TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOINFORMAATIKA ÕPPETOOL

Telomeeride keskmise pikkuse hindamine teise generatsiooni sekveneerimisandmetest

Bakalaureusetöö

Lõputöö maht (12 EAP)

Karl-Sander Erss

Juhendaja Tarmo Puurand

TARTU 2017

# **Sissejuhatus**

# Kirjanduse ülevaade

## Telomeerid

Telomeerid on spetsiaalsete valkudega seotud tandemkordused kromosoomide otstes, mis kaistsevad neid lagundamise ja kokkukleepumise eest. Telomeerid esinevad vaid lineaarsetes kromosoomides. (Witzany, 2008)

### Struktuur

Imetajate telomeerid koosnevad TTAGGG kordustest ja valgukompleksist nimega *shelterin*. (Martínez & Blasco, 2010)

Telomeeride pikkus on erinevatel liikidel erinev – inimesel 10-15 Kb, hiirtel 25-50 Kb. Telomeeri otsale on iseloomulik 150-200 nukleotiidi pikkune üleulatuv G-rikas 3’ ahel, mis keerab lõpuosa T-linguks. Üleulatuv osa tungib kaheahelalise osa vahele ja moodustub D-ling. (Martínez & Blasco, 2010)

*Shelterin* koosneb kuuest ühikust:

* telomeeri kordusseonduvad faktorid (*telomeric repeat binding factors*) TRF1 ja TRF2
* TRF-1 interakteeruv valk 2 TIN2
* kaitsev valk POT1
* POT1-TIN2-ograniseeriv valk TPP1
* repressor/aktivaator valk RAP1.

TRF1, TRF2 ja POT1 seonduvad otse telomeersele DNA-le. kusjuures viimane ainult üleulatuvale ahelale. TIN2 seondub ühe TRF valguga ja on vajalik TPP1-POT1 kompleksi seondumiseks.

### Funktsioon

### Otsa replikatsioon probleem

DNA polümeraasid paljundavad DNA-d vaid 5’ – 3’ suunal ja seega ei saa viivisahelat pidevalt paljundada. DNA paljundamisel toimub replikatsioonikahvlist tagasisuunalisel DNA ahelal DNA süntees lühikeste juppide, Okazaki fragmentide, kaupa. Selleks, et DNA pülmeraas saaks Okazaki fragmendi sünteesi alustada, seondub matriitsahelaga RNA-praimer. Nende praimerite asemele sünteesitakse hiljem 5’ – 3’ suunal DNA. Kuna pärast RNA eemaldamist DNA juppide sünteesiks läheb vaja 3’ OH otsa, aga kromosoomi otsas, viimase RNA praimeri järel rohkem DNA-d ei ole, jääb kromosoomi lõpust matriitsahelaga komplementaarne DNA sünteesimata. (Witzany, 2008)

### Telomeraas

Telomeraas on ribonukleoproteiin, mis koosneb telomeraasi RNA-st (TER) ja telomeraasi pöördtranskriptaasist (TERT). Telomeraas katalüüsib üleulatuva G-otsa uute telomeeri-korduste sünteesi. Telomeraasi aktiivsus on enamuses somaatilistes rakkudes madal või tuvastamatu, kuid umbes 85%-s vähirakkudes ülesreguleeritud. See telomeraasne aktiivsus panustab vähirakkude surematusse. (Tian, Chen, & Liu, 2009)

#### Telomeraasi aktiivsuse regulatsioon

Telomeraasi TER komponenti ekspresseeritakse kõikides rakkudes ühesugusel määral, kuid katalüütilist TERT subühikut ekspresseeritakse vaid vähirakkudes. TTAGGG korduste lisamine kromosoomi otsa toimub kahes etapis – esmalt sünteesitakse TER komponendiga komplementaarne DNA jupp ning pausi järel liigub ensüüm edasi. hTERT (*human TERT*) geen asub viienda kromosoomi lühema õla otsas. (Tian et al., 2009) Arvatakse, et TERT peamine regulatsioonimehhanism on transkriptsiooniline kontroll. Mitmete onkogeenide (nt c-Myc) ja kasvajate supressiooni geenide (WT1, p53) produktid avaldavad üleekspresseerimise korral mõju hTERT transkriptsioonile. (Horikawa & Barrett, 2003)

### Telomeeride pikkus

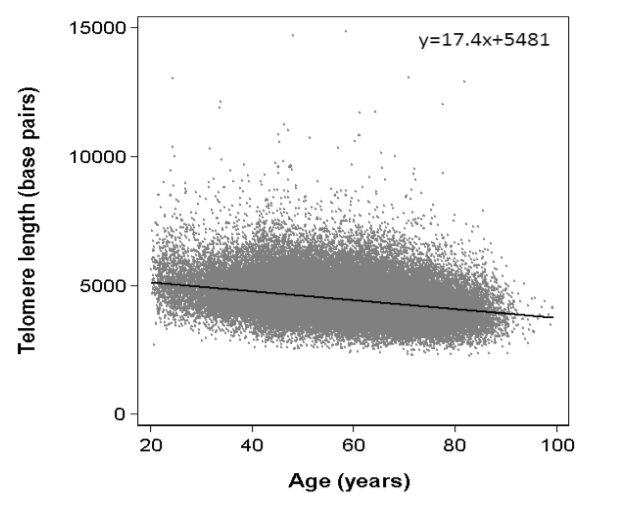
#### Telomeeride pikkus vananemise biomarkerina

Vananemise biomarkerite olulisus seisneb pigem inimese vanusest sõltuva tervise ja funktsionaalse oleku hindamises kui kronoloogilise vanuse määramises. Samuti vanusest tulenevate haiguste, suremisriski ning vananemisvastaste sekkumiste efektiivsuse hindamiseks.(Mather, Jorm, Parslow, & Christensen, 2011)

Telomeeride puhul on leitud, et telomeeride pikkus ja inimese vanus on negatiivselt korrelatsioonis. Ka on leitud seoseid telomeeride pikkuse ja muude vanusest sõltuvate näitajate, haiguste ning suremuse vahel, kuid tulemuste tõlgendused on ebaselged.

Näiteks üle 60 aastaste inimeste sead oli kõige lühemate telomeeridega grupil suurem suremus – kolm korda suurem suremus südamehaigustest ning kaheksa korda suurem nakkushaigustest. Kui sama uuringu andmeid analüüsiti aga vanusevahemike kaupa, ei olnud tulemus enam üle 74 aastaste seas statistiliselt oluline. Seda võib seletada ellujääja-efektiga – kui lühemate telomeeridega inimesed sureks varem, ei ole neid vanemates vahemikes. (Mather et al., 2011)

Longituuduuringute puhul on saadud väga varieeruvaid tulemusi – telomeeride pikkus võib aja jooksul nii suureneda kui väheneda. (Mather et al., 2011)



Joonis 1 Telomeeride pikkuse sõltuvus vanusest (Rode, Nordestgaard, & Bojesen, 2015)

### Telomeeride pikkust mõjutavad faktorid

Telomeeride pikkust võib mõjutada isa vanus järglase sünni ajal, põletikreaktsioonid, suitsetamine, füüsiline aktiivsus, sugu, klass, kehamassiindeks, multivitamiinide tarbimine, antioksüdantide tarbimine, alkoholi tarbimine, hormooniasendusteraapa ning rass. (Mather et al., 2011)

### Telomeeride pikkuse laboratoorne mõõtmine

#### Terminaalsete restriktsioonifragmentide analüüs - TRF

Selle meetodi kasutamise jaoks on vaja vähemalt 3ug puhastatud DNA-d. Järgmisena on oluline hinnata, kas eraldatud DNA on analüüsiks sobiv, kuna proovide kogumisel, hoiustamisel ning transpordil võib esineda mitmeid proove degradeerivaid asjaolusid. Selle jaoks analüüsitakse eraldatud DNA-d geelektroforeesil ning kinnitatakse, et proov visualiseerub tiheda ja mitte laialivalgunud vöödina. Katkiste DNA proovide analüüs selle meetodiga annab tulemuseks tegelikust lühema telomeeri pikkuse. (Kimura et al., 2010)

DNA lõigtakse restritsiooniensüümidega (*Hinf*I ja *RsaI*), millel pole äratundmiskohti ei telomeeri sees ega telomeeri-eelses alas. See protsess rikastab proovid pikkade telomeersete fraktsioonidega – ülejäänud genoom lõigatakse kuni 800bp pikkusteks tükkideks ning eraldatakse agaroosgeelil ning visualiseeritakse southern blot meetodiga. Visualiseerimiseks kasutatakse TTAGGG komplementaarseid märgistatud oligonukleotiide. Telomeeride pikkused saadakse *ladder*-DNA-ga võrdlemisel või eelnevalt valmistatud ruudustiku abil. (Kimura et al., 2010)

#### Kvantitatiivne PCR - qPCR

Vähem DNA materjali kui TRF analüüsi jaoks, kulub qPCR-põhiste meetoditega telomeeride pikkuse mõõtmiseks. qPCR meetodid põhinevad fluorofooride kasutamisel. Fluorofoorid annavad fluorestsents-signaali, kui huvipakkuv järjestus paljundatakase ning võimaldavad paljundatavat DNA-d kvantifitseerida. qPCR põhiste meetodite peamine keerukus seisneb selles, et telomeeri-spetsiifilised praimerid on üksteisega komplementaarsed ning moodustavad omavahel dimeere. (Montpetit et al., 2014)

Praimerite dimeeride tekkimise vastu aitas spetsiaalsete praimerite disain, mille puhul DNA polümeraas paljundas käivitus vaid siis, kui praimer oli seondunud telomeeriga, mitte teise praimeriga. Lisaks mõõdetakse selle meetodi puhul lisaks telomeeri amplifikatsiooni produktile (T) ka ühe *single-copy* geeni hulk (S). Nende põhjal saadakse T/S suhe, mis korreleerub keskmise telomeeri pikkusega. (Cawthon, 2002) Esimese qPCR põhise telomeeride pikkuse mõõtmise meetodi puuduseks oli mõõtmise ebatäpsus mis tekib T ja S võimenduste eraldi reaktsioonidest mõõtmisest. Selle ebatäpsuse vältimiseks tehakse meetodi edasiarendatud versioonis reaktsioon ühes tuubis – T signaal kogutakse PCR varajastes tsüklites, enne kui S signaal detekteerimispiiri ületab. (Cawthon, 2009)

#### Üksiku telomeeri pikkuse analüüs - STELA

TRF ja PCR põhiste meetoditega saab mõõta vaid telomeeride keskmist pikkust proovist. Kuna on näidatud, et ka üksiku või mõne telomeeri kriitiline lühenemine võib esile kutsuda rakujagunemise lõppemist või apoptoosi, võib üksikute telomeeride pikkuse mõõtmine olla rohkemate praktiliste kasutusaladega. STELA meetod täiendab tavalist qPCR põhist meetodit nii, et kasutab telomeeri-eelsele alale seonduvat praimerit. Selle meetodiga saab mõõta vaid nende kromosoomide telomeere, mille telomeeri-eelne ala on unikaalne (XpYp, 2p, 11q, 12q, and 17p). (Montpetit et al., 2014)

#### Hübridisatsiooni-kaitse meetod – HPA

See meetod võimaldab mõõta keskmist telomeeride pikkust nii puhastatud DNA-st kui ka rakulüsaadist. Erinevalt TRF meetodist, ei pea DNA intaktne olema. Meetod põhineb telomeersete korduste töötlemises komplementaarsete oligonukleotiididega, mis on märgistatud *acridinium* estri molekulidega. Hübridiseerumata märgistatud oligonukleotiidid inaktiveeritakse hüdrolüüsi-lahuse abil, kuid hübridiseerunud märgised on hüdrolüüsi eest kaitstud. Telomeerne DNA-lt mõõdetakse kemoluminestsents-signaal (T). Selleks, et saadud signaali põhjal telomeeride pikkust hinnata, mõõdetakse ka luminestsents-signaal, mis tekib mõne Alu-järjestusega komplementaarse märgistatud oligonukleotiidiga (A). Analoogselt qPCR meetodiga, saadakse TA-suhe. Kui võrreldi samast proovist saadud TRF analüüsi tulemust HPA meetodi TA suhtega, leiti, et Alu-element TGTAATCCCAGCACTTTGGGAGGC vastab TA suhe 0,01 umbes 2000 aluspaari pikkusele TR fragmendile. (Nakamura et al., 1999)

### Telomeeride pikkuse hindamine sekveneerimisandmetest

DNA pimejärjestamisega (*whole genome shotgun sequencing*) saadakse järjestused ka telomeersete osade kohta, kuid kuna telomeersed järjestused on väga korduvad, ei ole nende täpne referentsjärjestusega joondamine võimalik. Telomeeride pikkused on aga tuvastatavad TTAGGG korduste arvu põhjal. (Ding, Mangino, Aviv, Spector, & Durbin, 2014)

#### TelSeq

TelSeq tarkvara kasutab sisendiks BAM faili. TelSeq loeb kokku telomeerseid järjestusi sisaldavad *read*id ning hindab telomeeri füüsilist pikkust valemi *l=tksc* järgi, kus l on hinnatav pikkus, *tk* on telomeersete korduste hulk läve *k* juures, s on suurusfaktor, ning c on genoomi pikkus jagatud telomeeri otste (23×2) arvuga. Suurusfaktori väärtuseks valitakse nende readide arv, kus GC sisaldus on 48% kuni 52%. (Ding et al., 2014)

#### Computel

Computel tarkvara joondab *read*id telomeerse järjestuse indeksiga. Selleks, et täiesti kõik telomeersed *read*id arvutusel arvesse võetaks, koostatakse indeks nii, et loetaks ka neid *read*e, mis sisaldavad subtelomeerseid osi. Ka arvestab programm telomeersete järjestuste joondamisel sekveneerimsvigadega. Keskmine telomeeride pikkus arvutatakse valemiga MTL=(mean(rel.cov)) \* (rl+pl-1)/(2\*n\_chr), kus rel.cov on baaskatvuse (readide arv) \* (readi pikkus) / (genoomi pikkus) ning telomeerse katvuse (mitu read’i indeksiga joondati) suhe; rl on readi pikkus, pl on indeksi telomeerse osa pikkus ning n\_chr on kromosoomide arv. (Nersisyan & Arakelyan, 2015)

## Alu elemendid

Alu elemendid on inimese genoomis esinevad korduvjärjestused, mille koopiaarv on üle miljoni. Alu elemente on erinevaid ning need jaotuvad järjestuse järgi erinevatesse perekondadesse. Alu-elemendid on reptroposoonsed. (Roy-Engel et al., 2001)

## Teise põlvkonna sekveneerimisandmed

1970ndatel aastatel töötati välja DNA fragmenteerimisel ja ahela termineerimisel põhinev meetod DNA sekveneerimiseks. Seda meetodit nimetatakse Sangeri sekvneerimiseks. Aastaks 2004 sekveneeriti selle meetodiga inimese genoom. (Van Dijk, Lè Ne Auger, Jaszczyszyn, & Thermes, 2014)

Teise põlvkonna sekveneerimismeetodite (NGS – *next generation sequencing*) puhul kasutatakse DNA bakteriaalse kloonimise asemel rakuvaba süsteemi. Ka võimaldavad NGS tehnoloogiad miljoneid sekveneerimisreaktsioone paralleelselt jooksutada ning väljundi dekteerimiseks pole, erinevalt Sangeri meetodist, geelelektroforeesi vaja. Samas on NGS puuduseks lühikesed väljundjärjestused (ingk. *read*), mis teevad uute genoomide kokkupanemise või joondamise keerukaks. (Van Dijk et al., 2014)

Tänapäeval on populaarseim NGS platvorm Illumina sekveneerimismasinad. DNA paljundamiseks on vaja kolme komponenti – matriitsahelat, vabu nukleiinhappeid ning DNA polümeraasi. Illumina süsteem kasutab klaasplaadikest, kuhu on kinnitatud miljoneid erinevaid matriitsahelaid. Modifitseeritud DNA lõigud seonduvad plaadile kindlatesse kohtadesse, tugevama signaali saamiseks lõike paljundatakse. Peale võimendamissammu sünteesitakse fluorestseeruvalt märgistatud nukleotiidide abil DNA-le komplementaarne ahel. Peale iga nukleotiidi lisamist tehaks plaadist pilt ja loetakse flurestsentssignaalid. (Muzzey, Evans, & Lieber, 2015)

Kõikide Illumina uuemate seadmetega saab sekveneerimistulemuseks *paired-end read*id. See tähendab, et pikemast DNA lõigust sekveneeritakse ära kaks otsa ning on teada mitu nukleotiidi on nende kahe otsa vahel. See meetod vähendab NGS küllalt lühikeste readide joondamise keerukust ning võimaldab detekteerida insertsioone, deletsioone ning korduvaid järjestusi. (<https://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html>)

### Sekveneerimiskvaliteet

NGS protsessi viimane samm on fluorestsentssignaali põhjal nukleotiidi – A, T, C või G määramine, *base-calling*. Illumina sekveneerimisel esineb nii keemiast kui signaal mõõtmisest tulenevaid piiranguid. Kui sekveneerimistsükli jooksul jääb mõnele DNA ahelale nukleotiid lisamata, jääb see ahel teistest ahelatest maha. Mida rohkem ahelaid maha jääb, seda ebaselgemaks muutub fluorestsentssignaal, kuna ühe DNA tasemel esineb mitmeid erinevaid signaale. See efekt akumuleerub tsüklite jooksul ning seetõttu on nukleotiidi täpne määramine ahela lõpu poole ebatäpsem. Ka lisavad ebatäpsust tõigad, et märgiste emissioonispektrid kattuvad osaliselt ning värvuse intensiivsus on tugevam detekteerimispiirkonna keskme pool. (Ledergerber & Dessimoz, 2011)

Iga tsükli väljundiks on intensiivsus-skoorid, mis konverteeritakse base-calling tarkvara abil nukleotiidiks ning tõenäosuseks, et tegu on just selle nukleotiidiga. Nende tõenäosuste põhjal arvutatakse igale nukleotiidile kvaliteediskoor. (Cacho, Smirnova, Huzurbazar, & Cui, 2016) Levinuim viis kvaliteedi esitamiseks on Phred kvaliteediskoor *Q=-10log10P*, kus P on tõenäosus, et tegu on vale nukleotiidiga. Phred skoore esitatakse QUAL formaadis, mis koosneb iga *read*i kohta päisest ning täisarvude nimekirjast. (Cock, Fields, Goto, Heuer, & Rice, 2009)

Kõige populaarsem formaat NGS andmete esitamiseks on tekstipõhine FASTQ formaat, mis sisaldab nelja tüüpi ridu (ranges järjekorras):

1. @ märgiga algav päiserida
2. Järjestuse rida/read
3. + märgiga algav valikuline päiserea kordus
4. Kvaliteediskooride rida

Faili suuruse piiramiseks esitatakse kvaliteediskoorid vahemikus 0-93 ühe sümbolina ASCII vahemikust 33-126. See formaat võimaldab kvaliteediskoori väga täpset esitamist vahemikus 1.0 (vale nukleotiid) kuni 10-9.3 (väga täpselt määratud nukleotiid). (Cock et al., 2009)

### Sekveneerimiskatvus

Sekveneerimiskatvuse ehk katvuse all mõeldakse *read*ide arvu, mis on joondatult genoomil kohakuti. Kuna NGS tehnoloogiad teevad nukleotiidide määramisel vigu, on variantide määramiseks oluline, et üks positsioon genoomis oleks loetud mitu korda. (Muzzey et al., 2015)

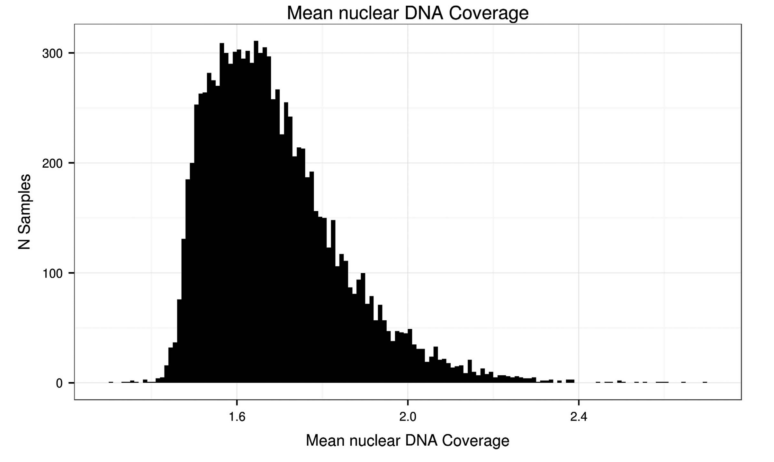
### Sequence Read Archive

Sequence Read Archive (SRA) on teise generatsiooni sekveneerimisandmete avalik arhiiv, mida haldab National Center for Biotechnology Information (NCBI). Arhiiv sisaldab töötlemata sekveneerimisandmeid ning metaandmeid. Arhiiv on ligipääsetav aadressil <https://www.ncbi.nlm.nih.gov/sra/>. (Kodama, Shumway, & Leinonen, 2012)

SRA andmete kasutamiseks on vajalik tööriistakomplekt nimega SRA Toolkit. See võimaldab arhiivifailid viia edasiseks analüüsiks sobivale kujule. Andmete allalaadimiseks tuleb kasutada tööriista *prefetch*. See laeb alla arhiivifaili ning referentsfailid millest arhiivifail sõltub. (<https://github.com/ncbi/sra-tools/wiki/HowTo:-Access-SRA-Data>)

### CONVERGE andmestik

CONVERGE andmestik on SRA andmebaasis avalikult saadaolev andmekogu, mis sisaldab 11670 Han Hiina naise madala katvusega täisgenoomi sekveneerimisandmeid. Andmed koguti depressioonihäire uurimise käigus. Andmekogu keskmine tuumagenoomi katvus on 1,7 kordne, inimeste vanused on vahemikus 30-60 ning alla laadides on andmete maht umbes 100 terabaiti. Sekveneerimiseks kasutati Illumina Hiseq seadmeid. (Cai et al., 2017)



**Joonis 2** CONVERGE andmestiku tuumagenoomi katvus. Keskmine katvus on 1.7X.

## K-mer metoodika sekveneerimisandmete analüüsis

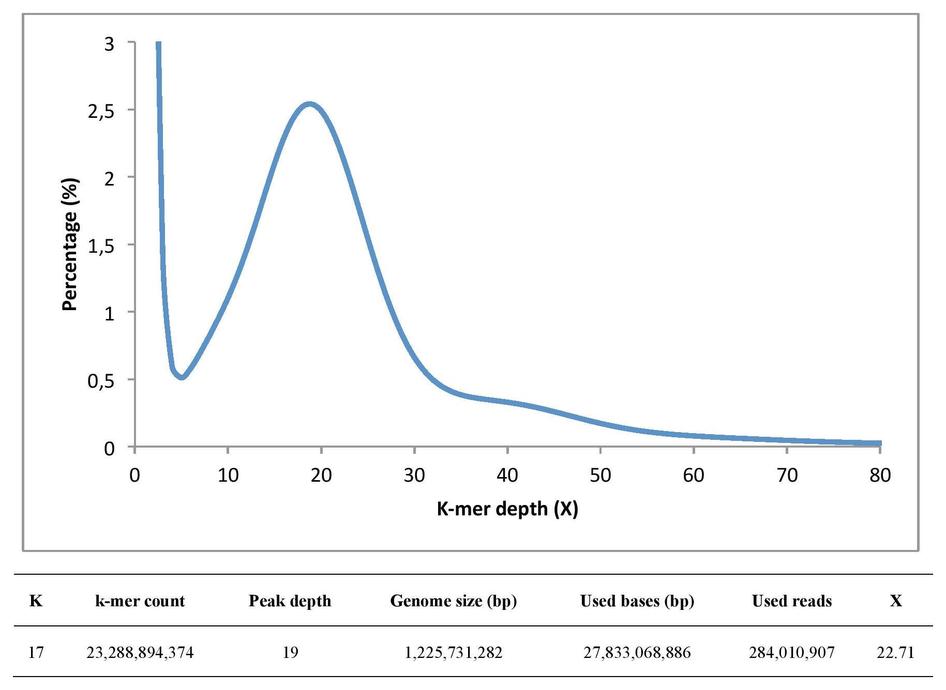
Traditsioonilised sekveneerimisandmete analüüsi meetodid põhinevad sekveneeritud lõikude referentsile joondamisel. Joondamisele põhinevate analüüside tulemused ning keerukus on väga tundlikud joondamisparameetrite valikule ja seega veaohtlikud. Samuti on lõikude referentsile joondamine arvutuslikult väga kulukas, mis võib viia selleni, et andmete kogumisvõimekus ületab analüüsivõimekuse. Samuti on probleemiks järjestused, mis sobivad referentsile mitmesse kohta. (Patro, Mount, & Kingsford, 2014)

Mõiste k-mer viitab kõikidele kindla pikkusega osasõnedele, mida sõne sisaldab. K on osasõne pikkus. (Compeau, Pevzner, & Tesler, 2011) Näiteks sõne GTAGAGCTGT 5-merid on GTAGA, TAGAG, AGAGC, GAGCT, AGCTG ja GCTGT.

K-mer põhilised analüüsimeetodid põhinevad k-meride lugemisel – mitu ühesugust k-meri andmetes esineb. See kokkulugemise protsess on üle 20 korra kiirem kui sekveneeritud lõikude referentsile joondamine. (Patro et al., 2014)

### K-meride abil sekveneerimise katvuse määramine

Suure katvusega genoomi puhul on võimalik sekveneerimise katvust hinnata k-meride hulkade jaotumise järgi (Joonis 1).



**Joonis 3** X-teljele on seatud hulgad – mitu korda identseid k-mere esineb, y-teljel on toodud osakaal, mitu protsenti selle sagedusega esinevad k-merid moodustavad kõikidest k-meridest. Vasakpoolse piigi moodustavad sekveneerimisvead. Parempoolne piik näitab k-meride katvust. (Lamichhaney et al., 2015)

### TODO: GenomeTester4

GenomeTester4 tarkvarakomplekt koosneb kolmest k-mer andmete töötlemise tööriistast – GListMaker, GListCompare ja GListQuery. Programmid salvestavad andmed binaarformaadis failidesse, kus k-merid on kodeeritud 64 bitisteks märgita täisarvudeks ning k-meride arvud on 32 bitised märgita täisarvud. Tarkvarapaketi failidesse salvestatakse ainult k-meri kanooniline vorm – üks sama täisarv tähistab nii järjestust kui selle pöördkomplementi. Kumba versiooni k-merist salvestatakse, otsustatakse selle põhjal, kumma täisarv-väärtus väiksem on. (Kaplinski, Lepamets, & Remm, 2015)

Selleks, et GenomeTester4 tööriistu kasutada, tuleb nukleotiidjärjestustega FASTA failist teha k-mer tabel. Selle jaoks on tööriist GListMaker, mis valmistab eelnevalt kirjeldatu list-faili. (Kaplinski et al., 2015)

# Eksperimentaalosa

## Töö eesmärgid

Selle töö eesmärgid on:

1. analüüsida ja arendada meetod suure hulga sekveneerimisandmete töötlemiseks
2. leida lihtne viis k-meride abil katvuse määramiseks
3. määrata teise generatsiooni madala katvusega sekveneerimisandmete kogu põhjal keskmised telomeeride pikkused

Selle töö peamine eesmärk on **määrata teise generatsiooni madala katvusega sekveneerimisandmete kogu põhjal keskmised telomeeride pikkused** (E3). Eelnevalt kirjeldatud telomeeride pikkuse määramise keemilised meetodid on aeganõudvad, inimtööjõust sõltuvad ning veaohtlikud. Samas peetakse telomeeride pikkust näitajaks, mis võib abiks olla mitmete haiguste ja tervislike seisundite määramisel ja hindamisel. Ka muutub terve genoomi sekveneerimine iga aastaga odavamaks. Nendest teadmistest johtuvalt on sekveneerimisandmetest telomeeride pikkuse määramiseks vajalik käepäraste, kiirete ja täpsete meetodite olemasolu.

Kuna telomeeride pikkuse hindamiseks on vaja sekveneerimisandmete katvust, on töö kõrvaleesmärgiks **leida lihtne viis k-meride abil katvuse määramiseks** (E2).

Lisaks, kuna meetodite testimiseks kasutatakse suurt hulka, üle 50 TB, andmeid, mille **allalaadimine ja töötlemine** on mittetriviaalne ülesanne, on töö eesmärgiks ka töö **jaoks sobiva rakenduse analüüs ja arendamine** (E1).

## Metoodika ja materjalid

### Arenduskeskkond ja tööriistad

Tänapäevane tarkvaraarendus peaks probleemivaldkonnast sõltumatult soodustama nii gruppide sisest kui -vahelist koostööd ning olema platvormi-agnostiline. Töövahendid valiti sellest printsiibist lähtuvalt. Kogu töö käigus valminud lähtekood on saadaval aadressil <https://github.com/karlerss/telomere-length>.

Kuna võimalustele vastavalt oli arenduskeskkonnaks Windows-arvuti, aga terve analüüsi teostamise keskkonnaks 64 bitine linux-arvuti, viidi keskkondade erinevusest tekkivate probleemide vältimiseks kogu arendus läbi Docker-konteineris (<https://github.com/karlerss/telomere-length/blob/master/Dockerfile>). Alussüsteemiks kasutati biocontainers/biocontainers tõmmise uusimat versiooni. SraTools 2.8.1 lisati docker-konteinerisse conda pakihaldussüsteemi abil. Python programmide toimimiseks vajalikud teegid lisati pip pakihalduri abil. GenomeTester4 binaarfailid kopeeriti käesoleva töö repositooriumisse ning lisati sealt docker-konteinerisse.

Andmete laadimisprogrammi käsurealiidese loomiseks kasutati click-teeki, mis hoolitseb käsureaargumentide ning seadete sõelumise ja valideerimise eest. (<http://click.pocoo.org/5/>)

Andmete agregeerimine viidi läbi CentOS 7 arvutis, millel on 32 protsessorituuma ning 512GB muutmälu.

Huvipakkuvate k-meride sageduste analüüs viidi läbi Jupyter interaktiivses keskkonnas (<http://jupyter.org/>). Tulemused visualiseeriti python teegi matplotlib abil (<https://matplotlib.org/>).

### TODO: K-meride valik

### TODO: Andmestik

## Tulemused

### Sekveneerimisandmete analüüsi meetod

K-mer metoodika abil uurimiseesmärkide saavutamiseks peavad ühe inimese andmed läbima järgnevad sammud (sulgudes välise programmi nimi):

1. Andmete allalaadimine internetist (prefetch)
2. Allalaetud andmete viimine fasta-formaati (fastq-dump)
3. *Read*ide fasta failist k-meride sagedustabeli tegemine (glistmaker)
4. Sagedustabelist huvipakkuvate k-meride sageduse päringu tegemine (glistquery)
5. Ühe inimese huvipakkuvate k-meride sageduste salvestamine edasiseks analüüsiks
6. Kettaruumi vabastamine

Lisaks analüüsi enda sammudele on oluline iga sammu logimine ning protsessi paralleliseeritavus. Ka on tähtis, et programm lõpetakse ülesande nii vähese ajaga kui võimalik. Selleks, et leida, milline arvutusressurss (allalaadimiskiirus, protsessorituumad või muutmälu) saab piiravaks, jooksutati käsurealt iga kasutatavat välist programmi. Tulemused tabelis Tabel 1.

Tabel 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Programm | Maks. mälukasutus | Väljundi suurus kettal | Kasutatud tuumasid | Kulunud aeg | Ressurss |
| prefetch | <1 GB | 2 GB | 1 | ~3 min | Võrk |
| fastq-dump | 1 GB | 5 GB | 1 | ~9 min | HDD |
| glistmaker | 35 GB | 25 GB | 2 | ~9 min | RAM |
| glistquery | <1 GB | ~300 KB | 1 | ~1s | - |

Nendest tulemustest selgub, et ainus samm, mis kasutab kogu saadaolevat ressurssi, on prefetch. Ka on näha, et allalaadimiskiirusest mittesõltuvad sammud võtavad umbes kuus korda nii palju aega, kui ühe uue lähtefaili laadimiseks kulub. Sellest johtuvalt ning et ülesanne lahendada maksimaalse kiirusega, peaks ühe allalaadimise kohta jooksma paralleelselt vähemalt kuus protsessi, kus toimuvad lokaalsed arvutustööd.

#### get\_srr.py

Eelnevaid asjaolusid arvesse võttes, arendati tööriist get\_srr.py. Programmi kasutamiseks tuleb käsureale kirjutada *python get\_srr.py [OPTIONS] RUN\_TABLE QUERY\_LIST*, kus *RUN\_TABLE* on SRA veebiliidesest saadud kokkuvõttefaili asukoht kõvakettal, näiteks ./SraRuntable.txt ning *QUERY\_LIST* on analoogselt tekstifail, kus on igal real üks k-mer. Lisaks on võimalik seadistada muid parameetreid *[OPTIONS]*, mis on toodud tabelis Tabel 2.

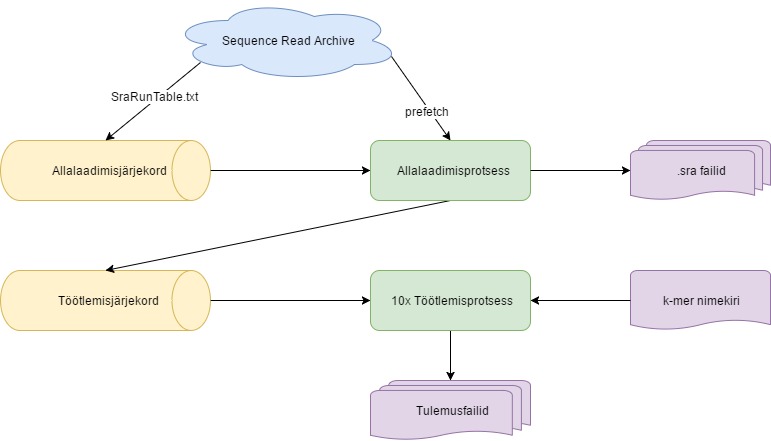
**Tabel 2** get\_srr.py lisavalikud.

|  |  |  |  |
| --- | --- | --- | --- |
| Lühike nimi | Pikk nimi | Tüüp | Kommentaar |
| -d | --data\_root | Rada | Kataloog, kuhu sisse tehakse ajutised kaustad ja väljundite kaust. Vaikimisi /data. |
| -n | --ncbi\_root | Rada | Kataloog, kus asuvad ncbi tööriistade tekitatud failid. Selle raja leiab linux keskkonnas tavaliselt failist ~/.ncbi/user-settings.mkfg. Vaikimisi /data/.ncbi. |
|  | --check\_fasta | Tõeväärtus | Lisavalik, mille lubamisel kontrollitakse enne prefetch käivitamist, kas on fasta fail juba genereeritud. Silumiseks/testimiseks. |
|  | --fasta\_limit | Täisarv | Lisavalik, millega saab piirata fastq-dump väljundi pikkust. Silumiseks/testimiseks. |
| -p | --processing\_cores | Täisarv | Valik, mis määrab, mitu arvutamisprotsessi allalaadimisega paralleelselt jooksutatakse. Vaikimisi 1. |
|  | --help |  | Kuvab abidokumendi. |

Programmi töö algab kahe riviloendi (multiprocessing.JoinableQueue) loomisega. Allalaadimisrivi on piiramata mahuga ning töötlemisrivi mahutab kuni 10 tööd. Töötlemisrivi suurus on piiratud, et vältida liiga suure hulga toorandmete ettelaadimist ja mäluseadme täitumist – kui töötlemisrivi saab täis, siis allalaadimisprotsess ootab enne uue laadimise alustamist, kuni töötlemisrivis tekib vaba koht. Selline riviloenditel põhinev tarkvara arhitektuur on kasutatav ka hajussüsteemi korral.

Programmi ühe katse andmete töötlemise loogika on koondatud *Job*-klassi. Klassi konstruktor võtab ühe argumendi, milleks on SRA identifikaator formaadis SRR0000000. Klassi isendil on meetodid *fetch()* ning *process()*.

Pärast riviloendite loomist sõelutakse *RUN\_TABLE* fail ning luuakse saadud SRA identifikaatorite põhjal Job isendid ning pannakse see allalaadimis-riviloendisse. Seejärel luuakse üks allalaadimisprotsess, mis kutsub allalaadimisrivist saadud isendil *fetch*-meetodit ning lisavalikutes määratud suurusega protsessi-*pool* (multiprocessing.Pool) töötlemisprotsessidega, mis kutsuvad töötlemisjärjekorrast saadud isenditel *process*-meetodit.



Joonis 4 get\_srr.py ülesehituse loogika.

TODO: kaua jooksis ja kuidas läks

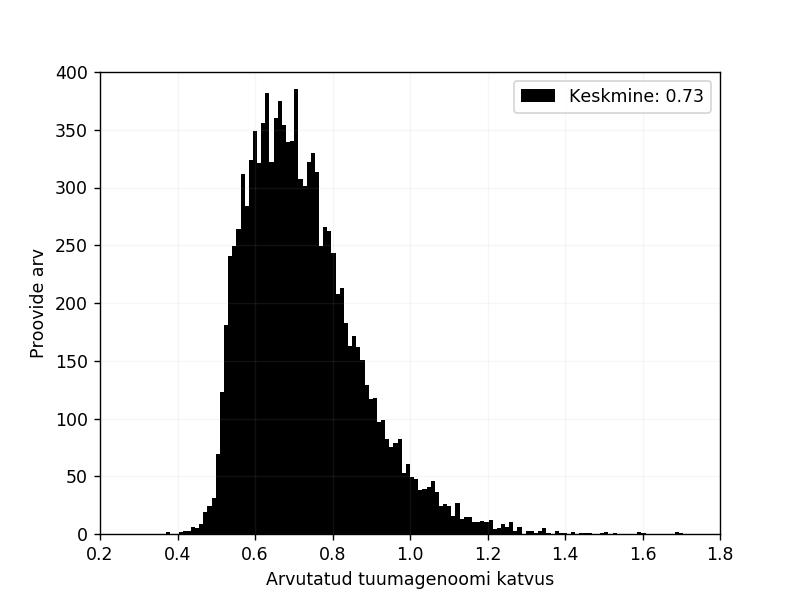
### Katvuse määramise meetod

Selleks, et luua lihtne k-meride põhine meetod sekveneerimiskatvuse leidmiseks, on esmalt vaja viisi uue meetodi hindamiseks. Katsetamiseks valitud andmestiku puhul saame kontrolliks kasutada näitajaid: publikatsioonis toodud keskmine katvus 1,7X, publikatsioonis näidatud katvuste jaotuse võrdlemine ning SRA andmebaasist saadud sekveneeritud nukleotiidide koguarv.

PCR ja HPA meetodiga telomeeride pikkuse määramiseks kasutatakse telomeeri pikkuse kvantifitseerimiseks võrldust mõne geenijupiga, mille koopianumber on teada. Analoogset lähenemist saaks kasutada ka sekveneerimiskatvuse määramiseks.

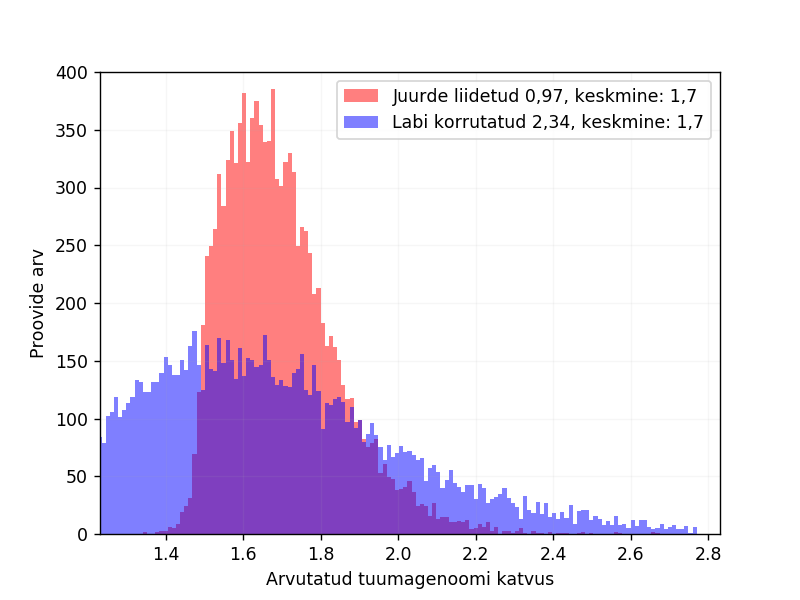
Selle jaoks valiti kõrge koopiaarvuga k-mer, sama mida kasutati eelnevalt kirjeldatud HPA meetodis (CTGTAATCCCAGCACTTTGGGAGGC). Selleks, et vältida võimalikke populatsioonide erinevusest tulenevaid ebatäpsusi, valiti koopiaarvu määramiseks inimene Han hiinlaste seast HGDP00778. Selle genoomi katvus määrati eelnevalt näidatud mediaankatvuse meetodil ning saadi tulemuseks 22. 22 kordse katvusega andmetes esines valitud k-meri 5273596 korda. See tähendab, et 1 kordse katvusega sekveneeritud genoomis esineks seda k-meri 5273596/22 = 239708 korda.

Sellest lähtuvalt uuriti CONVERGE andmestiku sekveneerimisandmetes sama k-meri esinemist. Selgus, et keskmiselt esineb valitud k-meri sekveneerimisandmetes 174265 korda, mis teeks keskmiseks arvutatud katvuseks 174265/239708 = 0,73. See tulemus erineb CONVERGE andmestiku artiklis toodud 1,7-st märgatavalt. Siiski on jooniselt (Joonis 5) näha, et arvutatud katvuste jaotus sarnaneb kirjanduses (Joonis2) tooduga.



Joonis 5

Kuigi esmane katvuse määramise meetod ei osutunud sobivaks, on telomeeride pikkuse määramiseks siiski katvust vaja. Kuna esmase tulemuse arvutatud katvuste jaotus sarnaneb väga andmekogu artiklis toodule, on võimalus arvutamise meetodit *ad-hoc* korrigeerida nii, et tulemus sarnaneks artiklis esitatud tulemusele. Selle jaoks prooviti kahte võimalust – leida kordaja, millega katvusi läbi korrutades saadakse keskmiseks katvuseks 1,7 ning liita arvutatud katvusele puudujääv osa juurde.



Joonis 6

Jooniselt (Joonis 6) on näha, puudujääva osa juurde liitmisel saadav jaotus on kirjanduses toodule palju sarnasem kui kordajaga läbi korrutamisel saadav jaotus.

### Keskmised telomeeride pikkused

Telomeeride k-meri (TTAGGGTTAGGGTTAGGGTTAGGGT) arvude uurimisel selgus, et andmestiku siseselt ei ole telomeeri k-mere ühepalju.

## Arutelu

### TODO: Sekveneerimisandmete analüüsi meetod

### TODO: Katvuse määramise meetod

* Vaja paremat meetodit
* 0.97 ~= 1?
* Erinev arusaam katvusest?
* Read’ide filtreerimise tulemus?
* Kas saadud tulemus sobib järgmise sammu jaoks?

### TODO: keskmised telomeeride pikkused

# Kokkuvõte

# Resümee

# Kasutatud kirjanduse loetelu

# Kasutatud veebiaadressid