TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOINFORMAATIKA ÕPPETOOL

Katvust mõjutavate parameetrite hindamine

Bakalaureusetöö

12 EAP

Carmen Oroperv

Juhendaja MSc Fanny-Dhelia Pajuste

TARTU 2019

Katvust mõjutavate parameetrite hindamine

Katvus ehk sekveneerimissügavus väljendab seda, mitu korda on üks nukleotiid sekveneeritud. Katvuse andmeid kasutatakse genoomianalüüsis nii indiviidi geneetiliste variatsioonide uurimiseks, geenidekspressiooni analüüsiks kui ka DNA kõrgema struktuuri uurimiseks. Peamiseks probleemiks seejuures on katvuse kõrvalekalded oodatud ühtlasest väärtusest. Käesoleva töö eesmärk on anda ülevaade katvuse rakendustest inimese genoomi analüüsides ja kirjeldada katvuse väärtust mõjutavaid tegureid ning eksperimentaalses osas hinnata GC-sisalduse, genoomipositsiooni ja kromosoomi mõju *k*-meeri katvuse väärtusele. Katvust mõjutavate parameetrite tuvastamine ning sobivad mudelid katvuse korrigeerimiseks võimaldavad täpsemalt analüüsida madalama katvusega sekveneeritud proove ning vähendada analüüside valepositiivsete ja –negatiivsete tulemuste hulka.

Märksõnad: katvus, *k*-meer, Illumina sekveneerimine, GC-sisaldus, lineaarne regressioonimudel

CERCS: B110 (Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika)

Evaluation of the parameters affecting sequencing coverage

Coverage expresses how many times each nucleotide is sequenced. Coverage data is used in genomic analyses to detect genetic variations, determine gene expression abundance or analyse the higher structure of DNA. The main problem of using coverage data is the deviation from the expected value. The purpose of this study is to give an overview of applications of coverage in human genome analyses, describe factors that cause deviation from the expected coverage value and in the practical part, evaluate the effect of GC content, position and chromosome on *k*-mer coverage. Finding the parameters that affect coverage and composing appropriate models to correct the bias permits to accurately analyse low-coverage sequencing samples and reduces the amount false positives and false negatives in the results.

Keywords: coverage, *k*-mer, Illumina sequencing, GC content, linear regression model

CERCS: B110 (Bioinformatics, medical informatics, biomathematics, biometrics)

SISUKORD

KASUTATUD LÜHENDID

ANOVA	*analysis of variance*	dispersioonianalüüs
bp	*base pair*	aluspaar
cDNA	*complementary DNA*	komplementaarne DNA
ChIP-seq	*chromatin immunoprecipitation followed by sequencing*	kromatiini immunosadestamine ja seejärel sekveneerimine
CNV	*copy number variant*	koopiaarvu variatsioon
EGV	*Estonian Genome Center*	Eesti geenivaramu
FPKM	*fragments per kilobase of transcript per million fragments mapped*	fragmente ühe kilobaasilise transkripti ja miljoni joondatud fragmendi kohta
GRC	*Genome Reference Consortium*	referentsgenoome haldav konsortsium
indel	*insertion and deletion*	insertsioon ja deletsioon
LOESS	*locally estimated scatterplot smoothing*	hajuvusdiagrammi silumine lokaalse regressiooniga
mRNA	*messenger RNA*	informatsiooni-RNA
NGS	*next-generation sequencing*	teise põlvkonna sekveneerimine
RNA-seq	*RNA sequencing*	RNA sekveneerimine
rRNA	*ribosomal RNA*	ribosoomi-RNA
SBS	*sequencing by synthesis*	sünteesil põhinev sekveneerimine
SNV	*single nucleotide variant*	ühenukleotiidiline variatsioon
WES	*whole exome sequencing*	eksoomi sekveneerimine
WGS	*whole genome sequencing*	täisgenoomi sekveneerimine

SISSEJUHATUS

Teise põlvkonna sekveneerimistehnoloogiad (NGS) on käesoleva sajandi jooksul muutnud sekveneerimise varasemast oluliselt kiiremaks ja odavamaks. Sekveneerimisandmete kasutamine on sealjuures saanud tavapäraseks osaks genoomianalüüsis ning kliinilistes rakendustes. Katvuse andmeid, mis väljendavad uuritava positsiooni esinemissagedust sekveneerimislugemites, kasutatakse näiteks geneetiliste variatsioonide määramiseks ja geenidekspressiooni ning DNA kõrgema struktuuri uurimiseks.

Katvuse põhjal saab määrata nii ühenukleotiidilisi polümorfisme, koopiaarvu variante kui ka suuremaid variatsioone nagu tri- ja monosoomiad. Olulisel kohal on sekveneerimisandmete ja katvuse kasutamine ka sünnieelses diagnostikas, võimaldades loote genoomi uurida invasiivsest testimisest ohutumate meetoditega.

Katvuse rakendamisel analüüsides on suurimaks probleemiks kõrvalekalded oodatud väärtusest, mis on mõjutatud mitmete tegurite poolt: GC-sisaldus uuritavas genoomis, lugemite joondamine ning sekveneerimisvead. Kuigi sekveneerimistehnoloogiad arenevad pidevalt edasi, andes järjest usaldusväärsemaid väljundandmeid, on teise põlvkonna lühikeste lugemite katvuse korrigeerimine olulisel kohal analüüside töövoos. Katvuse korrigeerimine võimaldab analüüse läbi viia madalama sekveneerimiskatvusega, parandada analüüside täpsust ja tagab odavama sekveneerimishinna ning seega paremad võimalused testide kliiniliseks kasutamiseks.

Käesolev töö annab ülevaate katvuse rakendustest genoomianalüüsides, koondab senised teadmised mõjuteguritest, mis tekitavad katvuses kõrvalekaldeid ning analüüsib *k*-meeri katvuse varieerumist GC-protsendi ning *k*-meeri asukoha põhjal.

1. KIRJANDUSE ÜLEVAADE

1.1. Katvus

Teoreetiline või eeldatav katvus näitab, mitu korda on üks nukleotiid keskmiselt sekveneeritud, sõltuvalt lugemite pikkusest ja arvust ning eeldades, et lugemid jaotuvad üle genoomi ühtlaselt. Katvuse mõiste võib viidata ka sellele, kui suurt osa genoomist lugemid protsentuaalselt katavad (Sims *et al*., 2014). Edaspidi aga keskendume katvusele kui sekveneerimissügavusele ehk kui palju lugemeid antud positsiooni joondub. Hetkel kasutatakse eeldatava katvusena kogu genoomi keskmist katvust, mida arvutatakse Lander-Watermanni valemiga $C = \frac{L*N}{G}$, kus *C* tähistab katvust, *L* lugemi pikkust, *N* lugemite arvu ja *G* genoomi pikkust (Lander ja Waterman, 1988). Lokaalne katvus on keskmine katvus huvipakkuvas piirkonnas näiteks geeni ümbruses või kindlas genoomi positsioonis, mis on arvutatud uuritavasse alasse kuuluvate nukleotiidide katvuse põhjal.

K-meeri (*k* nukleotiidi pikkuse oligomeeri) katvus on *k*-meeri esinemiste arv joondatud või joondamata sekveneerimislugemites, millele võib viidata ka kui *k*-meeri sagedusele lugemites (Kaplinski *et al*., 2015). Üldjuhul kasutatakse unikaalseid *k*-meere, mis esinevad genoomis ainult ühes kohas. Kuna iga *k*-meer tuvastatakse lugemites tervikuna, mitte ei määrata igale *k*-meeri nukleotiidile joondatud lugemite arv, on *k*-meeri katvus positsiooni keskmisest katvusest väiksem.

1.2. Sekveneerimisel tekkivad vead

Sekveneerimisel tekkivad vead mõjutavad nii joondatud lugemite põhjal määratud katvuse kui ka joondamata lugemitest arvutatud *k*-meeri katvuse väärtust. Põhjustades joondamisel valepaardumisi ning halvemal juhul lugemi joondumist valele asukohale või joondumata jäämist, tekitavad vead ühtlasest katvuse väärtusest kõrvalekaldeid. Joondusvabade meetodite korral võib vigade tulemusel unikaalsete *k*-meeride sagedus olla oodatust kõrgem või madalam (Laehnemann *et al*., 2016). Sekveneerimisandmete kasutamiseks edasistel analüüsidel ja valede järelduste vältimiseks on oluline teada tihedamini esinevaid vigu, nende osakaalu ja põhjuseid.

Illumina sekveneerimismeetod on hetkel turul domineeriv tehnoloogia ning põhineb järjestuse sünteesil (SBS). Vigade esinemissagedus Illumina sekveneerimisel on keskmiselt 5’ paarislugemis 0,0021 ühe nukleotiidi kohta ja 3’ paarislugemis 0,0042 ühe nukleotiidi kohta.

Samas ei ole vead lugemites ühtlaselt jaotunud ning vigade esinemissagedus on suurem teatud motiividele järgnevates positsioonides (Schirmer *et al*., 2016), näiteks kolmenukleotiidilise GGC järjestuse järel ja ümberpööratud korduste ümbruses (Nakamura *et al*., 2011). Kõige levinumaks veaks on vale nukleotiidi määramine järjestusse (Schirmer *et al*., 2015). Suurema sagedusega tekivad vead A ja T nukleotiidide määramisel, misjuhul asendadakse need enamasti G nukleotiidiga. Indeleid tekib sekveneerimise käigus harvem, kuid kõrge või madala G ja C nukleotiidide osakaaluga regioonides suureneb ka indelite esinemissagedus (Ross *et al*., 2013).

Vale nukleotiidi määramisel on mitmeid põhjused. Kui ensüümide töö pole täielik ja mõnelt sünteesitavalt ahelalt ei eemaldata terminaatormärgist, jääb see ahel võrreldes teiste ahelatega sildamplifikatsiooni käigus moodustatud klastris nukleotiidi võrra maha (*phasing*). Kui teised ahelad tuvastavad järgmistes tsüklites juba uusi nukleotiide, annab mahajäänud ahel varasemate positsioonide signaale. Sama probleem tekib kui ühe tsükli käigus lisatakse ahelale mitu nukleotiidi korraga, mille tulemusel antud ahel on teistest klastri ahelatest positsiooni võrra eespool (*pre-phasing*). Nende probleemide tõttu pole pildid fluorestsentssignaalidest täpsed ja tulemuseks võib olla vale nukleotiidi määramine (Cacho *et al*., 2016). Kuna *phasing* ja *pre-phasing* võivad toimuda sünteesi käigus korduvalt, on vigade sagedus suurem lugemite lõpus (Minoche *et al*., 2011; Schirmer *et al*., 2015). Teiseks vale nukleotiidi määramise põhjuseks võib olla fluorofooride emissioonispektrite kattumine, mille tulemusel ühe fluorofoori ergastumisel tuvastatakse osaliselt ka teise fluorofoori signaal (Laehnemann *et al*., 2016).

Saadaval on erinevaid programme nukleotiidide määramise kvaliteedi hindamiseks ja vigade eemaldamiseks (Laehnemann *et al*., 2016). Vastav algoritm valitakse sõltuvalt sekveneerimistehnoloogiast, kuna erinevate tehnoloogiate puhul on peamine vigade tüüp erinev ning vigade sagedus on suurim erinevates regioonides: kui Illumina metoodika põhjustab peamiselt nukleotiidide asendusi, mis on sagedasemad lugemite lõpus, siis näiteks 454 (Roche Diagnostics Corporation), Ion Torrent (Life Technologies Corporation) ja SMRT (Pacific Biosciences Inc.) tehnoloogiate puhul tekivad peamiselt indelid homopolümeersetes järjestustes (Laehnemann *et al*., 2016).

Toodud näited tekkivatest vigadest ja nende põhjustest on ainult osa sekveneerimisprotsessi keerukusest. Detailne meetodite valik sõltub iga analüüsi algmaterjalist ja sekveneerimise eesmärkidest; näiteks, kas kasutatakse üksik- või paarislugemeid, kui suured peaksid olema

fragmendid raamatukogu koostamisel, kas sekveneerimisele eelnev amplifikatsioon on vajalik (Bronner *et al*., 2013).

1.3. Lugemite joondamine referentsgenoomile

Selleks, et sekveneerimisandmeid edasisteks analüüsideks kasutada, joondatakse lugemid enamasti referentsgenoomile. Joondamine annab informatsiooni selle kohta, kuhu lugemid genoomil paigutuvad ja kuidas nad üksteise suhtes paiknevad. Andmete analüüsimisel on see arvutuslikult üks kõige ressursi- ja ajakulukamaid osasid (Reinert *et al*., 2015). Alternatiivid referentsgenoomile joondamisele on *de novo* assambleerimine või joondusvabade analüüsimeetodite kasutamine. Kui referentsile joondamist siiski edasisteks analüüsideks kasutatakse, mõjutavad lugemite paigutumist ning seega ka katvuse väärtust mitmed tegurid: referentsgenoom, kordusjärjestused, sekveneerimisel tekkivad vead ja variatsioonid sekveneeritud proovis.

Inimese genoomi uuringud toetuvad suuremas jaos referentsgenoomile, mis avaldati esmakordselt 2001. aasta veebruaris ja koostati mitmete anonüümsete geenidoonorite andmete põhjal. Genome Reference Consortium (GRC) avaldab teatud aja tagant järjestuse väiksemaid muudatusi või suuremaid järjestuste koordinaatide muutusi kaasavaid uuendusi. Kõige hilisem referentsjärjestuse versioon GRCh38.p13 avaldati 2019. aasta märtsis, mis sisaldab 875 joonduse vahet, mida pole suudetud lugemitega katta ja ka määramata nukleotiide (märgitud referentsjärjestuses tähega N).[1] Määramata järjestusega regioonidesse ei saa vastavad lugemid joonduda, mistõttu need võivad joonduda kas sarnasesse või identsesse asukohta mujal genoomis või jääda joondumata. Lisaks pole ükski genoom varasemalt uuritutega ega referentsiga täies ulatustes identne. Selleks, et variatsioone referentsile joondatud lugemitest tuvastada, tuleb lugemite paigutamisel lubada valepaardumisi või joonduse vahesid, mille tulemusel kõik lugemid ei paigutu referentsile täielikult (kõik lugemi nukleotiidid ei ole sarnased referentsgenoomiga) (Hung ja Weng, 2017).

Lugemite joondamine on kõige problemaatilisem kordusjärjestuste aladel, kuna korduva motiivi tõttu võib lugem referentsjärjestusel sobida võrdse tõenäosusega mitmesse kohta. Kui kordused on identsed, võib lühikeste lugemite algse asukoha leidmine olla võimatu (Reinert *et al*., 2015). Joondamist kordusjärjestustele lihtsustab pikemate lugemite kasutamine või paarislugemite joondamine, misjuhul on suurem tõenäosus, et üks lugemitest joondub

[1] https://www.ncbi.nlm.nih.gov/grc/human/data, 03.05.2019

unikaalselt kindlasse positsiooni (Hung ja Weng, 2017). Kuna aga inimese genoomist moodustavad peaaegu poole korduvad järjestused (Lander *et al*., 2001), ei pruugi ka paarislugemite joondamine alati üheselt võimalik olla. Üks võimalus on jätta mitmesse regiooni joonduvad lugemid joondamata, kuid sel juhul läheb suur osa kordusjärjestuste informatsioonist kaotsi ning katvuse väärtus üle genoomi on lõpptulemusel ebaühtlasem. Selle vältimiseks on kaks võimalust: valitakse parim joondus, võimalikult väheste valepaardumistega (võrdselt sobivate joonduste korral valitakse neist üks juhuslikult) või joondusalgoritm tagastab kõik võimalikud joonduse variandid. Kuigi toodud kordusjärjestustele joondamise meetodid võimaldavad määrata neis regioonides katvuse, ei ole järeldused nende põhjal alati täiesti usaldusväärsed. Näiteks parima joonduse valimisel võib jääda variatsioon tuvastamata, sest lugem joondub mõnesse teise genoomi regiooni küll vähesemate valepaardumistega, kuid lugemi tõene asukoht koos variatsiooniga välistatakse kehvema joonduse tõttu (Treangen ja Salzberg, 2012).

1.4. Katvuse rakendused inimese genoomi analüüsides

Kõrge läbilaskevõimega (*high throughput*) sekveneerimismeetodid on alates turule jõudmisest arenenud kiiresti ja sekveneerimine on muutunud aastatega järjest odavamaks. See on andnud võimaluse kasutada genoomi või RNA andmeid nii bioloogias kui ka meditsiinis palju ulatuslikumalt ning pannud aluse mitmetele uutele uurimissuundadele. Näiteks personaalmeditsiin, mille aluseks on geneetiliste variatsioonide tuvastamine, võib tulevikus aidata haigusi paremini ennetada ning muuta ravimite manustamist. Üldiste ettekirjutuste asemel võiks ravimi valimine ja ravimidooside määramine põhineda iga inimese individuaalsetel eripäradel (Ye *et al*., 2015). Sõltuvalt sellest, millise eesmärgi täitmiseks genoomi, genoomi osa või RNA-d sekveneeritakse, on sekveneerimisuuringute ülesehitus erinev.

DNA resekveneerimine on varasemalt sekveneeritud liigi erinevate isendite DNA sekveneerimine eesmärgiga uurida geneetilist varieeruvust indiviidide, perekondade või populatsioonide vahel. Kogu genoomi sekveneerimise (WGS) andmetest on võimalik määrata variatsioone kogu genoomi ulatuses. Kogu eksoomi sekveneerimine (WES) keskendub valke kodeerivate geenide uurimisele. Suunatud ehk ainult väiksema huvipakkuva ala resekveneerimine võimaldab sekveneerida võrdväärsete kuludega, kuid suurema katvusega, mis võib tagada suurema täpsuse variatsioonide tuvastamisel. Kuigi sekveneerimise kulud võivad olla suunatud sekveneerimisel väiksemad, seab see analüüside osas piiranguid,

võimaldades uurida ainult väiksemaid variatsioone, mis jäävad sekveneeritud järjestuste piirkonda (Sims *et al*., 2014). Samuti on näidatud, et WES on rohkem mõjutatud proovi GC-sisaldusest ning eksoomi järjestuste eraldamine DNA proovist ei pruugi tagada alati kogu eksoomi esindatust lugemites (Meienberg *et al*., 2016).

Eksoomilt transkribeeritud RNA analüüsiks on välja töötatud transkriptoomi sekveneerimine ehk RNA-seq, mis võimaldab katvuse abil täpsemalt hinnata erinevate transkriptide ekspressiooni taset ja alternatiivseid splaissinguvariante (Sims *et al*., 2014). Andmete analüüsil tuleb arvesse võtta, et transkriptid võivad ekspresseeruda ühes rakus väga erineval hulgal, ühest koopiast miljoniteni, sõltudes näiteks raku tüübist ja arengutasemest. Seetõttu varieerub ka transkriptide katvus. Madalamalt ekspresseeritud geenide transkripte on suhteliselt vähem kui kõrgelt ekspresseeritud geenide lugemeid ja nende detekteerimine on keerulisem (Halvardson *et al*., 2013).

Sekveneerimise eesmärgiks võib olla ka DNA-valk interaktsioonikohtade leidmine. Üheks levinumaks meetodiks on kromatiini immunosadestamine ja seejärel sekveneerimine (ChIP-seq). Interaktsiooni asukohtade tuvastamiseks vajalik lugemite arv ja seega ka katvuse väärtuse suurus sõltub sellest, kas uuritav faktor on kindla genoomiregiooni spetsiifiline või on genoomis laiemalt levinud ning seostub mitmete kohtadega. Mida rohkem on uuritaval valgul seondumiskohti DNA-ga, seda suurem lugemite arv on sekveneerimisel vajalik. (Landt *et al*., 2012).

Katvuse andmeid kasutatakse nii geneetiliste variatsioonide määramisel kui ka genoomi struktuuri ja geenide ekspressiooni puudutavates uurimisküsimustes. Peamine probleem andmete rakendamisel on katvuse kõrvalekalded oodatud ühtlasest väärtusest, mis tekitavad analüüsidel probleeme kahel põhjusel. Esiteks, oodatust madalamad katvuse väärtused ja nende põhjal tehtud edasised analüüsid on rohkem mõjutatud lugemites esinevatest sekveneerimisvigadest. Kui sekveneerimisprotsessi käigus on tekkinud vead, võivad need väheste joondatud lugemite informatsiooni põhjal olla edasiste analüüside käigus valesti interpreteeritud (Sims *et al*., 2014), mistõttu võib saada näiteks valepositiivseid tulemusi ühenukleotiidiliste variatsioonide (SNV) tuvastamisel (Gómez-Romero *et al*., 2018). Madala katvuse korral on suurem ka tõenäosus, et uuritaval positsioonil olevad lugemid esindavad ainult ühte kahest kromosoomikoopiast (Nielsen *et al*., 2011). Teiseks on keeruline ilma täiendava informatsioonita kindlaks teha, mis on oodatust madalama katvuse põhjuseks. Antud regiooni võib vähem lugemeid olla joondunud geeni puudumise või referentsist erineva

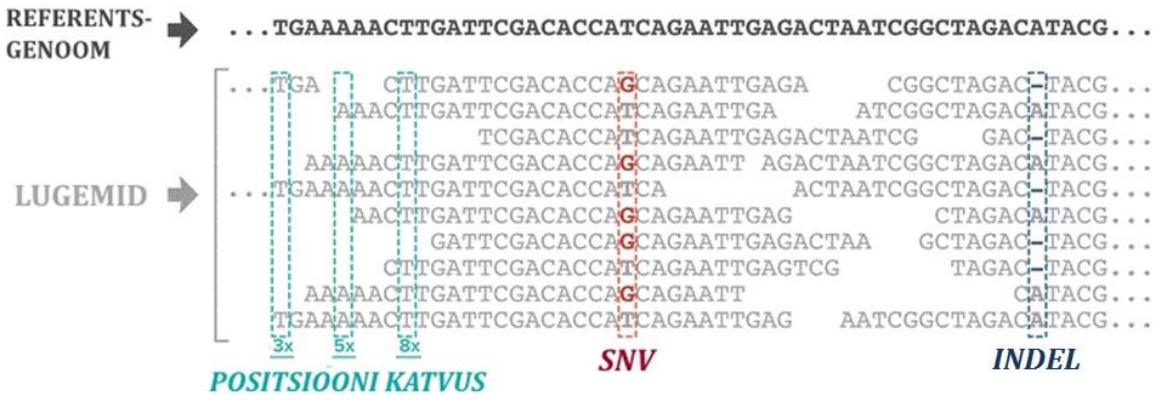
koopiaarvu tõttu. Samas võib vähemate lugemite joondumist põhjendada ka assambleerimise või joondamise probleemiga (Sims *et al*., 2014).

1.4.1. Variatsioonide tuvastamine

Katvuse väärtuse andmeid kasutatakse peamiselt erinevate geneetiliste variatsioonide tuvastamiseks: SNV-d, väiksemad insertsioonid ja deletsioonid (indelid), suuremad struktuurilised ümberkorraldused, mille hulka kuuluvad translokatsioonid ning koopiaarvu variatsioonid (CNV-d). Hoolimata variatsiooni pikkusest on peamine lähtekoht järelduste tegemisel katvuse väärtuse kõikumised, mis viitavad indiviidi geneetilistele variatsioonidele. Tõeste variatsioonide täpne tuvastamine on võimalik aga ainult juhul, kui need suudetakse eristada katvuse kõrvalekalletest, mis on tekkinud näiteks järjestuse omadustest või tehnoloogiast tulenevatel põhjustel (Ross *et al*., 2013).

1.4.1.1. Ühenukleotiidiliste variatsioonide määramine

Varasemad SNV-de tuvastamise ja genotüpiseerimise meetodid toetusid joondatud lugemite katvuse andmetele. Kui sekveneerimisel ei tekiks vigu ja katvus oleks üle genoomi ühtlane, oleks kõrge katvusega sekveneerimisel SNV-de tuvastamine sel meetodil õigustatud – heterosügootse SNV puhul erineksid pooled lugemid referentsjärjestusest, homosügootse puhul oleks uuritavas positsioonis kõik lugemid referentsist erinevad (Muzzey *et al*., 2015). Näiteks, kui referentsjärjestuses on nukleotiid T ja kui sekveneerimiskatvuse väärtus on 10, millest viis lugemit sisaldavad uuritavas positsioonis T nukleotiidi ja viis lugemit G nukleotiidi, võib uuritava indiviidi genotüübiks määrata TG (Joonis 1). Sama metoodika põhjal saab tuvastada ka lugemite pikkusest väiksemaid indeleid. Kuna aga sekveneerimise käigus tekib lugemitesse vigu ning lugemite joondamisel ei ole võimalik kõikide lugemite asukohta üheselt leida, pole SNV-de tuvastamine alati sel meetodil usaldusväärne. Kui 10-st joondatud lugemist seitse sisaldavad T nukleotiidi ning kolm G nukleotiidi on keerulisem järeldada, kas tegu on SNV või sekveneerimisveaga.



Joonis 1. Näide SNV ja indeli määramisest katvuse põhjal. (Muzzey *et al*., 2015, kohandatud, osaline)

Probleemi lahenduseks on välja töötatud n-ö tõenäosuslikud algoritmid, mis kaasavad eelnevat infot võimalikest tekkinud sekveneerimisvigadest, alleelide sagedustest ja ahelduse tasakaalutusest (*Linkage disequilibrium*), et anda iga SNV esinemise tõenäosus (Depristo *et al*., 2011; Li, 2011; McKenna *et al*., 2010).

Hiljuti avaldatud COBASI *(coverage-based single nucleotide variant identification*) meetod *de novo* SNV-de tuvastamiseks põhineb genoomis leiduvate unikaalsete *k*-meeride katvusel. *De novo* mutatsioonid on geneetilised variandid, mis ei ole pärilikud. Mutatsioonid on lapsel tekkinud esmakordselt ja vanematel antud geneetilist varianti ei esine. Variatsioonide määramiseks leitakse referentsjärjestusest unikaalsed *k*-meerid. *K*-meeri katvus lugemites on heterosügootse variatsiooni korral poole väiksem ning nullilähedane homosügootsete variatsioonide korral. Katvuse väärtuste kõikumise põhjal määratakse regioonid, kus potentsiaalselt esinevad SNV-d ning vastavaid *k*-meere sisaldavad lugemid joondatakse. *De novo* variatsioonide tuvastamiseks joondatakse ka ema ja isa varieeruva katvusega regioonide lugemid. Genotüüpe võrreldes tuvastatakse võimalikud *de novo* SNV-d. (Gómez-Romero *et al*., 2018)

Kõige ajakulukam osa SNV-de tuvastamisel on lugemite joondamine. Kiiremaks analüüsimiseks ning joondamisel tekkivate küsitavuste vältimiseks on välja töötatud metoodikad, mis ei vaja variatsioonide tuvastamiseks lugemite joondamist referentsile. Üks väljapakutud lahendustest (Kimura ja Koike, 2015) kasutab Burrows-Wheeler transformatsiooni ning määrab SNV-d minimaalse pikkusega unikaalsete *k*-meeride sageduste muutuste põhjal transformeeritud lugemite andmetest. Meetodi kiiruse tagab transformeeritud lugemitest sõnastiku loomine, mis võimaldab sarnaseid järjestusi (unikaalseid *k*-meere)

sisaldavaid lugemeid üheaegselt analüüsida. FastGT on samuti kiirem meetod SNV-de määramiseks, mis kasutab eelnevalt teadaolevate SNV-de põhjal koostatud unikaalsete *k*-meeride paare ja määrab genotüübi lugemites esinevate *k*-meeride sageduste põhjal (Pajuste et al., 2017).

1.4.1.2. Koopiaarvu variatsioonide määramine

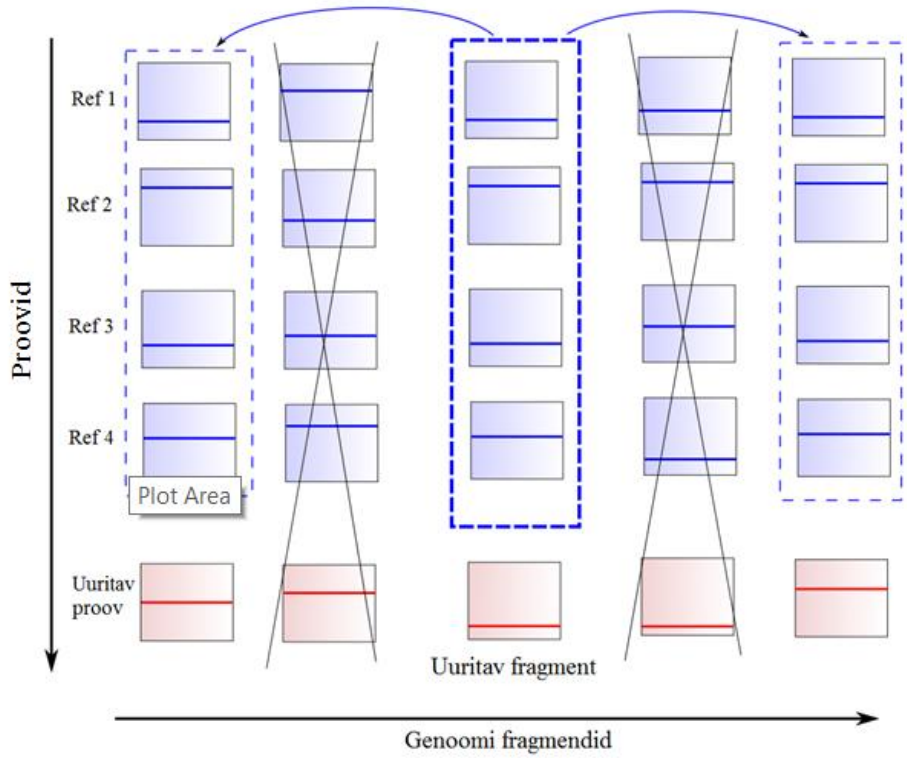
Koopiaarvu variatsioonideks loetakse DNA segmente, millel on uuritavas proovis referentsist erinev koopiaarv. Varasemalt määrati CNV-deks ühekilobaasilised või suuremad variatsioonid (Redon *et al*., 2006). Nüüdseks võetakse arvesse ka väiksemaid variatsioone alates 50 aluspaarist (MacDonald *et al*., 2014). CNV-de tuvastamiseks järgmise põlvkonna sekveneerimisandmetest on neli peamist lähenemist: paarislugemite joondamine (Korbel *et al*., 2007), *split*-lugemid (Zhang *et al*., 2011), katvuse andmete rakendamine (Alkan *et al*., 2009) ja *de novo* assambleerimine (Nijkamp *et al*., 2012). Mitmed meetodid kasutavad neid lähenemisi kombineeritult. (Medvedev *et al.*, 2010; Mills et al., 2011).

Kõige levinum eeltoodud lähenemisviisidest CNV-de määramiseks on katvuse andmete kasutamine, kuna see võimaldab edukamalt tuvastada ka suuremaid CNV-sid ja määrata lisaks CNV asukohale ka täpsema koopiaarvu. Põhiliseks eelduseks järelduste tegemisel on katvuse väärtuse seos genoomi regiooni koopiaarvuga – suurema koopiaarvu korral on katvus oodatust suurem (Yoon *et al.*, 2009). CNV-de määramiseks joondatakse lugemid referentsgenoomile ja arvutatakse katvus eelnevalt kindlaks määratud suurusega akendes. Koopiaarvude täpsemaks tuvastamiseks korrigeeritakse katvus võttes arvesse GC-sisaldust ja kordusjärjestusi. Lõpuks ühendatakse järjestikused sarnase koopiaarvuga genoomi regioonid (Magi *et al.*, 2012).

Sarnaselt SNV-de määramise meetoditele sõltub nende meetodite täpsus ühtlasest katvusest üle genoomi. Hoolimata katvuse normaliseerimisest GC-protsendi suhtes, on valepositiivsete tulemuste arv ebaühtlase katvuse tõttu kõrge (Monlong, *et al*., 2018a). Valepositiivsete tulemuste määra vähendamiseks on välja pakutud katvuse andmete kasutamine mitmetest proovidest. Cn.MOPS töövoog määrab katvuse väärtuse kõikumise põhjal regiooni CNV-ks kui kõrgem või madalam katvus esineb mitmes proovis. Rakendades Poissoni segumudelit igale genoomi regioonile eraldi, eristatakse CNV-dele viitavad katvuse kõikumised mürast, mis võib olla põhjustatud tehnilistest vigadest või järjestuse lokaalsetest omadustest. Regioone, kus ühe proovi katvuse põhjal võiks eeldada koopiaarvu variatsiooni, kuid mitmed proovid viitavad ühtlasele katvuse kõikumisele tehnilistel põhjustel, mitte konkreetse proovi

variatsioonile, ei määrata CNV-deks. (Klambauer *et al*., 2012). Meetodi kasutamist võib piirata võrdluseks vajalike proovide puudumine. Lisaks ei võta meetod arvesse üksikute proovide katvuse kõikumisi, mis takistab haruldaste CNV-de leidmist.

Mitmete proovide informatsioonile toetub ka PopSV metoodika. Erinevalt cn.MOPS töövoost korrigeeritakse katvus GC-sisalduse põhjal. Iga proovi analüüsitakse eraldi, kasutades teisi proove referentsidena. Genoomid fragmenteeritakse ja katvus ühtlustatakse igas fragmendis ja proovis eraldi, toetudes nendele fragmentidele, kus katvuse muster on referentside lõikes uuritava fragmendiga sarnane (Joonis 2). Igale fragmendile arvutatakse Z-skoor, mis näitab kui erinev on katvus uuritavas proovis võrreldes referentsidega. Kui CNV esineb juba mitmes referentsis, siis Z-skoori väärtus väheneb ja neid CNV-sid ei tuvastata (Monlong *et al*., 2018a). PopSV tuvastab varasematest meetoditest paremini harva esinevaid CNV-sid, kuid valepositiivsete määr jääb kõrgeks väiksemate kordusaladel esinevate CNV-de suhtes. Katvusel põhinevad metoodikad ei suuda alati täpselt määrata CNV-de murdekohti (*breakpoints*), mistõttu kindla koopiaarvu määramine väikeste CNV-de korral on keeruline. (Monlong *et al*., 2018b)



Joonis 2. Katvuse normaliseerimiseks sarnase katvuse mustriga fragmentide valimine. (Monlong *et al*., 2018b, kohandatud)

1.4.1.3. Sünnieelne diagnostika

Alates 2011. aastast on saadaval meetodid mitteinvasiivseks sünnieelseks skriininguks, mis kasutavad analüüsiks loote rakuvaba DNA-d ema vereplasmast (Lau *et al*., 2012). Loote DNA moodustab ema vereplasmast 3-20% (Lun *et al*., 2008), mis võimaldab DNA eraldamise ja sekvneerimise järel testida nii aneuploidiate kui ka väiksemate variatsioonide olemasolu. Sekveneeritud lugemid joondatakse ning igale kromosoomile joondunud lugemite arvu võrreldakse referentsgrupiga. Joondamisel filtreeritakse välja lugemid, mis ei joondu üheselt, sisaldavad valepaardumisi (*mismatch* - mittesarnaste nukleotiidide paar joonduses) või indeleid ning katvus korrigeeritakse GC-sisalduse mõju vähendamiseks. Trisoomia (monosoomia) esinemine määratakse Z-skoori põhjal kromosoomides, kuhu on lugemeid joondunud rohkem (vähem), kui euploidsuse korral oodatud (Bayindir *et al*., 2015; Jiang *et al*., 2012).

Sarnaselt SNV-de ja CNV-de määramise meetoditele suurendavad valepositiivsete tulemuste määra aneuploidiate analüüsil katvuse kõrvalekalded oodatud väärtusest. Lisaks joondamise, seknveneerimisvigade ja GC-sisalduse mõjule, tuleb rakuvaba DNA-analüüsimisel arvesse võtta ka DNA fragmentatsiooni mustreid. WGS andmete puhul ei avalda sekveneerimisele eelnev DNA fragmenteerimine katvusele olulist mõju (Benjamini ja Speed, 2012), kuid rakuvaba DNA korral võib fragmenteerumine olla mõjutatud bioloogilistest protsessidest nagu apoptoos (Chandrananda *et al*., 2014). Metoodikate täpsus sõltub ka loote rakuvaba DNA osakaalust veres ja selle määramise täpsusest. Madalamate väärtuste korral on aneuploidiate tuvastamine keeruline, kuna katvuse väärtuse erinevused on väiksemad (Jiang *et al*., 2012). Võttes arvesse, et rakuvaba loote DNA pärineb platsentast, mõjutab aneuploidiate tuvastamist ka geneetiline mosaiiksus. Juhul kui platsenta DNA on euploidne, kuid loote DNA on osaliselt või täielikult aneuploidne, ei suuda mitteinvasiivsed meetodid variatsioone tuvastada (Canick *et al.*, 2013).

Keeruka ja ajakuluka joondamisprotsessi vältimiseks, mis muudaks analüüsid kliiniliseks kasutamiseks lihtsamaks ja kättesaadavamaks, on välja pakutud *k*-meeride katvust rakendav meetod NIPTmer. Sarnaselt FastGT SNV-de määramise metoodikale (Pajuste *et al*., 2017) kasutab NIPTmer varem väljavalitud unikaalseid *k*-meere. Iga kromosoomi suhteline katvus avaldatakse proovist leitud kromosoomispetsiifiliste *k*-meeride arvu ja varasemalt koostatud loendis leiduvate *k*-meeride arvu suhtena. Kuigi *k*-meeride loendite koostamisel jäetakse välja *k*-meerid, mis kattuvad levinud polümorfismidega ja madala komplekssusega aladega (tsentromeerid, telomeerid), ei ole katvus euploidsete referentsproovide kromosoomides

ühtlane. Samas on kromosoomide ja katvuse vahel seos – kindlate kromosoomide katvus oodatust kõrgem või madalam, mis on osaliselt seletatav kromosoomide erineva GC-sisaldusega. NIPTmer kasutab iga kromosoomi oodatud katvuse arvutamisel keskmise katvuse asemel lineaarset mudelit, kus parameetriteks on proovi GC-sisaldus ja ülejäänud kromosoomide suhtelised katvused, mis on leitud referentsproovide põhjal. Samas jääb teatud varieeruvus iga proovi katvuse puhul alles ning valepositiivsete ja -negatiivsete tulemuste hulka mõjutab nii rakuvaba DNA osakaalu määramine, algne sekveneerimissügavus kui ka mosaiiksus (Sauk *et al*., 2018).

Erinevalt aneuploidiatest, mille esinemise oht suureneb ema vanusega, on patogeensete CNV-de esinemine vanusest sõltumatu, mistõttu on nende tuvastamine oluline ka noorematel lapseootel naistel. Katvust kasutatakse CNV-de määramiseks sünnieelses diagnostikas sarnaselt CNV-de määramisele täiskasvanud organismis, kuid see on tavalisest keerukam, kuna loote DNA moodustab uuritavast rakuvabast DNA-st vaid väikese osa. Mida suurem on loote DNA osakaal ja CNV suurus, seda suurema tundlikkusega on võimalik katvuse kõikumisi tuvastada (Zhao *et al*., 2015). Ka kõrge sekveneerimiskatvus tagab parema täpsuse, kuid see muudab analüüsid liiga kulukaks kliiniliseks kasutuseks (Benn ja Cuckle, 2014; Yu *et al*., 2013). Selleks, et väiksemaid CNV-sid edukamalt tuvastada ka madala sekveneerimiskatvuse abil, on oluline katvuse varieeruvuste kõrvaldamine võimalikult täpselt. Üks võimalus on vähendada katvuse varieeruvust lisaks GC-sisalduse mõju korrigeerimisele ka referentsproovide põhjal. Uurides katvuse varieerumist peakomponentanalüüsi abil euploidsetes referentsides, saab korrigeerimisel arvesse võtta esimeste peakomponentide kirjeldatud varieeruvust katvuse andmetes, mis CNV-dele ei viita (Zhao *et al*., 2015).

1.4.2. Geenieksпressiooni analüüsid

Katvuse andmeid RNA-sekveneerimisel saab kasutada nii transkribeeritud järjestuste tuvastamiseks kui ka nende ekspressioonitaseme määramiseks. Diferentsiaalse ekspressiooni analüüsi eesmärgiks on tuvastada erinevusi geenieksпressioonis, mis võivad olla tingitud näiteks arengutasemest või ravimi manustamise mõjust. Analüüsiks vajalike lugemite arv on RNA-sekveneerimisel määratud kõige madalamalt ekspresseeritud transkriptide põhjal (Sims *et al*., 2014). Selleks, et huvipakkuvaid, kuid madalamalt ekspresseeritud transkripte nagu näiteks mRNA analüüsida, tuleb eemaldada suuremal hulgal esinevad RNA-d, näiteks rRNA järjestused, mis moodustavad 90% kogu RNA-st imetajate rakkudes. mRNA võib teistest RNA-dest eraldada polü-A-sabade abil, mis kinnituvad immobiliseeritud deoksütümidiini oligojärjestustele ning ülejäänud RNA-d pestakse proovist välja (Kingston, 2001).

Transkribeeritud järjestuste määramiseks RNA-sekveneerimisandmetest pöördtranskribeeritakse RNA järjestused cDNA-ks, sekveneeritakse ning lugemid joondatakse referentsgenoomile. Cufflinks tarkvaraprogramm assambleerib joondunud lugemid transkriptideks annoteeritud transkriptide põhjal või *de novo* ning ekspressioonitase määratakse assambleeritud transkriptidele joondunud lugemite arvu põhjal. Joondamisel kasutatakse TopHat tarkvara (Trapnell *et* al., 2009), mis võimaldab lugemi erinevad osad joondada erinevatele eksonitele jättes nende vahele intronjärjestuse tühimiku. Tänu sellele saab hinnata erinevate alternatiivselt splaissitud RNA-de (isovormide) hulka. Juhul kui ühe geeni põhjal on transkribeeritud mitu erinevat isovormi, võib fragment joonduda ühele geenile erinevalt. Sel juhul on joondatud fragmendil sõltuvalt erinevast joondusest mitu võimalikku pikkust. Fragmentide pikkuste jaotusele toetudes hinnatakse, millisest isovormist on fragment suurima tõenäosusega pärit ning ekspressioonitasemete määramiseks valitakse kõige tõenäolisem lahendus, mis vastab kõige paremini andmetes esinevatele fragmentidele ning nende pikkustele. Ekspressioonitase esitatakse FPKM (*fragments per kilobase of transcript per million fragments mapped*) ühikutes, mis normaliseerib katvuse väärtuse transkripti pikkuse ja kõigi joondatud fragmentide suhtes (Trapnell *et al*., 2010). Alternatiivne meetod on määrata ekspressioonitase ilma eelneva transkriptide assambleerimiseta, misjuhul katvus leitakse lugemite põhjal, mis on eksonile joondunud (Anders *et al*., 2013).

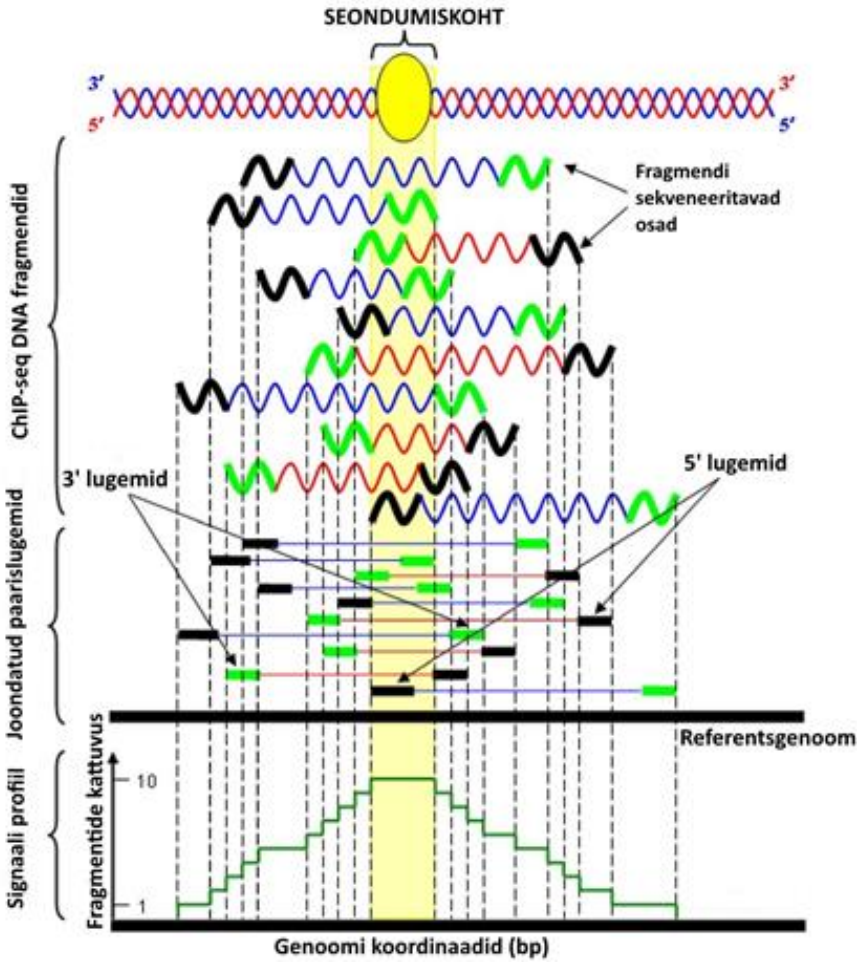
Sekveneerimistehnoloogia areng tagab järjest kiiremini uusi RNA-seq andmeid, kuid nende analüüsimine on ajakulukas, lugemeid ei ole võimalik alati üheselt joondada ning alternatiivselt splaissitud RNA-d muudavad lugemite joondamise veelgi keerulisemaks. Sarnaselt DNA sekveneerimisandmete analüüsile kasutatakse RNA andmete kiiremaks analüüsimiseks ja joondusel tekkivate küsitavuste vältimiseks joondusvabasid metoodikaid. Sailfish hindab ekspressioonitaset *k*-meeride sageduse põhjal toorlugemites (Patro *et al*., 2014). Transkriptide *k*-meerid määratakse referentstranskriptide põhjal, seega ei võimalda meetod tuvastada uusi annoteerimata transkripte, kuid on tunduvalt kiirem kui joondamist kasutavad meetodid ning võimaldab paremini vältida sekveneerimisvigadest tulenevaid mõjusid. Kui joondamisel mõjutavad sekveneerimisvead kogu lugemit, põhjustades näiteks selle joondumise valele asukohale, mõjutavad antud meetodi puhul vead ainult nendega kattuvaid *k*-meere ning ülejäänud lugemi *k*-meerid saab tuvastada vigadeta. Kallisto meetod kasutab unikaalseid *k*-meere ning koostab transkriptide *k*-meeridest de Bruijn graafi, mille põhjal saab määrata millisest transkriptist või isovormist lugemitest leitud *k*-meerid pärinevad ning seeläbi tagada täpsemad ekspressioonitaseme hinnangud (Bray *et al*., 2016).

1.4.3. DNA-valk interaktsioonikohtade määramine

Valgu ja DNA interaktsioonikohtade analüüsimise eesmärgiks on leida need genoomi regioonid, millega uuritav valk seondub. Üheks sagedasemaks rakenduseks on transkriptsioonifaktorite seondumiskohtade tuvastamine, mis võimaldab uurida, milliste geenide regulatsioonis faktorid osalevad. Oluliseks uurimisküsimuseks on ka tervete ja vähirakkude geeniregulatsiooni võrdlus. ChIP-seq tööprotsessis eelneb sekveneerimisele kromatiini fragmenteerimine ja immunosadestamine, mille käigus sadestatakse valguspetsiifiliste antikehade abil välja need kromatiini fragmendid, millega uuritav valk on seondunud. Seejärel fragmendid sekveneeritakse ja lugemid joondatakse referentsgenoomile (Johnson *et al*., 2007)

Seondumiskohtade tuvastamiseks on eelkõige oluline määrata tõeste signaalide täpne asukoht. Sekveneerimise tulemusel saadakse lugemid, millest suurem osa esindab neid regioone genoomis, kuhu valk on seondunud. Seondumiskohad on lugemites aga erinevatel positsioonidel ning seega on täpse interaktsioonikoha tuvastamiseks oluline leida regioon, kus katvus on maksimaalne. Samuti tuleb tõesed signaalid eristada mürast, mida põhjustavad need lugemid, mis interaktsioonikohti ei sisalda. Sõltuvalt immunosadestamise edukusest, võib valesignaali anda erinev hulk lugemeid. SIPeS (*Site Identification from Paired-end Sequencing*) meetod kasutab seondumiskohtade tuvastamiseks paarislugemeid. Joondatud lugemid määravad fragmendid, mille kattumise põhjal tuvastatakse signaali profiililt seondumise asukoht (Joonis 3) (Wang *et* al., 2010).

Üksiklugemite kasutamisel mõjutavad haripunkti tuvastamist ahela spetsiifilised kõrvalekalded. Positiivse ahela lugemid esindavad seondumiskohta sisaldava fragmendi 5’ otsa ning negatiivse ahela lugemid fragmendi 3’ otsa. Haripunkti leidmiseks on sel juhul kaks võimalust: (1) lugemeid nihutatakse ahela 3’ suunas või (2) lugemeid pikendatakse algse fragmendi pikkuseni (Wilbanks ja Facciotti, 2010). Nihutamisel leitakse signaali profiililt kahele ahelale vastavate lugemite haripunktid ja nendevaheline distants ning nihutatakse haripunktid poole distantsisuuruse võrra ahelate 3’ suunas (Zhang *et al*., 2008). Lugemite pikendamisel kasutatakse keskmist fragmendi pikkust ChIP raamatukogus ning signaali profiilid koostatakse pikendatud lugemite põhjal (Rozowsky *et al*., 2009).



Joonis 3. Valgu ja DNA seondumiskoha leidmine. Paarislugemid määravad fragmendid, mille kattumiste põhjal koostatakse signaali profiil. Seondumiskohana tuvastatakse profiili haripunkt, kus fragmentide kattumise väärtus on suurim. (Wang *et al*., 2010, kohandatud)

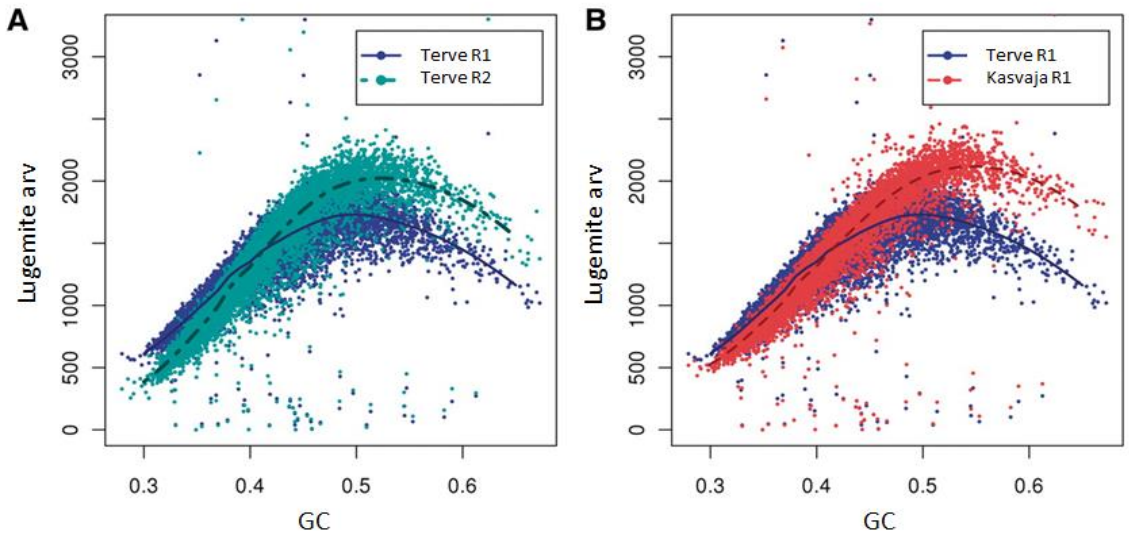
ChIP-seq analüüsil tuleb katvuse põhjal järelduste tegemisel lisaks tavapärastele mõjuteguritele arvesse võtta ka kromatiini struktuuri. Aktiivselt transkribeeritav ja lõdvemalt pakitud eukromatiin fragmenteerub sonikeerimisel edukamalt kui tihedalt pakitud heterokromatiin, mistõttu on eukromatiin pärast sobiva suurusega fragmentide valimist DNA raamatukogus rohkem esindatud (Auerbach et al., 2009). Seondumiskohtade lõplikuks tuvastamiseks võrreldakse leitud regioone kontroll-DNA andmetega, milleks on uuritava ChIP prooviga samadel tingimustel fragmenteeritud ja sekveneeritud, kuid immunosadestamata kromatiin. Iga tuvastatud seondumiskoha katvusest lahutatakse sama regiooni kontrollproovi katvus. See on vajalik valepositiivsete tulemuste määra vähendamiseks, mis tulenevad katvuse tehnilistest kõrvalekalletest ning põhineb eeldusel, et kontrollis ja uuritavas proovis esinevad sarnased katvuse kõikumised. See ei pruugi alati täielikult tõele vastata. BIDCHIPS meetod eristab tõese signaali teistest katvust mõjutavatest teguritest. Lisaks kontroll-DNA signaalile võtab BIDCHIPS arvesse ka GC-sisalduse,

kromatiini struktuuri, joondamise ja IgG antikehaga immunosadestatud kontrolli signaalid (Ramachandran *et al*., 2015).

1.5. GC-sisaldus

Varasemates töödes (Benjamini ja Speed, 2012; Cheung *et al*., 2011; Dohm *et al*., 2008) on katvuse väärtuse mõjutajana kõige rohkem tähelepanu pööratud GC-sisaldusele uuritavas proovis. Erinevalt joondamisest, mille mõju katvusele on võimalik vältida näiteks joondusvabade meetoditega, või sekveneerimisvigadest, mille mõju on võimalik vähendada tehnoloogiate täiustamisega, on GC-sisalduse mõju vähendamiseks oluline katvuse korrigeerimine sekveneerimisjärgselt. Esimesed uuringud tuvastasid lineaarse seose, kus G ja C nukleotiidide osakaalu kasvades suureneb ka katvus (Dohm *et al*., 2008). Hilisemad analüüsid on näidanud GC-sisalduse ja katvuse vahel unimodaalset seost, kus katvuse väärtus on kõrgeim, kui GC-sisaldus ümbritsevates regioonides on 40-55%. Nii kõrge GC- kui ka AT-sisaldusega lugemid on sekveneerimisandmetes seega alaesindatud (Benjamini ja Speed, 2012).

G ja C nukleotiidide osakaal ei ole erinevates genoomi regioonides ühtlane ja on tihti korreleeritud funktsionaalsusega. Oodatust madalama katvusega on näiteks mõned promootoralad ja geenide esimesed eksonid, kus on vastavalt ka GC-protsent kõrgem (Cheung *et al*., 2011; Ross *et al*., 2013). Probleemi muudab keerulisemaks asjaolu, et GC-protsendi ja katvuse seose omadused varieeruvad nii korduvates eksperimentides kui ka erinevates DNA raamatukogudes, mis on koostatud ühest proovist. Erinevused on nii seose haripunktis kui ka katvuse varieeruvuse ulatuses (Joonis 4) (Benjamini ja Speed, 2012).



Joonis 4. GC-sisalduse ja katvuse vaheline seos (A) erinevates DNA raamatukogudes, mis on koostatud ühe proovi põhjal; (B) vähkkasvaja ja vastava koe terves proovis. GC-sisaldus on arvutatud 10 kb mittekattuvates järjestikustes lõikudes (*bins*), mis on valitud juhuslikkuse alusel 1. kromosoomist. Värvidega on tähistatud sekveneerimisandmestikust leitud katvuse väärtused vastava GC-sisalduse juures ning LOESS (*locally estimated scatterplot smoothing*) jooned tähistavad ennustatud seost. Legendi tähised R1 ja R2 viitavad sama algmaterjali põhjal koostatud kahele erinevale DNA raamatukogule. (Benjamini ja Speed, 2012, kohandatud)

Kõrvalekalded katvuses sõltuvalt GC-sisaldusest võivad tekkida mitmetes protsessi osades, mistõttu on keeruline leida ühte konkreetset põhjust, mis GC-sisaldusest tulenevalt varieeruvusi tekitab. Näiteks võib mõju avaldada sekveneerimisele eelnev sobiva suurusega fragmentide valik. On leitud, et AT-rikkad fragmendid võivad olla DNA raamatukogus alaesindatud, kuna sobiva suurusega DNA fragmente sisaldava geelilõigu sulatamisel puhvris denatureeruvad need järjestused suurema tõenäosusega. Nendele fragmentidele vastavad genoomi regioonid on sel juhul pärast sekveneerimist madalama katvusega. Sulatamistemperatuuri langetamine aga kõrvaldab suuremal määral selle efekti. (Quail *et al*., 2008). RNA-sekveneerimisel mõjutab cDNA sünteesiks kasutatavate heksameersete praimerite mittejuhuslik seondumine RNA-ga sekveneerimislugemite 5’-otsa nukleotiidset koostist, mistõttu lugemid ei jaotu ekspresseeritavatele genoomiregioonidele ühtlaselt (Hansen *et al*., 2010).

GC-sisaldusest sõltuvalt on aga suurim mõju katvuse kõikumistele sekveneerimisele eelneval amplifikatsioonil (Aird *et al*., 2011; Benjamini ja Speed, 2012). PCR-i abil sekveneeritavate järjestuste paljundamine on vajalik algmaterjali koguse suurendamiseks ja nende fragmentide paljundamiseks, millele on adapterid edukalt mõlemasse otsa ligeerunud. PCR-i tsükleid viiakse läbi küll vähe (paarislugemite puhul 10-12) (Bentley *et al*., 2008), kuid hoolimata

sellest on väga kõrge ja madala G ja C nukleotiidide osakaaluga lugemid pärast amplifitseerimist DNA raamatukogus alaesindatud. Üks lahendus amplifitseerimisel tekkivate kõrvalekallete vältimiseks on PCR-i vaba sekveneerimine (Kozarewa *et al*., 2009). See ei pruugi aga võimalik olla, kui algmaterjali hulk on väike ja amplifikatsioon on vältimatu sekveneerimiseks vajaliku hulga DNA tagamiseks. Alternatiiv on PCR-i tingimuste muutmine, mis tagab kõrge GC-sisaldusega fragmentide eduka amplifikatsiooni, samas jäävad AT-rikkad fragmendid siiski alaesindatuks (Aird *et al*., 2011). PCR-i vaba sekveneerimine küll vähendab, aga ei eemalda GC-protsendi mõju täielikult ning Illumina sekveneerimisel jääb alles ka sildamplifikatsiooni etapi mõju (Ross *et al*., 2013), mistõttu on edasistel analüüsidel valepositiivsete ja -negatiivsete tulemuste vältimiseks vajalik katvuse väärtuse korrigeerimine.

GC-sisalduse mõju hindamisel ja korrigeerimisel on nüüdseks oluline osa järgmise põlvkonna sekveneerimise uuringutes. Enamik meetoditest kasutab korrigeerimiseks ühesugust lahendust: katvuse väärtused ning G ja C nukleotiidide arv määratakse valitud suurusega mittekattuvates lõikudes ning leitakse keskmine katvus, mis iga GC-sisalduse juures esineb. Lõigu pikkus valitakse GC-sisalduse arvutamiseks enamasti vastavalt hilisemate analüüside vajadustele (näiteks keskmise katvuse arvutamiseks CNV-de määramisel). GC-sisalduse põhjal määratud hinnanguid kasutatakse seejärel katvuse korrigeerimiseks. (Boeva *et al*., 2011; Miller *et al*., 2011; Yoon *et al*., 2009). Kuigi need meetodid korrigeerivad suures osas GC-sisalduse mõju, vaadatakse kohati mööda olulistest seose omadustest, nagu unimodaalsus. Seda näiteks juhul, kui GC-sisalduse arvutamiseks kasutatakse lõike, mis on pikemad kui 10 kb, sest sellistes akendes on inimese genoomis GC-sisaldus harva 50%-st suurem. Sel juhul võib ekslikult järeldada, et GC-sisalduse ja katvuse vaheline seos on lineaarne (Benjamini ja Speed, 2012).

Benjamini ja Speed uurisid GC-sisalduse ja katvuse vahelist seost senisest põhjalikumalt. Nad koostasid mudeli, mis võimaldab korrigeerida iga nukleotiidi katvuse eraldi. Katvuse ennustamiseks valitakse genoomist positsioonid, mis jaotatakse GC-sisalduse põhjal gruppidesse. Mudeli põhjal ennustatud katvuse väärtus leitakse jagades gruppi kuuluvatele positsioonidele joondunud lugemite arvu grupis olevate positsioonide arvuga. Ennustatud katvus vastab kõige paremini reaalsetele andmetele, kui GC-sisaldus arvutada kogu fragmendis, mille määravad paarislugemid ning nende vahele jääv genoomiala. Fragmendi GC-sisalduse ja katvuse seos toetab varasemaid (Aird *et al*., 2011) järeldusi PCR-i kui suure mõjuteguri kohta ning samuti leidsid nad seose fragmendi GC-sisalduse ja

sekveneerimisvigade tõenäosuse vahel (Benjamini ja Speed, 2012). Kuigi mudel ennustab katvuse varieeruvust paremini, kui varasemad meetodid, jäävad varieeruvused katvuse väärtuses väiksemal määral alles ka pärast katvuse korrigeerimist PCR-i vaba sekveneerimise andmetes ning katvuse ennustamine on piiratud nende positsioonidega, kuhu on joondunud lugemid unikaalselt (Benjamini ja Speed, 2012).

2. EKSPERIMENTAALOSA

2.1. Töö eesmärgid

Töö eesmärgiks oli hinnata, kas GC-sisaldus ja *k*-meeri asukoht genoomis mõjutavad *k*-meeride katvuse (edaspidi katvuse) väärtusi. Täpsemad eesmärgid olid:

- vaadata, kas katvuse väärtuse muutused järjestikustel *k*-meeridel indiviidide lõikes on sarnased või erinevad;
- uurida seost katvuse ja GC-sisalduse vahel;
    - leida optimaalne akna pikkus GC-sisalduse arvutamiseks, et korrelatsioon GC-sisalduse ja katvuse vahel oleks suurim;
    - teha kindlaks, kas Illumina Platinum indiviidi (NA12877) ja Eesti Geenivaramu (EGV) indiviidide andmestikus on optimaalne akna pikkus erinev;
- koostada lineaarne regressioonimudel, mille põhjal analüüsida, kui suure osa katvuse varieerumisest seletab GC-sisaldus;
- uurida, kas *k*-meeri kromosoomi ja genoomipositisooni parameetrite lisamisel kirjeldab mudel katvuse varieeruvusi paremini.

2.2. Materjal ja metoodika

2.2.1. Andmed

GC-sisaldus arvutati inimese referentsgenoomi versiooni GRCh37 (*release* 75) autosoomide järjestuse põhjal[2]. Katvuse analüüsimiseks kasutati Illumina Platinum indiviidi (NA12877) ja 50 EGV indiviidi (25 meest ja 25 naist) SNV-de *k*-meeride (unikaalsed *k*-meerid, mis sisaldavad varasemalt teadaolevaid ühenukleotiidilisi variatsioone) katvusi.

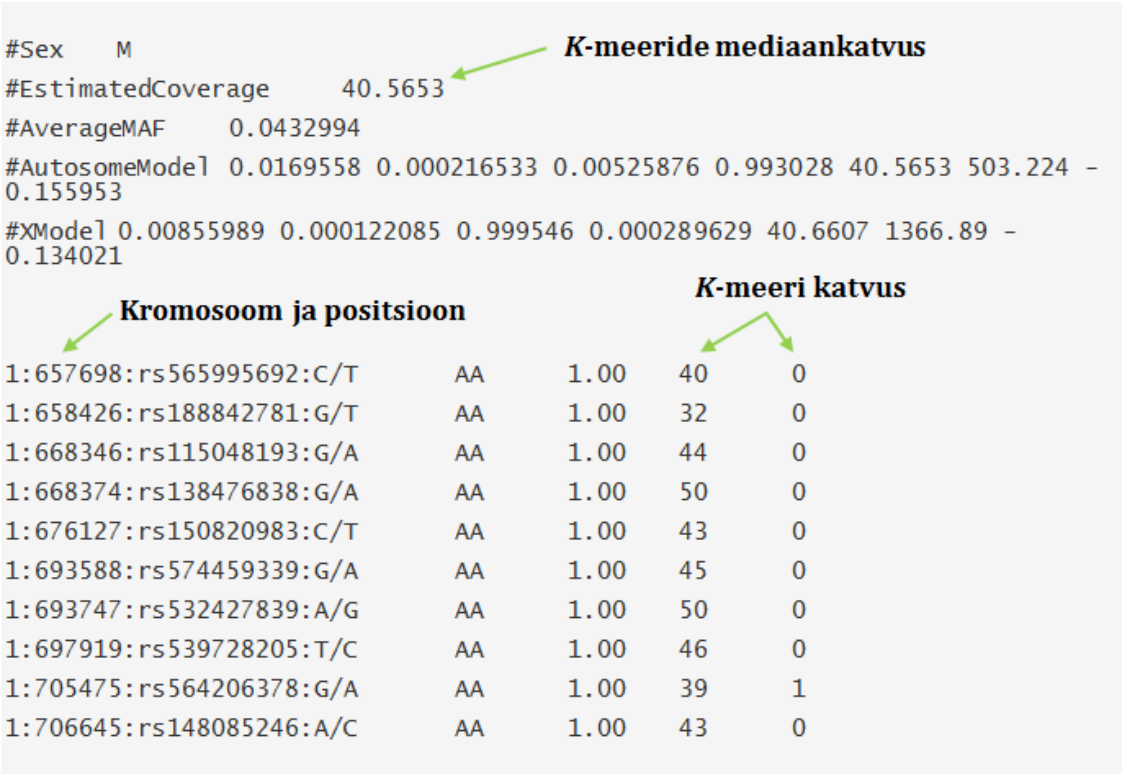
SNV-de *k*-meerid leidis referentsgenoomist Tartu Ülikooli molekulaar- ja rakubioloogia instituudi bioinformaatika õppetooli nooremteadur Fanny-Dhelia Pajuste. *K*-meeri pikkus oli 25 bp (*base pair*) ning *k*-meerid leiti paaridena, millest üks esindab referentsgenoomi ja teine alternatiivset (SNV-d sisaldavat) järjestust. Iga SNV jaoks sai algselt koostada 25 SNV-d sisaldavat *k*-meeri paari, millest filtreeriti välja paarid, mis (1) sisaldasid enam kui ühte SNV-d; (2) ei olnud genoomis unikaalsed ning (3) olid mõne EGV indiviidi puhul vähemalt kolm

[2] ftp://ftp.ensembl.org/pub/release-75/fasta/homo_sapiens/dna/, 24.05.2019

korda suurema katvusega, kui *k*-meeride mediaankatvus. Edasi valiti allesjäänud *k*-meeri paaridest kuni kolm üksteisest kõige kaugemal asuvat paari, millest mediaankatvusele kõige sarnasema katvusega *k*-meeri paari kasutati genotüpiseerimiseks. Lõpuks filtreeriti välja SNV-d, millele määrati bialleelsest genotüübist erinev genotüüp (Pajuste *et al*., 2017).

Filtreeritud *k*-meeri andmebaasi põhjal leidis Platinum indiviidi (NA12877) ja 50 EGV indiviidi joondamata sekveneerimislugemitest *k*-meeride katvused ja SNV genotüübid FastGT tarkvarapaketi (Pajuste *et al*., 2017) abil Tartu Ülikooli molekulaar- ja rakubioloogia instituudi bioinformaatika õppetooli professor Maido Remm. Genotüpiseerimisel kasutati iga SNV puhul filtreeritud andmebaasis olevast kolmest *k*-meeri paarist mediaankatvusele kõige sarnasema katvusega *k*-meeri paari.

K-meeride paare oli iga EGV indiviidi ja Platinum indiviidi failides 29 041 678. *K*-meeride failidest kasutati katvuse analüüsimiseks *k*-meeri kromosoomi ja positsiooni informatsiooni, *k*-meeride mediaankatvust ning *k*-meeride katvusi (Joonis 5). Positsioon viitab *k*-meeris sisalduva SNV positsioonile kromosoomis ning katvust kasutati kahele alleelile vastavate *k*-meeride katvuste summana. Kõik analüüsid viidi läbi normaliseeritud katvustega, mis leiti jagades igas proovis kõikide *k*-meeride katvused läbi *k*-meeride poole mediaankatvusega $C = \frac{C_{k-meer}}{\frac{1}{2} * C_{mediaan}}$.



Joonis 5. Näide *k*-meeride katvuste faili algusest. Platinum indiviidi andmed[3].

2.2.2. *K*-meeri katvuste kõikumiste hindamine

Selleks, et näha, kas erinevate indiviidide *k*-meeride katvused muutuvad lokaalselt sarnaselt, koostati programmid Python programmeerimiskeeles (versioon 3.6). Iga *k*-meeri katvuse standardhälve arvutati 50 EGV indiviidi katvuste põhjal *statistics* mooduli[4] *stdev()* funktsiooni abil. Selleks, et saada ülevaade *k*-meeride katvuste varieeruvusest indiviiditi koostati standardhälvete jaotus, kus igale ühe komakohani ümardatud standardhälbele vastab antud standardhälbega *k*-meeride arv. Edasi valiti genoomist regioonid, kus *k*-meeride katvused ei olnud kasutatud 50 indiviidi puhul suure varieeruvusega, et vaadata, kas katvuse väärtus muutub nendes piirkondades eri indiviididel sarnaselt. Selleks leiti genoomist kindla etteantud pikkusega piirkonnad (150 bp, 500 bp, 1000 bp, 10 000 bp), milles ühegi *k*-meeri standardhälve ei olnud suurem kui 0,5 ning valiti nendest juhuslikult piirkonnad varieeruvuste visuaalseks hindamiseks. Standardhälbe piirmäär valiti standardhälvete jaotuse põhjal, jättes välja väiksema osa *k*-meere, kus katvuse standardhälve on keskmisest suurem. Katvuse varieerumise hindamiseks koostati graafikud tarkvarapaketi Microsoft Office programmis Excel.

[3] http://bioinfo.ut.ee/FastGT/index.php?r=site/page&view=manual, 21.05.2019
[4] https://docs.python.org/3/library/statistics.html, 21.05.2019

2.2.3. GC-sisalduse ja katvuse vaheline seos, optimaalse akna suuruse leidmine

GC-sisalduse arvutamiseks koostati programm Python programmeerimiskeeles (versioon 3.6). GC-sisaldused arvutati akendes suurusega 101 bp, 301 bp ja 1001 bp, leides referentsgenoomist G ja C nukleotiidide osakaalu $GC\% = \frac{N_{G+C}}{N_{A+T+G+C}} * 100\%$. Akna algus- ja lõpp-positsiooni leidmiseks kasutati *k*-meeride failides olevaid positsioone: akna keskmiseks nukleotiidiks oli SNV positsioon kromosoomis. Määramata nukleotiide (tähistus N) protsendi arvutamisel arvesse ei võetud. Seose omaduste hindamiseks koostati graafikud Excelis ning seose tugevuse määramiseks kasutati Pearsoni korrelatsioonikordajat, mis arvutati Pythoni SciPy[5] paketi *pearsonr()* käsuga. Analüüs viidi eraldi läbi Platinum indiviidi ja EGV indiviidide katvusi kasutades. EGV indiviidide katvusi analüüsiti nii iga indiviidi andmete põhjal eraldi (10 indiviidi) kui ka *k*-meeri keskmiste katvuste põhjal (50 indiviidi). Esimeste leitud korrelatsioonikordajate põhjal korrati analüüsi optimaalse akna pikkuse täpsemaks määramiseks, kasutades GC-sisalduse arvutamiseks 101 bp kuni 301 bp suuruseid aknaid (10 bp sammuga).

2.2.4. Lineaarse regressioonimudeli koostamine

Lineaarsete regressioonimudelite koostatamiseks kirjutati tarkvarapaketi R programm (versioon 3.5.2)[6]. Mudelite koostamiseks kasutati ainult EGV indiviidide andmeid ning valiti välja *k*-meerid, mis oleks üksteisest vähemalt 251 bp kaugusel, et GC-sisalduse arvutamiseks kasutatavad aknad ei kattuks ning andmed ei oleks üksteisest sõltuvad. Pärast valikut jäi andmestikku 7 329 269 SNV *k*-meeri paari. GC-sisaldus arvutati referentsgenoomi põhjal 251 bp pikkuses aknas. Akna keskmiseks nukleotiidiks oli SNV positsioon kromosoomis. Mudelite funktsioontunnuseks *y* olid *k*-meeride keskmised katvused, mis arvutati 50 EGV indiviidi katvuste põhjal. Argumenttunnusteks olid GC-sisaldus, kromosoomi number ja positsioon. Koostati neli mudelit erinevate argumenttunnuste kombinatsioonidega: (1) GC-sisaldus; (2) GC-sisaldus, kromosoomi number; (3) GC-sisaldus, positsioon; (4) GC-sisaldus, kromosoomi number, positsioon. Kromosoomi number sisestati mudelisse faktortunnusena. *K*-meeri positsioonid, mis varasemalt vastasid positsioonile kromosoomis, teisendati genoomi positsioonideks. 2. kromosoomis asuvatele *k*-meeride koordinaatidele liideti 1. kromosoomi pikkus; 3. kromosoomi koordinaatidele 1. ja 2. kromosoomi pikkus jne. Positsioonid ja GC-

[5] http://www.scipy.org/, 21.05.2019
[6] https://www.R-project.org/, 21.05.2019

sisaldused lisati mudelisse kuupsplainina, *splines* paketi[7] *bs()* käsuga. Kuupsplain on sile polünoome sisaldav funktsioon, mis koosneb sisemiste sõlmede poolt määratud lõikudel koostatud kolmanda astme polünoomidest. Sõlmepunktides vastavad polünoomid teatud sileduse tingimustele[8]. Splaini sõlmede arv vastab vabadusastmete arvule, millest on lahutatud polünoomi aste (kuupsplaini puhul 3)[9]. Positsioonide splaini vabadusastmete arv oli 200 ning GC-sisalduse splaini vabadusastmete arv 7. Erinevate mudelite võrdlemiseks kasutati dispersioonianalüüsi testi (ANOVA).

Selleks, et hinnata, kas varasemalt leitud optimaalne akna pikkus GC-sisalduse arvutamiseks (251 bp) tagab ka kõige paremini katvuse varieeruvust kirjeldava mudeli (suurima determinatsioonikordaja), koostati kontrolliks ainult GC-sisalduse parameetrit sisaldavad mudelid ka optimaalsest aknast väiksemas (151 bp) ja suuremas (501 bp) aknas arvutatud GC-sisalduste põhjal.

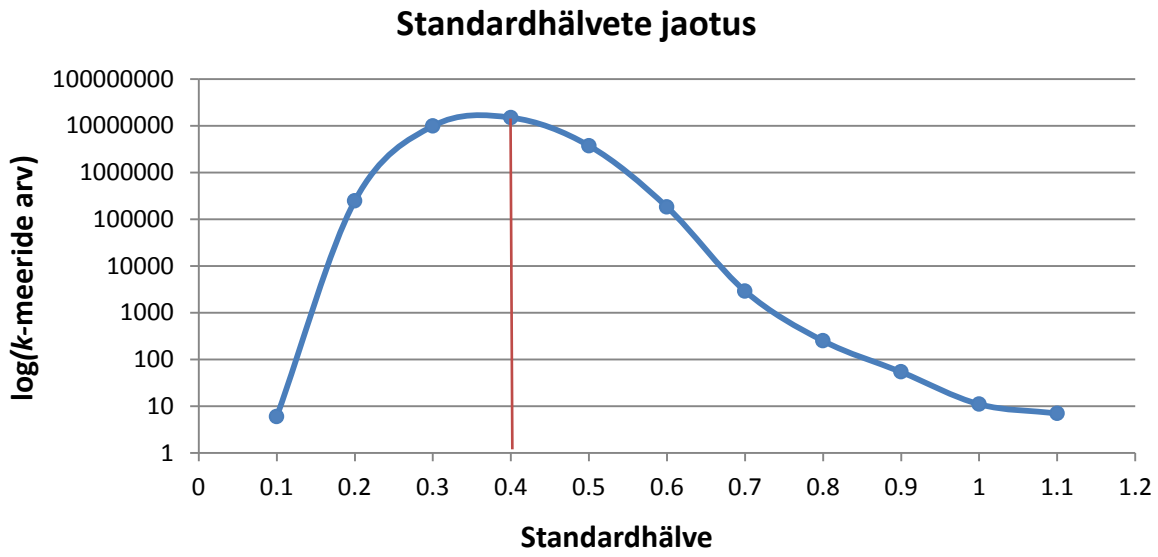
2.3. Tulemused

2.3.1. EGV indiviidide *k*-meeri katvuste kõikumised

Katvuste standardhälvete jaotuse põhjal (Joonis 6) valiti katvuse varieerumiste graafiku koostamiseks standardhälbe piirmääraks 0,5, jättes välja *k*-meerid, kus indiviidide katvuste varieerumise ulatus on keskmisest suurem.

Katvuse kõikumised indiviidide lõikes olid erinevad – ühe *k*-meeri katvus võis erinevatel indiviididel olla nii oodatust kõrgem kui ka madalam. Ühesuguseid kõikumisi esines ainult üksikute positsioonide kaupa. Joonisel 8 LISAS 1 on välja toodud nelja meessoost EGV indiviidi katvuse kõikumised 10 000 bp regioonis. Mustade punktidega märgitud positsioonidel võib näha mõnda näidet katvuse väärtuse ühesugusest kõikumisest kahel indiviidil, kuid sarnaseid trende suurema piirkonna lõikes ja kõikidel indiviididel korraga ei ole. Ebaühtlased kõikumised esinesid ka lühemate piirkondade ja teiste indiviidide andmete põhjal koostatud graafikutel.

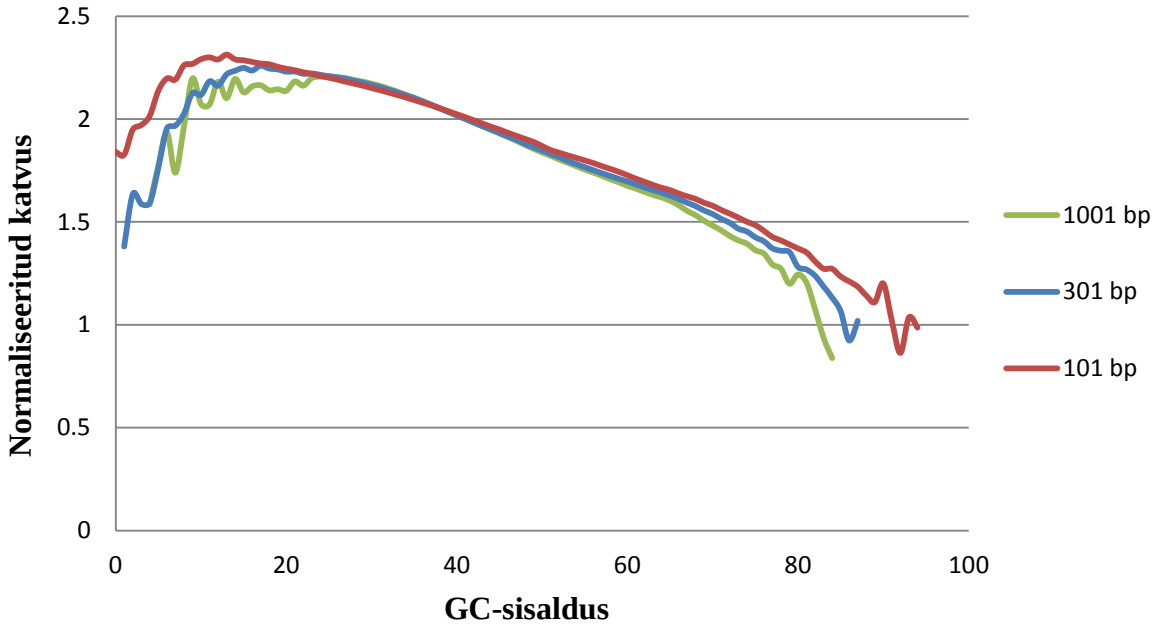
[7] https://www.R-project.org/, 21.05.2019
[8] http://www.tlu.ee/~tonu/Arvmeet/Splkonsp.pdf, 26.05.2019
[9] https://www.rdocumentation.org/packages/splines/versions/3.6.0/topics/bs, 26.05.2019



Joonis 6. EGV indiviidide katvuse andmete standardhälvete jaotus. *K*-meeride arv *y*-teljel on logaritmitud ning horisontaalne joon näitab jaotuse optimumi.

2.3.2. GC-sisalduse ja katvuse seos, optimaalne akna suurus

GC-sisalduse ja katvuse vahel esineb unimodaalne seos – optimumist kõrgemate või madalamate GC-sisalduste juures on katvus madalam. Platinum indiviidi katvuse ja GC-sisalduse seose optimum on ligikaudu 20% juures (Joonis 7) ning EGV indiviidide katvus on kõrgeim ligikaudu 25% juures. Korrelatsioonikordaja arvutamisel jäeti välja *k*-meerid, mille ümbruses on GC-sisaldus väiksem kui 20%, kuna vastava GC-sisaldusega *k*-meere oli vähe ning seos ei olnud selles piirkonnas lineaarne. Optimaalne akna suurus GC-sisalduse arvutamiseks on Platinum indiviidil on 171 bp (R^2 = 0,223). Optimaalne akna suurus 10 EGV indiviidil, kelle andmete analüüs viidi läbi eraldi, varieerus 241 bp-st 271 bp-ni ning EGV indiviidide keskmiste katvuste põhjal arvutatud optimaalne akna suurus oli 251 (R^2 = 0,346).



Joonis 7. GC-sisalduse ja katvuse seos erinevate akna pikkuste korral. *Y*-teljel on Platinum indiviidi *k*-meeride normaliseeritud keskmised katvused, *x*-teljel GC-sisaldus, joon illusreerib GC-sisalduse ja katvuse seost erinevates akna pikkustes arvutatud GC-sisalduste korral.

2.3.3. Lineaarne regressioonimudel

Kuigi muutus oli väike, kasvas lineaarse regressioonimudeli determinatsioonikordaja väärtus parameetrite lisamisel,. Mudeli, kus argumenttunnuseks oli ainult GC-sisaldus (arvutatud 251 bp aknas), kohandatud (*adjusted*) R^2 oli 0,3132 ehk mudel kirjeldas 31,32% keskmiste katvuste varieeruvusest. GC-sisalduse ja kromosoomi numbri, GC-sisalduse ja positsioonide ning GC-sisalduse, kromosoomi numbri ja positsioonide parameetritega mudelite kohandatud determinatsioonikordajad olid vastavalt 0,3138, 0,3148 ning 0,315. Mudelite valemid, kohandatud determinatsioonikordajad ning näited argumenttunnuste kordajatest koos usaldusintervallide ja *p*-väärtustega on LISAS 2. Samuti näitasid ANOVA testid, et kolme parameetriga mudel on ühe või kahe parameetriga mudelitest statistiliselt oluliselt parem – *p*-väärtused $<2*10^{-16}$. Optimaalsest aknast väiksemas (151 bp) ja suuremas (501 bp) aknas leitud GC-sisalduste põhjal koostatud mudelite kohandatud determinatsioonikordajad olid vastavalt 0,284 ja 0,274.

2.4. Arutelu

Varasemalt on GC-sisaldusest tulenevate katvuse kõrvalekallete peamise põhjusena välja toodud sekveneerimisele eelnev PCR. Samas on ka PCR-i vaba sekveneerimise puhul *k*-meeride katvuses varieeruvus, millest GC-sisaldus kirjeldab koostatud mudeli põhjal

ligikaudu 30%. Seega tulenevad GC-sisaldusest põhjustatud kõrvalekalded olulisel määral ka teistest teguritest ja katvuse andmete rakendamiseks edasistel analüüsidel on vajalik GC-sisalduse mõju korrigeerida.

Benjamini ja Speed näitasid, et GC-sisalduse mõju korrigeerimine mitte juhusliku, vaid kindla valitud suurusega aknas uuritava regiooni ümber on olulise tähtsusega. Joondatud lugemite puhul korrigeerib varieeruvust kõige paremini mudel, kus GC-sisaldus on arvutatud paarislugemite poolt määratud fragmendi ulatuses (Benjamini ja Speed, 2012). Seda järeldust toetavad ka siinse töö tulemused: optimaalne akna suurus GC-sisalduse arvutamiseks on suurem lugemi pikkusest, mis viitab, et GC-sisalduse mõju katvusele ei tulene ainult lugemi järjestuse sünteesist, vaid on ulatuslikum.

EGV indiviidide optimaalne akna suurus varieerus indiviiditi 241 bp-st 271 bp-ni ning võib seostuda indiviidi DNA raamatukogu fragmendi pikkusega. Kui lahutada iga indiviidi DNA raamatukogu keskmisest fragmendi pikkusest ühe lugemi pikkus, on tulemuseks korrelatsioonikordaja põhjal leitud optimaalse akna suurusele lähedane vaste, mis on optimaalsest aknast maksimaalselt 50 bp võrra erinev. 10 EGV indiviidi fragmentide ja optimaalsete akende pikkused on LISAS 3. Kuigi tundub ebatõenäoline näha sellist seost juhuslikult, on siinse töö andmete põhjal järeldus siiski hüpoteetiline. Selle tõestamiseks oleks vajalik läbi viia täpsem analüüs, kus saaks näiteks arvesse võtta ka *k*-meeri asukohta lugemites ning SNV asukohta *k*-meeris. Platinum indiviidi kohta DNA raamatukogu keskmise fragmendi pikkuse andmed puuduvad, kuid väiksem optimaalse akna suurus viitab, et sobivat universaalset akent, mille ulatuses saaks GC-sisalduse korrigeerida, ei ole ning sobiva akna suuruse peaks määrama iga andmestiku jaoks eraldi. Kontrolliks optimaalsest akna pikkusest suuremas ja väiksemas aknas leitud GC-sisalduste põhjal koostatud mudelite determinatsioonikordajad olid mõlemal juhul väiksemad, kui optimaalse akna GC-sisalduse andmetel koostatud mudelil. See kinnitab, et parima tulemuse saavutamiseks on oluline korrigeerida GC-sisaldus sobiva suurusega regioonis *k*-meeri ümber.

K-meeri asukoha arvesse võtmine lugemis võimaldaks edaspidi analüüsida ka sekveneerimisvigade mõju. Kuna vead tekivad suurema tõenäosusega lugemite lõpuosas, võiks mudeli põhjal leida, kas *k*-meeride madalam katvus võib osaliselt tuleneda *k*-meeridest, mis asusid lugemite lõpus ning jäid vigade tõttu lugemitest tuvastamata. See eeldab katvuse arvutamisel ka nende *k*-meeride arvesse võtmist, mis on lugemitest leitavad mõne valepaardumisega. Sekveneerimisvigu saaks sarnaselt arvesse võtta ka joondatud lugemite

põhjal leitud positsiooni katvuse korrigeerimisel. Erinevalt *k*-meeri katvuse korrigeerimisest, kus joondamisprotsess katvusele mõju ei avalda, tuleks joondatud lugemite katvuse korrigeerimisel arvesse võtta ka küsitavusi, mis võivad tekkida lugemi joondumisel mitmesse asukohta.

Koostatud mudelile kromosoomi parameetri lisamisel olid tunnuste mõjud statistiliselt olulised (*p*-väärtused LISAS 2), kuid mudeli kohandatud R^2 suurenes vaid 0,001 võrra. Tõenäoliselt on suure valimi korral lisatud parameeter statistiliselt oluline, kuid praktikas on muutused väikesed. Ühtlase GC-sisalduse juures on katvuse varieerumine kromosoomiti väike, kuid sisestades iga kromosoomi keskmise GC-sisalduse eraldi, esinevad erinevates kromosoomides suuremad varieeruvused – seega on katvuse varieerumine kromosoomide lõikes suuresti sõltuv GC-sisaldusest ning kromosoomide muud eripärad avaldavad väikest mõju (Joonis 9 LISAS 4). Sarnaselt kromosoomile olid ka positsiooni parameetri lisamisel splaini mõjud statistiliselt olulised (LISA 2), kuid praktikas on kõikumiste ulatus GC-sisalduse mõju eemaldamisel väike (Joonis 10 LISAS 5).

Töö raames koostatud mudeleid saaks edaspidi kasutada joondusvabades meetodites, mis rakendavad *k*-meeride katvust näiteks geneetiliste variatsioonide tuvastamiseks ning analüüsida, kui palju mudelite abil korrigeeritud katvuse väärtuste kasutamine meetodite täpsust parandaks. Kuigi praktikas võib olla eelistatud lihtsama (ainult GC-parameetriga) mudeli kasutamine, mis on kiirem, võiks parima mudeli leidmiseks edaspidi proovida katvust korrigeerida ka kromosoomi ja positsiooni parameetreid sisaldava mudeliga. See võimaldaks kindlaks teha, kas väikesed mõjud, mida positsioon ja kromosoom kirjeldavad, on katvuse väärtuse korrigeerimisel olulised ning kas keerulisema mudeli kasutamine oleks õigustatud. Eelkõige sobiksid mudelid katvuse korrigeerimiseks FastGT meetodi rakendamisel, kuna töös kasutati FastGT meetodil genotüpiseerimiseks kasutatavat *k*-meeride andmebaasi. Samas saaks mudeleid edaspidi ümber kohandada ka teistes meetodites kasutamiseks.

KOKKUVÕTE

Katvuse andmete rakendamine genoomi analüüsis võimaldab ilma eelteadmisteta uurida indiviidi geneetilisi variatsioone ja geeniekspressiooni eripärasid kogu genoomi või transkriptoomi ulatuses. Sealjuures on valepositiivsete ja –negatiivsete tulemuste vältimiseks määrava tähtsusega tõeste signaalide eristamine tehnilistest kõrvalekalletest. Siinse töö eesmärk oli anda ülevaade katvuse rakendusest, peamistest varasemalt teadaolevatest katvuse kõrvalekallete põhjustest ning eksperimentaalses osas hinnata GC-sisalduse, *k*-meeri asukoha ja kromosoomi mõju katvusele.

Varasemalt kõige põhjalikumalt analüüsitud GC-sisaldus seletab olulise osa katvuse varieerumisest. Siinse töö tulemused kinnitavad, et katvuse korrigeerimisel GC-sisalduse põhjal on parima seose leidmise puhul oluline GC-sisalduse arvutamine kindla suurusega aknas, mille pikkus tuleb määrata sõltuvalt iga proovi andmetest eraldi.

Koostatud lineaarseid regressioonimudeleid kasutati töös parameetrite mõju hindamiseks. GC-sisalduse mõju on selge ja tugev, kuid positsiooni ja kromosoomi mõjud seevastu väikesed, kuigi *p*-väärtuste põhjal statistiliselt olulised. Koostatud mudelite rakendamine analüüsidel võimaldaks edaspidi hinnata, kas positsiooni ja kromosoomi põhjustatud väikeste varieerumiste korrigeerimine parandaks katvuse andmete põhjal tehtud järeldusi näiteks CNV-de või SNV-de määramisel.

Siintoodud mudelid kirjeldasid ainult osa kõrvalekalletest (suurim R^2 = 0,315). Kuigi katvuse kõikumised on teatud määral põhjustatud ka andmetes esinevatest geneetilistest variatsioonidest, ei seleta need koos analüüsitud parameetritega kogu varieeruvust. Probleemi edasine uurimine ja võimalike mõju avaldavate parameetrite hindamine võib tulevikus mudelite ennustusvõimet parandada ja tagada analüüsideks järjest usaldusväärsemaid katvuse andmeid.

SUMMARY

Evaluation of the parameters affecting sequencing coverage

Carmen Oroperv

Summary

Coverage, which expresses the number of times each nucleotide is sequenced, is widely used for detection of genetic variations, gene expression analysis and DNA higher structure studies. Regardless of whether the coverage data represents reads aligned to a genomic position or the frequency of *k*-mers in raw reads, the main idea throughout the analyses is to detect differences in coverage value, which should be caused by genetic variations. Therefore, the main problem with making accurate conclusions is coverage deviation from the expected value, which is caused for example by the sequence content of the examined region, sequencing errors or, if aligned reads are used to determine the coverage value, read alignment.

The purpose of the theoretical part of this study was to give an overview of methods which use coverage to detect genetic variations, DNA-protein binding sites or analyse gene expression and discuss the main factors that affect coverage value. In the practical part, the aim was to evaluate the effect of GC content, genome position and chromosome on *k*-mer coverage by composing linear regression models.

The results show that the GC content of the regions surrounding the *k*-mer strongly affect coverage. Adjusted R^2 of the regression model indicates that GC content can explain up to 31,32% of the coverage variation and the length of the region, where the GC content is calculated, plays a crucial role to achieve the highest possible R^2 value. The optimal window size is different for each sample and seems to correlate with the read length and fragment size of DNA library.

Effects of the genomic position and chromosome are smaller, increasing the value of adjusted R^2 only by 0,001 and explaining the variations of coverage on a smaller scale. In the future, combined models can be used to correct coverage value, which can help to conclude whether the presented models are accurate and if taking into account the small effects of position and chromosome improves the detection of genetic variations.

KASUTATUD KIRJANDUSE LOETELU

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., … Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology, 12: R18. https://doi.org/10.1186/gb-2011-12-2-r18

Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., … Eichler, E. E. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. Nature Genetics, 41(10): 1061-7. https://doi.org/10.1038/ng.437

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., & Robinson, M. D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nature Protocols, 8: 1765–1786. https://doi.org/10.1038/nprot.2013.099

Auerbach, R. K., Euskirchen, G., Rozowsky, J., Lamarre-Vincent, N., Moqtaderi, Z., Lefrancois, P., … Snyder, M. (2009). Mapping accessible chromatin regions using Sono-Seq. Proceedings of the National Academy of Sciences, 106 (35): 14926-14931. https://doi.org/10.1073/pnas.0905443106

Bayindir, B., Dehaspe, L., Brison, N., Brady, P., Ardui, S., Kammoun, M., … Vermeesch, J. R. (2015). Noninvasive prenatal testing using a novel analysis pipeline to screen for all autosomal fetal aneuploidies improves pregnancy management. European Journal of Human Genetics, 23: 1286–1293. https://doi.org/10.1038/ejhg.2014.282

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research, 40(10): e72. https://doi.org/10.1093/nar/gks001

Benn, P., & Cuckle, H. (2014). Theoretical performance of non-invasive prenatal testing for chromosome imbalances using counting of cell-free DNA fragments in maternal plasma. Prenatal Diagnosis, 34(8); 778-783. https://doi.org/10.1002/pd.4366

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456: 53–59. https://doi.org/10.1038/nature07517

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J. P., Janoueix-Lerosey, I., Delattre, O., & Barillot, E. (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics, 27(2): 268–269. https://doi.org/10.1093/bioinformatics/btq635

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology, 34: 525–527. https://doi.org/10.1038/nbt.3519

Bronner, I. F., Quail, M. A., Turner, D. J., & Swerdlow, H. (2013). Europe PMC Funders Group Improved Protocols for Illumina Sequencing. Current Protocols in Human Genetics, 18(18.2). https://doi.org/10.1002/0471142905.hg1802s62.Improved

Cacho, A., Smirnova, E., Huzurbazar, S., & Cui, X. (2016). A comparison of base-calling algorithms for illumina sequencing technology. Briefings in Bioinformatics, 17(5): 786–795. https://doi.org/10.1093/bib/bbv088

Canick, J. A., Palomaki, G. E., Kloza, E. M., Lambert-Messerlian, G. M., & Haddow, J. E. (2013). The impact of maternal plasma DNA fetal fraction on next generation sequencing tests for common fetal aneuploidies. Prenatal Diagnosis, 33(7): 667-674. https://doi.org/10.1002/pd.4126

Chandrananda, D., Thorne, N. P., Ganesamoorthy, D., Bruno, D. L., Benjamini, Y., Speed, T. P., … Bahlo, M. (2014). Investigating and correcting plasma DNA sequencing coverage bias to enhance aneuploidy discovery. PLoS ONE, 9(1): e86993. https://doi.org/10.1371/journal.pone.0086993

Cheung, M. S., Down, T. A., Latorre, I., & Ahringer, J. (2011). Systematic bias in high-throughput sequencing data and its correction by BEADS. Nucleic Acids Research, 39(15): e103. https://doi.org/10.1093/nar/gkr425

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43: 491–498. https://doi.org/10.1038/ng.806

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research, 36(16): e105. doi: 10.1093/nar/gkn425

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. Molecular Ecology Resources, 11(5): 759-769. https://doi.org/10.1111/j.1755-0998.2011.03024.x

Gómez-Romero, L., Palacios-Flores, K., Reyes, J., García, D., Boege, M., Dávila, G., … Palacios, R. (2018). Precise detection of de novo single nucleotide variants in human genomes. Proceedings of the National Academy of Sciences, 115 (21): 5516-5521. https://doi.org/10.1073/pnas.1802244115

Halvardson, J., Zaghlool, A., & Feuk, L. (2013). Exome RNA sequencing reveals rare and novel alternative transcripts. Nucleic Acids Research, 41(1): e6. https://doi.org/10.1093/nar/gks816

Hansen, K. D., Brenner, S. E., & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Research, 38(12): e131. https://doi.org/10.1093/nar/gkq224

Hung, J. H., & Weng, Z. (2017). Mapping short sequence reads to a reference genome. Cold Spring Harbor Protocols, (2). https://doi.org/10.1101/pdb.prot093161

Jiang, F., Ren, J., Chen, F., Zhou, Y., Xie, J., Dan, S., … Zhang, X. (2012). Noninvasive Fetal Trisomy (NIFTY) test: An advanced noninvasive prenatal diagnosis methodology for fetal autosomal and sex chromosomal aneuploidies. BMC Medical Genomics, 5: 57. https://doi.org/10.1186/1755-8794-5-57

Kaplinski, L., Lepamets, M., & Remm, M. (2015). Genome Tester4: A toolkit for performing basic set operations - union, intersection and complement on k-mer lists. GigaScience, 4(1). https://doi.org/10.1186/s13742-015-0097-y

Kimura, K., & Koike, A. (2015). Ultrafast SNP analysis using the Burrows-Wheeler transform of short-read data. Bioinformatics, 31(10): 1577–1583. https://doi.org/10.1093/bioinformatics/btv024

Kingston, R. E. (2001). Preparation of Poly(A) $^{+}$ RNA. In Current Protocols in Molecular Biology, 21(1): 4.5.1-4.5.3. https://doi.org/10.1002/0471142727.mb0405s21

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D. A., Mitterecker, A., Bodenhofer, U., & Hochreiter, S. (2012). Cn.MOPS: Mixture of Poissons for discovering copy number

variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Research, 40(9): e69. https://doi.org/10.1093/nar/gks003

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., … Snyder, M. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science, 318(5849): 420-426. https://doi.org/10.1126/science.1149504

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., & Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nature Methods, 6: 291–295. https://doi.org/10.1038/nmeth.1311

Laehnemann, D., Borkhardt, A., & McHardy, A. C. (2016). Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. Briefings in Bioinformatics, 17(1): 154–179. https://doi.org/10.1093/bib/bbv029

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. Nature, 409: 860–921. https://doi.org/10.1038/35057062

Lander, E. S., & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. Genomics, 2(3): 231-239. https://doi.org/10.1016/0888-7543(88)90007-9

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., … Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research, 22: 1813-1831. https://doi.org/10.1101/gr.136184.111

Lau, T. K., Chan, M. K., Salome Lo, P. S., Connie Chan, H. Y., Kim Chan, W. S., Koo, T. Y., … Pooh, R. K. (2012). Clinical utility of noninvasive fetal trisomy (NIFTY) test early experience. Journal of Maternal-Fetal and Neonatal Medicine, 25(10): 1856-1859. https://doi.org/10.3109/14767058.2012.678442

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27(21): 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research, 18: 1851-1858. https://doi.org/10.1101/gr.078212.108

Lun, F. M. F., Chiu, R. W. K., Chan, K. C. A., Tak, Y. L., Tze, K. L., & Lo, Y. M. D. (2008). Microfluidics digital PCR reveals a higher than expected fraction of fetal DNA in maternal plasma. Clinical Chemistry, 54(10): 1664-1672. https://doi.org/10.1373/clinchem.2008.111385

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The Database of Genomic Variants: A curated collection of structural variation in the human genome. Nucleic Acids Research, 42(D1): D986–D992. https://doi.org/10.1093/nar/gkt958

Magi, A., Tattini, L., Pippucci, T., Torricelli, F., & Benelli, M. (2012). Read count approach for DNA copy number variants detection. Bioinformatics, 28(4): 470–478. https://doi.org/10.1093/bioinformatics/btr707

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20: 1297-1303. https://doi.org/10.1101/gr.107524.110

Medvedev, P., Fiume, M., Dzamba, M., Smith, T., & Brudno, M. (2010). Detecting copy number variation with mated short reads. Genome Research, 20: 1613-1622. https://doi.org/10.1101/gr.106344.110

Miller, C. A., Hampton, O., Coarfa, C., & Milosavljevic, A. (2011). ReadDepth: A parallel R package for detecting copy number alterations from short sequencing reads. PLoS ONE, 6(1): e16327. https://doi.org/10.1371/journal.pone.0016327

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., … Collins, F. S. (2011). Mapping copy number variation by population-scale genome sequencing. Nature, 470: 59–65 . https://doi.org/10.1038/nature09708

Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biology, 12(11): R112. https://doi.org/10.1186/gb-2011-12-11-r112

Monlong, J., Cossette, P., Meloche, C., Rouleau, G., Girard, S. L., & Bourque, G. (2018b). Human copy number variants are enriched in regions of low mappability. Nucleic Acids Research, 46(14): 7236–7249. https://doi.org/10.1093/nar/gky538

Monlong, J., Girard, S. L., Meloche, C., Cadieux-Dion, M., Andrade, D. M., Lafreniere, R. G., … Cossette, P. (2018a). Global characterization of copy number variants in epilepsy patients from whole genome sequencing. PLoS Genetics, 14(4): e1007285. https://doi.org/10.1371/journal.pgen.1007285

Muzzey, D., Evans, E. A., & Lieber, C. (2015). Understanding the Basics of NGS: From Mechanism to Variant Calling. Current Genetic Medicine Reports, 3: 158. https://doi.org/10.1007/s40142-015-0076-8

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., … Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. Nucleic Acids Research, 39(13): e90. https://doi.org/10.1093/nar/gkr344

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. Nature Reviews. Genetics, 12: 443–451. https://doi.org/10.1038/nrg2986

Nijkamp, J. F., Van Den Broek, M. A., Geertman, J. M. A., Reinders, M. J. T., Daran, J. M. G., & De Ridder, D. (2012). De novo detection of copy number variation by co-assembly. Bioinformatics, 28(24): 3195–3202. https://doi.org/10.1093/bioinformatics/bts601

Pajuste, F. D., Kaplinski, L., Möls, M., Puurand, T., Lepamets, M., & Remm, M. (2017). FastGT: An alignment-free method for calling common SNVs directly from raw sequencing reads. Scientific Reports, 7. https://doi.org/10.1038/s41598-017-02487-5

Patro, R., Mount, S. M., & Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature Biotechnology, 32: 462–464. https://doi.org/10.1038/nbt.2862

Piovesan, A., Pelleri, M. C., Antonaros, F., Strippoli, P., Caracausi, M., & Vitale, L. (2019). On the length, weight and GC content of the human genome. BMC Research Notes, 12: 106. https://doi.org/10.1186/s13104-019-4137-z

Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., … Turner, D. J. (2008). A large genome center's improvements to the Illumina sequencing system. Nature Methods, 5: 1005–1010. https://doi.org/10.1038/nmeth.1270

Ramachandran, P., Palidwor, G. A., & Perkins, T. J. (2015). BIDCHIPS: Bias decomposition and removal from ChIP-seq data clarifies true binding signal and its functional correlates Medicine. Epigenetics and Chromatin, 8:33. https://doi.org/10.1186/s13072-015-0028-2

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Hurles, M. E. (2006). Global variation in copy number in the human genome. Nature, 444: 444–454. https://doi.org/10.1038/nature05329

Reinert, K., Langmead, B., Weese, D., & Evers, D. J. (2015). Alignment of Next-Generation Sequencing Reads. Annual Review of Genomics and Human Genetics, 16: 133-151. https://doi.org/10.1146/annurev-genom-090413-025358

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., … Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. Genome Biology, 14(5): R51. https://doi.org/10.1186/gb-2013-14-5-r51

Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., … Gerstein, M. B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature Biotechnology, 27: 66–75. https://doi.org/10.1038/nbt.1518

Sauk, M., Žilina, O., Kurg, A., Ustav, E. L., Peters, M., Paluoja, P., … Kaplinski, L. (2018). NIPTmer: Rapid k-mer-based software package for detection of fetal aneuploidies. Scientific Reports 8. https://doi.org/10.1038/s41598-018-23589-8

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Research, 43(6): e37. https://doi.org/10.1093/nar/gku1341

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. Nature Reviews Genetics, 15: 121–132. https://doi.org/10.1038/nrg3642

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with

RNA-Seq. Bioinformatics, 25(9): 1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology, 28: 511–515. https://doi.org/10.1038/nbt.1621

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. Nature Reviews Genetics, 13: 36–46. https://doi.org/10.1038/nrg3117

Wang, C., Xu, J., Zhang, D., Wilson, Z. A., & Zhang, D. (2010). An effective approach for identification of in vivo protein-DNA binding sites from paired-end ChIP-Seq data. BMC Bioinformatics, 11: 81. https://doi.org/10.1186/1471-2105-11-81

Wilbanks, E. G., & Facciotti, M. T. (2010). Evaluation of algorithm performance in ChIP-seq peak detection. PLoS ONE, 5(7): e11471. https://doi.org/10.1371/journal.pone.0011471

Ye, H., Meehan, J., Tong, W., & Hong, H. (2015). Alignment of short reads: A crucial step for application of next-generation sequencing data in precision medicine. Pharmaceutics, 7(4): 523-541. https://doi.org/10.3390/pharmaceutics7040523

Yoon, S., Xuan, Z., Makarov, V., Ye, K., & Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Research, 19: 1586-1592. https://doi.org/10.1101/gr.092981.109

Yu, S. C. Y., Jiang, P., Choy, K. W., Chan, K. C. A., Won, H. S., Leung, W. C., … Chiu, R. W. K. (2013). Noninvasive Prenatal Molecular Karyotyping from Maternal Plasma. PLoS ONE, 8(4): e60968. https://doi.org/10.1371/journal.pone.0060968

Zhang, Z. D., Du, J., Lam, H., Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). Identification of genomic indels and structural variations using split reads. BMC Genomics, 12: 375. https://doi.org/10.1186/1471-2164-12-375

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., … Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biology, 9(9): R137. https://doi.org/10.1186/gb-2008-9-9-r137

Zhao, C., Tynan, J., Ehrich, M., Hannum, G., McCullough, R., Saldivar, J. S., … Deciu, C. (2015). Detection of fetal subchromosomal abnormalities by sequencing circulating cell-free DNA from maternal plasma. Clinical Chemistry, 61(4): 608-616. https://doi.org/10.1373/clinchem.2014.233312

KASUTATUD VEEBIAADRESSID

Archive EnsEMBL: GRCh37 release 75, kasutatud: 24.05.2019, ftp://ftp.ensembl.org/pub/release-75/fasta/homo_sapiens/dna/

Genome Reference Consortium: Human Genome Assembly GRCh38.p13, kasutatud: 03.05.2019, https://www.ncbi.nlm.nih.gov/grc/human/data

Jones E, Oliphant E, Peterson P, *et al.* SciPy: Open Source Scientific Tools for Python, kasutatud: 21.05.2019, http://www.scipy.org/

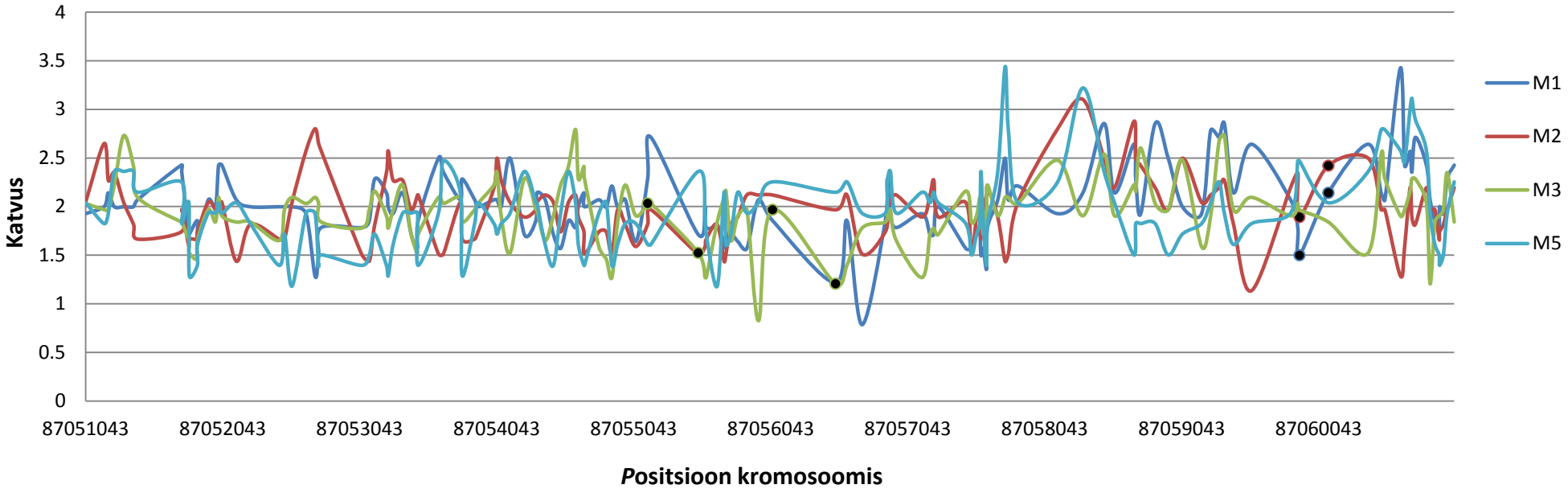
FastGT: from raw sequence reads to 30 million genotypes in less than an hour, kasutatud: 21.05.2019, http://bioinfo.ut.ee/FastGT/index.php?r=site/page&view=manual

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, kasutatud: 21.05.2019, https://www.R-project.org/

Steven D'Aprano. Statistics: mathematical statistics functions, kasutatud: 21.05.2019, https://docs.python.org/3/library/statistics.html

Tõnu Tõnso. Splainid. Tallinna Ülikool, kasutatud 26.05.2019, http://www.tlu.ee/~tonu/Arvmeet/Splkonsp.pdf

LISA 1



Joonis 8. Katvuse väärtuse muutused nelja meessoost EGV indiviidi (M) 8. kromosoomi keskel. Joonisel on 10 000 aluspaari pikkune regioon, 135 *k*-meeri ning katvuste standardhälbed (arvutatud 50 indiviidi katvuste põhjal) on väiksemad kui 0,5. Kõikumiste tihedus graafikul varieerub, kuna *k*-meeride on üksteisest erinevatel kaugustel. Mustade punktidega tähistatud katvuse muutused tähistavad sarnaseid katvuse väärtuse muutusi kahel erineval indiviidil. Kahel vasakpoolsel juhul on katvuse väärtused indiviididel võrdsed. Paremal pool tähistatud katvuse muutus on sarnase ulatusega ja katvus muutub nii M1 kui ka M2 indiviidil võrreldes eelmise *k*-meeriga kõrgemaks, kuid katvuste väärtus indiviidide vahel varieerub ligikaudu 0,5 võrra.

LISA 2

Tabel 1. Mudelite valemid, kohandatud determinatsioonikordajad (R^2) ja näited argumenttunnuste kordajatest usaldusintervallide ja *p*-väärtustega. Ainult GC-sisalduse parameetriga mudeli puhul on toodud kõik argumenttunnused, mille põhjal on näha, et splaini väärtuse kasvades kordaja langeb ehk kõrgemate GC-sisalduste juures on katvuse väärtus väiksem. β - tunnuse kordaja, 95% CI – kordaja usaldusintervalli alumine (2,5%) ja ülemine piir (97,5%).

Argumenttunnus	β	95% CI	*p*-väärtus
Katvus ~ bs(GC%, df = 7) ***R^2 =0,3132***			
Konstant	1,5614	1,543: 1,560	$< 2*10^{-16}$
GC splain 1	0,7982	0,776; 0,820	$< 2*10^{-16}$
GC splain 2	0,5859	0,568; 0,604	$< 2*10^{-16}$
GC splain 3	0,4774	0,459; 0,496	$< 2*10^{-16}$
GC splain 4	0,4117	0,394; 0,430	$< 2*10^{-16}$
GC splain 5	0,2131	0,195; 0,231	$< 2*10^{-16}$
GC splain 6	0,2564	0,238; 0,275	$< 2*10^{-16}$
GC splain 7	0,0774	0,057; 0,098	$4,39*10^{-14}$
Katvus ~ bs(GC%, df = 7) + factor(Chr) ***R^2 = 0,3138***			
Konstant	1,5613	1,543; 1,579	$< 2*10^{-16}$
GC splain 1	0,7945	0,773; 0,816	$< 2*10^{-16}$
GC splain 2	0,5831	0,565; 0,601	$< 2*10^{-16}$
...			
GC splain 7	0,0746	0,054; 0,095	$3,53*10^{-13}$
Chr 2	0,0031	0,003; 0,004	$< 2*10^{-16}$
Chr 3	0,0035	0,003; 0,004	$< 2*10^{-16}$
...			
Chr 22	-0,0042	-0,005; -0,003	$< 2*10^{-16}$
Katvus ~ bs(GC%, df = 7) + bs(position, df = 200) ***R^2 = 0,3148***			
Konstant	1,595	1,577; 1,614	$< 2*10^{-16}$
GC splain 1	0,7911	0,769; 0,813	$< 2*10^{-16}$
GC splain 2	0,5810	0,563; 0,599	$< 2*10^{-16}$
...			
GC splain 7	0,0782	0,058; 0,098	$2,38*10^{-14}$
Pos splain 1	-0,0437	-0,052; -0,036	$< 2*10^{-16}$
Pos splain 2	-0,0401	-0,045; -0,035	$< 2*10^{-16}$
...			
Pos splain 200	-0,0163	-0,022; -0,010	$1,84*10^{-7}$
Katvus ~ bs(GC%, df = 7) + bs(position, df = 200) + factor(Chr) ***R^2 = 0,315***			
Konstant	1,5956	1,577; 1,614	$< 2*10^{-16}$
GC splain 1	0,7905	0,769; 0,812	$< 2*10^{-16}$
GC splain 2	0,5806	0,563; 0,598	$< 2*10^{-16}$

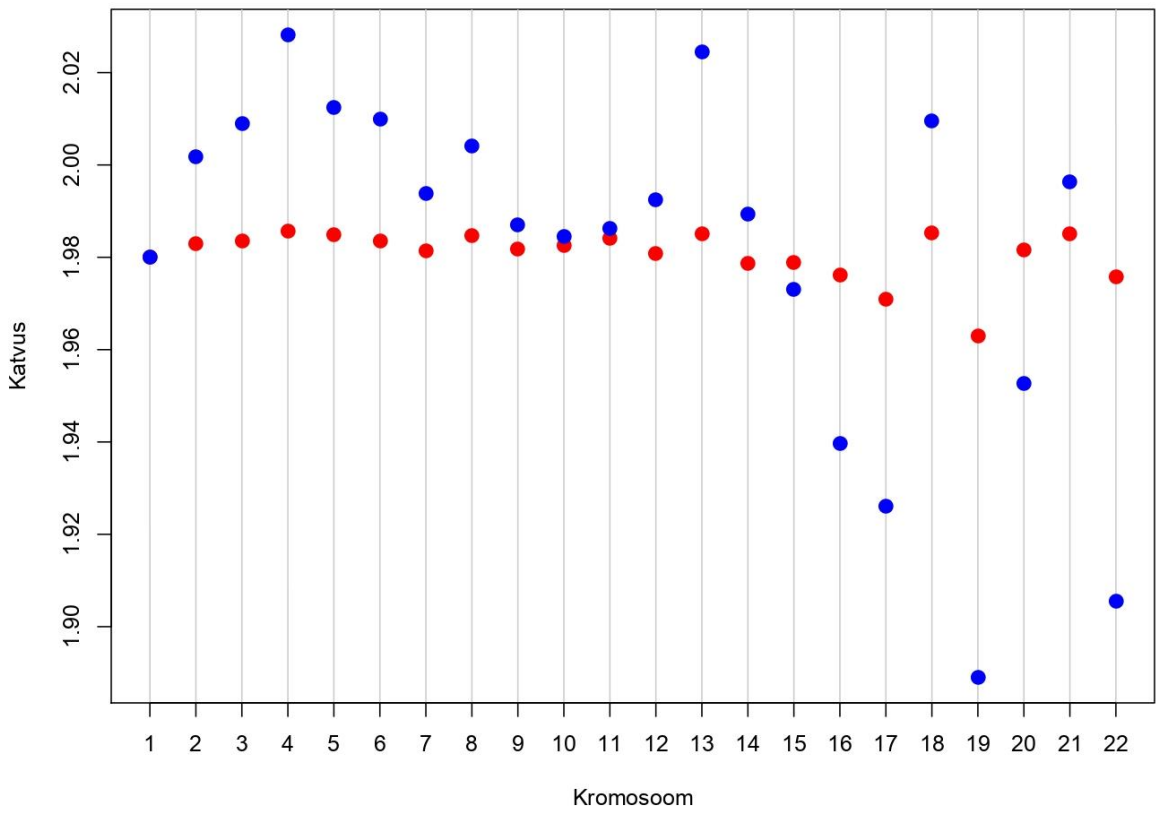
...			
GC splain 7	0,0795	0,059; 0,1	$8,60*10^{-15}$
Chr 2	0,0196	0,015; 0,024	$< 2*10^{-16}$
Chr 3	0,0140	0,008; 0,02	$1,22*10^{-6}$
...			
Chr 22	-0,4198	-0,46; -0,38	$< 2*10^{-16}$
Pos splain 1	-0,0437	-0,052; -0,036	$< 2*10^{-16}$
Pos splain 2	-0,0401	-0,045; -0,035	$< 2*10^{-16}$
...			
Pos splain 200	0,4016	0,361; 0,442	$< 2*10^{-16}$

LISA 3

Tabel 2. Optimaalse akna seos fragmendi pikkusega. Lugemi pikkus on ühe paarislugemi pikkus ning fragmendi pikkus on DNA raamatukogu keskmine fragmendi pikkus. N ja M tähistavad indiviide EGV andmestikus.

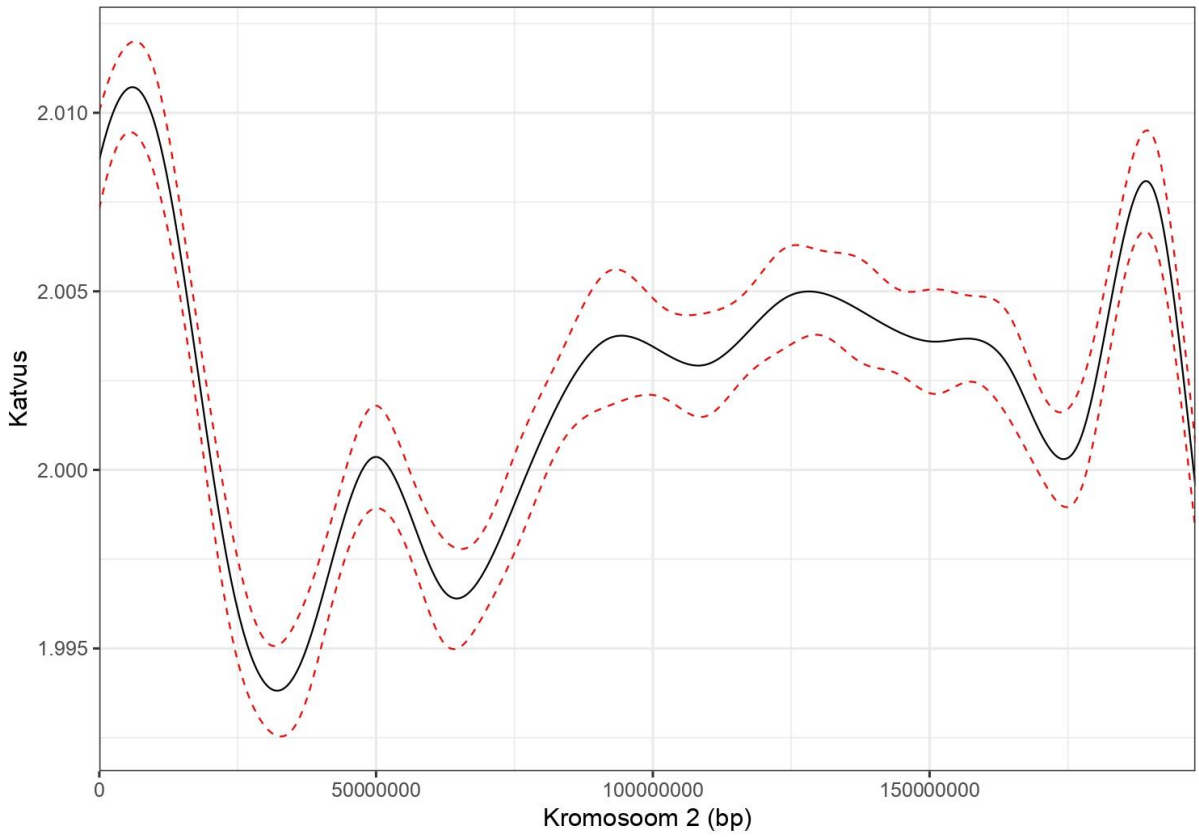
Indiviid	Keskmine fragmendi pikkus	Lugemi pikkus	Optimaalne akna suurus	Fragmendi pikkuse ja paarislugemi vahe
N1	385	151	251	234
N2	361	146	251	251
N3	392	151	271	241
N4	370	146	271	224
N5	372	151	251	221
M1	379	151	241	228
M2	396	151	251	245
M3	401	143	271	258
M4	406	151	241	255
M5	387	151	241	236

LISA 4



Joonis 9. Kromosoomi mõju katvusele ühtlase ja varieeruva GC-sisalduse korral. Punased täpid viitavad kromosoomi ja GC-sisalduse põhjal mudeli poolt ennustatud katvuse väärtustele, kui GC-sisaldus oleks kromosoomide lõikes konstantne. Sinised täpid viitavad mudeli poolt ennustatud katvuse väärtustele, kui GC-sisaldus on iga autosoomi keskmine. Konstantse GC-sisaldusena lisati mudelisse inimese autosoomide keskmine GC-sisaldus. Nii keskmine (41.74%) kui ka autosoomide GC-sisaldused on leitud referentsgenoomi versiooni GRCh37 põhjal (Piovesan *et al*., 2019).

LISA 5



Joonis 10. Positsiooni mõju katvusele. *X*-teljel on 2. kromosoomi koordinaadid, *y*-teljel normaliseeritud katvus. Konstantse GC-sisaldusena kasutati 2. kromosoomi keskmist GC-sisaldust (40.24%) (Piovesan *et al*., 2019). Joon iseloomustab mudeli põhjal ennustatud katvuse väärtusi 2. kromosoomi positsioonides, kui GC-sisaldus katvuse varieerumisele mõju ei avalda (on konstantne). Punased punktiirjooned näitavad mudeli põhjal ennustatud katvuse väärtuse 95% usaldusintervalle.

LIHTLITSENTS

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Carmen Oroperv (19.08.1997),

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose **Katvust mõjutavate parameetrite hindamine**, mille juhendaja on Fanny-Dhelia Pajuste, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.
2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Carmen Oroperv

27.05.2019