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 MSc Tarmo Puurand

TARTU 2017

Infoleht

Telomeeride keskmise pikkuse hindamine teise generatsiooni sekveneerimisandmetest

Telomeerid on korduvad järjestused kromosoomide otstes, mis kaitsevad kromosoome. Telomeeride pikkusega on seostatud mitmesuguseid haigusi ning inimese üldist tervislikku seisundit. Teise põlvkonna sekveneerimise tehnoloogiate hinna alanemine on viinud tekitatavate andmemahtude hüppelise suurenemiseni. Seetõttu on oluline nende andmete töötlemiseks kiirete ja üldiste meetodite olemasolu.

Uurimistöö eesmärk on leida meetod ja arendada vahendid teise põlvkonna sekveneerimisandmetest telomeeride pikkuse määramiseks. Eesmärgi saavutamiseks kastutatakse k-mer põhist metoodikat. Töö käigus arendati välja tööriist suure hulga sekveneerimisandmete töötlemiseks ning esitleti k-mer põhist meetodit sekveneerimiskatvuse ja telomeeride keskmise pikkuse määramiseks.

Märksõnad: telomeeride pikkus, Alu-elemendid, k-mer metoodika

CERCS kood: B110

Estimation of telomere length from next generation sequencing data

Telomeres are nucleoprotein structures that protect the ends of chromosome from fusion and degradation. Telomere length has been associated with several diseases and has been considered a marker for general health and aging. The explosion in quantities of next generation sequencing data introduces the need for fast and generic methods for data analysis.

A script which handled retrieval of data from Sequnece Read Archive and reducing it into lists of ~3000 k-mers was developed. The program can be used for similar future studies. The efficacy of the presented k-mer based methods of estimating sequencing coverage and telomere length was inconclusive.

Keywords: telomere length, Alu-elements, k-mer counting

CERCS code: B110

Sisukord

Kasutatud lühendid

bp — Ühik, aluspaari (*base pairs*)

FASTA — Failiformaat nukleotiidjärjestuse salvestamiseks

FASTQ — Failiformaat nukleotiidjärjestuse salvestamiseks koos sekveneerimise kvaliteediskooridega

HPA — Hübridisatsiooni-kaitse meetod (*hybridization protection assay*)

LINE — Pikad insertsioonilised hajuskorduselemendid (*long interspersed nuclear elements*)

NGS — Järgmise/teise põlvkonna sekveneerimine

qPCR — Kvantitatiivne polümeraasi ahelreaktsioon

SRA — *Sequence Read Archive*

STELA — Üksiku telomeeri pikkuse analüüs (*single telomere length analysis*)

TERT — Telomeraasi pöördtranskriptaas

TRF — Terminaalsete restriktsioonifragmentide analüüs (*terminal restriction fragment*)

Sissejuhatus

Telomeerid on korduvad nukleotiidjärjestused kromosoomide otstes, mis kaitsevad kromosoome. Telomeerid jäävad paljudes kudedes iga raku jagunemisega lühemaks. Telomeeride pikkusega on seostatud mitmesuguseid haigusi ning inimese üldist tervislikku seisundit.

Teise põlvkonna sekveneerimise tehnoloogiad on viinud genoomi sekveneerimise hinna väga madalaks. See omakorda on viinud tekitatavate andmemahtude hüppelise suurenemiseni. Seetõttu on oluline nende andmete töötlemiseks kiirete ja üldiste meetodite olemasolu.

Uurimistöö eesmärk on leida meetod ja arendada vahendid madala katvusega teise põlvkonna sekveneerimisandmetest telomeeride pikkuse määramiseks. Eesmärgi saavutamiseks kastutatakse k-mer põhist metoodikat. Kuna töös kasutatakse madala sekveneerimiskatvusega andmeid, määratakse sekveneerimiskatvus Alu-elementide abil, millel on suur koopiaarv ning koopiaarv varieerub inimeste vahel vähe.

1 Kirjanduse ülevaade

Käesoleva uurimistöö praktiline osa toetub mitmete eri valdkondade teadmistele – bioloogiale, biotehnoloogiale ning bioinformaatikale. See peatükk tutvustab telomeeride bioloogilist funktsiooni; telomeeride pikkuse mõõtmise olulisust ja meetodeid; teise põlvkonna sekveneerimist ja sekveneerimisandmeid; ning k-mer metoodikat NGS andmete analüüsis.

1.1 Telomeerid

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

1.1.1 Struktuur

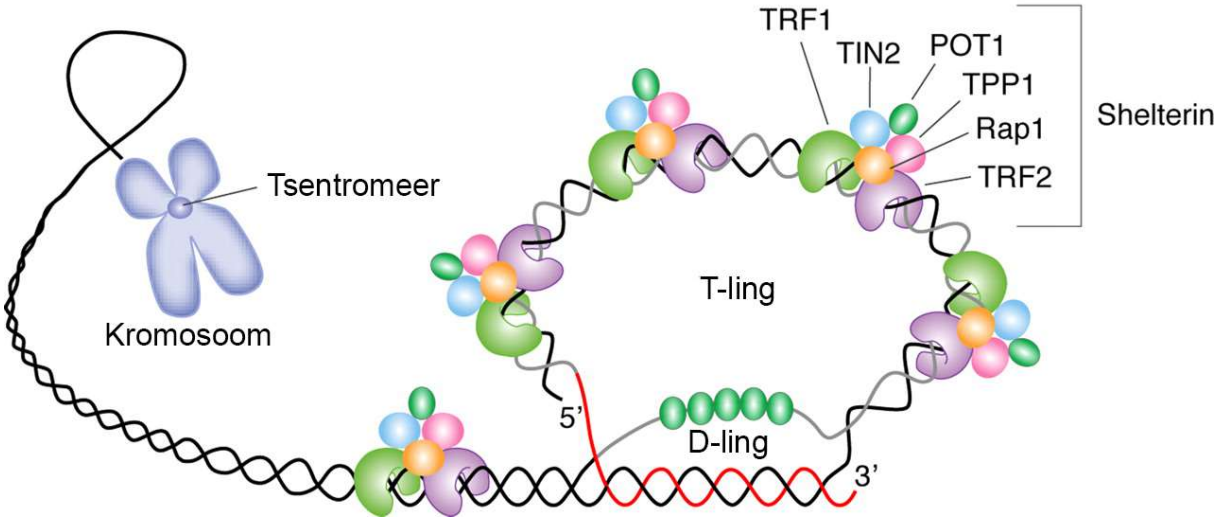
Imetajate telomeerid koosnevad TTAGGG kordustest ja valgukompleksist nimega *shelterin*. (Martínez ja Blasco, 2010)

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 ja Blasco, 2010) Telomeeri ja telomeeri otsa struktuur on välja toodud joonisel (Joonis 1).

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 DNAle. kusjuures viimane ainult üleulatuvale ahelale. TIN2 seondub ühe TRF valguga ja on vajalik TPP1-POT1 kompleksi seondumiseks.



Joonis 1 Telomeeri struktuuri skeem. Telomeerid asuvad kromosoomide otstes. Telomeeride 3' ots on üheahelaline üleulatuv osa, mis tungib telomeeri topeltheeliksi vahele ja annab DNA-le lingu kuju. Telomeeri valkude kompleksi (TRF1, TRF2, TPP1, POT1, TIN2 ja Rap1) nimetatakse *shelterin*iks. *Shelterin* takistab DNA-kahjustuste parandamise mehhanismide käivitumist. (Calado ja Young, 2008)

1.1.2 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 polümeraas 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. See põhjustab telomeeri järk-järgulise lühenemise, mida nimetatakse otsa replikatsiooni probleemiks. (Witzany, 2008)

1.1.3 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 et al., 2009)

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 ja Barrett, 2003)

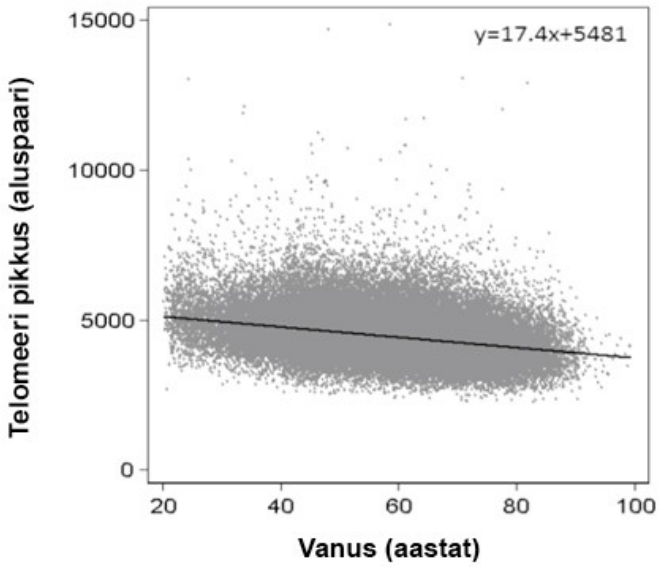
1.1.4 Telomeeride pikkus

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 et al., 2011)

Telomeeride puhul on leitud, et telomeeride pikkus ja inimese vanus on negatiivselt korrelatsioonis (Joonis 2). 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 seas 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 2 Telomeeride pikkuse sõltuvus vanusest. Perifeersete leukotsüütide telomeeride pikkus vanuse funktsioonina. Graafik põhineb 64637 inimese vereproovi analüüsil. Lineaarne regressioon näha musta joonena. Regressioonivalem y=-17.4x+5481 näitab, et keskmiselt väheneb telomeeri pikkus 17.4 bp võrra aastas. (Rode et al., 2015)

1.1.5 Telomeeride pikkust mõjutavad faktorid

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

1.1.6 Telomeeride pikkuse hindamine

Tabel 1 loetleb nii arvutuslikke kui keemilisi meetodeid telomeeride pikkuse määramiseks. Uurimistöö praktilises osas pakutakse välja keemilise HPA meetodiga analoogne arvutuslik meetod telomeeride keskmise pikkuse määramiseks.

Tabel 1 Telomeeride pikkuste hindamise meetodite ülevaade

Nimi	**Sisend ja põhimõte**	**Resolutsioon (kb)**	**Allikas**
TRF	DNA molekulide lõikamine ja visualiseerimine southern blotiga	1	(Kimura et al., 2010)
qPCR	PCR reaktsioonil fluorestsentsi mõõtmine	-	(Cawthon, 2009)

STELA	qPCR kromosoomi-spetsiifiliste praimeritega	0,1	(Baird et al., 2003)
HPA	Telomeerile ja Alu-järjestusele seonduvate fluorestseeruvate märgiste intensiivsuste võrdlus	-	(Nakamura et al., 1999)
TelSeq	NGS-andmetest TTAGGG kordust sisaldavate *read*ide lugemine	2,5-4	(Ding et al., 2014)
Computel	NGS-andmetest TTAGGG sisaldavate *read*ide lugemine koos sekveneerimisvigadega arvestamisega	2-3	(Nersisyan ja Arakelyan, 2015)

1.1.6.1 Terminaalsete restriktsioonifragmentide analüüs - TRF

Selle meetodi kasutamise jaoks on vaja vähemalt 3 µg 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 geelelektroforeesil 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 lõigatakse restriktsiooniensüümidega (*Hinf*I ja *Rsa*I), 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)

1.1.6.2 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 paljundatakse 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)

1.1.6.3 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)

1.1.6.4 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. Telomeerselt 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)

1.1.6.5 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 et al., 2014)

1.1.6.6 TelSeq

TelSeq tarkvara kasutab sisendiks BAM faili. TelSeq loeb kokku telomeerseid järjestusi sisaldavad *read*id ning hindab telomeeri füüsilist pikkust valemi

$$l = t_k sc$$

järgi, kus l on hinnatav pikkus, t_k 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 *read*ide arv, kus GC sisaldus on 48% kuni 52%. (Ding et al., 2014)

1.1.6.7 Computel

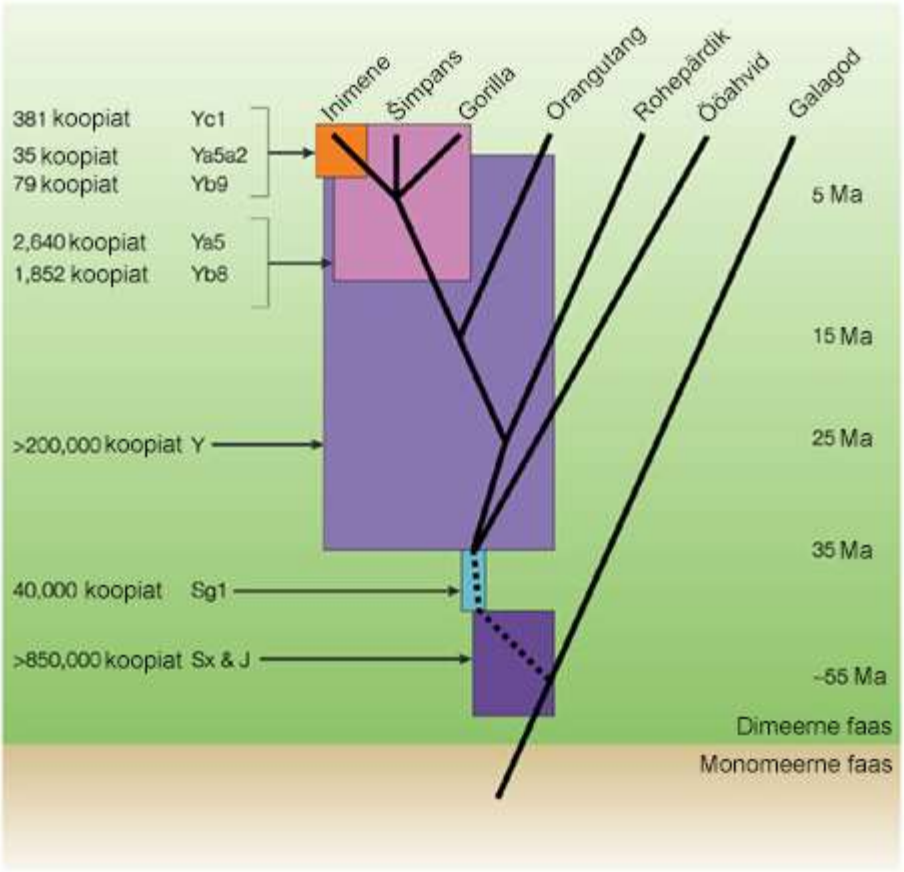
Computel tarkvara joondab *read*id telomeerse järjestuse indeksiga. Selleks, et kõik telomeersed *read*id arvutusel arvesse võetaks, koostatakse indeks nii, et loetaks ka neid *read*e, mis sisaldavad subtelomeerseid osi. Ka arvestab programm telomeersete järjestuste joondamisel sekveneerimsvigadega. Keskmine telomeeride pikkus arvutatakse valemiga

$$l_{MTL} = \overline{c_{rel}} \frac{l_r + l_p - 1}{2n_{chr}}$$

kus c_{rel} on baaskatvuse (*read*ide arv) * (*read*i pikkus) / (genoomi pikkus) ning telomeerse katvuse (mitu *read*i indeksiga joondati) suhe; l_r on *read*i pikkus, l_p on indeksi telomeerse osa pikkus ning n_{chr} on kromosoomide arv. (Nersisyan ja Arakelyan, 2015)

1.2 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 retroposoonsed. (Roy-Engel et al., 2001)



Joonis 3 Alu elementide koopiate areng. *Alu elemendid on arenenud viimase 65 miljoni aastaga primaatide genoomis. Aja jooksul on eristunud erinevad Alu-elementide perekonnad.* (Batzer ja Deininger, 2002)

Eristatakse ~5000 Alu-elementi, mis on inimese genoomi integreerunud viimase 4-6 miljoni aasta jooksul. Enamus neist integreerus inimese genoomi enne Aafrikast välja rändamist. Umbes 1200 Alu kordust on genoomi sisenenud küllalt hiljutisel ajal, et täpne inserteerumismuster erineb populatsiooniti. Populatsioonisiseselt on Alu-järjestuste varieerumine väga väike. (Batzer ja Deininger, 2002)

Et Alu kordusi esineb genoomis üle miljoni ja need on ka üle genoomi küllalt ühtlaselt jaotunud, kasutatakse töö praktilises osas Alu-elementide k-meride põhjal detekteeritud arvu sekveneerimiskatvuse ning telomeeri pikkuse hindamiseks. Alu elementide suur koopiaarv peaks vähendama madalast sekveneerimiskatvusest tulenevat ebatäpsust.

1.3 Pikad insertsioonilised hajuskorduselemendid LINE1

LINE elemendid on umbes 6kb pikad, moodustavad genoomist umbes 21% ning sisaldavad polümeraas II promootorit ning kodeerivad kahte avatud lugemisraami. Transleerumisel

kombineerub LINE RNA enda kodeeritud valguga ning liigub tuuma, kus teeb genoomi umbes 1000bp pikkuse sisestuse. LINEdest on aktiivsed veel vaid LINE1 elemendid. LINE1 elemendid on eukarüoodi genoomis juba 150 miljonit aastat. Inimese (ja eellaste) genoomis on LINE elementide aktiivsus viimase 35-50 miljoni aasta vältel langenud. LINE1 elemendid katavad inimese genoomi enamustes kromosoomides ühtlaselt ning X ja Y kromosoomides esinevad AT-rikastes regioonides sagedamini. (Lander et al., 2001)

Selle uurimistöö praktilises osas kasutatakse üht LINE1 elemendile iseloomulikku k-meri selleks, et võrrelda telomeeri-Alu esinemise suhet Alu-LINE1 esinemise suhtega. Kuna töös kasutatakse madala sekveneerimskatvusega andmeid, parandab Alu elementide suur koopiaarv katvuse arvutamise täpsust.

1.4 Teise põlvkonna sekveneerimine

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 et al., 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 detekteerimiseks 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 tehakse plaadist pilt ja loetakse fluorestsentssignaalid. (Muzzey et al., 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)

1.4.1 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 ja 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 et al., 2016) Levinuim viis kvaliteedi esitamiseks on Phred kvaliteediskoor $Q=-10log_{10}P$, 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 et al., 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)

Ka selles töös on kasutatud Illumina sekveneerimismasinast pärinevaid andmeid.

1.4.2 Sekveneerimiskatvus

Sekveneerimiskatvuse ehk katvuse all mõeldakse aluspaaride 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. Inimese genoomi kõrge kvaliteediga sekveneerimiseks peab sekveneerimiskatvus olema suurem kui 25. (Muzzey et al., 2015) Käesolevas töös kasutatud andmestik koosneb aga madala katvusega (1,7) sekveneerimisandmetest.

1.5 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 kogumise võimekus ületab analüüsivõimekuse. Samuti on probleemiks järjestused, mis sobivad referentsile mitmesse kohta. (Patro et al., 2013)

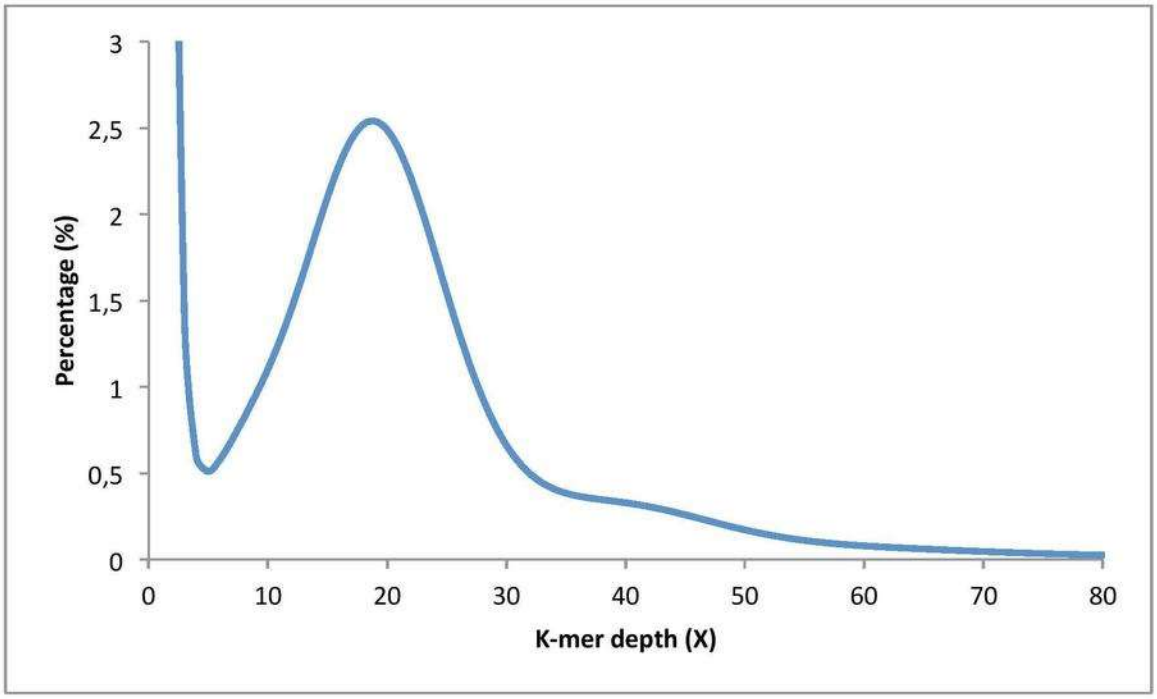
Mõiste k-mer viitab kõikidele kindla pikkusega osasõnedele, mida sõne sisaldab. K on osasõne pikkus. (Compeau et al., 2011) Näiteks sõne GTAGAGCTGT 5-merid on GTAGA, TAGAG, AGAGC, GAGCT, AGCTG ja GCTGT.

K-mer põhilised analüüsimeetodid põhinevad sekveneerimisandmetest 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., 2013)

Selles töös kasutatakse genoomide uurimiseks just k-mer metoodikat. Kasutatud on telomeeridele (TTAGGGTTAGGGTTAGGGTTAGGGT), ühele Alu-elemendile ja LINE1 elemendile omase k-meri loendamist.

1.5.1 K-meride abil sekveneerimise katvuse määramine

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



K	k-mer count	Peak depth	Genome size (bp)	Used bases (bp)	Used reads	X
17	23,288,894,374	19	1,225,731,282	27,833,068,886	284,010,907	22.71

Joonis 4 K-meride abil sekveneerimise katvuse määramine *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 sekveneerimiskatvust.* (Lamichhaney et al., 2015)

2 Eksperimentaalosa

2.1 Töö eesmärgid

Selle töö eesmärgid on:

1. analüüsida ja arendada meetod suure hulga (100TB) 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**. 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**.

Lisaks, kuna meetodite testimiseks kasutatakse suurt hulka, ligi 100 TB, andmeid, mille **allalaadimine ja töötlemine** on mittetriviaalne ülesanne, on töö eesmärgiks ka töö **jaoks sobiva rakenduse analüüs ja arendamine**.

2.2 Metoodika ja materjalid

2.2.1 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 kõige uuemat 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/).

K-meride lugemiseks ja analüüsimiseks kasutati GenomeTester4 komplekti tööriistu.

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 et al., 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)

GListQuery tööriista abil loetakse genereeritud binaarfailist kokku loetud k-meride arvud. Programmi sisendiks määratakse kas üks k-mer järjestus või tekstifail huvipakkuvate k-mer järjestustega. (Kaplinski et al., 2015)

HGDP genoomi katvus määrati käsu *glistquery HGDP00778_25_intrsec.list | awk '{print $2}' | sort -n | uniq -c* abil. Graafik visualiseeriti Microsoft Excel tarkvaraga ning loeti katvuse-piik.

2.2.2 K-meride valik

K-meride nimekirja faili valiti nelja tüüpi k-mere. Selleks, et uurida telomeeride pikkust, on esimesel real telomeeri k-mer. Võimalikeks negatiivseteks kontrollideks ja eksploratiivsel eesmärgil valiti ridadele 2-1319 juhuslik valik mitmesugustest k-meridest, mis esinevad erinevates geenides. Samas järjekorras k-meride nimekiri koos geenide ENSG-identifikaatoritega on kättesaadav aadressil https://github.com/karlerss/telomere-length/blob/master/tlenpy/kmer_sample.txt. Selleks, et määrata analoogselt HPA meetodiga sekveneerimiskatvust ja telomeeri pikkust, on ridadel 1320-1583 erinevatele Alu-elementidele omaste k-meride nimekiri, mis on eelnevalt uurimisgrupis koostatud. Ridadel 1584-3083 on mitmesugused telomeeri pikkusega assotsieeritud SNP-sid sisaldavad k-merid, mis selles töös genoomide madala katvuse tõttu kasutust ei leia. Terve k-meride nimekiri on kättesaadav aadressil https://github.com/karlerss/telomere-length/blob/master/tlenpy/kmer_sample_min.txt.

2.2.3 Andmestik

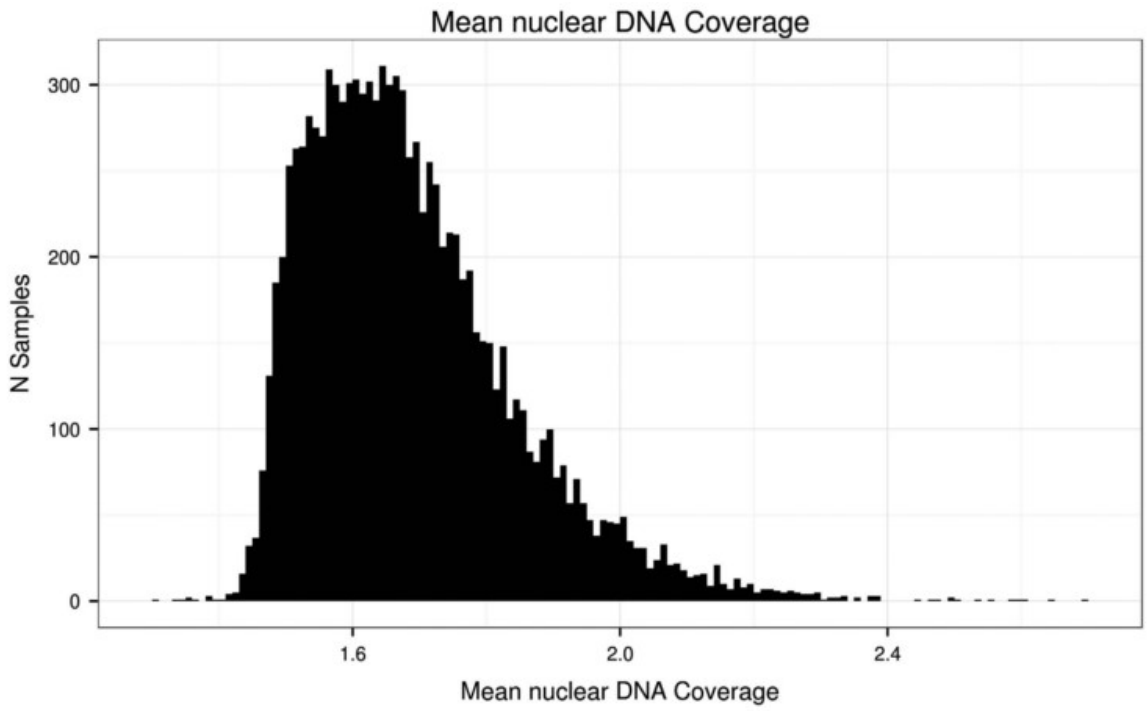
2.2.3.1 Sequence Read Archive

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 et al., 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)

2.2.3.2 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 5 CONVERGE andmestiku tuumagenoomi katvus. *Keskmine katvus on 1.7X.*

Töö jaoks kasutati tutvustatud CONVERGE andmekogu. Sekveneerimisandmetest eraldati valitud k-meride arvud get_srr.py programmi abil, mis väljastas iga genoomi kohta ühe tekstifaili, kus on igal real tabeldusmärgiga eraldatud k-meri järjestus ning esinemiskordade arv.

2.2.4 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. Prefetch salvestatud lähteandmete arvelt kettaruumi vabastamine

Lisaks analüüsi enda sammudele on oluline iga sammu logimine ning protsessi paralleliseeritavus. Ka on tähtis, et programm lõpetaks ülesande nii vähese ajaga kui võimalik. Selleks, et leida, milline arvutusressurss (allalaadimiskiirus, protsessorituumad või muutmälu) saab piiravaks, jooksutati käsurealt iga kasutatavat välist programmi. Tulemused on tabelis Tabel 3 Tööriistade analüüsi tulemused.

2.2.4.1 get_srr.py

Programmide analüüsi 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.

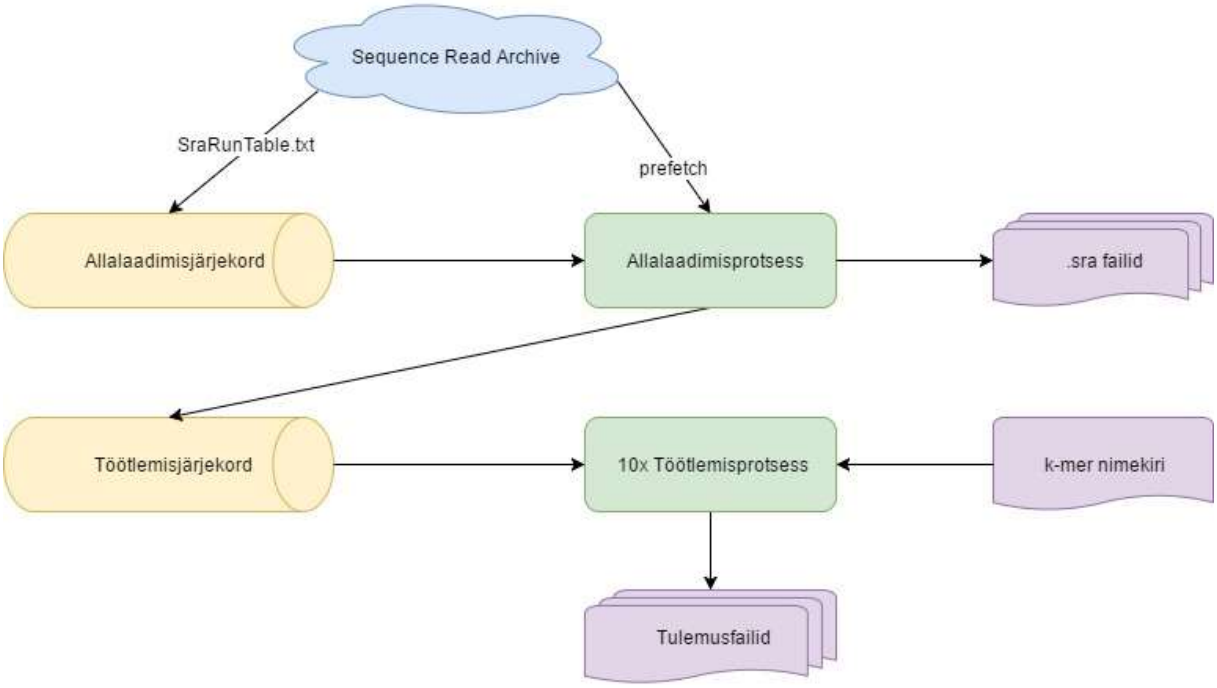
-n	--ncbi_root	Rada	Kataloog, kus asuvad ncbi tööriistade tekitatud failid. Selle raja leiab linux keskkonnas tavaliselt failist ~/.ncbi/user-settings.mkfg. Vaikimisi /data/.ncbi.
	--check_fasta	Tõeväärtus	Lisavalik, mille lubamisel kontrollitakse enne prefetch käivitamist, kas on fasta fail juba genereeritud. 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. Kui *fetch*-meetodis esineb erind, pannakse see isend rivi lõppu.



Joonis 6 get_srr.py ülesehituse loogika. *Käsitsi alla laetud SraRunTable.txt faili põhjal käivitatakse paralleelselt üks allalaadimisprotsess ning 10 töötlemisprotsessi. Ahela väljundiks on tulemusfailid k-meride arvudega.*

2.2.5 Katvuse määramise meetod

Selleks, et luua lihtne k-meride põhine meetod sekveneerimiskatvuse leidmiseks, on esmalt vaja viisi uue meetodi hindamiseks. Katsetamiseks valitud andmestiku puhul saab 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, sarnane sellele mida kasutati eelnevalt kirjeldatud HPA meetodis (CTGTAATCCCAGCACTTTGGGAGGC). Kuna HPA meetodi artiklis toodud Alu-oligonukleotiid oli 24 bp pikkune, kuid töös on k-väärtuseks kasutatud 25, lisati artiklis toodud järjestusele C-nukleotiid nii nagu see Alu elemendis ka esineb. Selleks, et vältida võimalikke populatsioonide erinevusest tulenevaid ebatäpsusi, valiti Alu elemendi koopiaarvu

määramiseks inimene Human Genome Diversity Project'i Han hiinlaste seast HGDP00778. Selle genoomi katvus määrati eelnevalt näidatud mediaankatvuse meetodil ning saadi tulemuseks 23. 23 kordse katvusega andmetes esines valitud k-meri 5273596 korda. See tähendab, et 1 kordse katvusega sekveneeritud genoomis esineks seda k-meri 5273596/23 = 229286 korda.

Sellest lähtuvalt uuriti CONVERGE andmestiku sekveneerimisandmetes valitud k-meri esinemist. Selgus, et keskmiselt esineb valitud k-meri sekveneerimisandmetes 174265 korda, mis teeks keskmiseks arvutatud katvuseks 174265/229286 = 0,76. See tulemus erineb CONVERGE andmestiku artiklis toodud 1,7-st märgatavalt. Siiski on jooniselt (Joonis 7, punane histogramm) näha, et arvutatud katvuste jaotus sarnaneb kirjanduses tooduga.

K-meride põhjal katvuse määramisel tuleb aga arvestada, et iga pikkusega *L read*i kohta saab *L-k+1* k-meri. Näiteks 10000 bp pikkuse genoomi 1X katvusega 100bp *read*idega sekveneerides saaks 100 *read*i. Kui tekitada ja lugeda kokku nende readide k-merid, saaks igalt readilt 100-25+1=76 k-meri ja kokku 7600 k-meri. 10000bp-se genoomi kataks aga 10000-25+1 = 9976 k-meriga. Kuna Alu-elemente mõõdetakse samuti samas vahekorras, jagati Alu-elementide põhjal määratud k-mer katvus läbi väärtusega (*L-k+1)/L*. Keskmisest k-meri katvusest saadi selle abil 0,77/((83-25+1)/83) = 1,07. Ka see tulemus erineb CONVERGE andmekogu artiklis toodud 1,7-st. Nende tulemuste jaotus on histogrammil (Joonis 7) sinine.

Kuna andmekogu SRA sissekande metaandmestikus on toodud ka sekveneeritud aluspaaride koguarv, saab seda kasutada k-meri põhise katvuse hindamise meetodite hindamiseks. Selle jaoks jagati toodud aluspaaride arvud läbi naise genoomi referents-pikkusega (3200mb) (Morton, 1991). Selle arvutuse tulemus (Joonis 7, roheline) sarnaneb teisendatud Alu-põhise katvuse tulemusega.

2.2.6 Keskmiste telomeeride pikkuste leidmise meetod

Andmestiku iseärasuse tõttu (Joonis 8) leiti telomeeride pikkused vaid poole andmete kohta.

Telomeeride keskmised pikkused arvutati valemi

$$l_{mtl} = \frac{25n_{tel}}{92c}$$

abil, kus n_{tel} on telomeeri k-meri arv ning c on sekveneerimiskatvus. 25 on valemis k-meri pikkus, 92 on telomeeride arv tuumas.

2.3 Tulemused

2.3.1 Tööriistade analüüsi tulemused

Tabel 2 loetleb programmide kasutatavate ressursside piirid, mida kasutati andmestiku paralleelse analüüsimise skripti vajaduste analüüsiks.

Tabel 3 Tööriistade analüüsi tulemused

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	-

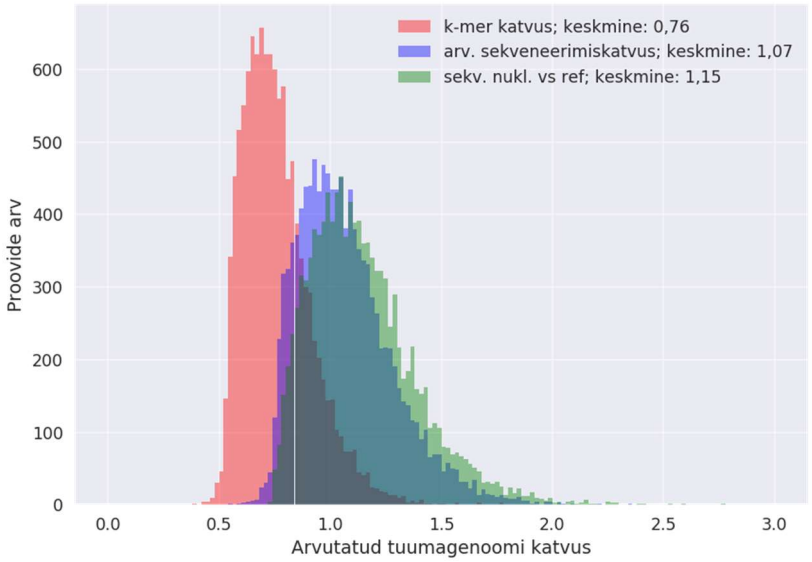
Nendest tulemustest selgub, et ainus samm, mis kasutab kogu saadaolevat ressurssi, on prefetch. Ka on näha, et allalaadimiskiirusest mittesõltuvad sammud võtavad umbes kuus korda nii palju aega, kui ühe uue lähtefaili laadimiseks kulub. Sellest johtuvalt ning et ülesanne lahendada maksimaalse kiirusega,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

2.3.2 get_srr.py töö tulemus

Programm get_srr.py käivitati 23. märtsil 2017 CONVERGE andmekogu sisend-tabeli ja *nohup* käsuga. Programmi töö lõppes 3. mail. Andmete allalaadimiseks ning töötlemiseks kulus 42 päeva. Väljundiks saadi 11670 faili, mille kogumaht oli 1004 megabaiti. SRA-failide kausta jäi 12 gigabaiti .sra.cache ja .sra.vdbcache faile, mille tekkimisega arendusprotsessis ei arvestatud.

2.3.3 Katvuse määramise meetodi tulemused

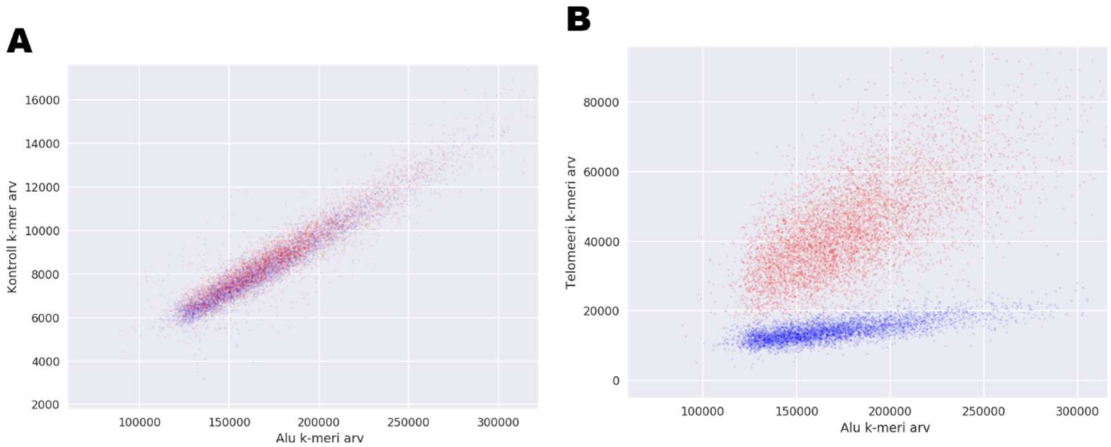
Andmestiku sekveneerimiskatvus leiti kolmel erineval meetodil ning visualiseeriti histogrammil (Joonis 7).



Joonis 7 Arvutatud tuumagenoomide katvused. *Erinevate meetoditega arvutatud sekveneerimiskatvuste jaotumine. Punane – Alu elemendi k-meri arvude suhe referentsväärtusega. Sinine - k-meride lugemisel tekkiva kao suhtes korrekteeritud tulemus. Roheline – sekveneeritud nukleotiidide koguarv ja referentsi pikkuse suhete abil saadud katvuste jaotus.*

2.3.4 Keskmised telomeeride pikkused

Telomeeride k-meri (TTAGGGTTAGGGTTAGGGTTAGGGT) arvude uurimisel selgus, et andmestiku siseselt jagunevad hulgad kahte gruppi (Joonis 8). A: Ühtlaselt üle genoomi jaotuva LINE-elemendi kmeri-arvu (TCTACATATGGCTAGCCAGTTTTCC) ja Alu k-meri suhe. B: Telomeeri k-meri ja Alu k-meri suhe. Punaseks on värvitud proovid, kus AvgSpotLen_l on 166, siniseks kus AvgSpotLen_l on mõni muu väärtusega (enamus 165) (Joonis 8).

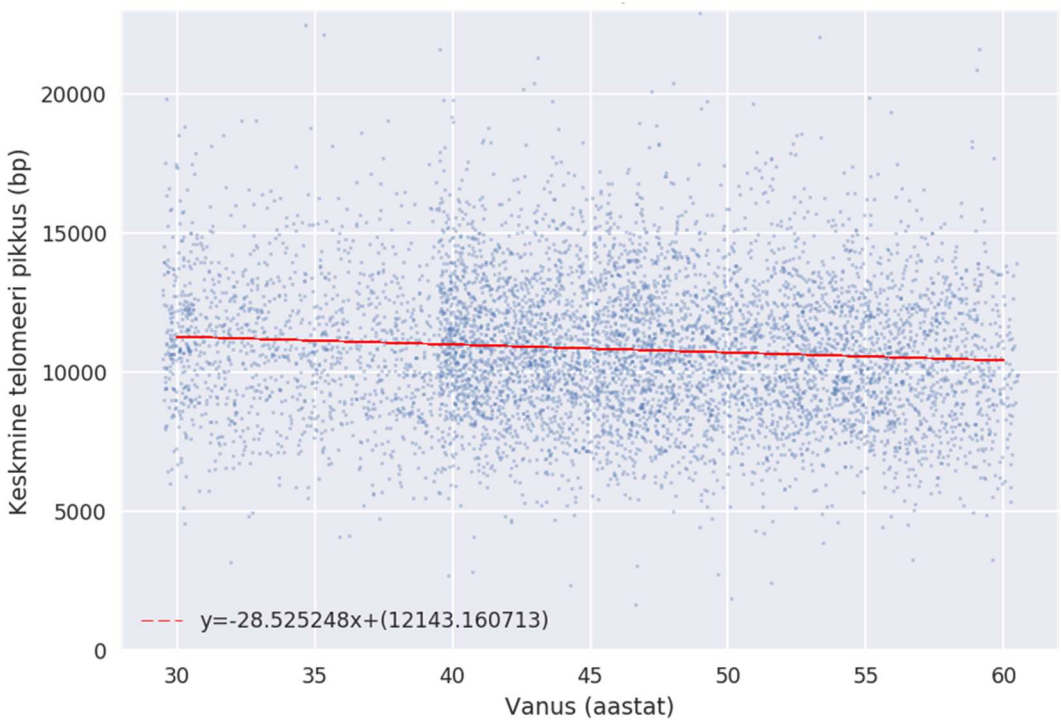


Joonis 8 Poole andmestiku info telomeeride kohta puudulik. *A: Ühtlaselt üle genoomi jaotuva LINE-elemendi kmeri-arvu (TCTACATATGGCTAGCCAGTTTTCC) ja Alu k-meri suhe. B: Telomeeri*

k-meri ja Alu k-meri suhe. Mõlemal graafikul on punaseks värvitud proovid, kus AvgSpotLen_l on 166, siniseks kus AvgSpotLen_l on mõni muu väärtusega (enamus 165).

Joonis 8 A osa illustreerib, et andmed vastavad teooria osas toodud väidetele - Alu ja LINE1 elemente on üle genoomi palju ning arvude lineaarne korrelatsioon näitab, et Alu ja LINE1 elemendi arv on korrelatsioonis sekveneerimiskatvusega. Joonis 8 B osas, kus on võrreldud Alu k-meri arvu ja telomeeri k-meri arvu esineb aga anomaalia. Oodatav tulemus oleks, et punktid esinevad ühtlase hajusa, kuid mõõduka positiivse korrelatsiooniga pilvena, sest telomeeride pikkused on individuaalselt erinevad. Selle asemel aga jagunevad telomeeri ja Alu k-meride suhted kaheks pilveks, kus ühel pilvel ei paista märkimisväärset telomeeride pikkusest tulenevat varieerumist. Lähteandmete uurimise abil leiti aga, et gruppidesse jagunemine toimub muutuja AvgSpotLen_l alusel. Mida see muutuja tähistab, ei ole teada.

Selle erisuse tõttu viidi telomeeride pikkuse analüüs läbi vaid nende proovidega, kus AvgSpotLen_l väärtus oli 166. Neid proove oli kokku 6863 tükki.



Joonis 9 Arvutatud telomeeride pikkuse sõltuvus vanusest. *Horisontaalteljel on inimese vanus ning vertikaalteljel keskmine telomeeri pikkus. Täisarvulisele vanusele on visualiseerimise tarvis liidetud juhuslik arv vahemikus -0.5 kuni 0.5. Trendijoon on arvutatud tõeliste vanuse-väärtustega. See vastab kirjanduses ja Joonis 2 toodud leiule.*

Hajuvusdiagrammilt on näha vanusega seotud valimi-efekti – 40-50 aastaseid on rohkem, kui 30-40 aastaseid.

2.4 Arutelu

2.4.1 Sekveneerimisandmete analüüsi meetod

Planeeritud ja arendatud SRA andmebaasist sekveneerimisandmete laadimise ja töötlemise programm töötas plaanitult ning lõpetas töö prognoositud ajaraami sees. Kogu andmehulga (11670 katset) laadimine õnnestus esimese korraga. Ei esinenud probleeme ei kettamahu ega ülemäärase arvutusressursi kasutamisega.

Kuna programmi võib lugeda töökindlaks ja eesmärki täitvaks, sobib see kasutamiseks ka tulevastes k-mer põhistes ja suurest andmehulgast sõltuvates uurimistöödes.

Programmi ülesehitus võimaldab selle edasiarendamist hajussüsteemidele sobivaks. Selle jaoks tuleks riviloendid asendada tsentraalse järjekorrateenusega, failihoid tsentraalse failihoiuga ning iga töötlemise või laadimise protsess käivitada eraldi. Arvutus- ja laadimisprotsesside mitmele arvutile jagamine võimaldaks protsessi piiravaks ressursiks teha andmekogu hoidja üleslaadimiskiiruse.

2.4.2 Katvuse määramise meetod

Katvuse määramiseks kasutatud Alu järjestuse k-meri sageduse põhise meetodiga ei õnnestunud jõuda sama tulemuseni, milleni jõudsid artikli (Cai et al., 2017) autorid. Kuna artikli autorid ei selgitanud katvuse määramise meetodit ega avaldanud kõikide proovide katvusi, on raske hinnata, kust tulevad erinevused. Ka ei saa k-mer katvuse sekveneerimiskatvuseks teisendamisel sisse tuua veamäära parameetrit, kuna seda pole andmekogu metaandmetes täpsustatud.

Arvutatud keskmise 1,07 ja publitseeritud 1,7 märkimisväärne vahe viitab sellele, et töös esitletud meetod ei sobi sellisel kujul madala katvusega sekveneerimisandmete k-mer põhiseks katvuse määramiseks.

Siiski väärib tähelepanu tõik, et Alu-elemendi k-meri põhjal arvutatud keskmine katvus on sarnane sekveneeritud aluspaaride arvu põhjal arvutatud keskmisele katvusele.

2.4.3 Keskmised telomeeride pikkused

Telomeersete k-meride ja Alu k-meride suhte uurimisel välja tulnud grupeerumise põhjus jääb CONVERGE andmestiku artiklist välja tulemata. Niisiis on tegu uue leiuga. Ka jääb küsimus, kas AvgSpotLen_l parameetri väärtus on kõrvalekalde põhjus või tulemus.

Graafikult on näha, et väljajääva (sinise) grupi liikmed hajuvad Alu-teljel punase grupi sarnaselt ent telomeeri teljel on väärtused kondenseerunud. See võib tähendada seda, et osale proovidest on rakendatud erinevat kvaliteedikontrolli – telomeeride korduvast iseloomust tulenevalt võib õigete nukleotiidide määramine olla ebatäpsem.

Sinise grupi punase asemel välja jätmist õigustab märkimisväärse negatiivse vanusest sõltuva trendi puudumine.

Valitud punases grupis esineb ootuspärane negatiivne telomeeri pikkuse sõltuvus vanusest, kuid keskmine pikkus on raporteeritust kõrgem. See võib tuleneda telomeeri k-meri korduvast iseloomust, mis annab tõenäoliselt ühe readi kohta rohkem vasteid kui peaks. Nende vastete arvu 25-ga korrutades saadakse ülemäära pikad telomeeride keskmised pikkused.

Selleks, et saada telomeeride keskmisteks pikkusteks kirjandusega sarnasemaid tulemusi, võiks valemile lisada ühe korrigeeriva kordaja. See võimaldaks hinnata üksikute proovide telomeeride päris-pikkust ning oleks kasutatav näitaja uuringu jaoks kus on ka fenotüübi info kättesaadav. Samas, andmekogu siseselt sobiks selleks näitajaks ka lihtsalt Alu/telomeeri k-meride suhe. Üldiselt kasutatava korrigeeriva kordaja lisamiseks tuleks leida seosed sekveneerimistehnika ja sekveneerimisandmete töötlemise parameetrist nii, et korrigeeriv kordaja oleks sekveneerimisvigade ja filtreerimisparameetrite funktsioon.

Ka võivad kirjandusele mittevastavad telomeeride pikkused tulla ebatäpselt määratud katvusest. Kui lisada korrigeeriv koefitsient katvuse arvutusele nii, et katvuseks tuleks artiklis toodud 1,7, tulevad määratavad telomeeri pikkused kirjanduses toodule vastavad. Siiski, ilma põhjenduseta, kuidas jõuda 1,7 katvuseni, ei ole lihtsalt katvuste korrigeerimine õigustatud.

Telomeeride keskmiste pikkuste k-meri põhiseks määramiseks tuleks k-meri põhist meetodit kontrollida mõne keemilise meetodiga.

Kokkuvõte

Töö käigus valmis tööriist Sequence Read Archive andmebaasist suure andmekogu allalaadimiseks ning k-mer metoodika tööriistadega töötlemiseks. Tööriist täitis eesmärgi täielikult ning sobib edasiseks kasutamiseks tulevastes uurimistöödes, kus uuritakse sarnases koguses andmeid sarnase metoodikaga.

K-mer metoodika abil sekveneerimiskatvuse metoodika leidmine õnnestus osaliselt – kontrollandmestikul saadi tulemus, mis tõenäoliselt küll korreleerub sekveneerimiskatvusega ent mis erines kontrollandmestikus toodud keskmisest sekveneerimskatvusest.

Sarnaselt katuvse määramise metoodikale, esines töös katsetatud telomeeride keskmise pikkuse määramise metoodikas puudujääke, mis võisid tuleneda nii ebakorrektsest katvuse määramisest kui ka metoodika enda ebasobivusest. Siiski saadi telomeeride keskmise pikkuse määramise metoodika abil arvulised tulemused, mis tõenäoliselt korreleeruvad telomeeride päris-pikkusega.

Resümee

Karl-Sander Erss

Summary

Telomeres are nucleoprotein strutctures that protect the ends of chromosome from fusion and degradation. Ends of telomeres can not be replicated in their entirety so telomeres get shorter with every cell cycle. Telomere length has been associated with several diseases and has been considered a marker for general health and aging.

Next generation sequencing (NGS) technologies have driven the cost of sequencing a human genome down to the $1000 mark. This has led to an explosion in sequencing data quantities which introduces the need for fast and generic methods for data analysis.

The main objective of this thesis was to develop a method to estimate the average telomere length from NGS-data derived k-mer lists. Additional methods for data retrieval and processing, also sequencing coverage estimation are developed and evaluated.

A script which handled retrieval of data from Sequnece Read Archive and reducing it into lists of ~3000 k-mers was developed. The program can be used for similar future studies.

The efficacy of the presented k-mer based methods of esitmating sequencing coverage and telomere length was inconclusive.

Kasutatud kirjanduse loetelu

Baird DM, Rowson J, Wynford-Thomas D, Kipling D (2003). Extensive allelic variation and ultrashort telomeres in senescent human cells. *Nat. Genet.* 33:203–207.

Batzer MA, Deininger PL (2002). ALU REPEATS AND HUMAN GENOMIC DIVERSITY. *Nat. Rev. Genet.* 3:370–379.

Cacho A, Smirnova E, Huzurbazar S, Cui X (2016). A comparison of base-calling algorithms for illumina sequencing technology. *Brief. Bioinform.* 17:786–795.

Cai N, Bigdeli TB, Kretzschmar WW, … Flint J (2017). 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. *Sci. data* 4:170011.

Calado RT, Young NS (2008). Telomere maintenance and human bone marrow failure. *Blood* 111.

Cawthon RM (2009). Telomere length measurement by a novel monochrome multiplex quantitative PCR method. *Nucleic Acids Res.* 37.

Cawthon RM (2002). Telomere measurement by quantitative PCR. *Nucleic Acids Res.* 30.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 38:1767–1771.

Compeau PEC, Pevzner PA, Tesler G (2011). How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* 29:987–991.

Van Dijk EL, Lè Ne Auger H, Jaszczyszyn Y, Thermes C (2014). Ten years of next-generation sequencing technology. *Trends Genet.* 30:418–426.

Ding Z, Mangino M, Aviv A, Spector T, Durbin R (2014). Estimating telomere length from whole genome sequence data. *Nucleic Acids Res.* 42.

Horikawa I, Barrett JC (2003). Transcriptional regulation of the telomerase hTERT gene as a target for cellular and viral oncogenic mechanisms. *Carcinogenesis* 24:1167–1176.

Kaplinski L, Lepamets M, Remm M (2015). GenomeTester4: a toolkit for performing basic set operations - union, intersection and complement on k-mer lists. *Gigascience* 4:58.

Kimura M, Stone RC, Hunt SC, Skurnick J, Lu X, Cao X, Harley CB, Aviv A (2010). Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths. *Nat. Protoc.* 5:1596–1607.

Kodama Y, Shumway M, Leinonen R (2012). The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Res.* 40.

Lamichhaney S, Fan G, Widemo F, … Andersson L (2015). Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nat. Genet.* 48:84–88.

Lander ES, Linton LM, Birren B, … Morgan MJ (2001). Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

Ledergerber C, Dessimoz C (2011). Base-calling for next-generation sequencing platforms. *Brief. Bioinform.* 12:489–497.

Martínez P, Blasco MA (2010). Role of shelterin in cancer and aging. *Aging Cell* 9:653–666.

Mather KA, Jorm AF, Parslow RA, Christensen H (2011). Is Telomere Length a Biomarker of Aging? A Review. *Journals Gerontol. Ser. A Biol. Sci. Med. Sci.* 66A:202–213.

Montpetit AJ, Alhareeri AA, Montpetit M, … Jackson-Cook CK (2014). Telomere length: a review of methods for measurement. *Nurs. Res.* 63:289–99.

Morton NE (1991). Parameters of the human genome (physical map/genetic map/genomic size). *Med. Sci.* 88:7474–7476.

Muzzey D, Evans EA, Lieber C (2015). Understanding the Basics of NGS: From Mechanism to Variant Calling. *Curr. Genet. Med. Rep.* 3:158–165.

Nakamura Y, Hirose M, Matsuo H, Tsuyama N, Kamisango K, Ide T (1999). Simple, rapid, quantitative, and sensitive detection of telomere repeats in cell lysate by a hybridization protection assay. *Clin. Chem.* 45:1718–1724.

Nersisyan L, Arakelyan A (2015). Computel: Computation of mean telomere length from whole-genome next-generation sequencing data. *PLoS One* 10.

Patro R, Mount SM, Kingsford C (2013). Sailfish: Alignment-free Isoform Quantification from RNA-seq Reads using Lightweight Algorithms.

Rode L, Nordestgaard BG, Bojesen SE (2015). Peripheral blood leukocyte telomere length and mortality among 64,637 individuals from the general population. *J. Natl. Cancer Inst.* 107:djv074.

Roy-Engel AM, Carroll ML, Vogel E, Garber RK, Nguyen S V, Salem AH, Batzer MA, Deininger PL (2001). Alu insertion polymorphisms for the study of human genomic diversity. *Genetics* 159:279–290.

Tian X, Chen B, Liu X (2009). Telomere and Telomerase as Targets for Cancer Therapy.

Witzany G (2008). The Viral Origins of Telomeres and Telomerases and their Important Role in Eukaryogenesis and Genome Maintenance. *Biosemiotics* 1:191–206.

Kasutatud veebiaadressid

https://www.ncbi.nlm.nih.gov/sra/

https://github.com/ncbi/sra-tools/wiki/HowTo:-Access-SRA-Data

http://click.pocoo.org/5/

http://jupyter.org/

https://matplotlib.org/

https://github.com/karlerss/telomere-length

Lisad

Lisa 1 - SraRunTable.txt – Sequnce Read Archive'i väljund uurimuse katsete kohta. Kasutati analüüsiprogrammi sisendina.

Kättesaadav veebiaadressilt:

https://github.com/karlerss/telomere-length/blob/master/tlenpy/SraRunTable.txt

Lisa 2 - kmer_sample_min.txt – Nimekiri huvipakkuvatest k-meridest. Kasutati analüüsiprogrammi sisendina.

Kättesaadav veebiaadressilt:

https://github.com/karlerss/telomere-length/blob/master/tlenpy/kmer_sample_min.txt

Lisa 3 – HGDP00778.bam

Kättesaadav veebiaadressilt:

http://cdna.eva.mpg.de/denisova/BAM/human/HGDP00778.bam

Lihtlitsents

Mina, Karl-Sander Erss

(sünnikuupäev: 11.12.1994)

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

Telomeeride keskmise pikkuse hindamine teise generatsiooni sekveneerimisandmetest,

mille juhendaja on Tarmo Puurand,

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, 28.05.17 (kuupäev)