UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN INFORMACIJSKE TEHNOLOGIJE

ZAKLJUČNA NALOGA

TESTIRANJE UČINKOVITOSTI BIOINFORMACIJSKIH ORODIJ ZA ZDRUŽEVANJE ZAPOREDIJ GENOMSKE DNA LAŠKEGA SMILJA (*HELICHRYSUM ITALICUM* (ROTH.) G. DON.), PRIDOBLJENIH S TEHNOLOGIJO ION TORRENT

BENJAMIN BOŽIČ

2020
ZAKLJUČNA NALOGA
BOŽIČ

UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga

Testiranje učinkovitosti bioinformacijskih orodij za združevanje zaporedij genomske DNA laškega smilja (*Helichrysum italicum* (Roth.) G. Don.), pridobljenih s tehnologijo Ion Torrent

(Testing efficiency of bioinformatic tools for DNA sequencing of Curry plant's DNA (*Helichrysum italicum* (Roth.) G. Don.), produced with Ion Torrent technology)

Ime in priimek: Benjamin Božič
Študijski program: Bioinformatika
Mentor: doc. dr. Matjaž Hladnik
Somentor: izr. prof. dr. Dunja Bandelj

Koper, september 2020

Ključna dokumentacijska informacija

Ime in PRIIMEK: Benjamin BOŽIČ
Naslov zaključne naloge:
Testiranje učinkovitosti bioinformacijskih orodij za združevanje zaporedij genomske DNA laškega smilja (*Helichrysum italicum* (Roth.) G. Don.), pridobljenih s tehnologijo Ion Torrent

Kraj: Koper
Leto: 2020
Število listov: 42 Število slik: 5 Število tabel: 4
Število referenc: 31
Mentor: doc. dr. Matjaž Hladnik
Somentor: izr. prof. dr. Dunja Bandelj
Ključne besede: laški smilj, *Helichrysum italicum*, zbirniki, DNA odčitki, združevanje odčitkov, BUSCO, QUAST, SPAdes
Izvleček:
V diplomski nalogi smo primerjali rezultate združenih zaporedij, pridobljenih z osmimi različnimi zbirniki. Izvorni set zaporedij genomske DNA laškega smilja (v datotečnem zapisu uBAM), pridobljenih s tehnologijo Ion Torrent, smo uporabili za pripravo dodatnih dveh setov. Prvi set podatkov je vseboval odčitke z odstranjenimi ostanki adapterjev (potrebnih za sekvenciranje), pri drugem setu pa smo odčitke filtrirali še glede na Q vrednost 20 z orodjem Cutadapt. Zbirnike (razen zbirnika MIRA, ki je potreboval več delovnega spomina) smo namestili na računalnik s 125 GB delovnega spomina in 32 logičnih jeder ter po navodilih za vsak zbirnik posebej izvedli združevanje odčitkov za oba seta podatkov (SPAdes dodatno še z uBAM datoteko).
Vsa končna združena zaporedja smo nato ovrednotili s pomočjo programov BUSCO (orodje za ovrednotenje deleža pokritosti genoma na osnovi skupin ortologni genov) in QUAST, ki za združena zaporedja izračuna več statističnih parametrov.
Glede na rezultate, pridobljene s prej omenjenima programoma, smo zaključili, da smo pridobili najboljše rezultate z uporabo zbirnika SPAdes.

Key document information

Name and SURNAME: Benjamin BOŽIČ

Title of the final project paper:

Testing efficiency of bioinformatic tools for DNA sequencing of Curry plant's DNA (*Helichrysum italicum* (Roth.) G. Don.), produced with Ion Torrent technology

Place: Koper

Year: 2020

Number of pages: 42 Number of figures: 5 Number of tables: 4

Number of references: 31

Mentor: Assist. Prof. Matjaž Hladnik, PhD

Co-Mentor: Assoc. Prof. Dunja Bandelj, PhD

Keywords: curry plant, DNA sequences, *Helichrysum italicum*, de novo assemblers, DNA assembly, BUSCO, QUAST, SPAdes

Abstract: In the final project paper we compared results of final assemblies produced by eight different de novo assemblers. The raw dataset (in uBAM format) of Curry plant's DNA was produced with Ion Torrent technology. From this dataset we created two others (in FASTQ format). For the first one we used Cutadapt and removed remaining adapter sequences and for the second one we additionally filtered DNA sequences with Q values below 20.

We installed de novo assemblers (excluding MIRA, which needed more RAM) on a computer with 125 GB RAM and 16 core / 32 threads. Then we followed the manuals for each assembler and ran them for both datasets (SPAdes was also tested with uBAM dataset). Final assemblies were evaluated with two software tools: BUSCO and QUAST. The first one is designed to evaluate genome coverage based on orthology groups and the last one performs statistical analysis for multiple metrics. Based on the evaluated results we concluded that the best assembly was produced with de novo assembler SPAdes.

ZAHVALA

V prvi vrsti bi se rad zahvalil mentorju doc. dr. Matjažu Hladiku za ves vložen trud, pripravljenost priskočiti na pomoč z razlago, predlogom ali novo idejo in vodenje skozi izvajanje tako praktičnega kot tudi pisnega dela diplome.

Zahvalil bi se tudi somentorici izr. prof. dr. Dunji Bandelj, ki me je spodbujala in mi ponujala odlične priložnosti tekom študija (tudi temo za to diplomsko nalogo), ter za njene nasvete pri pisanju diplomske naloge.

Zahvala gre tudi prof. dr. Jerneju Jakšetu, ki je opravil združevanje ter priskrbel končno zaporedje, pridobljeno z zbirnikom CLC Genomics Workbench.

Zahvaljujem se tudi staršem in Kristini, ki so mi vedno stali ob strani, me podpirali in spodbujali.

Na koncu bi se rad zahvalil še UP FAMNIT, ki mi je omogočil uporabo strojne opreme, brez katere tehnična izvedba diplomske naloge ne bi bila mogoča.

Najlepša hvala vsem!

KAZALO VSEBINE

KAZALO PREGLEDNIC

KAZALO SLIK IN GRAFIKONOV

1 UVOD

Danes so za določevanje zaporedja DNA na voljo različne tehnologije visoko pretočnega sekvenciranja. Odčitke (angl. reads) oziroma krajša zaporedja (nekaj 100 nukleotidov), ki jih sekvenatorji generirajo, je potrebno združiti v daljša zaporedja, s čimer rekonstruiramo zaporedje molekul DNA, prisotnih v proučevanem organizmu. Za ta namen je bilo razvito večje število bioinformacijskih orodij, ki pa lahko pripeljejo do različnih rezultatov, zato je izbira programa pomemben dejavnik za doseganje kakovosti in informativnosti raziskave.

Cilj zaključne naloge sta pregled in primerjava rezultatov različnih programov za združevanje zaporedij, pridobljenih iz genomske DNA laškega smilja (*Helichrysum italicum* (Roth.) G. Don) z uporabo sekvenatorja Ion Torrent Ion S5.

Analizo smo izvedli s setom nukleotidnih zaporedij laškega smilja, ki so jih predhodno pridobili raziskovalci UP FAMNIT[1]. Za primerjavo smo naredili izbor programov, ki glede na znanstveno literaturo veljajo za najbolj primerne pri delu z odčitki, pridobljenimi s tehnologijo Ion Torrent (SPAdes, MIRA, Abyss, Ray) ter nekaj novejših (MEGAHIT, GATB Minia pipeline). Dobljene rezultate, t.j. združena zaporedja smo analizirali z orodjem QUAST, ki omogoča ovrednotenje skupne dolžine združenih zaporedij ter izračun drugih parametrov, ki so v uporabi za ovrednotenje združenih zaporedij genomske DNA.

Z nalogo smo želeli ugotoviti, kako uporaba različnih programov vpliva na združevanje zaporedij, ter skozi primerjavo najti najustreznejšega za obdelavo omenjenih zaporedij. Ugotovitve naloge bodo koristne pri načrtovanju nadaljnjih raziskav.

1.1 Predstavitev rastline smilja

Laški smilj je sredozemsko zelišče, čigar latinsko ime *Helichrysum italicum* pomeni “zlato sonce”, kar zelo dobro opiše samo rastlino [2]. Rod *Helichrysum* spada v družino Asteraceae (Nebinovke) in vsebuje več kot 500 vrst, ki jih najdemo v Evropi, Afriki, Aziji in Avstraliji [3].
Helichrysum italicum zraste v grmičku višine od 30 do 70 cm in premera do enega metra. Grm prepoznamo po rumenih cvetovih v kobulastih socvetjih in ozkih črtalistih listih svetlozelene do srebrnkaste barve, ki so na steblu spiralasto ali premenjalno razvrščeni [4].
Uspeva v Sredozemlju, kjer je veliko sonca in milo podnebje. Ustrezajo mu odcedna tla [5].
Pri nas smilj najdemo ob morju, zelo pogosto pa ga najdemo vzdolž celotne Hrvaške obale, ki premore tudi veliko rastišč, iz katerih rastlinski material uporabijo za namene pridobivanja eteričnega olja [2].
Laški smilj se uporablja v različne namene: posušene rastline za okras, listi kot začimbe za razne jedi, najbolj pogosta in cenjena pa je njegova uporaba v zdravstvu in kozmetiki in sicer

v obliki eteričnega olja. Iz eteričnega olja se izdelujejo razne kreme za kožo, ki so zelo cenjene [6]. Prodajajo se s trgovskimi oznakami "anti-age" in "imortelle", saj kožo poživljajo in obnavljajo. Hidrolat pa kožo hladi in vlaži in je še posebej primeren za opekline ter po sončenju.

Za pridobivanje eteričnega olja so zaželeni zeleni deli rastline in cvetovi. Pri žetvi se odvzame neoleseneli del poganjkov. Postopek se nadaljuje s parno destilacijo in sicer najkasneje v roku enega dne po žetvi. Rastlinski material dajo v kotle iz nerjavečega jekla in skoznje spustijo vodno paro pod pritiskom 1,2 bara (pri t. i. parni destilaciji). Slednja se navzame različnih snovi in v hladilnem sistemu, ki sledi kotlu z rastlinskim materialom, se obogatenega para utekočini. Oblikujeta se dve fazi in sicer vodna faza (hidrolat) in oljna faza (eterično olje). V zbirnem kotlu olje ni edina uporabna stvar, saj so številne koristne učinkovine topne v vodi, nastalo raztopino pa imenujemo hidrolat oziroma cvetna voda. Za slednjega je pomembno, v kateri fazi destilacije ga shranijo [2].

1.2 Tehnologija Ion Torrent

Tehnologija Ion Torrent pravaja kemijski zapis nukleotidov v digitalno informacijo (binarni zapis). Tehnologija temelji na uporabi polprevodniškega čipa (angl. semiconductor chip). Slednji ima lahko tudi več deset milijonov majhnih reakcijskih/reaktorskih posodic, v katerih so posebne mikro oz. nano kroglice in na njih vezane številne enoverižne molekule DNA. S cikličnim izmenjevanjem dovajanja dNTP in tekočine za spiranje čipa poteka v vseh posodicah na čipu sekvenčna reakcija paralelno. Vsaka posodica posebej deluje kot izjemno majhen pH meter [7].

1.2.1 Postopek prevajanja kemijskega zapisa DNA v digitalni zapis

Vsakih 15 sekund se vse posodice na polprevodnem čipu napolnijo z raztopino, ki vsebuje enega od štirih nukleotidov. V luknjicah, ki vsebujejo kroglice z molekulami DNA, kjer je naslednja prosta baza komplementarna nukleotidu, v raztopini poteče reakcija vezave nukleotida na enoverižno DNA molekulo, pri kateri se odcepi 1 H^+ (vodikov ion). Posledica cepitve H^+ povzroči spremembo pH raztopine. Spremembo zazna ionska plast pod luknjico in izmerjeno vrednost spremeni v električno napetost, na podlagi katere čip določi, katera baza je naslednja v zaporedju DNA. V primeru, da se neki nukleotid v zaporedju ponovi (npr. AA ali TT), je izmerjena sprememba pH 2-kratna in posledično čip zazna dve enaki bazi. Podobno velja za daljša zaporedja sestavljena iz ene baze. V primeru, ko ne poteče reakcija (neujemanje nukleotidov v raztopini in na kratki DNA molekuli), se pH ne spremeni in posledično tudi ne električna napetost. Postopek se ponavlja dokler ne opravi v naprej določenega števila ciklov [7].

1.2.1 Prednosti Ion Torrent tehnologije

Ker je zaznavanje spremembe v pH in s tem spremembe napetosti neposredno, oziroma ker proces ne vključuje fotodetektorjev ali senzorjev za zaznavanje spremembe oziroma poteka reakcije, poteče celoten proces prevajanja kemijskega zapisa v digitalnega v le nekaj sekundah. Posledično je celoten proces določanja nukleotidnega zaporedja izjemno hiter.

1.3 Pregled literature

V preteklih desetletjih se je tehnologija sekvenciranja močno razvila. Relativno dolge in cenovno dražje odčitke, pridobljene s Sangerjevo metodo sekvenciranja, so zamenjali krajši in cenejši odčitki, pridobljeni s sekvenatorji za visokopretočno sekvenciranje. Slednji so postali zelo zmogljivi in natančni, glavni izziv pa ostaja združevanje krajših DNA zaporedij v daljše soseske ali kontige (angl. contig). Področje so preplavili številni zbirniki, zasnovani na različnih metodah združevanja in z inovativnimi algoritmi za odpravljanje napak. Še vedno pa se ni noben uveljavil kot najboljši, saj so rezultati različni tako glede na obdelavo odčitkov različnih sekvenatorjev ali orodij za simulacijo, kot tudi glede na vrsto proučevanega organizma [8]. Posledično so raziskovalci večkrat skozi omenjeno obdobje primerjali različne zbirnike.

Prvo večje testiranje zbirnikov za kratka zaporedja so opravili in objavili leta 2011 v okviru članka Assemblathon 1. Avtorji so pripravili več podatkovnih setov pridobljenih s simulatorjem odčitkov in le te poslali različnim skupinam. Nazaj so dobili 41 združenih genomov od 17 različnih raziskovalnih skupin. Rezultate so primerjali med seboj z različnimi statstičnimi metodami (N50 in NG50), s prekrivanjem večih zaporedij, glede na pokritost proučevanega genoma, pregledom števila napak pri združevanju in drugih parametrih. Ugotovili so, da so zbirniki za obdelavo sekvenc sposobni doseči visoko stopnjo pokritosti in natančnosti, vendar obstajajo velike razlike med zbirniki samimi [9].

Podobno so raziskovalci izvedli raziskavo Assemblaton 2, kjer so primerjali novejše zbirnike, ki so začeli za združevanje odčitkov uporabljati de Bruijnove grafe, na pravih podatkih (namesto simuliranih) treh različnih vrst živali (ptica, riba in kača). Tokrat so avtorji prejeli 43 primerov združenih odčitkov od 21 različnih sodelujočih raziskovalnih skupin.
Poleg razvoja zbirnikov skozi številne posodobitve in pojavov novih so se razvile tudi nove metode za vrednotenje rezultatov zbirnikov. Drugi problem pri sami primerjavi rezultatov je povzročalo dejstvo, da niso imeli na voljo še točno določnih referenčnih zaporedij DNA za preiskovane tri organizme.

Raziskovalci so prišli do zaključka, da noben zbirnik ni »najboljši«, saj so bili nekateri boljši po nekem merilu in med zadnjimi po drugem. Prav tako so ugotovili, da so nekateri zbirniki zelo dobri za združevanje odčitkov ene vrste in slabši za drugi dve vrsti [10].

Leta 2017 so E. Foroozan in sodelavci objavili članek, v katerem so primerjali devet popularnih nelicenčnih zbirnikov za združevanje kratkih sekvenc, pridobljenih iz genoma mikroorganizmov. Pri raziskavi so uporabili odčitke, pridobljene s tehnologijo Illumina in simulirane odčitke za primerjavo. Med drugim so uporabili naslednje zbirnike: SPAdes, Velvet, SOAPdenovo2, Ray, IDBA-UD in MaSuRCA. Prišli so do ugotovitve, da so nekateri zbirniki boljši pri delu s simuliranimi zaporedji in drugi s pravimi. Glavna razlika med vrstama podatkov je porazdelitev zaporedij – pri podatkih pridobljenih iz organizma se pogosto pojavijo odstopanja od normalne porazdelitve odčitkov. Raziskovalci so ugotovili, da na pravih podatkih delujejo bolje zbirniki, ki pri združevanju uporabljajo več k-merov (več različnih velikosti). Ugotovili so tudi, da se ta odstopanja povečujejo z večanjem odstopanja odčitkov. Na pravih podatkih sta se najbolje odrezala SPAdes in IDBA-UD, prvi pa je dosegel boljše rezultate na simuliranih podatkih v primerjavi z drugim. Po drugi strani pa je SPAdes potreboval več časa in delovnega spomina.
Raziskovalci so za primerjavo in vrednotenje rezultatov uporabili orodje QUAST. Glede na vrednosti NGA50 je dosegel najboljše rezultate zbirnik MaSuRCA, sledila pa sta mu Ray in ABySS. Po drugi strani pa je ravno zbirnik MaSuRCA napravil največ napak, najmanj pa SPAdes, IDBA-UD in Velvet. Slednji se je izkazal za najbolj konzervativni zbirnik (t.j. združno zaporedje je bilo najbolj podobno pravemu nukleotidnemu zaporedju DNA). Avtorji so zaključili, da sta najboljša zbirnika SPAdes in IDBA-UD, v primerih, ko je ključnega pomena čim manjše število napak, pa Velvet. Poudarili so tudi, da to velja le za verzije zbirnikov, ki so jih uporabili oni, in se lahko s časom in novimi verzijami testiranih orodij za združevanje odčitkov DNA ti rezultati spremenijo [11].

2 MATERIALI IN METODE

2.1 Raziskovalna oprema

Vsi programi so bili naloženi na zmogljiv računalnik (128 GB RAM, 2 16-jedrna procesorja / 32 logičnih jeder), ki ga poganja Ubuntu server 16.04.6 LTS. Zaradi slabše optimizacije porabe delovnega spomina nekaterih orodij za združevanje DNA zaporedij smo le ta namestili in testirali na drugem serverju z velikostjo delovenega spomina 750 GB RAM. Slednji ima na voljo 80 logičnih jeder in ga poganja Ubuntu server 18.04.2 LTS.

2.2 Priprava podatkovnih setov za preiskušanje zbirnikov

Za analizo združevanja kratkih zaporedij ali odčitkov z različnimi orodji smo uporabili set nukleotidnih zaporedij jedrne DNA laškega smilja, *Helichrysum italicum, ki* so bila predhodno pridobljena v laboratoriju UP FAMNIT s sekvenatorjem Ion Torrent Ions S5 [1]. V setu je bilo 17.025.076 odčitkov s povprečno dolžino 245 baznih parov (bp). Za analizo smo uporabili podatkovni set z obliko zapisa uBAM (angl. Unmapped BAM format) in v obliki zapisa FASTQ (tekstovna datoteka z informacijo o nukletodnem zaporedju in Q vrednostmi (tudi vrednosti Phred oziroma ocene zanesljivosti določitve posamezne baze (angl. Phred quality score)).
Najprej smo preverili kakovost odčitkov z orodjem FastQC verzija v0.11.9 [12]. Slednji je namenjen iskanju potencialnih težav v podatkih, kot npr. nizke vrednosti Q (tudi vrednosti Phred oziroma ocene zanesljivosti določitve posamezne baze (angl. phred quality score)), prisotnosti ostankov adapterjev, itd.
Kljub temu, da so nukleotidna zaporedja adapterjev odstranjena že pri obdelavi surovih odčitkov, pridobljenih iz sekvenatorja s priloženo programsko opremo, smo z orodjem Cutadapt verzije 2.9 [13] preverili, če so v podatkovnem setu ostali morebitni ostanki adapterjev. Cutadapt je namenjen iskanju in odstranjevanju zaporedij adapterjev, začetnih oligonukleotidov, poli A repov (v primeru sekvenciranja RNA) ter ostalih neželenih zaporedij [14]. Program je spisan v jeziku Python, uporabljeni algoritem za poravnavo pa zaradi hitrosti delovanja v C. Program predstavlja enostavno rešitev za uporabo in pripravo odčitkov na nadaljno analizo [14]. Z njim lahko odstranimo tudi baze glede na Phred vrednost. Za preizkušanje orodij za združevanje zaporedij smo z orodjem Cutadapt pripravili dva podatkovna seta, pri čemer smo prvega pridobili zgolj z ukazom za odstranjevanje adapterjev, za drugi set pa smo uporabili tudi ukaz za odstranitev baz s Phred vrednostjo nižjo od 20. Uporabili smo naslednja dva ukaza (vhodna datoteka je bila v obliki FASTQ):

- Za pridobitev podatkovnega seta, kjer so bili odstranjeni le deli odčitkov, ki so pripadali adapterjem (datoteko v nadaljevanju imenujemo *no_qual.*)

```
$ cutadapt -a GTCTCAGCCTCTCTATGGGCAGTCGGTGAT --cores=30 -g
CCATCTCATCCCTGCGTGTCTCCGACTCAG  Helichrysum_podatkovni_set.fastq >
no_qual.fastq 2> report_no_qual.txt
```

- Za pridobitev podatkovnega seta, kjer so bili poleg ostankov adapterjev odstranjeni tudi deli odčitkov glede na Q vrednost posamezne baze (datoteko v nadaljevanju imenujemo *qual_20*)

```
$ cutadapt -a GTCTCAGCCTCTCTATGGGCAGTCGGTGAT --cores=30 -g
CCATCTCATCCCTGCGTGTCTCCGACTCAG -q 20 Helichrysum_podatkovni_set.fastq >
qual_20.fastq 2> report_qual_20.txt
```

S parametrom *-a* smo določili komplementarno nukleotidno zaporedje P1B adapterja v smeri 5'-3', s parametrom *-g* pa nukleotidno zaporedje adapterja A, ki je prisotno na začetku (na 5' koncu) odčitkov.

Vsa orodja za združevanje odčitkov DNA, ki so predstavljena v nadaljevanju, razen CLC Genomics Workbench, smo pognali nad obema fastq datotekama. Orodje SPAdes pa smo pognali tudi z uporabo uBAM datoteke, ker poleg FASTQ podpira tudi omenjeno obliko zapisa. Odstranitev morebitnih ostankov adapterjev pred analizo s CLC Genomics Workbench je bila izvedena s filtrom, ki je vključen v CLC Genomics Workbench.

2.3 Orodja za združevanje odčitkov DNA

Zbirniki lahko za združevanje zaporedij DNA uporabljajo številne algoritme. Pri zbirnikih nove generacije je najpogosteje implementiran eden izmed naslednjih dveh algoritmov: prvi algoritem temelji na izgradnji de Bruijnovega grafa (DBG), drugi pa na soglasju prekrivanja (angl. overlap layout consensus (OLC)).
De Bruijnov graf je algoritem, ki temelji na uporabi k-merov – DNA odčitki razrezani na manjša zaporedja dolžine k (k je pozitivno število). Za vsako k-mero se nato zgradijo vozlišča grafa in ustvarijo povezave, kjer se pozamezne k-mere (deli odčitka) prekrivajo (dolžina prekrivanja k-1). Na koncu dobimo usmerjen graf [15].
Pri OLC algoritmu se najprej definirajo vsi pari prekrivajočih se odčitkov. Iz njih se nato zgradi graf, kjer vozlišča predstavljajo odčitke, povezave pa kateri odčitki se prekrivajo. Zaporedje DNA je nato določeno z iskanjem možnih poti po zgrajenem grafu [16].
Za oba algoritma velja, da se največ časa porabi pri iskanju prekrivanj odčitkov. Zbirniki DBG so hitrejši, OLC pa imajo boljše rezultate pri združevanju daljših odčitkov [17].
Večina zbirnikov, ki smo jih primerjali, temelji na DBG algoritmu. Vse razen CLC Genomics Workbench smo namestili na prej omenjeni računalnik. Zbirniki in delo z njimi je predstavljeno v nadaljevanju.

2.3.1 SPAdes

SPAdes je bil razvit kot rešitev za himerne povezave (ki nastanejo zaradi himernih odčitkov – to so tisti odčitki, kjer del zaporedja prilega na eno mesto, drugi del odčitka pa na neko drugo mesto združenega nukleotidnega zaporedja) in reševanje kompleksnejših primerov na de Bruijnovih grafih, kjer med dvema točkama obstaja več poti z vmesnimi točkami. Omenjene težave se pojavljajo pri združevanju zaporedij DNA, še posebej pri odčitkih, pridobljenih na ravni ene same celice (npr. pri sekvenciranju genomov bakterij, ki jih ni mogoče gojiti na gojiščih). Z razvitim algoritmom so avtorji pokazali možnost določitve genomov pri hkratnem sekvenciranju manjšega števila različnih celic (t.i. mini-metagenom) [18].

SPAdes ima od verzije 3.0 dalje na voljo dodatno možnost poteka združevanja, ki vključuje dodaten korak za odpravljanje napak pri združevanju z IonHammer algoritmom. Slednji je namenjen ravno odpravljanju specifičnih napak, ki so značilne za odčitke pridobljene s tehnologijo Ion Torrent. Algoritem je sestavljen iz dveh korakov: v prvem koraku predvidi katere hk-mere (homopolimerska k-mera (angl. homopolymerspace k-mer)) so dobre, v drugem pa se izvede odpravljanje napak. Algoritem poskusi vse slabše hk-mere popraviti in istočasno narediti čim manj sprememb [19].

2.3.1.1 Priprava programa

SPAdes je med drugim na voljo v obliki binarne datoteke. Slednjo smo prenesli na naš server s pomočjo ukaza wget. Datoteko je bilo potrebno pred uporabo programa le še razširiti. V mapi bin je bila skripta spades.py, s katero smo pognali SPAdes (v 3.14.1).

```
#prenos datoteke s programom
$ wget http://cab.spbu.ru/files/release3.14.1/SPAdes-3.14.1-Linux.tar.gz

#razširitev stisnjene datoteke
$ tar -xzf SPAdes-3.14.1-Linux.tar.gz

#navigacija v mapo, ki vsebuje skripto
$ cd SPAdes-3.14.1-Linux/bin/
```

2.3.1.2 Postopek za izvedbo združevanja odčitkov

```
# --iontorrent zastavica je potrebna za združevanje odčitkov pridobljenih s
tehnologijo Ion Torrent, omogoča tudi uporabo uBAM datateke
# -t parameter določa število niti
# -m parameter omeji porabo delovnega spomina (RAM)
# -k parameter sprejme listo k-mer

$ spades.py --iontorrent -s qual_20.fastq -t 26 -m 110 -k 21,33,55,77,99,127 -o
spades_rezultati_qual_20
```

```
$ spades.py --iontorrent -s no_qual.fastq -t 26 -m 110 -k 21,33,55,77,99,127 -o
spades_rezultati_no_qual

$ spades.py --iontorrent -s Helichrysum_podatkovni_set.bam -t 26 -m 110 -k
21,33,55,77,99,127 -o spades_rezultati_bam
```

2.3.2 ABySS

Prvotno je bil ABySS (ABySS 1.0) implementiran z MPI (add info). Njegova posebnost je bila možnost delovanja s kratkimi odčitki (50 baznih parov), vendar je za svoje delovanje potreboval velike količine delovnega spomina (več 100 GB RAM).
Naslednja verzija programa (ABySS 2.0) je opustila MPI in namesto njega implementirala Bloomov filter – verjetnostno podatkovno strukturo za reprezentacijo de Bruijnovega grafa. Slednja implementacija omogoča združevanje zaporedij DNA na manjših količinah delovnega spomina (< 32 GB) [20].

2.3.2.1 Priprava programa

Nalaganje ABySS (v.1.9.0) je enostavno, saj je na voljo v glavnem Ubuntu repozitoriju. Naložili smo ga z naslednjim ukazom:

```
$ sudo apt-get install abyss
```

2.3.2.2 Postopek za izvedbo združevanja odčitkov

ABySS je pri izvajanju analize nad FASTQ datoteko javil, da so v datoteki prisotne vrstice brez nukleotidnega zaporedja. Slednje smo odstranili s pomočjo programa Bioawk [21], ki je med drugim na voljo kot bioconda paket (v nadaljevanju bomo za tako pripravljeno datoteko uporabili ime *qual_20_bioawk*). Bioawk smo za odstranitev vrstic brez zaporedij pognali z naslednjim ukazom:

```
$ bioawk -cfastx 'length($seq) > 1 {print "@"$name"\n"$seq"\n+\n"$qual}'
qual_20.fastq > qual_20_bioawk.fastq
```

Nato smo pognali nad obema setoma (*no_qual* in *qual_20_bioawk*) podatkov ABySS z ukazoma:

- Za vhodno datoteko *qual_20_bioawk*
  ```
  $ abyss-pe np=8 k=96 name=helichrysum se=qual_20_bioawk.fastq
  ```
- Za vhodno datoteko *no_qual*
  ```
  $ abyss-pe np=8 k=96 name=helichrysum se=no_qual.fastq
  ```

2.3.3 MEGAHIT

MEGAHIT je ultra hitro in spominsko varčno orodje nove generacje za združevanje zaporedij DNA. Optimiziran je za delo z metagenomskimi vzorci, prav tako pa je zanesljiv za generično združevanje posameznih manjših genomov oziroma genomov velikosti sesalcev ter določanje zaporedja enoceličnih organizmov [22].

2.3.3.1 Priprava programa

MEGAHIT (v.1.2.9) je na voljo kot bioconda paket in smo ga naložili z ukazom:

```
$ conda install -c bioconda megahit
```

2.3.3.2 Postopek za izvedbo združevanja odčitkov

Program smo pognali s privzetimi nastavitvami. Določili smo le listo k-mer (21, 33, 55, 77, 99, 127).
Ukaza sta bila torej naslednja:

- Za podatkovni set *no_qual*

```
# --k-list paramtere sprejme listo k-mer

$ megahit -r no_qual.fastq --k-list 21,33,55,77,99,127 -o no_qual
```

- Za podatkovni set *qual_20*

```
$ megahit -r qual_20.fastq --k-list 21,33,55,77,99,127 -o qual_20
```

2.3.4 Minia

Minia je orodje za združevanje kratkih DNA zaporedij in temelji na de Bruijnovih grafih. Z orodjem je mogoče združiti celoten genom človeka v enem samem dnevu na domačem računalniku. Rezultati prvih različic programa so bili enakovredni ostalim zbirnikom, ki temeljijo na de Bruijnovih grafih, tako v natančnosti kot tudi glede na statistični parameter N50 (parameter, ki določa zadnjo najkrajšo dolžino združenega zaporedja oz. soseske, ki v seštevku z ostalimi daljšimi soseskami predstavlja polovico skupne dolžine vseh sosesk). Po letu 2015 so zbirnik razvili in izboljšali (večji N50). Posledično je izšla GATB-Minia-Pipeline, ki temelji na Minia, vendar podobno kot SPADes in MEGAHIT ter drugi lahko obdela podatke za več k-merov hkrati [23].

2.3.4.1 Priprava programa

Zbirnik Minia smo naložili na server z ukazom wget. Stisnjeno datoteko s programom smo pridobili z Github spletne strani v zavihku *Izdaje* (angl. releases). Uporabili smo Minia verzije 3.2.2.

2.3.4.2 Postopek za izvedbo združevanja odčitkov

Zbirnik smo pognali z naslednjima dvema ukazoma:

- za podatkovni set *qual_20*

```
$ ./minia -in qual_20.fastq -out qual_20
```

- za podatkovni set *no_qual*

```
$ ./minia -in no_qual.fastq -out no_qual
```

2.3.5 GATB-Minia-Pipeline

Kot omenjeno pri poglavju Minia, GATB-Minia-Pipeline temelji na zbirniku kratkih DNA zaporedij Minia. Torej prav tako uporablja algoritem, ki temelji na de Bruijnovih grafih. Kot prej omenjeno je glavna razlika izvajanje algoritma nad večjim številom k-merov, kar vodi do boljših rezultatov (N50).
Program je sestavljen iz treh delov: Minia 3 za združevanje zaporedij, Bloocoo za odstranjevanje napak in BESST za združevanje sosesk v ogrodja (angl. scaffolding) [24].

2.3.5.1 Priprava programa

Nalaganje GATB-Minia-Pipeline je bilo drugačno kot pri večini ostalih zbirnikov. Najprej smo naložiti okolje python2 za uporabnika BESST. Nato smo klonirali projekt z Githuba preko gita na server in v prenešeni mapi pognali zbirnik.

```
$ python2 -m pip install --user BESST
```

```
$ git clone --recursive https://github.com/GATB/gatb-minia-pipeline

$ cd gatb-minia-pipeline ; make test
```

2.3.5.2 Postopek za izvedbo združevanja odčitkov

Združevanje smo pognali z naslednjima ukazoma:

- za podatkovni set *no_qual*

```
$ ./gatb -s no_qual.fastq -o no_qual
```

- za podatkovni set *qual_20*

```
$ ./gatb -s qual_20.fastq -o qual_20
```

2.3.6 SOAPdenovo2

SoapDenovo2 je združevalnik kratkih odčitkov DNA nove generacije, ki lahko sestavi genome v velikosti človeškega. Program je specifično prilagojen za Illumina GA kratke odčitke. SOAPdenovo2 je odprl nove možnosti priprave referenčnih zaporedij in natančnejših analiz še nedoločenih DNA zaporedij (zaporedij organizmov, ki še niso bila določena) z nizkimi stroški [25].
SOAPdenovo2 je prinesel veliko izboljšav v primerjavi s prvotno verzijo. Novi algoritem potrebuje za svoje delovanje manjše količine delovnega spomina za izgradnjo de Bruinovega grafa, prepozna več regij s ponavljajočimi zaporedji, poveča pokritost in dolžino ogrodja (angl. scaffold length) ter je bolje optimiziran za večje genome [26].

2.3.6.1 Priprava programa

Zbirnik SOAPdenovo2 verzije 2.04 smo naložili na server z ukazom wget. Stisnjeno datoteko s programom smo pridobili z Github spletne strani v zavihku *Izdaje* (angl. releases). Slednjo smo razširili in pognali.

2.3.6.2 Postopek za izvedbo združevanja odčitkov

Program smo zagnali z naslednjima ukazoma:

- za podatkovni set *no_qual*

```
# all je del osnovnega ukaza, ki določa nabor parametrov
# -p parameter določa število niti

$ SOAPdenovo-127mer all -s no_qual.config -o no_qual -p 26 1>no_qual.log
2>20_qual.txt
```

- za podatkovni set *qual_20*

```
$ SOAPdenovo-127mer all -s qual_20.config -o qual_20 -p 26 1>qual_20.log
2>qual_20.txt
```

2.3.7 MIRA

MIRA je zbirnik za določevanje zaporedja celotnega genoma, EST in RNA zaporedja [27]. Orodje temelji na uporabi OLG algoritma in večkratne strategije določanja zaporedja nukleotidov. MIRA je tako edini testirani zbirnik, ki ne temelji na uporabi DBG.

2.3.7.1 Priprava programa

Zbirnik MIRA (v4.9.6) smo naložili kot conda paket.

```
$ conda install -c bioconda mira
```

2.3.7.2 Postopek za izvedbo združevanja odčitkov

Mira potrebuje za zagon konfiguracijsko datoteko (datoteka formata .conf), zato smo pripravili novi datoteki za oba podatkovna seta:

manifest_ qual_20.conf

```
project = qual_20                # ime mape, v kateri so shranjeni podatki
job = genome,denovo,accurate     # združi genom v natančnem načinu
parameters = -GE:not=4           # osnovni parametri: uporabi 4 niti
readgroup = shotgunlibrary       # oznaka
data = data/Helichrysum01.2018-07-24T11_29_18Z_ADAPT_TRIMMED_qual_20.fastq
technology = iontor              # odčitki so pridobljeni s tehnologijo Ion Torrent
```

manifest_no_qual.conf

```
project = no_qual
job = genome,denovo,accurate
parameters = -GE:not=4
readgroup = shotgunlibrary
data = data/Helichrysum01.2018-07-24T11_29_18Z_ADAPT_TRIMMED_no_qual.fastq
technology = iontor
```

Nato smo pognali združevanje z MIRA z naslednjima ukazoma:

```
# -t parameter določa število niti

$ mira -t 30 manifest_no_qual.conf >&log_assembly_no_qual.txt

$ mira -t 30 manifest_qual_20.conf >&log_assembly_qual_20.txt
```

Združevanje odčitkov z zbirnikom MIRA se je večkrat prekinilo (eno od sporočil opisuje napako »MAF datoteka se konča z odprtim odčitkom«). Prvotno smo sumili, da je programu zmanjkalo delovnega spomina, zato smo podatkovna seta prenesli na močnejši računalnik opisan v poglavju 2.1. Na slednjem smo uspeli združiti podatkovni set in sicer *qual_20*. Zaradi časovne kompleksnosti poganjanja zbirnika (večji del programa se izvaja zaporedno) nismo uspeli združiti seta *no_qual*.

2.3.8 CLC

CLC je licenčni zbirnik nove generacije in temelji na uporabi de Bruijnovih grafov. Isti zbirnik se uporablja v okviru CLC Genomics Workbench, CLC Genomics Server in CLC Assembly Cell 4.0. Uporablja se lahko za združevanje Illumina, 454, SOLiD, Ion Torrent in Sanger sequencing [28].

2.3.8.1 Obdelava DNA zaporedij

Odčitke laškega smilja je z zbirnikom CLC Genomics Workbench združil prof. dr. Jernej Jakše na Biotehniški fakulteti Univerze v Ljubljani. Uporabil je de-novo zbirnik (verzije 1.4), implementiran v CLC Genomics Server verzije 10.0.1 z opcijo nalaganja odčitkov na sestavljene soseske. Programu so bili podani naslednji parametri:

- Mapping mode = Map reads back to contigs (slow)
- Update contigs = Yes
- Automatic bubble size = Yes
- Minimum contig length = 200
- Automatic word size = Yes
- Mismatch cost = 2
- Insertion cost = 3
- Deletion cost = 3
- Length fraction = 0.5
- Similarity fraction = 0.8
- Create list of un-mapped reads = Yes

2.4 Primerjava statističnih parametrov združenih zaporedij s programom QUAST

QUAST je program namenjen vrednotenju kakovosti združenih zaporedij. QUAST ima implementirane metode za izračun parametrov kakovosti združenih zaporedij iz programov kot so Plantagora, GAGE, GeneMark.hmm in GlimmerHMM, dodatno pa vključuje tudi nekaj novih (npr. NA50 je nadgrajena statistična mera N50). Orodje deluje tako z referenčnim genomom, kot tudi brez njega. Slednja lastnost velja za prednost orodja QUAST. Za poravnavo in ovrednotenje le te uporablja Nucmerjev zbirnik [29].

2.4.1 Statistični parametri oziroma ocene kakovosti združenih zaporedij

Parametre, ki jih je mogoče ovrednotiti z orodjem QUAST, lahko razdelimo v več skupin.

2.4.1.1 Dolžina sosesk

Večina meritev za izračun ne potrebuje referenčnega genoma, ga pa lahko uporabi, če je le ta podan. Izjema je NGx. V to kategorijo spadajo:

- število sosesk
- najdaljša soseska
- skupna dolžina vseh sosesk (število baz v sestavljeni DNA)

- Nx – (kjer je 0 <= x <= 100) dolžina najdaljše soseske L, pri čemer soseske, daljše od L predstavljajo vsaj x % dolžine vseh sosesk
- NGx, Genome Nx – podobno kot pri Nx, le da primerjamo x procente z referenčnim genomom

2.4.1.2 Napake in strukturne variacije

Za izračun strukturnih variacij QUAST potrebuje referenčni genom. Do razlik med referenčnih genomom in ponovno sestavljenim istim genomom lahko pride zaradi načina združevanja odčitkov oziroma napak zbirnikov ali pa so posledica himernih odčitkov – torej posledica sekvenciranja. V primeru razlik pri primerjavi združenih zaporedij organizma, ki je soroden organizmu, za katerega je bil določen referenčni genom, pa so lahko te razlike prave strukturne variacije, kot npr. daljše insercije ali delecije, razlike v številu kopij določenega zaporedja itd. Vrednosti, ki jih program poda, so:

- število neujemanj oz. prekinitev, kot jih določa Plantagorjevo pravilo: soseska je pri nalaganju na referenčni genom prekinjena z regijo referenčnega genoma daljšo od 1 kb; soseska naleže na nasprotno orientirani verigi ali na različna kromosoma
- število sosesk s prekinitvami
- dolžina vseh sosesk, pri katerih so bile identificirane prekinitve
- število sosesk, ki jih ni bilo mogoče poravnati z referenčnim genomom
- število sosesk, ki jih je mogoče poravnati na več mest v referenčnem genomu

Poleg zgoraj naštetih vrednosti QUAST izdela poročilo, ki za vsako pregledano sosesko poda podatek vključno z: ali ga ni bilo mogoče poravnati z referenčnim genomom, ali ga je mogoče ravnati na več mest, ali je bil napačno poravnan ali poravnan pravilno.

2.4.1.3 Reprezentativnost genoma in njegovi funkcijski elementi

Tako kot parametre iz prejšnjega sklopa, je tudi večino teh (npr. kolikšen del referenčnega genoma predstavljajo soseske, povprečno število neujemanj na 100 kb poravnanih baz, število insercij in delecij na 100 kb poravnanih baz, koliko funkcijskih elementov, kot npr. geni in operoni, je prisotnih na soseskah) mogoče oceniti na osnovi primerjave z referenčnim genomom.

V primeru, da referenčni genom ni na voljo, je mogoče oceniti število genov na soseskah z orodjem BUSCO (analiza z orodjem BUSCO je predstavljena v nadaljevanju).

2.4.1.4 Zadnja skupina parametrov

Zadnjo skupino parametrov (različice N50 glede na poravnane dele sosesk), ki jih omenjajo, je mogoče izračunati le s pomočjo referenčnega genoma [29].

NAx in NGAx sta nova parametra, s katerima so avtorji orodja QUAST želeli odpraviti možne napake pri interpretaciji vrednosti N50. Obe omenjeni meritvi potrebujeta za izračun referenčni genom.
NAx je kombinacija prej omenjenih parametrov Nx in števila prekinitev. Izračun poteka v dveh korakih. V prvem koraku se soseske, pri katerih so bile identificirane prekinitve ali regije, ki jih ni bilo mogoče poravnati, na mestih prekinitev ali neporavnanih regij delijo na manjše, t.i. poravnane bloke. Soseske, ki vsebujejo napake združevanja, so razdeljene na manjše enote na mestih, kjer je prišlo do napak. Podobno so razdeljene tudi soseske, ki vsebujejo neporavnane regije. Pri teh se najprej odstranijo omenjene regije, nato pa se soseske razbije na bloke. V drugem koraku se na vseh blokih izvede statistična analiza Nx. Razlika je torej v analizi blokov in ne originalnih sosesk.

2.4.2 Grafična predstavitev rezultatov

QUAST na osnovi izračunanih parametrov izriše grafe v različnih formatih (PNG, PDF in SVG). Glede na podatke, ki jih grafi prikazujejo, so le-ti razdeljeni v 5 skupin.

2.4.2.1 Grafi Nx

Prikazujejo trende vrednosti Nx, NGx, Nax ali NGAx glede na spremenljivko x. Ti grafi so v primerjavi z meritvami N50 bolj informativni.

2.4.2.2 Kumulativni grafi

Soseske so urejene po dolžini (število baz) od največjega do najmanjšega za vse tipe obravnavanih kumulativnih grafov.

2.4.2.3 Grafi vsebnosti GC

Prikazujejo razporeditev vsebnosti GC v soseskah. Na x osi so odstotki (0 – 100) in na y osi število neprekrivajočih zaporedij dolžine 100 bp, ki vsebujejo GC v x odstotkih. Običajno krivulje kažejo Gaussovo porazdelitev, medtem, ko v primeru kontaminacij v podatkovnem setu pride do superpozicij večih Gaussovih razporeditev.

2.4.2.4 Grafi poravnave sosesk

Prikazujejo poravnavo sosesk glede na referenčni genom in mesta, kjer so bloki nepravilno združenih odčitkov.

2.4.2.5 Primerjalni histrogrami

Prikaz primerjalih histogramov za več različnih statističnih vrednosti. Med drugim lahko prikažemo histrogram za število celih operonov, število celih genov in genomska frakcija v procentih.

2.4.3 Priprava programa

Orodje QUAST smo na naš računalnik namestili v obliki conda paketa. Za namestitev smo uporabili naslednji ukaz:

```
$ conda install -c bioconda quast
```

2.4.4 Izvedba analize združenih zaporedij z orodjem QUAST

Vse FASTA datoteke združenih DNA odčitkov smo skopirali (datoteke smo pri kopiranju tudi preimenovali na način 'quast_' + ime zbirnika) v isto mapo in Quast pognali preko vseh FASTA datotek z naslednjim ukazom:

```
# -t parameter označuje število niti
# quast_* pomeni, da se analizirajo vsi končni združeni odčitki (vse .fasta
datoteke)

$ quast.py -t 30 quast_* -o quast_all_report
```

2.5 BUSCO

BUSCO (angl. Benchmarking Universal SingleCopy Orthologs) je bioinformacijsko orodje za vrednotenje združenih zaporedij ali genomov. Orodje je na voljo kot Python 3 paket in vključuje druge programe za anotacijo in primerjavo zaporedij (tBLASTn, AUGUSTUS Gene Predictor, Prodigal, HMMER, SEPP, itd.) z nukleotidnimi zaporedji genov iz podatkovne zbirke OrthoDB (BUSCO uporablja le set genov, ki so pri 90 % vrst prisotni le v eni kopiji). Program lahko sprejme in izvede analizo za tri glavne tipe podatkov in sicer za genomsko DNA pridobljeno z zbirnikom, anotirane genske sete in transkriptome. Odvisno od podatkov je potrebno pri poganjanju izbrati tudi ustrezen tip analize. Glede na slednjega se izvede zanj specifično zaporedje korakov obdelave podatkov. Za delovanje programski paket potrebuje tudi konfiguracijsko datoteko. To lahko prilagajamo svojim potrebam ali pa uporabimo osnovno verzijo. Na koncu BUSCO zmeraj vrne rezultate v enakem formatu. Med rezultati tako zmeraj dobimo podatke o številu kompletnih t.i. skupin BUSCO, oziroma natančneje kompletnih in enkrat indentificiranih skupin BUSCO ter kompletnih in podvojenih skupin BUSCO, delno identificiranih (razdrobljenih) skupin BUSCO ter

manjkajočih skupin BUSCO. Python paket zajema tudi vizualno predstavitev končnih podatkov. Do slednjih lahko pridemo tako, da zaženemo skripto, ki v okolju R uporabi paket ggplot2 za izdelavo grafov in na koncu vrne grafe za podane podatke [30].

2.5.1 Priprava programa

Programsko orodje Busco je na voljo na več načinov. Med glavnima dvema preko conda paketa in kot docker slika. Mi smo se odločili za namestitev na strojno opremo v obliki bioconda paketa.

```
# namestitev Busco (skupaj s potrebnim paketom augustus)
$ conda install -c bioconda -c conda-forge busco=4.1.2 augustus=3.3.3

# aktivacija okolja, v katerem lahko poganjamo orodje Busco
$ conda activate base
```

2.5.2 Izvedba analize z orodjem BUSCO

Podobno kot pri uporabi orodja Quast smo vse FASTA datoteke kopirali v novo mapo, ki smo jo tokrat poimenovali Busco. V njej smo za vsako datoteko posebej pognali naslednji ukaz, kjer je parameter *input* bila FASTA datoteka in parameter *output* ime zbirnika + '*_busco'*, s katerim so bili odčitki združeni:

```
# -m parameter določa način analize glede na vhodne podatke (genom / proteini /
transkriptom)
# -l parameter določa podatkovni set ortologov, prisotnih v sorodnih vrstah
# -t parameter določa število niti

$ busco -m genome -i input -o output -l eudicots_odb10 -t 30
```

Orodje Busco je za vsako datoteko naredil novo mapo imenovano po *output* parametru. V njej se po končani analizi nahaja tudi txt datoteka z rezultati. Slednje smo nato zbrali v tabelo 4, ki je prikazana v poglavju 3.

3 REZULTATI

3.1 Cutadapt

Tabela 1: Poročilo orodja Cutadapt za pripravo podatkovnih setov no_qual in qual_20

podatkovni set	*no_qual*		*qual_20*	
enota	število	odstotek	število	odstotek
število vseh odčitkov	17025150	100	17025150	100
število odčitkov z adapterjem	383177	2,3	396936	2,3
število zapisanih odčitkov	17025150	100	17025150	100
število vseh baznih parov	4218692879	100	4218692879	100
število bp z nizko vrednostjo Q	/	/	81971750	1,9
število zapisanih baznih parov	4218692879	100	4135395085	98

Iz tabele 1 je razvidno, da je bilo v začetnem podatkovnem setu 17.025.150 odčitkov skupne dolžine 4.218.692.879 bp. Set *no_qual* je imel 383.177 odčitkov z ostanki adapterjev, set *qual_20* pa 396.936.

Pri obeh končnih setih je število odčitkov (zapisanih odčitkov) ostalo enako (tabela 1), vendar smo kasneje ugotoivli, da orodje Cutadapt podatkov za odčitke, ki so bili s filtrom odstranjeni v celoti, po privzetih nastavitvah ne odstrani (več v naslednjem poglavju). Pri podatkovnem setu *qual_20* je bilo 81.971.750 bp manj kot v setu *no_qual*. Toliko baz je bilo odstranjenih zaradi dodatne filtracije glede na Q vrednost (slika 1 in 2).

3.2 FastQC

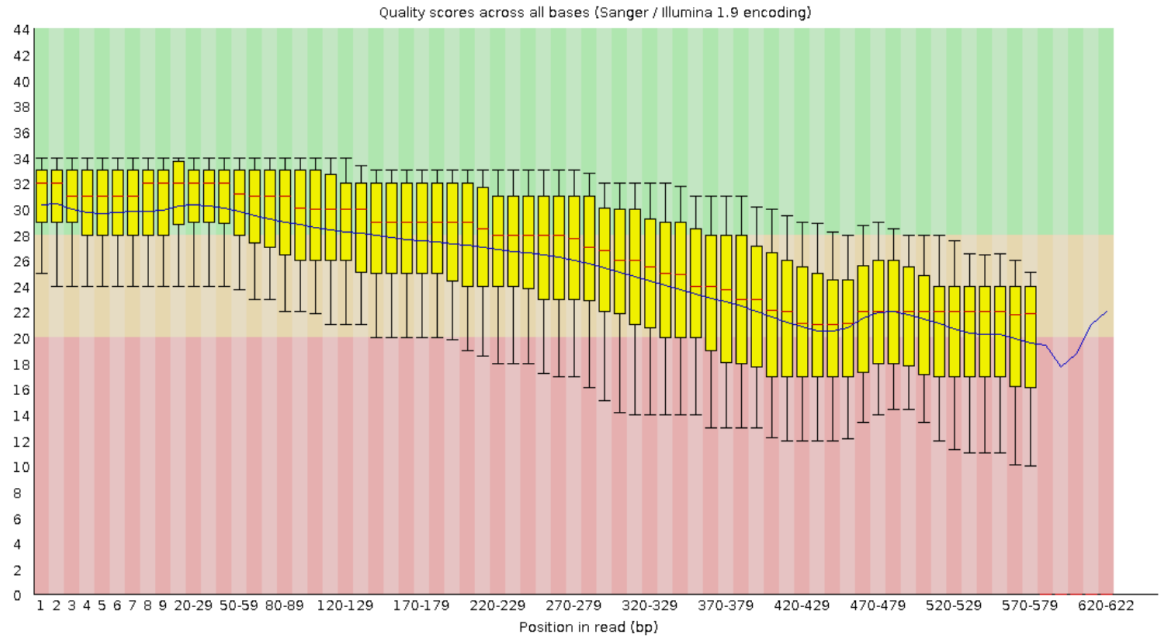
Tabela 2: Primerjava podatkov pridobljenih z orodjem FastQC

	Začetni podatki	*no_qual*	*qual_20*	*qual_20_bioawk*
Število odčitkov	17025150	17025150	17025150	17013344
Število očitkov nizke kakovosti	0	0	0	0
Dolžine odčitkov	25-622	17-622	0-622	2-622
%GC	36	36	36	36

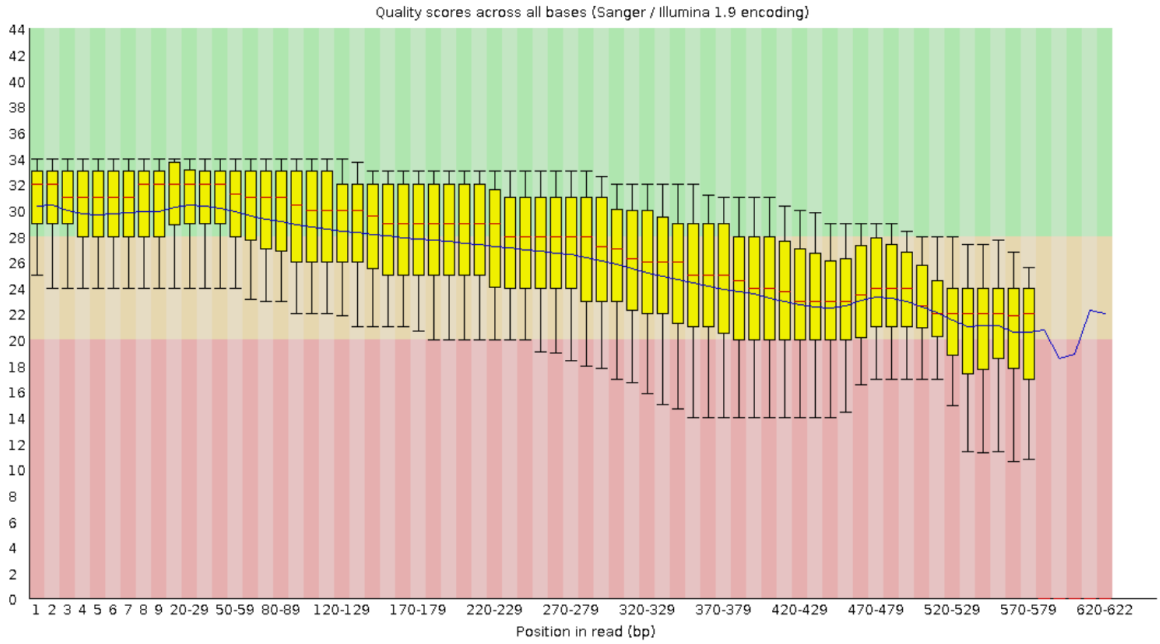
Kot omenjeno v poglavju 3.1 smo po filtraciji baz glede na Q vrednost pridobili podatek, da je ostalo enako število odčitkov v podatkovnih setih *no_qual* in *qual_20* (tabela 2). Vendar kot je razvidno iz preglednice, so bili v to število šteti tudi podatki o odčitkih, pri katerih je bilo v celoti odstranjeno nukleotidno zaporedje. Po odstranitvi le-teh z orodjem Bioawk smo prodobili točno število vseh preostalih odčitkov (17.013.344). Z omenjenim programom smo torej odstranili 11.806 praznih zaporedij (zaporedja dolžine 0).

V nobenem podatkovnem setu ni bilo zaporedij nizke kakovosti (glede na parametre orodja FastQC), delež nukleotidov gvanina in citozina pa je bil pri vseh setih 36 %.

Dolžine odčitkov so se pri surovih odčitkih gibale med 25 bp in 622 pb. Po odstranitvi zaporedij adapterjev s programom Cutadapt so bile dolžine med 17 in 622 baznih parov. Po filtriranju odčitkov glede na Phred oceno kakovosti 20, so se dolžine gibale med 0 bp (preostali le podatki o odčitkih brez nukleotidnih zaporedij) in 622 bp. Ko smo na koncu odstranili prazna zaporedja z orodjem Bioawk, pa so se dolžine zaporedij gibale med 2 in 622 baznih parov.



Slika 1: Prikaz razporeditve Q vrednosti glede na dolžino zaporedij za podatkovni set *no_qual* po odstranitvi adapterjev z orodjem Cutadapt

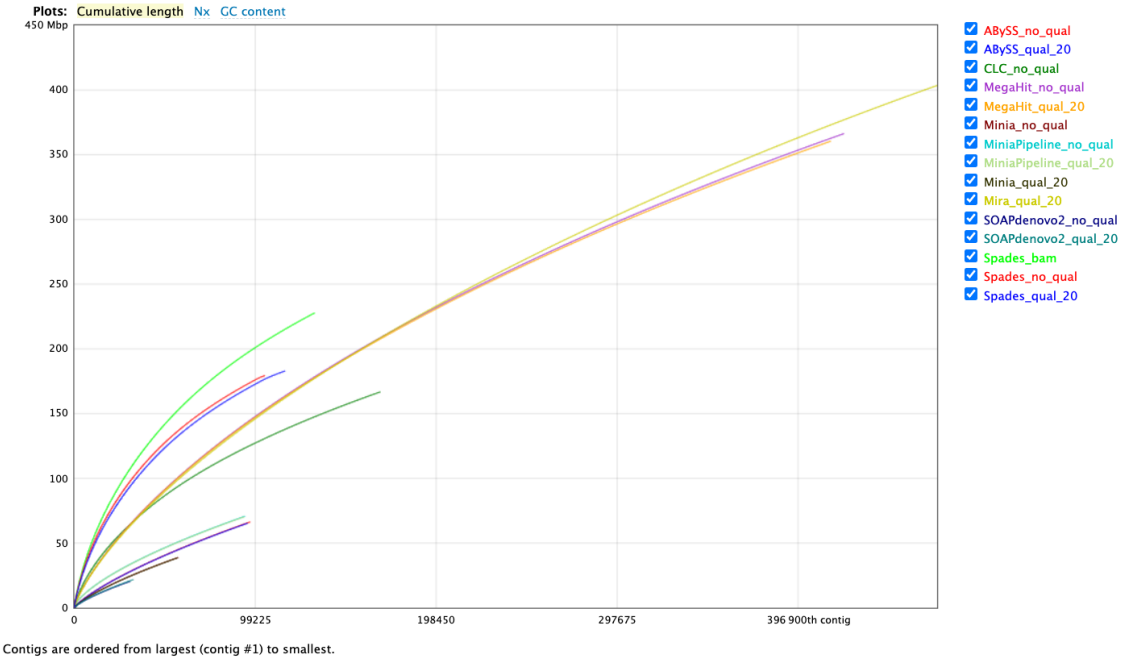


Slika 2: Prikaz rezporeditve Q vrednosti glede na dolžino zaporedij za podatkovni set *qual_20* (po opravljeni filtraciji s parametrom -q 20 z orodjem Cutadapt in odstranitvi adapterjev)

3.3 QUAST

Tabela 3: Rezultati programskega orodja QUAST

Statistics without reference	# contigs	# contigs (>= 0 bp)	# contigs (>= 1000 bp)	# contigs (>= 5000 bp)	# contigs (>= 10000 bp)	# contigs (>= 25000 bp)	Largest contig	Total length	Total length (>= 0 bp)	Total length (>= 1000 bp)	Total length (>= 5000 bp)	Total length (>= 10000 bp)	Total length (>= 25000 bp)	N50	N75	L50	L75	GC (%)
AbySS no_qual	96160	1750726	7315	2	0	0	5224	66202387	412499256	9034958	10263	0	0	666	568	37968	65001	35.36
AbySS qual_20	94935	1730116	7129	2	0	0	5224	65284727	408320894	8805733	10263	0	0	665	567	37514	64199	35.38
CLC no_qual	167699	450590	52205	512	11	1	48844	166744105	258617590	89242172	3156407	175666	48844	1069	690	46526	95964	34.33
MegaHit no_qual	421999	1041143	100048	154	6	0	15821	366311549	587364708	148611294	957595	73482	0	872	648	137150	260212	34.36
MegaHit qual_20	414719	1008898	98593	154	8	0	17265	360422604	572811942	146572017	971842	103170	0	873	649	134689	255672	34.38
Minia no_qual	56999	2531846	3942	2	1	0	10722	38895374	282189818	4856553	16053	10722	0	660	566	22667	38657	34.96
Minia Pipeline no_qual	93536	104168	12236	119	16	4	40855	70707784	75718025	18117685	996744	326346	139340	717	582	33028	60607	35.33
Minia Pipeline qual_20	93026	103596	12098	113	14	4	40855	70172810	75153804	17834496	948384	301044	139305	716	582	32939	60340	35.34
Minia qual_20	56399	2486571	3837	2	1	0	10722	38441495	279278622	4722984	16053	10722	0	659	565	22449	38264	34.98
Mira qual_20	473376	714713	102843	168	12	0	16732	403515790	498074210	149639483	1055921	141533	0	846	650	159741	296689	35.52
SOAPdenovo2 no_qual	30388	22750855	1837	0	0	0	3336	20279395	1030356921	2270482	0	0	0	640	557	12222	20752	35.17
SOAPdenovo2 qual_20	32326	22090413	2123	0	0	0	3573	21722027	1001860582	2640110	0	0	0	643	558	12914	22018	35.09
Spades bam	131750	149096	90066	4058	217	1	27794	227631626	231136821	192202014	26982520	2632741	27794	1989	1216	33736	70564	34.33
Spades no_qual	104503	186905	69869	3182	123	1	26177	179415781	196237679	150797679	20845862	1485910	26177	2019	1218	26453	55144	34.45
Spades qual_20	115425	619416	70987	2625	63	0	16578	182884675	293883106	148200739	16725185	731152	0	1877	1140	29320	60641	34.45



Slika 3: Graf komulativnih dolžin

Rezultati analize s programskim orodjem za vrednotenje kakovosti združenih zaporedij QUAST so prikazani v tabeli 3. Iz slednje je razvidno, da je najdaljšo sosesko združil program CLC in sicer dolžine 48.844 baznih parov. Sledila sta zbirnika GATB Minia Pipeline in SPAdes z dolžinama 40.855 in 27.794 baznih parov. Najkrajšo sosesko je sestavil program SOAPdenovo2.

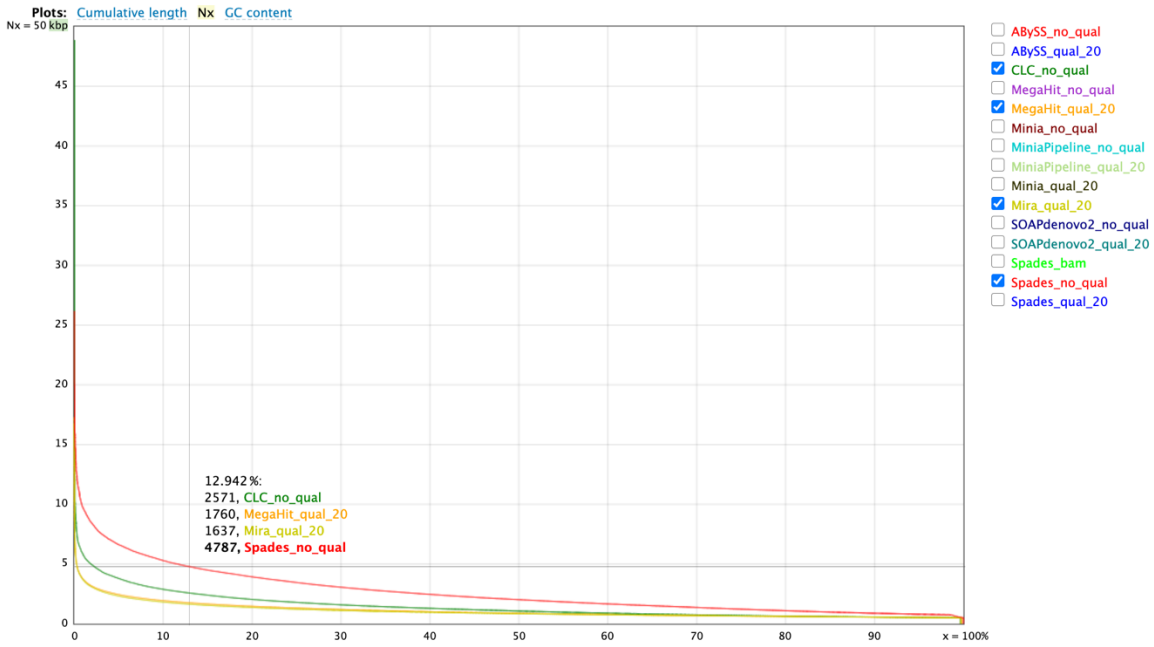
Dolžina celotnega sestavljenega DNA zaporedja je bila najdaljša pri obdelavi odčitkov z orodjem MEGAHIT in sicer 366.311.549 pri obdelavi *no_qual* seta in nekoliko krajše – 360.422.604 baznih parov – pri zaporedjih podatkovnega seta *qual_20*. Drugo najdaljše zaporedje smo pridobili z orodjem SPAdes. Iz podatkov v uBAM datoteki je končna dolžina merila 227.631.626, pri ostalih dveh setih pa 179415781 bp in 182884675 bp. Na tretjem mestu je orodje CLC s skupno dolžino 166.744.105 baznih parov. Najkrajše zaporedje DNA je združilo orodje SOAPdenovo2 in sicer skupnih dolžin 2.270.482 bp in 2.640.110 bp.

Vrednost analize N50 je bila najvišja pri uporabi orodja SPAdes in sicer se je gibala med 1.877 in 2.019. Sledili sta vrednosti orodij CLC in MEGAHIT z rezultatoma 1.069 in 873. Najnižji rezultat N50 je dosegel program SOAPdenovo2 z vrednostjo 640. Preostali zbirniki so dosegli podobne rezultate pri merjenju vrednosti N75. Rezultati si torej sledijo v istem zaporedju, vse vrednosti pa so ne glede na zbirnik nekoliko nižje.

Orodja za združevanje DNA odčitkov so dosegla slabše rezultate glede na parametra L50 in L75. Najnižjo vrednost je dosegel program SOAPdenovo2 z rezultatoma 12.222 in 12.914. Sledila sta programa Minia in Spades z gibanjem rezultatov okoli 22.500 in 27.500. Najvišjo vrednost in s tem najslabši rezultat je dosegel program MEGAHIT in sicer približno 135.000. Zaporedje rezultatov L50 in L75 je bilo ponovno v istem zaporedju, vrednosti pri vseh programih pa so bile višje (po večini približno 2-kratne vrednosti meritev L50).

Delež nukleotidov gvanina in citozina je bil pri vseh obdelavah podatkov približno enak in se je gibal med 34,33 in 35,38 %.

Komulativni graf na sliki 3 prikazuje komulativno dolžino vseh sosesk glede na ševilo sosesk. Iz njega lahko razberemo, da sta zbirnika MEGAHIT in MIRA imela največje število sosesk in največjo skupno dolžino. Najmanj sosesk in najmanjšo skupno dolžino pa je imel SOAPdenovo2.



Slika 4: Graf Nx za najboljše štiri zbirnike

3.4 Busco

Tabela 4: Primerjava rezultatov analize z orodjem BUSCO

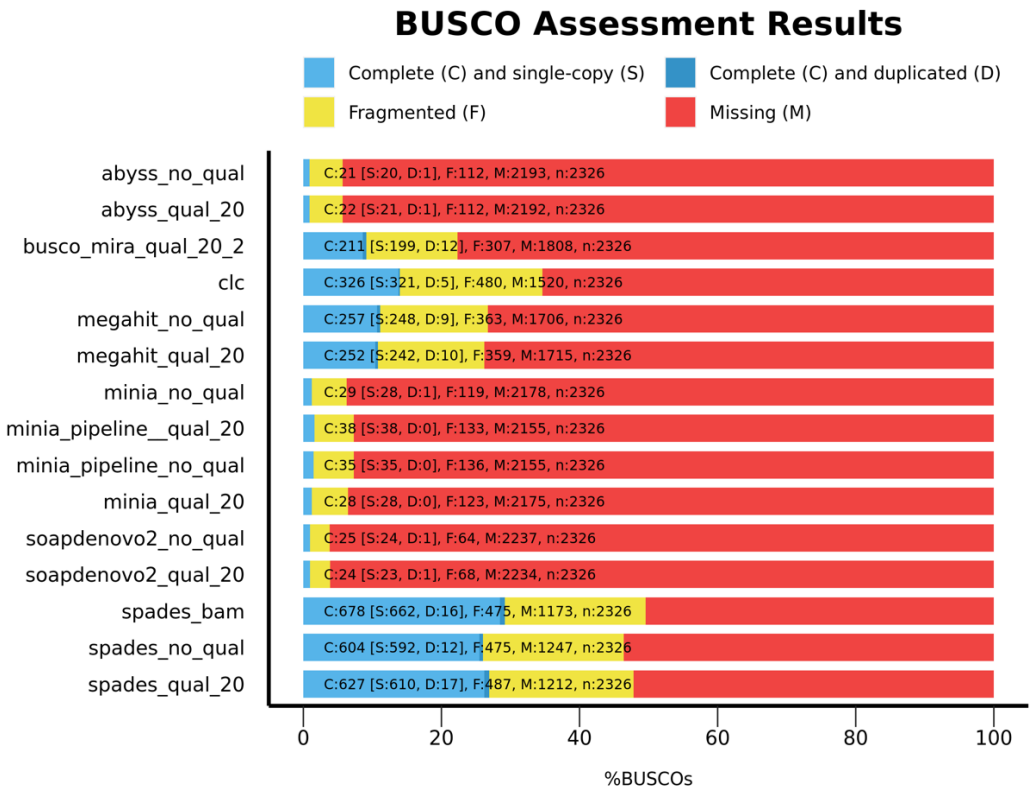
Zbirnik	*Enota*	*Kompletne skupine BUSCO*	*Kompletne in enkrat Indentificirane skupine BUSCO*	*Kompletne in podvojene skupine BUSCO*	*Razdrobljene skupine BUSCO*	*Manjkajoče skupine BUSCO*	*Število preiskanih skupin BUSCO*
ABySS - no qual	Število	21	20	1	112	2193	2326
	%	0,9	0,9	0	4,8	94,3	100
ABySS - qual 20	Število	22	21	1	112	2192	2326
	%	0,9	0,9	0	4,8	94,3	100
MEGAHIT - no qual	Število	257	248	9	363	1706	2326
	%	11,1	10,7	0,4	15,6	73,3	100
MEGAHIT - qual 20	Število	252	242	10	359	1715	2326
	%	10,8	10,4	0,4	15,4	73,8	100
Minia - no qual	Število	29	28	1	119	2178	2326
	%	1,2	1,2	0	5,1	93,7	100
Minia - qual 20	Število	28	28	0	123	2175	2326
	%	1,2	1,2	0	5,3	93,5	100
GATB Minia Pipeline - no qual	Število	35	35	0	136	2155	2326
	%	1,5	1,5	0	5,8	92,7	100
GATB Minia Pipeline - qual 20	Število	38	38	0	133	2155	2326
	%	1,6	1,6	0	5,7	92,7	100
SOAPdenovo 2 - no qual	Število	25	14	1	64	2237	2326
	%	1	1	0	2,8	96,2	100
SOAPdenovo 2 - qual 20	Število	24	23	1	68	2234	2326
	%	1	1	0	2,9	96,1	100
SPAdes - no qual	Število	604	592	12	475	1247	2326
	%	26	25,5	0,5	20,4	53,6	100
SPAdes - qual 20	Število	627	610	17	487	1212	2326
	%	26,9	26,2	0,7	20,9	52,2	100
SPAdes - bam	Število	678	662	16	475	1173	2326
	%	29,2	28,5	0,7	20,4	50,4	100
CLC	Število	326	321	5	480	1520	2326
	%	14	13,8	0,2	20,6	65,4	100
MIRA - no qual	Število	211	199	12	307	1808	2326
	%	9,1	8,6	0,5	13,2	77,7	100

V tabeli 3 so prikazani rezultati analize z orodjem BUSCO. Iz nje je razvidno, da so razlike med podatkovnima setoma *qual_20* in *no_qual* zelo majhne, natančneje manj kot 1,4 %. Pri vseh združenih zaporedjih je bilo preiskanih 2.326 skupin BUSCO. Pri večini združevanj z zbirniki je podatkovni set *qual_20* imel zaznanih več skupin BUSCO (kompletnih, kompletnih-enkratnih in kompletnih-ponovljenih) in manj manjkajočih v primerjavi z *no_qual*. V vsaki skupini je prišlo do nekaj odstopanj in sicer MEGAHIT in Minia sta združila več kompletnih skupin BUSCO pri nefiltriranem setu (*no_qual*). Podobno velja za ostali dve skupini kompletnih skupin BUSCO. V skupini enkrat identificiranih je bilo zaznanih več skupin BUSCO pri podatkovnem setu *no_qual* le pri programu MEGAHIT in pri podvojenih le pri Minia. Za delno prisotne skupine BUSCO se vzorec rezultatov ponovi: nefiltriran podatkovni set ima več odkritih le pri orodju MEGAHIT. Podobni rezultati so tudi za zadnjo skupino – manjkajoče skupine BUSCO. Tudi tukaj so podatki slabši za filtrirane podatke, nasprotno velja le pri obdelavi zaporedij z zbirnikom MEGAHIT.

Iz tabele 3 lahko razberemo tudi, da je največ kompletnih skupin BUSCO program zaznal pri odčitkih združenih z orodjem za združevanje DNA odčitkov SPAdes. Med tremi podatkovnimi seti jih je bilo največ identificiranih pri podatkovnem setu odčitkov v uBAM datoteki in sicer 678, kar predstavlja 29,2 %. Sledi set *qual_20* s 627 (26,9 %) in *no_qual* s 604 (26 %). Drugo največje število kompletnih skupin je združil zbirnik CLC in sicer 326 oziroma 14 %. Za njim sledi MEGAHIT s 257 in 252 kompletnimi skupinami BUSCO, kar predstavlja 11,1 in 10,8 %. Pridobljeni rezultati prikazujejo samo majhne spremembe med kompletnimi in kompletnimi – enkrat identificiranimi skupinami BUSCO. Podatki za kompletne – podvojene skupine sledijo podobnemu vzorcu. Največje število takih genov je bilo zaznanih pri uporabi zbirnika SPAdes (12 - 17). Drugi po vrsti je bil zbirnik MIRA z rezultatom 12 kompletnih – podvojenih skupin. Sledilo je orodje MEGAHIT z rezultatoma 9 in 10. Rezultati za ostala orodja za združevanje DNA zaporedij so bili po večini 0 ali 1 z izjemo CLC (5 podvojenih skupin BUSCO).

Fragmentiranih skupin je bilo največ pri podatkovnem setu *qual_20,* obdelanim z zbirnikoma SPAdes in CLC (487 in 480). Sledila sta nefiltrirana podatkovna seta, združena s SPAdes. Zbirnik MEGAHIT je ponovno združil tretje največje število skupin BUSCO in sicer 363 za nefiltriran in 359 za filtriran set. Sledijo MIRA, GATB Minia Pipeline, Minia, ABySS in SOAPdenovo2 z rezultati 307, 136, 123, 112 in 68 v tem vrstnem redu.

Manjkajočih skupin BUSCO je bilo zaznanih najmanj pri odčitkih združenih z orodjem SPAdes (~ 50 %). Sledili so zbirniki CLC s 65,4 %, MEGAHIT z dobrimi 73 % in nato MIRA s 77 % manjkajočih skupin. Pri rezultatih ostalih programov je bilo manjkajočih skupin BUSCO več kot 92 %. Največ manjkajočih je bilo pri združevanju z orodjem SOAPdenovo2 in sicer dobrih 96 %.



Slika 5: Grafični prikaz rezultatov pridobljenih z orodjem BUSCO

4 DISKUSIJA

4.1 QUAST

4.1.1 Statistična parametra Nx in Lx

N50 parameter označuje dolžino soseske, za katero velja, da soseske, ki so enako dolge ali daljše, pokrivajo vsaj polovico (>= 50 %) celotnega združenega zaporedja odčitkov. Podobno velja za N75, le da v tem primeru soseske pokrivajo vsaj 75 % združenih odčitkov. L50 predstavlja minimalno število sosesk potrebnih za pokrivanje 50 % celotnega združenega zaporedja odčitkov. Podobno velja za L75, kjer število L75 pomeni, koliko sosesk je potrebno združiti za 75 % delež celotnega združenega zaporedja.

V tabeli 3 lahko vidimo, da je zbirnik SPAdes dosegel najboljši rezultat tako za N50 z vrednostmi med 1.877 in 2.019, kot tudi za N75 z vrednostmi med 1.140 in 1.218 (daljše zaporedje oz. manj združenih sosesk pomeni boljši rezultat). Na drugo mesto se je uvrstil CLC z rezultatoma 1.069 pri merjenju N50 in 690 pri N75. Tretje mesto je glede na N50 dosegel zbirnik MEGAHIT, glede na N75 pa MIRA. Iz grafa na sliki 4 je razvidno tudi, da je Nx vrednost za zbrinik SPAdes pri vseh vrednostih x (0 – 100) najvišja, sledi krivulja za CLC in na zadnjih dveh mestih sta krivulji za MEGAHIT in MIRA.

Prav tako lahko iz tabele 3 razberemo, da je najboljše rezultate L50 (12.222 in 12.914) in L75 (20.752 in 22.018) dosegel zbirnik SOAPdenovo2, ki mu je sledila Minia z rezultatoma 22.667 in 22.449 pri L50 ter 38.657 in 38.264 pri merjenju L75, vendar zaradi krajše skupne dolžine združenega zaporedja v primerjavi z ostalimi boljšimi zbirniki menimo, da ti rezultati niso primerni za primerjavo.. Na tretje mesto se je z nekoliko slabšimi rezultati uvrstil zbirnik SPAdes (L50 *no_qual*: 26.453 in L75 *no_qual*: 55.144).

4.1.2 Dolžine sosesk in združenih odčitkov

Najdaljšo sosesko smo s programom QUAST zaznali pri zbirniku CLC (48.844). Sledila sta GATB-Minia-Pipeline (40.855 pri obeh setih) in SPAdes (set bam 27.794 in set *no_qual* 26.177).
Najdaljše zaporedje nukleotidov je združil zbirnik MIRA in sicer dolžine 403.515.790. Sledil je zbirnik MEGAHIT s skupnima dolžinama 366.311.549 in 360.422.604. Na tretjem mestu se je ponovno znašel SPAdes z dolžinami 227.631.626, 179.415.781 in 182.884.675.

V tabeli 3 lahko vidimo tudi, da so razlike med podatkovnimi seti zelo majhne. To pomeni, da filtriranje odčitkov, ki so imeli vrednost Q manjšo od 20, ni imelo večjega učinka na združevanje zaporedij laškega smilja. Q vrednosti za pozamezni podatkovni set lahko

vidimo na sliki 1 (vrednosti za set *no_qual*) in sliki 2 (vrednosti za set *qual_20*). Iz tabele 1 lahko razberemo, da je bilo pri odstranjevanju nukleotidov s Q vrednostjo manjšo od 20 odstranjenih 81.971.750 nukleotidov, kar predstavlja 1,9 %. Glede na majhna odstopanja rezultatov za oba seta lahko predvidevamo, da bi manjkajoča seta podatkov za CLC in MIRA imela prav tako podobne rezultate kot uporabljena seta podatkov.

Glede na vrednosti, pridobljene z orodjem za primerjavo združenih odčitkov, lahko zaključimo, da je za združevanje odčitkov laškega smilja najboljše rezutate dosegel zbirnik SPAdes, saj je dosegel daleč najboljše rezultate glede na statistična parametra N50 in N75. Poleg tega je bil med najboljšimi tremi tudi glede na dolžino končnega združenega zaporedja odčitkov in vrednosti L50 ter L75. Zbirnik SPAdes je imel sicer prednost pred ostalimi, saj zbirnik vsebuje algoritem za popravljanje napak, ki je specializiran za delo z odčitki, pridobljenimi s tehnologijo Ion Torrent [19]. Po drugi strani pa so pri testiranju devetih zbirnikov raziskovalci pri združevanju odčitkov, pridobljenih s tehnologijo Illumina, prišli prav tako do zaključka, da je najbolj primeren zbirnik za obdelavo pravih (nesimuliranih) podatkov SPAdes [11].

4.2 BUSCO

4.2.1 Kompletne skupine BUSCO

V skupino kompletnih skupin BUSCO spadajo vsi tisti geni, pri katerih so bili rezultati ujemanja v pričakovanem intervalu vrednosti in pričakovanem intervalu dolžin poravnanih zaporedij. V določnih primerih, ko nek ortolog ni prisoten ali je delno prisoten (zelo razdrobljen) v podatkovnem setu in slednji vsebuje veliko enakih nukleotidov na istem mestu (t.j. se ujema po strukturi) kot kompleten homologni gen, obstaja majhna možnost, da je ta homolog pomotoma zaznan kot pripadnik skupine BUSCO [30]. V to skupino spadata tudi obe podskupini: kompletne in enkrat identificirane ter kompletne in podvojene skupine BUSCO.
Rezultati, ki smo jih dobili pri analizi vseh združenih zaporedij z orodjem BUSCO, kažejo na veliko razliko v številu kompletnih skupin BUSCO med različnimi zbirniki. SPAdes, ki je združil največje število kompletnih skupin (604 – 678 zaznanih genov oz. 26 – 29,2 %), je v primerjavi z drugouvrščenim CLC (326 oz. 14 % zaznanih genov) združil približno 2-krat toliko ortologov. Zbirnika MEGAHIT in MIRA sta združila nekoliko manj celotnih skupin BUSCO kot CLC (257 – 211 oz. 11,1 – 9,1 %). Pri ostalih štirih smo zaznali manj kot 40 (1,6 %) kompletnih skupin BUSCO.
Število identificiranih kompletnih skupin je lahko posledica biološkega ali tehničnega izvora. Ker smo vse zbirnike primerjali glede na združevanje istih podatkov, lahko v tem primeru razlike pripišemo delovanju zbirnikov.

Velika večina kompletnih skupin BUSCO je bila v združenih zaporedjih identificirana enkrat, kar istočasno pomeni, da smo imeli zelo majhno število podvojenih skupin. Največ podvojenih skupin BUSCO je bilo sicer zaznanih pri uporabi SPAdes, MIRA in MEGAHIT, vendar ni nikjer preseglo 1 %.

4.2.2 Razdrobljene skupine BUSCO

Med razdrobljene skupine BUSCO spadajo vsi geni, ki so po rezultatih ujemanja v pričakovanem intervalu vrednosti, vendar niso znotraj intervala dolžin poravnav. Za naš preučevani genom to pomeni, da je tak gen lahko prisoten delno v genomu ali pa je prisoten v celoti in le ni bil združen v celoti. Določevanje razdrobljenih skupin poteka v dveh krogih in sicer tisti, ki so v prvem krogu označeni kot razdrobljeni, so v drugem krogu iskani z novimi parametri, ki so trenirani na podatkih kompletnih skupin BUSCO [30].
Kot je razvidno iz rezultatov, je število razdrobljenih skupin BUSCO pri vseh zbirnikih višje od števila kompletnih skupin z izjemo SPAdes. Slednji in CLC sta imela po rezultatih največ razdrobljenih skupin BUSCO (487 – 475 ter 480). Sledita MEGAHIT in MIRA z več kot 300 razdrobljenimi skupinami. Najmanj razdrobljenih skupin BUSCO je bilo zabeleženih pri orodju SOAPdenovo2 (68 in 64), pri ostalih zbirnikih pa se je rezultat gibal med 112 in 136.

4.2.3 Manjkajoče skupine BUSCO

Skupina BUSCO je označena kot manjkajoča v dveh primerih:

1. zaznana ujemanja niso bila v pričakovem intervalu ali ujemanju,
2. ujemanj sploh ni bilo zaznanih.

Povodov za uvrstitev v manjkajoče skupine BUSCO je več. Iskan ortolog lahko v genomu proučevanega organizma sploh ni prisoten ali pa je v združenih odčitkih premalo zastopan oz. pokrit, da ni zaznan in uvrščen niti med razdrobljene skupine BUSCO. Obstaja tudi majhna možnost, da je iskani gen prisoten, vendar je zelo kompleksen ali pa obstaja več različnih varianc ali različic tega gena in ga BUSCO ne prepozna. Podobno kot pri iskanju razdrobljenih skupin, se tudi v tem primeru ponovi iskanje s primerjanjem genov z že zaznanimi kompletnimi skupinami BUSCO [30].
Kot je razvidno iz rezultatov tabele 3 in slike 1, je večina zbirnikov združila manj kot 10 % kompletnih in razdrobljenih skupin skupaj. Najmanj ortologov je bilo manjkajočih pri združevanju DNA zaporedij z zbirnikom SPAdes, sledijo CLC, MEGAHIT in MIRA. Glede na odstotke, je najmanj manjkajočih skupin BUSCO pri vzrocu *spades_bam* in sicer 50,4 % (1.173 manjkajočih genov).

Glede na podatke pridobljenih z orodjem BUSCO, je najboljše rezultate za združevanje odčitkov laškega smilja dosegel zbirnik SPAdes. Klub najvišjemu številu duplikatov je bilo število kompletnih genov daleč največje. Prav tako je dosegel največje število razdrobljenih fragmentov in skupaj sestavil 1.153 (49,6 %) ortologov iz BUSCO podatkovenga seta *eudicots_odb10*, ki vsebuje 2.326 genov. Na drugo mesto se je uvrstil zbirnik CLC, ki je skupaj sestavil 806 (34,6 %) ortologov, in na tretje mesto MEGAHIT s 620 (26,7 %) identificiranimi skupinami BUSCO.

Za genom laškega smilja so raziskovalci ocenili, da vsebuje 3.247 Mpb oz. 1.660 Mbp za haploidni set kromosomov [31]. Glede na podatke prikazane v tabeli 3 lahko zaključimo, da smo z najdaljšimi tremi združenimi zaporedji pokrili med 13 % in 25 % celotnega genoma laškega smilja. Nizka fizična pokritost genoma z odčitki pa je najverjetneje razlog, da smo dobili nizke povprečne odstotke identificiranih skupin BUSCO (glede na vse primerjane zbirnike).

5 ZAKLJUČEK

V diplomski nalogi smo s pomočjo večih orodij za vrednotenje kakovosti združnih zaporedij ugotovili, da je bil pri združevanju odčitkov laškega smilja s tehnologijo Ion Torrent najboljši zbirnik SPAdes. Slednji je bil daleč najbolj uspešen glede na število identificiranih skupin BUSCO in glede na statistični parameter N50 (pridobljen z orodjem QUAST).

V prihodnje bi bilo potrebno preveriti, če bi lahko z optimizacijo parametrov za posamezni zbirnik pridobili boljše rezultate. Eden od bolj pomembnih parametrov (razen za MIRA, ki ne temelji na DBG) je ravno velikost k-mer, ki se pri večini zbirnikov lahko istočasno uporablja le ena.

Zanimivo bi bilo tudi primerjati iste zbirnike z istimi parametri na drugem podatkovnem setu, pridobljenem s tehnologijo Ion Torrent. Tako bi lahko preverili, kakšen pomen je imel na rezultate uporabljen podatkovni set odčitkov laškega smilja.

6 LITERATURA IN VIRI

[1] M. Hladnik, A. Baruca Arbeiter, T. Knap, J. Jakše, and D. Bandelj, "The complete chloroplast genome of *Helichrysum italicum* (Roth) G. Don (Asteraceae)," *Mitochondrial DNA Part B*, vol. 4, no. 1, pp. 1036–1037, Jan. 2019, doi: 10.1080/23802359.2019.1580156.

[2] M. Jeršek, "Vojna za smilj," *Gea*, vol. 26, no. 9, pp. 40–43, 2016.

[3] R. Perrini, V. Alba, C. Ruta, I. Morone-Fortunato, A. Blanco, and C. Montemurro, "An evaluation of a new approach to the regeneration of Helichrysum italicum (Roth) G. Don, and the molecular characterization of the variation among sets of differently derived regenerants," *Cell. Mol. Biol. Lett.*, vol. 14, no. 3, pp. 377–394, Sep. 2009, doi: 10.2478/s11658-009-0007-3.

[4] M. Galbany-Casals, J. M. Blanco-Moreno, N. Garcia-Jacas, I. Breitwieser, and R. D. Smissen, "Genetic variation in Mediterranean Helichrysum italicum (Asteraceae; Gnaphalieae): do disjunct populations of subsp. microphyllum have a common origin?," *Plant Biol.*, vol. 13, no. 4, pp. 678–687, Jul. 2011, doi: 10.1111/j.1438-8677.2010.00411.x.

[5] R. Tundis *et al.*, "Influence of environmental factors on composition of volatile constituents and biological activity of Helichrysum italicum (Roth) Don (Asteraceae)," *Nat. Prod. Res.*, vol. 19, no. 4, pp. 379–387, Jun. 2005, doi: 10.1080/14786410420002619 69.

[6] G. Appendino *et al.*, "Arzanol, an anti-inflammatory and anti-HIV-1 phloroglucinol α-pyrone from Helichrysum italicum ssp. microphyllum," *J. Nat. Prod.*, vol. 70, no. 4, pp. 608–612, Apr. 2007, doi: 10.1021/np060581r.

[7] "Semiconductor Sequencing Technology | Thermo Fisher Scientific - SI." https://www.thermofisher.com/si/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html (accessed Aug. 15, 2020).

[8] S. L. Salzberg *et al.*, "GAGE: A critical evaluation of genome assemblies and assembly algorithms," *Genome Res.*, vol. 22, no. 3, pp. 557–567, Mar. 2012, doi: 10.1101/gr.131383.111.

[9] D. Earl *et al.*, "Assemblathon 1: A competitive assessment of de novo short read assembly methods," *Genome Research*, vol. 21, no. 12. Cold Spring Harbor Laboratory Press, pp. 2224–2241, Dec. 2011, doi: 10.1101/gr.126599.111.

[10] K. R. Bradnam *et al.*, "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species," 2013. doi: 10.1186/2047-217X-2-10.

[11] E. Forouzan, M. S. M. Maleki, A. A. Karkhane, and B. Yakhchali, "Evaluation of nine popular de novo assemblers in microbial genome assembly," *J. Microbiol. Methods*, vol. 143, pp. 32–37, Dec. 2017, doi: 10.1016/j.mimet.2017.09.008.

[12] "Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput

Sequence Data." https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed Aug. 10, 2020).

[13] "marcelm/cutadapt at 9ce76e87d2f96c5369b054dbbd6ad83fa0c15f34." https://github.com/marcelm/cutadapt/tree/9ce76e87d2f96c5369b054dbbd6ad83fa0c15f34 (accessed Aug. 05, 2020).

[14] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, no. 1, p. 10, May 2011, doi: 10.14806/ej.17.1.200.

[15] A. Bankevich *et al.*, "SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing," *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–477, May 2012, doi: 10.1089/cmb.2012.0021.

[16] Y. T. Huang and C. F. Liao, "Integration of string and de Bruijn graphs for genome assembly," *Bioinformatics*, vol. 32, no. 9, pp. 1301–1307, May 2016, doi: 10.1093/bioinformatics/btw011.

[17] A. R. Khan, M. T. Pervez, M. E. Babar, N. Naveed, and M. Shoaib, "A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective," *Evol. Bioinforma.*, vol. 14, Feb. 2018, doi: 10.1177/1176934318758650.

[18] S. Nurk *et al.*, "Assembling genomes and mini-metagenomes from highly chimeric reads," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 7821 LNBI, pp. 158–170, doi: 10.1007/978-3-642-37195-0_13.

[19] V. Ershov, A. Tarasov, A. Lapidus, and A. Korobeynikov, "IonHammer: Homopolymer-Space Hamming Clustering for IonTorrent Read Error Correction," *J. Comput. Biol.*, vol. 26, no. 2, pp. 124–127, Feb. 2019, doi: 10.1089/cmb.2018.0152.

[20] S. D. Jackman *et al.*, "ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter," *Genome Res.*, vol. 27, no. 5, pp. 768–777, May 2017, doi: 10.1101/gr.214346.116.

[21] "lh3/bioawk: BWK awk modified for biological data." https://github.com/lh3/bioawk (accessed Jul. 20, 2020).

[22] D. Li *et al.*, "MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices," *Methods*, vol. 102. Academic Press Inc., pp. 3–11, Jun. 01, 2016, doi: 10.1016/j.ymeth.2016.02.020.

[23] "GATB/minia: Minia is a short-read assembler based on a de Bruijn graph." https://github.com/GATB/minia (accessed Jul. 25, 2020).

[24] "GATB/gatb-minia-pipeline: GATB Minia assembly pipeline." https://github.com/GATB/gatb-minia-pipeline (accessed Jul. 25, 2020).

[25] "aquaskyline/SOAPdenovo2: Next generation sequencing reads de novo assembler." https://github.com/aquaskyline/SOAPdenovo2 (accessed Jul. 28, 2020).

[26] R. Luo *et al.*, "SOAPdenovo2: An empirically improved memory-efficient short-read

de novo assembler," *GigaScience*, vol. 1, no. 1. Oxford University Press, p. 18, Dec. 27, 2012, doi: 10.1186/2047-217X-1-18.

[27] B. Chevreux *et al.*, "Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs," *Genome Res.*, vol. 14, no. 6, pp. 1147–1159, Jun. 2004, doi: 10.1101/gr.1917404.

[28] "White paper on de novo assembly in CLC Assembly Cell 4.0." https://digitalinsights.qiagen.com/files/whitepapers/whitepaper-denovo-assembly.pdf (accessed Aug. 18, 2020).

[29] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "QUAST: Quality assessment tool for genome assemblies," *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, Apr. 2013, doi: 10.1093/bioinformatics/btt086.

[30] M. Seppey, M. Manni, and E. M. Zdobnov, "BUSCO: Assessing genome assembly and annotation completeness," in *Methods in Molecular Biology*, vol. 1962, Humana Press Inc., 2019, pp. 227–245.

[31] S. Garcia *et al.*, "New data on genome size in 128 Asteraceae species and subspecies, with first assessments for 40 genera, 3 tribes and 2 subfamilies," *Plant Biosyst.*, vol. 147, no. 4, pp. 1219–1227, 2013, doi: 10.1080/11263504.2013.863811.