TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Kaupo Koppel

PheWAS ja selle praktiline läbiviimine Tartu Ülikooli Eesti Geenivaramu andmete põhjal

Bakalaureusetöö (9 EAP)

Juhendajad:

Kristi Läll, MSc

Silva Kasela, MSc

TARTU 2015

PheWAS ja selle praktiline läbiviimine TÜ Eesti Geenivaramu andmete põhjal

Fenotüübi-põhine assotsiatsiooniuuring (PheWAS, *Phenome-Wide Association Study*) on kvantitatiivsetel meetoditel põhinev uuringutüüp, mis oma olemuses otsib mõne konkreetse geeniga assotsieerunud haigusi. PheWAS tugineb andmete saamisel isikute digitaalsetele tervisekaartidele ja kliinilistele andmeladude ning suudab koostöös ülegenoomsete assotsiatsiooniuuringutega (GWAS – *Genome-Wide Association Study*) pakkuda uut kliinilist teavet ning tuvastada võimalike pleiotroopsete omadustega geene. Antud bakalaureusetöö eesmärgiks oli tutvustada PheWAS metoodikat ning teoreetilist tausta, samuti viidi läbi Tartu Ülikooli Eesti Geenivaramu (TÜ EGV) andmete põhjal rasvumisega seostatud FTO-geeni variandiga rs8050136 seotud PheWAS, mis tugineb Robert M. Cronini 2014. aasta samalaadsele uuringule. Kehamassiindeksiga kohandamata PheWAS tuvastas kolm SNP-haigus assotsiatsiooni, teiste seas ka ülekaalu ning rasvumisega. Kohandamise järgselt Bonferroni paranduse mõistes olulisi seoseid polnud.

Märksõnad: geneetilised assotsiatsiooniuuringud, ühenukleotiidsed polümorfismid, fenotüüp, PheWAS, fenotüübiline muutlikkus

PheWAS in theory and in practise, based on data from Estonian Genome Center, University of Tartu

Phenome-wide association study (PheWAS) is a quantitative research technique trying to find disease association with a given gene. Data is extracted from Electronical Health Records (EHR) and Clinical Data Warehouses (CDW). PheWAS and GWAS (genome-wide association studies), when studied togheter, can provide new clinical insights and detect genetic variants with pleiotropic properties. The purpose of this bachelor thesis was to introduce methodology, theoretical backround and to conduct PheWAS on obesity associated FTO-gene variant rs8050136, using data from Estonian Genome Center, University of Tartu. This PheWAS was based on similar study by Robert M. Cronin (2014). PheWAS unadjusted for body mass index was able to detect three SNP-disease associations, which included overweight and obesity. PheWAS adjusted for body mass index did not detect any associations, when Bonferroni correction was applied.

Keywords: genetic association studies, single nucleotid polymorphisms, phenotype, PheWAS, phenotypic variation

Sisukord

Sissejuhatus

Viimase kümnendi jooksul on genotüüp-fenotüüp seoste otsimisel ning tuvastamisel ühe meetodina kasutatud ülegenoomseid assotsiatsiooniuuringuid. Paaril viimasel aastal on selliste seoste leidmiseks kasutatud ka alternatiivset, ent seotud lähenemist pakkuvat fenotüübi-põhist assotsiatsiooniuuringut.

Käesoleva bakalaureusetöö eesmärgiks on tutvustada fenotüübi-põhiste assotsiatsiooniuuringute metoodikat ning teostada statistikaprogrammi R paketi PheWAS abil praktiline analüüs Tartu Ülikooli Eesti Geenivaramu andmestiku põhjal, uurimaks FTO-geeni variandi rs8050136 võimalikke seoseid uuringus vaadeldavate haigustega.

Töö koosneb kolmest suuremast osast, enne mida on defineeritud peamised töös kasutatavad geneetika-alased terminid esimeses peatükis. Teises peatükis tutvustatakse referatiivselt PheWAS metoodikat ning põhimõtteid. Vaadeldakse PheWASi kujunemist, antakse ülevaade omadustest võrdluses ülegenoomsete assotsiatsiooniuuringutega, selgitatakse fenotüüpide kodeerimiseks kasutatavat rahvusvahelist ICD-9 haiguste klassifikatsiooni ning kirjeldatakse ka kvantitatiivsetele mõõtmistele tuginevat PheWASi.

Kolmandas peatükis antakse ülevaade kasutatavatest statistilistest meetoditest ning väljakutsetest. Tutvustatakse χ^2-testi ning logistilist regressiooni, samuti keskendutakse mitmese testimise probleemi selgitamisele ning meetoditele selle lahendamiseks.

Neljandas peatükis tutvustatakse R paketti PheWAS ning viiakse TÜ EGV andmetega läbi praktiline PheWAS, mille tulemusi analüüsitakse võrdluses Robert M. Cronini 2014. aastal teostatud uuringuga.

Töö vormistamiseks on kasutatud tarkvaraprogrammi MS Word 2013, programmikood on koostatud statistikapaketiga R (versioon 3.1.3), analüüs on teostatud R-i lisapaketiga R PheWAS. Kasutatud allikatele on viidatud nurksulgude abil.

Autor tänab siiralt käesoleva bakalaureusetöö juhendajaid, Tartu Ülikooli Eesti Geenivaramu spetsialiste Kristi Lälli ja Silva Kaselat intrigeeriva teemapüstituse, rohkete täpsustuste ja suunamiste, pühendatud aja ning jätkuva vahvuse eest.

1 Geneetika põhimõisted

Aminohape - Orgaaniline ühend, mis sisaldab amino- (NH2-) ja karboksüül-(COOH-) rühma. 20 standardsest aminohappest koosnevad valgud. [1, lk 969]

Alleel - geeniteisend, geeni esinemisvorm. [2, lk 7]

DNA - desoksüribonukleiinhape. Geneetilist informatsiooni kandev polümeer, millest koosnevad geenid. [1, lk 981]

Fenotüüp - organismi vaadeldavad tunnused, mis on määratud tema genotüübi ja keskkonnategurite koostoimes. [1, lk 991]

Geen - DNA lõik, mis määrab ära ühe valgu molekuli sünteesi. [1, lk 915]

Genoom - liigiomane ühekordses kromosoomikomplektis sisalduv geneetiline materjal. [2, lk 6]

Genotüüp - organismi geneetiline struktuur. [1, lk 998]

Haplotüüp - Tihedalt aheldunud geneetiliste elementide (nt. eri lookuste kindlate alleelide) järjestus kromosoomis, mis pärandub ühtse üksusena. [1, lk 999]

Hardy-Weinbergi printsiip/tasakaal - populatsiooni geneetilise tasakaalu seadus, mille kohaselt püsivad alleeli- ja genotüübisagedused populatsioonis muutumatuna, kui puuduvad migratsioon, valik, mutatsioonid ning ristumine indiviidide vahel on täiesti juhuslik.[2, lk 9]

Koodon - Ühele aminohappele vastav mRNA molekuli nukleotiidikolmik geneetilises koodis. [1, lk 263]

Lookus - Kindel koht kromosoomis, kus asub geen (üks tema alleelidest). [1, lk 1029]

Nukleotiid - Nukleotiidid on DNA- ja RNA-molekuli alaüksused, mis koosnevad lämmastikalusest (N-alus), suhkrust (pentoos, riboos või desoksüriboos) ja fosfaatrühmast. [1, lk 1044]

Pleiotroopsus - Geeni mitmene efekt, üksik geen mõjutab mitme tunnuse avaldumist. [1, lk 1050]

Populatsioon - Üksteisega ristuvate organismide kogum, kes kuuluvad ühte taimede või loomade gruppi ning kes elavad ühes geograafilises elupaigas. [1, lk 1052]

SNP (*Single Nucleotide Polymorfism)* ehk üksiku nukleotiidi polümorfism on DNA ahela teisend, mis seisneb ühe nukleotiidi muutuses kindlas positsioonis DNA ahelas ja esineb rohkem kui 1% populatsioonist. SNPid asuvad genoomis nii valku kodeerivates, regulatoorsetes kui ka mittekodeerivates alades. Kodeerivas regioonis asuv SNP võib omakorda olla sünonüümne (muudab koodoni järjestust, ent ei põhjusta aminohappe asendumist) või mittesünonüümne (muudab koodoni järjestust selliselt, et kodeeritakse uus aminohape). SNPid enamasti haigusi ei põhjusta, ent aitavad kindlaks määrata geene, mis seostuvad mõne haigusega. [1][2, lk 7]

2 PheWAS

2.1 PheWAS kujunemisest

2.1.1 Ülegenoomne assotsiatsiooniuuring

Ülegenoomsed assotsiatsiooniuuringud (*Genome-Wide Association Study*, GWAS) püüavad tuvastada seost geneetilise polümorfismi (*single nucleotide polymorphism*, SNP) ning uuritava tunnuse või haiguse esinemise vahel. Sõltuvalt sellest, kas uuritav tunnus on binaarne või pidev, kasutatakse GWASi läbiviimisel vastavalt kas logistilist regressiooni ning juht-kontrolluuringut (levinum võimalus) või lineaarset regressiooni. Lihtsustatult, kui analüüsi käigus tuvastatakse, et mõni SNP esineb haigust põdevate inimeste seas tihedamini, kui oleks oodatav juhuslikult, peetakse antud SNPi seostatuks uuritava tunnuse või haigusega.

Tavaliselt raporteeritakse GWAS uuringus statistiline olulisus, et teatud marker on seotud uuritava fenotüübiga; efekti suurus ehk kui palju vastav alleel uuritavat tunnust mõjutab; statistilised veahinnangud. Samas ei saa GWASi tulemuste põhjal kinnitada assotsiatsiooni bioloogilist olulisust, samuti ei saa sageli teavet ka assotsieerunud geeni kohta, kuna paljud markerid asuvad mittekodeerivates alades[3]. GWASis on vaatluse all palju SNPe, ent kuna enamasti uuritakse vaid nende peamõju, jättes koosmõju analüüsimata, jäävad potentsiaalsed SNP-fenotüüp seosed avastamata. Kui geneetiline tegur mõjutab fenotüüpi komplekse mehhanismi kaudu, mis hõlmab arvukalt teisi geene ning võimalikke keskkonnategureid, võib koosmõjude mittevaatlemisel konkreetse geeni mõju märkamata jääda. Praktikas ei ole teada, kui tihti esineb juhtumeid, kus koosmõju on oluline, kuid peamõju mitte. [4]

Esimene GWAS uuring avaldati 2005. aastal. Kümne aasta jooksul on teostatud üle 1500 taolise uuringu ning leitud üle 6000 SNPi assotsiatsiooni ligikaudu 250 haigusega, seejuures on täiendatud paljusid varem arvatud SNP-fenotüüp assotsiatsioone. Taoliste näidete alla kuuluvad 9p21.3 (algselt seostati südamerabandusega, hiljem aordi kõhu aneurüsmiga) ja SNP R602W (esialgu seotud madalama Crohni tõve kõrgema riskiga, hiljem reumatoidartriidi ja teiste autoimmuunhaiguste kõrgema riskiga). [5, lk 3]

2.1.2 Mis on PheWAS

Alternatiivne ning täiendav lähenemine genotüüp-fenotüüp assotsiatsioonide leidmiseks ning võimaliku pleiotroopsuse tuvastamiseks on fenotüübi-põhine assotsiatsiooniuuring (*Phenome Wide Association Study*, PheWAS). Kui GWAS uuringud vastavad üldistatult küsimusele

„Milline SNP on seotud vaatlusaluse fenotüübiga?“, siis PheWAS püüab vastukaaluks leida mõne konkreetse SNPiga assotsieerunud haigust või tunnust.

Esimene PheWAS avalikustati 2010. aastal Denny *et al*. poolt ajakirjas *Bioinformatics* [6]. Uuringusse kaasati viis SNPi ning kontrolliti ülegenoomsete assotsiatsiooniuuringute käigus seni leitud haigus-SNP assotsiatsioone. Peamise tulemusena kinnitati nelja SNP-haigus seose olemasolu seitsmest, samuti tuvastas PheWAS 19 varasemalt tundmatut statistiliselt olulist assotsiatsiooni vaatlusaluste SNPide ning haiguste vahel.

Enamus PheWAS uuringuid põhineb mõnel iseseisval ajaliselt varem läbi viidud GWASil, ent tihti viiakse neid uuringuid läbi korraga ühe projekti raames. Näiteks Denny *et al*. 2011 aastal avalikustatud artiklis viidi esmalt läbi esimest tüüpi kilpnäärme alatalitluse-teemaline GWAS, millele järgnevalt teostati PheWAS üle 13 000 inimese andmeid kasutades eelneva ülegenoomse uuringute põhjal leitud ning nüüd huvipakkuvas lookuses [7]. Hiljutisema sama projekti raames läbi viidud GWASi ja PheWASi, keskendudes südame rütmihäirele, viisid aastal 2013 läbi Ritchie *et al*., kes identifitseerisid kodade fibrillatsiooni ning arütmiaga seotud uued markerid [8].

2.1.3 Eelised võrdluses ülegenoomsete assotsiatsiooniuuringutega

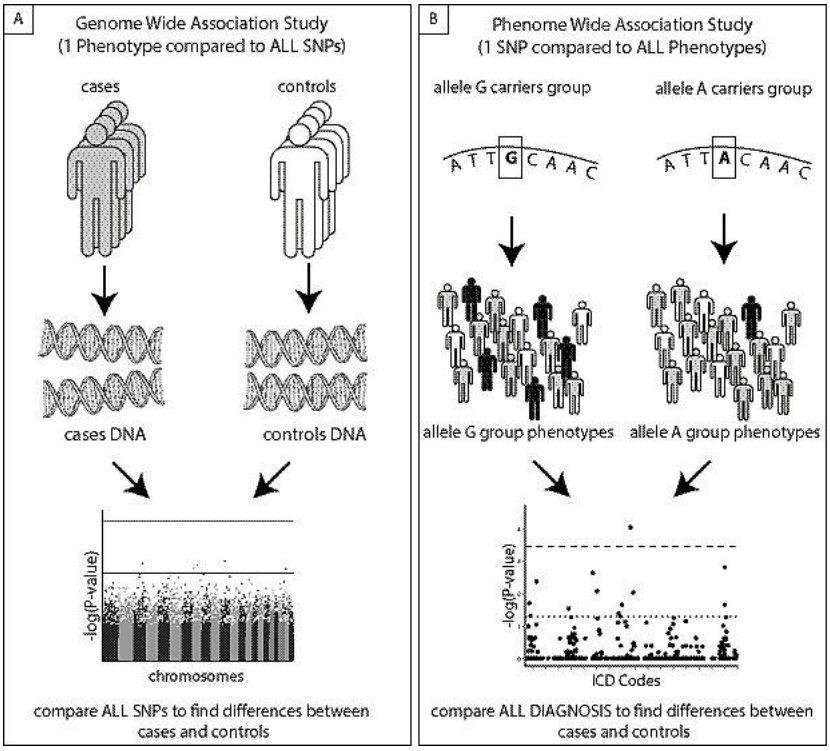
GWAS on osutunud edukaks leidmaks haigusega assotsieerunud SNPe eelkõige nende vaadeldavate haiguste (või tunnuste) korral, mis on mõjutatud ühe või vähese arvu geenide poolt. Vaadeldes aga haigusi, mis on seotud paljude geneetiliste variatsioonidega (eriti mitmed laialt levinud kesk- ning vanemas eas esinevad haigused), on GWAS lähenemine kuni hiljutise ajani pakkunud loodetust vähem sisukat teavet. [9]

Ligikaudu 13% GWASis oluliseks osutunud SNPidest on assotsieerunud kahe või enama fenotüübiga. Üheks selliseks näiteks on rs1260326 teises kromosoomis. Mittesünonüümse SNPina on rs1260326 assotsieerunud 17 GWASi põhjal 12 erineva fenotüübiga. See ilmestab, et kuigi GWAS suudab tuvastada võimalikku pleiotroopsust, on selleks vaja mitme uuringu läbiviimist. PheWAS on sellele omadusele viitamisel ulatuslikum ning efektiivsem. HLA-DBR1*1501 (rs3135388) on ülegenoomsete assotsiatsiooniuuringute põhjal seostatud hulgiskleroosiga (MS), mida kinnitas ka Denny *et al.* 2010 läbiviidud PheWAS [10]. Iseseisev jätku-uuring Hebbring *et al.* (2013) poolt kinnitas veelkord seose olemasolu, ent tuvastas omakorda uue assotsiatsiooni HLA-DBR1*1501 ning nahapunetusega seotud haigustega (*rosacea)* ja alkohoolse maksatsirroosi ehk sidekoestumise vahel (viimane haigus jäetud Denny uuringust kõrvale potentsiaalselt suure keskkonnamõju tõttu) [11]. Ühiste

geneetiliste haiguste tekkepõhjuste mõistmine nagu MSi ning *rosacea* puhul, on PheWASi poolne tähelepanuväärne edusamm, pakkudes teadmisi, mis võivad viia uute kuluefektiivsete ravistrateegiate kujunemisele. Näiteks võivad *rosacea* ravimid mõjuda efektiivselt ka MSi ravimiseks. [10, lk 162]

PheWAS uuringud võimaldavad fenotüüpe soovi korral defineerida mitmel eri viisil, ilma tervet uuringut uuesti läbi viimata [9]. Haiguste ümberdefineerimine ning liigitamine sarnaseid haigused koondavatesse gruppidesse võimaldab suuremate juhtgruppide ning vähemate võimalike fenotüüpide näol tõsta uuringu võimet leida kehtivaid assotsiatsioone. Samas lisab selline grupeerimine uuringule subjektiivse faktori. [10, lk 159]

Keskendudes geenidele ning geenimutatsioonidele, mille kohta on bioloogiast ning meditsiinist teada olulisel määral taustinformatsiooni (vt ka ptk 2.3), saab vastavat lisateavet otseselt kasutada uute PheWAS tulemuste täpsemal tõlgendamisel. [10, lk 162]



Joonis 2.1 Ülegenoomsete assotsiatsiooniuuringute (A) ning PheWAS (B) uuringute võrdlus. [12]

Joonisel 2.1 on Neuraz *et al*. (2013) poolt kõrvutatud GWAS ning PheWAS uuringuid, vastavalt paneelides A ja B. GWASi puhul alustatakse kindla fenotüübiga (nt haigusega) isikute (juht-)grupist, mida võrreldakse kontrollgrupiga, kellel antud tunnus/haigus puudub. Kasutatakse teadaolevaid genotüüpseid andmeid, et leida süstemaatilisi genoomseid erinevusi

kahe grupi vahel. PheWAS uuringutes võrreldakse kindla alleeliga (juht-)gruppi sama geeni mõne teise alleeli esindajatega. Teadaolevad fenotüübilisi andmeid kasutatakse leidmaks süstemaatilisi fenotüübilisi erinevusi juht- ning kontrollgrupi vahel sõltuvalt alleelidest jaotusest.

2.1.4 Piirangud ning puudused

Sarnaselt ülegenoomsetele assotsiatsiooniuuringutele on ka PheWAS hüpoteese genereeriv lähenemine, olles mõjutatud mitmese testimise probleemidest (vt ka peatükk 3.3.1). Kui näiteks soovitakse testida 17 000 fenotüübi võimalikku seotust ühe SNPiga, oleks katseviisilise vea $\alpha = 0{,}05$ korral Bonferroni parandust rakendades seose näitamiseks vajalik p-väärtus $2{,}9 * 10^{-6}$. Seejuures on eeldatud, et PheWASis analüüsitakse vaid ühte SNPi.

Ühtlasi komplitseerib mitmese testimise olemust asjaolu, et paljud omavahel mitte suguluses ega hierarhilises suhtes olevad haigused võivad erinevatel bioloogilistel põhjustel olla omavahel siiski seotud. Näiteks võib kõrge kolesterool omakorda põhjustada arterite lupjumist ja/või südamelihase põletikku. [10,lk 161]

PheWAS on olulisel määral sõltuv haiguste defineerimise ja liigitamise detailsusest, mis on meditsiinis muutumas järjest täpsemaks ning spetsiifilisemaks. Näiteks võib see sisuliselt tähendada ühe konkreetse haiguse jagamist peenete nüansside tõttu mitmeks erinevaks haiguseks. PheWAS mõistes on viimane aluseks keskmiste juhtude arvu vähenemisele, millega omakorda kaasneb oluline võimsuse vähenemine assotsiatsiooni tuvastamiseks.

Ühe võimaliku lahendusena on olulisuse nivoo määramisel kasutada fenotüüpide omavahelisi suhteid arvestavaid meetodeid (nt permutatsioonidel põhinevad testid).

Sarnaselt ülegenoomsetele uuringutele, on ka PheWAS puhul oluline tulemuste valideerimine. Spetsiifilise fenotüübi valideerimiseks on parim lahendus iseseisev juht-kontrolluuring, eriti just väikeste juhtumite arvuga gruppide korral võimaldab valideerimine aidata tuvastada kehtivaid assotsiatsioone, mis algse uuringu puhul selgelt ei ilmnenud. Seejuures on iga PheWASi korral kriitiline arvestada populatsioonide erinevustega. Konkreetne SNP võib ühes populatsioonis mõju avaldada, ent see ei tähenda automaatselt sarnase mõju kehtimist ka alternatiivse päritoluga rahvastikus. Näiteks ei oleks GWASis ootamatu saada mittekattuvaid tulemusi Aafrika- ning Euroopa päritoluga inimeste seas. Sama mure esineb mõneti võimendatud kujul ka PheWASis. Kui valim on moodustatud erinevatest populatsioonidest

ilma populatsioonistratifikatsioone arvestamata, võib osutuda raskeks replitseerida PheWAS tulemusi teises sõltumatus kohordis. Populatsioonidest tulenevad erinevused ei pruugi olla vaid geneetilise taustaga, lahknevused võivad tekkida nii arstiabi olemuse, praktiseerimise kui ka haiguste üles märkimise ning liigitamise erinevustest. [10, lk 162]

2.2 ICD-9 kodeering

Fenotüüpsete uuringute läbiviimise ning haldamisega on otseselt seotud rahvusvaheline haiguste klassifikatsioon ICD-9 *(International Classification of Disease, Ninth Revision).*

ICD-9 on süsteem haiguste, sümptomite, vigastuste ning diagnooside kodeerimiseks ning grupeerimiseks. Haiguse või sümptomi koodid koosnevad kolmekohalisest numbrist ehk kategooriast, millele järgneb enamasti üks või kaks täpsustavat numbrit. Kokku sisaldab süsteem üle 14 000 haiguse koodi, mis on kategooriates veel omakorda hierarhiliselt grupeeritud alamsektsioonideks ja peatükkideks. [6, lk 2]

Kuna ICD-9 terminoloogia ning süsteem on mõeldud peamiselt administratiivseteks eesmärkideks, on PheWAS uuringute läbiviimiseks defineeritud olemasolevate ICD-9 kodeeringutel põhinevad spetsiaalsed analüüsis kasutatavad ICD-9 juht- ning kontrollgrupid nn PheWAS koodi grupid. Igale haiguskoodile on määratud temale lähedased haigused, mille olemasolu korral inimest vaatlusaluse haigusega seotud analüüsi ei kaasata. Sellised tunnused on tavaliselt toodud välja andmetabelis.

ICD-9 süsteemi kasutatakse kontrollgruppide moodustamiseks kõikide juhtude gruppide jaoks iga PheWAS koodi põhjal:

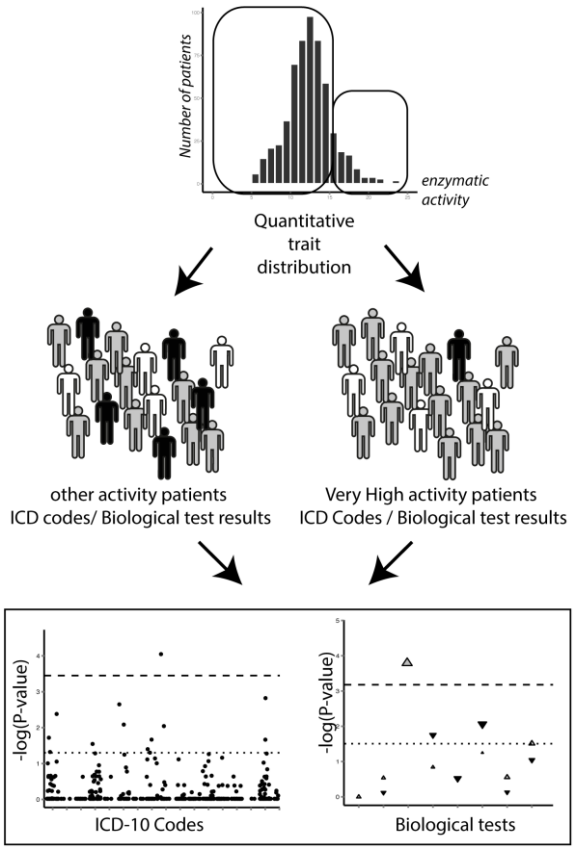
- Indiviid on juhtgrupis, kui tal on ICD-9 kood, mis kuulub vastavasse PheWAS koodi piirkonda.
- Isik on kontrollgrupis, kui tal ei ole antud ICD-9 koodiga haigust ega ka ühtegi sellega sarnast haigust.

2.3 Kvantitatiivsete tunnuste PheWAS

PheWAS uuringuid saab teostata ka kvantitatiivsete tunnuste (nt bioloogilised testiskoorid) korral, seostades konkreetsete testiskooride erinevaid tasemeid PheWAS analüüsiks kohandatud ICD-9 diagnoosikoodidega. Taolise analüüsi teostamiseks vaadeldakse tüüpiliselt selle (kvantitatiivse) omaduse ekstreemsete väärtustega indiviide, keda võrreldakse kõikide ülejäänud testiskoori omavate isikutega.[12]

Kvantitatiivne lähenemine pakub kolme eelist:

- Kvantitatiivselt mõõdetud tunnused on kättesaadavad osana kliinilisest andmestikust.
- Kvantitatiivse tunnuste PheWAS hõlmab oma olemustelt rohkem uuringusisendeid kui tavaline PheWAS, kuna lisaks genoomsetele andmetele on kasutada arvulised mõõtmistulemused, vabakirjalised kommentaarid (*free-text reports*) ning teave manustatud ravimite kohta. Viies paralleelselt läbi standardse ICD-koodidele tugineva kui ka bioloogilistel testiskooridel põhineva PheWASi, pakuvad kaks lähenemist koostöös sisukamat informatsiooni potentsiaalsetest assotsiatsioonidest ning võimalust tulemuste valideerimiseks. Vabakirjas kommentaarid ning teave ravimite kohta võimaldavad omakorda soovi korral täpsemat andmekvaliteedi kontrolli ning tulemuste sisulise analüüsi mõistmist ja lihtsustamist.
- Arvuliselt mõõdetud testiskoor sisaldab endas potentsiaalselt täpsemat teavet kui binaarne tunnus, muutes seeläbi võimaliku seose avastamise võimsuse tõstmise abil lihtsamaks.



Joonis 2.2 PheWAS kvantitatiivse tunnuse korral. [12]

Joonis 2.2 illustreerib kvantitatiivse tunnuse PheWASi ideed Neuras *et al. (2013)* artikli näitel. Vaadeldakse tiopuriinmetüültransferaasi (TPMT) ensüümi aktiivsust ning patsiendid on jagatud kaheks sõltuvalt TPMT aktiivsuse tasemest, eraldatud on kõrge aktiivsusega indiviidid. Analüüsiti nii bioloogilisi testiskoore kui ICD-10 haiguskoodide esinemist, leidmaks kahe grupi vahelisi erinevusi sõltuvalt TPMT ensüümi aktiivsusgrupist.

3 Statistilised meetodid ja väljakutsed PheWAS analüüsi puhul

PheWAS uuringute eesmärgiks on leida mõne konkreetse SNPiga assotsieerunud haigust või tunnust. Käesolevas peatükis tutvustatakse selleks kasutatavaid statistilisi meetodeid ning esilekerkivaid probleeme, kui lähemalt vaadatakse mitmese testimise probleeme ning võimalikke lahendusi PheWAS kontekstis.

3.1 Hii-ruut test

Testimaks, kas alleelide sagedused on juhtude ja kontrollide grupis erinevad, kasutatakse χ^2-testi. χ^2-test võrdleb konkreetsete andmete alusel konstrueeritud sagedustabelit teoreetilise ning sõltumatuse juhule vastava sagedustabeliga.

χ^2-statistik avaldub kujul

$$\chi^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i} = \sum_{i}^{m} \sum_{j}^{k} \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n})^2}{n_{i.} n_{.j}/n},$$

kus *O* – vaadeldud, mõõdetud (*observed*),

E – oodatud, teoreetiline (*expected*),

n_{ij} – *i*-nda rea *j*-ndas veerus olev (inimeste) arv,

$n_{.j}$ – *j*-nda veeru summa,

$n_{i.}$ – *i*-nda rea summa.

Teststatistik on nullhüpoteesi kehtides ligikaudu χ^2-jaotusega vabadusastmete arvuga $(m-1)(k-1)$, kus m ja k on uuritavate tunnuste erinevate väärtuste arvud (vastavalt ridade ja veergude arv andmetabelis). Binaarse uuritava tunnuse ja kaheväärtuselise faktortunnuse korral on teststatistik vabadusastmete arvuga üks. [13]

Tüüpiline sagedustabel χ^2-testi teostamiseks on esitatud järgnevalt:

Tabel 3.1 Juhud ning kontrollid genotüübiti aditiivse mudeli korral [14, lk 147]

	AA	AC	CC	Kokku
Juhud	*a*	*b*	*c*	n_{juht}
Kontrollid	*d*	*e*	*f*	$n_{kontroll}$
Kokku	n_{AA}	n_{AC}	n_{CC}	*n*

Antud valimis, mahuga *n*, on n_{juht} haigust omavat juhtu ning $n_{kontroll}$ kontrollgrupis olevat inimest, *a* .. *f* tähistavad vastavas rühmas olevat kindla genotüübiga inimeste konkreetset arvu. χ^2 –statistik saab kuju $\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i} = \sum_1^6 \frac{(O_i - E_i)^2}{E_i}$, kus $O_i \in \{a, .., f\}$. Teststatistiku vabadusastmete arvuks tuleb $(3-1)(2-1) = 2$.

3.2 Logistilise regressiooni mudel

Logistilise regressiooni mudeliga prognoositakse uuritava sündmuse toimumise tõenäosust sõltuvalt mõõdetud seletavate argumenttunnuse väärtuse muutumisest. Binaarse uuritava tunnuse korral kasutatakse peamiselt parema interpreteeritavuse tõttu *logit* seosefunktisooni, ent võimalikud on ka näiteks *probit* või *CLog-Log*.

$$\eta = logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k \mathrm{x_{ki}} + \varepsilon_\mathrm{i}, \qquad (3.2.1)$$

kus $\pi = P(Y = 1 \mid \mathrm{X})$ on sündmuse esinemise tõenäosus,

$\Pi = \frac{\pi}{1-\pi}$ on sündmuse šanss ehk sündmuse esinemise tõenäosuse ja tema mitteesinemise tõenäosuse suhe,

β_0 on vabaliige,

β_j on regressioonikordajad (*j*= 1, ..., *k*, *k* - argumenttunnuste arv),

x_i on argumenttunnused,

ε_i on juhuslik viga,

i = 1, ... , *n*, *n*-valimimaht.

Logit seosest saab avaldada sündmuse esinemise tõenäosuse ehk prognoosi tõenäosusele

$$\pi = \frac{e^\eta}{1 + e^\eta}.$$

Šansside suhe defineeritakse kui kahe isiku šansside suhe

$$OR = \frac{\Pi_i}{\Pi_j} = \frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_j}{1-\pi_j}}.$$

Parameetrite β_i olulisust hinnatakse χ^2-statistikuga, mis näitab, kui palju suureneks mudeli hälbimus, kui vastav argument mudelist välja jätta. Kui χ^2–statistkule vastav olulisustõenäosus on väike ($p < 0{,}05$), siis kirjeldab see argument uuritava tunnuse hajuvusest olulise osa ja ta tuleb jätta mudelisse. Parameetri β interpretatsioon toimub põhimõttel: ühikulise argumendi muutusega kaasneb šansside muutus $e^{\hat{\beta}}$ korda ning kui argument muutub *c* ühikut, siis kaasneb šansside muutus $e^{c\hat{\beta}}$ korda. Seejuures eeldatakse mitme argumendiga mudeli korral, et teised argumendid ei muutu. [15]

Näitame siinkohal ära parameetri β interpretatsioonipõhimõtte matemaatilise kehtivuse, arvestades mittevaadeldavate argumentide fikseeritust.

Logistilise regressiooni võrrandist (3.2.1) tuleneb otseselt võrdus

$$\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = e^{\hat{\beta}_0+\hat{\beta}x}.$$

Suurendame argumenti *x* ühe ühiku võrra ning teisendame saadud avaldist

$$e^{\hat{\beta}_0+\hat{\beta}(x+1)} = e^{\hat{\beta}_0+\hat{\beta}x+\hat{\beta}} = e^{\hat{\beta}_0}e^{\hat{\beta}x}e^{\hat{\beta}} = e^{\hat{\beta}_0+\hat{\beta}x}e^{\hat{\beta}}.$$

Asendades tagasi $\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = e^{\hat{\beta}_0+\hat{\beta}x}$, saame viimasest

$$e^{\hat{\beta}_0+\hat{\beta}x}e^{\hat{\beta}} = \left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)e^{\hat{\beta}}.$$

Seega on näha, et

$$e^{\hat{\beta}_0+\hat{\beta}(x+1)} = \left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)e^{\hat{\beta}}$$

ehk teisisõnu on näidatud, et argumendi *x* ühikulise suurenemisega kaasneb sündmuse toimumise šanssi muutus $e^{\hat{\beta}}$ korda.

3.3 Statistilised väljakutsed

3.3.1 Mitmene testimine

Statistilised testid lähtuvad põhimõttest, mille korral nullhüpotees kummutatakse, kui leitud seose või erinevuse saamise tõenäosus jääb väiksemaks kui etteantud olulisuse nivoo α, milleks on tavaliselt võetud α = 0,05. Teisisõnu tähendab see valik, et kuni 5% testidest võivad anda valepositiivse tulemuse.

Kui korraga tahetakse kontrollida mitut hüpoteeside paari, viib see mitmese testimise probleemini. Näiteks, kui läbi viia *k* sõltumatut testi samal olulisuse nivool α, s.t üksiku testi korral on I liiki vea (nn võrdlusviisilise vea, *experiment-wise error rate*) tegemise tõenäosus α, siis tõenäosus, et vähemalt ühe hüpoteesipaari korral tehti viga (nn katseviisiline viga, *family-wise error rate*), on:

$$P(\textit{vähemalt ühe hüpoteesi korral tehti I liiki viga}) = 1 - (1 - \alpha)^k .$$

Näiteks saja testi ning $\alpha = 0{,}05$ korral on vähemalt ühe valepositiivse testi saamise tõenäosus ligikaudu 99,4%. [2, lk 12]

3.3.2 Bonferroni parandus

Kõige lihtsamaks, ent ühtlasi ka konservatiivsemaks meetodiks katseviisilise vea kontrollimiseks on itaalia matemaatiku C. E. Bonferroni järgi nimetatud Bonferroni parandus. Meetodi idee põhineb võrdlusviisilise vea võtmisel α/k, kus k on võrdluspaaride või testide arv.

Korrektsioon kehtib nii sõltumatute kui ka sõltuvate testide puhul, kuna tõenäosus, et vähemalt üks sündmustest $A_1, \ldots, A_n$ leiab aset, on $P(\bigcup_i A_i) \leq \sum_i P(A_i)$. Seega, kui A_j on sündmus, et nullhüpotees H_j kummutatakse ja kui kõik nullhüpoteesid on tõesed, siis vähemalt ühe nullhüpoteesi kummutamise tõenäosus on väiksem või võrdne kui $\sum_j^k \alpha/k = \alpha$.

Bonferroni paranduse puuduseks on selle liigne konservatiivsus [2][6]. Vähendades valepositiivsete tulemuste/seoste hulka, vähendab korrektsioon ühtlasi ka tegelikult kehtivate seoste leidmise võimalust ehk testi võimsus on väike. Näiteks $k = 1000$ (testimaks, kas mõni 1000 fenotüübist seostub uuritava SNPiga) ja tavapärase olulisuse nivoo 0,05 puhul tuleks Bonferroni parandusega arvutusi teha olulisuse nivool 0,05/1000. See suurendab omakorda II tüüpi vea tegemise tõenäosust ja vähendab testi võimsust, muutes tegelikult oluliste erinevuste avastamise raskeks.

Samas ei nõua Bonferroni meetod, et testidele vastavaid *p*-väärtuseid kõrvutataks konstantselt võrdsete olulisuse nivoodega ehk igale testile määratud olulisusnivoo ei pea olema α/k, vajalik on, et nende summa oleks kokku α. See on otstarbekas juhul, kui eelistada ühe konkreetse hüpoteesipaari vastuvõtmist teiste sisuliste hüpoteeside vastuvõtmisele. Olgu näiteks katseviisiline veamäär 0,05 ning eelistatud hüpoteesile määratud võrdlusviisiline viga 0,04, siis kõikide teiste paaride veamääraks tuleb $\frac{0,05-0,04}{k-1} = \frac{0,01}{k-1}$. [16, lk 8]

3.3.3 FDR

Tihti on genoomikas siiski mõttekas lubada teatud hulga valepositiivseid tulemusi, et avastada paremini tegelikkuses kehtivaid seoseid. Valeavastusmäär (*False Discovery Rate*, FDR) on statistiline meetod, mis seab ülempiiri valepositiivsete testitulemuste osakaalule. Kui *p*-väärtus 0,05 tähendab, et esimest liiki viga tehakse keskmiselt 5% testide korral, siis *q*-väärtus (FDR analoog *p*-väärtusele) 0,05 määrab, et kuni 5% kõikidest olulistest tulemustest on valepositiivsed. Näiteks olles saanud $q = 0,05$ korral 1000 statistiliselt olulist tulemust, s.o. vastu võtnud H_1, on nende tulemuste seas kuni 50 valepositiivset otsust.

FDR meetod ei ole vaid üheselt defineeritud, eri autorite käsitluses sellest mitmeid versioone. Neist üks sagedamini kasutatav on Benjamin-Hochbergi meetod, mille puhul järjestatakse testide olulisustõenäosused kasvavalt ($p_{(1)} \leq \cdots \leq p_{(k)}$) ning võrreldakse neid olulisusnivoodega vastavalt ($\frac{\alpha}{k}, \frac{2\alpha}{k}, \cdots, \frac{k\alpha}{k} = \alpha$) [17].

3.3.4 SimpleM

Geneetikas sageli rakendatavat SimpleM meetodit võib mõneti pidada täienduseks Bonferroni parandusele. Antud meetodiga püütakse leida efektiivsete võrlduspaaride arvu, saavutamaks Bonferroni korrektsioonist väiksemat vähem konservatiivsemat võrdlusviisilist viga.

SimpleM tugineb tunnuste (näiteks SNPid või haigused) korrelatsioonimaatriksi põhjal leitavatele omaväärtustele ning nendevahelisele varieeruvusele. Suurem tunnuste vaheline korreleeritus tähendab ka suuremat omaväärtuste vahelist varieeruvust.[18]

Kui kõik tunnused on vastastikku maksimaalselt korreleeritud, on esimene omaväärtus võrdne vaadeldavate tunnuste arvuga *M* ning teised omaväärtused saavad võrdseks nulliga. Kui kõik tunnused on sõltumatud, saavad kõik omaväärtused võrdseks ühega ning nende vahel varieeruvus puudub.

Kasutatavate efektiivsete tunnuste arv leitakse valemiga[19]:

$$M_{eff} = M\left(1 - \frac{(M-1)V_{\lambda obs}}{M^2}\right),$$

kus

M_{eff} – iseseisvate tunnuste (efektiivne) arv, mis muutub vahemikus 1 ... M,

M – tunnuste arv maatriksis,

$V_{\lambda obs}$ – omaväärtuse varieeruvus, mis leitakse valemiga $V_{\lambda obs} = \sum_{i=1}^{M}(\lambda_i - 1)^2/(M-1)$.

Teades M_{eff} saame kogu analüüsi võrdlusviisiliseks vea:

$$\alpha = \frac{Katseviisiline\ viga}{Meff}.$$

Olgu meil näiteks 20 tunnust, katseviisiline viga $\alpha = 0{,}05$ ning esimesel juhul omaväärtuste varieeruvus $V\lambda_{obs} = 5$. Asendades väärtused valemisse, saame efektiivste tunnuste arvuks $M_{eff} = 10{,}5$ ning vastavaks võrdlusviisiliseks veaks $\alpha = \frac{0{,}05}{10{,}5} \approx 0{,}0047$.

4 Praktiline läbiviimine TÜ EGV andmete põhjal

Käesoleva töö praktiline pool tugineb Robert M. Cronin *et al*. poolt 2014. aastal avalikustatud artiklile „*Phenome-wide association studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index*“, millest antakse lühiülevaade peatükis 4.3. Artiklis läbiviidud PheWASi korratakse lihtsustatud kujul Tartu Ülikooli Eesti Geenivaramu (TÜ EGV) poolt saadud andmetega, üritades viimaste põhjal leida assotsiatsioone SNPiga rs8050136. Analüüsi sooritamiseks kasutatavat tarkvara tutvustatakse peatükis 4.1 ja analüüsi kood on lisas 1.

4.1 Ülevaade R paketist PheWAS

PheWAS uuringute läbiviimiseks kasutatakse vabavaralist tarkvara R (konkreetse töö puhul versioon R.3.1.3) ning tema lisapaketti PheWAS, mille autoriteks on Robert Carroll, Josh Denny ning Lisa Bastarache. Eristamaks edaspidi konkreetset R-i paketti samanimelisest uuringutüübist, kasutatakse selguse mõttes tarkvarapaketi nimena „R PheWAS“.

R PheWAS ei ole R-i põhiosasse sisseehitatud pakett, esmakordsel kasutamisel tuleb kõigepealt installida ning laadida paketikogum „devtools“ ning selle järel installida „R PheWAS“. Paketi kasutamiseks tuleb ta laadida käsuga *library(PheWAS).*

R PheWAS sisaldab meetodeid/käske/funktsioone sobival kujul andmetabelite loomiseks ja defineerimiseks, PheWAS analüüsi läbiviimiseks ning tulemuste graafiliseks väljastamiseks. Lühike ülevaade olulisematest käskudest:

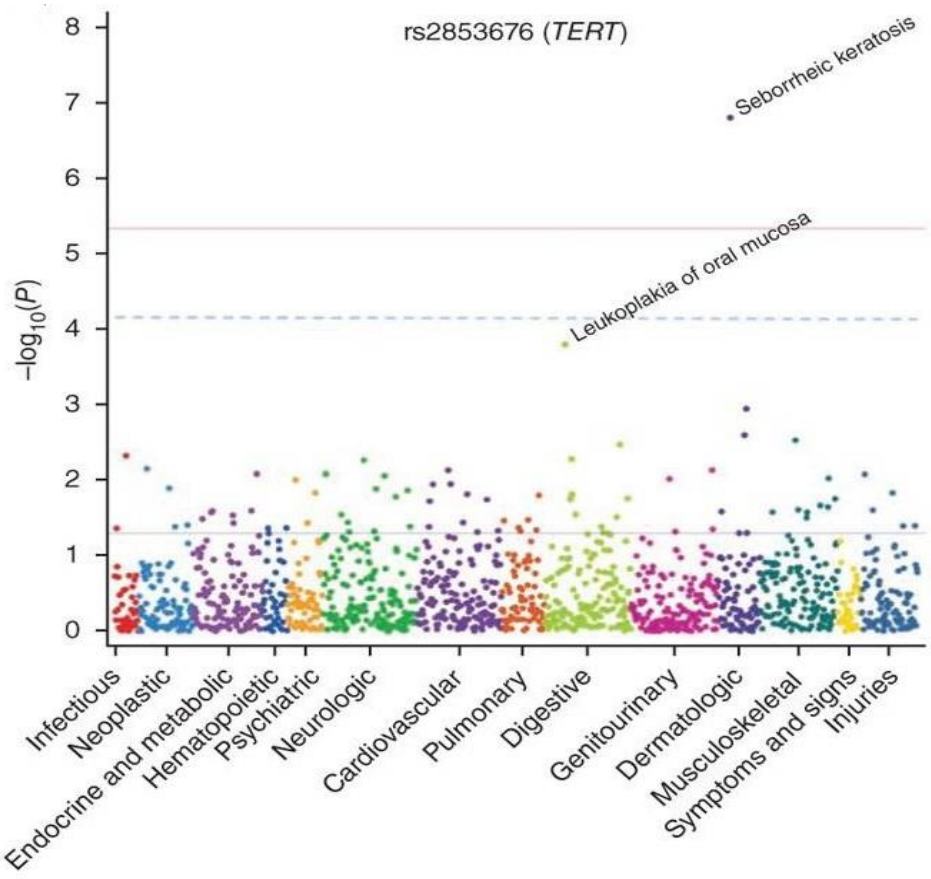
- *mapICD9toPheWAS()* – teisendab vajadusel kasutaja andmetabelis etteantud meditsiinis kasutatavad ICD-9 koodid PheWAS analüüsiks sobivateks ICD-9 kooditeks nn PheWAS koodideks.
- *createPhewasTable()* - viib analoogselt eelmisele funktsioonile koodid analüüsiks sobivale kujule, defineerides juhud ja kontrollid iga PheWAS koodi jaoks, arvestades seejuures lisaks ka edasises analüüsis kontrollgruppide moodustamisel lähedaste haiguste välistamisega (andmeveerg „*exclusion*“).
- *Phewas()* – viib läbi PheWASi, sisenditena on võimalik ette anda olemasolevad fenotüübid, SNPd, andmetabelis olevad kovariaadid, analüüsis kasutatavad mitmese testimise korrektsioonid nagu Bonferroni, FDR või SimpleM meetodi kasutamine, minimaalne haiguste arv, mille korral indiviid kaasatakse, soovi korral ka χ^2-testi ning t-testi kasutamine ning mudelite väljastamine.

- *phewasManhattan()* – väljastab tulemustest Manhattan graafiku, vt järgnev peatükk.

4.2 Tulemuste väljastamine ja kirjeldavad graafikud

Manhattan graafik (*Manhattan plot)* on PheWASi puhul enim kasutatav hajuvusgraafik. PheWAS uuringu puhul annab Manhattan graafik ülevaate uuritavate tunnuste või haiguste ning SNPi vahelisest seosest. Punktid graafikul tähistavad uuritavaid fenotüüpe, tavaliselt konkreetseid haiguseid. Horisontaalteljel on erinevate värvivööndite abil kujutatud haigusgruppe, mis on moodustatud ICD-9 peatükis seletatud põhimõttel. Vertikaalteljel on iga haiguse assotsiatsiooni *p*-väärtuse negatiivne kümnendlogaritm. Alumine horisontaaljoon tähistab olulisuse nivood 0,05, ülemine esindab vastavat Bonferroni parandust.

Haigused, mille puhul on assotsiatsioon uuritava tunnusega, omavad väiksemat *p*-väärtust ning seega on nende puhul võrreldes teiste haigustega ka negatiivne kümnendlogaritm *p*-väärtusest suurem.



Joonis 4.1 SNP rs2853676 (TERT) Manhattan graafik. [5]

Antud joonis 4.1 on saadud, kui on testitud 1358 erineva fenotüübi assotsiatsiooni SNPiga rs2853676 (TERT). Ülemine punane joon tähistab FDR q väärtust 0,1 ehk FDR olulisuse

nivood 4,6*10^{-6}, alumine sinine olulisuse nivood 0,05 ning katkendlik joon keskel Bonferroni olulisuse nivood $p = 0{,}05/1358$. Joonise põhjal on ainus statistiliselt oluline assotsiatsioon vastava SNPiga healoomulisel nahkkasvajal seborroiline keratoos (*seborrheic keratosis*), oletuslikult saaks suurema valimi korral näidata seost ka leukoplaakiaga (*leukoplakia of oral mucosa,* valged laigud suu limaskestadel).

4.3 Cronin *et al*. 2014 uurimus FTO-geenist

FTO-geeni on kehamassiindeksiga (KMI) ning ülekaalulisusega lähemalt seostatud alates 2007. aastast [20]. Seost on kinnitanud mitmed GWASid, mis on teiste hulgas toonud välja seoseid FTO geenis asuva geneetilise variandi rs8050136 ja teist tüüpi diabeedi (T2D) ning rasvumise (KMI > 30) vahel [21], ent eksisteerib ka arvukalt uurimusi, mis neid assotsiatsioone ei tuvasta [22]. Samuti on mitmed tööd andnud põhjust oletusele FTO-geenis paiknevate SNPide pleiotroopsusele [23,24 jpt]. Sellest lähtuvalt seadsid Cronin *et al*. eesmärgiks kindlaks määrata, kas EHR-põhine PheWAS suudab tuvastada võimalikku pleiotroopsust, mis traditsioonilistes geeniuuringutes jääb märkamata [25, lk 2].

Andmed olid pärit kahest erinevast kogumist: eMERGE Network, kust koguti infot 10 487 inimese kohta ning BioVU DNA teabeladu, kust uurimise alla võeti 13 711 isikut. Mõlemal juhul olid inimesed Euroopa päritolu. Kahe kogumi täpsemad kirjeldused on toodud lisas 2.

Pärast andmekontrolle teostati kaks iseseisvat analüüsi ning järgnev meta-analüüs. Fenotüüp kaasati uuringusse, kui ta leidus vähemalt 20 indiviidil. PheWAS viidi läbi nii KMIga kohandatuna kui ka ilma, kusjuures kasutati kehamassiindekseid vahemikus 15 – 70. Hilisem meta-analüüs viidi läbi üheksa ühise SNP ning 1010 levinud fenotüübiga. Bonferroni korrektsiooni kasutades saadi võrdlusviisiliseks veaks $4{,}95 * 10^{-5}$.

KMIga kohandamata PheWAS rs8050136 jaoks:

Nii BioVU kui eMERGE populatsioonis leiti pärast Bonferroni paranduse arvestamist statistiliselt oluline assotsiatsioon SNP rs8050136 ning rasvumise, T2D ja OSA (obstruktiivne uneapnoe ehk üle 10 sekundi kestev hingamisseisak une ajal) vahel. Lisaks osutus BioVU populatsioonis oluliseks ka NAFLD (krooniline mitte-alkohoolne rasvmaks). Ka metaanalüüsis leiti assotsiatsioonid rasvumise, tervisele ohtliku rasvumise (KMI > 40), T2D, NAFLDi ning vaadeldava SNP vahel.

KMIga kohandatud PheWAS rs8050136 jaoks:

Pärast kehamassiindeksiga kohandamist ei osutunud ükski seos SNP ning haiguste vahel statistiliselt oluliseks. Rohkem suurenesid olulisustõenäosused SNPi ning rasvumise, tervisele ohtliku rasvumise ja OSA vahel, vähemal määral suurenes ka NAFLD ning rs8050136 vahelise seose olulisustõenäosus. Seevastu vähenes assotsiatsiooni *p*-väärtus SNP ning fibrotsüstiliste rinnahaiguste, stafülokok- ja streptokokinfektsioonide (teatavad bakteriaalsed nakkused), osteomüeliidi (mädane luupõletik) ja liigesefusiooni (liigesepaistetus/liigne vedelik liigestes) vahel. Ükski nimetatud seos ei olnud Bonferroni parandust kasutades statistiliselt oluline. Tulemusi illustreerivad allolev tabel ning Manhattan graafik lisas 3.

Tabel 4.1. rs8050136 meta-analüüs sõltuvalt KMIga kohandatusest. [25]

Phenotype	Cases	Not adjusted for BMI		Adjusted for BMI	
		$p^{\dagger}$	*OR* (95% CI)	$p^{\dagger}$	*OR* (95% *CI*)
Overweight	3943	1.38×10^{-8}	1.17 (1.11–1.24)	0.185	1.05 (0.98–1.12)
Obesity	1662	2.10×10^{-9}	1.25 (1.16–1.35)	0.017	1.11 (1.02–1.22)
Morbid obesity	756	1.07×10^{-7}	1.34 (1.20–1.48)	0.016	1.17 (1.03–1.33)
Type 2 diabetes	3936	2.34×10^{-6}	1.14 (1.08–1.21)	4.56×10^{-4}	1.09 (1.03–1.15)
Sleep apnea	2335	3.33×10^{-5}	1.14 (1.07–1.22)	0.040	1.07 (1.00–1.15)
Cystic mastopathy	967	2.00×10^{-4}	0.82 (0.74–0.91)	4.75×10^{-4}	0.84 (0.75–0.92)
Chronic Nonalcoholic Liver disease	684	2.22×10^{-4}	1.23 (1.10–1.37)	1.86×10^{-3}	1.19 (1.07–1.33)
Chronic Ulcer of Leg or Foot	768	8.31×10^{-4}	1.19 (1.08–1.32)	2.55×10^{-3}	1.17 (1.06–1.30)
Acute Renal Failure	2047	1.12×10^{-3}	1.12 (1.05–1.20)	3.74×10^{-3}	1.11 (1.03–1.19)
Staphylococcus infections	723	2.44×10^{-3}	1.18 (1.06–1.31)	5.76×10^{-3}	1.16 (1.04–1.29)
Superficial cellulitis and abscess	2861	5.65×10^{-3}	1.09 (1.02–1.15)	0.039	1.06 (1.00–1.13)
Streptococcus infection	428	4.26×10^{-3}	1.21 (1.05–1.39)	6.56×10^{-3}	1.21 (1.05–1.39)
Osteomyelitis	352	6.15×10^{-3}	1.23 (1.06–1.43)	0.011	1.21 (1.04–1.41)
All gram positive infections	1095	6.21×10^{-4}	1.16 (1.07–1.27)	1.3×10^{-3}	1.15 (1.06–1.26)
Joint effusions	387	2.35×10^{-3}	1.25 (1.08–1.44)	6.90×10^{-3}	1.22 (1.06–1.41)

Tabelis 4.1 on välja toodud 15 fenotüüpi, mille *p*-väärtus enne kehamassiindeksiga kohandamist on väiksem kui $1{,}0 * 10^{-3}$. Iga sellise fenotüübi kohta on toodud tema leidumise arv ning KMIga kohandamise eelne ning järgne olulisustõenäosus $\alpha = 0{,}05$ korral on siin Bonferroni paranduse *p*-väärtus $4{,}95 * 10^{-4}$ ning FDR *q*-väärtus $2{,}48 * 10^{-4}$.

4.4 Andmetöötlus ning TÜ EGV kohordi kirjeldus

Autori poolt kasutatav TÜ EGV andmestik (EGV kohort) sisaldab teavet 8103 inimese kohta, iga isiku teadaolevad tunnused on: sugu (1 – mees, 2 - naine), vanus (täisaastates), kaal (kilogrammides), pikkus (sentimeetrites), kehamassiindeks, konkreetsel isikul tuvastatud haigused (kasutades ICD-10 klassifikatsiooni [26]), suitsetamine (1 – suitsetab, 0 – mitte), SNP rs8050136 – riskialleel A sagedusega 44%, genotüübid: 0 – (C;C), 1 – (A;C), 2 – (A;A).

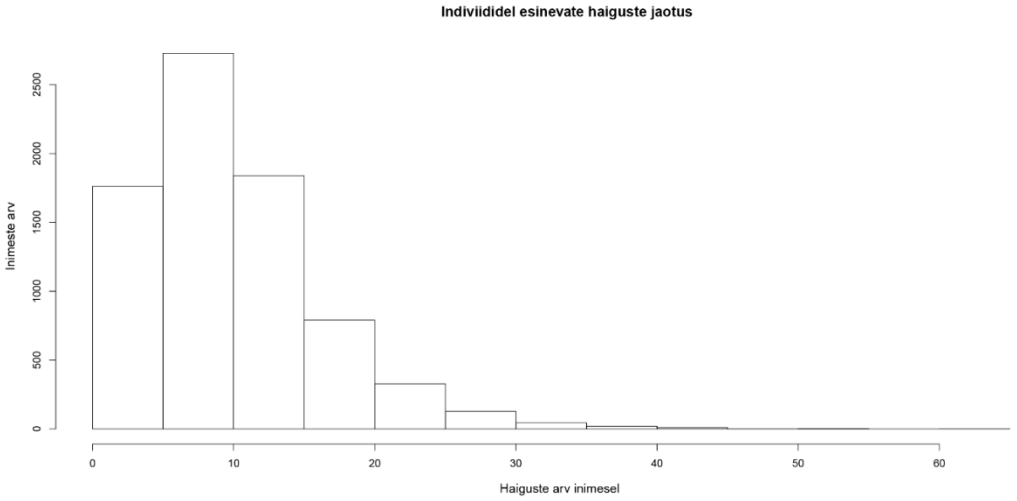
Andmetabelist eemaldati esmalt inimesed, kelle kehamassiindeks andmestikus puudub või ei kuulu vahemikku $15 \ldots 70$. Kuna leidus uuritavaid, kellel oli andmestikus vähemalt üks haiguskood esitatud vahemikuna, eemaldati vastavad isikud järgnevast analüüsist.

Fenotüübi põhise assotsiatsiooniuuringu läbiviimiseks sooritati üleminek ICD-10 klassifikatsioonilt ICD-9 süsteemile, nagu on kirjeldatud Neuraz *et al*. poolt [12]. Analüüsi jaoks sobiva andmetabeli loomiseks ning *phewas()* käsu kasutamiseks viidi tunnused sobivasse formaati, mille järel teostati PheWAS.

Uuringusubjektidest 3425 (u 44,7%) olid mehed ning 4231 (u 55,3%) naised. Keskmine vanus oli 51,2 eluaastat (standardhälve 20,3), noorim isik oli 18-aastane, vanim 103-aastane. Keskmiseks pikkuseks osutus 170 cm, kusjuures tulemused olid vahemikus $140 \ldots 206$ cm. Uuritavate keskmine kaal oli 77,3 kg, jäädes vahemikkus $37 \ldots 164$ kg. Keskmine kehamassiindeks oli 26,7, olles vahemikus $15{,}06 \ldots 59{,}63$.

Lõplikus tabelis on 284 PheWAS haiguskoodi mis esinevad vähemalt 20 indiviidil. Levinumad on ülekaal (*overweight*, 4376 inimest ehk 57,1% uuritavatest), viirusnakkused (*viral infection*, 3618, 47,2%), tuulerõuged (*varicella infection*, 2691, 35,1%), hüpertensioon ehk püsivalt normaalsest kõrgem arteriaalne vererõhk (*hypertension*, 2542, 33,2%), kõrgvererõhutõbi (*essential hypertension*, 2520, 32,9%) ning bakternakkused (*bacterial infection NOS*, 2149, 28,1%). Teisi haiguseid esines alla 2000 korra.

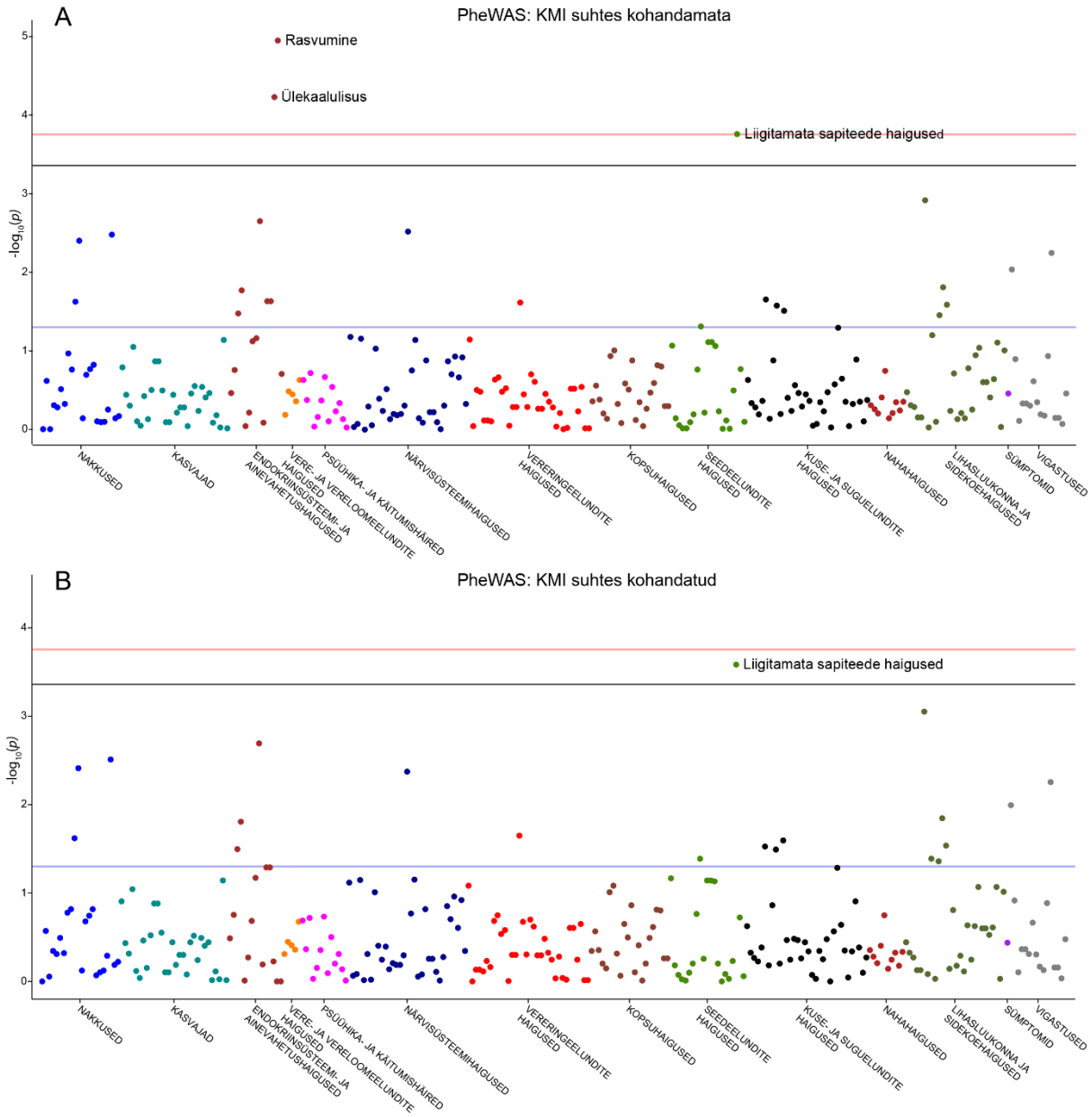
Keskmiselt oli ühel inimesel kümme PheWAS haiguskoodi, ent esines ka 62 täiesti tervet isikut. 16 inimesel leidus üle 40 tuvastatud haiguse, maksimaalselt oli ühel inimesel 61 koodi. Vaata ka joonis 4.2 indiviididel esinevate haiguste jaotuse kohta.



Joonis 4.2 Inimestel esinevate haiguste jaotus.

4.5 Tulemuste analüüs

Fenotüübi-põhine assotsiatsiooniuuring viidi Cronin *et al*. uuringust lähtudes läbi kahel juhul: kehamassiindeksiga kohandamata ning seda kovariaadina arvestades.



Joonis 4.3. PheWAS Manhattan graafikud TÜ EGV andmete põhjal. Joonise A osas on KMI suhtes kohandamata analüüsi ja joonise B osas KMI suhtes kohandatud analüüsi tulemused. Ülemine punane horisontaaljoon esindab Bonferroni paranduse olulisuse nivoo $1{,}75 * 10^{-4}$, tema all olev must joon SimpleM olulisuse nivoo $4{,}1 * 10^{-4}$ ja alumine hallikas joon tähistab olulisuse nivood $0{,}05$.

Joonise 4.3 paneelilt A ilmneb, et kui KMIga kohandamist ei toimunud, leidis PheWAS SNPi rs8050136 ning vaatlusaluste fenotüüpide vahel kolm assotsiatsiooni, mis osutusid oluliseks ka Bonferroni korrektsiooniga leitud olulisuse nivood $1{,}75 * 10^{-4}$ arvestades: rasvumine

($p = 1{,}12 * 10^{-5}$, $OR = 1.21$), ülekaalulisus ($p = 5{,}86 * 10^{-5}$, $OR = 1{,}15$) ning liigitamata sapiteede haigused ($p = 1{,}74 * 10^{-4}$, $OR = 1{,}52$) (vt ka tabel 4.2). Kui rasvumine ja ülekaalulisus näitasid assotsiatsiooni olemasolu ka Cronin *et al.* uurimuses, siis sapiteede haiguste korral PheWAS seost ei tuvastanud. Samas ei viita käesoleva uuringu tulemused assotsiatsioonidele SNPi ning uneapnoe ja tervisele ohtliku rasvumise vahel, mille olemasolu Cronin *et al.* töös tuvastati.

Tabel 4.2 TÜ EGV kohordi PheWAS rs8050136 jaoks (KMI suhtes kohandamata).

PheWAS kood	Haigus	*β*	*OR*	*SE*	*n*	juhud	kontrollid	*p*
278.1	Rasvumine	0,19	1,21	0,04	5089	1810	3279	$1{,}12*10^{-5}$
278	Ülekaal	0,13	1,15	0,03	7655	4376	3279	$5{,}86*10^{-5}$
575	Sapiteede haigused	0,42	1,52	0,11	7006	163	6843	$1{,}73*10^{-4}$
721	Lülijäikused	-0,31	0,73	0,01	6235	246	5989	$1.21*10^{-3}$
271	Metabolismi-häired	0,51	1,67	0,17	7655	73	7582	$2{,}24*10^{-3}$
362	Võrkkesta haigused	0,52	1,67	0,17	7482	68	7414	$3{,}02*10^{-3}$
133	Arboviirused	0,56	1,75	0,19	7380	56	7324	$3{,}30*10^{-3}$
079.1	Tuulerõuged	-0,11	0,89	0,04	6304	2691	3613	$3{,}97*10^{-3}$
840	Nihestused ja venitused	0,50	1,64	0,18	7655	63	7592	$5{,}66*10^{-3}$
800	Alajäsemete murrud	- 0,45	0,64	0,17	7477	75	7402	$9{,}12*10^{-3}$

Tabelis 4.2 on väljastatud PheWASist leitud kümme olulisemat SNP-haigus assotsiatsiooni, mis on tabelis järjestatud *p*-väärtuse põhjal. Seejuures on teistest on eraldatud kolm haigust, mis osutusid oluliseks ka Bonferroni paranduse olulisuse nivood $1{,}75 * 10^{-4}$ arvestades.

Teiste kovariaatide olulisuse uurimiseks koostati nimetatud kolme haiguste korral logistilise regressiooni mudel. Selgus, et rasvumise mudelis on oluline seos ka vanusel ($\beta = 0{,}03$, $p < 2 * 10^{-16}$), ent mitte sool. Ülekaalulisuse mudelis on oluline mõju nii vanusel ($\beta = 0{,}03$, $p < 2 * 10^{-16}$) kui ka sool ($\beta = -0{,}21$, $p < 1{,}84 * 10^{-5}$). Ka sapiteede korral osutuvad oluliseks nii vanus ($\beta = 0{,}02$, $p < 1{,}21 * 10^{-7}$) kui sugu ($\beta = 1{,}35$, $p < 3{,}17 * 10^{-11}$). Mudelite väljundid on toodud ka lisas 4.

Pärast KMIga kohandamist ei osutunud Bonferroni paranduse mõistes oluliseks ükski assotsiatsioon (vt joonise 4.3 paneel B). Väikseimad p-väärtused tulid esile sapiteede haiguste ($p = 2{,}60 * 10^{-4}, OR = 1{,}51$), lülijäikuste ($p = 8{,}91 * 10^{-4}, OR = 0{,}73$) ning metabolismihäirete ($p = 2{,}03 * 10^{-3}, OR = 1{,}68$) korral (vt ka tabel 4.3). Kasutades mitmese võrdluse korrektsioonina SimpleM meetodit, saadi korrigeeritud olulisuse nivooks $4{,}31 * 10^{-4}$ ning oluliseks osutub seos sapiteede haigustega. Suurema valimi korral võib statistiliselt oluline assotsiatsioon SNPiga potentsiaalselt esineda ka lülijäikuste ning metabolismi korral. Seega kerkib ka KMIga kohandamata PheWASi korral praeguses uuringus esile sapiteede haiguste võimalik assotsieeritus SNPiga, mida Cronin *et al.* artiklis ei nähtud. Samas ei ole EGV kohordis viiteid seoseid fibotsüstiliste rinnahaiguste või Cronin *et al.* uuringus mainitud bakteriaalsete haigustega.

Tabel 4.3 TÜ EGV kohordi PheWAS rs8050136 jaoks (KMI suhtes kohandatud).

PheWAS kood	Haigus	*β*	*OR*	*SE*	*n*	juhud	kontrollid	*p*
575	Sapiteede haigused	0,41	1,51	0,11	7006	163	6843	$2{,}60*10^{-4}$
721	Lülijäikused	-0,32	0,73	0,10	6235	246	5989	$8{,}91*10^{-4}$
271	Metabolismi-Häired	0,52	1,68	0,17	7655	73	7582	$2{,}03*10^{-3}$
133	Arboviirused	0,57	1,76	0,19	7380	56	7324	$3{,}07*10^{-3}$
079.1	Tuulerõuged	-0,11	0,89	0,04	6304	2691	3613	$3{,}83*10^{-3}$
362	Võrkkesta haigused	0,50	1,64	0,17	7482	68	7414	$4{,}22*10^{-3}$
840	Nihestused ja venitused	0,50	1,64	0,18	7655	63	7592	$5{,}55*10^{-3}$
800	Alajäsemete murrud	-0,45	0,64	0,17	7477	75	7402	$9{,}94*10^{-3}$
727	Liiges-, kõõlus-vigastused	0,31	0,73	0,13	6693	131	6562	0,014
244	Kilpnäärme alatalitlus	-0,29	0,75	0,12	7171	152	7019	0,015

Tabelis 4.3 on välja toodud kümme olulisemat seost, kohandades KMI suhtes. Ükski assotsiatsioon ei osutu oluliseks Bonferroni paranduse mõttes, ent SimpleM meetodiga saadud olulisuse nivoo arvestades on SNPil oluline assotsiatsioon sapiteede haigustega.

Kokkuvõte

Käesoleva bakalaureusetöö eesmärk oli tutvustada PheWAS metoodikat ning ühtlasi viia vabatarkvara R paketiga PheWAS läbi praktiline uuring, kasutades Tartu Ülikooli Eesti Geenivaramu andmeid.

Töö teoreetilises osas selgitati fenotüübi-põhiste assotsiatsiooniuuringute olemust ning uurimismetoodikat nii bioloogilises kui ka matemaatilises mõistes. Teiste seas keskenduti võrdlusele ülegenoomsete assotsiatsiooniuuringutega, kvantitatiivsetele mõõtmistele tugineva PheWASi tutvustamisele ning mitmese võrdlemise probleemile antud töö kontekstis.

Bakalaureusetöö praktilises osas teostati üle 7 500 inimese andmete põhjal PheWAS, mille sihiks oli leida FTO-geeni variandiga rs8050136 seotud haigusi. Uuring teostati kahel korral sõltuvalt kehamassiindeksiga kohandatusest ning saadud tulemusi kõrvutati praktilise poole aluseks oleva Robert M. Cronin *et al*. samalaadse uuringuga.

Kui KMI suhtes kohandamist ei toimunud, tuvastas läbiviidud PheWAS kolm statistiliselt olulist assotsiatsiooni vaadeldava SNPi ning fenotüüpide vahel: seos leiti ülekaalulisusega, rasvumisega ning sapiteede haigustega, neist kaks esimest tuvastas ka ülalmainitud uuring. KMId kovariaadina kasutades Bonferroni paranduse mõistes olulisi seoseid ei ilmnenud, ent võimalikule assotsiatsioonile sapiteede haigustega viitab SimpleM meetodi kasutamine olulisuse nivoo määramisel.

Kasutatud kirjandus

[1] Heinaru, A. 2012 *Geneetika*. Tartu. TÜ kirjastus.

[2] Kasela, S. 2011. Ülegenoomne assotsiatsiooniuuring ja selle praktiline läbiviimine TÜ Eesti Geenivaramu andmete põhjal. Bakalaureusetöö. Tartu: Tartu Ülikool, matemaatilise statistika instituut.

[3] Haller, T. 2014. Gwas praktikas. Tartu: Tartu Ülikooli geenivaramu.

[4] Cordell, H. J. 2010. „Detecting gene-gene interactions that underlie human diseases". *Nat Rev Genet.* 2009 Jun; 10(6): 392–404.

Kättesaadav: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2872761/

[5] Denny *et al.* 2013. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013 December ; 31(12): 1102–1110.

Kättesaadav: http://www.nature.com/nbt/journal/v31/n12/full/nbt.2749.html

[6] Denny *et al.* 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics.* Vol. 26 no. 9 2010, pages 1205–1210.

Kättesaadav: http://bioinformatics.oxfordjournals.org/content/26/9/1205.full.pdf

[7] Denny 2011 Variants Near FOXE1 Are Associated with Hypothyroidism and Other Thyroid Conditions: Using Electronic Medical Records for Genome- and Phenome wide Studies. *Am J Hum Genet.* 2011 Oct 7;89(4):529-42.

Kättesaadav: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3188836/

[8] Ritchie *et al*. (2013) Genome- and Phenome-Wide Analysis of Cardiac Conduction Identifies Markers of Arrhythmia Risk. *Circulation.* 2013 Apr 2; 127(13): 1377–1385.

Kättesaadav: http://circ.ahajournals.org/content/127/13/1377.long

[9] TCP Innovations. 2013. PheWAS – the tool that's revolutionizing drug development that you've likely never heard of.

http://www.tcpinnovations.com/drugbaron/phewas-the-tool-thats-revolutionizing-drug-development-that-youve-likely-never-heard-of/ (vaadatud 01.03.2015)

[10] S. J. Hebbring. 2013 The challenges, advantages and future of phenome-wide association studies. *Immunology*, 141, 157–165.

Kättesaadav: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3904236/

[11] S. J. Hebbring, S. J. Schrodi, Z. Ye, Z Zhou, D. Page, M. H Brilliant. 2013. A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun.* 2013 Apr;14(3):187-91.

Kättesaadav: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3637423/

[12] Neuraz A, Chouchana L, Malamut G, Le Beller C, Roche D, *et al*. (2013). Phenome-Wide Association Studies on a Quantitative Trait: Application to TPMT Enzyme Activity and Thiopurine Therapy in Pharmacogenomics. *PLoS Comput Biol 9(12): e1003405*.

Kättesaadav: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003405

[13] Kaart T. *Binaarsete tunnuste analüüsimeetodid*. Eesti Maaülikool 2012.

[14] Lewis, C. M., 2002. Genetic Association Studies: Design, Analysis and Interpretation. *Briefings in Bioinformatics* 3(2).

Kättesaadav: http://bib.oxfordjournals.org/content/3/2/146.long

[15] Käärik, E., 2014. *Andmeanalüüs II. Loengukonspekt*. Tartu: Tartu Ülikool, matemaatilise statistika instituut.

[16] Tamme, L. 2014. *Mitmese testimise probleem.* Bakalaureusetöö. Tartu: Tartu Ülikool, matemaatilise statistika instituut. Kättesaadav:

http://dspace.utlib.ee/dspace/bitstream/handle/10062/42526/tamme_liina_bsc_2014.pdf?sequence=1

[17] Benjamini Y and Hochberg Y (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Series B (Methodological),* Vol.57, No. 1: 289-300.

[18] X. Gao, J. Starmer, E. R. Martin (2008). A Multiple Testing Correction Method for Genetic Association Studies Using Correlated Single Nucleotide Polymorphisms. *Genetic Epidemiology Genetic Epidemiology* 32: 361-369.

http://onlinelibrary.wiley.com/doi/10.1002/gepi.20310/epdf

[19] J. M. Cheverud (2001) A simple correction for multiple comparisons in interval mapping genome scans. Kättesaadav: http://www.nature.com/hdy/journal/v87/n1/full/6889010a.html

[20] TM Frayling *et al*. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316, 889–894.

Kättesaadav: http://www.ncbi.nlm.nih.gov/pubmed/17434869

[21] Hubacek J.A. (2008). The FTO gene and obesity in a large Eastern European population sample: the HAPIEE study. *Obesity (Silver Spring)* . 2008;16:2764–2766.

Kättesaadav: http://onlinelibrary.wiley.com/doi/10.1038/oby.2008.421/full

[22] Li, H., *et al*. (2008). Variants in the fat mass- and obesity-associated (FTO) gene are not associated with obesity in a Chinese Han population. *Diabetes* 57, 264–268.

Kättesaadav: http://www.ncbi.nlm.nih.gov/pubmed/17959933

[23] Keller, L. *et al*. (2001). The obesity related gene, FTO, interacts with APOE, and is associated with Alzheimer's disease risk: a prospective cohort study. *J. Alzheimers* Dis. 23, 461–469.

[24] Lurie, G., *et al*. (2011) The obesity-associated polymorphisms FTO rs9939609 and MC4R rs17782313 and endometrial cancer risk in non-Hispanic white women. *PLoS ONE* 6:e16756.

Kättesaadav: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0016756

[25] Cronin RM *et al*., (2014). Phenome-wide association studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index. *Front Genet*. 2014 Aug 5;5:250. eCollection 2014.

Kättesaadav: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4134007/

[26] ICD10 Data. [online] Kättesaadav:

http://www.icd10data.com/ICD10CM/Codes [vaadatud 21.04.2015]

Lisad

Lisa 1: Programmikood

```
load("C:\\Users\\Kasutaja\\Andmed") #EGV andmestiku „fenot“ sisselugemiseks

#ICD-10 – ICD-9 seoste tabel
seosed = read.table("c:/seostetabel.txt", header=T, sep="\t", colClasses =
"character",quote="")

#Seoste tabelis esines vigu, leidus tühi kood ning sama täpsustusega ICD10
#koode. Parandused kirjanduse põhjal

seosed=seosed[- which(nchar(seosed[,1])==0),]
seosed=seosed[!duplicated(seosed), ]
exc.seosed=c(5652,6596,6590,6595,6580,6601,4288,770,781)
#B36 B360 I241  M32 M320 M321 M328 M329 N398
seosed=seosed[!(rownames(seosed) %in% exc.seosed),]

rownames(seosed)=seosed[,1]

#Andmete puhastamine
fenot=fenot[fenot$vanus>=18,]
fenot=fenot[fenot$bmi>=15,]
fenot=fenot[!is.na(fenot$bmi),]
valja=which(grepl("-", fenot$haig))
fenot=fenot[-valja,]

## Tunnus ID, mille määrame reanimedeks
fenot$ID = paste("ID", 1:nrow(fenot), sep="_")
rownames(fenot) = fenot$ID

#Andmestikus olevate ICD-10 koodide teisendamine ICD-9 omadeks.
#Eemaldame punktid ja alakriipsud, võttes neli esimest sümbolit, selle
jaoks vahemuutuja "haig2".
#Funktsioonid seoste teisendamiseks kirjutasid K. Läll ning S. Kasela.

library("stringr")
fenot$haig2=NA
for ( i in 1:dim(fenot)[1]){
      abi=unlist(strsplit(as.character(fenot[i,"haig"]),split=c("; ")))
      abi2=str_replace_all(abi, "[^[:alnum:]]", "")
      fenot$haig2[i]=paste(substr(abi2,1,4),collapse=";")
      }
fenot$ICD9=NA
for(i in 1:dim(fenot)[1])
{
koodid=unlist(strsplit(as.character(fenot[i,"haig2"]),split=c(";")))
abi=NULL
if (length(koodid)>0){              # Tervetele ICD9 haiguse koode ei otsi
      for (j in 1:length(koodid)){
                                    # Kasutame seoste tabelis olevaid ICD10
                                    # koode
            if(koodid[j] %in% seosed$ICD10_PHEWAS_CODE){
                  abi = c(abi, seosed[koodid[j], "PHEWAS_CODE"])
}
                                    #ICD10     -     kolmekohaliseks     ning
                                    #kontrollime kas on seoste tabelis
```

```
            if     (!(koodid[j]     %in%     seosed$ICD10_PHEWAS_CODE)     &
nchar(koodid[j])==4  &  substr(koodid[j],1,3)  %in%  seosed$ICD10_PHEWAS_CODE
){
                  abi= c(abi, seosed[substr(koodid[j],1,3), "PHEWAS_CODE"])
}
}
}
fenot[i,]$ICD9=paste(abi,collapse=";")
}

### Defineerime oma andmestikku juurde ülekaalulisusega seotud haiguskoodid

fenot$over=ifelse(fenot$bmi>25,1,0)
fenot$obese=ifelse(fenot$bmi>30,1,0)
fenot$morbid=ifelse(fenot$bmi>40,1,0)
fenot$ICD9=ifelse(fenot$over==1,paste("278.02",fenot$ICD9,sep=";"),
                  fenot$ICD9)
fenot$ICD9=ifelse(fenot$obese==1,paste("278.00",fenot$ICD9,sep=";"),fenot$I
CD9)
fenot$ICD9=ifelse(fenot$morbid==1,paste("278.01",fenot$ICD9,sep=";"),fenot$
ICD9)

###Andmed analüüsiks sobivasse formaati, kasutades muutujat „uus"
uus = NULL
for(i in 1:nrow(fenot)){
 if(fenot$ICD9[i] !=""){

      id = rownames(fenot)[i]
      count = 1
      haig  = unlist(strsplit(as.character(fenot$ICD9)[i],split=c(";")))
      for(j in 1:length(haig)){
            uus = rbind(uus, c(id, haig[j], count))
      }
 }
}
#Teisendame andmetüübid ning nimetame veerud.
colnames(uus) = c("id", "icd9", "count")
uus = data.frame(uus, stringsAsFactors=F)
uus$count = as.integer(uus$count)

library(PheWAS) #laeme paketi

#Koostame analüüsiks sobiva PheWAS andmetabeli.
phewastabel=createPhewasTable(uus, min.code.count = 1)

terved = which(fenot$ICD9 == "")
terved_andmestik = data.frame(id=rownames(fenot)[terved],
                              matrix(FALSE,nrow=62,ncol=1619))
colnames(terved_andmestik)=colnames(phewastabel)
phenotypes = rbind(phewastabel,terved_andmestik)

#Defineerime analüüsiks kovariaadid
kovariaadid=cbind(phenotypes$id,fenot[phenotypes$id,c("sugu","vanus")])
kovariaadid_bmi=cbind(phenotypes$id,fenot[phenotypes$id,c("sugu","vanus","b
mi")])

#Teisendame sugu 0,1 tunnuseks
kovariaadid$sugu = kovariaadid$sugu -1
kovariaadid_bmi$sugu = kovariaadid_bmi$sugu -1
colnames(kovariaadid)[1]="id"
colnames(kovariaadid_bmi)[1]="id"
```

```
#Defineerime genotüüpide andmetabeli
genotypes = data.frame("id"=phenotypes$id,
                       "rs8050136" = fenot[phenotypes$id,"rs8050136"])

#Viime läbi PheWAS analüüs sõltuvalt kehamassiindeksiga kohandatusest
results=phewas(phenotypes, covariates=kovariaadid, genotypes,
               significance.threshold=c("bonferroni") )
#Vajadusel viimast vahetada „SimpleM" või „fdr"

results_bmi=phewas(phenotypes,covariates=kovariaadid_bmi, genotypes,
                   significance.threshold=c("bonferroni"))
#Väljastame graafikud
#KMI kohandamata
phewasManhattan(results, annotate.angle=0,
title="KMI kohandamata PheWAS",y.axis.interval=1,x.axis.label="Fenotüübid")
+geom_hline(yintercept=3.36)

#KMI kohandatud
phewasManhattan(results_bmi, annotate.angle=0,
title="KMI kohandatud PheWAS",y.axis.interval=1, annotate.level=0.000431,
x.axis.label="Fenotüübid") +geom_hline(yintercept=3.36)

# Tulemused kood PheWAS koodide kirjeldustega
results_1 = addPhewasDescription(results)
results_2 = addPhewasDescription(results_bmi)

#Olulised tulemused
results_1[results_1$bonferroni & !is.na(results_1$p),]
results_2[results_2$bonferroni & !is.na(results_2$p),]

#Kümme tähtsaimat seost
results_1[order(results_1$p)[1:10],]
results_2[order(results_2$p)[1:10],] #BMI

#Kümme levinuimat haigust
results_1[order(results_1$n_cases,decreasing = TRUE)[1:10],]

#Täpsem teave konkreetse haiguse kohta
results_1[results_1$phewas_description =="Morbid obesity",]
results_2[results_2$phewas_description =="Overweight",]

# Uurime kui palju on inimestel haigusi?
h = rowSums(phenotypes[,-1], na.rm=T)
summary(h)
table(h)
hist(h, main="Indiviididel esinevate haiguste jaotus",xlab="Haiguste arv
inimesel",ylab="Inimeste arv")

feno= 1*phenotypes[,"278.1"] #MUDEL RASVUMISE JAOKS
summary(glm(feno[,1] ~ genotypes$rs8050136 + kovariaadid$sugu +
kovariaadid$vanus, family=binomial))

feno2=1*phenotypes[,"278"] #MUDEL ÜLEKAALU JAOKS
summary(glm(feno2[,1] ~ genotypes$rs8050136 + kovariaadid$sugu +
kovariaadid$vanus, family=binomial))

feno3=1*phenotypes[,"575"] #MUDEL SAPITEEDE JAOKS
summary(glm(feno3[,1] ~ genotypes$rs8050136 + kovariaadid$sugu +
kovariaadid$vanus, family=binomial))
```

Lisa 2: Cronin *et al.* (2014) poolt kasutatavate andmestike peamised karakteristikud

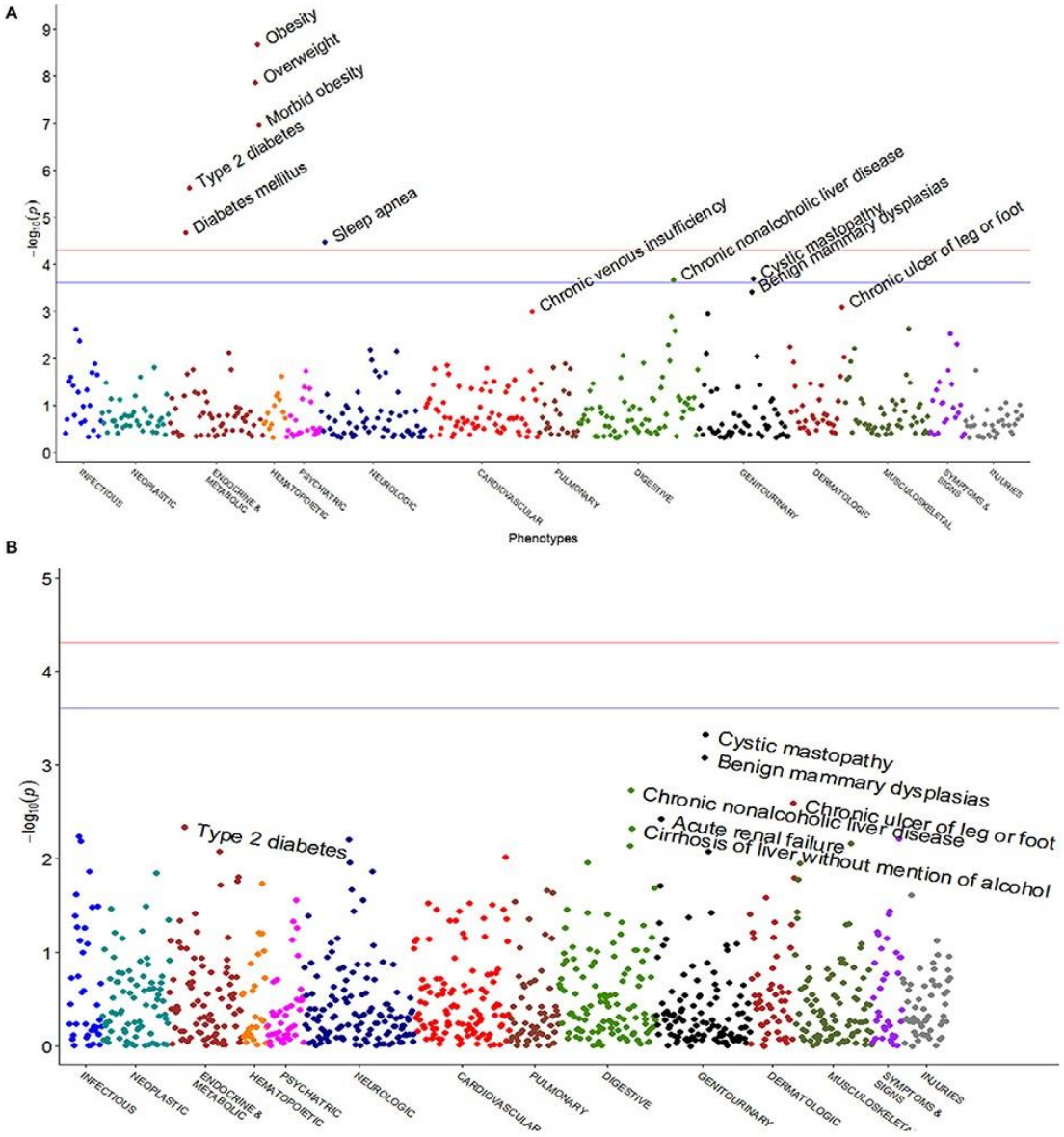
Table 1 | Characteristics of the study sets.

	eMERGE ***n* = 10,487**	**BioVU** ***n* = 13,711**	**Combined** ***n* = 24,198**
Genotyping Platform	Illumina Human660W-Quadv1_A	Illumina HumanExome	
Number of SNPs	54	9	9
Total number of phenotypes	1094	1254	1010
Median age (IQR)	58 (48–68)	60 (47–72)	59 (48–70)
Female (%)	52.24	54.31	53.35
BMI (average ± *SD*)	30.86 ± 7.48	28.43 ± 6.44	29.54 ± 7.04
Most frequent diagnoses	Hypertension (66%) Hyperlipidemia (61%) Pain in limb (47%) Malaise and fatigue (39%) Abdominal/pelvic symptoms (36%)	Hypertension (63%) Malaise and fatigue (51%) Eye infection, viral (50%) Hyperlipidemia (40%) Pain in limb (39%)	Hypertension (64%) Hyperlipidemia (49%) Malaise and fatigue (46%) Pain in limb (43%) GERD (34%)

This table shows the main characteristics of the study populations of European ancestry, including age, sex, BMI and the five most significant PheWAS phenotypes observed in the datasets. The sample size included 10,487 from the eMERGE population and 13,711 from the BioVU population for a total of 24,198 people. For a given phenotype, in the combined dataset our maximum number of cases was 14,592 in hypertension and the minimum number of cases was 44.

Tabelis on toodud teave nii eMERGE kui BioVU andmestike kohta, sisaldades vastavaid inimeste koguarve, neid kirjeldavaid peamisi statistikuid ning levinumaid haiguseid.

Lisa 3: Cronin *et al.* (2014) Manhattan graafik rs8050136 kohta



Joonise A osas ei ole analüüsis KMI kohandatust, B osas on vastav kohandamine tehtud. Mõlemal joonisel tähistab roosa horisontaaljoon Bonferroni parandust, $kus\ p = 4{,}95 * 10 - 5$ ning sinine horisontaaljoon FDR *q*-väärtust 0,05 (p=2,48 * 10^{-4}). Ilmneb, et kehamassiindeksiga kohandamise järel ei ole ükski fenotüüp vaatlusaluse SNPiga assotsieerunud.

Lisa 4: R-i väljund ülekaalulisuse ja rasvumise mudelitest.

```
###RASVUMINE
Call:
glm(formula = feno[, 1] ~ genotypes$rs8050136 +
kovariaadid$sugu +
    kovariaadid$vanus, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6190  -0.9023  -0.6620   1.2144   1.9699
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.369899   0.102593 -23.100  < 2e-16 ***
genotypes$rs8050136  0.191795   0.043670   4.392 1.12e-05 ***
kovariaadid$sugu     0.024276   0.062157   0.391    0.696
kovariaadid$vanus    0.031142   0.001537  20.258  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 6624.7  on 5088  degrees of freedom
Residual deviance: 6160.1  on 5085  degrees of freedom
  (2566 observations deleted due to missingness)
AIC: 6168.1
Number of Fisher Scoring iterations: 4

###ÜLEKAAL
Call:
glm(formula = feno2[, 1] ~ genotypes$rs8050136 +
kovariaadid$sugu +
    kovariaadid$vanus, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9995  -1.0961   0.7594   1.0139   1.5566
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.169772   0.076362 -15.319  < 2e-16 ***
genotypes$rs8050136  0.137112   0.034123   4.018 5.86e-05 ***
kovariaadid$sugu    -0.207720   0.048493  -4.284 1.84e-05 ***
kovariaadid$vanus    0.028861   0.001233  23.411  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10454.3  on 7654  degrees of freedom
Residual deviance:  9835.9  on 7651  degrees of freedom
AIC: 9843.9
Number of Fisher Scoring iterations: 4
```

```
###SAPITEED
Call:
glm(formula = feno3[, 1] ~ genotypes$rs8050136 +
kovariaadid$sugu +
    kovariaadid$vanus, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5013  -0.2528  -0.1834  -0.1270   3.3099

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -6.199436   0.316813 -19.568  < 2e-16 ***
genotypes$rs8050136  0.421832   0.112350   3.755 0.000174 ***
kovariaadid$sugu     1.354510   0.204034   6.639 3.17e-11 ***
kovariaadid$vanus    0.020736   0.003919   5.291 1.21e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1548.2  on 7005  degrees of freedom
Residual deviance: 1448.0  on 7002  degrees of freedom
  (649 observations deleted due to missingness)
AIC: 1456

Number of Fisher Scoring iterations: 7
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Kaupo Koppel (sünnikuupäev 28.12.2991),

annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „PheWAS ja selle praktiline läbiviimine TÜ Eesti Geenivaramu andmete põhjal“, mille juhendajateks on Kristi Läll ja Silva Kasela

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

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

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

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

Tartus, 27.04.2015