TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

GENOOMIKA INSTITUUT

Erinevate referentspaneelide mõju ülegenoomsete assotsiatsiooniuuringute tulemustele

Bakalaureusetöö

12 EAP

Alissa Kazmin

Juhendajad PhD Kristi Läll

PhD Katri Pärna

TARTU 2025

“Erinevate referentspaneelide mõju ülegenoomsete assotsiatsiooniuuringute tulemustele”

Levinud metoodikaks genoomsetes uuringutes on andmete imputatsioon, kus imputeerimise täpsus sõltub suuresti õigesti valitud referentspaneelist. Ülegenoomsetes assotsiatsiooniuuringutes imputatsioonivead võivad viia valepositiivsete seosteni. Käesoleva töö eesmärk oli võrrelda Eesti-põhise EstREF ja rahvusvahelise HRC referentspaneelide valiku tähtsust imputeerimisel. Referentspaneelide täpsust võrreldi ülegenoomsete uuringute efektisuuruste hinnangute kaudu, kasutades selle jaoks PheMED metoodikat. Valimiks kasutati üle 200 000 Eesti Biopanga doonorite geeniandmeid ja kümmet erinevat peakomponenti.

Kahe referentspaneeli võrdlus näitas, et Eesti-põhisel EstREF referentspaneelil imputeeritud andmed ei erinenud statistiliselt rahvusvahelise HRC referentspaneelil imputeeritud andmetest. Antud töö demonstreerib PheMED metoodika kasulikkust ülegenoomsete assotsiatsiooniuuringute metaanalüüsis. Bakalaureusetöö oli teostatud Tartu Ülikooli Genoomika instituudi geneetilise epidemioloogia uurimisgrupis.

Märksõnad: ülegenoomsed assotsiatsiooniuuringud, referentspaneelide võrdlus, EstREF referentspaneel, HRC referentspaneel, PheMED metoodika

CERCS-kood: B220 Geneetika, tsütogeneetika

“The impact of different imputation panels on genome-wide association study results”

Imputation is a widely used methodology in genomic studies, and its accuracy depends on a correctly selected reference panel. In genome-wide association studies, imputation errors can lead to false-positive associations. The aim of this work was to compare the importance of selecting the Estonian-based EstREF and the international HRC reference panels for imputation. The accuracy of the reference panels was compared by estimating genome-wide association studies effect sizes using the PheMED methodology. The dataset comprised genetic data and ten different principal components from more than 200,000 Estonian Biobank donors.

The comparison showed that the data imputed with the Estonian-based EstREF reference panel didn’t differ statistically from data imputed with the international HRC reference panel. This work highlights the utility of the PheMED methodology in

meta-analysis of genome-wide association studies. The bachelor's thesis was conducted in the Genetic Epidemiology Research Group of the Institute of Genomics, University of Tartu.

Keywords: genome-wide association studies, comparison of reference panels, EstREF reference panel, HRC reference panel, PheMED methodology

CERCS code: B220 Genetics, cytogenetics

SISUKORD

Kasutatud lühendid

DNA – desoksüribonukleiinhape

EstBB – Eesti Biopank

EstREF – Eesti täisgenoomide sekveneerimisandmestik

GWAS – ülegenoomsed assotsiatsiooniuuringud

HRC – Haplotype Reference Consortium

KMI – kehamassiindeks (ingl *body mass index*)

LD – aheldustasakaalutus

MAF – minoorse alleeli sagedus

NPV – negatiivne prognoosiväärtus

PCR – polümeraasi ahelreaktsioon

PheMED – fenotüüpi iseloomustava efektiivse lahjendatuse hindamise tööriist

PPV – positiivne prognoosiväärtus

SNP – üksiku nukleotiidi polümorfism

WGS – kogu genoomi sekveneerimine

Põhimõisted

Aluspaar (ingl *base pair*) vesiniksidemetega ühendatud kaks DNA nukleotiidi (A-T ja G-C).

Deletsioon (ingl *deletion*) mutatsiooni tüüp, mille puhul eemaldatakse DNA molekuli järjestusest üks või mitu nukleotiide.

DNA desoksüribonukleiinhape (ingl *deoxyribonucleic acid*) on geneetilise informatsiooni kandja, mis koosneb nukleotiididest.

Efektialleel (ingl *effect allele*) alleel, mis põhjustab teatud tunnuse muutust või organismi omadusi.

Fenotüüp (ingl *phenotype*) organismi tunnus või omadus, mille aluseks on pärilikkus tegurite ja keskkonna koostoime.

Geenivariant (ingl *genetic variant*) eri vorm ühest ja samast geenist, sünonüüm allelliga.

Genoom (ingl *genome*) kogu geneetiline materjal organismis, mis sisaldab kogu tema DNA-d.

Genotüpiseerimine (ingl *genotyping*) protsess, mille käigus määratakse organismi geneetilised variatsioonid kindlates DNA piirkondades.

GWAS ülegenoomne assotsiatsiooniuuring (ingl *genome-wide association study*) on statistiline uuring, mis uurib seoseid geneetiliste variantide ja tunnuste või haiguste vahel populatsioonis.

Haplotüüp (ingl *haplotype*) tihedalt aheldunud geenipaaride kogum kromosoomis, mis päranduvad koos.

Imputatsioon (ingl *imputation*) statistiline meetod puuduvate genotüüpide tuletamiseks, kasutades populatsioonis teadaolevaid haplotüüpe.

Insertsioon (ingl *insertion*) mutatsiooni tüüp, kus nukleotiidipaari(de) lisandumine DNA molekuli, mis põhjustab nukleotiidijärjestuse muutusi.

Kohort (ingl *cohort*) inimeste rühm, keda uuritakse kindla ajavahemiku jooksul.

Kromosoom (ingl *chromosome*) rakutuumas paiknev DNA molekuli ja valkude kompleks, mis määrab geenide pärandumise.

Lookus (ingl *locus*) kindel koht kromosoomis , kus geen asub (üks tema alleelidest).

MAF (ingl *minor allele frequency*) vähemesineva alleeli sagedus, mis näitab, kui sageli esineb populatsioonis sageduselt teine alleel.

Manhattan graafik (ingl *Manhattan plot*) graafik, mida kasutatakse ülegenoomse assotsiatsiooniuuringu tulemuste visualiseerimiseks. X-teljel on märgitud kromosoomid ja

Y-teljel assotsieerunud markerite p-väärtuste negatiivsed kümnendlogaritmid, mis näitavad, kui statistiliselt oluline on seos mingi geneetilise variatsiooni ja uuritava tunnuse vahel.

Metaanalüüs (ingl *meta-analysis*) meetod, mille abil kombineeritakse varasemate väiksemate uuringute tulemused ühte suure uuringusse, et teha täpsemaid ja usaldusväärsemaid järeldusi.

NGS (ingl *next generation sequencing*) järgmise põlvkonna sekveneerimine kiire ja kõrge läbilaskevõimega DNA või RNA sekveneerimise tehnoloogia, mis võimaldab samaaegselt määrata miljoneid kuni miljardeid lühikesi DNA järjestusi.

PCR (ingl *polymerase chain reaction*) polümeraasi ahelreaktsioon on laboratoorne lämmastikaluste amplifitseerimise tehnika.

PheMED (ingl *Phenotypic Measurement of Effective Dilution*) fenotüüpi iseloomustava efektiivse lahjendatuse tööriist, mis aitab hinnata tulemuste usaldusväärsust geneetilistes uuringutes hinnates ja korrigeerides fenotüübi määratluse ebatäpsust.

P-väärtus (ingl *p-value*) statistiline näitaja, mis hindab kui tõenäoline on saada uuritavate andmete tulemusi nullhüpoteesi kehtimisel.

Referentspaneel (ingl *reference panel*) DNA järjestuse andmebaas, mis koosneb eri doonorite geneetilistest andmetest, mida kasutatakse omakorda teiste isikute geneetiliste markerite imputeerimiseks.

Rekombinatsioon (ingl *recombination*) on meioosi profaasis toimuv protsess, mille käigus toimub mõlema vanema kromosoomide vahel geneetilise materjali vahetus, luues seeläbi uusi geenide kombinatsioone.

Sekveneerimine (ingl *sequencing*) DNA või RNA nukleotiidide täpse järjestuse määramine.

SNP (ingl *single nucleotide polymorphism*) ühenukleotiidne polümorfism, mis esineb DNA ahelas üksiku nukleotiidi erinevusena.

Sõeluuring (ingl *screening*) elanikkonna teatud sihtrühma (nt kindel vanuserühm) inimeste uurimine, et avastada mingi haiguse põdejaid või alleeli – geeni üks võimalikest vormidest, mis määrab konkreetse tunnuse variatsiooni.

Varjatud Markovi mudel (ingl *Hidden Markov Model*) kvantitatiivne modelleerimismeetod mingi protsessi kirjeldamiseks ajas, kus uuritavad jaotatakse seisunditesse vastavalt protsessi arengu etapile ja/või rakendatavatele sekkumistele ning seisundite vahelised liikumised toimuvad vastavalt üleminekutõenäosustele.

WGS (ing k *whole genome sequencing*) kogu genoomi sekveneerimine on protsess, mille käigus määratakse organismi genoomi DNA järjestus täielikult.

Sissejuhatus

Geneetika teadusharu juured ulatuvad 19. sajandisse, mil Austria munk Gregor Mendel uuris tunnuste pärandumist ehk pärilikkust herneste abil. Mendel avastas, et tunnused päranduvad kindlate seaduspärasuste järgi, mis on tänapäeval tuntud kui Mendeli seadused (W. E. Castle, 1903).

Sellele järgnes geneetika teadusharu kiire areng 20. sajandil, mil leiti, et desoksüribonukleiinhappe molekul (DNA) on päriliku info kandja ning teadlased Watson ja Crick avastasid 1953.aastal DNA molekuli kaheahelalise struktuuri (Watson & Crick, 2003). DNA molekul koosneb nukleotiididest, mis omakorda koosnevad kolmest komponendist: fosfaatrühm, viiesüsinikuline suhkur (desoksüriboos) ja lämmastikalus. Kaheahelalist DNA molekuli hoiavad omavahel koos lämmastikaluste paarid, mille vahel on vesiniksidemed. Lämmastialuste paar adeniin (A) ja tümiin (T) on seotud kahe vesiniksidemega ning guaniini (G) tsütosiini (C) paar moodustab kolme vesiniksidemega ühenduse (Avery & McCARTY, n.d.).

Pärast Gregor Mendeli katsete taasavastamist ja mõistmist, et DNA on päriliku info kandja, keskenduti üha enam konkreetsete DNA lõikude ehk geenide uurimisele. Geen on kindla nukleotiidse järjestusega DNA lõik. Geen kannab infot ühe valgu valmistamiseks, et uurida, kuidas need määravad organismi tunnuseid (Gayon, 2016). Organismi kogu DNA, sealhulgas geenid moodustavad genoomi.

Genoomi genotüpiseerimine on protsess, milles määratakse kindlaks DNA järjestuses esinevad variatsioonid, eesmärgiga tuvastada indiviidide vahelised geneetilised erinevused. Genotüpiseerimise ajalugu ulatub 20. sajandi lõppu (Galas & McCormack, n.d.), kui hakati teostama esimesi katseid geneetilise info määramiseks, mis keskendusid üksikutele geenidele ja nende variatsioonidele (International Human Genome Sequencing Consortium et al., 2001). 1970. aastatel töötati välja DNA järjestuse määramiseks Sangeri sekveneerimine, mis põhineb DNA ahela sünteesi katkestamisel dideoksünukleotiidide abil, võimaldades seeläbi määrata nukleotiidide täpse järjestuse. Sekveneerimisega sai kindlaks teha kogu genoomi järjestuse, mis omakorda võimaldas täpsemalt uurida erinevaid geeni variatsioone (Hutchison, 2007) nagu üksiku nukleotiidi polümorfismid (SNP-d), insertsioonid, deletsioonid või struktuurilised ümberkorraldused (Sayitoğlu, 2016).

Kõige enam uuritakse SNP-e, (ühenukleotiidsed variatsioonid DNA järjestuses), kuna need moodustavad ligikaudu 90% kogu genoomi variatsioonist (Collins et al., 1998). SNP-d on olulised haigusriskide ning pärilike omaduste uurimisel (Do et al., 2012). Nähtust, kus geneetilised variatsioonid esinevad populatsioonis koos sagedamini kui juhuslikult arvatud nimetatakse aheldustasakaalutuseks (Sved & Hill, n.d.). Aheldustasakaalutus mängib olulist rolli geneetilistes uuringutes, sealhulgas haiguste pärilikkuse mõistmises (Kruglyak, 1999). Kuigi enamik SNP-dest on levinud erinevates kohortides (Hinds et al., 2005), on ka seesuguseid, mis on spetsiifilised teatud populatsioonidele või geograafilistele piirkondadele (Choudhury et al., 2014).

Geneetilised variandid mõjutavad organismi omadusi ning võivad olla erinevate haiguste põhjustajad (Ramírez-Bello, 2023). Tervikpildi geneetilistest variantides moodustab kogu genoomi sekveneerimine ehk WGS (ingl *whole genome sequencing*) (Bagger et al., 2024). Esimene inimese täielik genoomi sekveneerimine viidi lõpule 2003. aastal Inimese genoomi projekti (ingl *The Human Genome Project*) raames (The Human Genome Project, 2025). Sekveneerimise projekt kestis 13 aastat ja maksis ligikaudu 2,7 miljardit dollarit (Human Genome Project, 2024). Tänapäeva tehnoloogiad, nagu järgmise põlvkonna sekveneerimine ehk NGS (ingl *next generation sequencing*), on viinud genotüpiseerimise kulud märkimisväärselt madalamalele- 1 miljardist dollarist (The Cost of Sequencing a Human Genome, 2021) kuni vähem kui sada dollarit proovi kohta (Campbell et al., 2015), muutes see kättesaadavamaks (Pruneri et al., 2021).

DNA esmane kasutus meditsiinis algas geneetiliste haiguste diagnostikas ja personaalmeditsiini vallas (Passarge, 2021). Üks esimesi rakendusi oli sünnieelne diagnostika, kus analüüsiti loote geneetilist profiili Downi sündroomi ja teiste kromosoomianomaaliate tuvastamiseks (Mckusick, n.d.). Hiljem laienes kasutus onkoloogiasse, farmakogeneetikasse ja nakkushaiguste tuvastamisse (Sadee et al., 2023).

Aastast 1993. on alustatud Eestis vastsündinute sõeluuringut fenüülketonuuria avastamiseks (Tartu Ülikooli Kliinikum, 2015). Tänaseks päevaks võimaldab uuring diagnoosida 22 päriliku haiguse suhtes (Tartu Ülikooli Kliinikum, 2024). Imikute sõeltestimine esimestel elupäevadel aitab tuvastada pärilikke haigusi, mis võivad ilma ravita kahjustada lapse tervist või arengut. Protseduuri käigus võetakse vastsündinu kannast või varbast verd, mis kantakse testkaardile (Pitt, 2010). Testitakse mitmete tõsiste haiguste

esinemist nagu näiteks: vitamiin B12 puudulikkus, spinaalne lihasatroofia, kaasasündinud hüpotüreoos, aminohapete ainevahetushäired, klassikaline galaktoseemia (Tartu Ülikooli Kliinikum, 2024).

Teaduspõhiste sõeluuringute aluseks on teadmised haigust põhjustavate geneetiliste variatsioonide kohta. Tänapäeva genotüpiseerimistehnoloogiad võimaldavad koguda ulatuslikke genoomi andmeid, mille analüüsimine aitab paremini mõista geneetilisi variatsioone. Samas nii suure hulga geneetilise info analüüsimisel tuleb arvestada andmete kvaliteedi mõju tulemustele. Andmete kvaliteedi hindamiseks on loodud statistilised programmid, mis parandavad tulemuste usaldusväärsust luues aluse sõeluuringute teostamiseks.

Levivaks metoodikaks genoomsetes uuringutes on andmete imputatsioon, kus imputeerimise täpsus sõltub suuresti õigesti valitud referentspaneelist. Ülegenoomsetes assotsiatsiooniuuringutes imputatsioonivead võivad viia vale-positiivsete seosteni. Käesoleva töö eesmärk oli uurida referentspaneelide valiku tähtsust imputeerimisel läbi ülegenoomse uuringute efektisuuruste võrdlemise, kasutades PheMED metoodikat Eesti Biopanga andmetel. Antud töö oli teostatud Genoomika instituudis.

1. Kirjanduse ülevaade

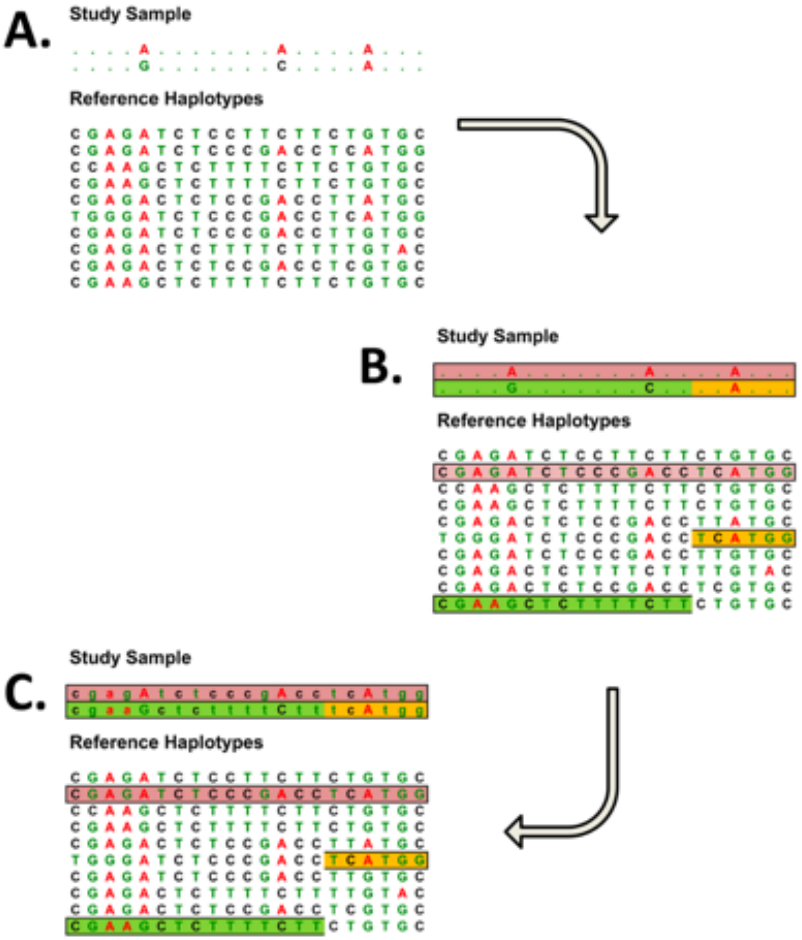
1.1 Imputatsioon

Genoomi või selle osade kaardistamiseks kasutatakse mitmesuguseid meetodeid, millest levinuim on genotüpiseerimine (Casals et al., 2012). Imputatsioon kuulub statistiliste meetodite hulka, mis ennustab DNA variante lookustes, mis esialgsetes genotüpiseeritud andmetes puudusid (Joonis 1). Imputatsiooni teostamiseks kasutatakse referentspaneele – andmestikud, mis on loodud inimeste sekveneeritud geeniandmetest ning sisaldavad teadaolevaid haplotüüpe (the Haplotype Reference Consortium, 2016). Nende kasutamisel tuvastatakse geneetilisi mustreid ja seoseid uuritava indiviidi ja referentspaneeli haplotüüpide vahel (Rubinacci et al., 2023). Imputatsioon on parim alternatiiv täisgenoomi sekveneerimisele, võimaldades kuluefektiivselt ennustada puuduvaid genotüüpe (Martin et al., 2021).

See on võimalik tänu SNP-idele, mis eristavad ühte proovi (isikut/genoomi) teistest ning mis paiknevad üle kogu genoomi. SNP-del on tavaliselt kaks alleeli. Enamasti

raporteeritakse harvemini esineva alleeli sagedust (ingl *minor allele frequency, MAF*) (Bush et al., 2012.). Alleeli sagedusega 0.05 (5%) ja rohkem defineeritakse sagedaselt levinud variandiks ja alleeli esinemistõenäosusega 0.01 (1%) või vähem harvem esinevaks variandiks populatsioonis.

Kuigi kogu genoomi sekveneerimist peetakse standardiks haruldaste variantide detekteerimisel (Quick et al., 2020), on imputatsioon aja- ja kuluefektiivsem viis rohkemate geneetiliste markerite kaasamiseks uuringutes (Shi et al., 2023). Seesugust imputatsiooni on rakendatud UK Biopanga geneetilistel andmetel, mille tulemusel saadi 825,927 genotüpiseeritud markeri asemel 96 miljonit markerit. (Bycroft et al., 2018). Imputatsiooni meetodeid on aastate jooksul välja arendatud mitmeid (Alwateer et al., 2024), kõige täpsemad genotüübi imputatsiooni meetodid põhinevad varjatud Markovi mudelil (ingl *Hidden Markov Model*). Imputatsiooni töövoos joondatakse uurimisrühma (genotüpiseeritud aluspaarid) genotüübid referentspaneeli haplotüüpidega (Howie et al., 2011). Imputatsiooni teostamisel peame silmas pidama, milline on meie sihtpopulatsiooni genoomide struktuur (aheldustasakaalutus) ja toetudes sellele infole, kasutama kõige sarnasemat imputatsiooni referentspaneeli. Sealhulgas sõltub imputatsiooni täpsus märkimisväärselt referentspaneeli suurusest ja selle geneetilisest kaugusest sihtpopulatsioonist. (Quick et al., 2020).



Joonis 1. Illustratiivne näidis genotüübi imputatsioonist omavahel kaugete isikute puhul. Pildil A on kujundatud uuritava proovi osalisi SNP-i andmeid: punktid tähistavad positsioone, kus alleel kattub referentspaneelis sagedamini esineva alleeliga ning A, G ja C positsioonid märgivad SNP-e, mis erinevad referentspaneeli sagedaseimast alleelist. Uuritava proovi andmed kõrvutatake teadaolevate referentspaneeli võimalike haplotüüpidega. Pildil B leitakse referentshaplotüübid, mis kattuvad kõige paremini uuritava proovi teadaolevate SNP-dega. Selle etapi käigus kasutatakse algoritme, mis võimaldavad määrata kõige tõenäolisemad haplotüübid uuritava genotüübi alusel. Pilt C demonstreerib, kuidas valitud haplotüüpide alusel ennustatakse imputatsiooniga kogu proovi järjestus, sealhulgas ka algselt teadmata nukleotiidid. Tulemusena saadakse kaks järjestatud haplotüüpi, mis esindavad kumbagi vanemalt saadud kromosoomi koopiat. See lähenemine tugineb referentspaneelile ja võimaldab konstrueerida täieliku genotüübi osaliste andmete põhjal (Li et al., 2009).

1.1.2 Referentspaneelid

Referentspaneelid on eelnevalt sekveneeritud genoomide andmekogumid, mida kasutatakse imputeerimiseks (Das et al., 2018). Referentspaneelid sisaldavad paljude inimeste sekveneeritud genoome (või genoomiosade järjestusi) esindades laialdast geneetilist varieeruvust eri populatsioonides (The 1000 Genomes Project Consortium et al., 2015).

Kasutades referentspaneeli teostatakse imputeerimist, mille käigus täidetakse puuduvad geneetilised andmed (SNP-d), suurendades seeläbi analüüsi täpsust ja võimekust tuvastada haruldasi või harva esinevaid variante (the Haplotype Reference Consortium, 2016). Analüüsi täpsus sõltub eelkõige referentspaneeli õigest valikust (O'Connell et al., 2021). Geneetikas tuntud referentspaneelid nagu *Haplotype Reference Consortium* (HRC), 1000 *Genomes Project* (1000G) ja *TOPMed* erinevad omavahel valimi suuruse, populatsiooni katvuse, variantide tiheduse ning haruldaste variantide katvuse poolest (Tabel 1) (Sengupta et al., 2023). Imputeerimise edukaks teostamiseks tuleb jälgida, kui hästi valitud referenetspaneel esindab uuritavat populatsiooni (Schurz et al., 2019). Järjest enam arendatakse populatsioonispetsiifilisi ja mitmekesisemaid paneele, tagamaks paremaid võimalusi haiguste uurimiseks (Gurdasani et al., 2019).

Tabel 1. Levinumate referentspaneelide võrdlus

Referentspaneel	SNP-de arv	Haplotüüpide arv	MAF-ide väärtus (%)	Proovide arv
The Haplotype Reference Consortium (HRC)	39,235,157	64,976	> 1%	32,470
1000 Genomes Project (1000G)	84,700,000	5,008	> 1%	2,504
TOPMed	445,600,184	194,512	>0,01%	53,831

Tabel 1. Ülevaade levinuimatest referentspaneelidest. Võrdlemiseks on välja toodud iga paneeli kohta tuvastatud SNP-de koguarv, haplotüüpide arv, MAF-ide sagedus ning referentsandmestike proovide arv. The Haplotype Reference Consortium (HRC) sisaldab 39,235,157 SNP-d ja 64,976 haplotüüpi, mille MAF on üle 1%, põhinedes 32,470 isiku andmetel (the Haplotype Reference Consortium, 2016). 1000 Genomes Project (1000G) hõlmab 84,700,000 SNP-d ja 5,008 haplotüüpi, sisaldades nii sagedasi kui ka haruldasemaid variante (MAF>1%), tuginedes 2,504 isikul (Byrska-Bishop et al., 2022). TOPMed referentspaneel sisaldab 445,600,184 SNP-d ja 194,512 haplotüüpi, hõlmates ka väga haruldasi variante (MAF>0,01%) ning põhineb 194,000 kogutud andmetel (Taliun et al., 2021).

1.3 Ülegenoomsed assotsiatsiooniuuringud

Ülegenoomne assotsiatsiooniuuring (Genome-Wide Association Study, lühend GWAS) on meetod, millega uuritakse geneetiliste variatsioonide seoseid erinevate haiguste või tunnustega. Uuringute käigus analüüsitakse suure hulga inimeste genoome, et tuvastada nende statistiliselt olulisi seoseid SNPide ja kindlate fenotüüpide või haiguste vahel. Selleks kasutatakse regressioonimudeleid, mis võtavad sisendiks indiviidide genotüübi- ja fenotüübiandmed (Uffelmann et al., 2021). Regressioonimudelid arvutavad väljundina iga SNP kohta alleeli efektisuuruse ning selle kohta käiva p-väärtuse, mis iseloomustab SNP seost uuritava tunnusega. Ülegenoomselt oluliseks loetakse SNP-d, mille p-väärtus jääb alla 5×10^-8 (Jannot et al., 2015).

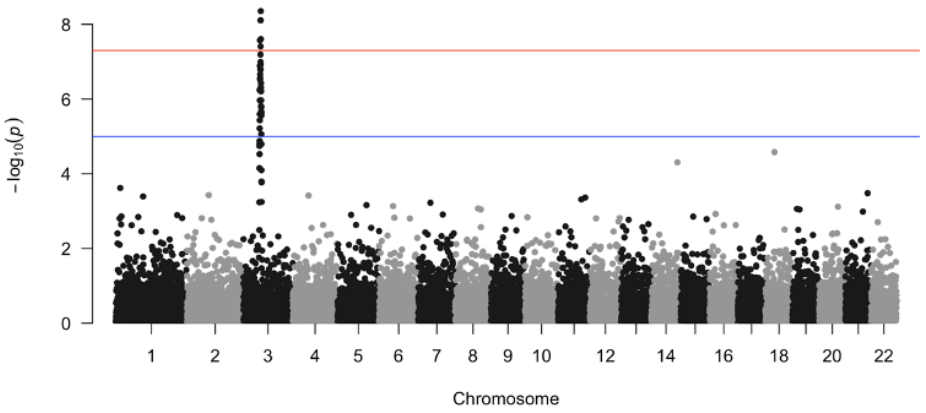
GWAS-i kasutatakse laialdaselt erinevate haiguste, nagu diabeet (Suzuki et al., 2024), Alzheimeri tõbi (Wightman et al., 2021) ja südamehaigused (Tcheandjieu et al., 2022), geneetiliste faktorite tuvastamiseks. Uuringus võrreldakse SNP-de esinemissagedust haigust põdevate inimeste ja kontrollrühma isikute vahel.

Analüüsiprotsess algab sobivate uurimis- ja kontrollrühmade valimisega, järgneb SNP-de sihipärane genotüpiseerimine, seejärel imputeerimine ning lõpuks viiakse läbi statistiline assotsiatsioonianalüüs, et tuvastada võimalikud haigusega seotud genoomipiirkonnad ehk lookused.

GWAS-i tulemusi visualiseeritakse Manhattan graafikuna (ingl *Manhattan plot*), kus vertikaalteljel on kujutatud statistilise olulisuse tasemed ning horisontaalteljel SNP-de kromosomaalne asukoht (Wang et al., 2019) (Joonis 2). Olulised seosed klasterduvad tihti piirkonniti, mis sisaldavad omavahel korreleeruvaid SNP-klastreid (Xu et al., 2019). Suurendamaks analüüsitulemuste usaldusväärsust, kasutatakse varem kogutud andmeid ja metaanalüüsimeetodeid, mis aitavad vähendada valepositiivsete seoste osakaalu ning suurendada statistilist võimsust (Zhou et al., 2023).

Tuleb rõhutada, et ainuüksi GWAS-i analüüsi tulemused ei ole piisavad haigust põhjustavate SNP-de täpseks määramiseks. See piirang tuleneb sellest, et GWAS-i tuvastatud SNP-d võivad olla tegelike haigust põhjustavate variantidega aheldunud (Schaid et al., 2018). Lisaks peame arvestama pleiotroopia võimalusega, mil geneetiline variant on seotud mitme fenotüübiga, ilma et see oleks kõigi nende tunnuste otsene põhjustaja. Näiteks SNP, mis on seotud kehamassiindeksiga (KMI), võib olla statistiliselt oluline ka II-tüüpi diabeedi puhul, kuid ei mõjuta otseselt diabeedi teket (Solovieff et al., 2013). Sellest hoolimata pakuvad

GWAS-i tulemused tähtsat lähtepunkti edasistele uuringutele, mis aitavad määrata SNP-ide täpset rolli haiguste kujunemisel.



Joonis 2. Manhattan graafik. Manhattan graafikut kasutatakse GWAS-uuringutes, illustreerimaks geneetiliste variantide (üldjuhul SNP-de) ja uuritava fenotüübi vahelisi seoseid. X-teljel on kromosoomid 1 kuni 22. Y-teljel on iga geneetilise variandi p-väärtuse negatiivne kümnendlogaritm. Sinine horisontaaljoon märgib soovituslikku statistilise olulisuse taset (p=1*10^-5). Punane horisontaaljoon märgib ülegenoomset statistilise olulisuse piiri (p=5*10^-8.), millest lähtudes tehakse järeldusi geneetilise variandi rollist fenotüübi kujunemisel. Geneetilised variandid, mis ületavad statistilise olulisuse piiri viitavad genoomi piirkondadele, mis võivad olla seotud uuritava fenotüübiga (Holtz, 2021) .

1.4 Linkage disequilibrium patterns

Aheldustasakaalutus (ingl *linkage disequillibrium*, lühend LD) seletab alleelide mittejuhuslikku seost erinevates lookustes. Teisisõnu, teatud alleelide kombinatsioonid esinevad populatsioonis kas sagedamini või harvemini, kui oleks ootuspärane täiesti juhusliku kombinatsiooni korral (Slatkin, 2008). Aheldustasakaalutust saab kirjeldada matemaatilise parameetri D abil, mis näitab tegelike ja eeldatavate haplotüüpide sageduste vahet, juhul kui alleelide edasikandumine toimuks sõltumatult. Kui D = 0, on süsteem tasakaalus; D ≠ 0 korral viitab see aheldustasakaalutusele (Lewontin & Kojima, 1960).

LD mängib keskset rolli GWAS uuringutes tänu genoomipiirkonna assotsiatsiooni kaardistamisele, kus tegelik põhjuslik SNP võib jääda tuvastamata (Weiss & Silverman, n.d.). Tugevad LD-seosed võivad mõjutada fenotüübi avaldumist viisil, mida genotüübilised andmed otseselt ei peegelda (Aissani, 2014). Siinkohal tuleb mõista, et teatud aheldunud

alleelide koosmõju, eriti olukordades, kus üks neist on epistaatiline - ehk takistab teise lookuse alleeli avaldumist (Phillips, 2008), võib jääda klassikalistes assotsiatsiooniuuringutes märkamata (Levitan, 1955).

LD mustrid varieeruvad märkimisväärselt erinevate inimpopulatsioonide ja kohortide vahel, mõjutades oluliselt GWAS-uuringute tulemuste tõlgendatavust ja replitseeritavust eri populatsioonides (Teo et al., 2009). Tuleb arvestada, et LD mustrid ei ole eri populatsioonide vahel otseselt võrreldavad, kuna alleelisagedused varieeruvad. Varieeruvuse tõttu tuleb GWAS-uuringute kavandamisel kasutada kohortidele spetsiifilisi LD mustreid (Goddard et al., 2000). Seevastu rekombinatsioonisagedused on eri rahvastikes üldiselt sarnased (Serre et al., 2005). Rekombinatsioonisageduse määrad mõjutavad otseselt LD mustrite kujunemist ning tasub assotsiatsiooniuuringutes kaaluda täpseid rekombinatsioonikaardistusi (Evans & Cardon, 2005) .

1.5 Fenotüübi efektiivse lahjendatuse hindamine

GWAS valimite suurenemine on osutunud võimalikuks tänu suurte biopankade loomisele, kuid andmehulkade eksponentsiaalne kasv toob endaga paratamatult kaasa ka rohkem müra – ebakvaliteetseid või ebatäpseid andmeid. Kuna GWAS-uuringute tulemuste usaldusväärsus sõltub suuresti algandmete korrektsusest (Laurie et al., 2010), siis näiteks fenotüübi ebatäpne määratlemine võib oluliselt moonutada nende põhjal tehtud järeldusi ja seega piirata nende praktilist väärtust (Barendse, 2011). Näiteks võib metaanalüüsis fenotüübi ebatäpsus tuleneda sellest, et ühes kohordis on see indiviidi enda poolt raporteeritud ja teises määratud meditsiinitöötaja poolt (Zhou et al., 2022). Ka imputatsioonipaneelide valik mõjutab GWAS uuringute tulemusi, sest paneelide erinev katvus ja täpsus mõjutavad imputatsiooni kvaliteeti ning omakorda võivad mõjutada edasisi genoomiüleste andmeanalüüside efektisuuruste hinnanguid (Chundru et al., 2019).

PheMED (ingl *Phenotypic Measurement of Effective Dilution*) on statistiline meetod ja tööriist, mis võimaldab hinnata GWAS sisendandmete kvaliteeti, hinnates fenotüübi ebatäpset määratlemist/mõõtmist ja sellest tingitud efektisuuruste lahjendatust. Statistiline mudel võimaldab igale individuaalsele GWAS-i uuringule määrata fenotüübi relatiivse

usaldusväärsuse, parandades niiviisi lõpliku metaanalüüsi statistilist täpsust (Burstein et al., 2023).

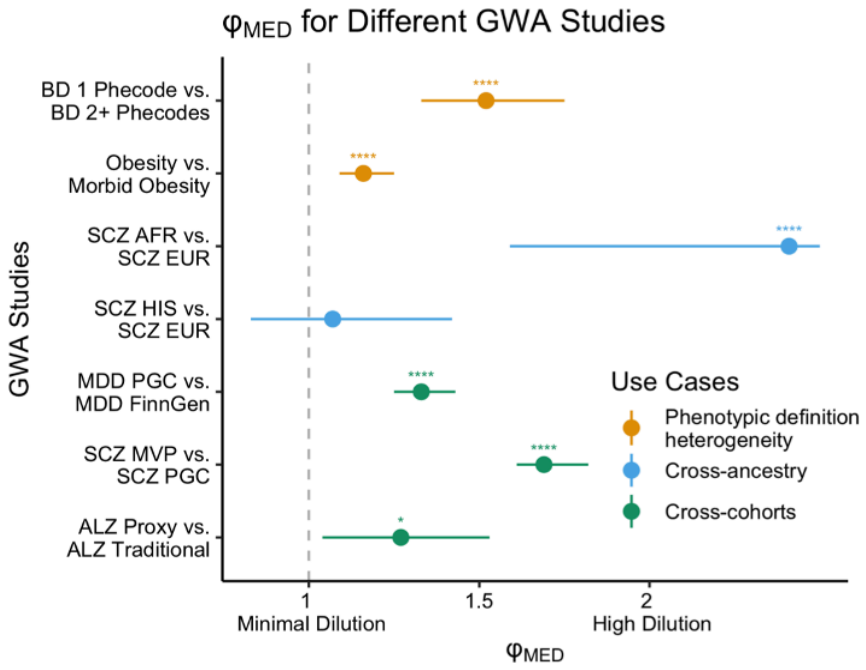
PheMEDi abil saab läbi viia ka korrigeeritud metaanalüüsi, seda algoritmi nimetatakse DAW-iks (ingl *dilution-adjusted weights*). See põhineb suurima tõepära meetodil, kus GWAS-i uuringutele määratakse kaalud vastavalt sellele, kui suur hinnanguliselt fenotüübi ebatäpsus igas GWASis on (Burstein et al., 2023). PheMEDi tugevus seisneb selles, et see töötab ainult GWAS-i kokkuvõtvate statistikute põhjal ja ei vaja geneetlisi indiviidi tasandil sisendandmeid. GWASi kokkuvõtvates statistikutes on on meil teada iga SNP-i vastav efektisuurus ehk beeta (β) uuritava fenotüübi suhtes. PheMED eeldab lihtsustatult, et iga meile teadaolev beeta võrdub SNP-i tegeliku efekti suuruse ja lahjendusteguri jagatisega (beeta = tegelik / lahjenduskordaja (φ)).

PheMED algoritmi keskmes on arusaam, et fenotüübi ebatäpne klassifitseerimine vähendab SNP-de eeldatavaid efektisuurusi läbi kogu andmestiku ühtlase kordajaga (Duffy, 2004). Lahjenduskordajate (φ) ja tegelike efektisuuruste leidmiseks kasutab PheMED baasuuringut, mille lahjenduskordajaks määratakse väärtus 1. Teiste GWAS-uuringute lahjenduskordajad leitakse selle baasuuringu suhtes, ehk lahjendatus ja fenotüübi määratluse täpsuse hinnang on alati relatiivne baasuuringu suhtes (Joonis 3). Lahjenduskordaja (φ) arvutamiseks saaks rakendada järgnevat valemit:

$$\varphi_{MED,1,3} = \frac{(PPV_1 + NPV_1 - 1)}{(PPV_3 + NPV_3 - 1)} ,$$

kus valemis:
1 - määrab GWAS-uuringut, mille jaoks leitakse lahjenduskordaja;
3 - tähistab baas GWAS-uuringut;
$\varphi_{MED,1,3}$ - on baasuuringu (3) põhjal määratud lahjenduskordaja uuritava GWAS-i (1) jaoks;

PPV (ingl *positive predictive value*) positiivne prognoosiväärtus on tõenäosus, et diagnostilise testiga positiivse tulemuse saanud isikul on uuritav omadus (nt haigus);
NPV (ingl *negative predictive value*) negatiivne prognoosiväärtus on tõenäosus, et negatiivse testitulemuse saanud isikul seda omadust/haigust ei esine.



Joonis 3. Erinevatest genoomiülestest andmeanalüüsidest saadud fenotüüpide lahjenduskordajate võrdlused. Graafiku X-teljel on märgitud PheMED programmi lahjendatuse hinnang, kus väärtus 1 tähistab minimaalset lahjendatust ja väärtus 2 kõrget lahjendatust (φ_MED). Graafiku y-teljel ülevalt alla on on märgitud võrdlused kahe erineva GWAS-si vahel.
Kasutatud lühendid: **BD** - bipolaarne häire; **Phecode** - fenotüübi kood; **Obesity** - ülekaalulisus; **Morbid obesity**- haiguslik rasvumine; **SCZ** - skisofreenia; **AFR** - Aafrika päritolu; **EUR** - Euroopa päritolu; **HIS** - Hispaania päritolu; **MDD** - depressioon; **PGC** - Psühhiaatrilise Genoomika Konsortium; **FinnGen** - Soome biopankade geeniprojekt; **MVP** - USA veteranide geeniprojekt; **ALZ** - Alzheimeri tõbi; **Proxy** - kaudselt määratud fenotüüp, näiteks pereliikme haiguse põhjal määratud fenotüüp; **Traditional** - fenotüüp määratud otse meditsiiniliste andmete alusel.
Erinevad metoodilised rakendused:
Oranž- φ_MED kahe erineva fenotüübi vahel (haiguslik rasvumine versus rasvumine).
Sinine- φ_MED kahe eri populatsiooni GWAS-de vahel (AFR versus EUR).
Roheline- φ_MED eri kohortide skisofreenia GWAS-de vahel (PGC versus MVP).
Joonise algallikas: Burstein et al., 2023.

1.5.1 Lahjenduskordaja arvutus kattuvate valimite korral

Olukorras, kus tahetakse võrrelda kahte uuringut, milles sisalduvad samad indiviidid, on vajalik kolmas uuring, mis on eelnevast kahest sõltumatu. Efektiivse lahjendatuse hindamine esimese ja kolmanda uuringu vahel järgmise valemi järgi:

$$\varphi_{MED,1,3} = \frac{(PPV_1 + NPV_1 - 1)}{(PPV_3 + NPV_3 - 1)} \ .$$

Lahjendatuse hindamine teise ja kolmanda uuringu vahel:

$$\varphi_{MED,2,3} = \frac{(PPV_2 + NPV_2 - 1)}{(PPV_3 + NPV_3 - 1)} .$$

Eelnevate uuringute väärtuste jagatis annab lahjendatuse hinnangu esimese ja teise uuringu vahel:

$$\varphi MED, 1, 3 / \varphi MED, 2, 3 = \frac{(PPV_1 + NPV_1 - 1)}{(PPV_2 + NPV_2 - 1)} = \varphi_{MED,1,2^*}$$ (valem 1).

2. Eksperimentaalosa

2.1 Töö eesmärgid

Käesoleva bakalaureusetöö peamine eesmärk oli võrrelda kahe imputatsiooni referentspaneeli erinevusi genoomiüleste andmeanalüüside tulemustel PheMED metoodika ja tarkvara abil Eesti Biopanga andmetel. Töö alaeesmärgiks oli võrrelda GWAS tulemuste erinevusi sagedaste ja harudlaste SNP-de kategooriate vahel.

2.2 Materjal ja metoodika

2.2.1 Eesti Biopanga andmete ülevaade

Eesti Biopank (EstBB) on rahvastikupõhine geenivaramu, mis koosneb enam kui 200 000 täiskasvanu andmetest. EstBB sisaldab mitmekesised andmekihte - fenotüübi ja genotüübi andmed; erinevad kliinilised mõõtmised; digiretseptide, raviprotseduuride ja haigusdiagnooside ajalugu; mikrobioomi andmed jms (Milani et al., 2025). Geenidoonorid liitusid EstBB-ga kahes suuremas kogumislaines:

Esimene laine (2002–2011): liitus umbes **52 000 indiviidi**, kellelt koguti elustiili, tervisekäitumise, haigusloo andmed ja koguti bioloogilist materjali.
Teine laine (2018–2019): liitus ligi **150 000 uut doonorit**. Küsimustik, millele liituja vastas, oli palju lühem, enamik andmeid saadi tänu EstBB andmete ühendamisele erinevate registritega (nagu näiteks Tervisekassa, vähiregister, müokardiinfarktiregister ja Eesti kahe suurima haigla Põhja-Eesti Regionaalhaigla ja Tartu Ülikooli Kliinikumi andmestikega) ja läbi erinevate uuringuküsimustike (nagu näiteks COVID-19 uuring, vaimse tervise ja heaolu uuring HEVT, isiksuseuuring PS21 ja ravimite ning vaktsiinide kõrvaltoimete uuring ADE-Q). Registritest saadakse infot isiku raviretseptide, raviarvete, diagnooside ja epikriiside kohta.

See bakalaureusetöö viidi läbi 204,742 EstBB indiviidi andmetel. See valim oli juba eelneva uuringu jaoks puhastatud ja igale indiviidile oli arvutatud kehamassiindeks (lühend KMI, $\frac{1{,}3 \cdot kehakaal}{pikkus}$). Kaastud olid veel tunnused nagu vanus, sugu ja geneetilised peakomponedid. Uuritavate isikute hulgas oli 133 634 naist (65,3%) ja 71 108 meest

(34,7%). Uuritavate keskmine vanus oli 43,6 aastat ning keskmine kehamassiindeks 26,10 (25,74 naiste ja 26,77 meeste puhul).

2.2.2 Ülegenoomne uuring EstBB andmetel

Eesti biopanga andmetel viidi läbi neli GWASi, milles uuritavaks tunnuseks oli kehamassiindeks (KMI). Genotüübi andmed olid imputeeritud kahe eri referentspaneeliga – Eesti geenidoonorite (n = 2244) (Mitt et al., 2017) täisgenoomide sekveneerimisandmestiku põhjal koostatud referentspaneeliga (EstREF) (Milani et al., 2025) ja HRC referentspaneeliga imputeeritud andmetel (n = 32 470) (Michigan Imputation Server 2, 2024). GWAS viidi läbi REGENIE (versioon 3.6p) programmiga (Regenie documentation, 2021), kus lisaks uuritavatele SNP-idele olid mudeli sõltumatuteks tunnusteks kümme geneetilist peakomponenti, sugu ja vanus ning uuritavaks tunnuseks oli KMI. GWASi tehti segamudeli abil, mis võtab arvesse, et geenidoonorid on omavahel suguluses. GWAS-uuringute tulemfaili formaat näeb välja järgmine (Tabel 2).

Tabel 2. GWAS-uuringu tulemfaili päise näidis

CHR	BP	SNP	A1	A2	N	SE	P	BETA	INFO	MAF
1	756604	rs3131962	A	G	388028	0,00302	0,48317	0,99789	0,89056	0,36939
1	768448	rs12562034	A	G	388028	0,00329	0,83481	1,00069	0,89590	0,33685
1	779322	rs4040617	G	A	388028	0,00303	0,42897	0,99760	0,89751	0,37737

CHR – kromosoom, kus assotsiatsiooniuuringutesse kaasatud SNP asub; **BP** – SNP-i kromosoomi koordinaat (ingl *base pair*); **SNP** – SNP-i tunnuskood, tavaliselt kujul rs-ID; **A1** – SNP-i efektialleel; **A2** – SNP-i mitte-efektialleel; **N** – proovide arv, mida kasutati efektisuuruse hindamiseks; **SE** – efektisuuruse standardviga; **P** – SNP-i genotüübi ja uuritava fenotüübi suhteline p-väärtus; **BETA** – efektisuurus riskialleeli kohta; **INFO** – imputatsiooni kvaliteediskoor; **MAF** – SNP-i harvem esineva alleeli sagedus.

Antud bakalaureusetöös jagati esialgne valim suurusega n = 204742 kaheks: 1) treening- (n = 102371) ja 2) testandmed (n = 102371). Töös edaspidi nimetatud EstBBtreening ja EstBBtest. Andmestiku kaheks jagamine on vajalik, sest fenotüübi efektiivse lahjendatuse hindamiseks peavad võrreldavad uuringud olema teineteisest sõltumatud. Seega tehti neli GWAS-si: 1) EstBBtreening andmetel, mis olid imputeeritud EstREF referentspaneeli abil (lühend EstBBtreening-EstREF), 2) EstBBtreening andmetel, mis olid imputeeritud HRC referentspaneeli abil (lühend EstBBtreening-HRC), 3) EstBBtest

andmetel, mis olid imputeeritud EstREF referentspaneeli abil (EstBBtest-EstREF) ja 4) EstBBtest andmetel, mis olid imputeeritud HRC referentspaneeliga (EstBBtest-HRC).

2.2.3 Fenotüübilise lahjendatuse hindamine PheMED abil EstBB andmetel

Selleks, et lahjendatust hinnata, viidi PheMED analüüs läbi kaks korda - esimene kord EstBBtreening–HRC versus EstBBtest–EstREF ($\varphi_{MED,1,3} = \frac{(PPV_1+NPV_1-1)}{(PPV_3+NPV_3-1)}$) ja teine kord EstBBtreening–EstREF versus EstBBtest–EstREF ($\varphi_{MED,2,3} = \frac{(PPV_2+NPV_2-1)}{(PPV_3+NPV_3-1)}$). Selleks, et nüüd eri imputatsiooni referentpaneelide võimalikku lahjendatust hinnata, saab rakendada (Burstein et al., 2023) valemit 1, mis on toodud leheküljel 16, ehk φ_med,1,2 hindamiseks kasutame suhet φ_med 1,3 / φ_med 2,3. Efektisuuruste lahjendatust hinnati kolmes eri kategoorias kasutades valemit 1:

1. Kogu GWAS-i SNP-d
2. GWAS SNP-de alamhulk, kus on MAF<=5%
3. GWAS SNP-de alamhulk, kus on MAF<=1%

2.3 Tulemused ja arutelu

EstBBtest-EstREF ning EstBBtreening-HRC tehtud GWAS-i võrdlus näitas, et mõlemas analüüsis kattus 14 124 553 SNP-i. Esmalt võrreldakse, kas kahe erineva referentspaneeli abil imputeeritud andmetel on SNP-del samad alleelid (A1 ja A2) valitud, et nende efektisuurusi oleks võimalik otse võrrelda. Kuna mittekattuvaid alleele (A1 ja A2) oli kahe referentspaneeli korral vähe (~33 000), siis ümberkodeerimist ei tehtud ning kõik mittekattuvad alleelid eemaldati enne edasist analüüsi. Beetade korrelatsioon kahe GWAS-i vahel oli kõikide SNP-de lõikes 0,014 (Tabel 3). Nii madal korrelatsiooniväärtus näitab, et kahe erineva referentspaneeliga saadud SNP-de efektisuurused on väga erinevad. Analüüs, milles piirasime mõlemas andmestikus haruldase alleeli sageduse väärtusega ≥5%, tõusis korrelatsioon 0,35 peale. Valides ainult MAF vahemikus 1%...5%, langes korrelatsiooni väärtus 0,0134 peale ja kui MAF<0.1%, siis on korrelatsioon 0.013. See viitab haruldaste SNP-de madalale imputatsiooni kvaliteediskoorile ja suuremale varieeruvusele nende

efektide osas. Kui vaadata väiksemat osa kogu SNP-dest (MAF≥5% ja nende imputatsiooni kvaliteediskoor >0,8 ja mille p-väärtus on < 0,1), siis on korrelatsioon beetade vahel palju kõrgem, 0,6, kinnitades, et peamine põhjus madalaks üleüldiseks korrelatsiooniks betade vahel on just madala infoskooriga haruldased SNPid. Võrreldi ka test- ja treeningandmete efektialleeli sagedust, ning selle võrdlusel korrelatsioon oli 0,999.

Tabel 3. EstBBtest-EstREF GWAS-i ja EstBBtreening-HRC GWAS-i beetade korrelatsioonid kõikide GWAS tulemuste ja minoorse alleelisageduse alamrühmades

MAF-i väärtus (%)	Korrelatsioonikordaja
Kõik SNP-id	0,014
≥ 5%	0,35
<5%	0,0134
< 1%	0,013

GWASide tulemused koondati paarikaupa ühte faili ja sooritati PheMED analüüs. Kui PheMed lahjenduskordaja ($\varphi_{MED,1,3} = \frac{(PPV_1+NPV_1-1)}{(PPV_3+NPV_3-1)}$) on suurem kui 1, siis see näitab, et võrdlusuuringus on keskmiselt väiksemad beetad kui baasuuringus. Usaldusvahemik (95%CI) näitab, kas erinevus on statistiliselt oluline - kui väärtus "1" on usaldusvahemikus sees, siis statistiliselt olulist erinevust beetade keskmise osas ei ole.

Tabel 4. PheMED φ hinnangud erinevate baas- ja võrdlusuuringute lõikes				
Baasuuring	**Võrdlusuuring**	**MAF**	**PheMED φ**	**95% CI**
EstBBtest-EstREF	EstBBtreening-EstREF	kõik	1,029	(1,004;1,056)
EstBBtest-EstREF	EstBBtreening-HRC	kõik	1,025	(1,004;1,056)
EstBBtest-EstREF	EstBBtreening-EstREF	<5%	1,055	(0,993;1,118)
EstBBtest-EstREF	EstBBtreening-HRC	<5%	1,062	(1,001;1,126)
EstBBtest-EstREF	EstBBtreening-EstREF	<1%	1,254	(1,042;1,510)
EstBBtest-EstREF	EstBBtreening-HRC	<1%	1,231	(0,990;1,521)

Tabelist 4 on näha, et võrreldes GWAS tulemusi EstBBtest-EstREF versus EstBBtreening-EstREF ja EstBBtest-EstREF versus EstBBtreening-HRC, siis olid tulemused statistiliselt oluliselt erinevad ($p<0,05$). Mõlemas analüüsis oli treeningandmestiku beetad veidi väiksemad kui testandmete beetad ($\varphi>1$). Vaadeldes väiksemat osa SNP-dest (MAF<5%), siis efektisuuruste lahjendatus oli statistiliselt oluline EstBBtest-EstREF vs EstBBtreening-HRC puhul. Vaadeldes SNP-e MAF<1%, siis oli efektisuuruste lahjendatuse hinnang ~1,2, mis näitab, et EstBBtreening-EstREF kui ka EstBBtreening-HRC beetad olid testandmetega võrreldes väiksemad. Kuna 95% usaldusvahemik ulatub EstBBtreening-HRC puhul 0,990 kuni 1,521, siis efektisuuruse lahjendatuse hinnangust ei saa väita, et beetade väärtused oleksid väiksemad, kuna erinevus pole statistiliselt oluline. EstREF referentspaneeli puhul on erinevus statistiliselt oluline - analüüsist lähtub, et EstBBtest-EstREF ja EstBBtest-HRC beetad on veidi suuremad kui EstBBtreening-EstREF ja EstBBtreening-HRC beetad seda isegi juhul kui kahel erineval GWAS-l on kasutatud sama imputatsiooni referentspaneeli.

Võttes tulemused kokku, leitud PheMED hinnangud näitavad, et fenotüübi ebatäpne määratlemine on kõigi variantide lõikes minimaalne, umbes 1,03, kuid suureneb haruldaste alleelide puhul (MAF <1%), kus treeningandmete (EstBBtreening-EstREF ja EstBBtreening-HRC) puhul SNP-de efektid suurenevad 1,25 korda võrreldes testandmetega (EstBBtest-EstREF). See viitab, et erinevuse allikaks võib olla andmestiku jagamine treeninguks ja testiks, mitte referentspaneelide EstREF või HRC valik. Oluline on arvestada haruldaste SNP-de võimaliku ebavõrdse jagunemisega andmestikkude lõikes. Tulemustest ei ole võimalik hetkel haruldaste SNP-de puhul midagi märkimisväärset järeldada.

Kasutades PheMED analüüsi juures valemit, kus φ_med_2,1=φ_2,3/φ_1,3, saame järgmised tulemused referentspaneelide võrdluses:

Tabel 5. EstBBtreening-HRC ja EstBBtreening-EstREF vahelised φ-hinnangud

Baasuuring	Võrdlusuuring	MAF	φ_med_1,2
EstBBtreening-EstREF	EstBBtreening-HRC	kõik	0,9960015
EstBBtreening-EstREF	EstBBtreening-HRC	<5%	1,007188
EstBBtreening-EstREF	EstBBtreening-HRC	<1%	0,9815138

Kasutades kõiki SNP-e, siis on φ-hinnang 0,996, mis on väga lähedane väärtusele 1. See näitab, et keskmiselt on mõlemad referentspaneelid beetad samas suurusjärgus. Uurides MAF <5% vahemikku, saame φ-hinnanguks 1,007. MAF-i <1% puhul saame φ-hinnanguks 0,982 - see näitab, et keskmiselt on HRC referentspaneeli kasutades GWAS beetad natuke suuremad kui EstREF referentspaneeli kasutades, kuid statistilist olulisust olemasoleva valemi abil hinnata ei saa. Seega vaadeldes vaid phemedi hinnanguid, mis on ühele väga lähedased, võiks eeldada, et EstBB genoomi andmete puhul EstREF ja HRC referentspaneelide imputatsioonikvaliteet ja täpsus omavahel sarnased nii haruldaste kui ka sagedaste SNP-ide lõikes, kuigi statistilise olulisuse kohta midagi öelda ei saa.

Kokkuvõte

Käesolev bakalaureusetöö käsitleb erinevate imputatsiooni referentspaneelide mõju kehamassiindeksit uurivate ülegenoomsete assotsiatsiooniuuringute (GWAS) tulemustele. Referentspaneel koosneb sekveneeritud genoomidest, milles sisalduvate haplotüübide abil ennustatakse genotüpiseerimata jäänud SNP-e. Lisaks käistletakse antud töös PheMED-i statististilist mudelit, mis hindab GWAS-ides uuritava fenotüübi ebatäpsust. PheMED-i saab rakendada järelanalüüsis nii imputeeritud andmete kvaliteedi kui ka sellest tuleneva fenotüübi ebtäpsuse hindamiseks.

Töö peamine eesmärk oli võrrelda Eesti populatsioonipõhise EstREF-i ja rahvusvahelise HRC referentspaneeli mõju kehamassiindeksi GWAS-i efektisuurustele ning hinnata fenotüübilist lahjendatust PheMED-i abil Eesti Biopanga andmetel. Selleks jagati 204 743 geenidoonori andmestik pooleks - treening- ja testandmeteks, kus genotüpiseerimata jäänud SNP-d olid imputeeritud ühel juhul EstREF ja teisel juhul HRC referentspaneeliga. Seejärel viidi läbi neli sõltumatut GWAS-i, kus lisaks uuritavatele SNP-idele võeti arvesse vanus, sugu ja kümme geneetilist peakomponenti.

EstBBtest-EstREF ja EstBBtreening-HRC lõikes kattus 14 124 553 miljonit SNP-i. Ligikaudu 33 000 mittekattuvat alleeli eemaldati enne edasist võrdlust. Kõigi SNP-ide puhul oli beetade korrelatsioon mõlema paneeli andmete vahel madal (b = 0,014), kuid sagedaste variantide (MAF≥5%; INFO>0,8; p<0,1) puhul tõusis see 0,6-ni, osutades paneelide suuremale sarnasusele sagedaste SNP-ide imputeerimisel.

PheMED-i analüüs, mille referentsiks oli EstBBtest-EstREF, näitas, et treeningandmete beetad olid keskmiselt väiksemad kui testandmetel (φ = 1,029; p<0,05). EstBBtreening-EstREF (MAF<5%) võrdlusuuringute puhul oli tulemus sarnane, kuid ei olnud enam statistiliselt oluline. Haruldaste variantide (MAF<1%) puhul oli EstBBtreening-EstREF lahjendus referentsuuringu suhtes φ = 1,25. EstBBtreening-HRC võrdlusuuringute puhul näeme samu mustreid, kuid haruldaste variantide puhul ei olnud efekt enam statistliselt oluline.

Otsevõrdlus EstBBtreening-HRC ja EstBBtreening-EstREF-i vahel (φ_MED EstREF vs HRC) andis kõigi SNP-ide lõikes tulemuse φ = 0,996, mis näitab, et beetad on samas suurusjärgus. Sagedaste variantide korral oli φ = 1,007, kuid haruldaste SNP-ide puhul

langes väärtus 0,982-ni, viidates EstREF ja HRC referentspaneelide sarnasusele. Suuri erinevusi üldiselt ega ka MAF-ide erikategooriates ei õnnestunud leida. Töös rakendatud metoodika ei võimaldanud otsevõrdlustele statistlise olulisuse vahemikku arvutada.

Kokkuvõttes saadud tulemuste põhjal ei saa väita, et populatsioonispetsiifiline EstREF ega HRC referentspaneel mõjutaks KMI GWAS-ide beetade absoluutväärtuseid. Antud töö näitab PheMED-i kasulikkust fenotüübilist lahjendust arvestava kvaliteedikontrollina, mis võib suurendada biopankade GWAS-ide statistilist võimsust ja metaanalüüside usaldusväärsust.

The impact of different imputation panels on genome-wide association study results

Alissa Kazmin

Summary

The aim of this bachelor´s thesis is to determine the influence of different imputation reference panels on genome-wide association studies (GWAS) of body-mass index (BMI). A reference panel is a set of sequenced genomes, the haplotypes of which are used to predict ungenotyped SNPs. In addition, here is introduced the PheMED statistical model, which estimates the phenotypic misclassification of GWASs, and can be applied in post-analysis to assess both the quality of imputed data and the resulting phenotypic misclassification.

The main goal of the work was to compare the impact of the Estonian-specific EstREF and the international HRC reference panels on effect sizes in an Estonian Biobank (EstBB) body mass index (BMI) GWASs, and to assess phenotypic dilution with PheMED. The dataset of 204 743 biobank participants was divided into test and training sets and each dataset was imputed with both EstREF and HRC reference panels. Four independent GWASs were conducted, incorporating age, sex, and ten genetic principal components alongside the SNPs of interest.

A total of 14 124 553 SNPs overlapped between the EstBBtest-EstREF dataset and the EstBBtrain-HRC dataset. Approximately 33 000 non-overlapping alleles were removed before further comparison. For all SNPs, the correlation of beta coefficients between the data from both reference panels (EstREF and HRC) was low ($r = 0{,}014$), but increased to 0,6 for frequent variants (MAF≥5%; INFO>0,8; $p<0{,}1$), indicating similar performance between the panels when imputing frequent SNPs.

The Phemed analysis using EstBBtest-EstREF as reference, showed that training set beta values were, on average, smaller than those in the test set ($\varphi = 1{,}029$; $p<0{,}05$). For the EstBBtrain-EstREF (MAF<5%) comparison, the result was similar, but was no longer statistically significant. For rare variants (MAF<1%), the dilution relative to EstBBtraining-EstREF to the reference study was $\varphi = 1{,}25$. EstBBtrain-HRC comparisons displayed similar trends, though the effect for rare variants was no longer statistically significant as well.

A direct comparison between EstBBtrain-EstREF and EstBBtrain-HRC (φ_MED EstREF vs HRC) yielded a result of φ=0,996 for all SNPs, indicating that betas are of the same order of magnitude. For frequent variants φ equaled 1,007, but for rare SNPs φ dropped to 0,982, suggesting similarity between the EstREF and the HRC reference panels. The methodology used in the work did not allow to calculate statistical significance range for direct comparisons.

Overall, thesis results don't confirm that the population-specific EstREF and HRC reference panels affect the absolute values of betas in BMI GWAS. This thesis also demonstrates the utility of PheMED as a quality control that considers for phenotype dilution, potentially increasing the statistical power of biobank GWASs and the reliability of meta-analyses.

Tänusõnad

Soovin siiralt tänada Kristi Lälli ja Katri Pärna väga hea koostöö, juhendamise ja kiire ning sisuka tagasiside eest. Suur tänu Kristile, kes aitas mind lõputöö eksperimentaalse osa programmide käivitamisega.

Teie teadmised, kannatlikkus ja toetus on olnud hindamatud lõputöö kirjutamisel ning olen siiralt tänulik meie koostöö eest.

Soovin samuti tänada Maido Remmi, kes suunas mind Genoomika instituuti.

Kasutatud kirjandus

Alwateer, M., Atlam, E.-S., El-Raouf, M. M. A., Ghoneim, O. A., & Gad, I. (2024). Missing Data Imputation: A Comprehensive Review. *Journal of Computer and Communications*, *12*(11), 53–75. https://doi.org/10.4236/jcc.2024.1211004

Aissani, B. (2014). Confounding by linkage disequilibrium. *Journal of Human Genetics*, *59*(2), 110–115. https://doi.org/10.1038/jhg.2013.130

Avery, O. T., & McCARTY, M. (n.d.). *STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES*.

Bagger, F. O., Borgwardt, L., Jespersen, A. S., Hansen, A. R., Bertelsen, B., Kodama, M., & Nielsen, F. C. (2024). Whole genome sequencing in clinical practice. *BMC Medical Genomics*, *17*(1), 39. https://doi.org/10.1186/s12920-024-01795-w

Barendse, W. (2011). The effect of measurement error of phenotypes on genome wide association studies. *BMC Genomics*, *12*(1), 232. https://doi.org/10.1186/1471-2164-12-232

Bush, W. S., Moore, J. H. (2012). *Chapter 11: Genome-Wide Association Studies*

Burstein, D., Hoffman, G., Mathur, D., Venkatesh, S., Therrien, K., Fanous, A. H., Bigdeli, T. B., Harvey, P. D., Roussos, P., & Voloudakis, G. (2023). *Detecting and Adjusting for Hidden Biases due to Phenotype Misclassification in Genome-Wide Association Studies*. Genetic and Genomic Medicine. https://doi.org/10.1101/2023.01.17.23284670

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Paul Flicek, Germer, S., Brand, H., Hall, I. M., Talkowski, M. E., … Xiao, C.

(2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, *185*(18), 3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004

Campbell, N. R., Harmon, S. A., & Narum, S. R. (2015). Genotyping‐in‐Thousands by sequencing (GT‐seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, *15*(4), 855–867. https://doi.org/10.1111/1755-0998.12357

Casals, F., Idaghdour, Y., Hussin, J., & Awadalla, P. (2012). Next-generation sequencing approaches for genetic mapping of complex diseases. *Journal of Neuroimmunology*, *248*(1–2), 10–22. https://doi.org/10.1016/j.jneuroim.2011.12.017

Castle, W. E. (1903). Mendel's Law of Heredity. *Science*, *18*(456), 396–406. https://doi.org/10.1126/science.18.456.396

Choudhury, A., Hazelhurst, S., Meintjes, A., Achinike-Oduaran, O., Aron, S., Gamieldien, J., Jalali Sefid Dashti, M., Mulder, N., Tiffin, N., & Ramsay, M. (2014). Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. *BMC Genomics*, *15*(1), 437. https://doi.org/10.1186/1471-2164-15-437

Chundru, V. K., Marioni, R. E., Prendergast, J. G. D., Vallerga, C. L., Lin, T., Beveridge, A. J., SGPD Consortium, Gratten, J., Hume, D. A., Deary, I. J., Wray, N. R., Visscher, P. M., & McRae, A. F. (2019). Examining the Impact of Imputation Errors on Fine-Mapping Using DNA Methylation QTL as a Model Trait. *Genetics*, *212*(3), 577–586. https://doi.org/10.1534/genetics.118.301861

Collins, F. S., Brooks, L. D., & Chakravarti, A. (1998). A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation: Table 1. *Genome Research*, *8*(12), 1229–1231. https://doi.org/10.1101/gr.8.12.1229

Das, S., Abecasis, G. R., & Browning, B. L. (2018). Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics*, *19*(1), 73–96. https://doi.org/10.1146/annurev-genom-083117-021602

Do, C. B., Hinds, D. A., Francke, U., & Eriksson, N. (2012). Comparison of Family History and SNPs for Predicting Risk of Complex Disease. *PLoS Genetics*, *8*(10), e1002973. https://doi.org/10.1371/journal.pgen.1002973

Duffy, S. W. (2004). A simple model for potential use with a misclassified binary outcome in epidemiology. *Journal of Epidemiology & Community Health*, *58*(8), 712–717. https://doi.org/10.1136/jech.2003.010546

Evans, D. M., & Cardon, L. R. (2005). A Comparison of Linkage Disequilibrium Patterns and Estimated Population Recombination Rates across Multiple Populations. *The American Journal of Human Genetics*, *76*(4), 681–687. https://doi.org/10.1086/429274

Galas, D. J., & McCormack, S. J. (n.d.). *An Historical Perspective on Genomic Technologies*.
Gayon, J. (2016). From Mendel to epigenetics: History of genetics. *Comptes Rendus. Biologies*, *339*(7–8), 225–230. https://doi.org/10.1016/j.crvi.2016.05.009

Goddard, K. A. B., Hopkins, P. J., Hall, J. M., & Witte, J. S. (2000). Linkage Disequilibrium and Allele-Frequency Distributions for 114 Single-Nucleotide Polymorphisms in Five Populations. *The American Journal of Human Genetics*, *66*(1), 216–234. https://doi.org/10.1086/302727

Gurdasani, D., Carstensen, T., Fatumo, S., Chen, G., Franklin, C. S., Prado-Martinez, J., Bouman, H., Abascal, F., Haber, M., Tachmazidou, I., Mathieson, I., Ekoru, K., DeGorter, M. K., Nsubuga, R. N., Finan, C., Wheeler, E., Chen, L., Cooper, D. N., Schiffels, S., … Sandhu, M. S. (2019). Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell*, *179*(4), 984-1002.e36. https://doi.org/10.1016/j.cell.2019.10.004

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A., & Cox, D. R. (2005). Whole-Genome Patterns of Common DNA Variation in Three Human Populations. *Science*, *307*(5712), 1072–1079. https://doi.org/10.1126/science.1105436

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype Imputation with Thousands of Genomes. *G3 Genes|Genomes|Genetics*, *1*(6), 457–470. https://doi.org/10.1534/g3.111.001198

Hutchison, C. A. (2007). DNA sequencing: Bench to bedside and beyond. *Nucleic Acids Research*, *35*(18), 6227–6237. https://doi.org/10.1093/nar/gkm688

International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research:, Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Jannot, A.-S., Ehret, G., & Perneger, T. (2015). $P < 5 \times 10{-}8$ has emerged as a standard of statistical significance for genome-wide association studies. *Journal of Clinical Epidemiology*, *68*(4), 460–465. https://doi.org/10.1016/j.jclinepi.2015.01.001

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, *22*(2), 139–144. https://doi.org/10.1038/9642

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., Boehm, F., Caporaso, N. E., Cornelis, M. C., Edenberg, H. J., Gabriel, S. B., Harris, E. L., Hu, F. B., Jacobs, K. B., Kraft, P., Landi, M. T., Lumley, T., Manolio, T. A., McHugh, C., … for the GENEVA Investigators. (2010). Quality control and quality assurance in genotypic data for genome‐wide association studies. *Genetic Epidemiology*, *34*(6), 591–602. https://doi.org/10.1002/gepi.20516

Levitan, M. (1955). STUDIES OF LINKAGE IN POPULATIONS. I. ASSOCIATIONS OF SECOND CHROMOSOME INVERSIONS IN *DROSOPHILA ROBUSTA*. *Evolution*, *9*(1), 62–74. https://doi.org/10.1111/j.1558-5646.1955.tb01514.x

Lewontin, R. C., & Kojima, K. (1960). THE EVOLUTIONARY DYNAMICS OF COMPLEX POLYMORPHISMS,,. *Evolution*, *14*(4), 458–472. https://doi.org/10.1111/j.1558-5646.1960.tb03113.x

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. *Annual Review of Genomics and Human Genetics*, *10*, 387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242

Marchini JL. Haplotype Estimation and Genotype Imputation. In: Handbook of Statistical Genomics. 4th ed.; 2019.

Martin, A. R., Atkinson, E. G., Chapman, S. B., Stevenson, A., Stroud, R. E., Abebe, T., Akena, D., Alemayehu, M., Ashaba, F. K., Atwoli, L., Bowers, T., Chibnik, L. B., Daly, M. J., DeSmet, T., Dodge, S., Fekadu, A., Ferriera, S., Gelaye, B., Gichuru, S., … Zingela, Z. (2021). Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *The American Journal of Human Genetics*, *108*(4), 656–668. https://doi.org/10.1016/j.ajhg.2021.03.012

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E., & Marchini, J. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, *53*(7), 1097–1103. https://doi.org/10.1038/s41588-021-00870-7

Mckusick, V. A. (n.d.). *The Growth and Development of Human Genetics as a Clinical Discipline*.

Passarge, E. (2021). Origins of human genetics. A personal perspective. *European Journal of Human Genetics*, *29*(7), 1038–1044. https://doi.org/10.1038/s41431-020-00785-7

Milani, L., Alver, M., Laur, S., Reisberg, S., Haller, T., Aasmets, O., Abner, E., Alavere, H., Allik, A., Annilo, T., Fischer, K., Hofmeister, R., Hudjashov, G., Jõeloo, M., Kals, M., Karo-Astover, L., Kasela, S., Kolde, A., Krebs, K., … Metspalu, A. (2025). The Estonian Biobank's journey from biobanking to personalized medicine. *Nature Communications*, *16*(1), 3270. https://doi.org/10.1038/s41467-025-58465-3

Mitt, M., Kals, M., Pärn, K., Gabriel, S. B., Lander, E. S., Palotie, A., Ripatti, S., Morris, A. P., Metspalu, A., Esko, T., Mägi, R., & Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics*, *25*(7), 869–876. https://doi.org/10.1038/ejhg.2017.51

O'Connell, J., Yun, T., Moreno, M., Li, H., Litterman, N., Kolesnikov, A., Noblin, E., Chang, P.-C., Shastri, A., Dorfman, E. H., Shringarpure, S., 23andMe Research Team, Aslibekyan, S., Babalola, E., Bell, R. K., Bielenberg, J., Bryc, K., Bullis, E., Coker, D., … McLean, C. Y. (2021). A population-specific reference panel for improved genotype imputation in African Americans. *Communications Biology*, *4*(1), 1269. https://doi.org/10.1038/s42003-021-02777-9

Phillips, P. C. (2008). Epistasis—The essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, *9*(11), 855–867. https://doi.org/10.1038/nrg2452

Pitt. J. J. (2010). Newborn Screening. *The Clinical Biochemist Reviews, 31(2),57-68*

Pruneri, G., De Braud, F., Sapino, A., Aglietta, M., Vecchione, A., Giusti, R., Marchiò, C., Scarpino, S., Baggi, A., Bonetti, G., Franzini, J. M., Volpe, M., & Jommi, C. (2021). Next-Generation Sequencing in Clinical Practice: Is It a Cost-Saving Alternative to a Single-Gene Testing Approach? *PharmacoEconomics - Open*, *5*(2), 285–298. https://doi.org/10.1007/s41669-020-00249-0

Quick, C., Anugu, P., Musani, S., Weiss, S. T., Burchard, E. G., White, M. J., Keys, K. L., Cucca, F., Sidore, C., Boehnke, M., & Fuchsberger, C. (2020). Sequencing and imputation in GWAS: Cost‐effective strategies to increase power and genomic coverage across diverse populations. *Genetic Epidemiology*, *44*(6), 537–549. https://doi.org/10.1002/gepi.22326

Ramírez-Bello, J. (2023). Role of genetic variability in Mendelian and multifactorial diseases. *Gaceta Médica de México*, *155*(5), 3728. https://doi.org/10.24875/GMM.M20000333

Reich, D. E., Gabriel, S. B., & Altshuler, D. (2003). Quality and completeness of SNP databases. *Nature Genetics*, *33*(4), 457–458. https://doi.org/10.1038/ng1133

Rubinacci, S., Hofmeister, R. J., Sousa Da Mota, B., & Delaneau, O. (2023). Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. *Nature Genetics*, *55*(7), 1088–1090. https://doi.org/10.1038/s41588-023-01438-3

Sadee, W., Wang, D., Hartmann, K., & Toland, A. E. (2023). Pharmacogenomics: Driving Personalized Medicine. *Pharmacological Reviews*, *75*(4), 789–814. https://doi.org/10.1124/pharmrev.122.000810

Sayitoğlu, M. (2016). Clinical Interpretation of Genomic Variations. *Turkish Journal of Hematology*, *33*(3), 172–179. https://doi.org/10.4274/tjh.2016.0149

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

Schurz, H., Müller, S. J., Van Helden, P. D., Tromp, G., Hoal, E. G., Kinnear, C. J., & Möller, M. (2019). Evaluating the Accuracy of Imputation Methods in a Five-Way Admixed Population. *Frontiers in Genetics*, *10*, 34. https://doi.org/10.3389/fgene.2019.00034

Sengupta, D., Botha, G., Meintjes, A., Mbiyavanga, M., Hazelhurst, S., Mulder, N., Ramsay, M., & Choudhury, A. (2023). Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations. *Cell Genomics*, *3*(6), 100332. https://doi.org/10.1016/j.xgen.2023.100332

Serre, D., Nadon, R., & Hudson, T. J. (2005). Large-scale recombination rate patterns are conserved among human populations. *Genome Research*, *15*(11), 1547–1552. https://doi.org/10.1101/gr.4211905

Shi, M., Tanikawa, C., Munter, H. M., Akiyama, M., Koyama, S., Tomizuka, K., Matsuda, K., Lathrop, G. M., Terao, C., Koido, M., & Kamatani, Y. (2023). Genotype imputation accuracy and the quality metrics of the minor ancestry in multi-ancestry reference panels. *Briefings in Bioinformatics*, *25*(1), bbad509. https://doi.org/10.1093/bib/bbad509

Slatkin, M. (2008). Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*(6), 477–485. https://doi.org/10.1038/nrg2361

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., & Smoller, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nature Reviews Genetics*, *14*(7), 483–495. https://doi.org/10.1038/nrg3461

Suzuki, K., Hatzikotoulas, K., Southam, L., Taylor, H. J., Yin, X., Lorenz, K. M., Mandla, R., Huerta-Chagoya, A., Melloni, G. E. M., Kanoni, S., Rayner, N. W., Bocher, O., Arruda, A. L., Sonehara, K., Namba, S., Lee, S. S. K., Preuss, M. H., Petty, L. E., Schroeder, P., … Zeggini, E. (2024). Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature*, *627*(8003), 347–357. https://doi.org/10.1038/s41586-024-07019-6

Sved, J. A., & Hill, W. G. (n.d.). *One Hundred Years of Linkage Disequilibrium*.

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., … Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290–299. https://doi.org/10.1038/s41586-021-03205-y

Tcheandjieu, C., Zhu, X., Hilliard, A. T., Clarke, S. L., Napolioni, V., Ma, S., Lee, K. M., Fang, H., Chen, F., Lu, Y., Tsao, N. L., Raghavan, S., Koyama, S., Gorman, B. R., Vujkovic, M., Klarin, D., Levin, M. G., Sinnott-Armstrong, N., Wojcik, G. L., … Assimes, T. L. (2022). Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nature Medicine*, *28*(8), 1679–1692. https://doi.org/10.1038/s41591-022-01891-3

Teo, Y. Y., Fry, A. E., Bhattacharya, K., Small, K. S., Kwiatkowski, D. P., & Clark, T. G. (2009). Genome-wide comparisons of variation in linkage disequilibrium. *Genome Research*, *19*(10), 1849–1860. https://doi.org/10.1101/gr.092189.109

the Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

†The International HapMap Consortium. (2003). The International HapMap Project. *Nature*, *426*(6968), 789–796. https://doi.org/10.1038/nature02168

The 1000 Genomes Project Consortium, Corresponding authors, Auton, A., Abecasis, G. R., Steering committee, Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Travers, A., & Muskhelishvili, G. (2015). DNA structure and function. *The FEBS Journal*, *282*(12), 2279–2295. https://doi.org/10.1111/febs.13307

Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 59. https://doi.org/10.1038/s43586-021-00056-9

Wang, M. H., Cordell, H. J., & Van Steen, K. (2019). Statistical methods for genome-wide association studies. *Seminars in Cancer Biology*, *55*, 53–60. https://doi.org/10.1016/j.semcancer.2018.04.008

Watson, J. D., & Crick, F. H. C. (2003). Reprint: Molecular Structure of Nucleic Acids. *Annals of Internal Medicine*, *138*(7), 581–582. https://doi.org/10.7326/0003-4819-138-7-200304010-00015

Weiss, S. T., & Silverman, E. K. (n.d.). *Pro: Genome-Wide Association Studies (GWAS) in Asthma*.

Wightman, D. P., Jansen, I. E., Savage, J. E., Shadrin, A. A., Bahrami, S., Holland, D., Rongve, A., Børte, S., Winsvold, B. S., Drange, O. K., Martinsen, A. E., Skogholt, A. H., Willer, C., Bråthen, G., Bosnes, I., Nielsen, J. B., Fritsche, L. G., Thomas, L. F., Pedersen, L. M., … Posthuma, D. (2021). A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nature Genetics*, *53*(9), 1276–1282. https://doi.org/10.1038/s41588-021-00921-z

Xu, Y., Xing, L., Su, J., Zhang, X., & Qiu, W. (2019). Model-based clustering for identifying disease-associated SNPs in case-control genome-wide association studies. *Scientific Reports*, *9*(1), 13686. https://doi.org/10.1038/s41598-019-50229-6

Zhou, G.-L., Xu, F.-J., Qiao, J.-K., Che, Z.-X., Xiang, T., Liu, X.-L., Li, X.-Y., Zhao, S.-H., & Zhu, M.-J. (2023). E-GWAS: An ensemble-like GWAS strategy that provides effective control over false positive rates without decreasing true positives. *Genetics Selection Evolution*, *55*(1), 46. https://doi.org/10.1186/s12711-023-00820-3

Kasutatud veebilehekülje

Holtz, Y. (2021). Manhattan plot in R: a review. *The R Graph Gallery*, 16. märts. Kasutatud 30.04.2025, https://r-graph-gallery.com/101_Manhattan_plot.html

Human Genome Project. (2024). *National Human Genome Research Institute*, 13. juuni. Kasutatud 12.04.2025, https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project

Mbatchou, J., Barnard, L., Backman, J. et al. (2021). Regenie documentation. *Github*, 2021. Kasutatud 14.05.2025, https://rgcgithub.github.io/regenie/

Reference Panels. (2024). *GitHub*, 19. detsember. Kasutatud 14.05.2025, https://genepi.github.io/michigan-imputationserver/reference-panels/

The Human Genome Project. (2025). *National Human Genome Research Institute*, 19. märts. Kasutatud 12.04.2025, https://www.genome.gov/human-genome-project

The Cost of Sequencing a Human Genome. (2021). *National Human Genome Research Institute*, 1. november. Kasutatud 12.04.2025, https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Vastsündinute kaasasündinud haiguste sõeluuring. (2024). *Tartu Ülikooli Kliinikum*, 2024.

Kasutatud 12.04.2025,

https://www.kliinikum.ee/patsiendiinfo-andmebaas/vastsundinute-kaasasundinud-haiguste-soeluuring/

Vastsündinute uus sõeluuring. (2015). *Kliinikumi Leht*, 22. jaanuar 2015. Kasutatud 12.04.2025, https://www.kliinikum.ee/leht/esilehe-uudis/866-vastsuendinute-uus-soeluuring

Lihtlitsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

1. Mina, Alissa Kazmin, annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose “Erinevate referentspaneelide mõju ülegenoomsete assotsiatsiooniuuringute tulemustele”,

   mille juhendajad on Kristi Läll ja Katri Pärna,

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kui 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 4.0, mis lubab autorile viidates teost reprodutseerida, leviada ja üldsusele suunata ning keelab luua 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.

Alissa Kazmin

26.05.2025