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)

### 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

## 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 matriistahelaid. 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>)

### Sekveneerimiskatvus

### 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)

## K-mer metoodika sekveneerimisandmete analüüsis

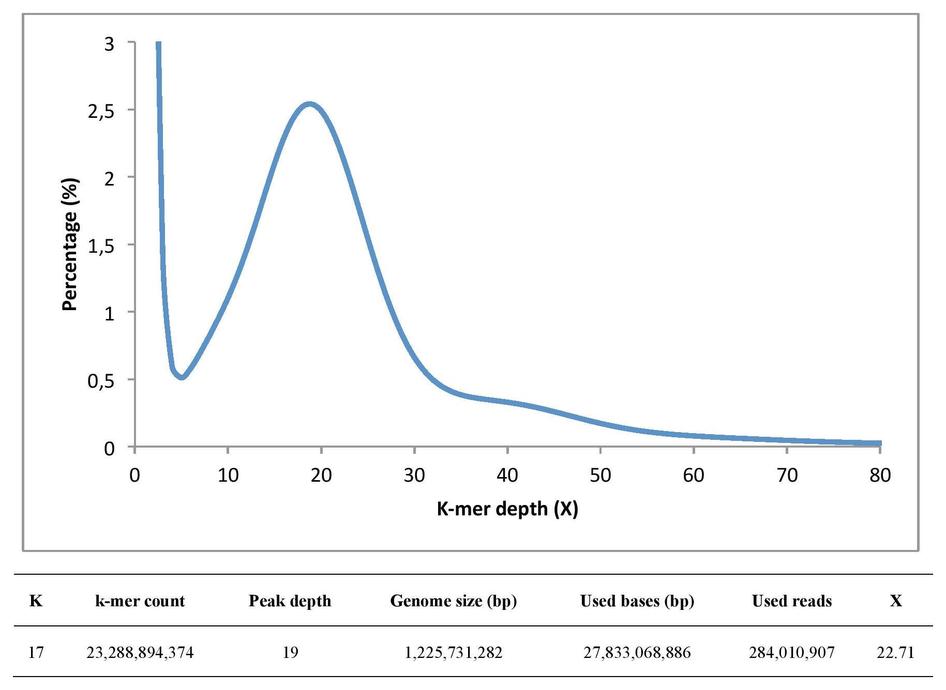
Traditsioonilised sekveneerimisandmete analüüsi meetodid põhinevad sekveneeritud lõikude referentsile joondamisel. Joondamisele põhinevate analüüside tulemused ning keerukus on väga tundlikud joondamisparameetrite valikule ja seega veaohtlikud. Samuti on lõikude referentsile joondamine arvutuslikult 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 1** 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)

### Jellyfish 2

### GenomeTester4

# Eksperimentaalosa

## Töö eesmärgid

Selle töö eesmärgid on:

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

Selle töö peamine eesmärk on **määrata teise generatsiooni madala katvusega sekveneerimisandmete kogu põhjal keskmised telomeeride pikkused** (E1). 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** (E3).

## Metoodika

Kasutame andmeid uuringust…

### 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/>).

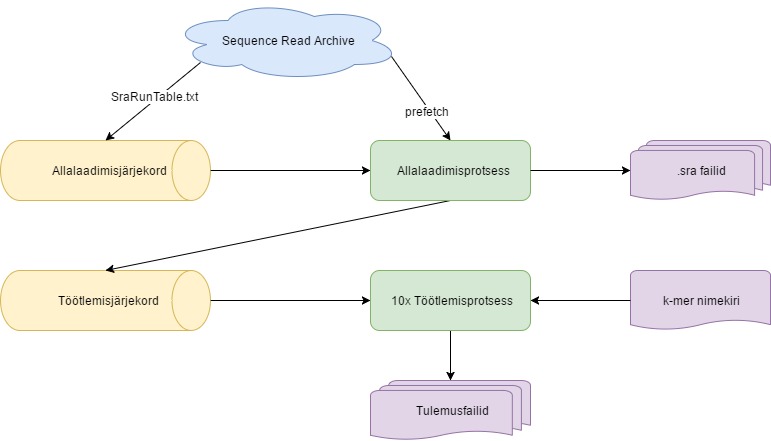
Huvipakkuvate k-meride sageduste analüüs viidi läbi Jupyter interaktiivses keskkonnas (<http://jupyter.org/>). Tulemused visualiseeriti matplotlib abil (<https://matplotlib.org/>).

## Tulemused

### Keskmised telomeeride pikkused

### Katvuse määramise meetod

### Sekveneerimisandmete analüüsi meetod



Joonis 2

K-mer metoodika abil uurimiseesmärkide saavutamiseks peavad ühe inimese andmed läbima järgnevad sammud:

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.

Kuna valitud andmed asuvad SRA andmebaasis, on andmete allalaadimiseks otstarbekas kasutada SRA toolkit komponente.

## Arutelu

# Kokkuvõte

# Resümee

# Kasutatud kirjanduse loetelu

# Kasutatud veebiaadressid