TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR-JA RAKUBIOLOOGIA INSTITUUT

Sobivate inimese genoomi regioonide leidmine loote rakuvaba DNA osakaalu määramiseks ema vereproovist

Bakalaureusetöö

12 EAP

Alvin Meltsov

Juhendajad

Priit Palta, PhD

Kaarel Krjutškov, PhD

TARTU 2019

Sobivate inimese genoomi regioonide leidmine loote rakuvaba DNA osakaalu määramiseks ema vereproovist

Mitte-invasiivne prenataalne test (NIPT) on mugav ja turvaline sõeluuring loote kromosoomhaiguste riski määramiseks, analüüsides loote päritolu rakuvaba DNAd lapseootel naise vereproovist. NIPTi jaoks välja arendatud meetodid kasutavad erinevaid tehnoloogiaid, et tuvastada loote kromosoomide väärarenguid ehk aneuploidiaid ja anda kvalitatiivne hinnang testile loote DNA osakaalu ehk lootefraktsiooni määramise läbi. Loote rakuvaba DNA on segunenud ema enda rakuvaba DNAga ning seetõttu on oluline, et loote aneuploidsuse hinnang pärineks usaldusväärsest analüüsist, kus loote päritolu rakuvaba DNA on tuvastatav ja kõrgem kui teatud piirnorm. Käesoleva töö eesmärgiks on leida minimaalse suurusega inimese genoomi piirkonnad, mis toimiks kui loote DNA fraktsiooni määramise potentsiaalsed biomarkerid.

Märksõnad: Biomarker, bioinformaatika, loote DNA fraktsioon, NIPT, meditsiiniinformaatika, biomeetrika

CERCS: B110

Finding suitable regions in the human genome for determination of fetal fraction from maternal cell-free DNA

Non-invasive prenatal test (NIPT) is a convenient and safe alternative screening method for detecting fetal chromosomal abnormalities requiring only a sample of pregnant woman's blood sample. Test methods developed for NIPT use various technologies to detect aneuploidies of fetuses and to give a qualitative assessment of the test in the form of the relation of fetal DNA to the total cell-free DNA (cfDNA) in pregnant women's blood, called fetal fraction. In the analysed blood sample, fetal cfDNA is mixed with mother's cfDNA, and it is important to distinguish DNA of fetal origin. The goal of this thesis is to find biomarkers, that are relevant to fetal fraction calculation, and to use the resulting list of regions to successfully calculate fetal fraction.

Keywords: Biomarkers, bioinformatics, fetal fraction, NIPT, medical informatics, biometrics

CERCS: B110

Sisukord

Kasutatud lühendid

- BAM - *Binary SAM,* binaarne failiformaat SAM failide salvestamiseks.
- cfDNA - *cell-free DNA*, rakuvaba DNA.
- cffDNA - *cell-free fetal DNA*, rakuvaba loote DNA.
- CVS - *chorionic villus sampling*, koorionibiopsia.
- NIPT - *non-invasive prenatal test*, mitteinvasiivne sünnieelne sõeluuring ema verest.
- FASTQ – tekstipõhine failiformaat nukleotiidijärjestuste salvestamiseks koos kvaliteediskooriga.
- FF - *fetal fraction*, loote DNA fraktsioon.
- NGS - *Next-Generation Sequencing*, teise põlvkonna sekveneerimine.
- PCR - *Polymerase Chain Reaction*, polümeraas-ahelreaktsioon.
- qPCR – *quantitative-PCR*, kvantitatiivne PCR.
- SAM - *Sequence Alignment/Map*, referentsgenoomiga kaardistatud järjestusefail.
- SNP - *Single-Nucleotide Polymorphisms*, ühenukleotiidsed polümorfismid.

Sissejuhatus

Loote kromosomaalsed väärarengud ehk aneuploidiad on lootele enamasti fataalsed, lõppedes raseduse enneaegse katkemisega. Seevastu kromosoomide 21, 18, 13 ja sugukromosoomide aneuploidsus ei põhjusta tingimata loote hukkumist ja sünnib kromosoomhaigusega laps. Kõige sagedasem on kromosoom 21 trisoomia ehk Downi sündroom, mis esineb 0,1% kõikidest sündidest alla 30 aastaste naistel. Sünnitaja vanuse kasvades Downi sündroomi risk suureneb, esinedes ligi 20% rasedustest üle 40 aasta naistel (Moorthie *et al.*, 2018). Eestis on tänu riiklikule rasedate esimese trimestri sõeluuringule vähenenud Downi sündroomi laste sündimus oluliselt võrreldes näiteks paarikümne aasta taguse olukorraga, jäädes tänapäeval umbes viie sünnini aastas (Lokko *et al.*, 2016). Käesoleva töö koostamise ajal näeb Eesti rasedate sõeluuringu ravijuhis ette vereseerumi biomarkerite ja ultraheli kombineeritud sõeluuringut. Kõrgenenud kromosoomhaiguse riski korral suunatakse patsient täiendavale invasiivsele diagnostilisele uuringule, mille jaoks võetakse loote proov kas looteveest (amniotsentees) või platsentast (koorionbiopsia). Kombineeritud meetod on aga suhteliselt ebatäpne ja piiratud usaldusväärsusega, andes valepositiivseid tulemusi 5% juhtudest ja selle tundlikkus on 10% trisoomia 21 puhul ja 50-25% muude kromosomaalsete haiguste korral (Dan *et al.*, 2012). Eestis jääb näiteks hinnanguliselt iga seitsmes Downi sündroomiga loode sõeluuringu käigus leidmata ja enam kui 90% kõrge riskiga sõeluuringu leidudest osutuvad vale-positiivseteks (NIPT taotlusvorm)**.** Looteveest või platsentast proovi võtmine on ebameeldiv ja riskantne protseduur, mis võib komplikatsioonide korral viia kas ema tervise halvenemiseni või raseduse katkemiseni. Uuringud on näidanud, et raseduse katkemine amniotsenteesi läbinutel esineb ligikaudu 0,1% ja koorionbiopsia läbinutel 0,2% (Akolekar *et al.*, 2015) (Eddleman *et al.*, 2006) ning komplikatsioonide risk suureneb naise vanuse tõusuga (Papantoniou *et al.*, 2001).

Riskide vähendamiseks ja invasiivsete protseduuride kõrvalnähtude vältimiseks on alates 2011 aastast kasutusele võetud mitte-invasiivne prenataalne (genoomi) test (*Non-Invasive Prenatal Test* ehk NIPT). NIPT põhineb avastusel, et lisaks naise rakuvabale DNAle (*cell-free DNA* ehk cfDNA), on raseda veres vähesel määral loote, eelkõige platsenta päritolu rakuvaba DNAd (*cell-*

free fetal DNA ehk cffDNA) (Lo *et al.*, 1997). Loote cfDNA osakaal kogu raseda cfDNAst on tuvastatav alates 4-5. rasedusnädalast ja tõuseb raseduse kasvades. Umbes 10. rasedusnädalal on cffDNA osakaal 3-6% ja raseduse kolmandas trimestris kuni 22% (Lo *et al.*, 1998) (Lun *et al.*, 2008). Loote cfDNA kui geneetiline materjal on väga sarnane ema cfDNAle, mistõttu on vajalik edasine bioinformaatiline analüüs lootespetsiifilise rakuvaba DNA eristamiseks. Rakuvaba DNA sekveneerimise andmetest määrab NIPT analüüs eripärasusi loote kromosoomispetsiifiliste DNA lõikude jaotuses ema ja loote materjali vahel, et selle põhjal hinnata, millised kromosoomid või kromosoomide piirkonnad võivad olla lootel kas ala- või üle-esindatud.

NIPT test vajab usaldusväärse hinnangu andmiseks võimalikult täpset loote DNA osahulga ehk lootefraktsiooni (*fetal fraction ehk* FF) mõõtu, sest ilma selleta pole võimalik hinnata, kas loote DNA hulk on piisav, et määrata usaldusväärset võimalikku aneuploidia esinemist või puudumist.

NIPT test on saanud võimalikuks tänu mass-sekveneerimise meetodite (*Next Generation Sequencing* ehk NGS) olemasolule ja kättesaadavusele. Lootefraktsiooni leidmiseks NIPT testis kasutatakse mitmeid lähenemisi. Esimene on ühenukleotiidsete polümorfismide ehk SNPde (*Single Nucleotide Polymorphisms* ehk SNP) genotüpiseerimine ja lootefraktsiooni leidmine läbi genotüübi proportsioonide (Pergament *et al.*, 2014). Lisaks kasutatakse rakuvaba DNA pikkuste analüüsi ja metülatsiooni erinevusi (Straver *et al.*, 2016), et leida võimalikult täpne FF hinnang. Puudusena on leitud, et mainitud meetodid ei ole alati usaldusväärsed ja on SNPde näol populatsioonispetsiifilised. Kokkuvõttes võib järeldada, et eeltoodud lähenemised ei ole rakendatavad madala katvusega üle genoomi NIPT uuringutes (Whole Genome Sequecning ehk WGS) (Dan *et al.*, 2012). Lootefraktsiooni usaldusväärseks määramiseks on arendatud bioinformaatiline meetod SeqFF, mis on võimeline ainult sekveneerimislugemite hulka kasutades hindama FF määra (Kim *et al.*, 2015).

Käesoleva töö eesmärgiks on leida, kas madala katvusega sekveneerimisandmete põhjal on võimalik leida lühemaid genoomseid piirkondi, mida saaks tulevikus kasutada FF määramiseks suunatud sekveneerimisega. Käesolev töö on valminud Tervisetehnoloogiate Arenduskeskus ASis.

1. Kirjanduse ülevaade

1.1 Kromosoomianomaaliad üldiselt

Kromosoomianomaaliad on ulatuslikud mutatsioonid, mis hõlmavad rohkem kui ühte geeni või kitsast genoomi regiooni. Kromosomaalne struktuurne mutatsioon tekib mitoosi või meioosi käigus ja võib olla nii deletsioon, duplikatsioon, translokatsioon, inversioon, insertsioon või isomerisatsioon. Vaatamata sellele, et tegemist on ulatusliku kromosomaalse kõrvalekaldega, ei mõjuta kromosomaalsed anomaaliad alati fenotüüpi. Fenotüüpi mõjutavad kromosoomianomaaliad võivad põhjustada loote arengu jooksul tõsiseid arenguhälbeid või sünnijärgselt lapse surma (Theisen and Shaffer, 2010).

Valdav enamus võimalikest kromosoomianomaaliatest lõppevad lootele fataalselt ehk iseenesliku abordiga. Erandiks on sugukromosoomide või suhteliselt väikeste autosoomide 21, 18 ja 13 trisoomiad. Autosoomsetest trisoomiatest on kõige suurema elumusega 21. kromosoomi trisoomia ehk Downi sündroom, mis väljendub sarnaselt teistele aneuploidiatele kaasasündinud terviseriketes, aeglustunud kasvus, anatoomilistes hälvetes ja madalas intelligentsivõimes (Theisen and Shaffer, 2010).

1.2 cffDNA ja cfDNA

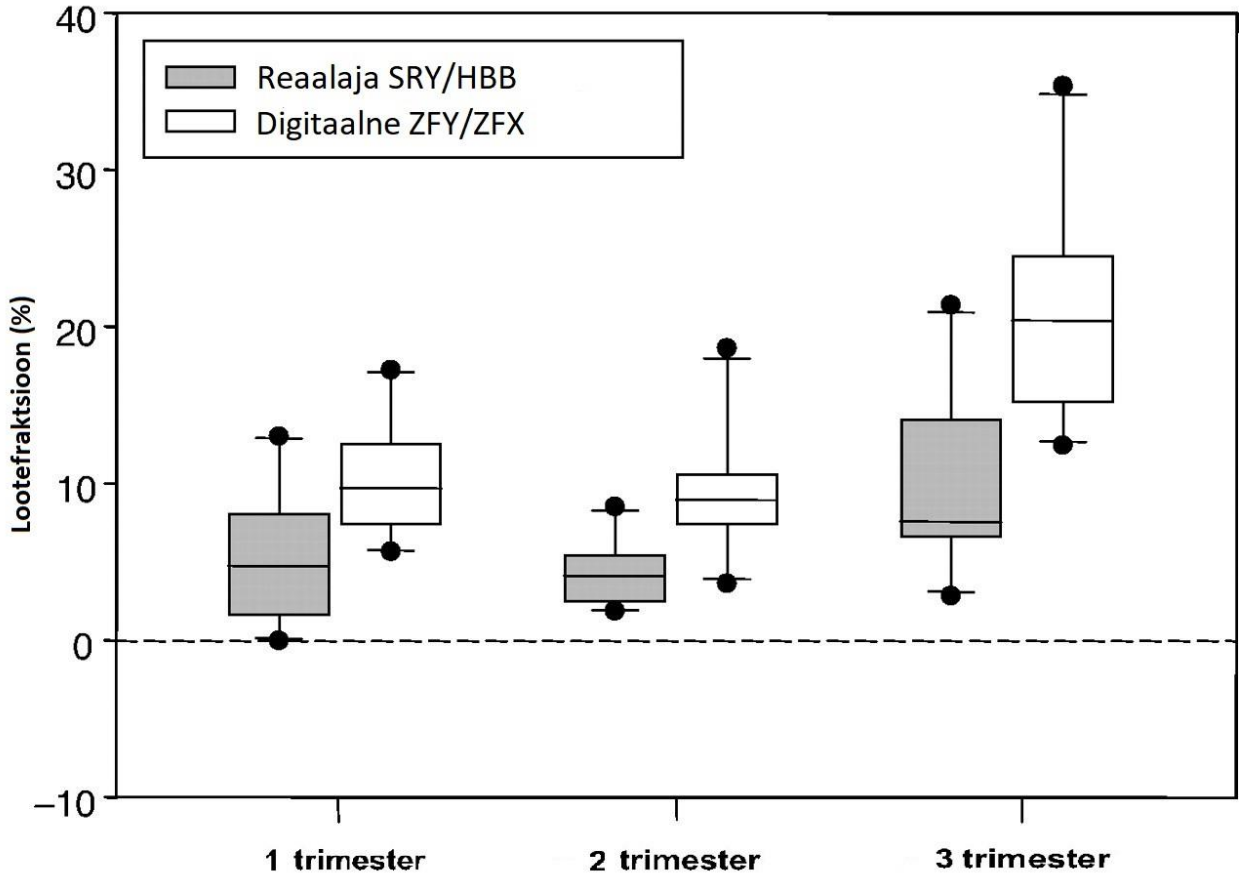
Rakuvaba DNA ehk cfDNA on fragmenteerunud genoomne DNA vereplasmas, mille olemasolu inimese veres kirjeldasid esimest korda Mandel ja Metais poolt 1948. aastal (Metais ja Metais, 1948). CfDNA vabaneb vereringesse surnud rakkudest nende lagunemise protsessi käigus, kuid cfDNAd võib leida ka teistest kehavedelikest, näiteks uriinist. Suurem osa cfDNAd vabaneb tuumaga lümfotsüütide apoptoosi tagajärjel (Rogers *et al.*, 1972). Teiseks oluliseks cfDNA allikaks peetakse nekrootilisi rakke, mis erinevalt apoptootilistest rakkudest ei ole fagotsütoosi käigus alati täielikult lagundatud (Gravina *et al.*, 2016) (García-Olmo *et al.*, 2010). Need allikad põhjustavad rakuvaba DNA olemasolu ka tervete inimeste vereringes. On leitud, et cfDNA tase võib muutuda suurenenud füüsilise aktiivsuse tagajärjel või vanuse kasvades. On spekuleeritud, et cfDNA võib läbida elusate rakkude seina ning mobiilsete geneetiliste elementidena

integreeruda suvaliselt genoomi, suurendades mutageneesi ja vähkkasvajate riski (Schwarzenbach *et al.*, 2011).

Rakuvaba DNA jaotatakse päritolu järgi kaheks: (i) kasvaja päritoluga rakuvaba DNA (ctDNA, circulating tumor DNA) või (ii) loote päritolu rakuvaba DNA (cffDNA, cell-free fetal DNA). Kasvaja rakuvaba DNA on oluline uurimisobjekt kasvajate uuringutes, pärinedes kas kasvaja rakkude apoptoosist, nekroosist või on eritatud elusate liikuvate vähirakkude poolt. Rakuvaba DNA suurenenud osakaal vähihaigete veres on tingitud puudulikust fagotsütoosist (Stroun *et al.*, 2001). Loote rakuvaba DNA pärineb peamiselt platsenta trofoblastide lagunenud või tervetest rakkudest, mis on sattunud ema verre läbi ema-loote barjääri (Gupta *et al.*, 2004). Seda, et cffDNA on pärit just platsenta trofoblastidest, mitte lootest, on näidanud uuring, kus uuriti anembrüonaalseid rasedusjuhte. Leiti, et kuigi loodet ei olnud, sisaldus raseda naise veres ikkagi normaalses hulgas cffDNA-d (Alberry *et al.*, 2007). Seega võib raseda naise verest eraldatud rasedale naisele mitteomane geneetiline info loote somaatilise mosaiiksuse või anembrüonaalsete raseduse puhul erineda reaalsest loote DNAst.

1.2.1 cffDNA eristamine ülejäänud rakuvabast DNA'st *in vitro*

Vereplasmas leidub rakuvaba DNAd, millest lootefraktsioon on raseduse varajases staadiumis algsel hinnangul 3-4% ja hilisemas staadiumis 6% (Lo *et al.*, 1998). Järgnenud uuringud on määranud lootefraktsiooni kõrgemaks - keskmiselt 10% 1-2 trimestri jooksul ja kuni 20% raseduse hilisemas staadiumis, nagu näidatud joonisel 1 (Lun *et al.*, 2008). Seega on lootespetsiifiliste kromosoomianomaaliatele viitavate DNA järjestuste hulga tuvastamine laboritehniliselt ja bioinformaatiliselt väljakutsuv ülesanne.



Joonis 1. Lootefraktsiooni osakaalud ema vereproovis raseduse trimestrite kaupa mõõdetuna qPCR ja digitaalse PCR-iga. Uuring keskendus kahele soospetsiifilistele geenile SRY/HBB ja ZFY/ZFX. Kastisisene joon tähistab 10 proovi mediaani, ja kastid ulatuvad 25-protsentiilist ehk alumisest kvartiilist ehk 1. kvartiilist 75-protsentiilini ehk ülemise kvartiilini ehk 3. kvartiilini. Punktid tähistavad erindeid ehk võõrväärtusi (Lun *et al.*, 2008).

On hinnatud, et usaldusväärse aneuploidsuse või selle puudumise hinnangu andmiseks on vajalik vähemalt 4% cffDNA-d ema veres leiduvast geneetilisest materjalist (Ashoor *et al.*, 2013), mis on tavaliselt saavutatav esimese trimestri 9-10 nädalal. NIPTi katse puhul sekveneeritakse vereplasmast eraldatud rakuvaba DNA teise põlvkonna NGS tehnoloogiaga ja joondatakse referentsgenoomi vastu. Saadud lugemid kvantifitseeritakse kromosoomipõhiselt ning erinevused normaalsest lugemite jaotusest huvipakkuvate kromosoomide puhul on indikatiivne suurenenud kromosoomide arvu hulgale. On näidatud, et kirjeldatud kvalitatiivne meetod on võimeline määrama Downi sündroomiga looteid 99% sensitiivsusega ja 95% spetsiifilisusega

(Palomaki *et al.*, 2011). Kirjeldatud protsessi järel ei saa aga tõestada, et proovis oli cffDNA-d, sest mõõdetakse ainult sekveneerimisandmete kromosoomipõhist hulka, mitte ei määrata indiviidi spetsiifiliste järjestuste põhjal. Seega on vajalik täpne cffDNA osakaalu hinnangu ehk lootefraktsiooni leidmise metoodika, mida saaks kasutada nii tulemuse kvaliteedi parandamiseks kui üldise proovi kvaliteedi hindamiseks.

CffDNA paremaks eraldamiseks cfDNA-st oleks lootepäritolu nukleotiidjärjestuste osakaalu suurendamine. Selleks on kirjeldatud kahte võimalikku meetodit. Esimeses kasutatakse ära cffDNA kõrget fragmenteeritust, mis tähendab seda, et lootepäritolu fragmendid on ema veres väiksema keskmise pikkusega kui 313 aluspaari (Chan *et al.*, 2004). Selle eelduse põhjal on välja pakutud, et rikastades pikkusespetsiifiliselt cffDNA optimaalse pikkusega fragmente, on võimalik tõsta lootefraktsiooni 70%-ni cfDNA-st (Li *et al.*, 2004).

Teise meetodina on võimalik cffDNA osakaalu tõsta läbi formaalaldehüüdi lisamise, mis takistab ema rakkudel vereplasmas edasise tegevuse ning võimaliku DNA rakuvälisesse ruumi paiskamise peale proovi võtmist (Dhallan *et al.*, 2004). Seda meetodit on aga kritiseeritud, sest formaalaldehüüdiga rikastamise tulemused on raskelt reprodutseeritavad (Chinnapapagari *et al.*, 2005). Biotehnoloogiliste võtetega ei suudeta seega cffDNA osakaalu suurendada piisavalt, et see ületaks mitmekordselt ema cfDNA osakaalu, mistõttu on jätkuvalt vajalik lootefraktsiooni leidmine diagnostiliseks hinnanguks.

1.3 Teise põlvkonna sekveneerimismeetodid

1.3.1 Teise põlvkonna meetodid

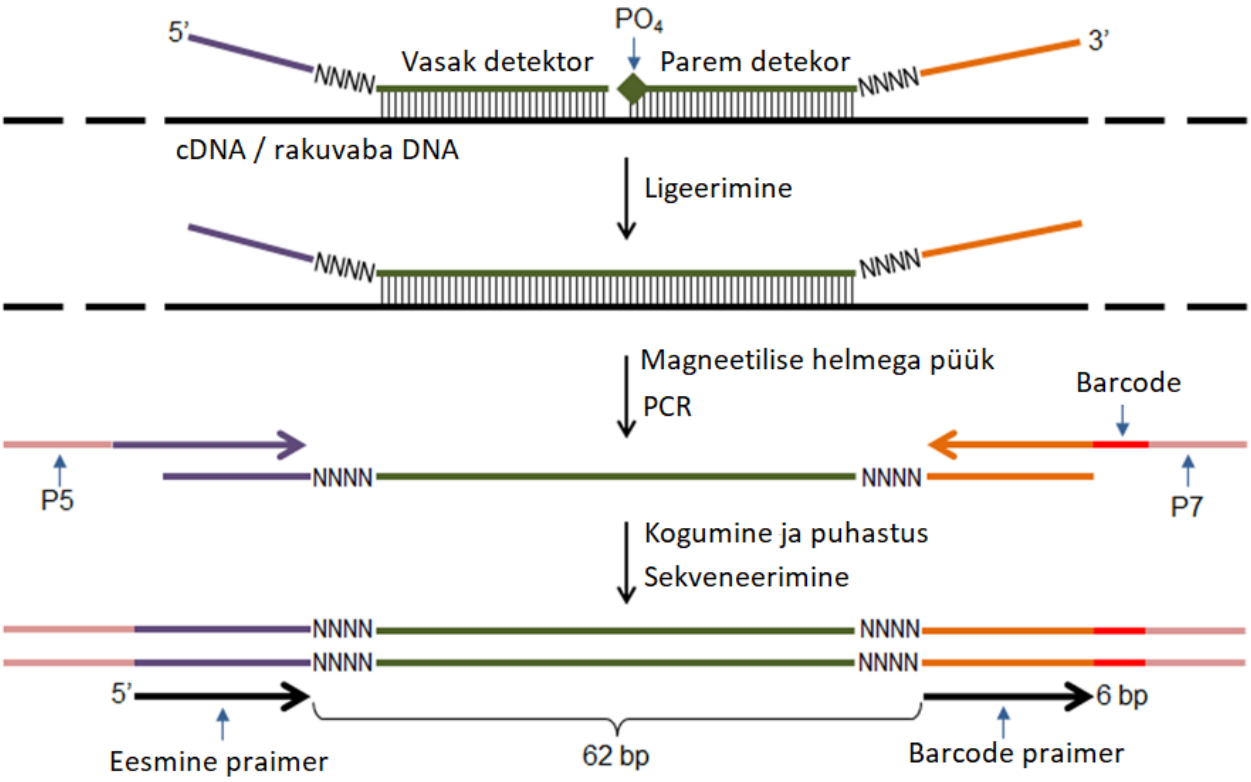
Teise põlvkonna sekveneerimismeetodite all mõistetakse erinevaid DNA järjestamisesüsteeme, mis muutusid laboritele kättesaadavaks 2000. aastate jooksul ning on siiamaani laialdaselt kasutusel. Neid meetodeid iseloomustab kõrgem analüüsikiirus, oluliselt odavam hind võrreldes esimese põlvkonna ehk Sangeri meetodiga (Sanger *et al.*, 1977), suur nukleotiidide eristamise täpsus ja paralleelsusest tingitud läbilaskevõime. Erinevalt varasematest meetoditest, mis järjestasid edukalt kuni 1000 nukleotiidi pikkusi ahelaid, piirduvad NGS meetodid tavaliselt kuni paarisaja nukleotiidi pikkuste lugemitega. Puudujääke kompenseerib aga suur paralleelne

läbilaskevõime, mis võimaldab genereerida hulganisti osaliselt ülekattes olevaid lühikesi lugemite järjestusi, mis seejärel arvutuslikult ühendatakse täielikuks genoomiks (Anderson ja Schrijver, 2010). NGS laiatarbeplatvorme nagu Illumina *sequencing-by-synthesis*, 454 pürosekveneerimine, ligeerimispõhine sekveneerimine SOLiD platvorm ja Nanopore tehnoloogiat kasutatakse tänapäeval laialdaselt genoomika uurimisvaldkonnas (Slatko *et al.*, 2018). Näidetena võib tuua genoomi või transkriptoomi sekveneerimine ja ekspressiooniprofiili määramine, krüptiliste liikide või ökoloogiate genoomika tuvastamine ja suunatud sekveneerimine ehk valikuline sekveneerimine (Morozova ja Marra, 2008).

Suurest läbilaskevõimekusest tuleneva infokülluse tõttu sõltuvad NGS meetodid tugevalt suurel määral infotehnoloogia ja informaatika võimekusest. Esimese inimese täisgenoomi järjestamiseks kulus ligikaudu 10 aastat, samas kui tänapäeval on see võimalik 2-3 päevaga. Samuti suurendab NGS rakendamine täppis- ja personaalmeditsiini analüüside täpsust ning avab testimiseks võimalused, mida varem tervishoius ei nähtud (Shendure ja Ji, 2008).

1.3.2 Suunatud sekveneerimine TAC-seq meetodil

NGS meetodid on odavama hinna ja kiirema tulemuse tõttu kasutusele võetud laialdaselt täppis- ja personaalmeditsiinis, kasutades ära NGS meetodite kõrget paralleliseeritavust. Meie töögrupp on välja arendanud kõrge spetsiifilisuse ja tundlikkusega ligeerimispõhise suunatud sekveneerimisel põhineva TAC-seq meetodi, mille tööpõhimõte on näidatud joonisel 2. TAC-seq on NGS meetod biomarkerite suunatud analüüsiks, mis võimaldab vähendada laboris järjestuste amplifikatsiooni sammus tekkida võivaid vigu. Selleks kasutatakse unikaalseid järjestusi ehk molekulaarseid indekseid. TAC-seq ühendab nii NGS sensitiivsuse ja ligatsioon-PCRi spetsiifilisuse ühe molekuli tasemel, olles samal ajal suunatud sekveneerimise tõttu odavama hinnaga kui traditsioonilised NGS meetodid (Teder *et al.*, 2018)



Joonis 2. TAC-seqi meetodi tööpõhimõte tuvastamaks spetsiifilist mRNA või cfDNA lõiku. Ülemises osas on näha parema ja vasaku detektor-oligonukleotiidide paardumine spetsiifilistel tingimustel sihtmärk molekuliga. Detektorjärjestused koosnevad 27 bp pikkusest spetsiifilisest piirkonnast, 4 bp UMI järjetusest ja universaalsete järjestuste sabast 3' ja 5' otstes. Samuti on näha fosforüleeritud 5' otsa spetsiifilise parema detektori otsas, mis järgmise sammuna ligeeritakse vasaku detektoriga. Kolmandal sammul püütakse paramagneetilise keraga paardunud detektor-sihtmärkpiirkond kompleks ning paljundatakse PCRiga. Viimasena tehakse PCR produkti puhastus ja analüüs Illumina sekveneerimise tehnoloogiaga (Teder *et al.*, 2018)

1.4 cffDNA osakaalu leidmise metoodikad

Olemasolevaid cffDNA osakaalu leidmise meetodeid võib jagada kaheks – sekveneerimispõhisteks ehk füüsiliseks ja DNA analüüsipõhisteks ehk analüütiliseks. Peamine vahe seisneb erinevate biotehnoloogiliste võtete kasutamises. Sekveneerimispõhiste meetodite puhul proovitakse leida FF ainult sekveneerimistulemuste põhjal (van Beek *et al.*, 2017). DNA analüüsipõhised meetodid aga nõuavad, et peale sekveneerimise, et tuvastada kromosomaalseid anomaaliaid, leitakse veel lootefraktsioon kas digitaal-PCR (Traeger-Synodinos, 2006) või reaal-aja PCR-i kasutades (qPCR) (Lun *et al.*, 2008). Kuigi need meetodid on FF leidmiseks

usaldusväärsemad, siis on nende puuduseks NIPT lisakulud, sest lisaks sekveneerimisele on ka vajalikud katsed kas mikrokiibi või qPCR-i näol. Hinnatundliku NIPTi jaoks on aga oluline, et hind oleks võimalikult madal ning seeläbi kvaliteetne sõeluuringu test patsientidele kättesaadav.

Enamuse meetodite puhul kasutatakse usaldusväärse FF leidmiseks Y-kromosoomi geenide lugemite osahulga leidmist cfDNA-st. Lähenemine kehtib ainult poissloodete korral. NIPT laboriprotokollid on suuresti kohandatud madala katvusega sekveneerimisandmetele, et sellest andmehulgast leida FF bioinformaatiliste tööriistadega. Järjestused on NGS platvormil sekveneeritud tavaliselt lühikeste pikkustega (40-80 bp), jaotatud regioonidesse, filtreeritud välja duplikaadid ja parandatud GC-osakaalu kasutades LOESS lokaalset regressiooni (van Beek *et al.*, 2017) (Bayindir *et al.*, 2015).

1.4.1 DEFRAG-i meetod

DEFRAG arvutab lootefraktsiooni kasutades kahte näitajat - Y-kromosoomile joondatud normaliseeritud lugemite osahulga ($DEFRAG_a$) ja Y-kromosoomile unikaalsete piirkondade joondatud lugemeid ($DEFRAG_b$) jaotatuna fikseeritud suurustega regioonidesse. Olles Y-kromosoomi põhine, saab seda meetodit kasutada ainult poissloodete lootefraktsiooni arvutamiseks. Sellise meetodi rakendamiseks mõõdetakse kõigepealt proovide Y kromosoomi nullprotsent, kasutades naissoost loodete proove (% $Y_{XX\ loode}$). Seejärel määratakse etteteada meessoost loodete Y protsent (% $Y_{XY\ mees}$). Saadud tulemused on piirväärtusteks valemis:

$$\text{DEFRAG}_a = \frac{\%Y_{XY\ loode} - \%Y_{XX\ loode}}{\%Y_{XY\ mees}}$$

Teise poole arvutamiseks kasutatakse unikaalseid Y-kromosoomi piirkondi. Unikaalsete piirkondade leidmiseks tuleb välistada üle genoomi korduvad piirkonnad, milleks jaotatakse tüdrukloodete lugemid madala katvuse kompenseerimiseks fikseeritud suurusega 1 Mb regioonidesse ja sorteeritakse välja need, millesse langes rohkem kui 1 järjestust. Saadud alamgrupp esindab Y-spetsiifilisi piirkondi, mille lugemite mediaani kasutatakse lootefraktsiooni esitamiseks. (van Beek *et al.*, 2017).

1.4.2 Bayindir-i meetod

Lootefraktsiooni täpsemaks määramiseks on madala katvusega sekveneerimise puhul piiravaks faktoriks kättesaadavate joondatud järjestuste arv ema verest. Sellega võitlemiseks kasutatakse suurematesse piirkondadesse lahterdamist, kuid ka arvestatava piirkonna suurendamist. Bayindir on välja töötanud meetodi, mis ei ole tingimata Y-spetsiifiline NIPT, sest kasutab kõiki autosoomseid ja sugukromosoome trisoomiate tuvastamiseks, kuid siiski suudab lootefraktsiooni arvutada ainult Y-kromosoomi järgi (Bayindir *et al.*, 2015). Järjestused jaotatakse 5Mb regioonidesse 50kb sammuga ning saadud lugemid iga regiooni kohta jagatakse summaarse lugemite arvuga. Seejärel arvutatakse z-väärtused kasutades autosomaalsete regioonide (Chr_{auto}) ning X-kromosoomi (Chr_X) mediaane:

$$BAYINDIRz = \frac{med(Chr_{auto}) - med(Chr_X)}{med(Chr_{auto})} * 2$$

Saadud esinduslikud arvud kasutatakse *z*-skoori ehk normaliseeritud lugemite differentsiaalsusemäära arvutamiseks üle kromosoomi. Saadud *z*-skooriga hinnatakse huvipakkuva kromosoomi osahulka võrreldes ülejäänud autosoomidega. Kõikumised eeldatust on indikatsiooniks aneuploidiale (Bayindir et al., 2015).

Tulemuste valideerimiseks ja soo määramiseks arvutatakse välja ka lootefraktsioon Y-kromosoomi põhjal, seega ainult poisslastel. Siinkohal leitakse sarnaselt DEFRAG-ile unikaalsed sugukromosoomide piirkonnad ning arvutatakse lootefraktsioon:

$$BAYINDIR_{ff} = \frac{med(Chr_Y^{mask})}{med(Chr_{auto})} * 2$$

1.4.3 SANEFALCON-i meetod

SANEFALCON (*Single reAds Nucleosome-basEd FetAL fraCtiON – üksik-lugemi nukleosoomipõhine lootefraktsioon*) määrab lootefraktsiooni kasutades ära vähem-degradeerunud cfDNA, mida leidub nukleosoomide ümber. Seetõttu on see meetod sõltumatu loote soost, ehk ei ole vaja Y-kromosoomi (Straver *et al.*, 2016). SANEFALCON meetod loob nukleosoomide profiili üle genoomi jälgides ülegenoomsete järjestuste positsioonide ja nukleosoomide poolt mõjutatud

teadaolevate piirkondade kokkulangevusi ning loetakse joondatud järjestused nendes kohtades kokku. Ülegenoomsed nukleosoomiprofiilid on seejärel tuletatud keskmistest lugemitest nendes piirkondades. Profiilidel on aga erinevad väärtused samas kohas, mis tuleneb varem kirjeldatud erinevustest cfDNA ja cffDNA vahel, peamiselt fragmendipikkuse tõttu. Nende erinevuste tuvastamiseks kasutab SANEFALCON hinnangulise lootefraktsiooni andmiseks lineaarregressiooni mudelit.

1.4.4 SeqFF

Selles töös käsitletav SeqFF meetod on neist kõige uuem ning kõige levinum meetod FF arvutamiseks. SeqFF kasutab kahte regressioonimudelit - elastilist närvivõrku (*Efficient neural network* - Enet) ja kaalutud järgupõhist valikut (*Weighted Rank Selection Criterion* - WRSC) - et disainida mudel, mis on võimeline ennustama lootefraktsiooni ülegenoomsete sekveneerimislugemite pealt. Selle meetodi treenimiseks kasutati 25 312 ema vereseerumi proovi ja kinnitamiseks kasutati 505 vereseerumi proovi (Kim *et al.*, 2015). Meetod põhineb kaalutud regressioonimudelil, mis loeb lugemeid 50 kbp järjestikuselt korduvates regioonides ehk regioonides üle autosoomide ning ennustab Y-kromosoomi regioonide lugemite arvu arvestades ka kovariatsiooni või üldistatud Y-kromosoomi osahulka. Lootefraktsiooniks on kahe tulemuse keskmine. Seega on tegemist andmetega, mis on treenitud poissloodete andmete peal, kuid mis on edukalt lootefraktsiooni ennustanud ka tüdrukloodete puhul.

SeqFF originaalses uurimuses leiti piirkonnad, mis demonstreerivad suurt eripära emapoolse ja loote cfDNA vahel. Seega saab sekveneeritud proovi piirkondade lugemite ja treenitud mudeli kordajate põhjal väita, milline on lootefraktsioon suurus. Selliseid piirkondi leiti vastavalt uurimusele järgnevalt: (i) täisealise ema ja areneva loote erinevalt metüleeritud DNA piirkondade vahe, (ii) tugevalt heterosügootsete SNP piirkondade spetsiifiliste variantide lugemite vahe ja (iii) erineva pikkusega cfDNA fragmentide vahe.

On tähele pandud, et erineva pikkusega cfDNA fragmendid, eriti enamjaolt lootepäritolu fragmendid pikkusega <150 bp, on positiivselt korreleeritud piirkonda jäävate eksonite arvuga, mis omakorda on seotud suurema GC osakaaluga. SeqFF algoritmis on esmaseks tähendusrikaste

piirkondade kriteeriumiks kõrge GC ja eksonite sisaldus, mille kriitilised tasemed määrasid Kim ja tema kolleegid (Kim *et al.*, 2015).

2 Eksperimentaalosa

2.1 Töö eesmärgid

NIPTi eelis kombineeritud esimese trimestri sõeluuringu ees on oluliselt kõrgem sensitiivsus ja spetsiifilisus. NIPTi jaoks on kriitiline, et analüüsi käigus saaks usaldusväärselt määrata loote DNA fraktsiooni hulka analüüsitud raseda vereproovist. See annab võimaluse hinnata loote sekveneeritud geneetilise info paikapidavust ning anda usaldusväärsushinnangu igale analüüsitud proovile.

Käesoleva töö üldiseks eesmärgiks on leida, kas cfDNA põhise analüüsi puhul leidub genoomseid DNA piirkondi, mis on lootefraktsiooni seisukohast informatiivsed ning mida oleks otstarbekas TAC-seq meetodil (lisaks aneuploidiate määramisele) analüüsida loote DNA fraktsiooni määramiseks. Eesmärgini jõudmiseks jaotati tööd järgmisteks eraldiseisvateks alameesmärkideks:

1. uurida olemasoleva SeqFF mudeli treenitud korrelatsioonikordajaid ja tõlgendada neid käesoleva töö seisukohast;
2. hinnata Eestist kogutud madala katvusega sekveneeritud proovide lugemite katvust erinevate regioonisuuruste juures ja minimaalse kõlbliku regioonisuuruse leidmine;
3. kasutades Eestis sekveneeritud NIPT testi madala katvusega proovide andmeid, leida järjend kõlbliku regioonisuurusega piirkondadest, mis on kõige olulisemad FF arvutamisel;
4. kontrollida, kui hästi sobivad leitud piirkonnad ennustama võrreldes algse SeqFF meetodiga lootefraktsiooni käesolevate proovide korral leides korrelatsiooni originaalse SeqFF meetodi ja käesolevas töös välja arendatud meetodi vahel;

Kirjanduse ülevaates kirjeldatud kriteeriumite tõttu keskendutakse käesolevas töös suunatud järgmise põlvkonna sekveneerimise meetodile TAC-seqile, mis õigete sihtmärkpiirkondade puhul suudaks hinnata lisaks aneuploidiatele ka lootefraktsiooni (Teder *et al.*, 2018).

2.2 Materjal ja metoodika

2.2.1 Andmete päritolu

Töö viidi läbi Tervisetehnoloogiate Arengute Keskuses ja hõlmas 368 järjestatud genoomi, mis on kogutud Tartu Ülikooli Kliinikumist ja Ida-Tallinna Keskhaiglast. Proovide võtmiseks andsid doonorid kirjaliku loa. Kasutatud proovide doonorite keskmine vanus oli 33 aastat. Sekveneerimiseks kasutati Illmina NextSeq 500 platvormi, keskmise katvusega 0.32x ja keskmise lugemi pikkusega 85 aluspaari. Töös kasutatud proovid on sekveneeritud ühe-otsalise (*single-end*) sekveneerimise meetodil. Käesoleva töö autor ei osalenud proovide võtmise ja laboratoorse töötlemise protsessides.

Andmeanalüüsiks vajalik andmetöötlus läbis järgmised etapid:

1. Toorsekveneerimisandmete demultipleksimine kasutades Illumina kommertsiaalseid programme ja platvorme (*bcl2fastq*). Programmi väljundina saadakse iga proovi kohta nelja raja kaupa GZ kompressiooniga FASTQ formaadis kvaliteediskooriga geneetiline järjestus.

2. FASTQ failid (kogumaht 123GB) pakitakse lahti ja kaardistatakse indekseeritud GRCh37/hg19 Homo Sapiens referentsgenoomi vastu kasutades programmi *bowtie2* v2.3.4.3.

3. Järjestatud lugemite alguspositsiooni genoomis leidmine ja lugemite filtreerimine toimub *samtools* v1.6 programmiga. Selle etapi väljundiks on filtreeritud lugemite stardipositsioonide list formaadis SAM.

4. Positsioneeritud lugemite fail loetakse sisse R keeles kirjutatud skripti, mis lahterdab lugemid ja kasutades eelmääratud SeqFF meetodi parameetreid arvutab sisendproovi lootefraktsiooni väärtuse. (Kim *et al.*, 2015)

2.2.2 Bowtie2

Bowtie2 on programm, mis järjestab sekveneerimise lugemid pikkadele referentsjärjestustele kasutades Burrows-Wheeler Transformatsiooni (BWT) algoritmi. Programmi eesmärgiks on tagada kiirus, juhumälusäästlikkus ja kerge skaleeritavus, utiliseerides paralleelprotesseerimise võimsust (Langmead and Salzberg, 2012). Arvestades sekveneerimisandmete kirjeldatud omadusi, on optimaalne kasutada *bowtie2*-d, sest kaardistada on vaja keskmiselt pikki lugemeid suhteliselt suurele referentsgenoomile. *Bowtie2* on kaardistamisel laialdasemalt kasutatud kui alternatiivsed programmid, näiteks *Bowtie*, *BWA*, *SOAP2*, *MAQ*, *RMAP*, *GSNAP* või *Novoalign* (Hatem *et al.*, 2013). Lisaks on programm optimiseeritud Illumina sekvenaatorite andmetele.

Käesolevas töös on lugemite kaardistamine läbi viidud järgneva käsuga:

bowtie2 --very-sensitive -X 500 -q $INP --norc -x $REF --no-unal -p 10

Kus vastavad valikud toimivad järgnevalt:

1. --very-sensitive

   Märgib, et kaardistamine toimub väga tundlikus režiimis, mis välistab lokaalse järjestamise, s.t. kaardistatavad fragmendid peavad täielikult vastama referentsgenoomile. See käsk on ka lühend eelseadistatud valikutele, mis välistab valepaardumiste lubamise (`-N 0`), seab otsisõnade pikkuseks 20 (`-L 20`), mitu kordusotsingut lubab teha (`-R 3`) ja mitu korda võib otsing ebaõnnestuda (`-D 20`).

2. -X/--maxins

   Tähistab kui pikk võib kaardistatav DNA fragment olla, käesolevas töös on see väärtus 500nt. Miinimumväärtus (-I) on vaikeväärtusena 0.

3. -q

   Sisendfailide (FASTQ) asukoht arvutis.

4. -norc

   Määrab ära, et programm ei järjesta paardumata lugemeid Crick'i referentsi järgi

5. -x

   Referentsgenoomi asukoht arvutis.

6. --no-unal

   Programm ei väljasta SAM formaadis infot lugemitest, mida ei õnnestunud kaardistada.

7. -p

   Paralleelse töötlemise kasutamiseks määratud lõimede arv.

2.2.3 Samtools

Samtools on programm SAM (Sequence Alignment/Map) failide kiireks manipuleerimiseks ja käitlemiseks (Li *et al.*, 2009). Töötlus toimub binaarse SAM-i ehk BAM (Binary SAM) formaadis, mis teeb töötluse programmaatiliselt kiiremaks ja efektiivsemaks, lisaks on binaarsel kujul talletatud info nii püsimälus kui töö käigus vahemälus kompaktsem. Samuti kasutati *samtools*-i kaardistamise käigus *bowtie2*-ga määratud kvaliteediskoorile põhinedes lugemite filtreerimiseks mis on vajalik madala katvusega sekveneerimisproovide puhul võimalike valelugemite müra vähendamiseks. *Samtoolsi* kasutati lugemite stardipositsioonide failide väljastamiseks edasiseks allavoolu andmetöötluseks, kasutades funktsionaalsust *view* ja selle alamfunktsioone:

8. -q

   Filtreerib välja lugemid väiksema kvaliteediskooriga kui määratud. Käesolevas töös oli -q väärtuseks valitud 30, seega edasiselt kasutati ainult lugemeid kaardistamiskvaliteediskooriga (*MAPQ*) üle 30.

9. -S

   Valik, mis määrab *samtools* programmi ühilduma vanemate *samtools*-i versioonidega.

2.2.4 Programmeerimiskeel R

Selle töö tulemused on leitud kasutades programmeerimiskeelt R (v3.5.1). R on funktsionaalne programmeerimiskeel, milles on võimalik kasutada imperatiivseid elemente. R-keelt kasutatakse

statistikas ja andmeteaduses peamiselt selle kiiruse ja sarnasuse tõttu matemaatilise kirjastiiliga, samuti leidub palju andmebaasihalduskeeltest tuletatud funktsionaalsusi. Enamjaolt on funktsioonid kirjutatud madalama taseme C-keeles, mis universaalse kasutuse tõttu igast operatsioonisüsteemis tagab R-keele portatiivsuse igale platvormile. Seega on töös kasutatud kood kasutatav ka igas tavalises personaalarvutis.

Lisaks R-keele baasfunktsionaalsustele on käesolevas töös veel kasutatud järgnevaid lisapakette:

- *Tidyverse* pakette, mille eesmärgiks on koodi loetavuse ja kiiruse parandamine, ning mille alla kuuluvad järgnevad paketid:
    - *ggplot2* visualiseerimistööriistad.
    - *dplyr* mugavad ja kiired andmete manipuleerimistööriistad.
- *Snow / foreach / doMC* paralleelarvutusi võimaldavad pakette, mis kiirendavad oluliselt suurte andmemahtude korral triviaalsete arvutuste tegemist. Need paketid on vajalikud suuremahulisel andmetöötlusel R-keelega, sest standardne R ei ole võimeline kasutama paralleelset töötlust.

2.3 Tulemused

2.3.1 SeqFF mudeli genoomseid regioone kirjeldavate sisendparameetrite uurimine ja tõlgendamine

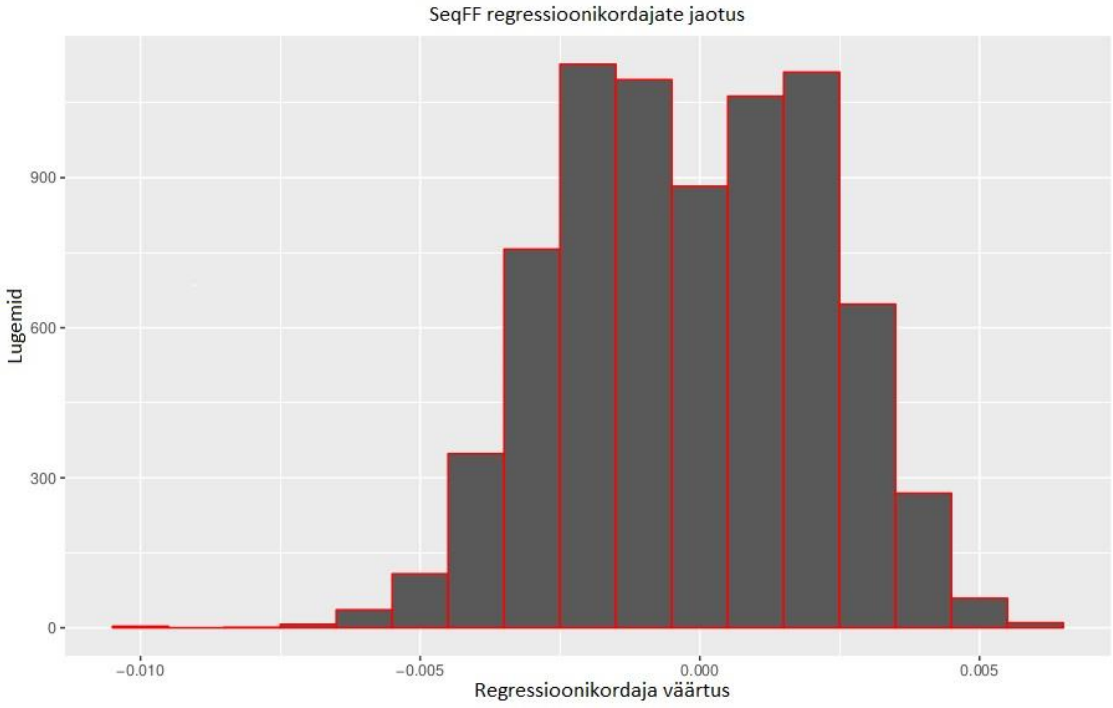
Käesolevas töös oluliste piirkondade filtreerimistingimusena kasutati SeqFF arendajate Kim-i ja teiste poolt treenitud Enet-i mudeli genoomi regioonide koefitsiente ehk regressioonikordajaid. Regressioonikordaja tuleneb Enet-i toimimispõhimõttest, mis leiab suhte tulemusmuutuja ja ennustajamuutujate vahel, milleks olid vastavalt Y-kromosoomi-põhised lootefraktsiooni arvutused ja normaliseeritud autosoomide 50kbp regioonide lugemid. Viimasel juhul normaliseeriti SeqFF autorite poolt hinnatava sekveneeritud proovi lugemid arvestades G ja C nukleotiidide osakaalu piirkonnas. Saadud korrelatsioonikordajad olid seejärel kasutatavad järgnevate 50kbp regioonidesse jaotatud ülegenoomsete sekveneerimislugemite põhjal lootefraktsiooni arvutamiseks, mistõttu need SeqFF programmiga kaasasolevad

korrelatsioonikordajad on käesolevas töös heaks indikaatoriks, et hinnata, millised genoomsed regioonid on lootefraktsiooni arvutamise seisukohalt olulised ehk informatiivsed.

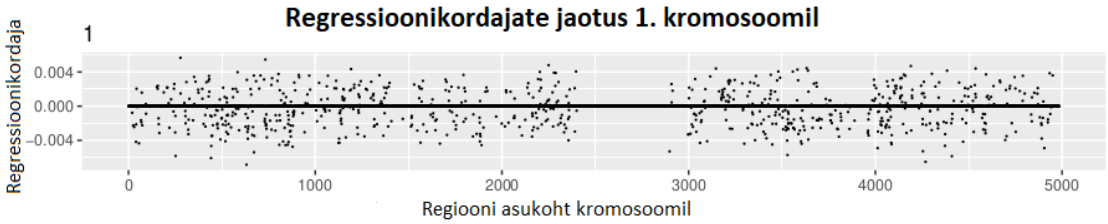
Lihtsustades lähtudes Eneti loojate Hui Zou ja Trevor Hastie järgi (Zou ja Hastie, 2005), on regressioonikordajad järgneva lineaarse regressiooni mudeli kordajad:

$y = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p$

kus y on tulemusmuutuja ja $x_1, \ldots, x_p$ ennustajamuutujad. β_p on sel juhul kordaja, mida Elnet lähenemine leiab, andes tulemuseks järjendi piirkondadele vastavatest koefitsentidest. SeqFF-i puhul on $x_1, \ldots, x_p$ järjestikused 50 kbp piirkonnad ja y Y-kromosoomipõhine lootefraktsiooni tulemus. Lihtsustatud valemist võib ka järeldada, et regressioonikordajad on indikatooriks, millised genoomsed piirkonnad on keskmiselt olnud olulisemad ehk informatiivsemad loote DNA fraktsiooni hindamiseks. Seetõttu on käesoleva töö raames huvipakkuvad peamiselt suurema positiivse ja negatiivse väärtusega kordajad, sest need näitavad, et kordajale vastava 50 kbp piirkonna lugemite hulga muutus SeqFF treeningandmetes on olnud informatiivsed loote DNA fraktsiooni arvutamisel.



Joonis 3. Joonisel on kujutatud kordajate jaotuvust üle genoomi. Jooniselt on näha regressioonikordajate jaotuvuse keskpunkte ±0.002 juures. Suurim regressioonikordaja väärtus on 0.006309316 ja väikseim -0.01047021, keskmine väärtus -1.957425e-05.



Joonis 4. Ülevaade regressioonikordajate jaotusest üle esimese kromosoomi stardipositsioonide järgi, jaotatuna 50 kbp piirkondadeks.

Jooniselt 4 ja lisa 2 jooniselt on näha, et regressioonikordajad jaotuvad regioonist sõltumatult, ehk ei ole seotud lugemite hulga või pikkusega. Katmata jäävad sugukromosoomid, tsentromeeride piirkonnad ja kromosoomid 13, 18, 21, mis tulenevad GC sisalduse ja keskmise fragmendi pikkuse järgi filtreerimisest ja ülesobituse vältimisest. Mainitud kromosoomid tekitavad ülesobitust, sest SeqFF uurimuse käigus treenitud mudelisse kaasamise puhul esineks uuritavad ebanormaalsused nii tulemustes kui lähteandmetes.

2.3.2 Minimaalse informatiivse regioonisuuruse leidmine

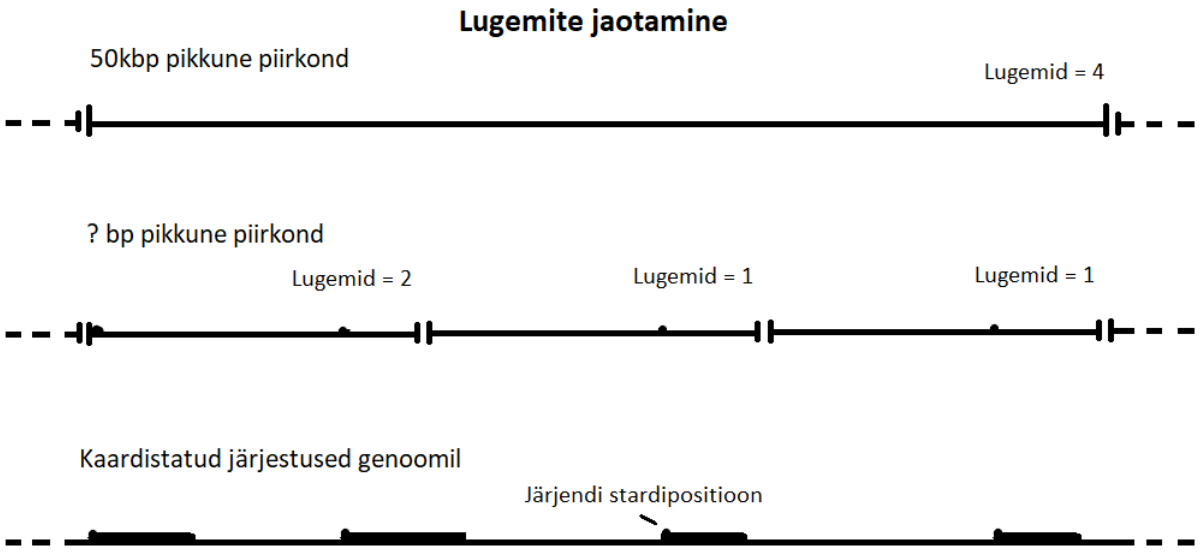
Suunatud ligeerimispõhised sekveneerimismeetodid suudavad saavutada molekulitasemel resolutsiooni DNA korduste tuvastamisel, kuid tuginevad PCR meetodil, mis toimib ainult spetsiifiliselt disainitud praimeritega konkreetselt defineeritud genoomsetes regioonides. Sellised meetodid on efektiivsed geeni- või SNP-põhise ekspressiooniprofiili saamiseks, kuid on liiga kulukad ülegenoomsetel uuringutel. Seega pole SeqFF meetodis kasutatud 50 kbp pikkused piirkonnad piisavalt hästi defineeritud ehk piisavalt lühikesed käesoleva töö eesmärgi - suunatud sekvneerimise meetodi - rakendamiseks, sest ligeerimispõhised meetodid vajavad alla 150 bp pikkuseid amplikone. SeqFF meetodi kohandamiseks on seega vaja leida piisavalt pikkasid genoomseid regioone, mis oleksid madala katvusega genoomide põhjal piisavalt kõrge lugemite

katvusega, kuid ka piisavalt lühikesed, et saaks disainida regiooni sobivaid praimereid ja amplikone.

Vereplasmas leiduva DNA puhul on suurimaks piiranguks geneetilise materjali vähene kontsentratsioon ja selle kõrge fragmenteeritus, seega ühe proovi leitud sihtmärkregioonides ei pruugi olla tuvastatavaid lugemeid teises cfDNA proovis. Madala katvusega proovide puhul peab seega leidma minimaalse informatiivse regioonisuuruse, mille põhjal oleks võimalik piisava lugemite hulga pealt ennustada lootefraktsiooni. Lugemite piirkondade jaotamiseks jaotati olemasolevate 368 proovi referentsgenoomile joondatud lugemid järjestikult korduvatesse x_{suurus} suurusega piirkondadesse stardipositsioonide järgi, ehk iga lugem, mille s_{pos} jääb n_{genoom} suurusega genoomis vahemikku

$$\frac{x_{genoom}}{x_{suurus}} * x_{num} \leq s_{pos} < \frac{n_{genoom}}{x_{suurus}} * (x_{num} + 1),$$

kus x_{num} on piirkonna number.

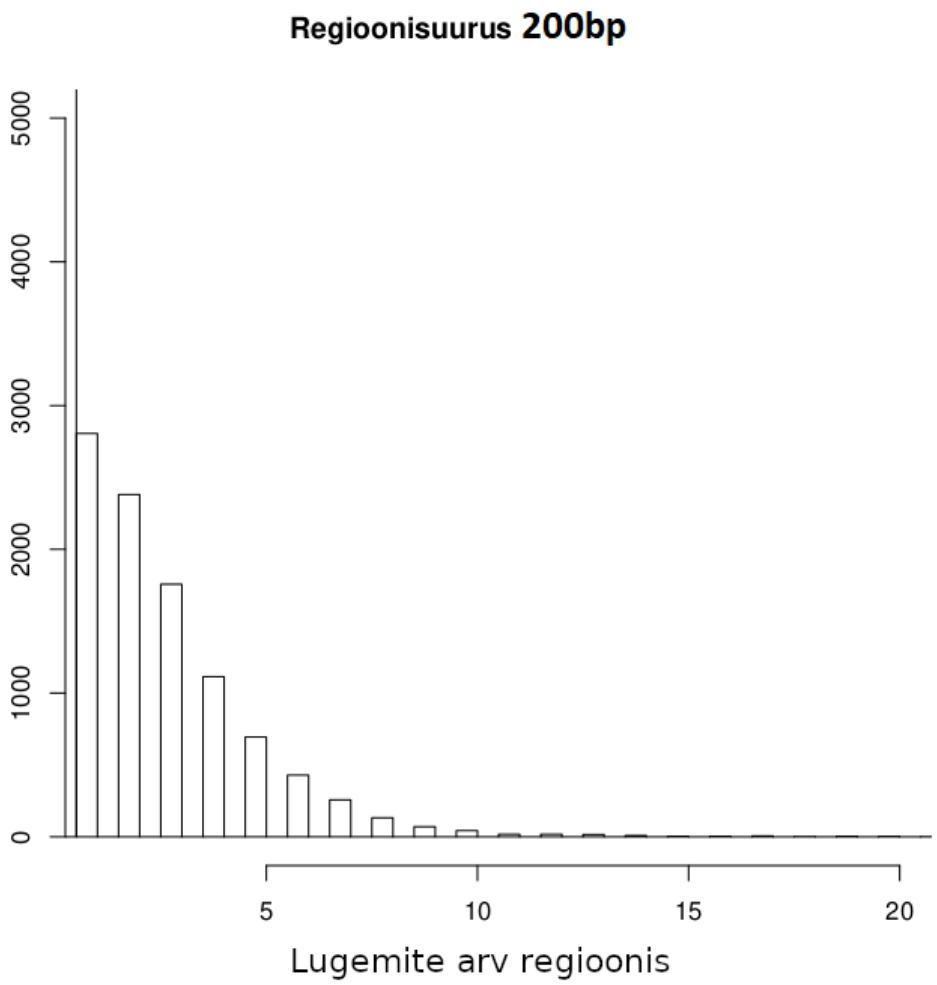


Joonis 5. Sekveneeritud ja joondatud lugemite regioonidesse jaotamine.

Sobiva vahemiku leidmiseks liideti kõikide käesolevas töös kasutatud 368 proovi kaardistatud lugemid ja loodi histogrammid tulemuste visuaalseks uurimiseks. Antud histogrammid kujutavad

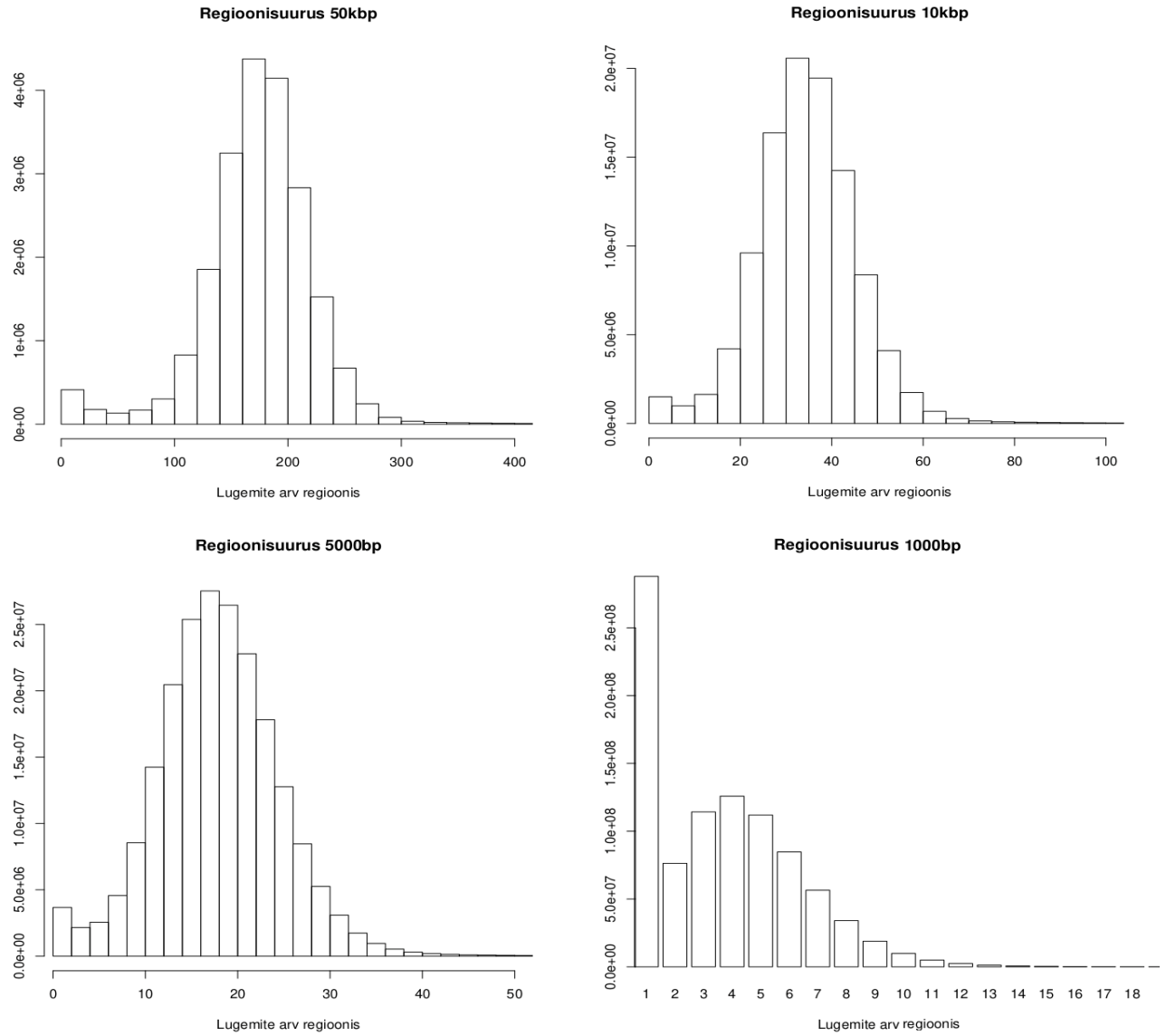
käesolevas töös valitud suurusega piirkondadesse jaotuvate lugemite hulka. Nende jaotuste põhjal on võimalik hinnata, kas valitud regioonisuuruse juures esineb piisavalt kõrge katvusega ja märkimisväärselt eristuvate lugemite piirkondi.

Esimeseks katseks kasutatakse 200 bp pikkuseid regioone ühe proovi näitel, sest sellise pikkusega piirkonnad oleksid optimaalsed TAC-seq meetodil edasi uurimiseks. Sellise lahutuse juures oleks võimalik leida tähendusrikkaid piirkondi, mis on tingitud DNA kaksikheeliksi pakkimisest ümber histoonide 146 bp kaupa ja on seetõttu degradeerumisele vastupidavamad (Luger *et al.*, 1997). Kogu genoomi ühtlase regioonideks jaotamise korral pole histoonidest piirkondade paiknemine vastavuses, sest DNA pakkimine ei sõltu kromosoomide pikkusest, kuid üle paljude genoomide peaks histoonidest tingitud järjestuste säiluvus olema märgatav jaotuse muutja.



Joonis 6. Histogramm kujutamas 200 bp pikkuseid regioone ühe proovi näitel, arvestades välja 0-väärtused.

Joonisel 6 on näha 200 bp pikkuste piirkondade lugemite jaotus, millest võib näha, et regiooni suurus on liiga väike, et leiduks piisavalt kõrgelt kaetud ja informatiivseid piirkondi madala katvusega sekveneeritud proovidest. Suurem osa regioone langeb siin väljapoole lahutust, ehk lugemite jaotus ei ole genoomselt unikaalne, vaid tõenäosus saada lugemi vastavasse piirkonda on ligikaudselt ühtlane üle kogu genoomi. Samuti puudub mõju lugemitele mainitud degradatsiooni vähendavatelt teguritelt. Seega tuleb kasutada suuremat regioonisuurust, et leida tähelepanuväärseid piirkondi.

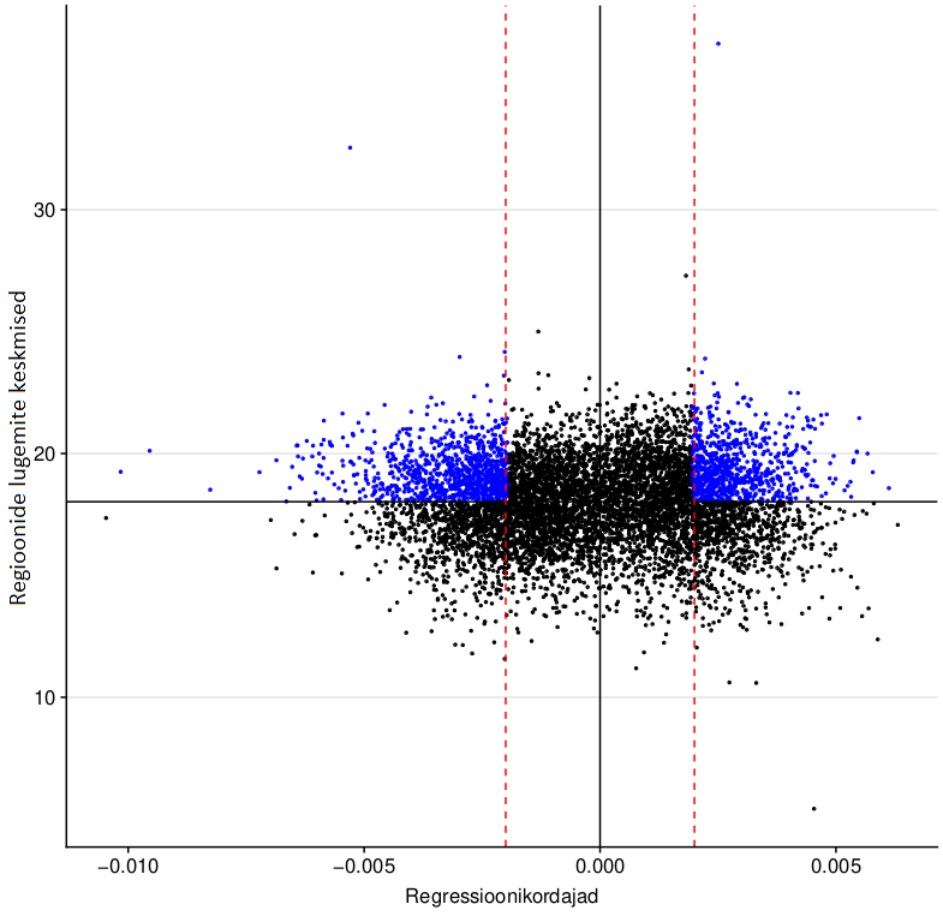


Joonis 7. Histogrammid kujutamas 50000 bp, 10000 bp, 5000 bp ja 1000 bp pikkuste järjestikuste regioonide lugemite jaotust üle genoomi, arvestades välja regioonid, kuhu langes 0 lugemit.

Joonisel 7 võib näha, et SeqFF-is kasutatavate 50 kb pikkuste piirkondade korral on lähedane normaaljaotusele, kus kõige enam esineb regioone, kuhu on jäänud 170-180 lugemit. Samuti esineb ka palju genoomseid regioone, kus katvus jääb 0-10 lugemi lähedale, mis on tingitud madalast sekveneerimissügavusest ja cfDNA lugemite ebaühtlasest jaotusest üle genoomi. Selle põhjuseks võib olla väiksem metüleerituse või pakituse tase, mis suurendab degradeerumist nendes piirkondades, andes vähem lugemeid. Järgmiste 10 kbp ja 5 kbp histogrammide puhul on märgatav keskmise lugemite hulgaga regioonide väärtuse langemine, mis on tingitud väiksemast tõenäosusest, et kaardistatud lugem langeks regiooni. Arvestades antud regioonipikkuste lugemite hulka meie andmetes ja TAC-seqi proovide disainimiseks sobivate optimaalsete regioonide pikkust, otsustati käesolevas töös jätkata 1000 bp pikkuste kandidaat-regioonide uurimisega.

2.3.3 Sekveneerimislugemitega kõrgelt kaetud genoomsete regioonide leidmine Eesti NIPT-i madala katvusega läbiviidud täisgenoomide sekveneerimisandmetest

Kasutades sobiva lugemite jaotuse piirkonna pikkust 1000 bp, mindi töös edasi regioonide jaotuse uurimisega olemasolevate regressioonikordajate suhtes. Selleks visualiseeriti kõigepealt 50 kbp regioonisuurusega jaotatud Eesti NIPT'i madala katvusega sekveneeritud proovide lugemite jaotus SeqFF regressioonikordajate suhtes.



Joonis 8. Suhe 50 kbp regioonidesse jaotatud Eesti proovide lugemite keskmise ja vastava piirkonna regressioonikordajate vahel. Joonisel on märgitud abtsissteljel joon, mis kujutab lugemite keskmiste keskmist 18,02 ja ordinaatteljel regressioonikordaja väärtusel 0 ja punase katkendliku joonega väärtusel ±0.002. Välja on jäetud lugemid, mille kordaja on 0.

Joonisel 8 on näha 50 kbp regioonide keskmiste lugemite ja regressioonikordaja suhet, millest võib näha, et enamik regressioonikordajatega regioone omavad >10 lugemit. Seetõttu tekib kogum abtsissteljel regressioonikordaja 0-väärtuse ja ordinaatteljel joonis 3 jaotusest nähtuna lugemite keskmise ümber. Samuti on märgatav joonise 3 eeskujul tihedam kogum regioone regressioonikordaja ±0.002 juures. Kõige kaalukamad piirkonnad FF arvutamise seisukohast SeqFF mudelis joonisel 8 on piirkonnad sinise värviga, kus regressioonikordaja absoluutväärtus on suurem (>0.002) ja keskmine lugemite arv on kõrgem kui keskmiste lugemite keskmine indiviidi järgi. Viimane on oluline, sest mida suurem on keskmine lugemite arv, seda suurem on tõenäosus, et regiooni satub järjestus, mida enamjaolt leidub cfDNA'st. Seega saab lühemate

regioonisuuruste puhul kasutada esmase regioonide filtrina 50 kbp regressioonikordajate nullist erineva väärtuse olemasolu.

Pärast sobiva sihtmärkregioonide suuruse (1000 bp) leidmist eelmises peatükis, oli järgmiseks sammuks leida 368 Eestist kogutud sekveneeritud rakuvaba DNA proovide andmetest parameetrid, mille järgi leida lootefraktsiooni arvutamiseks sobivad genoomsed piirkonnad, mida kasutada edasises analüüsis. Selleks jaotati sarnaselt eelnevale (joonis 5) iga proovi leitud järjestused järjestikustesse 1000 bp pikkustesse regioonidesse ja kõikide cfDNA proovide jaotatud lugemid normaliseeriti võrdlemise jaoks kasutades valemit

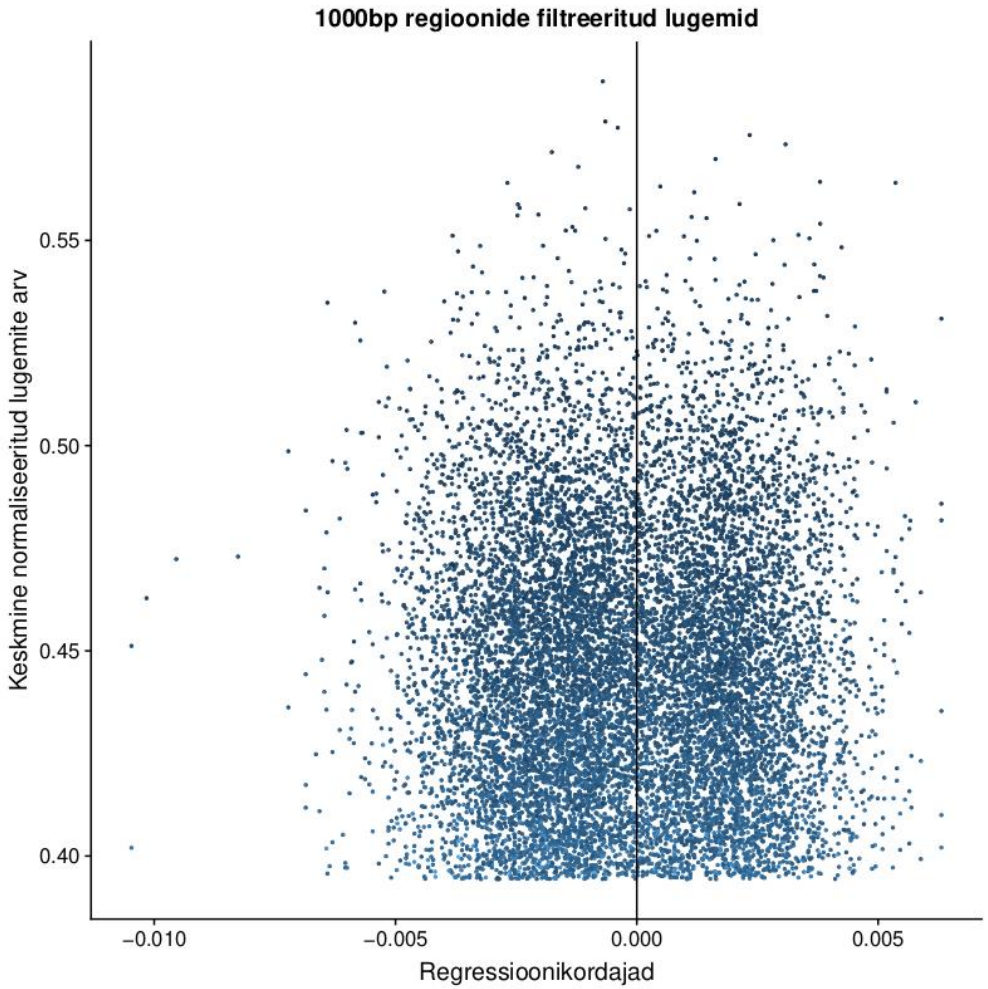
$$CPM_i = \frac{X_i}{N} * 10^6,$$

kus CPM_i on normaliseeritud lugem, X_i algne lugemi väärtus ja N ühe sekveneeritud proovi kõikide lugemite summa.

Selekteerimiseks valitud parameetrid olid järgmised:

- 50 kbp regioonisuurusest päritud regressioonikordajad, seega igal viiekümnel 1000 bp regioonil on sama regressioonikordaja, mis on talle vastaval 50 kbp regioonil. Nendest filtreeriti välja regressioonikordaja väärtustega „0“ piirkonnad.
- Eelnevalt kirjeldatud 50 kbp pikkuste regioonide filter, mille pärivad ka 1000 bp piirkonnad.
- Üle kõikide proovide lahterdatud lugemite aritmeetilised keskmised, millest valiti välja regioonid, millel leidus üle keskmise lugemeid regioonide järgi.
- Regioonipõhine standardhälve. Selle parameetri põhjal valiti välja regioonid, mis olid väiksemad, kui keskmine regiooni standardhälbe väärtus.
- Iga 1000 bp normaliseerimata regiooni kohta üle proovide leitud 0-lugemite arv, millest valiti välja piirkonnad, millel leidus 0-lugemite osakaal alla keskmise.

2.3.4 Saadud piirkondade andmete katvus SeqFF-i genoomseid regioone iseloomustavate regressioonikordajate kontekstis



Joonis 9. Graafik näitab suhet lahterdatud filtreeritud regioonide normaliseeritud lugemite keskmist üle proovide ja vastava piirkonna regressioonikordajate näitel. Joonisel on märgitud helesinisega piirkonnad, mis on suurema 0 lugemiga regioonide osakaaluga ja tumedamalt väiksema osakaaluga.

Sarnaselt joonisele 8 on joonisel 9 samuti märgatav regressioonikordajatele omane jaotus. Peale filtri rakendamist jäi algsest 2 535 909 regioonist alles 12 119, ehk kaetud on ligikaudu 12 miljonit nukleotiidi tervest inimese genoomist võrreldes algse 2,5 miljardi nukleotiidiga. Filtreerimistingimustest tingivalt on normaliseeritud lugemite keskmiste alampiir 0.3943503. Jooniselt on näha ka 0-lugemitega regioonide osakaal algsetest normaliseerimata lugemitest,

millest võib järeldada, et kõrgema keskmise arvuga ja regressioonikordajaga regioonid on ka parema katvusega 1000bp suuruse regiooni korral. Saadud filtreeritud piirkonnad moodustavad regioonide maski, mida saab kasutada edasi olemasolevate proovide lootefraktsiooni arvutamisel.

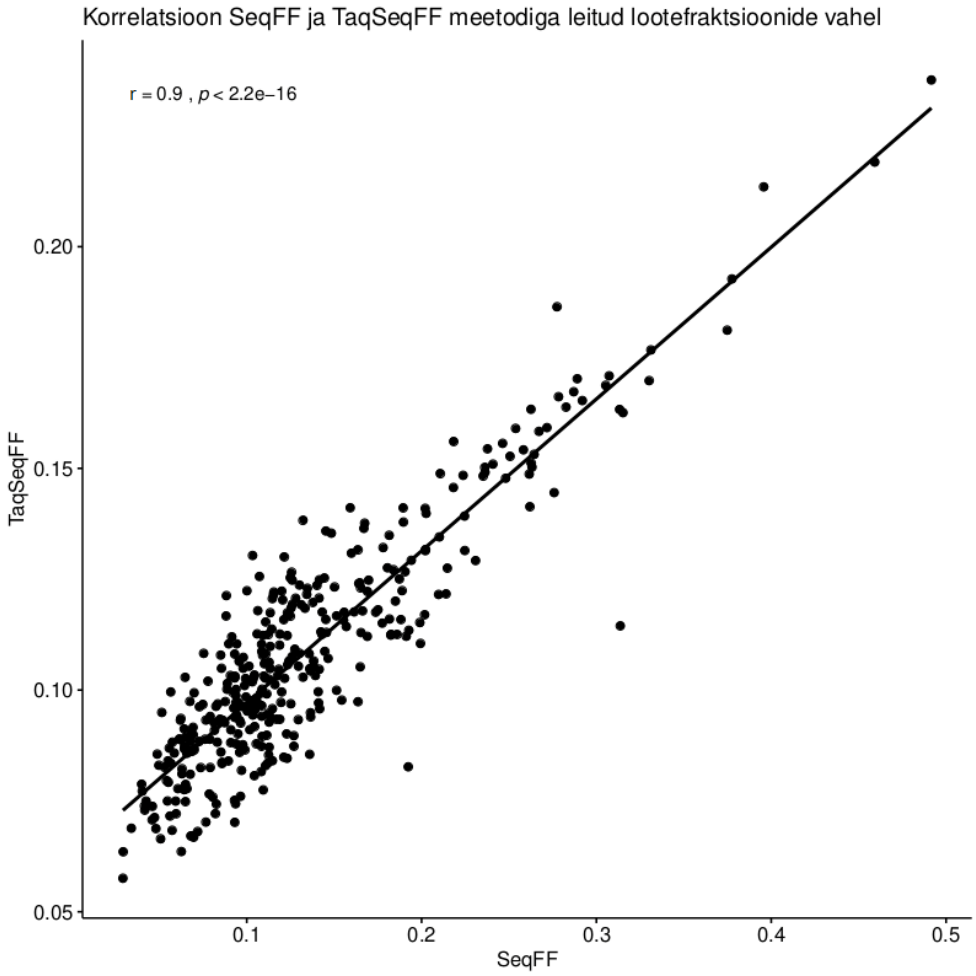
2.3.5 Filtreeritud piirkondade efektiivsus lootefraktsiooni hindamisel

Piirkondadesse jaotatud lugemitest tuletatud uus filter on kasutatav SeqFF algoritmis kirjeldamaks tähendusrikkaid 1000 bp pikkuseid piirkondi. SeqFF algoritm, mis on algselt loodud 50 kbp piirkondade põhjal lootefraktsiooni arvutamiseks, ei ole otseselt üle kantav väiksemate lahutuste kasutamiseks, sest hinnang põhineb tehisnärvivõrgu meetodil eeltreenitud regressiooniväärtustel. Seetõttu tuleb saadud regioonide filtri efektiivsuse võrdlemiseks SeqFF meetodiga leida iga 50 filtreeritud 1000 bp piirkonna kohta suurim filtreeritud piirkonna lugemite arv, mida kasutada vastava 50 kbp piirkonna iseloomustamiseks.

Selleks kirjutati programm R keeles, mis põhineb SeqFF meetodil. Programm kasutab sisendina *samtools'i* poolt genereeritud uuritava sekveneeritud ja kaardistatud proovi kromosoom – alguspositsioon paare, mis seejärel jaotatakse soovitud suurusega regioonidesse nagu kirjeldatud joonisel 5. Seejärel loetakse mällu SeqFF uurimusest pärit andmed 50 kbp regioonide kohta, mis koosneb regressioonikordajate ja filtreerimistähiste tabelist. Mälus olevad proovi lahterdatud lugemid seejärel filtreeritakse kasutades selles töös leitud piirkondade loendit. Filtreeritud proovid ühendatakse 50 kbp regioonide tabeliga põhimõttel, et iga 50 kbp regiooni info omistatakse sinna kuuluvatele 1000 bp regioonile. Järgmisena leitakse iga 50 kbp regiooni kohta temas olevatest 1000 bp piirkondadest maksimum ja kasutatakse saadud regioonide lugemite tabelit SeqFF meetodil lootefraktsiooni arvutamisel, mis koosneb autosoomide filtreerimisest, lugemite skaleerimisest ning WRSC ja elastilise tehisnärvivõrgu mudeli ennustatud lootefraktsiooni keskmise leidmisest.

Kirjeldatud programmiga analüüsiti kõiki 368 madala katvusega ülegenoomselt sekveneeritud Eesti raseda proovi ja leiti järgi korrelatsioon, mis võrdleb käesolevas töös välja töötatud ja olemasolevat SeqFF meetodil leitud lootefraktsioone. Saadud võrdlus on toodud joonisel 10, kus on märgata head vastavust (r=0.9) saadud filtreeritud piirkondade ja originaalse SeqFF vahel.

Suhteliselt kõrge korrelatsioon võrreldes originaalse SeqFF meetodiga saaduga näitab, et selekteeritud väiksemate regioonisuuruste ja seetõttu ka vähemate lugemitega regiooni kohta on esialgsete tulemuste põhjal võimalik ennustada lootefraktsiooni piisava täpsusega.



Joonis 10. Korrelatsioonigraafik töös välja töötatud meetodi nimetusega TaqSeqFF ja SeqFF leitud lootefraktsioonide vahel.

2.4 Arutelu

Tuginedes käesolevas töös leitud regioonidele ja nende põhjal arvutatud lootefraktsiooni hinnangutele (Joonis 10), võib väita, et olemasolevat SeqFF meetodit saab kasutada lootefraktsiooni hindamiseks ka 50× lühemate genoomsete regioonide ja vaid spetsiaalselt väljavalitud informatiivsete regioonide pealt. Piirkondade selektsioon viib seega efektiivse lootefraktsiooni arvutamiseks vajaliku summaarse genoomse piirkonna suuruse 2,5 Gbp-lt alla 12 Mbp-ni. Küll aga on saadud tulemus sõltuv proovist, ehk tulemuse täpsuse tagab piiratud paindlikkus analüüsitavate kandidaatpiirkondade valimisel suuremas 50kbp regiooniaknas. See tähendab, et kuigi kasutatud filtreeritud piirkonnad on jätkuvalt tähenduslikud lootefraktsiooni arvutamiseks, on nende seast veel edasi vaja valida alamhulk fikseeritud genoomseid regioone, mida töö üldisema eesmärgi saavutamiseks kasutada.

Leitud genoomsed piirkonnad oleks seejärel kasutatavad edasises analüüsis, kus disainitakse uus NIPT testi meetod põhinedes TAC-seq meetodil, mis hõlmaks huvipakkuvatele piirkondadele detektor-proovide disaini ning saadud lugemite analüütilise poole loomist. Uus meetod alandaks NIPT testi hinda, mis on töö kirjutamise hetkel 250-400 eurot (olenevalt pakkujast ja täpsest meetodist), sest pole vajalik enam ülegenoomne sekveneerimine. Arendatava testi juures oleks oluline, et saadud test oleks sarnaselt SeqFFi-le soost mittesõltuv ehk toimib ka tüdrukloote puhul.

Spetsiifiliste biomarkerite või genoomsete DNA piirkondade leidmiseks tuleks kasutada kas kõrgema katvusega sekveneerimist või sarnaselt SeqFF meetodile treenida elastset tehisnärvivõrku cfDNA sekveneerimisandmetega, kasutades 50 kbp pikkuste regioonide asemel lugemite 1000 bp pikkustesse piirkondadesse jaotamist. Sel viisil oleks võimalik leida regressioonikordajad iga piirkonna kohta, ning teha nende hulgast valik, kuhu disainida detektor-proovid TAC-seq meetodi rakendamiseks. Samuti võib olla väiksemate piirkondade peal treenitud mudel täpsem kui SeqFF. Tehisnärvivõrgu treenimiseks on aga vaja suurt valimit, mistõttu on täpsete biomarkerite määramiseks vajalik edasine uuring suurema andmemahu pealt.

Kokkuvõte

Käesoleva töö eesmärgiks oli leida, kas madala katvusega cfDNA sekveneerimisandmete põhjal on võimalik leida oluliselt lühemaid lootefraktsiooni hindamiseks täpseks sobivaid genoomseid piirkondi. Piirkondade leidmiseks kasutatakse 368 Eestist kogutud ja madala katvusega sekveneeritud vereseerumist eraldatud cfDNA sekveneerimisproovi. Sekveneerimisandmete töötluseks kasutati vabavaralisi geneetikas laialt kasutatavaid programme *Bowtie2* ja *Samtools*, mis on vastavalt katsedisainile optimaalselt seadistatud. Saadud cfDNA lugemitest leiti minimaalse pikkusega piirkonnad, mis on üle valimi märkimisväärselt lugemite arvukusega esindatud.

Selle töö tulemusena leiti lootefraktsiooni seisukohast olulised minimaalse suurusega piirkonnad, kust on potentsiaal leida sobivaid genoomseid regioone, mida kasutada TAC-seq põhise NIPTi väljatöötamisel loote DNA fraktsiooni hindamiseks. Selleks kasutati 368 Eestist kogutud raseda naise madala katvusega sekveneeritud cfDNA andmeid. Uurimise käigus selgitati välja minimaalne informatiivne genoomse piirkonna suurus 1000 bp olemasolevate andmete põhjal ja uuriti saadud jaotust SeqFF mudeli parameetreid kasutades. Seejärel leiti saadud pikkusega lootefraktsiooni arvutamise seisukohast oluliste biomarkeri piirkondade järjend ja kirjutati programm, mis arvutab leitud piirkondi kasutades lootefraktsiooni.

Töö järeldusena saab olemasolevat SeqFF meetodit täiendada, kui kasutada ligi 50× suuremat ülegenoomset lahutust, säilitades originaalsele SeqFF mudelile sarnase lugemite jaotuse ja lootefraktsiooni arvutamise täpsuse. Siiski on saadud biomarkeri piirkondade puhul täpsemalt määratletud FF arvutamiseks olulised järjestused ning edasise analüüsiga võib neid piirkondi veelgi kitsendada.

Resümee

Finding suitable regions in the human genome for determination of fetal fraction from maternal cell-free DNA

Alvin Meltsov

Summary

Non-invasive prenatal testing (NIPT) has been established as reliable and safe method for chromosomal abnormality screening in high-risk pregnancies. Compared to clinical tests chorionic villus sampling (CVS) or amniocentesis, NIPT does not affect the child directly, but analyses the blood of the mother. This is done either by various PCR methods targeting population-specific regions or microfluidics tests using an array. These methods are therefore limited by either population, where no suitable biomarkers are found. Besides detection of chromosomal abnormalities, the relation of cell-free fetal DNA (cffDNA) to cell-free DNA (cfDNA) or fetal fraction (FF) is used as qualitative measure of the reliability of NIPT test outcome.

Methods based on next-generation sequencing techniques have been developed which resolved the limitations of previous methods of fetal fraction calculation. These include Bayindr's, SANEFALCON, DEFRAG, and SeqFF. While many are still mostly reliable on Y-chromosome, SeqFF offers a new approach using machine learning to train a multivariate regression model, which can be used to predict fetal fraction of cell-free DNA in mothers blood using low coverage sequencing. Therefore, with new methods, NIPT tests perform with high sensitivity and specificity using only sequencing reads and has no gender-specificity. Still, the price for sequencing is relatively high, and new methods are developed to addressing this issue using targeted sequencing technologies.

This thesis focuses on improving existing NIPT method TAC-seq, which is a targeted sequencing method used for detecting chromosomal anomalies. This means that whole genome sequencing is no longer required. Nevertheless, this method has no qualitative measure because TAC-seq does not have the capability to calculate precise fetal fraction. Using SeqFF multivariate model, this thesis explores if there are significant regions in cfDNA genome that can be reliably used for quantification and FF calculation using targeted sequencing techniques.

Blood samples from 368 pregnant woman in Estonia were collected and sequenced with low coverage, mapped to GRCh37/hg19 reference genome using *bowtie2* and quantified using *Samtools* to different lengths of sequentially repeating genomic regions or bins. The read count distributions of bins with different sizes were analysed and bin length of 1000bp per bin was selected. Using this binning method, most significant regions in relevance to FF were filtered out using the data available and SeqFF multivariate model regression coefficients. Resulting list of filtered regions was used to calculate FF and results were compared to the original SeqFF predicted fetal fractions, which resulted in a correlation coefficient of 0.9.

In conclusion, SeqFF method can be used to detect fetal fraction using 50 times smaller bin sizes, but further research is needed to find suitable genomic regions for targeting in TAC-seq.

Tänusõnad

Soovin tänada oma juhendajaid Priit Paltat ja Kaarel Krjutškovi suurepärase ja otsekohese juhendamise eest. Samuti soovin tänada Tervisetehnoloogiate Arenduskeskuse tiimi, eriti Hindrek Tederit ja Priit Paluojat, kes olid suureks abiks R-keele, andmete visualiseerimise ja serveris töötamise õppimisel.

Kasutatud kirjanduse loetelu

Anderson, M.W., Schrijver, I., 2010. Next Generation DNA Sequencing and the Future of Genomic Medicine. Genes 1, 38–69. https://doi.org/10.3390/genes1010038

Ashoor, G., Syngelaki, A., Poon, L.C.Y., Rezende, J.C., Nicolaides, K.H., 2013. Fetal fraction in maternal plasma cell-free DNA at 11-13 weeks' gestation: relation to maternal and fetal characteristics. Ultrasound Obstet Gynecol 41, 26–32. https://doi.org/10.1002/uog.12331

Bayindir, B., Dehaspe, L., Brison, N., Brady, P., Ardui, S., Kammoun, M., Van der Veken, L., Lichtenbelt, K., Van den Bogaert, K., Van Houdt, J., Peeters, H., Van Esch, H., de Ravel, T., Legius, E., Devriendt, K., Vermeesch, J.R., 2015. Noninvasive prenatal testing using a novel analysis pipeline to screen for all autosomal fetal aneuploidies improves pregnancy management. Eur. J. Hum. Genet. 23, 1286–1293. https://doi.org/10.1038/ejhg.2014.282

Chan, K.C.A., Zhang, J., Hui, A.B.Y., Wong, N., Lau, T.K., Leung, T.N., Lo, K.-W., Huang, D.W.S., Lo, Y.M.D., 2004. Size Distributions of Maternal and Fetal DNA in Maternal Plasma. Clinical Chemistry 50, 88–92. https://doi.org/10.1373/clinchem.2003.024893

Chinnapapagari, S.K.R., Holzgreve, W., Lapaire, O., Zimmermann, B., Hahn, S., 2005. Treatment of Maternal Blood Samples with Formaldehyde Does Not Alter the Proportion of Circulatory Fetal Nucleic Acids (DNA and mRNA) in Maternal Plasma. Clinical Chemistry 51, 652–655. https://doi.org/10.1373/clinchem.2004.042119

Dan, S., Wang, Wei, Ren, J., Li, … , Su, Y., Zhang, Xiuqing, 2012. Clinical application of massively parallel sequencing-based prenatal noninvasive fetal trisomy test for trisomies 21 and 18 in 11 105 pregnancies with mixed risk factors. Prenatal Diagnosis 32, 1225–1232. https://doi.org/10.1002/pd.4002

Dhallan, R., Au, W.-C., Mattagajasingh, S., Emche, S., Bayliss, P., Damewood, M., Cronin, M., Chou, V., Mohr, M., 2004. Methods to increase the percentage of free fetal DNA recovered from the maternal circulation. JAMA 291, 1114–1119. https://doi.org/10.1001/jama.291.9.1114

Hatem, A., Bozdağ, D., Toland, A.E., Çatalyürek, Ü.V., 2013. Benchmarking short sequence mapping tools. BMC Bioinformatics 14, 184. https://doi.org/10.1186/1471-2105-14-184

Kim, S.K., Hannum, G., Geis, J., Tynan, J., Hogg, G., Zhao, C., Jensen, T.J., Mazloom, A.R., Oeth, P., Ehrich, M., Boom, D. van den, Deciu, C., 2015. Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts. Prenatal Diagnosis 35, 810–815. https://doi.org/10.1002/pd.4615

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Y., Zimmermann, B., Rusterholz, C., Kang, A., Holzgreve, W., Hahn, S., 2004. Size separation of circulatory DNA in maternal plasma permits ready detection of fetal DNA polymorphisms. Clin. Chem. 50, 1002–1011. https://doi.org/10.1373/clinchem.2003.029835

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., Richmond, T.J., 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature 389, 251. https://doi.org/10.1038/38444

Lun, F.M.F., Chiu, R.W.K., Chan, K.C.A., Leung, T.Y., Lau, T.K., Lo, Y.M.D., 2008. Microfluidics Digital PCR Reveals a Higher than Expected Fraction of Fetal DNA in Maternal Plasma. Clinical Chemistry 54, 1664–1672. https://doi.org/10.1373/clinchem.2008.111385

Pergament, E., Cuckle, H., Zimmermann, B., … , Rabinowitz, M., 2014. Single-Nucleotide Polymorphism–Based Noninvasive Prenatal Screening in a High-Risk and Low-Risk Cohort. Obstet Gynecol 124, 210–218. https://doi.org/10.1097/AOG.0000000000000363

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. PNAS 74, 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135–1145. https://doi.org/10.1038/nbt1486

Slatko, B.E., Gardner, A.F., Ausubel, F.M., 2018. Overview of Next Generation Sequencing Technologies. Curr Protoc Mol Biol 122, e59. https://doi.org/10.1002/cpmb.59

Straver, R., Oudejans, C.B.M., Sistermans, E.A., Reinders, M.J.T., 2016. Calculating the fetal fraction for noninvasive prenatal testing based on genome-wide nucleosome profiles. Prenat. Diagn. 36, 614–621. https://doi.org/10.1002/pd.4816

Teder, H., Koel, M., Paluoja, P., Jatsenko, T., Rekker, K., Laisk-Podar, T., Kukuškina, V., Velthut-Meikas, A., Fjodorova, O., Peters, M., Kere, J., Salumets, A., Palta, P., Krjutškov, K., 2018. TAC-seq: targeted DNA and RNA sequencing for precise biomarker molecule counting. npj Genomic Medicine 3, 34. https://doi.org/10.1038/s41525-018-0072-5

Theisen, A., Shaffer, L.G., 2010. Disorders caused by chromosome abnormalities. Appl Clin Genet 3, 159–174. https://doi.org/10.2147/TACG.S8884

Traeger-Synodinos, J., 2006. Real-time PCR for prenatal and preimplantation genetic diagnosis of monogenic diseases. Mol. Aspects Med. 27, 176–191. https://doi.org/10.1016/j.mam.2005.12.004

van Beek, D.M., Straver, R., Weiss, M.M., Boon, E.M.J., Huijsdens-van Amsterdam, K., Oudejans, C.B.M., Reinders, M.J.T., Sistermans, E.A., 2017. Comparing methods for fetal fraction determination and quality control of NIPT samples. Prenat Diagn 37, 769–773. https://doi.org/10.1002/pd.5079

Zou, H., Hastie, T., 2005. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society, Series B 67, 301–320.

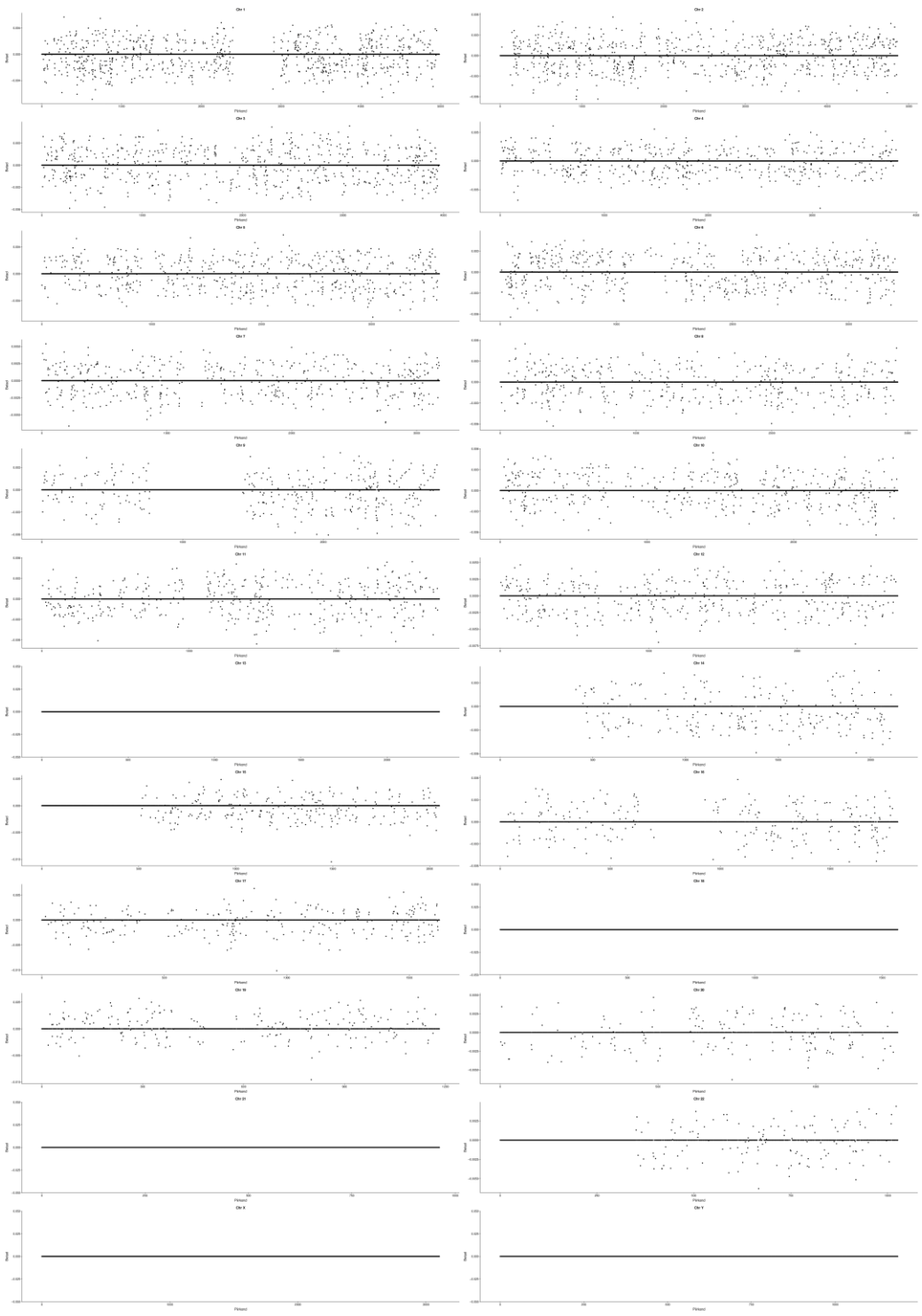
Kasutatud veebiaadressid

„Mitteinvasiivne sünnieelne sõeluuring loote trisoomiate 21, 18 ja 13 suhtes, kasutades loote rakuvaba DNA analüüsimist ema verest“ taotlusvorm haigekassale, 2016. Külastamise kuupäev: 22.05.2019. Lehekülg: https://www.haigekassa.ee/sites/default/files/TTL/2019/1195_taotlus.pdf

Lisad

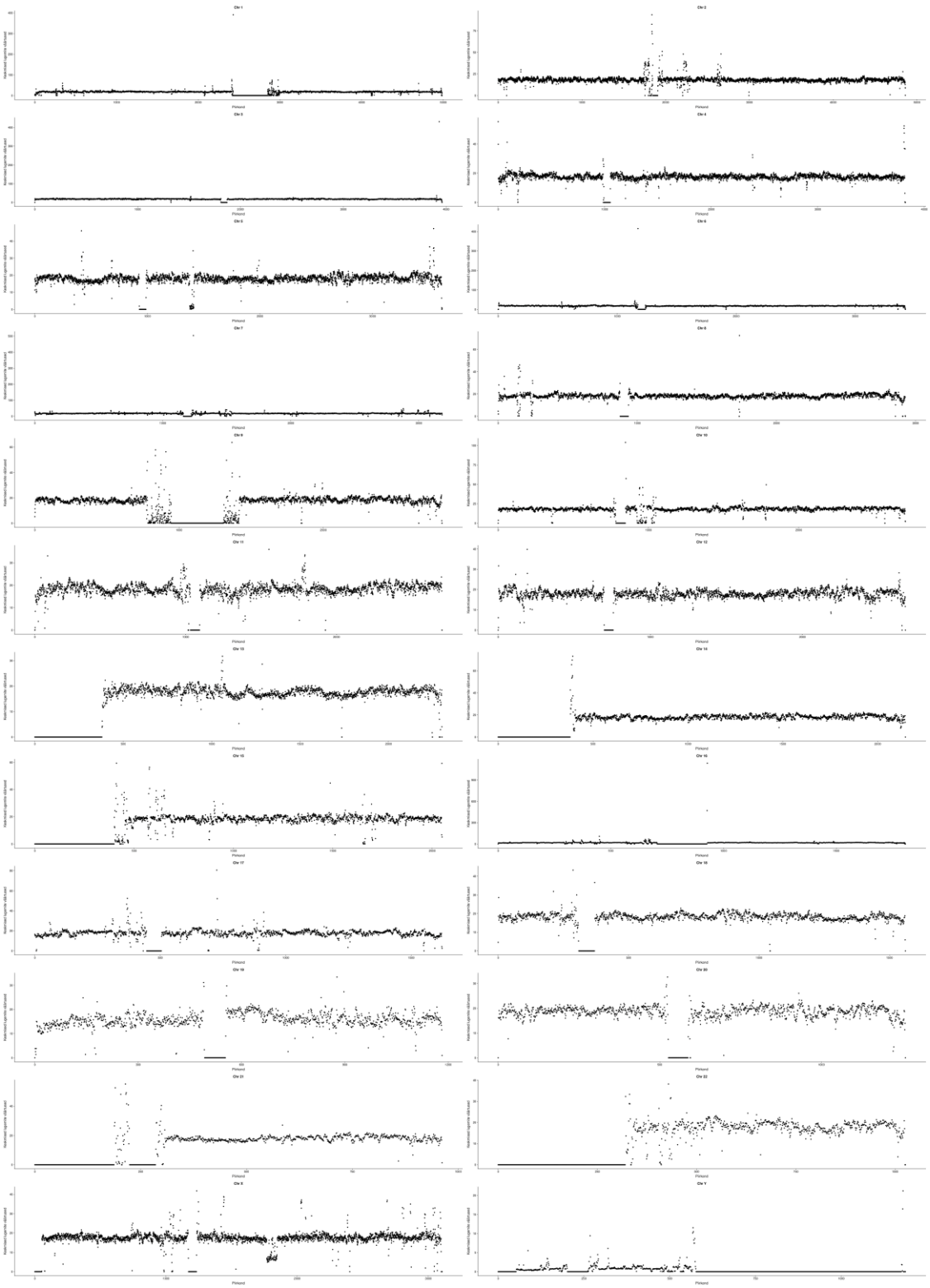
Lisa 1

Regressioonikordajate jaotus regiooni järgi üle genoomi



Lisa 2

Katsevalimi summeeritud lugemite jaotus regiooni järgi üle genoomi



Lihtlitsents

Lihtlitsents lõputööreprodutseerimiseks ja lõputööüldsusele kättesaadavaks tegemiseks

Mina, Alvin Meltsov (sünnikuupäev: 14.08.1995)

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

„Sobivate inimese genoomi regioonide leidmine loote rakuvaba DNA osakaalu määramiseks ema vereproovist“,

mille juhendajad on Priit Palta ja Kaarel Krjutškov,

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.

Alvin Meltsov

27.05.2019