TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOINFORMAATIKA ÕPPETOOL

Bakteritüvede tuvastamine sekveneerimise toorlugemitest kindla pikkusega oligomeeride abil

Magistritöö

Geenitehnoloogia

30 EAP

Mihkel Vaher

Juhendaja: Märt Roosaare, M.Sc.

Tartu

2016

Bakteritüvede tuvastamine sekveneerimise toorlugemitest kindla pikkusega oligomeeride abil

Teise põlvkonna sekveneerimine võimaldab anda hulgaliselt infot proovide liigilise koosluse kohta, olgu selleks näiteks inimese mikrobioomi- või keskkonnaproov. Uuemad bakterite tuvastamise programmid kasutavad määramiseks lühikesi kindla pikkusega oligomeere, olles seeläbi kiired, kuid ka mitmete puudustega.

Antud töö käigus loodi tarkvara StrainSeeker, mis tuvastab sekveneerimise toorandmetest bakterid tüve tasemeni. StrainSeeker analüüsib kõigi lugemite koondinfot, mitte üksikuid lugemeid eraldi. Lisaks suudab StrainSeeker tuvastada ka andmebaasist puuduvaid organisme ning kiirus ja paindlikkus teevad selle võrreldavaks parimate avaldatud bakterite tuvastamise programmidega.

StrainSeeker on saadav ja veebis kasutatav aadressil www.bioinfo.ut.ee/strainseeker.

Märksõnad: bakterid, identifitseerimine, lugemid, assambleerimata, k-meer
CERCS: B110 (Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika)

Identifying bacterial strains from unassembled sequencing reads using fixed-length oligomers

Making use of next generation sequencing, metagenomic studies give valuable insights how different microbial communities affect human health. Recent bacterial identification programs use short, fixed-length oligomers but lack the option to customize the database and analyze individual short reads.

A bacterial identification software called StrainSeeker was designed, that can identify bacteria at the strain level by taking the info from many reads into account. Also, StrainSeeker provides options for database customization and can detect organisms not present in the database. StrainSeeker is comparable to any state-of-the-art bacterial identification tool, outperforming many in regard to speed and accuracy.

StrainSeeker is available and can be used online at www.bioinfo.ut.ee/strainseeker.

Keywords: bacteria, identification, reads, unassembled, k-mer
CERCS: B110 (Bioinformatics, medical informatics, biomathematics, biometrics)

SISUKORD

KASUTATUD LÜHENDID

16S – ribosoomi väikest subühikut kodeeriv ja seda ümbritsev ala

ap – aluspaar

C1 ja *C2* – käesoleva sõlme otsesed alamsõlmed (*child*)

E – eeldatud leitud sõlme k-meeride osakaal kõigist sõlme unikaalsetest k-meeridest protsentides (*expected*)

kontiig – assambleerimise käigus ülekattes olevate lugemite ühendamisel saadud pikem järjestus

k-meerid – järjestuse kõikvõimalikud kindla pikkusega *k* alamjärjestused

N – käesolev sõlm (*node*)

O – käesolevasse sõlme määratud k-meeride osakaal kõigist sõlme unikaalsetest k-meeridest protsentides (*observed*)

obs – sõlmele määratud k-meeride arv (*observed*)

S – tüvetase otsingualgoritmis; leht (*strain*)

tot – sõlme kõigi k-meeride arv (*total*)

SISSEJUHATUS

Mikroobikoosluste kirjeldamine aitab mõista nende rolli erinevates keskkondades – alates bakteritüvede poolt põhjustatud haigustest kuni mõlemale osapoolele kasuliku kooseluni inimese soolestiku mikrofloora puhul (Zeevi *et al*., 2015). Tihti toimub bakterite määramine perekonna või liigi tasemel, kuid sellest infost ei piisa, kui tüvede omadused üksteisest oluliselt erinevad. Näiteks *Escherichia coli* enamik liigi tüvesid ei ole ohtlikud ning vaid mõni üksik tüvi on patogeenne (Karch *et al*., 2005).

Kui isolaadi määramine on lihtne, siis paljudest bakteritest koosneva metagenoomse proovi iseloomustamine on keeruline mitmekesise koostise ja mikroobihulkade suure varieeruvuse tõttu. Peale selle ei pruugi paljud proovis leiduvad bakterid olla laboritingimustes kultiveeritavad (S. G. Tringe ja Rubin, 2005), seega on vaja iseloomustada kogu proovi korraga. Bakterite tuvastamise standardiks on kujunenud 16S geenipiirkonna amplifitseerimine ning sekveneerimine, kuid selle tulemused võivad olla PCR-i tõttu kallutatud ja ebapiisava resolutsiooniga eristamaks organisme tüve tasemel (B. B. Ward, 2002). Koosluse kohta annab rohkem infot kogu proovi sekveneerimine, mille puhul tehakse proovis leiduv DNA lühikesteks tükkideks ning määratakse nende kõigi järjestus, saades suurel hulgal lühikesi lugemeid. See teeb andmestiku suhteliselt keeruliseks ning väga mahukaks – üheks suurimaks väljakutseks tänapäeva bioinformaatikas ongi just andmehulga kiire kasv, samal ajal kui analüüsivõime areneb aeglasemalt (Hunter *et al*., 2012). Senised mikroorganismide klassifitseerimisprogrammid põhinevad joondamisel ning vajavad tihti ka assambleerimist, mille käigus liidetakse lühikesed lugemid pikemateks lõikudeks (kontiigideks) või terviklikuks genoomiks. Suuremahuliste teise põlvkonna sekveneerimisandmete puhul on aga nii joondamine kui ka assambleerimine väga aeganõudvad protsessid.

Antud töö käigus loodi bakterite tuvastamiseks tarkvara nimega StrainSeeker, mis lühikesi DNA oligomeere (k-meere) kasutades identifitseerib assableerimata sekveneerimisandmetest seal leiduvad bakteritüved.

1. KIRJANDUSE ÜLEVAADE

1.1 Mikroobide tuvastamine amplikonide järgi

Amplikon on DNA või RNA lõik, mis on kunstliku amplifikatsiooni või replikatsiooni produkt – näiteks PCR-i produktid. Amplikonide sekveneerimine on üks enamkasutatav meetod mikroobikoosluse kirjeldamiseks. Uuritava proovi (näiteks vee-, mulla- või koeproovi) kõikidest rakkudest eraldatakse DNA. Seejärel valitakse taksonoomiliselt informatiivne geneetiline marker, mis on ühine kõigil uuritavatel organismidel, ning amplifitseeritakse see PCR-ga. Saadud amplikonid sekveneeritakse ja järjestusi võrreldakse andmebaasides olevatega, et kindlaks määrata, millised organismid ning millises suhtelises hulgas proovis esinesid. Bakterite ja arhede puhul on markerjärjestuseks tavaliselt ribosoomi väikest subühikut kodeeriv ala (16S) (Hugenholtz ja Pace, 1996). Võib kasutada ka teisi markereid, kuid oluline on tähele panna, et need oleksid kõigis uuritavates mikroobides olemas ning amplifitseeritavad.

Erinevatel genoomi piirkondadel on erinev taksonite lahutusvõime (Liu *et al*., 2008; Schloss, 2010; D. V. Ward *et al*., 2012) ning seda tuleb analüüsil arvestada. 16S sobib kõrgemate taksonite (perekond, sugukond) eristamiseks, kuid tavaliselt ei sisalda piisavalt infot tüvede eristamiseks. On näidatud, et organismid, mis on 16S info põhjal identsed või klasterduvad kokku, võivad reaalselt olla väga erinevad ning isegi liikide määramiseks ei ole see meetod alati sobiv (B. B. Ward, 2002). Lisaks võib see lähenemine anda kallutatud tulemusi mikroobide hulga kohta proovis, sest 16S geeni koopiaarv võib varieeruda ka väga lähedastes liikides (Větrovský ja Baldrian, 2013). Suureks probleemiks amplikonide, kaasa arvatud 16S kasutamisel, on PCR-i etapp, mis võib tekitada tulemuste kallutatuse ning kimäärseid järjestusi (Acinas *et al*., 2005). Kuigi 16S piirkondade võrdlemist peetakse bakterite identifitseerimisel universaalseks, on mõned liigid niivõrd divergeerunud, et üldkasutatavad praimerid nende 16S-le ei seondu ja seetõttu jäävad need organismid proovist tuvastamata (Brown *et al*., 2015).

Vaatamata ülaltoodud puudustele kasutatakse 16S andmete võrdlemist endiselt väga laialdaselt prokarüootide eristamiseks, sest 16S on olemas igas prokarüoodis ning võimaldab ka fülogeneetilist analüüsi (Pace, 1997; Stackebrandt ja Goebel, 1994;

Woese *et al*., 1990). On näidatud, et kogu genoomi ja 16S analüüsi põhjal koostatud fülogeneesipuud on sarnased (Bansal ja Meyer, 2002). Lisaks on odavam sekveneerida kogu genoomi asemel vaid 16S piirkond.

16S sekveneeritakse kas terviklikult või keskendutakse selle varieeruvamatele aladele (S. G. Tringe ja Hugenholtz, 2008) ja seejärel klassifitseeritakse järjestused vastavalt nende sarnasusele andmebaasides olevatega (Wang *et al*., 2007). 16S informatsiooni saab analüüsida näiteks programmipaketiga QIIME (Caporaso *et al*., 2010) koos GreenGenes andmebaasiga (DeSantis *et al*., 2006). QIIME on mõeldud eelkõige amplikonide andmete uurimiseks, kuid seda on katsetatud ka terve genoomi andmete analüüsimiseks (Lindgreen *et al*., 2015).

1.2. Kogu proovi DNA sekveneerimine

1.2.1. Ülevaade

Kogu proovi DNA sekveneerimise korral ei valita amplifitseerimiseks mõnda kindlat geenilookust, vaid fragmenteeritakse ning sekveneeritakse kogu leiduv DNA. Saadud järjestused (lugemid) võivad olla nii taksonoomiliselt informatiivsed kui ka anda infot bioloogiliste funktsioonide kohta.

Isolaadi sekveneerimisel (proovis vaid üks tüvi) on kõik lugemid eeldatavasti pärit kindlast genoomist ja sellise proovi analüüs on võrdlemisi lihtne. Metagenoomse proovi korral pärinevad lugemid erinevate genoomide paljudest piirkondadest ning analüüsi teeb keerukaks lugemite seostamine kindlate organismidega. Proovi koosluse mitmekesisuse suurenemine vähendab omakorda iga eraldiseisva organismi katvust, sest piiratud arv lugemeid on jaotatud paljude organismide vahel. Nii võib tekkida olukord, kus mõni osa või kogu genoom jääb sekveneerimata ja seetõttu, olenemata klassifitseerimisalgoritmide võimekusest, pole võimalik organismi tuvastada. Sekveneerimissügavuse planeerimisel tuleb arvestada ka soovimatu peremeesorganismi DNA-ga, eriti kui tegu on mikrobioomi uuringuga. Kuigi sekveneerimise hind jätkuvalt alaneb, on ebaefektiivne ja kallis sekveneerida peremehe DNA-d, mis on analüüsiks ebavajalik. Samuti on peremeesorganismi genoomist saadud lugemitel potensiaalne oht mõjutada analüüsi tulemusi – näiteks juhul kui inimese lugem määratakse kõrge sarnasuse tõttu bakteri omaks. Soovimatu

DNA eraldamiseks on kasutatud näiteks hübridiseerimist (Chew ja Holmes, 2009) ja mikroobide eelnevat väljafiltreerimist (Garcia-Garcerà *et al*., 2013).

Selleks, et metagenoomse proovi uurimisel saada parimaid tulemusi, on soovituslik luua teadaoleva koostisega kunstlik proov. See peaks nii mitmekesisuse kui ka bakterite omavaheliste hulkade poolest olema võimalikult sarnane uuritava prooviga. Kunstliku proovi abil saab hinnata, kui palju kalduvad tulemused tegelikkusest kõrvale, kas sekveneerimissügavus on piisav ning kui suure osa proovist moodustab peremeesorganismi DNA. Arvestada tuleks ka sellega, et üksikute lugemite määramise tundlikkus kasvab lugemi pikenemisega ning suurim hüpe toimub vahemikus 100 kuni 250 ap (aluspaari). Seega on sekveneerimisel oluline saada lugemid, mis on vähemalt 250 ap pikkused. (Peabody *et al*., 2015)

Proovist organismide tuvastamiseks vaid sekveneerimisest ei piisa – saadud lühikesed lugemid tuleb viia vastavusse kindlate organismidega, kellelt need pärinevad. Selleks on loodud mitmesuguseid algoritme, mis suudavad sekveneerimisandmed läbi töötada ning anda vastuse, millised organismid proovis esinesid või referentsandmete puudumisel grupeerida lugemid omavahel.

1.2.2. Bakterite tuvastamine andmebaasi ja joondamise abil

Teise põlvkonna sekveneerimine annab suvalises järjekorras lugemeid proovis esinenud organismide DNA kohta. Iga üksik lugem on lühike ning esindab vaid väikest osa genoomist. Andmete koondamiseks kasutatakse lugemite lahterdamist (ingl. *binning*). Selleks võrreldakse lugemit andmebaasis, näiteks NCBI RefSeq-is (Tatusova *et al*., 2014), olevate referentsjärjestustega ning viiakse see lahtrisse, mille saadud vaste on ühine kõigile sellesse taksonisse kuuluvatele järjestustele (Patil *et al*., 2011). Näiteks kui lugem sobib igale *E. coli* tüvele andmebaasis, lahterdatakse see liigi tasemele (*E. coli*). Kirjeldatult töötab näiteks MEGAN (Huson *et al*., 2007), mis algselt kasutas päringute tegemiseks BLAST´i paketi programme (Altschul *et al*., 1990), kuid nüüd kasutab kiiremaid algoritme nagu DIAMOND (Buchfink *et al*., 2014). Üldiselt lahterdatakse üksikuid lugemeid, kuid kuna enamike algoritmide täpsus kasvab fragmendi pikenedes, siis püütakse kasutada ka assambleerimisel saadud kontiige, et saada paremaid tulemusi (Sharpton, 2014).

See, kui täpselt fragmendid suudetakse jaotada taksonoomilistesse lahtritesse, sõltub

mitmest faktorist. Esiteks on oluline tegur fragmendi pikkus. Lühemad ja vähemspetsiifilised fragmendid määratakse palju ebatäpsemalt kui 2000 ap või pikemad lõigud (Patil *et al*., 2011). Lahterdamise täpsust mõjutab ka proovi mitmekesisus. Keerukast kooslusest (näiteks mullast) pärit metagenoomse proovi sekveneerimine annab igale organismile küllaltki madala katvuse, sest lugemite koguarv on jaotatud väga paljude organismide vahel. Seetõttu on sellise proovi assambleerimisel võimalik saada vaid lühikesi kontiige. Peale selle suureneb paljude lahtrite (taksonite) korral tõenäosus, et mõni fragment määratakse valesti (Dröge ja Mchardy, 2012). Suurte andmehulkade puhul võivad paljud valesti määratud fragmendid kokkuvõttes anda valepositiivseid lõpptulemusi.

1.2.3. Bakterite tuvastamine markerjärjestuste abil

Paljud genoomsed piirkonnad on kas väga laialdaselt levinud (eriti lühemad osad lugemist), andmebaasiga võrreldes tundmatud või sisaldavad sama järjestust, mis andmebaasis. Sellest lähtuvalt on mõnikord otstarbekas hoida andmebaasis vaid taksonoomiliselt informatiivseid järjestusi, milleks võivad olla konkreetsed markergeenid või muud genoomsed markerid. Kuna markerite andmebaas on küllaltki väike, on otsinguprotsess kiire. (Peabody *et al*., 2015)

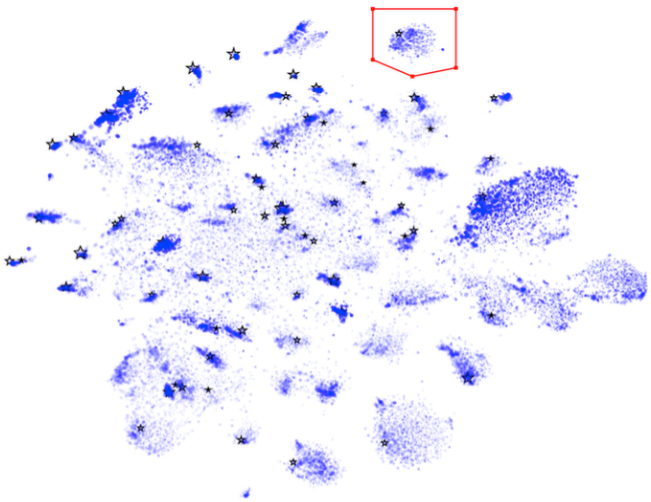
Organismi tuvastamiseks markergeeni põhjal võrreldakse proovist leitud geenijärjestusi andmebaasis olevatega. Terve genoomi sekveneerimisel saab analoogselt amplikonide sekveneerimisele võrrelda näiteks 16S järjestusi. Markergeenide valikul tuleb arvestada seda, et valitud geen peab olemas olema kõigil uuritavatel organismidel – mida vähem levinud on geen, seda kitsamat organismide gruppi saab selle põhjal analüüsida. Potensiaalseid markergeene on identifitseerinud mitmed uurimisgrupid (Darling *et al*., 2014; Segata *et al*., 2012; Wu *et al*., 2013).

MetaPhlAn (Segata *et al*., 2012) kasutab taksonoomiliseks määramiseks umbes üht miljonit markerit, mis pärinevad kodeerivatelt aladelt ja on lühemad kui terviklikud geenid. Markerite kogum koosneb erinevatele klaadidele (näiteks liik või sugukond) spetsiifilistest järjestustest. Markerjärjestused peavad olema klaadisiseselt väga konserveerunud, kuid väljaspool klaadi puuduma. Taksonoomilise määramise jaoks joondatakse kõik lugemid erinevate markerite vastu Bowtie2 programmiga (Langmead ja Salzberg, 2012).

Markeripõhise lähenemise korral määratakse vaid markerit sisaldavate lugemite päritolu, ülejäänud jäävad kõrvale. Selle tõttu väheneb küll tundlikkus, kuid täpsus on kõrge. Kuna uuritakse konkreetseid markergeene ja mitte lugemeid, saab täpsemalt hinnata ka organismide hulki proovis. Lugemite arv võib anda kogusest vale ettekujutuse, sest suurtest genoomidest tuleb rohkem lugemeid. (Peabody *et al*., 2015)

1.2.4. Lugemite klasterdamine omavahelise sarnasuse alusel

Järjestuste andmebaasi abil on proovist võimalik leida vaid neid organisme, mis on andmebaasis esindatud või neile väga sarnased. Kui sekveneeritud kooslus on referentsandmetest väga kauge, võib suurem osa organismidest tuvastamata jääda. Sellises olukorras on kasulikum grupeerida lugemid omavahelise sarnasuse alusel, mitte teadaolevate järjestuste põhjal. Eeldades, et genoomne koostis on küllaltki ühtlane üle terve genoomi, on ka ühest genoomist pärit lugemid omavahel sarnasemad. Fragmentide klasterdamine koostise järgi võib toimuda näiteks GC sisalduse, koodonkasutuse või lühikeste, 4-6 ap pikkuste oligomeeride järgi. On leitud, et sellised omadused varieeruvad eri genoomide vahel ning on seeläbi andnud aluse terminile genoomi signatuur (Deschavanne *et al*., 1999; Karlin ja Burge, 1995). Sellised signatuurid esinevad ka kõrgematel klaadidel, mistõttu saab neid kasutades määrata lugemeid erinevatel tasemetel (McHardy *et al*., 2007).



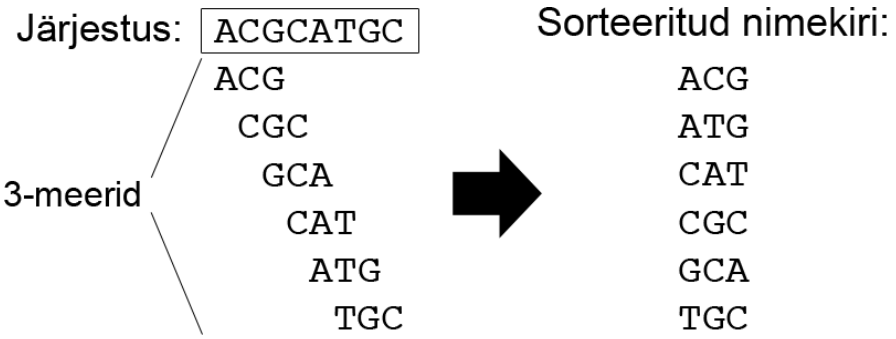
Joonis 1. Programmi VizBin visualiseeritud metagenoomne andmestik. Programm klasterdab järjestused koostise järgi, võttes arvesse lühikeste kindla pikkusega oligomeeride sagedusi. Klastrid kujutatakse visuaalselt ning kasutajad saavad ise märkida ära huvipakkuvad klastrid ning eraldada valitud fragmendid edasiseks uuringuks. (Laczny *et al*., 2015)

Koostise põhjal klasterdamise meetodid vajavad sageli assambleerimist, kuna määratav fragment peab olema piisavalt pikk (>1000 ap), et selle genoomne signatuur piisavalt selge oleks. Lühemate fragmentide korral hakkavad klastrid omavahel liialt kattuma (Laczny *et al*., 2015, joonis 1) ning joondusel põhinevad meetodid annavad tavaliselt paremaid tulemusi (Brady ja Salzberg, 2009).

1.3. K-meer - kindla pikkusega oligomeer

1.3.1 Ülevaade

K-meerid on kõikvõimalikud ühe kindla pikkusega (*k*) oligomeerid, mis saadakse järjestusest ühese sammuga liikudes (joonis 2). Kaks järjestikku asuvat k-meeri erinevad teineteisest kahe nukleotiidi võrra (ühe k-meeri algusest ja teise k-meeri lõpust ühe nukleotiidi võrra). Koostatud k-meeride nimekirjast toimub otsing tavaliselt täpsete vastete (ingl. *exact match*) põhimõttel, mis teeb selle kiiremaks kui joondamine, sest ei toimu k-meeri pikendamist (Altschul *et al*., 1990; Kaplinski *et al*., 2015).



JOONIS 2. Järjestus ja selle koostis k-meeridena. Järjestus tehakse k-meerideks (*k=3*) ning seejärel sorteeritakse tähestikulises järjekorras otsingu kiirendamiseks.

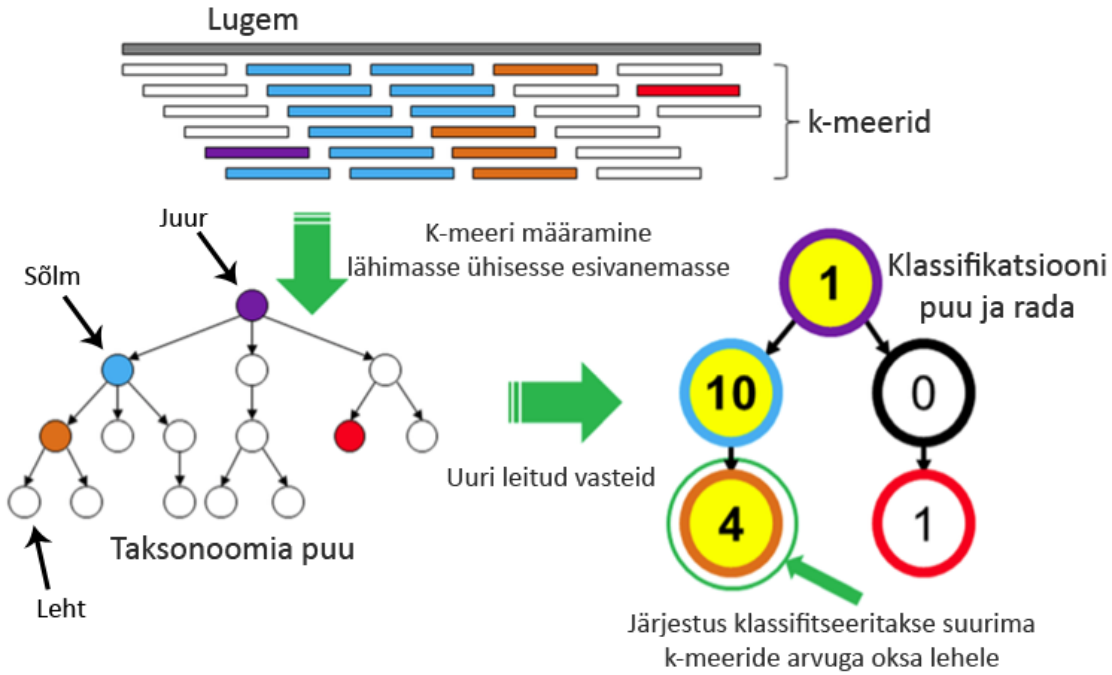
Kui k-meere oli varasemalt rakendatud näiteks joondamisel ning assambleerimisel seemnejärjestustena (ingl. *seed*), siis on üha enam proovitud geneetilist infot (terveid genoome, kontiige, lugemeid) kujutada k-meeride kogumitena (Wood ja Salzberg, 2014). Sellega kaob küll positsiooniline informatsioon, sest üldjuhul ei salvestata k-meeri asukohta järjestuses, kuid kui uuritakse vaid mingi järjestuse olemasolu või hulka, pole see ka vajalik. Sellise lähenemise eeliseks on kiirus, sest k-meeridega töötades kaob andmete analüüsil vajadus nii assambleerimise kui ka joondamise järgi

ning need on võimalik asendada arvutuslikult kiire ja lihtsa k-meeride lugemisega. Heaks näiteks k-meeride poolt pakutavatest lihtsustustest on kogu genoomi k-meeride põhjal fülogeneetilist distantsi arvutavad programmid (Fan *et al*., 2015; Ondov *et al*., 2015). Traditsioonilised joondustel põhinevad meetodid on väga arvutusmahukad ning vajavad assambleerimist, kuid sarnane tulemus on võimalik saavutada erinevate genoomide k-meeride hulkade ühisosasid võrreldes.

1.3.2 Organismide tuvastamine k-meeridega

Varasemalt on lühikesi, 4-6 ap pikkusi k-meere kasutatud järjestuste klasterdamiseks (vt. 1.2.4. Lugemite klasterdamine omavahelise sarnasuse alusel), kuid üha enam on võetud kasutusele ka pikemad k-meerid (15-32), mida võib käsitleda lühikeste markeritena. Otsinguprotsessis ei joondata lugemeid andmebaasis olevate markerite referentsjärjestustega, vaid tehakse nii lugemid kui ka andmebaasis olevad markerid k-meerideks ning võrreldakse neid omavahel, otsides täpseid vasteid, mis on arvutuslikult oluliselt kiirem. Organismide tuvastamiseks luuakse eelnevalt spetsiifiliste k-meeride andmebaas, asendades genoomid nende k-meeride nimekirjadega. Spetsiifilised k-meerid leiduvad vaid ühes või kindlas grupis genoomides. Lahterdamise korral võrreldakse lugemi k-meere andmebaasis olevatega ning olenevalt algoritmi eripärast määratakse lugem taksonoomilisse lahtrisse.

Kraken (Wood *et al*., 2014) on k-meeripõhine klassifitseerimise programm, mille andmebaasi taksonoomilised lahtrid põhinevad NCBI taksonoomial. K-meeride nimekirjad on igal taksonoomilise puu lehel ja sõlmel (juur, sõlm ja leht näidatud joonisel 3; ka leht ja juur on sõlmed). Esialgu tehakse k-meeride nimekirjad igale lehele ning seejärel liigutatakse iga k-meer, mis on kahel lehel ühine, nende lähimasse ühisesse sõlme. Seega moodustavad väga kaugete organismide ühised k-meerid puu juurelähedased sõlmed. Tulemuseks on andmebaas, kus igale k-meerile vastab üks kindel sõlm. Lugemi klassifitseerimisel (joonis 3) tehakse see k-meerideks, nende hulgast otsitakse välja spetsiifilised, mis seejärel puule paigutatakse. Nii tekib „klassifikatsiooni puu“, kus lugem määratakse kõige tipmisele sõlmele, mille k-meere lugemis leidus. Oksa hargnemise korral valitakse suurima kaaluga (leitud k-meeride arvuga) oks.



Joonis 3. Krakeni klassifitseerimisalgoritm. Järjestuse klassifitseerimiseks paigutatakse iga järjestuses olev k-meer madalaimasse ühise eelasega sõlme, mis andmebaasis leidub. Seejärel leitakse suurima kaaluga oks (lilla, sinine, oranž) ning järjestus klassifitseeritakse kõige tipmisele sõlmele. (Wood *et al*., 2014, kohandatud)

Krakenile sarnaselt töötab CoMeta (Kawulok ja Deorowicz, 2015), mis aga k-meeride arvu lugemise asemel hindab, mitu lugemi nukleotiidi spetsiifilised k-meerid katsid. CoMeta kasutab samuti NCBI taksonoomiat. Veidi lihtsam on CLARK (Ounit *et al*., 2015), mis ei kasuta puukujulist taksonoomiat – korduvad k-meerid lihtsalt eemaldatakse andmebaasist selle loomise käigus. Lugem määratakse sellesse lahtrisse, mille k-meere see kõige rohkem sisaldas. Tavaliselt on andmebaasis palju väga sarnaseid või identseid tüvesid (näiteks *E. coli* K-12), mille enamik k-meeridest on ühised, seega on neil väga vähe spetsiifilisi k-meere. Uute tüvede lisamisega andmebaasi väheneb spetsiifiliste k-meeride arv veelgi, kuni mõne tüve määramine pole ühel hetkel enam võimalik. Seetõttu on mõttekam andmebaasi ehitamisel kasutada puu struktuuri, kus mitme tüve ühiseid k-meere ei kaotata ära, vaid viiakse nende ühisesse sõlme.

TABEL 1. Erinevate identifitseerimisprogrammide võrdlus. (Lindgreen *et al*., 2015, osaline)

Programm	**Määratud lugemeid**	**Valepositiivseid**	**Tööaeg (min)**
Kraken	71.98%	0.00%	60.95
CLARK	73.32%	0.02%	211.50
QIIME	58.23%	0.28%	8.88
MEGAN	42.21%	0.49%	2489.65
MetaPhlAn	5.09%	0.75%	108.51

K-meeripõhistel määrajatel on teiste meetoditega võrreldes olulisi tugevusi nii kiiruse kui ka täpsuse osas. Hiljutises võrdluses näidati, et traditsioonilised joondamisel põhinevad meetodid võivad üht andmestikku analüüsida rohkem kui päeva, k-meeripõhised programmid nagu Kraken ja CLARK seevastu töötasid sama andmestiku kallal vaid mõned tunnid. Samuti andsid Kraken ja CLARK vähe valepositiivseid, määrates samal ajal ära rohkem lugemeid kui konkureerivad programmid (Lindgreen *et al*., 2015; tabel 1).

2. PRAKTILINE OSA

1. Töö eesmärgid

Käesoleva töö eesmärgiks oli luua k-meeripõhine bakterite tuvastamise tarkvara, mis võimaldab baktereid leida assambleerimata sekveneerimislugemitest. Programmi andmebaasi loomiseks peab saama kasutada suvalisi assambleeritud bakterigenoome ning nende järgi tehtud juhtpuud, toetumata seejuures ühelegi olemasolevale taksonoomiale.

2. Metoodika

2.1. Programmi testimiseks kasutatud andmestikud

Andmebaas koostati RefSeq 69 versiooni järgi (Tatusova *et al*., 2014) ning see koosnes kokku 4324 genoomist (bakterid ja arhed). "Must nimekiri" koosneb k-meeridest, mida ei soovita andmebaasi ning mis võivad määramisel probleeme tekitada. "Must nimekiri" koosnes inimese genoomi (GRCh38) ja bakterite plasmiidide (RefSeq versioon 65 bakterite plasmiidid) k-meeridest (*k=32*).

Programmi testimiseks kasutati kunstlikku proovi, mis koosnes 6,6 miljonist 75 ap Illumina lugemist (NCBI SRA andmebaas, proov SRR172902). Teine kunstlik andmestik pärines Peabody *et al*. (2015) tööst ("FW *in vitro*"), milles oli ligikaudu 150 000 lugemit (Illumina) keskmise pikkusega 223 ap. Suurim kasutatud andmestik oli saadud simuleeritud andmetega ning pärines Lindgreen *et al*. (2015) tööst, koosnedes umbes 58 miljonist 100 ap lugemist.

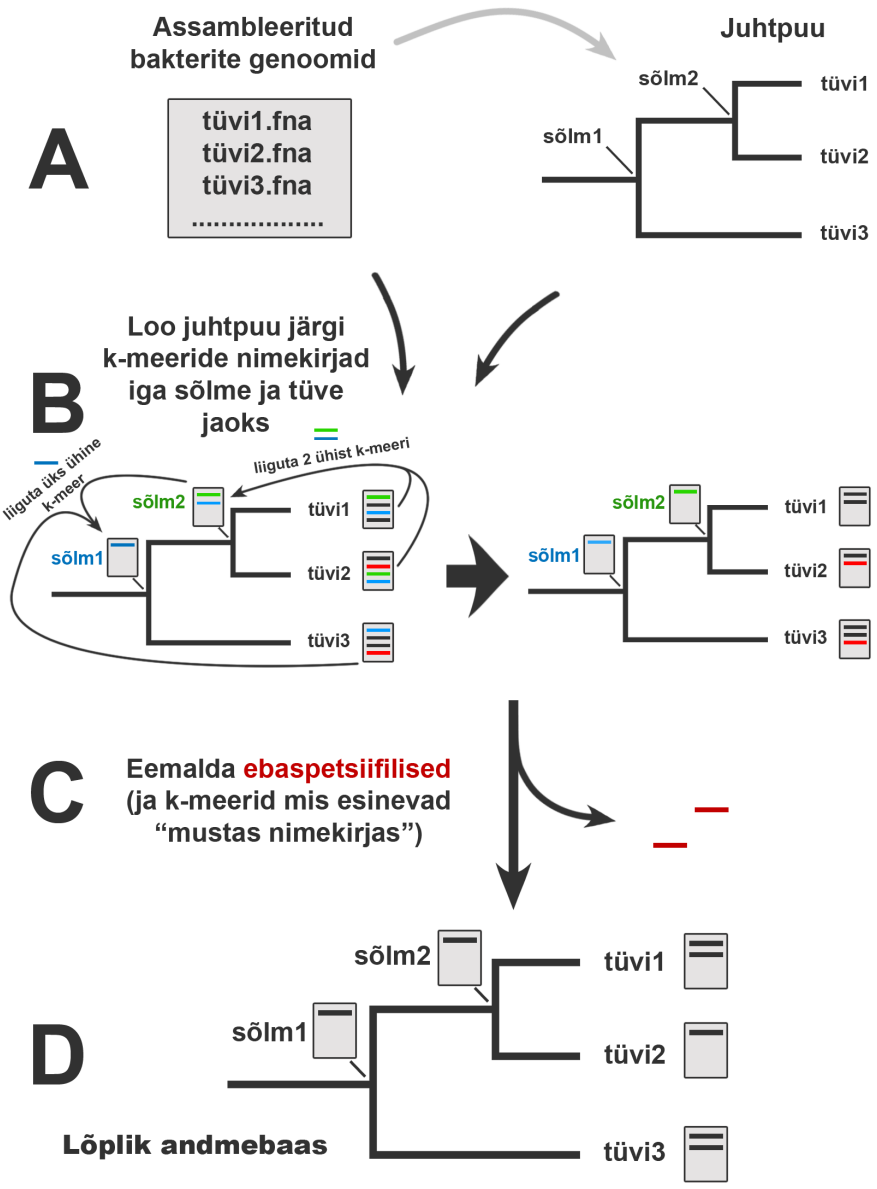
2.2. Juhtpuu tegemine

Juhtpuu jaoks arvutati distantsmaatriks programmiga MASH (Ondov *et al*., 2015), kasutades parameetrit -s 10000. Puu konstrueeriti programmiga MEGA6 (Tamura *et al*., 2013) UPGMA meetodil vaikimisi parameetreid kasutades.

2.3. Andmebaasi loomine

Andmebaas luuakse kasutaja poolt etteantud genoomide ning juhtpuu põhjal (joonis 4A), mis kirjeldab antud genoomide omavahelisi seoseid. Mida lähedasemad erinevad tüved puul on, seda enam sisaldavad nad ühiseid k-meere. Esiteks koostatakse k-meeride (*k=32*) nimekiri igale puu lehele antud järjestuste alusel. Nii nimekirjade loomiseks kui ka nendega manipuleerimiseks kasutati programmipaketti GenomeTester4 (Kaplinski *et al*., 2015). Kui kahes alamsõlmes/-lehes on sama k-meer, siis liigutatakse see vanemsõlme. Sõlmede täitmine k-meeridega toimub järk-järgult, liikudes lehtedelt juure suunas (joonis 4B). Tulemuseks on k-meeride nimekirjad iga lehe ja sõlme kohta. Selles etapis on veel võimalik kõik ühe genoomi k-meerid kokku koguda, kui liita omavahel kõikide k-meeride nimekirjad, mis jäävad juurest leheni (joonis 4B paremal).

Selleks, et nimekirjades olevad k-meerid oleksid spetsiifilised vaid ühele sõlmele, eemaldatakse kõik k-meerid, mis asuvad selles etapis endiselt mitmes sõlmes (joonis 4C). Lisaks eemaldadatakse andmebaasist ka k-meerid, mis asuvad ebasoovitavate k-meeride „mustas nimekirjas“. Selles töös olid nendeks inimese genoomis ja bakterite plasmiidides leiduvad k-meerid. Tulemuseks on k-meeride nimekirjad igale lehele ja sõlmele, mis võivad olla väga erineva suurusega (joonis 4D). Andmebaasi suuruse vähendamiseks eemaldati suurtest nimekirjadest juhuslikult k-meere, kuni igas sõlmes oli maksimaalselt 100 000 k-meeri. Lõpliku andmebaasi suurus oli umbes 18 GB.



JOONIS 4. Andmebaasi koostamine. (A) Eeltööna kogutakse kokku andmebaasi lisatavad täisgenoomid ja konstrueeritakse nende põhjal juhtpuu. (B) Andmebaasi tegemine algab genoomide k-meerideks konverteerimisega. Seejärel liigutatakse ühised k-meerid juurepoolsemate sõlmede nimekirjadesse: tüvedest 1 ja 2 liigutatakse roheline ja sinine k-meer sõlme 2 ning seejärel viiakse sõlmest 2 ja tüvest 3 sinine k-meer sõlme 1. Punane k-meer jäetakse liigutamata, sest tüve 1 genoom seda ei sisalda. (C) Punane k-meer on olemas nii tüves 2 kui ka tüves 3, seepärast on see ebaspetsiifiline ning eemaldatakse andmebaasist. Lisaks eemaldatakse kõik k-meerid, mis on "mustas nimekirjas" (näiteks inimese genoomis leiduvad k-meerid). (D) Lõplik andmebaas koosneb sõlme- ja tüvespetsiifilistest k-meeride nimekirjadest, kus iga k-meer esineb vaid ühes nimekirjas. (Roosaare *et al.*, 2016, kohandatud)

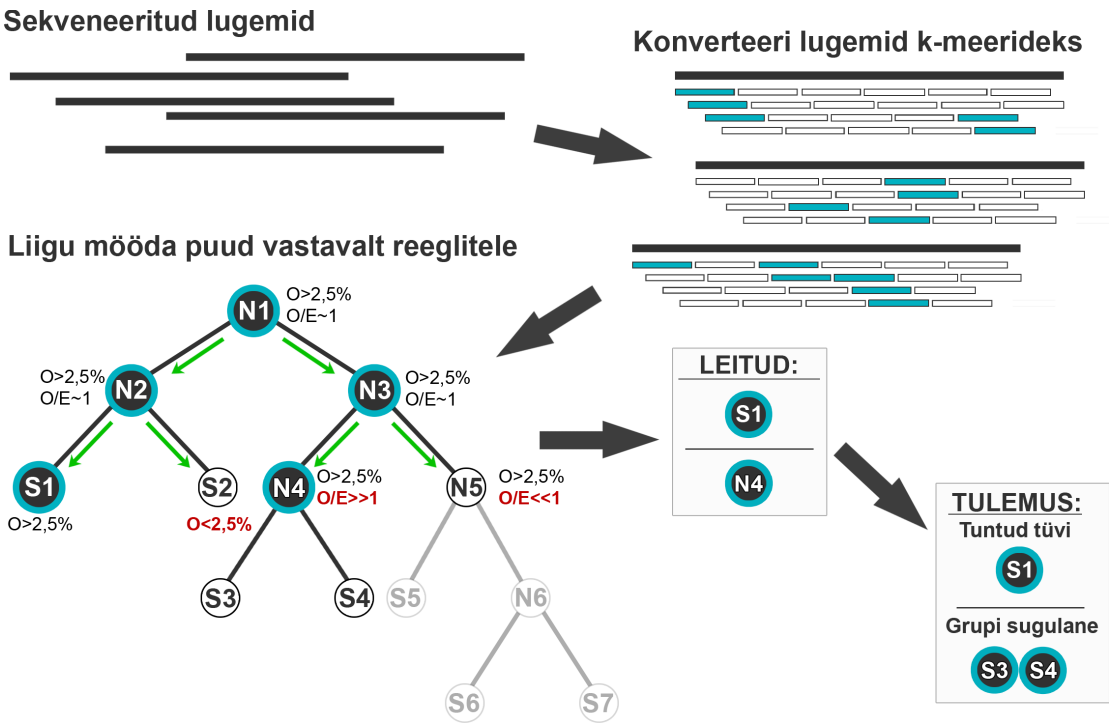
2.4. Otsinguprotsess

Otsinguprotsessi eeltööks on kõikide proovi lugemite konverteerimine üheks suureks k-meeride nimekirjaks, seega ei määrata iga üksikut lugemit eraldi, vaid otsused tehakse kogu proovi k-meeride andmestiku põhjal. Seejärel liigutakse rekursiivselt mööda puu sõlmi vastavalt etteantud reeglitele ja loetakse kokku, kui palju antud nimekirja k-meere esines proovis. Kasutati kahte põhilist reeglit.

Esiteks leitakse, kui suur on proovist leitud sõlme k-meeride N_{obs} osakaal O kõigi sõlme k-meeride N_{tot} suhtes ($O = N_{obs}/N_{tot}$; O – leitud osakaal (*observed*); N_{obs} – käesolevast sõlmest leitud unikaalsete k-meeride arv (*node observed*); N_{tot} – käesoleva sõlme kõigi unikaalsete k-meeride arv (*node total*)). Selleks, et minna teise reegli juurde ning jätkata otsinguga, peab O olema piisav (käesolevas töös $O > 2{,}5\%$). Liiga väheste k-meeride leidmise korral jääb otsing selles sõlmes seisma (joonis 5, sõlm N5).

Teiseks – kui proovis on piisavalt sõlme k-meere (1. reegel), leitakse osakaalu O ja eeldatava osakaalu E suhe (O/E), kus $E = C1_{obs}/C1_{tot} + C2_{obs}/C2_{tot} - C1_{obs}/C1_{tot} * C2_{obs}/C2_{tot}$, kus $C1$ ja $C2$ on alamsõlmed (*child), obs* proovist leitud vastava sõlme k-meeride arv ja *tot* sõlme kogu k-meeride arv. Kui $O/E \sim 1$, siis on eeldatav k-meeride hulk sarnane leituga ning liigutakse edasi lehtede poole või väljastatakse leheni jõudes tulemuseks proovis leiduv tüvi (joonis 5, leht S1). $O/E >> 1$ korral nähti sõlmes palju rohkem k-meere kui võiks eeldada alamsõlmede põhjal (joonis 5, sõlm N4). See viitab andmebaasist puuduvale organismile, millel puuduvad lahknemise tõttu osad alamsõlmede k-meerid (joonis 5, sõlm N4). Sellises olukorras antakse tulemuseks nimekiri tüvedest, mis jäävad antud sõlme alla ning on võõra tüve sugulased (joonis 5, tulemuses grupi S3, S4 sugulane). $O/E << 1$ on teoreetiline juhus, kus alamsõlmedest leiti rohkem k-meere kui käesolevast sõlmest ning võib tekkida rohkete sekveneerimisvigade tõttu. O/E väärtuse ja katvuse hindamiseks kasutati statistilist testi (Roosaare *et al*., 2016; lisa statistilise testi kohta[1]).

[1] http://bioinfo.ut.ee/strainseeker/publication/Additional_data_file_1.pdf



JOONIS 5. Otsingualgoritm. Kõik sekveneeritud lugemid konverteeritakse üheks k-meeride nimekirjaks. Otsing algab puu juure k-meeride nimekirjast (N1). Iga uue sõlme juurde liikudes tehakse kindlaks, kas proovis on piisavalt selle nimekirja k-meere ($O > 2{,}5\%$), vastasel korral peatub otsing (S2). Leheni jõudes märgitakse vastav tüvi tulemustes (S1). Vaadeldud ja eeldatud k-meeride suhte (O/E) järgi otsustatakse, kas otsing jätkub (rohelised nooled) või peatub. $O/E << 1$ tulemust ei anta ning $O/E >> 1$ viitab andmebaasist puuduvale organismile (kuulub N4 alla, on S3 ja S4 lähisugulane). (Roosaare *et al.*, 2016, kohandatud)

2.5. Implementeerimine

StrainSeeker on kirjutatud programmeerimiskeeles PERL. K-meeride nimekirjade tegemisel ja manipuleerimisel kasutati GenomeTester4 paketti. *O/E* hindamiseks ja katvuse leidmiseks kasutati programmeerimiskeeles R koostatud skripte (Roosaare *et al.*, 2016).

3. Tulemused

Väljatöötatud metoodika põhjal loodi programm nimega StrainSeeker, mis koosneb kahest osast: andmebaasi ehitamise tööriist ning sekveneerimislugemitest andmebaasis leiduvaid järjestusi otsiv programm. Andmebaasi loomise sisendfailideks on assambleeritud bakterite ja arhede genoomid (kokku 4324) ja nende põhjal konstrueeritud juhtpuu. K-meeri pikkuseks valiti 32. Otsinguprotsessi sisendandmeteks on lugemid ning andmebaas, mida otsingul kasutatakse. Väljundina antakse kasutajale tabuleeritud kujul fail, kus on kirjas kõik määratud organismid koos nende osakaaludega proovis. Lisaks on märgitud ka see, kas leitu oli andmebaasis olemas või esines vaid selle sugulane – sel puhul on näidatud ka lähimad organismid leitud organismile (vt. lisa 1 ja lisa 2).

Antud töös testiti StrainSeekerit kahe kunstliku ja ühe simuleeritud andmestikuga. Iga andmestiku analüüsi korral märgiti üles nii StrainSeekeri kui ka Krakeni (Wood *et al*., 2014) kasutatud aeg. Väiksemate andmestikega on kiirem Kraken, kuid mitte märkimisväärselt. Suurema andmestiku korral on StrainSeeker selgelt kiirem. Ilmselt tulevad erinevused algoritmide ülesehitusest ja eripäradest. Peale selle on StrainSeekeri andmebaas suurem nii mahult (18 GB) kui ka genoomide arvult (4324) võrreldes MiniKrakeni andmebaasiga (4 GB ja 2256 genoomi).

3.1. StrainSeekeri võrdlemine teiste programmidega

Avaldatud on mitmed töid, mis võrdlevad erinevate klassifitseerimisalgoritmide võimekust (Lindgreen *et al*., 2015; Peabody *et al*., 2015). Tavaliselt võrreldakse kogu määratud lugemite osakaalu ning vaadatakse, kui suur hulk neist määrati õigesti. StrainSeeker üksikuid lugemeid ei klassifitseeri, seega ei saa võrdlusel sellist lähenemist kasutada, vaid analüüsitakse kas ja kui täpselt organism proovist tuvastati. StrainSeekeri tulemusi võrreldi eelkõige Krakeniga (Wood *et al*., 2014), kuna see on laialt kasutust leidnud nii metagenoomsetes uuringutes (Hiraoka *et al*., 2016), analüüsivate töövoogude osana (Kim *et al*., 2016) kui ka peamise otsingualgoritmina andmebaasis (Forster *et al*., 2015). Kiiruse võrdlemiseks analüüsiti andmestikke samas serveris vaikimisi parameetritega, kuid tulemuste tõlgendamisel toetuti avaldatud infole, kus see oli saadaval (Peabody *et al*. (2015) ning Lindgreen *et al*. (2015) tööd).

3.1.1. StrainSeekeri võrdlus Krakeniga kasutades sekveneerimisandmeid

StrainSeekeriga analüüsiti SRA andmebaasist pärinevat proovi SRR172902, mida kasutati Inimese Metagenoomi Projektis (The NIH HMP Working Group, 2009) sekveneerimisprotokollide kontrollimiseks ja täpsuse hindamiseks eri keskuste vahel. Tegemist oli kunstliku prooviga, kus teadaolevate bakteritüvede DNA segati destilleeritud vees ning seejärel sekveneeriti. Proov sisaldas 22 organismi, millest StrainSeeker tuvastas 20, enamiku neist tüve (lisa 1). Leidmata jäi *Candida albicans*, mis on pärm ja seega puudus bakterite-arhede andmebaasist. Lisaks jäi tuvastamata *Actinomyces odontolyticus* – andmebaasis oli küll kaks selle perekonna esindajat, kuid need moodustasid mõlemad omaette alampuu. *E. coli* puhul väljastati tulemustes kõik tüve K-12 esindajad (proovis esines K-12 alamtüvi MG1655, mis oli ka andmebaasis olemas). *Listeria monocytogenes* tuvastati liigi tasemel, kuid määrati vale tüvi. *Staphylococcus aureus´*e puhul anti tulemused ühe tüve asemel mitme eri grupina ning polnud võimalik öelda, milline tüvi proovis esines. StrainSeeker analüüsis andmestikku 3,75 minutit ning Kraken (MiniKraken andmebaasiga) 3 minutit. Krakeni tulemustes esinesid kõik õiged organismid, enamik neist suure määratud lugemite arvuga. Samas oli väga paljudele organismidele, mida proovis ei esinenud, määratud üksikud lugemid, mille tõttu ei saa tulemusi kasutada ilma eelneva töötluseta. Tõlgendamisel on kohati probleemiks valesti määratud lugemite kõrge arv (näiteks *Serratia marcescens* WW4 – 492 lugemit) ning õigesti määratud lugemite madal arv (näiteks *Streptococcus agalactiae* 2603V/R – 343 lugemit). Kui kasutada kindlat piirmäära tulemuste töötlemisel, kaovad koos valepositiivsetega ka mõned õiged tulemused.

Teine kunstlik andmestik pärineb Peabody *et al*. (2015) tööst, kus võrreldakse omavahel mitmeid erinevaid programme nii kunstlike proovide kui ka simuleeritud andmetega töötamisel. Käesolevas töös kasutati "FW *in vitro*" andmestikku, mis on saadud kunstlikku proovi sekveneerides (sarnaselt SRR172902 andmestikule). Proovi loomiseks kasvatati 11 bakteri puhaskultuurid, eraldati neist DNA, viidi see võrdsetes kogustes destilleeritud vette ning saadud segu sekveneeriti. StrainSeeker suutis tuvastada kõik 11 liiki (lisa 2), andes liigi tasemel ühe valepositiivse tulemuse (*Burkholderia* grupp), millele anti ka madalaim osakaal proovis (0,36%). Tüve tasemel jagunes *E. coli* kolmeks ning õiget alamtüve ei leitud. *Pseudomonas putida* kohta anti välja kaks tüve, millest õige oli väiksema osakaaluga kui vale. StrainSeeker

analüüsis andmestikku umbes 3 minutit ning Kraken (MiniKraken andmebaasiga) 50 sekundit. Kraken leidis samuti kõik 11 liiki (Peabody *et al*., 2015), kuid ilma piirmäärata leiti lisaks veel 327 liiki, mis raskendab valepositiivsete seast õigete tulemuste leidmist sarnaselt eelnevalt kasutatud SRR172902 andmestikule.

3.1.2. StrainSeekeri võrdlus Krakeniga kasutades simuleeritud andmeid

Lindgreen *et al*. (2015) koostasid simuleeritud andmestiku, mis sisaldas 417 erinevat perekonda (koos bakteritega ka seitse eukarüoodi perekonda). Andmestik pidi olema keerukuselt ja suuruselt võimalikult sarnane reaalsetele andmetele, kuid erinevalt reaalsete proovide kasutamisest võimaldab simuleerimine täpselt öelda, millisest organismist mingi lugem pärines. Kõikidest proovis olevatest lugemitest moodustasid 70% bakterite ja arhede genoomid, 5% *in silico* muudetud genoomid, 5% eukarüootide genoomid ja 20% juhuslikult segatud järjestusega genoomid, mis ei tohiks tulemust anda. *In silico* muutmine imiteeris evolutsioneerumist ning andis tulemuseks andmebaasis olevast genoomist mingil määral erineva järjestuse. Kuna kogu võrdlus põhineb lugemite jaotamisel, siis StrainSeekeriga saab võrrelda vaid lõpptulemusi. Analüüsimiseks valiti "setA" kolm andmestikku, mis andsid sarnased tulemused (perekond *Cupriavidus* esines antud kahe andmestiku tulemustes grupi osana, ühel iseseisvana – loeti õigesti määratuks). Tulemusi hinnati perekonna tasemel. StrainSeekeril kulus analüüsiks 14 minutit ja Krakenil (MiniKraken andmebaasiga) samas serveris 36 minutit. StrainSeekeri tundlikkus antud andmestikuga oli 0,95 ning täpsus 1 (tabel 2). Krakeni tundlikkus oli 0,8998 ning täpsus ~1 (Lindgreen *et al*., 2015), kuid arvestada tuleb kindlasti sellega, et arvutamiselt kasutati lugemite, mitte tulemuseks saadud perekondade arvu.

TABEL 2. StrainSeekeri analüüsi tulemused perekonna tasemel. Simuleeritud andmestik pärineb Lindgreen *et al*. (2015) tööst.

<table>
<tr><td>Tõeseid positiivseid</td><td>395</td><td colspan="3"></td></tr>
<tr><td rowspan="4">Valenegatiivseid</td><td rowspan="4">22</td><td rowspan="2">Esineb andmebaasis</td><td rowspan="2">4</td><td>Esineb teine liik</td><td>3</td></tr>
<tr><td>Esineb sama tüvi</td><td>1</td></tr>
<tr><td rowspan="2">Puudub andmebaasist</td><td rowspan="2">18</td><td>Eukarüoote</td><td>8</td></tr>
<tr><td>Baktereid/arhesid</td><td>10</td></tr>
<tr><td>Valepositiivseid</td><td>0</td><td colspan="3"></td></tr>
<tr><td>Kokku</td><td>417</td><td colspan="3"></td></tr>
</table>

3.2. Kättesaadavus ja veebiversioon

StrainSeeker on tasuta allalaetav aadressilt www.bioinfo.ut.ee/strainseeker. Samal aadressil on võimalik kasutada ka programmi veebiversiooni, millega saab analüüsida väiksemaid andmestikke. Tulemused kuvatakse nii tabeli kui ka sektordiagrammi kujul.

ARUTELU

Teise põlvkonna sekveneerimine toodab suurtes kogustes toorandmeid, mida on vaja analüüsida. Metagenoomsetest uuringust saadud sekveneerimisandmete suur hulk on BLAST´i-põhiste analüüsimeetodite puhul pudelikaelaks, mille lahendamine arvutusliku võimsuse tõstmisega ei ole pikas perspektiivis mõistlik (Angiuoli *et al*., 2011). Lootustandvad on k-meeridel põhinevad organismide klassifitseerimismeetodid, mis suudavad kiirelt analüüsida suurt hulka andmeid ja on väga täpsed (Lindgreen *et al*., 2015). K-meeride kasutamine on kiire eelkõige tänu arvutuslikult efektiivsele täpsete vastete otsimisele.

Andmehulga kiire kasv tekitab ka vajaduse spetsiaalselt kohandatud ja kiirelt uuenevate andmebaaside järele, sest enamik programme sõltub mingil hetkel saadaval olnud andmetest (Kraken – RefSeq andmebaas, MetaPhlAn – Inimese Metagenoomi Projekti andmed) ning ei võimalda kasutajal teha suuri muudatusi nendega kaasas olevates andmebaasides.

Nimetatud tähelepanekuid silmas pidades loodi assambleerimata teise põlvkonna sekveneerimisandmetest bakterite genoomseid järjestusi tuvastav tarkvara, mis nimetati StrainSeekeriks. Sarnaselt teistele programmidele kasutab ka StrainSeeker juhtpuud (eelised välja toodud punktis 1.3.2). Krakeni ja CoMeta juhtpuud põhinevad NCBI taksonoomial, mis piirab andmebaasis olevaid tüvesid. Peale selle on NCBI taksonoomias mõningased vastuolud – näiteks on perekond *Shigella* eraldi, kuigi peaks kuuluma liigi *E. coli* alla (Lan ja Reeves, 2002). Samuti on liigist madalamal, tüvede tasemel, organismide omavahelised suhted halvasti kirjeldatud. Selleks, et andmebaasi saaks lisada ükskõik millise tüve genoome, tuleks juhtpuu ehitada andmebaasi loomisel sisendina antud genoome kasutades. StrainSeekeri puhul ongi see üks eeltöö etappidest (vt 2.2 Juhtpuu tegemine). Lindgreen *et al*. andmestiku (417 perekonda) analüüsil andis StrainSeeker esmapilgul mitmeid valepositiivseid tulemusi, mis aga lähemal uurimisel osutusid sünonüümsete nimedega organismideks, millest mitmed olid ümber klassifitseeritud. Ühtse nimetamise süsteemi puudumisel võib programmi kasutaja tulemusi valesti tõlgendada, mistõttu peaks tulemusi analüüsides võõraid või esmapilgul valesid tüvesid lähemalt uurima. Samuti seab see kahtluse alla NCBI taksonoomia kasutamise juhtpuuna, kus olulised pole mitte

organismide nimed, vaid nende omavaheline paiknemine puul.

Andmebaasi konstrueerimisel on üks olulistest parameetritest k-meeri pikkus. Mida pikemaid k-meere kasutada, seda spetsiifilisemad need on, kuid ühtlasi mõjutavad neid rohkem ka sekveneerimisvead (Patro *et al*., 2014). Väga lühikesed k-meerid seevastu esinevad suure tõenäosusega juhuslikult ka teistes genoomides, tehes need vähespetsiifiliseks. Puukujulise hierarhiaga k-meeride andmebaasi puhul paiknevad k-meerid vastavalt pikkuse kasvule tüvepoolsetes sõlmedes kuni olukorrani, kus kõik sõlmed peale tüvede on tühjad. Antud töös kasutati k-meeri pikkust 32, mis oli vajalik tüvede eristamiseks, kuid mille puhul jäid mitmed juurepoolsed sõlmed puhul tühjaks ning otsing teostati iga tekkinud alampuuga eraldi. Tühjad sõlmed võisid mõningal määral tundlikkust alandada. Lahenduseks võib olla puu jagamine erinevateks tasemeteks, millest igal kasutatakse erinevat k-meeri pikkust. Peale selle võib tundlikkus paraneda, kui nõutud proovist leitud sõlme k-meeride hulk muutuks dünaamiliselt koos sõlme suuruse ning proovi mitmekesisusega.

Mitmed klassifitseerimisprogrammid peavad määramiseks vajaliku andmebaasi mällu lugema. Väiksema mälumahuga arvutitel võib see probleemiks osutuda ning seetõttu on näiteks Krakeni põhiandmebaasi (70 GB) kõrvale loodud väiksem, vähendatud andmebaas – MiniKraken (4 GB). Krakeni andmebaasis on k-meerid ühes suures nimekirjas ning vähendamiseks jäetakse alles vaid iga 19. k-meer (Wood *et al*., 2014). MiniKrakeni andmebaas annab sarnaseid tulemusi Krakeni põhiandmebaasile, kuid k-meere võib kaduda ka sellistest sõlmedest, kus neid oli juba varasemalt vähe. Kuigi StrainSeeker andmebaasi mällu ei loe, siis Krakeni näitel pole täissuuruses andmebaas samuti vajalik (üle 100 GB). Selle asemel, et eemaldada k-meere igast sõlmest, võeti StrainSeekeri andmebaasis need vaid sealt, kus neid oli juba algselt väga palju (mõnes ligi 2 miljonit k-meeri) kuni igas sõlmes oli maksimaalselt 100 000 k-meeri. Tüvede tuvastamise tundlikkus kärpimise tagajärjel ei muutunud.

Genoomide tuvastamiseks proovitakse tavaliselt klassifitseerida igat lugemit eraldi. Sellisel lähenemisel saab kindel olla, et iga terviklik lugem on pärit ühest kindlast organismist. StrainSeeker ei klassifitseeri üksikuid lugemeid, sest eraldi võttes ei pruugi need sisaldada piisavalt infot, et nende päritolu korrektselt määrata. Erinevaid lugemeid võib määrata näiteks vaid ühe k-meeri või muu markeri järgi, samas kui mõne teine lugem sisaldab paljusid. Probleem süveneb, kui üht spetsiifilist markerit

kasutatakse paljude lugemite klassifitseerimiseks või on üht lugemit sekveneerimisandmetes mitmekordselt. Samuti võivad probleeme tekitada proovis esinevad genoomid, mida andmebaasis pole, kuid mis juhuslikult sisaldavad mõnd andmebaasis olevat markerit ja põhjustavad valepositiivseid tulemusi.

Kogu proovi k-meere uurides ning juhtpuud kasutades on mõningal määral võimalik ennustada andmebaasist puuduvate tüvede olemasolu proovis (*O*/*E* kasutamisest punktis 2.4.), mida lugemite klassifitseerimisel oleks keeruline teha. Lugemid jaotuksid mitme erineva tüve vahel ning on raske kindlaks teha, kas proovis oli tundmatu organism või mitmed teadaolevad tüvesid väikses koguses. StrainSeekeri piiranguks tundmatute tüvede tuvastamisel on nende paiknemine olemasoleva juhtpuu suhtes. Kuna andmebaasis on juhtpuu jaotatud väiksemateks alampuudeks, võib juhtuda, et tundmatu organism on lahkenud varasemalt kui on lähim alamjuur (juhtpuu juure poolt) ning seega pole võimalik seda tuvastada. Probleemi ei tohiks tekkida, kui andmebaas sisaldab piisavalt tundmatu organismile lähedasi sugulasi. Programmi testimisel on selliseks näiteks *Actinomyces odontolyticus´*e mitteleidmine, sest andmebaasis olid vaid sama perekonna kaugemad liigid, mis moodustasid omaette alampuu. Samal põhjusel jäi ka Lindgreen *et al*. (417 perekonda) andmestikust leidmata kolm perekonda, kus andmebaasist puudus proovis olev liik, kuid oli olemas sama perekonna teine liik. Parim viis tundmatute organismide avastamiseks proovist oleks ilmselt lugemite eelnev assambleerimine. Piisava katvuse korral kõrvaldaks see ka sekveneerimisvead ja tulemust saaks üsna täpselt olemasolevate järjestustega võrrelda – ka nukleotiidi täpsusega. K-meeride võrdlemisel pole aga võimalik vigu efektiivselt eemaldada ning seega ei saa eristada tundmatule genoomile kuuluvat k-meeri sekveneerimisviga sisaldavast k-meerist.

StrainSeekeri testimiseks ning võrdlemiseks teiste programmidega kasutati kunstlikke ja simuleeritud proove seetõttu, et nende koostis on teada. Reaalse proovi kasutamisel puuduks erinevate programmide tulemuste võrdlusel mõte, sest ei ole teada, mis on õige ja mis vale. Näiteks Peabody *et al*. (2015) töö näitab, et isegi 11 liiki sisaldava proovi sekveneerimisel varieeruvad töötlemata (*no cutoff* – arvestati kõiki tulemusi) tulemused väga palju. Lisaks 11-le õigele liigile sisaldasid programmide tulemused lisaks kuni 1226 valepositiivset liiki, kuigi enamik kasutatud 11-st tüvest olid andmebaasides olemas. Piirmäära tõstmine ühe protsendini küll parandas tulemusi,

kuid enamasti kadus koos valepositiivsetega ka mõni õige (jäeti välja kõik organismid, mille lugemite arv oli alla 1% kõigist lugemitest). Ainsate eranditena andsid töödeldud tulemustes kõik 11 liiki ilma valepositiivseteta Kraken ja CLARK (Peabody *et al*., 2015). Artikli autorid tõid välja ka selle, et teistega võrreldes parima tulemuse andnud Kraken eksis enim *E. coli* ja *Bacillus cereus´*e eristamisel, mis on põhjustatud ajalooliselt väljakujunenud NCBI taksonoomiast. Vigase juhtpuu eest hoiatavad ka mitmed teised tööd (Koslicki ja Falush, 2016). StrainSeeker leidis proovist kõik 11 liiki koos ühe valepositiivse tulemusega. Piirmäära tõstmisega kaoks esimesena valepositiivne tulemus, kuid et StrainSeekeril on piirangud juba otsinguprotsessis, siis pole see võrreldav üksikute lugemite arvuga teiste programmide kasutamisel. StrainSeekeri tundlikkus 0,95 perekonna tasemel on võrreldav CLARK-iga ning on parem kui Krakenil (Peabody *et al*., 2015). Täpsust programmidel ei arvutatud ning kuna töös võrreldakse õigesti ja valesti määratud lugemite arvu, mitte lõpptulemusi, pole need tulemused StrainSeekeriga võrreldavad.

Vaatamata sellele, et simuleeritud andmetega võrreldi vaid perekonna taset, andis StrainSeeker selle andmestikuga täpsemaid tulemusi kui reaalsete sekveneerimisandmetega. Sama märgati ka Peabody *et al*. (2015) töös, kus samu genoome nii sekveneeriti kui ka simuleeriti. Erinevus võib tulla simuleerimistarkvara ebatäpsuses imiteerida reaalselt sekveneerimises tekkivaid vigu.

StrainSeeker ei suuda eristada tüvesid, millel on vaid mõnenukleotiidilised erinevused. Nii täpsel tuvastamisel on suur roll sekveneerimisel tekkinud vigadel, mis omakorda tekitavad vigaseid k-meere. Mingil määral on võimalik veaga k-meere eristada hulga järgi: õige k-meeri sagedus on kõrgem kui veaga k-meeril. Kui aga katvus on madal (<1), sarnanevad õigete ja vigadega k-meeride sagedused niivõrd, et neid pole võimalik eristada. Selleks, et näiteks ühenukleotiidset erinevust kahe tüve vahel leida, on vajalik nii assambleerimine (või referentsile joondamine) kui ka väga kõrge katvus, mis üldiselt eeldab isolaatide kasutamist (Walker *et al*., 2013).

StrainSeekeri andmebaas, mida kasutati käesolevas töös, koosneb bakterite ja arhede genoomidest, kuid metoodika võib töötada ka suuremate eukarüootsete genoomidega. Kuna paljud genoomid on saadaval vaid lugemite või kontiigide kujul, aitaks StrainSeekeri andmebaasi täiustamisele kaasa ka võimalus kasutada assableerimata

genoome. Põhiliseks takistuseks selle lähenemise korral on vigaste k-meeride eristamine õigetest.

Sekveneerimistehnoloogia arenguga muutuvad lugemid üha pikemaks ja neid saab üha suurema täpsusega joondada. Sellegipoolest, tänapäeval on enim levinud Illumina sekvenaatorid, mille andmete analüüsimisel on k-meeripõhised klassifitseerijad kiiremad ning kuigi programmide tundlikkus kasvab lugemi pikkusega, jääb lugemi määramise täpsus üsna muutumatuks (Peabody *et al*., 2015).

KOKKUVÕTE

Bakterite tuvastamiseks proovist kasutatakse tavaliselt 16S rRNA geenipiirkonna amplikone, kuid üha enam ka kogu proovi DNA sekveneerimist. Enamik teise põlvkonna sekveneerimisandmete saadud lugemite analüüsiks kasutatavaid programme põhinevad joondamisel, mis on ressursimahukas ja koos assambleerimisega võib muuta analüüsi väga ajakulukaks. Lahenduseks võivad olla k-meeripõhised programmid, mis on kordades kiiremad kui joondamisel põhinevad programmid.

K-meerid on lühikesed kindla pikkusega oligomeerid, mida organismide tuvastamisel on kasutatud seni eelkõige kompositsiooni uurimisel (sarnaneb GC%-le). Hiljutised programmid, mis kasutavad pikemaid k-meere, on märkimisväärselt kiiremad kui joondamisalgoritmid ja vähemalt sama täpsed. K-meeridel põhinevad algoritmid kasutavad enamasti puukujulise struktuuriga andmebaasi, mis põhineb küllaltki staatilisel NCBI taksonoomial. Peale selle tehakse otsus, kas organism proovis esines, vaid mõne leitud k-meeri põhjal, mis lugemis leidusid.

Käesoleva töö käigus loodi k-meeridel põhinev bakterite määraja nimega StrainSeeker, mis kasutab sisendina assambleerimata teise põlvkonna sekveneerimisandmeid. Selle asemel, et iga lugemit eraldi vaadata, uuritakse kogu proovi k-meere koos. Kasutajal on võimalik andmebaasi loomisel kasutada oma assambleeritud genoome ja juhtpuud, andes võimaluse kohandada andmebaas vastavalt vajadustele.

Algoritmi testiti nii kunstlike proovide kui ka simuleeritud andmestikega, see suutis tuvastada praktiliselt kõik proovis olnud organismid, mis olid andmebaasis, olles võrreldav parimate seni avaldatud programmidega.

StrainSeeker on allalaetav ja veebis kasutatav aadressil www.bioinfo.ut.ee/strainseeker.

Identifying bacterial strains from unassembled sequencing reads using fixed-length oligomers

Mihkel Vaher

SUMMARY

Second generation sequencing of metagenomic samples provides valuable information about the organisms present in the sample. However, the information is in the form of short reads, which represent a very small part of a certain genome. Traditionally, every read is mapped to a reference genome with a programm such as BLAST. As the amount of sequencing data is rapidly growing, there is a need for more efficient algorithms.

Exact matching of short oligonucleotides (k-mers) for read classification has an accuracy comparable to BLAST while being much faster. Current k-mer–based classification algorithms mostly use a guide tree in their search process. This tree is usually built according to a trusted, publicly available taxonomy source such as NCBI taxonomy, constraining the user to a given set of genomes. Moreover, the decision to assign the read to a genome might be based only on a few k-mers that could in turn lead to false positive results.

A k-mer–based identification algorithm - named StrainSeeker - was designed to tackle the problems presented, giving users the ability to construct the database and guide tree from any assembled genomes. Also, the decision whether the sample contains an organism is made by taking all the sample k-mers into account.

The algorithm was tested on two mock samples (real sample with known composition) and on an *in silico* simulated dataset. In the first mock sample (containing 22 different species), StrainSeeker identified all the species that were present in its database (20), most with strain-level accuracy. The results for the second mock sample (11 species) contained a false positive result along with the 11 true ones. Overall, StrainSeeker is comparable with the best identification algorithms available, such as Kraken, and is more accurate than most of the other programs tested on the same mock and simulated samples.

StrainSeeker can be used online and downloaded at www.bioinfo.ut.ee/strainseeker.

TÄNUSÕNAD

Tänan oma juhendajat Märt Roosaaret, kellega koostöös käesolev töö valmis ning kes alati abistas nii nõu kui jõuga. Tänan ka Märt Mölsi, kes StrainSeekeri jaoks lihtsate piirarvude asemele statistilised testid koostas ning Maarja Lepametsa ja Lauris Kaplinskit, kelle loodud tarkvaral StrainSeeker töötab. Tänan veel Reidar Andersoni, kes andis töö kohta väärtuslikku tagasisidet. Lisaks tänan kõiki töögrupi ja TÜMRI töötajaid, kes aitasid töö valmimisele kaasa. Tänamata ei saa jätta ka Maido Remmi, kes andis mulle võimaluse töötada tema töögrupis.

KASUTATUD KIRJANDUS

Acinas, S. G., Sarma-rupavtarm, R., Polz, M. F., Acinas, S. G., Sarma-rupavtarm, R., Klepac-ceraj, V., Polz, M. F. (2005). PCR-Induced Sequence Artifacts and Bias : Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample PCR-Induced Sequence Artifacts and Bias : Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Sam, *71*(12), 8966–8969. http://doi.org/10.1128/AEM.71.12.8966

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–10. http://www.ncbi.nlm.nih.gov/pubmed/2231712

Angiuoli, S. V., White, J. R., Matalka, M., White, O., Fricke, W. F. (2011). Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. *PLoS ONE*, *6*(10).

Bansal, A. K., Meyer, T. E. (2002). Evolutionary analysis by whole-genome comparisons. *Journal of Bacteriology*, *184*(8), 2260–2272. http://doi.org/10.1128/JB.184.8.2260-2272.2002

Brady, A., Salzberg, S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, *6*(9), 673–6. http://doi.org/10.1038/nmeth.1358

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H., Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, *523*(7559), 208–211. http://doi.org/10.1038/nature14486

Buchfink, B., Xie, C., Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. http://doi.org/10.1038/nmeth.3176

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. a, Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. a, Mcdonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. a, Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R. (2010). correspondence QIIME allows analysis of high- throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. *Nature Publishing Group*, *7*(5), 335–336. http://doi.org/10.1038/nmeth0510-335

Chew, Y. V., Holmes, A. J. (2009). Suppression subtractive hybridisation allows selective sampling of metagenomic subsets of interest. *Journal of Microbiological Methods*, *78*(2), 136–143. http://doi.org/10.1016/j.mimet.2009.05.003

Darling, A. E., Jospin, G., Lowe, E., Matsen, F. A., Bik, H. M., Eisen, J. A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, *2*, e243. http://doi.org/10.7717/peerj.243

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072. http://doi.org/10.1128/AEM.03006-05

Deschavanne, P. J., Giron, a, Vilain, J., Fagot, G., Fertil, B. (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Molecular Biology and Evolution*, *16*(10), 1391–1399.

Dröge, J., Mchardy, A. C. (2012). Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Briefings in Bioinformatics*, *13*(6), 646–655. http://doi.org/10.1093/bib/bbs031

Fan, H., Ives, A. R., Surget-Groba, Y., Cannon, C. H. (2015). An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics*, *16*(1), 522. http://doi.org/10.1186/s12864-015-1647-5

Forster, S. C., Browne, H. P., Kumar, N., Hunt, M., Denise, H., Mitchell, a., Finn, R. D., Lawley, T. D. (2015). HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. *Nucleic Acids Research*, *44*(11), 604–609. http://doi.org/10.1093/nar/gkv1216

Garcia-Garcerà, M., Garcia-Etxebarria, K., Coscoll, M., Latorre, A., Calafell, F. (2013). A New Method for Extracting Skin Microbes Allows Metagenomic Analysis of Whole-Deep Skin. *PLoS ONE*, *8*(9), 1–12. http://doi.org/10.1371/journal.pone.0074914

Hiraoka, S., Machiyama, A., Ijichi, M., Inoue, K., Oshima, K., Hattori, M., Yoshizawa, S., Kogure, K., Iwasaki, W. (2016). Genomic and metagenomic analysis of microbes in a soil environment

affected by the 2011 Great East Japan Earthquake tsunami. *BMC Genomics*, *17*(1), 53. http://doi.org/10.1186/s12864-016-2380-4

Hugenholtz, P., Pace, N. R. (1996). Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends in Biotechnology*, *14*(6), 190–7. http://doi.org/10.1016/0167-7799(96)10025-1

Hunter, C. I., Mitchell, A., Jones, P., Mcanulla, C., Pesseat, S., Scheremetjew, M., Hunter, S. (2012). Metagenomic analysis: The challenge of the data bonanza. *Briefings in Bioinformatics*, *13*(6), 743–746. http://doi.org/10.1093/bib/bbs020

Huson, D., Auch, A., Qi, J., Schuster, S. (2007). MEGAN analysis of metagenome data. *Gennome Res.*, *17*, 377–386. http://doi.org/10.1101/gr.5969107.

Kaplinski, L., Lepamets, M., Remm, M. (2015). GenomeTester4: a toolkit for performing basic set operations - union, intersection and complement on k-mer lists. *GigaScience*, *4*(1), 58. http://doi.org/10.1186/s13742-015-0097-y

Karch, H., Tarr, P. I., Bielaszewska, M. (2005). Enterohaemorrhagic Escherichia coli in human medicine. *International Journal of Medical Microbiology*, *295*(6-7), 405–418. http://doi.org/10.1016/j.ijmm.2005.06.009

Karlin, S., Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics*, *11*(7), 283–290. http://doi.org/10.1016/S0168-9525(00)89076-9

Kawulok, J., Deorowicz, S. (2015). CoMeta: Classification of Metagenomes Using k-mers. *Plos One*, *10*(4), e0121453. http://doi.org/10.1371/journal.pone.0121453

Kim, M., Zhang, X., Ligo, J. G., Farnoud, F., Veeravalli, V. V., Milenkovic, O. (2016). MetaCRAM: an integrated pipeline for metagenomic taxonomy identification and compression. *BMC Bioinformatics*, *17*(1), 94. http://doi.org/10.1186/s12859-016-0932-x

Koslicki, D., Falush, D. (2016). MetaPalette: A K-mer painting approach for metagenomic taxonomic profiling and quantification of novel strain variation. *bioRxiv*. http://biorxiv.org/content/early/2016/02/17/039909.abstract

Laczny, C. C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H., Coronado, S., der Maaten, L., Vlassis, N., Wilmes, P. (2015). VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*, *3*(1), 1. http://doi.org/10.1186/s40168-014-0066-1

Lan, R., Reeves, P. R. (2002). Escherichia coli in disguise: Molecular origins of Shigella. *Microbes and Infection*, *4*(11), 1125–1132. http://doi.org/10.1016/S1286-4579(02)01637-4

Langmead, B., Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*(4), 357–359.

Lindgreen, S., Adair, K. L., Gardner, P. (2015). An evaluation of the accuracy and speed of metagenome analysis tools. *bioRxiv*, 017830. http://doi.org/10.1101/017830

Liu, Z., Desantis, T. Z., Andersen, G. L., Knight, R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Research*, *36*(18).

McHardy, A. C., García Martín, H., Tsirigos, A., Hugenholtz, P., Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, *4*(1), 63–72. http://dx.doi.org/10.1038/nmeth976

Ondov, B. D., Treangen, T. J., Mallonee, A. B., Bergman, N. H., Koren, S., Phillippy, A. M. (2015). Fast genome and metagenome distance estimation using MinHash. *bioRxiv*, 029827. http://doi.org/10.1101/029827

Ounit, R., Wanamaker, S., Close, T. J., Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, *16*(1), 236. http://doi.org/10.1186/s12864-015-1419-2

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science (New York, N.Y.)*, *276*(5313), 734–740. http://doi.org/10.1126/science.276.5313.734

Patil, K. R., Haider, P., Pope, P. B., Turnbaugh, P. J., Morrison, M., Scheffer, T., McHardy, A. C. (2011). Taxonomic metagenome sequence assignment with structured output models. *Nat Meth*, *8*(3), 191–192. http://dx.doi.org/10.1038/nmeth0311-191

Patro, R., Mount, S. M., Kingsford, C. (2014). Seq Reads Using Lightweight Algorithms, *32*(5), 462–464. http://doi.org/10.1038/nbt.2862.Sailfish

Peabody, M. A., Van Rossum, T., Lo, R., Brinkman, F. S. L. (2015). Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics*, *16*(1), 363. http://doi.org/10.1186/s12859-015-0788-5

Roosaare, M., Vaher, M., Kaplinski, L., Mols, M., Andreson, R., Lepamets, M., Koressaar, T., Naaber, P., Koljalg, S., Remm, M. (2016). StrainSeeker: fast identification of bacterial strains from unassembled sequencing reads using user-provided guide trees. *bioRxiv*.

http://biorxiv.org/content/early/2016/02/19/040261.abstract
Schloss, P. D. (2010). The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. *PLoS Comput Biol*, *6*(7), 1–16. http://doi.org/10.1371/journal.pcbi.1000844
Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, *9*(8), 811–4. http://doi.org/10.1038/nmeth.2066
Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, *5*(June), 209. http://doi.org/10.3389/fpls.2014.00209
Stackebrandt, E., Goebel, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic Bacteriology*, *44*(4), 846–849. http://doi.org/10.1099/00207713-44-4-846
Tamura, K., Stecher, G., Paterson, D., Filipski, A., Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis software version 6.0. *Comput Appl Biosci*, *30*, 2725–2729. http://doi.org/10.1093/molbev/msm092
Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K., Tolstoy, I. (2014). RefSeq microbial genomes database: New representation and annotation strategy. *Nucleic Acids Research*, *42*(D1).
The NIH HMP Working Group. (2009). The NIH Human Microbiome Project. *Genome Research*, *19*(12), 2317–2323. http://doi.org/10.1101/gr.096651.109
Tringe, S. G., Hugenholtz, P. (2008). A renaissance for the pioneering 16S rRNA gene. *Current Opinion in Microbiology*, 442–446. http://doi.org/10.1016/j.mib.2008.09.011
Tringe, S. G., Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews. Genetics*, *6*(11), 805–14. http://doi.org/10.1038/nrg1709
Větrovský, T., Baldrian, P. (2013). The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS ONE*, *8*(2), 1–10. http://doi.org/10.1371/journal.pone.0057923
Walker, T. M., Ip, C. L. C., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., Eyre, D. W., Wilson, D. J., Hawkey, P. M., Crook, D. W., Parkhill, J., Harris, D., Walker, A. S., Bowden, R., Monk, P., Smith, E. G., Peto, T. E. A. (2013). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: A retrospective observational study. *The Lancet Infectious Diseases*, *13*(2), 137–146. http://doi.org/10.1016/S1473-3099(12)70277-3
Wang, Q., Garrity, G. M., Tiedje, J. M., Cole, J. R. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. http://doi.org/10.1128/AEM.00062-07
Ward, D. V., Gevers, D., Giannoukos, G., Earl, A. M., Methé, B. A., Sodergren, E., Feldgarden, M., Ciulla, D. M., Tabbaa, D., Birren, B. W. (2012). Evaluation of 16s rDNA-based community profiling for human microbiome research. *PLoS ONE*, *7*(6). http://doi.org/10.1371/journal.pone.0039315
Ward, B. B. (2002). How many species of prokaryotes are there? *Proceedings of the National Academy of Sciences of the United States of America*, *99*(16), 10234–10236.
Woese, C. R., Kandler, O., Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(12), 4576–4579. http://doi.org/10.1073/pnas.87.12.4576
Wood, D. E., Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. http://doi.org/10.1186/gb-2014-15-3-r46
Wu, D., Jospin, G., Eisen, J. A. (2013). Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. *PLoS ONE*, *8*(10), e77033. http://doi.org/10.1371/journal.pone.0077033
Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., Suez, J., Mahdi, J. A., Matot, E., Malka, G., Kosower, N., Rein, M., Zilberman-Schapira, G., Dohnalová, L., Pevsner-Fischer, M., Bikovsky, R., Halpern, Z., Elinav, E., Segal, E. (2015). Personalized Nutrition by Prediction of Glycemic Responses. *Cell*, *163*(5), 1079–1095. http://doi.org/10.1016/j.cell.2015.11.001

LISA 1

StrainSeekeri tulemused proovile SRR172902 SRA andmebaasist.

```
28.30%  KNOWN   Deinococcus_radiodurans_R1_chromosome_1
8.86%   KNOWN   Acinetobacter_baumannii_ATCC_17978-mff
6.06%   RELATED
        Staphylococcus_aureus_strain_M121,Staphylococcus_aureus_subsp_aureus_strain_USA300_2014C01,Staphylococcus_a
ureus_USA300-
ISMMS1,Staphylococcus_aureus_strain_CA15,Staphylococcus_aureus_strain_V2200,Staphylococcus_aureus_subsp_aureus_USA
300_FPR3757,Staphylococcus_aureus_subsp_aureus_strain_USA300_2014C02,Staphylococcus_aureus_subsp_aureus_USA300_TCH
1516,Staphylococcus_aureus_strain_25b_MRSA,Staphylococcus_aureus_strain_27b_MRSA,Staphylococcus_aureus_strain_33b,
Staphylococcus_aureus_strain_26b_MRSA,Staphylococcus_aureus_strain_29b_MRSA,Staphylococcus_aureus_strain_UA-
S391_USA300,Staphylococcus_aureus_strain_31b_MRSA
5.64%   KNOWN   Streptococcus_pneumoniae_TIGR4_chromosome
5.24%   KNOWN   Bacteroides_vulgatus_ATCC_8482
4.70%   KNOWN   Staphylococcus_epidermidis_ATCC_12228_chromosome
4.67%   KNOWN   Propionibacterium_acnes_KPA171202
4.46%   RELATED Helicobacter_pylori_strain_26695-1MET,Helicobacter_pylori_26695-1CL_DNA,Helicobacter_pylori_26695-
1,Helicobacter_pylori_26695-1CH_DNA,Helicobacter_pylori_26695-1_DNA
3.39%   KNOWN   Streptococcus_mutans_UA159_chromosome
3.27%   KNOWN   Staphylococcus_aureus_subsp_aureus_CN1
3.24%   RELATED Staphylococcus_aureus_strain_MSSA476,Staphylococcus_aureus_subsp_aureus_MW2_DNA
3.03%   RELATED Staphylococcus_aureus_subsp_aureus_strain_GR2
3.03%   KNOWN   Neisseria_meningitidis_MC58_chromosome
2.82%   RELATED Staphylococcus_aureus_subsp_aureus_ST772-MRSA-
V_strain_DAR4145,Staphylococcus_aureus_strain_SA564,Staphylococcus_aureus_strain_502A,Staphylococcus_aureus_subsp_
aureus_JH9,Staphylococcus_aureus_subsp_aureus_ST228_complete_genome,Staphylococcus_aureus_subsp_aureus_ED98,Staphy
lococcus_aureus_04-02981,Staphylococcus_aureus_subsp_aureus_ECT-
R_2_complete_genome,Staphylococcus_aureus_subsp_aureus_Mu50_DNA,Staphylococcus_aureus_strain_FCFHV36,Staphylococcu
s_aureus_subsp_aureus_Mu3_DNA,Staphylococcus_aureus_subsp_aureus_N315_DNA,Staphylococcus_aureus_subsp_aureus_JH1
2.32%   KNOWN   Streptococcus_mutans_UA159-FR
1.96%   RELATED
        Rhodobacter_sphaeroides_241_chromosome_1,Rhodobacter_sphaeroides_ATCC_17029_chromosome_1,Rhodobacter_sphaer
oides_WS8N_chromosome_chrI,Rhodobacter_sphaeroides_KD131_chromosome_1
1.84%   RELATED Listeria_monocytogenes_strain_SLCC2372,Listeria_monocytogenes_FSL_R2-
561,Listeria_monocytogenes_strain_SLCC2479
1.61%   KNOWN   Clostridium_beijerinckii_NCIMB_8052
1.38%   KNOWN   Enterococcus_faecalis_OG1RF
1.35%   RELATED Escherichia_coli_K-12_GM4792_Lac-
,Escherichia_coli_BW25113,Escherichia_coli_strain_CQSW20,Escherichia_coli_strain_K-
12_substrain_MG1655_TMP32XR2,Escherichia_coli_strain_K-
12_substrain_MG1655_TMP32XR1,Escherichia_coli_KLY,Escherichia_coli_K-
12_strain_ER3435,Escherichia_coli_BW2952,Escherichia_coli_ER2796,Escherichia_coli_DH1,Escherichia_coli_strain_RR1,
Escherichia_coli_K-12_strain_ER3476,Escherichia_coli_str_K-12_substr_MG1655_chromosome,Escherichia_coli_K-
12_strain_ER3466,Escherichia_coli_str_K-12_substr_MG1655,Escherichia_coli_strain_SQ2203,Escherichia_coli_K-
12_strain_ER3413,Escherichia_coli_genome_assembly_EcRV308Chr,Escherichia_coli_K-
12_strain_ER3445,Escherichia_coli_K-
12_genome_assembly_EcoliK12AG100,Escherichia_coli_strain_DH1Ec104,Escherichia_coli_strain_DH1Ec169,Escherichia_col
i_str_K12_substr_DH10B,Escherichia_coli_K-
12_strain_ER3475,Escherichia_coli_str_K12_substr_W3110_DNA,Escherichia_coli_strain_SQ37,Escherichia_coli_str_K-
12_substr_MC4100_complete_genome,Escherichia_coli_strain_SQ88,Escherichia_coli_K-
12_strain_ER3454,Escherichia_coli_K-
12_strain_ER3446,Escherichia_coli_strain_DH1Ec095,Escherichia_coli_DH1_ME8569_DNA,Escherichia_coli_genome_assembly
_EcHMS174Chr,Escherichia_coli_K-12_strain_ER3440
1.11%   KNOWN   Methanobrevibacter_smithii_ATCC_35061
1.09%   KNOWN   Pseudomonas_aeruginosa_PAO1_chromosome
0.45%   KNOWN   Bacillus_cereus_ATCC_10987
0.14%   KNOWN   Streptococcus_agalactiae_2603VR_chromosome
0.05%  KNOWN        Lactobacillus_gasseri_ATCC_33323
```

LISA 2

StrainSeekeri tulemused proovile FW *in vitro* Peabody *et al*. (2015) tööst.

16.74% KNOWN Micrococcus_luteus_NCTC_2665_uid59033
9.59% KNOWN Bacillus_amyloliquefaciens_DSM_7_uid53535
9.17% KNOWN Escherichia_coli_K_12_substr__DH10B_uid58979
9.16% KNOWN Rhodobacter_capsulatus_SB_1003_uid47509
7.97% KNOWN Escherichia_coli_BW2952_uid59391
7.48% KNOWN Pseudomonas_fluorescens_Pf_5_uid57937
6.74% KNOWN Frankia_CcI3_uid58397
6.58% RELATED Escherichia_coli_K_12_substr__W3110_uid161931
6.56% KNOWN Pseudomonas_putida_KT2440_uid57843
6.28% KNOWN Pseudomonas_aeruginosa_UCBPP_PA14_uid57977
4.84% KNOWN Burkholderia_cenocepacia_J2315_uid57953
3.89% KNOWN Bacillus_cereus_ATCC_14579_uid57975
3.71% KNOWN Streptomyces_coelicolor_A3_2__uid57801
0.93% KNOWN Pseudomonas_aeruginosa_PAO1_uid57945
0.36% RELATED Burkholderia_phytofirmans_PsJN_uid58729, Burkholderia_xenovorans_LB400_uid57823

LIHTLITSENTS

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Mihkel Vaher (08.04.1991)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

Bakteritüvede tuvastamine sekveneerimise toorlugemitest kindla pikkusega oligomeeride abil,

mille juhendaja on Märt Roosaare,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace-i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 27.05.2016