TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Anette Maria Kuklane

Valimi suuruse mõju ekspressiooni kvantitatiivsete tunnuste lookuste täppiskaardistamisele lümfoblastoidrakuliinides

Bakalaureusetöö (9 EAP)

Juhendaja: Kaur Alasoo, PhD

Tartu 2021

Valimi suuruse mõju ekspressiooni kvantitatiivsete tunnuste lookuste täppiskaardistamisele lümfoblastoidrakuliinides

Lühikokkuvõte:

Inimese genoomis leiduvad kodeerivad osad, mida nimetatakse geenideks. Geenides olevate geneetiliste variantide järjestuste põhjal sünteesitakse geeniekspressiooni käigus valke, mis määravad ära organismi fenotüübi. Geeniekspressiooni mõjutavad geneetilised variandid asuvad ekspressiooni kvantitatiivsete tunnuste lookustes ja avaldavad valgusünteesi kaudu olulist mõju tunnuste, sealhulgas haiguste ja häirete, avaldumisele. Sageli on taoliste haigusi ja häireid põhjustavate variantide tuvastamine ning nende poolt mõjutatavatest geneetilistest mehhanismidest arusaamine variantide omavahelise korrelatsiooni tõttu keeruline. Kasutades täppiskaardistamist, on võimalik süstemaatiliselt hinnata geneetiliste variantide põhjuslikkuse tõenäosust, leides tunnuse signaalile vastavalt vähemalt ühte põhjuslikku varianti sisaldavad usaldusväärsete variantide hulgad. Senini ei ole aga uuritud valimi suuruse mõju täppiskaardistamise täpsusele põhjuslike variantide leidmisel. Siinse bakalaureusetöö eesmärk oli uurida, kuivõrd suuremate valimite kasutamine võimaldab tõsta statistilist usaldusväärsust põhjuslike variantide leidmisel lümfoblastoidrakuliinides. Töö käigus leiti, et valimi suurenedes muutub täppiskaardistamine täpsemaks, kuna suureneb nii statistiline usaldusväärsus põhjuslike variantide leidmisel kui ka statistiline võimsus põhjuslikke variante sisaldavaid usaldusväärsete variantide hulkasid tuvastada. Täpsemalt leiti suuremate valimite korral rohkem täppiskaardistatud usaldusväärsete variantide hulkasid, mis sageli sisaldasid väiksemal hulgal tõenäoliselt põhjuslikke variante.

Võtmesõnad:

Valimi suurus, geeniekspressioon, eQTL, täppiskaardistamine, LCL

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Effect of Sample Size on Fine Mapping Expression Quantitative Trait Loci in Lymphoblastoid Cell Lines

Abstract:

In the human genome, coding regions called genes can be found. The sequence of genetic variants in the genes encode for proteins that are synthesised during gene expression. In

turn, these proteins determine the phenotype of the organism. Genetic variants that affect gene expression are found in regions called expression quantitative trait loci and have a significant effect on the manifestation of traits, including diseases and disorders. Due to the correlation between the genetic variants associated with such diseases and disorders, identifying these variants and understanding the genetic mechanisms they affect remains a challenge. Using fine mapping, it is possible to systematically assess the probability of a genetic variant to be causal. This is done by finding sets of credible variants containing at least one causal variant according to the signal of the trait. Currently, however, the effect of sample size on the accuracy of fine mapping in finding causal variants has not been researched. The aim of this bachelor's thesis was to investigate the extent to which the use of a larger sample size increases the statistical power to find causal variants in lymphoblastoid cell lines. It was determined that larger sample sizes increase the resolution of fine mapping as both the statistical significance of finding causal variants and the statistical power to identify the number of credible sets containing causal variant increase. Specifically, increasing sample size led to the detection of more fine mapped credible sets, and these credible sets often contained a smaller set of putatively causal genetic variants.

Keywords:

Sample size, gene expression, eQTL, fine mapping, LCL

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Sisukord

1. Sissejuhatus

Kõik elusolendid koosnevad rakkudest, mis sisaldavad pärilikku informatsiooni, mida edastatakse vanematelt järglastele [1]. Enamikus organismides hoitakse seda teavet kromosoomideks kokku pakitud desoksüribonukleiinhappe (ingl *deoxyribonucleic acid, DNA*) molekulidena, mis sisaldavad juhiseid organismi toimimiseks [1]. DNA järjestatud osi, mille põhjal toodetakse geenieksprессiooni käigus ribonukleiinhapet (ingl *ribonucleic acid, RNA*) ja sellest omakorda valke, nimetatakse geenideks [1]. Valgud määravad organismi tunnuste avaldumise, kusjuures ühe tunnuse avaldumist võivad mõjutada erinevad geenid ja keskkonnategurid [1]. Erinevate geneetiliste järjestuste tulemuseks on erinevad tunnused [1], mistõttu võivad taolised erinevused suurendada mitmete haiguste ja häirete tekke riski.

Inimese genoomis on valke kodeerivaid geene umbes 20 000 [2, 3], mis moodustab vaid umbkaudu 1% tervest genoomist [3]. Neid genoomi osasid, kus asuvaid geneetilisi variante seostatakse sellise kvantitatiivse ehk mõõdetava tunnusega nagu geenieksprессioon, nimetatakse ekspressiooni kvantitatiivsete tunnuste lookusteks (ingl *expression quantitative trait loci*, e*QTL*) [4]. Lookuses asuvad geneetilised variandid võivad mõjutada nii transkriptsiooni kui ka translatsiooni – geenist transkribeeritud RNA kogust või RNA molekulide valgusünteesi määra [5]. Seetõttu on lookustel oluline seos kvantitatiivsete tunnustega ning nendega seonduvate haiguste ja häirete tekkega [4, 6, 7].

Kvantitatiivseid tunnuseid mõjutavate geneetiliste variantide analüüsimiseks ja tuvastamiseks kasutatakse täppiskaardistamist (ingl *fine mapping*) [7]. See võimaldab leida omavahel seotud ehk mittetasakaalulises ahelduses (ingl *linkage disequilibrium*, *LD*) olevate variantide hulgad ning seejärel kõige tõenäolisema põhjusliku variandi [7–9]. Põhjuslike variantide leidmise üheks piiravaks teguriks on eQTL-uuringutes kasutatava valimi suurus [4, 7, 9, 10]. Seniajani ei ole teada, täpselt kuidas valimi suurus variantide täppiskaardistamist mõjutab, kuid võib eeldada, et valimi kasvades muutuvad ka täppiskaardistamise tulemused täpsemaks.

Käesoleva bakalaureusetöö eesmärk on uurida, millisel määral mõjutab valimi suurus täppiskaardistamise tulemusi. See aitaks mõista, kuivõrd oleks suuremate valimite korral võimalik saavutada suuremat statistilist usaldusväärsust haigust tekitavate põhjuslike variantide leidmiseks ja nende eristamiseks LD-s olevate variantide hulgast. Selleks

analüüsitakse töö käigus 358 indiviidi proovidest koosnevat andmestikku valimi nelja erineva suuruse korral ja kolmest andmestikust kokku pandud andmestikku 966 indiviidi andmetega. Analüüsis kasutatakse vaid Euroopa päritolu indiviidide proove, et populatsioonidevaheliste erinevuste asemel hinnata selgemalt valimi suuruse mõju täppiskaardistamisele. Indiviide proovide põhjal on võimalik leida igale geenile vastavad, vähemalt ühte põhjuslikku varianti sisaldavad, usaldusväärsete variantide hulgad (ingl *credible sets*, *CS*) ning võrrelda hulkade suuruste ja hulgas sisalduvate variantide muutumist erinevate valimite puhul.

Bakalaureusetöö esimeses peatükis „Kirjanduse ülevaade" selgitatakse geeniekspressiooni, geneetilist variatsiooni, eQTL-ide olulisust ja mõju haigusriskile ning põhjuslike variantide leidmist täppiskaardistamise meetodil. Töö teises peatükis „Metoodika" kirjeldatakse tööprotsessis kasutatud andmestikke ja nende eeltöötlust, eQTL-ide täppiskaardistamiseks vajalikke töövooge ja ressursse. Viimaks analüüsitakse peatükis „Tulemused" täppiskaardistamise tulemusel leitud variantide hulkasid ja hulkades sisalduvate variantide statistilist olulisust erinevate valimite korral ning pakutakse välja võimalusi edasisteks uurimusteks.

2. Mõisted ja terminid

Aposterioorne kaasamise tõenäosus (ingl *posterior inclusion probability*, *PIP*) on tõenäosus, mis näitab, kuivõrd võib geneetiline variant olla põhjuslik ehk mõjutada tunnuse avaldumist [7].

Distaalne eQTL (ingl *trans-eQTL*) on geneetiline variant, mis avaldab mõju samal või erineval kromosoomil, kokkulepitud piirist (üldjuhul 1 megabaasist) kaugemal asuvale ühele või rohkemale kvantitatiivset tunnust mõjutavale geenile [11].

Ekspressiooni kvantitatiivse tunnuse lookus (ingl *expression quantitative trait loci*, *eQTL*) on piirkond genoomis, kus asuvad variandid mõjutavad tunnuse avaldumist geenikspressiooni kaudu [4].

Fenotüüp (ingl *phenotype*) on indiviidi vaadeldavate tunnuste kogum, mis sõltub nii indiviidi genotüübist kui ka keskkonnateguritest [1].

Geenikspressioon (ingl *gene expression*) on protsess, mille käigus toodetakse geneetilisest informatsioonist valke [1].

Genotüüp (ingl *genotype*) on indiviidi kõikide geneetiliste variantide kogum [1].

Haplotüüp (ingl *haplotype*) on genoomi mingis piirkonnas olev omavahel tugevalt korreleeritud SNP-de kogum, mis päritakse ühelt vanemalt [1].

Juhtvariant (ingl *lead SNP*) on geneetiline variant, mis on statistiliselt kõige olulisema mõjuga tunnuse avaldumisele [7].

Lokaalne eQTL (ingl *cis-eQTL*) on geneetiline variant, mis avaldab mõju lähedal, kokkulepitud kauguse (üldjuhul 1 megabaasi) piires asuvale ühele või rohkemale kvantitatiivset tunnust mõjutavale geenile [11].

Lümfoblastoidrakuliin (ingl *lymphoblastoid cell line*, *LCL*) on rakkude kogum, mis on saadud Epsteini-Barri viirusega nakatatud B-rakkudest [12].

Mittetasakaaluline aheldus (ingl *linkage disequilibrium*, *LD*) on nähtus, kus lähedal asuvad geneetilised variandid päranduvad koos edasi ja on seetõttu omavahel korrelatsioonis [7].

Täppiskaardistamine (ingl *fine mapping*) on protsess, mis võimaldab kindlaks määrata mingi tunnuse avaldumist mõjutava ühe või rohkema geneetilise variandi [7].

Usaldusväärsete variantide hulk (ingl *credible set, CS*) on hulk, mis sisaldab vähemalt ühte konkreetse tunnuse avaldumisega seotud põhjuslikku varianti [13].

Üksiknukleotiidi polümorfism (ingl *single nucleotide polymorphism*, *SNP*) on mutatsiooni tagajärjel tekkinud ühe nukleotiidi vahetus genoomis, mis on populatsioonis edasi kandunud [1].

Üksikute mõjude summa (ingl *sum of single effects*, *SuSiE*) on meetod geneetiliste variantide valimiseks, mis võimaldab luua mingi tunnuse signaali seletavate variantide hulgad [13].

3. Kirjanduse ülevaade

Järgnevas peatükis kirjeldatakse geenide ja geeniekspressiooni funktsiooni ja olulisust. Lisaks selgitatakse geneetilise variatsiooni ja geeniekspressiooni mõjutavate eQTL-ide seost haigusi ja häireid põhjustavate geneetiliste variantidega. Viimaks antakse ülevaade põhjuslike variantide leidmisest, kasutades täppiskaardistamise meetodit, ja seda mõjutavatest teguritest.

3.1 Geenid ja geeniekspressioon

Kõik organismid on üles ehitatud rakkudest, mis sisaldavad geneetilist informatsiooni, võimaldades rakul funktsioneerida ja organismi pärilikku informatsiooni järglastele edasi anda [1]. Eukarüootide, sealhulgas inimese organismis asub see informatsioon rakutuumas DNA molekulidena [1]. DNA koosneb kahest lineaarsest ahelast, mille alamüksusteks on nukleotiidid – adeniin (A), tsütosiin (C), guaniin (G) ja tümiin (T) [1].

Inimorganismi igas rakus on umbkaudu 3 miljardit nukleotiidipaari [2, 3], mis kokku moodustavad inimese genoomi [1]. Genoomi suuruse tõttu on DNA enamjaolt kondenseerunud valgumolekulide ümber kompaktseks struktuuriks, mida kutsutakse kromosoomideks [1]. Inimeste kromosoomide arv paarikaupa on 23, mistõttu pärib järglane vanematelt sama geeni kaks varianti ehk alleeli [1]. Organismi kõikide alleelide kogumit nimetatakse genotüübiks, mis koos keskkonnateguritega mõjutab organismi fenotüüpi ehk vaadeldavaid tunnuseid [1], sh haigusi ja häireid.

DNA võib jagada kaheks – kodeeriv ja mittekodeeriv DNA [1]. Kodeerivaks DNA-ks on geenid – järjestatud DNA osad, mis hoiavad valkude sünteesimiseks vajalikku informatsiooni valkude struktuuri kohta, mis omakorda määrab ära mingi tunnuse [1]. Seevastu mittekodeeriv DNA sisaldab regulatoorseid elemente, mis mõjutavad DNA-st valkude tootmist ehk geeniekspressiooni ja moodustab genoomist 99% [3].

Geeniekspressioon koosneb kahest protsessist – transkriptsioon ja translatsioon [1]. Transkriptsiooni ajal tehakse ühest DNA ahelast koopia – sõnumi-RNA (ingl *messenger RNA*, *mRNA*) ahel. RNA, sarnaselt DNA-le, on geneetilise informatsiooni kandjaks, kuid erineb selle poolest, et RNA suhkruks on riboos ja ta sisaldab tümiini (T) asemel uratsiili (U) [1]. Seejärel luuakse mRNA informatsiooni põhjal translatsiooni käigus polüpetiidahel [1].

Valgud koosnevad ühest või mitmest polüpeptiidahelast, moodustades spetsiifilise kolmemõõtmelise struktuuri [14]. Inimese organismis leidub kümneid tuhandeid valke, millest sõltub organismi funktsioneerimine. Näiteks on valkude ülesandeks organismi kaitsmine (antikehad), keemiliste reaktsioonide kiirendamine (ensüümid) ja ainete transportimine (hemoglobiin) [14]. Seetõttu on oluline, et organism suudaks toota õiges koguses funktsionaalseid valke.

3.2 Geneetiline variatsioon

Ükskõik millise kahe inimese genoom sarnaneb umbes 99,9% ulatuses [2]. Genoomidevaheliste erinevuste ehk geneetilise variatsiooni peamisteks põhjusteks on rekombinatsioon ja mutatsioon [1, 15], millest viimane esineb kõige sagedamini üksiknukleotiidi polümorfismina (ingl *single nucleotide polymorphism*, *SNP*) [1, 5, 16]. SNP-d on nukleotiididevahelised erinevused mingites genoomi piirkondades [1, 5]. Ilmnedes genoomis üldiselt kord 100–300 aluspaari kohta, moodustavad need 90% geneetilisest variatsioonist [5]. SNP-d tekivad üksiknukleotiidi mutatsiooni tagajärjel ja kanduvad populatsioonis vanematelt järglastele [1].

Genoomis asuvad piirkonnad, kus rekombinatsioon toimub tavapärasest sagedamini ning mida kutsutakse genoomi kuumadeks punktideks [1]. Nende punktide vahel leiduvad ahelduspiirkonnad, kus DNA järjestused rekombinatsiooni tagajärjel ei muutu ja päranduvad koos edasi [1]. Ahelduspiirkonnas olevate kindlate SNP-de kombinatsiooni nimetatakse haplotüübiks (ingl *haplotype*), kusjuures ühes haplotüübiplokis varieerub SNP-de arv mõnest mõnekümneni [1]. Haplotüübiplokke on inimesel teada üle 200 000 ja nende kombinatsioonid võivad populatsioonis esineda suuremal või vähemal määral [1]. Seega võib mingit tunnust põhjustava geneetilise variandi leidmine olla ebatäpsem, kuna ühes haplotüübiplokis olevad SNP-d on LD-s ehk omavahelises mittejuhuslikus seoses.

Enamjaolt esinevad haigustega seotud SNP-d mittekodeerivas DNA-s, kuid neid leidub ka geenides ehk kodeeriva DNA lõikudes [5, 16]. Paljusid SNP-sid loetakse neutraalseteks, kuna need ei avalda mõju inimese bioloogilistele funktsioonidele [16]. Seevastu funktsionaalsed SNP-d, mis asuvad geeni lähedal või sees, mõjutavad valgusünteesi tulemust [16]. Viimast mõjutavad SNP-d võivad muuta nii polüpeptiidahela järjestust kui ka geenieksressiooni määra [5]. Põhjustades varieeruvust fenotüübis, seostatakse selliseid geneetilisi erinevusi mitmete haiguste ja häiretega [1, 4, 16]. Nende hulka kuuluvad näiteks

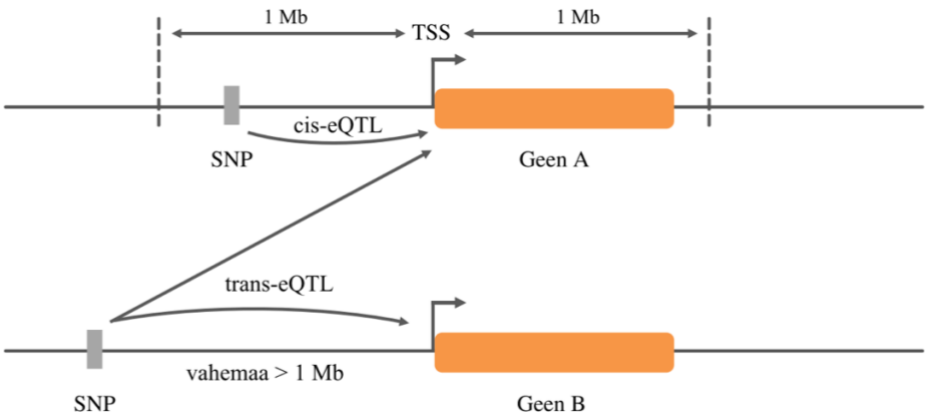
hemofiilia [1], Alzheimeri tõbi, skisofreenia ning bipolaarne häire [10]. Näiteks Alzheimeri tõve puhul on üheks põhjuslikuks SNP-ks rs117618017, mille tõttu suureneb geeni APH1B ekspressioon ja seetõttu ka haigusrisk [10].

3.3 Ekspressiooni kvantitatiivsete tunnuste lookused

Kvantitatiivseteks tunnusteks nimetatakse tunnuseid, mida on võimalik kvantitatiivselt hinnata – mõõta ja kaaluda [1]. Nende tunnuste avaldumine ja varieeruvus sõltub mitmest geenist ning geenide ja keskkonnafaktorite koosmõjust [6]. Sellised tunnused on näiteks inimese pikkus, kaal [1, 6], ensüümide aktiivsus [1] ja vererõhk [6].

Kvantitatiivse tunnuse lookuse all mõeldakse alleelide kogumit DNAs, mis mõjutab variatsiooni mingi konkreetse kvantitatiivse tunnuse puhul [6]. Kui see lookus mõjutab variatsiooni geeniekspressioonis (mRNA ekspressioonitaset), siis kutsutakse seda eQTL-iks [11]. Enamjaolt seostatakse neid geeniekspressiooni mõjutavaid geneetilisi variante SNP-dega.

Enamik eQTL-idest asuvad mittekodeerivas DNA-s ja võivad mõjutada geeniregulatsiooni [9]. Need variandid saab, sõltuvalt variandi kaugusest ja eeldatavast mõjust geenidele, jagada kaheks – lokaalsed ja distaalsed [4, 11]. Lokaalsed eQTL-id (*cis-eQTLs*) mõjuvad lähedal olevatele geenidele, asudes üldiselt transkriptsiooni alguskohtade (ingl *transcription start site*, *TSS*) juures [4, 11]. Seevastu distaalsed eQTL-id (*trans-eQTLs*) võivad mõju avaldada nii samal kromosoomil kaugemal asuvatele geenidele kui ka geenidele, mis asuvad teistel kromosoomidel [11]. Peamiselt on tuvastatud cis-eQTL-e, sest trans-eQTL-ide leidmiseks on SNP-geenide vahelisi assotsiatsioone vaja testida palju suuremal määral ja see eeldab suurema valimi olemasolu [4, 11]. Cis- ja trans-eQTL-e on kujutatud joonisel 1, kus cis-eQTL on geeni lähedal, kokkuleppeliselt kuni 1 megabaasi (ingl *megabase*, *Mb*) ehk 1 000 000 aluspaari kaugusel geeni TSS-ist ja trans-eQTL mõjutab geeni A ja geeni B kaugemalt kui 1 Mb.



Joonis 1. Geeniekspressiooni mõjutavad lokaalsed ja distaalsed geneetilised variandid.

Genoomiülene assotsiatsiooniuuring (ingl *Genome-wide association studies*, *GWAS*) võimaldab vaadata üle terve genoomi geneetilisi variante, testides iga varianti ja selle statistilist seost mingi konkreetse haiguse riskiga [17]. Siiski ei anna ainuüksi variandi teadmine piisavat informatsiooni geneetilistest mehhanismidest, mis kujundavad haigusriski. eQTL-analüüs sarnaneb GWAS-iga, kuid vaadeldavaks tunnuseks ei ole mitte haigus, vaid geeniekspressioon. Integreerides GWAS- ja eQTL-analüüsi, on võimalik hinnata, kas konkreetse geeni ekspressioon võiks põhjuslikult mõjutada haiguse riski.

Analüüsi tulemusel leitakse sageli mitu haigusega seotud varianti, mis on omavahel LD kaudu seotud [18]. Seejuures võib erinevate variantide omavaheline korrelatsioon olla suurem kui 0,99 [13], mistõttu on analüüside tulemuseks leitud variante sisaldav lookus, mitte tingimata üks konkreetne variant. Põhjusliku variandi (või põhjuslike variantide) leidmiseks on seetõttu vajalik järgnev analüüs – täppiskaardistamine.

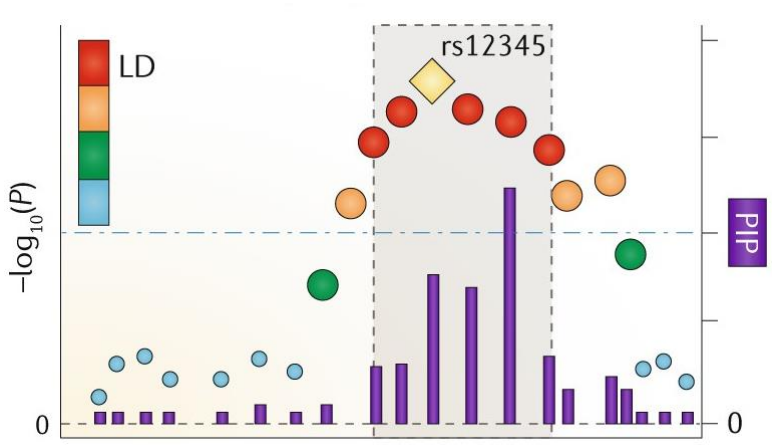
3.4 Täppiskaardistamine

Täppiskaardistamise eesmärk on kindlaks määrata tunnuse avaldumisega seotud üks või mitu geneetilist varianti [7–9, 13, 18]. Eelnevate analüüside – GWAS-i või eQTL-i – tulemusel leitakse SNP-d, mis on statistiliselt kõige olulisema mõjuga tunnuse avaldumisele ehk väikseima p-väärtusega [7, 18]. Leitud SNP-sid nimetatakse juhtvariantideks (ingl *lead SNPs*), mis sellest olenemata ei pruugi osutuda põhjuslikuks variandiks [7–9].

Geneetiliste variantide täppiskaardistamine Bayesi mitmese lineaarse regressiooni üksikute mõjude summa (ingl *sum of single effects*, *SuSiE*) meetodil leitakse juhtvariantide lookustes usaldusväärsete variantide hulgad ehk CS-id [13]. Mudel arvestab ja analüüsib eraldi juhtvariantide juures olevaid geene ja geneetiliste variantide omavahelist korrelatsiooni

eesmärgiga hinnata, milline variantide kombinatsioon seletab kõige paremini tunnuse avaldumise signaali ehk on kõige tõenäolisemalt põhjuslik [13].

CS-id luuakse variantide sorteerimisel vastavalt iga variandi aposterioorse kaasamise tõenäosusele (ingl *posterior inclusion probability*, *PIP*), lisades variante hulka senikaua, kuni tõenäosuste kumulatiivne summa ületab mingi lävendi [7, 8, 18]. Seejuures luuakse hulgad võimalikult väikesed, kuid sellised, mis sisaldavad vähemalt ühte põhjuslikku varianti [7, 13]. Joonis 2 kujutab väikseima p-väärtusega varianti ehk juhtvarianti (rs12345), sellega LD-s olevaid variante ning nendest variantidest koosnevat CS-i.



Joonis 2. Bayesi täppiskaardistamise [7] meetodil juhtvariandi (rs12345) lookuses loodud CS, kuhu kuuluvad variandid on esitatud hallil taustal. Joonise x-telg kujutab lookuses olevaid variante vastavalt positsioonile ja vasak y-telg variantide p-väärtusi, mida on kujutatud joonisel punktidena. Parem y-telg tähistab variantidele täppiskaardistamisel leitud PIP väärtusi, kus igale variandile vastab PIP väärtust kujutav tulp. Juhtvariandi ja selle lookuses asuvate variantide LD tugevust on kujutatud värviskaalal punasest siniseni, kus punane tähistab tugevat LD-d ja sinine vastavalt nõrka LD-d. Samuti on siin oluline välja tuua, et suurim PIP väärtus ei ole mitte suurima p-väärtusega variandil ehk juhtvariandil (rs12345), vaid sellega tugevas LD-s oleval variandil.

SuSiE puhul on oluline välja tuua, et mudel kasutab otsimisstrateegiana iteratiivset Bayesi astmelist valikut (ingl *Iterative Bayesian Stepwise Selection*, *IBSS*) [13], kus variantide valikut ja täppiskaardistamise tulemusi mõjutavad nii variantide arv vaadeldavas piirkonnas, põhjuslike variantide arv ja mõju tunnusele [7], variantidevaheline LD [7, 10] kui ka valimi suurus [4, 7, 9, 10].

Olgugi et mitmed eQTL-uuringud [4, 7, 9, 10] toovad välja valimi suuruse olulisuse, on siiski ebaselge, kui suurel määral omab see piiravat mõju täppiskaardistamise tulemuste

täpsusele. Seda on võimalik uurida, kasutades andmestikke, kus geneetilistel variantidel on statistiliselt oluline mõju geenieksпressioonile, teisisõnu statistiliselt olulised efektisuurused, ja variantide vahel eksisteerivad LD mustrid. Võrreldes tulemuseks saadud geenidele vastavaid CS-e, on võimalik välja selgitada, kas hulgad lähevad valimi suurenedes väiksemaks ja tuvastatud hulkade arv kasvab.

4. Metoodika

Järgnevas peatükis kirjeldatakse kasutatud andmestikke ja nende eeltöötlust, samuti antakse ülevaade põhjuslike variantide leidmiseks kasutatud töövoogudest ja ressurssidest.

4.1 Andmestikud

Bakalaureusetöös kasutatud andmestikest – GEUVADIS [19], GENCORD [20] ja TwinsUK [21] – on valitud indiviidide andmed, mis pärinevad lümfoblastoidrakuliinist (ingl *lymphoblastoid cell line*, *LCL*) ehk Epsteini-Barri viirusega (EBV) nakatunud B-rakkudest [12]. Viirusega nakatunud rakud immortaliseeruvad, võimaldades neid kasutada rakukultuurimudelina mitmes uuringus [12]. Min jt [22] toob välja, et LCL on kergesti ligipääsetav allikas üksikute rakkude saamiseks, mida keskkonnast või muudest variatsioonidest põhjustatud tegurid oluliselt ei mõjuta. Seetõttu väidab autor, et eQTL-uuringutes kasutatud LCL-idel on oluline mõju komplekssete tunnuste geneetilise regulatsiooni uurimisele. Tunnuseid loetakse kompleksseteks, kui nende avaldumist mõjutavad mitmed geenid ja keskkonnategurid ning kuhu hulka kuuluvad ka kvantitatiivsed tunnused [7].

Projektis GEUVADIS (Genetic European Variation in Disease) analüüsisid Lappalainen jt [23] 462 indiviidi LCL-idest pärit RNA andmeid 1000 genoomi projektist (ingl *1000 Genomes Project*). Andmed pärinevad viiest erinevast populatsioonist: CEPH (Põhja- ja Lääne-Euroopa päritolu Utahi populatsioon, CEU), Soome (FIN), Suurbritannia (GBR), Toscani (Itaalia, Toscana piirkonna populatsioon, TSI) ja Yoruba (Nigeeria, Ibadani piirkonna populatsioon, YRI). Töös kasutatava Euroopa alamhulga (EUR), kuhu kuulub 373 indiviidi, moodustavad eelmainitud populatsioonidest kõik peale viimase, mis on osa Aafrika (AFR) alamhulgast.

Töös kasutatakse ka Gutierrez-Arcelusi jt [24] poolt uuritud andmestikku GENCORD (Geneva Cord), mis koosneb Kesk-Euroopa päritolu indiviidide andmetest. Täpsemalt analüüsiti 204 vastsündinu nabanöörist saadud fibroblastide, T-rakkude ja LCL-ide proove [24]. Kokku sisaldab andmestik 567 proovi, millest 192 on töös kasutatavad LCL-id.

Lisaks kasutatakse töös Buili jt [25] poolt analüüsitud TwinsUK andmestikust pärineva 856 Euroopa päritolu indiviidi RNA andmeid. Andmestikus leidus 2505 proovi neljast koest: rasvkoest, epiteelkoest, verest ja LCL-ist, millest viimast esines 764 korral.

4.2 Andmestike eeltöötlus

Siinse töö eesmärk oli hinnata valimi suuruse mõju eQTL-ide täppiskaardistamisele. Selleks kasutati 100, 200, 300 indiviidi proovidest koosnevaid juhuslikult valitud alamhulkasid, et vaadelda täppiskaardistamise tulemusi ka väiksemate valimite puhul, ja 358 Euroopa päritolu indiviidi hulka andmestikust GEUVADIS. Lisaks neljale eelmainitud indiviidide hulgale kasutati ka 966-st Euroopa päritolu indiviidist koosnevat valimit, mille andmed on kokku pandud GEUVADIS-e, GENCORD-i ja TwinsUK andmestikest. Igale andmestikule vastab indiviidide genotüüpe sisaldav fail, metaandmefail ning geenieksрessioonimaatriksist koosnev fail.

Geenijärjestuse variatsioone hoitakse VCF (ingl *Variant Call Format*) formaadiga failis, mis sisaldab teavet proovide genotüübi kohta vastavalt konkreetsele positsioonile genoomis [26]. Täpsemalt vastavad need positsioonidele, kus indiviidide genoomid erinevad [26], kusjuures variatsiooni tuvastamiseks kasutatakse viitegenoomi GRCh38 [27]. Tööriista BCFtools [28] abil moodustati andmestike ühine VCF-fail, kus sisaldusid kõigi andmestike proovid vastavalt viitegenoomi tuvastatud geneetilise variatsiooni positsioonidele genoomis ning kust eemaldati puuduvad väärtused ja mitmealleelsed variandid.

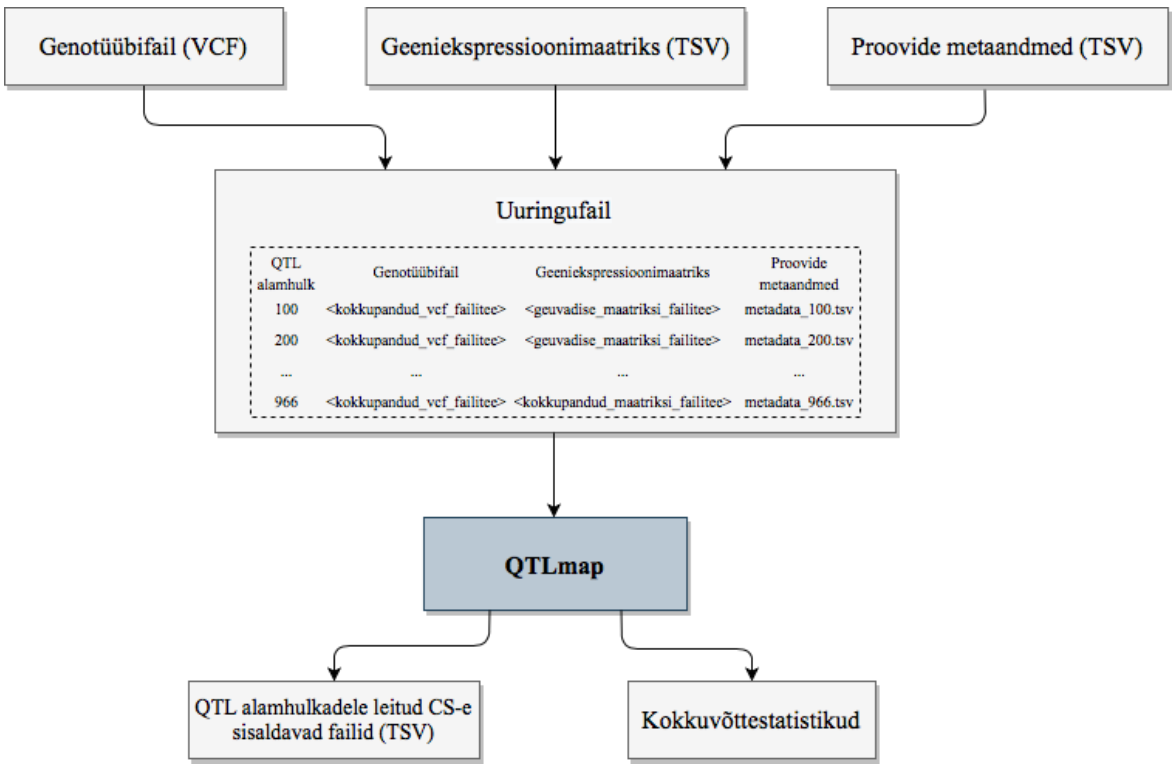
Metaandmetest koosnev fail sisaldab muuhulgas nii indiviidide proovi ja genotüübi unikaalset identifikaatorit, indiviidi sugu kui ka üldjuhul rakutüüpi või kude, millest proov võeti. Failide töötlemisel valiti GEUVADIS-e andmestiku puhul indiviidid, kes vastasid ülempopulatsioonikoodile „EUR” ja kelle proov oli läbinud nii RNA kui ka genotüübiga seotud kvaliteedikontrolli [29]. Andmestikest GENCORD ja TwinsUK filtreeriti välja indiviidid, kelle proovid vastasid LCL koetüübile ning olid samuti läbinud RNA ja genotüübi kvaliteedikontrolli. Töö käigus koostati igale valimi suurusele vastav metaandmete fail.

Geenieksрessioonimaatriksis on igale geenile olemas geenieksрessiooni normaliseeritud väärtus kõikides proovides. Faili töödeldi vaid kolmest andmestikust koosneva valimi jaoks, pannes kokku kõigi andmestike maatriksid. Selle käigus jäeti alles kõik proovid, mis vastasid metaandmete failis olevatele proovidele. Täpsemalt valiti välja proovid, mis vastasid LCL koetüübist võetud ja kvaliteedikontrolli läbinud Euroopa päritolu indiviidide proovidele.

4.3 Töövood ja ressursid

Nextflow [30] on raamistik, mida kasutades saab protsesse jooksutada ühe töövoona (ingl *workflow*), võimaldades sealjuures paralleelsust, teisaldatavust ning tööprotsessi korratavust [30, 31]. Siinses töös kasutati Nextflow raamistikku töövoo QTLmap [32] osana.

Kerimov jt [29] töötasid välja ühise eQTL kataloogi erinevatest eQTL-uuringutest, mille käigus arendati välja kataloogi kuuluvate andmestike töötlemiseks töövoog QTLmap. Viimane võimaldab leida statistiliselt olulisi seoseid tunnuste ja geneetiliste variantide vahel ehk QTL-e kaardistada [29, 31, 32]. Käesolevas bakalaureusetöös rakendati QTLmapi SuSiE täppiskaardistamise meetodit, mis andis väljundina iga analüüsitud geeni kohta põhjuslikku varianti sisaldava CS-i. Joonisel 3 on esitatud lihtsustatud kujul töövoo QTLmap sisendid ja väljundid.



Joonis 3. QTLmapi poolt sisendina saadud uuringufail, kus kirjeteks on valimi suurusele vastavad failid ja väljundiks valimitele vastavad CS-e sisaldavad failid ning kokkuvõttestatistikud.

Töövoogude jooksutamiseks kasutati Tartu Ülikooli teadusarvutuste keskuse ressursse. Teadusarvutuste keskuse ja GitHub koodivaramu kaudu oli olemas ka ligipääs valimi suuruse mõju analüüsimiseks kasutatud andmestikele GEUVADIS, GENCORD ja TwinsUK. Töö käigus valminud kood, QTLmapi töövoo leitud tulemusfailid ja

genotüübifailide töötlemiseks kirjutatud skript on avalikult kättesaadav GitHub-i koodivaramus akuklane/bakalaureusetoo [33]. Kood on kirjutatud valdavalt programmeerimiskeeles Python, kuid üksikute geenide CS-ide muutumist erinevate valimi suuruste juures analüüsiti keeles R.

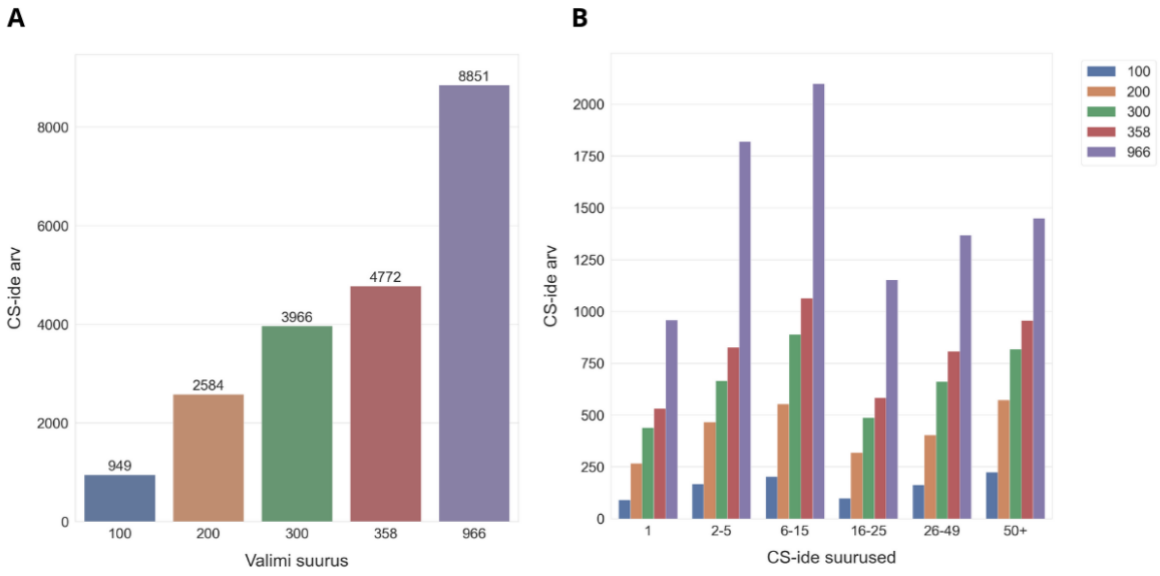
5. Tulemused

Käesolevas peatükis antakse ülevaade täppiskaardistamise tulemustest. Täpsemalt analüüsitakse CS-ide koguarvu ja suuruse muutumist erinevate valimi suuruste puhul ning teostatakse ülekattuvate hulkade võrdlemiseks komponentide analüüs. Seejärel pakutakse välja võimalusi edasiseks uurimiseks.

5.1 Töö tulemuste analüüs

QTLmapi SuSiE täppiskaardistamine leidis igale geenile vastavad CS-id GEUVADIS-e andmestiku alamhulkadele 100, 200, 300 ja 358 ning GEUVADIS-e, GENCORD-i ja TwinsUK andmetest kokku pandud andmestikule 966 indiviidiga.

QTLmapi tulemustest võib näha, et valimi kasvades suurenes CS-ide arv. Näiteks võrreldes valimeid suurusega 100 ja 966, kasvas CS-ide arv peaaegu kümme korda, kusjuures hulkade arvud suurenesid erinevates suurusvahemikes ühtlasel määral. Geenieksprессioon on kvantitatiivne tunnus, mis on sageli mõjutatud mitme väikse efektisuurusega eQTL-ist. Mida rohkem on tunnust mõjutavaid lookuseid, seda keerulisem on neid kõiki kaardistada. Selline hulkade arvu suurenemine näitab võimekuse kasvu tuvastada, eriti väiksemate efektisuurustega, lookuseid, mis mõjutavad geene ja seeläbi ka tunnuste avaldumist. Joonisel 4**A** on esitatud CS-ide koguarv erinevatel valimi suurustel; joonisel 4**B** on kujutatud kindla suurusega või suurusvahemikes olevate CS-ide koguarvu erinevate valimite puhul.



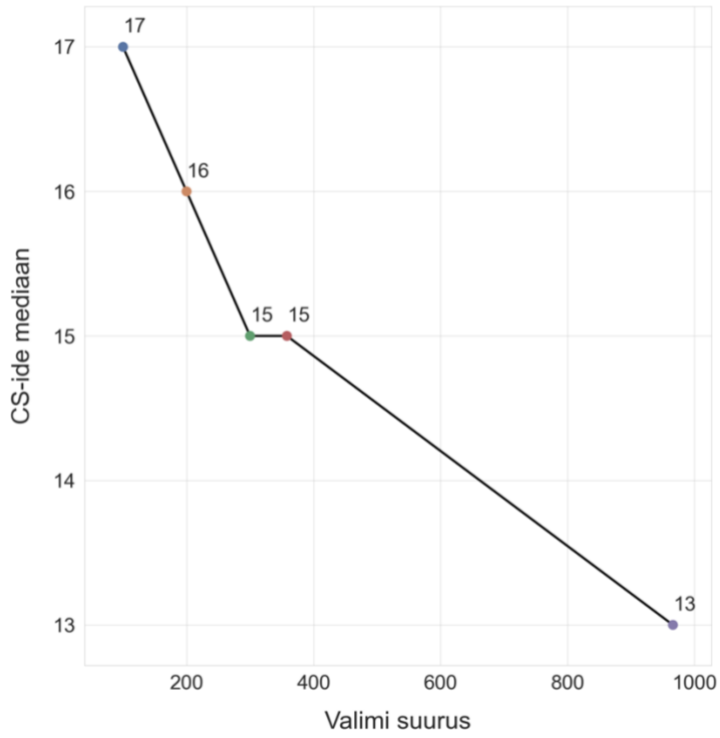
Joonis 4. **A** CS-ide koguarv erinevate valimi suuruste korral tulpdiagrammina. **B** Kindla suurusega (1) või suurusvahemikes (2–5, 6–15, 16–25, 26–49, 50+) olevate CS-ide arv erinevate valimi suuruste korral tulpdiagrammina.

Väiksemate efektisuurustega eQTL-ide tuvastamise kasvu kinnitab ka geenide arvu suurenemine, millele leidus valimi kasvades rohkem kui üks CS. Kuna iga hulk sisaldab põhjuslikku varianti, siis järeldub, et kasvas geenide arv, mida mõjutab rohkem kui üks variant. Kusjuures 966 indiviidi andmetest koosneva valimi puhul leidus ka geene, millele vastas koguni kuus kuni kaheksa CS-i. Täpsemad tulemused on välja toodud tabelis 1, kus veerud L1–L8 on geeni erinevad CS-id vastavalt tuvastamise järjekorrale. Taoline kasvutrend viitab võimalusele, et geeniekspressiooni väiksemal määral mõjutavate eQTL-ides olevate variantide tuvastamine ei ole veel haripunkti lähedale jõudnud. Seega võib leiduda veel mitmeid geene, mis on mõjutatud erinevates lookuses olevate variantide poolt, kuid mille tuvastamine on piiratud suuremate valimite kättesaadavuse tõttu.

Tabel 1. Erinevatel valimi suurustel geenidele leitud erinevate CS-ide (L1–L8) koguarvud.

	L1	**L2**	**L3**	**L4**	**L5**	**L6**	**L7**	**L8**
100	871	75	3	0	0	0	0	0
200	2316	242	22	3	1	0	0	0
300	3449	460	50	7	0	0	0	0
358	4096	588	70	16	2	0	0	0
966	6855	1602	289	70	19	7	6	3

Lisaks statistilise võimsuse kasvule tuvastada geeniekspressiooni mõjutavaid variantide hulkasid, kasvas suuremate valimite juures võimsus põhjuslikke variante eristada teistest hulgas olevatest variantidest, mis on LD tõttu seotud. Võrreldes valimi suurusi 100 ja 966, vähenes CS-ide mediaan 17 variandilt 13 variandini. See annab alust oletada, et suuremad valimid vähendavad korrelatsiooni tõttu tekkinud müra, parandades seeläbi täppiskaardistamise täpsust. Siiski on oluline arvesse võtta, et välja toodud hulkade mediaani vähenemine võib olla seotud uute CS-ide leidmisega ega pruugi tegelikkuses peegeldada ühe hulga suuruse muutumist erinevate valimi suuruste puhul. CS-ide mediaani vähenemist valimi suurenedes esitab joonis 5.



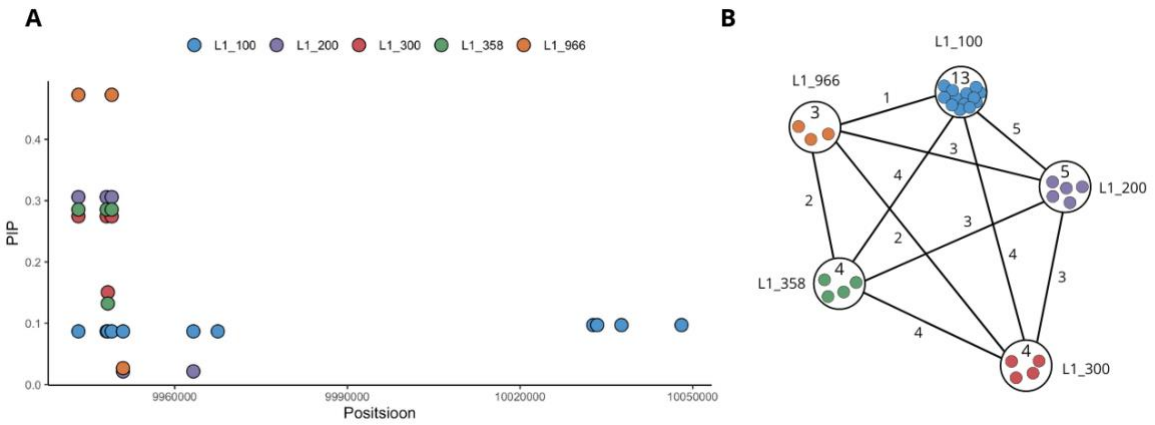
Joonis 5. CS-ide mediaan erinevatel valimi suurustel joondiagrammina.

Olgugi et QTLmapi poolt leitud CS-ide analüüsimine annab ülevaate kõikide geenide hulkade muutumisest erinevate valimi suuruste puhul, ei võimalda see näha tendentsi hulkade suuruste muutumises üksiku hulga tasandil. Selle jaoks tehti komponentide analüüs, mille käigus moodustati komponendid paarikaupa ülekattuvatest CS-idest. Analüüsides konkreetse geeni komponentides olevate CS-ide muutumist, on võimalik leida, kas ühes hulgas sisalduvate variantide arv väheneb valimi suuruse kasvades.

Ülekattes olevate hulkade kindlaksmääramiseks, leiti PyRangesi [34] paketi abil paarikaupa ülekattuvad variandid ehk CS-ides olevad variantide paarid vastavalt nende variantide positsioonile genoomis. Kuna siinne töö analüüsib eraldi iga geeni hulkade muutumist, jäeti ülekattuvatest paaridest alles vaid seosed sama geeni hulkades olevate variantide vahel. Selle tulemusel jäi 58412 ülekattes olevast paarist alles 42565 paari. See võimaldas leida ülekattes olevad hulgad, kus geeni ühe CS-is olevatest variantidest vähemalt üks sisaldus teises hulgas.

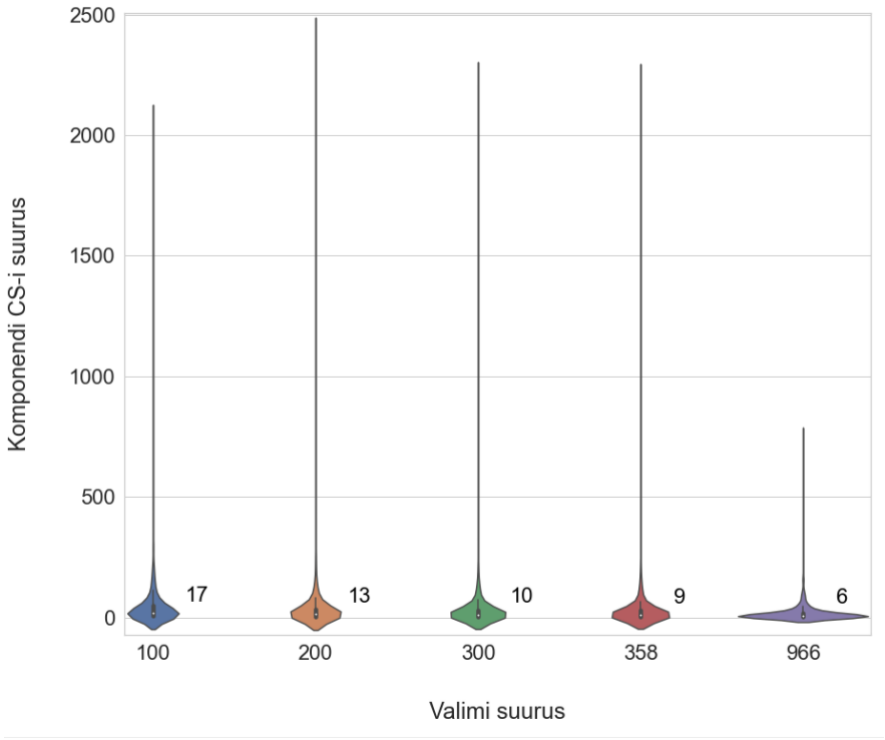
Käsitledes osaliselt või täielikult ülekattuvaid hulkade paare kui graafi tippe ja ülekattuvusi kui servi, võimaldas paketi python-igraph [35] kasutamine visualiseerida hulkade ülekattuvusi 42565 tipust ja 21122 servast koosneva graafina. Täpsemalt oli tegemist mittesidusa graafiga, mis koosnes 9752 komponendist. Igale geenile võis vastata üks või

rohkem komponenti, kusjuures komponendid võisid koosneda erinevatest CS-idest (L1–L8). Näiteks geenile LZIC vastas üks komponent, mis koosnes ühest CS-ist (L1) erinevatel valimi suurustel. Joonise 6 **A**-osas on kujutatud geeni LZIC CS-is (L1) olevaid variante erinevatel valimi suurustel ja **B**-osas valimi suurustele vastavatest hulkadest PyRanges ja python-igraph abil moodustatud komponenti.



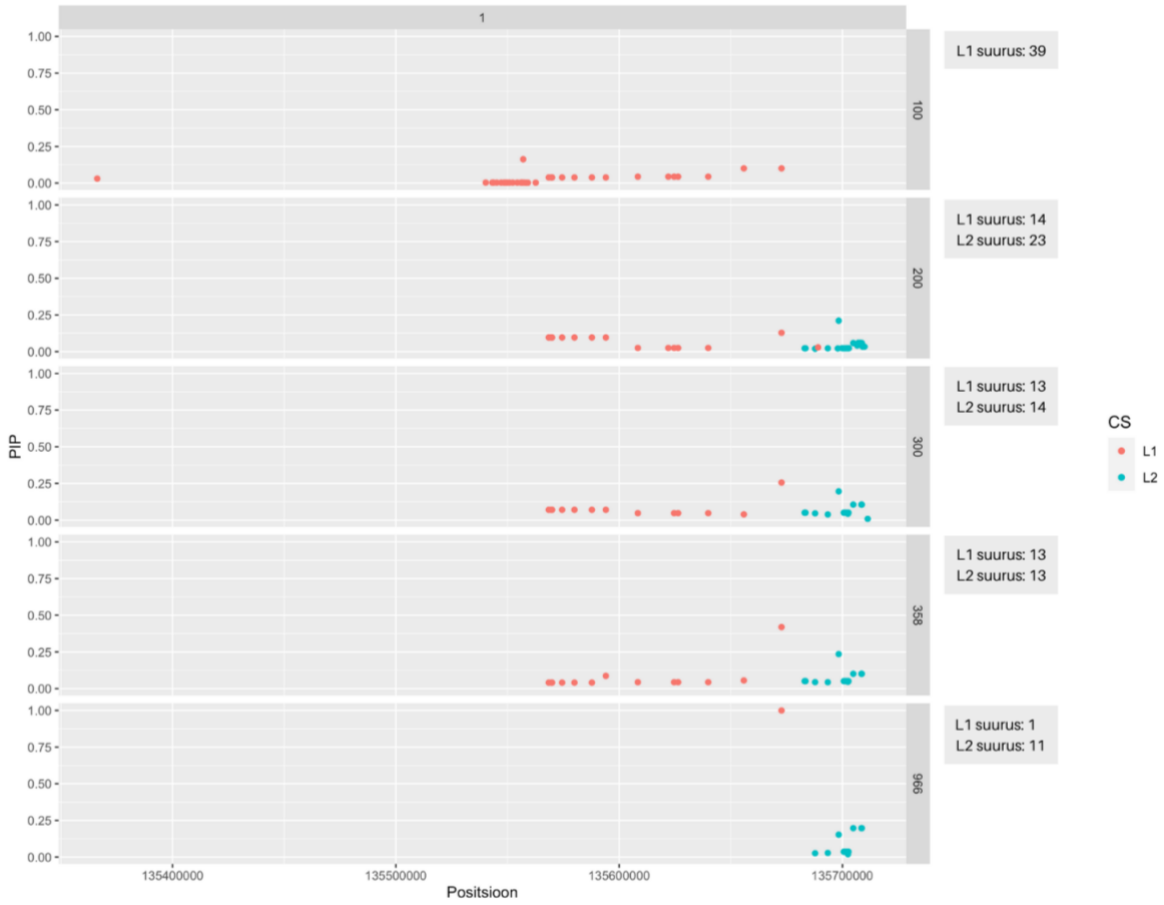
Joonis 6. **A** Geeni LZIC hulkades sisalduvad usaldusväärsed variandid vastavalt valimi suurusele, variantide positsioonidele kromosoomil 1 ja PIP väärtustele. **B** Geeni LZIC hulkadest moodustatud komponent, kus tippudeks on erinevatel valimi suurustel leitud usaldusväärsetest variantidest koosnevad L1 hulgad ja servadeks on hulkade vahel jagatud variandid.

Esialgu prooviti töö käigus graafi komponentidesse kuuluvate hulkade suuruse muutumist visualiseerida joondiagrammina, kuid statistilise võimsuse kasvust tuleneva hulkade arvu üleüldise suurenemise tõttu ei olnud sel viisil võimalik ülekattes olevate hulkade suuruse muutumist selgelt visualiseerida. Selle asemel võrreldi komponentides, millele leidus igal valimi suurusel üks CS, mediaani muutumist valimi kasvamisel, mis võimaldab selgemalt välja tuua valimi suuruse mõju CS-ide suurustele. Joonisel 7 on kujutatud viielistes komponentides sisalduvate CS-ide mediaani muutumist valimi suurenedes.



Joonis 7. Viiest CS-ist koosnevate komponentide mediaani muutumine erinevatel valimi suurustel viiuldiagrammina.

Paarikaupa ülekattuvatest variantide hulkadest koosnevas graafis oli igal komponendil kuni seitse CS-i. Kõige rohkem esines graafis komponente suurustega üks ja neli, vastavalt 5077 ja 1459 korral. Seejuures komponendid suurustega kuus ja seitse, mida oli graafis vastavalt üheksa ja kuus korda, koosnesid vähemalt kahest erineva indeksiga geenile vastavast CS-ist. Lisaks leidus geene, millele vastas kaks või enam erinevatest hulkadest koosnevat komponenti. Näiteks geenile SLC13A4 vastas kaks komponenti suurustega viis ja neli. Ühes komponendis olid L1 ja teises L2 hulgad erinevatel valimi suurustel. Joonis 8 kujutab geeni SLC13A4 komponentides olevaid CS-e erinevatel valimi suurustel. Viimaselt on võimalik näha, et valimi suurenedes vähenes nii esimese kui ka teise komponendi hulkades olevate variantide arv. Suuremad valimid võimaldasid esimeses komponendis (L1) LD-s olevaid variante paremini eristada. Viimast kinnitab L1-s valimi suurusel 966 leitud usaldusväärsete variantide hulk, mis sisaldas vaid ühte, ning seetõttu ka suure tõenäosusega põhjuslikku, geneetilist varianti.



Joonis 8. Geeni SLC13A4 komponentide CS-ides olevate variantide kujutamine erinevatel valimi suurustel vastavalt variandi positsioonile kromosoomil 7 ja variandi PIP väärtusele.

Valimi suurenemine tõstis võimsust leida seoseid geenide ja neid mõjutavate variantide hulkade vahel. Eriti võimaldab see lisaks tugevamatele signaalidele üles leida ka väiksemate efektisuurustega geenieksressiooni mõjutavaid variante, mida näitas rohkem kui ühe CS-iga geenide arvu kasv. Hulkade mediaani vähenemine viitab, et suuremate valimite korral on võimalik paremini eristada põhjuslikke variante tugevalt korreleeritud variantide hulgast. Seega on valimi suurusel oluline mõju täppiskaardistamise tulemustele, kuna suuremad valimid tõstavad täppiskaardistamise resolutsiooni, võimaldades täpsemalt määrata geene mõjutavaid põhjuslikke variante.

5.2 Edasised võimalused

Siinse töö tulemuste põhjal on võimalik teha mitu olulist järeldust valimi suuruse mõjust täppiskaardistamisele, kuid leidub veel mitmeid võimalusi analüüsi parandamiseks. Töö raames analüüsiti vaid Euroopa päritolu indiviidide proove, et täpsemalt hinnata valimi suuruse mõju täppiskaardistamisele, piirates populatsioonidevaheliste erinevuste tõttu

tekkida võivat müra põhjuslike signaalide leidmisel. Siiski ei võimalda see leida, mis piirini kasvab leitavate lookuste ja geenile mõjuvate variantide arv, mistõttu saab täpsuse muutumisele anda vaid limiteeritud hinnangu. Täiendavalt aitaks erinevatest populatsioonidest proovide agregeerimine edasi analüüsida täppiskaardistamise tulemusi suuremate valimite korral. Niisamuti on suuremate valimite mõju uurimiseks lisaks töös kasutatud andmestikele võimalik veel analüüsida selliseid andmestikke nagu näiteks GTEx [36], CAP [37] jt.

Töös analüüsiti vaid indiviidide LCL rakutüübist pärinevaid proove. Kuna geeniekspressioon varieerub vastavalt koele [7], mistõttu enamik tunnuseid avalduvad vaid teatud kudedes, siis oleks parema ülevaate saamiseks kasulik analüüs läbi viia mitmetes kudedes või rakutüüpides. Sageli asuvad kvantitatiivseid tunnuseid mõjutavad geneetilised variandid just erinevates kudedes, mistõttu võimaldaks erinevate kudede või koetüüpide analüüs neid variante paremini kaardistada.

Komponentide analüüsi käigus jäeti välja 15847 seost, kus paarikaupa ülekattuvad geneetilised variandid pärinesid erinevate geenide CS-idest. Siiski moodustavad nende variantide, mis mõjutavad rohkem kui ühte geeni, vahelised seosed umbkaudu veerandi leitud seostest ning on oluliseks osaks geneetiliste mehhanismide mõistmisel. Seetõttu on analüüsi võimalik täiendada, võttes arvesse kõiki paarikaupa ülekattuvate variantide seoseid geenist sõltumata. See annaks võimaluse vaadelda täpsemat trendi valimi suuruse ja täppiskaardistamise täpsuse vahel.

6. Kokkuvõte

Genoomis asuvad piirkonnad, milles olevad geneetilised variandid võivad mõjutada geeniekspressiooni ja kujundada erinevate haiguste riski. Täppiskaardistamise meetodil on võimalik leida geeniekspressiooni mõjutavaid põhjuslikke variante, kuid selle täpsust mõjutab kasutatava valimi suurus.

Käesoleva bakalaureusetöö eesmärk oli uurida, kuivõrd võimaldab suuremate valimite kasutamine saavutada suuremat statistilist usaldusväärsust põhjuslike variantide leidmisel, muutes täpsemaks täppiskaardistamise tulemusi. Töö käigus analüüsiti kokku viit erineva suurusega valimit, leides igale geenile ühe või rohkem CS-i, mis geeniekspressiooni mõjutab.

Töös leiti, et suuremate valimitega kasvab CS-ide arv ja väheneb hulkade keskmine suurus nii üleüldiselt kui ka üksiku hulga puhul erinevatel valimi suurustel. See näitab, et valimi suurenedes kasvab nii statistiline usaldusväärsus kui ka statistiline võimsus, mistõttu on võimalik leida ka väiksema efektisuurusega geeniekspressiooni mõjutavaid variante ning on paremini võimalik eristada korrelatsioonis olevaid variante. Tööd on võimalik täiendada vastavalt välja pakutud edasistele võimalustele (ptk 5.2), et edasi analüüsida valimi suuruse mõju täppiskaardistamise tulemustele.

7. Viidatud kirjandus

[1] Heinaru A. Geneetika. Õpik kõrgkoolile. Tartu: Tartu Ülikooli Kirjastus. 2012.

[2] Chial H. DNA sequencing technologies key to the Human Genome Project. *Nature Education*, 2008, Vol. 1, Issue 1, p. 219. https://www.nature.com/scitable/topicpage/dna-sequencing-technologies-key-to-the-human-828/ (15.02.2021).

[3] Zhao R. F. ENCODE: Deciphering Function in the Human Genome. 2012. https://www.genome.gov/27551473/genome-advance-of-the-month-encode-deciphering-function-in-the-human-genome (15.02.2021).

[4] Nica A. C., Dermitzakis E. T. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2013, Vol. 368, Issue 1620. https://doi.org/10.1098/rstb.2012.0362.

[5] Single Nucleotide Polymorphism (SNP). Encyclopedia of Public Health. Ed. by W. Kirc. Dordrecht: Springer, 2008, p. 1305. https://doi.org/10.1007/978-1-4020-5614-7_3214.

[6] Members of the Complex Trait Consortium. The nature and identification of quantitative trait loci: a community's view. *Nature Reviews Genetics*, 2003, Vol. 4, Issue 11, pp. 911–916. https://doi.org/10.1038/nrg1206.

[7] Schaid D. J., Chen W., Larson N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 2018, Vol. 19, Issue 8, pp. 491–504. https://doi.org/10.1038/s41576-018-0016-z.

[8] Hutchinson A., Watson H., Wallace C. Improving the coverage of credible sets in Bayesian genetic fine-mapping. *PLoS Computational Biology*, 2020, Vol. 16, Issue 4. https://doi.org/10.1371/journal.pcbi.1007829.

[9] Spain S. L., Barrett J. C. Strategies for fine-mapping complex traits. *Human Molecular Genetics*, 2015, Vol. 24, Issue R1, pp. R111–R119. https://www.doi.org/10.1093/hmg/ddv260.

[10] Zeng B., Bendl J., Kosoy R., Fullard J. F., Hoffman G. E., Roussos P. Trans-ethnic eQTL meta-analysis of human brain reveals regulatory architecture and candidate causal variants for brain-related traits. *medRxiv*, 2021. https://doi.org/10.1101/2021.01.25.21250099.

[11] Shan N., Wang Z., Hou L. Identification of trans-eQTLs using mediation analysis with multiple mediators. *BMC Bioinformatics*, 2019, Vol. 20, Issue 3, p. 126. https://doi.org/10.1186/s12859-019-2651-6.

[12] Yap H.-Y., Siow T.-S., Chow S.-K., Teow S.-Y. Epstein-Barr Virus- (EBV-) Immortalized Lymphoblastoid Cell Lines (LCLs) Express High Level of CD23 but Low CD27 to Support Their Growth. *Advances in Virology*, 2019, Vol. 2019. https://doi.org/10.1155/2019/6464521.

[13] Wang G., Sarkar A., Carbonetto P., Stephens M. A simple new approach to variable selection in regression, with application to genetic fine-mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2020, Vol. 82, Issue 5, pp. 1273–1300. https://doi.org/10.1111/rssb.12388.

[14] Urry L. A., Cain M. L., Wasserman S. A., Minorsky P. V., Reece J. B. Campbell biology. 11th edition. New York: Pearson. 2017. https://www.pearson.com/content/dam/one-dot-com/one-dot-com/us/en/higher-ed/en/custom-product/urry-campbell-biology-11e/pdf/urry11e-ch5.pdf (13.02.2021).

[15] Griffiths A. J. F., Miller J. H., Suzuki D. T., Lewontin R. C., Gelbart W. M. An Introduction to Genetic Analysis. 7th edition. New York: W. H. Freeman. 2000. https://www.ncbi.nlm.nih.gov/books/NBK22012/ (16.02.2021).

[16] Bello J., Jiménez M. Functional implications of single nucleotide polymorphisms (SNPs) in protein-coding and non-coding RNA genes in multifactorial diseases. *Gaceta medica de Mexico*, 2017, Vol. 153, Issue 2, pp. 218–229. https://www.researchgate.net/publication/319880708_Functional_implications_of_single_nucleotide_polymorphisms_SNPs_in_protein-coding_and_non-coding_RNA_genes_in_multifactorial_diseases (17.02.2021).

[17] Bush W. S. Genome-Wide Association Studies. Encyclopedia of Bioinformatics and Computational Biology. Ed. by Ranganathan S., Gribskov M., Nakai K., Schönbach C. Oxford: Academic Press, 2019, pp. 235–241. https://doi.org/10.1016/B978-0-12-809633-8.20232-X.

[18] Hutchinson A., Asimit J., Wallace C. Fine-mapping genetic associations. *Human Molecular Genetics*, 2020, Vol. 29, Issue R1, pp. R81–R88. https://doi.org/10.1093/hmg/ddaa148.

[19] IGSR: The International Genome Sample Resource. GEUVADIS. https://www.internationalgenome.org/data-portal/data-collection/geuvadis (10.04.2021).

[20] European Genome-phenome Archive. GENCORD2 GENOTYPES. https://ega-archive.org/datasets/EGAD00001000428 (10.04.2021).

[21] TwinsUK. https://twinsuk.ac.uk/resources-for-researchers/access-our-data/ (10.04.2021).

[22] Min J. L., Taylor J. M., Richards J. B., *et al*. The Use of Genome-Wide eQTL Associations in Lymphoblastoid Cell Lines to Identify Novel Genetic Pathways Involved in Complex Traits. *PLoS ONE*, 2011, Vol. 6, Issue 7. https://doi.org/10.1371/journal.pone.0022070.

[23] Lappalainen T., Sammeth M., Friedländer M., *et al*. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 2013, Vol. 501, Issue 7468, pp. 506–511. https://doi.org/10.1038/nature12531.

[24] Gutierrez-Arcelus M., Lappalainen T., Montgomery S. B., *et al*. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife*, 2013, Vol. 2. https://doi.org/10.7554/eLife.00523.

[25] Buil A., Brown A., Lappalainen T., *et al*. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nature Genetics*, 2015, vol 47, pp 88–91. https://doi.org/10.1038/ng.3162.

[26] The Variant Call Format Specification. https://samtools.github.io/hts-specs/VCFv4.3.pdf (10.04.2021).

[27] Genome Reference Consortium. GRCh38.p13. https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39 (11.04.2021).

[28] BCFtools. http://samtools.github.io/bcftools/bcftools.html (11.04.2021).

[29] Kerimov N., Hayhurst J. D., Peikova K., *et al*. eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs. *bioRxiv*, 2021. https://doi.org/10.1101/2020.01.29.924266.

[30] Nextflow. https://www.nextflow.io/ (07.04.2021).

[31] Kerimov N. Designing a robust and portable workflow for detecting genetic variants associated with molecular phenotypes across multiple studies. TÜ arvutiteaduse instituudi magistritöö. 2019. https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=66782&year=2019.

[32] Kerimov N. Bioinformatics analysis pipeline for QTL Analysis. https://github.com/kerimoff/qtlmap (15.03.2021).

[33] Kuklane A. M. Bakalaureusetöö analüüs valimi suuruse mõjust eQTL-ide täppiskaardistamisele lümfoblastoidrakuliinides. https://github.com/akuklane/bakalaureusetoo (07.05.2021).

[34] Stovner E. B., Sætrom P. PyRanges: efficient comparison of genomic intervals in Python. *Bioinformatics*, 2020, Vol. 36, Issue 3, pp. 918-919. https://doi.org/10.1093/bioinformatics/btz615.

[35] python-igraph. https://igraph.org/python/ (25.03.2021).

[36] GTEx Portal. GTEx Analysis V8. https://www.gtexportal.org/home/datasets (26.04.2021).

[37] dbGAP. Cholesterol and Pharmacogenetics (CAP) Study. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000481.v3.p2 (26.04.2021).

8. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Anette Maria Kuklane**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „**Valimi suuruse mõju ekspressiooni kvantitatiivsete tunnuste lookuste täppiskaardistamisele lümfoblastoidrakuliinides**", mille juhendaja on **Kaur Alasoo**, 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.

Anette Maria Kuklane

07.05.2021