TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Kristin Jesse

Ülegenoomsete geen-geen koosmõjude hindamise meetodi valideerimine

Bakalaureusetöö (6 EAP)

Juhendajad:

Toomas Haller, PhD

Krista Fischer, PhD

Tartu 2014

Ülegenoomsete geen-geen koosmõjude hindamise meetodi valideerimine

Käesoleva bakalaureusetöö eesmärgiks on uue meetodi valideerimine geen-geen koosmõjude uurimiseks ülegenoomsetes assotsiatsiooniuuringutes. Esmalt tuuakse ülevaade ülegenoomsetest uuringutest ja seal kasutatavast statistilisest metoodikast ning bioloogilise ja statistilise koosmõju mõistetest. Järgnevalt kirjeldatakse kahte meetodit, mille abil koosmõjusid määrata, ning rakendatakse neid simuleeritud andmetel ja Tartu Ülikooli Eesti Geenivaramu andmetel. Esimene meetod reaalsetel andmetel ei tööta, kuid teine meetod annab rahuldavaid tulemusi. Teist meetodit plaanitakse rakendada uues geen-geen koosmõjude määramise programmis.

Märksõnad: geneetilised assotsiatsiooniuuringud, interaktsioonid, statistilised meetodid, valideerimine

Validation of a method for estimating genome wide gene-gene interactions

The goal of this thesis is to validate a new method for studying gene-gene interactions in genome wide association studies. Firstly, an overview of genome wide studies, statistical methods and biological and statistical interactions is given. Next, two methods for determining interactions are described, and applied to simulated data and the data of the Estonian Genome Center of the University of Tartu. The first method does not work on real data, but the second method offers satisfactory results. The second method will be used in a new program for detecting gene-gene interactions.

Keywords: genetic association studies, interactions, statistical methods, validation

Sisukord

Sissejuhatus

Ülegenoomsed assotsiatsiooniuuringud on viimase kümnendi jooksul muutunud oluliseks genotüüpide ning fenotüüpide vaheliste seoste uurimise meetodiks. Enamasti vaadatakse korraga ühe geenimarkeri mõju fenotüübile, kuid sel juhul ei arvestata võimalike koosmõjudega mitme markeri vahel. Seetõttu on vaja uurida ka geen-geen koosmõjusid, kuid senised meetodid ei võimalda seda suurte andmestike korral piisavalt kiiresti teha.

Käesolevas bakalaureusetöös on Tartu Ülikooli Eesti Geenivaramu andmete põhjal testitud uut meetodit geen-geen koosmõjude tuvastamiseks. Valideeritud meetodit kavatsetakse rakendada uue geen-geen koosmõjude määramise programmi ühe osana. Programm võimaldaks analüüsida senisest suuremaid andmestikke ja teha analüüsi kiiremini kui varem. Andmestik oli autori kasutuses konfidentsiaalsena ning kooskõlas kõigi Tartu Ülikooli Eesti Geenivaramu regulatsioonidega.

Töö esimeses peatükis on antud ülevaade ülegenoomsete assotsiatsiooniuuringute läbiviimisest, seal kasutatavatest statistilistest meetoditest ning bioloogilise ja statistilise koosmõju mõistetest. Töö teises peatükis on kirjeldatud uut meetodit geen-geen koosmõjude määramiseks, probleemi ilmnemisel pakutud välja parandatud meetod ning see valideeritud. Bakalaureusetöös kirjeldatud meetodite implementeerimiseks ning jooniste tegemiseks on kasutatud statistikapaketti R. Kasutatud R-i skriptid on Lisas 1.

Autor tänab käesoleva bakalaureusetöö juhendajaid, Tartu Ülikooli Eesti Geenivaramu vanemteadureid Toomas Hallerit ja Krista Fischerit põneva probleemipüstituse ning rohkete nõuannete eest.

1. Kirjanduse ülevaade

1.1. Geneetika põhimõisted

Desoksüribonukleiinhape ehk DNA on päriliku informatsiooni kandja rakus. DNA on polümeer, mis koosneb desoküriboosist, lämmastikalusest ja forforhappejäägist. Selle elementaarlülideks on nukleotiidid. Esineb nelja erinevat nukleotiidi: adeniin (A), guaniin (G), tsütosiin (C) ja tümiin (T).

Rakkude sees moodustab DNA kromosoomid. Kromosoomis paiknevad geenid ja geenidevahelised alad. Geeniks nimetatakse teatud nukleotiidijärjestust, mis on enamasti aluseks valkude sünteesile.

Ühenukleotiidsed polümorfismid (*single nucleotide polymorphism*, SNP, loeme *snipp*) on DNA järjestuse variatsioonid, mis on tekkinud ühe nukleotiidi asendumisel teisega. SNP-d on inimgenoomi suurim geneetilise varieeruvuse allikas ning geeniuuringutes kasutatakse SNP-e kui genoomipiirkonna markereid. (Bush ja Moore, 2012, lk 1)

Geeni eri vorme nimetatakse alleelideks. SNP-del on enamasti kaks alleeli ehk kaks erinevat võimalust, millised neljast nukleotiidist DNA ahelas sellel kohal paikneda saavad. SNP-de esinemissagedus antakse vähem levinud alleeli ehk minoorse alleeli sageduse kaudu. Näiteks kui populatsioonis on minoorseks alleeliks G, siis SNP-i sagedus 0,4 näitab, et 40% populatsioonis esinevatest alleelidest on G-d. (Bush ja Moore, 2012, lk 1)

Alleelidoosiks nimetatakse minoorse alleeli esinemiste arvu genotüübis. Kui SNP-i alleelid on näiteks A ja G ning minoorne alleel on G, siis on genotüübi AA alleelidoos 0, AG alleelidoos 1 ning GG alleelidoos 2. Statistilistes

analüüsides ei pruugi alleelidoos olla täisarv, kui ta ei ole täpselt teada, vaid talle on antud statistiline hinnang.

Fenotüübiks nimetatakse avaldunud tunnust, mis on kujunenud genotüübi ja keskkonna koosmõju tulemusena. Fenotüübiks võivad olla näiteks füüsilised tunnused (pikkus, silmade värv), biokeemilised näitajad (ensüümide aktiivsus, kusihappe kontsentratsioon veres) või tervisliku seisundi hinnangud (kas inimesel on diabeet või mitte).

Fenotüübid jagunevad üldiselt kaheks: kvalitatiivsed (nt südamehaiguse esinemine) ja pidevad (nt vere kolesteroolitase). Statistilistes analüüsides eelistatakse pidevaid tunnuseid, kuna nende puhul on geneetilise efekti avastamine lihtsam ning tihti on tulemusi võimalik selgemini tõlgendada. (Bush ja Moore, 2012, lk 5)

1.2. Statistilised meetodid ühe SNP-i ja fenotüübi vahelise assotsiatsiooni uurimiseks

Ülegenoomne assotsiatsiooniuuring (*genome wide association study*, GWAS) kujutab endast järjestikuseid teste kontrollimaks fenotüübi ja geneetilise markeri vahelist assotsiatsiooni.

1.2.1. Pidevad tunnused

Pidevate tunnuste analüüsimiseks kasutatakse üldiseid lineaarseid mudeleid. Üheks võimaluseks on läbi viia dispersioonanalüüs, kus faktortunnuseks on genotüüp. Nullhüpoteesiks on see, et erinevate genotüüpide korral on fenotüübiväärtuste keskmised võrdsed. (Bush ja Moore, 2012, lk 6)

Esinegu lookuses kolm erinevat genotüüpi AA, AB ja BB. Olgu fenotüübiväärtuste keskmised vastavates gruppides μ_{AA}, μ_{AB} ja μ_{BB}. Soovime kontrollida järgmist hüpoteeside paari:

$$\begin{cases} H_0 : \mu_{AA} = \mu_{AB} = \mu_{BB}, \\ H_1 : \mu_{AA} \neq \mu_{AB} \text{ või } \mu_{AA} \neq \mu_{BB} \text{ või } \mu_{AB} \neq \mu_{BB}. \end{cases}$$

Erinevate keskväärtustega mõõtmistulemused saame esitada mudeliga

$$Y_{ij} = \mu + \alpha_1 + \varepsilon_{ij},$$

kus Y_{ij} on fenotüübiväärtus j. indiviidil, kellel on i. genotüüp, μ on fiktiivne fenotüübiväärtuste üldkeskmine, α_i on i. genotüübi põhjustatud muutus fenotüübis ja ε_{ij} on juhuslik viga. Mudeli ühese määratuse jaoks eeldame, et $\sum_{i=1}^{3} \alpha_i = 0$. (Parring, Vähi ja Käärik, 1997, lk 259)

Kuna iga genotüübi jaoks võib olla tehtud erinev arv fenotüübiväärtuste mõõtmisi, siis on tegemist tasakaalustamata mudeliga. Olgu genotüüpide tasemetele vastavad valimimahud n_{AA}, n_{AB} ja n_{BB}. Vaatluste koguarv on seega $N = n_{AA} + n_{AB} + n_{BB}$. (Parring, Vähi ja Käärik, 1997, lk 270)

Iga valimi keskväärtuse saame leida valemist

$$\bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

ja üldkogumi keskväärtuse valemist

$$\bar{Y}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{3} \sum_{j=1}^{n_i} Y_{ij}.$$

(Parring, Vähi ja Käärik, 1997, lk 270)

Dispersioonanalüüsi jaoks vajalikud statistikud saame koondada järgnevasse tabelisse.

Tabel 1. Dispersioonanalüüs (Parring, Vähi ja Käärik, 1997, lk 271)

Varieeruvuse allikas	Hälvete ruutude summa	Vabadusastmete arv	Keskruut	F-statistik
Genotüüp	$S_A^2 = \sum_{i=1}^{3} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	2	$MS_A^2 = \frac{S_A^2}{2}$	$F = \frac{MS_A^2}{MS^2}$
Viga	$S^2 = \sum_{i=1}^{3} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2$	$N-3$	$MS^2 = \frac{S^2}{N-3}$	
Üldine	$S_Y^2 = \sum_{i=1}^{3} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$	$N-1$		

Nullhüpoteesi kehtides on F-statistik F-jaotusest, $F \sim F(2, N-3)$. Dispersioonanalüüsi mudeli tegemiseks peavad olema täidetud järgnevad eeldused:

- juhuslikud vead on erinevate vaatluste korral sõltumatud,
- juhuslike vigade keskväärtus on null,
- juhuslike vigade standardhälve on konstantne.

(Parring, Vähi ja Käärik, 1997, lk 266)

Teine võimalus pidevate tunnuste analüüsimiseks on lineaarne regressioon alleelidoosi kaudu. Regressioonisirge saame leida samadel eeldustel, mis peavad kehtima dispersioonanalüüsi korral.

Üldkogumit kirjeldav regressiooniseos on esitatav valemiga

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

kus Y on fenotüübiväärtus, β_0 ja β_1 on regressioonikordajate väärtused üldkogumis, x on alleelidoos ning ε on juhuslik viga. Regressioonikordajate leidmiseks kasutatakse kõige sagedamini vähimruutude meetodit. (Parring, Vähi ja Käärik, 1997, lk 233)

Samaaegselt võib hinnata ka mitme genotüübi mõju fenotüübile. Eeldades fenotüübi lineaarset seost alleelidoosidega, saame mudelit k genotüübitunnuse $X_1, \ldots, X_k$ korral üldistada kujul

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon.$$

(Parring, Vähi ja Käärik, 1997)

1.2.2. Olulisusnivoo korrektsioonid

Kui ühe testi olulisusnivoo on α, siis võime ekslikult nullhüpoteesi kummutada tõenäosusega α. Seljuhul on tõenäosus I liiki viga mitte teha $1 - \alpha$. Kui teeme k sõltumatut testi, siis tõenäosus, et ühegi testi tulemus ei ole valepositiivne, on $(1-\alpha)^k$ ja tõenäosus teha vähemalt üks valepositiivne otsus on $1-(1-\alpha)^k$. Seega, et vea tegemise tõenäosus ei ületaks α, tuleks Bonferroni võrratuse $1-(1-\alpha)^k \leq k\alpha$ kohaselt kasutada iga üksiku testi jaoks olulisusnivood $\frac{\alpha}{k}$. Sellist olulisusnivoo parandust nimetatakse Bonferroni meetodiks. (Abdi, 2007)

Selline korrektsioon on võimalikest kõige konservatiivsem, kuna tehakse eeldus, et kõik assotsiatsioonitestid on üksteisest sõltumatud. Ahelduse mittetasakaalulisuse tõttu see eeldus enamasti ei kehti. (Bush ja Moore, 2012, lk 7)

Ahelduse mittetasakaalulisuseks (*linkage disequilibrium*, LD) nimetatakse SNP-de omadust, mis kirjeldab, kui tugevalt on ühe SNP-i alleel teise SNP-i alleeliga korreleeritud. LD mõiste on seotud geneetilise aheldatusega, kus kaks kromosoomil lähestikku paiknevat markerit päranduvad järglastele koos raga edasi. Igas põlvkonnas pärandub järglastele osa kromosoomist emalt ja osa isalt. Juhusliku ristumisega populatsioonis lahknevad kromosoomid aja

jooksul, kuni lõpuks on kõik alleelid populatsioonis sõltumatud ehk leiab aset ahelduse tasakaalulisus. (Reich jt, 2001)

Teine võimalus olulisusnivood parandada, on kindlaks teha valepositiivsete tulemuste arv (*false discovery rate*, FDR). Hochbergi ja Benjamini (1990) väljatöötatud FDR meetodi korral järjestatakse testide olulisustõenäosused kasvavalt ($p_{(1)} \leq \cdots \leq p_{(k)}$) ning võrreldakse neid vastavalt olulisusnivoodega $\frac{\alpha}{k}, \frac{2\alpha}{k}, \frac{3\alpha}{k}, \ldots, \frac{k\alpha}{k} = \alpha$.

Kolmas võimalus sobivat olulisusnivood leida on kasutada permutatsioonitesti. Igale indiviidile määratakse juhuslikult samast andmestikust võetud teise indiviidi fenotüüp. Sellega kaotatakse seos genotüübi ja fenotüübi vahel. Selliseid andmestikke luuakse N tükki. Iga selliselt saadud andmestik on üks võimalik valim nullhüpoteesi kehtides. Nende andmestike pealt saab leida statistiku empiirilise jaotuse. Näiteks $N = 1000$ korral saame leida empiirilise p-väärtuse $\frac{1}{1000}$ täpsusega. Selline meetod on aga väga arvutusmahukas ja reaalsete andmestike peal seda enamasti ei kasutata. (Bush ja Moore, 2012, lk 7)

Veel üks tihti kasutuselolev meetod on ülegenoomne olulisusnivoo (*genome-wide significance*). Mingi populatsiooni LD jaotuse põhjal saab leida sõltumatute geenoomipiirkondade „efektiivse" arvu ja seega saab ka teada, mitut statistiku väärtust peab tegelikult korrigeerima. Kõige tüüpilisemalt kasutusel olev selliselt hinnatud olulisusnivoo on 5×10^{-8}. (Dudbridge ja Gusnanto, 2008, lk 228-232)

1.2.3. Juht-kontroll-tüüpi tunnused

Kahe väärtusega nn juht-kontroll-tüüpi tunnuste analüüsimiseks kasutatakse sagedustabeleid või logistilist regressiooni. Juht-kontroll-tüüpi tunnuseks on näiteks südamehaiguse esinemine. Kui inimene põeb südamehaigust, siis nimetatakse teda juhuks, ning vastasel korral on tegu kontrolliga. (Bush ja Moore, 2012, lk 6-7)

Esinegu SNP-l kaht erinevat alleeli A ja B. Olgu valimis n_{juht} juhtu ja n_{kontr} kontrolli. Isikud saab genotüüpide järgi esitada järgneva tabelina.

Tabel 2. Juhud ja kontrollid genotüübiti (Lewis, 2002, lk 147)

	AA	AB	BB	Kokku
Juhud	a	b	c	n_{juht}
Kontrollid	d	e	f	n_{kontr}
Kokku	n_{AA}	n_{AB}	n_{BB}	n

Sagedustabeli abil esitatud tunnuse analüüsimiseks kasutatakse enamasti χ^2-statistikut $H = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$, kus $O_i \in \{a, b, c, d, e, f\}$. E_i on mingi genotüübiga juhtude või kontrollide oodatav arv. Näiteks $E_1 = \frac{n_{AA} n_{juht}}{n}$. Suure vaatluste arvu ja nullhüpoteesi õigsuse korral on juhuslik suurus H χ^2-jaotusega vabadusastmete arvuga 2.(Lewis, 2002, lk 147)

Eeldusel, et alleel B suurendab haiguse riski või et haiguse riski suurendamiseks on vaja kahte B-alleeli, saame teha χ^2-testi ka järgnevate 2×2 tabelite peal (vastavalt).

Tabel 3. Dominantne mudel (alleel B suurendab riski) (Lewis, 2002, lk 147)

	AA	AB + BB	Kokku
Juhud	a	b + c	n_{juht}
Kontrollid	d	e + f	n_{kontr}
Kokku	n_{AA}	$n_{AB} + n_{BB}$	n

Tabel 4. Retsessiivne mudel (riski suurendamiseks vaja kahte B-alleeli) (Lewis, 2002, lk 147)

	AA + AB	BB	Kokku
Juhud	a + b	c	n_{juht}
Kontrollid	d + e	f	n_{kontr}
Kokku	$n_{AA} + n_{AB}$	n_{BB}	n

Multiplikatiivse mudeli korral eeldatakse, et iga B alleel r-kordistab haiguse riski: AB genotüübiga isiku haigestumise risk on r korda suurem ja BB genotüübiga isiku risk r^2 korda suurem kui AA genotüübiga inimesel. Sellisel juhul analüüsitakse andmestikku alleelide kaupa. Sagedustabelis (vt. Tabel 5) on antud juhtude ja kontrollide A- ja B-alleelide esinemissagedused. Kui multiplikatiivne mudel on andmetega vastavuses, siis on nii juhtude kui kontrollide genotüübid Hardy-Weinbergi tasakaalus. (Lewis, 2002, 148)

Tabel 5. Multiplikatiivne mudel (Lewis, 2002, lk 147)

	A	B	Kokku
Juhud	2a + b	b + 2c	n_{juht}
Kontrollid	2d + e	e + 2f	n_{kontr}
Kokku	n_A	n_B	n

Hardy-Weinbergi seadus ütleb, et väga suures populatsioonis, kus isenditevaheline ristumine on täiesti juhuslik ning puuduvad mutatsioonid, migratsioon ja looduslik valik, püsivad genotüübi- ja alleelisagedused põlvkonniti muutumatutena. Kahe alleeliga (A ja B) lookuse korral kehtib seos $(p+q)^2 = 1$, kus p on alleeli A sagedus ja $q = 1-p$ on alleeli B sagedus. Genotüüpide sagedused avalduvad sellisel juhul kujul $P(AA) = p^2$, $P(AB) = 2pq$ ja $P(BB) = q^2$. Alleelidoosi jaotuseks on binoomjaotus $\mathrm{Bin}(2, p)$, kus p on minoorse alleeli sagedus. (Guo ja Thompson, 1992)

Neljas võimalik geneetiline mudel on aditiivne mudel, kus eeldatakse, et inimesel genotüübiga AB on r võrra kõrgem ja isikul genotüübiga BB $2r$ võrra kõrgem risk haigestuda kui inimesel genotüübiga AA. Selle mudeli paikapidavust saab testida rakendades Tabel 2 peal Cochrani-Armitage'i testi. (Lewis, 2002; Armitage, Berry ja Matthews, 2008)

Juht-kontroll-tüüpi tunnuseid saab analüüsida ka logistilise regressiooni abil. Logistilise regressiooni jaoks kodeeritakse tunnus Y (haiguse esinemine) väärtustega 0 ja 1, kus 1 tähistab sündmuse toimumist (haiguse olemasolu). Tunnus Y on binoomjaotusest $B(n, \pi)$, kus n on valimi suurus ja π on meid huvitava sündmuse $Y = 1$ toimumise tõenäosus. Logistilise mudeliga hinnatakse šansi logaritmi $\ln \dfrac{\pi}{1-\pi} = \beta_0 + \beta_1(genot = AA) + \beta_2(genot = AB)$.

(Käärik, 2012, lk 110)

Juht-kontroll-tüüpi uuringute probleemiks on see, et tihti ei ole võimalik indiviide kahte kindlasse gruppi jagada. Bioloogilised fenotüübid ei allu sageli sellisele lihtsale jaotusele ja tekib viga.

1.3. Geen-geen koosmõjud

Ülegenoomsed assotsiatsiooniuuringud ei avasta tihti SNP-de seoseid fenotüüpidega, kuna uuritakse vaid üksikute SNP-de mõju ning jäetakse kõrvale erinevate SNP-de koosmõjud. Kui geneetiline tegur mõjutab fenotüüpi keerulise mehhanismi kaudu, mis hõlmab teisi geene ning keskkonnategureid, siis võib selle geeni mõju jääda märkamata, kui koosmõjusid ei vaadelda. Praktikas ei ole teada, kui tihti esineb juhtumeid, kus koosmõju on oluline, kuid peamõjud mitte. (Cordell, 2009, lk 1)

1.3.1. Statistilised koosmõjud

Assotsiatsiooniuuringu eesmärk on uurida tunnuse Y ja geneetiliste ning keskkonnategurite, vastavalt $G = (g_1, g_2, \ldots, g_m)$ ja $E = (z_1, z_2, \ldots, z_k)$, vahelist seost. Pideva tunnuse jaoks saab selle seose kirja panna järgneva mudeli abil:

$$E(Y) = \eta(G; E) = \eta(g_1, g_2, \ldots, g_m; z_1, z_2, \ldots, z_k),$$

kus $E(Y)$ on tunnuse Y oodatav väärtus ja η on tundmatu funktsioon. (Yi, 2010, lk 445)

Paljude geneetiliste ja keskkonnategurite korral on võimalik vaadelda kolme liiki koosmõjusid:

- geen-geen koosmõjud,
- geen-keskkonnategur koosmõjud,
- keskkonnategur-keskkonnategur koosmõjud. (Yi, 2010, lk 445)

Käesolevas töös tegeletakse ainult geenide vaheliste koosmõjudega.

Kui kahe geneetilise teguri g_1 ja g_2 korral saab funktsiooni $\eta(g_1, g_2)$ esitada lihtsamal kujul

$$\eta(g_1, g_2) = \eta_1(g_1) + \eta_2(g_2), \tag{1}$$

siis ei esine g_1 ja g_2 vahel koosmõju ehk lookuse g_1 mõju tunnusele Y ei sõltu lookusest g_2. Kui aga (1) ei kehti, siis esineb nende tegurite vahel koosmõju. Kuna teguritel g_1 ja g_2 on kummalgi kolm võimalikku genotüüpi, siis saame koosmõjuga mudeli esitada kujul

$$\eta(g_{1i}, g_{2j}) = \mu + g_{1i} + g_{2j} + \delta_{ij}, \tag{2}$$

kus $i, j = 1, 2, 3$, g_{1i} on faktori g_1 peamõju tasemel i, g_{2j} on faktori g_2 peamõju tasemel j ja δ_{ij} on faktorite g_1 ja g_2 koosmõju vastavalt tasemetel i ja j. Selles mudelis on teguri g_1 mõju fenotüübile Y kujul $\mu + g_{1i} + \delta_{ij}$, mis sõltub g_2-st. (Yi, 2010, lk 445)

Ilma kitsendusteta ei ole võrrandite süsteem (2) üheselt lahenduv. Tavaliselt jäetakse ühese lahenduvuse huvides mudelist välja g_1 ja g_2 esimene tase, mida nimetatakse baastasemeks. Selle kitsendusega vähendatakse iga faktori peamõjude arv kaheni ja kahe faktori vaheliste koosmõjude arv neljani. Mudeli (2) saab nüüd reparametriseerida kujule

$$\begin{aligned}\eta(g_1, g_2) = \mu + (x_{a1}a_1 + x_{d1}d_1) + (x_{a2}a_2 + x_{d2}d_2) + (x_{a1}x_{a2}aa_{12} \\ + x_{a1}x_{d2}ad_{12} + x_{d1}x_{a2}da_{12} + x_{d1}x_{d2}dd_{12}),\end{aligned} \tag{3}$$

kus $x_{ak} = 1$, kui $g_k = 2$, $x_{ak} = 0$ mujal, ja $x_{dk} = 1$, kui $g_k = 3$, $x_{dk} = 0$ mujal; a_k ja d_k on peamõjud ning aa_{12}, ad_{12}, da_{12} ja dd_{12} on koosmõjud. (Yi, 2010, lk 445)

Käesolevas töös kasutatakse lineaarsete koosmõjudega mudelit

$$Y = \mu + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 X_2 + \varepsilon, \tag{4}$$

mis on mudeli (2) erijuht, kus $g_i = (i-1)\beta_1$, $g_j = (j-1)\beta_2$ ja $\delta_{ij} = (i-1)(j-1)\gamma$. Järgnev tabel iseloomustab fenotüübitunnuse keskväärtusi, kui kehtib mudel (4).

Tabel 6. Fenotüübi keskväärtus kahe SNP-i korral, kui kehtib mudel (4)

		X_1		
		0	1	2
X_2	0	μ	$\mu + \beta_1$	$\mu + 2\beta_2$
	1	$\mu + \beta_2$	$\mu + \beta_1 + \beta_2 + \gamma$	$\mu + 2\beta_1 + \beta_2 + 2\gamma$
	2	$\mu + 2\beta_2$	$\mu + \beta_1 + 2\beta_2 + 2\gamma$	$\mu + 2\beta_1 + 2\beta_2 + 4\gamma$

1.3.2. Bioloogilised koosmõjud

Statistiliste koosmõjude uurimine on vajalik selleks, et leida olukorrad, mida asuda bioloogiliselt süvitsi uurima. Geen-geen koosmõju ehk epistaasi mõiste oli algselt kasutusel kirjeldamaks olukorda, kus ühel geenil on pärssiv mõju teise lookuse geeni avaldumisele. Selline definitsioon erineb statistilise koosmõju mõistest. (Yi, 2010, lk 454)

Ei ole teada, mil määral statistilised koosmõjud näitavad bioloogilisi ja vastupidi. VanderWeele (2010) on näidanud, et küllaltki tugevatel eeldus-

tel vastavad statistilised koosmõjud bioloogilistele, kuid pakutud meetodid võivad viia kasulike strateegiateni, kuidas bioloogilisi koosmõjusid uurida.

1.3.3. Koosmõjutestide kiirendamine

Kui uuringus vaadeldakse 500 000 SNP-i, siis nendevahelisi koosmõjusid on vaatluse all peaaegu 125 miljardit. Kõigi selliste paariviisiliste võrdluste tegemine on ka väga võimsatel arvutitel võimatu. Seega tuleks enne analüüsi tegemist SNP-e filtreerida, et võrdluste arvu vähendada. (Bush ja Moore, 2012, lk 7-8)

Üks võimalus on valida välja vaid sellised SNP-d, mille peamõjud on dispersioonanalüüsis olulised. Niimoodi filtreerides saadakse valitud SNP-de hulgas usaldusväärsed tulemused ning ka arvutusmahukuselt on töö palju väiksem. Samas aga ei võimalda selline lähenemine avastada mudeleid, kus peamõjud on statistiliselt ebaolulised, kuid kahe SNP-i koosmõju on oluline. (Bush ja Moore, 2012, lk 7-8)

Teine võimalus on vaadelda vaid selliseid SNP-de kombinatsioone, millel on kindel bioloogiline sisu, näiteks teatud bioloogiline rada või valkude perekond. (Bush ja Moore, 2012, lk 8)

Kolmas moodus koosmõjutestide kiirendamiseks on meetodite efektiivne implementeerimine. Selleks võib kasutada erinevaid efektiivseid algoritme ning lähendamist.

Lisaks kahe SNP-i vahelisele koosmõjule saab vaadelda ka kolme, nelja ja rohkema SNP-i vahelisi koosmõjusid, kuid nende hindamine nõuab väga suuri andmestikke ning saadud tulemuste interpreteerimine on keeruline. (Cordell, 2009, lk 3)

1.4. Olemasolevad programmid geen-geen koosmõjude määramiseks

1.4.1. PLINK

PLINK on avatud lähtekoodiga programm genotüübi info haldamiseks ja uurimiseks, sealhulgas ka ülegenoomsete assotsiatsiooniuuringute läbiviimiseks. Haiguse esinemist näitavate tunnuste korral on PLINK-i abil võimalik testida kahe SNP-i vahelist koosmõju. Testida saab kõiki paariviisilisi SNP-de koosmõjusid. Alternatiivina võib tehtavate testide hulka vähendada näiteks valides välja ainult need SNP-d, mille peamõjud on statistiliselt kõige olulisemad. (Purcell, 2009)

Vaikimisi kasutatakse testimiseks lineaarset või logistilist regressiooni olenevalt sellest, kas fenotüüp on pidev või binaarne (Purcell, 2009).

PLINK-is on võimalik kasutada valikut `fast-epistasis`. Nii juhtude kui kontrollide jaoks tehakse genotüüpide sagedustabel (vt Tabel 7). Sellistest 3×3-tabelitest saadakse alleelide 2×2-tabelid (vt Tabel 8). (Purcell, 2009)

Tabel 7. Genotüüpide sagedustabel

	BB	Bb	bb
AA	a	b	c
Aa	d	e	f
aa	g	h	i

Tabel 8. Alleelide sagedustabel

	B	b
A	$4a + 2b + 2d + e$	$4c + 2b + 2f + e$
a	$4g + 2h + 2d + e$	$4i + 2h + 2f + e$

Selle 2×2-sagedustabeli põhjal leitakse statistik

$$Z = \frac{\log(R) - log(S)}{\sqrt{s(R) + s(S)}},$$

kus R ja S on vastavalt juhtude ja kontrollide jaoks leitud šansisuhted

$$\frac{(4a + 2b + 2d + e)(4c + 2b + 2f + e)}{(4g + 2h + 2d + e)(4i + 2h + 2f + e)}$$

ning $s(R)$ ja $s(S)$ on juhtude ja kontrollide jaoks leitud šansisuhete standardhälbed

$$\frac{1}{4a + 2b + 2d + e} + \frac{1}{4c + 2b + 2f + e} + \frac{1}{4g + 2h + 2d + e} + \frac{1}{4i + 2h + 2f + e}.$$

Z-statistik on nullhüpoteesi kehtides standardnormaaljaotusest. (Purcell, 2009)

1.4.2. INTRAPID

INTRAPID kasutab koosmõjude uurimiseks kaheetapilist meetodit, kus alguses rakendatakse kiiret interaktsioonimeetodit. Binaarse tunnuse korral on see samaväärne PLINK-i `fast-epistasis` valikuga. Pidevad tunnused jagatakse mediaani juurest pseudo-juhtudeks ja -kontrollideks. Simuleeritud andmete peal on näidatud, et see ei vähenda oluliselt testi võimsust. (Bhattacharya, Mägi ja Morris, 2014)

2. Meetodi valideerimine

2.1. Andmestiku kirjeldus

Andmestikus, millel meetodit lõpuks rakendati, oli 2505 SNP-i inimese 4. kromosoomist. Nende SNP-de alleelidoosid olid leitud 867 indiviidi jaoks. Fenotüübiks oli kusihappe kontsentratsioon veres. Andmestik oli töö autori kasutuses konfidentsiaalsena ning kooskõlas kõigi Tartu Ülikooli Eesti Geenivaramu regulatsioonidega. Kasutati Oxfordi GEN ja SAMPLE formaadis faile (Marchini, 2011).

GEN-failis on ühe SNP-i kohta üks rida, kus esimeses viies tulbas on vastavalt kromosoomi number, rs-number, SNP-i asukoht kromosoomil, esimene alleel (A, G, C või T) ning teine alleel (A, G, C või T). Järgmises kolmes tulbas on esimese indiviidi genotüüpide tõenäosused, näiteks kui tema genotüüp on AG, siis on tulpades vastavalt arvud `0 1 0`, sest genotüübi AA tõenäosus on 0, AG tõenäosus 1 ja GG tõenäosus 0. Siin ei pruugi olla täisarvud, kui genotüüp ei ole teada ning on seetõttu imputeeritud, näiteks `0.1 0.9 0`. Neile järgnevates tulpades on teise indiviidi genotüüpide tõenäosused jne. Failitüübi näidis on Tabelis 9. GEN-faili abil saab leida alleelidoosid.

Tabel 9. GEN formaadis faili näidis

4	rs11724390	831082	A	G	0	0	1	1	0	0	0	1	0
4	rs1134921	843508	C	T	0	0	1	0	0	1	0	0	1
4	rs11248052	859634	C	G	1	0	0	0	1	0	0	1	0

SAMPLE-failis on igal real ühe indiviidi kohta käivad andmed. Failil on

kaks päiserida, millest esimeses on kirjas tulpade nimed ning teises, millist tüüpi andmed selles tulbas paiknevad. Esimeses kolmes tulbas on identifikaatorid ning puuduvate andmete osakaal. Neile tulpadele järgnevad kovariaadid ja seejärel pidevad ning diskreetsed tunnused. Käesolevas andmestikus on neli kovariaati ning vanus, sugu, vere kusihappesisaldus (logaritmitud) ja GWAS-ist saadud regressioonimudeli jäägid. Failitüübi näidis on Tabelis 10.

Tabel 10. SAMPLE formaadis faili näidis

ID _1	ID _2	missing	C1	C2	C3	C4	vanus	sugu	ln _UA	QnormRes
0	0	0	C	C	C	C	D	D	P	P
V32239	V32239	0.005	-0.005	0.002	-0.004	0.009	21	2	5.488	0.139
V32238	V32238	0.013	0.003	0.011	-0.017	0.008	20	2	5.645	0.931
V32237	V32237	0.009	0.002	0.005	0.005	0.009	37	2	5.247	-1.269

2.2. Algne meetod

2.2.1. Meetodi kirjeldus

Olgu meil fenotüüp Y ning n sõltumatut markerit ja neile vastavad alleelidoosid $X_1, X_2, \ldots, X_n$. Olgu alleelidoosid X_j, $j = 1, \ldots, n$, skaleeritud nii, et $E(X_j) = 0$. Kui X_i ja X_j $(i \neq j)$ on sõltumatud, siis

$$E(X_i X_j) = E(X_i)E(X_j) = 0,$$

$$E(X_i|X_j) = E(X_i) = 0.$$

Kui kehtib mudel

$$Y = \mu + \beta_i X_i + \beta_j X_j + \gamma_{ij} X_i X_j + \varepsilon, \tag{5}$$

kus $E(\varepsilon|X_i, X_j) = 0$, siis

$$\begin{aligned} E(Y|X_j) &= E(\mu + \beta_i X_i + \beta_j X_j + \gamma_{ij} X_i X_j + \varepsilon | X_j) \\ &= E(\mu|X_j) + E(\beta_i X_i|X_j) + E(\beta_j X_j|X_j) \\ &\quad + E(\gamma_{ij} X_i X_j|X_j) + E(\varepsilon|X_j) \\ &= \mu + \beta_i E(X_i|X_j) + \beta_j X_j + \gamma_{ij} X_j E(X_i|X_j) \\ &= \mu + \beta_j X_j. \end{aligned}$$

Analoogiliselt saame

$$E(Y|X_i) = \mu + \beta_i X_i$$

ja kokkuvõttes $E(Y) = \mu$.

Seega on β_i ja β_j hinnatavad eraldi regressioonimudelitest igale markerile. Kui $\beta_i,\ i = 1, \ldots, n$, on regressiooniparameetrid mudelist

$$Y = \mu + \beta_i X_i + \varepsilon,$$

siis $Y_n = Y - \sum_{i=1}^{n} \beta_i X_i$ (skaleerituna nii, et $E(Y_n) = 0$) on marginaalselt sõltumatu kõigist markeritest $X_i,\ i = 1, \ldots, n$. Järelikult

$$E(Y_n X_i) = E(Y_n)E(X_i) = 0.$$

Mudeli (5) kehtides tuleb hüpoteesi $\gamma_{ij} = 0$ kontrollimiseks testida, kas $E(Y_n X_i X_j) = 0$.

Selleks tuleb läbi viia järgnevad sammud:

1. Määrata GWAS-i tulemuste põhjal n markerit, mille vahel koosmõjusid otsima hakata.

2. Skaleerida markerite alleelidoosid lahutades iga markeri alleelidoosist keskmise selle markeri alleelidoosi. Olgu skaleeritud alleelidoosid X_1, $\ldots, X_n$.

3. Hinnata iga $i = 1, \ldots, n$ korral mudel

$$Y = \mu + \beta_i X_i + \varepsilon.$$

   Olgu $\hat{\beta}_i$ saadud hinnang parameetrile β_i.

4. Leida geneetiline riskiskoor $\hat{S}_n = \sum_{i=1}^{n} \hat{\beta}_i X_i$.

5. Leida jääk $Y_n = Y - \hat{S}_n$ ja see skaleerida.

6. Viia läbi $\frac{n(n-1)}{2}$ testi kontrollimaks hüpoteesi $E(Y_n X_i X_j) = 0$ iga $i < j$ korral, kus $i = 1, \ldots, (n-1)$, $j = 2, \ldots, n$.

   Teststatistikuks on

$$\sqrt{N} \frac{\overline{Y_n X_i X_j}}{s(Y_n X_i X_j)},$$

   kus N on indiviidide arv valimis, $\bar{x}$ tähistab vektori x valimikeskmist ja $s(x)$ tema standardhälvet. Siin on tegu ühe valimi t-testiga.

2.2.2. Simuleeritud andmed

Esmalt testiti meetodit simuleeritud andmete peal. Selleks genereeriti 1000 indiviidile kaks erinevat SNP-i alleelidoosidega X_1 ja X_2 vastavalt binoomjaotustest $B(2, p_1)$ ja $B(2, p_2)$ ning skaleeriti nad nii, et keskväärtused oleksid nullid. Seejärel genereeriti fenotüübiväärtused Y mudeliga

$$Y = \mu + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 X_2 + \varepsilon,$$

kus $\varepsilon \sim N(0, 1)$.

Toodud näidetes on võetud $p_1 = 0{,}2$, $p_2 = 0{,}3$, $\mu = 2$, $\beta_1 = 0{,}11$, $\beta_2 = 0{,}12$ ja γ vastavalt 0 ja 0,15.

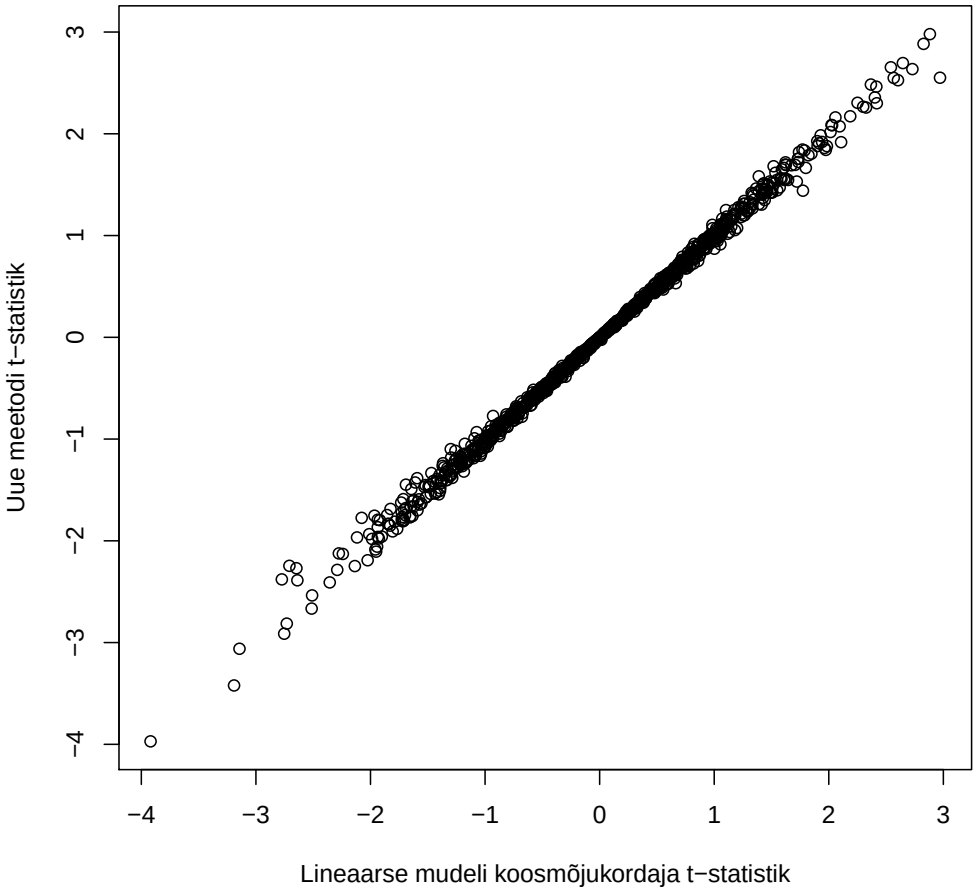
Seejärel hinnati lineaarsest regressioonimudelist parameetrid β_1 ja β_2, leiti geneetiline riskiskoor ja lõpuks skaleeritud jäägid, nagu punktis 2.2.1 kirjeldatud algoritm ette näeb.

Saadud andmete peal viidi läbi ühe valimi t-test. Võrdluseks leiti ka lineaarsest regressioonimudelist

$$Y = \mu + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 X_2 + \varepsilon,$$

hinnatud koosmõjukordaja t-statistik. Neid samme korrati 1000 korda.

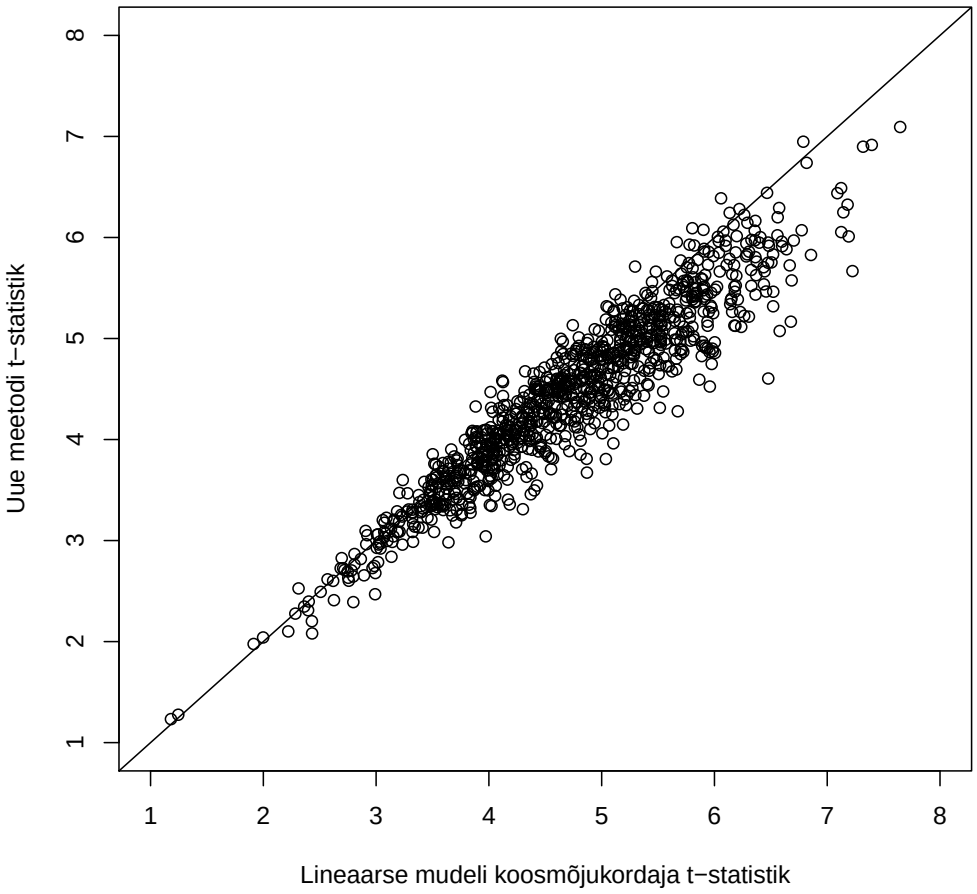
Kui fenotüübiväärtused genereeriti nii, et koosmõjukordaja oli 0, oli uue meetodi t-statistiku olulisustõenäosus olulisusnivoost 0,05 suurem 955 korral 1000-st. Ülejäänud 45 korda, mil nullhüpoteesi kummutataks ja loetaks koosmõjukordaja nullist oluliselt erinevaks, jäävad lubatud 5%-se vea piiresse. Joonisel 1 on näha t-statistikute võrdlus selliselt genereeritud valimite korral. Korrelatsioon traditsioonilise meetodiga leitud t-statistiku ning uue meetodiga leitud t-statistiku vahel oli 0,9983 usaldusvahemikuga (0,9981...0,9985).



Joonis 1. Uue meetodi t-statistiku ja lineaarse mudeli koosmõjukordaja t-statistiku võrdlus, kui koosmõju puudub

Kui andmestik genereeriti selliselt, et koosmõjukordajaks oli 0,15, saadi nullhüpotees uuel meetodil leitud t-statistiku põhjal olulisusnivool 0,05 kummutada 998 korral 1000-st. Korrelatsioon t-statistikute vahel oli 0,9435 usaldusvahemikuga (0,9362...0,9499). Jooniselt 2 on näha, et kui koosmõju-

kordaja erineb oluliselt nullist, siis uue meetodi t-statistik on pigem väiksem kui lineaarse mudeli koosmõjukordaja t-statistik ehk uus meetod on konservatiivsem.



Joonis 2. Uue meetodi t-statistiku ja lineaarse mudeli koosmõjukordaja t-statistiku võrdlus, kui koosmõjukordaja on 0,15

2.2.3. Empiirilised andmed

Pärast simulatsiooni läbiviimist rakendati meetodit reaalsetel andmetel. Taas leiti punktis 2.2.1 kirjeldatud algoritmi kohaselt teststatistikud ning võrdluseks lineaarse regressioonimudeli koosmõjukordaja t-statistikud.

Algses andmestikus oli 867 indiviidi ning igaühel neist olid antud 2505 SNP-i alleelidoosid. Sellest andmestikust eraldati 380 SNP-i, mille omavahelised korrelatsioonid on paarikaupa väiksemad kui 0,1. Selleks kirjutati R-i skript (vt Lisa 1), mis kontrollis kõiki SNP-e paarikaupa ja kui paari korrelatsioon ületas 0,1, siis eemaldas andmestikust selle paari esimese SNP-i. Uuest andmestikust võeti juhuslikud valimid, kuhu kuulus 10, 20, 50, 100, 200 ja 300 SNP-i, ning testiti nende SNP-de koosmõjusid fenotüübiväärtusele.

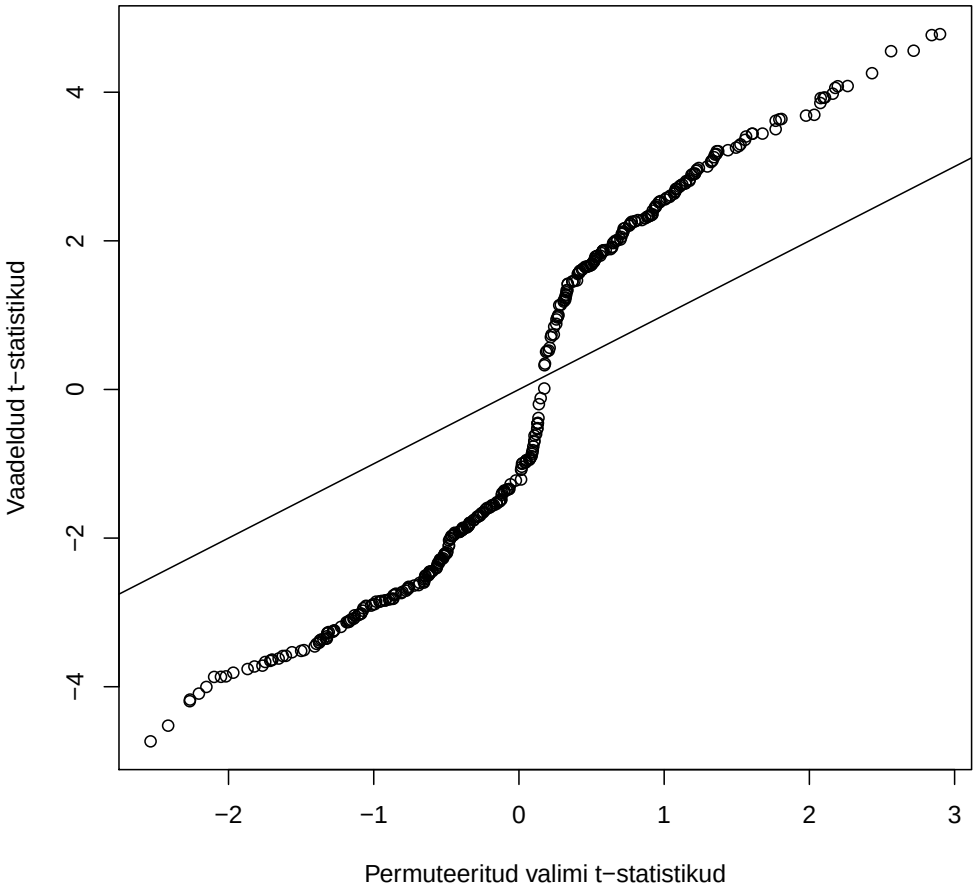
Ideaalsel juhul oleks uue ja klassikalise meetodi t-statistikute korrelatsioon 1. Nagu võib näha Tabelist 11, siis mida enam SNP-e vaatluse alla võetakse, seda nõrgemaks muutub korrelatsioon t-statistikute vahel. Statistikuid võrdlevad joonised on Lisas 2.

Tabel 11. Korrelatsioonid t-statistikute vahel

SNP-de arv n	Paaride arv $\frac{n(n-1)}{2}$	Korrelatsioon	95% usaldusvahemik
10	45	0,989	(0,981 . . . 0,994)
20	190	0,968	(0,957 . . . 0,976)
50	1225	0,884	(0,871 . . . 0,896)
100	4950	0,727	(0,713 . . . 0,740)
200	19900	0,315	(0,302 . . . 0,327)
300	44850	-0,082	(-0,091 . . . -0,073)

Kui vaadelda 10 SNP-i, siis tuleb kõigi paariviisiliste koosmõjude testimiseks läbi viia 45 t-testi. Statistikutevaheline korrelatsioon oli antud juhul 0,9893. Kui aga vaadelda 300 SNP-i, siis tuleb läbi viia 44850 testi. Sel juhul on näha, et korrelatsioon on muutunud väga nõrgaks ning isegi negatiivseks.

Probleem tekib jäägi arvutamisel valemist $Y_n = Y - \sum_{i=1}^{n} \hat{\beta}_i X_i$. Eeldasime, et markerite paarid X_i ja X_j on teineteisest lineaarselt sõltumatud, kuid tegelikkuses nad seda ei ole. Vähese arvu nõrgalt korreleeritud markerite korral probleem ei avaldu, kuna iga marker jääb nõrgalt korreleerituks ülejäänute lineaarkombinatsiooniga. Suure arvu markerite korral tekivad aga tugevamad korrelatsioonid ning seega ei ole jääk üksikutest markeritest sõltumatu, mida näitab ka Joonis 3. Sõltumatute markerite korral peaksid linaarsest regressioonist saadud t-statistikud olema t-jaotusest, mille korral umbes 95% väärtustest peaks jääma -2 ja 2 vahele. Siin on aga näha, et paljud statistikute väärtused jäävad sellest vahemikust välja.



Joonis 3. Vaadeldud ja permuteeritud t-statistikute võrdlus

2.3. Parandatud meetod

2.3.1. Meetodi kirjeldus

Eelpool kirjeldatud meetodi korral tekkis probleem, et leitud jääk ei olnud üksikutest markeritest sõltumatu. Probleemi lahendamiseks leitakse nüüd

jäägid nii, et eemaldatakse fenotüübiväärtusest ainult nende kahe markeri mõjud, millevahelist koosmõju parasjagu uuritakse:

$$Y_{(ij)} = Y - \hat{\beta}_i X_i - \hat{\beta}_j X_j.$$

Seejärel testitakse t-testi abil, kas $E(Y_{(ij)} X_i X_j) = 0$.

Seega tuleb nüüd läbi viia järgnevad sammud:

1. Määrata GWAS-i tulemuste põhjal n markerit, mille vahel koosmõjusid otsima hakata.

2. Skaleerida markerite alleelidoosid lahutades iga markeri alleelidoosist keskmise selle markeri alleelidoosi. Olgu skaleeritud alleelidoosid X_1, $\ldots, X_n$.

3. Hinnata iga $i = 1, \ldots, n$ korral mudel

   $$Y = \mu + \beta_i X_i + \varepsilon.$$

   Olgu $\hat{\beta}_i$ saadud hinnang parameetrile β_i.

4. Valida kaks markerit X_i ja X_j ning leida jääk $Y_{(ij)} = Y - \hat{\beta}_i X_i - \hat{\beta}_j X_j$. Jääk skaleerida.

5. Viia läbi t-test kontrollimaks hüpoteesi $E(Y_{(ij)} X_i X_j) = 0$. Teststatistikuks on

   $$\sqrt{N} \frac{\overline{Y_{(ij)} X_i X_j}}{s(Y_{(ij)} X_i X_j)}.$$

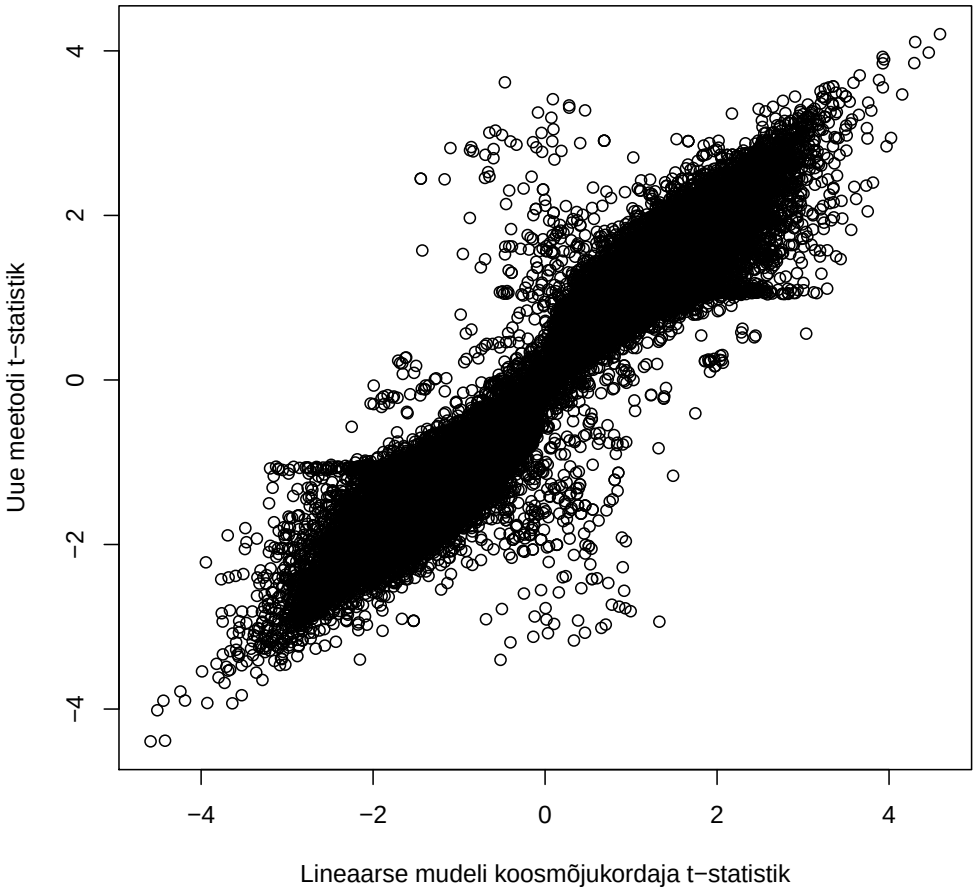
6. Korrata $\frac{n(n-1)}{2}$ korda samme 4 ja 5 , et vaadata läbi kõik võimalikud markerite paarid X_i ja X_j, kus $i < j$.

2.3.2. Empiirilised andmed

Parandatud meetodit rakendati taas 867 indiviidi ja 2505 markeriga andmestikul. Andmestikust eemaldati markerid, mis ei olnud fenotüübiga oluliselt seotud ehk mille korral mudelist $Y = \mu + \beta_i X_i + \varepsilon$ hinnatud regressioonikordaja olulisustõenäosus $p > 0{,}05$. Seejärel eemaldati markerid, mille korrelatsioon talle eelneva või järgneva markeriga oli üle 0,5. Alles jäi 542 markerit.

Saadud andmestikule rakendati punktis 2.3.1 kirjeldatud samme. Saadud t-statistikut võrreldi lineaarse regressiooni abil hinnatud koosmõjukordaja t-statistikuga. Statistikutevaheline korrelatsioon oli 0,975. Seega annab kirjeldatud meetod väga lähedasi tulemusi tavaliselt kasutusel oleva lineaarse regressiooniga.

Joonisel 4 on näha t-statistikute võrdlus. Kui lineaarse regressiooni koosmõjukordaja t-statistik on nulli lähedal, siis näeme, et uue meetodi t-statistik on mõndadel juhtudel nullist väga erinev. Seda võib põhjustada SNP-de omavaheline sõltuvus, kuna valisime välja SNP-d, mille korrelatsioon ulatus 0,5-ni. Vähendades lubatud korrelatsiooni, on uue meetodi t-statistik ka nulli ümbruses lineaarse regressiooni koosmõjukordaja t-statistikule lähemal.



Joonis 4. Parandatud meetodi t-statistiku ja lineaarse mudeli koosmõjukordaja t-statistiku võrdlus

Kokkuvõte

Käesoleva bakalaureusetöö eesmärgiks oli tutvustada olemasolevaid võimalusi ning valideerida uus meetod geen-geen koosmõjude uurimiseks ülegenoomsetes assotsiatsiooniuuringutes.

Töös kirjeldati meetodit, kus fenotüübiväärtusest lahutatakse maha kõigi uuritavate markerite mõjud ning kontrollitakse ühe valimi t-testi abil, kas niimoodi saadud jääk on sõltumatu kahe markeri koosmõjust. Esmalt rakendati meetodit simuleeritud andmetel kahe markeri korral. Selliselt leitud t-statistiku korrelatsioon traditsiooniliselt koosmõjude hindamiseks kasutusel oleva lineaarse regressiooni koosmõjukordaja t-statistikuga oli väga tugev. Seejärel rakendati meetodit ka reaalsetel andmetel, mis pärinevad Tartu Ülikooli Eesti Geenivaramust. Mida enam SNP-e vaatluse alla võeti, seda nõrgemaks muutus t-statistikute vaheline korrelatsioon. Kuna meetodis eeldati markerite sõltumatust, kuid tegelikult nad seda ei ole, siis suure arvu markerite korral ei ole leitud ka jääk üksikutest markeritest sõltumatu ning kirjeldatud meetodit rakendada ei saa.

Probleemi kõrvaldamiseks muudeti meetodit nii, et jäägi arvutamisel lahutatakse fenotüübiväärtusest maha vaid parasjagu vaatluse all oleva kahe markeri mõjud. Nii on saadud jääk kummastki markerist sõltumatu isegi siis, kui markerite vahel esineb nõrk korrelatsioon. Parandatud meetodit rakendati 542 SNP-l ja t-statistiku korrelatsioon lineaarse regressiooni koosmõjukordaja t-statistikuga oli 0,975. Seega võib väljapakutud meetodit kasutada traditsioonilise lineaarse regressioonimeetodi asemel.

Töö edasiarendamiseks saab muu hulgas hinnata uue meetodi kiirust võr-

reldes lineaarse regressiooniga ning enne meetodi rakendamist vähendada uuritavate SNP-de arvu. Valideeritud meetodit kavatsetakse rakendada uue geen-geen koosmõjude määramise programmi ühe osana.

Viited

Abdi, H., 2007. The Bonferroni and Šidák Corrections for Multiple Comparisons. *Encyclopedia of Measurement and Statistics.*

Armitage, P., Berry, G. ja Matthews, J. N. S., 2008. *Statistical Methods in Medical Research.* John Wiley and Sons.

Bhattacharya, K., Mägi, R. ja Morris, A., 2014. [URL] `http://www.well.ox.ac.uk/INTRAPID/index.shtml` [Vaadatud 04.05.2014]

Bush, W. S. ja Moore, J. H., 2012. Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology* 8(12).

Cordell, H. J., 2009. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 10(6).

Dudbridge, H. J. ja Gusnanto, A., 2008. Estimation of Significane Thresholds for Genomewide Association Scans. *Genetic Epidemiology* 32(3).

Guo, S. W. ja Thompson, E. A., 1992. Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles. *Biometrics* 48(2).

Hochberg, Y. ja Benjamini, Y., 1990. More Powerful Procedures for Multiple Significance Testing. *Statist Med* 9(7).

Käärik, E., 2013. *Andmeanalüüs II. Loengukonspekt.*

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

Marchini, J., 2011. [URL]
http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html
[Vaadatud 04.05.2014]

Parring, A.-M., Vähi, M. ja Käärik E., 1997. *Statistilise andmetöötluse alg-õpetus.* Tartu Ülikooli Kirjastus.

Purcell, S., 2009. [URL]
http://pngu.mgh.harvard.edu/~purcell/plink
[Vaadatud 04.05.2014]

Reich, D. E jt, 2001. Linkage Disequilibrium in the Human Genome. *Nature* 411(6834).

VanderWeele, T. J, 2010. Epistatic Interactions. *Statistical Applications in Genetics and Molecular Biology* 9(1).

Yi, N., 2010. Statistical Anlaysis of Genetic Interactions. *Genetics Research* 92(5-6).

Lisad

Lisa 1. Programmikoodid

Simulatsioon algsele meetodile

```
N <- 1000
nsim <- 1000
d <- 0.15
out <- NULL
for(i in 1:nsim) {
 x1 <- rbinom(N, 2, 0.2)
 x2 <- rbinom(N, 2, 0.3)
 X <- cbind(x1,x2)
 Xs <- scale(X) #skaleerimine
 Y <- 2 + 0.11*Xs[,1] + 0.12*Xs[,2] + d*Xs[,1]*Xs[,2] +
     rnorm(N,0,1)
 k <- ncol(Xs)
 b <- rep(NA, ncol(Xs))
 for (j in 1:ncol(Xs)) {
  lm <- lm(Y~Xs[,j])
  b[j] <- lm$coef[2]
 }
 S <- c(Xs%*%b) #geneetiline riskiskoor
 Y_k <- scale(Y-S) #j"a"ak
 nint <- k*(k-1)/2
 n<-length(Y_k)
```

```
 for (j in 1:(k-1)) for (l in (j+1):k) {
  stat <- sqrt(n)*mean(Y_k*Xs[,j]*Xs[,l])/apply(Y_k*Xs[,j]*
      Xs[,l], 2, sd) #teststatistik
  pv <- 2*pt(abs(stat), n-1, lower.tail=F) #p-v"a"artus
  m <- summary(lm(Y~Xs[,j]*Xs[,l]))$coef #lin mudel
      koosmojuga
  beta <- m[4,1]
  stat2 <- m[4,3] #koosmojukordaja
  pv2 <- m[4,4] #koosmoju olulisus
  out <- rbind(out, c(j, l, cor(Xs[,j], Xs[,l]), b[j], b[l
      ], beta, stat, pv, stat2, pv2))
 }
 colnames(out) <- c("i","j","corr","bi","bj", "beta", "T","
     pv","Tc","pv2")
 out <- as.data.frame(out)
}
pdf("fail1.pdf")
plot(out$Tc, out$T, xlab="Lineaarse mudeli koosmojukordaja
    t-statistik", ylab="Uue meetodi t-statistik", xlim=c
    (1,8), ylim=c(1,8)) #t-statistikute vordlus
abline(0,1)
dev.off()
table(out$pv<0.01,out$pv2<0.01)
table(out$pv<0.05,out$pv2<0.05)
cor.test(out$Tc, out$T)
save(out, file="fail2.RData")
```

Korreleeritud SNP-de eemaldamine

```
load(file="fail1.RData")
X <- lg$out
i <- 1
n <- ncol(X)
while ( i < n) {
 a <- NULL
 for ( j in (i+1):n ) {
  if ( abs(cor(X[,i], X[,j]))>0.1 ) {
   a <- c(a,j)
  }
 }
 X <- X[,setdiff(1:n, a)]
 n <- ncol(X)
 i <- i+1
}
save(X, file="fail2.RData")
```

Valimite võtmine ja algse meetodi rakendamine

```
load(file="fail1.RData")
X <- lg$out
Y <- as.vector( read.table("fail2.txt", header=TRUE)[,1] )
samp <- sample(X, 10)
Xs <- scale(samp) #skaleerimine
#regressioonikordajad
```

```
b <- rep(NA, ncol(Xs))
for (j in 1:ncol(Xs)) {
    lm <- lm(Y~Xs[,j])
    b[j] <- lm$coef[2]
}
k <- ncol(Xs)
S <- c(Xs %*% b) #geneetiline riskiskoor
Y_k <- scale(Y-S) #j"a"ak
nint <- k*(k-1)/2
out <- NULL
n <- length(Y_k)
for (i in 1:(k-1)) for (j in (i+1):k) {
 stat <- sqrt(n)*mean(Y_k*Xs[,i]*Xs[,j])/apply(Y_k*Xs[,i]*
     Xs[,j],2,sd) # teststatistik
 pv <- 2*pt(abs(stat),n-1,lower.tail=F) # p-v"a"artus
 m <- summary(lm(Y~Xs[,i]*Xs[,j]))$coef # lineaarne mudel
     koosmojuga
 if (dim(m)[1]==4){
  beta <- m[4,1] #koosmojukordaja
  stat2 <- m[4,3] # t-statistik
  pv2 <- m[4,4] # koosmoju olulisus
  ut <- rbind(out,c(i,j,cor(Xs[,i],Xs[,j]),b[i],b[j],beta,
      stat,pv,stat2,pv2))
 }
}
colnames(out) <- c("i","j","corr","bi","bj","beta","T","pv"
    ,"Tc","pv2")
```

```
out <- as.data.frame(out)
cor.test(out$T,out$Tc)
pdf("fail3.pdf")
plot(out$Tc,out$T, xlab="Lineaarse mudeli koosmojukordaja t
   -statistik", ylab="Uue meetodi t-statistik") # "oige" ja
    meie T-statistiku vordlus
dev.off()
save(out, file="fail4.RData")
```

Parandatud meetodi rakendamine

```
load(file="fail1.RData")
X <- lg$out
Y <- as.vector( read.table("fail2.txt", header=TRUE)[,1] )
Y <- scale(Y)  # skaleerime Y
Xs <- scale(X)  # skaleerime X
### 1) j"atame X-maatriksisse ainult Y-ga oluliselt seotud
   (p<0.05) markerid
### 2) eemaldame koik markerid, mis on eelnevaga voi j"
   argnevaga korrelatsioonis "ule 0.5
### X korrastamise algus
cm <- bb <- bs <- rep(NA,ncol(Xs))
for (i in 1:ncol(Xs)) {
 mud <- summary(lm(Y~Xs[,i]))$coef
 bb[i] <- mud[2,1]
 bs[i] <- mud[2,2]
```

```
  if (i>1) cm[i]<-cor(Xs[,i],Xs[,i-1])
}
cm[1] <- 0
plot(cm)
Xs <- Xs[,abs(bb/bs)>2]  # esimesel korral k"aivita ainult
    see rida
Xs <- Xs[,abs(cm)<0.5]   # teisel ja j"argmistel kordadel k
    "aivita see rida
dim(Xs)
length(bb)
### X korrastamise lopp
k <- ncol(Xs)
nint <- k*(k-1)/2
out <- NULL
n <- length(Y)
for (i in 1:(k-1)) for (j in (i+1):k) {
 Y_k <- Y - bb[i]*Xs[,i]-bb[j]*Xs[,j]   # eemaldame i-nda
     ja j-nda X moju
 stat <- sqrt(n)*mean(Y_k*Xs[,i]*Xs[,j])/apply(Y_k*Xs[,i]*
     Xs[,j],2,sd) # teststatistik
 pv <- 2*pt(abs(stat),n-1,lower.tail=F) # p-v"a"artus
 m <- summary(lm(Y~Xs[,i]*Xs[,j]))$coef # lineaarne mudel
     koosmojuga
 if (dim(m)[1]==4){
  beta <- m[4,1] #koosmojukordaja
  stat2 <- m[4,3] # t-statistik
  pv2 <- m[4,4] # koosmoju olulisus
```

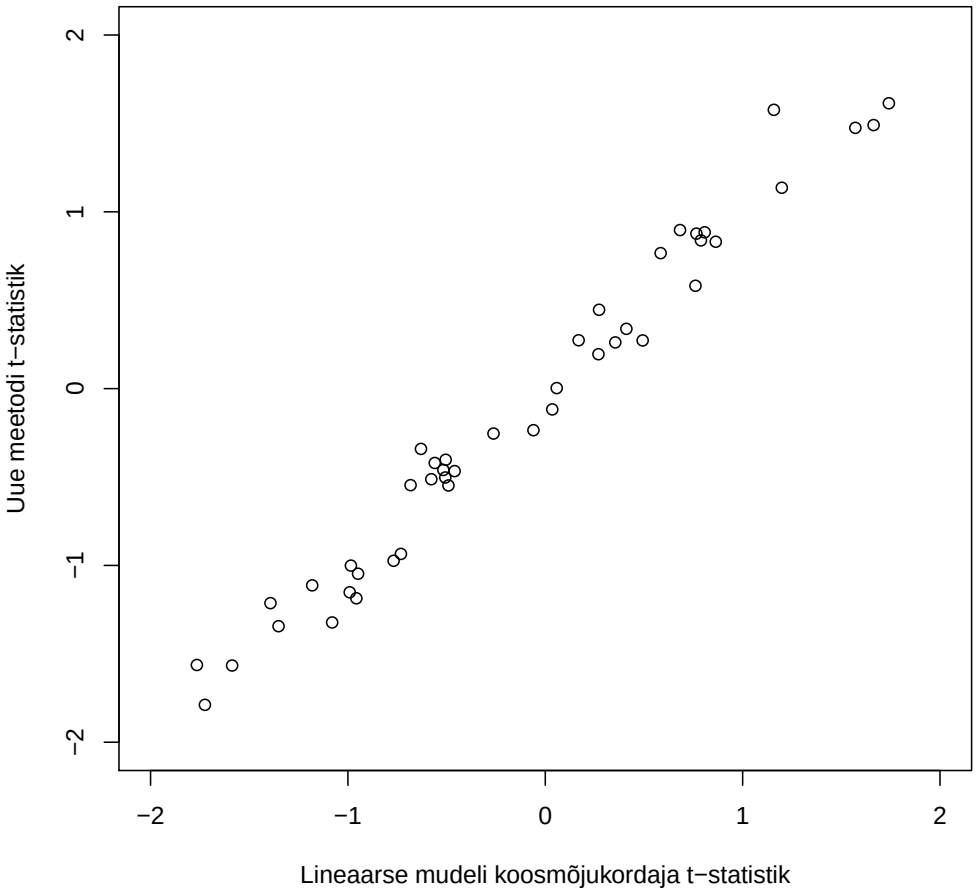
```
   out <- rbind(out,c(i,j,cor(Xs[,i],Xs[,j]),bb[i],bb[j],
      beta,stat,pv,stat2,pv2))
 }
}
colnames(out) <- c("i","j","corr","bi","bj","beta","T","pv"
    ,"Tc","pv2")
out <- as.data.frame(out)
table(out$pv<0.01,out$pv2<0.01)
table(out$pv<0.05,out$pv2<0.05)
cor(out$T,out$Tc)
pdf("fail3.pdf")
plot(out$Tc,out$T,xlab="Lineaarse mudeli koosmojukordaja t-
    statistik", ylab="Uue meetodi t-statistik") # "oige" ja
    meie T-statistiku vordlus
dev.off()
save(out, file="fail4.RData")
```

Graafik sõltumatuse uurimiseks

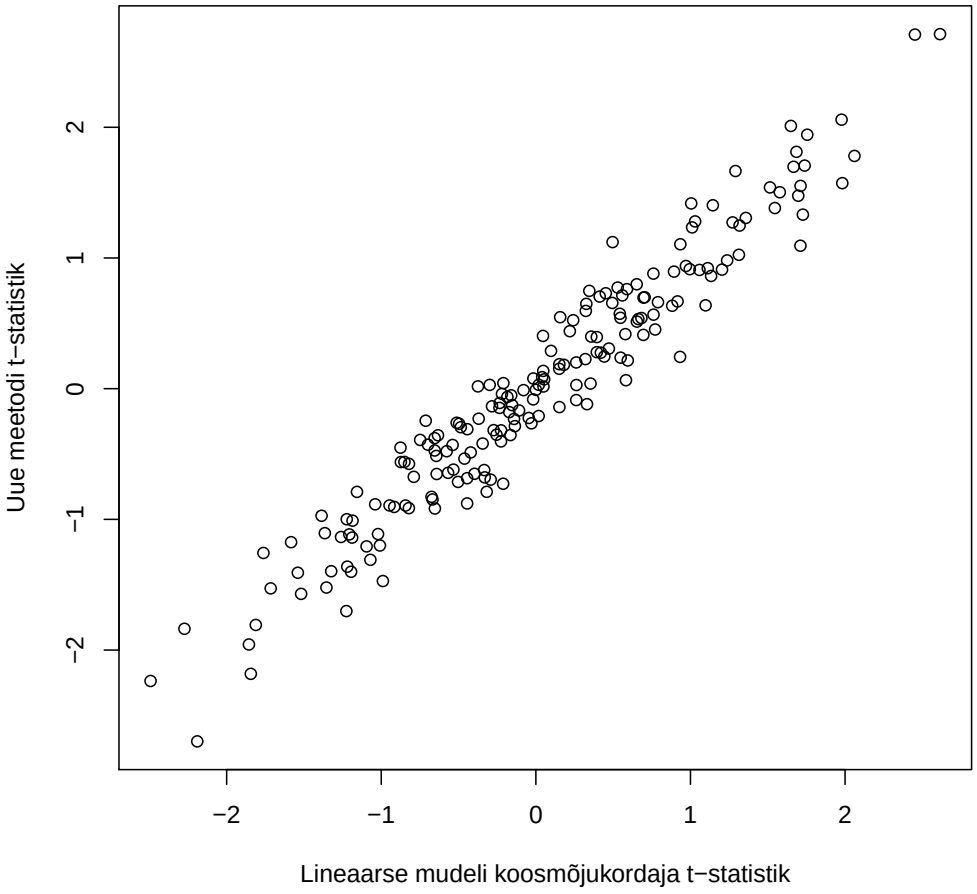
```
Y_k <- scale(Y-S) # S arvutada nii nagu esialgses skriptis
Y_s<-sample(Y_k)  # juhuslikult permuteeritud Y_k vaartused
t1 <- rep(NA,ncol(Xs))
t2 <- rep(NA,ncol(Xs))
for (i in 1:ncol(Xs)) {
 mud1<-summary(lm(Y_k~Xs[,i]))$coef
 mud2<-summary(lm(Y_s~Xs[,i]))$coef
```

```
  t1[i]<- mud1[2,3]
  t2[i]<-mud2[2,3]
}
pdf("fail.pdf")
plot(sort(t2),sort(t1),xlab="Permuteeritud valimi t-
    statistikud",ylab="Vaadeldud t-statistikud")
abline(0,1)
dev.off()
# kui seos X-dega oleks eemaldatud, peaksid punktid joonel
    asuma
```

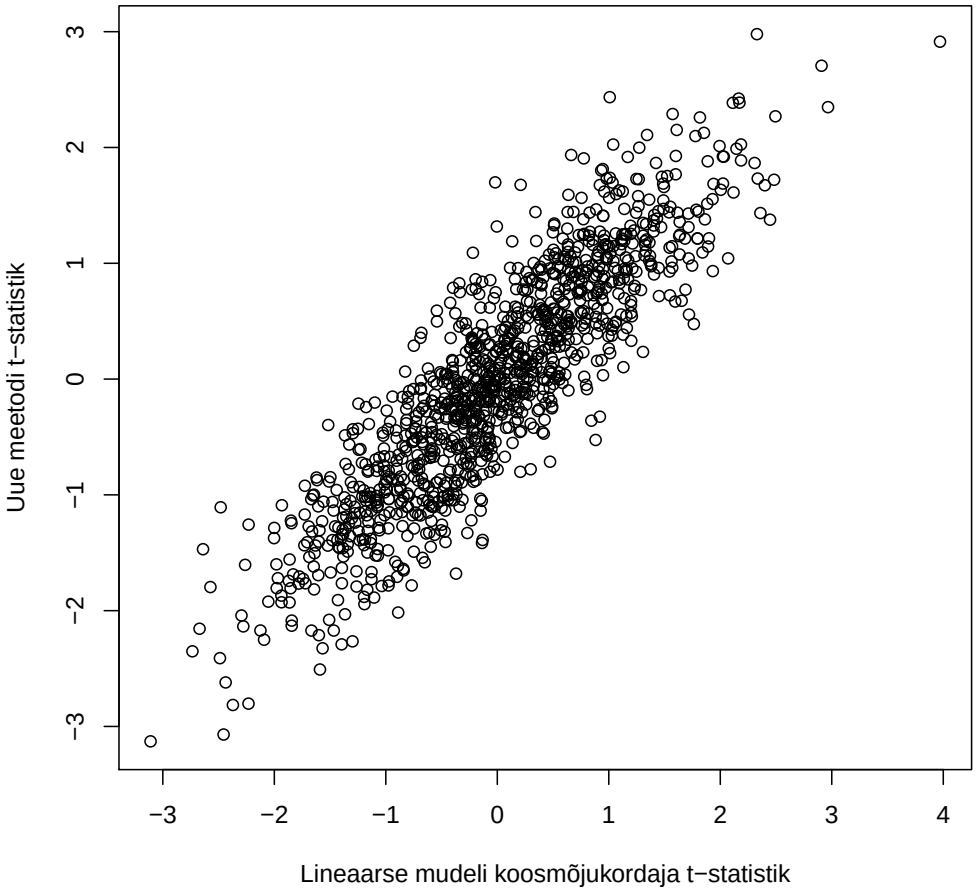
Lisa 2. Uue meetodi t-statistikut ja lineaarse regressiooni koosmõjukordaja t-statistikut võrdlevad joonised



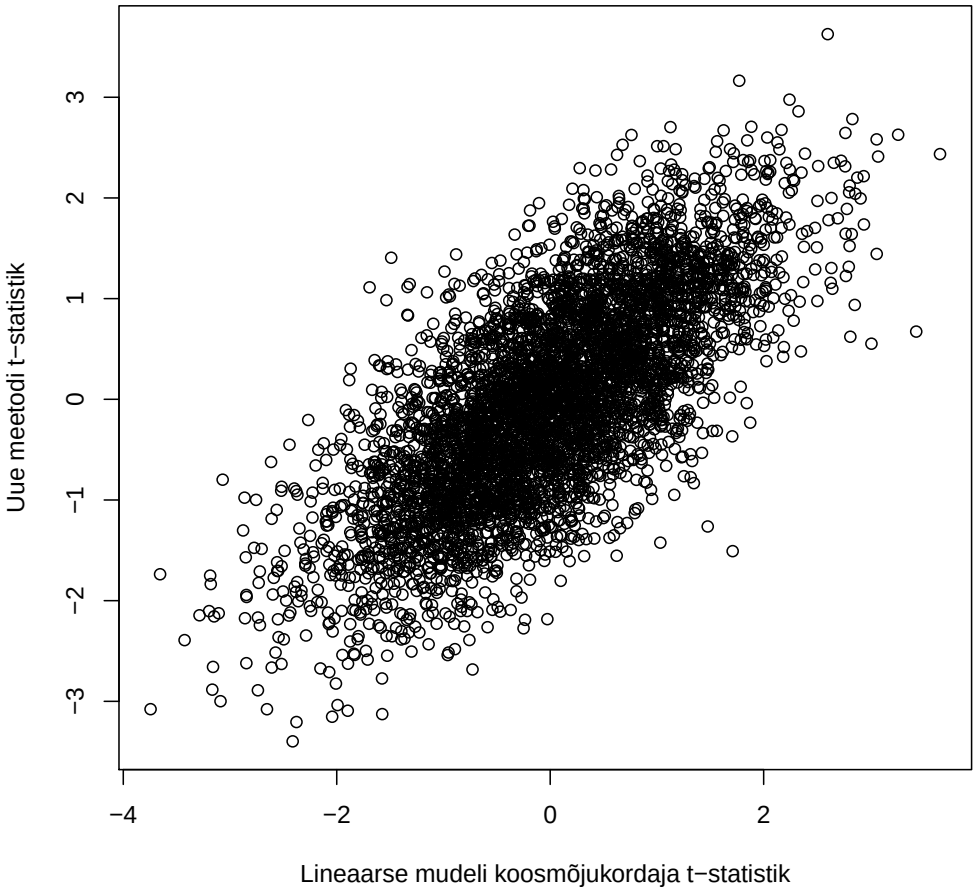
Joonis L1. Uue meetodi t-statistiku ja lineaarse regressiooni koosmõjukordaja t-statistikut võrdlus 10 SNP-i korral



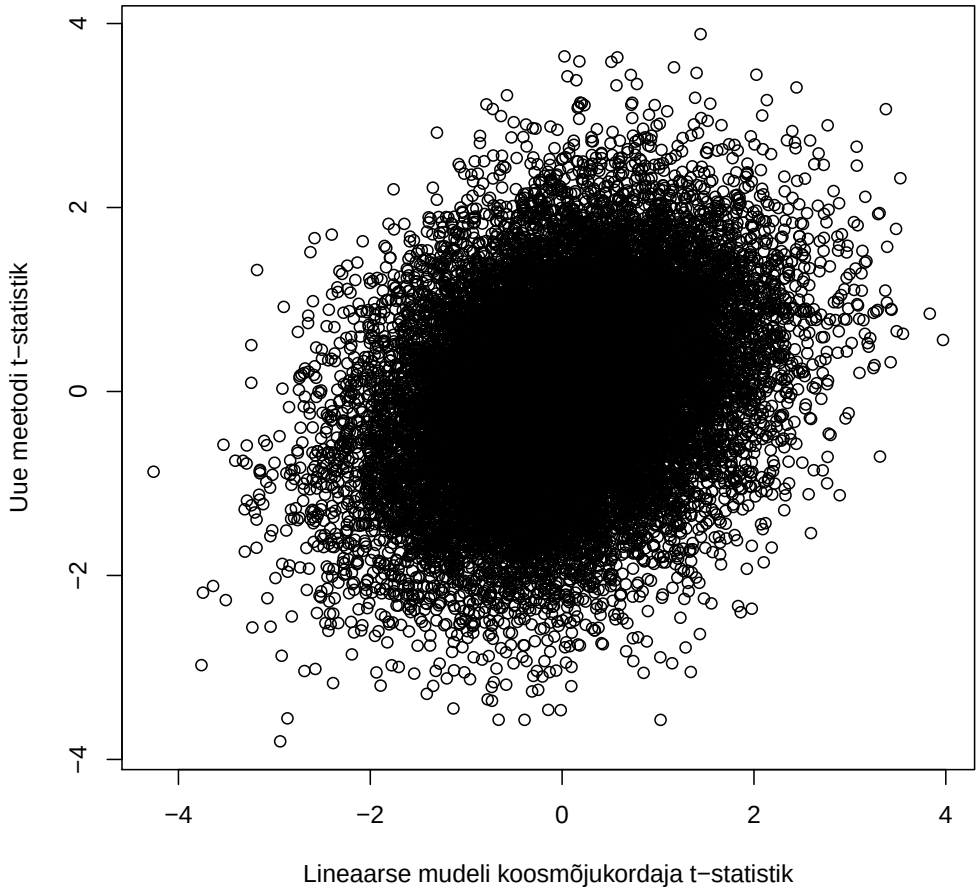
Joonis L2. Uue meetodi t-statistiku ja lineaarse regressiooni koosmõjukordaja t-statistikut võrdlus 20 SNP-i korral



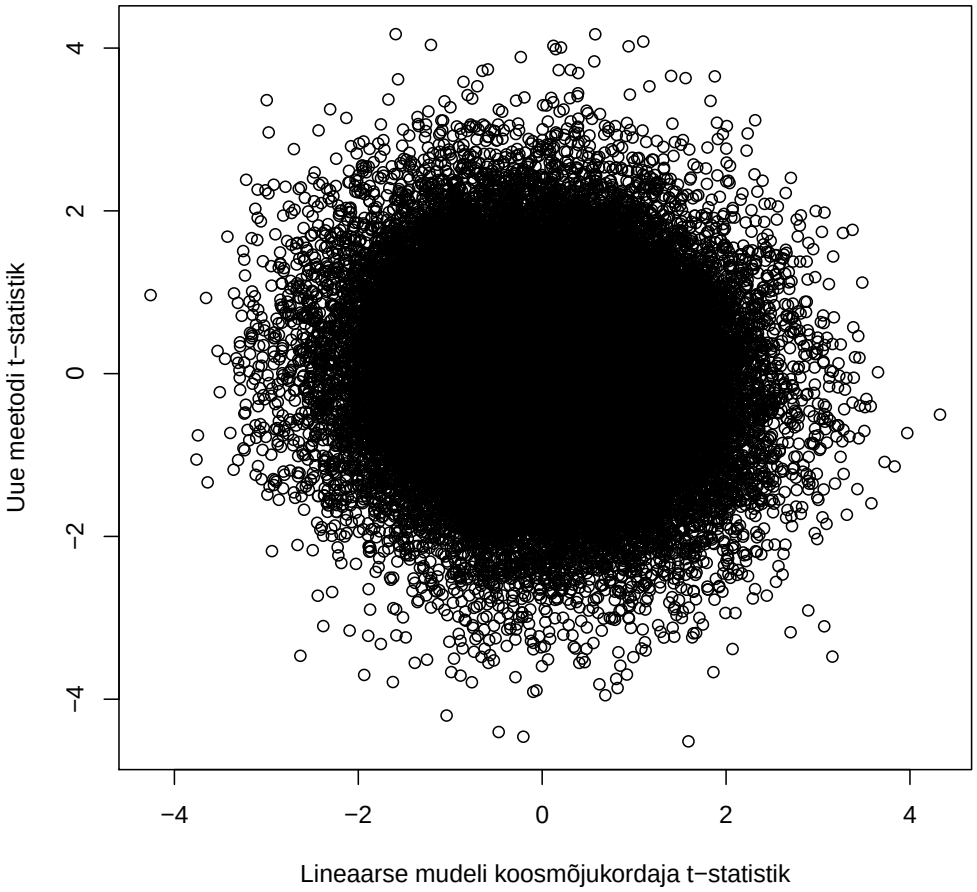
Joonis L3. Uue meetodi t-statistiku ja lineaarse regressiooni koosmõjukordaja t-statistikut võrdlus 50 SNP-i korral



Joonis L4. Uue meetodi t-statistiku ja lineaarse regressiooni koosmõjukordaja t-statistikut võrdlus 100 SNP-i korral



Joonis L5. Uue meetodi t-statistiku ja lineaarse regressiooni koosmõjukordaja t-statistikut võrdlus 200 SNP-i korral



Joonis L6. Uue meetodi t-statistiku ja lineaarse regressiooni koosmõjukordaja t-statistikut võrdlus 300 SNP-i korral

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Kristin Jesse (sünnikuupäev 20.08.1992)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Geengeen koosmõjude hindamine ülegenoomsetes uuringutes", mille juhendajad on Toomas Haller ja Krista Fischer,

   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