnice s bazom bakterijskih genoma. BLAST je alat za pretraživanje baza podataka nukleinskih i proteinskih sljedova prema sličnosti sa slijedom od interesa. Baze sljedova su pred-procesirane na način koji omogućava brzo nalaženje klica (k-mera) zajedničkih sa sljedovima od interesa, te ukoliko postoji dovoljan broj zajedničkih klica sljedovi se poravnavaju Smith-Waterman algoritmom.

Koristila sam zadane postavke programa blastn (BLAST za pretraživanje nukleotidne baze nukleotidnim sljedovima) uz maksimalnu dopuštenu e-vrijednost pogotka 1.

Tablica 3 Podatci o knjižnicama korištenima za procjenu razine kontaminacije

Knjižnica	**Duljina sljedova**	**Broj sekvenciranih fragmenata**	**Izolirano iz**
Miseq2015	251	15728104	Cijela spužva
Macrogen	151	190744050	Stanice primorfa

Sljedovi u knjižnici Miseq2015 su duljine 251 nt a u knjižnici Macrogen su dugački 151 nt, te je knjižnica Macrogen za cijeli red veličine veća od knjižnice Miseq2015.

Zbog ovih razlika izravna usporedba knjižnica nije moguća te je pretpostavljeno da se, ukoliko je više od 95% dužine slijeda prekriveno BLAST pogotkom, radi o „pravom pogotku“, odnosno da slijed potječe iz bakterije, neovisno o duljini slijeda.

Kako bih skratila vrijeme pretraživanja, umjesto cijelih knjižnica koristila sam nasumične triplikate podskupova knjižnica veličine 1000, 5000 i 10000 sljedova.

Za svaki podskup izbrojila sam pogotke kojima je prekriveno više od 95% slijeda iz knjižnice i izračunala sam omjer pronađenih pogodaka za podskupove određene veličine u knjižnicama Miseq2015 i Macrogen.

Provela sam dvostruki test ANOVA kako bih ustvrdila ovisi li broj pogodaka normaliziran na veličinu podskupa (broj pogodaka je podijeljen s veličinom podskupa) o veličini podskupa i o knjižnici iz koje je podskup uzorkovan. Testirane su nul-hipoteze:

1. Ne postoji razlika u srednjim vrijednostima broja pogodaka za skupove različitih veličina.
2. Ne postoji razlika u srednjim vrijednostima broja pogodaka za skupove uzorkovane iz različitih knjižnica.

Uz alternativne hipoteze:

1. Postoji razlika u srednjim vrijednostima broja pogodaka za skupove različitih veličina.
2. Postoji razlika u srednjim vrijednostima broja pogodaka za skupove uzorkovane iz različitih knjižnica.

3.3.2 Ispravak sljedova iz nanopora pomoću sljedova dobivenih tehnologijom Illumina

Metoda sekvenciranja nanoporama je tehnologija u razvitku i sljedovi dobiveni ovom metodom imaju visok udio pogreški. Posebno su česte insercije i delecije, što je rijetko u metodama sekvenciranja prve i druge generacije.

Kvalitetu sljedova iz nanopora moguće je poboljšati koristeći kratke sljedove visoke točnosti Illumina. Za ispravljanje sljedova iz nanopora koristila sam program NaS. Algoritam korišten u programu NaS opisan je u uvodu.

Pokrenula sam NaS u načinu rada „*sensitive*“. Za ispravljanje sljedova iz nanopora koristila sam knjižnicu Miseq2015 te podskup knjižnice sljedova iz nanopora dužih od 500 nt veličine 2000 sljedova.

Za ispravak ovog skupa sljedova iz nanopora odabrala sam knjižnicu Miseq2015 umjesto veće i manje kontaminirane knjižnice Macrogen upravo zbog visoke razine kontaminacije. Naime, vjerojatnost nalaženja genomskih sljedova kontaminanata u bazi nukleotidnih sekvenci je puno veća nego za spužvine sljedove. Usporedba ispravljenih sljedova iz nanopora sa sljedovima iz baze omogućiti će evaluaciju točnosti ispravljenih i neispravljenih sljedova.

Prije samog pokretanja programa NaS prilagodila sam ulazne datoteke pomoću programskog jezika Python. Datoteke formata FASTQ koje sadrže knjižnicu Miseq2015 prilagodila sam tako da je za svaki slijed izmijenjena opisna linija. Primjer neispravno formatirane linije je:

@MISEQ:282:000000000-AACHF:1:1101:16069:1338 1:N:0:

a primjer ispravno formatirane linije je:

@MISEQ:282:000000000-AACHF:1:1101:16069:1338/1

Ulazne datoteke sa sljedovima iz nanopora formata FASTA prilagodila sam tako da su u opisnim linijama za svaki slijed uklonjeni razmaci.

3.3.3 Ispitivanje točnosti ispravljenih sljedova

Da bismo odredili kvalitetu ili točnost sekvenciranog slijeda potrebno je sa sigurnošću znati kako taj slijed izgleda. Prepreka kod rada sa genomima slabo istraženih vrsti kao što su spužve je činjenica da su njihovi genomski sljedovi uglavnom nepoznati.

Umjesto usporedbe sa nekolicinom poznatih spužvinih sljedova stoga sam za ispitivanje točnosti ispravljenih sljedova koristila bazu neredundantnih proteinskih sljedova. Iako broj sljedova porijeklom iz spužve u ovoj bazi nije velik, prisutnost kontaminanata čiji su sljedovi dostupni u uzorku i u bazi omogućiti će usporedbu sljedova.

Za uspoređivanje sljedova ispravljenih programom NaS te neispravljenih sljedova sa bazom neredundantnih proteinskih sekvenci koristila sam program DIAMOND. Pomoću programa DIAMOND moguće je usporediti nukleotidne sljedove sa bazom proteinskih sljedova u sučelju vrlo sličnom programu BLAST. Bitno je napomenuti da se DIAMOND može koristiti isključivo na proteinskim bazama, iz čega slijedi da se njime mogu istražiti samo kodirajući

sljedovi u genomu. Prednost programa DIAMOND naspram BLAST je značajno smanjenje vremena izvođenja uz jednaku osjetljivost.

DIAMOND pretražuje bazu proteinskih sekvenci tako što identificira slične sljedove prema broju zajedničkih klica, i potom produžuje sravnjenje koristeći Smith-Waterman algoritam. U većini programa za sravnjenje sljedova koji koriste klice (BLAST, BBmap (Bushnell, sourceforge.net/projects/bbmap/)), klice su kratki sljedovi, i za pronalaženje pogotka potrebno je da dva slijeda dijele nekoliko identičnih klica. Klice u DIAMOND-u se razlikuju od uobičajenih klica u dva ključna obilježja. DIAMOND koristi reduciranu abecedu, što znači da se aminokiseline sličnih svojstava tretiraju kao isti znak, tako da će klice koje nisu identične (ali su slične s obzirom na svojstva aminokiselina) biti tretirane kao iste, što povećava osjetljivost pretraživanja. Na ovaj način je abeceda aminokiselina u DIAMOND-u smanjena sa 20 aminokiselina na 11. Druga razlika između klica u DIAMOND-u i drugih programa je korištenje razmaknutih klica (eng. *spaced seeds*). Takve klice sadrže pozicije koje se ne uzimaju u obzir, odnosno za pronalaženje pogotka zahtijevaju da su samo neke od pozicija u klici identične u klicama dvaju sljedova. DIAMOND koristi razmaknute klice veće duljine od klica u uobičajenim programima kako bi se osigurala točnost, a korištenjem različitih oblika klica (permutacijama pozicija koje se uzimaju u obzir) zadržava osjetljivost jednaku onoj koja se postiže korištenjem kraćih klica.

Pokrenula sam DIAMOND koristeći zadane parametre (korištena je supstitucijska matrica BLOSUM62 te kazne za otvaranje i produženje insercije 11 i 1).

Cilj pretrage bio je identificirati sljedove koji uistinu odgovaraju pogotku u bazi te usporediti pogotke za parove sljedova ispravljeni pomoću programa NaS – neispravljeni slijed iz nanopora. Većina sljedova ima više od jednog pogotka u bazi, stoga sam među ispravljenim sljedovima filtrirala sve pogotke kojima je prekriveno manje od 80% sličnog proteina iz baze, te sam od preostalih pogodaka za svaki ispravljeni slijed uzela pogodak sa najvećim postotkom identiteta.

Među pogodcima na neispravljene sljedove odabrala sam sve pogotke parova ispravljenih sljedova koji su prošli filtriranje, i uzela pogodak kojim je prekriven najveći postotak proteina iz baze.

Potom sam parove sljedova odnosno pogodaka ispravljeni – neispravljeni uredila u tablici kako bi se mogli pregledno usporediti. Za nasumično odabran par pogodaka ispravljeni – neispravljeni napravila sam sravnjenje pomoću programa Needle iz programskog paketa

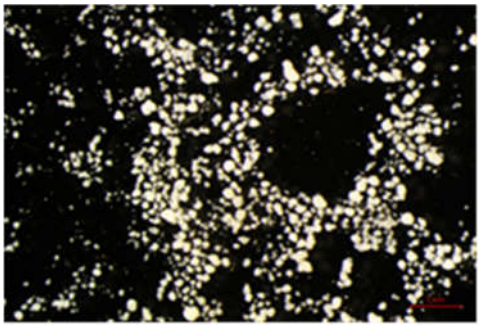
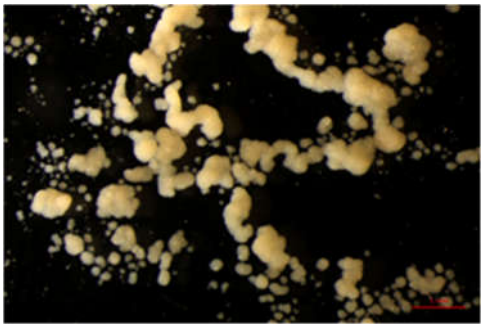
EMBOSS dostupnog na www.ebi.ac.uk (Rice *i sur.*, 2000.). Pritom sam koristila supstitucijsku matricu EDNAFULL te kažnjavanje započinjanja i produživanja insercije jednim bodom. Završno započinjanje i produživanje insercije (na početku i kraju poravnanja) kažnjeni su s 10 i 0.5 boda.

Naposljetku, napravila sam lokalna poravnanja između ispravljenih i neispravljenih sljedova koristeći program BLAST (blastn). Za usporedbu ovih parova sljedova je lokalno poravnanje pogodnije od globalnog jer se neki od parova sljedova značajno razlikuju u duljini. Za svaki par sljedova ispravljeni-neispravljeni uzela sam najduže lokalno poravnanje.

4 Rezultati

4.1 Uspostava kulture spužvinih stanica (primorfi)

Uspostava kulture spužvinih stanica bila je uspješna. Na slici 11 prikazana je uspostava kulture spužvinih stanica nakon 15 minuta i 24 sata snimljena pod lupom u povećanju od 10X.

A)  B) 

Slika 11 Stanična kultura spužve *Eunapius subterraneus nakon 15 minuta i 24 sata*

4.2 Priprema uzoraka za sekvenciranje

4.2.1 Izolacija genomske DNA

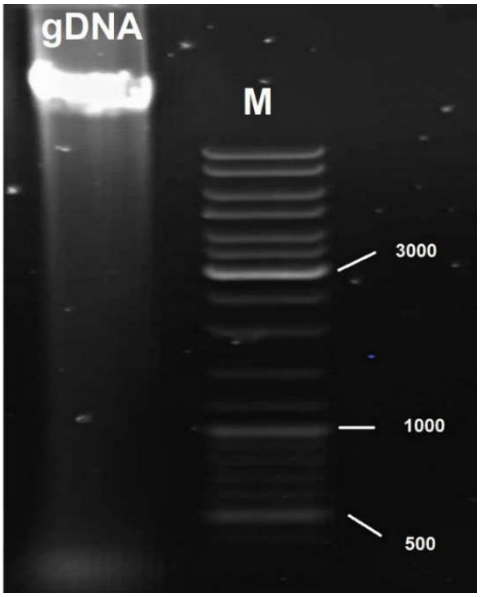
Uspješno sam izolirala visokomolekularnu genomsku DNA iz kulture stanica spužve *Eunapius subterraneus* (2 puta $1x10^7$ st) pomoću kompleta QIAGENBlood & Cell Culture DNA. Dobivenu DNA otopila sam u 50 ml vode bez nukleaza.

U tablici 4 prikazani su omjeri apsorbancija pri valnim duljinama 260 nm i 280 nm te pri valnim duljinama 260 nm i 230 nm koji su unutar prihvatljivih granica, dok je koncentracija DNA prilično visoka.

Tablica 4 Koncentracija i čistoća DNA izolirane iz primorfa *E. subterraneus*

Uzorak	**C (ng/µL)**	**A(260/280)**	**A(260/230)**
E. *subterraneus*	406,66	1,87	2,27

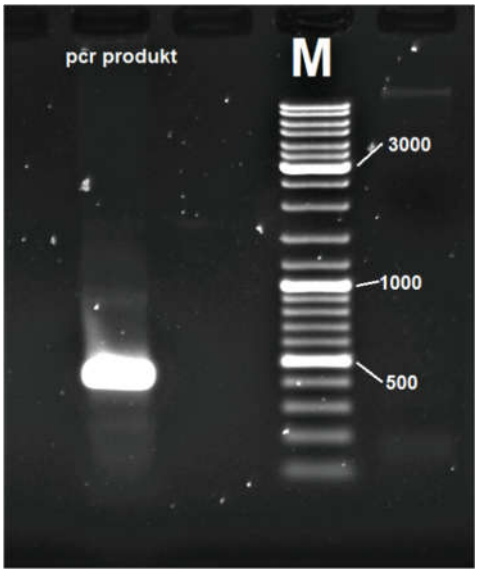
Na slici 12 prikazan je 0,8%-tni agarozni gel nakon agarozne gel-elektroforeze sa nanesenim uzorkom izolirane genomske DNA *Eunapius subterraneus* i DNA biljegom. Na gelu je jasno uočljiva visokomolekularna genomska DNA.



Slika 12 Slika agaroznog gela nakon elektroforeze genomske DNA *E. subterraneus* i DNA biljeg GeneRuler (Fermentas), veličine označenih fragmenata izražene su u (nt)

4.2.2 Genetička karakterizacija

Umnažanje fragmenta jezgrinog ITS2 lančanom reakcijom polimerazom bilo je uspješno. Na slici 13 prikazan je agarozni gel sa vidljivim PCR fragmentom (ITS2) očekivane veličine.



Slika 13 0,8%-tni agarozni gel nakon elektroforeze umnoženog PCR produkta (jezgrinog ITS2 spužve *E. subterraneus*) i DNA-biljega GeneRuler (Fermentas), veličine označenih fragmenata izražene su u (nt)

Dobiveni fragment sam izrezala iz gela, pročistila pomoću kompleta *QIAquick Gel Exstraction Kit* (poglavlje *QIAquick PCR Purification Kit using a Microcentrifuge*) i odredila mu slijed nukleotida sekvenciranjem Sangerovom dideoksi metodom.

Dobiven je slijed

TTTAGTTAGGTTAAATTTCAGCGGGTAGTCACGCCTGAGCTGAGGTCCAGGATGGAAGAGCTTT
CCCCGCTCTTTTAGGGGGAGAAAGGCTCCTCGCGTATTTAACCGACTTGTTTTTGTTCTTCGGAA
CGACGGGCCCCTCGCGAACGAGGTTAACCTCCTCTCCCTCCTTGTCATCGTCGAGTTCCCGACG
CGCAAGGTGGGAAAAGGAGGGCACTCGAGCCGCGAGCGAGTCCTTCCGAACCGGAGCGCTTC
GCACTTGAAGGGACGCCCGTTTCCGGACGACGCCTCAACGCGCCTCCGCGTTTAAACGAACGTT
TAAACAAATAAACGCTCGCCCGCGCGGGGAGACACAAACGAAACGGACGCTCAGACAGACGT
GCCCCCGGCTTCAAACCGAGGGCGCATTTTGCGTTCAAAGACTCATTGATTCACGGAATT

Pretragom baze podataka nukleotidnih sljedova dostupne na www.ncbi.nlm.nih.gov pomoću alata BLAST potvrdila sam da se uistinu radi o vrsti *Eunapius subterraneus*. U tablici 5 prikazano je 5 najboljih BLAST pogodaka.

Tablica 5 Pet najboljih pogodaka dobivenih pretraživanjem baze neredundantnih nukleinskih sljedova sekvenciranim slijedom ITS2 pomoću alata BLAST.

Ime pogotka	**Identifikator**	**Postotak identiteta**	**E-vrijednost**
Eunapius subterraneus ITS2	FJ715436.1	98.403	1e-151
Ephydatia fluviatilis izolat R klon 1 ribosomalni geni	KC244038.1	82.299	3e-92
Ephydatia fluviatilis izolat V2 klon 9 ribosomalni geni	KC244034.1	82.299	3e-92
Ephydatia fluviatilis izolat V2 klon 5 ribosomalni geni	KC244030.1	82.299	3e-92
Ephydatia fluviatilis izolat Pe1 klon 9 ribosomalni geni	KC244010.1	82.299	3e-92

4.3 Računalna analiza sljedova

4.3.1 Procjena razine kontaminacije knjižnica

U tablici 6 navedeni su brojevi „pravih“ pogodaka u bazi bakterijskih genoma za sve nasumično izvučene podskupove analiziranih knjižnica.

Tablica 6 Broj pronađenih „pravih“ pogodaka za podskupove knjižnica Macrogen i Miseq2015 veličine 1000, 5000 i 10000 sljedova, svaki podskup je nasumično izvučen 3 puta

	Knjižnica	
Veličina podskupa	**Macrogen**	**Miseq2015**
1000	6, 7, 6	108, 123, 108
5000	36, 24, 44	578, 563, 559
10000	57, 65, 65	1207, 1168, 1189

Za svaku veličinu podskupa omjer pronađenih kontaminacija u knjižnicama Miseq2015 i Macrogen je 95:5 (kada se postotci uzimaju s preciznošću od 2 decimalne točke).

Rezultati dvostrukog ANOVA testa (Tablica 7) pokazuju da normaliziran broj pronađenih kontaminacija u podskupu ne ovisi o veličini podskupa (nul-hipoteza 1 se ne odbacuje), ali ovisi o tome je li podskup uzet iz knjižnice Miseq2015 ili Macrogen (nul-hipoteza 2 se odbacuje u korist alternativne hipoteze 2).

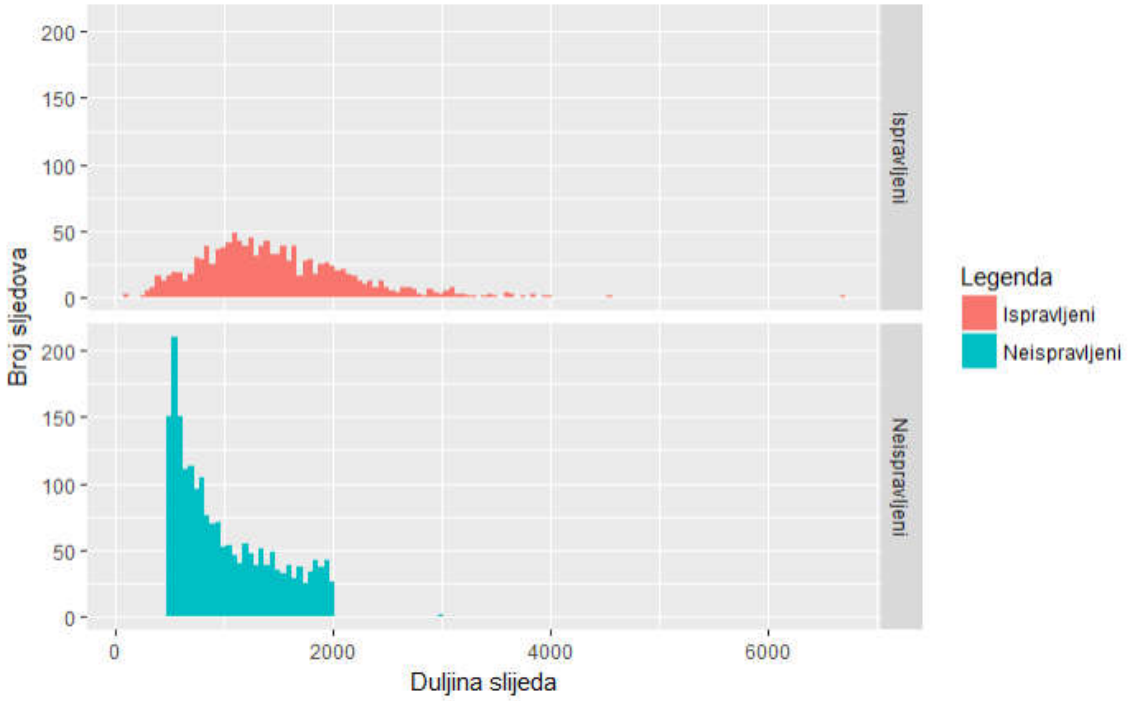
Tablica 7 Rezultati dvostruke ANOVA-e kojom je provjereno ovisi li broj pronađenih kontaminacija normaliziran na veličinu podskupa o veličini podskupa i knjižnici iz koje je podskup uzet.

	stupnjevi slobode	**suma kvadrata**	**varijanca**	**F-vrijednost**	**p-vrijednost**
Veličina podskupa	1	0.00003	0.00003	1.788	0.201
Knjižnica	1	0.05302	0.05302	3718.848	<2e-16
Greška	15	0.00021	0.00001		

4.3.2 Ispravak sljedova iz nanopora

Od 2000 sljedova, pomoću programa NaS uspješno je ispravljeno 1204. Ukupna duljina ispravljenih sljedova je 431774 nt sa srednjom vrijednošću duljine slijeda 1468.3 i medijanom 1369.5, dok je srednja vrijednost duljine slijeda za neispravljene sljedove 1100.3, a medijan 1009.

Na slici 14 su prikazane raspodjele duljine sljedova za ispravljene i neispravljene sljedove.



Slika 14 Raspodjele duljina ispravljenih i neispravljenih sljedova iz nanopora sljedova

4.3.3 Ispitivanje točnosti ispravljenih sljedova

U tablicama 8 i 9 prikazane su informacije o pogodcima dobivenima usporedbom ispravljenih i neispravljenih sljedova iz nanopora s bazom neredundantnih proteinskih sljedova. Za gotovo sve sljedove ispravljeni i neispravljeni slijed odgovaraju istoimenom proteinu, ali drugačijeg porijekla (iz druge vrste). Pritom je postotak identiteta za pogotke na ispravljenim sljedovima iznad 90%, dok se u neispravljenim sljedovima kreće oko 40%.

Sivo su označeni sljedovi kod kojih se ne podudaraju rezultati za ispravljeni i neispravljeni slijed.

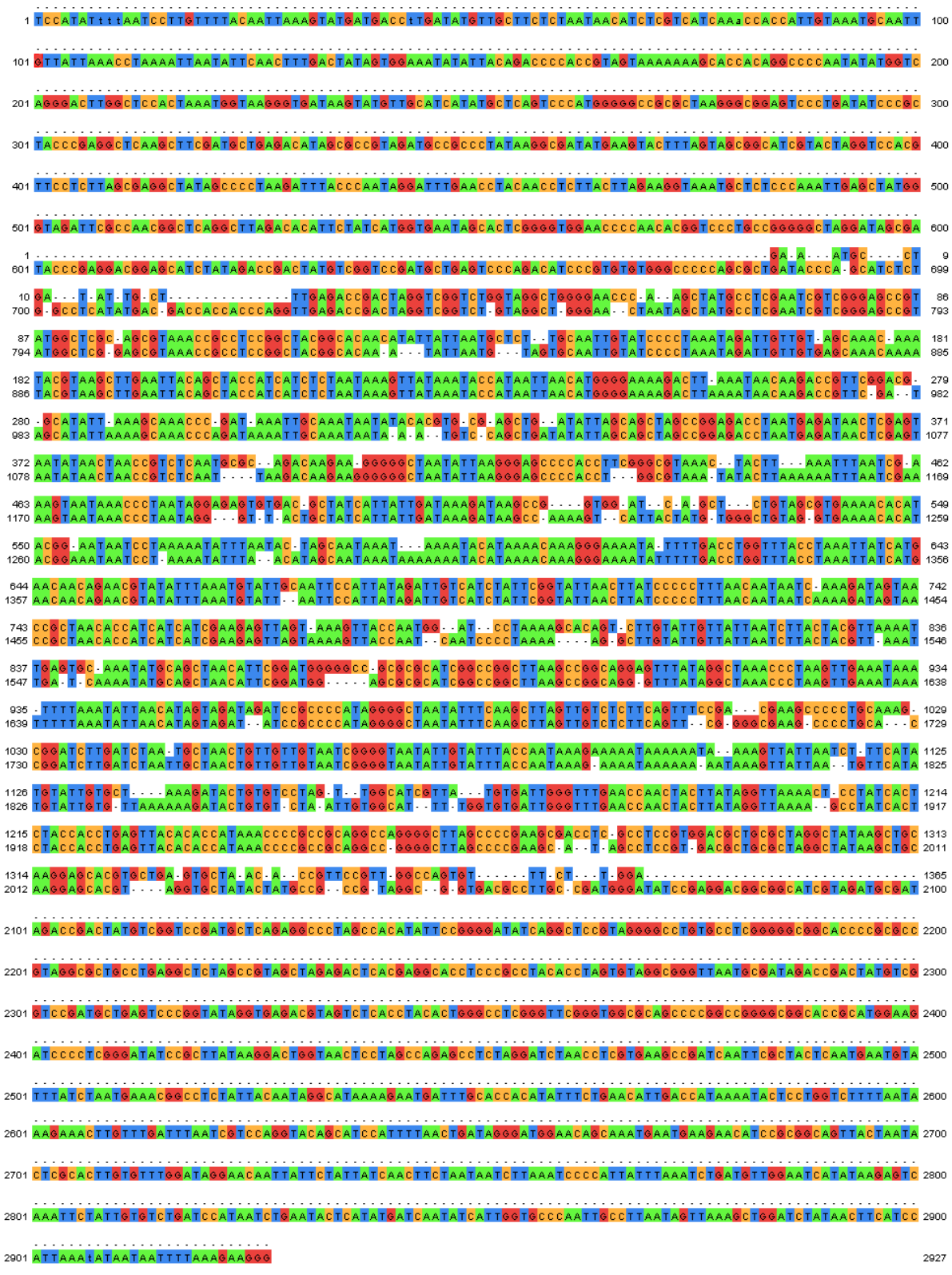
Tablica 8 Pogodci sljedova ispravljenih programom NaS na neredundantnu bazu proteinskih sljedova

	Duljina sljedova iz nanopora	**Duljina proteina**	**Duljina pogotka**	**Postotak identiteta**	**RefSeq identifikator**	**Naziv proteina**	**Vrsta**
1	1250	296	296	92.2	WP_011465792.1	acetylglutamate kinase	Rhodoferax ferrireducens
2	1820	366	347	90.2	PKO32824.1	ferredoxin--NADP(+) reductase	Betaproteobacteria bacterium
3	2447	118	118	98.3	YP_003412025.1	NADH dehydrogenase subunit 3 (mitochondrion)	Lubomirskia baicalensis
4	3048	249	249	97.6	YP_004935618.1	cytochrome c oxidase subunit II (mitochondrion)	Eunapius subterraneus
5	2155	244	244	98.8	YP_004935620.1	ATP synthase F0 subunit 6 (mitochondrion)	Eunapius subterraneus
6	3726	244	244	98.8	YP_004935620.1	ATP synthase F0 subunit 6 (mitochondrion)	Eunapius subterraneus
7	1212	131	131	97.7	WP_091313809.1	glyoxalase	Flavobacterium terrigena
8	2594	381	381	98.7	YP_004935622.1	cytochrome b (mitochondrion)	Eunapius subterraneus
9	2446	118	118	98.3	YP_003412025.1	NADH dehydrogenase subunit 3 (mitochondrion)	Lubomirskia baicalensis
10	2746	634	622	98.9	YP_004935631.1	NADH dehydrogenase subunit 5 (mitochondrion)	Eunapius subterraneus
11	2927	244	244	98.8	YP_004935620.1	ATP synthase F0 subunit 6 (mitochondrion)	Eunapius subterraneus
12	3108	526	494	93.7	YP_004935628.1	cytochrome c oxidase subunit I (mitochondrion)	Eunapius subterraneus
13	3034	494	496	96.4	AEK94599.1	NADH dehydrogenase subunit 2 (mitochondrion)	Ephydatia fluviatilis
14	891	290	254	96.5	WP_073371385.1	succinate--CoA ligase subunit alpha	Flavobacterium fluvii
15	3517	328	328	98.5	YP_004935629.1	NADH dehydrogenase subunit (mitochondrion)	Eunapius subterraneus
16	2892	494	496	96.4	AEK94599.1	NADH dehydrogenase subunit 2 (mitochondrion)	Ephydatia fluviatilis
17	3285	262	262	96.2	YP_003412020.1	cytochrome c oxidase subunit III (mitochondrion)	Lubomirskia baicalensis
18	3844	494	496	96.4	AEK94599.1	NADH dehydrogenase subunit 2 (mitochondrion)	Ephydatia fluviatilis
19	2853	262	262	96.2	YP_003412020.1	cytochrome c oxidase subunit III (mitochondrion)	Lubomirskia baicalensis
20	2311	630	507	98	AMV74149.1	NADH dehydrogenase subunit 5 (mitochondrion)	Spongilla lacustris

Tablica 9 Pogodci neispravljenih sljedova iz nanopora na neredundantnu bazu proteinskih sljedova

	Duljina sljedova iz nanopora	**Duljina proteina**	**Duljina pogotka**	**Postotak identiteta**	**RefSeq identifikator**	**Naziv proteina**	**Vrsta**
1	1082	136	89	48.3	PJC11579.1	polysaccharide biosynthesis protein, partial	Comamonadaceae bacterium
2	817	342	215	47	ALK92296.1	Ferredoxin--NADP reductase	Limnohabitans sp
3	781	195	131	53.4	AFH09312.1	NADH dehydrogenase subunit 6 (mitochondrion)	Swartschewskia papyracea
4	1562	244	203	37.4	YP_009158751.1	cytochrome c oxidase subunit II (mitochondrion)	Petrosia ficiformis
5	797	241	152	40.8	YP_001648411.1	ATP synthase F0 subunit 6 (mitochondrion)	Cinachyrella kuekenthali
6	1907	247	204	45.1	ATI10782.1	cytochrome c oxidase subunit II (mitochondrion)	Plenaster craigi
7	NA	NA	NA	NA	NA	NA	NA
8	721	382	150	49.3	AFH09310.1	apocytochrome b (mitochondrion)	Swartschewskia papyracea
9	1060	196	41	73.2	AFH09353.1	NADH dehydrogenase subunit 6 (mitochondrion)	Corvomeyenia sp
10	863	627	234	44.9	YP_003412030.1	NADH dehydrogenase subunit 5, partial (mitochondrion)	Lubomirskia baicalensis
11	1365	244	250	30.4	YP_214863.1	ATP synthase F0 subunit 6 (mitochondrion)	Axinella corrugata
12	1364	479	148	43.9	ASB15604.1	cytochrome c oxidase subunit I, partial (mitochondrion)	Goniopora sp
13	1408	476	368	30.7	YP_001648475.1	NADH dehydrogenase subunit 2 (mitochondrion)	Ephydatia muelleri
14	1332	290	259	29	PKP52908.1	succinate--CoA ligase subunit alpha	Bacteroidetes bacterium HGW-Bacteroidetes-1
15	1974	324	324	36.1	AFH09316.1	NADH dehydrogenase subunit (mitochondrion)	Swartschewskia papyracea
16	1274	494	233	47.6	AEK94599.1	NADH dehydrogenase subunit 2 (mitochondrion)	Ephydatia fluviatilis
17	1826	244	248	41.9	YP_009158753.1	ATP synthase F0 subunit 6 (mitochondrion)	Petrosia ficiformis
18	1959	476	369	39.8	YP_001648475.1	NADH dehydrogenase subunit 2 (mitochondrion)	Ephydatia muelleri
19	NA	NA	NA	NA	NA	NA	NA
20	889	615	290	36.9	YP_001648476.1	NADH dehydrogenase subunit 5, partial (mitochondrion)	Ephydatia muelleri

Na slici 15 prikazano je sravnjenje ispravljenog i neispravljenog slijeda pod brojem 11 (iz tablica 8 i 9).



Slika 15 Sravnjenje neispravljenog slijeda (gornji) i ispravljenog slijeda (donji) pod brojem 11 u tablicama 8 i 9

Za lokalna poravnanja parova ispravljeni-neispravljeni slijed izračunala sam minimum, maksimum, 1. i 3. kvartil te medijan i srednju vrijednost postotka supstitucija, insercija/delecija i identičnih nukleotida u poravnanju (Tablica 10).

Tablica 10 Deskriptori distribucije profila greške lokalnih poravnanja između ispravljenih i neispravljenih sljedova

	Minimum	**1. kvartil**	**Medijan**	**Srednja vrijednost**	**3. kvartil**	**Maksimum**
identični	0,6452	0,8071	0,8492	0,8435	0,8831	1
supstitucije	0,0000	0,05	0,06995	0,07899	0,1	0,28570
insercije/delecije	0,0000	0,05769	0,07833	0,07755	0,09872	0,25980

5 Rasprava

Istraživanja na ne-modelnim organizmima predstavljaju izazov zbog nedostatka ustaljenih protokola i metoda, no ta istraživanja su nužna kako bismo preciznije odgovorili na mnoga biološka pitanja. Kod istraživanja genoma ogulinske spiljske spužvice *Eunapius subterraneus* kontaminacija mikroorganizmima predstavlja velik problem.

Kod analize knjižnica sekvenciranih tehnologijom Illumina broj pogodaka pronađenih u bazi bakterijskih genoma te heuristika korištena za određivanje „pravih pogodaka“ ne prikazuju nužno stvarnu razinu kontaminacije u analiziranim knjižnicama. Moguće je da postoje kontaminacije koje nisu bakterijskog podrijetla ili potiču od organizama čiji genomi još nisu sekvencirani a evolucijski su bliski istraživanom organizmu. Unatoč tome, usporedbom broja pogodaka među knjižnicama možemo zaključiti koja od njih je manje kontaminirana.

Analiza varijance pokazala je da broj pronađenih kontaminacija ne ovisi o veličini uzorkovanog podskupa već samo o knjižnici iz koje su sljedovi uzeti (Miseq2015 ili Macrogen). Iz omjera broja kontaminacija u ove dvije knjižnice očito je da knjižnica Macrogen ima manji broj kontaminacija, što ukazuje na to da izolacija DNA iz primorfa značajno smanjuje razinu kontaminacije u uzorcima u usporedbi s izolacijom DNA iz cijelih spužvi. Za preostale ne-spužvine sljedove u knjižnici Macrogen možemo pretpostaviti da potječu iz endosimbiontskih organizama kojima spužvine stanice obiluju i koji se ne mogu eliminirati uzgojem primorfa.

Dodatan izazov pri pripremi uzoraka za sekvenciranje na uređaju ONT MinION predstavlja dobivanje visokomolekularne genomske DNA, kako bi se sekvenciranjem dobili što duži sljedovi. Kako bi se smanjila fragmentacija DNA potrebno je u svakom koraku pažljivo postupati s uzorkom i izbjegavati postupke poput vorteksiranja tokom kojih može doći do fragmentacije DNA. Uz takav postupak, moguće je izolirati visokomolekulanu genomsku DNA korištenjem Blood & Cell Culture DNA kompleta tvrtke QIAGEN.

Nakon izolacije i sekvenciranja visomolekularne i relativno nekontaminirane DNA iz kulture spužvinih stanica, prije pristupanja sklapanju genoma potrebno je obratiti pažnju na moguće prepreke.

Eukariotski genomi obiluju ponavljajućim sljedovima. Kada se takvi sljedovi ponavljaju u tandemu, teško je odrediti koja je duljina ponavljajućeg dijela sekvence. Ukoliko se, pak, ista ponavljanja nalaze na više mjesta u genomu teško je točno posložiti sljedove uzvodno i

nizvodno od ponavljanja. Upravo dugački sljedovi iz nanopora su potencijalno rješenje ovih problema, no njihova kvaliteta je puno niža od sljedova dobivenih tehnologijama druge generacije i stoga analiza takvih sljedova zahtijeva drugačije pristupe. Posebno je problematična činjenica da su, za razliku od većine drugih tehnologija sekvenciranja u kojima su supstitucije najčešće pogreške u sekvenciranju, kod sljedova iz nanopora učestalosti insercija/delecija i supstitucija gotovo jednake.

Ispravljanje sljedova iz nanopora pomoću kratkih sljedova dobivenih tehnologijama sekvenciranja druge generacije može pomoći u zaobilaženju ovog problema.

Program NaS čiji je algoritam opisan u uvodu uspješno je ispravio 1204 od 2000 sljedova iz nanopora. Zbog toga što ne dolazi do izravnog ispravljanja sljedova iz nanopora već se skup kratkih Illumina sljedova iterativno proširuje i potom sklapa može doći i do skraćivanja i do produživanja ispravljenih sljedova iz nanopora u odnosu na neispravljene, što je vidljivo iz raspodjele duljina ispravljenih i neispravljenih sljedova na slici 14.

Za provjeru točnosti ispravljenih sljedova iz nanopora odabrano je pretraživanje baze proteinskih sekvenci zato što su aminokiselinski sljedovi evolucijski bolje očuvani od nukleotidnih te je na taj način moguće pronaći pogotke koji ne bi bili uočljivi na razini DNA. Mogućnost prepoznavanja teže uočljivih sličnosti među sljedovima je u ovom slučaju od velike važnosti, jer je pretraživanje provedeno s pretpostavkom da je kvaliteta sljedova iz nanopora vrlo niska.

Među 20 pogodaka filtriranih na način opisan u poglavlju 3.3.3, kod gotovo svih pogodci na ispravljeni i neispravljeni slijed predstavljaju istoimeni protein. To znači da je pri sklapanju Illumina sljedova uistinu rekonstruiran originalni slijed. Činjenica da je postotak identiteta za ispravljene sljedove veći te da su organizmi u kojima su pronađeni pogodci za ispravljene sljedove većinom *E. subterraneus* ili taksonomski bliske slatkovodne spužve govori o tome da su ispravljeni sljedovi veće kvalitete od neispravljenih. Ne treba se zabrinjavati oko pogodaka koji ne odgovaraju *E. subterraneus* nego drugim spužvama jer to ukazuje ili na to da odgovarajući slijed iz *E. subterraneus* nije dostupan u bazi ili na to da je zbog grešaka u slijedu pukim slučajem poravnanje sa sljedovima iz srodnih vrsta bilo bolje.

Dva pogotka kod kojih pogodci na ispravljeni i neispravljeni slijed ne predstavljaju isti protein (sivo označeno u tablicama 8 i 9) potrebno je pomnije pregledati.

U pogotku 1 ispravljeni slijed je najsličniji acetilglutamat kinazi, dok je neispravljeni slijed sličniji proteinu biosinteze polisaharida. U genomu *Rhodoferax ferrireducens* (NC_007908.1), u kojem je pronađen pogodak za ispravljeni slijed, ti se geni nalaze na pozicijama 3920254-3921144 i 2947642-2949585, dakle previše su udaljeni da bi se mogli zajedno nalaziti na slijedu duljine 1082 nt. Pretraga baze podataka neredundantnih nukleotidnih sljedova u alatu BLAST dostupnom na internetu pokazuje da postoji sličnost neispravljenog slijeda i genoma bakterije *Rhodoferax ferrireducens* u dijelu genoma na kojemu se nalazi acetilglutamat kinaza, 3920254-3921144.

Ovo upućuje na zaključak da je prisutnost proteina biosinteze polisaharida u neispravljenom slijedu artefakt korištenja enzima transpozaze pri pripremi knjižnica za sekvenciranje uređajem ONT MinION, te da slijed uistinu potječe iz *Rhodoferax ferrireducens* ili neke srodne vrste.

Drugi slučaj u kojem ispravljeni i neispravljeni protein ne odgovaraju istom slijedu je slijed 6, gdje ispravljeni slijed ima najveću sličnost sa podjedinicom 6 ATP sintaze F0, dok neispravljeni slijed ima najveću sličnost sa podjedinicom II citokrom c oksidaze. Ovaj slučaj je lako objašnjiv, jer se ovi geni nalaze jedan pored drugoga u mitohondrijskom genomu *E. subterraneus* (GU086203.1). Kod ispravljenog slijeda je pogodak na ATP sintazu prema kriterijima filtriranja bio bolji nego pogodak na citokrom c oksidazu, no oba proteina su sadržana unutar slijeda.

Potrebno je napomenuti kako je postotak identiteta za ispravljene sljedove dostupan u tablici 8 pouzdan, no za neispravljene sljedove (Tablica 9) ne možemo iz postotka identiteta zaključiti puno o njihovoj kvaliteti. Uzrok tome je što su ispravljeni sljedovi sklopljeni od kratkih Illumina sljedova, kod kojih su najčešće greške u sekvenciranju supstitucije, a insercije i delecije su vrlo rijetke. Ovakav profil greške je pogodan za usporedbu translatiranog slijeda s bazom proteinskih sljedova, jer ne dolazi do pomaka okvira čitanja i sličnost je očuvana. Neispravljeni sljedovi obiluju insercijama i delecijama, koji su najčešće greške kod sekvenciranja tehnologijom nanopora. Zbog toga se pri translaciji sljedova često gubi okvir čitanja i ukoliko bismo procijenili kvalitetu sljedova prema tablici 9, značajno bismo ju podcijenili.

Iz tog razloga je usporedba s ispravljenim sljedovima na nukleotidnoj razini, ukoliko prihvatimo zaključak da su ispravljeni sljedovi dovoljno točni, bolja procjena kvalitete neispravljenih sljedova iz nanopora, a ujedno nam omogućuje i evaluaciju ispravnosti sljedova koji nisu dostupni u bazi podataka. Usporedba s ispravljenim sljedovima je možda bolja i stoga što smo za njih sigurni da su sklopljeni iz sljedova prisutnih u uzorku, dok najbolji pogodak u

bazi podataka može odgovarati slijedu koji je najsličniji neispravljenom slijedu, ali ne odgovara stvarnom slijedu iz uzorka.

Iz tablice 10 možemo očitati da postotak identiteta u lokalnim poravnanjima između ispravljenih i neispravljenih sljedova varira oko 85% dok je učestalost supstitucija i insercija/delecija približno ista (7.8%). Na primjeru sravnjenja ispravljenog i neispravljenog slijeda na slici 15 možemo dobiti i vizualni dojam o opisanom profilu greške. Sljedovi ispravljeni programom NaS nisu 100% točni te ne možemo precizno odrediti profil greške iz njih, no napravljene usporedbe daju dobru sliku o očekivanom profilu greške i točnosti sljedova iz nanopora.

Program NaS je prvenstveno namijenjen sklapanju bakterijskih genoma i njegovo korištenje za sklapanje genoma viših eukariota nije izgledno zbog dugog vremena izvođenja, no u ovom radu je NaS iskorišten kako bi se procijenila kvaliteta i profil greške sljedova iz nanopora. Poznavanje profila greške moglo bi biti ključno za uspješno korištenje sljedova iz nanopora u daljnjim primjenama. Ovo je posebno očito na primjeru iz poglavlja 4.3.3, gdje se sravnjenje sljedova na proteinskoj razini pokazalo neprikladno zbog prisutnosti velikog broja insercija i delecija u neispravljenim sljedovima iz nanopora.

Kako bi se sljedovi iz nanopora uspješno koristili za sklapanje genoma potrebno je razvijati specifične algoritme i programe temeljene na pristupima koji se mogu nositi sa neobičnim profilom greške sljedova iz nanopora.

6 Zaključak

- Izolacijom genomske DNA iz stanične kulture spužve dobiva se značajno manje kontaminirana DNA nego izolacijom DNA iz cijele spužve
- Program NaS vrlo uspješno ispravlja sljedove iz nanopora koristeći Illumina sljedove, no ima predugo vrijeme izvođenja za korištenje na velikim genomima viših eukariota
- Profil greške sljedova iz nanopora takav je da se supstitucije i insercije/delecije pojavljuju gotovo jednakom učestalošću, što predstavlja problem za analizu kodirajućih sljedova

7 Literatura

Bilandžija H, Bedek J, Jalžić B, Gottstein S (2007). the Morphological Variability , Distribution Patterns and Endangerment in the Ogulin Cave Sponge Eunapius Subterraneus Sket & Velikonja , 1984 (Demospongiae. *Nat Hist* **16**: 1–17.

Bouri L, Lavenier D (2017). *Evaluation of long read error correction software*at <https://hal.inria.fr/hal-01463694/>.

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

Bushnell B BBMap. at <sourceforge.net/projects/bbmap/>.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, i sur. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics* **10**: .

Chaisson MJ, Pevzner PA (2008). Short read fragment assembly of bacterial genomes. *Genome Res* **18**: 324–330.

Chen F, Dong M, Ge M, Zhu L, Ren L, Liu G, i sur. (2013). The History and Advances of Reversible Terminators Used in New Generations of Sequencing Technology. *Genomics, Proteomics Bioinforma* **11**: 34–40.

Chernogor LI, Denikina NN, Belikov SI, Ereskovsky A V. (2011). Long-Term Cultivation of Primmorphs from Freshwater Baikal Sponges Lubomirskia baikalensis. *Mar Biotechnol* **13**: 782–792.

Goodwin S, McPherson JD, McCombie WR (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**: 333–351.

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

Heather JM, Chain B (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**: 1–8.

Ip CLC, Loose M, Tyson JR, Cesare M de, Brown BL, Jain M, i sur. (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research*

doi:10.12688/f1000research.7201.1.

Jain M, Tyson JR, Loose M, Ip CLC, Eccles DA, O'Grady J, i sur. (2017). MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Research* **6**: 760.

Kent WJ (2002). BLAT — The BLAST -Like Alignment Tool. *Genome Res* **12**: 656–664.

Kielbasa SM, Wan R, Sato K, Kiebasa SM, Horton P, Frith MC (2011). Adaptive seeds tame genomic sequence comparison Adaptive seeds tame genomic sequence comparison. *Genome Res* **21**: 487–493.

Lavrov AI, Kosevich IA (2016). Sponge cell reaggregation: Cellular structure and morphogenetic potencies of multicellular aggregates. *J Exp Zool Part A Ecol Genet Physiol* **325**: 158–177.

Leggett RM, Clark MD (2017). A world of opportunities with nanopore sequencing. *J Exp Bot* **68**: 5419–5429.

Leys SP, Hill A (Elsevier Ltd.: 2012). *The Physiology and Molecular Biology of Sponge TissuesAdv Mar Biol* **62**: .

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.

Love GD, Grosjean E, Stalvies C, Fike DA, Grotzinger JP, Bradley AS, i sur. (2009). Fossil steroids record the appearance of Demospongiae during the Cryogenian period. *Nature* **457**: 718.

Lu H, Giordano F, Ning Z (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics Bioinforma* **14**: 265–279.

Madoui MA, 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**: 1–11.

Maillet N, Lemaitre C, Chikhi R, Lavenier D, Peterlongo P (2012). Compareads: comparing huge metagenomic experiments. *BMC Bioinformatics* **13**: .

Müller WEG, Müller IM (2003). Origin of the metazoan immune system: identification of the

molecules and their functions in sponges. *Integr Comp Biol* **43**: 281–92.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, i sur. (2000). A whole-genome assembly of Drosophila. *Science (80-)* **287**: 2196–2204.

Nagarajan N, Pop M (2013). Sequence assembly demystified. *Nat Rev Genet* **14**: 157–167.

Reuter JA, Spacek D V., Snyder MP (2015). High-Throughput Sequencing Technologies. *Mol Cell* **58**: 586–597.

Rice P, Longden I, Bleasby A (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.

Sebé-Pedrós A, Mendoza A De, Lang BF, Degnan BM, Ruiz-Trillo I (2011). Unexpected repertoire of metazoan transcription factors in the unicellular holozoan capsaspora owczarzaki. *Mol Biol Evol* **28**: 1241–1254.

Simpson JT, Pop M (2015). The Theory and Practice of Genome Sequence Assembly. *Annu Rev Genomics Hum Genet* **16**: 153–172.

Sket B, Velikonja M (1986). Troglobitic freshwater sponges (Porifera, Spongillidae) found in Yugoslavia. *Stygologia* **2**: 254–266.

Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, i sur. (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* **466**: 720–726.

Wilson H V. (1910). Development of sponges from dissociated tissue cells. *Bull Bur Fish* 1–30.

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**: .

Životopis

OBRAZOVANJE

Rujan 2018 **Ljetna škola NGSchool 2018**

Nanopore sequencing & personalized medicine

2016-2018 **Diplomski studij molekularne biologije**

Prirodoslovno-matematički fakultet,
Sveučilište u Zagrebu

2013-2016 **Preddiplomski studij molekularne biologije**

Prirodoslovno-matematički fakultet,
Sveučilište u Zagrebu

RADNO ISKUSTVO

Travanj 2018 **Stručna praksa**

OmicX, Rouen, Francuska

2017 - 2018 **Stručna praksa i izrada diplomskog rada**

Kristian Vlahoviček, Zagreb (Hrvatska)
Grupa za bioinformatiku, Biološki odsjek, Prirodoslovni-matematički fakultet Sveučilišta u Zagrebu

2017 **Izrada diplomskog rada**

Helena Ćetković, Zagreb (Hrvatska)
Laboratorij za molekularnu genetiku, Institut Ruđer Bošković

2016 **Stručna praksa**

Miroslav Plohl, Zagreb (Hrvatska)
Laboratorij za strukturu i funkciju heterokromatina, Institut Ruđer Bošković
Pod nadzorom Brankice Mravinac

2014-2015 **Demonstrator**

Mladen Kučinić, Zagreb
Kolegij "Zoologija" na Biološkom odsjeku, Prirodoslovno-matematički fakultet Sveučilišta u Zagrebu

2010-2011 **Volonter**

Rebeka Bulat, Rijeka
Rad s djecom s poteškoćama u razvoju u Centru za ranu dijagnostiku OKOlonaOKOlo

DODATNE INFORMACIJE

Završni rad

Naslov: DNA transpozoni u ljudskom genomu

Mentor: prof. dr. sc. Miroslav Plohl

Popularizacija znanosti

Swap-shop na ljetnoj školi S3++ u Požegi 2017: "Solving a mistery: the origins of mixed-up DNA"

Noć biologije (2014., 2015., 2016., 2017., 2018.)

Otvoreni dan kemije (2014.)

Đir po Ruđeru (2014.)

Nagrade

Posebna rektorova nagrada za akademsku godinu 2014./2015 za istraživačko-edukacijski Projekt „Grabovača 2014."

Članstva

BIUS, Udruga studenata biologije, 2014.

VJEŠTINE

Jezici: Engleski (C1), Njemački (B2), Francuski (A1)

Informatičke vještine: R, Python (osnove), Bash (Unix)