Sukulaisuussuhteiden vaikutukset tyypin 1 ja 2 virheisiin ja niiden minimointi genominlaajuisissa assosiaatiokartoituksissa

Tilastotieteen pro gradu -tutkielma

Mikko Väänänen
Matematiikan ja tilastotieteen laitos
Jyväskylän yliopisto
11. marraskuuta 2023

JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Väänänen, **Mikko**: Sukulaisuussuhteiden vaikutukset tyypin 1 ja 2 virheisiin ja niiden minimointi genominlaajuisissa assosiaatiokartoituksissa

Tilastotieteen pro gradu -tutkielma, 32 sivua, 3 liitettä (13 sivua)
11. marraskuuta 2023

Tiivistelmä

Genominlaajuisilla assosiaatiokartoituksilla (*genome-wide association studies*, GWAS) tutkitaan assosiaatioita geno- ja fenotyyppien välillä genominlaajuisesti tilastollisin menetelmin. GWAS-tutkimusten tuloksille on monia sovelluskohteita esimerkiksi lääkekehityksessä, periytyvyyden tutkimisessa ja geneettisten terveysriskien arvioimisessa. GWAS on tutkimusalana kuitenkin nuori ja kehittyy nopeasti, sillä vasta 2000-luvun aikana merkittävästi kehittyneet DNA:n luentamenetelmät ovat mahdollistaneet riittävän suuret tutkimusaineistot assosiaatioiden tutkimiselle genominlaajuisesti.

Yksilöiden välisten sukulaisuussuhteiden on havaittu vääristävän assosiaatiotestien tuloksia GWAS-tutkimuksissa, minkä vuoksi viime vuosina on pyritty kehittämään toimivia menetelmiä vääristymien korjaamiseksi. Tässä tutkielmassa tutkitaan, miten hyvin lineaarinen sekamalli (LMM), lineaarinen regressiomalli, jossa käytetään genomitiedoista muodostettuja pääkomponentteja kovariaatteina (LMP) ja lineaarinen regressiomalli (LM) huomioivat sukulaisuussuhteita assosiaatiotesteissä. Menetelmien toimivuutta arvioidaan sillä, kuinka vähän kukin menetelmä tuottaa tyypin 1 ja 2 virheitä. Menetelmien tuottamien tyypin 1 ja 2 virheiden määriä arvioitiin soveltamalla menetelmiä simuloituihin aineistoihin.

Tutkielman tulosten mukaan LMM- ja LMP-menetelmät korjaavat sukulaisuussuhteiden aiheuttamia tyypin 1 ja 2 virheitä hyvin verrattuna LM-menetelmään. LMM ei tuottanut simuloinnissa yhtään tyypin 1 virhettä ja huomattavasti vähemmän tyypin 2 virheitä kuin LM. Hyvin valitulla kovariaatteina käytettävien pääkomponenttien määrällä LMP oli LMM-menetelmän tasoa tyypin 1 ja 2 virheiden määrällä mitattuna.

Avainsanat: GWAS, sekamalli, regressiomalli, pääkomponentti, SNP, simulointi, tyypin 1 virhe, tyypin 2 virhe

Sisällys

1 Johdanto

Genominlaajuiset assosiaatiokartoitukset (*genome-wide association studies*, GWAS) pyrkivät löytämään assosiaatioita eliöiden geno- ja fenotyyppien väliltä (Uffelmann ym. 2021). Genotyypillä tarkoitetaan yksilön geneettistä rakennetta (Torres-Duque, Garcia-Rodriguez, ja Gonzalez-Garcia 2016) ja fenotyypillä yksilön genotyypin havaittavaa ilmentymää (Wojczynski ja Tiwari 2008). Yleisesti GWAS-tutkimus tarkoittaa nimensä mukaisesti, että assosiaatioita tutkitaan genomin eli eliön koko perimäaineksen laajuisesti. Tässä tutkielmassa rajaudutaan kuitenkin vain ihmisiä käsitteleviin GWAS-tutkimuksiin, joissa tutkitaan assosiaatioita yksittäisten snippien (*single nucleotide polymorfism*, SNP) ja fenotyypin välillä. Snipeistä kerrotaan tarkemmin luvussa 2.2.

Aineiston yksilöiden väliset sukulaisuussuhteet aiheuttavat usein vääristymiä assosiaatiotestien tuloksiin GWAS-tutkimuksissa (Chen ym. 2016). Esimerkiksi Sul, Martin ja Eskin (2018) havaitsivat sukulaisuussuhteiden kasvattavan tyypin 1 virheiden määrää assosiaatiotesteissä. Sukulaisuussuhteet voidaan jakaa kryptisiin sukulaisiin ja geneettisiin populaatioihin. Kryptisellä sukulaisuudella tarkoitetaan, että otoksessa on yksilöitä, jotka ovat läheistä sukua toisilleen, mutta sukulaisuussuhteet eivät ole tutkijan tiedossa, ja geneettisellä populaatiolla ryhmää, jonka yksilöt ovat kaukaisen syntyperänsä takia sukua toisilleen (Sul, Martin, ja Eskin 2018; Astle ja Balding 2009). Kun aineistoon kerätään yksilöitä eri geneettisistä populaatioista, muodostuu aineistoon geneettisiä populaatiorakenteita eli ryhmiä, joiden yksilöt ovat geneettisesti samankaltaisempia ryhmien sisällä kuin ryhmien välillä. Otoksessa voi siis olla yksilöitä, jotka ovat joko kaukaista tai läheistä sukua toisilleen, ja sukulaisuussuhteiden huomiotta jättäminen aiheuttaa usein vääristyneitä tuloksia assosiaatioiden testaamisessa geno- ja fenotyyppien välillä.

Tämän tutkielman tarkoitus on tutkia miten kolme eri tilastollista menetelmää ottavat huomioon kryptiset sukulaisuussuhteet ja geneettiset populaatiot testattaessa assosiaatiota geno- ja fenotyyppien välillä. Tutkittavat menetelmät ovat lineaarinen sekamalli (*linear mixed model*, LMM), lineaarinen regressiomalli, jossa käytetään yksilöiden genomitiedoista muodostettuja pääkomponentteja kovariaatteina (*linear model with principal components*, LMP) ja lineaarinen regressiomalli (*linear regression model*, LM). Menetelmien toimivuutta tutkitaan simuloimalla geneettisiä populaatiorakenteita ja kryptisiä sukulaisia sisältäviä geno-fenotyyppiaineistoja, ja arvioimalla miten hyvin nämä menetelmät minimoivat tyypin 1 ja 2 virheitä assosiaatiotesteissä. Geno-fenotyyppiaineistolla tarkoitetaan aineistoa, joka sisältää tutkittavan fenotyypin tiedot ja tutkittavien snippien genotyypit. LMM- ja LMP-menetelmien on havaittu ottavan hyvin huomioon geneettiset populaatiorakenteet assosiaatioiden testaamisessa, mutta näistä kahdesta vain LMM on pystynyt ottamaan hyvin huomioon myös kryptisen sukulaisuuden (Li ja Zhu 2013). Tosin Chen (2016) ei puolestaan ole yhtä vakuuttunut LMM-menetelmän toimivuudesta. LM ei huomioi yksilöiden välistä sukulaisuutta, joten sen pitäisi antaa huonompia tuloksia kuin LMM ja LMP, jos sukulaisuussuhteet oikeasti aiheuttavat virheitä tuloksiin ja LMM- ja LMP-menetelmät todella korjaavat niitä. LM-menetelmää käytetään siis pääasiassa vain verrokkimenetelmänä LMM- ja LMP-menetelmille ja sen tutkimiseen, että aiheuttavatko sukulaisuussuhteet todella tuloksiin virheitä. Sukulaisuuden on tosin hyvin monesti jo havaittu aiheuttavan

vääristyneitä tuloksia tutkittaessa assosiaatioita geno- ja fenotyyppien välillä (Chen ym. 2016; Li ja Zhu 2013; Astle ja Balding 2009).

Genetiikka eli perinnöllisyystiede on keskeisessä roolissa GWAS-tutkimuksissa, minkä vuoksi tutkielman toinen luku kertoo genetiikasta niiltä osin kuin se on tarpeen tämän tutkielman ymmärtämisen kannalta. Tutkielma keskittyy enimmäkseen GWAS-tutkimuksen vaiheeseen, jossa testataan assosiaatioita geno- ja fenotyypin välillä. Kuitenkin, jotta lukija saisi kokonaisvaltaisemman käsityksen GWAS-tutkimuksesta, luvussa 3 esitellään lyhyesti GWAS-tutkimuksen muitakin vaiheita. Luvut 2 ja 3 sisältävät siis perustietoa GWAS-tutkimuksesta ja genetiikasta, eivätkä käsittele vielä tutkielman pääaihetta, joka on vain pieni osa GWAS-tutkimusta. Sen vuoksi tämä tutkielma on myös sopiva luettavaksi niille, jotka eivät ole millään lailla aiemmin tutustunut genetiikkaan tai GWAS-tutkimukseen, mutta ne, joille nämä aiheet ovat jo tuttuja voivat edetä suoraan lukuun 4. Luvusta 4 alkaen tutkielma keskittyy pääaiheeseensa eli tutkimaan LMM-, LMP- ja LM-menetelmien toimivuutta assosiaatioiden testaamisessa sellaisilla aineistoilla, jotka sisältävät yksilöiden välistä sukulaisuutta. Luku 4 käsittelee yksilöiden välisiä sukulaisuussuhteita ja geneettistä vaihtelua. Luvussa 5 esitellään LMM-, LMP- ja LM-menetelmien teoriaa ja luvussa 6 simuloinnin toteutusta. Luvussa 7 kerrotaan simuloinnin tuloksista ja luvussa 8 pohditaan simuloinnin toteutusta ja tuloksia.

2 Ihmisen DNA

Ihmisen genomista suurin osa on DNA:ta eli GWAS on ihmisen kohdalla oikeastaan assosiaatiotutkimusta DNA:n ja fenotyyppien välillä. Tässä luvussa esitellään ihmisen DNA:n rakenne, kerrotaan sen vaihtelusta ihmisten välillä sekä selitetään käsitteet snippi ja genotyyppi, jotka ovat keskeisiä käsitteitä GWAS-tutkimuksissa. Snipin ja genotyypin käsitteet ovat luvun keskeisimmät asiat, mutta etenkin genetiikkaan täysin tutustumattoman kannattaa lukea luku kokonaan, sillä muut luvussa kerrottavat asiat tukevat käsitteiden ymmärtämistä. Luku perustuu pääosin kirjaan *Biology. A global approach. Global edition* (Campbell ym. 2018)

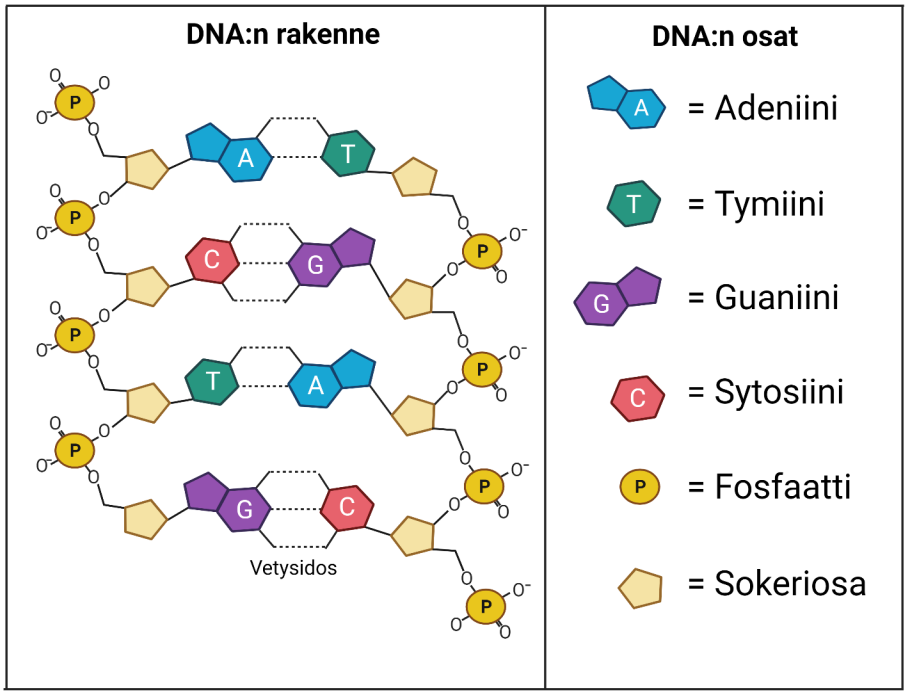
2.1 DNA:n rakenne

DNA eli deoksiribonukleiinihappo on geneettistä materiaalia eli periytyvää materiaalia ja se vaikuttaa ihmisen ominaisuuksiin. Ihmisen DNA koostuu noin 6 miljardista nukleotidiparista. Nukleotidipari muodostuu kahdesta toisiinsa kiinnittyneestä nukleotidista ja yksi nukleotidi koostuu kolmesta komponentista, jotka ovat emäsosa, sokeriosa ja fosfaatti. Nukleotidin rakenteen vaihtelu tulee sen emäsosasta, joten emästen tutkiminen on GWAS-tutkimuksissa mielenkiinnon kohteena. Emäsosa voi olla joko adeniini, tymiini, guaniini tai sytosiini, joita merkitään vastaavassa järjestyksessä kirjaimin A, T, G ja C.

DNA-molekyyli koostuu kahdesta nukleotidien muodostamasta juosteesta, jotka ovat kiinnittyneet toisiinsa vetysidoksilla muodostaen rakenteen, jota kutsutaan kaksoiskierteeksi. Vetysidokset muodostuvat emäsosien välille, joista vain A ja T sekä C ja G voivat muodostaa

keskenään vetysidoksen eli vastakkaisissa juosteissa samassa kohtaa DNA-molekyyliä ovat aina pareina joko emäkset A ja T tai G ja C. Kuvassa 1 havainnollistetaan DNA:n rakenne.

Ihmisen 6 miljardin nukleotidiparin muodostama DNA on jakautunut 46 kromosomiin, joista jokainen koostuu yhdestä DNA-molekyylistä. Nämä 46 kromosomia jakautuvat 23 kromosomipariin ja joka parista toinen kromosomi on peritty äidiltä ja toinen isältä. Kromosomiparin kromosomeja kutsutaan toistensa vastinkromosomeiksi. Vastinkromosomit eivät ole täysin kopioita toisistaan, mutta ne ovat saman pituisia ja määrittävät samoja geneettisiä ominaisuuksia. Poikkeuksena miehillä 23. kromosomipari koostuu kromosomeista X ja Y, joista Y-kromosomi on X-kromosomia lyhyempi.



Kuva 1: Kaksiulotteinen kuva neljän nukleotidiparin mittaisesta DNA-molekyylistä. Toisen juosteen emäkset G, T, C ja A ylhäältä alaspäin luettuna ovat kiinnittyneet vetysidoksilla vastinjuosteen emäksiin C, A, G ja T. Oikealla puolella kuvaa ovat DNA:n osien nimet. Kuva on tehty sivulla BioRender.com.

2.2 Geneettinen vaihtelu ja genotyyppi

GWAS-tutkimuksissa ollaan kiinnostuneita yksittäisten emästen vaihtelusta yksilöiden välillä mikä tarkoittaa, että tietyssä sijainnissa DNA:ta eri yksilöillä voi olla eri emäkset. Jos tietyssä sijainnissa DNA:ta emäksen vaihtelua esiintyy yli yhdellä prosentilla väestöstä, puhutaan yhden emäksen monimuotoisuudesta eli snipistä (*single nucleotide polymorfism*, SNP). Ihmisen DNA:ssa esiintyy snippejä noin 100–300 nukleotidiparin välein eli snippejä on kaikkiaan nykytiedon mukaan muutama miljoona. Yleinen tapa nimetä löydettyjä snippejä

ovat rsID:t (*reference SNP cluster ID*). RsID alkaa kirjaimilla rs jonka perässä on vaihteleva määrä numeroita.

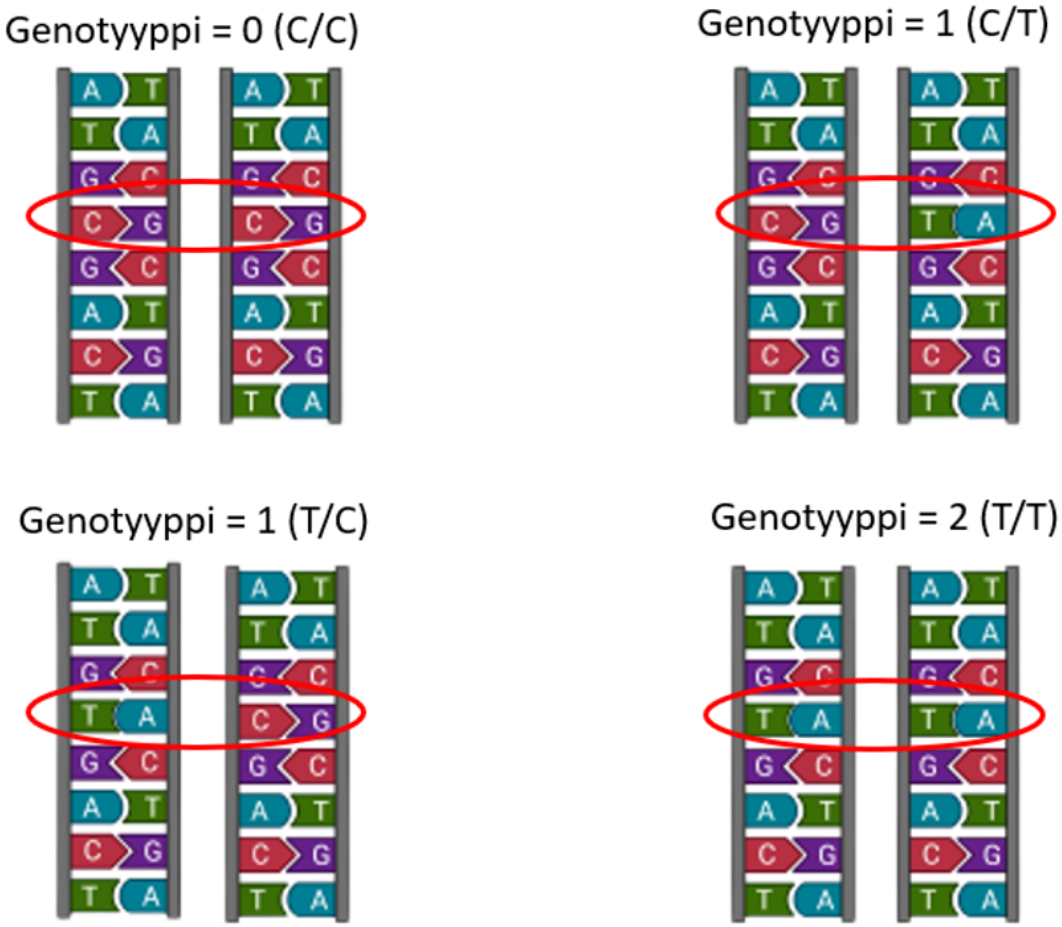
Yhden snipin mahdollisia eri ilmentymiä kutsutaan alleeleiksi. Yhdellä snipillä voi olla emästen lukumäärän mukaan neljä eri alleelia, mutta usein snipeillä on vain kaksi alleelia ja tässä tutkielmassa keskitytään vain kaksialleelisiin snippeihin. Esimerkiksi snipin rs429358 alleelivaihtoehdot ovat C ja T (Wu ym. 2020), mikä tarkoittaa, että yksilöllä voi olla snipin määräämässä kohdassa toisessa juosteessa DNA-molekyyliä joko emäs C tai T. Alleelivaihtoehdot samalle snipille voisi myös ilmaista olevan G ja A, jos alleelit luettaisiin DNA-molekyylin toisesta juosteesta, mutta koska emäsparit ovat aina A ja T tai C ja G niin snipin alleeleiksi riittää ilmaista vain toisen juosteen emäkset. Kuvalla 2 havainnollistetaan mikä on kaksialleelinen snippi ja sen alleelit.

Kuva 2: Henkilöiden 1 ja 2 kahdentoista nukleotidiparin mittaiset DNA-molekyylit, joissa on ympyröidyssä kohdassa eri emäsparit G-C (Henkilö 1) ja A-T (Henkilö 2). Kyseinen kohta DNA:ta on kaksialleelinen snippi, jos molempia emäspareja esiintyy siinä kohdassa DNA:ta yli yhdellä prosentilla väestöstä ja muita emäspareja ei ollenkaan. Tämän snipin alleelit olisivat tällöin ylemmästä juosteesta luettuna G ja A tai alemmasta juosteesta luettuna C ja T. Kuva on tehty sivulla BioRender.com

Jokainen snippi löytyy sekä äidiltä että isältä peritystä kromosomista. Yksilön genotyyppi ilmaistaan kaksialleelisen snipin kohdalla snipin toisen alleelin lukumäärän mukaan, jotka yksilö on vanhemmiltaan perinyt. Sillä ei ole väliä kumman alleelin lukumäärän mukaan genotyyppi ilmaistaan tutkimusaineistossa, kunhan kaikkien tutkimusaineistossa olevien yksilöiden genotyypit ilmaistaan saman alleelin lukumäärän mukaan.

Tarkastellaan esimerkkinä jälleen snippiä rs429358, jonka alleelit ovat C ja T. Jos valitaan, että snipille rs429358 genotyyppi lasketaan perittyjen C alleelien lukumäärän mukaan, niin

kyseisen snipin genotyypit saavat arvot seuraavasti: T/T = 0, T/C = 1, C/T = 1, C/C = 2. Merkintätavalla X_1/X_2 tarkoitetaan, että alleeli X_1 on peritty toiselta vanhemmista ja alleeli X_2 toiselta. Kuvassa 3 havainnollistetaan mitä genotyyppi tarkoittaa yhden kaksialleelisen snipin kohdalla.



Kuva 3: Kuvassa havainnollistetaan, miten yksilön genotyyppi määräytyy kaksialleelisen snipin kohdalla sen mukaan, mitä alleelleja yksilö on vanhemmiltaan perinyt. Jokaisessa neljässä tilanteessa on kaksi kahdeksan nukleotidiparin mittaista DNA-molekyyliä. Vasemmanpuoleinen DNA-molekyyli on toiselta vanhemmalta peritty ja oikeanpuoleinen toiselta. DNA-molekyyleistä on ympyröity kohta, jossa on snippi, jonka alleelit ovat C ja T vasemman reunan juosteesta luettuna. Kuvan genotyypit on laskettu perittyjen T alleelien lukumäärän mukaan. Kuva on tehty sivulla BioRender.com

3 GWAS-tutkimus

Tässä luvussa kerrotaan GWAS-tutkimuksen eri vaiheista mukaan lukien assosiaatioiden testaaminen, jossa käytettävät menetelmät ovat tässä tutkielmassa suurimman kiinnostuksen

kohteena. Luvun tarkoitus on esitellä lukijalle yleisesti assosiaatioiden testaamisen vaihe GWAS-tutkimuksissa sekä muut keskeiset vaiheet koko GWAS-tutkimuksesta. Tutkielma rajautuu vain pieneen osaan GWAS-tutkimusta, joten muidenkin vaiheiden avaaminen lyhyesti on hyödyksi tutkielman ymmärtämisen kannalta. Päälähteenä käytetään Uffelmannin (2021) tiivistelmää GWAS-tutkimuksista.

3.1 Aineiston kerääminen ja käsittely

Tutkimus alkaa kohdepopulaation valinnalla ja aineiston keräämisellä valitusta populaatiosta. Tutkija voi kerätä täysin uuden aineiston, mutta aineiston keräämisessä voi hyödyntää myös jo ennalta kerättyjä aineistoja, joita on nykyään paljon saatavilla esimerkiksi biopankeista kuten UK Biobank ja Keski-Suomen Biopankki. Valmiiden datalähteiden hyödyntäminen on yleensä kannattavaa, sillä usein merkitsevien assosiaatioiden löytämiseksi tarvitaan suuria aineistoja ja ilman valmiiden datalähteiden hyödyntämistä riittävän suuri aineisto voi olla liian työlästä kerätä. Esimerkiksi Alzheimerin tautia tutkivassa meta-analyysissä (Wightman ym. 2021) aineisto koostui jopa 1126563 yksilöstä. Aineiston koon tarve kuitenkin vaihtelee tutkimuskohtaisesti.

GWAS-tutkimuksen kannalta keskeisimmät tiedot aineistossa ovat yksilöiden geno- ja fenotyyppitiedot, joiden välisiä assosiaatioita tutkitaan. Yksilön genotyyppitiedot saadaan esimerkiksi sylki- tai verinäytteestä. Genotyyppitietojen luentaan tarvittavia teknologioita on useita erilaisia, mutta niiden ymmärtäminen ei ole oleellista tämän tutkielman kannalta, sillä yksilöiden genotyypit luodaan tässä tutkielmassa simuloimalla. Genotyyppausteknologioihin voi kuitenkin halutessaan tutustua Kockumin, Huangin ja Stridhin (2023) artikkelista. Geno- ja fenotyyppitietojen lisäksi yksilöiltä kerätään tutkimuksesta riippuen muita tarpeellisia tietoja, kuten esimerkiksi kliiniset tiedot, taustatiedot kuten ikä, sukupuoli ja asuinpaikka ja genotyyppaukseen liittyvät tiedot.

Aineistolle suoritetaan monia laadunvarmistuksia ennen kuin edetään assosiaatioiden tutkimiseen. Aineiston laadunvarmistus käsittää useita vaiheita, joiden tavoitteena on varmistaa, että genotyyppitiedot ovat tarkkoja ja luotettavia. Esimerkiksi snippien genotyyppijakaumista tarkistetaan ovatko ne Hardy-Weinberg tasapainossa. Hardy-Weinberg tasapaino tarkoittaa, että snippien genotyypit noudattavat $\text{Bin}(2, f)$ jakaumaa, missä f on snipin genotyypin määräävän alleelin alleelifrekvenssi. Alleelifrekvenssillä tarkoitetaan genetiikkaa käsittelevissä tutkimuksissa alleelin esiintymän osuutta. Hardy-Weinberg tasapainosta poikkeaminen viittaa usein genotyyppausvirheeseen, minkä vuoksi snippi on usein syytä jättää tällöin pois analyyseistä. Tästä ja muista laadunvarmistuksen vaiheista kerrotaan enemmän Mareesin (2018) artikkelissa.

Yksilöiden genomien luennassa tapahtuu monesti luentavirheitä, minkä vuoksi puuttuvia genotyyppitietoja usein myös imputoidaan. Puuttuvien genotyyppitietojen imputoimiseen käytetään haplotyypin estimointimenetelmiä (*haplotype phasing*). Haplotyyppi tarkoittaa usean eri alleelin yhdistelmää, jotka peritään yhdessä (Kockum, Huang, ja Stridh 2023). Esimerkiksi viisi eri kaksialleelista snippiä voisivat teoriassa muodostaa $2^5 = 32$ erilaista haplotyyppiä, mutta koska kaikki alleelit eivät periydy toisistaan riippumattomasti, niin haplotyyppien määrä on usein huomattavasti pienempi. Jos siis esimerkiksi viiden snipin

alleeleista puuttuu yksi, niin se voidaan imputoida neljän muun tiedossa olevien alleelien avulla. Haplotyyppien estimointimenetelmistä voi lukea tarkemmin Browningin (2011) artikkelista. Imputoinnin käyttö kuitenkin vaihtelee tutkimuksesta riippuen. Esimerkiksi Marees (2018) ohjeistaa poistamaan sellaiset snipit aineistosta, joiden alleelit ovat jääneet monelta aineiston yksilöltä määrittämättä.

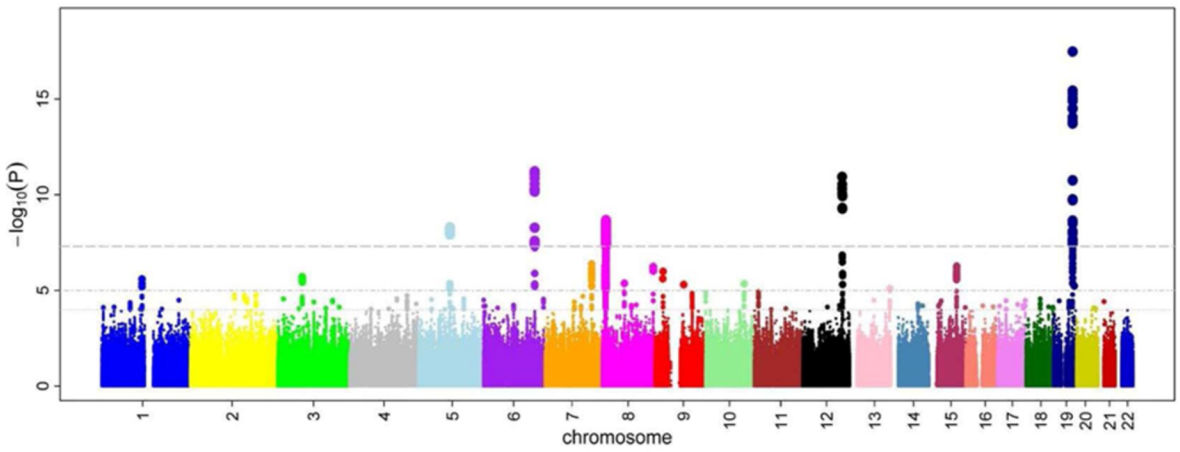
3.2 Assosiaatioiden testaaminen

Assosiaatioiden testausvaiheessa GWAS-tutkimuksissa pyritään löytämään tilastollisin menetelmin assosiaatioita geno- ja fenotyyppien välillä. Vaihtoehtoja käytettävälle menetelmälle on useita, joista kolmea (LMM, LMP ja LM) tämä tutkielma tarkastelee tarkemmin. Käytettävä menetelmä riippuu tutkimuksesta, mutta Uffelmanin (2021) mukaan lineaarinen regressiomalli on tyypillisimmin käytetty menetelmä testaamiseen jatkuvan vasteen tapauksessa ja logistinen regressiomalli binäärisen vasteen tapauksessa. Kuitenkin esimerkiksi LMM (Li ja Zhu 2013) ja logistinen sekamalli (Chen ym. 2016) ovat viime aikoina yleistyneet assosiaatioiden testaamisessa GWAS-tutkimuksissa.

Mallit vaihtelevat tutkimusten välillä myös sen mukaan käytetäänkö additiivista vai ei-additiivista mallia. Pirastun (2016) mukaan additiiviset mallit ovat yleisempiä GWAS-tutkimuksissa kuin ei-additiiviset, mutta hän ei pidä additiivisia malleja välttämättä parempina ja käytti kahvinkulutusta käsittelevässä GWAS-tutkimuksessa ei-additiivista mallia. Xue (2022) puolestaan toteutti sikojen kasvuominaisuuksia tutkivassa GWAS-tutkimuksessa assosiaation testaamisen sekä additiivisella että ei-additiivisella mallilla. Tässä tutkielmassa kuitenkin rajoitutaan additiivisiin malleihin.

GWAS-tutkimuksissa assosiaatioita testataan erikseen useiden, yleensä noin miljoonan toisistaan riippumattomien snippien kohdalla (Uffelmann ym. 2021). Tämän takia testaamisessa käytettävä merkitsevyystaso on huomattavasti pienempi kuin perinteisesti tilastollisissa testeissä käytettävä $p < 0.05$, jotta vältetään tyypin 1 virheitä. GWAS-tutkimuksissa yleisimmin käytettävä merkitsevyysraja on $p < 5 \cdot 10^{-8}$. Tämä raja perustuu Bonferronin korjaukseen (Bonferroni 1936) testattaessa miljoonan toisistaan riippumattoman snipin assosiaatiota (Johnson ym. 2010). Myös tässä tutkielmassa merkitsevyysrajana käytetään kyseistä rajaa.

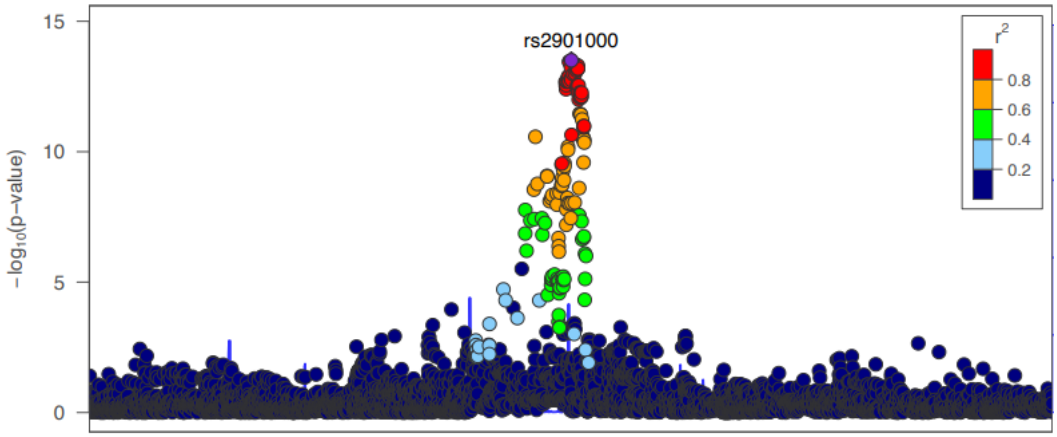
Valituista menetelmistä tai merkitsevyysrajoista riippumatta testien tulokset esitetään usein Manhattan-kuviolla (*Manhattan plot*). Manhattan-kuviossa esitetään kromosomeittain järjestyksessä kunkin snipin assosiaatiotestin p-arvot. Kuvassa 4 on esimerkki Manhattan-kuviosta (Ikram ym. 2010). Manhattan kuviosta kerrotaan tarkemmin Wangin (2022) artikkelissa.



Kuva 4: Manhattan kuviossa esitetään kunkin snipin assosiaatiotestin p-arvo $\log_{10}$-skaalalla. Yksi piste vastaa kuvassa yhtä testattua snippiä. Snipit ovat kuvassa sijaintinsa mukaan järjestyksessä ja kromosomit on eroteltu kuvassa väreillä. Ylin katkoviiva on GWAS-tutkimuksissa yleisimmin käytetyn merkitsevyystason $5 \cdot 10^{-8}$ korkeudella (Ikram ym. 2010). Kuva julkaistu Creative Commons lisenssillä http://creativecommons.org/licenses/by/4.0/.

Kuvan 4 assosiaatiotesteissä on löydetty viisi DNA-jaksoa, joissa on havaittu useita merkitseviä snippejä, kun käytetään merkitsevyysrajaa $5 \cdot 10^{-8}$. Merkitsevien snippien tarkkaa määrää on kuvasta vaikea arvioida suoraan, koska pisteet menevät päällekkäin. Tämä johtuu siitä, että samassa kromosomissa toisiaan lähellä olevien snippien alleelit ovat usein vahvasti korreloituneita, minkä vuoksi merkitseviä snippejä sisältävässä DNA-jaksossa merkitseviä snippejä on yleensä myös useampi. Jos toisiaan lähellä olevien snippien alleelit ovat korreloituneita niin sanotaan, että ne ovat kytkentäepätasapainossa (*linkage disequilibrium*, LD) (Nordborg ja Tavaré 2002). Toisiaan lähellä olevien snippien alleelien korreloituneisuus johtuu erilaisista biologisista ominaisuuksista, mutta tässä tutkielmassa rajaudutaan tutkimaan toisistaan riippumattomia snippejä, joten kytkentäepätasapainoa ei käsitellä tarkemmin.

Niitä DNA-jaksoja, joista assosiaatiotestit ovat havainneet merkitseviä snippejä, tarkastellaan usein tarkemmin zoomatuilla lokuskuvioilla (Pruim ym. 2010). Zoomatusta lokuskuviosta näkee tarkemmin DNA-jakson merkitsevien snippien määrän ja siinä havainnollistetaan usein myös kytkentäepätasapainoa snippien välillä. Kuvassa 5 on esimerkki lokuskuviosta (Mitchell ym. 2022).



Kuva 5: Zoomatulla lokus kuviolla esitetään merkitsevät snipit ja p-arvot merkitseviä snippejä sisältävälle DNA-jaksolle. Yksi piste kuvassa vastaa yhtä snippiä ja pisteet ovat sijaintinsa mukaan järjestyksessä vaaka-akselilla. Lisäksi kuvassa on havainnollistettu väreillä, miten vahvasti kukin snippi on kytkentäepätasapainossa merkitsevimmän eli kaikista pienimmän p-arvon omaavan snipin kanssa (Mitchell ym. 2022). Kuva julkaistu Creative Commons Attribution 4.0 International lisenssillä http://creativecommons.org/licenses/by/4.0/.

3.3 Kausaalisten snippien etsintä

Luvun 3.2 lopussa todettiin, että toisiaan lähellä olevat snipit ovat usein vahvasti korreloituneita ja sen vuoksi assosiaatiotestit havaitsevat monesti DNA-jaksoissa useita merkitseviä snippejä. Tämän vuoksi moni merkitsevistä snipeistä ei välttämättä ole kuitenkaan kausaalisesti yhteydessä fenotyyppiin. Kausaalisten snippien etsinnästä käytetään GWAS-kontekstissa termiä hienokartoitus (*fine-mapping*). Hienokartoitus on siis yleinen käsite kausaalisten snippien etsimiseen eikä yksittäinen menetelmä. Hutchinson (2020) esittelee artikkelissaan hienokartoitusmenetelmiä.

4 Yksilöiden väliset sukulaisuussuhteet

Tässä luvussa kerrotaan geneettisistä populaatiorakenteista ja kryptisestä sukulaisuudesta, joiden on havaittu aiheuttavan tulosten vääristymiä geneettisissä assosiaatiotutkimuksissa (Chen ym. 2016; Li ja Zhu 2013; Astle ja Balding 2009). Tulosten vääristymien korjaamiseksi on kehitetty menetelmiä, joista tässä tutkielmassa tutkitaan LMM- ja LMP-menetelmiä. Menetelmiä tutkitaan soveltamalla niitä simuloimalla tuotettuihin sukulaisuutta sisältäviin aineistoihin. Yksilöiden välinen sukulaisuus ilmenee geno-fenotyyppiaineistossa monilla eri tavoilla, minkä vuoksi tämän luvun tarkoituksena on antaa perusteluja sille, millä tavalla sukulaisuus on simuloitu aineistoihin tässä tutkielmassa.

4.1 Geneettiset populaatiorakenteet

Geneettisellä populaatiorakenteella tarkoitetaan sitä, että otoksessa on ryhmiä, joiden yksilöt ovat kaukaisten yhteisten esivanhempien takia sukua toisilleen (Sul, Martin, ja Eskin 2018; Astle ja Balding 2009). Tämä johtaa siihen, että yhteisiä esivanhempia omaavat yksilöt ovat geneettisesti keskimäärin samankaltaisempia muihin verrattuna. Geneettinen samankaltaisuus liittyy snippien tapauksessa samojen alleelien lukumäärään. Mitä enemmän samoja snippien alleeleja kahdella yksilöllä on, sitä samankaltaisempia yksilöt ovat geneettisesti. Eli samaan geneettiseen populaatioon kuuluvat yksilöt jakavat keskimäärin enemmän samoja alleeleja verrattuna muiden geneettisten populaatioiden yksilöihin. Geneettisten populaatioiden erot ilmenevät siis alleelifrekvenssien eroina.

Ihmisten geneettistä jakautumista on tutkittu paljon. Kansainvälisessä tutkimuksessa 1000 Genome Project (1000 Genomes Project Consortium ym. 2015) selvitettiin ihmisten geneettistä vaihtelua koko genomin ja maailman väestön laajuudelta. Rosenberg (2011) tutki myös maailmanlaajuisesti ihmisten geneettistä vaihtelua, mutta vain pieneltä osin genomia. Hannelius (2008) tutki puolestaan ihmisten geneettisiä eroja Suomen ja Ruotsin väestön osalta. Näiden tutkimusten mukaan geneettisten populaatioiden väliset erot alleelifrekvensseissä vaihtelevat riippuen snipistä ja verrattavista geneettisistä populaatioista. Tietyn snipin alleelin alleelifrekvenssi voi olla toisessa geneettisessä populaatiossa hyvinkin yleinen, mutta toisessa geneettisessä populaatiossa sitä ei välttämättä esiinny ollenkaan ja kolmannessa alleelifrekvenssi on jotain siltä väliltä. Joidenkin snippien kohdalla alleelifrekvenssi voi olla geneettisten populaatioiden välillä myös samansuuruinen. Geneettiset populaatiot ovat usein myös linjassa ihmisten maantieteellisen sijainnin kanssa. Populaatiorakenteiden tuominen simuloituihin aineistoihin perustuu tässä tutkielmassa edellä mainittujen tutkimusten havaintoihin.

Siihen, miten geneettisten populaatioiden erot syntyvät on löydetty kolme päämekanismia, jotka ovat luonnonvalinta, geneettinen ajautuminen ja geenivirta. Seuraavakasi esitellään nämä kolme mekanismia käyttäen lähteenä Campbelin (2018) kirjaa *Biology. A global approach. Global edition*.

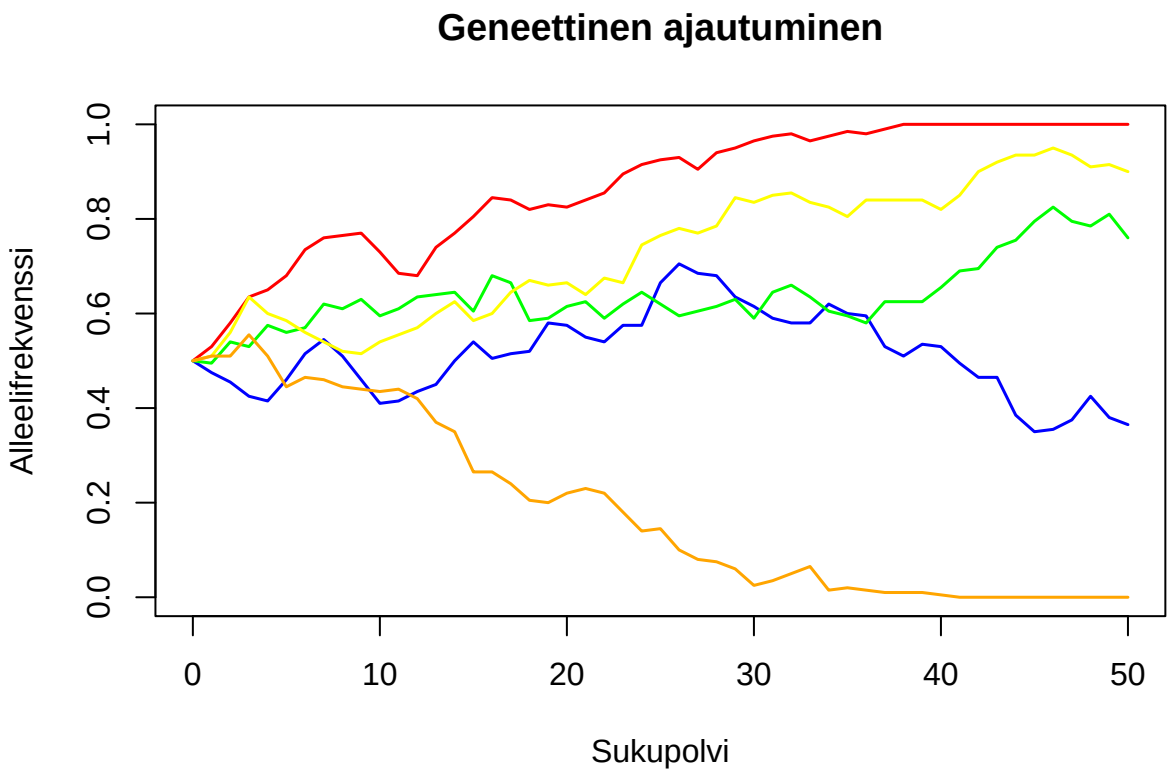
Luonnonvalinta (*natural selection*) tarkoittaa, että sellaiset yksilöt yleistyvät, joiden ominaisuudet tekevät yksilön selviytymiskyvyn ja lisääntymismahdollisuudet muita yksilöitä paremmiksi. Snippien kohdalla tämä tarkoittaa, että ne snipin alleelit yleistyvät, jotka lisäävät yksilön selviytymiskykyä ja lisääntymismahdollisuuksia. Esimerkiksi *D. melanogaster* kärpäsillä esiintyy alleeli, joka antaa vastustuskykyä DDT-hyönteismyrkkyä vastaan. 1930-luvun alkupuolella, ennen DDT:n kehittämistä havaittiin, että kärpäsistä kerätyssä laboratoriokannassa suojaavan alleelin esiintyvyys oli 0 %, mutta 1960-luvun jälkeen, jolloin DDT:tä oli käytetty jo yli 20 vuotta, kerätyssä laboratoriokannassa suojaavan alleelin esiintyvyys oli jopa 37 %. Todennäköisesti siis DDT-hyönteismyrkyltä suojaavaa alleelia kantavat kärpäset selviytyivät muita kärpäsiä paremmin ja sen seurauksena suojaava alleeli on yleistynyt.

Geneettinen ajautuminen (*genetic drift*) tarkoittaa alleelifrekvenssien muuttumista sukupolvesta toiseen pelkän sattuman kautta. Kaksialleelisen snipin tapauksessa se tarkoittaa, että toinen alleeleista yleistyy toiseen verrattuna alleelien satunnaisen

periytyvyyden takia. Geneettinen ajautuminen ilmenee voimakkaimmin pienissä populaatioissa.

Liitteessä 1 havainnollistetaan geneettistä ajautumista jonkin verran todellisuutta yksinkertaistavalla simuloinnilla. Simuloinnissa viisi sadan yksilön populaatiota eristäytyvät toisistaan 50 sukupolven ajaksi. Simuloinnissa tarkastellaan yhden kaksialleelisen snipin toisen alleelin frekvenssin muutosta sukupolvien välillä. Snipin alleelit periytyvät simuloinnissa jälkeläiselle yhtä suurella todennäköisyydellä. Tarkasteltavaa alleelia merkitään simuloinnissa numerolla 1. Ensimmäisen sukupolven kohdalla alleelifrekvenssi on kunkin populaation kohdalla 50 % ja puolet yksilöistä on miehiä ja puolet naisia. Simulointi olettaa seuraavat asiat, jotka voivat poiketa todellisuudesta.

1. Vain saman sukupolven nainen ja mies voivat saada jälkeläisen keskenään.
2. Jälkeläisen isä ja äiti valitaan satunnaisesti edellisen sukupolven miehistä ja naisista riippumatta siitä, kuinka monta ja kenenkä kanssa saatuja jälkeläisiä isällä ja äidillä jo on tai tulee olemaan.
3. Jälkeläisen sukupuoli on yhtä suurella todennäköisyydellä mies tai nainen.
4. Populaation koko ei muutu sukupolvien välillä.



Kuva 6: Simuloitu esimerkki geneettisestä ajautumisesta. Kukin käyrä vastaa yhtä eristäytynyttä populaatiota. Ensimmäisen sukupolven kohdalla (sukupolvi = 0) snipin alleelifrekvenssi on 0.5 jokaisella populaatiolla. Sattumanvaraisen alleelien periytymisen takia populaatioiden alleelifrekvenssit erkanevat, kun populaatiot eristäytyvät toisistaan 50 sukupolven ajaksi. Jokaisen populaation koko on jokaisen sukupolven kohdalla 100 yksilöä.

Kuvasta 6 nähdään miten alleelifrekvenssit erkanevat toisistaan eristäytyneiden populaatioiden välillä, kun alleelit periytyvät sattumavaraisesti sukupolvelta toiselle. Selvät erot alleelifrekvensseissä alkavat muodostua jo muutaman sukupolven jälkeen. Kahden populaation kohdalla toinen alleeleista häviää jopa kokonaan populaatiosta 50 sukupolven aikana.

Geenivirralla (*gene flow*) tarkoitetaan alleelien siirtymistä geneettisestä populaatiosta toiseen, koska yksilöt siirtyvät geneettisestä populaatiosta toiseen ja lisääntyvät keskenään. Geenivirrat vähentävät geneettisten populaatioiden välisiä eroja.

Geenivirtaa on havaittu esimerkiksi Eriejärvellä Pohjois-Amerikassa. Järven alueella elää vesikäärmepopulaatio (*Nerodia sipedon*), joista mantereen puolella elää pääosin raidallisia ja järvellä olevilla saarilla raidattomia käärmeitä, koska raidattomuus sopii paremmin naamioitumiseen saarien kivikkoisessa maastossa ja raidallisuus mantereen suomaastoon. Raidat muodostuvat käärmeille muutaman eri alleelin seurauksena. Luonnonvalinnan mukaan saarella elävien käärmeiden tulisi olla tällöin raidattomia, mutta näin ei kuitenkaan

ole, koska joka vuosi mantereelta ui saarille raidallisia käärmeitä kantaen mukanaan raidallisuuden aiheuttavia alleeleja, mikä estää luonnonvalintaa karsimasta raidallisia käärmeitä saarilta.

Ihmisten kohdalla geenivirroilla on nykypäivänä voimakas vaikutus geneettiseen vaihteluun. Ihmiset muuttavat paljon paikasta toiseen ympäri maailmaa, minkä vuoksi geenivirrat vähentävät merkittävästi geneettisten populaatioiden välisiä eroja.

4.2 Kryptinen sukulaisuus

Kryptisellä sukulaisuudella tarkoitetaan, että aineistossa on toiselleen läheistä sukua olevia yksilöitä, mutta sukulaisuussuhteet eivät ole tutkijan tiedossa. Verrattuna geneettisiin populaatiorakenteisiin, kryptiset sukulaiset muodostavat aineistoon paljon pienempiä ihmisryhmiä, jopa vain pareja, ja sukulaisuussuhteet johtuvat huomattavasti viimeaikaisemmista yhteisistä esivanhemmista (Sul, Martin, ja Eskin 2018; Astle ja Balding 2009).

Mitä läheisempää sukua yksilöt ovat toisilleen, sitä enemmän ne ovat keskimäärin syntyperän perusteella identtisiä (*identical by descent*, IBD) (Speed ja Balding 2015). Yksittäisen snipin alleelia tarkasteltaessa kahden ihmisen alleelit ovat IBD, jos ne on peritty yhteiseltä esivanhemmalta ja ovat samat. IBD-osuus tarkoittaa osuutta kaikkien snippien alleeleista, jotka ovat yksilöiden välillä IBD. Visscher (2006) havaitsi, että sisarusten välisten genomien IBD-osuudet vaihtelivat 37.4 % ja 61.7 % välillä. Lapsi perii vanhemmalta puolet genomistaan, joten vanhemman ja lapsen välinen IBD-osuus on 50 %. Identtiset kaksoset jakavat saman genomin, joten niiden IBD-osuus on 100 %. Nämä sukulaisuussuhteet ovat kaikista läheisimpiä sukulaisuussuhteita, mikä antaa ylärajat yksilöiden välisten genomien IBD-osuuksille, mikä on keskeinen oletus luvussa 6 toteutettavalle simuloinnille.

5 Tutkittavat menetelmät

Tässä luvussa käsitellään LMM-, LMP- ja LM-menetelmien teorioita. Teorioita ei käsitellä yleisesti vaan siten, miten menetelmiä käytetään tämän tutkielman assosiaatioiden testaamisessa.

Assosiaatioiden testaaminen suoritetaan kullakin menetelmällä snipeille j, $j = 1, \ldots, n_{\text{snp}}$ erikseen, missä n_{snp} on testattavien snippien lukumäärä. Assosiaatiotestin j nollahypoteesi on H_0: $\beta_j = 0$, missä β_j on testattavan snipin j vaikutuksen suuruutta kuvaava parametri. Nollahypoteesi olettaa siis, että testattava snippi j ei ole assosioitunut fenotyypin kanssa. Vastahypoteesi H_1: $\beta_j \neq 0$ puolestaan olettaa, että testattava snippi j on assosioitunut fenotyypin kanssa (Sul, Martin, ja Eskin 2018).

Snippien testaaminen erikseen vaikuttaa siihen, mikä tulkitaan testeissä merkitseväksi assosiaatioksi. Tässä tutkielmassa käytetään GWAS-tutkimuksissa yleisesti käytettyä rajaa $p < 5 \cdot 10^{-8}$. Rajaa käytetään GWAS-tutkimuksissa tyypin 1 virheiden vähentämisen takia ja se perustuu siihen, että nykytiedon mukaan ihmisellä on noin miljoona toisistaan

riippumatonta snippiä ja Bonferronin korjauksen mukaan käyttämällä rajaa $p < 5 \cdot 10^{-8}$ testattaessa erikseen miljoonaa toisistaan riippumatonta snippiä, joilla ei ole todellista assosiaatiota fenotyypin kanssa, saadaan testeissä yksi tai enemmän tyypin 1 virheitä todennäköisyydellä $p < 0.05$ (Johnson ym. 2010; Uffelmann ym. 2021).

Geno-fenotyyppiaineisto, johon menetelmiä sovelletaan sisältää yksilöiden $i,\ i = 1, \ldots, n$ fenotyypit $\mathbf{y} = (y_1, \ldots y_n)$, missä n on geno-fenotyyppiaineiston yksilöiden lukumäärä, ja jokaisen yksilön i jokaisen testattavan snipin j genotyypit $\mathbf{X} = [\mathbf{x}_j] = [x_{ij}]$. Matriisi $\mathbf{X}$ on siis $n \times n_{snp}$ matriisi, jonka rivillä i ja sarakkeella j on yksilön i snipin j genotyyppi x_{ij}. Fenotyyppi y_i on jatkuva-arvoinen muuttuja ja genotyyppi $x_{ij} \in \{0, 1, 2\}$.

5.1 Lineaarinen regressiomalli

Lineaarinen regressiomalli on ollut perinteinen tapa testata assosiaatioita geno- ja fenotyyppien GWAS-tutkimuksissa. Malli on muotoa

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_j \mathbf{x}_j + \boldsymbol{\epsilon}$$

missä $\mathbf{1}$ on ykkösiä sisältävä $n \times 1$ vektori, β_0 on tuntematon populaatiokeskiarvo ja $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma_\epsilon \mathbf{I})$ on vektori satunnaisvirheistä (Sul, Martin, ja Eskin 2018). Mallin parametrit β_0 ja β_j estimoidaan iteratiivisesti painotetun pienimmän neliösumman (*iteratively reweighted least squares*, IWLS) menetelmällä (Charnes, Frome, ja Yu 1976).

5.2 Lineaarinen sekamalli

Lineaarisen sekamallin (LMM) päälähteenä käytetään Sulin, Martin ja Eskin (2018) artikkelia. Lineaarinen sekamalli on regressiomalli, joka yhdistää kiinteitä vaikutuksia ja satunnaisvaikutuksia. Malli on muotoa

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_j \mathbf{x}_j + \boldsymbol{\mu} + \boldsymbol{\epsilon},$$

missä $\boldsymbol{\mu} \sim \mathrm{N}(\mathbf{0}, \sigma_\mu^2 \mathbf{K})$ on satunnaisvektori, $\mathbf{K}$ on sukulaisuusmatriisi (*kinship matrix*), σ_μ hajontaparametri ja mallin muut komponentit ovat samat kuin luvun 5.1 mallissa.

Satunaisvektorin $\boldsymbol{\mu}$ tarkoitus on ottaa huomioon muiden kuin testattavana olevan snipin vaikutus ja sen vuoksi sitä kutsutaan GWAS-tutkimusten yhteydessä usein polygeeniseksi vaikutukseksi (*polygenic effect*). Muiden snippien yhteenlaskettua vaikutusta voidaan mallintaa satunnaisvaikutuksella $\boldsymbol{\mu}$, koska sellaisilla yksilöillä, joilla on samankaltainen genomi, on myös keskimäärin samansuuruinen koko muun genomin yhteenlaskettu vaikutus. Yksilöiden välisten genomien samankaltaisuus otetaan huomioon sukulaisuusmatriisin $\mathbf{K}$ avulla. Sukulaisuusmatriisin $\mathbf{K}$ alkiot kuvaavat pareittain yksilöiden genomien välistä samankaltaisuutta. Malli ottaa siis huomioon yksilöiden välisen sukulaisuuden, koska toisilleen enemmän sukua olevien yksilöiden genomit ovat keskenään keskimäärin samankaltaisempia kuin toisilleen vähemmän sukua olevien yksilöiden

genomit. Sukulaisuusmatriisin $\mathbf{K}$ määrittämiseen on eri tapoja, mutta tässä tutkielmassa sukulaisuusmatriisi $\mathbf{K}$ lasketaan kaavalla

$$\mathbf{K} = \frac{\mathbf{Z}\mathbf{Z}^T}{n_{\text{snp}}},$$

missä matriisi $\mathbf{Z}$ sisältää yksilöiden standardoitut genotyypit, eli $\mathbf{K}$ on korrelaatiomatriisi.

Parametrit σ_ϵ ja σ_μ estimoidaan EMMA-algoritmilla (*efficient mixed-model association*) (Kang ym. 2008) ja parametrien β_0 ja β_j estimointi toteutetaan yleistetyllä pienimmän neliösumman (*generalized least squares*, GLS) menetelmällä (Aitken 1936).

5.3 Lineaarinen regressiomalli pääkomponenteilla

Tässä tutkielmassa lineaarinen regressiomalli pääkomponenteilla (LMP) tarkoittaa lineaarista regressiomallia, jossa käytetään kovariaatteina yksilöiden genotyyppitiedoista muodostettuja pääkomponentteja. Lähteenä käytetään Abegazin (2019) artikkelia. Malli on muotoa

$$\mathbf{y} = \beta_0\mathbf{1} + \beta_j\mathbf{x}_j + \mathbf{p}_1\alpha_1 + \mathbf{p}_2\alpha_2 + ... + \mathbf{p}_c\alpha_c + \boldsymbol{\epsilon},$$

missä $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_c$ ovat malliin valitut ensimmäiset c kappaletta pääkomponentteja ja $\alpha_1, \alpha_2, \ldots, \alpha_c$ ovat niiden tuntemattomat vaikutukset. Muut mallin komponentit ovat samat kuin luvun 5.1 mallissa.

Mallin pääkomponentit muodostetaan genotyypit sisältävästä matriisista $\mathbf{X}$ käyttämällä singulaariarvohajotelmaa. Olkoon matriisi $\mathbf{G}$ keskistetty matriisi $\mathbf{X}$, jonka aste on w. Matriisin $\mathbf{G}$ singulaariarvohajotelma on

$$\mathbf{G} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}},$$

missä $\boldsymbol{\Lambda}$ on $w \times w$ diagonaalimatriisi, jonka diagonaalilla ovat matriisin $\mathbf{G}$ singulaariarvot $\lambda_1, \ldots, \lambda_w$, joille pätee $\lambda_1 \geq \lambda_2 \geq, \ldots, \geq \lambda_w$. $\mathbf{U}$ on ortogonaalinen $n \times w$ matriisi ja $\mathbf{V}$ on ortogonaalinen $n_{snp} \times w$ matriisi. Merkitään matriisin $\mathbf{V}$ sarakkeita $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_w$, joista $\mathbf{v}_1$ on ensimmäinen sarake, $\mathbf{v}_2$ toinen sarake ja niin edelleen. Pääkomponentit $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_c,\ c \leq w$ saadaan tällöin kaavalla

$$\mathbf{p}_d = \mathbf{G}\mathbf{v}_d, \quad d = 1, ..., c.$$

Mallin parametrit β_0 ja β_j estimoidaan iteratiivisesti painotetun pienimmän neliösumman menetelmällä.

6 Simulointi

Luvussa 5 esiteltyjen menetelmien toimivuutta assosiaatioiden testaamisessa geno- ja fenotyyppien välillä tutkitaan simuloimalla. Tässä luvussa kerrotaan, miten simulointi on toteutettu. Kerrotaan simuloinnin toteutuksesta ensin lyhyesti pääpiirteittäin.

Simuloimalla muodostettiin 5000 erilaista aineistoa, jotka sisälsivät yksilöiden geno- ja fenotyyppitiedot. Osa aineistojen snipeistä määritettiin vaikuttavan fenotyyppiin ja aineistoihin sisällytettiin yksilöiden välistä sukulaisuutta. Perustuen tämän tutkielman aiempiin lukuihin, aineistoista pyrittiin luomaan sellaisia, mitä geno-fenotyyppiaineistot todellisuudessa ovat. Simuloiduille aineistoille toteutettiin assosiaatioiden testaus kullakin tutkittavalla menetelmällä. Menetelmien toimivuutta arvioitiin laskemalla, kuinka pienen osuuden menetelmät havaitsivat merkitsevänä snipeistä, joille ei ollut määritetty vaikutusta fenotyyppiin. Samoin arvioitiin myös, kuinka suuren osan kukin menetelmä havaitsi merkitsevänä snipeistä, joille oli määritetty vaikutus fenotyyppiin. Eli tilastollisen testaamisen termeillä sanoen toimivuutta arvioitiin sillä, kuinka vähän kukin menetelmä teki assosiaatiotesteissä tyypin 1 ja 2 virheitä.

Simulointi toteutettiin R-ohjelmointikielellä (R Core Team 2023) ja sen suoritus kesti noin 2 vuorokautta. Simuloinnin toteuttaneen tietokoneen tekniset tiedot ovat liitteessä 2 ja simulointikoodi on liitteessä 3. Seuraavaksi kerrotaan simuloinnin vaiheista yksityiskohtaisemmin.

6.1 Geno-fenotyyppiaineistojen simulointi

Simuloinnin ensimmäinen vaihe oli simuloida aineistoja, jotka sisälsivät yksilöiden geno- ja fenotyyppitiedot. Aineistoista pyrittiin tekemään todellisuuden kaltaisia geno-fenotyyppiaineistoja, jotta tutkielman tuloksia olisi mahdollista hyödyntää käytännössä GWAS-tutkimuksissa. Tutkielman simuloidut geno-fenotyyppiaineistot eivät kuitenkaan kata kaikenlaisia geno-fenotyyppiaineistoja, mitä GWAS-tutkimuksissa käytetään. Esimerkiksi kaikissa tämän tutkielman simuloiduissa geno-fenotyyppiaineistoissa fenotyyppi on jatkuva muuttuja, jolloin tutkielman tulokset eivät ole sovellettavissa binäärisiä fenotyyppimuuttujia koskevissa GWAS-tutkimuksissa. Lukijan on siis tärkeä ymmärtää, minkälaisia geno-fenotyyppiaineistoja simuloimalla tuotettiin. Geno-fenotyyppiaineistoja tuotettiin 5000 kappaletta, joista jokaisen simulointi toteutettiin seuraavaksi esiteltävällä tavalla.

Kerrotaan aluksi geno-fenotyyppiaineiston geneettisistä populaatioista ja läheisistä sukulaisista, koska niillä oli useassa kohtaa vaikutusta simuloinnin toteutukseen. Aineiston jokainen yksilö i kuului aina johonkin ja vain yhteen geneettiseen populaatioon k, $k = 1, ..., n_{\text{pop}}$, missä n_{pop} on aineiston geneettisten populaatioiden lukumäärä. Merkitään, että $k(i)$ on se geneettinen populaatio k, johon yksilö i kuuluu. Geneettisiä populaatiota otettiin aineistoon yksi kappale 30 % todennäköisyydellä ja enemmän kuin yksi 70 % todennäköisyydellä. Todennäköisyydet valittiin sen perusteella, että saatiin riittävän isot otoskoot kaikkiin tutkittaviin osa-aineistoihin. Se, että aineistossa on vain yksi geneettinen populaatio tarkoittaa siis sitä, että geno-fenotyyppiaineistossa ei ole geneettisiä populaatiorakenteita. Jos geneettisten populaatioiden määrä n_{pop} oli enemmän kuin yksi, se simuloitiin diskreetistä tasajakaumasta $\text{DU}(2, 10)$. Merkintä $\text{DU}(\lambda_1, \lambda_2)$ tarkoittaa tasajakaumaa joukossa $\{\lambda_1, \lambda_1 + 1, \lambda_1 + 2, \ldots, \lambda_2 - 1, \lambda_2\}$, missä λ_1 ja λ_2 ovat kokonaislukuja ja $\lambda_2 > \lambda_1$. Geneettisten populaatioiden lukumäärä rajoitettiin maksimissaan kymmeneen simuloinnin nopeuden takia.

Jotkin aineistojen yksilöistä olivat läheistä sukua toisilleen. Assosiaatioiden testaamisessa ei kuitenkaan hyödynnetä tietoa siitä, mitkä yksilöt ovat toisilleen läheistä sukua, jolloin ne luovat aineistoon kryptistä sukulaisuutta. Simuloinnissa oletetaan, että kaikki toisilleen läheistä sukua olevat yksilöt kuuluvat samaan geneettiseen populaatioon. Aluksi valittiin otetaanko läheisiä sukulaisia ollenkaan aineistoon vai ei. Läheisiä sukulaisia otettiin aineiston mukaan 70 % todennäköisyydellä. Todennäköisyys valittiin jälleen sen perusteella, että saatiin riittävän isot otoskoot kaikkiin tutkittaviin osa-aineistoihin.

Jos läheisiä sukulaisia otettiin mukaan aineistoon, niiden kokonaismäärä n_{krypt} simuloitiin diskreetistä tasajakaumasta $\text{DU}(2, 400)$, minkä jälkeen ne jaettiin satunnaisesti geneettisiin populaatioihin. Välillä satunnaisuuden takia geneettiseen populaatioon päätyi vain yksi läheinen sukulainen, mikä ei kuitenkaan ole mahdollista, koska geneettisessä populaatiossa täytyy olla vähintään yksi toinen yksilö, jolle olla sukua. Näissä tilanteissa geneettisen populaation läheisten sukulaisten määräksi asetettiin kaksi, mikä kasvatti myös koko geno-fenotyyppiaineiston läheisten sukulaisten määrää n_{krypt} yhdellä. Kunkin geneettisen populaation läheisistä sukulaisista muodostettiin h ryhmää, joiden yksilöillä oli mahdollisuus olla läheisiä sukulaisia vain keskenään. Ryhmien $r = 1, \ldots, h-1$ koot n_r simuloitiin diskreetistä tasajakaumasta $\text{DU}(2, 20)$ ja ryhmän h koko n_h saatiin vähentämällä geneettisen populaation läheisten sukulaisten kokonaismäärästä ryhmien $r = 1, \ldots, h-1$ kokojen n_r summa. Rajaus maksimissaan 20 yksilön kokoisiin ryhmiin oli subjektiivinen valinta. Jotta vältyttiin yhden ja 21 yksilön kokoisilta ryhmiltä niin ryhmän $r = h-1$ koko n_r simuloitiin uudelleen aina, jos ennen uusintaa ryhmien $r = 1, \ldots, h-1$ kokojen n_r summa oli yhden pienempi kuin geneettisen populaation läheisten sukulaisten kokonaismäärä ja geneettisen populaation läheisten sukulaisten kokonaismäärän ja ryhmien $r = 1, \ldots, h-2$ kokojen n_r erotus oli 21.

Läheisten sukulaisten simuloinnin jälkeen geneettisiin populaatioihin lisättiin yksilöitä, jotka eivät ole läheistä sukua kenellekään aineistosta. Niiden kokonaismäärä simuloitiin diskreetistä tasajakaumasta $\text{DU}(500 - n_{\text{krypt}}, 1000 - n_{\text{krypt}})$, jolloin otoskooksi n saatiin 500–1000 yksilöä, mikä oli hyvä suuruus simuloinnin tehokkuuden ja tarkkojen tulosten saamisen kannalta. Yksilöt, jotka eivät olleet sukua kenellekään toiselle aineiston yksilölle, jaettiin satunnaisesti geneettisiin populaatioihin siten, että niitä oli jokaisessa vähintään 1.

Jos läheisiä sukulaisia ei otettu mukaan aineistoon niin aineiston yksilöiden kokonaismäärä n simuloitiin diskreetistä tasajakaumasta $\text{DU}(500, 1000)$ ja yksilöt jaettiin geneettisiin populaatioihin siten, että jokaiseen geneettiseen populaatioon tuli vähintään yksi yksilö.

Siitä, miten geneettisiä populaatioita ja läheisiä sukulaisia tarkalleen esiintyy todellisissa GWAS-tutkimusten aineistoissa, ei ole tarkkaa tietoa saatavilla ja aineistojen rakenteet vaihtelevat sukulaisuussuhteiden osalta tutkimusten välillä. Sen vuoksi geneettiset populaatiot ja läheiset sukulaiset määritettiin geno-fenotyyppiaineistoihin yllä mainitulla tavalla, jotta saataisiin mahdollisimman monenlaisia rakenteita aineistoon sukulaisuussuhteiden kannalta. Rakenteiden vaihtelevuuden ansiosta sukulaisuuden vaikutuksia assosiaatiotestien tuloksiin pystyttiin tutkimaan monella tapaa. Lisäksi simuloinnissa oli tärkeää, että kryptisten sukulaisten ja geneettisten populaatioiden määrä simuloitiin toisistaan riippumattomasti, jotta niiden vaikutuksia assosiaatiotestaukseen voidaan tutkia erikseen. Esitetään seuraavaksi kaava, jolla yksilöiden fenotyyppien arvot määritettiin. Sen

jälkeen kerrotaan tarkasti, miten kaavan komponentit määritettiin ja miksi ne määritettiin kyseisellä tavalla.

Tämän tutkielman simuloiduissa geno-fenotyyppiaineistoissa yksilön i fenotyypin y_i arvot määritettiin kaavalla

$$y_i = \sum_{j=1}^{n_{\text{snp}}} s_{ij} b_j + a_i + e_i,\ i = 1, \dots n$$

Summa $\sum_{j=1}^{n_{\text{snp}}} s_{ij} b_j$ on yksilön genotyyppien yhteenlaskettu vaikutus fenotyyppiin. Kerrotaan seuraavaksi komponenttien $n_{\text{snp}},\ s_{ij}$ ja b_j määrittäminen ja edelleen komponenttien a_i ja e_i määrittäminen.

Testattavien snippien lukumäärä n_{snp} simuloitiin diskreetistä tasajakaumasta $\text{DU}(3000, 6000)$. Testattavien snippien lukumäärä on huomattavasti pienempi kuin monissa GWAS-tutkimuksissa, mutta simulointia ei olisi ollut mahdollista toteuttaa paljoa suuremmalle snippien määrälle simuloinnin hitauden takia. Vähäinen testattavien snippien määrä ei kuitenkaan tehnyt menetelmien toimivuuden arvioinnista huonompaa, koska tuloksia tarkasteltiin suhteessa testattavien snippien lukumäärään ja snippien genotyypit simuloitiin toisistaan riippumattomasti.

Symboli s_{ij} on yksilön i snipin j genotyypin arvo. Genotyyppien simulointi toteutettiin kahdella eri tavalla riippuen siitä, onko yksilö läheinen sukulainen toiselle aineiston yksilölle vai ei. Käsitellään ensin niiden yksilöiden genotyyppien simulointi, jotka eivät ole läheistä sukua toiselle aineiston yksilölle. Näiden yksilöiden tapauksessa yksilön i snippien ensimmäinen ja toinen alleeli simuloitiin binomijakaumasta $\text{Bin}(1, f_{jk(i)})$, missä $f_{jk(i)}$ on snipin j alleelifrekvenssi geneettisessä populaatiossa $k(i)$. Genotyypit laskettiin summaamalla kunkin snipin alleelit yhteen. Geneettisten populaatioiden snippien alleelifrekvenssit simuloitiin ensimmäiselle geneettiselle populaatiolle $k = 1$ tasajakaumasta $\text{U}(0.05, 0.95)$. Muiden geneettisten populaatioiden $k > 1$ snippien alleelifrekvenssit simuloitiin katkaistusta normaalijakaumasta $\text{TN}(f_{j1}, \sigma_g^2, 0.05, 0.95)$, missä f_{j1} on jakauman sijaintiparametri, σ_g on hajontaparametri, ja 0.05 ja 0.95 ovat katkaisun rajat. Hajontaparametri σ_g simuloidaan tasajakaumasta $\text{U}(0.1, 0.3)$. Alleelifrekvenssit rajoitettiin 0.05–0.95 välille, jotta saatiin riittävästi vaihtelua yksilöiden genotyyppien arvoille.

Niiden yksilöiden, jotka eivät ole läheistä sukua toiselle aineiston yksilölle, genotyypit simuloitiin yllä esitetyllä tavalla perustuen lukuun 4.1. Luvussa 4.1 mainittiin, että geneettisten populaatioiden snippien alleelifrekvenssit vaihtelevat populaatioiden välillä, joten käyttämällä geneettisten populaatioiden $k > 1$ alleelifrekvenssien simuloinnissa sijaintiparametrina geneettisen populaation $k = 1$ alleelifrekvenssiä f_{1j} ja satunnaistamalla hajontaparametri σ_g, saatiin erisuuruista alleelifrekvenssivaihtelua geneettisten populaatioiden välille.

Aineiston läheisten sukulaisten genotyypit simuloitiin toisella tavalla, koska läheisten sukulaisten genotyypit eivät ole toisistaan riippumattomia. Tämän luvun alussa kerrottiin, miten geno-fenotyyppiaineiston yksilöt jaettiin ryhmiin $r,\ r = 1, \dots, h$. Kunkin ryhmän r

yksilöiden genotyypit simuloitiin seuraavalla tavalla. Merkitään ryhmän r yksilöä kirjaimella $t,\ t = 1, ..., n_r$ ja sitä geneettistä populaatiota, johon ryhmän yksilöt kuuluvat kirjaimella l. Ryhmän r ensimmäisen yksilön $t = 1$ snippien ensimmäinen ja toinen alleeli simuloitiin binomijakaumasta $\text{Bin}(1, f_{jl})$. Osuus $1 - q_t$ ryhmän r muiden yksilöiden $t > 1$ alleeleista simuloitiin vastaavasta jakaumasta, mutta osuuden q_t verran alleeleita kopioitiin ryhmän r satunnaisesti valitulta yksilöltä $d,\ d < t$. Osuus q_t simuloitiin tasajakaumasta $\text{U}(0.05, 0.60)$. Genotyypit laskettiin summaamalla kunkin snipin alleelit yhteen.

Aineistojen kryptiset sukulaiset simulointiin luvun 4.2 perusteella. Luvussa 4.2 todettiin, että läheistä sukua olevat yksilöt perivät samoja alleeleja yhteisiltä esivanhemmiltaan, minkä vuoksi heidän genominsa ovat suuremmalta osin identtisiä verrattuna yksilöihin, jotka eivät ole läheistä sukua toisilleen. Tämän vuoksi osuus q_t yksilön alleelista kopioitiin toisen yksilön alleeleista tehden niistä lähisukulaisia. Luvussa 4.2 todettiin myös, että lukuun ottamatta identtisiä kaksosia, keskimääräisesti suurin IBD-osuus yksilöiden genomeista on sisaruksilla sekä vanhemmalla ja lapsella. Sisarusten sekä vanhemman ja lapsen keskimääräiset IBD-osuudet ovat yleensä 37.4 % ja 61.7 % välillä ja niitä kaukaisemmissa sukulaisuussuhteissa IBD-osuus laskee. Tämän perusteella osuus q_t simuloitiin tasajakaumasta $\text{U}(0.05, 0.60)$. Alaraja 0.05 osuudelle q_t oli subjektiivinen valinta. Identtisiä kaksosia ei simuloitu aineistoon.

Parametri b_j on snipin j additiivinen vaikutus fenotyyppiin. Eli snipin j genotyypin 0 vaikutus eroaa genotyypistä 1 yhden b_j verran ja genotyypistä 2 kahden b_j verran. Fenotyyppiin vaikuttavien b_j kertoimien lukumäärä simuloitiin diskreetistä tasajakaumasta $\text{DU}(1, 10)$ ja suuruus tasajakaumasta $\text{U}(2, 5)$. Vaikuttamattomien snippien kertoimet b_j asetettiin nollaksi.

Parametri a_i lisää yksilölle i geneettisen populaation $k(i)$ genomin ulkopuoliset vaikutukset fenotyyppiin. Yksilölle i vaikutuksen a_i suuruus simuloitiin normaalijakaumasta $\text{N}(\mu_{k(i)}, \sigma^2_{k(i)})$. Odotusarvo μ_k simuloitiin kullekin geneettiselle populaatiolle k tasajakaumasta $\text{U}(-10, 10)$ ja keskihajonta σ_k simuloitiin tasajakaumasta $\text{U}(1, 3)$. Jokaiselle yksilölle i lisättiin vaikutus a_i, koska luvussa 4.1 geneettisten populaatioiden todettiin jakautuvan usein maantieteellisesti samoihin paikkoihin, minkä vuoksi on syytä olettaa, että ainakin ympäristön vaikutukset fenotyyppiin ovat usein keskimäärin samankaltaisempia samaan geneettiseen populaatioon kuuluvilla yksilöillä verrattuna toisen geneettisen populaation yksilöihin. Erojen suuruus geneettisten populaatioiden välillä satunnaistettiin satunnaistamalla μ_k, ja yksilöiden välisten erojen suuruus geneettisten populaatioiden sisällä satunnaistettiin satunnaistamalla σ_k.

Parametri e_i kuvaa yksilötason vaihtelua ja sen arvo simuloitiin normaalijakaumasta $\text{N}(0, \sigma^2_e)$, jolle σ_e simuloidaan tasajakaumasta $\text{U}(1, 3)$.

6.2 Testaaminen ja simulointiaineisto

Geno-fenotyyppiaineistojen muodostamisen jälkeen geno- ja fenotyyppien assosiaatioita testattiin LMM-, LMP- ja LM-menetelmillä. Testien tuloksista ja simuloinnista kerättiin talteen tietoja simulointiaineistoon jälkianalyysejä varten. Seuraavaksi kerrotaan assosiaatioiden testaamisen toteutuksesta kullakin menetelmällä ja simulointiaineistosta.

Assosiaationtestaus LMM-menetelmällä toteutettiin hyödyntämällä R-pakettia *statgenGWAS* (van Rossum ja Kruijer 2022). Paketista hyödynnettiin funktioita *kinship*, *createGData*, *runSingleTraitGwas*. Funkiolla *kinship* muodostettiin sukulaisuusmatriisi $\mathbf{K}$, *createGData*-funktiolla muokattiin testattava aineisto *runSingleTraitGwas*-funktiolle sopivaan muotoon ja *runSingleTraitGwas*-funktiolla suoritettiin testit. LMP- ja LM-menetelmien assosiaationtestaus toteutettiin R:n funktiolla *glm* ja kovariaatteina käytettävät pääkomponentit muodostettiin LMP-menetelmässä R:n funktiolla *prcomp*. Kovariaatteina käytettävien pääkomponenttien määrä simuloitiin diskreetistä tasajakaumasta $\mathrm{DU}(1, 10)$. Pääkomponenttien määrä satunnaistettiin, jotta voitiin tutkia myös niiden määrän vaikutusta LMP-menetelmän toimivuuteen.

Simuloinnista muodostettiin simulointiaineisto, johon kerättiin tietoja kultakin simulointikierrokselta. Simulointiaineisto sisälsi tietoa assosiaatiotestien tyypin 1 ja 2 virheiden määrästä. Virheiden määrä laskettiin jokaisella simulointikierroksella suhteessa niiden teoreettiseen maksimimäärään, mikä on tyypin 1 virheiden tapauksessa vaikuttamattomien snippien määrä ja tyypin 2 virheiden tapauksessa vaikuttavien snippien määrä. Lisäksi simulointiaineistoon kerättiin geneettisten populaatioiden lukumäärät, kryptisten sukulaisten lukumäärät ja kovariaatteina käytettyjen pääkomponenttien määrät.

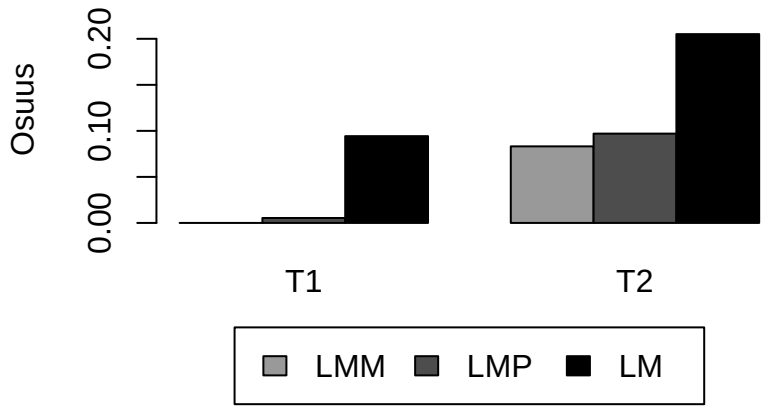
6.3 Simulointiaineiston analysointi

Menetelmien vertailuun valittiin tunnusluvuksi simulointiaineiston tyypin 1 ja 2 virheiden osuuksien keskiarvot. Käytetään jatkossa tyypin 1 virheiden osuuden keskiarvosta lyhennettä T1 ja tyypin 2 virheiden osuuden keskiarvosta T2. Simulointiaineistosta laskettiin T1- ja T2-osuuksia koko simulointiaineistolle ja sen osa-aineistoille. Sukulaisuussuhteiden vaikutukset assosiaatiotestien tuloksiin olivat suurimman mielenkiinnon kohteena, joten niiden selvittämiseksi T1- ja T2-osuuksia laskettiin geneettisten populaatioiden ja kryptisten sukulaisten lukumäärien määräämissä osa-aineistoissa.

Käytettäessä LMP-menetelmää, tutkijan täytyy valita montako pääkomponenttia malliin otetaan mukaan kovariaatiksi. Siihen kuinka monta pääkomponenttia kannattaa ottaa mukaan malliin ei ole olemassa yleispätevää vastausta, joten sen vuoksi osuuksia laskettiin myös kovariaatteina käytettyjen pääkomponenttien määrään perustuvissa osa-aineistoissa.

7 Tulokset

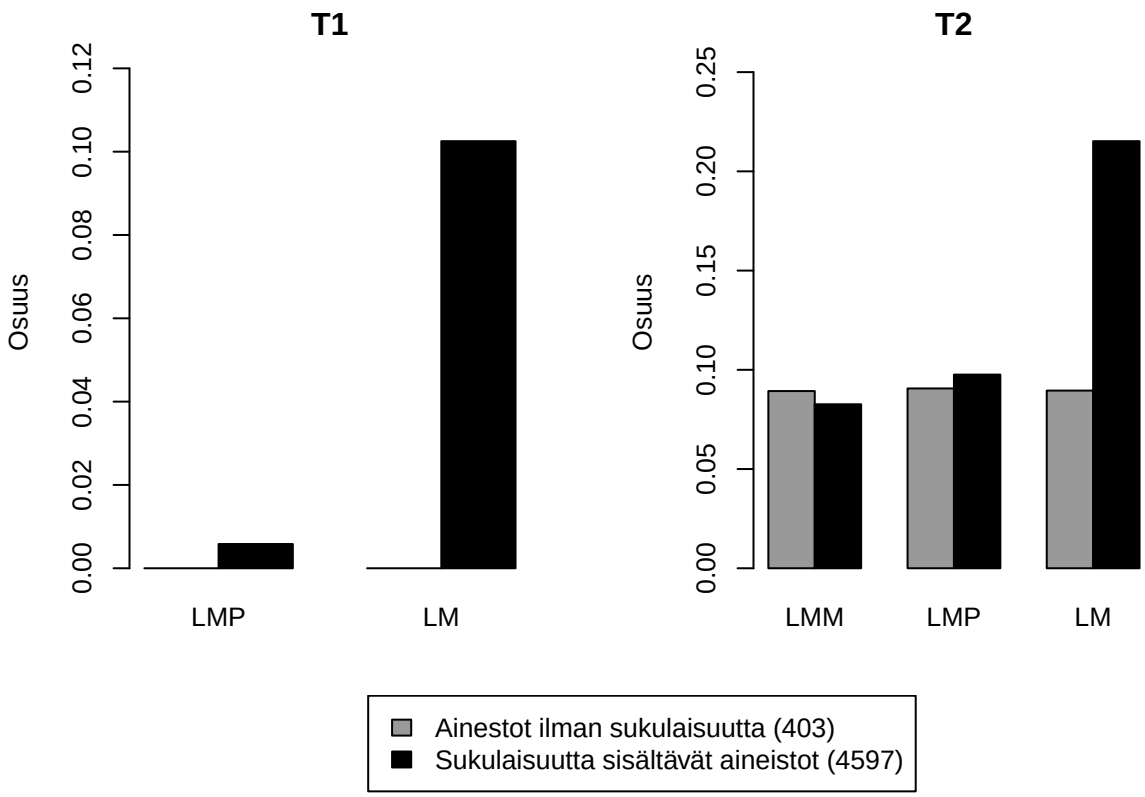
Tässä luvussa kerrotaan simuloinnin tulokset. Kunkin menetelmän paremmuutta arvioitiin siten, miten hyvin ne minimoivat T1- ja T2-osuuksia. Osuudet laskettiin koko simulointiaineistosta ja edellisessä luvussa esitetyissä osa-aineistoista. Tulokset esitetään pylväsdiagrammeilla. Kuvasta 7 nähdään koko simulointiaineistosta lasketut T1- ja T2-osuudet.



Kuva 7: Koko simulointiaineistosta lasketut T1- ja T2-osuudet LMM-, LMP- ja LM-menetelmille.

Kuvasta 7 nähdään, että LM tuotti selvästi eniten tyypin 1 ja 2 virheitä. LMP-menetelmän kohdalla T1- ja T2-osuudet olivat vain hieman suurempia verrattuna LMM-menetelmän osuuksiin. LMM ei tuottanut ollenkaan tyypin 1 virheitä ja myös tyypin 2 virheiden määrä oli pienin. Se, että simuloinnissa LMM ei tuottanut yhtään tyypin 1 virhettä ei pois sulje sitä, etteikö niitä voisi tulla, jos geno-fenotyyppi aineistoja simuloitaisiin lisää. Kuitenkin nyt koko simulointiaineistosta laskettujen T1- ja T2-osuuksien perusteella voisi siis sanoa, että LMM olisi paras, LMP toiseksi paras ja LM huonoin menetelmä. LMP-menetelmän toteutus kuitenkin vaihtelee simulointikierrosten välillä, koska kovariaatteina käytettyjen pääkomponenttien lukumäärä valittiin satunnaisesti. Myöhemmin tässä luvussa huomataan, että valitsemalla oikein käytettävien pääkomponenttien määrän, LMP ei tee myöskään yhtään tyypin 1 virhettä ja tyypin 2 virheet ovat LMM-menetelmän tasoa.

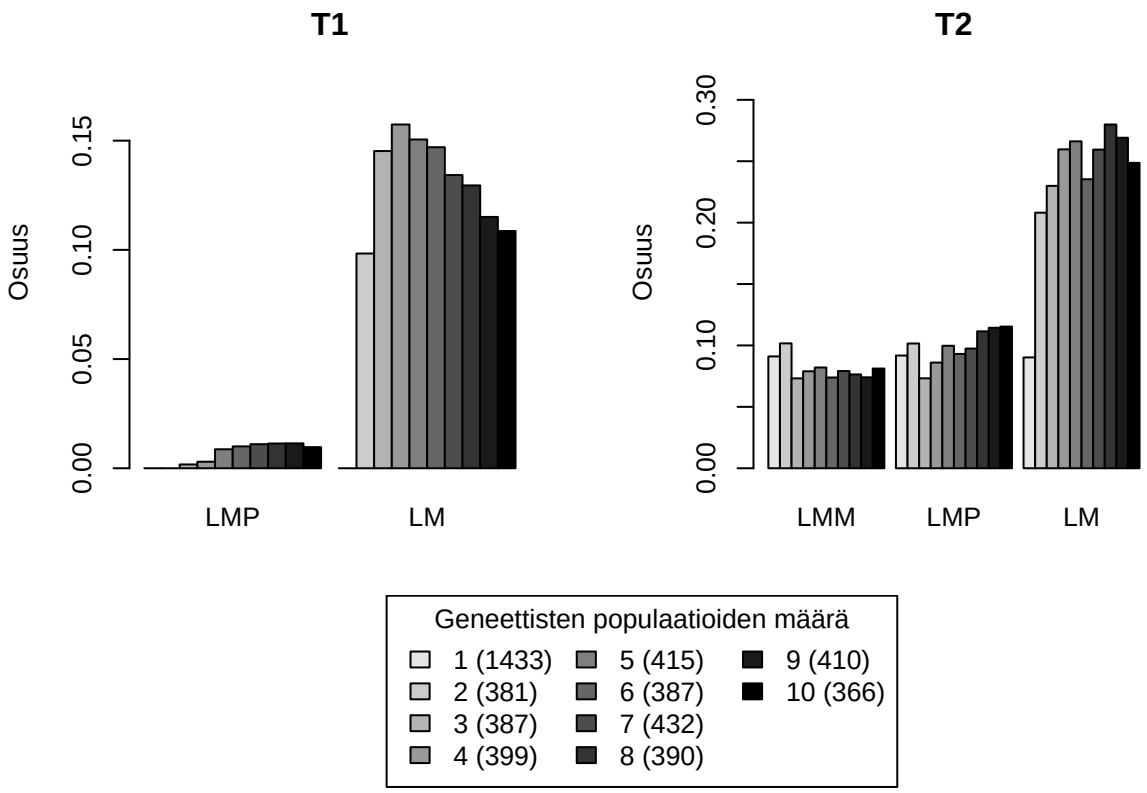
T1- ja T2- osuuksien laskeminen koko simulointiaineistoista ei kuitenkaan kerro siitä, miten hyvin menetelmät huomioivat aineiston yksilöiden väliset sukulaisuussuhteet. Selvitetään siis seuraavaksi vaikuttavatko sukulaisuussuhteet testien tuloksiin ja miten hyvin kukin menetelmä huomioi niitä. Tämän selvittämiseksi verrataan ensin T1- ja T2-osuuksia sukulaisuutta sisältävien ja sisältämättömien aineistojen välillä (Kuva 8). Sukulaisuutta sisältämätön aineisto tarkoittaa siis aineistoa, jossa on vain yksi geneettinen populaatio ja ei yhtään kryptistä sukulaista. Sukulaisuutta sisältävä aineisto tarkoittaa puolestaan taas aineistoa, jossa on joko enemmän kuin yksi geneettinen populaatio tai vähintään yksi kryptinen sukulainen. LMM-menetelmän T1-osuuksia ei sisällytetä muihin tämän luvun kuviin, koska kuvan 7 perusteella tiedetään jo, että ne olisivat 0. Siitä voi siis päätellä että, jos sukulaisuus ylipäätään vääristää tuloksia niin LMM osaa käsitellä sen erittäin hyvin tyypin 1 virheiden osalta.



Kuva 8: T1- ja T2-osuudet sukulaisuutta sisältävissä ja sisältämättömissä aineistoissa LMM-, LMP- ja LM-menetelmillä. Suluissa kerrotaan kuinka monen simulointikierroksen assosiaatiotestien tuloksista T1- ja T2-osuudet on laskettu.

Kuvan 8 mukaan vaikuttaa siltä, että tyypin 1 virheet johtuvat vain sukulaisuussuhteista, koska niitä ilmeni vain testattaessa sukulaisuutta sisältäviä aineistoja. Tyypin 2 virheet näyttävät kasvavan sukulaisuuden takia, jos sukulaisuutta ei huomioida testaamisessa, koska LM-menetelmän T2-osuus on huomattavasti suurempi sukulaisuutta sisältävissä aineistoissa. LMP-menetelmän kohdalla sukulaisuus kasvattaa T2-osuutta vain hieman ja LMM-menetelmän kohdalla laskee sitä. Kuvan 8 mukaan ja sen perusteella, että LMM ei tuota ollenkaan tyypin 1 virheitä, LMM käsittelee sukulaisuuden parhaiten, LMP toiseksi parhaiten ja LM huonoiten.

Sukulaisuus koostui simuloinnin geno-fenotyyppiaineistoissa geneettisistä populaatioista ja kryptisistä sukulaisista. Tutkitaan T1- ja T2-osuuksia seuraavaksi erikseen geneettisten populaatioiden ja kryptisten sukulaisten määrän mukaisissa osa-aineistoissa. Kuvasta 9 nähdään T1- ja T2-osuudet geneettisten populaatioiden määrän mukaisissa osa-aineistoissa.

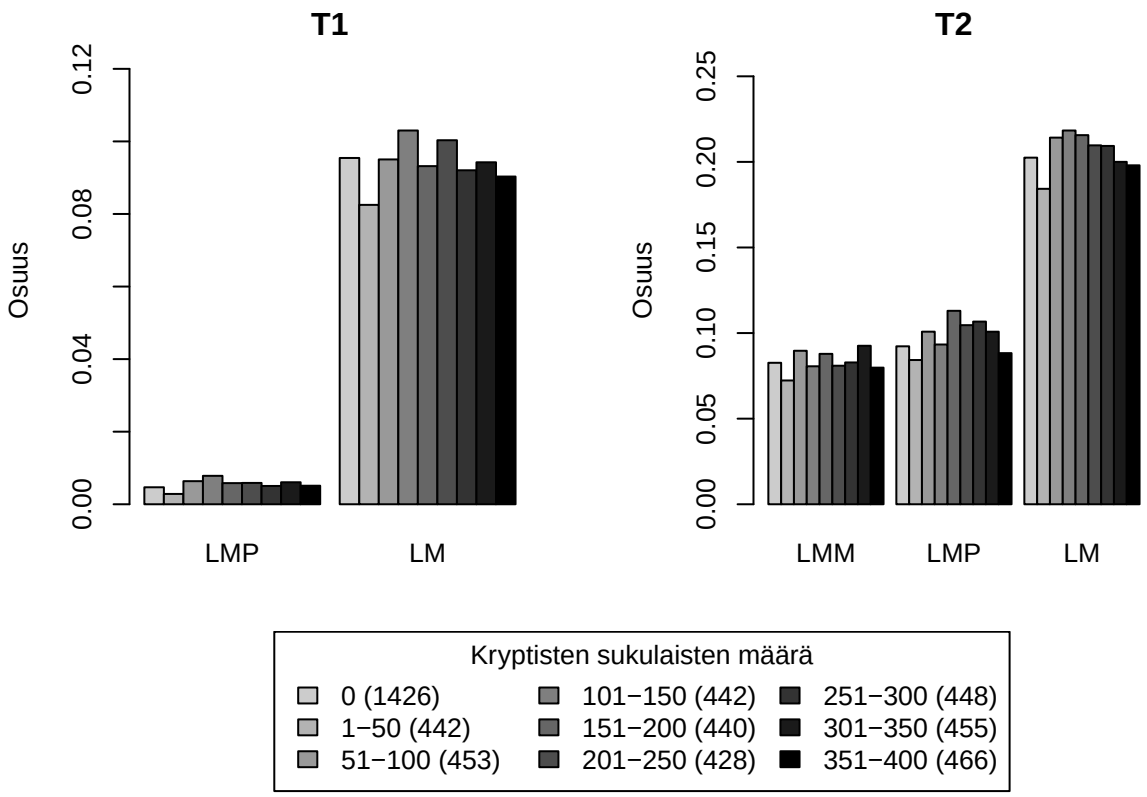


Kuva 9: T1- ja T2-osuudet geneettisten populaatioiden lukumäärän mukaisissa osa-aineistoissa LMM-, LMP- ja LM-menetelmillä. Suluissa kerrotaan kuinka monen simulointikierroksen assosiaatiotestien tuloksista T1- ja T2-osuudet on laskettu.

LMP-menetelmän T1-osuudet kasvavat geneettisten populaatioiden lukumäärän kasvaessa. LM-menetelmän tapauksessa yhdestä geneettisestä populaatiosta neljään tyypin 1 virheiden määrä kasvaa, mutta neljästä geneettisestä populaatiosta kymmeneen nähdään puolestaan laskua T1 osuuksissa. Olisi mielenkiintoista selvittää laskisivatko tyypin 1 virheet edelleen, jos geneettisten populaatioiden määrää vielä kasvatettaisiin, mutta tässä simuloinnissa rajoituttiin kymmeneen geneettiseen populaatioon, jotta saatiin simuloinnista tarpeeksi nopea.

Geneettisten populaatioiden määrällä ei näytä olevan selkeää vaikutusta T2-osuuksiin LMM-menetelmän kohdalla. LMP-menetelmän kohdalla T2-osuus näyttää kasvavan hieman kolmesta geneettisestä populaatiosta kymmeneen. LM-menetelmän kohdalla yhdestä geneettisestä populaatiosta viiteen T2-osuudet kasvavat selvästi, mutta sen jälkeen vaikutus on epäselvempää.

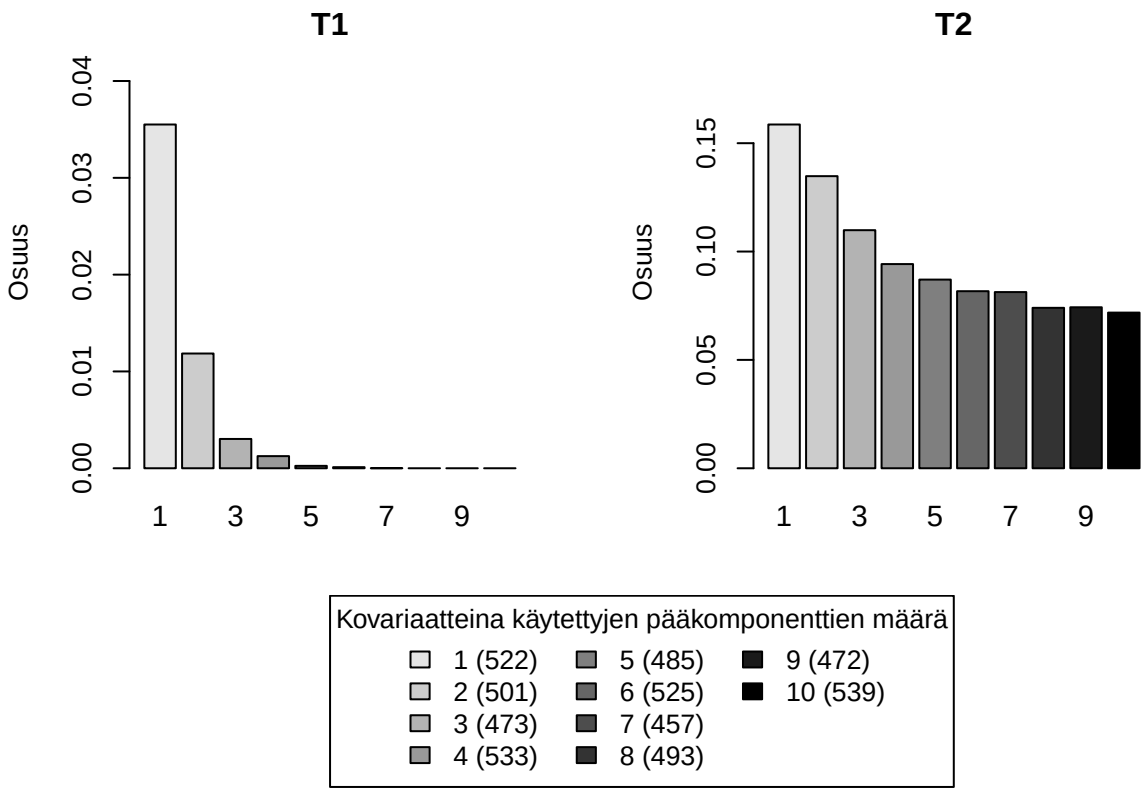
Tarkastellaan seuraavaksi kryptisten sukulaisten määrän vaikutuksia T1- ja T2-osuuksiin. Kryptisten sukulaisten määriä tarkastellaan 50 kryptisen sukulaisen välein muodostetuissa luokissa (Kuva 10).



Kuva 10: T1- ja T2-osuudet kryptisten sukulaisten lukumäärän mukaisissa osa-aineistoissa LMM-, LMP- ja LM-menetelmillä. Suluissa kerrotaan kuinka monen simulointikierroksen assosiaatiotestien tuloksista T1- ja T2-osuudet on laskettu.

Kryptisten sukulaisten määrällä ei näytä olevan vaikutusta tyypin 1 tai 2 virheisiin minkään menetelmän kohdalla.

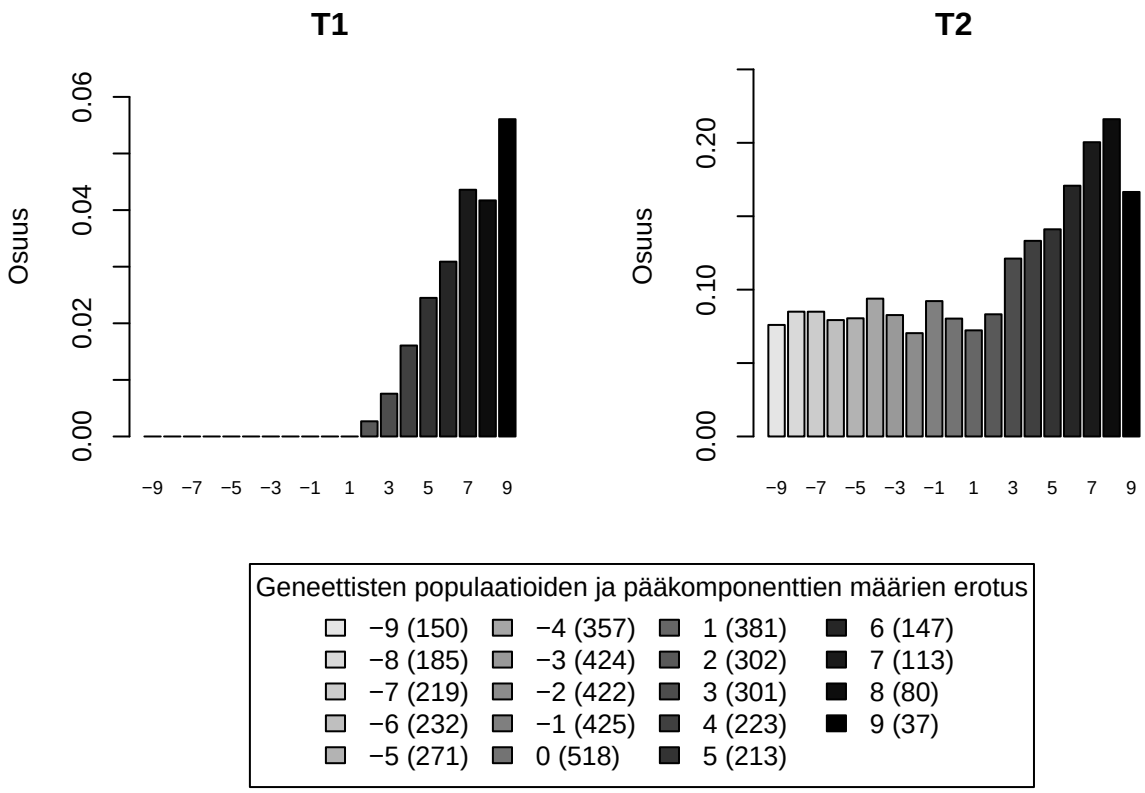
Tutkitaan vielä kovariaatteina käytettyjen pääkomponenttien lukumäärän vaikutusta virheiden määrään. Tähän mennessä LMP-menetelmän T1- ja T2-osuudet on laskettu osa-aineistoissa, joissa pääkomponenttien määrä on vaihdellut satunnaisesti 1–10 pääkomponentin välillä. Se ei kuitenkaan ole paras tapa arvioida LMP-menetelmän toimivuutta, sillä todellisuudessa tutkija ei arvo satunnaisesti pääkomponenttien määrää vaan voi valita sen itse. Seuraavaksi yritetään siis selvittää, onko mahdollista valita optimaalista pääkomponenttien määrää tyypin 1 ja 2 virheiden minimoimiseksi käytettäessä LMP-menetelmää. Kuvasta 11 nähdään T1- ja T2-osuudet kovariaatteina käytettyjen pääkomponenttien lukumäärän mukaisissa osa-aineistoissa.



Kuva 11: LMP-menetelmän T1- ja T2-osuudet kovariaatteina käytettyjen pääkomponenttien lukumäärän mukaisissa osa-aineistoissa. Suluissa kerrotaan kuinka monen simulointikierroksen assosiaatiotestien tuloksista T1- ja T2-osuudet on laskettu.

Kovariaatteina käytettävien pääkomponenttien määrän kasvattaminen vähentää selvästi T1- ja T2-osuuksia. Tyypin 1 virheet laskevat huomattavasti noin 3.5 %:sta aina 0 %:iin asti 1–10 pääkomponentin välillä. Tyypin 2 virheet laskevat myös selvästi 1–10 pääkomponentin välillä.

Kovariaatteina käytettyjen pääkomponenttien määrällä on siis selvästi vaikutusta virheiden määrään käytettäessä LMP-menetelmää. Kuvasta 9 nähtiin, että myös geneettisten populaatioiden lukumäärä vaikutti virheiden määrään. Tutkitaan sen vuoksi vielä geneettisten populaatioiden ja kovariaatteina käytettävien pääkomponenttien määrän erotuksen vaikutuksia virheisiin (Kuva 12).



Kuva 12: LMP-menetelmän T1- ja T2-osuudet geneettisten populaatioiden ja kovariaatteina käytettyjen pääkomponenttien lukumäärän erotuksen mukaisissa osa-aineistoissa. Suluissa kerrotaan kuinka monen simulointikierroksen assosiaatiotestien tuloksista T1- ja T2-osuudet on laskettu.

LMP ei tuota ollenkaan tyypin 1 virheitä, jos geneettisten populaatioiden ja pääkomponenttien määrien erotus on pienempi kuin 2. Erotuksesta 2 alkaen tyypin 1 virheiden määrä taas kasvaa huomattavasti. T2-osuudet eivät juurikaan muutu erotuksesta -9 erotukseen 2, mutta kasvavat paljon erotuksien 2–9 välillä. Molempien virheiden minimoimiseksi pääkomponentteja tulisi siis kuvan 12 mukaan valita suunnilleen saman verran kuin geno-fenotyyppiaineistossa on geneettisiä populaatioita tai sitä enemmän.

8 Pohdinta

Tutkielmassa pyrittiin selvittämään miten hyvin LMM-, LMP- ja LM-menetelmät käsittelevät geno-fenotyyppiaineistoissa olevien yksilöiden sukulaisuussuhteita assosiaatiotestauksessa GWAS-tutkimuksissa. Menetelmien paremmuuden mittarina käytettiin tyypin 1 ja 2 virheiden minimointia. Lisäksi selvitettiin tekijöitä, jotka vaikuttivat virheiden määrään ja pyrittiin selvittämään LMP-menetelmässä kovariaatteina käytettävien pääkomponenttien optimaalista määrää virheiden minimoimiseksi. Näihin ongelmiin saatiin hyviä vastauksia,

mutta ennen niiden pohdintaa on hyvä pohtia ensin simuloinnin luotettavuutta, koska sillä on vaikutusta tutkielman kaikkiin tuloksiin.

Simulointi oli hyvä tapa lähestyä tutkimusongelmaa seuraavista syistä. Simuloimalla geno-fenotyyppiaineistot pystyttiin tietämään kaikki aineiston ominaisuudet tarkasti. Tiedettiin esimerkiksi mitkä snipit vaikuttivat fenotyyppiin ja mitkä eivät, jolloin T1- ja T2-osuudet pystyttiin laskemaan tarkasti. Aineiston geneettisten populaatioiden ja läheisten sukulaisten määrät olivat tiedossa, jolloin niiden määrien vaikutusta oli mahdollista tutkia. Lisäksi tunnettiin muuttujien väliset riippuvuussuhteet, minkä ansiosta tiedettiin, että geneettisten populaatioiden ja kryptisten sukulaisten määrät simuloidussa geno-fenotyyppiaineistoissa olivat täysin satunnaisia riippumatta muista muuttujista. Sen ansiosta voitiin olla varmoja, että populaatioiden ja kryptisten sukulaisten määrien mahdolliset vaikutukset tyypin 1 ja 2 virheiden määriin johtuivat juuri niiden määrästä, eivätkä jostain muusta sekoittavasta tekijästä. Lisäksi simulointi mahdollisti tarpeeksi suuren otoskoon tarkkojen tulosten saamiseksi.

Simuloinnin huono puoli oli puolestaan se, että ei voitu tietää tarkasti, kuinka hyvin simuloidut geno-fenotyyppiaineistot vastasivat oikeissa GWAS-tutkimuksissa käytettäviä geno-fenotyyppiaineistoja. Simuloitujen geno-fenotyypiaineistojen tulisi olla sen kaltaisia, joita todellisissa GWAS-tutkimuksissa on käytetty tai tullaan käyttämään, jotta tutkielman tuloksia voisi soveltaa käytännössä. Todellisuuden kaltaisia geno-fenotyyppiaineistoja pyrittiin simuloimaan aiempien tutkimusten perusteella, jotka käsittelivät geno-fenotyyppiaineistojen ominaisuuksiin vaikuttavia tekijöitä, kuten sukulaisuussuhteiden vaikutuksia ihmisen perimään ja geneettistä vaihtelua ihmisten välillä. Kaikista ominaisuuksista ei kuitenkaan ollut saatavilla tarkkaa tietoa, mikä tuotti epävarmuutta simuloinnin luotettavuuteen.

Geno-fenotyyppiaineistojen ominaisuuksiin vaikuttavia parametrejä satunnaistettiin, jotta saatiin paljon erilaisia geno-fenotyyppiaineistoja, jolloin tutkielman tuloksia on mahdollista soveltaa laajemmin. Joidenkin parametrien arvoja kuitenkin rajoitettiin simuloinnissa ja sen seurauksena geno-fenotyyppiaineistot eivät kattaneet kaikenlaisia geno-fenotyyppiaineistoja, mitä todellisuudessa voi olla. Esimerkiksi geneettisten populaatioiden määrää rajoitettiin, minkä takia tutkimustulokset eivät sovellu rajojen ulkopuolelle jääviin geno-fenotyyppiaineistoihin välttämättä niin hyvin. Tuloksia tulkitessa on siis tärkeää tietää minkälaisia geno-fenotyyppiaineistoja simulointi tuotti.

Simulointiin liittyi siis asioita, jotka vähentävät tulosten luotettavuutta. Käsitellään seuraavaksi tuloksia kuitenkin olettaen, että simulointi on hyvin toteutettu. Tarkkaan ottaen tulokset kuitenkin pätevät varmuudella hyvin vain simuloinnissa tuotettuihin geno-fenotyyppiaineistoihin.

Tutkielman tulokset osoittivat, että yksilöiden väliset sukulaisuussuhteet aiheuttavat tyypin 1 ja 2 virheitä ja LMM- ja LMP-menetelmät vähentävät niitä LM-menetelmään verrattuna. Se, että sukulaisuus aiheuttaa vääristymiä tuloksiin osoitettiin sillä, että LM-menetelmä, joka ei huomioi sukulaisuutta, tuotti sukulaisuutta sisältävissä aineistoissa huomattavasti enemmän tyypin 1 ja 2 virheitä verrattuna aineistoihin, jotka eivät sisältäneet sukulaisuutta. Se, että LMM- ja LMP-menetelmät korjasivat sukulaisuuden aiheuttamia virheitä, osoitettiin sillä, että sukulaisuutta sisältävissä aineistoissa T1- ja T2-osuudet olivat

LMM- ja LMP-menetelmillä huomattavasti pienemmät verrattuna LM-menetelmään. LMM korjasi virheitä hieman paremmin kuin LMP, jos menetelmiä verrataan tilanteissa, joissa kovariaatteina käytettävien pääkomponenttien määrä valitaan satunnaisesti. Kuitenkin, jos pääkomponenttien määrä osataan valita siten, että niitä on suunnilleen saman verran kuin geno-fenotyyppiaineistossa on geneettisiä populaatioita tai sitä enemmän, niin myös LMP-menetelmän kohdalla T1-osuus on LMM-menetelmän tapaan 0 ja T2-osuus on LMM-menetelmän tasolla.

Sukulaisuus koostui geneettisistä populaatiosta ja kryptisistä sukulaisista. Näistä kahdesta ainoastaan geneettisten populaatioiden havaittiin aiheuttavan vääristymiä assosiaatiotestien tuloksiin. Geneettisten populaatioiden vaikutus osoitettiin sillä, että LM-menetelmällä tyypin 1 ja 2 virheiden määrät kasvoivat selvästi, kun geno-fenotyyppiaineiston geneettisten populaatioiden määrä kasvoi yhdestä useampaan, ja kryptisten sukulaisuuden vaikuttamattomuus osoitettiin sillä, että virheiden määrät eivät kasvaneet kryptisten sukulaisten määrän kasvaessa. Mielenkiintoisena havaintona huomattiin myös se, että neljästä geneettisestä populaatiosta kymmeneen tyypin 1 virheet kuitenkin vähenivät LM-menetelmällä. T1-osuus oli kuitenkin vielä yli 10 % niissä aineistoissa, joissa oli simuloinnin maksimimäärän kymmenen verran geneettisiä populaatioita. Mielenkiintoista olisi kuitenkin selvittää millainen vaikutus geneettisillä populaatioilla olisi, jos niiden määrää kasvatettaisiin kymmenestä ylöspäin. Vähenisikö T1-osuus esimerkiksi nollaan jollain isommalla geneettisten populaatioiden määrällä.

Tämän tutkielman tulokset olivat osittain linjassa aiempien tutkimusten tulosten kanssa ja osittain ristiriitaiset. Tosin jo aiemmissakin tutkimuksissa on ollut ristiriitaisuuksia. Esimerkiksi Chen (2016) havaitsi, että sekamalleja käytettäessä assosiaatiotestit tuottavat tyypin 1 virheitä, mikä ei ole linjassa tämän tutkimuksen kanssa. Sulin (2018) ja Lin (2013) tulokset ovat puolestaan linjassa sen tuloksen kanssa, että sekamallien käyttö vähentää tyypin 1 virheitä. Lin (2013) tulokset ovat linjassa myös sen tuloksen kanssa, että LMP vähentää tyypin 1 virheiden määrää, mutta hän pitää sekamalleja parempana menetelmänä, koska hän väittää, että LMP osaa huomioida vain pienen määrän geneettisiä populaatioita ja ei ollenkaan kryptistä sukulaisuutta. Myös Price (2006) puoltaa sitä, että LMP vähentää tyypin 1 virheitä. Se miten paljon geneettisiä populaatioita LMP huomioi riippui tämän tutkielman tulosten mukaan siitä, miten monta pääkomponenttia valittiin malliin. Sitä puolestaan miten hyvin LMP huomioi kryptistä sukulaisuutta ei voitu tutkia, koska kryptinen sukulaisuus ei tämän tutkielman mukaan ylipäätään vääristä tuloksia. Sille, että kryptinen sukulaisuus ei aiheuta vääristymiä tuloksiin, ei ole ainakaan kirjoittajan tiedossa yhtään samassa linjassa olevaa aiempaa tutkimusta. Sukulaisuuden aiheuttamille vaikutuksille tyypin 2 virheisiin ei ole kirjoittajan käsityksen mukaan ollenkaan aiempia tutkimuksia.

Vaikka tuloksena havaittiin, että LMM- ja LMP-menetelmät vähentävät hyvin geneettisten populaatioiden aiheuttamia tyypin 1 ja 2 virheitä, niin menetelmän valintaa ei kuitenkaan pidä perustaa pelkästään virheiden minimointiin. Esimerkiksi tämän tutkielman tulosten mukaan LMM olisi helppo valinta käytettäväksi menetelmäksi virheiden minimoinnin kannalta, jos geno-fenotyyppiaineiston geneettisten populaatioiden lukumäärästä ei ole minkäänlaista käsitystä. Kuitenkin, jos analyysien nopeus olisi merkittävä kriteeri

menetelmän valinnalle, niin silloin LMP voisi olla LMM-menetelmää parempi valinta, koska LMM on isoilla aineistoilla huomattavasti hitaampi kuin LMP. Joskus GWAS-tutkimuksessa voidaan myös haluta vain rajata mahdollisesti fenotyyppiin vaikuttavien snippien joukkoa, jolloin tyypin 1 virheiden esiintyminen ei ole niin vakavaa. Tällaisissa tapauksissa jopa LM voisi olla riittävä menetelmä. Tutkielman tulokset eivät siis anna yleispätevää vastausta oikean menetelmän valitsemiseen GWAS-tutkimuksissa. Yhtenä menetelmän valintakriteerinä tulisi kuitenkin olla se miten menetelmä korjaa sukulaisuus-suhteiden aiheuttamia tyypin 1 ja 2 virheitä ja sillä kriteerillä mitattuna menetelmät LMM- ja LMP ovat tämän tutkielman tulosten mukaan hyviä menetelmiä.

Tutkielmaa on mahdollista laajentaa kehittämällä simulointia. Koska tutkielman nykyinen simulointikoodi luo geno-fenotyyppiaineistoja perustuen säädettäviin parametrien arvoihin niin yksinkertainen tapa laajentaa tutkielmaa on säätää parametrien arvoja uusia tutkimusongelmia vastaaviksi. Simulointiin on mahdollista lisätä myös täysin uusia tekijöitä, joiden vaikutuksia assosiaatiotestien tuloksiin on kiinnostavaa tutkia. LMM-, LMP- ja LM-menetelmien lisäksi assosiaatiotestejä on mahdollista toteuttaa koskemaan muita menetelmiä. Menetelmien toimivuutta voisi arvioida muillakin tavoilla kuin tyypin 1 ja 2 virheiden määrällä, kuten esimerkiksi nopeuden perusteella. Lisäksi simulointia voisi toistaa useaan kertaan, jolloin saataisiin arvioitua myös simuloinnin tulosten jakaumia. Simulointikoodia on mahdollista suorittaa rinnakkain, joten simuloinnin toistaminen useaan kertaan olisi myös ajallisesti järkevästi toteutettavissa.

Lähteet

1000 Genomes Project Consortium ym. 2015. ”A global reference for human genetic variation”. *Nature* 526 (7571): 68.

Abegaz, Fentaw, Kridsadakorn Chaichoompu, Emmanuelle Génin, David W Fardo, Inke R König, Jestinah M Mahachie John, ja Kristel Van Steen. 2019. ”Principals about principal components in statistical genetics”. *Briefings in Bioinformatics* 20 (6): 2200–2216.

Aitken, Alexander C. 1936. ”IV.—On least squares and linear combination of observations”. *Proceedings of the Royal Society of Edinburgh* 55: 42–48.

Astle, William, ja David J Balding. 2009. ”Population structure and cryptic relatedness in genetic association studies”. *Statistical Science* 24 (4): 451–71.

Bonferroni, Carlo. 1936. ”Teoria statistica delle classi e calcolo delle probabilita”. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8: 3–62.

Browning, Sharon R, ja Brian L Browning. 2011. ”Haplotype phasing: existing methods and new developments”. *Nature Reviews Genetics* 12 (10): 703–14.

Campbell, Neil A, Lisa A Urry, Michael L Cain, Steven A Wasserman, Peter V Minorsky, ja Jane B Reece. 2018. ”Biology. A global approach. Global edition”. Boston: Pearson.

Charnes, Abraham, Edward L Frome, ja Po-Lung Yu. 1976. ”The equivalence of generalized least squares and maximum likelihood estimates in the exponential family”. *Journal of the American Statistical Association* 71 (353): 169–71.

Chen, Han, Chaolong Wang, Matthew P Conomos, Adrienne M Stilp, Zilin Li, Tamar Sofer, Adam A Szpiro, ym. 2016. ”Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models”. *The American Journal of Human Genetics* 98 (4): 653–66.

Hannelius, Ulf, Elina Salmela, Tuuli Lappalainen, Gilles Guillot, Cecilia M Lindgren, Ulrika von Döbeln, Päivi Lahermo, ja Juha Kere. 2008. ”Population substructure in Finland and Sweden revealed by the use of spatial coordinates and a small number of unlinked autosomal SNPs”. *BMC Genetics* 9: 1–12.

Hutchinson, Anna, Jennifer Asimit, ja Chris Wallace. 2020. ”Fine-mapping genetic associations”. *Human Molecular Genetics* 29 (R1): R81–88.

Ikram, M Kamran, Sim Xueling, Richard A Jensen, Mary Frances Cotch, Alex W Hewitt, M Arfan Ikram, Jie Jin Wang, ym. 2010. ”Four novel Loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo”. *PLoS Genetics* 6 (10): e1001184.

Johnson, Randall C, George W Nelson, Jennifer L Troyer, James A Lautenberger, Bailey D Kessing, Cheryl A Winkler, ja Stephen J O’Brien. 2010. ”Accounting for multiple comparisons in a genome-wide association study (GWAS)”. *BMC Genomics* 11: 1–6.

Kang, Hyun Min, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, ja Eleazar Eskin. 2008. ”Efficient control of population structure in model organism association mapping”. *Genetics* 178 (3): 1709–23.

Kockum, Ingrid, Jesse Huang, ja Pernilla Stridh. 2023. ”Overview of Genotyping Technologies and Methods”. *Current Protocols* 3 (4): e727.

Li, Gengxin, ja Hongjiang Zhu. 2013. ”Genetic studies: the linear mixed models in genome-wide association studies”. *The Open Bioinformatics Journal* 7 (1).

Marees, Andries T, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis,

Cynthia Marie-Claire, ja Eske M Derks. 2018. "A tutorial on conducting genome-wide association studies: quality control and statistical analysis". *International Journal of Methods in Psychiatric Research* 27 (2): e1608.

Mitchell, Brittany L, Jake R Saklatvala, Nick Dand, Fiona A Hagenbeek, Xin Li, Josine L Min, Laurent Thomas, ym. 2022. "Genome-wide association meta-analysis identifies 29 new acne susceptibility loci". *Nature Communications* 13 (1): 702.

Nordborg, Magnus, ja Simon Tavaré. 2002. "Linkage disequilibrium: what history has to tell us". *Trends in Genetics* 18 (2): 83–90.

Pirastu, Nicola, Maarten Kooyman, Antonietta Robino, Ashley Van Der Spek, Luciano Navarini, Najaf Amin, Lennart C Karssen, Cornelia M Van Duijn, ja Paolo Gasparini. 2016. "Non-additive genome-wide association scan reveals a new gene associated with habitual coffee consumption". *Scientific Reports* 6 (1): 1–6.

Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, ja David Reich. 2006. "Principal components analysis corrects for stratification in genome-wide association studies". *Nature Genetics* 38 (8): 904–9.

Pruim, Randall J, Ryan P Welch, Serena Sanna, Tanya M Teslovich, Peter S Chines, Terry P Gliedt, Michael Boehnke, Gonçalo R Abecasis, ja Cristen J Willer. 2010. "LocusZoom: regional visualization of genome-wide association scan results". *Bioinformatics* 26 (18): 2336–37.

R Core Team. 2023. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rosenberg, Noah A. 2011. "A population-genetic perspective on the similarities and differences among worldwide human populations". *Human Biology* 83 (6): 659.

Speed, Doug, ja David J Balding. 2015. "Relatedness in the post-genomic era: is it still useful?" *Nature Reviews Genetics* 16 (1): 33–44.

Sul, Jae Hoon, Lana S Martin, ja Eleazar Eskin. 2018. "Population structure in genetic studies: confounding factors and mixed models". *PLoS Genetics* 14 (12): e1007309.

Torres-Duque, Carlos A, Maria Carmen Garcia-Rodriguez, ja Mauricio Gonzalez-Garcia. 2016. "Is chronic obstructive pulmonary disease caused by wood smoke a different phenotype or a different entity?" *Archivos de Bronconeumologia (English Edition)* 52 (8): 425–31.

Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, Jantina De Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, ja Danielle Posthuma. 2021. "Genome-wide association studies". *Nature Reviews Methods Primers* 1 (1): 59.

van Rossum, Bart-Jan, ja Willem Kruijer. 2022. *statgenGWAS: genome wide association studies*. https://CRAN.R-project.org/package=statgenGWAS.

Wang, Jiabo, Jianming Yu, Alexander E Lipka, ja Zhiwu Zhang. 2022. "Interpretation of Manhattan plots and other outputs of genome-wide association studies". Teoksessa *Genome-Wide Association Studies*, 63–80. Springer.

Wightman, Douglas P, Iris E Jansen, Jeanne E Savage, Alexey A Shadrin, Shahram Bahrami, Dominic Holland, Arvid Rongve, ym. 2021. "A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease". *Nature Genetics* 53 (9): 1276–82.

Visscher, Peter M, Sarah E Medland, Manuel A R Ferreira, Katherine I Morley, Gu Zhu, Belinda K Cornes, Grant W Montgomery, ja Nicholas G Martin. 2006. "Assumption-

free estimation of heritability from genome-wide identity-by-descent sharing between full siblings". *PLoS Genetics* 2 (3): e41.

Wojczynski, Mary K, ja Hemant K Tiwari. 2008. "Definition of phenotype". *Advances in Genetics* 60: 75–105.

Wu, Heming, Qingyan Huang, Zhikang Yu, Hailing Wu, ja Zhixiong Zhong. 2020. "The SNPs rs429358 and rs7412 of APOE gene are association with cerebral infarction but not SNPs rs2306283 and rs4149056 of SLCO1B1 gene in southern Chinese Hakka population". *Lipids in Health and Disease* 19: 1–9.

Xue, Yahui, Shen Liu, Weining Li, Ruihan Mao, Yue Zhuo, Wenkai Xing, Jian Liu, ym. 2022. "Genome-wide association study reveals additive and non-additive effects on growth traits in Duroc pigs". *Genes* 13 (8): 1454.

Liitteet

Liite 1: Geneettisen ajautumisen simulointikoodi

```
set.seed(170823)
# Toisistaan eristäytyvien populaatioiden koot.
populaatiokoko <- 100
# Simuloitavien sukupolvien määrä.
sukupolvienMaara = 50
# Toisistaan eristäytyvien populaatioiden määrä.
populaatioMaara = 5
# Alustetaan matriisi populaatioiden sukupolvien alleelifrekvensseille.
alleeliFrekvenssit = matrix(NA,
           ncol = sukupolvienMaara + 1,
           nrow = populaatioMaara)
# Simuloidaan populaatio kerrallaan alleelifrekvenssien
# muutos sukupolvesta seuraavaan.
for (populaatio in 1:populaatioMaara) {
  # Alustetaan alleelit ensimmäiselle sukupolvelle
  # ja määritetään yksilöiden sukupuolet.

  # Molempia alleeleja alussa samanverran.
  alleelit = c(rep(0, populaatiokoko), rep(1, populaatiokoko))
  # Laitetaan alleelit satunnaiseen järjestykseen, jotta saadaan yksilöitä,
  # joiden genotyypit ovat satunnaisesti 0, 1 tai 2.
  satunnainenJarjestys <- sample(1:(populaatiokoko*2), populaatiokoko*2)
  jarjestetytalleelit <- alleelit[satunnainenJarjestys]
  # Alustetaan matriisi ensimmäisen sukupolven yksilöiden
  # alleeleille ja sukupuolelle.
  sukupolvenAlleelit <- matrix(nrow = populaatiokoko, ncol = 3)
  # Yksilöiden ensimmäinen alleeli
  sukupolvenAlleelit[,1] <- jarjestetytalleelit[1:populaatiokoko]
  # Yksilöiden toinen alleeli.
  sukupolvenAlleelit[,2] <-
    jarjestetytalleelit[(populaatiokoko+1):(populaatiokoko*2)]
  # Yksilöiden sukupuolet.
  sukupolvenAlleelit[,3] <-
    c(rep(0, populaatiokoko/2,), rep(1, populaatiokoko/2))
  # Ensimmäisen sukupolven alleelifrekvenssi.
  alleeliFrekvenssit[populaatio, 1] <-
    mean(c(sukupolvenAlleelit[,1], sukupolvenAlleelit[,2]))
  # Simuloidaan populaation loppujen sukupolvien alleelifrekvenssit.
  for (i in 1:sukupolvienMaara) {
```

```
    # Matriisi uuden sukupolven yksilöiden alleeleille ja sukupuolelle.
    uudenSukupolvenAlleelit <-  matrix(nrow = populaatiokoko, ncol = 3)
    # Poimitaan vain mahdolliset isät.
    isat <- sukupolvenAlleelit[sukupolvenAlleelit[,3] == 1,]
    # Poimitaan vain mahdolliset äidit.
    aidit <- sukupolvenAlleelit[sukupolvenAlleelit[,3] == 0,]
    # Simuloidaan seuraavan sukupolven yksilöiden alleelit ja sukupuolet.
    for (j in 1:populaatiokoko) {
      # Arvotaan jälkelaisen isä.
      isa <- isat[sample(1:nrow(isat), 1),]
      # Arvotaan jälkelaisen isältä perittävä alleeli.
      isaltaPerittyAleeli <- isa[sample(c(1,2),1)]
      # Arvotaan jälkelaisen äiti.
      aiti <- aidit[sample(1:nrow(aidit), 1),]
      # Arvotaan jälkelaisen äidiltä perittävä alleeli.
      aidiltaPerittyAleeli <- aiti[sample(c(1,2),1)]
      # Arvotaan jälkelaisen sukupuoli.
      jalkelaisenSp <- sample(c(0,1), 1)
      # Kerätään talteen jälkeläisen tiedot.
      uudenSukupolvenAlleelit[j,] <-
        c(isaltaPerittyAleeli, aidiltaPerittyAleeli, jalkelaisenSp)
    }
    # Uuden sukupolven alleelifrekvenssi talteen.
    alleeliFrekvenssit[populaatio, i + 1] <-
      mean(c(uudenSukupolvenAlleelit[,1], uudenSukupolvenAlleelit[,2]))
    # Otetaan viimeisimpänä simuloidun sukupolven alleelit talteen
    # seuraavan sukupolven alleelien simulointia varten.
    sukupolvenAlleelit <- uudenSukupolvenAlleelit
  }
}
# Kuvaajan piirto
plot(
  NULL,
  xlab = "Sukupolvi",
  ylab = "Alleelifrekvenssi",
  xlim = c(0, sukupolvienMaara),
  ylim = c(0, 1),
  main = "Geneettinen ajautuminen"
)
varit <- c("red", "blue", "green", "yellow", "orange")
for (populaatio in 1:populaatioMaara) {
  lines(0:sukupolvienMaara,
        alleeliFrekvenssit[populaatio, ],
        lwd = 1.5,
```

```
        col = varit[populaatio])
}
```

Liite 2: Simuloinnin toteuttaneen laitteen tekniset tiedot

Suoritin: Intel(R) Core(TM) i5-1035G1 CPU @ 1.00GHz
Keskusmuisti: 8,00 Gt
Levymuisti: 475 Gt
Järjestelmätyyppi: 64-bittinen käyttöjärjestelmä, x64-suoritin

Liite 3: Simulointikoodi geno-fenotyyppiaineistojen muodostamiseen ja assosiaatiotestaukseen

```
set.seed(8637932)
# Määritetään funktio, joka luo geno- ja fenotyyppitiedot sekä
# sukulaisuusmatriisin sukulaisuutta sisältäville tai
# sisältämättömille ainiestoille.
simuloiGwasData <- function(populaatiot, snpN, betat) {
  # Alustetaan matriisi, johon tallennetaan kaikki genotyypit.
  genotyyppiDataKinship <- matrix(ncol = snpN)
  # Nimetään snipit "rs1", "rs2", "rs3"....
  snpNumero <- seq(1, snpN)
  snpNimi <- c(paste0("rs", snpNumero))
  colnames(genotyyppiDataKinship) <- snpNimi
  # Simuloidaan geneettinen populaatio kerrallaan
  for (populaatio in 1:length(populaatiot$populaatiokoko)) {
    # Poimitaan populaation koko.
    populaationKoko <- populaatiot$populaatiokoko[populaatio]
    # Poimitaan populaation kryptisten sukulaisten määrä.
    populaationKryptisetSukulaisetN <-
      populaatiot$populaationKryptisetsukulaiset[populaatio]
    # Alustetaan matriisi, johon tallennetaan yhden populaation
    # yksilöiden genotyypit.
    populaationGenotyypit <-
      matrix(nrow = populaationKoko, ncol = snpN)
    # Nimetään rivit tyylillä
    # ty_p[POPULAATIONID]_h[YKSILÖNIDPOPULAATIOSSA].
    hId <- seq(1, populaationKoko)
    pId <- rep(populaatio, populaationKoko)
    henkiloNimi <- c(paste0("ty_p", pId, "_h", hId))
    rownames(populaationGenotyypit) <- henkiloNimi
    # Poimitaan populaation alleelifrekvenssit.
    populaationAlleeliFrekvenssit <-
      populaatiot$alleeliFrekvenssit[,populaatio]
    # Populaation ei sukua toisilleen olevat yksilöt.
    polaationEiKryptisetSukulaisetN <-
      populaationKoko - populaationKryptisetSukulaisetN
    # Simuloidaan jokaiselle populaation toisilleen ei sukua
    # oleville yksilöille genotyypit snippi kerrallaan.
    for (snp in 1:snpN) {
      # Snipin ensimmäiset alleelit.
      alleeli1 <-
        rbinom(polaationEiKryptisetSukulaisetN, 1,
```

```
            populaationAlleeliFrekvenssit[snp])
    # Snipin toiset alleelit.
    alleeli2 <- rbinom(polaationEiKryptisetSukulaisetN, 1,
                       populaationAlleeliFrekvenssit[snp])
    # Lasketaan snipin genotyypit.
    genotyypit <- alleeli1 + alleeli2
    # Kerätään genotyypit talteen.
    populaationGenotyypit[1:polaationEiKryptisetSukulaisetN,snp] <-
      genotyypit
  }
  # Lisätään populaatioon kryptisiä sukulaisia, jos niitä
  # on enemmän kuin 0.
  if (populaationKryptisetSukulaisetN > 0) {
    # Jaetaan populaation kryptisten sukulaisten lukumäärä 2-20
    # yksilön ryhmiin, jotka voivat olla keskenään sukulaisia.
    ryhmaKoot <- c()
    ryhmanYlaraja <- 20
    while (TRUE) {
      # Arvotaan ryhmän koko.
      ryhmaKoko <- sample(2:ryhmanYlaraja, 1)
      # Tarkastetaan onko kaikki populaation kryptiset sukulaiset
      # jo jaettu ryhmiin.
      if ((populaationKryptisetSukulaisetN - sum(ryhmaKoot)) == 0) break
      # Tarkastetaan, onko aiempien ryhmien ja uuden ryhmän kokojen
      # summa pienempi kuin kryptisten sukulaisten lukumäärä
      # vähennettynä yhdellä. Jos on, niin uusi ryhmä voi olla
      # määritetyn kokoinen, koska silloin jaettavia kryptisiä
      # sukulaisia jää vähintään 2 seuraavaan ryhmään (yhden kokoinen
      # ryhmä ei ole sallittu, koska ei voi olla vain itselleen sukua).
      # Jos ei ole, niin uuden ryhmän kooksi määritetään se,
      # paljonko kryptisiä sukulaisia on jakamatta ryhmiin, paitsi jos
      # niitä on jakamatta yhden yli ryhmän ylärajan verran, jolloin
      # ryhmän koko arvotaan uudestaan.
      if (sum(ryhmaKoot, ryhmaKoko) <
          (populaationKryptisetSukulaisetN - 1)) {
        ryhmaKoot <- c(ryhmaKoot, ryhmaKoko)
      } else {
        if ((populaationKryptisetSukulaisetN - sum(ryhmaKoot)) ==
            (ryhmanYlaraja + 1)) next
        ryhmaKoot <-
          c(ryhmaKoot,
            (populaationKryptisetSukulaisetN - sum(ryhmaKoot)))
      }
    }
```

```
    # Otetaan käyttöön indeksi, jonka mukaan tallennetaan genotyypit.
    indeksi <- polaationEiKryptisetSukulaisetN + 1
    # Simuloidaan kryptisten sukulaisten genotyypit ryhmä kerrallaan.
    for (j in 1:length(ryhmaKoot)) {
      # Alustetaan matriisi, johon tallennetaan alleelit.
      alleelit <- matrix(nrow = snpN, ncol = ryhmaKoot[j]*2)
      # Ryhmän ensimmäisen yksilön alleelit.
      alleelit[,1] <- rbinom(snpN, 1, populaationAlleeliFrekvenssit)
      alleelit[,2]<- rbinom(snpN, 1, populaationAlleeliFrekvenssit)
      # Lasketaan ensimmäisen yksilön genotyypit ja otetaan ne talteen.
      genotyypit <- alleelit[,1] + alleelit[,2]
      populaationGenotyypit[indeksi, 1:snpN] <- genotyypit
      # Indeksin päivitys.
      indeksi <- indeksi + 1
      # Simuloidaan ryhmän loppujen yksilöiden genotyypit.
      for (yksilo in 2:ryhmaKoot[j]) {
        # Arvotaan aluksi yksilön alleelit perustuen populaation
        # alleelifrekvensseihin.
        alleelit[,(yksilo*2 - 1)] <-
          rbinom(snpN, 1, populaationAlleeliFrekvenssit)
        alleelit[,(yksilo*2)] <-
          rbinom(snpN, 1, populaationAlleeliFrekvenssit)
        # Ryhmän yksilö, jonka alleeleita kopioidaan.
        kopioitavaYksilo <- sample(1:(yksilo-1), 1)
        # Kopioitavien alleelien osuus ja määrä.
        kopioitavienAlleelienOsuus <- runif(1, 0.05, 0.6)
        kopioitavienAlleelienMaara <-
          round(kopioitavienAlleelienOsuus*snpN*2)
        # Kopioitavien alleelien sijainnit.
        kopioitavatAlleelit <-
          sample(1:(snpN*2), kopioitavienAlleelienMaara)
        kopioitavatAlleelit1 <-
          kopioitavatAlleelit[kopioitavatAlleelit <= snpN]
        kopioitavatAlleelit2 <-
          kopioitavatAlleelit[kopioitavatAlleelit > snpN] - snpN
        # Kopioidaan alleeleja.
        alleelit[kopioitavatAlleelit1,(yksilo*2 - 1)] <-
          alleelit[kopioitavatAlleelit1, (kopioitavaYksilo*2 - 1)]
        alleelit[kopioitavatAlleelit2,(yksilo*2)] <-
          alleelit[kopioitavatAlleelit2, (kopioitavaYksilo*2)]
        # Lasketaan yksilön genotyypit ja otetaan ne talteen.
        genotyypit <-
          alleelit[,(yksilo*2 - 1)] + alleelit[,(yksilo*2)]
        populaationGenotyypit[indeksi, 1:snpN] <- genotyypit
```

```
        # Nimetään yksilö siten, että tunnistaa kenelle se
        # on eniten sukua.
        yksilonimi <-
          paste0(rownames(populaationGenotyypit)[indeksi],
                 "_", kopioitavaYksilo)
        rownames(populaationGenotyypit)[indeksi] <- yksilonimi
        # Indeksin päivitys.
        indeksi <- indeksi + 1
      }
    }
  }
  # Otetaan talteen populaation genotyypit.
  genotyyppiDataKinship <-
    rbind(genotyyppiDataKinship, populaationGenotyypit)
}
# Poistetaan alustuksen takia tullut ensimmäinen turha rivi.
genotyyppiDataKinship <- genotyyppiDataKinship[-1,]
# Lasketaan sukulaisuusmatriisi.
kinship <- kinship(genotyyppiDataKinship, method = "IBS")
# Valitaan mitä snipeistä testataan.
genotyyppiDataAnalyysi <- genotyyppiDataKinship
# Map-datan muodostaminen createGData funktiota varten. Muuten nimillä
# ja sijainneilla ei tässä ole varsinaista merkitystä, mutta ne pitää
# määritellä joksikin funktion toiminnan takia.
kromosomi <- sort(sample(1:22, snpN, replace = TRUE))
sijainti <- seq(1, snpN)
map <- data.frame("chr" = kromosomi, "pos" = sijainti)
rownames(map) <- snpNimi
# Fenotyyppien laskeminen
# Lasketaan yskilöiden genomin vaikutus.
snippienVaikutus <- genotyyppiDataAnalyysi %*% betat
# Lasketaan yksilön geneettiseen populaatioon kuulumisen vaikutus.
populaationVaikutusKa <-
  rep(populaatiot$populaationVaikutusFenotyyppiin,
      populaatiot$populaatiokoko)
sdPopulaatio <- runif(populaatioMaara, 1, 3)
sdPopulaatio <- rep(sdPopulaatio, populaatiot$populaatiokoko)
otoskoko <- sum(populaatiot$populaatiokoko)
populaationVaikutus <-
  rnorm(otoskoko, populaationVaikutusKa, sdPopulaatio)
# Arvotaan yksilöille satunnainen vaikutus.
sdYksilo <- runif(1, 1, 3)
yksilonVaihtelu <- rnorm(otoskoko, mean = 0, sd = sdYksilo)
# Lasketaan fenotyyppien arvo.
```

```
  fenotyyppi <-
    snippienVaikutus + populaationVaikutus + yksilonVaihtelu
  # Kootaan data createGData-fuktiolle sopivaan muotoon.
  fenotyyppiData <-
    data.frame("genotype" = rownames(genotyyppiDataAnalyysi),
               "fenotyyppi" = fenotyyppi, row.names = NULL)
  # Tehdään datasta runSingleTraitGwas-funktiolle sopiva.
  gwasSyote <-
    createGData(geno = genotyyppiDataAnalyysi, map = map,
                pheno = fenotyyppiData, kin = kinship)
  # Palautetaan data.
  return(gwasSyote)
}

# Simulointi

# Montako simulointikierrosta suoritetaan.
simulointiMaara <- 5000
# Alustetaan matriisi kerättäville simulointitiedoille.
simulointiTiedot <- matrix(nrow = simulointiMaara, ncol = 16)
# Suoritetaan simulointi.
for (i in 1:simulointiMaara) {
  # Tulostetaan suoritettava simulointikierros seurantaa varten.
  cat("------------------------------------------------\n")
  cat(paste("Simulointi", i, "\n"))
  # Arvotaan testattavien snippien määrä.
  snpN = sample(3000:6000, 1)
  # Arvotaan kunkin snipin beeta-kertoimet eli yhden alleelin
  # vaikutus fenotyyppiin. Snippi ei vaikuta fenotyypin arvoon, jos
  # beeta-kerroin on 0 ja vaikuttaa, jos beeta-kerroin poikkeaa
  # nollasta. Vaikuttavien snippien määrä on maksimissaan 10 ja
  # minimissään 1.
  betat = rep(0, snpN)
  merkittavienSnippienMaara <- sample(1:10, 1)
  betat[1:merkittavienSnippienMaara] <-
    runif(merkittavienSnippienMaara, 2, 5)
  # Arvotaan geneettisten populaatioiden määrä, joihin aineiston
  # yksilöt kuuluvat.
  otetaankoPopulaatioita <-
    sample(c(TRUE, FALSE), 1, prob = c(0.7, 0.3))
  if (otetaankoPopulaatioita) {
    populaatioMaara <- sample(2:10, 1)
  } else {
    populaatioMaara <- 1
```

```
}
# Muodostetaan geneettiset populaatiot kryptisistä sukulaisista ja
# ei kryptisistä sukulaisista.

# Arvotaan otetaanko kryptisiä mukaan ollenkaan.
otetaankoKryptisia <- sample(c(TRUE, FALSE), 1, prob = c(0.7, 0.3))
if (otetaankoKryptisia) {
  # Arvotaan kryptisten sukulaisten määrä. Määrä voi heittää aluksi
  # määritellystä hieman, jos satunnaisuuden takia populaatiolle
  # arvotaan 1 kryptinen sukulainen, mikä ei ole mahdollista.
  # Tällöin populaation 1 kryptinen sukulainen muutetaan kahdeksi
  # ja kokonaismäärä kasvaa silloin myös yhdellä.
  kryptisetsukulaisetMaara <- sample(2:(400), 1)
  # Jaetaan kryptiset sukulaiset geneettisiin populaatioihin.
  populaationKryptisetsukulaiset <-
    diff(c(0, sort(sample(1:(kryptisetsukulaisetMaara-1),
                          populaatioMaara - 1, replace = TRUE)),
           kryptisetsukulaisetMaara))
  # Muutetaan ykköset kakkosiksi, koska yksi kryptinen sukulainen ei
  # ole mahdollinen, koska täytyy olla toinen, jolle on sukua.
  populaationKryptisetsukulaiset[populaationKryptisetsukulaiset ==
                                   1] <- 2
  # Arvotaan ei kryptisten sukulaisten määrä. Määrä arvotaan siten,
  # että otoskooksi saadaan 500-1000.
  eiKryptisetsukulaisetMaara <-
    round(sample((500 - kryptisetsukulaisetMaara):
                   (1000 - kryptisetsukulaisetMaara), 1))
  # Jaetaan ei kryptiset sukulaiset satunnaisesti populaatoihin.
  populaationEiKryptisetsukulaiset <-
    diff(c(0, sort(sample(1:(eiKryptisetsukulaisetMaara-1),
                          populaatioMaara - 1)),
           eiKryptisetsukulaisetMaara))
  #Lasketaan populaatioiden koot.
  populaatioKoot <- populaationKryptisetsukulaiset +
    populaationEiKryptisetsukulaiset
} else {
  # Arvotaan ei kryptisten sukulaisten määrä.
  eiKryptisetsukulaisetMaara <- sample(500:1000, 1)
  # Jaetaan ei kryptiset sukalaiset populaatioihin.
  populaatioKoot <-
    diff(c(0, sort(sample(1:(eiKryptisetsukulaisetMaara - 1),
                          populaatioMaara - 1)),
           eiKryptisetsukulaisetMaara))
  # Asetetaan kryptisten sukulaisten lukumäärä nollaksi.
```

```
  populaationKryptisetsukulaiset <- rep(0, populaatioMaara)
}
# Arvotaan geneettisten populaatioiden alleelifrekvenssit.
# Ensimmäisen geneettisen populaation alleelifrekvenssit arvotaan
# tasajakaumasta U(0.05, 0.95). Muiden geneettisten populaatioiden
# alleelifrekvenssit arvotaan katkaistusta normaalijakaumasta,
# jossa keskiarvona käytetään ensimmäisen geneettisen populaation
# alleelifrekvenssejä ja hajontaparametrinä tasajakaumasta
# U(0.1, 0.3) arvottua lukua. Hajontaparametri arvotaan erikseen
# kullekin populaatiolle. Mitä suurempi hajontaparametri
# sitä enemmän arvotut alleelifrekvenssit keskimäärin poikkeavat
# ensimmäisen geneettisen populaation alleelifrekvensseistä.
populaatiodenAlleeliFrekvenssit <-
  matrix(nrow = snpN, ncol = populaatioMaara)
populaatiodenAlleeliFrekvenssit[,1] <- runif(snpN, 0.05, 0.95)
if (populaatioMaara > 1) {
  sd <- runif((populaatioMaara-1), 0.1, 0.3)
  for (k in 2:(populaatioMaara)) {
    populaatiodenAlleeliFrekvenssit[,k] <-
      rtruncnorm(snpN, 0.05, 0.95,
                 mean = populaatiodenAlleeliFrekvenssit[,1],
                 sd = sd[k-1])
  }
}
populaatiodenAlleeliFrekvenssit <-
  data.frame(populaatiodenAlleeliFrekvenssit)
names(populaatiodenAlleeliFrekvenssit) <-
  c(paste0("p", 1:populaatioMaara, "Frekvenssit"))
# Arvotaan kuhunkin geneettiseen populaatioon kuulumisen vaikutus
# fenotyyppiin. Simulointi olettaa, että samaan geneettiseen
# populaatioon kuuluvilla yksilöillä voi olla yhteisiä fenotyyppiin
# vaikuttavia tekijöitä samankaltaisen genomin lisäksi. Esimerkiksi
# ympäristövaikutukset. Arvonta mahdollistaa kuitenkin myös sen,
# että tällaisia vaikutuksia ei ole geneettisellä populaatiolla,
# koska vaikutus arvotaan tasajakaumasta U(-10, 10).
populaationVaikutusFenotyyppiin <- runif(populaatioMaara, -10, 10)
# Kootaan popoulaatioiden tiedot listaan.
populaatiot <-
  list(populaatiokoko = populaatioKoot,
       alleeliFrekvenssit = populaatiodenAlleeliFrekvenssit,
       populaationVaikutusFenotyyppiin = populaationVaikutusFenotyyppiin,
       populaationKryptisetsukulaiset = populaationKryptisetsukulaiset)
# Simuloidaan geno-fenotyyppiaineisto simuloiGwasData-funktiolla.
# SimuloiGwasData-funktio luo geno-fenotyyppiaineiston, joka sopii
```

```
# parametriksi runSingleTraitGwas-funktiolle.
gwasSyote <- simuloiGwasData(populaatiot = populaatiot,
                             snpN = snpN,
                             betat = betat)
# Suoritetaan assosiaatioiden testaus lineaarisella sekamallilla
# käyttämällä statgenGWAS-kirjaston funktiota runSingleTraitGwas.
gwasTulokset <- runSingleTraitGwas(gData = gwasSyote)
# Suoritetaan assosiaatiotestaus samalle aineistolle perinteisellä
# lineaarisella regressiolla ja lineaarisella regressiolla, jossa
# on genomitiedoista muodostettuja pääkomponentteja kovariaatteina.
# Poimitaan fenotyyppien arvot.
feno <-
  gwasSyote$pheno$fenotyyppiData$
  fenotyyppi[order(gwasSyote$pheno$fenotyyppiData$genotype)]
# Muodostetaan pääkomponentit. Pääkomponenttien määrä arvotaan
# diskreetistä tasajakaumasta jonka alaraja on 1 ja yläraja 10.
paakomponenttienMaara <- sample(1:10, 1)
pca <- prcomp(gwasSyote$markers, rank. = paakomponenttienMaara)
# Alustetaan vektorit, joihin kerätään testien p-arvot.
pArvot <- rep(0, snpN)
pArvotPc <- rep(0, snpN)
# Suoritetaan testit yksi snippi kerrallaan ja kerätään
# p-arvot talteen.
for (j in 1:snpN) {
  snp <- gwasSyote$markers[,j]
  malli <- glm(feno ~ snp)
  malliPc <- glm(feno ~ snp + pca$x)
  pArvot[j] <- summary(malli)$coefficients[2,4]
  pArvotPc[j] <- summary(malliPc)$coefficients[2,4]
}
# Kerätään tarpeelliset simulointitiedot jälkianalyysejä varten
simulointiTiedot[i,1] <- i
simulointiTiedot[i,2] <-
  sum(gwasTulokset$GWAResult$fenotyyppiData$
        pValue[1:merkittavienSnippienMaara] < 5*10^{-8})
simulointiTiedot[i,3] <-
  simulointiTiedot[i,2]/merkittavienSnippienMaara
simulointiTiedot[i,4] <-
  sum(gwasTulokset$GWAResult$fenotyyppiData$
        pValue[-(1:merkittavienSnippienMaara)] < 5*10^{-8})
simulointiTiedot[i,5] <-
  simulointiTiedot[i,4]/(snpN - merkittavienSnippienMaara)
simulointiTiedot[i,6] <-
  sum(pArvot[1:merkittavienSnippienMaara] < 5*10^{-8})
```

```
  simulointiTiedot[i,7] <-
    simulointiTiedot[i,6]/merkittavienSnippienMaara
  simulointiTiedot[i,8] <-
    sum(pArvot[-(1:merkittavienSnippienMaara)] < 5*10^{-8})
  simulointiTiedot[i,9] <-
    simulointiTiedot[i,8]/(snpN - merkittavienSnippienMaara)
  simulointiTiedot[i,10] <-
    sum(pArvotPc[1:merkittavienSnippienMaara] < 5*10^{-8})
  simulointiTiedot[i,11] <-
    simulointiTiedot[i,10]/merkittavienSnippienMaara
  simulointiTiedot[i,12] <-
    sum(pArvotPc[-(1:merkittavienSnippienMaara)] < 5*10^{-8})
  simulointiTiedot[i,13] <-
    simulointiTiedot[i,12]/(snpN - merkittavienSnippienMaara)
  simulointiTiedot[i,14] <- populaatioMaara
  simulointiTiedot[i,15] <- paakomponenttienMaara
  simulointiTiedot[i,16] <- sum(populaationKryptisetsukulaiset)
}
```