TARTU ÜLIKOOL
Matemaatika-informaatikateaduskond
Matemaatilise statistika instituut

Riho Klement

Geneetiliste markerite mõju muutlikkust kirjeldavad meta-analüüsi mudelid

Magistritöö

Juhendaja: Märt Möls, lektor

Tartu 2013

Sisukord

SISSEJUHATUS

Üha arenevate teadussaavutuste toel viiakse läbi aina rohkem ja rohkem uuringuid üle kogu maailma pea kõigis eluvaldkondades. On selge, et kogu maailma inimesi hõlmavat uuringut ei ole võimalik korraldada. Küll aga on võimalik uurida ühte ja sama asja väga paljudes erinevates inimrühmades. Tulemuseks on mingi väiksema üldkogumi kohta käivad tulemused ja seosed. Sageli on kasutusel väga erinev metoodika, millest tulenevalt on uuringutel erinev hindamistäpsus. Samuti võivad uuringutulemusi mõjutada asukohamaast tulenevad iseärasused. Et oleks võimalik kõiki ühelaadseid uuringutulemusi kuidagi üldistada, selleks ongi kasutusel meta-analüüs. Meta-analüüsi mõiste defineeris 1976. aastal Gene Glass, kui suure hulga üksikuuringute tulemuste statistilise analüüsi eesmärgiga võtta kokku uuringute tulemusi. Meta-analüüsi tehnikaid kasutati aga juba palju varem. Karl Pearson (1904) rakendas tüüfuse vaktsiini uuringutes leitud korrelatsioonikoefitsientide kombineerimiseks kohaldatud meetodit. Leonhard Henry Caleb Tippet (1931) ja Ronald Fisher (1932) esitlesid meetodeid p-väärtuste kombineerimiseks[1, lk 1].

Käesolevas töös tutvustatakse meta-analüüsi läbiviimiseks sobivaid meetodeid ning kasutades tutvustatud meetodeid, uuritakse inimese pikkusega seotud geneetiliste markerite mõjusid. Uuritakse, miks erinevates uuringutes näib sama geneetilise markeri mõju olevat erinev ning kas ka erinevate markerite hinnatud mõjude vahel võib olla korreleeritust.

Magistritöö koosneb kolmest peatükist. Esimeses peatükis tuuakse ülevaade meta-analüüsis kasutatavast teoreetilisest taustast. Teine peatükk keskendub praktiliseks analüüsiks olemasolevate andmete kirjeldamisele. Viimases peatükis kirjeldatakse läbiviidud analüüsi käiku ja sellest saadud tulemusi.

Analüüsi läbiviimiseks ning illustreerivate jooniste tegemiseks on kasutatud tarkvarapaketti ***R***. Autori koostatud programmikood on ära toodud töö lõpus olevas lisas 2. Töö kirjutamisel on kasutatud tekstitöötlusprogrammi ***MiKTeX***.

Töös kasutatud kirjandusele viitamiseks kasutatakse nurksulgi, kus vajadusel on lisaks numbrile, mis näitab viite järjekorda kasutatud kirjanduse loetelus, ära toodud ka leheküljed, kust kasutatud informatsioon on võetud.

1 META-ANALÜÜS

1.1 Fikseeritud mõjudega mudel

Järgnev põhineb suuresti allikal [1, lk 58-59] Olgu meil r sõltumatut uuringut, igas uuringus hinnatakse mingi töötluse mõju. Kõige lihtsama lähenemise korral eeldame, et uuritava töötluse tegelik mõju on θ. Iga uuringu tulemus (hinnang töötluse mõjule) olgu y_i. Siis saame uuringu tulemuste jaoks kirjutada välja mudeli:

$$y_i = \theta + \varepsilon_i,$$

kus ε_i, i=1,...,r on töötluse mõju hindamisel tehtud vead, mis on realisatsioonid normaaljaotusest keskväärtusega 0 ja dispersiooniga σ_i^2. Seega

$$y_i \sim N(\theta, \sigma_i^2).$$

Olgu w_i= $1/\sigma_i^2$. Eeldame, et

$$y_i \sim N(\theta, w_i^{-1}), i = 1, ..., r.$$

Kui kõikides uuringutes on ühesugune töötluse mõju ja eeldades, et uuringud on omavahel sõltumatud saame, et

$$\sum_{i=1}^{r} y_i w_i \sim N\left(\theta \sum_{i=1}^{r} w_i, \sum_{i=1}^{r} w_i\right),$$

millest saame üldise töötluse mõju hinnangu

$$\hat{\theta} = \frac{\sum_{i=1}^{r} y_i w_i}{\sum_{i=1}^{r} w_i}$$

ning 95-protsendiline usaldusvahemik θ -le leitakse valemiga

$$\hat{\theta} \pm 1,96 \sqrt{\frac{1}{\sum_{i=1}^{r} w_i}}.$$

Vaatame mudelit üldisel kujul :

$$Y = X\beta + \boldsymbol{\varepsilon}, \tag{1}$$

kus antud juhul Y on r-elemendiline uuringutulemuste vektor, X on plaanimaatriks, β on parameetrite vektor ning $\boldsymbol{\varepsilon}$ on jääkide vektor. Kui kõigis uuringutes on töötluse mõju sama, siis näeb mudel 1 välja selline:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_r \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \theta + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_r \end{pmatrix}.$$

Et uuringud eeldame olevat sõltumatud siis

$$\boldsymbol{\varepsilon} \sim N(0, R),$$

kus

$$R = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r^2 \end{pmatrix}.$$

Seega

$$Y \sim N(X\beta, R)$$

ehk antud juhul on vektori Y dispersioonimaatriks V=R.

Üldistatud vähimruutude hinnang parameetrite vektorile näeb välja järgmine:

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y. \tag{2}$$

Pannes selles hinnangus asemele vastavad väärtused saame:

$$\hat{\theta} = \left(\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}^T \begin{pmatrix} \sigma_1^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r^{-2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}^T \begin{pmatrix} \sigma_1^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r^{-2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_r \end{pmatrix},$$

millest järeldub, et

$$\begin{aligned}\hat{\theta} &= \left(\sum_{i=1}^{r} \frac{1}{\sigma_i^2}\right)^{-1} \cdot \sum_{i=1}^{r} \frac{y_i}{\sigma_i^2} \\ &= \left(\sum_{i=1}^{r} w_1\right)^{-1} \cdot \sum_{i=1}^{r} y_i w_i,\end{aligned}$$

millest saame, et

$$\hat{\theta} = \frac{\sum_{i=1}^{r} y_i w_i}{\sum_{i=1}^{r} w_i}. \tag{3}$$

See on aga juba tuttav hinnang. Saadud hinnang on sisuliselt uuringutulemuste kaalutud keskmine, kus kaaludeks on

$$\frac{w_i}{\sum_{i=1}^{r} w_i}, i = 1, ..., r.$$

Suuruste w_i definitsioonist näeme, et mida täpsem on uuringutulemus, ehk mida väiksem on vaadeldava töötluse mõju hinnangu dispersioon, seda suurem kaal sellele antakse.

1.2 Juhuslike mõjudega mudel

Järgnev põhineb allikal [1, lk 88-95] Eelnevalt tehtud eeldus, et meil on igas uuringus (igas uuritud populatsioonis) ühesugune tegelik töötluse mõju on reaalselt tihti liiga lihtsustav. Mõnikord on uuringud tehtud väga erinevates populatsioonides ning juba populatsiooni eripära ise mõjutab töötluse mõju. Olgu θ endiselt töötluse tegelik mõju ning olgu γ_i töötluse eripära i. populatsioonis. Tähistades töötluse mõju i. populatsioonis $\theta + \gamma_i$, saame uuringutulemuste jaoks mudeli kujul:

$$y_i = \theta + \gamma_i + \varepsilon_i,$$

kus ε_i, i=1,...r on endiselt töötluse mõju hindamisel tehtud vead keskväärtusega 0 ja dispersiooniga σ_i^2 ning γ_i, i=1,...,r on i-nda uuringu eripära, mida eeldatakse olevat normaaljaotusega keskväärtusega 0 ning dispersiooniga τ^2. Eeldades et γ_i ja ε_i on sõltumatud, saame:

$$y_i \sim N(\theta, \tau^2 + \sigma_i^2).$$

Tavaliselt eeldatakse, et uuringute (populatsioonide) mõjud on omavahel sõltumatud

$$\gamma_i \perp \gamma_j \forall i, j, i \neq j.$$

Sellisel juhul on uuritava tunnuse Y kovariatsioonimaatriks V kujul

$$V = \begin{pmatrix} \tau^2 + \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \tau^2 + \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tau^2 + \sigma_r^2 \end{pmatrix}.$$

Harilikult ei ole τ^2 teada, vaid tuleb hinnata andmetelt. Eeldame, et

$$y_i \sim N(\theta, \hat{\tau}^2 + \sigma_i^2),$$

kus $\hat{\tau}^2$ on hinnang τ^2 -le. Tähistades $w_i^\star$:=$(w_i^{-1} + \hat{\tau}^2)^{-1}$ saame

$$y_i \sim N(\theta, (w_i^\star)^{-1}).$$

Kui $(w_i^\star)^{-1}$ on y_i õige dispersioon, siis hinnang θ-le on arvutatav valemiga

$$\hat{\theta}^\star = \frac{\sum_{i=1}^{r} y_i w_i^\star}{\sum_{i=1}^{r} w_i^\star},$$

95%-usaldusintervall θ-le avaldub kujul

$$\hat{\theta}^\star \pm 1,96 \sqrt{\frac{1}{\sum_{i=1}^{r} w_i^\star}}.$$

Eelduse kohaselt on Y jaotus normaaljaotus:

$$f(Y) = |2\pi V|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(Y - X\beta)^T V^{-1}(Y - X\beta)\right).$$

Parameetri τ^2 hinnangu leidmiseks vaatame tõepärafunktsiooni. Logaritmiline tõepärafunktsioon näeb antud juhul välja

$$l = -\frac{1}{2}\log|2\pi V| - \frac{1}{2}(Y - X\beta)^T V^{-1}(Y - X\beta).$$

Tulles tagasi konkreetse juhu juurde on ühe uuringu panus tõepärasse

$$L(\theta, \tau^2|y_i) = \frac{1}{\sqrt{2\pi(w_i^{-1} + \tau^2)}} \exp\left\{\frac{-(y_i - \theta)^2}{2(w_i^{-1} + \tau^2)}\right\}$$

ning r sõltumatu uuringu puhul näeb logaritmiline tõepärafunktsioon välja selline:

$$l(\theta, \tau^2|Y) = C - \frac{1}{2}\sum_{i=1}^{r}\log(w_i^{-1} + \tau^2) - \frac{1}{2}\sum_{i=1}^{r}\frac{(y_i - \theta)^2}{(w_i^{-1} + \tau^2)},$$

kus C on konstant. Suurima tõepära hinnangud θ-le ja τ^2-le on leitavad iteratiivse skeemi alusel, kus igal iteratsioonil esmalt τ^2 koheldakse kui fikseeritut ning arvutatakse θ väärtus, mis maksimiseerib tõepärafunktsiooni. Seejärel võetakse saadud θ väärtus fikseerituks ning arvutatakse τ^2 väärtus, mis maksimiseerib tõepärafunktsiooni. Nii saadakse θ hinnanguks t+1 sammul

$$\hat{\theta}^*_{t+1} = \frac{\sum_{i=1}^{r} y_i w^*_{it}}{\sum_{i=1}^{r} w^*_{it}}, t = 0, 1, ...$$

kus w^*_{it}=$(w_i^{-1} + \tau^2_{M,t})$ ning $\tau^2_{M,t}$ on suurima tõepära hinnang t-ndal iteratsioonil. Suurima tõepära hinnang t+1 sammul τ^2-le leitakse Newton-Rapshoni meetodiga. Esimesel sammul võib aga τ^2 fikseeritud väärtusena kasutada momentide meetodiga saadud hinnangut, mille tuletame järgnevalt.

Juhuslike efektidega mudeli korral on fikseeritud efektidega mudeli juures lei-

tud hinnangu

$$\hat{\theta} = \frac{\sum_{i=1}^{r} y_i w_i}{\sum_{i=1}^{r} w_i}$$

keskväärtus θ ning dispersioon esitub valemiga

$$\begin{aligned}
\mathrm{D}(\hat{\theta}) &= \frac{\sum_{i=1}^{r} w_i^2 \mathrm{D}(y_i)}{\left(\sum_{i=1}^{r} w_i\right)^2} \\
&= \frac{\sum_{i=1}^{r} w_i^2 (w_i^{-1} + \tau^2)}{\left(\sum_{i=1}^{r} w_i\right)^2} \\
&= \frac{1}{\sum_{i=1}^{r} w_i} + \frac{\tau^2 \sum_{i=1}^{r} w_i^2}{\left(\sum_{i=1}^{r} w_i\right)^2}
\end{aligned}$$

Heterogeensuse testimiseks kasutatakse statistikut mida nimetatakse ka Cochrani Q-ks (vaata peatükk 1.4). See on defineeritud järgnevalt:

$$\begin{aligned}
Q &= \sum_{i=1}^{r} w_i (y_i - \hat{\theta})^2 \\
&= \sum_{i=1}^{r} w_i (y_i - \theta)^2 - \left(\sum_{i=1}^{r} w_i\right) (\hat{\theta} - \theta)^2.
\end{aligned}$$

Detailne tuletuskäik viimase võrduse kohta on saadaval [2, lk 20]. Kuna

$$\begin{aligned}
E(Q) &= \sum_{i=1}^{r} w_i D(y_i) - \left(\sum_{i=1}^{r} w_i\right) D(\hat{\theta}) \\
&= \sum_{i=1}^{r} w_i (w_i^{-1} + \tau^2) - \left(\sum_{i=1}^{r} w_i\right) \left\{ \frac{1}{\left(\sum_{i=1}^{r} w_i\right)} + \frac{\tau^2 \sum_{i=1}^{r} w_i^2}{\left(\sum_{i=1}^{r} w_i\right)^2} \right\} \\
&= (r-1) + \tau^2 \left(\sum_{i=1}^{r} w_i - \frac{\sum_{i=1}^{r} w_i^2}{\sum_{i=1}^{r} w_i} \right),
\end{aligned}$$

(detailsem tuletus [2, lk 21]), siis on τ^2 nihketa hinnanguks

$$\hat{\tau}^2 = \frac{Q - (r-1)}{\sum_{i=1}^{r} w_i - \frac{\sum_{i=1}^{r} w_i^2}{\sum_{i=1}^{r} w_i}} \tag{4}$$

1.3 Sõltuvate uuringute meta-analüüs

Ka eeldus, et uuringud on sõltumatud, võib reaalses kontekstis olla veidi lihtsustav. Töötluse mõju hindamiseks võidakse teha erinevaid uuringuid üle kogu maailma ning üksteisele lähemal paiknevates populatsioonides on erinevad varjatud mõjud töötlusele, näiteks populatsiooni käitumisharjumused ja mõõtmismetoodika, sarnased. Selliseid peidetud faktoreid võiks kuidagi arvesse võtta ka uuringute tulemuste kombineerimisel. Üheks võimaluseks on hinnata kovariatsioonimaatriks, kus kahes uuringus saadavate töötluste mõjude vaheline kovariatsioon sõltub kahe uuritava populatsiooni kaugusest. Näiteks tavaline geograafiline kaugus uuringute läbiviimise kohtade vahel või geneetiline kaugus kahe uuringu all olevate indiviidide grupi vahel. Tutvustame siinkohal kahte näidet, mis pärinevad allikast [3]. Esiteks eksponentsiaalne kovariatsioonistruktuur kujul

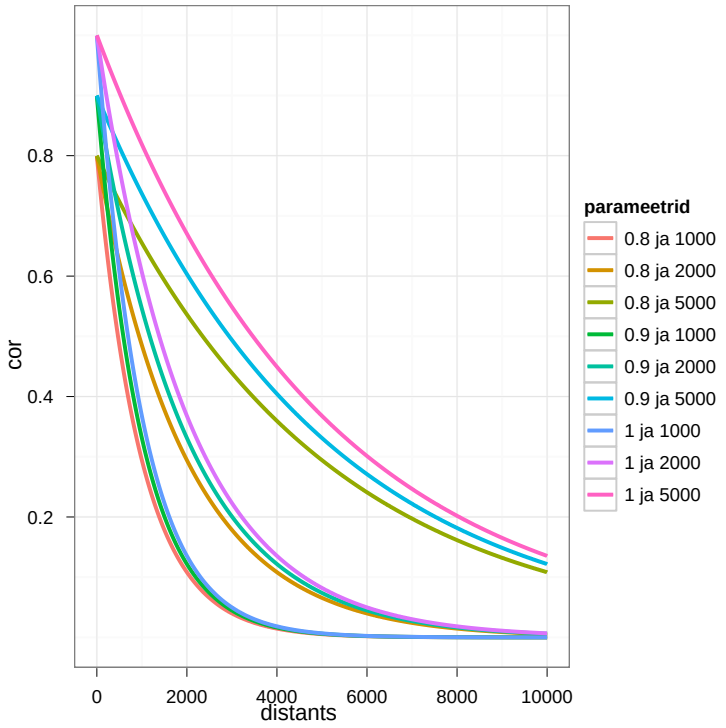
$$\begin{aligned} \sigma_{ij} &= \text{cov}(\theta + \gamma_i, \theta + \gamma_j) \\ &= \tau^2 \cdot \exp\left(\frac{-d_{ij}}{\theta}\right), \end{aligned}$$

kus d_{ij} on kahe uuringu vaheline kaugus ning θ on hinnatav parameeter. Nüüd juhul kui kaks uuringut toimuvad näiteks samas riigis, siis kirjeldatud struktuur eeldab, et nende uuringute tegelike mõjude vaheline korrelatsioon on 1. See ei pruugi aga alati nii olla. Võib esineda olukord, kus samas kohas läbi viidud uuringutes olid vaatluse all erinevas vanuses või erinevate käitumisharjumus-

tega inimesed, erinevad meetodid, või veel mõned taolised faktorid, mis põhjustavad selle, et töötluse mõju pole isegi samas kohas tehtud uuringute puhul täpselt sama. Antud olukord esineb ka näiteks käesolevas töös, kus kahe eraldi uuringuna käsitletakse ühes kohas läbi viidud uuringut eraldi naiste ja meeste kohta. Sellise situatsiooni tarbeks võetakse kasutusele nn kamakaefekt. Sellisel juhul avaldub kovariatsioonistruktuur kujul

$$\sigma_{ij} = (1 - \rho) \cdot \tau^2 \cdot \exp\left(\frac{-d_{ij}}{\theta}\right),$$

kus ρ tähistabki nn kamakaefekti.



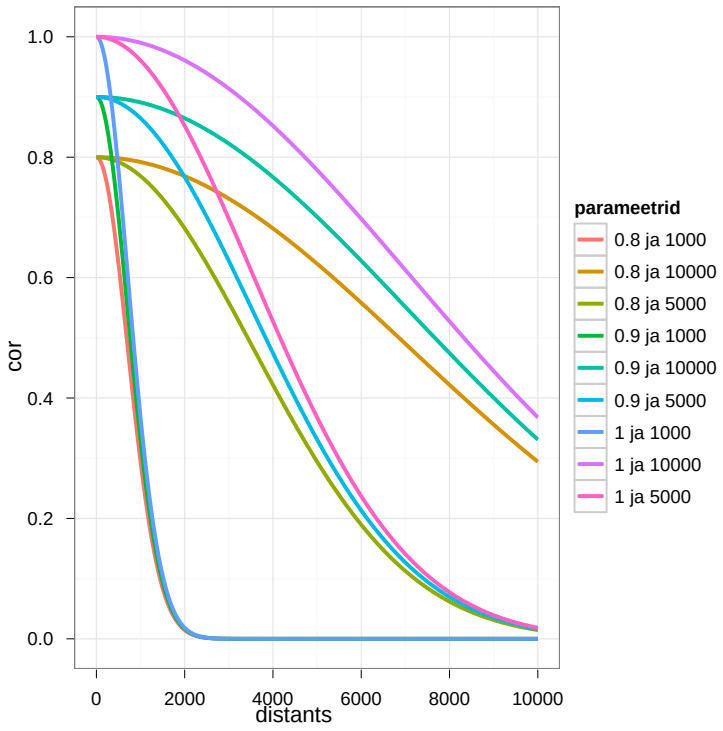
Joonis 1: Eksponentsiaalne korrelatsioonistruktuur

Kirjeldatu illustreerimiseks on joonisel 1 toodud korrelatsioonid sõltuvalt distantsist eksponentsiaalse kovariatsioonistruktuuri põhjal. Kasutatud on kolme erinevat parameetri θ väärtust (1000, 2000 ja 5000) ning igaühte neist on oma-

korda kujutatud kolme erineva kordaja 1-ρ korral (1; 0,9; 0,8). Joonisel 2 on kujutatud analoogne situatsioon Gaussi kovariatsioonistruktuuri korral. Gaussi kovariatsioonistruktuuri puhul avaldub kahe uuringu efektide vaheline kovariatsioon kujul

$$\sigma_{ij} = (1 - \rho) \cdot \tau^2 \cdot \exp\left(\frac{-d_{ij}^2}{\theta^2}\right),$$

kus ρ on taas kamakakordaja, d_{ij} on kahe uuringu vaheline distants ning θ on hinnatav parameeter.



Joonis 2: Gaussi korrelatsioonistruktuur

Kirjeldatud kovariatsioonistruktuuride korral on uuritava tunnuse Y kova-

riatsioonimaatriks V kujul

$$V = \begin{pmatrix} \sigma_1^2 + \tau^2 & \sigma_{12} & \cdots & \sigma_{1r} \\ \sigma_{21} & \sigma_2^2 + \tau^2 & \cdots & \sigma_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{r1} & \sigma_{r2} & \cdots & \sigma_r^2 + \tau^2 \end{pmatrix}$$

Töötluse tegelik mõju võib sõltuda erinevatest uuringupopulatsiooni kirjeldavatest tunnustest ja tundmatutest parameetritest β (vaata 1). Senises käsitluses oleme vaadanud ainult juhtu, kus meil on vaid üks parameeter θ ehk töötluse mõju tegelik väärtus. Sellise juhu kirjeldamiseks eeldatakse normaaljaotust keskväärtusega $X\beta$.

$$Y \sim (X\beta, V)$$

üldistatud vähimruutude hinnang parameetrile β on toodud valemiga (2). Leiame selle hinnangu dispersiooni

$$\begin{aligned} D\hat{\beta} &= D\left[\left(X^T V^{-1} X\right)^{-1} X^T V^{-1} Y\right] \\ &= \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} DY \left[\left(X^T V^{-1} X\right)^{-1} X^T V^{-1}\right]^T \\ &= \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} V \left(V^{-1}\right)^T X \left[\left(X^T V^{-1} X\right)^{-1}\right]^T \\ &= \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} X \left[\left(X^T V^{-1} X\right)^{-1}\right]^T \\ &= \left[\left(X^T V^{-1} X\right)^{-1}\right]^T \\ &= \left(X^T V^{-1} X\right)^{-1} \end{aligned}$$

Nüüd juhul kui meil on palju seletavaid tunnuseid, tekib küsimus, mille põhjal valida, milline tunnus jätta mudelisse ja milline mitte. Selleks kasutame tõepärasuhte testi. Tõepärasuhte teststatistik on kujul

$$\Lambda(x) = -2 \cdot \log \lambda(x) = 2 \cdot (l_1 - l_0),$$

kus l_1 tähistab keerulisema mudeli tõepärafunktsiooni maksimaalset väärtust ning l_0 lihtsama mudeli oma. Põhimõte seisneb selles, et enamasti suurendab rohke-

mate tunnuste kasutamine mudelis tõepärafunktsiooni väärtust, ehk siis keerulisem mudel ei ole kehvem kui lihtsam mudel. Ülesanne on näidata, kas keerulisem mudel on ka statistiliselt oluliselt parem. On teada, et tõepärasuhte teststatistik on asümptootiliselt hii-ruut jaotusega

$$\Lambda(x) \xrightarrow{D} \chi^2_p,$$

kus

$$p = df_1 - df_0$$

kus df_1 on keerulisema mudeli ning df_0 on lihtsama mudeli hinnatavate parameetrite arv.

1.4 Cochrani Q

Heterogeensust testitakse selleks, et otsustada, kas oleks targem kasutada fikseeritud mõjudega mudelit või juhuslike mõjudega mudelit. Statistik Q on uuringus saadud töötluse mõjude ja töötluste mõjude kaalutud keskmiste vahede ruutude kaalutud summa.

$$Q = \sum_{i=1}^{r} w_i(y_i - \hat{\theta})^2,$$

kus $\hat{\theta}$ on toodud valemiga 3. Kui see summa on väga suur, siis see viitab sellele, et kõikides uuringutes ei pruugi olla töötluse tegelik mõju ühesugune. Kui aga kehtib nullhüpotees, kehtib fikseeritud mõjudega mudel, siis on statistik Q asümptootiliselt hii-ruut jaotusega vabadusastmete arvuga r-1, kus r on uuringute arv. See tuleneb järgnevast [4, lk 210-211]:

Lause.

Olgu Y mitmemõõtmeline K-vektor jaotusega

$$Y \sim N(\mu, \Sigma),$$

kus Σ on teadaolev mittesingulaarne diagonaalmaatriks pöördmaatriksiga $W = \Sigma^{-1}$ ja μ on K-mõõtmeline vektor komponentidega μ_k, k=1,...,K. Tähistame W

k-nda diagonaalielemendi tähega w_k ja defineerime

$$p_k = \frac{w_k}{\sum_{j=1}^{K} w_j},$$
$$\mu_w = \sum_{k=1}^{K} p_k \mu_k,$$
$$\bar{Y}_w = \sum_{k=1}^{K} p_k y_k.$$

Siis statistik S on mittetsentraalse hii-ruut jaotusega:

$$S = \sum_{k=1}^{K} w_k \left(y_k - \bar{Y}_w\right)^2 \sim \chi^2_{K-1}(\lambda),$$

kus mittetsentraalsuse parameeter on kujul

$$\lambda = \sum_{k=1}^{K} w_k \left(\mu_k - \mu_w\right)^2.$$

Järgnevalt on allikast pärit tõestus autori poolt detailsemalt lahti seletatud.

Tõestus.

Olgu P selline diagonaalmaatriks, mille k-s element on p_k, I olgu ühikmaatriks mõõtmetega K*K ning J olgu ühtedest koosnev maatriks, samuti mõõtmetega K*K. Nüüd

$$C = (I - JP)^T W (I - JP)$$

on sümmeetriline maatriks. Vaatame ruutvormi

$$Y^T C Y = Y^T (I - JP)^T W (I - JP) Y$$

Näeme, et

$$(I - JP) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} - \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_K \end{pmatrix}$$
$$= \begin{pmatrix} 1-p_1 & -p_2 & \cdots & -p_K \\ -p_1 & 1-p_2 & \cdots & -p_K \\ \vdots & \vdots & \ddots & \vdots \\ -p_1 & -p_2 & \cdots & 1-p_K \end{pmatrix}.$$

Nüüd

$$
\begin{aligned}
Y^TCY &= \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{pmatrix}^T \begin{pmatrix} 1-p_1 & -p_1 & \cdots & -p_1 \\ -p_2 & 1-p_2 & \cdots & -p_2 \\ \vdots & \vdots & \ddots & \vdots \\ -p_K & -p_K & \cdots & 1-p_K \end{pmatrix} \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_K \end{pmatrix} \begin{pmatrix} 1-p_1 & -p_2 & \cdots & -p_K \\ -p_1 & 1-p_2 & \cdots & -p_K \\ \vdots & \vdots & \ddots & \vdots \\ -p_1 & -p_2 & \cdots & 1-p_K \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{pmatrix} \\
&= \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{pmatrix}^T \begin{pmatrix} \frac{1-p_1}{\sigma_1^2} & \frac{-p_1}{\sigma_2^2} & \cdots & \frac{-p_1}{\sigma_K^2} \\ \frac{-p_2}{\sigma_1^2} & \frac{1-p_2}{\sigma_2^2} & \cdots & \frac{-p_2}{\sigma_K^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-p_K}{\sigma_1^2} & \frac{-p_K}{\sigma_2^2} & \cdots & \frac{1-p_K}{\sigma_k^2} \end{pmatrix} \begin{pmatrix} y_1 - \sum_{k=1}^K p_k y_k \\ y_K - \sum_{k=1}^K p_k y_k \\ \vdots \\ y_K - \sum_{k=1}^K p_k y_k \end{pmatrix} \\
&= \begin{pmatrix} \frac{y_1 - \sum_{k=1}^K p_k y_k}{\sigma_1^2} & \frac{y_2 - \sum_{k=1}^K p_k y_k}{\sigma_2^2} & \cdots & \frac{y_K - \sum_{k=1}^K p_k y_k}{\sigma_K^2} \end{pmatrix} \begin{pmatrix} y_1 - \sum_{k=1}^K p_k y_k \\ y_K - \sum_{k=1}^K p_k y_k \\ \vdots \\ y_K - \sum_{k=1}^K p_k y_k \end{pmatrix} \\
&= \sum_{k=1}^K w_k (y_k - \bar{Y}_w)^2 \\
&= S.
\end{aligned}
$$

Nüüd järeldub standardsest tulemusest [5, lk 128-129], et kui

$$Y \sim N(\mu, \Sigma),$$

kus Σ on mittesingulaarne ja kui C on sümmeetriline (mida meie konstrueeritud C on), siis ruutvormil Y^TCY on mittetsentraalne hii-ruut jaotus siis ja ainult siis, kui

$$\Sigma C \Sigma C \Sigma = \Sigma C \Sigma$$

ning sellisel juhul on vabadusastmete arv maatriksi $C\Sigma$ jälg ja mittetsentraalsuse parameeter on $\mu^T C \mu$. Paneme tähele, et

$$
\begin{aligned}
C\Sigma &= (I - JP)^T W (I - JP)\Sigma \\
&= \begin{pmatrix} 1-p_1 & -p_1 & \cdots & -p_1 \\ -p_2 & 1-p_2 & \cdots & -p_2 \\ \vdots & \vdots & \ddots & \vdots \\ -p_K & -p_K & \cdots & 1-p_K \end{pmatrix} \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_K \end{pmatrix} \begin{pmatrix} 1-p_1 & -p_2 & \cdots & -p_K \\ -p_1 & 1-p_2 & \cdots & -p_K \\ \vdots & \vdots & \ddots & \vdots \\ -p_1 & -p_2 & \cdots & 1-p_K \end{pmatrix} \begin{pmatrix} w_1^{-1} & 0 & \cdots & 0 \\ 0 & w_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_K^{-1} \end{pmatrix} \\
&= \begin{pmatrix} w_1(1-p_1) & -w_2 p_1 & \cdots & -w_K p_1 \\ -w_1 p_2 & w_2(1-p_2) & \cdots & -w_K p_2 \\ \vdots & \vdots & \ddots & \vdots \\ -w_1 p_K & -w_2 p_K & \cdots & w_K(1-p_K) \end{pmatrix} \begin{pmatrix} w_1^{-1}(1-p_1) & -w_2^{-1} p_2 & \cdots & -w_K^{-1} p_K \\ -w_1^{-1} p_1 & w_2^{-1}(1-p_2) & \cdots & -w_K^{-1} p_K \\ \vdots & \vdots & \ddots & \vdots \\ -w_1^{-1} p_1 & -w_2^{-1} p_2 & \cdots & w_K^{-1}(1-p_K) \end{pmatrix}.
\end{aligned}
$$

Vaatame nüüd saadud maatriksi $C\Sigma$ elemente

$$\begin{aligned}
(C\Sigma)_{1,1} &= (1-p_1)^2 + \frac{w_2}{w_1}p_1^2 + \cdots + \frac{w_K}{w_1}p_1^2 \\
&= 1 - 2p_1 + p_1^2 + \frac{w_2}{w_1}\frac{w_1}{\sum_{j=1}^{K} w_j}p_1 + \cdots + \frac{w_K}{w_1}\frac{w_1}{\sum_{j=1}^{K} w_j}p_1 \\
&= 1 - 2p_1 + p_1^2 + p_2 \cdot p_1 + \cdots + p_K p_1 \\
&= 1 - 2p_1 + (p_1 + p_2 + \cdots + p_K)p_1 \\
&= 1 - p_1, \\
(C\Sigma)_{2,1} &= -p_2(1-p_1) - \frac{w_2}{w_1}(1-p_2)p_1 + \cdots + \frac{w_K}{w_1}p_2 p_1 \\
&= -p_2 + p_1 p_2 - p_2(1-p_2) + \cdots + p_K p_2 \\
&= (-1 + p_1 - 1 + p_2 + \cdots + p_K)p_2 \\
&= -p_2.
\end{aligned}$$

Analoogselt tulevad ka kõik teised elemendid, ning kokku saame, et

$$\begin{aligned}
C\Sigma &= \begin{pmatrix} 1-p_1 & -p_1 & \cdots & -p_1 \\ -p_2 & 1-p_2 & \cdots & -p_2 \\ \vdots & \vdots & \ddots & \vdots \\ -p_K & -p_K & \cdots & 1-p_K \end{pmatrix} \\
&= (I - JP)^T \\
&= I - PJ.
\end{aligned}$$

Kuna

$$JPJ = J$$

siis

$$\begin{aligned}\Sigma C\Sigma C\Sigma &= \Sigma(I-PJ)(I-PJ))\\ &= \Sigma(I-PJ-PJ+PJPJ)\\ &= \Sigma(I-PJ-PJ+PJ)\\ &= \Sigma(I-PJ)\\ &= \Sigma C\Sigma.\end{aligned}$$

Vabadusastmete arv statistiku S jaotuse jaoks tuleb

$$\begin{aligned}\text{tr}(C\Sigma) &= \text{tr}(I-PJ)\\ &= 1-p_1+1-p_2+\cdots+1-p_K\\ &= K-1\end{aligned}$$

ning mittetsentraalsuse parameeter on

$$\begin{aligned}\lambda &= \mu^T C\mu\\ &= \sum_{k=1}^{K} w_k(\mu_k-\mu_w)^2,\end{aligned}$$

millest näeme, et statistiku S hii-ruut jaotus on tsentraalne parajasti siis, kui kõik tegelikud efektid μ_k on võrdsed. ■

2 ANDMESTIKU KIRJELDUS

Andmestikuna kasutatakse erinevate läbi viidud geeniuuringute tulemusi, kus on muuhulgas uuritud kuidas teatud geneetilised markerid mõjutavad inimeste pikkust. Geneetiliseks markeriks nimetatakse DNA järjestust, mille asukoht kromosoomis on teada. Markeriks võib olla osa mingist geenist või lihtsalt mingi DNA osa, millel ei pruugi olla ühtegi teadaolevat funktsiooni. SNP (*Single-nucleotide polymorphism*) on kindlaks tehtud koht DNA ahelas, kus erinevatel indiviididel võib asuda erineva lämmastikalusega nukleotiid. Lämmastikalus on üks kolmest ühendist (lämmastikalus, suhkur ja fosforhappe jääk), mille omavahelisel liitumisel moodustub nukleotiid. Nukleotiidide liitumise tulemusel tekibki DNA ahel. DNA ehituses esineb nelja erinevat lämmastikalust: adeniin (A), guaniin (G), tümiin (T) ja tsütosiin (C). Näide SNP-ist:

...AAGCCTA... ja ...AAGCTTA...

Siin on näha et lühike DNA ahela osa võib kahel inimesel erineda vaid ühe nukleotiidi poolest. SNP-id on ühed enamlevinumad geneetilised markerid ning ka vaatluse all olevates uuringutes on kasutatud just neid. [6, lk 7-8]

Andmestik pärineb konsortsiumilt GIANT (Genetic Investigation of ANthropometric Traits), mis tegeleb läbi rahvusvahelise koostöö geneetiliste markerite otsimisega, mis mõjutavad inimese keha suurust ja kuju. GIANT konsortsium on läbi viinud meta-analüüsi, kasutades vaid fikseeritud efektidega mudelit [7]. Käesolevas töös püütakse uurida ka heterogeensust ja üritatakse kirjeldada geneetiliste markerite mõju sõltuvust erinevatest tunnustest.

Andmestik koosnes 100 uuringu tulemustest. Kõikide uuringute peale kokku oli vaatluse all 180 SNP-i, kuid mõningates uuringutes oli mõni SNP puudu. SNP-id ei olnud valitud juhuslikult, vaid nendel SNP-idel oli varasematest assotsiat-

siooniuuringutest teadaolev seos inimese skeleti kasvuga. Väidetavalt peaksid need SNP-id kirjeldama umbes 10% kogu fenotüübilisest pikkuse hajuvusest. Iga uuringu kohta oli raporteeritud iga SNP-i kohta tema mõju hinnang, hinnangu standardviga, alleel, mille suhtes efekti on mõõdetud, samal kohal asuv teine alleel, see kumba DNA ahelat on vaadatud, valimi maht. Uuringute kohta olid veel eraldi andmed selles uuringus osalenud inimeste kohta. Tunnuste *kehakaal*, *pikkus*, *kehamassiindeks* ja *vanus* kohta oli teada keskmine väärtus, maksimaalne ja minimaalne väärtus, mediaan, standardhälve, korrelatsioon *pikkusega* ja korrelatsioon *kehamassiindeksiga*. Samuti oli teada ka uuringus osalenute *sugu*. Lisaks oli olemas informatsioon imputeerimise kohta, sest uuringus ei olnud mitte kõik SNP-id mõõdetud vaid nii mõnelgi juhul oli kasutatud imputeerimist (geneetilise markeri väärtus oli prognoositud tema läheduses paiknevate SNP-ide järgi. Kuna valdavalt olid meeste ja naiste uuringud eraldi, siis sellest tulenevalt eemaldasime valimist kõik uuringud, kus olid mõõdetud nii mehi kui naisi korraga, sest plaan oli uurida ka seda, kuidas inimese sugu mõjutab seda, kuidas mingi SNP mõjutab inimese pikkust. Lisaks oli väga mitmeid uuringuid, kus uuringus osalenute kohta käivaid andmeid ja uuringu tulemuste kohta käivaid andmeid ei õnnestunud kokku viia. Kokkuvõttes õnnestus analüüsiks vajalikule kujule viia 54 uuringut.

Enne analüüsi alustamist tuli algandmeid teisendada. Erinevates uuringutes oli sama SNP-i väärtus mõõdetud erineva DNA-ahela pealt. Kuna DNA koosneb kahest ahelast, mis on omavahel kokku keerdunud topeltspiraali kujuliselt, nimetatakse kokkuleppeliselt üht ahelat „+ " -ahelaks ja teist ahelat „-" -ahelaks. Kaks ahelat püsivad omavahel koos komplementaarsusprintsiibi alusel, mis määrab selle, et ühe ahela adeniini vastas on teise ahela peal alati tümiin, ning guaniini vastas tsütosiin. Lisaks avastasime, et osades uuringutes on SNP-i mõju mõõdetud ühe alleeli suhtes ja ülejäänud uuringutes teise alleeli suhtes. Ka need viisime kõik ühele kujule muutes teise alleeli suhtes mõõdetud mõju hinnangu märki.

3 ANALÜÜSI KÄIK JA TULEMUSED

Esmalt uurisime, kas andmetes esineb heterogeensust. Selleks kasutasime Cochrani teststatistikut ja tema jaotust. Arvutasime teststatistiku Q iga SNP-i kohta ning kasutades hii-ruut jaotust vabadusastmete arvuga r-1, kus r on uuringute arv, leidsime p-väärtuse. Tulemuseks saime, et 180-st markerist 11 puhul tulnuks vastu võtta alternatiivne hüpotees, ehk et uuringute vahel esineb heterogeensust. Kuna kasutasime olulisusnivood 0,05, siis nullhüpoteesi kehtides võiks meil selliseid väärtusi olla 0,05 · 180=9. Kuna 11 moodustab 180-st 6,11%, siis võiks sarnast tulemust oodata ka juhul, kui iga markeri tegelikud mõjud kõigis populatsioonides oleksid olnud ühesugused. Väikeste p-väärtustega markerid on toodud tabelis 1.

Tabel 1: Markerid, mille puhul võib uuringute vahel esineda heterogeensust

JRK	SNP	Q	p-väärtus
58	"rs17780086"	74,0505	0,0119
67	"rs2110001"	75,8271	0,0136
91	"rs2871865"	77,2795	0,0102
95	"rs3129109"	69,0632	0,0468
96	"rs3764419"	77,3868	0,0161
99	"rs3812163"	77,3563	0,0162
108	"rs4640244"	69,8773	0,0407
137	"rs7155279"	71,3831	0,0469
145	"rs7466269"	93,0955	0,0006
150	"rs7567851"	80,5780	0,0086
180	"rs9969804"	73,0106	0,0356

Uurime neid geneetilisi markereid lähemalt. Kuna just need on sellised, mille puhul erinevates uuringutes võib tegelik efekti väärtus olla erinev, siis uurime,

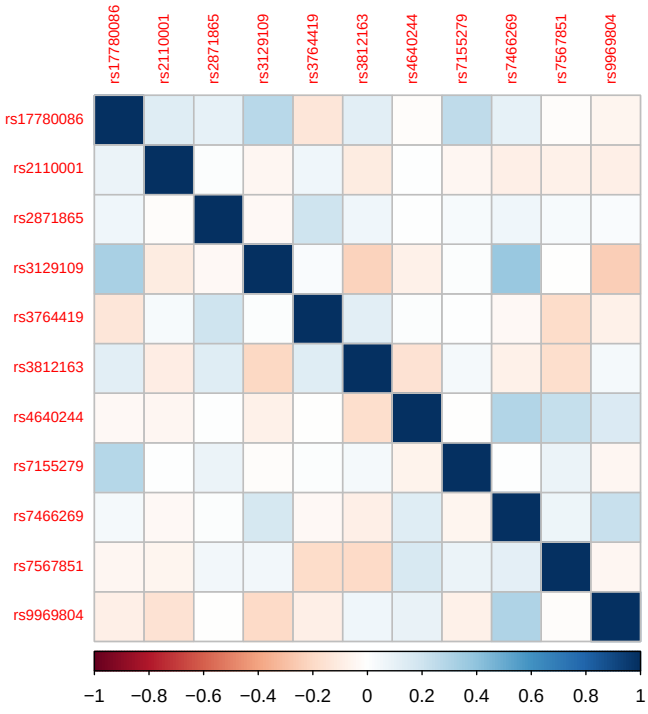
kuidas on nende markerite efektid korreleeritud omavahel. Selleks vaatame just tegelikke efekte. Olgu meil kahe markeri tegelikud efektid M_1 ja M_2. Siis

$$M_1 = c_0 + c_1 \cdot M_2 + \varepsilon.$$

Nüüd M_1 ja M_2 vaheline korrelatsioon avaldub kujul

$$\text{cor}(M_1, M_2) = \frac{c_1 \cdot \text{D}M_2}{\sqrt{\text{D}M_1 \cdot \text{D}M_2}},$$

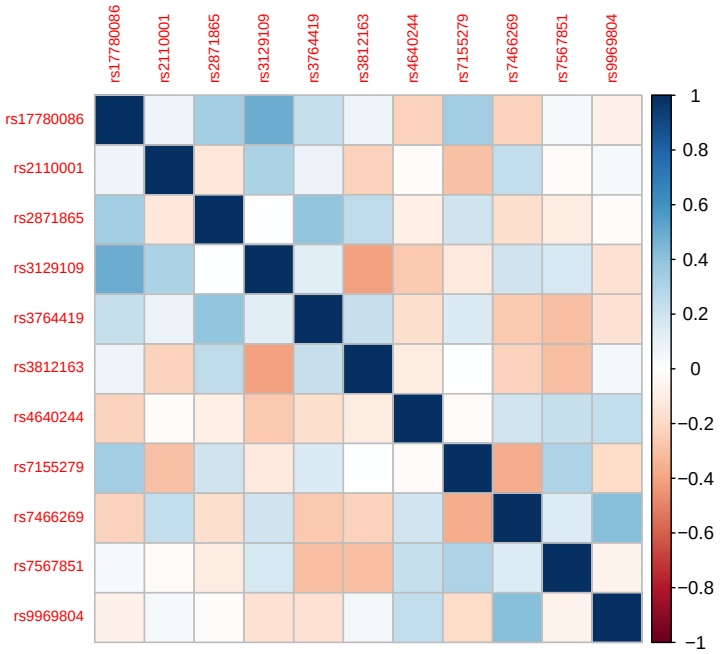
kus markeri mõju tegelik dispersioon on hinnatav valemiga 4. Saadud korrelatsioonid on kujutatud joonisel3.



Joonis 3: Valitud markerite vahelised korrelatsioonid

Jooniselt on näha, et väga suuri korrelatsioone ei esine. Suurim hinnatud korrelatsioon oli 0,36. Samuti torkab silma, et hinnatud maatriks ei ole päris sümmeetriline. Põhjustatud on see ilmselt sellest, et regressioonikordaja c_1 on alahinnatud, sest pole arvestatud teise markeri tegeliku mõju M_2 hindamisviga. Keerukaks teeb olukorra see, et seletava tunnuse rolli pandud markerite tegelikud

efektid pole teada. Teeme ühe graafiku veel, kus vaatleme samade markerite mõjuhinnangute omavahelisi korrelatsioone, mis on leitud programmi ***R*** käsuga *cor*. Tulemused on kantud joonisele 4. Sellelt jooniselt on näha juba natukene tugevamaid korrelatsioone. Selliselt leitud korrelatsioonimaatriks näitab pigem seoseid markerite efektide hinnangute mõõtmistäpsuse vahel.



Joonis 4: Valitud markerite vahelised korrelatsioonid

Toome nüüd analüüsi sisse mõned uued tunnused. Uurime kas mõni geneetiline marker mõjutab inimese pikkust erinevalt sõltuvalt sellest, kas tegu on mehe või naisega. Lisaks toome sisse veel tunnused *keskmine pikkus*, *keskmine kaal* ja *keskmine vanus*. Kasutame kõige lihtsamat juhtu, kus tunnuse Y kovariatsioonimaatriks sisaldab peadiagonaalil uuringutes raporteeritud standardvigu. Kasutame tõepärasuhte testi. Kuna iga tunnuse puhul on tegu ühe hinnatava parameetriga siis statistilise olulisuse näitamiseks kasutame hii-ruut jaotust vabadusastmete arvuga 1. Testisime tunnuseid ükshaaval ning tulemuste liht-

samaks ülevaateks arvutasime välja p-väärtused. Tulemused on kantud tabelisse 2 (vaata lisa 1) vasakpoolsesse ossa. Tulemuste uurimisel selgus, et markereid, mille korral tunnuse $sugu$ olulisust näitav p-väärtus oli alla 0,05 on meil valimis 7. $Keskmise\ pikkuse$ korral on see number 6, $keskmisel\ kaalul$ samuti 6 ning $keskmisel\ vanusel$ 10. Kuna meil on 180 geneetilist markerit ja olulisuse nivoo on 0,05, siis nullhüpoteesi kehtides võiks selliseid markereid, mille p-väärtus on väiksem kui 0,05 olla oodatavalt 9 tükki. Seega võib öelda, et tunnuste $sugu$, $keskmine\ pikkus$ ja $keskmine\ kaal$ korral on väikese p-väärtusega geneetilisi markereid isegi nullhüpoteesi kehtimise korral oodatust veidi vähem, $keskmine\ vanus$ aga vastab enam-vähem oodatule. Tegime sarnase analüüsi läbi niimoodi, et kõiki nimetatud tunnuseid sisaldavat mudelit võrreldi mudeliga, mis ühte neist ei sisaldanud. Need tulemused on kantud tabelisse 2 parempoolsetesse tulpadesse. Taaskord kasutasime olulisuse nivood 0,05 ja tunnuste kaupa on markerite arvud, mis osutusid statistiliselt olulisteks: 10,10,8 ja 13. Taaskord võrdleme saadud tulemusi arvuga 9. Võime öelda, et $keskmise\ vanuse$ korral ilmnes väikeseid p-väärtuseid oodatavast veid sagedamini, teiste tunnuste puhul oli tulemus üsna oodatav.

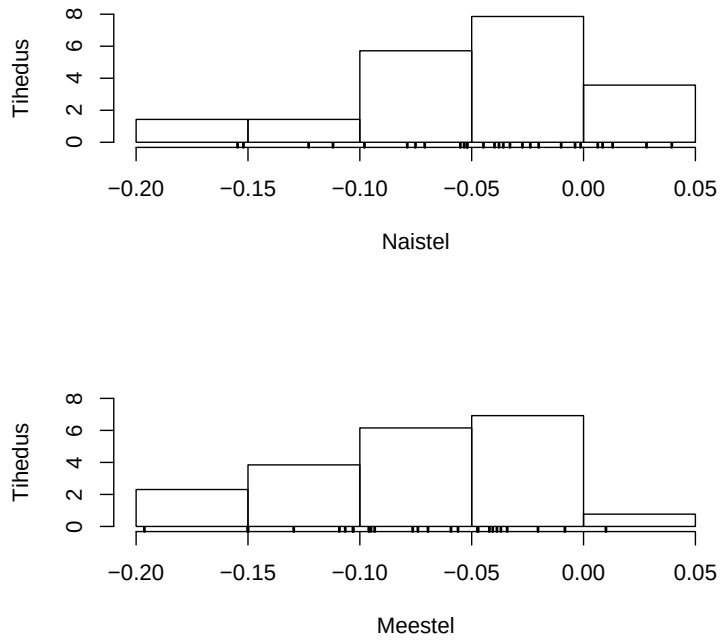
Tabeli põhjal valisime mõned huvitavamad markerid, mida uurisime veidi lähemalt. Esiteks võtsime vaatluse alla SNP-i nr 19, „rs11259936". Pannes mudelisse vaid $sugu$ ja testides tõepärasuhte testiga selle olulisust saime p-väärtuseks 0,0082, $keskmise\ kaalu$ korral oli p-väärtus 0,016. Samuti pakkus huvi see kui kõik vaatluse all olevad neli tunnust lisati mudelisse ning testiti ükshaaval nende olulisust. Ka siin osutusid $sugu$ ja $keskmine\ kaal$ olulisteks (p-väärtused 0,0418 ja 0,0376). See näitab, et vaadeldava tunnuse (näiteks $keskmine\ kaal$) seos geneetilise markeri mõjusuurusega pole tingitud teiste vaadeldud tunnuste ($sugu$, $keskmine\ pikkus$, ...) segavast mõjust. Näiteks, on loomulik, et mehed on tavaliselt keskmiselt suurema kaaluga kui naised. Juhul kui selline olukord kehtib, siis lisades tunnust $sugu$ sisaldavasse mudelisse veel tunnuse $kaal$ võib juhtuda, et

$kaal$ osutub statistiliselt mitteoluliseks, sest tema mõju uuritavale tunnusele on tingitud vaid tunnuse $sugu$ kaudsest mõjust. Tabelit 2 vaadates näeme, et näiteks markeri nr 16, „rs11118346" puhul on $sugu$ üksi mudelisse panduna osutunud oluliseks, aga kui ka $keskmine\ pikkus$, $keskmine\ kaal$ ja $keskmine\ vanus$ on arvesse võetud, siis $sugu$ justkui enam midagi juurde ei annaks. Samuti on näha ka vastupidiseid olukordi, näiteks markeri nr 20, „rs1159750" , korral on näha, et $sugu$ üksinda mudelisse pannes oluliseks ei osutu. Seega justkui meestel ja naistel oleks selle markeri mõju pikkusele ühesugune. Samas kui panna ülejäänud tunnused ka mudelisse, osutub $sugu$ oluliseks, mis näitab, et ühesuguse vanuselise, pikkuselise ja kehamassiga inimeste grupi puhul on meestel ja naistel markeri mõju pikkusele veidi erinev.

Pöördume nüüd tagasi markeri nr 19 juurde. Proovisime lisaks $soole$ panna mudelisse ka teisi tunnuseid. $Keskmise\ kaalu$ puhul tuli p-väärtuseks 0,0539. Teised tunnused osutusid selgelt mitteolulisteks, aga $kaalu$ otsustasime siiski sisse jätta. Tulemuseks siis mudel

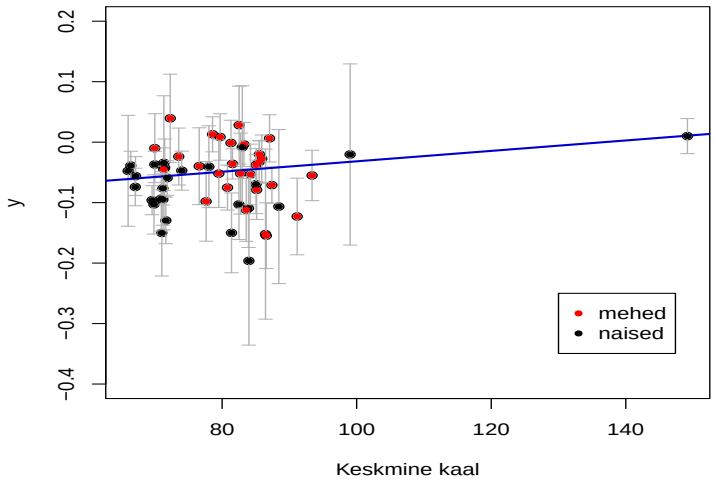
$$y = -0,1177 + 0,0244 \cdot SUGU_{naine} + 0,0007 \cdot keskmine.kaal + \varepsilon$$

Joonisel 5 on kujutatud vaadeldava markeri hinnatud efektid inimese pikkusele meeste ja naiste kaupa. Näha on, et naistel on rohkem punkte nulli ümbruses kui meestel. Seega naistel on vastava SNP-i efekt väiksem, kuid mehed, kelle genoomis paikneb SNP-i „rs11259936" kohal „A" on lühemad, kui mehed, kellel on samas kohas genoomis „C".

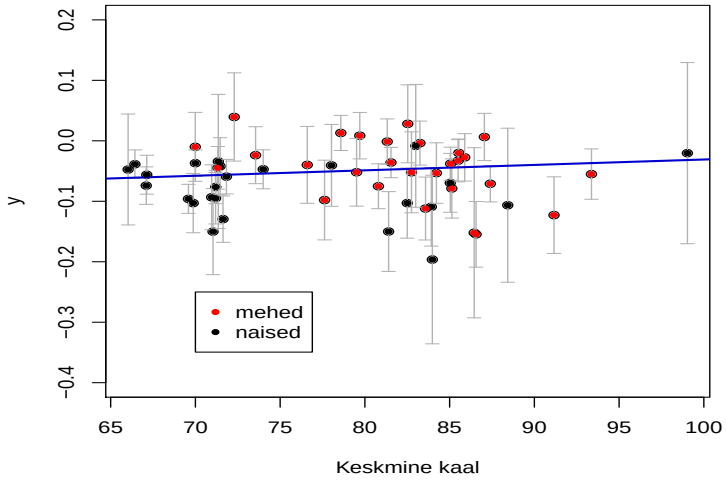


Joonis 5: Markeri „rs11259936" efektid naistel ja meestel

Joonisel 6 on näidatud seost $keskmise\ kaalu$ ja hinnatud efekti vahel (joonis kirjeldab mudelit, kus teisi tunnuseid mudelis ei ole). Joonisel torkab silma üks väga suure keskmise kehakaaluga inimeste seas tehtud uuring, mis ilmselt mõjutab ka seose olulisust. Joonisel 7 on see, uuring analüüsist välja jäetud. Näha on, et seal on efekti ja kaalu vahelist seost kirjeldav joon üsna x-teljega paralleelne. Seda kinnitab ka uuesti hinnatud p-väärtus, mis $keskmise\ kaalu$ puhul tuleb 0,2220. Seega võib öelda, et ilmselt mõjutab see uuring ka mitmeid teisi p-väärtusi. Samas teiste tunnuste juures erindlikke uuringuid ei paistnud, seega otsustasime olukorda, kus üks uuring on välja jäetud, sügavamalt mitte uurida.



Joonis 6: Markeri „rs11259936" efektid sõltuvalt uuritava kohordi kaalust



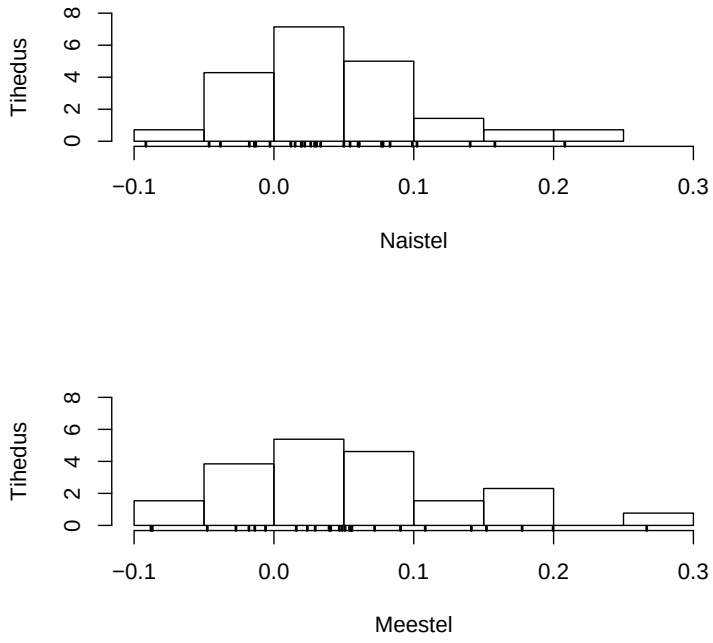
Joonis 7: Markeri „rs11259936" efektid sõltuvalt uuritava kohordi kaalust

Võtsime vaatluse alla veel ühe SNP-i, nr 177, „rs9844666". Nagu tabelist 2 näha, on sellel markeril oluline $keskmine\ pikkus$. $Sugu$ on oluline siis kui ka teised tunnused on mudelis. Kui vaatame mudelit, kus sees on uuritavate keskmine pikkus ja sugu, siis $soole$ tuleb p-väärtuseks 0,0511. Jätame siiski ka peaaegu

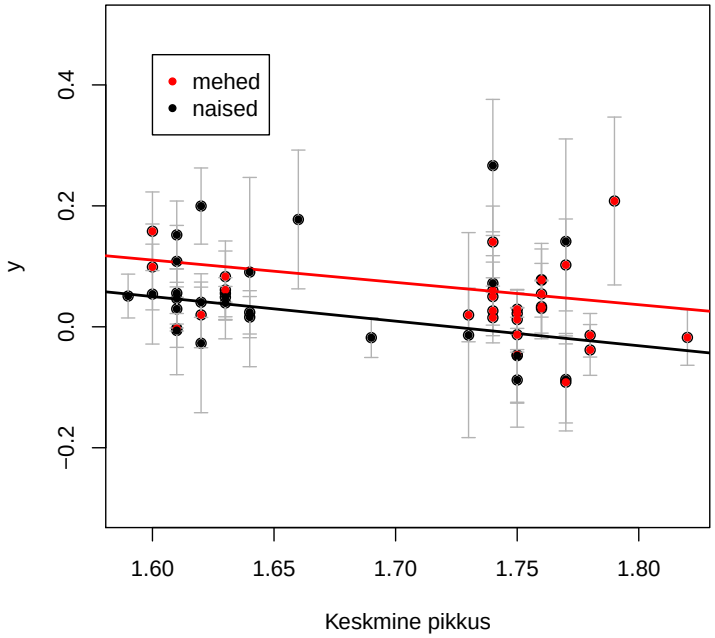
statistiliselt olulise tunnuse $sugu$ mudelisse, kuid eemaldame teised, statistiliselt mitteolulised tunnused.

$$y = 0,7009 - 0,0377 \cdot SUGU_{naine} - 0,4068 \cdot keskmine.pikkus + \varepsilon$$

Teeme ka seekord analoogsed joonised.



Joonis 8: Markeri „rs9844666" efektid sõltuvalt soost



Joonis 9: Markeri „rs9844666" efektid sõltuvalt uuritava kohordi kaalust

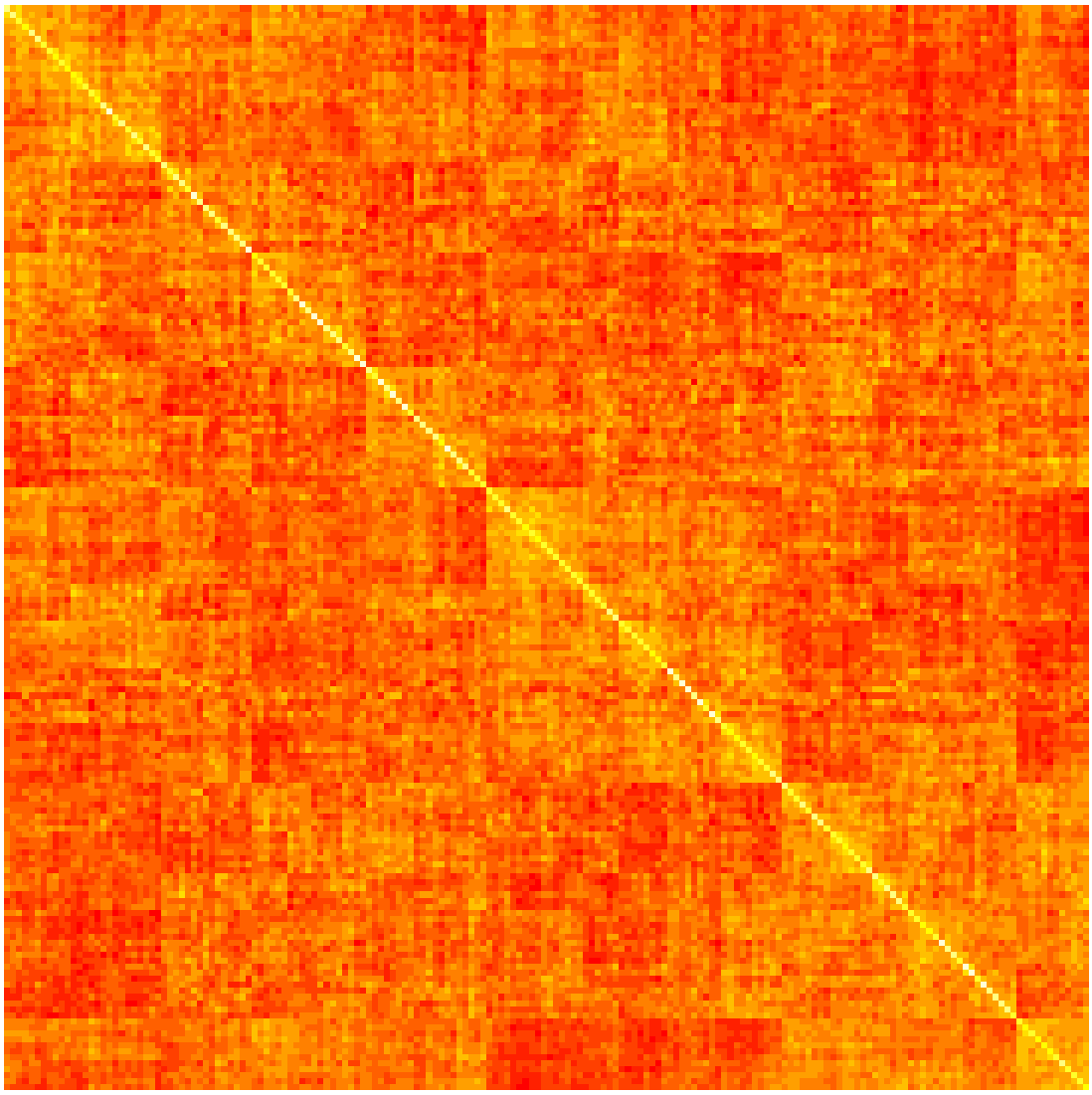
Jooniselt 8 on selgelt näha, miks $sugu$ üksinda mudelis ei ole oluliseks osutunud.Nii naistel kui meestel paiknevad hinnangud üsna sarnaselt. Jooniselt 9 võib näha, et nendes uuringutes, kus keskmine pikkus on olnud suurem, on selle markeri mõju hinnatud veidi väiksemaks, kui nendes uuringutes, kus pikkus on väiksem. Siiski kõik uuringud päris ilusti joone ümber ei ole ja väga täpseid järeldusi antud juhul teha ei saa.

Veel uurisime, kas geneetiliste markerite efektid võivad olla omavahel korreleeritud. Võib ju olla väga võimalik, et kui ühe markeri mõju pikkusele on suurem, siis on kohe suurem ka mõne teise markeri mõju pikkusele. Selleks hindasime mudeli, kus seletava tunnusena on sees ühe geneetilise markeri hinnatud efektid samades populatsioonides. Leidsime suurima tõepära hinnangu kordajale c, mille korrutasime läbi seletava tunnuse rollis oleva geneetilise markeri hinnatud

efektide dispersiooniga

$$\begin{aligned}\mathrm{cov}(X, Y) &= \mathrm{cov}(X, c \cdot X + \varepsilon) \\ &= \mathrm{cov}(X, c \cdot X) = c \cdot \mathrm{D}X.\end{aligned}$$

Tulemuseks saime kovariatsioonimaatriksi mõõtmetega 180*180. Sellest üritasime välja lugeda geneetilisi markereid, mida võiks omavahel kuidagi grupeerida. Selgus, et nähtavaid mustreid ei esinenud. Seejärel otsustasime leida Pearsoni korrelatsioonimaatriksi markerite efektide hinnangutele, et näha, kas nende vahel on seost. See seos võib tekkida näiteks ühesugusest mõõtmisveast, või sarnasest kasutatavast metoodikast. Tulemused on kantud joonisele 10. Joonisel on korrelatsioone kujutatud värvitud ruudukestena.Värvid varieeruvad valgest kuni punaseni läbi kollase. Mida valgem seda positiivsem on korrelatsioon ning mida punasem, seda negatiivsem on korrelatsioon. Kollased ruudud tähistavad nullilähedasi korrelatsioone. Joonisel on markerite järjekorda muudetud leidmaks sarnaselt käituvate markerite gruppe. Nagu näha on päris mitu sellist gruppi ka tekkinud. Tuleb tähele panna, et meie vaatluse all olevad uuringud on mõned sellised, kus uuring on läbi viidud ühes kohordis, aga selle kohta on meil eraldi andmed naiste ja meeste, juhtude ja kontrollide kohta. Samas on näha et nii mõnigi grupp on ka veidi suurem, kuigi väga selgeid mustreid siiski ei ole.



Joonis 10: Markerite efektide hinnangute vahelised korrelatsioonid

Kokkuvõte

Töös viidi läbi meta-analüüs, mille eesmärgiks oli üldistada läbiviidud 54 uuringu tulemusi. Kõikides uuringutes oli raporteeritud geneetiliste markerite mõjud inimese pikkusele. Kokku oli vaatluse all 180 geneetilist markerit, mis olid välja valitud eelnevate assotsiatsiooniuuringute põhjal.

Eelnevalt kirjeldatud teooriat kasutades uuriti reaalse andmestiku põhjal, kas markerite efektide juures esineb uuringutevahelist heterogeensust, kasutades selleks Cochrani Q testi. Ilmnes, et heterogeensust esines vaid 11 markeri korral 180st, mis oli oodatud piirides, see tähendab, et võttes arvesse mitmest testimist, võib ligikaudu nii palju markereid heterogeenuse poolest oluliseks osutuda ka nullhüpoteesi kehtides. Seda arvestades kasutati mudelite hindamisel fikseeritud efektidega mudelit. Püüti uurida, kas geneetilise markeri mõju inimese pikkusele sõltub ka uuringupopulatsiooni soost, keskmisest pikkusest, keskmisest kaalust või keskmisest vanusest. Iga markeri kohta hinnati mudelid, mis sisaldasid ühte mainitud tunnustest ning hinnati tunnuse statistilist olulisust näitav p-väärtus. Selgus, et oluliseks osutunud tunnustega mudeleid oli ligikaudu nii palju nagu oleks võinud oodata ka juhul, mil ükski neist seostest poleks tegelikult statistiliselt oluline. Kõike seda silmas pidades hinnati siiski kahe markeri jaoks mudel, mis sisaldas rohkem kui ühte uuringupopulatsiooni kirjeldavat tunnust. Lisaks uuriti, kas markerite mõjude hinnangud on omavahel korreleeritud.

Kokkuvõttes võib öelda, et geneetiliste markerite mõju võib sõltuda nii teiste geneetiliste markerite mõjudest, kui ka erinevatest uuringupopulatsiooni kirjeldavatest tunnustest. Kuigi antud töös väga tugevaid seoseid välja ei tulnud, ilmnes vihjeid selle kohta, et antud seoseid on mõtet kindlasti ka edaspidistes uuringutes vaatluse alla võtta.

Modelling the effect of genetic markers with meta-analysis methods

Master's thesis

Riho Klement

Summary

The purpose of the present master' s thesis is to introduce methods for meta-analysis. Some of these methods accept possible heterogeneity between studies.

When the results of different studies are collected together, we would like to generalize them. The paper describes different possible covariance structures of the collected data. The first one is for the assumption that all studies have the same real effect of the treatment. The second structure accepts heterogeneity between studies and the third one describes the covariance structure, that depends on the distance between studies. The paper also gives an explanation of Cochran's Q test for heterogeneity between studies.

In the last chapter, data from GIANT consortium is used to model the effect of genetic markers to human height using methods described in the paper. Altogether 54 studies are involved in the analysis.

Kasutatud kirjandus

[1] Whitehead A. 2002. *Meta-Analysis of Controlled Clinical Trials*, John Wiley and sons, UK

[2] Kool, P. 2010. *Sõltuvate uuringute meta-analüüs*, magistritöö, Tartu Ülikool [WWW]
http://dspace.utlib.ee/dspace/bitstream/handle/10062/15090/Kool_Pille.pdf
(20.05.2013)

[3] SAS online doc, *The MIXED Procedure*, *Repeated Statements*, *Spatial Covariance Structures* [WWW]
http://support.sas.com/onlinedoc/913/getDoc/en/statug.hlp/mixed_sect19.htm#idxmix0
(20.05.2013)

[4] Kulinskaya E. 2008., Morgenthaler S. , Staudte R. G. *Meta Analysis. A Guide to Calibrating and Combining Statistical Evidence*, John Wiley and sons (Wiley series in probability and statistics)

[5] Serfling, R. J. 2002. *Approximation theorems of mathematical statistics*, John Wiley and sons (Wiley series in probability and statistics)

[6] Klement, R. 2011 *Ülegenoomne assotsiatsiooniuuring hariduse omandamist ennustavate geneetiliste markerite leidmiseks*, bakalaureusetöö, Tartu Ülikool.

[7] GIANT consortium [WWW]
http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium
(20.05.2013)

Lisa 1: p-väärtuste tabel

Tabel 2: p-väärtused

Nr	SNP	Sugu	Pikkus	Kaal	Vanus	Sugu	Pikkus	Kaal	Vanus
1	"rs10010325"	0,8740	0,132	0,0564	0,344	0,0841	0,0441	0,1323	0,1165
2	"rs10037512"	0,6234	0,8747	0,0242	0,0361	0,1469	0,3744	0,0207	0,0119
3	"rs1013209"	0,7962	0,1465	0,4152	0,2202	0,3487	0,1354	0,738	0,4103
4	"rs10152591"	0,9732	0,5054	0,1957	0,5064	0,3944	0,1932	0,1042	0,6682
5	"rs1043515"	0,1190	0,1318	0,8417	0,9691	0,5314	0,6141	0,819	0,9364
6	"rs1046934"	0,7138	0,8127	0,3077	0,9195	0,8279	0,8237	0,2599	0,9838
7	"rs1046943"	0,5152	0,5415	0,8865	0,8865	0,8049	0,8318	0,9757	0,899
8	"rs1047014"	0,3750	0,4151	0,984	0,1565	0,4274	0,9136	0,7773	0,129
9	"rs10748128"	0,8923	0,7116	0,4414	0,1334	0,906	0,8425	0,329	0,1319
10	"rs10770705"	0,8537	0,9775	0,5167	0,8499	0,8057	0,9431	0,515	0,9297
11	"rs10799445"	0,8016	0,8189	0,2048	0,0712	0,7709	0,8437	0,1846	0,0571
12	"rs10838801"	0,8567	0,9394	0,7098	0,3803	0,5249	0,5111	0,6836	0,3134
13	"rs10863936"	0,6941	0,4446	0,7508	0,9172	0,7239	0,3902	0,5644	0,8134
14	"rs10874746"	0,1557	0,8184	0,8189	0,8665	0,0563	0,2141	0,8204	0,7492
15	"rs11107116"	0,7853	0,8534	0,609	0,1456	0,7788	0,7009	0,508	0,1774
16	"rs11118346"	0,0426	0,0013	0,7974	0,6513	0,3758	0,0032	0,1978	0,3711
17	"rs11144688"	0,2653	0,1606	0,7182	0,2596	0,6332	0,147	0,2608	0,1835
18	"rs11205277"	0,1615	0,3486	0,5841	0,8957	0,2974	0,8658	0,6548	0,994
19	"rs11259936"	0,0082	0,0823	0,0160	0,7061	0,0418	0,4098	0,0376	0,9466
20	"rs11599750"	0,1750	0,7796	0,622	0,4727	0,0170	0,0358	0,4565	0,97
21	"rs11648796"	0,6344	0,762	0,3983	0,3085	0,8971	0,9835	0,3996	0,2952
22	"rs11684404"	0,4599	0,4622	0,4925	0,5319	0,938	0,532	0,3524	0,5391
23	"rs1173727"	0,7117	0,8396	0,3913	0,7356	0,6472	0,6861	0,4011	0,6946
24	"rs11830103"	0,5550	0,2598	0,4115	0,3556	0,8212	0,2849	0,2112	0,4327
25	"rs11867479"	0,9876	0,8258	0,7743	0,9763	0,7679	0,7492	0,8269	0,945
26	"rs11958779"	0,2603	0,1268	0,5409	0,2515	0,7129	0,2099	0,835	0,1822
27	"rs12153391"	0,5476	0,2531	0,2134	0,6317	0,6221	0,348	0,3417	0,4623
28	"rs12470505"	0,3256	0,0974	0,757	0,2083	0,3925	0,0767	0,8495	0,1145
29	"rs12474201"	0,9506	0,5711	0,5347	0,0904	0,2863	0,2573	0,6319	0,0485
30	"rs12534093"	0,8475	0,8239	0,6489	0,2848	0,7727	0,8219	0,7245	0,3562
31	"rs1257763"	0,5222	0,5918	0,873	0,0476	0,8242	0,614	0,9938	0,0499
32	"rs12680655"	0,6387	0,8858	0,994	0,6986	0,3177	0,3868	0,9515	0,5249
33	"rs12694997"	0,6388	0,4161	0,7182	0,7719	0,732	0,4543	0,9046	0,6591
34	"rs12902421"	0,7067	0,5199	0,9782	0,928	0,8421	0,5508	0,7746	0,9806
35	"rs12982744"	0,1308	0,8086	0,4121	0,1847	0,0170	0,1209	0,3733	0,0544
36	"rs13088462"	0,3881	0,4175	0,1881	0,1211	0,9903	0,3498	0,1081	0,1276

37	"rs13177718"	0,0788	0,0681	0,6503	0,4114	0,7624	0,3639	0,9763	0,3983
38	"rs1325598"	0,7994	0,312	0,214	0,3994	0,305	0,2064	0,3258	0,2248
39	"rs1330"	0,0574	0,8635	0,2191	0,6077	0,0110	0,1224	0,2403	0,8347
40	"rs1351164"	0,4847	0,4655	0,9314	0,2756	0,63	0,9235	0,8717	0,261
41	"rs1351394"	0,1946	0,2637	0,0611	0,0148	0,9538	0,1898	0,0272	0,0168
42	"rs143384"	0,6972	0,612	0,4698	0,2472	0,7215	0,4024	0,3632	0,2142
43	"rs1468758"	0,4901	0,6874	0,5079	0,0528	0,277	0,589	0,3964	0,0312
44	"rs1490384"	0,9023	0,1447	0,0747	0,1412	0,2549	0,1604	0,194	0,3498
45	"rs1570106"	0,7008	0,6691	0,7767	0,3351	0,8501	0,592	0,6813	0,3141
46	"rs1582931"	0,2503	0,2777	0,29	0,9127	0,626	0,8879	0,4402	0,9498
47	"rs1659127"	0,9228	0,3702	0,0098	0,6111	0,3886	0,4135	0,0155	0,3573
48	"rs16942341"	0,5349	0,9541	0,9512	0,8997	0,3678	0,5145	0,9967	0,9417
49	"rs16964211"	0,1659	0,4187	0,579	0,4718	0,3288	0,9676	0,4497	0,615
50	"rs17081935"	0,5623	0,864	0,4269	0,0066	0,9446	0,9821	0,3816	0,0075
51	"rs1708299"	0,2881	0,2158	0,2211	0,761	0,7738	0,7166	0,3771	0,8051
52	"rs17318596"	0,4563	0,4079	0,8968	0,7341	0,9515	0,5802	0,7247	0,7164
53	"rs17346452"	0,2689	0,4832	0,7786	0,371	0,5127	0,9616	0,887	0,4465
54	"rs1738475"	0,1445	0,3628	0,5186	0,632	0,1998	0,788	0,3595	0,4561
55	"rs17391694"	0,8221	0,9236	0,7053	0,9783	0,6179	0,6096	0,6506	0,9293
56	"rs1741344"	0,9430	0,4264	0,383	0,7136	0,3324	0,2623	0,5269	0,4824
57	"rs17511102"	0,7977	0,5719	0,4257	0,8249	0,3377	0,3387	0,5328	0,9342
58	"rs17780086"	0,4054	0,9464	0,7199	0,2299	0,1109	0,1835	0,7133	0,1218
59	"rs17782313"	0,5442	0,5011	0,2956	0,0797	0,1923	0,2323	0,4334	0,207
60	"rs17806888"	0,3573	0,586	0,774	0,6929	0,3391	0,6245	0,847	0,5374
61	"rs1814175"	0,7709	0,7468	0,8327	0,5832	0,4911	0,4985	0,9353	0,7389
62	"rs1950500"	0,2057	0,1349	0,3019	0,8708	0,8064	0,5352	0,5392	0,9554
63	"rs2066807"	0,0314	0,2174	0,6992	0,3428	0,0942	0,5715	0,8958	0,6078
64	"rs2072153"	0,1619	0,2538	0,7867	0,8713	0,3898	0,9703	0,9744	0,7703
65	"rs2079795"	0,8550	0,1567	0,5408	0,0581	0,1108	0,0251	0,2388	0,176
66	"rs2093210"	0,3981	0,1501	0,216	0,3522	0,7963	0,182	0,0857	0,4376
67	"rs2110001"	0,5459	0,6065	0,4292	0,2135	0,9676	0,815	0,4797	0,2148
68	"rs2145272"	0,0978	0,156	0,9061	0,6733	0,321	0,8863	0,7844	0,5727
69	"rs2145998"	0,9313	0,8916	0,1522	0,6322	0,6219	0,4362	0,1228	0,5874
70	"rs2154319"	0,7776	0,577	0,5376	0,0869	0,5037	0,5023	0,7128	0,1568
71	"rs2237886"	0,8523	0,2249	0,2318	0,6977	0,1214	0,0809	0,4211	0,8962
72	"rs2247341"	0,9356	0,5527	0,6211	0,3838	0,4276	0,3457	0,7412	0,2688
73	"rs2256183"	0,2584	0,8879	0,4899	0,9323	0,0508	0,0883	0,3948	0,6714
74	"rs227724"	0,3928	0,2949	0,7068	0,9035	0,9388	0,4681	0,4678	0,9397
75	"rs2279008"	0,6025	0,7294	0,8978	0,1793	0,496	0,735	0,9915	0,1562
76	"rs2284746"	0,7008	0,5464	0,0829	0,8941	0,1984	0,3029	0,1142	0,5725
77	"rs2336725"	0,1670	0,6285	0,661	0,8445	0,1213	0,3594	0,6761	0,8765

78	"rs2341459"	0,3408	0,1714	0,9267	0,5991	0,7759	0,2247	0,6114	0,4954
79	"rs237743"	0,4342	0,5983	0,2952	0,3968	0,3907	0,5757	0,3518	0,3363
80	"rs2580816"	0,6991	0,794	0,1011	0,9567	0,3855	0,5736	0,1148	0,7142
81	"rs2597513"	0,9821	0,3228	0,6657	0,7662	0,2431	0,1444	0,9441	0,8917
82	"rs2629046"	0,4924	0,2541	0,4616	0,4523	0,622	0,2968	0,6655	0,324
83	"rs2638953"	0,5182	0,7467	0,9663	0,3305	0,7181	0,9985	0,919	0,3821
84	"rs2665838"	0,8759	0,6605	0,121	0,2419	0,9436	0,8897	0,1452	0,2832
85	"rs26868"	0,9724	0,4304	0,9628	0,4742	0,3843	0,2127	0,8887	0,3209
86	"rs274546"	0,2202	0,3984	0,9133	0,1674	0,2071	0,6549	0,8746	0,1103
87	"rs2778031"	0,5952	0,6889	0,7869	0,901	0,7426	0,9763	0,8572	0,9511
88	"rs2780226"	0,6068	0,6315	0,0834	0,3832	0,9908	0,9699	0,0864	0,328
89	"rs2834442"	0,3675	0,254	0,6662	0,6633	0,9201	0,4352	0,9181	0,5759
90	"rs2856321"	0,1227	0,3106	0,5217	0,5804	0,1732	0,6278	0,7091	0,4293
91	"rs2871865"	0,0556	0,2043	0,5952	0,2029	0,055	0,375	0,7993	0,0968
92	"rs310405"	0,2493	0,2235	0,0029	0,426	0,5142	0,7429	0,0063	0,459
93	"rs3110496"	0,1845	0,0783	0,4294	0,391	0,8411	0,3878	0,7975	0,4784
94	"rs3118905"	0,3744	0,599	0,763	0,6289	0,519	0,8687	0,8477	0,722
95	"rs3129109"	0,7022	0,1541	0,7157	0,3963	0,0098	0,0037	0,9036	0,109
96	"rs3764419"	0,5028	0,2241	0,1019	0,0685	0,9026	0,7465	0,204	0,0981
97	"rs3782089"	0,4697	0,4974	0,512	0,2292	0,923	0,7471	0,5462	0,2152
98	"rs3791675"	0,2479	0,9432	0,7762	0,7546	0,097	0,2495	0,8123	0,8924
99	"rs3812163"	0,7199	0,8111	0,2973	0,0424	0,8211	0,8203	0,2758	0,0364
100	"rs4072910"	0,0356	0,0136	0,5695	0,3588	0,5052	0,2854	0,942	0,5054
101	"rs42235"	0,2362	0,339	0,0145	0,6494	0,351	0,5726	0,0219	0,6176
102	"rs422421"	0,1031	0,2271	0,2091	0,4693	0,2113	0,9818	0,0918	0,3179
103	"rs425277"	0,2372	0,3592	0,1893	0,8228	0,449	0,7375	0,098	0,6843
104	"rs4282339"	0,5035	0,9078	0,886	0,3731	0,4922	0,6607	0,8581	0,488
105	"rs4470914"	0,6636	0,3884	0,6594	0,1309	0,0229	0,0121	0,4458	0,0314
106	"rs4601530"	0,1980	0,4162	0,5957	0,9301	0,2814	0,7153	0,7327	0,7769
107	"rs4605213"	0,8923	0,4754	0,6914	0,0476	0,1013	0,084	0,8589	0,0165
108	"rs4640244"	0,8485	0,7317	0,9125	0,0418	0,5876	0,4669	0,8994	0,0314
109	"rs4665736"	0,2385	0,2476	0,3892	0,793	0,7502	0,4768	0,2231	0,8228
110	"rs473902"	0,6100	0,8279	0,1707	0,9777	0,2689	0,5018	0,1682	0,8221
111	"rs4800452"	0,8140	0,6707	0,3752	0,409	0,9224	0,9971	0,446	0,4512
112	"rs4821083"	0,4646	0,4083	0,9393	0,4381	0,9228	0,4982	0,758	0,4
113	"rs494459"	0,9671	0,5656	0,292	0,6375	0,5808	0,3106	0,1823	0,7544
114	"rs4965598"	0,7650	0,9539	0,2484	0,9894	0,651	0,5907	0,2264	0,9499
115	"rs4986172"	0,7150	0,5791	0,5822	0,8405	0,9999	0,7826	0,6823	0,9044
116	"rs5017948"	0,5920	0,2156	0,4151	0,5694	0,6039	0,2949	0,6674	0,7437
117	"rs526896"	0,8842	0,46	0,7821	0,6168	0,4248	0,2802	0,9614	0,4465
118	"rs543650"	0,8442	0,9421	0,484	0,5282	0,6428	0,7464	0,4721	0,429

119	"rs572169"	0,1337	0,5305	0,8118	0,461	0,1674	0,5999	0,7062	0,7063
120	"rs5742915"	0,5513	0,0490	0,6798	0,8634	0,1878	0,0214	0,887	0,4662
121	"rs634552"	0,5910	0,9127	0,1002	0,9836	0,4584	0,4218	0,0864	0,9257
122	"rs6439167"	0,3039	0,5515	0,9347	0,7727	0,3264	0,6965	0,9499	0,6269
123	"rs6449353"	0,6568	0,3328	0,4954	0,6863	0,493	0,297	0,7258	0,4925
124	"rs6457620"	0,9084	0,8264	0,6219	0,75	0,6924	0,5925	0,537	0,8348
125	"rs6457821"	0,4170	0,5035	0,3855	0,3378	0,393	0,5739	0,5005	0,2512
126	"rs6470764"	0,5663	0,5541	0,1033	0,3638	0,9916	0,905	0,1184	0,3176
127	"rs6473015"	0,5060	0,7058	0,14	0,5331	0,4738	0,9735	0,0941	0,4061
128	"rs654723"	0,1256	0,7174	0,4024	0,9246	0,0452	0,2419	0,3804	0,5122
129	"rs6569648"	0,4499	0,0712	0,4419	0,2812	0,4302	0,051	0,1347	0,4734
130	"rs6684205"	0,7363	0,2654	0,1043	0,0283	0,1878	0,1272	0,1645	0,0092
131	"rs6699417"	0,9230	0,8506	0,7436	0,6536	0,9785	0,8796	0,6727	0,6599
132	"rs6714546"	0,7799	0,6443	0,8702	0,6193	0,2788	0,26	0,9958	0,4065
133	"rs6879260"	0,9650	0,5456	0,3838	0,2988	0,3485	0,3212	0,4681	0,1817
134	"rs6959212"	0,5064	0,882	0,5789	0,996	0,1818	0,2113	0,489	0,7276
135	"rs7027110"	0,2282	0,6921	0,4452	0,0364	0,0622	0,0771	0,2818	0,1213
136	"rs7112925"	0,2740	0,131	0,0595	0,5807	0,9499	0,5834	0,1401	0,7223
137	"rs7155279"	0,8715	0,593	0,7539	0,0829	0,9335	0,8732	0,9103	0,0982
138	"rs7178424"	0,8108	0,7142	0,9664	0,1509	0,7451	0,9518	0,8489	0,1464
139	"rs720390"	0,9032	0,3813	0,0568	0,0174	0,6041	0,6345	0,1111	0,0419
140	"rs724016"	0,0118	0,053	0,8209	0,2612	0,1921	0,75	0,787	0,3672
141	"rs7274811"	0,4202	0,4835	0,749	0,4553	0,0526	0,0487	0,4958	0,8383
142	"rs7319045"	0,1142	0,0984	0,5239	0,377	0,791	0,4089	0,8406	0,3732
143	"rs7332115"	0,5005	0,4184	0,107	0,3009	0,8966	0,7119	0,1404	0,2499
144	"rs7460090"	0,1354	0,4898	0,5326	0,3631	0,1965	0,5489	0,5852	0,5411
145	"rs7466269"	0,0203	0,0122	0,5517	0,0591	0,8686	0,1057	0,9293	0,0425
146	"rs7507204"	0,7629	0,2357	0,826	0,733	0,3978	0,1361	0,5195	0,9813
147	"rs751543"	0,7221	0,7525	0,7327	0,9657	0,8791	0,8608	0,6442	0,992
148	"rs7532866"	0,2324	0,6861	0,6289	0,2126	0,2885	0,5807	0,641	0,3209
149	"rs7567288"	0,3395	0,3655	0,2711	0,9778	0,6672	0,9345	0,396	0,9831
150	"rs7567851"	0,2469	0,1514	0,0874	0,4921	0,9103	0,2525	0,0264	0,5109
151	"rs7689420"	0,6349	0,7744	0,5656	0,2565	0,1658	0,1577	0,4832	0,1498
152	"rs7697556"	0,6404	0,4275	0,4149	0,2027	0,1836	0,1847	0,6143	0,4164
153	"rs7759938"	0,8058	0,7414	0,7134	0,3438	0,5797	0,4991	0,5792	0,4431
154	"rs7763064"	0,0397	0,0218	0,1651	0,2461	0,8952	0,2281	0,4301	0,1907
155	"rs7849585"	0,2078	0,1968	0,6307	0,5338	0,5755	0,5871	0,377	0,503
156	"rs7853377"	0,9019	0,4101	0,4555	0,3055	0,5686	0,2675	0,276	0,4186
157	"rs7864648"	0,7945	0,7701	0,3027	0,2163	0,7241	0,861	0,3597	0,2992
158	"rs788867"	0,9757	0,5825	0,7399	0,4306	0,5926	0,4237	0,5689	0,5521
159	"rs7909670"	0,6871	0,6684	0,5125	0,9235	0,9174	0,738	0,4148	0,9114

160	"rs7926971"	0,1100	0,249	0,6059	0,4657	0,1864	0,7139	0,839	0,3512
161	"rs7971536"	0,1105	0,2785	0,1616	0,2751	0,4142	0,6791	0,088	0,3919
162	"rs798489"	0,4635	0,0967	0,6038	0,3441	0,3059	0,0631	0,9805	0,1979
163	"rs8052560"	0,6075	0,7206	0,1496	0,3752	0,7746	0,8627	0,1406	0,3609
164	"rs806794"	0,6185	0,7754	0,9378	0,319	0,4285	0,464	0,8127	0,4442
165	"rs8181166"	0,1829	0,1003	0,9243	0,7768	0,8504	0,3145	0,5498	0,8323
166	"rs822552"	0,1451	0,4381	0,1953	0,9603	0,1486	0,3976	0,2407	0,6781
167	"rs862034"	0,8407	0,2613	0,9894	0,1833	0,4755	0,217	0,6586	0,3137
168	"rs889014"	0,1121	0,0571	0,1123	0,2015	0,5496	0,6382	0,299	0,2318
169	"rs891088"	0,8706	0,3732	0,5438	0,4266	0,5975	0,3923	0,7477	0,597
170	"rs9360921"	0,9517	0,4595	0,0093	0,2893	0,5849	0,6932	0,0154	0,4584
171	"rs9428104"	0,7936	0,911	0,7391	0,3196	0,5955	0,7592	0,6953	0,2687
172	"rs9456307"	0,4106	0,1735	0,3914	0,9304	0,6225	0,2874	0,642	0,8523
173	"rs9472414"	0,8197	0,3906	0,4888	0,5982	0,6046	0,4145	0,6792	0,7662
174	"rs955748"	0,9285	0,9422	0,9054	0,3384	0,9808	0,9966	0,8557	0,3506
175	"rs961764"	0,6328	0,092	0,6971	0,7439	0,2188	0,0467	0,8506	0,8715
176	"rs9835332"	0,0870	0,9773	0,2163	0,8436	0,0104	0,0995	0,2243	0,5961
177	"rs9844666"	0,6909	0,0292	0,4569	0,5283	0,0233	0,0019	0,9753	0,1488
178	"rs9863706"	0,6096	0,965	0,3905	0,1188	0,197	0,3567	0,3665	0,0583
179	"rs9967417"	0,0682	0,7699	0,3251	0,4776	0,0166	0,0726	0,2825	0,9237
180	"rs9969804"	0,9462	0,6568	0,5089	0,7949	0,5839	0,5646	0,6018	0,9574

Lisa 2: Kasutatud programmikood

```
#Loeme sisse geeniandmeid sisaldavad failid
tee="D:/HeightSNP/Lahti"
#oigest kaustast txt laiendiga failid.
failid=list.files(path = tee, pattern="[[:alnum:]].txt")
n=length(failid)
#loeme sisse esimese faili ja teemeneed R-le arusaadavale kujule
i=1
andmed1=read.table(file=paste(tee, "/", failid[i], sep=""), header=TRUE)
names(andmed1)=toupper(names(andmed1))
if(sum(names(andmed1)=="INFORMATION")==0) andmed1$INFORMATION=rep(NA, dim(andmed1)[1])
if(sum(names(andmed1)=="INFORMATION_TYPE")==0) andmed1$INFORMATION_TYPE=rep(NA, dim(andmed1)[1])
if(sum(names(andmed1)=="IMPUTATION_TYPE")==0) andmed1$IMPUTATION_TYPE=rep(NA, dim(andmed1)[1])
if(sum(names(andmed1)=="IMPUTATION")==0) andmed1$IMPUTATION=rep(NA, dim(andmed1)[1])
#paneme paika soo, faili nimes on enamasti olemas.
if (length(grep(".WOMEN.", toupper(failid[i])))>0) { naised=2
 }else if (length(grep(".MEN.", toupper(failid[i])))>0) {naised=1
 }else  { naised=3}
sugu=rep(c("mees","naine","paranda")[naised], dim(andmed1)[1])
jrk=rep(i, dim(andmed1)[1])
andmed1$SUGU=sugu
andmed1$JRK=jrk
andmed=andmed1
#Nuud loeme sisse ka ulejaanud andmed
for (i in 2:n){
and=read.table(file=paste(tee, "/", failid[i], sep=""), header=TRUE)
names(and)=toupper(names(and))
if(sum(names(and)=="INFORMATION")==0) and$INFORMATION=rep(NA, dim(and)[1])
if(sum(names(and)=="INFORMATION_TYPE")==0) and$INFORMATION_TYPE=rep(NA, dim(and)[1])
if(sum(names(and)=="IMPUTATION_TYPE")==0) and$IMPUTATION_TYPE=rep(NA, dim(and)[1])
if(sum(names(and)=="IMPUTATION")==0) and$IMPUTATION=rep(NA, dim(and)[1])
  if (length(grep(".WOMEN.", toupper(failid[i])))>0) { naised=2
 }else if (length(grep(".MEN.", toupper(failid[i])))>0) {naised=1
 }else  { naised=3}
  sugu=rep(c("naine","mees","paranda")[naised], dim(and)[1])
  jrk=rep(i, dim(and)[1])
  and$SUGU=sugu
  and$JRK=jrk
  andmed=rbind(andmed, and) }
andmed
dim(andmed)
names(andmed)
#Salvestame oma andmed
write.table(andmed, file="D:/Magistritoo/HeightSNP/andmed.txt", row.names=F,)
k=read.table("D:/HeightSNP/andmed.txt", header=T)
#Viskame valja mittesobivad andmed
indeks=(1:100)
v2lja=c(7,16,17,24,25,26,27,28,29,33,34,35,53,54,55,56,61:84,86,89,93,94,100)
length(v2lja)
indeks=indeks[-v2lja]
length(indeks)
uusandmestik=k[k$JRK %in% indeks,]
uusandmestik$JRK
#Uhel failil paistab sugu puudu olevat
uusandmestik$SUGU[uusandmestik$JRK==95]="naine"
uusandmestik$SUGU[uusandmestik$JRK==95]
```

```
#Sisse vaja lugeda ka uuringuid kirjeldavad andmed
teisedandmed=read.table("D:/Magistritoo/HeightSNP/laadimiseks3.txt", header=T, sep="\t", dec=",")
names(teisedandmed)=toupper(names(teisedandmed))
#Teised andmed on pikal kujul, meil vaja teha laiaks
uus=reshape(teisedandmed,idvar=c("A","NIMI", "JRK"), timevar="B" , direction="wide",)
#Paneme jarjekorranumbri alusel jarjekorda
sorditud=uusandmestik[order(uusandmestik$JRK),]
sordituduus=uus[order(uus$JRK),]
#uuringu andmed on vaja paljundada iga markeri kohta
indeks2=1:55
abi=as.vector(table(sorditud$JRK))
indeksid=rep(indeks2,abi)
lisamiseks=sordituduus[indeksid,]
#Uhendame andmestikud
valmis=cbind(sorditud, lisamiseks)
#Nuud on vaja koigil andmetel strand plussiks teha
valmis2=valmis
abi2=(valmis2$STRAND=="-")&(valmis2$EFFECT_ALLELE=="A")&(valmis2$OTHER_ALLELE=="C")
valmis2$STRAND[abi2]="+"
valmis2$EFFECT_ALLELE[abi2]="T"
valmis2$OTHER_ALLELE[abi2]="G"
abi3=(valmis2$STRAND=="-")&(valmis2$EFFECT_ALLELE=="T")&(valmis2$OTHER_ALLELE=="C")
valmis2$STRAND[abi3]="+"
valmis2$EFFECT_ALLELE[abi3]="A"
valmis2$OTHER_ALLELE[abi3]="G"
abi4=(valmis2$STRAND=="-")&(valmis2$EFFECT_ALLELE=="A")&(valmis2$OTHER_ALLELE=="G")
valmis2$STRAND[abi4]="+"
valmis2$EFFECT_ALLELE[abi4]="T"
valmis2$OTHER_ALLELE[abi4]="C"
abi5=(valmis2$STRAND=="-")&(valmis2$EFFECT_ALLELE=="T")&(valmis2$OTHER_ALLELE=="G")
valmis2$STRAND[abi5]="+"
valmis2$EFFECT_ALLELE[abi5]="A"
valmis2$OTHER_ALLELE[abi5]="C"
abi6=(valmis2$STRAND=="-")&(valmis2$EFFECT_ALLELE=="C")&(valmis2$OTHER_ALLELE=="A")
valmis2$STRAND[abi6]="+"
valmis2$EFFECT_ALLELE[abi6]="G"
valmis2$OTHER_ALLELE[abi6]="T"
abi7=(valmis2$STRAND=="-")&(valmis2$EFFECT_ALLELE=="G")&(valmis2$OTHER_ALLELE=="A")
valmis2$STRAND[abi7]="+"
valmis2$EFFECT_ALLELE[abi7]="C"
valmis2$OTHER_ALLELE[abi7]="T"
abi8=(valmis2$STRAND=="-")&(valmis2$EFFECT_ALLELE=="C")&(valmis2$OTHER_ALLELE=="T")
valmis2$STRAND[abi8]="+"
valmis2$EFFECT_ALLELE[abi8]="G"
valmis2$OTHER_ALLELE[abi8]="A"
abi9=(valmis2$STRAND=="-")&(valmis2$EFFECT_ALLELE=="G")&(valmis2$OTHER_ALLELE=="T")
valmis2$STRAND[abi9]="+"
valmis2$EFFECT_ALLELE[abi9]="C"
valmis2$OTHER_ALLELE[abi9]="A"
abi10=(valmis2$STRAND=="-")&(valmis2$EFFECT_ALLELE=="A")&(valmis2$OTHER_ALLELE=="T")
valmis2$STRAND[abi10]="+"
valmis2$EFFECT_ALLELE[abi10]="T"
valmis2$OTHER_ALLELE[abi10]="A"
abi11=(valmis2$STRAND=="-")&(valmis2$EFFECT_ALLELE=="T")&(valmis2$OTHER_ALLELE=="A")
valmis2$STRAND[abi11]="+"
valmis2$EFFECT_ALLELE[abi11]="A"
valmis2$OTHER_ALLELE[abi11]="T"
abi12=(valmis2$STRAND=="-")&(valmis2$EFFECT_ALLELE=="C")&(valmis2$OTHER_ALLELE=="G")
valmis2$STRAND[abi12]="+"
```

```
valmis2$EFFECT_ALLELE[abi12]="G"
valmis2$OTHER_ALLELE[abi12]="C"
abi13=(valmis2$STRAND=="-")&(valmis2$EFFECT_ALLELE=="G")&(valmis2$OTHER_ALLELE=="C")
valmis2$STRAND[abi13]="+"
valmis2$EFFECT_ALLELE[abi13]="C"
valmis2$OTHER_ALLELE[abi13]="G"
#Salvestame oma andmed
write.table(valmis, file="D:/HeightSNP/valmis.txt", row.names=F,)
write.table(valmis2, file="D:/HeightSNP/valmis2.txt", row.names=F,)
#Analuusiks andmed sisse:
valmis2=read.table("D:/Magistritoo/HeightSNP/valmis2.txt", header=T)
table(valmis2$SUGU)
#Tuleb valja, et uks sugu on puudulik, selle viskame valja
indeksalles=!valmis2$SUGU=="paranda"
valmis3=valmis2[indeksalles,]
#votame valja markerite nimekirja
markerid=names(table(valmis3$MARKERNAME))
#Defineerime funktsiooni efektide korrastamiseks
efektkorda=function(andmestik){
teineandmestik=andmestik
abi=!andmestik$EFFECT_ALLELE==andmestik$EFFECT_ALLELE[1]
teineandmestik$EFFECT_ALLELE[abi]=andmestik$EFFECT_ALLELE[1]
teineandmestik$OTHER_ALLELE[abi]=andmestik$OTHER_ALLELE[1]
teineandmestik$BETA[abi]=-andmestik$BETA[abi]
return(teineandmestik)}
#Uuriks efektidevahelisi korrelatsioone
korrelatsioonid=matrix(rep(NA,180*180),180,180)
for (i in 1:180){
andmed=valmis3[valmis3$MARKERNAME==markerid[i],]
#Teeme nii, et koigil oleks efekt sama alleeli suhtes moodetud
andmed2=efektkorda(andmed)
for(j in 1:180){
andmed3=valmis3[valmis3$MARKERNAME==markerid[j],]
andmed4=efektkorda(andmed3)
#puuduvad andmed valja
andmed5=andmed2[andmed2$JRK %in% andmed4$JRK,]
andmed6=andmed4[andmed4$JRK %in% andmed5$JRK,]
korrelatsioonid[i,j]=cor(andmed5$BETA,andmed6$BETA)}}
#salvestame ka need tulemused
write.table(korrelatsioonid, file="D:/Magistritoo/HeightSNP/korrelatsioonid.txt", row.names=F,)
#uritame illustreerida saadud tulemusi, graafiku ilusaks teha
abiasi=heatmap(korrelatsioonid)
abiasi
u=rev(abiasi$colInd)
uu=korrelatsioonid[,u]
uu=uu[abiasi$rowInd,]
heatmap(uu,Rowv=NA, Colv=NA, marg=c(0,0))
#####
#Leiame p-vaartused tunnustele sugu, keskmine pikkus, keskmine kaal ja keskmine vanus
pv22rtused=matrix(rep(NA, 180*4),180,4)
for (i in 1:180){
#votame esimese uuringufaili
andmed= valmis3[valmis3$MARKERNAME==markerid[i],]
#Teeme nii, et koigil oleks efekt samapidine
andmed2=efektkorda(andmed)
# votame valja Y tunnuse
Y=andmed2$BETA
#moodustame mudelimaatriksi tunnusega sugu
mudel=lm(Y~andmed2$SUGU, data=andmed2)
```

```
X=model.matrix(mudel)
#lihtsama variandi kovariatsioonimaatriks:
V1=diag((andmed2$SE)**2)
#lihtsama variandi logaritmilise toeparafunktsiooni vaartus:
c1=unlist(determinant(2*pi*V1))[1]
beta1=solve(t(X)%*%solve(V1)%*%X)%*%t(X)%*%solve(V1)%*%Y
toep1=-1/2*(c1)-(1/2)*t(Y-X%*%beta1)%*%solve(V1)%*%(Y-X%*%beta1)
#leiame ka teiste tunnuste jaoks toeparafunktisoonide vaartused
mudel3=lm(Y~andmed2$MEAN.Height)
X3=model.matrix(mudel3)
beta3=solve(t(X3)%*%solve(V1)%*%X3)%*%t(X3)%*%solve(V1)%*%Y
toep3=-1/2*(c1)-(1/2)*t(Y-X3%*%beta3)%*%solve(V1)%*%(Y-X3%*%beta3)
mudel4=lm(Y~andmed2$MEAN.Weight)
X4=model.matrix(mudel4)
beta4=solve(t(X4)%*%solve(V1)%*%X4)%*%t(X4)%*%solve(V1)%*%Y
toep4=-1/2*(c1)-(1/2)*t(Y-X4%*%beta4)%*%solve(V1)%*%(Y-X4%*%beta4)
mudel5=lm(Y~andmed2$MEAN.Age)
X5=model.matrix(mudel5)
beta5=solve(t(X5)%*%solve(V1)%*%X5)%*%t(X5)%*%solve(V1)%*%Y
toep5=-1/2*(c1)-(1/2)*t(Y-X5%*%beta5)%*%solve(V1)%*%(Y-X5%*%beta5)
# veel on vaja nullmudeli toeparafunktsiooni vaartust
mudel2=lm(Y~1)
X2=model.matrix(mudel2)
beta2=solve(t(X2)%*%solve(V1)%*%X2)%*%t(X2)%*%solve(V1)%*%Y
toep2=-1/2*(c1)-(1/2)*t(Y-X2%*%beta2)%*%solve(V1)%*%(Y-X2%*%beta2)
#toeparasuhtetestid ja pvaartused
teststatistik1=2*(toep1-toep2)
teststatistik3=2*(toep3-toep2)
teststatistik4=2*(toep4-toep2)
teststatistik5=2*(toep5-toep2)
p1=1-pchisq(teststatistik1,df=1)
p3=1-pchisq(teststatistik3,df=1)
p4=1-pchisq(teststatistik4,df=1)
p5=1-pchisq(teststatistik5,df=1)
?pchisq
pv22rtused[i,1]=p1
pv22rtused[i,2]=p3
pv22rtused[i,3]=p4
pv22rtused[i,4]=p5}
round(pv22rtused,4)
sum(pv22rtused[,1]<0.05)
sum(pv22rtused[,2]<0.05)
sum(pv22rtused[,3]<0.05)
sum(pv22rtused[,4]<0.05)
######
#Teeme sarnase asja nii, et vordleme mudeleid, kus tervet mudelit vordleme sellega
#kust uks on valja visatud
pv22rtused2=matrix(rep(NA, 180*4),180,4)
for (i in 1:180){
andmed= valmis3[valmis3$MARKERNAME==markerid[i],]
#Teeme nii, et koigil oleks efekt samapidine
andmed2=efektkorda(andmed)
# votame valja Y tunnuse
Y=andmed2$BETA
#moodustame mudelimaatriksi koiki tunnuseid sisaldavale maatriksile
mudel=lm(Y~andmed2$SUGU+andmed2$MEAN.Height+andmed2$MEAN.Weight+andmed2$MEAN.Age)
X=model.matrix(mudel)
#lihtsama variandi kovariatsioonimaatriks:
V1=diag((andmed2$SE)**2)
```

```
#lihtsama variandi logaritmilise toeparafunktsiooni vaartus:
c1=unlist(determinant(2*pi*V1))[1]
beta1=solve(t(X)%*%solve(V1)%*%X)%*%t(X)%*%solve(V1)%*%Y
toep1=-1/2*(c1)-(1/2)*t(Y-X%*%beta1)%*%solve(V1)%*%(Y-X%*%beta1)
#Nuud leiame toeparafunktsioonide vaartused juhul kus uks tunnus on valja visatud
mudel2=lm(Y~andmed2$MEAN.Height+andmed2$MEAN.Weight+andmed2$MEAN.Age)
X2=model.matrix(mudel2)
beta2=solve(t(X2)%*%solve(V1)%*%X2)%*%t(X2)%*%solve(V1)%*%Y
toep2=-1/2*(c1)-(1/2)*t(Y-X2%*%beta2)%*%solve(V1)%*%(Y-X2%*%beta2)
mudel3=lm(Y~andmed2$SUGU+andmed2$MEAN.Weight+andmed2$MEAN.Age)
X3=model.matrix(mudel3)
beta3=solve(t(X3)%*%solve(V1)%*%X3)%*%t(X3)%*%solve(V1)%*%Y
toep3=-1/2*(c1)-(1/2)*t(Y-X3%*%beta3)%*%solve(V1)%*%(Y-X3%*%beta3)
mudel4=lm(Y~andmed2$SUGU+andmed2$MEAN.Height+andmed2$MEAN.Age)
X4=model.matrix(mudel4)
beta4=solve(t(X4)%*%solve(V1)%*%X4)%*%t(X4)%*%solve(V1)%*%Y
toep4=-1/2*(c1)-(1/2)*t(Y-X4%*%beta4)%*%solve(V1)%*%(Y-X4%*%beta4)
mudel5=lm(Y~andmed2$SUGU+andmed2$MEAN.Height+andmed2$MEAN.Weight)
X5=model.matrix(mudel5)
beta5=solve(t(X5)%*%solve(V1)%*%X5)%*%t(X5)%*%solve(V1)%*%Y
toep5=-1/2*(c1)-(1/2)*t(Y-X5%*%beta5)%*%solve(V1)%*%(Y-X5%*%beta5)
#leiame p-vaartused
teststatistik2=-2*(toep2-toep1)
teststatistik3=-2*(toep3-toep1)
teststatistik4=-2*(toep4-toep1)
teststatistik5=-2*(toep5-toep1)
p2=1-pchisq(teststatistik2,df=1)
p3=1-pchisq(teststatistik3,df=1)
p4=1-pchisq(teststatistik4,df=1)
p5=1-pchisq(teststatistik5,df=1)
pv22rtused2[i,1]=p2
pv22rtused2[i,2]=p3
pv22rtused2[i,3]=p4
pv22rtused2[i,4]=p5}
round(pv22rtused2,4)
sum(pv22rtused2[,1]<0.05)
sum(pv22rtused2[,2]<0.05)
sum(pv22rtused2[,3]<0.05)
sum(pv22rtused2[,4]<0.05)
### Testime heterogeenust Cochrani Q abil.
n=length(markerid)
Qd=rep(NA,180)
pv=rep(NA,180)
for (i in 1:n){
#votame uhe marekeri jagu andmeid
andmed=valmis3[valmis3$MARKERNAME==markerid[i],]
#Teeme nii, et koigil oleks efekt samapidine
andmed2=efektkorda(andmed)
#Votame valja oma Y tunnuse
Y=andmed2$BETA
#moodustame mudelimaatriksi
mudel=lm(Y~1)
X=model.matrix(mudel)
# kovariatsioonimaatriks:
V1=diag((andmed2$SE)**2)
#vajame hinnangut parameetrile teeta
c1=unlist(determinant(2*pi*V1))[1]
beta1=solve(t(X)%*%solve(V1)%*%X)%*%t(X)%*%solve(V1)%*%Y
teeta=beta1
```

```
oomega=1/(andmed2$SE**2)
#mitu uuringut on
K=length(Y)
#arvutame statistiku
Q=sum(oomega*(Y-teeta)**2)
pv[i]=1-pchisq(Q,df=K-1)
Qd[i]=Q}
sum(pv<0.05)
ind=pv<0.05
round(pv[ind],4)
# Puuame teha korrelatsioonimaatriksi nende 11 valitud markeri
# toeliste efektide kohta
valitudmarkerid=markerid[ind]
valitudQ=Qd[ind]
maatriks= matrix(rep(NA,11*11),11,11)
hinnatudtaud=matrix(rep(NA,11*11),11,11)
# Puuan leida SNPde vahelised korrelatsioonid. Selleks leian esmalt
#parameetrihinnangud
for ( i in 1:11){
andmed= valmis3[valmis3$MARKERNAME==valitudmarkerid[i],]
#Teeme nii, et koigil oleks efekt samapidine
andmed2=efektkorda(andmed)
#uuringute arv ja ja hinnatud dispersioonide poordvaartused
r1=length(andmed2$BETA)
oomega1=1/((andmed2$SE)**2)
# votame X tunnuseks uhe teise SNP
for (j in 1:11){
andmed3=valmis3[valmis3$MARKERNAME==valitudmarkerid[j],]
andmed4=efektkorda(andmed3)
r2=length(andmed4$BETA)
oomega2=1/((andmed4$SE)**2)
#Nuud oleks vaja Y ja seletav SNP uhepikkusteks ka veel teha.
andmed5=andmed2[andmed2$JRK %in% andmed4$JRK,]
andmed6=andmed4[andmed4$JRK %in% andmed5$JRK,]
#Votame valja oma Y tunnuse
Y=andmed5$BETA
seletav=andmed6$BETA
DXhinnatud=var(seletav)
#keerulisema variandi kovariatsioonimaatriks:
V1=diag((andmed5$SE)**2)
mudel=lm(Y~seletav)
X=model.matrix(mudel)
d=length(andmed5$JRK)
#funktsioon tau hindamiseks STP meetodil
f=function(tauruut){
V2=V1+diag(rep(tauruut,d))
c2=unlist(determinant(2*pi*V2))[1]
beta2=solve(t(X)%*%solve(V2)%*%X)%*%t(X)%*%solve(V2)%*%Y
v22rtus=-1/2*(c2)-(1/2)*t(Y-X%*%beta2)%*%solve(V2)%*%(Y-X%*%beta2)
v22rtus}
#leiame maksimumi
m=optimize(f, lower=0, upper=1000, maximum=T)
#maksimaalne toeparafunktsiooni vaartus
toep2=m$objective
#maksimumpunkti tau
hinnatudtaud[i,j]=m$maximum
#hindame
#arvutame tegeliku SNP efekti dispersiooni hinnangu
DYteg=(valitudQ[i]-(r1-1))/(sum(oomega1)-sum(oomega1**2)/sum(oomega1))
```

```
DXteg=(valitudQ[j]-(r2-1))/(sum(oomega2)-sum(oomega2**2)/sum(oomega2))
hinnang2=solve(t(X)%*%solve(V1+diag(rep(m$maximum,d)))%*%X)%*%t(X)%*%solve(V1+diag(rep(m$maximum,d)))%*%Y
#korrelatsioonid
maatriks[i,j]=hinnang2[2]*DXteg/(sqrt(DYteg)*sqrt(DXteg))}}
#uus pakett sisse
#install.packages("corrplot")
library(corrplot)
colnames(maatriks)=valitudmarkerid
rownames(maatriks)=valitudmarkerid
corrplot(maatriks,method="color", addcolorlabel="b")
#Leiame tavalised korrelatsioonid ka
korrelatsioonid2=matrix(rep(NA,11*11),11,11)
for (i in 1:11){
andmed=valmis3[valmis3$MARKERNAME==valitudmarkerid[i],]
#Teeme nii, et koigil oleks efekt samapidine
andmed2=efektkorda(andmed)
for(j in 1:11){
andmed3=valmis3[valmis3$MARKERNAME==valitudmarkerid[j],]
andmed4=efektkorda(andmed3)
andmed5=andmed2[andmed2$JRK %in% andmed4$JRK,]
andmed6=andmed4[andmed4$JRK %in% andmed5$JRK,]
korrelatsioonid2[i,j]=cor(andmed5$BETA,andmed6$BETA)}}
colnames(korrelatsioonid2)=valitudmarkerid
rownames(korrelatsioonid2)=valitudmarkerid
corrplot(korrelatsioonid2, method="color")
####Uurime markerit nr 19 lahemalt
andmed= valmis3[valmis3$MARKERNAME==markerid[19],]
markerid
#Teeme nii, et koigil oleks efekt samapidine
andmed2=efektkorda(andmed)
Y=andmed2$BETA
mudel=lm(Y~andmed2$SUGU)
X=model.matrix(mudel)
#lihtsama variandi kovariatsioonimaatriks:
V1=diag((andmed2$SE)**2)
#lihtsama variandi logaritmilise toeparafunktsiooni vaartus:
c1=unlist(determinant(2*pi*V1))[1]
beta1=solve(t(X)%*%solve(V1)%*%X)%*%t(X)%*%solve(V1)%*%Y
toep1=-1/2*(c1)-(1/2)*t(Y-X%*%beta1)%*%solve(V1)%*%(Y-X%*%beta1)
#lihtsama variandi parameetrite  hinnang ja selle dispersioon
hinnang1=beta1
#hinnangudisp1=solve(t(X)%*%solve(V1)%*%X)
#Nuud sugu ja kaalu sisaldav mudel
mudel2=lm(Y~andmed2$SUGU+andmed2$MEAN.Weight)
X2=model.matrix(mudel2)
beta2=solve(t(X2)%*%solve(V1)%*%X2)%*%t(X2)%*%solve(V1)%*%Y
toep2=-1/2*(c1)-(1/2)*t(Y-X2%*%beta2)%*%solve(V1)%*%(Y-X2%*%beta2)
hinnang2=beta2
teststatistik=2*(toep2-toep1)
#kas kaal on oluline lisaks soole
p=1-pchisq(teststatistik, df=1)
p
hinnang2
#Nuud lihtsalt kaalu sisaldav mudel
mudel3=lm(Y~andmed2$MEAN.Weight)
X3=model.matrix(mudel3)
beta3=solve(t(X3)%*%solve(V1)%*%X3)%*%t(X3)%*%solve(V1)%*%Y
toep3=-1/2*(c1)-(1/2)*t(Y-X3%*%beta3)%*%solve(V1)%*%(Y-X3%*%beta3)
hinnang3=beta3
```

```
teststatistik=2*(toep2-toep3)
p=1-pchisq(teststatistik,df=1)
#kas sugu on oluline
p
sugu=factor(andmed2$SUGU)
plot(sugu,Y)
sss=andmed2$SE
plot(andmed2$MEAN.Weight,Y,ylim=c(-0.4, 0.2),xlab="Keskmine kaal",ylab="y")
arrows(andmed2$MEAN.Weight, Y-sss, andmed2$MEAN.Weight, Y+sss, length=0.05, angle=90, code=3, col="gray70")
abline(coef=beta3[,1], lwd=2, col="blue3")
points(andmed2$MEAN.Weight,Y, pch=20, col=sugu)
legend(130, -0.25, c("mehed", "naised"), col = c(2, 1), pch = c(20, 20))
par(mfrow=c(2,1))
hist(andmed2$BETA[sugu=="naine"], prob=T, main="", xlab="Naistel", ylab="Tihedus", ylim=c(0,8))
rug(andmed2$BETA[sugu=="naine"], ticksize=0.03, lwd=2)
hist(andmed2$BETA[sugu=="mees"], prob=T, main="", xlab="Meestel", ylab="Tihedus", ylim=c(0,8))
rug(andmed2$BETA[sugu=="mees"],ticksize=0.03, lwd=2)
###Prooviks uhe uuringu valja visata
andmed= valmis3[valmis3$MARKERNAME==markerid[19],]
#Teeme nii, et koigil oleks efekt samapidine
andmed2=efektkorda(andmed)
andmed3=andmed2[andmed2$MEAN.Weight<140,]
Y=andmed3$BETA
mudel=lm(Y~andmed3$SUGU)
X=model.matrix(mudel)
#lihtsama variandi kovariatsioonimaatriks:
V1=diag((andmed3$SE)**2)
#lihtsama variandi logaritmilise toeparafunktsiooni vaartus:
c1=unlist(determinant(2*pi*V1))[1]
beta1=solve(t(X)%*%solve(V1)%*%X)%*%t(X)%*%solve(V1)%*%Y
toep1=-1/2*(c1)-(1/2)*t(Y-X%*%beta1)%*%solve(V1)%*%(Y-X%*%beta1)
#Nuud
mudel2=lm(Y~andmed3$SUGU+andmed3$MEAN.Weight)
X2=model.matrix(mudel2)
beta2=solve(t(X2)%*%solve(V1)%*%X2)%*%t(X2)%*%solve(V1)%*%Y
toep2=-1/2*(c1)-(1/2)*t(Y-X2%*%beta2)%*%solve(V1)%*%(Y-X2%*%beta2)
hinnang2=beta2
teststatistik=2*(toep2-toep1)
p=1-pchisq(teststatistik, df=1)
p
hinnang2
#Nuud
mudel3=lm(Y~andmed3$MEAN.Weight)
X3=model.matrix(mudel3)
beta3=solve(t(X3)%*%solve(V1)%*%X3)%*%t(X3)%*%solve(V1)%*%Y
toep3=-1/2*(c1)-(1/2)*t(Y-X3%*%beta3)%*%solve(V1)%*%(Y-X3%*%beta3)
hinnang3=beta3
teststatistik=2*(toep2-toep3)
p=1-pchisq(teststatistik,df=1)
p
####Uurime markerit nr 177 lahemalt
andmed= valmis3[valmis3$MARKERNAME==markerid[177],]
#Teeme nii, et koigil oleks efekt samapidine
andmed2=efektkorda(andmed)
Y=andmed2$BETA
mudel=lm(Y~andmed2$SUGU)
X=model.matrix(mudel)
#lihtsama variandi kovariatsioonimaatriks:
V1=diag((andmed2$SE)**2)
```

```
#lihtsama variandi logaritmilise toeparafunktsiooni vaartus:
c1=unlist(determinant(2*pi*V1))[1]
beta1=solve(t(X)%*%solve(V1)%*%X)%*%t(X)%*%solve(V1)%*%Y
toep1=-1/2*(c1)-(1/2)*t(Y-X%*%beta1)%*%solve(V1)%*%(Y-X%*%beta1)
#lihtsama variandi parameetrite  hinnang ja selle dispersioon
#hinnang1=beta1
#hinnangudisp1=solve(t(X)%*%solve(V1)%*%X)
#Nuud mudel, kus nii sugu kui pikkus
mudel2=lm(Y~andmed2$SUGU+andmed2$MEAN.Height)
X2=model.matrix(mudel2)
beta2=solve(t(X2)%*%solve(V1)%*%X2)%*%t(X2)%*%solve(V1)%*%Y
toep2=-1/2*(c1)-(1/2)*t(Y-X2%*%beta2)%*%solve(V1)%*%(Y-X2%*%beta2)
hinnang2=beta2
teststatistik=2*(toep2-toep1)
p=1-pchisq(teststatistik, df=1)
p
hinnang2
#Nuud mudel, kus vaid pikkus
mudel3=lm(Y~andmed2$MEAN.Height)
X3=model.matrix(mudel3)
beta3=solve(t(X3)%*%solve(V1)%*%X3)%*%t(X3)%*%solve(V1)%*%Y
toep3=-1/2*(c1)-(1/2)*t(Y-X3%*%beta3)%*%solve(V1)%*%(Y-X3%*%beta3)
hinnang3=beta3
beta3
teststatistik=2*(toep2-toep3)
p=1-pchisq(teststatistik,df=1)
p
mudel4=lm(Y~andmed2$SUGU+andmed2$MEAN.Height+andmed2$MEAN.Age)
X4=model.matrix(mudel4)
beta4=solve(t(X4)%*%solve(V1)%*%X4)%*%t(X4)%*%solve(V1)%*%Y
toep4=-1/2*(c1)-(1/2)*t(Y-X4%*%beta4)%*%solve(V1)%*%(Y-X4%*%beta4)
hinnang4=beta2
teststatistik=2*(toep4-toep2)
p=1-pchisq(teststatistik, df=1)
p
hinnang2
sugu=factor(andmed2$SUGU)
plot(sugu,Y)
sss=andmed2$SE
plot(andmed2$MEAN.Height,Y,ylim=c(-0.3, 0.5),xlab="Keskmine pikkus", main="", ylab="y")
arrows(andmed2$MEAN.Height, Y-sss, andmed2$MEAN.Height, Y+sss, length=0.05, angle=90, code=3, col="gray70")
koef=c(beta2[1,1], beta2[2,1]*1+beta2[3,1])
koef2=c(beta2[1,1], beta2[2,1]*0+beta2[3,1])
abline(coef=koef, lwd=2, col="red")
abline(coef=koef2, lwd=2, col="black")
points(andmed2$MEAN.Height,Y, pch=20, col=sugu)
legend(1.6, 0.45, c("mehed", "naised"), col = c(2, 1), pch = c(20, 20))
par(mfrow=c(2,1))
hist(andmed2$BETA[sugu=="naine"], prob=T, main="", xlab="Naistel", xlim=c(-0.1,0.3), ylim=c(0,8), ylab="Tihedus")
rug(andmed2$BETA[sugu=="naine"], ticksize=0.03, lwd=2)
hist(andmed2$BETA[sugu=="mees"], prob=T, main="", xlab="Meestel",ylim=c(0,8), ylab="Tihedus")
rug(andmed2$BETA[sugu=="mees"],ticksize=0.03, lwd=2)
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Riho Klement

(sünnikuupäev: 25.06.1989)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
Geneetiliste markerite mõju muutlikkust kirjeldavad meta-analüüsi mudelid,

mille juhendaja on Märt Möls,

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, 20.05.2013