TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOINFORMAATIKA ÕPPETOOL

Kaarel Koitne

Stabiilsete *k*-meeride hulga valimine mitteinvasiivseks sünnieelseks testimiseks

Magistritöö (30 EAP)

Geenitehnoloogia

Juhendaja PhD Lauris Kaplinski

TARTU 2016

INFOLEHT

Stabiilsete *k*-meeride hulga valimine mitteinvasiivseks sünnieelseks testimiseks

Traditsiooniliste meetoditega loote pärilike haiguste sünnieelsel tuvastamisel on mitmeid puudusi ning seetõttu otsitakse pidevalt uusi ja paremaid lahendusi. Üheks perspektiivikamaks suunaks peetakse rakuvaba DNA analüüsimist, mis võiks olla ohutu ning täpne asendus praegu kasutatavatele meetoditele.

Magistritöö kirjanduse ülevaates kirjeldatakse rakuvaba DNA avastamist, selle põhjal tehtavaid analüüse ning põhilisi kitsaskohti. Lisaks võrreldakse lugemite referentsgenoomile paigutamisel ja *k*-meeride sageduste lugemisel põhinevat genoomianalüüsi. Töö eksperimentaalse osa eesmärgiks oli leida stabiilsete *k*-meeride hulk mitteinvasiivse sünnieelse testimise jaoks.

Märksõnad: loote rakuvaba DNA, teise põlvkonna sekveneerimine, mitteinvasiivne sünnieelne testimine, *k*-meerid.

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Choosing a set of stable *k*-mers for Non-Invasive Prenatal Testing

Traditional methods for prenatal diagnosis have many shortcomings and therefore new and better solutions are being developed. Analysis of cell-free DNA is a rapidly evolving field and it is thought to be one of the most promising replacement for the methods which are currently being used.

Literature review of this thesis describes the discovery of cell-free DNA, current applications and main challenges for analysing cell-free DNA. Also, read mapping and methods based on *k*-mer counting for analysing genomic data are compared. The purpose of the experimental part was to choose a set of stable *k*-mers for Non-Invasive Prenatal Testing.

Keywords: cell-free fetal DNA, Next-Generation Sequencing, Non-Invasive Prenatal Testing, *k*-mers.

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

SISUKORD

KASUTATUD LÜHENDID

BAM – binaarne joondus/paigutus (*Binary Alignment/Map*)

cffDNA – loote rakuvaba DNA (*cell-free fetal DNA*)

CNV – koopiaarvu variatsioon (*Copy-Number Variation*)

CPM – platsenta mosaiiksus (*Confined Placental Mosaicism*)

dbSNP – ühenukleotiidsete polümorfismide andmebaas (*Single Nucleotide Polymorphism Database*)

FASTA – failiformaat nukleotiid- või peptiidjärjestuste salvestamiseks

FASTQ – failiformaat nukleotiidjärjestuste salvestamiseks, milles on informatsioon ka kvaliteediskoori kohta

ff – loote fraktsioon (*fetal fraction*)

GB – gigabait ehk 10^9 baiti (gigabyte)

Indel – insertsioon või deletsioon (*insertion or deletion*)

IUPAC – Rahvusvaheline Puhta Keemia ja Rakenduskeemia Liit (*International Union of Pure and Applied Chemistry*)

MAF – minoorse alleeli sagedus (*Minor Allele Frequency*)

MB – megabait ehk 10^6 baiti (megabyte)

NGS – teise põlvkonna sekveneerimine (*Next-Generation Sequencing*)

NIPT – mitteinvasiivne sünnieelne testimine (*Non-Invasive Prenatal Testing*)

SNP – ühenukleotiidne polümorfism (*Single Nucleotide Polymorphism*)

VCF – failiformaat geenijärjestuste varieeruvusinfo salvestamiseks (*Variant Call Format*)

SISSEJUHATUS

Loote pärilike haiguste sünnieelne tuvastamine on oluline nii lapse kui ka tema vanemate elukvaliteedi tagamiseks.

Käesoleval ajal on kasutusel mitmed erinevad meetodid pärilike haiguste, eelkõige 21. kromosoomi trisoomia ehk Downi sündroomi, sünnieelseks tuvastamiseks. Mitteinvasiivsed meetodid hõlmavad endas ultraheliuuringut, millega mõõdetakse loote kuklavoldi paksust ja kontrollitakse siseorganite normaalset arengut, ja vereanalüüsi, millega mõõdetakse platsentaarse päritoluga hormoonide taset. Nende meetodite puuduseks on suur valepositiivsete ja -negatiivsete diagnooside osakaal. Invasiivseteks meetoditeks on koorioni hattude biopsia ja amniotsentees ehk lootevee uuring. Need on küll täpsed, ent miinuseks on suhteliselt väike, kuid siiski reaalne risk nii ema tervisele kui ka raseduse katkemisele.

Viimasel kümnendil on traditsiooniliste meetodite rakendamise kõrval võetud kliinilisse kasutusse ka mitteinvasiivne sünnieelne testimine, mis põhineb loote rakuvaba DNA analüüsil. Rakuvaba DNA testimise eelisteks võrreldes varasemate meetoditega on täpsus ja ohutus. Lisaks on sellega võimalik analüüse sooritada raseduse varasemates perioodides kui traditsioonilisi meetodeid kasutades.

Magistritöö eksperimentaalse osa eesmärgiks oli leida stabiilsete *k*-meeride hulk, mille abil oleks tulevikus võimalik analüüsida mitteinvasiivse sünnieelse testimise jaoks kogutud sekveneerimisandmeid ning seeläbi tuvastada kui suur on risk Downi sündroomiga lapse sünniks.

1 KIRJANDUSE ÜLEVAADE

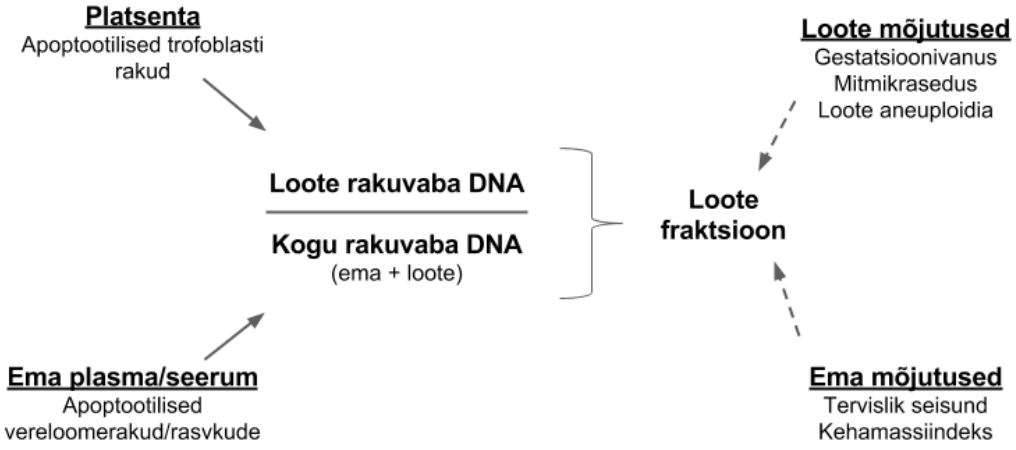
1.1 Loote rakuvaba DNA

Ema vereseerumis ja -plasmas ringlev loote rakuvaba DNA (cffDNA – ingl k *cell-free fetal DNA*) avastati 1997. aastal. Lo *et al*. (1997) analüüsisid 43 rasedalt ja 10 mitterasedalt naiselt võetud plasma ja seerumi proove ning leidsid, et neist võib leida lootelt pärinevaid Y-kromosoomile iseloomulikke järjestusi. 30-st XY-sugukromosoomistikuga lootega raseda naise plasmaproovist detekteeriti Y-kromosoomile iseloomulikke järjestusi 24 juhul ning seerumiproovidest 21 juhul. 13-st XX-sugukromosoomistikuga lootega raseda ning 10 mitteraseda naise vereproovist Y-kromosoomile iseloomulikke järjestusi ei leitud (Lo *et al*., 1997).

Alberry *et al*. (2007) tegid kindlaks, et suurem osa cffDNA-st pärineb platsentast, täpsemalt apoptootilistest trofoblasti rakkudest (Alberry *et al*., 2007). Ema enda rakuvaba DNA pärineb seevastu enamasti vereloomerakkudest ja/või rasvkoest (Lui *et al*., 2002; Haghiac *et al*., 2012). Detekteeritava taseme saavutab cffDNA ema veres juba viiendal rasedusnädalal (Rijnders *et al*., 2003).

Nüüdseks on teada, et keskmiselt 10–20% rakuvabast DNA-st ema vereplasmas pärineb platsentast (Lun *et al*., 2008). See protsent, mida nimetatakse loote fraktsiooniks (ff – ingl k *fetal fraction*), varieerub indiviiditi ning üldjuhul suureneb raseduse vältel (Lo *et al*., 1998; Lun *et al*., 2008) kuni sünnituseni, pärast mida loote DNA kaob ema vereplasmast mõne tunniga (Lo *et al*., 1999). Preeklampsia, platsenta eesasetuse ja loote kromosomaalsete häirete korral on ff kõrgem kui normaalse raseduse korral (Masuzaki *et al*., 2004). Ff-i võib aga alandada ema suurenenud kehamassiindeks, kuna ülekaalulisus põhjustab suuremat DNA eraldumist rasvkoest ja seega ka ema rakuvaba DNA taseme tõusu (Wang *et al*., 2013) (vt Joonis 1).

Erinevalt intaktsest genoomsest DNA-st, esineb rakuvaba DNA plasmas fragmentidena. CffDNA fragmendid on üldiselt lühemad kui ema rakuvaba DNA fragmendid (Chan *et al*., 2004). Ema rakuvaba DNA fragmentide suuruseks on keskmiselt 166 aluspaari, cffDNA fragmentidel aga 143 aluspaari. Nii cffDNA kui ka ema rakuvaba DNA fragmendid jaotuvad enam-vähem võrdselt üle kogu genoomi (Lo *et al*., 2010).



Joonis 1. Loote fraktsioon ehk ff on loote rakuvaba DNA ja kogu rakuvaba DNA suhe. Ff-i mõjutavad raseduse vältel mitmed lootelt ja emalt pärinevad faktorid (Taglauer *et al*., 2014; modifitseeritud).

Ema veres ringleva cffDNA avastamine pani aluse DNA-põhisele mitteinvasiivsele sünnieelsele testimisele (NIPT – ingl k *Non-Invasive Prenatal Testing*), kuid siiani pole täpselt teada, mis ja kuidas mõjutab cffDNA produktsiooni, metabolismi ja eemaldumist vereringest (Taglauer *et al*., 2014).

Teise põlvkonna sekveneerimistehnoloogiad (NGS – ingl k *Next-Generation Sequencing*) on rakuvaba DNA analüüsimist märkimisväärselt hõlbustanud. NGS võimaldab miljonite kuni miljardite plasmas leiduvate nukleiinhapete molekulide suure läbilaskevõimega ning tootlikkusega sekveneerimist (Lo *et al*., 2013; Chiu *et al*., 2013).

1.2 Teise põlvkonna sekveneerimine

NGS-i kasutuselevõtt on muutnud täisgenoomide massilise sekveneerimise üldlevinud praktikaks, kuna see on kiirem ja odavam kui Sangeri sekveneerimine. Miinuseks on ebatäpsus ja oluliselt väiksem lugemite pikkus, kuid seda on võimalik osaliselt kompenseerida, suurendades sekveneerimiskatvust.

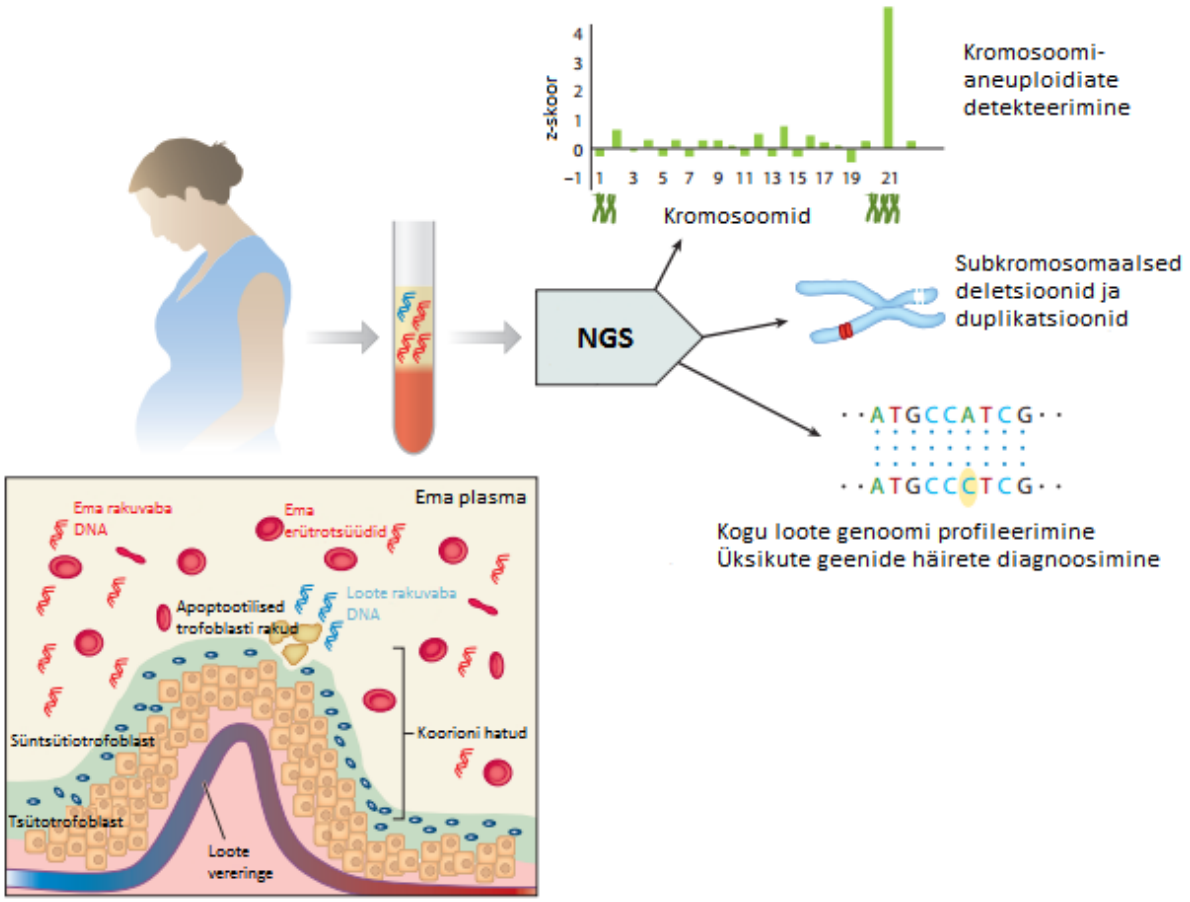
Sekveneerimiskatvus on NGS-i puhul oluline parameeter. See näitab, mitu lugemit katab keskmiselt ühte aluspaari genoomses järjestuses. Olenevalt diagnostilisest protseduurist, on vaja suurema või väiksema katvusega sekveneerimist. Suurema katvusega sekveneerimine võimaldab avastada haruldasemaid ühenukleotiidseid polümorfisme (SNP – ingl k *Single Nucleotide Polymorphism*) ning eristada neid sekveneerimisvigadest (Muzzey *et al*., 2015).

Kromosoomianeuploidiate korral piisab ka väiksema katvusega sekveneerimisest, mis on odavam kui suurema katvusega sekveneerimine (Sims *et al*., 2014). Kuna rakuvaba DNA on veres väikeste fragmentidena, siis on lühikeste lugemitega NGS sobiv tehnoloogia selle sekveneerimiseks (Lo *et al*., 2010).

1.3 Mitteinvasiivne sünnieelne testimine

NIPT tähendab loote kromosomaalse või geneetilise seisundi kontrollimist ema vereplasmas või -seerumis ringleva cffDNA kaudu (Wong ja Lo, 2016). Traditsiooniliste invasiivsete meetoditega, amniotsenteesi või koorioni hattude biopsiaga, on loote kromosomaalse või geneetilise konditsiooni diagnoosimine vajanud koeproove lootest või platsentast. Nende meetodite kasutamisel on aga risk nii ema tervisele kui ka raseduse katkemisele (Mujezinovic ja Alfirevic, 2007). Traditsiooniliste mitteinvasiivsete meetodite puuduseks on aga ebatäpsus ja sellega kasnev suur valenegatiivsete ja -positiivsete tulemuste arv (Norwitz ja Levy, 2013).

NIPT-i näol on tegemist kiiresti areneva valdkonnaga, mille praegused kliinilised rakendused hõlmavad loote soo ja reesusgrupi määramist, üksikute geenide häirete tuvastamist, täpset aneuploidsuste skriinimist ning subkromosomaalsete kõrvalekallete tuvastamist. Tulevikuperspektiiviks on monogeensete haiguste diagnoos ja loote kogu genoomi profileerimine (vt Joonis 2). NIPT-i kasutuselevõtt kliinilises praktikas pakub suuri eeliseid, kuid nõuab mitmete tehniliste ja ka eetiliste väljakutsete ületamist (Daley *et al*., 2014). Käesoleval ajal peetakse NIPT-i kõrgtehnoloogiliseks kontrolltestiks, mis võib harvadel juhtudel siiski valesid tulemusi anda. Seega on vajalik positiivsed tulemused invasiivsete meetoditega üle kontrollida. Otsust raseduse katkestamise kohta ei tohiks teha üksnes positiivse NIPT-i tulemuse põhjal (ACOG, 2012).



Joonis 2. Lihtsustatud skeem erinevatest käesoleval ajal rakuvaba DNA analüüsi baasil tehtavatest mitteinvasiivse sünnieelse testimise rakendustest (Wong ja Lo, 2016; modifitseeritud).

Järgnevalt kirjeldatakse peamiseid käesoleval ajal rakuvaba DNA analüüsi baasil tehtavaid teste ning nende põhilisi kitsaskohti.

1.3.1 Kromosoomianeuploidiad

Levinumaks kromosomaalseks aneuploidiaks on 21. kromosoomi trisooma ehk Downi sündroom, mille esinemissagedus on 1,1 juhtu 1000 elussünni kohta (Loane *et al*., 2013). Downi sündroom põhjustab vaimseid arenguhäireid, kaasasündinud südamerikkeid, lisaks suurendab kasvajate ja nakkushaiguste tekkimise tõenäosust, soole avanematust ning teisi haigusi (Korenberg *et al*., 1994). Teisteks levinumateks aneuploidiateks on 18. kromosoomi trisoomia ehk Edwardsi sündroom (esinemissagedusega 0,1 juhtu 1000 elussünni kohta) ja 13. kromosoomi trisoomia ehk Patau sündroom (esinemissagedusega 0,05 juhtu 1000 elussünni kohta) (Loane *et al*., 2013).

Kaasajal on arenenud riikides Downi sündroomi ja teiste aneuploidiate diagnoosimine

muutunud aina olulisemaks probleemiks, seoses sünnitajate keskmise vanuse tõusuga. Nimelt on näidatud, et ema vanuse kasvades tõuseb oluliselt aneuploidiate tõenäosus (Chiang *et al*., 2011).

Traditsiooniliste kontrollimismeetodite tundlikkus on Downi sündroomi tuvastamisel kuni 90% ja valepositiivsete diagnooside määr 5% (Driscoll ja Gross, 2009). Need meetodid detekteerivad loote patoloogilist fenotüüpi, kuid mitte kromosomaalset trisoomiat kui patoloogia põhjust. Seevastu NIPT analüüs on disainitud tuvastama ebanormaalset kromosoomi doosi põhimõttel, et suhteliselt koguselt on 21. kromosoomi trisoomia korral sellest kromosoomist pärinevaid DNA järjestusi ema veres rohkem kui euploidse loote korral. Järjestuste arvu erinevus on siiski küllaltki väike, sest cffDNA "lahjeneb" ema rakuvaba DNA taustal (Wong ja Lo, 2016).

1.3.1.1 Valepositiivsed ja -negatiivsed tulemused

Põhjuseks, miks NIPT-iga ja sellele järgneva invasiivse meetodiga (nt amniotsenteesiga) määratud karüotüüp omavahel ei ühti, võib olla mitu seletust. Tegemist võib olla eksimusega vereproovide käsitlemisel, sekveneerimistehnoloogiast tuleneva või bioinformaatilisel töötlusel tehtud veaga, kuid tihti võib põhjus olla hoopis bioloogilist päritolu (Brady *et al*., 2015).

Üheks valepositiivsete ja ka valenegatiivsete diagnooside bioloogiliseks põhjuseks on platsenta mosaiiksus (CPM – ingl k *Confined Placental Mosaicism*) (Taglauer *et al*., 2014). CPM-i korral on tegemist aneuploidse platsenta, kuid euploidse lootega või vastupidi. Kuna cffDNA pärineb platsenta trofoblasti rakkudest, annab selline ebakõla NIPT-iga vale tulemuse. CPM-i on tuvastatud umbes 1% rasedatel (Gardner ja Sutherland, 2004).

Teiseks bioloogiliseks põhjuseks võib olla mitmikrasedus. On näidatud, et NIPT-iga on võimalik detekteerida aneuploidiaid ka kaksikraseduse korral (Huang *et al*., 2014). CffDNA osakaal on suurem kaksikraseduse korral (Attilakos *et al*., 2011), kuid analüüsi täpsus ei pruugi olla nii hea kui üksikraseduse puhul (Bevilacqua *et al*., 2015). Üheks ebatäpse tulemuse põhjuseks võib olla nähtus nimega "haihtuv kaksik" (Grömminger *et al*., 2014), mis tähendab, et kaksikraseduse ajal üks loode emakas sureb ja teine absorbeerib selle osaliselt või täielikult (Landy *et al*., 1986).

Kuna enamik rakuvabast DNA-st pärineb emalt, võivad ema genoomi koopiaarvu

variatsioonid (CNVs – ingl k *Copy-Number Variations*) raskendada NIPT-i tulemuste analüüsimist ja tõlgendamist. Ema CNV-d võivad tekitada nii valepositiivseid kui ka -negatiivseid tulemusi (Brady *et al*., 2015).

Ka ema somaatiline X-kromosoomi aneuploidia võib olla valepositiivsete NIPT-i tulemuste põhjuseks. Tsütogeneetilised uuringud on näidanud, et vananedes kasvab naistel X-kromosoomi vererakkudest kadumise kiirus, mis võib anda tulemuseks valepositiivse X-kromosoomi monosoomia (Russell *et al*., 2007). Wang *et al*. on näidanud, et 8,6% positiivsetest NIPT-i tulemustest sugukromosoomide aneuploidiate kohta on põhjustatud ema geneetilisest mosaiiksusest (Wang *et al*., 2014).

Emalt pärinevate kasvajarakkude apoptoos võib olla põhjuseks ebanormaalse kromosoomikomplekti levimisel rakuvaba DNA-na. On avastatud mitmeid juhtumeid, kus NIP-testi väära tulemuse põhjuseks olid pahaloomulised kasvajad (Osborne *et al*., 2013).

1.3.2 Subkromosomaalsed kõrvalekalded

Enamik NIPT metoodikatest on keskendunud kromosoomianeuploidiate (täpsemalt kromosoomide 13, 18, 21, X ja Y aneuploidiate) detekteerimisele, ent need moodustavad ainult 30% elussündide kromosoomianomaaliatest. Mitmed struktuursed kromosoomimuutused, näiteks mikrodeletsioonid ja -duplikatsioonid, on palju levinumad, kuid neid on raskem tuvastada (Zhao *et al*., 2015).

Subkromosomaalsed kõrvalekalded on levinud põhjuseks kaasasündinud füüsilisele ja vaimsele alaarengule. Jaotades kromosoomi väiksemateks, näiteks 1 Mb pikkusteks regioonideks, saab mikrodeletsioonide ja -duplikatsioonide detekteerimisel kasutada samasugust põhimõtet nagu aneuploidiate korral (Wong ja Lo, 2016). Jensen *et al*. (2012) tuvastasid 3 Mb suuruse deletsiooni 22. kromosoomi pikas õlas, mis põhjustab DiGeorge-i sündroomi. Tulemused näitasid uuritavates proovides 22.q11.2 regioonis statistiliselt olulist lugemite arvu vähenemist (z-skoor < -3). Saavutamaks statistiliselt olulist erinevust, oli aga vaja suhteliselt suure katvusega sekveneerimist, umbes 200 miljonit lugemit proovi kohta (Jensen *et al*., 2012).

Seega põhimõtteliselt on NIPT-i abil võimalik tuvastada ka suuri duplikatsioone ning deletsioone, kuid see nõuab võrreldes kromosoomianeuploidiate detekteerimisega suurema katvusega sekveneerimist (Wong ja Lo, 2016).

1.3.3 Monogeensed haigused ja kogu genoomi profileerimine

Vaatamata märkimisväärselt suurele vajadusele meditsiinis, on NIPT-i kasutuselevõtt monogeensete haiguste diagnoosiks olnud aeglane (Daley *et al*., 2014). Hetkel on olemas teste, mis tuvastavad monogeensetest haigustest näiteks akondroplaasiat, talasseemiat ja tsüstilist fibroosi (Lench *et al*., 2013).

Praegu kliinilises praktikas kasutusel olevad testid võimaldavad tuvastada monogeenseid tunnuseid ja häireid, mille alleelid emal puuduvad. Nende detekteerimine on sarnane loote soo määramisele ja reesusfaktori genotüpiseerimisele. Mõlemal juhul on otsitavad omadused lootespetsiifilised (Wong ja Lo, 2016). On kaks praktilist piirangut, mis takistavad laialdasemat monogeensete haiguste tuvastamist. Esiteks, oleks kasulik, kui isa genotüüp on teada ning et seda oleks võimalik ema genotüübist selgelt eristada. Teiseks, emalt pärinevate alleelide DNA fragmentide eristamine ema enda rakuvabast DNA-st on tehniliselt keerukas (Benn, 2014).

Siiski, väga suure sekveneerimiskatvuse korral on võimalik määrata ka loote genotüüp ja läbi selle võimalikud emalt päritud monogeensed haigused.

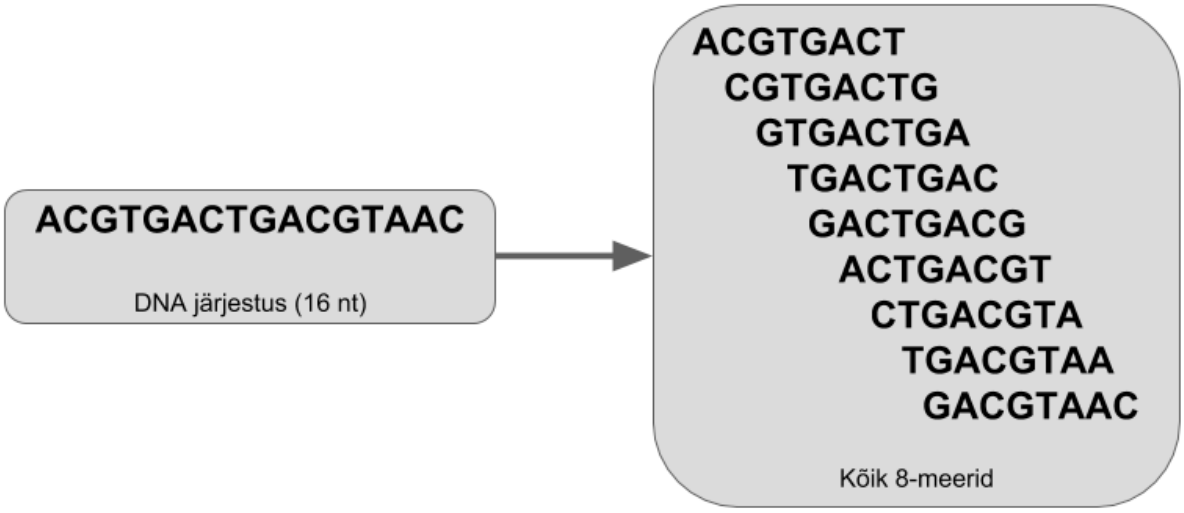
Mitmed uurimisrühmad on kasutanud vanemate genotüüpide sekveneerimist ja nende haplotüüpide võrdlemist, et seeläbi tuletada kogu loote genoom (Lo *et al*., 2010). Sellisel lähenemisel on võimalik diagnoosida ühe testiga mitmeid geneetilisi seisundeid, kuid standardse rakenduse leidmine lähitulevikus on vähetõenäoline, kuna tehnilised protseduurid ja andmeanalüüs on aeganõudvad ning rahaliselt väga kulukad (Daley *et al*., 2014). Veel enam – taoline lähenemine tõstatab ka mitmeid eetilisi ning ühiskondlikke probleeme, mis tuleks lahendada, et rajada teed tulevastele arengutele. Näiteks tuleks välistada võimalus raseduse katkestamiseks üksnes loote ebasoovitava soo teadasaamisel (Skirton ja Patch, 2013).

1.4 Lugemite paigutamine referentsgenoomile *vs* *k*-meeride sageduste lugemisel põhinev analüüs

Praegu kasutusel olevad NIPT metoodikad põhinevad sekveneerimislugemite paigutamisel (ingl k *mapping*) referentsgenoomile (Dondorp *et al*., 2015). Lugemite paigutamine on aga aeganõudev ning veaaldis protsess. Inimese genoom on suhteliselt suur, koosnedes ligikaudu 3 miljardist aluspaarist ning seetõttu on lühikesi lugemeid keeruline korrektselt referentsile paigutada. Raskendavateks asjaoludeks on ka genoomis esinevad kordusjärjestused ning

võimalikud sekveneerimisvead (Hatem *et al*., 2013).

K-meerideks nimetatakse *k* nukelotiidi pikkusi oligomeere (vt Joonis 3). Näiteks 20-meer tähistab 20 nukleotiidi pikkust oligomeeri. Paljudel analüüsidel, eelkõige kui on vaja võrrelda teatavate järjestuste suhtelisi või absoluutseid esinemissagedusi, on *k*-meeride sageduste lugemisel põhinev analüüs tehniliselt lihtsam, kiirem ja robustsem kui lugemite paigutamisel põhinevad analüüsid (Patro *et al*., 2014; Wood ja Salzberg, 2014).



Joonis 3. Näide 16 nukleotiidi pikkuselt DNA järjestuselt 8-meeride loendamisest. Antud järjestuses esinevad 8-meerid on kõik unikaalsed, mis tähendab, et kõiki esineb ainult üks kord.

1.5 Tarkvara *k*-meeride loendamiseks

On välja töötatud mitmeid tarkvarapakette, mis võimaldavad sekveneerimisel saadud toorlugemitest või juba assambleeritud genoomidest koostada *k*-meeride tabelfaile (Pérez, 2016).

Järgnevalt tutvustatakse lühidalt kolme käsurealt töötavat *k*-meeride loendusprogrammi.

1.5.1 Jellyfish 2

Jellyfish 2 anti välja 2011. aastal ning toona oli tegemist kiire ning efektiivse mälukasutusega *k*-meeride loendusprogrammiga.

Programmile on sisendiks võimalik anda üks või enam FASTA või FASTQ formaadis DNA järjestusi sisaldav fail. Väljundfail on binaarses formaadis, mida on võimalik teisendada inimloetavasse tekstiformaati ning millest on võimalik leida ka mingi kindla *k*-meeri sagedus

või koostada histogramm *k*-meeride esinemissageduste kohta. Operatsioonisüsteemidest on toetatud Linux, OS X ja Windows (Marçais ja Kingsford, 2011; veebileht 1).

1.5.2 KMC 2

KMC (K-mer Counter) 2 on Sileesia Tehnikaülikoolis välja töötatud *k*-meeride loendusprogramm, mille plussideks on kiirus ning vähene mälukasutus.

KMC-s kasutusel olev algoritm kasutab eeskätt kõvakettaruumi mitte muutmälu ning see võimaldab KMC-d ka tänapäevastel personaalarvutitel kasutada. Alates versioonist 2.3.0 on KMC-ga võimalik sooritada *k*-meeride tabelfailidega hulgateoreetilisi tehteid (leida ühendit, ühisosa ja vahet) ning *k*-meere sorteerida ja vastavalt arvukusele filtreerida ehk eemaldada liiga haruldased ja/või sagedased *k*-meerid.

Operatsioonisüsteemidest on toetatud Linux, OS X ja Windows (Deorowicz *et al*., 2015; veebileht 2).

1.5.3 GenomeTester 4

GenomeTester 4 on Tartu ülikooli Bioinformaatika õppetoolis välja töötatud tarkvarapakett, mille eeliseks võrreldes analoogsete programmidega on võimalus sooritada *k*-meeride tabelfailidega hulgateoreetilisi operatsioone (KMC viimases versioonis on ka see võimalus lisatud).

GenomeTester 4 koosneb kolmest programmist: GListMaker, GListCompare ja GListQuery. GListMaker on mõeldud FASTA või FASTQ formaadis failidest *k*-meeride loendamiseks. GListCompare sooritab *k*-meeride binaarkujul tabelfailidega hulgateoreetilisi tehteid – võimalik on arvutada ühendit, ühisosa ja vahet.

GListMaker-i ja GListCompare-i väljundfailid on samas binaarses formaadis. Failid sisaldavad *k*-meeride järjestusi ja nende arvu, lisaks on kirjas *k*-meeride koguarv ja unikaalsete *k*-meeride arv. *K*-meerid on sorteeritud tähestikulises järjekorras ning salvestatud on ainult kanoonilised *k*-meerid. See tähendab, et kui järjestuses esineb nii *k*-meer kui ka temaga pöördkomplementaarne *k*-meer, siis väljundis on neist ainult üks.

GListQuery-ga on võimalik teha *k*-meeride kohta päringuid binaarkujul tabelfailidest. Operatsioonisüsteemidest on toetatud Linux (Kaplinski *et al*., 2015).

2 EKSPERIMENTAALOSA

2.1 Töö eesmärgid

Magistritöö eksperimentaalse osa eesmärgiks oli leida stabiilsete *k*-meeride hulk, millega oleks võimalik analüüsida mitteinvasiivse sünnieelse testimise jaoks saadud sekveneerimisandmeid ja seeläbi tuvastada, kui suur on risk Downi sündroomiga lapse sünniks. Selleks töötati välja töövoog, mis jagunes järgmisteks etappideks:

- Analüüsiks mittesobivate *k*-meeride eemaldamine ehk *blacklist*i koostamine.
- Inimese iga autosoomi jaoks stabiilsete *k*-meeride hulga valimine ehk *whitelist*i koostamine.
- *K*-meeride hulkade testimine sekveneerimisandmetel ja analüüsiks parima *k*-meeride hulga valimine.

2.2 Materjal ja metoodika

Tabel 1. Kasutatud DNA järjestused.

DNA järjestus	Allikas
Inimese referentsgenoom versioon 37 *patch* 13	Veebileht 3
25 indiviidi täisgenoomi sekveneerimisandmed (vt Lisa 2)	*Estonian Center of Excellence in Genomics and Translational Medicine, project No.* 2014-2020.4.01.15-0012
51 rasedalt naiselt saadud sekveneerimisandmed (vt Lisa 3)	*SARM – Endometrial and Embryonic Genomics, Searching for Biomarkers in Assisted Reproduction*

Tabel 2. Kasutatud arvutiprogrammid ja skriptid. Skriptid faililaiendiga py on kirjutatud programmeerimiskeeles Python (veebileht 4), skriptid faililaiendiga pl on kirjutatud programmeerimiskeeles Perl (veebileht 5).

Nimi	Autor
GenomeTester 4	Kaplinski *et al*., 2015
hapl_generator.py	Käesolev töö

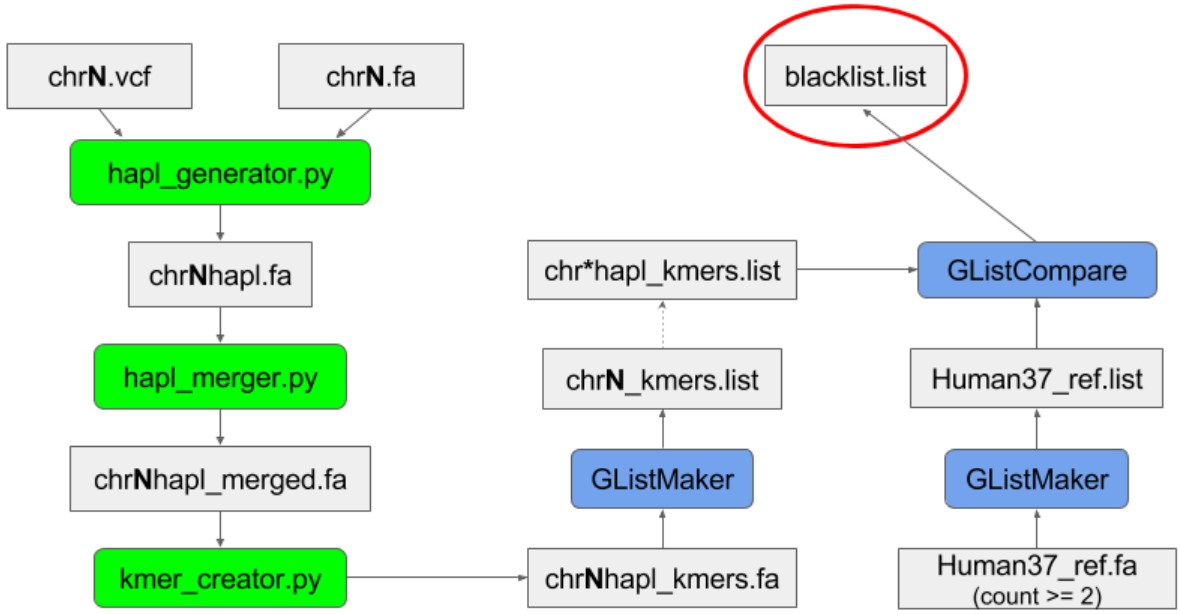
Nimi	Autor
hapl_merger.py	Käesolev töö
kmer_creator.py	Käesolev töö
kmer_filter.py	Käesolev töö
pipe.pl	Lauris Kaplinski, avaldamata
MakeIntersection.pl	Maarja Lepamets, avaldamata

Töös on kaldkirjas välja toodud lihtsustatud kujul käsurea käsud. Neis esinevate märgistuste tähendused on ära toodud tabelis 3.

Tabel 3. Kasutatud käsurea käskude märgistused.

Märgistus	Tähendus
N	Autosoomi number
K	*K*-meeri pikkus
*	Kõik nimetusele vastavad failid

2.2.1 *K*-meeride *blacklist*i koostamine



Joonis 4. Töövoog *k*-meeride *blacklist*i koostamiseks. **N** tähistab inimese autosoome, * kõiki nimetusele vastavaid faile. Rohelisega tähistatud skriptid on tehtud käesoleva töö raames, tumesinisega tähistatud GListMaker ja GListCompare kuuluvad GenomeTester 4 tarkvarapaketti. Hallides kastides on programmide sisend- ja väljundfailid.

Eksperimentaalse töö esimeseks etapiks oli *blacklist*i koostamine, mis sisaldaks inimese autosoomides esinevate teadaolevate SNP-ide ja lühikeste insertsioonide või deletsioonidega (*indels* – ingl k *insertions or deletions*) ülekattes olevaid ja inimese genoomis 2 või enam korda esinevaid *k*-meere. Töövoog *k*-meeride *blacklist*i koostamiseks on toodud joonisel 4 ning see jagunes järgmisteks osadeks:

1) Skriptiga hapl_generator.py inimese referentsautosoomide ja dbSNPv147 (veebileht 6) põhjal FASTA formaadis (pseudo)haplotüüpide failide genereerimine. Töö lihtsustamiseks kasutati PyVCF moodulit (veebileht 7), mis on mõeldud VCF (ingl k *Variant Call Format*) failide analüüsimiseks.

   DbSNP andmebaasist valiti ainult need polümorfismid, mille minoorse alleeli sagedus (MAF – ingl k *Minor Allele Frequency*) vähemalt ühes populatsioonis on >= 0,01 ja mis esineks vähemalt kahel erinevast perekonnast pärit indiviidil. Kuna kasutatud dbSNP andmebaasis sugukromosoomide polümorfismid puudusid ja sugukromosoomide aneuploidiate diagnoosimine ei olnud ka eesmärgiks, jäid X- ja Y-kromosoom polümorfismide *blacklist*i koostamisel välja. SNP-ide puhul kasutati

kettaruumi kokkuhoidmise eesmärgil alleelide tähistamisel Rahvusvahelise Puhta Keemia ja Rakenduskeemia Liidu (IUPAC – ingl k *International Union of Pure and Applied Chemistry*) mitmetähenduslike nukleotiidide märgistust (veebileht 8) (vt Joonis 5).

*hapl_generator.py -v chr**N**.vcf -i chr**N**.fa -k **K** -o chr**N**hapl.fa*

2) Skriptiga hapl_merger.py ühendati eelnevalt genereeritud failides kirjed, kus polümorfsed positsioonid asusid üksteisesele lähemal kui valitud *k*-meeri pikkus. See oli vajalik, et *blacklist*ist ei jääks välja *k*-meere, mis on ülekattes mitme SNP-i või indeliga (vt Joonis 5).

*hapl_merger.py chr**N**hapl.fa chr**N**hapl_merged.fa*

Lihtkirjed
>1:48156-48193 rs564373876 48174 A -1
ACCAAACATGTTCACATCGTGTGCGTTCCATTTTCCTA
>1:48156-48193 rs564373876 48174 B -1
ACCAAACATGTTCACATCGTTGTGCGTTCCATTTTCCTA
>1:48161-48199 rs529040510 48180 X
ACATGTTCACATCGTGTGCRTTCCATTTTCCTAATGGAA

Liitkirjed
>1:48156-48199 rs564373876/rs529040510 48174 A -1 48180 X
ACCAAACATGTTCACATCGTGTGCRTTCCATTTTCCTAATGGAA
>1:48156-48199 rs564373876/rs529040510 48174 B -1 48180 X
ACCAAACATGTTCACATCGTTGTGCRTTCCATTTTCCTAATGGAA

Joonis 5. Näide skriptide hapl_generator.py (Lihtkirjed) ja hapl_merger.py (Liitkirjed) väljundfailidest. Päises (read mis algavad ">" sümboliga) on kirjas kromosoomi number, järjestuse algus- ja lõppkoordinaadid kromosoomil, polümorfismi ID, polümorfse positsiooni koordinaat (SNP-id) või polümorfsele järjestusele eelneva nukleotiidi koordinaat (indelid), polümorfismi tüüp (A/B = indel, X = SNP) ja indeli korral selle pikkus (miinus viitab insertsioonile referentsi suhtes, pluss deletsioonile). Punasega on tähistatud polümorfsed positsioonid (R tähistab nukleotiidi A või G).

3) Skriptiga *kmer_creator.py* genereeriti eelmises etapis saadud ühendatud kirjetega failidest FASTA formaadis *k*-meeride nimekirjad, kus mitmetähenduslikud nukleotiidid asendati igas positsioonis kõigi võimalike neile vastavate nukleotiididega (A, C, G või T-ga). See oli vajalik, sest GListMaker-iga ei ole võimalik genereerida *k*-meere järjestustest, kus esinevad mitmetähenduslikud nukleotiidid.

*kmer_creator.py chr**N**hapl_merged.fa chr**N**hapl_kmers.fa **K***

4) Saadud *k*-meeride FASTA formaadis nimekirjad konverteeriti binaarses formaadis *k*-meeride tabelfailideks, kasutades selleks GenomeTester 4 tarkvarapaketis olevat GListMaker programmi.

   *glistmaker chr**N**hapl_kmers.fa -w **K** -o chr**N**hapl_kmers.list*

5) GListMaker-iga tehti inimese referentsgenoomist *k*-meeride tabelfail, mis sisaldaks kõiki *k*-meere, mis esinevad genoomis 2 või enam korda.

   *glistmaker Human37_ref.fa -w **K** -c 2 -o Human37_ref.list*

6) GListCompare-iga ühendati 4. ja 5. punktis saadud tabelfailid üheks failiks.

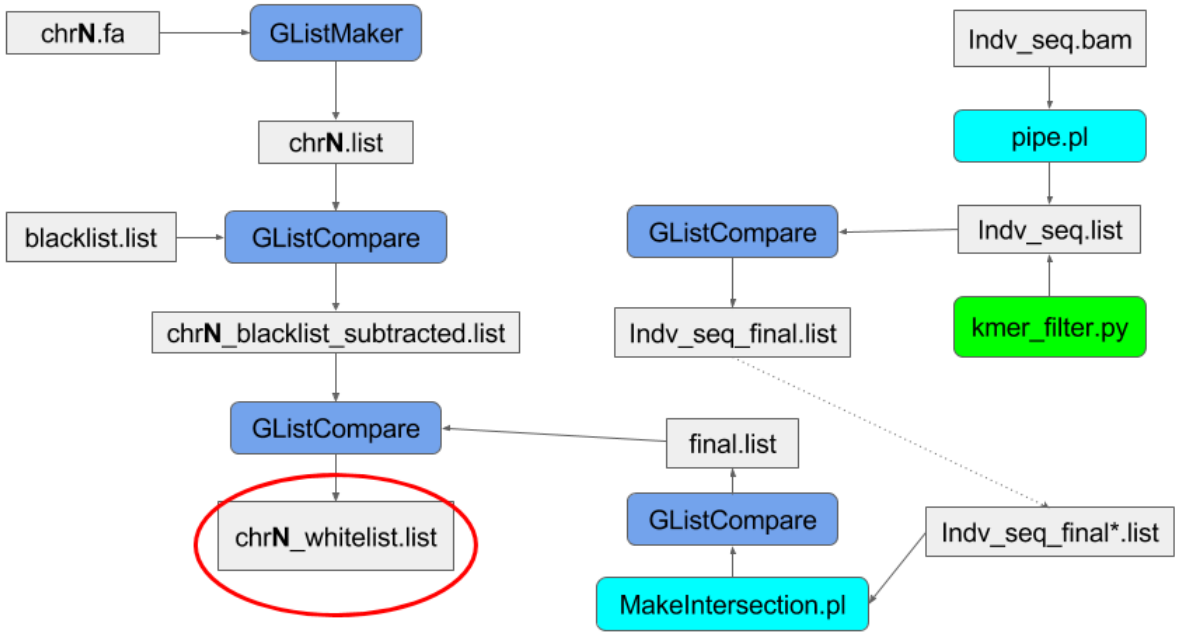
   *glistcompare chr*hapl_kmers.list Human37_ref.list -u -o blacklist.list*

Tulemuseks saadi *blacklist*, mis sisaldab inimese autosoomides polümorfsete positsioonidega ülekattes olevaid ja mitteunikaalseid ehk genoomis 2 või enam korda esinevaid *k*-meere. *Blacklist*ide suurused on toodud tabelis 4.

Tabel 4. *Blacklist*ides ja referentsgenoomis esinevate unikaalsete *k*-meeride arvud.

	***Blacklist*i *k*-meeride arv**	**Referentsgenoomi *k*-meeride arv**
16-meerid	639 266 507	798 662 373
20-meerid	1 275 070 644	2 186 447 047
32-meerid	2 113 346 891	2 509 328 883

2.2.2 *K*-meeride *whitelist*ide koostamine



Joonis 6. Töövoog *k*-meeride *whitelist*ide koostamiseks. **N** tähistab inimese autosoome, * kõiki nimetusele vastavaid faile. Rohelisega tähistatud skript on tehtud käesoleva töö raames, tumesinisega tähistatud programmid GListMaker ja GListCompare kuuluvad GenomeTester 4 tarkvarapaketti. Hallides kastides on programmide sisend- ja väljundfailid.

Töö teiseks etapiks oli inimese referentsautosoomidest *blacklist*is esinevate *k*-meeride eemaldamine ja tegelike genoomide sekveneerimisandmete (vt Lisa 2) põhjal *k*-meeride filtreerimine. Eesmärgiks oli saada inimese kõigi autosoomide jaoks stabiilsete ehk enamike indiviidide genoomis üks ja ainult üks kord esinevate *k*-meeride hulk. Töövoog stabiilsete *k*-meeride saamiseks on toodud joonisel 6 ning see jagunes järgmisteks osadeks:

1. Inimese referentsautosoomidest loodi GListMaker-iga *k*-meeride tabelfailid.

   *glistmaker chr**N**.fa -w **K** -o chr**N**.list*

2. Kõikidest referentsautosoomide tabelfailidest eemaldati *blacklist*is esinevad *k*-meerid. GListCompare-iga leiti referentsautosoomide *k*-meeride ja *blacklist*is esinevate *k*-meeride vahe.

   *glistcompare chr**N**.list blacklist.list -d -o chr**N**_blacklist_subtracted.list*

3. BAM formaadis sekveneerimisandmete failide konverteerimine *k*-meeride tabelfailideks.

Lähteandmeteks võeti 5, 10 või 25 (vt Lisa 2) indiviidi BAM formaadis sekveneerimisandmete failid, mis konverteeriti *pipe.pl* skriptiga binaarses formaadis *k*-meeride tabelfailideks.

*pipe.pl Indv_seq.bam **K***

4. Indiviidide täisgenoomi sekveneerimisandmete põhjal *k*-meeride filtreerimine.

   Skriptiga kmer_filter.py arvutati iga indiviidi sekveneerimise mediaankatvus ning Poissoni jaotusele ($p < 0,01$) vastavad minimaalne ja maksimaalne lugemite arv, mida võiks näha kui mingit *k*-meeri esineks sekveneeritud genoomis täpselt üks kord. Ideaalsel sekveneerimisel peaks genoomis unikaalsete *k*-meeride sageduste jaotus vastama Poissoni jaotusele (λ = katvus) (veebileht 9). Mingile *k*-meerile vastavate lugemite väike või väga suur arv viitavad kas sekveneerimistehnoloogia kallutatusele, saastusele või senikirjeldamata polümorfismidele genoomis. Kuna kõik sellised *k*-meerid võivad põhjustada NIP testil kromosoomide katvuste üle- või alahindamist, on mõistlik nad analüüsist eemaldada (vt Joonis 7).

   kmer_filter.py Indv_seq.list

   Saadud minimaalse ja maksimaalse väärtuse järgi filtreeriti *k*-meeride tabelfaile:

   Esiteks, leiti ühisosa, nii et minimaalsest *k*-meeride sagedusest madalama sagedusega *k*-meerid jääksid välja.

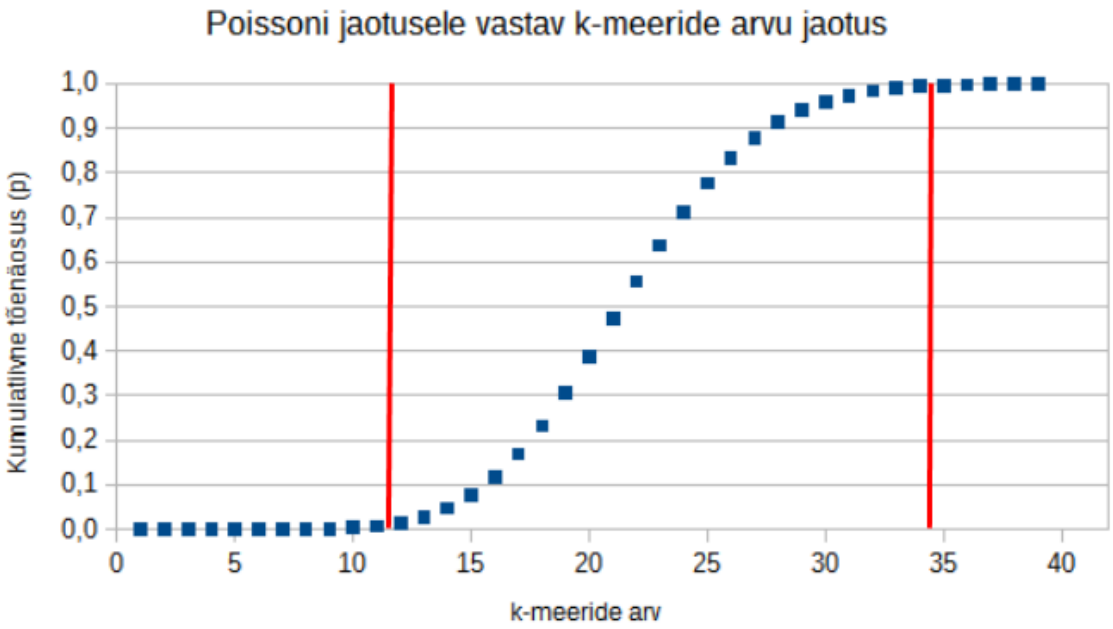
   glistcompare Indv_seq.list Indv_seq.list -i -c <min> -o Indv_seq_min.list

   Teiseks, leiti ühisosa, nii et maksimaalsest *k*-meeride sagedusest kõrgema sagedusega *k*-meerid jääksid välja.

   glistcompare Indv_seq.list Indv_seq.list -i -c <max+1> -o Indv_seq_max.list

   Viimaks arvutati kahe tabelfaili vahe.

   glistcompare Indv_seq_min.list Indv_seq_max.list -d -o Indv_seq_final.list



Joonis 7. Näide *k*-meeride filtreerimisest kui mediaankatvus (λ) on 22. Välja filtreeritakse *k*-meerid, mille esinemissageduse kumulatiivne tõenäosus on <0,01 ehk joonisel vasakpoolsest punasest joonest vasakule (arvukus < 12) ja parempoolsest punasest joonest paremale (arvukus > 34) jäävate arvukustega *k*-meerid.

5. Indiviidide sekveneerimisandmete *k*-meeride ühisosa leidmine.

   Indiviidide filtreeritud *k*-meeride tabelfailidest leiti GListCompare programmiga *k*-meeride ühisosa. Kasutati skripti *MakeIntersection.pl*, mis lihtsustas enam kui kahe tabelfaili ühisosa leidmist.

   MakeIntersection.pl Indv_seq_final.list*

   Väljundfailiks oli *final.list*, mis sisaldas kõigi indiviidide filtreeritud *k*-meeride ühisosa.

6. Indiviidide sekveneerimisandmete *k*-meeride ühisosast ja punktis 2 saadud iga autosoomi *k*-meeride hulgast leiti omakorda ühisosa.

   *glistcompare final.list chr**N**_blacklist_subtracted.list -i -o chr**N**_whitelist.list*

   Tulemuseks saadi inimese iga autosoomi jaoks *whitelist*, millega analüüsiti NIPT-i jaoks saadud sekveneerimisandmeid.

2.2.3 Mitte-invasiivseks sünnieelseks testimiseks saadud sekveneerimisandmete analüüs

Stabiilsete *k*-meeride hulkadega analüüsiti NIPT-i jaoks kogutud 51 raseda naise sekveneerimisandmeid. Töö jagunes järgmisteks osadeks:

1. Sekveneerimisproovide FASTQ formaadis failidest *k*-meeride tabelfailide koostamine.

   *glistmaker NIPT.fastq -w **K** -o NIPT.list*

2. GListCompare programmi kasutades leiti NIPT proovi ja iga autosoomi *whitelist*i *k*-meeride ühisosa. Sealjuures määrati GListCompare reegliks, et saadud tabelfailides olevad *k*-meeride arvud on pärit NIPT-i proovist (*-r first*).

   *glistcompare NIPT.list chr**N**_whitelist.list -i -r first -o chr**N**.list*

3. Statistiline töötlust tehti iga proovi autosoomide *k*-meeride koguarvudega ja unikaalsete *k*-meeride arvuga ning see koosnes järgmistest etappidest:

   3.1. Proovist leitud iga autosoomi *k*-meeride arv jagati vastava autosoomi *whitelist*is esinevate *k*-meeride arvuga. Saadi proovi autosoomide keskmine katvus.

   $O = Oi / Ei$

   3.2. Arvutati kõikide autosoomide keskmiste katvuste, välja arvatud 21. kromosoomi, keskmine. Saadi proovi keskmine katvus.

   $K = AVG(O)$

   3.3. Arvutati iga proovi 21. kromosoomi keskmise katvuse ja kogu vastava proovi keskmise katvuse suhe. Saadi 21. kromosoomi normaliseeritud katvus.

   $N_{21} = O_{21} / K$

   3.4. Arvutati negatiivsete proovide 21. kromosoomi normaliseeritud katvuste keskmine ja standardhälve.

   $AVG(N_{21}\_NEG)$; $STDEV(N_{21}\_NEG)$

3.5. Valiti positiivsetest proovidest minimaalse 21. kromosoomi normaliseeritud katvusega proov, leiti vahe punktis 3.4 saadud keskmisega ning jagati punktis 3.4 saadud standardhälbega.

$$(MIN(K_{21}\_POS) - AVG(N_{21}\_NEG)) / STDEV(N_{21}\_NEG)$$

Tulemuseks saadi negatiivsete proovide keskmisest kõige vähem eristunud positiivse proovi suhteline erinevus standardhälbe ühikutes (edaspidi skoor [z]), mida kasutati kvantitatiivse parameetrina *k*-meeride hulkade hindamiseks.

Tabel 5. Statistilise töötluse kirjeldamisel kasutatud tähised ja nende tähendused.

Tähis	Tähendus
Oi	Proovi autosoomi *k*-meeride arv
Ei	*Whitelist*i autosoomi *k*-meeride arv
O	Autosoomi keskmine katvus
K	Proovi keskmine katvus
N	Autosoomi normaliseeritud katvus
AVG	Keskmine
STDEV	Standardhälve

2.3 Tulemused ja arutelu

Magistritöö eksperimentaalse osa raames töötati välja töövoog mitteinvasiivse sünnieelse testimise jaoks kogutud sekveneerimisandmete analüüsimiseks, et leida stabiilne *k*-meeride hulk loote Downi sündroomi riski hindamiseks.

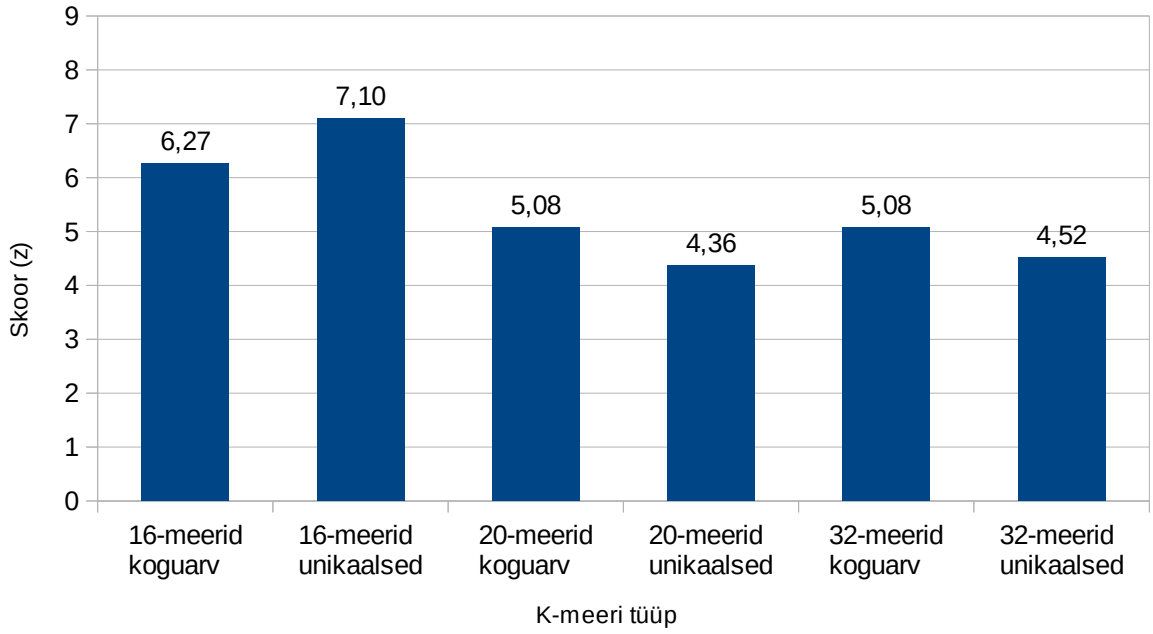
Töövoo abil testiti 18 erinevat *k*-meeride hulka (vt Lisa 1) 51 rasedalt naiselt saadud sekveneerimisandmetel, millest 4 puhul oli teada, et need on positiivsed Downi sündroomiga loote suhtes (vt Lisa 3).

Analüüsid teostati 16-, 20- ja 32-meeridega. *K*-meeride hulgad varieerusid järgmiste parameetrite poolest: polümorfsete positsioonidega ülekattes olevate *k*-meeride eemaldamine *vs* eemaldamata jätmine ja *k*-meeride filtreerimine tegelike genoomide põhjal *vs* filtreerimata jätmine. Lisaks võrreldi tulemuste erinevust skveneerimisandmetest leitud *k*-meeride koguarvu ja leitud unikaalsete *k*-meeride arvu korral.

Parimaks *k*-meeride hulgaks valiti hulk, millega sekveneerimisandmeid analüüsides saadi positiivsetest proovidest madalaima 21. kromosoomi normaliseeritud katvusega prooviga kõrgeim skoor (vt lehekülg 25 paragrahv 3.5.).

2.3.1 Polümorfsete positsioonidega ülekattes olevate *k*-meeride allesjätmine

Saamaks kinnitust, kas polümorfsete positsioonidega ülekattes olevate *k*-meeride eemaldamine *k*-meeride hulgast on mõttekas, arvutati esmalt skoorid kui referentsautosoomide *k*-meeride tabelfailidest olid eemaldatud ainult genoomis 2 või enam korda esinevad *k*-meerid. Sel juhul koosnes *blacklist* ainult *k*-meeridest, mis saadi etapis, mis on kirjeldatud leheküljel 20 paragrahvis 5.



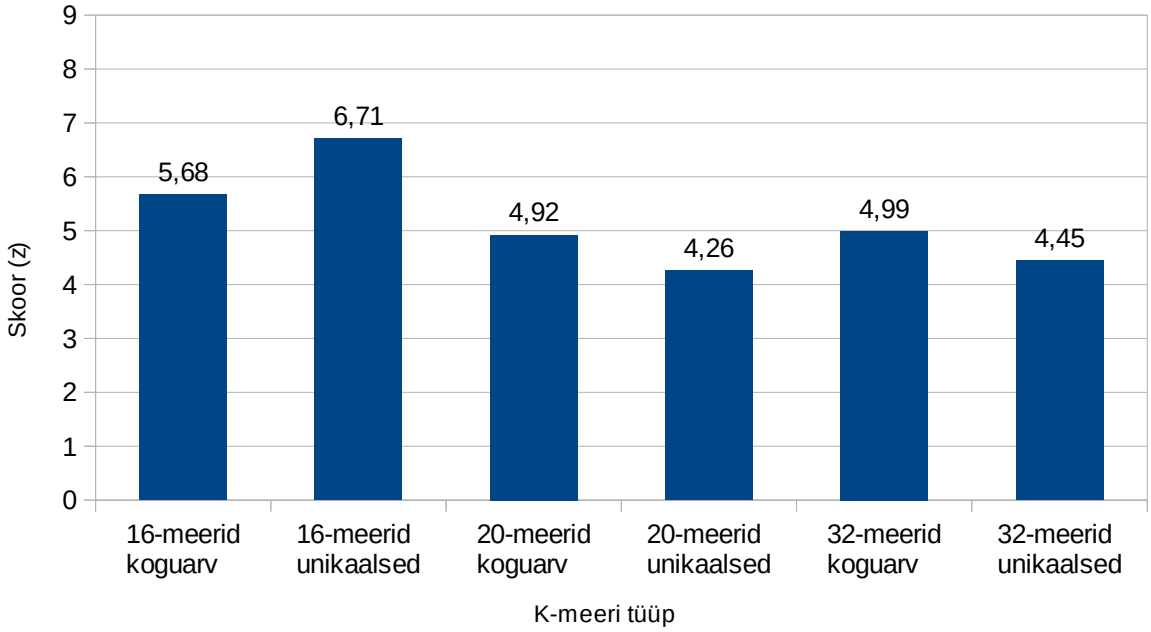
Joonis 8. Skoorid analüüsil *k*-meeride hulkadega, kus referentsautosoomidest on eemaldatud ainult genoomis 2 või enam korda esinevad *k*-meerid.

Kõrgeima skoori (7,10) andis sekveneerimisandmete analüüs proovist leitud unikaalsete 16-meeridega. Proovist leitud 16-meeride koguarvude analüüsi põhjal oli skoor 6,27.

20- ja 32-meeride puhul olid skoorid madalamad. Koguarvude analüüs andis mõlemal skooriks 5,08. Proovist leitud unikaalsete 20-meeride arvude põhine analüüs andis skooriks 4,36. Unikaalsete 32-meeride arvude põhisel analüüsil oli skooriks 4,52.

2.3.2 Polümorfsete piirkondadega ülekattes olevate *k*-meeride eemaldamine

Teisena teostati analüüs *k*-meeride hulkadega, kust lisaks genoomis 2 või enam korda esinevatele *k*-meeridele olid eemaldatud ka polümorfsete piirkondadega ülekattes olevad *k*-meerid.



Joonis 9. Skoorid analüüsil *k*-meeride hulkadega, kust on eemaldatud genoomis 2 või enam korda esinevad ja polümorfsete piirkondadega ülekattes olevad *k*-meerid.

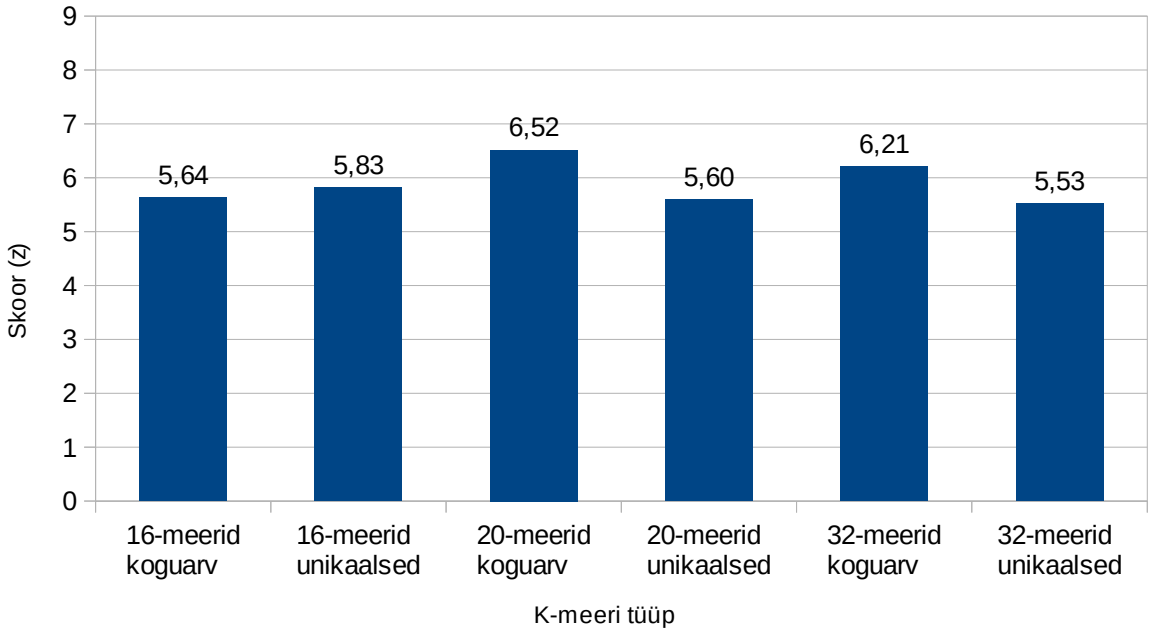
Kõrgeima skoori (6,71) andis sekveneerimisandmete analüüs proovist leitud unikaalsete 16-meeridega. Proovist leitud 16-meeride koguarvude analüüsi põhjal oli skoor 5,68.

20- ja 32-meeride puhul olid skoorid madalamad. Analüüs koguarvude põhjal andis 20-meeridega skooriks 4,92, 32-meeridega 4,99. Proovist leitud unikaalsete 20-meeride arvude põhine analüüs andis skooriks 4,26. Unikaalsete 32-meeride korral oli skooriks 4,45.

Kõik saadud skoorid olid madalamad kui polümorfsete piirkondadega ülekattes olevaid *k*-meere mitte eemaldades (vt Joonis 8), mis viitab sellele, et hulkadest eemaldati (ka) liiga palju analüüsiks sobilikke *k*-meere.

2.3.3 Populatsiooni tegeliku varieeruvuse arvestamine

Järgmisena vaadati, kas skoore mõjutab *k*-meeride hulkade filtreerimine indiviidide sekveneerimisandmete põhjal. Esmalt teostati filtreerimine viie indiviidi põhjal.

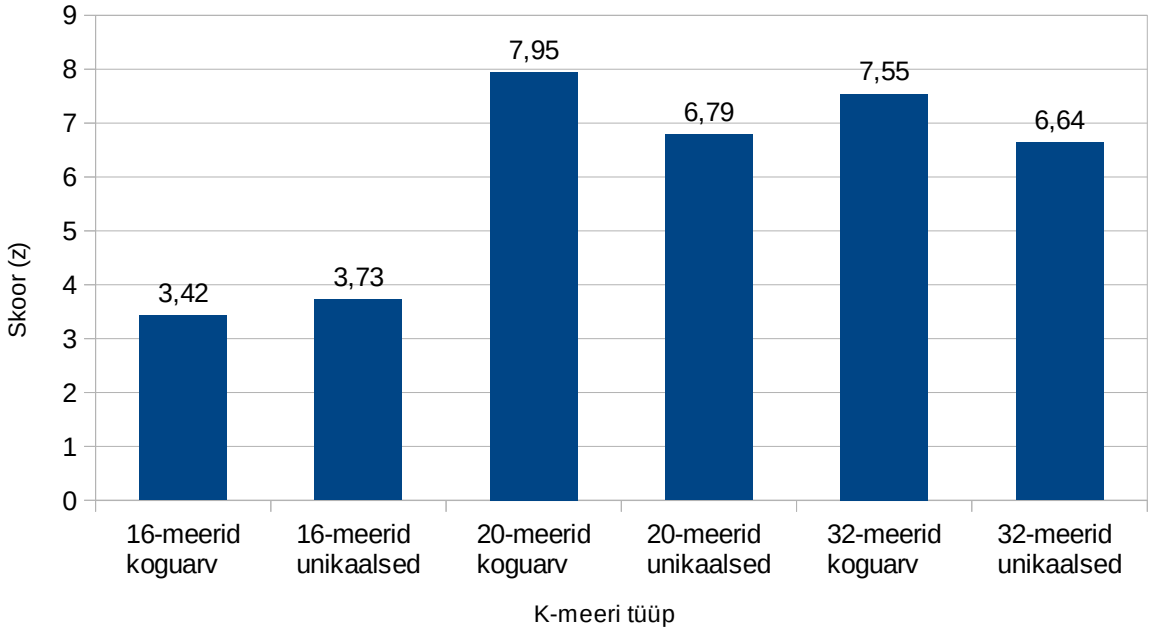


Joonis 10. Skoorid analüüsil *k*-meeride hulkadega, kust on eemaldatud genoomis 2 või enam korda esinevad ja polümorfsete piirkondadega ülekattes olevad *k*-meerid ning mis on filtreeritud viie indiviidi sekveneerimisandmete põhjal.

Kõrgeima skoori andis analüüs proovist leitud 20-meeride koguarvude põhjal, ent saadud skoor oli siiski madalam (6,52 vs 7,10) kui polümorfsete positsioonidega ülekattes olevaid *k*-meere mitte eemaldades ning mitte filtreeritud unikaalsete 16-meeride põhjal saadud skoor (vt Joonis 8).

32- ja 16-meeride puhul olid skoorid madalamad. Analüüs koguarvude põhjal andis 32-meeridega skooriks 6,21, 16-meeridega 5,64. Proovist leitud unikaalsete 32-meeride arvude põhine analüüs andis skooriks 5,53. Unikaalsete 16-meeride korral oli skooriks 5,83.

Järgmisena teostati filtreerimine kümne indiviidi sekveneerimisandmete põhjal.

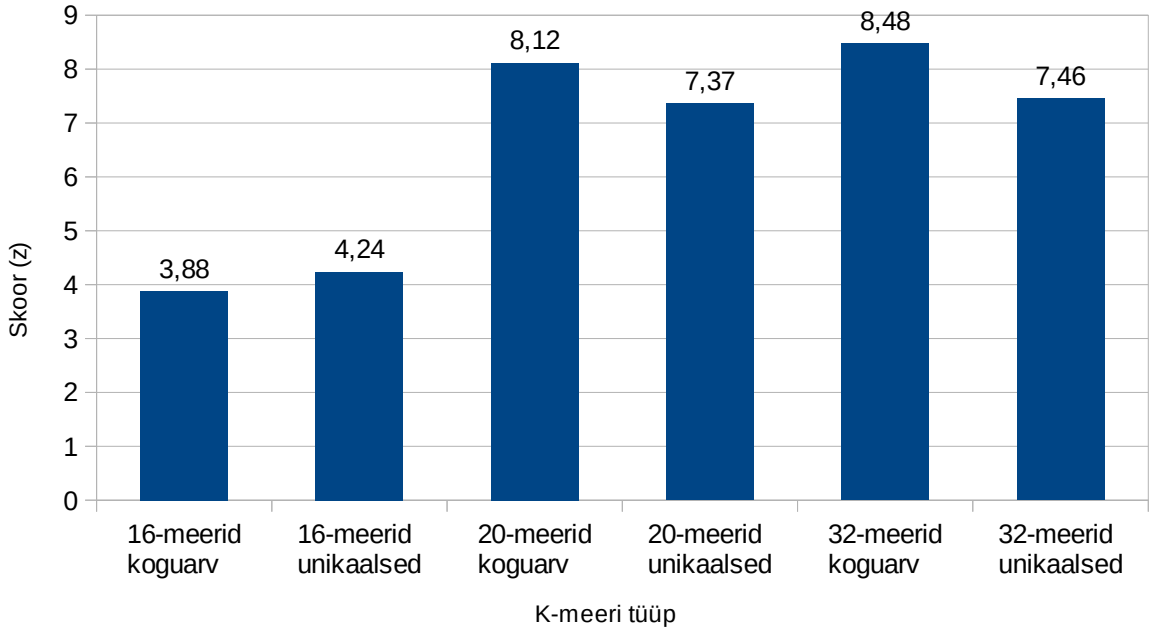


Joonis 11. Skoorid analüüsil *k*-meeride hulkadega, kust on eemaldatud genoomis 2 või enam korda esinevad ja polümorfsete piirkondadega ülekattes olevad *k*-meerid ning mis on filtreeritud kümne indiviidi sekveneerimisandmete põhjal.

Kõrgeima skoori (7,95) andis analüüs proovist leitud 20-meeride koguarvude põhjal. Analüüsil 32-meeride koguarvude põhjal oli skooriks 7,55.

16-meeride analüüsil saadud skoorid olid ligikaudu kaks korda madalamad (koguarvude põhjal 3,42 ja unikaalsete arvu põhjal 3,73), selle põhjuseks on tõenäoliselt analüüsiks liiga väheste *k*-meeride allesjäämine.

Kuna filtreerimata *k*-meeride hulkade puhul olid skoorid kõrgemad polümorfsete piirkondadega ülekattes olevaid *k*-meere mitte eemaldades (võrdle Joonis 8 ja Joonis 9), vaadati, milline on efekt filtreerimise ja polümorfsete piirkondadega ülekattes olevate *k*-meeride allesjätmise koosmõjul.



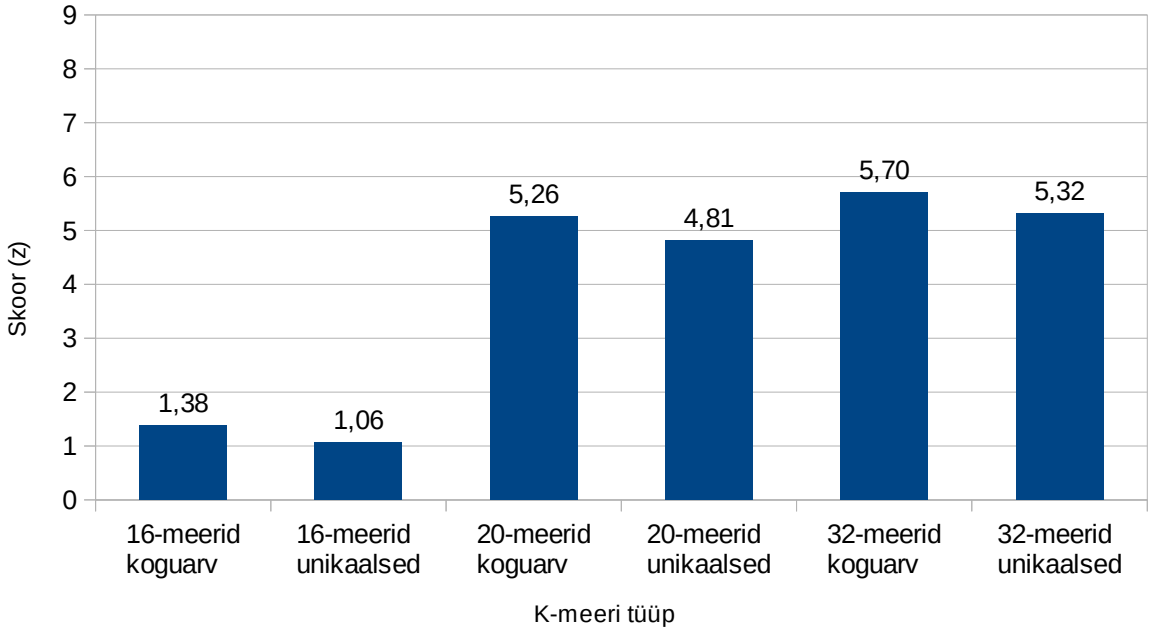
Joonis 12. Skoorid analüüsil *k*-meeride hulkadega, kust on eemaldatud ainult genoomis 2 või enam korda esinevad *k*-meerid ja mis on filtreeritud kümne indiviidi sekveneerimisandmete põhjal.

Kõrgeima skoori (8,48) andis analüüs 32-meeride koguarvu põhjal. Analüüsil 20-meeride koguarvude põhjal oli skooriks 8,12.

16-meeride analüüsil saadud skoorid olid ligikaudu kaks korda madalamad (koguarvude põhjal 3,88 ja unikaalsete arvu põhjal 4,24).

Kõik saadud skoorid olid kõrgemad kui kümne indiviidi sekveneerimisandmete põhjal filtreeritud ja polümorfsete positsioonidega ülekattes olevate *k*-meeride eemaldamisel saadud skoorid (vt Joonis 11).

Viimasena teostati filtreerimine 25 indiviidi sekveneerimisandmete põhjal.



Joonis 13. Skoorid analüüsil *k*-meeride hulkadega, kust on eemaldatud genoomis 2 või enam korda esinevad ja polümorfsete piirkondadega ülekattes olevad *k*-meerid ning mis on filtreeritud 25-e indiviidi sekveneerimisandmete põhjal.

Kõrgeima skoori (5,70) andis proovist leitud 32-meeride koguarvude põhine analüüs. Proovist leitud unikaalsete 32-meeride analüüsi põhjal oli skoor 5,32.

20- ja 16-meeride puhul olid skoorid madalamad. Analüüs koguarvude põhjal andis 20-meeridega skooriks 5,26, 16-meeridega 1,38. Proovist leitud unikaalsete 20-meeride arvude põhine analüüs andis skooriks 4,81. Unikaalsete 16-meeride korral oli skooriks 1,06.

Kõikide *k*-meeride hulkade puhul olid skoorid madalamad kui kümne indiviidi sekveneerimisandmete põhjal filtreerides.

2.3.4 Programmide tööaja mõõtmine

Töös kirjeldatud arvutuslik töö tehti Tartu ülikooli Bioinformaatika õppetooli CentOS 7.1.1503 Linuxi serveris, millel on 32 tuumaline protsessor taktsagedusega 2,27 GHz ja 512 GB muutmälu.

Programmide tööaegu mõõdeti Linuxi süsteemse *time* programmiga. Esitatud tööajad on arvutatud kui programmide kasutaja- ja süsteemiaegade summa. Reaalne tööaeg sõltub lisaks veel serveri koormusest programmide töötamise ajal (vt Tabel 6).

Tabel 6. Töös kasutatud programmide tööajad ja andmemahud on toodud 10. kromosoomi ja 20-meeride näitel (välja arvatud kmer_filter.py). Tööajad on kolme eraldiseisva katse keskmine.

Programmi nimi	Sisendmaht	Väljundmaht	Tööaeg
hapl_generator.py	351 MB (VCF) + 130 MB (FASTA)	158 MB	4 min 23.16 s
hapl_merger.py	158 MB	168 MB	38.98 s
kmer_creator.py	168 MB	4 GB	23 min 39.96 s
kmer_filter.py	36 GB	-	1 h 49 min 19.52 s

2.4 Järeldused

Töö eesmärgiks oli leida stabiilsete *k*-meeride hulk, millega analüüsida mitteinvasiivse sünnieelse testimise jaoks kogutud sekveneerimisandmeid.

On selge, et mida suurem on analüüsil kasutavate *k*-meeride hulk, seda väiksem on tulemuste juhuslik kõikumine. Samas, osad *k*-meerid pärinevad genoomsetelt regioonidelt, mis on populatsioonis varieeruvad. Seega on vajalik leida optimaalne hulk *k*-meere, mille eemaldamine vähendab maksimaalselt populatsiooni heterogeensusest tingitud varieeruvust, aga ei vähenda allesjäävate *k*-meeride hulka nii palju, et sekveneerimise juhuslik varieeruvus hakkaks tulemusi halvendama.

Magistritöö eksperimentaalse osa raames testiti 18 *k*-meeride hulka. Parimaks osutus 32-meeride hulk, millest oli eemaldatud mitteunikaalsed ehk genoomis 2 või enam korda esinevad *k*-meerid ja mis olid seejärel filtreeritud 10 indiviidi täisgenoomi sekveneerimisandmete põhjal.

Tulemused halvenesid, kui *k*-meeride hulkadest eemaldati polümorfsete piirkondadega ülekattes olevad *k*-meerid. Tõenäoliselt on selle põhjuseks liiga paljude stabiilsete *k*-meeride analüüsist väljajätmine. Eemaldati *k*-meerid, mis olid ülekattes polümorfismidega, mille MAF on >=0,01. Enamik selliseid polümorfisme on populatsiooni enamikel indiviididel ühe haplotüübina. Võimalik, et parema tulemuse annaks, kui kasutada suuremat MAF väärtust.

Suur osa *k*-meere eemaldati ka seetõttu, et polümorfismidega ülekattes olevate *k*-meeride *blacklist*i tegemisel ei kasutatud haplotüüpide andmeid. Seega eemaldati kõikvõimalikele

polümorfismide alleelikombinatsioonidele vastavad *k*-meerid, kuigi tegelikkuses on lähedaste polümorfismide alleelid omavahel aheldunud.

Edasises töös tuleks välja selgitada, kas oleks kasulik langetada tegelike genoomide põhjal *k*-meeride filtreerimisel *p*-väärtust, kuid kasutada seevastu rohkemate indiviidide sekveneerimisandmeid. Lisaks, on võimalik, et analüüsi tundlikkus kasvab kui *k*-meeride hulgast eemaldada ainult *k*-meerid, mis on ülekattes polümorfsete piirkondadega, mille MAF on >>0,01. Sel juhul eemaldataks analüüsist vähem stabiilseid *k*-meere.

KOKKUVÕTE

Traditsioonilistel loote pärilike haiguste sünnieelse tuvastamise meetoditel on mitmeid puudusi. Mitteinvasiivsed meetodid, vereanalüüs ja ultraheliuuring, ei ole piisavalt täpsed ning annavad palju valepositiivseid ja -negatiivseid tulemusi. Invasiivsete meetodite, amniotsenteesi ja koorioni hattude biopsia miinuseks on aga risk raseduse katkemisele ja ema tervisele. Seetõttu on oluline uute ja paremate pärilike haiguste sünnieelse tuvastamise meetodite väljatöötamine.

Üheks perspektiivikamaks suunaks peetakse rakuvaba DNA analüüsimist. Rasedatel naistel ringleb veres platsentaarse päritoluga DNA fragmente, mida on võimalik kasutada loote geneetilise konditsiooni skriinimiseks. Alates eelmise kümnendi keskpaigast on loote rakuvaba DNA analüüsi põhiseid meetodeid hakatud ka kliiniliselt rakendama. Tegemist on mitteinvasiivsete ning potentsiaalselt väga täpsete meetoditega.

Rakuvaba DNA analüüsimist on hõlbustanud teise põlvkonna sekveneerimistehnoloogiate areng. Täisgenoomide sekveneerimine on muutunud piisavalt odavaks, et seda massiliselt teha.

Enamik praegu kasutusel olevatest rakuvaba DNA analüüsi metoodikatest põhinevad sekveneerimislugemite paigutamisel referentsgenoomile. See on aga ajamahukas ning veaaldis protsess. Seepärast töötati käesoleva magistritöö raames välja *k*-meeride loendamisel põhinev meetod.

Magistritöö raames töötati välja töövoog leidmaks optimaalne *k*-meeride hulk mitteinvasiivse sünnieelse testimise jaoks saadud sekveneerimisandmete analüüsiks. Edasises töös oleks vaja analüüsida enamate *k*-meeride hulkadega rohkem sekveneerimisandmeid, et leida analüüside teostamiseks parim *k*-meeride hulk.

Choosing a set of stable *k*-mers for Non-Invasive Prenatal Testing

Kaarel Koitne

SUMMARY

Methods currently used for prenatal diagnosis have many shortcomings. Non-invasive methods, blood test and nuchal scan, are not very accurate and produce many false-positive and -negative results. On the other hand, invasive methods, such as amniocentesis and chorionic villus sampling carry a small, but significant risk for fetal loss and maternal morbidity. Therefore, it is important to develop better methods for prenatal testing.

One of the most promising are methods based on analyzing cell-free DNA. During pregnancy, cell-free DNA which originates from the placenta circulates in the maternal plasma. This DNA, which is called cell-free fetal DNA can be used for screening of fetal genetic condition. Methods based on cell-free DNA analysis are non-invasive and potentially very accurate.

Advances in sequencing technologies have significantly accelerated the development of methods based on cell-free DNA analysis. Whole genome sequencing is nowadays available on a massive scale.

Most current methods which facilitate the analysis of cell-free DNA are based on read mapping. However, that is a slow and error-prone process. Therefore a new method, based on *k*-mer counts is being proposed.

In conclusion, a pipeline for choosing a set of stable *k*-mers for analysing sequencing data acquired for Non-Invasive Prenatal Testing was developed. Still, further work is needed to find the best set of *k*-mers for Non-Invasive Prenatal Testing.

KIRJANDUSE LOETELU

Alberry M., Maddocks D., Jones M. *et al.* (2007). Free fetal DNA in maternal plasma in anembryonic pregnancies: confirmation that the origin is the trophoblast. Prenatal Diagnosis. 27: 415–18.

American College of Obstetricians and Gynecologists (2012). Non-invasive prenatal testing for fetal aneuploidy. Committee Opinion No. 545. Obstetetrics and Gynecology. 120: 1532–4.

Attilakos G., Maddocks D. G., Davies T., Hunt L., Avent N., Soothill P., Grant S. (2011). Quantification of free fetal DNA in multiple pregnancies and relationship with chorionicity. Prenatal Diagnosis 31(10): 967–972.

Benn P. (2014). Non-Invasive Prenatal Testing Using Cell Free DNA in Maternal Plasma: Recent Developments and Future Prospects. Journal of Clinical Medicine 3(2): 537–565.

Bevilacqua E., Gil M. M., Nicolaides K. H. *et al.* (2015). Performance of screening for aneuploidies by cell-free DNA analysis of maternal blood in twin pregnancies. Ultrasound in Obstetrics & Gynecology 45(1): 61–66.

Brady P., Brison N., Van Den Bogaert K., de Ravel T., Peeters H., Van Esch H., Devriendt K., Legius E. and Vermeesch J. (2015). Clinical implementation of NIPT – technical and biological challenges. Clinical Genetics 89(5): 523–530.

Chan K. C., Zhang J., Hui A. B. *et al.* (2004). Size distributions of maternal and fetal DNA in maternal plasma. Clinical Chemistry 50: 88–92.

Chiang T., Schultz R., Lampson M. (2011). Meiotic Origins of Maternal Age-Related Aneuploidy. Biology of Reproduction 86(1): 1–7.

Chiu R. W., Lo Y. M. (2013). Clinical applications of maternal plasma fetal DNA analysis: translating the fruits of 15 years of research. Clinical Chemistry and Laboratory Medicine 51: 197–204.

Daley R., Hill M., Chitty L. S. (2014). Non-invasive prenatal diagnosis: progress and potential. Archives of Disease in Childhood Fetal Neonatal Edition 99(5): 426–430.

Deorowicz S., Kokot M., Grabowski S., Debudaj-Grabysz A. (2015). KMC 2: fast and resource-frugal k-mer counting. Bioinformatics 31(10): 1569–1576.

Dondorp W., de Wert G., Bombard Y., Bianchi D., Bergmann C., Borry P., Chitty L., Fellmann F., Forzano F., Hall A., Henneman L., Howard H., Lucassen A., Ormond K., Peterlin B., Radojkovic D., Rogowski W., Soller M., Tibben A., Tranebjærg L., van El C., Cornel M. (2015). Non-invasive prenatal testing for aneuploidy and beyond: challenges of responsible innovation in prenatal screening. European Journal of Human Genetics 23(11): 1438–1450.

Driscoll D., Gross S. (2009). Prenatal Screening for Aneuploidy. New England Journal of Medicine 360(24): 2556–2562.

Gardner R. J. M., Sutherland G. R. (2004). Chromosome abnormalities and genetic counseling, 3rd edn. UK: Oxford University Press.

Grömminger S., Yagmur E., Erkan S., Nagy S., Schöck U., Bonnet J., Smerdka P., Ehrich M., Wegner R., Hofmann W., Stumm M. (2014). Fetal Aneuploidy Detection by Cell-Free DNA Sequencing for Multiple Pregnancies and Quality Issues with Vanishing Twins. Journal of Clinical Medicine 3(3): 679–692.

Haghiac M., Vora N. L., Basu S., Johnson K. L., Presley L., Bianchi D. W. *et al.* (2012). Increased death of adipose cells, a path to release cell-free DNA into systemic circulation of obese women. Obesity (Silver Spring) 20(11): 2213–9.

Hatem A., Bozdağ D., Toland A., Çatalyürek Ü. (2013). Benchmarking short sequence mapping tools. BMC Bioinformatics 14(1): 184.

Huang X., Zheng J., Chen M., Zhao Y., Zhang C., Liu L., Xie W., Shi S., Wei Y., Lei D., Xu C., Wu Q., Guo X., Shi X., Zhou Y., Liu Q., Gao Y., Jiang F., Zhang H., Su F., Ge H., Li X., Pan X., Chen S., Chen F., Fang Q., Jiang H., Lau T., Wang W. (2014). Noninvasive prenatal testing of trisomies 21 and 18 by massively parallel sequencing of maternal plasma DNA in twin pregnancies. Prenatal Diagnosis 34(4): 335–340.

Jensen T. J., Dzakula Z., Deciu C. *et al.* (2012). Detection of microdeletion 22q11.2 in a fetus by next-generation sequencing of maternal plasma. Clinical Chemistry 58: 1148–51.

Kaplinski L., Lepamets M., Remm M. (2015). GenomeTester4: a toolkit for performing basic set operations – union, intersection and complement on *k*-mer lists. GigaScience 4(1).

Korenberg J. R., Chen X. N., Schipper R., Sun Z., Gonsky R., Gerwehr S., Carpenter N., Daumer C., Dignan P., Disteche C. *et al.* (1994). Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences 91(11): 4997–5001.

Landy H. J., Weiner S., Corson S. L., Batzer F. R. (1986). The "vanishing twin": ultrasonographic assessment of fetal disappearance in the first trimester. American Journal of Obstetrics and Gynecology 155(1): 14–19.

Lench N., Barrett A., Fielding S. *et al.* (2013). The clinical implementation of non-invasive prenatal diagnosis for single-gene disorders: challenges and progress made. Prenatal Diagnosis 33: 555–62.

Lo Y. M. (2013). Non-invasive prenatal testing using massively parallel sequencing of maternal plasma DNA: from molecular karyotyping to fetal whole-genome sequencing. Reproductive BioMedicine Online 27: 593–98.

Lo Y. M., Chan K. C., Sun H. *et al.* (2010). Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Science Translational Medicine 2: 61–91.

Lo Y. M., Tein M. S., Lau T. K. *et al.* (1998). Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. American Journal of Human Genetics 62: 768–75.

Lo Y. M., Zhang J., Leung T. N. *et al.* (1999). Rapid clearance of fetal DNA from maternal plasma. American Journal of Human Genetics 64: 218–24.

Lo Y., Corbetta N., Chamberlain P., Rai V., Sargent I., Redman C. and Wainscoat J. (1997). Presence of fetal DNA in maternal plasma and serum. The Lancet 350(9076): 485–487.

Loane M., Morris J. K., Addor M. C. *et al.* (2013). Twenty-year trends in the prevalence of Down syndrome and other trisomies in Europe: impact of maternal age and prenatal screening. European Journal of Human Genetics 21: 27–33.

Lui Y. Y., Chik K. W., Chiu R. W., Ho C. Y., Lam C. W., Lo Y. M. (2002). Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. Clinical Chemistry 48(3): 421–7.

Lun F. M., Chiu R. W., Chan K. C. *et al.* (2008). Microfluidics digital PCR reveals a higher than expected fraction of fetal DNA in maternal plasma. Clinical Chemistry 54: 1664–72.

Marcais G., Kingsford C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27(6): 764–770.

Masuzaki H., Miura K., Yoshiura K. I. *et al.* (2004). Detection of cell free placental DNA in maternal plasma: direct evidence from three cases of confined placental mosaicism. Journal of Medical Genetics 41: 289–92.

Mujezinovic F., Alfirevic Z. (2007). Procedure-related complications of amniocentesis and chorionic villous sampling: a systematic review. Obstetrics and Gynecology 110: 687–94.

Muzzey D., Evans E., Lieber, C. (2015). Understanding the Basics of NGS: From Mechanism to Variant Calling. Current Genetic Medicine Reports 3(4): 158–165.

Norwitz E. R., Levy B. (2013). Noninvasive prenatal testing: the futuure is now. Reviews in Obstetrics and Gynecology 6:48–62.

Osborne C., Hardisty E., Devers P., Kaiser-Rogers K., Hayden M., Goodnight W., Vora N. (2013). Discordant noninvasive prenatal testing results in a patient subsequently diagnosed with metastatic disease. Prenatal Diagnosis 33(6): 609–611.

Patro R., Mount S., Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature Biotechnology 32(5): 462–464.

Pérez N., Gutierrez M., Vera, N. (2016). Computational Performance Assessment of *k*-mer Counting Algorithms. Journal of Computational Biology 23(4): 248–255.

Rijnders R. J., Van Der Luijt R. B., Peters E. D. *et al.* (2003). Earliest gestational age for fetal sexing in cell-free maternal plasma. Prenatal Diagnosis 23: 1042–44.

Russell L. M., Strike P., Browne C. E., Jacobs P. A. (2007). X chromosome loss and ageing. Cytogenetic and Genome Research 116(3): 181–185.

Sims D., Sudbery I., Ilott N., Heger A., Ponting C. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics 15(2): 121–132.

Skirton H., Patch C. (2013). Factors affecting the clinical use of non-invasive prenatal testing: a mixed methods systematic review. Prenatal Diagnosis 33: 532–41.

Taglauer E., Wilkins-Haug L., Bianchi D. (2014). Review: Cell-free fetal DNA in the maternal circulation as an indication of placental health and disease. Placenta 35: 64–68.

Wang E., Batey A., Struble C., Musci T., Song K., Oliphant A. (2013). Gestational age and maternal weight effects on fetal cell-free DNA in maternal plasma. Prenatal Diagnosis 33(7): 662–666.

Wang Y., Chen Y., Tian F. *et al.* (2014). Maternal mosaicism is a significant contributor to discordant sex chromosomal aneuploidies associated with noninvasive prenatal testing. Clinical Chemistry 60(1): 251–259.

Wong F., Lo Y. (2016). Prenatal Diagnosis Innovation: Genome Sequencing of Maternal Plasma. Annual Review of Medicine 67(1): 419–432.

Wood D., Salzberg S. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology 15(3): R46.

Zhao C., Tynan J., Ehrich M. *et al.* (2015). Detection of fetal subchromosomal abnormalities by sequencing circulating cell-free DNA from maternal plasma. Clinical Chemistry 61: 608–16.

VEEBILEHED

1) http://www.genome.umd.edu/jellyfish.html
2) http://sun.aei.polsl.pl/REFRESH/index.php?page=projects&project=kmc&subpage=about
3) http://www.gencodegenes.org/releases/19.html
4) https://www.python.org/
5) https://www.perl.org/
6) ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b147_GRCh37p13/VCF/
7) https://pyvcf.readthedocs.org/en/latest/
8) http://droog.gs.washington.edu/parc/images/iupac.html
9) http://www.illumina.com/science/education/sequencing-coverage.html

LISA 1

Tabel 7. Autosoomide *k*-meeride hulkade suurused, millega analüüsiti mitteinvasiivne sünnieelse testimise jaoks saadud sekveneerimisandmeid. "No poly" tähendab, et *k*-meeride hulgast ei ole eemaldatud polümorfsete positsioonidega ülekattes olevaid *k*-meere. Sulgudes olevad numbrid (0, 5, 10, 25) näitavad, mitme indiviidi täisgenoomi põhjal *k*-meeride hulka filtreeriti.

kromosoom / *whitelist*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
16-meerid no_poly (0)	27013995	28928255	23136044	20986120	20739411	19390736	17666873	17333508	13733835	16147758	15921517	15596453	11418751	10693714	9947809	10480322	9830715	9329747	7234023	8338966	4340123	4937528
20-meerid no_poly (0)	157978327	175371136	144811103	137629740	129442409	120427319	105708227	105324112	79899809	92218720	92737412	93308134	72652692	63712270	56328130	53032675	49504496	56360556	32134800	44195165	24715449	23081814
32-meerid no_poly (0)	186596902	205471525	170391564	162835872	152371110	141781056	125818669	123810534	94240346	108632668	109566328	111068195	85179482	75292907	66636012	63361496	59270504	65555019	40935081	51679426	29177677	27381556
16-meerid (0)	18806728	20990690	15974821	14362578	14312387	13364877	12188116	11913518	9494671	11155386	11025684	10830840	7841680	7404835	6911555	7307375	6942476	6419017	5155257	5815137	2985172	3460224
20-meerid (0)	122493026	143506969	111831980	105221367	99799660	92573803	80550801	79688937	60506943	70244458	70939463	72144880	55696694	48996353	43084970	39136175	38177643	43131870	24090129	33907877	18530683	17294188
32-meerid (0)	129287432	155286755	117444230	110809209	104757991	97087923	85261418	83349069	63283837	73480367	74538250	76596167	58156105	51553485	45343870	41343396	40715462	44637577	27130538	35303526	19356216	18147822
16-meerid (5)	10895087	11808417	8866150	7669797	7931707	7385467	6862466	6658132	5445811	6466143	6366525	6156737	4248799	4215226	4034026	4411268	4288636	3588082	3292778	3509140	1690702	2179924
20-meerid (5)	79762026	89799875	69662434	62812955	61833042	57137370	50529026	49949038	38978656	45693996	46172831	45895249	33444303	31187057	28307293	26680548	26559674	26935188	17462957	23199833	11545652	12518627
32-meerid (5)	88262964	102523278	77903610	71325742	69176174	63918642	56646654	55431642	42730828	50115964	50874682	51491829	37565549	34602680	31034997	28808225	28959743	29631487	19551089	24962983	12669413	13224963
16-meerid (10)	5605225	5759374	4200633	3421000	3749203	3484764	3364669	3200112	2758538	3291628	3267881	3061579	1937317	2113220	2085969	2452643	2454949	1712646	2040190	1904328	850279	1297981
20-meerid (10)	49787543	53498939	40910159	34625211	36003090	33134363	29976724	29426412	24006689	28422158	28733650	27758795	18683263	18986082	17940993	17882027	18184234	15784001	12757941	15380536	6946623	8991628
32-meerid (10)	62551743	70439263	53138257	46583988	46853631	43201493	38805290	37845245	30032910	35496750	35976254	35751411	24732892	24099740	22261505	21409964	21778870	20172038	15200796	18481507	8735662	10299363
16-meerid no_poly (10)	7389431	7272250	5562404	4560694	4968614	4625121	4464504	4271373	3651167	4376151	4318437	4038293	2576217	2797883	2758222	3244268	3209398	2280014	2651883	2520403	1134361	1714100
20-meerid no_poly (10)	61731867	63320830	50857491	43473523	44931150	41414356	37755484	37380017	30299615	35712402	35885166	34467837	23408110	23716271	22478957	23100366	22620573	19797922	16126931	19242062	8869442	11434889
32-meerid no_poly (10)	85497693	89557623	72978864	64729347	64689198	59724664	54115283	53298131	42106847	49436524	49741200	49024480	34261268	33301713	30869289	30799766	29913295	28003132	21325259	25581746	12418492	14539514
16-meerid (25)	443912	451812	322943	258212	288223	271921	265024	247572	215248	262195	260327	243361	148525	168300	165341	197359	200290	133806	172632	152097	66147	106997
20-meerid (25)	16685322	16965775	12633392	10015498	11034434	10103596	9516804	9174587	7910598	9419833	9599090	8975223	5500920	6203890	6015110	6459978	6811308	4883622	5216394	5462685	2253925	3550423
32-meerid (25)	29017277	31340699	23346662	19503587	20462807	18795903	17279169	16703275	13747261	16402207	16624360	16190266	10474011	10951113	10379136	10484079	10891091	8878379	7983998	8998275	3942941	5355566

LISA 2

Tabel 8. Töös kasutatud *Estonian Center of Excellence in Genomics and Translational Medicine, project No.* 2014-2020.4.01.15-0012 raames sekveneeritud Eesti populatsiooni täisgenoomid.

ID	Kasutatud ***k*-meeride hulkade filtreerimisel**
V00055	25
V00082	25
V00124	25
V00152	5, 10, 25
V00163	25
V03972	25
V04319	5, 10, 25
V09325	25
V10544	10, 25
V11595	10, 25
V13046	5, 10, 25
V15189	25
V16830	25
V18265	10, 25
V23198	10, 25
V23241	25
V24473	25
V24705	25
V25393	5, 10, 25
V33201	25
V33718	25
V34816	25
V35248	25
V39919	5, 10, 25
V40478	10, 25

LISA 3

Tabel 9. Töös kasutatud mitteinvasiivse sünnieelse testimise jaoks saadud sekveneerimisandmed.

ID	Positiivne Downi sündroomi suhtes
NIPT-NK-013-25183583	
NIPT-NK-014-25183584	
NIPT-NK-015-25183585	+
NIPT-NK-016-25183588	+
NIPT-NK-017-27216311	
NIPT-NK-018-27220288	
NIPT-NK-019-27209348	
NIPT-NK-020-27215326	
NIPT-NK-021-27208351	
NIPT-NK-022-27217324	
NIPT-NK-023-27228239	
NIPT-NK-024-27217325	
NIPT-NK-025-27228240	+
NIPT-NK-026-27226303	
NIPT-NK-027-27228241	
NIPT-NK-028-27217326	
NIPT-NK-029-30312286	
NIPT-NK-030-30318288	
NIPT-NK-031-30306282	
NIPT-NK-032-30310281	
NIPT-NK-033-30297269	
NIPT-NK-034-30318289	
NIPT-NK-035-30296267	
NIPT-NK-036-30304277	
NIPT-NK-037-30308283	

ID	Positiivne Downi sündroomi suhtes
NIPT-NK-038-30311282	
NIPT-NK-039-30317288	
NIPT-NK-040-30305277	
NIPT-NK-041-30296268	
NIPT-NK-042-30308284	
NIPT-NK-043-30315286	
NIPT-NK-044-30306283	
NIPT-NK-045-30313284	
NIPT-NK-046-30297270	
NIPT-NK-047-30305278	
NIPT-NK-048-30798827	
NIPT-NK-049-30800843	+
NIPT-NK-050-30801833	
NIPT-NK-051-30780224	
NIPT-NK-052-30800844	
NIPT-NK-053_2-32034149	
NIPT-NK-053-30801834	
NIPT-NK-054-30780225	
NIPT-NK-055-32032137	
NIPT-NK-056-32038129	
NIPT-NK-057-32052112	
NIPT-NK-058-32054098	
NIPT-NK-059-32044119	
NIPT-NK-060-32039153	
NIPT-NK-061-32056109	
NIPT-NK-062-32056111	

LIHTLITSENTS

Mina, Kaarel Koitne (sünnikuupäev: 19.09.1989)

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

Stabiilsete *k*-meeride hulga valimine mitteinvasiivseks sünnieelseks testimiseks,

mille juhendaja on Lauris Kaplinski,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace-i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 27.05.2016