TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Maia Arge

Markerite sõltumatuse olulisus geneetilistes riskiskoorides

Bakalaureusetöö (6 EAP)

Juhendajad:

Reedik Mägi, PhD

TÜ Eesti Geenivaramu vanemteadur

Kristi Läll, MSc

TARTU

2014

Markerite sõltumatuse olulisus geneetilistes riskiskoorides

Geneetiline riskiskoor on geneetiliste variantide (kõige sagedamini SNPide ehk ühenukleotiidsete polümorfismide) efektisuurustest moodustatav arvuline väärtus, mis võimaldab hinnata haigestumise riski või tunnuse väärtust. Riskiskoori valitakse ülegenoomsest assotsiatsioonianalüüsist (GWASist) leitud sõltumatud markerid, mis on seotud uuritava tunnusega. Antud bakalaureusetöö eesmärk oli uurida, kuidas lineaarsed regressioonikordajad sõltuvad tunnuste omavahelistest korrelatsioonidest ning kuidas riskiskoori kui ühe seletava tunnuse regressiooniparameetri hinnang sõltub sellest, kui tugevalt on skoori lisatud SNPid omavahel korreleeritud. Riskiskooride koostamiseks kasutati Tartu Ülikooli Eesti Geenivaramu andmestikku, mis koosnes 7944st geenidoonorist ja GWASist saadud geenimarkereid ning moodustati kehamassiindeksi mudelid, kuhu üheks seletavaks tunnuseks lisati riskiskoor. Teisteks seletavateks tunnusteks olid vanus ja sugu. Kui SNPid on sõltumatud ja nende efektide suurused on korrektsed, siis on oodatavaks riskiskoori kordajaks ligikaudu üks. Selgus, et mida suurem oli riskiskooris olevate SNPide omavaheline korrelatsioon, seda enam erines riskiskoori kordaja ühest. Riskiskoori kordaja tugevalt korreleeritud SNPidega oli nullilähedane. Seetõttu on oluline valida riskiskoori sõltumatud SNPid.

Märksõnad: regressioonanalüüs, standardiseeritud kordaja, multikollineaarsus, geneetilised assotsiatsiooniuuringud, ühenukleotiidsed polümorfismid

Importance of independence of markers in genetic risk scores

The genetic risk score is a numerical value formed of the effect sizes of genetic variants (most commonly single nucleotide polymorphisms - SNPs), and used to assess disease risk or trait value. Independent markers with effect sizes from genome-wide association study (GWAS) are used to construct risk score. The purpose of this bachelor thesis was to find out how risk score's, as an independent variable's estimation of regression parameter, depends on the level of correlation among SNPs. Data from the Genome Center of the University of Tartu with 7944 individuals and genetic markers from GWAS were used to compose risk scores to be applied in body mass index models. Other independent variables were age and sex. Risk score coefficient is approximately one if SNPs are independent from each other and their effect sizes are correct. It was revealed that the higher the correlation between SNPs, the more the coefficient of risk score deviated from one. Coefficient of risk score with highly correlated SNPs was close to zero. Therefore, it is important to use independent markers in risk scores.

Keywords: regression analysis, standardized coefficients, multicollinearity, genetic association studies, single nucleotide polymorphisms

Sisukord

Sissejuhatus 4

1 Mõisted 6

2 Teoreetiline osa 7

2.1 Lineaarne regressioonimudel 7

2.1.1 Lihtne lineaarne regressioonimudel 7

2.1.2 Mitmese regressiooni mudel 9

2.1.3 Kahe tunnusega mudeli parameetrite interpreteerimine 10

2.1.4 Korrelatsioon 10

2.2 Standardiseeritud kordaja 11

2.2.1 Ümardamisviga 11

2.2.2 Tunnuse standardiseerimine ja korrelatsioonitransformatsioon 11

2.2.3 Standardiseeritud regressioonimudel 12

2.2.4 Maatriks $X^T X$ transformeeritud tunnustega 13

2.2.5 Regressioonikordajad korrelatsioonikordajate kaudu avaldatuna 15

2.2.6 Multikollineaarsus 17

3 Praktiline osa 19

3.1 Kasutatavate andmestike kirjeldused 20

3.1.1 Fenotüüpide andmestik 20

3.1.2 Genotüüpide andmestik 20

3.1.3 GWASi tulemuste tabel 21

3.2 Näide: regressiooniparameetrite muutumine tugevalt korreleeritud tunnustega mudelis 22

3.3 Mudelite koostamine 23

3.3.1 Riskiskooride moodustamine 23

3.4 Tulemused 24

Kokkuvõte 28

Kasutatud kirjandus 29

Lisad 31

Lisa 1. Programmikood 31

Lisa 2. SNPide jaotus kromosoomides 36

Lisa 3. Korrelatsioonide joonised mõnedes kromosoomides 37

Lisa 4. Efektide märkide õigsuse kontrollimise joonis 40

Sissejuhatus

Tänapäeva meditsiin paneb aina enam rõhku haiguste ennetamisele, püüdes kasutada kõiki võimalikke inimese terviseandmeid. Ühe võimalusena on hakatud uurima, kas praegusi geneetikaalaseid teadmisi saaks kuidagi rakendada haigusriskide hindamiseks. Praeguseks on leitud mõnesid fenotüübilisi (vaadeldavaid) tunnuseid ja haiguseid, kus geneetilise riskiskoori abil on võimalik hinnata haigestumise riski või tunnuse väärtust indiviidil [1] [2].

Riskiskoor moodustatakse geneetilistest markeritest (SNPidest) ja see on sisuliselt haiguse või tunnusega seotud alleelide (riskialleelide) kaalutud summa, kus kaalumiseks kasutatakse nende markerite efektisuuruseid. Üldiselt valitakse riskiskoori sõltumatud markerid, millede efektisuurused on saadud ülegenoomsetest assotsiatsiooniuuringutest (GWAS), kus miljoneid SNPe analüüsitakse ükshaaval kasutades logistilise- või lineaarse regressiooni mudelit. Käesoleva bakalaureusetöö eesmärk on välja selgitada, kuidas riskiskoori kui seletava tunnuse kordaja muutub, kui riskiskoor sisaldab erineva tugevusega korreleeritud SNPe.

Töö koosneb teoreetilisest ja praktilisest osast. Teoreetilise osa esimeses peatükis antakse ülevaade lihtsa lineaarse regressioonimudeli ja mitmese lineaarse regressioonimudeli parameetrite hinnangutest ja interpreteerimisest ning kahe tunnuse vahelisest korrelatsioonist.

Teoreetilise osa teises peatükis keskendutakse standardiseeritud regressioonikordajale ja multikollineaarsuse mõjule. Vaadeldakse, kuidas teisendatakse algseid tunnuseid korrelatsioonitransformatsiooniga ja milline näeb välja standardiseeritud lineaarne mudel ning uuritakse, kuidas regressiooniparameetrid sõltuvad tunnustevahelisest korreleeritusest.

Praktilises osas antakse ülevaade riskiskoorist, kirjeldatakse kasutatavaid andmestikke ning tehakse teoreetilise osa olulisema tulemuse tähtsuse selgitamiseks läbi üks näide, kasutades kahte andmestikus olevat SNPi. Seejärel moodustatakse riskiskoorid ja vaadeldakse kehamassiindeksi mudelites riskiskooride kordajaid, kasutades Tartu Ülikooli Eesti Geenivaramu andmestikku. Kui SNPide efektid on tõepärased ehk näitavad tõelist seost SNPi ja uuritava tunnuse vahel ning SNPid on riskiskooris korreleerimata, siis eeldatakse, et riskiskoori kordaja on ligikaudu üks. Kui mainitud eeldused on täidetud, kuid kordaja erineb ühest oluliselt, võib see viidata valideerimiseks kasutatava kohordi eripärale.

Töö vormistamiseks on kasutatud tekstitöötlusprogrammi Microsoft Word 2010. Programmikood on koostatud statistikapaketiga R (versioon 2.15.3) [3].

Kasutatud allikatele on viidatud nurksulgude abil. Maatriksid ja vektorid on kaldkirjas. Sõnu „SNP“, „geneetiline marker“ ja „geneetiline variant“ kasutatakse sünonüümidena.

Autor tänab Tartu Ülikooli Eesti Geenivaramut andmestike kasutamise loa ja juhendaja Reedik Mäge instruktsioonide, nõuannete ja geneetikaga seotud mõistete selgitamise eest ja kaasjuhendajat Kristi Lälli rohkete täienduste, paranduste ja töö struktuuri puudutavate ideede eest ning Krista Fischerit näpunäidete eest.

1 Mõisted

Alleel on geeni või mistahes genoomse lookuse esinemise vorm. [4, lk 7]

Fenotüüp on organismi vaadeldavad (vahetult või tehniliselt ja eksperimentaalselt) tunnused, mis on määratud tema genotüübi ja keskkonnategurite koostoimes. [5, lk 991]

Genotüpiseerimine on kogu genoomi ehk geneetilise materjali analüüs indiviidide omavahelise erinevuse leidmiseks. [6]

Genotüüp on organismi geneetiline struktuur (geenilookuste alleelne koosseis). Inimese genotüübi igas geenipaaris on üks alleel pärinenud isalt ja teine emalt. Näiteks võib kindlale SNPile vastata genotüüp AG, mis tähendab, et populatsioonis võib esineda genotüüpe AA, AG ja GG. [5, lk 998] [4, lk 6]

GWASis ehk ülegenoomses assotsiatsiooniuuringus (*Genome-Wide Association Study*) püütakse tuvastada seost geneetilise variatsiooni ja haiguse või vaadeldava tunnuse vahel. Kui uuritav tunnus on binaarne, siis kasutatakse juht-kontrolluuringut ja logistilist regressiooni, kus testitakse geneetilise variatsiooni alleelisageduste statistilist erinevust gruppide vahel. Kui tegemist on pideva tunnusega, mis võib olla mõjutatud ka muudest tunnustest (sugu, vanus), siis kasutatakse geneetilise variatsiooni ja uuritava tunnuse vahelise seose tuvastamiseks lineaarset regressioonanalüüsi. Ülegenoomseks assotsiatsiooniuuringuks loetakse uuringut, kus on sõltumatult testitud vähemalt 100 000 geneetilist varianti. [4]

Referentsalleel on alleel, mis on referentsiks võrrelduna riski määramiseks kasutatava alleeliga. [4, lk 22]

Riskialleel on alleel, mille arvu näitab numbriliselt kodeeritud genotüüp. Geneetilise variandi ja haiguse vahelist seost kirjeldav efektisuurus näitab riskialleeli efekti võrrelduna referentsalleeliga. [4, lk 22]

Riskiskoorina käsitletakse antud töös riskialleelide kaalutud summat, mis võimaldab hinnata indiviidi haigestumise riski või tunnuse esinemist. Iga riskialleeli kaaluna kasutatakse ta efektisuurust, mis on leitud ülegenoomses assotsiatsiooniuuringus. [7]

SNP ehk üksiku nukleotiidi polümorfism (*Single Nucleotide Polymorphism*) on DNA järjestuse teisend, mis seisneb ühe genoomi nukleotiidi (A, T, G, C) muutumises ning esineb sagedusega rohkem kui 1% populatsioonis. [4, lk 7]

2 Teoreetiline osa

2.1 Lineaarne regressioonimudel

Lineaarse mudeli abil määratakse, milline on uuritava tunnuse ja seletavate tunnuste vaheline lineaarne seos. Lineaarne regressioonimudel avaldub maatrikskujul järgmiselt:

$$Y = X\beta + \varepsilon, \tag{2.1.1}$$

kus Y on $n \times 1$ mõõtmeline vektor, mis sisaldab n vaadeldavat väärtust, X on $n \times p$ mõõtmeline plaanimaatriks, β on $p \times 1$ mõõtmeline tundmatute parameetrite vektor ning ε on $n \times 1$ mõõtmeline juhuslike vigade vektor. [8, lk 222]

2.1.1 Lihtne lineaarne regressioonimudel

Vaatleme mudeli (2.1.1) ühte alamjuhtu, olgu selleks ühe seletava tunnusega regressioonimudel kujul

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{2.1.1.1}$$

kus β_0, β_1 on mudeli parameetrid, X_i on seletava tunnuse i-s väärtus ning ε_i on i-s juhuslik viga. [8, lk 9]

Juhuslike vigade kohta eeldatakse järgmiste eelduste täidetust: [9]

1) juhuslikud vead ε_i -d on keskväärtusega $E(\varepsilon_i) = 0$ ja konstantse dispersiooniga $D(\varepsilon_i) = \sigma^2, \ \forall i,$
2) jääkide ε_i ja ε_j koviariatsioon $cov(\varepsilon_i, \varepsilon_j) = 0$ iga i,j; $i \neq j$, $i = 1, \ldots, n$,
3) juhuslikud vead on normaaljaotusega.

Regressiooniparameetrite β_0 ja β_1 leidmiseks saab kasutada vähimruutude meetodit. Iga vaatluse (X_i, Y_i) korral vaadeldakse mõõdetud väärtuse Y_i ja oodatud väärtuse $\beta_0 + \beta_1 X_i$ vahet $Y_i - (\beta_0 + \beta_1 X_i)$. Vähimruutude meetodi korral on β_0 ja β_1 hinnanguteks vastavalt b_0 ja b_1, mis minimeerivad avaldise $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$ antud vaatluste jaoks. [8, lk 15]

Kui kasutada analüütilist lähenemist, siis punktihinnangud b_0 ja b_1 mudeli (2.1.1.1) parameetritele saab leida järgnevatest võrranditest:

$$\sum Y_i = nb_0 + b_1 \sum X_i \tag{2.1.1.2}$$

ja

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum {X_i}^2. \tag{2.1.1.3}$$

Võrrandite (2.1.1.2) ja (2.1.1.3) saamiseks võetakse avaldisest $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$ tuletised β_0 ja β_1 järgi ja võrdsustatakse nulliga. [8, lk 17-18]

Esmalt avaldame hinnangu b_0 võrrandist (2.1.1.2), viies b_0 vasakule poole:

$$b_0 = \frac{\sum Y_i - b_1 \sum X_i}{n}.$$

Järgmiseks avaldame hinnangu b_1 võrrandist (2.1.1.3):

$$\sum X_i Y_i = \left(\frac{\sum Y_i - b_1 \sum X_i}{n}\right) \sum X_i + b_1 \sum {X_i}^2 = \left(\frac{\sum Y_i \sum X_i - b_1 (\sum X_i)^2}{n}\right) + b_1 \sum {X_i}^2 =$$

$$= \frac{\sum Y_i \sum X_i}{n} - b_1 \frac{(\sum X_i)^2}{n} + b_1 \sum {X_i}^2.$$

Toome b_1 vasakule poole ja avaldame selle:

$$b_1 \left(\sum {X_i}^2 - \frac{(\sum X_i)^2}{n} \right) = \sum X_i Y_i - \frac{\sum Y_i \sum X_i}{n}$$

$$b_1 = \frac{\sum X_i Y_i - \frac{\sum Y_i \sum X_i}{n}}{\sum {X_i}^2 - \frac{(\sum X_i)^2}{n}}.$$

Teisendame samaaegselt murru lugejat ja nimetajat, liites lugejasse $-\bar{Y} + \bar{Y}$ ja nimetajasse $-\bar{X} + \bar{X}$ ning kasutame teadmist, et $\sum (X_i - \bar{X}) = 0$

$$b_1 = \frac{\sum X_i Y_i - \frac{\sum Y_i \sum X_i}{n}}{\sum {X_i}^2 - \frac{(\sum X_i)^2}{n}} = \frac{\sum X_i Y_i - \bar{X} \sum Y_i}{\sum {X_i}^2 - \bar{X} \sum X_i} = \frac{\sum (X_i Y_i - \bar{X} Y_i)}{\sum \left({X_i}^2 - \bar{X} X_i \right)} = \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X}) X_i} =$$

$$= \frac{\sum \left[(X_i - \bar{X}) \left((Y_i - \bar{Y}) + \bar{Y} \right) \right]}{\sum \left[(X_i - \bar{X}) \left((X_i - \bar{X}) + \bar{X} \right) \right]} = \frac{\sum [(X_i - \bar{X})(Y_i - \bar{Y}) + (X_i - \bar{X}) \bar{Y}]}{\sum [(X_i - \bar{X})(X_i - \bar{X}) + (X_i - \bar{X}) \bar{X}]} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}.$$

Kokkuvõttes, punktihinnangud b_0 ja b_1 avalduvad kujul

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \quad (2.1.1.4)$$

ja

$$b_0 = \frac{1}{n}(\sum Y_i - b_1 \sum X_i) = \bar{Y} - b_1\bar{X}.$$

2.1.2 Mitmese regressiooni mudel

Üldine lineaarne regressioonimudel esitub kujul

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad (2.1.2.1)$$

kus $\beta_0, \beta_1, \dots, \beta_{p-1}$ on parameetrid, p on parameetrite arv mudelis, $X_{i1}, \dots, X_{i,p-1}$ on teadaolevad konstandid ning ε_i on sõltumatud vead normaaljaotusega $N(0, \sigma^2)$, $i = 1, \dots, n$. Maatrikskujul lineaarse regressioonimudeli (2.1.1) regressiooniparameetrite hinnang vähimruutude printsiibil avaldub maatrikskujul järgmiselt:

$$b = (X^T X)^{-1} X^T Y,$$

kus b on regressioonikordajate hinnangute vektor,

$$b = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{pmatrix}.$$

[8, lk 223]

Regressioonikordajate hinnangute leidmiseks vähimruutude meetodil minimeeritakse avaldis

$$min_\beta \sum_{i=1}^{n} e_i^2 = min_\beta e^T e = min_\beta (Y - X\beta)^T (Y - X\beta)$$

nii, et prognoosivigade $e = Y - X\beta$ ruutude summa oleks minimaalne. Miinimumi leidmiseks võetakse tuletis β järgi ja saadud avaldis võrdsustatakse nulliga. Ühene lahend leidub, kui eksisteerib pöördmaatriks $(X^T X)^{-1}$. Kui aga ei leidu pöördmaatriksit $(X^T X)^{-1}$, siis eksisteerib rohkem kui üks erinevat lahendit. [10]

2.1.3 Kahe tunnusega mudeli parameetrite interpreteerimine

Kahe tunnusega lineaarse regressioonimudeli, mis on kujul

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \qquad (2.1.3.1)$$

parameetrite interpreteerimine on mudeli sisuline tõlgendamine. Parameeter β_1 näitab uuritava tunnuse muutust tunnuse X_{i1} ühikulise muutuse korral, kui tunnust X_{i2} hoitakse konstantsena. Samuti, parameeter β_2 näitab uuritava tunnuse muutust tunnuse X_{i2} ühikulise muutuse korral, kui tunnust X_{i1} hoitakse konstantsena. Vabaliige β_0 on regressioonitasandi lõikepunkt y-teljega. Kui väärtused $X_{i1} = 0$ ja $X_{i2} = 0$, siis β_0 näitab uuritava tunnuse keskmist. Vastasel korral ei oma mudeli vabaliige β_0 erilist tähendust. [8, lk 216]

2.1.4 Korrelatsioon

Kahe pideva normaaljaotusega tunnuse vahelise lineaarse seose ehk korrelatiivse sõltuvuse tugevuse hindamiseks kasutatakse lineaarset ehk Pearsoni korrelatsioonikordajat. Valimi põhjal tunnuste X ja Y jaoks arvutatud Pearsoni korrelatsioonikordaja tähistatakse r_{XY}. [9]

Korrelatsioonikordaja esitub kujul

$$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{[\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2]^{1/2}}.$$

Valimi korrelatsioonikordaja r_{XY} on hinnanguks üldkogumi korrelatsioonikordajale $\rho_{X,Y}$, st

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y},$$

kus $cov(X,Y)$ on kovariatsioon tunnuste X ja Y vahel, σ_X on tunnuse X standardhälve üldkogumis ning σ_Y on tunnuse Y standardhälve üldkogumis. [9]

2.2 Standardiseeritud kordaja

2.2.1 Ümardamisviga

Ümardamisviga (*roundoff error*) on erinevus täpse matemaatilise väärtuse ja arvutatud ligikaudse väärtuse vahel. Pöördmaatriksi $(X^TX)^{-1}$ arvutamisel kipub ümardamisviga sisse tulema.

Oht $(X^TX)^{-1}$ arvutamisel ümardamisviga teha on eriti suur, kui:

1) X^TX determinant on nullilähedane;
2) X^TX elemendid erinevad magnituudiliselt.

Esimene variant esineb siis, kui mõned või kõik seletavad tunnused on korreleeritud. Teine variant esineb siis, kui tunnuste väärtuste vahemik on lai. Näiteks olgu mudelis seletavateks tunnusteks inimese vanus ja palk. Esimese tunnuse väärtuste vahemik võiks olla 0,...,100, teise tunnuse väärtuste vahemik aga 0,...,∞ (ekstreemsetel juhtudel võib palk olla väga suur). Maatriksi X^TX väärtused on seetõttu laiast vahemikust. Lahenduseks on tunnused transformeerida. [8, lk 271]

2.2.2 Tunnuse standardiseerimine ja korrelatsioonitransformatsioon

Selleks, et ümardamisveast tekkivaid probleeme ennetada, kasutatakse tunnuse standardiseerimist. Tunnuse standardiseerimine on kõikidest väärtustest keskmise lahutamine ja jagamine tunnuse standardhälbega. Uuritava tunnuse Y ja seletavate tunnuste $X_1, \dots, X_{p-1}$ standardiseerimine esitub kujul

$$\frac{Y_i-\bar{Y}}{s_Y} \text{ ja } \frac{X_{ik}-\bar{X}_k}{s_k}, (k = 1, \dots, p-1),$$

kus $\bar{Y}$ on Y keskmine, $\bar{X}_k$ on X_k keskmine ning s_Y ja s_k on valimi standardhälbed.

Valimi standardhälbed s_Y ja s_k on defineeritud järgnevalt:

$$s_y = \sqrt{\frac{\sum_i (Y_i-\bar{Y})^2}{n-1}} \text{ ja } s_k = \sqrt{\frac{\sum_i (X_{ik}-\bar{X}_k)^2}{n-1}}, \text{ kus } k = 1, \dots, p-1.$$

Standardiseerimisel saadud tunnuseid kasutatakse järgmise teisenduse läbiviimiseks, milleks on korrelatsioonitransformatsioon. Vastavad valemid on kujul

$$Y_i' = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i-\bar{Y}}{s_Y}\right) \text{ ja } X_{ik}' = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik}-\bar{X}_k}{s_k}\right),\ k = 1, \dots, p-1.$$

[8, lk 272-273]

Korrelatsioonitransformatsiooni abil saadud uusi tunnuseid, mida tähistame tildega (näiteks X_{i1}'), kasutatakse standardiseeritud regressioonimudelis.

2.2.3 Standardiseeritud regressioonimudel

Regressioonimudelit transformeeritud tunnustega Y_i' ja X_{ik}' nimetatakse standardiseeritud regressioonimudeliks ning see avaldub järgmiselt:

$$Y_i' = \beta_1' X_{i1}' + \cdots + \beta_{p-1}' X_{i,p-1}' + \varepsilon_i',$$

kus $\beta_1', \dots, \beta_{p-1}'$ on parameetrid, $p-1$ on parameetrite arv mudelis, $X_{i1}', \dots, X_{i,p-1}'$ on teadaolevad seletavad tunnused ning ε_i' on sõltumatud vead standardsest normaaljaotusest $N(0,1)$, $i = 1, \dots, n$. [8, lk 273]

Standardiseeritud vabaliikme teoreetiline hinnang on kujul

$$\beta_0' = \bar{Y}' - \beta_1'\bar{X}_{i1}' - \cdots - \beta_{p-1}'\bar{X}_{i,p-1}' = 0 - 0 - \cdots - 0 = 0,$$

sest standardiseeritud kordaja Y' keskväärtus on 0 ehk $\bar{Y}' = 0$ ja samuti $\bar{X}_{ik}' = 0\ \forall k$, $k = 1, \dots, p-1$. [11]

Järgnevalt uurime, kuidas on standardiseeritud kordajad seotud algsete regressioonikordajatega mudelis (2.1.2.1).

LAUSE. Üldise lineaarse regressioonimudeli (2.1.2.1) parameetrid $\beta_0, \beta_1, \dots, \beta_{p-1}$ ja uued parameetrid $\beta_1', \dots, \beta_{p-1}'$ on omavahel seotud valemitega

$$\beta_k = \left(\frac{s_Y}{s_k}\right)\beta_k', \quad (k = 1, \dots, p-1) \tag{2.2.3.1}$$

$$\beta_0 = \bar{Y} - \beta_1\bar{X}_1 - \cdots - \beta_{p-1}\bar{X}_{p-1}. \tag{2.2.3.2}$$

[8, lk 273]

Tõestuse idee pärineb allikast [11].

TÕESTUS. Seosest (2.2.3.1) saab avaldada $\beta_k' = \left(\frac{s_k}{s_Y}\right)\beta_k$. Selleks lahutame mudeli (2.1.2.1) mõlemalt poolt uuritava tunnuse keskväärtuse $\bar{Y}$.

$$Y_i - \bar{Y} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i - \bar{Y}.$$

Asendame β_0 vabaliikme teoreetilise hinnanguga (2.2.3.2).

$$Y_i - \bar{Y} = \bar{Y} - \beta_1 \bar{X}_{i1} - \cdots - \beta_{p-1} \bar{X}_{i,p-1} + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i - \bar{Y}.$$

Võtame regressioonikordajad sulgude ette

$$Y_i - \bar{Y} = \beta_1 (X_{i1} - \bar{X}_{i1}) + \cdots + \beta_{p-1}\left(X_{i,p-1} - \bar{X}_{i,p-1}\right) + \varepsilon_i.$$

Korrelatsioonitransformatsiooni läbinud tunnuste saamiseks jagame avaldise liikmeid standardhälvete korrutisega $s_Y s_1 \cdot \ldots \cdot s_{p-1}$ ja korrutame väärtusega $\frac{1}{\sqrt{n-1}}$. Uue i-nda juhusliku vea tähistame ε_i^*.

$$\frac{Y_i - \bar{Y}}{s_Y s_1 \cdot \ldots \cdot s_{p-1}} \frac{1}{\sqrt{n-1}} = \frac{\beta_1 (X_{i1} - \bar{X}_{i1})}{s_Y s_1 \cdot \ldots \cdot s_{p-1}} \frac{1}{\sqrt{n-1}} + \cdots + \frac{\beta_{p-1}\left(X_{i,p-1} - \bar{X}_{i,p-1}\right)}{s_Y s_1 \cdot \ldots \cdot s_{p-1}} \frac{1}{\sqrt{n-1}} + \varepsilon_i^*$$

$$= \frac{\beta_1 X_{i1}'}{s_Y s_2 \cdot \ldots \cdot s_{p-1}} + \cdots + \frac{\beta_{p-1} X_{i,p-1}'}{s_Y s_1 \cdot \ldots \cdot s_{p-2}} + \varepsilon_i^*.$$

Korrutame võrduse pooli läbi avaldisega $s_1 \cdot \ldots \cdot s_{p-1}$.

$$Y_i' = \beta_1 \frac{s_1}{s_Y} X_{i1}' + \cdots + \beta_{p-1} \frac{s_{p-1}}{s_Y} X_{i,p-1}' + \varepsilon_i^*.$$

Järeldub, et

$$\beta_k' = \left(\frac{s_k}{s_Y}\right)\beta_k.$$

■

2.2.4 Maatriks $X^T X$ transformeeritud tunnustega

Kui uued tunnused X_{ik}' on moodustatud korrelatsioonitransformatsiooniga, siis on oluline uurida, mis juhtub maatriksiga $X^T X$.

Seletavate tunnuste $n \times (p-1)$ mõõtmeline maatriks X on kujul

$$X = \begin{pmatrix} X'_{11} & \cdots & X'_{1,p-1} \\ X'_{21} & \cdots & X'_{2,p-1} \\ \vdots & \ddots & \vdots \\ X'_{n1} & \cdots & X'_{n,p-1} \end{pmatrix}.$$

Korrelatsioonimaatriks mõõtmetega $(p-1) \times (p-1)$ on kujul

$$r_{XX} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1,p-1} \\ r_{21} & 1 & \cdots & r_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1,1} & r_{p-1,2} & \cdots & 1 \end{pmatrix},$$

kus r_{12} tähistab korrelatsioonikordajat tunnuste X_1 ja X_2 vahel. Peadiagonaal koosneb ühtedest, sest tunnuse korrelatsioon iseendaga on 1. Korrelatsioonimaatriks on sümmeetriline, st $r_{ks} = r_{sk},\ k, s = 1, \dots, (p-1)$ [9].

Uuritava tunnuse Y ja seletavate tunnuste X vahelistest korrelatsioonikordajatest koosneva vektori tähistame r_{YX}-ga:

$$r_{YX} = \begin{pmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Y,p-1} \end{pmatrix}.$$

Vaatleme $(p-1) \times (p-1)$ mõõtmelist transformeeritud tunnustega maatriksit $X^T X$. Näitame, et maatriks $X^T X$ on korrelatsioonimaatriks, st $X^T X = r_{XX}$.

Teeme selle läbi juhul, kui $X^T X$ on 2×2 maatriks:

$$X^T X = \begin{pmatrix} X'_{11} & X'_{21} & \cdots & X'_{n1} \\ X'_{12} & X'_{22} & \cdots & X'_{n2} \end{pmatrix} \begin{pmatrix} X'_{11} & X'_{12} \\ X'_{21} & X'_{22} \\ \vdots & \vdots \\ X'_{n1} & X'_{n2} \end{pmatrix} = \begin{pmatrix} \sum (X'_{i1})^2 & \sum X'_{i1} X'_{i2} \\ \sum X'_{i2} X'_{i1} & \sum (X'_{i2})^2 \end{pmatrix}$$

Uurime lähemalt maatriksis asuvaid elemente. Paneme tähele, et $X^T X$ ülemises vasakpoolses nurgas on $\sum (X'_{i1})^2 = \sum \left(\frac{X_{i1} - \bar{X}_1}{\sqrt{n-1} * s_1} \right)^2 = \frac{\sum (X_{i1} - \bar{X}_1)^2}{n-1} \div {s_1}^2 = 1.$

Esimese rea teises tulbas on aga

$$\sum X'_{i1} X'_{i2} = \sum \left(\frac{X_{i1} - \bar{X}_1}{\sqrt{n-1} * s_1} \right) \left(\frac{X_{i2} - \bar{X}_2}{\sqrt{n-1} * s_2} \right) = \frac{1}{n-1} \frac{\sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{s_1 s_2} = \frac{\sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{[\sum (X_{i1} - \bar{X}_1)^2 \sum (X_{i2} - \bar{X}_2)^2]^{1/2}} = r_{12}.$$

Näeme, et 2×2 mõõtmelise maatriksi $X^T X$ korral on tegemist korrelatsioonimaatriksiga r_{XX}.

Selgub, et maatriksi $X^T X$ elemendid ühtivad korrelatsioonimaatriksi r_{XX} elementidega $\forall\, X^T X$ korral ehk

$$X^T X = r_{XX} \quad \forall\, X^T X. \tag{2.2.4.1}$$

Lisaks saab näidata, et transformeeritud tunnustega maatriksi $X^T Y$ elemendid ühtivad korrelatsioonimaatriksi r_{YX} elementidega, st

$$X^T Y = r_{YX} \quad \forall\, X^T Y\,. \tag{2.2.4.2}$$

[8, lk 274-275]

2.2.5 Regressioonikordajad korrelatsioonikordajate kaudu avaldatuna

Järgmisena näitame, kasutades mudelit (2.1.1.1), et tunnuse X_1 regressioonikordaja hinnang ei muutu, kui lisame mudelisse tunnuse X_2, eeldusel, et X_1 ja X_2 pole omavahel korreleeritud.

Kahe seletava tunnuse ja uuritava tunnuse vahelised korrelatsioonimaatriksid on kujul

$$r_{XX} = \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix},$$

$$r_{YX} = \begin{pmatrix} r_{Y1} \\ r_{Y2} \end{pmatrix},$$

$$r_{XX}{}^{-1} = \frac{1}{1 - r_{12}{}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{21} & 1 \end{pmatrix},$$

kus r_{Y1} on korrelatsioonikordaja tunnuste Y ja X_1 vahel, r_{Y2} on korrelatsioonikordaja tunnuste Y ja X_2 vahel ning r_{12} on korrelatsioonikordaja tunnuste X_1 ja X_2 vahel.

Kasutades seoseid (2.2.4.1) ja (2.2.4.2), avalduvad standardiseeritud regressioonimudeli parameetrite hinnangud vähimruutude meetodil kujul

$$b = (X^T X)^{-1} X^T Y = r_{XX}{}^{-1} r_{YX} = \frac{1}{1 - r_{12}{}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix} \begin{pmatrix} r_{Y1} \\ r_{Y2} \end{pmatrix} = \frac{1}{1 - r_{12}{}^2} \begin{pmatrix} r_{Y1} - r_{12} r_{Y2} \\ r_{Y2} - r_{12} r_{Y1} \end{pmatrix},$$

kus pöördmaatriks $r_{XX}{}^{-1} = \frac{1}{|r_{XX}|} \cdot \widetilde{r_{XX}} = \frac{1}{1 - r_{12}{}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{21} & 1 \end{pmatrix}$ ja $\widetilde{r_{XX}}$ on r_{XX} algebraliste täiendite transponeeritud maatriks.

Standardiseeritud regressiooniparameetrite hinnangute vektor b on

$$b = \begin{pmatrix} b_1' \\ b_2' \end{pmatrix}.$$

Regressiooniparameetrite hinnangud b_1' ja b_2' avalduvad siis kujul

$$b_1' = \frac{r_{Y1} - r_{12} r_{Y2}}{1 - {r_{12}}^2} \qquad (2.2.5.1)$$

ja

$$b_2' = \frac{r_{Y2} - r_{12} r_{Y1}}{1 - {r_{12}}^2}, \qquad (2.2.5.2)$$

[8, lk 275-276]

Lisaks seostele (2.2.5.1) ja (2.2.5.2) kehtib ka

$$b_k = \left(\frac{s_Y}{s_k}\right) b_k'. \qquad (2.2.5.3)$$

Kasutades seost (2.2.5.3), avalduvad mudeli (2.1.3.1) regressioonikordajate hinnangud b_1 ja b_2 kujul:

$$b_1 = \left(\frac{s_Y}{s_1}\right) b_1' = \left(\frac{[\sum(Y_i - \bar{Y})^2/(n-1)]^{\frac{1}{2}}}{[\sum(X_{i1} - \bar{X}_1)^2/(n-1)]^{\frac{1}{2}}}\right) \frac{r_{Y1} - r_{12} r_{Y2}}{1 - {r_{12}}^2} = \frac{\left(\frac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i1} - \bar{X}_1)^2}\right)^{\frac{1}{2}} r_{Y1} - \left(\frac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i1} - \bar{X}_1)^2}\right)^{\frac{1}{2}} r_{12} r_{Y2}}{1 - {r_{12}}^2} =$$

$$= \frac{\left(\frac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i1} - \bar{X}_1)^2}\right)^{\frac{1}{2}} \cdot \frac{\sum(X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{(\sum(X_{i1} - \bar{X}_1)^2 \sum(Y_i - \bar{Y})^2)^{\frac{1}{2}}} - \left(\frac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i1} - \bar{X}_1)^2}\right)^{\frac{1}{2}} r_{12} r_{Y2}}{1 - {r_{12}}^2} = \frac{\frac{\sum(X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\sum(X_{i1} - \bar{X}_1)^2} - \left(\frac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i1} - \bar{X}_1)^2}\right)^{\frac{1}{2}} r_{12} r_{Y2}}{1 - {r_{12}}^2}$$

ja

$$b_2 = \left(\frac{s_Y}{s_2}\right) b_2' = \left(\frac{[\sum(Y_i - \bar{Y})^2/(n-1)]^{\frac{1}{2}}}{[\sum(X_{i2} - \bar{X}_2)^2/(n-1)]^{\frac{1}{2}}}\right) \frac{r_{Y2} - r_{12} r_{Y1}}{1 - {r_{12}}^2} = \frac{\left(\frac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i2} - \bar{X}_2)^2}\right)^{\frac{1}{2}} r_{Y2} - \left(\frac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i1} - \bar{X}_2)^2}\right)^{\frac{1}{2}} r_{12} r_{Y1}}{1 - {r_{12}}^2} =$$

$$= \frac{\left(\frac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i2} - \bar{X}_2)^2}\right)^{\frac{1}{2}} \cdot \frac{\sum(X_{i2} - \bar{X}_2)(Y_i - \bar{Y})}{\left(\sum(X_{i2} - \bar{X}_2)^2 \sum(Y_i - \bar{Y})^2\right)^{\frac{1}{2}}} - \left(\frac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i1} - \bar{X}_2)^2}\right)^{\frac{1}{2}} r_{12} r_{Y1}}{1 - {r_{12}}^2} = \frac{\frac{\sum(X_{i2} - \bar{X}_2)(Y_i - \bar{Y})}{\sum(X_{i2} - \bar{X}_2)^2} - \left(\frac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i2} - \bar{X}_2)^2}\right)^{\frac{1}{2}} r_{12} r_{Y1}}{1 - {r_{12}}^2}.$$

Kui X_1 ja X_2 vahel korrelatsioon puudub, siis $r_{12} = 0$, ning punktihinnangud b_1 ja b_2 taanduvad eespool näidatud kujule (2.1.1.4):

$$b_1 = \frac{\sum(X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\sum(X_{i1} - \bar{X}_1)^2}, \; b_2 = \frac{\sum(X_{i2} - \bar{X}_2)(Y_i - \bar{Y})}{\sum(X_{i2} - \bar{X}_2)^2}.$$

[8, lk 281]

2.2.6 Multikollineaarsus

Kui seletavad tunnused on omavahel tugevalt (tavaliselt alates korrelatsioonikordajast 0,8) korreleeritud, siis öeldakse, et eksisteerib multikollineaarsus. Kui seletavad tunnused ei ole omavahel korreleeritud, siis regressioonimudelis jäävad parameetrite hinnangud samaks, olenemata millised seletavad tunnused veel mudelisse võtta.

Multikollineaarsuse mõju regressiooniparameetritele: [8, lk 283]

1. Kui seletavad tunnused on tugevalt korreleeritud, siis regressioonikordajad on valimites suurte hajuvustega ja seetõttu erinevad regressioonikordajad erinevate valimite puhul üsna palju. Tulemuseks võib olla ebatäpne informatsioon tõelise regressioonikordaja kohta.
2. Multikollineaarsuse esinemise korral ei saa regressioonikordajat interpreteerida nii, nagu tavaliselt, st regressioonikordaja näitab uuritava tunnuse muutust, mis kaasneb vastava seletava tunnuse ühikulisele muutusele, kui teised tunnused ei muutu. Näiteks viljakasvu vihma ja päikesetundide järgi ennustavas regressioonimudelis ei saa ühte tunnust muuta, samal ajal teist tunnust konstantsena hoides.
3. Regressioonikordaja hinnang ei pruugi olla statistiliselt oluline, kuigi seletava ja uuritava tunnuse vahel on seos [12, lk 78].

Kui seletavad tunnused on korreleeritud, siis mistahes tunnuse regressioonikordaja sõltub sellest, millised tunnused on veel mudelisse lisatud ja millised on välja jäetud. Korreleeritud seletavate tunnustega mudelis ei näita regressioonikordaja seletava ja uuritava tunnuse vahelist efekti, vaid ainult marginaalset/osalist efekti olenevalt sellest, millised teised korreleeritud tunnused on mudelisse lisatud. [12, lk 47]

Näitame, et kahe seletava tunnusega regressioonimudeli (2.1.3.1) korral on regressioonikordajate hinnangute hajuvused lõpmata suured, kui tunnustevaheline korrelatsioon läheneb ühele.

Mudel (2.1.3.1) standardiseeritud kordajatega avaldub kujul

$$Y_i' = \beta_1' X_{i1}' + \beta_2' X_{i2}' + \varepsilon_i'.$$

Pöördmaatriks $(X^T X)^{-1}$ selle standardiseeritud mudeli korral on

$$(X^T X)^{-1} = {r_{XX}}^{-1} = \frac{1}{1-{r_{12}}^2}\begin{pmatrix} 1 & -r_{12} \\ -r_{21} & 1 \end{pmatrix}.$$

Variatsiooni-kovariatsiooni maatriks on kujul

$$\sigma^2(b) = (\sigma')^2 {r_{XX}}^{-1} = (\sigma')^2 \frac{1}{1-{r_{12}}^2}\begin{pmatrix} 1 & -r_{12} \\ -r_{21} & 1 \end{pmatrix},$$

kus b on regressiooniparameetrite hinnangute vektor ja $(\sigma')^2$ on standardiseeritud mudeli MSE ehk keskmine ruutviga.

Regressioonikordajate hinnangute b_1' ja b_2' hajuvus on

$$\sigma^2(b_1') = \sigma^2(b_2') = \frac{(\sigma')^2}{1-{r_{12}}^2},$$

mis muutub suuremaks, kui X_1 ja X_2 vaheline korrelatsioon suureneb. Kui X_1 ja X_2 korrelatsioon läheneb ühele, siis b_1' ja b_2' hajuvused on lõpmata suured.

[8, lk 288]

Oleme näidanud, et kui kahe tunnusega mudelis (2.1.3.1) on seletavad tunnused omavahel korreleeritud ja korrelatsioonikordaja läheneb ühele, siis regressioonikordajate hinnangute hajuvused on lõpmata suured. Mida rohkem korreleeritud tunnuseid mudelisse lisatakse, seda ebatäpsemaks muutuvad regressiooniparameetrite hinnangud [8, lk 286].

3 Praktiline osa

Polügeensete haiguste korral võivad haigusega olla seostunud palju geneetilisi variante, kus aga iga üksiku geneetilise variandi efektisuurus on väike. Haiguse geneetilise varieeruvuse paremaks kirjeldamiseks saab selliseid variante kombineerida kokku üheks. [13]

Riskiskooride arvutamisel kasutatakse tihti ülegenoomsete assotsiatsiooniuuringute (GWAS) abil leitud geneetilisi variante, mis on seotud uuritava tunnuse või haigusega. Üldiselt valitakse riskiskoori üksteisest sõltumatud geneetilised variandid. Kui riskiskoori on valitud sõltumatud markerid ja riskiskoori lineaarse regressiooni võrrandis kasutades erineb selle kordaja oluliselt ühest, on võimalik, et kohort, mille peal skoori efekti valideeritakse, erineb GWASis kasutatud kohortidest. Korreleerituse piir on kokkuleppeline ehk oleneb tavaliselt uurijast [14] [2].

Riskiskooris kasutatavate SNPide valimiseks on mitmeid meetodeid, nagu efekti suuruse või *p*-väärtuse järgi valimine. Antud töös on kasutatud ülegenoomselt olulisi SNPe.

Riskiskoor i-ndale indiviidile on kujul

$$Riskiskoor_i = \beta_1 SNP_{i1} + \cdots + \beta_n SNP_{in},$$

kus $\beta_1, \dots, \beta_n$ on GWASist saadud regressioonikordajad ja $SNP_{i1}, \dots, SNP_{in}$ on i-nda indiviidi SNPide „doosid“, st iga SNPi riskialleelide arvud.

Praktilise osa eesmärk on näidata, kuidas riskiskoori regressioonikordaja hinnang muutub vastavalt sellele, kui tugevalt on geneetilised markerid omavahel korreleeritud. Selleks koostame mudelid kehamassiindeksile, kuhu üheks seletavaks tunnuseks lisame riskiskoori, mis koosneb kehamassindeksiga seotud SNPidest, ning seejärel võrdleme riskiskoori regressiooniparameetrite hinnanguid.

3.1 Kasutatavate andmestike kirjeldused

3.1.1 Fenotüüpide andmestik

Kasutatavas Tartu Ülikooli Eesti Geenivaramu (TÜEGV) fenotüüpide andmestikus on kolm tunnust: vanus, sugu ja KMI ehk kehamassiindeks. KMI valem on kujul

$$KMI = \frac{kaal\ (kg)}{pikkus\ (m)^2}\ .$$

Kehakaal loetakse normaalseks, kui KMI jääb vahemikku 19 (vahel 18,5) kuni 25. Kui KMI on suurem kui 25, siis see võib viidata ülekaalule või rasvumisele (KMI>30). [15]

Andmestikus on 7944 geenidoonori andmed, kellest 4418 on naised ja 3526 on mehed, kes moodustavad vastavalt 55,6% ja 44,4% analüüsitavatest indiviididest (vt tabel 3.1.1.1). Keskmine vanus on 51,4 aastat ja keskmine kehamassiindeks on 26,7. Keskmise vanuse ja kehamassiindeksi jaotust sooti vaata tabelist 3.1.1.1.

Tabel 3.1.1.1 *Geenidoonorite sooline jaotus, keskmine vanus ja KMI fenotüüpide andmestikus.*

	Indiviidide arv	Keskmine vanus (sd)	Keskmine KMI (sd)
Naised	4418	52,0 (20,8)	26,7 (5,6)
Mehed	3526	50,6 (19,5)	26,8 (4,8)
Kokku	7944	51,4 (20,2)	26,7 (5,3)

3.1.2 Genotüüpide andmestik

Genotüüpide andmestikus on info 742 markeri genotüübi kohta igal inimesel, mis näitavad riskialleelide hulka ühes kindlas geenilookuses. Näiteks, kui markeri referentsalleel on A ja riskialleel on G ning indiviidil esineb DNA järjestuse teisend GG, siis andmestikus on see kodeeritud kahega. Vastavalt riskialleelide arvule saavad SNPide väärtused olla kas 0, 1 või 2, mida nimetatakse SNPi doosiks. Markerite väärtused, mida ülegenoomse markeripaneeliga ei määratud, on ennustatud, kasutades samas regioonis genotüpiseeritud markereid geneetilise imputeerimise meetodil [16]. Sellisel puhul kasutatakse genotüüpide esinemise tõenäosuseid ning G alleeli „doos" on pidev tunnus vahemikus 0-2.

3.1.3 GWASi tulemuste tabel

GWASi geenimarkerite tabelis on 742 markeri kirjeldused: markeri nimi, referentsalleel, riskialleel, markeri *p*-väärtus ja efekti suurus KMI jaoks. KMIga seotud markerid pärinevad metaanalüüsist [17], kus kasutati väga paljude Euroopa päritolu kohortide assotsiatsiooniuuringute tulemusi, millest olid eemaldatud TÜEGV geenidoonorite assotsiatsiooniuuringute tulemused. Analüüsid olid koostatud meestele ja naistele eraldi, mis metaanalüüsi käigus analüüsiti kokku. Lisaks viidi läbi pöörd-normaaljaotuse transformatsioon (*inverse normal transformation*), mille tulemusena uuritav tunnus KMI on standardiseeritud normaaljaotusega. Selleks, et metaanalüüsist saadud SNPe saaks kasutada TÜ Eesti Geenivaramu andmestikul, tuleb KMI eelnevalt transformeerida, kasutades pöörd-normaaljaotuse transformatsiooni (vt lisa 1).

Iga markeri *p*-väärtus ja efekti suurus pärineb mudelist

$$KMI = \beta_0 + \beta_1 \cdot vanus + \beta_2 \cdot vanus^2 + \beta_3 \cdot SNP,$$

kus efekti suurus on β_3 hinnang ja *p*-väärtus on β_3 hinnangu olulisustõenäosus. Andmestikku valiti kehakaaluga seotud markerid, mille *p*-väärtus on väiksem kui $5 \cdot 10^{-8}$, mida loetakse kokkuleppeliselt ülegenoomselt oluliseks seoseks. [18]

Enne efektide kasutamist riskiskooris vaadeldakse, kas metaanalüüsist [17] saadud markerite riskialleelid ühtivad TÜEGV andmestiku samade SNPide riskialleelidega. Kui markeri riskialleelid andmestikes ei ühti, siis muudetakse SNPi efekti märk vastupidiseks (vt lisa 1). Efektide märkide õigsuse graafiliseks kontrollimiseks saab kasutada joonist lisas 4.

3.2 Näide: regressiooniparameetrite muutumine tugevalt korreleeritud tunnustega mudelis

Teoreetilise osa olulisema tulemuse (regressioonikordajate muutumine ebatäpseteks, kui tunnused on mudelis korreleeritud) tähtsuse selgitamiseks teeme läbi ühe näite. Valime andmestikust kaks kõige väiksema *p*-väärtusega markerit, mille omavaheline korrelatsioon ei ole täpselt 1 ning koostame kolm mudelit, kasutades TÜEGV andmeid. Uuritavaks tunnuseks on KMI pöörd-normaaljaotusele transformeeritud kujul. Kriteeriumi järgi valitud markeriteks osutusid rs1421085 (SNP1), mille *p*-väärtus on $8{,}05 \cdot 10^{-62}$ ja rs17817449 (SNP2), mille *p*-väärtus on $7{,}77 \cdot 10^{-61}$ (vt lisa 1).

Mudel, kus seletavaks tunnuseks on SNP1, on kujul

$$KMI = -0{,}066 + 0{,}073 \cdot SNP1 \tag{3.2.1}$$

Vabaliige ja regressioonikordajad on mudelis olulised olulisusenivool $\alpha = 0{,}05$. Korrelatsioon KMI ja SNP1 vahel on $r_{BMI,SNP1} = 0{,}0472$, st mudeli headus on 0,0022 ehk mudel kirjeldab pöörd-normaaljaotusele transformeeritud kehamassiindeksi hajuvusest 0,22%.

Mudel tunnusega SNP2 on kujul

$$KMI = -0{,}061 + 0{,}069 \cdot SNP2 \tag{3.2.2}$$

Ka siin on vabaliige ja tunnus SNP2 olulised. Kehamassindeksi ja SNP2 vaheline korrelatsioon on $r_{BMI,SNP2} = 0{,}0449$. Mudel kirjeldab 0,20% kehamassiindeksi hajuvusest.

Mitmese regressiooni mudel seletavate tunnustega SNP1 ja SNP2 on kujul

$$KMI = -0{,}066 + 0{,}112 \cdot SNP1 - 0{,}041 \cdot SNP2$$

SNPide vaheline korrelatsioon on väga tugev: $r_{SNP1,SNP2} = 0{,}974$. Teoreetilisest osast on teada, et sellisel juhul on b_1 ja b_2 ehk SNP1 ja SNP2 regressiooniparameetrite hinnangud suurte hajuvustega ning ei ole sarnased mudelite (3.2.1) ja (3.2.2) regressiooniparameetrite hinnangutele.

Pöörd-normaaljaotusele transformeeritud kehamassiindeksi hajuvusest kirjeldab mudel 0,22%. SNP1 ja SNP2 efektide olulisustõenäosused on vastavalt 0,109 ja 0,562 (95%-lised usaldusintervallid on (-0,025;0,249) ja (-0,178;0,096)), mistõttu SNP1 ja SNP2 ei ole mudelis olulised ja see on teadaolev multikollineaarsuse probleem [12, lk 78].

3.3 Mudelite koostamine

Eelnevalt läbi tehtud näite põhjal selgus, et SNPide efektide suurus sõltub teistest mudelis olevatest SNPidest ja nende omavahelistest korrelatsioonidest. Järgnevalt uurime, kuidas käitub riskiskoori kordaja sõltuvalt sellest, kui tugevalt on riskiskooris olevad SNPid omavahel korreleeritud.

3.3.1 Riskiskooride moodustamine

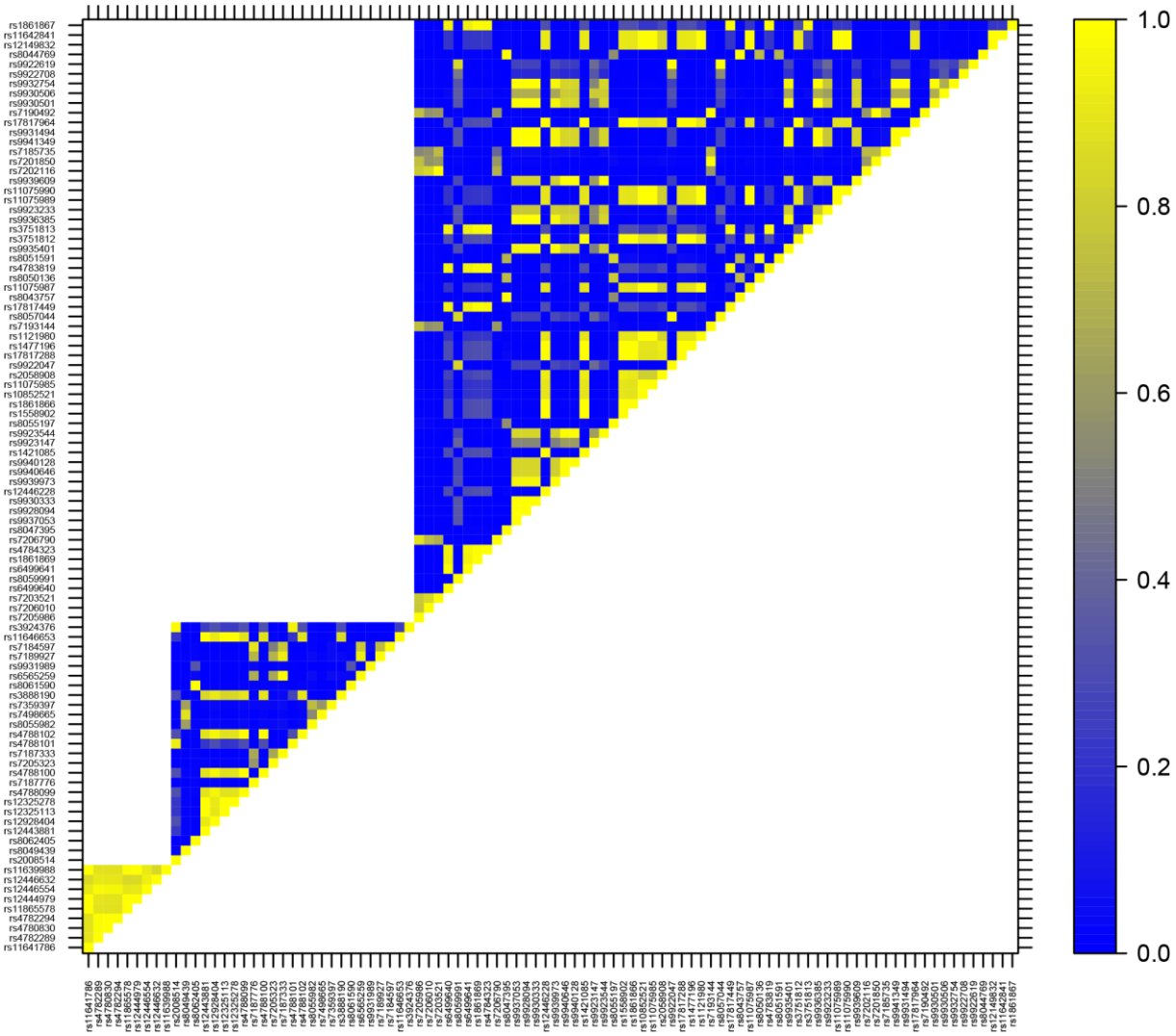
Nagu eelpool mainitud, osutusid 742 kehamassiindeksiga seotud markerit GWASis olulisteks. SNPid asuvad erinevates kromosoomides ja on teada SNPide füüsiline asukoht. Omavahelisi korrelatsioone vaatleme SNPidel, mille füüsiline kaugus (füüsiliste asukohtade vahe) on väiksem või võrdne 500 000 aluspaariga ja mis asuvad ühes kromosoomis, sest erinevates kromosoomides asuvate markerite ja samal kromosoomil, ent suurema distantsiga markerite koospärandumist võib lugeda juhuslikuks tänu meioosis toimuvale kromosoomide rekombineerumisele. SNPide jaotust erinevates kromosoomides vt lisas 2.

Riskiskoorid moodustame SNPide vaheliste korrelatsioonide absoluutväärtuste järgi (vt joonis 3.3.1.1 ja lisa 3), alustades poollõigust [0;0,05), st esimesena võtame riskiskoori need SNPid, mille korrelatsiooni absoluutväärtus mistahes teise sama kromosoomi SNPiga on väiksem kui 0,05 ning füüsiline kaugus väiksem või võrdne kui 500 000. Järgmisena võtame poollõigu [0;0,1) jne. Viimaseks poollõiguks on [0;1).

Seega riskiskoore, mis on moodustatud erineva arvu SNPide põhjal olenevalt SNPide vahelisest korrelatsioonist, kasutame 20 mudelis

$$KMI_i = \beta_{0i} + \beta_{1i} \cdot vanus \ + \ \beta_{2i} \cdot sugu \ + \ \beta_{3i} \cdot riskiskoor_i,$$

kus $i = 1, \dots, 20$. Sugu on binaarne tunnus (naine=1, mees=0).



Joonis 3.3.1.1 *SNPide vahelised korrelatsioonid kromosoomis 16. Korrelatsioonidest on võetud absoluutväärtus. Valgega on tähistatud korrelatsioonid, mida ei arvutatud, sest markerid olid rohkem kui 500 000 aluspaari kaugusel.*

3.4 Tulemused

Vanuse ja soo suhtes kohandatud 20 mudeli tulemused, kus hinnatakse riskiskoori, mis on moodustatud erineva arvu SNPide põhjal olenevalt SNPide vahelisest korrelatsioonist, mõju kehamassiindeksile on esitatud tabelis 3.4.1.

Tabel 3.4.1 *Riskiskoori kordajad ja 95%-lised usaldusintervallid korrelatsiooniintervallides.*

Korrelatsiooni-intervall	SNPide arv	Riskiskoori kordaja	95%-line usaldusintervall
[0;0,05)	20	0,868	(0,678;1,058)
[0;0,1)	20	0,868	(0,678;1,058)
[0;0,15)	21	0,833	(0,646;1,020)
[0;0,2)	22	0,793	(0,613;0,973)
[0;0,25)	22	0,798	(0,619;0,977)
[0;0,3)	23	0,731	(0,560;0,903)
[0;0,35)	24	0,729	(0.564;0,893)
[0;0,4)	25	0,691	(0,530;0,851)
[0;0,45)	29	0,622	(0,475;0,769)
[0;0,5)	31	0,568	(0,430;0,705)
[0;0,55)	35	0,476	(0,357;0,596)
[0;0,6)	37	0,423	(0,311;0,534)
[0;0,65)	42	0,382	(0,281;0,483)
[0;0,7)	44	0,371	(0,275;0,466)
[0;0,75)	51	0,308	(0,228;0,388)
[0;0,8)	63	0,208	(0,150;0,266)
[0;0,85)	74	0,177	(0,129;0,225)
[0;0,9]	93	0,129	(0,093;0,164)
[0;0,95)	128	0,100	(0,073;0,127)
[0;1)	567	0,020	(0,015;0,025)

Mudel, kus riskiskoori SNPide vahelised korrelatsioonid on väiksemad, kui 0,05, on kujul

$$KMI_1 = -0{,}357 - 0{,}001 \cdot vanus - 0{,}067 \cdot sugu + 0{,}868 \cdot riskiskoor_1.$$

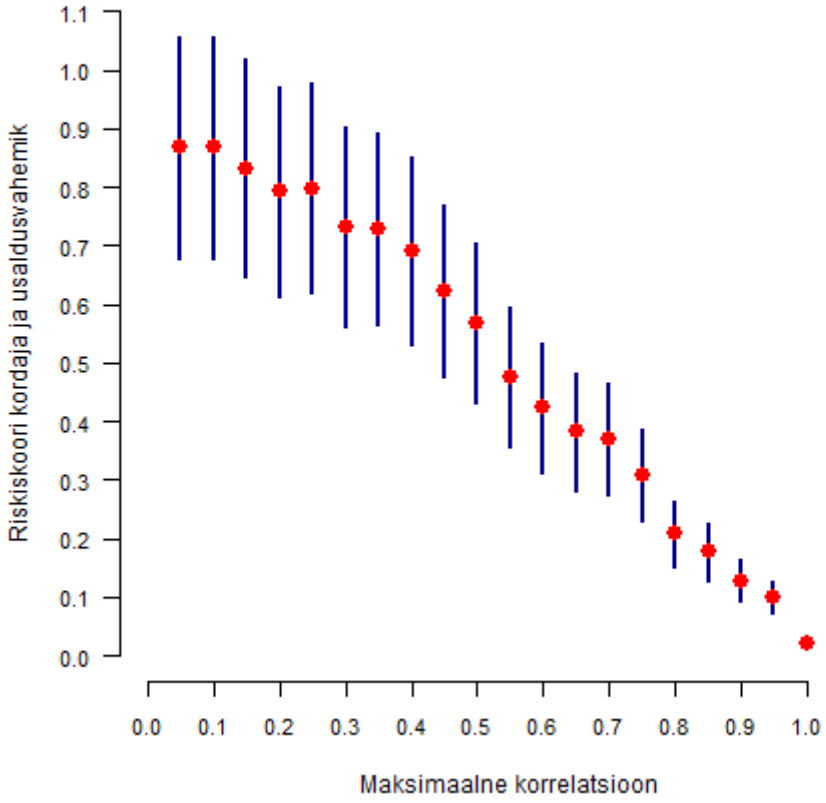
Mudel kirjeldab pöörd-normaaljaotusele transformeeritud kehamassindeksi hajuvusest 1,15%. Vanus ei ole mudelis oluline: olulisustõenäosus on 0,108 ja parameetri usaldusvahemik on (-0,002;0,0002). Sugu ja riskiskoor on olulised: *p*-väärtused on vastavalt 0,003 ja <0,001 ning usaldusvahemikud on (-0,111;-0,023) ja (0,678;1,058).

Mudel, kus riskiskoori SNPide vahelised korrelatsioonid on poollõigust [0,1), on kujul

$$KMI_{20} = -0{,}146 - 0{,}001 \cdot vanus - 0{,}067 \cdot sugu + 0{,}020 \cdot riskiskoor_{20}.$$

Mudel kirjeldab pöörd-normaaljaotusele transformeeritud kehamassiindeksi hajuvusest 0,79%. Ka selles mudelis ei ole vanus oluline, sest olulisustõenäosus on 0,084 ja 95%-line usaldusvahemik on (-0,002;0,0001). Statistiliselt olulised tunnused on sugu ja riskiskoor, olulisustõenäosuste ja usaldusvahemikega vastavalt 0,003 ja <0,001 ning (-0,111;-0,023) ja (0,015;0,025).

Riskiskoori kordajate ja usaldusvahemike muutumist illustreerib joonis 3.4.1.



Joonis 3.4.1 *Riskiskoori kordajad ja vastavad usaldusvahemikud erinevate maksimaalsete markeritevaheliste korrelatsioonide korral. Viimase kordaja usaldusvahemik on väga kitsas ja ei paista punkti tagant välja.*

Eeldusel, et efektid on tõelähedased (kirjeldavad SNPi tõelist mõju uuritavale tunnusele) ja SNPid on üksteisest sõltumatud, on riskiskoori kordaja väärtus ligilähedane ühele. Jooniselt 3.4.1 näeme, et tulemused on ootuspärased – sõltumatutest SNPidest koosneva riskiskoori kordaja on ligikaudu üks. Mida tugevam on korrelatsioon SNPide vahel, seda väiksem ühest on riskiskoori regressiooniparameetri hinnang. Seda seetõttu, et skoori lisatakse korreleerituse kasvades üha rohkem SNPe, mis kirjeldavad sama osa uuritava tunnuse varieeruvusest, ning samal ajal kasutatakse iga markeri puhul GWASi efekti, mis ei võta arvesse võimalikke omavahelisi korrelatsioone. Sellise teguviisi tulemusel on aga riskiskoori väärtused inimestel tegelikkusest kõrgemad ning seetõttu skoori regressioonikordaja üha väheneb. Väärib mainimist, et sellises olukorras ei ole skoori regressioonikordaja kohta võimalik eestlaste populatsiooni eripära kohta midagi väita.

Seega, kui riskiskoori efekti soovitakse valideerida haiguse riski või tunnuse väärtuse hindamiseks, siis on oluline valida sõltumatud SNPid.

Kokkuvõte

Antud töös kasutati geneetilist riskiskoori, mis on SNPidest ehk ühenukleotiidsetest polümorfismidest koosnev arvuline väärtus, mis võimaldab hinnata haigestumise riski või tunnuse väärtust ning moodustatakse GWASist saadud sõltumatute markerite riskialleelide summana, mis on kaalutud efektisuurustega. Käesoleva bakalaureusetöö eesmärk oli välja selgitada, kuidas riskiskoori kui seletava tunnuse kordaja muutub, kui riskiskoor on moodustatud erineva tugevusega korreleeritud SNPidest.

Töö teoreetilises osas uuriti, kuidas lineaarse regressioonimudeli parameetrid muutuvad, kui seletavad tunnused on omavahel korreleeritud. Selgus, et kui tunnustevaheline korrelatsioon läheneb ühele, siis parameetrite hajuvused on lõpmata suured.

Praktilises osas kasutati Tartu Ülikooli Eesti Geenivaramu andmestikku, mis koosnes 7944 geenidoonori andmetest, ja GWASist saadud geenimarkereid ning moodustati kehamassiindeksi mudelid, kuhu üheks seletavaks tunnuseks lisati riskiskoor. Teisteks seletavateks tunnusteks olid vanus ja sugu.

Kui SNPid on sõltumatud ja nende efektid on tõelised, siis riskiskoori kordaja on ligikaudu üks. Selgus, et mida suurem on riskiskooris olevate SNPide omavaheline korrelatsioon, seda väiksem ühest on riskiskoori kordaja. Sellises mudelis, kus SNPide vaheline maksimaalne korrelatsioon oli väiksem kui 0,05, oli riskiskoori kordaja 0,87 (usaldusvahemik (0,678;1,058)). Mudelis, kuhu oli lisatud SNPe korrelatsioonidega poollõigust [0;1), oli riskiskoori kordaja 0,02 (usaldusvahemik (0,015;0,025)). Liigselt korreleeritud markerite lisamine riskiskoori muudab kordaja ühest oluliselt erinevaks ning eestlaste populatsiooni eripära kohta ei ole võimalik regressioonikordajast järeldusi teha. Seetõttu on oluline valida riskiskoori sõltumatud markerid.

Kasutatud kirjandus

[1] Andersson, E. A., et al.2013. Genetic risk score of 46 type 2 diabetes risk variants associates with changes in plasma glucose and estimates of pancreatic β-cell function over 5 years of follow-up. *Diabetes.* 2013 Oct, 62(10). Artikkel PubMed Central'ist. Kättesaadav: http://www.ncbi.nlm.nih.gov/pubmed/23835328 [Vaadatud: 04.05.2014]

[2] Belsky, D. W., et al., 2013. Development and Evaluation of a Genetic Risk Score for Obesity. *Biodemography and Social Biology* [online] Kättesaadav: http://www.tandfonline.com/doi/full/10.1080/19485565.2013.774628#tabModule

[Vaadatud: 04.05.2014]

[3] R, 2014. *The R-project for statistical computing.* [online] Kättesaadav:

http://www.r-project.org/ [Vaadatud: 04.05.2014]

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

[5] Heinaru, A., 2012. *Geneetika*. Tartu: Tartu Ülikooli Kirjastus.

[6] Syvänen A. C., 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature reviews. Genetics.* Dec; 2(12). Artikkel PubMed Central'ist. Kättesaadav: http://www.ncbi.nlm.nih.gov/pubmed/11733746 [Vaadatud: 05.05.2014]

[7] Liu, S., Song, Y., 2010. Building Genetic Scores to Predict Risk of Complex Diseases in Humans: Is It Possible? *Diabetes*, [online] Kättesaadav:

http://diabetes.diabetesjournals.org/content/59/11/2729.full.pdf+html [Vaadatud: 29.04.2014]

[8] Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W., 2004. *Applied Linear Statistical Models*. 5th ed. New York: McGraw-Hill/Irwin.

[9] Käärik, E., 2013. *Andmeanalüüs II*. Loengukonspekt. Tartu: Tartu Ülikool, matemaatilise statistika instituut.

[10] Möls, M., 2010. *Lineaarsed mudelid*. Loengukonspekt. Tartu: Tartu Ülikool, matemaatilise statistika instituut.

[11] Williams, R., 2004. *Standardized Coefficients*, [pdf]. University of Notre Dame. Kättesaadav: http://www3.nd.edu/~rwilliam/stats1/x92.pdf [Vaadatud:29.04.2014]

[12] Janssen, P., 2012. Part 1: Regression, *0377 Foundations of linear models.* University of Hasselt,unpublished.

[13] Dudbridge, F., 2013. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, [online] Kättesaadav:

http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1003348 [Vaadatud: 29.04.2014]

[14] Morris, A. P., et al., 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* Sep; 44(9). Artikkel PubMed Central'ist. Kättesaadav: http://www.ncbi.nlm.nih.gov/pubmed/22885922 [Vaadatud 05.05.2014]

[15] Vikipeedia, 2014. Kehamassiindeks [online] (15. november 2013) Kättesaadav: http://et.wikipedia.org/wiki/Kehamassiindeks [Vaadatud 29.04.2014]

[16] Marchini J., Howie B., 2010. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics.* Jul; 11(7). Artikkel PubMed Central'ist. Kättesaadav: http://www.ncbi.nlm.nih.gov/pubmed/20517342 [Vaadatud: 05.05.2014]

[17] Speliotes, E. K. *et al.* 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* [online] Kättesaadav:

http://www.nature.com/ng/journal/v42/n11/full/ng.686.html#methods [Vaadatud: 29.04.2014]

[18] Bush W. S., Moore J. H, 2012. Chapter 11: Genome-wide association studies. *PloS Computational Biology*. 8(12). Artikkel PubMed Central'ist. Kättesaadav: http://www.ncbi.nlm.nih.gov/pubmed/23300413 [Vaadatud: 05.05.2014]

Lisad

Lisa 1. Programmikood

```
# FAILID

#load workspace korrelatsioonid, omni
m_g2 <- read.table(file =
"C:\\Users\\kasutaja\\Desktop\\materjal\\puuduvad_kromosoomid.txt", header
= TRUE, sep = "\t")
m_gws <- read.table(file =
"C:\\Users\\kasutaja\\Desktop\\materjal\\GWS_gc.txt", header = TRUE, sep =
"\t")

#faili "andmed"  ja "info" kopeerimine
andmed2 <- andmed
info2<-info

#PUUDUVATE KROMOSOOMIDE ASENDAMINE FAILIS info2
info2$RS<-as.character(info2$RS)
m_g2$MarkerName<-as.character(m_g2$MarkerName)
for (i in 1:742) {
    if (info2$chr[i]=="---")  {
        rs=info2$RS[i]
        for (j in 1:743){
            if (m_g2$MarkerName[j]==rs) { rs_jrknr=j }
            }
    info2$chr[i]=m_g2$CHR[rs_jrknr]
        }
}

#SNPIDE ANDMETABELID ERINEVATES KROMOSOOMIDES
info2_chr1= info2[info2$chr=="1",]
info2_chr2= info2[info2$chr=="2",]
info2_chr3= info2[info2$chr=="3",]
info2_chr4= info2[info2$chr=="4",]
info2_chr5= info2[info2$chr=="5",]
info2_chr6= info2[info2$chr=="6",]
info2_chr9= info2[info2$chr=="9",]
info2_chr11= info2[info2$chr=="11",]
info2_chr12= info2[info2$chr=="12",]
info2_chr14= info2[info2$chr=="14",]
info2_chr15= info2[info2$chr=="15",]
info2_chr16= info2[info2$chr=="16",]
info2_chr18= info2[info2$chr=="18",]
info2_chr19= info2[info2$chr=="19",]


#KORRELATSIOONIMAATRIKSEID ARVUTAV FUNKTSIOON

#positsioonide järgi sorteerida
library(plyr)
info2_chr16 <-arrange(info2_chr16, POS)
rownames(info2_chr16)=info2_chr16$POS

#korrelatsioonimaatrikseid arvutav funktsioon
kormatarvutaja=function(chr_tabel){
kormat=matrix(nrow=nrow(chr_tabel), ncol=nrow(chr_tabel))
colnames(kormat)<-chr_tabel$RS
```

```
rownames(kormat)<-chr_tabel$RS
    for(i in 1:(nrow(chr_tabel))) {
        for(j in 1:(nrow(chr_tabel))) {
            chr_tabel$POS=as.numeric(chr_tabel$POS)
            chr_tabel$RS=as.factor(chr_tabel$RS)

            if (abs(chr_tabel$POS[i]-chr_tabel$POS[j])<500000) {
                kormat[i,j]=cor(andmed2[,chr_tabel$RS[i]],
andmed2[,chr_tabel$RS[j]], use="everything", method="pearson")
                }
        }
    }
return(kormat)
}
korrelatsioonimaatriks=kormatarvutaja(info2_chr16)

#NEGATIIVSED KORRELATSIOONID POSITIIVSETEKS
for(i in 1:nrow(korrelatsioonimaatriks)){
    for(j in 1:ncol(korrelatsioonimaatriks)) {
        if (korrelatsioonimaatriks[i,j]<0
& !is.na(korrelatsioonimaatriks[i,j]) ){
            korrelatsioonimaatriks[i,j]=korrelatsioonimaatriks[i,j]*(-1)
            korrelatsioonimaatriks[i,j]
        }
    }
}
#võtame ülemise kolmnurkmaatriksi
korrelatsioonimaatriks[lower.tri(korrelatsioonimaatriks, diag=F)] <-NA

#KORRELATSIOONIPILDID
pdf('chr16_levelplot.pdf)
library(lattice)
rgb.palette <- colorRampPalette(c("blue", "yellow"), space = "rgb")
levelplot(korrelatsioonimaatriks, scales=list(x=list(cex=0.4,
rot=90),y=list(cex=0.4)),xlab="", ylab="",col.regions=rgb.palette(120),
cuts=100, at=seq(0,1,0.01))
dev.off()


#failide korrastamine
m_gws_sort <-arrange(m_gws, P_meie)

#2 olulist snippi
snp1=m_gws_sort$MARKER[1]
snp2=m_gws_sort$MARKER[3]

#sorteerime mõlemad rs numbri järgi
info_sort=arrange(info2,RS)
gws_sort=arrange(m_gws, MARKER)

#EFEKTIDELE ÕIGED MÄRGID

as.character(gws_sort$MARKER)
gws_sort <- gws_sort[ gws_sort$MARKER %in% info_sort$RS,]
gws_sort$poorBETA=gws_sort$BETA

for(i in 1:nrow(gws_sort)){
    if(as.character(info_sort$A2[i])!=as.character(gws_sort$EA[i])){
    gws_sort$poorBETA[i]=(-1)*gws_sort$BETA[i]
    }
    else{}
```

```
}

#EFEKTIDE MÄRKIDE ÕIGSUSE KONTROLLIMISE JOONIS
png("betade muutmine.png")
plot(gws_sort$Freq1,info_sort$eaf,col=c("green",
"yellow")[(sign(gws_sort$poorBETA)==sign(gws_sort$BETA))*1+1],
xlab="Metaanalüüsi riskialleelide sagedused", ylab="TÜEGV andmestiku
riskialleelide sagedused")
dev.off()

# BMI teisendamine
fenot4$invBMI=NA
model1 <- lm(fenot4$bmi ~ fenot4$vanus+(fenot4$vanus)^2)
resBMI <- resid(model1)
invBMI<-qnorm((rank(resBMI ,na.last="keep")-0.5)/sum(!is.na(resBMI)))
##4490, 3154, 3320, 141
invBMI <- append(invBMI, NA, 140)
invBMI <- append(invBMI, NA, 3153)
invBMI <- append(invBMI, NA, 3319)
invBMI <- append(invBMI, NA, 4489)
length(invBMI)
fenot4$invBMI=invBMI

# MUDELID näite jaoks
mud_snp1 = lm(fenot4$invBMI~kokku2$rs1421085)
summary(mud_snp1)
coefficients(mud_snp1)

mud_snp2=lm(fenot4$invBMI~kokku2$rs17817449)
summary(mud_snp2)
coefficients(mud_snp2)

kor_snp1_snp2=cor(kokku2$rs1421085, kokku2$rs17817449, use="everything",
method="pearson")
kor_snp1_snp2

mud_snp1_snp2=lm(fenot4$invBMI~kokku2$rs1421085+kokku2$rs17817449)
summary(mud_snp1_snp2)
coefficients(mud_snp1_snp2)
confint(mud_snp1_snp2)

#RISKISKOORIDE MOODUSTAMINE

#KASUTATAVAD FAILID
#failist gws_sort võtta p-väärtused ja panna failile info2
info2["pvalue"]<-NA
info2$RS <- as.character(info2$RS)
gws_sort$MARKER <-  as.character(gws_sort$MARKER)

for(i in 1:nrow(info2)){
    for(j in 1:nrow(gws_sort)){
        if(info2$RS[i]==gws_sort$MARKER[j]){
            info2$pvalue[i]<-gws_sort$P_meie[j]}
      }
}


#Sorteerime info2 p-väärtuse järgi
info2=arrange(info2,pvalue)
```

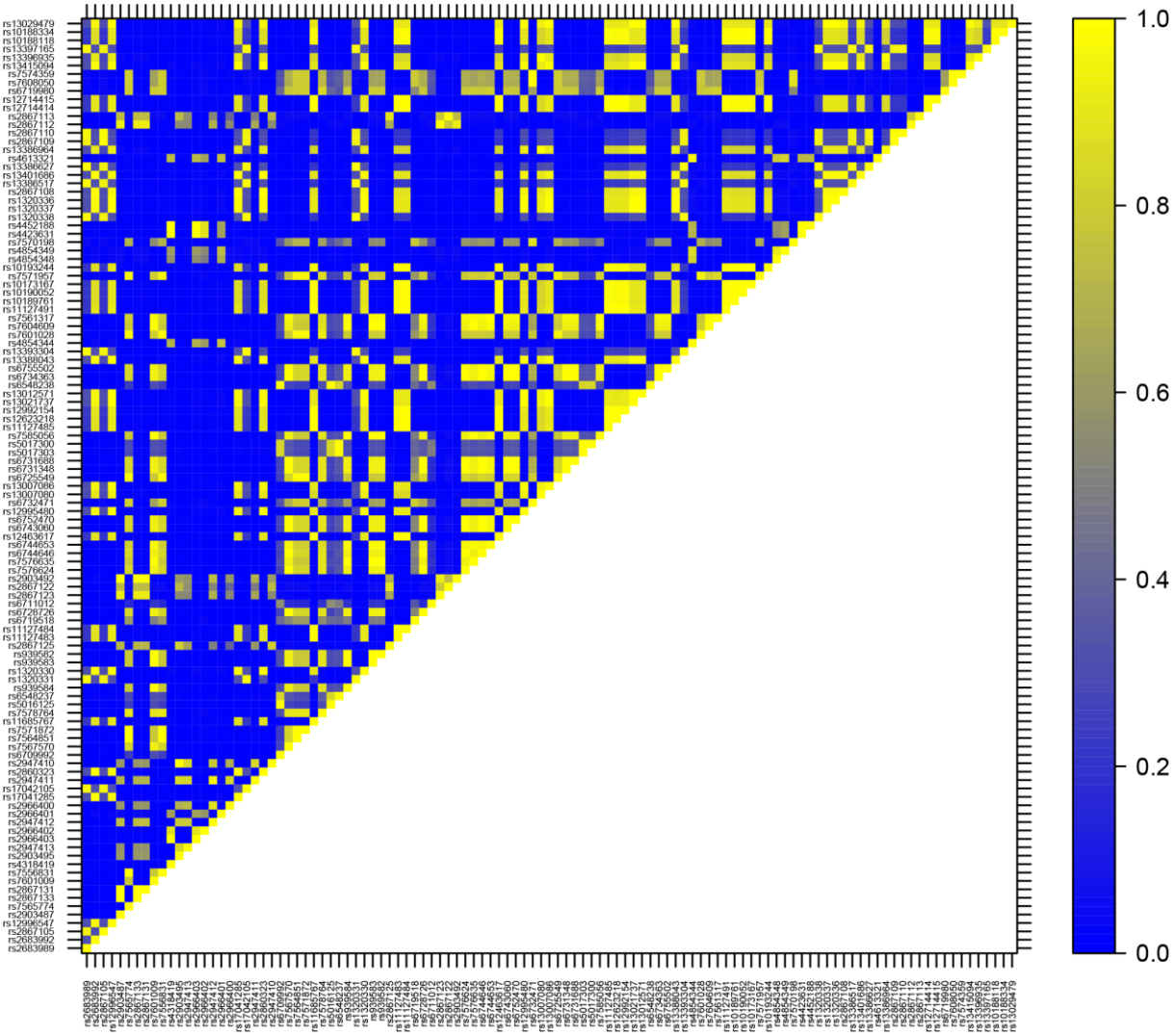
```
#FUNKTSIOON RISKISKOORIDE MOODUSTAMISEKS. Funktsiooni koodi kirjutas Kristi
Läll. Töö autor modifitseeris.
soltumatud=function(info,andmed2,kor){
k=1
kaasata2=rbind()
while( dim(info)[1]>0){
    koht=which((abs(info$POS[k]-info$POS)<=500000) &
(as.numeric(info$chr[k])==as.numeric(info$chr)))

    if (length(koht)>1){
    maatr=rep(0,length(koht)-1)
    for (i in 1:length(maatr)){

maatr[i]=abs(cor(andmed2[,as.character(info$RS[koht[k]])],andmed2[,as.chara
cter(info$RS[koht[i+1]] )])) }
        koht2=koht[which(maatr**2>=(kor^2))+1]
    kaasata2=rbind(kaasata2,info$RS[k])
    info=info[-c(k,koht2),]
    }
if (length(koht)==1){
kaasata2=rbind(kaasata2, info$RS[koht])
info=info[-koht,]}
}
return (kaasata2)}

# Riskiskooride setid
set1=soltumatud(info2,andmed2, 0.05)
set2=soltumatud(info2,andmed2, 0.1)
set3=soltumatud(info2,andmed2, 0.15)
set4=soltumatud(info2,andmed2, 0.2)
set5=soltumatud(info2,andmed2, 0.25)
set6=soltumatud(info2,andmed2, 0.3)
set7=soltumatud(info2,andmed2, 0.35)
set8=soltumatud(info2,andmed2, 0.4)
set9=soltumatud(info2,andmed2, 0.45)
set10=soltumatud(info2,andmed2, 0.5)
set11=soltumatud(info2,andmed2, 0.55)
set12=soltumatud(info2,andmed2, 0.6)
set13=soltumatud(info2,andmed2, 0.65)
set14=soltumatud(info2,andmed2, 0.7)
set15=soltumatud(info2,andmed2, 0.75)
set16=soltumatud(info2,andmed2, 0.8)
set17=soltumatud(info2,andmed2, 0.85)
set18=soltumatud(info2,andmed2, 0.9)
set19=soltumatud(info2,andmed2, 0.95)
set20=soltumatud(info2,andmed2, 1)

# RISKISKOORIDE ARVUTAMINE
skoorid=cbind()
setid<-data.frame(matrix(NA,ncol=20, nrow=567))
setid[1:20,1]=set1
setid[1:20,2]=set2
setid[1:21,3]=set3
setid[1:22,4]=set4
setid[1:22,5]=set5
setid[1:23,6]=set6
setid[1:24,7]=set7
setid[1:25,8]=set8
setid[1:29,9]=set9
setid[1:31,10]=set10
setid[1:35,11]=set11
```

```
setid[1:37,12]=set12
setid[1:42,13]=set13
setid[1:44,14]=set14
setid[1:51,15]=set15
setid[1:63,16]=set16
setid[1:74,17]=set17
setid[1:93,18]=set18
setid[1:128,19]=set19
setid[,20]=set20

skoorid=cbind()
rownames(gws_sort)=gws_sort$MARKER
for(i in 1:20){
sk=(as.matrix(kokku2[,setid[,i][which(!is.na(setid[,i]))]]))%*%(gws_sort[as
.character(setid[,i][which(!is.na(setid[,i]))]),"poorBETA"])
    skoorid=cbind(skoorid,sk)
}

# MUDELID

#mudel 1
m1=lm(fenot4$invBMI~fenot4$vanus+fenot4$sugu+ skoorid[,1])
summary(m1)
coefficients(m1)
confint(m1)

#salvestame beetad ja usaldusintervallid vektoritesse
betad=rbind()
alumised=rbind()
ulemised=rbind()
for (i in 1:20){
mudel=lm(fenot4$invBMI~fenot4$vanus +fenot4$sugu+ skoorid[,i])
betad=rbind(betad,coefficients(mudel)[4])
alumised=rbind(alumised,confint(mudel)[4,1])
ulemised=rbind(ulemised,confint(mudel)[4,2])
}

m=cbind(betad,alumised,ulemised)
# JOONIS

xtelg=seq(0.05,1,0.05)
library(ggplot2)

png('joonis.png')
plot(xtelg,m[,1], xlim=c(0,1.1),ylim=c(0,1.1), xlab="Maksimaalne
korrelatsioon", ylab="Riskiskoori kordaja ja usaldusvahemik", type="p",
axes=F, col="red", pch=16, cex=1)
for(i in 1:20){
    lines(c(xtelg[i],xtelg[i]),c(m[i,2],m[i,3]), col="darkblue", lwd=2)
}
points(xtelg,m[,1], , col="red", pch=16, cex=1.5)
par(cex=0.9)
axis(1, at=seq(0,1,0.1))
axis(2, at=seq(0,1.1,0.1), las=1)
dev.off()
```

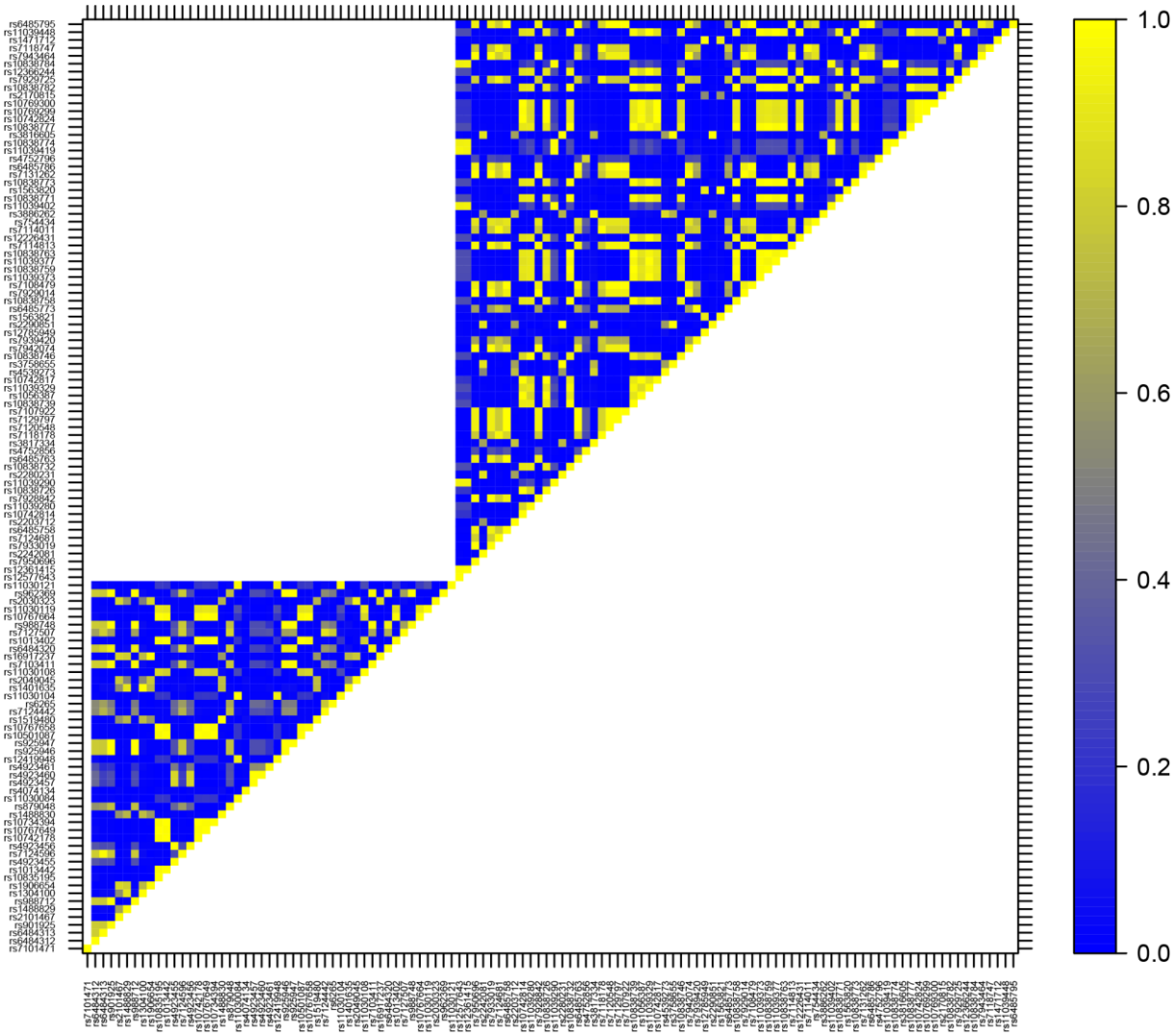
Lisa 2. SNPide jaotus kromosoomides

Kromosoom	SNPide arv kromosoomis
1	152
2	111
3	15
4	11
5	1
6	12
9	7
11	118
12	10
14	4
15	35
16	96
18	159
19	11
Kokku	742

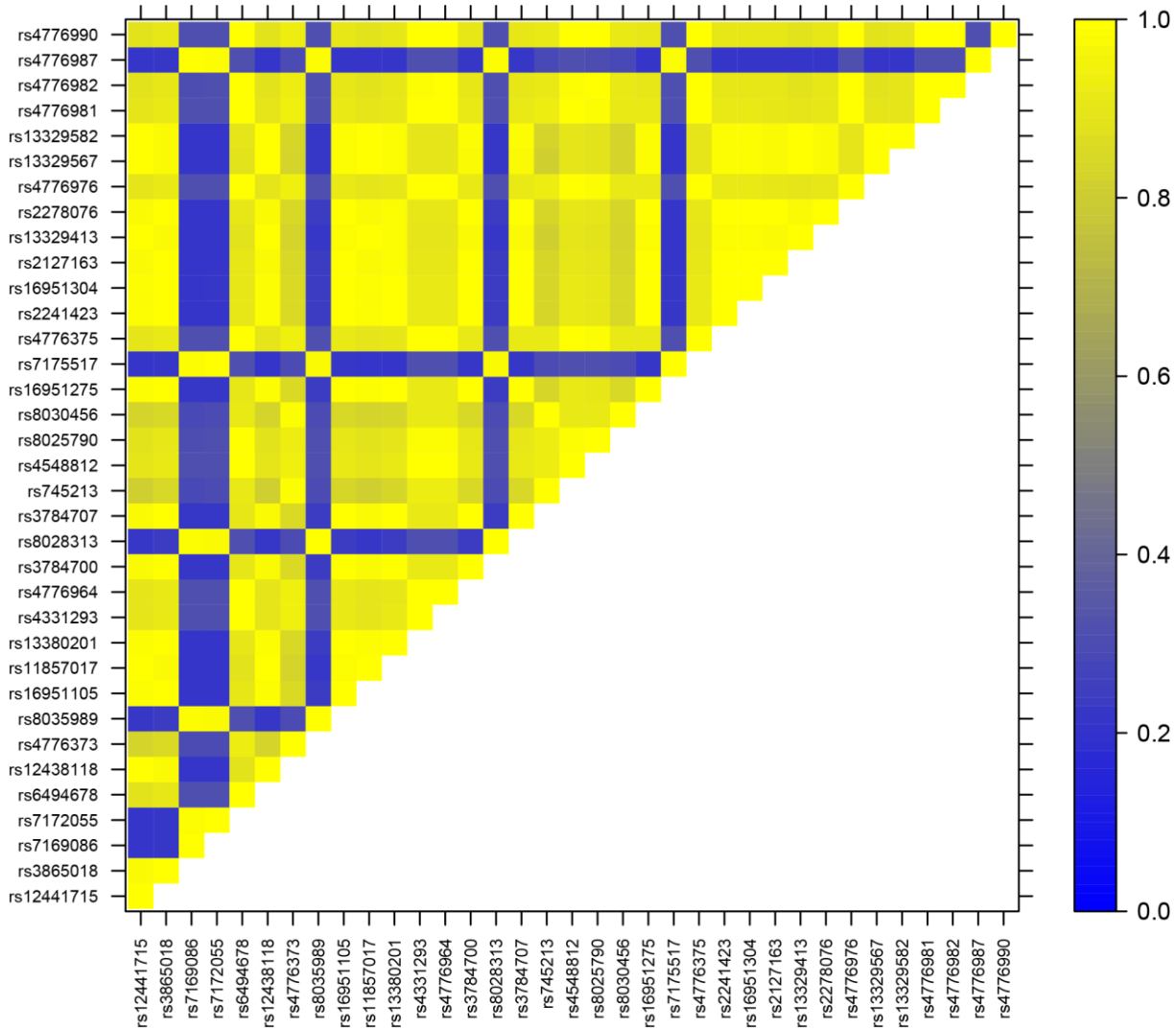
Lisa 3. Korrelatsioonide joonised mõnedes kromosoomides



Korrelatsioonid kromosoomis 2.

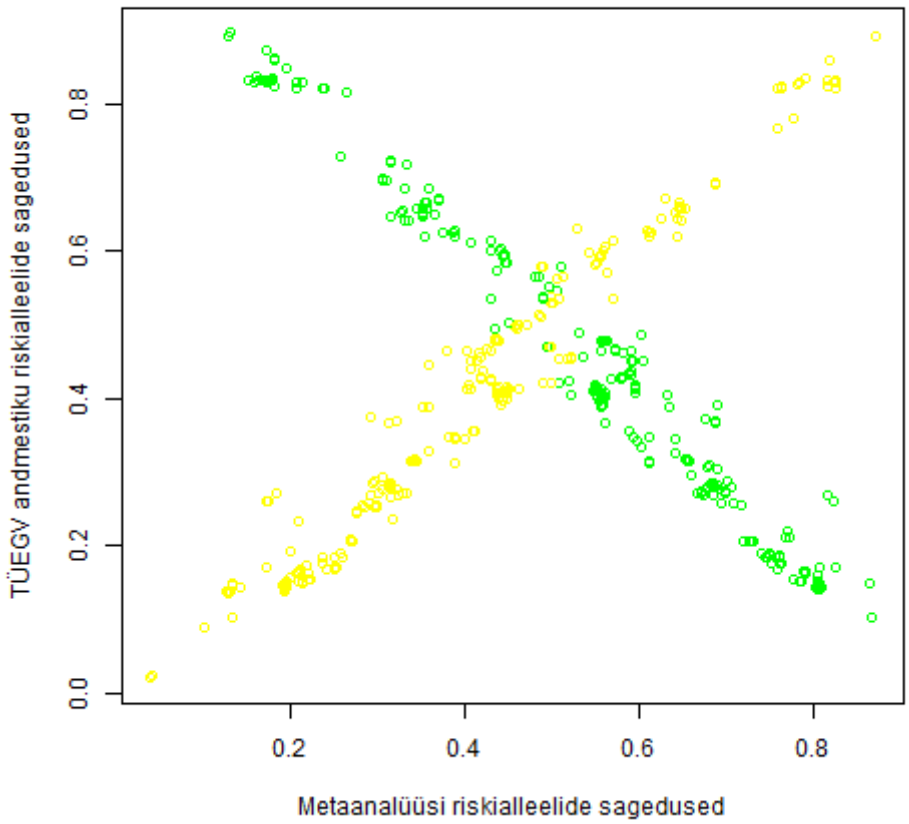


Korrelatsioonid kromosoomis 11.



Korrelatsioonid kromosoomis 15.

Lisa 4. Efektide märkide õigsuse kontrollimise joonis



Kollasega on tähistatud need SNPid, mille efektid ei vaja muutmist. Rohelisega tähistatud SNPide efektide märgid muudetakse vastupidiseks.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Maia Arge,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Markerite sõltumatuse olulisus geneetilistes riskiskoorides“, mille juhendajateks on Reedik Mägi ja Kristi Läll,

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, 05.05.2014