Kromosomien välisten genotyyppiassosiaatioiden etsintä

Teppo Niinimäki

Helsinki 23.6.2010

Pro gradu -tutkielma

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen tiedekunta	Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Teppo Niinimäki

Työn nimi — Arbetets titel — Title

Kromosomien välisten genotyyppiassosiaatioiden etsintä

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	23.6.2010	68 sivua

Tiivistelmä — Referat — Abstract

Perimän eri kohdissa sijaitsevat genotyypit ovat assosioituneita, jos niiden välillä on tilastollinen riippuvuus. Tässä tutkielmassa esitellään ja vertaillaan menetelmiä kromosomien välisten genotyyppiassosiaatioiden etsintään. Saatavilla olevista genotyyppiaineistoista voidaan muodostaa miljardeja kromosomien välisiä ehdokkaita mahdollisesti assosioituneiksi genotyyppipareiksi. Etsintätehtävä voidaan jakaa kolmeen erilliseen osaan: assosiaation voimakkuutta kuvaavan tunnusluvun valinta, tuloksen merkitsevyyden laskeminen sekä tarpeeksi merkitsevien tulosten valinta.

Tunnusluvun valintaan ja merkitsevyyden laskemiseen liittyen tutkielmassa esitellään pari alleeliassosiaation mittaamiseen tarkoitettua perinteistä alleeliassosiaatiomittaa sekä yleisempiä riippumattomuustestejä kuten khii-toiseen-testi, G-testi ja erilaisia satunnaiseen näytteenottoon perustuvia testaustapoja. Lisäksi ehdotetaan kahta menetelmää tarkkaan merkitsevyyden laskemiseen: genotyyppikohtaista tarkkaa testiä ja maksimipoikkeamatestiä. Merkitsevien tulosten valintaan liittyen tutustutaan koekohtaista virhetodennäköisyyttä rajoittavaan Bonferroni-korjaukseen, hylkäysvirheastetta rajoittavaan FDR-kontrollointiin sekä näiden muunnelmiin.

Lopuksi kokeillaan muutamaa esiteltyä menetelmää sekä keinotekoisesti tuotetulla että aidolla genotyyppiaineistolla ja analysoidaan löydettyjä assosiaatioita. Koetuloksista on havaittavissa joukko vahvasti merkitseviä assosiaatioita kromosomien välillä. Osa näistä on selitettävissä populaation sisäisillä osapopulaatioilla, ja muutamat näyttäisivät olevan seurausta aineistossa väärin sijoitelluista markkereista. Suuri osa riippuvuuksista aiheutuu kolmesta sukupuolen kanssa vahvasti assosioituneesta perimän kohdasta. Näiden lisäksi jäljelle jää joukko assosiaatioita, joiden syyt ovat tuntemattomia.

ACM Computing Classification System (CCS):
G.3 [Probability and Statistics]: Contingency table analysis / Statistical computing
J.3 [Life and Medical Sciences]: Biology and genetics

Avainsanat — Nyckelord — Keywords
kromosomien välinen kytkentäepätasapaino, moninkertainen merkitsevyystestaus

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Sisältö

1 Johdanto

Ihmisten ominaisuudet siirtyvät jälkeläisille perimän välityksellä. Perimästä tunnetaan paljon kohtia, joissa esiintyy vaihtelua. Näitä kohtia kutsutaan *markkereiksi*. Suuri osa markkereista sijaitsee geeneissä, perimän informaatiota sisältävissä osissa. Sanomme, että kahden markkerin välillä on assosiaatio, jos niiden arvojen välillä on tilastollinen riippuvuus. Markkerien välisten assosiaatioiden mittaamiseen on perinteisesti liittynyt kaksi erilaista tavoitetta: yhtäältä yksittäisien tautigeenien paikantaminen, toisaalta taas ihmisen perimän rakenteeseen ja populaatioihin liittyvän yleisen tietämyksen lisääminen.

Ensimmäiset tieteelliset teoriat geenien välisistä assosiaatioista ilmestyivät 1900-luvun alkupuolella [Mor07]. Aineistoihin perustuvat havainnot geenien välillä ilmenevistä assosioitumisesta johtivat spekulointiin geenien keskinäisestä vuorovaikutuksesta. Läheisten geenien välisten assosiaatioiden tutkimisesta muodostui vuosisadan kuluessa oma tieteenalansa ja alleeliassosiaation suuruuden havainnointiin kehitettiin useita eri mittalukuja. Teoreettisen tutkimuksen lisäksi assosiaatioiden mittaamiselle alettiin löytää käytännön hyötykohteita. Kerem ja kumppanit [KRB+89] onnistuivat 1980-luvun lopulla paikantamaan kystisen fibroosin aiheuttajageenin alleeliassosiaation avulla. Myöhemmin lukuisia muita tautigeenejä on paikannettu samaan tapaan. Viime aikoina erityisesti kansainvälisessä HapMap-projektissa [Con03] on mm. assosiaatioita mittaamalla avulla pyritty muodostamaan parempi kuva ihmisen perimän sisäisestä rakenteesta ja vaihtelusta.

Assosiaatiota on tutkittu pääasiassa yksittäisten kromosomien sisällä lähekkäisten markkerien välillä. Toisaalta pidemmälle ulottuvista ja jopa kromosomirajat ylittävistä assosiaatioistakin on olemassa havaintoja. Tässä tutkielmassa keskitymmekin juuri näiden kromosomien välisten assosiaatioiden etsimiseen. Painotamme menetelmiä, jotka toimivat hyvin laajasti saatavilla olevalla perimäaineistotyypillä. Lisäksi pyrimme kiinnittämään huomiota menetelmien käytännön tehokkuuteen; tavoitteena on suorittaa koko genomin laajuinen haku kaikkien kromosomien välisten markkeriparien kesken.

Tutkielman rakenne tästä eteenpäin on seuraavanlainen. Johdantoluvun loppuosassa käymme läpi ihmisen DNA:han ja sen periytymiseen liittyviä peruskäsitteitä sekä tutustumme saatavilla olevaan aineistoon. Luvussa 2 käsittelemme markkerien välisen assosiaation mittaamista eri menetelmillä. Pääosin keskitymme genotyyppien välisiin assosiaatioihin tavallisesti mitattavan alleeliassosiaation sijaan. Luvussa 3

tutustumme moninkertaiseen testaukseen liittyvään merkitsevyyskynnyksen valintaan. Lopuksi neljännessä luvussa sovellamme lukujen 2 ja 3 menetelmiä assosiaatioiden etsimiseen sekä keinotekoisesti muodostetusta että aidosta markkeriaineistosta ja analysoimme saatuja tuloksia.

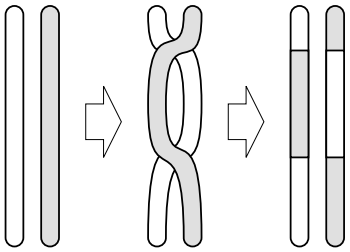
1.1 Johdatus kromosomeihin ja periytymiseen

Tässä ja seuraavassa osiossa käymme läpi tarvittavat perustiedot ihmisen perimästä ja saatavilla olevasta aineistosta. Seuraavat tiedot perustuvat pääosin Strachanin ja Readin kirjaan [SR96] sekä Collinsin kirjan johdantoartikkeliin [Col07].

Ihmisen perimä on on varastoituna *DNA-ketjuun*. Suurin osa DNA:sta on puolestaan jakautunut 23 *kromosomiin*, joista kukin koostuu yhdestä pitkästä DNA-rihmasta. Kromosomeista 22 on *autosomeja*, ja niihin viitataan yleensä numeroilla 1–22. Jäljelle jäävä 23. kromosomi on *sukupuolikromosomi*, ja siitä on kahta eri tyyppiä: X ja Y. Normaalisti ihmisellä on kaksi lähes identtistä versiota jokaisesta kromosomista: yhteensä siis 46 kromosomia, jotka muodostavat 22 autosomiparia ja yhden sukupuolikromosomiparin. Naisilla kummatkin sukupuolikromosomit ovat X-kromosomeja, kun taas miehillä toinen on X- ja toinen Y-kromosomi. Kromosomipari voi tarvittaessa kiinnittyä yhteen tietystä kullekin parille ominaisesta kohdasta, jota kutsutaan *sentromeeriksi*.

DNA-rihman sisältämä tieto on varastoitunut miljooniin peräkkäisiin *emäspareihin*. Erilaisia mahdollisia emäspareja on neljä, ja näitä merkitään kirjaimin A, C, G ja T. Vain arviolta kymmenyksellä emäspareista on periytyviä ominaisuuksia koodaava merkitys; loppuja kutsutaan usein roska-DNA:ksi. Yhdessä koodaavat emäsparit muodostavat *geenejä* eli *perintötekijöitä*, joita voidaan pitää eräänlaisina periytymisen perusyksiköinä. Geenistä voi olla olemassa useita hieman toisistaan poikkeavia muotoja, jotka eroavat toisistaan yhdessä tai useammassa emäsparissa. Tietyn geenin sijainnista kromosomissa käytetään nimitystä *lokus*. Kromosomiparin jokaisessa lokuksessa on kaksi *vastingeeniä* eli *alleelia*.

Kromosomin sisäistä rakennetta käsiteltäessä on usein tarpeellista mitata lokusten välisiä etäisyyksiä. Kromosomien pituus vaihtelee kymmenistä miljoonista satoihin miljooniin emäspareihin. Etäisyyden yksikkönä käytetään usein *kiloemästä* (kb, engl. kilobase) eli tuhannen emäsparin pituista DNA-jaksoa.



Kuva 1: Tekijäinvaihdunnassa kromosomiparin vastinkromosomit vaihtavat keskenään geneettistä materiaalia. Näin syntyy uusia geeniyhdistelmiä.

Periytyminen ja kytkentä

Ihmisen jokainen solu sisältää oman kopionsa koko perimästä. Poikkeuksen tähän muodostavat sukusolut, joilla on vain yksi versio kustakin kromosomista. Sukusolut muodostuvat *meioosiksi* kutsutussa solunjakautumisessa, jossa ne saavat kopion jokaisen kromosomiparin jommastakummasta vastinkromosomista. Mukaan tulevat kromosomit valikoituvat satunnaisesti, joten pelkästään tämän ansiosta mahdollisia sukusolun saamia kromosomiyhdistelmiä on vähintään 2^{23}. Hedelmöityksessä munasolun ja siittiön perimät yhdistyvät ja muodostavat kokonaisen kromosomiston. Siten jälkeläiset perivät kustakin kromosomiparista toisen kromosomin äidiltään ja toisen isältä.

Usein kromosomit eivät kuitenkaan periydy sellaisenaan. Sukusolujen muodostumisen yhteydessä saattaa meioosin ensimmäisessä vaiheessa kromosomiparin välillä tapahtua *tekijäinvaihduntaa* (geenienvaihdunta, engl. crossing-over). Siinä vastinkromosomit vaihtavat keskenään osia kytkeytymällä ristiin yhdestä tai useammasta kohdasta (katso kuva 1). Todennäköisyys, että kahden lokuksen välillä tapahtuu vaihdos, on karkeasti ottaen suoraan verrannollinen lokusten väliseen etäisyyteen. Tekijäinvaihdunta aiheuttaa siten etenkin toisistaan kaukana olevien lokusten uudelleenyhdistelyä. Toisaalta läheisten lokusten välillä tapahtuva vaihdunta on harvinaista, ja ne periytyvät useimmiten yhdessä. Tätä läheisten lokusten yhdessä periytymisen ilmiötä kutsutaan *kytkennäksi* (engl. linkage).

Lokuksien välillä olevasta kytkennästä saattaa seurata eri alleeliyhdistelmien poikkeavia esiintymismääriä verrattuna tilanteesta, jossa kytkentää ei olisi. *Alleeliassosiaatio* eli *LD* (kytkentäepätasapaino, engl. linkage disequilibrium) tarkoittaa tällaista kahden lokuksen alleelien välillä ilmenevää assosiaatiota. Assosiaatioita aiheuttavia tekijöitä ovat muun muassa sisäsiittoisuus, mutaatiot ja luonnonvalinta. Näiden li-

säksi assosiaatioita voivat aiheuttaa esimerkiksi geneettinen ajautuminen eli satunnaisuudesta aiheutuva muutos alleelien suhteellissa frekvensseissä, sekä geenivirta eli alleelien leviäminen eri populaatioiden välillä. Kytkentä ei itsessään muodosta uusia alleeliassosiaatioita, vaan ylläpitää jo muodostunutta epätasapainoa. Kytkentään liittyvä alleeliassosiaatio ulottuu normaalisti kymmenien tai satojen kiloemästen etäisyyksille. Tekijäinvaihduntatahti ja siten myös kytkennän voimakkuus kuitenkin vaihtelee voimakkaasti kromosomien sisällä. Tämä vaikuttaa merkittävästi LD:n ulottuvuuteen, ja se onkin selvästi vähäisempää kohdissa, joissa tekijäinvaihdunta on tiheää. Lisäksi assosiaatioetäisyydet riippuvat suurelta osin populaation iästä sekä sen perustajien määrästä. Varsinkin eristyksissä olleilla populaatioilla LD:n on havaittu ulottuvan tavanomaista pidemmälle. Alleeliassosiaatioita on jonkin verran havaittu myös kytkennän vaikutusetäisyyden ulkopuolella sekä kromosomien sisällä että niiden välillä [Hed05, PGC$^+$05].

1.2 Saatavilla oleva perimäaineisto

Tässä tutkielmassa esiteltävät assosiaatioiden etsintään liittyvät menetelmät käsittelevät *markkeriaineistoa*, jota on suhteellisen runsaasti saatavilla. *Markkerit* eli *geenimerkit* ovat tunnettuja DNA-ketjun kohtia, joissa esiintyy merkittävää populaation sisäistä vaihtelua. Niiden fyysinen sijainti kromosomissa tiedetään, ja DNA luetaan näistä sijainneista. Markkereiden tiheys kromosomin sisällä saattaa vaihdella, ja erityisesti sentromeerien ympäristöstä niitä ei normaalisti ole saatavilla.

Markkereita on olemassa useaa eri tyyppiä. Yleisimpiä ovat *yhden emäksen monimuotoisuudet* eli *SNP*:t (engl. single nucleotide polymorphism). Nämä ovat yhden emäsparin lokuksia, joissa esiintyy useaa eri alleelia. Tyypillisesti SNP:t voivat saada kaksi eri arvoa, esimerkiksi A ja G. Muita markkerityyppejä ovat esimerkiksi *mikrosatelliitit*, jotka puolestaan ovat lyhyitä vaihtelevan määrän toistuvia emäsjaksoja. Tästä eteenpäin markkereilla viitataan nimenomaan kaksialleelisiin SNP-markkereihin. Markkereiden kahta mahdollista arvoa merkitsemme isolla ja pienellä kirjaimella, esimerkiksi kirjaimilla A ja a tai kirjaimilla B ja b.

Haplotyyppi ja genotyyppi

Usean markkerin arvojen yhdistelmää yhdessä vastinkromosomissa sanotaan *haplotyypiksi*. Esimerkiksi kaksi markkeria, jotka saavat arvoja A/a ja B/b, voivat yhdessä muodostaa neljä eri haplotyyppiä: AB, Ab, aB ja ab. Näissä ensimmäinen

kirjain tarkoittaa ensimmäisen markkerin sisältämää arvoa ja vastaavasti toinen kirjain toisen markkerin arvoa. Vastaavalla tavalla voidaan muodostaa pidempiäkin haplotyyppejä, esimerkiksi $ABcDe$.

Genotyyppi on kahdesta alleelista muodostettu järjestämätön pari, jolla merkitään markkerin arvoa kromosomiparissa. Markkerin genotyyppi voi siten saada kolme arvoa: $\frac{A}{A}$, $\frac{A}{a}$ sekä $\frac{a}{a}$. Koska genotyypin sisältämän alleeliparin järjestyksellä ei ole merkitystä, ovat $\frac{A}{a}$ ja $\frac{a}{A}$ sama genotyyppi. Näistä kahdesta pyrimme käyttämään ensimmäistä merkintää. Haplotyypin tapaan genotyyppi yleistyy usealle markkerille jonoksi järjestämättömiä pareja. Tässä on huomattava, että esimerkiksi merkinnät $\frac{AB}{a\,b}$ ja $\frac{A\,b}{a\,B}$ tarkoittavat siten täsmälleen samaa kahden markkerin genotyyppiä. Tämä genotyyppi voi kuitenkin muodostua kahdesta erilaisesta haplotyyppiparista. Genotyyppien perusteella ei näin ollen voi kaikissa tapauksissa päätellä alla piileviä haplotyyppejä. Toisaalta jos yksilön kummatkin haplotyypit tiedetään, saadaan genotyypit unohtamalla kunkin markkerin alleeliparin sisäinen järjestys. Jos haluamme viitata *haplotyyppipariin*, käytämme genotyyppimerkintää, jossa haplotyypit on erotettu viivalla. Tässä tapauksessa siis merkinnät $\frac{AB}{a\,b}$ ja $\frac{A\,b}{a\,B}$ vastaavat eri haplotyyppipareja, vaikka niitä vastaavat genotyypit ovatkin samoja.

Tarvittaessa käytämme symbolia $*$ merkitsemään mielivaltaista alleelia tai genotyyppiä. Esimerkiksi $A*$ sisältää siis haplotyypit AB sekä Ab ja haplotyyppipari $\frac{*\,B}{a\,*}$ puolestaan kattaa neljä mahdollista yhdistelmää: $\frac{AB}{a\,B}$, $\frac{AB}{a\,b}$, $\frac{a\,B}{a\,B}$ sekä $\frac{a\,B}{a\,b}$.

Jos populaation parinmuodostuksen oletetaan olevan satunnaista, ovat yksilölle periytyneen kromosomiparin kromosomit toisistaan riippumattomia. Tästä seuraa suoraan, että kunkin yksittäisen markkerin vastinalleelit ovat riippumattomia. Näin ollen, jos eri alleelien frekvenssit populaatiossa ovat tiedossa, voidaan näiden perusteella laskea myös vastaavien genotyyppien frekvenssit. Tällöin sanotaan, että markkeri on *Hardy–Weinberg-tasapainossa* (käytämme myös lyhennettä *HW-tasapaino*).

Aineistokokoelmat ja niiden käyttö

Nykyisin yleisesti käytettävät markkeriaineiston keräysmenetelmät lukevat markkereiden genotyypit eivätkä osaa erottaa haplotyyppejä toisistaan. Tämä on ongelmallista alleeliassosiaation mittaamisen kannalta, sillä ilmiö on nimenomaan yksittäisen kromosomin sisäinen, ei niinkään koko kromosomipariin liittyvä. Yksilön haplotyyppijako voidaan ainakin osittain selvittää sopivien sukulaisten genotyyppien avulla [Sev04, MCP+06]. Usein sukulaisten genotyyppejä ei kuitenkaan ole saatavilla, tai

niiden hankkiminen on kallista. Myös tähän tapaukseen on myös olemassa erilaisia menetelmiä todennäköisten haplotyyppien päättelemiseksi genotyyppiaineistosta [SD03, MCP+06]. Kyseistä prosessia kutsutaan *vaiheistamiseksi* (engl. phasing) ja siihen käytetään laajalti etenkin PHASE-nimistä ohjelmistoa [SD03]. Vaiheistamismenetelmät toimivat tyypillisesti hyvin kuitenkin vain lyhyille pätkille genomia [Sev04, MCP+06]. Lisäksi on kehitetty tapoja mitata alleeliassosiaatiota suoraan genotyyppien perusteella [Wei79]. Sen sijaan mahdollisiin puhtaasti genotyyppien välisiin assosiaatioihin ei ole samassa määrin kiinnitetty huomiota. Tietynlaiset genotyyppiassosiaatiot eivät välttämättä erotu alleeliassosiaatiota mitattaessa, joten tähän on tähän voidaan kuitenkin nähdä perusteita.

Nykyiset tehokkaat markkeriaineiston keräysmenetelmät ovat mahdollistaneet uusia tutkimustapoja ja -kohteita. Tärkeä esimerkki ovat koko genomin laajuiset assosiaatiotutkimukset, joissa LD:n avulla pyritään paikantamaan esimerkiksi taudeille altistavia geenejä [CM98]. Kansainvälinen *HapMap-projekti* [Con03] puolestaan on pyrkinyt kartoittamaan ihmisen genomin eri kohtien vaihtelua ja siinä sivussa luomaan koko genomin laajuisen LD-kartan. Projektia varten on kerätty noin sadan–kahdensadan henkilön aineistokokoelmia useista eri populaatioista eri mantereilta ja kerätyt kokoelmat on asetettu vapaasti saataville.

Tässä tutkielmassa esiteltävissä kokeissa on käytetty pääosin aineistokokoelmaa, johon viittaamme jatkossa `NFBC`:nä (Northern Finland Birth Cohort). Kokoelma sisältää 335 118 markkerin genotyypit 5363:lta vuonna 1966 syntyneeltä pohjoissuomalaiselta henkilöltä. Jotkin kokeet on vertailukohdan saamiseksi ja tulosten yleistymisen tarkistamiseksi suoritettu myös muutamalla HapMap-projektin populaatiokokoelmalla[1]. Nämä sisältävät genotyypit noin miljoonalle markkerille. Kokeissa tosin on rajoituttu 264 680 markkeriin, jotka esiintyvät myös `NFBC`-kokoelmassa. Käytetyt HapMap-kokoelmat ovat `CHB` (137 näytettä; ei-sukulaisia, Han-kiinalaisia, Beijing, Kiina), `MEX` (86 näytettä; vanhemmat-lapsi-kolmikkoja, meksikolaista alkuperää, Los Angeles, Kalifornia), `MKK` (184 näytettä; ei-sukulaisia sekä vanhemmat-lapsi-kolmikkoja, Maasai-kansaa, Kinyawa, Kenia) ja `YRI` (203 näytettä; vanhemmat-lapsi-kolmikkoja, Joruba-kansaa, Ibadan, Nigeria).

[1]Käytimme vapaasti saatavilla olevia HapMap 3 release 3 -julkaisun aineistoja, jotka ovat ladattavissa osoitteessa `http://www.sanger.ac.uk/humgen/hapmap3/`.

2 Assosiaation mittaaminen

Tässä luvussa käsittelemme erilaisia menetelmiä markkerien välisten assosiaatioiden mittaamiseen. Assosiaatiota käytämme synonyymina tilastolliselle riippuvuudelle. Markkereita ajattelemme eräänlaisina diskreetteinä satunnaismuuttujina, jotka voivat tilanteesta riippuen saada arvoiksi joko alleeleita tai genotyyppejä. Sanomme kahden markkerin olevan assosioituneita, jos niiden saamien arvojen välillä on tilastollinen riippuvuus.

Markkeriparit ja piirreassosiaatiot

Assosiaation mittaamisella voidaan yleisesti viitata paitsi kahden markkerin välisten riippuvuuksien, myös markkerin ja ulkoisen piirteen välisen yhteyden tutkimiseen. Jos tutkittava piirre on selkeästi nominaalinen eli saa vain muutamia diskreettejä arvoja, voidaan näihin tyypillisesti soveltaa samoja menetelmiä. Nominaalinen piirre voi olla esimerkiksi perinnöllinen sairaus. Vahva motivaatio piirre–markkeri-assosiaatioiden mittaamisen takana onkin esimerkiksi tiettyyn sairauteen altistavien geenien etsiminen [CM98, Sev04]. Tässä hyödynnetään kromosomin sisäistä kytkentää; jos geenin ja piirteen välillä on selkeä riippuvuus, heijastuu se usein myös geeniä lähellä sijaitsevien markkerien assosioitumiseen piirteen kanssa.

Markkeri–markkeri-assosiaatioiden avulla puolestaan voidaan esimerkiksi pyrkiä tutkimaan perimän sisäistä rakennetta ja organisoitumista. Esimerkiksi kytkennän voimakkuuden vaihtelua kuvaavat LD-kartat perustuvat tyypillisesti läheisten markkerien välisen alleeliassosiaation mittaamiseen. Yleistä LD:n voimakkuutta eri kromosomeissa on käytetty pohjana myös vaikkapa arvioitaessa efektiivistä populaation kokoa eli parametria, joka määrittää geneettisen muuntelun ja geneettisen ajautumisen määrän populaatiossa [TNH$^+$07]. Tässä tutkielmassa tavoitteenamme on etsiä kromosomien välisiä riippuvuuksia, joten jatkossa assosiaatioilla viitataan nimenomaan markkerien välisiin assosiaatioihin.

Koko kromosomiston markkeriaineisto saattaa sisältää satoja tuhansia tai miljoonia markkereita. Jotta kaikki mahdollisesti olemassa olevat assosiaatiot löydettäisiin, halutaan luultavasti käydä kaikki markkeriparit läpi. Etsinnän rajoittaminen eri kromosomeissa sijaitsevien markkerien välille pienentää hieman läpikäytävien parien määrää. Vähennys ei kuitenkaan ole merkittävä, ja joka tapauksessa testattavia markkeripareja saattaa olla kymmenistä miljardeista jopa biljooniin. Esimerkiksi kokeissa pääasiassa käytetty `NFBC`-aineisto sisältää yhteensä 335 118 markkeria. Kro-

mosomien välisiä markkeripareja näistä muodostuu noin $5.3 \cdot 10^{10}$. Testien suuresta määrästä johtuen on toivottavaa, että yksittäinen markkeriparia koskeva testi on mahdollisimman yksinkertainen ja nopea suorittaa. Tämä asettaa rajoituksia käytettäville tekniikoille.

Useamman markkerin assosiaatiot

Myös monimutkaisempia esimerkiksi kolmen tai neljän markkerin interaktioita voitaisiin haluta etsiä. Kaiken kaikkiaan eri kolmikkoja tai nelikkoja alkaa kuitenkin olla liian paljon läpikäytäväksi. Toisaalta esimerkiksi kahden markkerin sekä piirteen välisiä assosiaatioita on jonkin verran tutkittu[MDC05, EMMC06, MSL+07, KL08]. Tällaisessa tapauksessa kaikkien mahdollisten markkeriparien läpikäyminen saattaa vielä olla tilanteesta riippuen mahdollista ja järkevää. Vaihtelevalla menestyksellä on kokeiltu myös strategiaa, että markkeripareihin valitaan ainoastaan sellaisia markkereita, jotka jo yksinään osoittavat jonkin asteista assosiaatiota piirteen kanssa.

Samankaltaista ahnetta menetelmää voitaisiin luonnollisesti soveltaa myös useiden geenien välisten riippuvuuksien hakuun: Ensin haettaisiin kaikki valitun merkitsevyyskynnyksen ylittävät markkeriparin assosiaatiot. Merkitsevyyskynnyksen asettamisen sijaan voitaisiin myös valita ennalta päätetty lukumäärä vahvimpia assosiaatioita. Tämän jälkeen kaikista löydetyistä pareista muodostettaisiin markkerikolmikkoja. Tämä tehtäisiin esimerkiksi liittämällä kuhunkin pariin vuorollaan kaikki muut yksittäiset markkerit. Siis parista XY muodostettaisiin kaikki kolmikot muotoa XYZ, missä Z on mielivaltainen. Vaihtoehtoisesti voitaisiin vaatia, että myös YZ, XZ tai molemmat olisivat löytyneiden assosiaatioiden joukossa. Muodostetuille kolmikoille suoritettaisiin tämän jälkeen vastaava assosiaatiotestaus kuin markkeripareille. Samaan tapaan voitaisiin jatkaa neljä- ja korkeampiasteisiin assosiaatioihin.

Ahne haku perustuu oletukseen, että korkeampiasteiset assosiaatiot havaittavissa määrin heijastuvat myös matalimpiin asteisiin, ja näin varmasti onkin useissa tapauksissa. Ainakin teoriassa on kuitenkin mahdollista muodostaa assosiaatioita, joissa tällaista ei ilmene. Esimerkiksi tällaisesta on pariteettifunktio: Olkoot X, Y ja Z markkereita, jotka kukin voivat saada kaksi arvoa. Merkitään kaikkien kahta mahdollista arvoa kirjaimilla A ja a. Kuvitellaan, että aineistossa esiintyy tasaisesti sellaisia näiden kolmen markkerin yhdistelmiä, joissa arvojen A määrä on parillinen. Sen sijaan parittomia A:n määriä ei esiinny ollenkaan. Tällöin kolmikon keskinäinen assosiaatio on erittäin vahva; kolmannen markkerin arvo voidaan päätellä varmasti, kun kahden muun arvo tiedetään. Sen sijaan esimerkiksi X:n arvo ei yksinään kerro

mitään Y:n arvosta, joten assosiaatio ei heijastu markkeriparien välille.

Usean markkerin assosiaatiot olisivat mielenkiintoinen tutkimuskohde. Tässä tutkielmassa keskitymme kuitenkin ainoastaan kahden eri sijainnin välisiin riippuvuuksiin.

Markkeriryhmien väliset assosiaatiot

Kytkennän ansiosta haplotyyppien monimuotoisuus kromosomin lyhyellä osalla on usein rajoittunut muutamaan populaatiossa esiintyvään alleelisekvenssiin [Sev04]. Esimerkiksi voi olla, että kolmessa peräkkäisessä markkerissa esiintyy kaikista kahdeksasta mahdollisesta haplotyypistä vain kolmea: ABc, Abc ja aBC. Tyypillisesti nämä tietyssä kromosomin kohdassa esiintyvät eri haplotyyppisekvenssit on tunnistettavissa muutaman lähekkäisen markkerin arvojen perusteella, mutta mikään yksittäinen markkeri ei tähän identifioimiseen välttämättä riitä. Ehkäpä muutamasta tällaisesta lähekkäin sijaitsevasta markkereista koostuva genotyyppiyhdistelmä saattaisi siten olla assosioitunut toisessa kromosomissa sijaitsevan markkerin tai markkeriryhmän kanssa, vaikka yksittäisten markkerien välillä ei olisi havaittavissa merkitsevää riippuvuutta. Tämän idean motivoimana kokeilemme assosiaatioiden hakemista yksittäisten markkerien lisäksi myös muutamasta peräkkäisestä markkerista koostuvilla markkeriryhmillä.

Markkeriryhmien välisiä assosiaatioita voisi äkkiseltään pitää pelkästään karsittuna versiona yllä mainitusta useamman markkerin assosiaatiosta; mukaan olisi otettu ainoastaan sellaisia markkerien kokoelmia, jotka muodostavat markkeriryhmiä. Erona on, että toisin kuin useamman markkerin assosiaatioissa, tässä tapauksessa mittauksen kohteena on ainoastaan ryhmien välinen assosiaatio, ja sisäinen assosiaatio jätetään huomiotta. LD-ilmiön ansiosta nimenomaan ryhmän sisäinen vierekkäisten markkerien riippuvuus on hyvin todennäköistä. Tästä emme kuitenkaan kromosomien välillä tapahtuvia assosiaatioita mitatessamme ole kiinnostuneita.

Ristiintaulu ja assosiaation havaitseminen

Assosiaatioita mitattaessa on luonnollista olettaa genotyyppinäytteiden olevan riippumattomia ja samoin jakautuneita (iid). Koska näytteiden järjestyksellä ei ole väliä, voidaan kaikki markkeriparin kannalta oleellinen tieto tällöin esittää *ristiintaulun* (usein myös *kontingenssitaulu*) avulla. Ristiintaulu on taulukko, joka sisältää käsiteltävien muuttujien kaikkien eri arvoyhdistelmien esiintymismäärät eli frekvenssit. Se siis kuvaa muuttujien empiirisesti havaittua yhteisjakaumaa.

Taulukko 1: Kaksi esimerkkiä mahdollisista markkerien X ja Y ristiintaulusta: ensimmäisessä esiintymismäärät on laskettu haplotyypeille, toisessa genotyypeille. Laidoille on summattu kummankin markkerin reunafrekvenssit ja oikeassa alakulmassa on näytteiden kokonaismäärä.

(a) Alleelit

X	Y		
	B	b	$*$
A	6	2	8
a	2	5	7
$*$	8	7	15

(b) Genotyypit

X	Y			
	$\frac{B}{B}$	$\frac{B}{b}$	$\frac{B}{b}$	$*$
$\frac{A}{A}$	43	54	22	119
$\frac{A}{a}$	16	101	34	151
$\frac{a}{a}$	49	63	19	131
$*$	108	218	75	401

Taulukot 1a ja 1b ovat esimerkkejä alleeleihin ja genotyyppeihin perustuvista ristiintauluista. Ensimmäiseen taulukkoon on kirjattu markkeriparin kaikkien neljän haplotyypin havaitut frekvenssit. Haplotyyppejä AB ja ab esiintyy aineistossa selvästi enemmän kuin haplotyyppejä Ab ja aB, vaikka yksittäin alleelit ovat kummassakin markkerissa suurin piirtein yhtä yleisiä. Tämän perusteella näyttäisi, että markkereiden välillä olisi selvä assosiaatio. Toinen taulukko on puolestaan esimerkki mahdollisista genotyyppien frekvensseistä. Myös tässä taulussa vaikuttaisi ilmenevän jonkin asteista epätasapainoa: etenkin genotyyppiä $\frac{AB}{aB}$ esiintyy selvästi vähemmän, kuin mitä reunafrekvenssien perusteella saattaisi odottaa.

Kuten olemme todenneet, tyypillisesti assosiaatiotestauksissa on keskitytty alleelien ja piirteen tai kahden alleelin välisiin riippuvuuksiin. Genotyyppejä käsittelevä ja erityisesti ne huomioon ottava tutkimus sitä vastoin on jäänyt merkittävästi vähemmälle huomiolle. Genotyyppien välillä voisi kuitenkin ainakin teoriassa olla assosiaatioita, jotka eivät näy alleeleihin perustuvassa analyysissä. Taulukko 1b on esimerkki tämän tyyppisestä assosiaatiosta, kuten myöhemmin näemme. Lisäksi koska saatavilla oleva perimäaineisto on pääasiassa genotyyppiaineistoa ja vaiheistamisella saadaan tyypillisesti vain approksimaatio haplotyypeistä, keskitymme genotyyppiassosiaatioihin.

Taulukoissa 2a ja 2b on `NFBC`-aineistosta satunnaisesti valituille muuttujapareille muodostetut ristiintaulut. Ensimmäisen taulun muuttujat ovat yksittäisiä markkereita, toisessa taulussa puolestaan kolmen peräkkäisen markkerin muodostamia

Taulukko 2: Kaksi todellisesta genotyyppiaineistosta satunnaisesti poimittua ristiintaulua. Muuttujien eri arvoja (genotyyppejä) on merkitty numeroin. Jälkimmäisestä taulukosta on jätetty pois ne rivit ja sarakkeet, joita vastaavia muuttujien arvoja ei esiinny aineistossa (eli reunafrekvenssit olisivat nollia). Taulut ovat samasta aineistosta vaikka ruutujen yhteenlaskettu summa näissä poikkeaakin. Ero johtuu aineiston epätäydellisyydestä: testattavissa markkereissa puuttuvia arvoja sisältävät näytteet on jätetty laskuista pois. Ensimmäisessä taulussa muuttujat ovat yksittäisiä markkereita, toisessa puolestaan kolmen peräkkäisen markkerin muodostamia markkeriryhmiä. Jälkimmäinen taulu on suhteellisen harva: osa reunafrekvensseistä on pieniä, ja siten useiden solujen odotusarvo on erittäin lähellä nollaa. Esimerkiksi vasemman yläkulman yhdistelmän $X = 1$ ja $Y = 1$ frekvenssin odotusarvo on $3 \cdot 1/5350 = 0.00056$.

(a) Yksittäiset markkerit

X	Y			
	1	2	3	$*$
1	21	102	179	285
2	99	682	1204	1790
3	165	1006	1897	3280
$*$	302	1985	3068	5355

(b) Kolmen peräkkäisen markkerin ryhmät

X	Y																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	$*$
1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	3
2	0	0	0	9	2	0	11	0	6	15	8	11	22	4	15	13	116
3	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	2
4	0	1	0	9	0	0	14	0	16	24	11	18	17	9	20	10	149
5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
6	0	0	0	2	0	0	1	0	5	5	1	0	9	1	4	3	31
7	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	3
8	0	1	0	15	2	1	34	0	31	39	21	31	36	19	43	23	296
9	0	0	0	3	0	0	10	0	3	4	3	7	6	6	9	3	54
10	0	0	0	1	0	0	0	0	1	0	0	1	1	0	1	0	5
11	0	1	1	22	2	2	62	1	50	82	37	81	91	28	83	48	591
12	0	0	0	22	2	1	65	0	42	76	32	54	81	19	83	40	517
13	0	1	0	17	1	1	23	0	34	28	13	29	38	12	46	24	267
14	0	0	1	13	2	2	20	0	28	31	19	29	32	16	32	24	249
15	0	0	0	9	0	0	19	0	11	20	14	17	30	9	18	16	163
16	0	0	0	6	0	1	7	0	5	6	3	8	7	6	11	4	64
17	0	0	0	0	0	0	0	0	0	0	0	0	4	2	2	0	8
18	0	2	0	30	1	2	58	0	44	70	34	60	80	23	77	55	536
19	0	4	0	25	3	1	64	0	50	89	37	76	88	33	83	62	615
20	0	0	0	6	1	2	9	0	10	9	5	9	9	6	13	4	83
21	1	1	0	20	1	0	45	0	26	41	25	43	59	26	65	38	391
22	0	2	0	39	3	4	85	0	73	101	43	78	117	40	122	85	792
23	0	0	1	19	4	2	39	0	46	53	24	45	56	27	46	52	414
$*$	1	13	3	267	24	20	566	1	483	694	330	600	783	286	773	506	5350

markkeriryhmiä. Huomattavaa on, että jälkimmäisessä taulussa on paljon nollafrekvenssejä. Tämä on seurausta pienistä reunafrekvensseistä, jotka aiheuttavat nollan lähellä olevia odotusarvoja taulukon soluille. Tällaista taulua sanotaan harvaksi. Markkeriryhmiä muuttujina käytettäessä voivat muuttujat saada suuren määrän eri arvoja, ja siten ristiintauluista muodostuu tyypillisesti suuria ja harvoja.

Assosiaation vahvuus ja merkitsevyys

Assosiaatioita tutkittaessa tulisi erottaa kaksi asiaa toisistaan: assosiaation vahvuus ja sen tilastollinen merkitsevyys. Vahvuudella tarkoitetaan tässä suhteellista poikkeamaa riippumattomuudesta. Merkitsevyys puolestaan viittaa havainnon harvinaisuuteen. Esimerkiksi taulukossa 1a markkerien välillä näyttäisi olevan varsin vahva riippuvuus. Koska näytteiden määrä on kuitenkin pieni, voi tällainen näennäinen assosiaatio kuitenkin hyvin olla sattumaa, eikä poikkeama ole siten tilastollisesti kovinkaan merkitsevä. Assosiaatioita etsiessämme emme halua, että mahdolliset löydöksemme sekoittuvat tällaiseen satunnaiskohinaan, joten tarvitsemme keinon mitata merkitsevyyttä. Tavoitteista riippuen voidaan haluta etsiä esimerkiksi tietyn vahvuuden ylittäviä assosiaatioita, tai sitten vahvuus voidaan jättää kokonaan huomiotta. Tässä tutkielmassa meitä kiinnostaa pääasiassa assosiaatioiden olemassaolo, eikä niinkään niiden vahvuus, joten keskitymme ainoastaan merkitsevyyteen.

Jotta voisimme mitata merkitsevyyttä, kiinnitämme nollahypoteesin. Se on tapauksessamme oletus, että tarkasteltavat markkerit ovat riippumattomia. Vaihtoehtoisen hypoteesin mukaan puolestaan jonkinlainen riippuvuus on olemassa. Mahdollisen riippuvuuden tyyppiin emme ota kantaa. Hypoteesin testausta varten täytyy kuitenkin määritellä tarkemmin mitä tarkoitamme markkerien tai markkeriryhmien välisellä riippumattomuudella. Ristiintaulujen tuottamiseen voidaan liittää kaksi mallia [Zar99].[2] *Vapaiden reunojen mallissa* oletetaan näytteiden kokonaislukumäärä n ennalta valituksi ja reunafrekvenssit pääsevät vaihtelemaan vapaasti. Nollahypoteesin vallitessa näytteet ovat riippumattomia ja samoin jakautuneita. Tällöin kumpaankin muuttujaan liittyvät reunafrekvenssit noudattavat multinomijakaumaa. Ristiintaulun sisäsolut noudattavat näitä vastaavaa *tulomultinomijakaumaa* (engl. product multinomial distribution) eli multinomijakaumaa, jonka todennäköisyydet saadaan

[2]Oikeastaan ristiintauluihin liittyviä tyypillisiä malleja on ainakin neljä: Kolmannessa mallissa vain toinen marginaali oletetaan ennalta valituksi [Zar99]. Tapauksessamme muuttujat ovat tasaarvoisia, joten kolmannelle mallille ei ole perusteita. Neljännessä mallissa puolestaan myös näytteiden kokonaismäärä voi vaihdella ja solujen arvot ovat Poisson-jakautuneita [Agr02]. Oletamme kuitenkin näytteiden määrän olevan ennalta valittu.

reunatodennäköisyyksien tuloina. *Kiinnitettyjen reunojen mallissa* oletetaan näytteiden määrän lisäksi myös reunafrekvenssit ennalta kiinnitetyiksi, jolloin nollahypoteesin vallitessa taulukon sisäsolujen frekvenssit noudattavat *moninkertaista hypergeometrista jakaumaa* (engl. multiple hypergeometric distribution). Kyseinen jakauma saadaan, kun tulomultinomijakauma ehdollistetaan havaituilla reunafrekvensseillä. Kiinnitettyjen reunojen mallissa näytteet eivät oikeastaan ole täysin riippumattomia, vaikka ovatkin samoin jakautuneita. Isolla näytekoolla tällä ei kuitenkaan pitäisi olla suurta merkitystä.

Yllä mainituista vapaiden reunojen malli tuntuisi vastaavan tilannettamme paremmin. Siihen liittyy kuitenkin tuntemattomia parametreja, nimittäin yksittäisen muuttujan (markkerin) arvojen todennäköisyydet. Toisaalta kiinnitettyjen reunojen tapauksessa nollajakauma määräytyy suoraan tunnettujen reunafrekvenssien perusteella, mistä on apua jatkossa. Tätä myös käytetään yleisesti vaikka todellisuudessa aineisto noudattaisi vapaiden reunojen mallia. Esimerkiksi yleisesti käytetty satunnaisotosmenetelmä, näytteenotto ilman takaisinpanoa tuottaa näytteitä lukittujen reunojen mallin mukaisesta jakaumasta. Jatkossa esittelemme testejä kummallekin mallille.

Tilastollisen merkitsevyyden mittana käytetään p-arvoa, joka on reaaliluku väliltä $[0, 1]$. Merkitsemme p-arvoa symbolilla p. Havainnon p-arvo kertoo todennäköisyyden saada vähintään yhtä äärimmäinen tai poikkeava havainto, kun nollahypoteesi oletetaan todeksi. Pienet p-arvot kuvaavat siis erityisen poikkeavia havaintoja, tapauksessamme siis tilanteita, joissa markkereiden välillä näyttäisi olevan jokin riippuvuus. Ongelmana on ”äärimmäisyyden” määritteleminen formaalisti siten, että p-arvo on mahdollista laskea – mieluusti vieläpä helposti ja nopeasti. Yksinkertaisissa tapauksissa havainnon äärimmäisyyttä voidaan mitata esimerkiksi suoraan sen tapahtumistodennäköisyydellä, tai äärimmäisyyttä kuvaava arvo voidaan jopa määritellä erikseen jokaiselle mahdolliselle havainnolle. Tyypillisesti kuitenkin käytetään havainnon perusteella laskettavaa *tunnuslukua*, tilastollista suuretta, jonka mittaa äärimmäisyyttä halutulla tavalla. Merkitsemme jatkossa havaittua tunnuslukua kirjaimella t ja nollahypoteesin vallitessa vastaavaa satunnaismuuttujaa T. Tavallisesti äärimmäisinä tapauksina pidetään niitä, joilla tunnusluku on tarpeeksi suuri (tai pieni). Siis p-arvo on $\mathrm{p} = \Pr(T \geq t)$ tai $\mathrm{p} = \Pr(T \leq t)$. Tässä merkintä $\Pr(\cdot)$ tarkoittaa tapahtuman todennäköisyyttä nollahypoteesin mukaisessa tilanteessa. Kaksipuoleisessa testissä äärimmäisyyttä mitataan kumpaankin suuntaan. Tällöin p-arvo voidaan määritellä usealla tavalla [Agr02]. Tässä tutkielmassa käytämme määritelmää $\mathrm{p} = 2\min(\Pr(T \geq t), \Pr(T \leq t))$.

Jatkuva-arvoisten havaintojen tapauksessa p-arvon on suoraan määritelmän mukaan tasaisesti jakautunut. Koska tapauksessamme erilaisten ristiintaulujen määrä on rajattu, on p-arvon jakauma kuitenkin diskreetti. Käytännössä tämä näkyy p-arvon konservatiivisuutena, mikä johtaa p-arvojen jakauman painottumiseen lähemmäs lukua yksi. Diskreetin tunnusluvun tapauksessa valitsemamme kaksipuoleisen p-arvon laskumenetelmä saattaa saada ykköstä suuremman arvon. Näissä tapauksissa pyöristämme luvun alaspäin ykköseksi. Koska mielenkiintomme kohteena ovat pääasiassa pienet p-arvot, ei asialla ole juuri merkitystä.

Jos tunnusluvun jakauma nollahypoteesin vallitessa tiedetään, voidaan havainnon perusteella laskea vastaava yksi- tai kaksipuoleinen p-arvo. Tarkasti laskettavia p-arvoja käsittelemme luvuissa 2.4 ja 2.5. Usein tunnetaan vain approksimaatio tunnusluvun todellisesta jakaumasta, mutta sen avulla saadaan riittävän tarkka arvio p-arvosta. Luvut 2.1 ja 2.2 liittyvät tämän tyyppisiin tunnuslukuihin. Mikäli tunnusluvun jakauma ei ole tiedossa, voidaan sitä yrittää arvioida näytteenotolla. Tätä käsitellään luvussa 2.3.

2.1 Alleeliassosiaatio

Alleeliassosiaation mittaamiseen on vuosien varrella esitelty lukuisia eri tunnuslukuja. Eri mittoja on vertailtu esimerkiksi Devlinin ja Rischin artikkelissa [DR95]. Tutustumme tässä yleisesti käytettyyn Δ-mittaan. Olkoot tarkasteltavana kaksi SNP-markkeria X ja Y. Markkeri X voi saada arvon A tai a ja vastaavasti Y voi saada arvon B tai b. Erilaisia mahdollisia haplotyyppejä on siis neljä: AB, Ab, aB sekä ab. Tarkastellaan aineistoa, joka sisältää yhteensä n näytettä. Merkitään haplotyyppien esiintymismääriä n_{AB}, n_{Ab}, n_{aB} sekä n_{ab}. Nämä muodostavat markkerien havaitun yhteisjakauman. Vastaavasti merkitään yksittäisten alleelien A, a, B ja b kokonaismääriä n_A, n_a, n_B ja n_b. Nyt esimerkiksi $n_A = n_{AB} + n_{Ab}$ ja toisaalta $n = n_A + n_a$. Käytetyt merkinnät on koottu taulukkoon 3a.

Voidaan ajatella, että kukin näyte on poimittu riippumattomasti alla piilevästä nollahypoteesin mukaisesta todennäköisyysjakaumasta. Merkitään haplotyyppien ja alleelien todennäköisyyksiä samaan tapaan kuin frekvenssejä käyttäen n:n sijaan p:tä: siis esimerkiksi p_{Ab} ja p_B. Jakamalla havaintojen lukumäärät n:llä saadaan arviot kunkin haplotyypin ja alleelin todennäköisyyksistä. Käytetään näistä merkintöjä $\hat{p}_{Ab} = n_{Ab}/n$, $\hat{p}_B = n_B/n$, jne. Kiinnitettyjen reunojen mallissa saadut arviot tarkkoja, ja vapaiden reunojen mallissa ne ovat suurimman uskottavuuden estimaatteja alla oleville todennäköisyyksille.

Taulukko 3: Merkinnät markkerien X ja Y arvojen havaituille frekvensseille sekä kummankin markkerin reunafrekvensseille. Vastaavat empiiriset todennäköisyydet saadaan jakamalla havaintojen määrät n:llä.

(a) Frekvenssit haplotyypeille

X	Y		
	B	b	$*$
A	n_{AB}	n_{Ab}	n_A
a	n_{aB}	n_{ab}	n_a
$*$	n_B	n_b	n

(b) Frekvenssit genotyypeille

X	Y			
	$\frac{B}{B}$	$\frac{B}{b}$	$\frac{B}{b}$	$*$
$\frac{A}{A}$	n_{AB}^{AB}	$n_{A\,b}^{AB}$	n_{Ab}^{Ab}	n_A^A
$\frac{A}{a}$	$n_{a\,B}^{AB}$	$n_{a\,b}^{AB}$	$n_{a\,b}^{Ab}$	n_a^A
$\frac{a}{a}$	n_{aB}^{aB}	$n_{a\,b}^{aB}$	n_{ab}^{ab}	n_a^a
$*$	n_B^B	n_b^B	n_b^b	n

Nyt jos markkerit X ja Y ovat toisistaan riippumattomia, saadaan tietyn haplotyypin todennäköisyys alleelien todennäköisyyksien tulona: $p_{AB} = p_A \cdot p_B$. Merkitään poikkeamaa $D = p_{AB} - p_A \cdot p_B$. Voidaan helposti havaita, että poikkeaman itseisarvo on sama kaikille neljälle haplotyypille:

$$\begin{aligned} D &= p_{AB} - p_A \cdot p_B \\ &= p_{ab} - p_a \cdot p_b \\ &= -p_{Ab} + p_A \cdot p_b \\ &= -p_{aB} + p_a \cdot p_B. \end{aligned}$$

Määritellään korrelaation vahvuutta kuvaava Δ-mitta normalisoimalla D:

$$\Delta = \frac{D}{\sqrt{p_A p_a p_B p_b}}. \tag{1}$$

Ajatellaan seuraavaksi alleelien lukuarvoiksi $A = B = 1$ ja $a = b = 0$. Huomataan, että

$$\Delta = \frac{p_{AB} - p_A \cdot p_B}{\sqrt{p_A(1-p_A)}\sqrt{p_B(1-p_B)}} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \mathrm{Corr}(X,Y).$$

Siis Δ on itse asiassa markkerien X ja Y välinen korrelaatiokerroin ja vaihtelee siten välillä $[-1, 1]$. Jos $\Delta = 0$, ovat X ja Y riippumattomia. Muussa tapauksessa Δ:n itseisarvo kertoo markkereiden välisen korrelaation suuruuden. Selkeästi Δ:aa voidaan käyttää nimenomaan aiemmin mainitun assosiaation vahvuuden mittaamiseen,

kunhan tuntemattomien todennäköisyyksien p tilalta käytetään arvioita $\hat{p}$. Tyypillisesti saatu Δ korotetaan toiseen, jolloin päästään eroon ”turhasta” etumerkistä. Jos neliö kerrotaan vielä näytteiden määrällä n, saadaan tuloksena Pearsonin khii toiseen -testin tunnusluku, joka X:n ja Y:n ollessa riippumattomia noudattaa likimain χ^2-jakamaa yhdellä vapausasteella [Wei79]:

$$T_\Delta = n\Delta^2 \overset{approx}{\sim} \chi^2(1).$$

Havaituille poikkeamille saadaan tunnetun jakauman ansiosta laskettua helposti p-arvot. Esitellyn mitan avulla voidaankin siten kätevästi mitata sekä markkerien välisen korrelaation vahvuutta että sen merkitsevyyttä. Esimerkiksi taulukosta 1a saadaan $\Delta = 0.4643$ ja edelleen käyttämällä χ^2-jakaumaa p = 0.0721. Esimerkin assosiaatio on suhteellisen vahva, kuten aiemmin todettiin, mutta p-arvo ei alita tyypillistä merkitsevyysrajaa 0.05.

Alleeliassosiaatio genotyypeille

Yllä esitelty Δ lasketaan haplotyyppien frekvenssien perusteella. Haluaisimme kuitenkin mitata assosiaatiota vaiheistamattomalla genotyyppiaineistolla. Pääasialliseksi ongelmaksi muodostuu tässä se, ettei esimerkiksi haplotyyppiyhdistelmiä $\frac{AB}{a\ b}$ ja $\frac{A\ b}{a\ B}$ voida erottaa toisistaan. Näin ollen haplotyypin AB frekvenssiä ei voida laskea. Merkitään jatkossa genotyyppien ja haplotyyppiparien frekvenssejä ja todennäköisyyksiä samaan tapaan kuin haplotyypeilläkin: genotyypille $\frac{AB}{a\ b}$ esimerkiksi $n^{\frac{AB}{a\ b}}$ ja $p^{\frac{AB}{a\ b}}$ ja haplotyyppiparille $\frac{AB}{a\ b}$ vastaavasti $n_{\frac{AB}{a\ b}}$ ja $p_{\frac{AB}{a\ b}}$. Genotyyppien merkinnät on koottu taulukkoon 3b.

Jos markkeri on HW-tasapainossa, ovat alleeliparin osapuolet riippumattomia. Siten genotyyppien todennäköisyydet saadaan $p_A^A = {p_A}^2$, $p_a^A = 2p_Ap_a$ ja $p_a^a = {p_a}^2$. Markkerin Hardy–Weinberg-poikkeama (HW-poikkeama) määritellään samaan tapaan kuin kytkentäpoikkeama: $D_A = p_A^A - {p_A}^2$.

Weirin [Wei79] esittelemä komposiittipoikkeama D^c on summa suoran alleeliparin $\frac{* *}{AB}$ (eli haplotyypin AB) poikkeamasta sekä ristikkäisen alleeliparin $\frac{*B}{A*}$ vastaavasta poikkeamasta. Tätä puolestaan voidaan arvioida genotyyppien avulla seuraavasti:

$$\begin{aligned} D^c &= (p_{\frac{* *}{AB}} - p_Ap_B) + (p_{\frac{*B}{A*}} - p_Ap_B) \\ &= (p_{\frac{* *}{AB}} + p_{\frac{*B}{A*}}) - 2p_Ap_B \\ &= 2p_{AB}^{AB} + p_{A\,b}^{AB} + p_{a\,B}^{AB} + \frac{1}{2}p_{a\ b}^{AB} - 2p_Ap_B. \end{aligned}$$

Nyt määritellään edellä esiteltyä Δ-mittaa vastaava *komposiittikorrelaatio*

$$\Delta^c = \frac{D^c}{\sqrt{(p_{A*}p_{a*} + D_A)(p_{*B}p_{*b} + D_B)}}.$$

Jos kumpikin markkeri on Hardy–Weinberg-tasapainossa, on $D_A = D_B = 0$, ja jakaja on siten sama kuin Δ:ssa. Lisäksi tällöin $p^{*B}_{A*} = p_{A*}p_{*B}$, joten $D^c = D$ ja siis $\Delta^c = \Delta$. HW-tasapainon vallitessa komposiittikorrelaatio palautuu siten yksittäisen haplotyypin sisäistä assosiaatiota mittaavaan Δ-korrelaatioon.

Samoin kuin aiemmin alleelien tapauksessa, annetaan genotyypeille nyt vastaavalla tavalla lukuarvot $\frac{A}{A} = \frac{B}{B} = 0$, $\frac{A}{a} = \frac{B}{b} = \frac{1}{2}$ ja $\frac{a}{a} = \frac{b}{b} = 1$. Jälleen voidaan osoittaa, että $\Delta^c = \mathrm{Corr}(X, Y)$ [Zay04]. Komposiittikorrelaatio siis mittaa markkerien alleelien lukumäärien välistä *lineaarista riippuvuutta*. Samoin kuin $n\Delta^2$, myös $T_{\Delta^c} = n(\Delta^c)^2$ on riippumattomuusoletuksen vallitessa likimain $\chi^2(1)$ -jakautunut [ZPW08].

Komposiittikorrelaatio mahdollistaa LD:n vahvuuden ja merkitsevyyden laskemisen genotyyppiaineistolle ilman vaiheistusta ja toimiikin siinä tehtävässä varsin hyvin. Se on tarkoitettu kuitenkin edelleen vain alleeliassosiaation mittaamiseen, eikä siten välttämättä havaitse kaikkia puhtaasti genotyyppeihin perustuvia assosiaatioita. Tämä seuraa suoraan yllä mainitusta korrelaatiotulkinnasta: kaikki muuttujien väliset riippuvuudet eivät välttämättä heijastu lineaarista assosiaatiota mittaavaan korrelaatiokertoimeen. Esimerkiksi taulukosta 1b komposiittikorrelaation arvoksi saadaan $\Delta^c = 0.0753$. Tästä puolestaan saadaan p-arvoksi 0.1313, joten taulukon suhteellisen selkeä assosioituminen ei siis juurikaan heijastu komposiittikorrelaatioon.

Alleeliassosiaation yleistyksiä

Eri LD-mitoista on kehitetty yleistyksiä esimerkiksi markkeriryhmille [NFR02, BB07], useammalle kuin kahdelle markkerille [KFZ08] sekä markkereille, jotka voivat saada useampia kuin kahta eri arvoa [ZPW08]. Nämä kaikki on kohdennettu kuitenkin nimenomaan alleeliassosiaation mittaamiseen, eivätkä tyypillisesti joko yleisty genotyypeille tai toimi muuten hyvin genotyyppiassosiaation mittaamisessa. Tästä syystä emme käsittele niitä enempää. Seuraavissa luvuissa tutustumme yleisempiin ristiintauluja koskeviin riippumattomuustesteihin, joita voidaan helposti käyttää myös genotyyppien välisen assosiaation havaitsemiseen.

2.2 Tavalliset asymptoottiset riippumattomuustestit

Edellä tutustuimme nimenomaan alleeliassosiaation mittaamista varten suunniteltuihin tunnuslukuihin. Seuraavaksi käymme lävitse muutaman yleisemmän kahden muuttujan välisen assosiaation mittaamiseen tarkoitetun assosiaatiotestin, ja perehdymme näiden hyviin ja huonoihin puoliin sovelluskohteemme kannalta. Esiteltävät testit ovat paljolti toisiaan muistuttavat khii-toiseen-testi ja G-testi. Kummatkin testit toimivat sekä vapaiden, että kiinnitettyjen reunojen mallissa, joten emme kiinnitä tarkasteltavaa mallia. Tiedot perustuvat Zarin [Zar99] ja Agrestin [Agr02] kirjojen osioihin aiheesta.

Khii-toiseen-testi

Ehkä yleisin ristiintaulujen yhteydessä käytetty riippumattomuutta mittaava tilastollinen testi on *Pearsonin khii-toiseen-testi*, johon viittaamme jatkossa myös lyhyemmin *khii-toiseen-testinä* tai χ^2*-testinä*. Olkoot X ja Y muuttujia, jotka voivat saada vastaavasti r ja c eri arvoa. Merkitään muuttujien saamia arvoja $1, 2, \ldots, r$ ja $1, 2, \ldots, c$. Tarkastellaan n:n näytteen otosta muuttujaparin arvoista ja merkitään näytteen $k \in \{1, 2, \ldots, n\}$ havaittuja arvoja x_k ja y_k. Otoksen perusteella saadaan muodostettua ristiintaulu, joka sisältää r riviä ja c saraketta. Merkitään taulukossa rivillä i ja sarakkeessa j olevaa arvoa n_{ij}. Siis n_{ij} on niiden näytteiden $k \in \{1, \ldots, n\}$ lukumäärä, joille pätee $x_k = i$ ja $y_k = j$. Käytetään n_{ij}:tä vastaavasta satunnaismuuttujasta isolla kirjaimella varustettua merkintää N_{ij}.

Määritellään $\hat{e}_{ij} = n_{i*}n_{*j}/n$. Kiinnitettyjen reunojen mallissa N_{ij} noudattaa nollahypoteesin vallitessa hypergeometrista jakaumaa, joten $\hat{e}_{ij}$ on N_{ij}:n odotusarvo. Vapaiden reunojen mallissa odotusarvo olisi $np_{i*}p_{*j}$, missä p_{i*} ja p_{*j} ovat tuntemattomia multinomijakauman parametreja. Toisaalta tiedetään, että $\hat{p}_{i*} = n_{i*}/n$ ja $\hat{p}_{*j} = n_{*j}/n$ ovat suurimman uskottavuuden estimaatteja näille reunatodennäköisyyksille. Siten $\hat{e}_{ij} = n\hat{p}_{i*}\hat{p}_{*j}$ on tässäkin tapauksessa hyvä arvio N_{ij}:n odotusarvosta.

Nyt Pearsonin khii-toiseen-testin tunnusluku määritellään seuraavasti:

$$T_{\chi^2} = \sum_i \sum_j \frac{(N_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}.$$

Koska arvot $\hat{e}_{ij}$ ovat positiivisia, on saatu tunnusluku aina ei-negatiivinen. Jos havaitut frekvenssit vastaavat suurin piirtein odotusarvojaan eli $N_{ij} \approx \hat{e}_{ij}$ kaikilla ij, niin tunnusluku saa arvon läheltä nollaa. Toisaalta mitä enemmän havainnot N_{ij}

poikkeavat nollahypoteesin mukaisista odotusarvoistaan, sitä suurempi arvosta T_{χ^2} tulee.

Koska reunafrekvenssit säilyttäviä ristiintauluja on rajallinen määrä, on saadun tunnusluvun T_{χ^2} jakauma diskreetti. Fisher [Fis22] osoitti, että riippumattomuuden vallitessa se lähestyy asymptoottisesti khii-toiseen-jakaumaa $(r-1)(c-1)$ vapausasteella kun näytteiden lukumäärä n kasvaa. Siis, jos n on suuri, niin

$$T_{\chi^2} \overset{approx}{\sim} \chi^2((r-1)(c-1)).$$

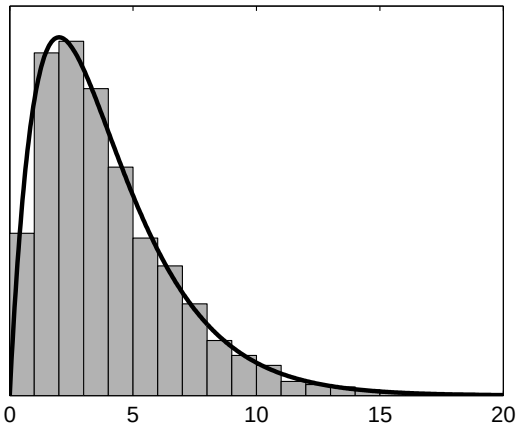
Tämä pätee siis sekä vapaiden että kiinnitettyjen reunojen mallissa. Approksimointi toimii varsin hyvin, kun kaikkien frekvenssien arvot $\hat{e}_{ij}$ ovat tarpeeksi suuria. Kuvassa 2a on verrattu teoreettista jakaumaa havaittuun todelliseen jakaumaan 3×3 -ristiintaululla, jossa on käytetty taulukon 2a reunafrekvenssejä. Valitut reunafrekvenssit ovat varsin suuria. Perinteisesti riittävänä alarajana frekvenssien odotusarvoille on pidetty lukua viisi, vaikkakin myös löyhempiä rajoja on esitetty. Kuitenkin kun odotusarvot painuvat riittävän alas, huonontuu approksimaatio merkittävästi. Erityisesti muuttujien saamien arvojen lukumäärän lisääminen pienentää reunafrekvenssejä ja edelleen odotusarvoja, kun näytteiden määrä n pidetään vakiona. Näin käy, jos muuttujiksi valitaan yksittäisten markkerien sijaan muutaman markkerin ryhmiä. Kuvassa 2b on tehty vastaava vertailu taulukon 2b reunafrekvenssien mukaisille todelliselle ja teoreettiselle jakaumalle. Approksimaatio vaikuttaa selvästi liian huonolta. Tämä havaitaan myös vertaamalla jakaumista saatavia p-arvoja: esimerkiksi tapaukselle $t_{\chi^2} = 450$ teoreettinen jakauma antaa p-arvoksi $1,2 \cdot 10^{-5}$. kun se todellisuudessa on noin $4,3 \cdot 10^{-2}$. Erityisesti koska todellisen jakauman häntä on selvästi paksumpi kuin teoreettisella jakaumalla, ovat suurilla tunnusluvun arvoilla teoreettisesta jakaumasta saadut p-arvot merkittävästi liian pieniä.

G-testi

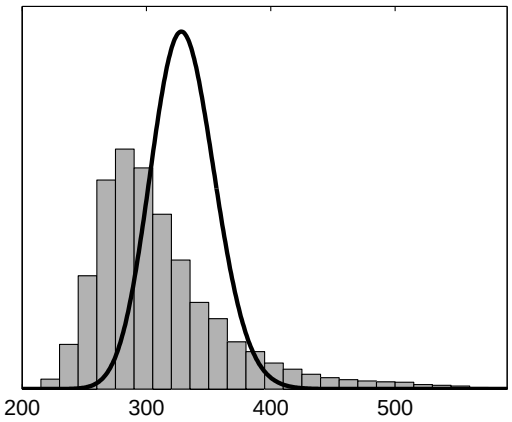
Khii-toiseen-testiä vastaava toinen yleinen merkitsevyyden mittaukseen käytetty testi on *G-testi* (usein myös *uskottavuusosamäärätesti*). Sen tunnusluku määritellään seuraavasti:

$$T_G = 2 \sum_i \sum_j N_{ij} \ln \frac{N_{ij}}{\hat{e}_{ij}}.$$

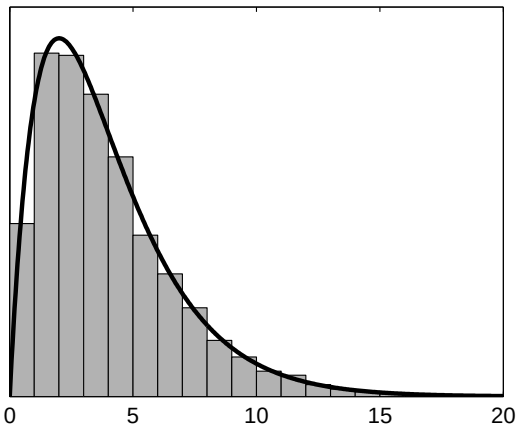
G-testin tunnuslukua on luonnollista ajatella informaatioteoreettisesta näkökulmasta. Merkitsemällä $n_{ij} = \hat{p}_{ij} n$ sekä $\hat{e}_{ij} = \hat{p}_{i*}\hat{p}_{*j} n$ ja sijoittamalla määritelmään saa-



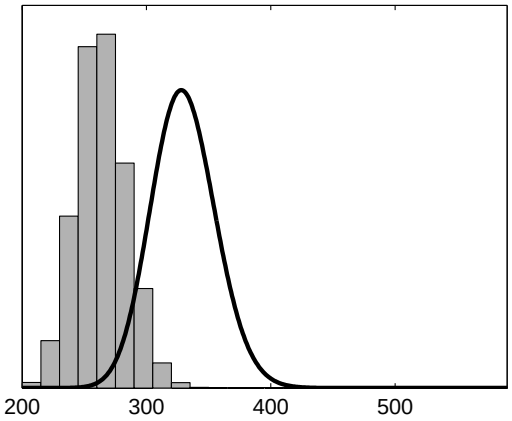
(a) T_{χ^2} taulukon 2a reunafrekvensseillä.



(b) T_{χ^2} taulukon 2b reunafrekvensseillä.



(c) T_G taulukon 2a reunafrekvensseillä.



(d) T_G taulukon 2b reunafrekvensseillä.

Kuva 2: Khii-toiseen-testin tunnusluvun T_{χ^2} sekä G-testin tunnusluvun T_G todelliset jakauma (harmaa histogrammi) sekä teoreettinen jakauma (musta käyrä) kahdella eri muuttujaparilla. Muuttujien arvojen frekvensseiksi on valittu taulukkojen 2a ja 2b reunafrekvenssit.

daan

$$\begin{aligned} t_G &= 2n \sum_{ij} \hat{p}_{ij} \ln \frac{\hat{p}_{ij}}{\hat{p}_{i*}\hat{p}_{*j}} \\ &= 2n \left(\sum_{ij} \hat{p}_{ij} \ln \hat{p}_{ij} - \sum_{i} \hat{p}_{i*} \ln \hat{p}_{i*} - \sum_{j} \hat{p}_{*j} \ln \hat{p}_{*j} \right) \\ &= 2n(H(x) + H(y) - H(x,y)), \end{aligned}$$

missä $H(x)$ on muuttujan X havaitun jakauman entropia ja vastaavasti $H(x,y)$ muuttujien X ja Y yhteisjakauman entropia. Tiedetään, että muuttujan arvojen koodaamiseen vaadittava minimitallennustila saadaan kertomalla muuttujan jakauman entropia tallennettavien arvojen määrällä. Näin ollen T_G kertoo oleellisesti paljonko tallennustilaa säästetään, kun muuttujien X ja Y arvojen koodaamisessa voidaan käyttää hyödyksi tietoa näiden välisistä assosiaatioista.

Kuten khii-toiseen-testin tunnusluku, myös G-testin tunnusluku noudattaa sekä vapaiden että kiinnitettyjen reunojen mallissa likimäärin samaa khii-toiseen-jakaumaa [Wil35, Agr02]:

$$T_G \overset{approx}{\sim} \chi^2((r-1)(c-1)).$$

Saadulla tunnusluvulla T_G on kuitenkin sama ongelma kuin khii-toiseen-testin tunnusluvulla: Approksimaatio ei toimi, jos frekvenssien odotusarvot ovat liian pieniä. Kuvissa 2c ja 2d on laskettu T_G:lle jakaumat vastaavilla reunafrekvensseillä kuin kuvissa 2a ja 2b tunnusluvulle T_{χ^2}. Jälkimmäisessä tapauksessa esimerkiksi havainnon $t_G = 350$ todellinen p-arvo olisi $5.0 \cdot 10^{-5}$, mutta approksimointi χ^2-jakaumalla antaa arvoksi 0.22. Toisin kuin khii-toiseen-tunnusluvun tapauksessa, tällä kertaa teoreettisesta jakaumasta saadut p-arvot ovat selvästi liian suuria.

2.3 Satunnaistestaus

Kuten edellä huomasimme, sopivan tunnusluvun löytäminen saattaa olla hankalaa etenkin, jos käsiteltävät ristiintaulut ovat harvoja. Ongelmana on tunnusluvun jakauman tarkka selvittäminen riittävällä tarkkuudella. Jakauman analyyttisen johtamisen sijaan toinen vaihtoehto on pyrkiä approksimoimaan tunnusluvun jakaumaa näytteenotolla. Tällöin assosiaatiota mittaava tunnusluku voidaan valita merkittävästi vapaammin. Huonona puolena on näytteenoton raskaus; tarkkaa approksimaatiota varten täytyy tyypillisesti arpoa suuri määrä nollahypoteesin mukaisia näytteitä. Seuraavassa tutustumme muutamaan erilaiseen näytteenottoon perustuvaan p-arvon approksimointitapaan.

P-arvon approksimointi

Olkoot X ja Y jälleen muuttujat, joiden välistä assosiaatiota pyrimme mittaamaan. Tarkastellaan mielivaltaista tunnuslukua. Olkoon t havaittu tunnusluvun arvo. Nyt jos käytössä on menetelmä riippumattomien näytteiden arpomiseen nollahypoteesin mukaisesta tunnusluvun jakaumasta, voidaan p-arvoa arvioida seuraavasti [AWB79]:

1. Arvo tunnusluvusta satunnaisesti m nollahypoteesin mukaista riippumatonta otosta $t_1, t_2, \ldots, t_m$.
2. Arvioi p-arvoa: $\hat{\mathrm{p}} = h/m$, missä $h = \#\{t_i \geq t \mid i = 1, 2, \ldots, m\}$ on havaittua tunnuslukua t suurempien näytteiden t_i lukumäärä.

Lemma 1. *Saatu arvio* $\hat{\mathrm{p}} = h/m$ *on suurimman uskottavuuden estimaatti p-arvolle.*

Todistus. P-arvon määritelmän mukaan $\Pr(T_i \geq t) = \mathrm{p}$. Näin ollen h on noudattaa binomijakaumaa parametrein p ja m. Siispä p-arvon uskottavuus on

$$L(\mathrm{p}) = \binom{m}{h} \mathrm{p}^h (1 - \mathrm{p})^{m-h}.$$

Voimme ottaa logaritmin puolittain, sillä aidosti kasvavana funktiona se säilyttää maksimin sijainnin. Saadaan

$$\ln L(\mathrm{p}) = \ln \binom{m}{h} + h \ln \mathrm{p} + (m - h) \ln(1 - \mathrm{p}).$$

Saatu log-uskottavuusfunktio on konkaavien funktioiden summana konkaavi. Siispä se saa suurimman arvonsa derivaatan nollakohdassa, eli kun

$$\frac{h}{\mathrm{p}} - \frac{m - h}{1 - \mathrm{p}} = 0$$
$$\mathrm{p} = \frac{h}{m}.$$

□

Arvion $\hat{\mathrm{p}} = h/m$ sijaan p-arvolle saatetaan joissain tapauksissa käyttää myös approksimaatiota $\hat{\mathrm{p}} = (h+1)/(m+1)$ [BC91, DH97]. Joka tapauksessa kun näytteiden lukumäärä on m, voi estimaatti $\hat{\mathrm{p}}$ saada $m+1$ eri arvoa. Nämä ovat jakautuneet tasaisin välein yksikkövälille $[0, 1]$. Jos haluamme arvioida p-arvoa vaikkapa viiden desimaalin tarkkuudella, tarvitaan näytteitä siis ainakin 10^5. Riippumatta näytteiden tuottamismenetelmästä on näin suuren määrän läpikäyminen jo itsessään työlästä. Koko genomin laajuisessa assosiaatioiden etsimisessä tarvittava tarkkuus on vielä selvästi tätä suurempi, luokkaa 10–15 desimaalia. Tämän vaatiman näytemäärän tuottaminen jokaiselle testattavalle parille ei selvästikään ole mahdollista.

Adaptiivinen näytteenotto

Tietyissä tapauksissa voimme helposti muuttaa näytteenoton adaptiiviseksi. Olkoon α valittu kynnysarvo nollahypoteesin hylkäämiselle. Hylkäys siis tapahtuu, jos korkeintaan $a = \lfloor \alpha m \rfloor$ näytteen tunnusluku on suurempi kuin t. Mikäli tiedämme kynnysarvon jo valmiiksi, voidaan näytteenotto keskeyttää heti, jos $h_k > a$, missä h_k on t:tä suurempien näytteiden määrä k:n näytteen jälkeen. Olkoon L satunnaismuuttuja, joka kertoo tarvittavien näytteiden lukumäärän. Nollahypoteesin vallitessa näytteiden määrälle saadaan seuraava odotusarvo [BC91]:

Lemma 2. *Mikäli näytteenotto keskeytetään heti, jos $h_k > a$, tarvitaan nollahypoteesin vallitessa odotusarvoisesti noin $\alpha m(1 - \ln \alpha)$ näytettä.*

Todistus. Nollahypoteesin vallitessa havaittu tunnusluku T noudattaa näytteiden $T_1, T_2, \ldots$ kanssa samaa jakaumaa. Olkoon $l \in \{1, 2, \ldots, m\}$. Tapahtuma $L \geq l$ edellyttää, että T on a:n suurimman joukossa arvoista $T, T_1, T_2, \ldots, T_{l-1}$. Symmetrisyydestä seuraa siten, että arvoille $l = 1, 2, \ldots, m$ on

$$\Pr(L \geq l) = \begin{cases} 1 & \text{jos } l \leq a \\ a/l & \text{jos } l > a. \end{cases}$$

Nyt näytteiden määrän odotusarvoksi saadaan

$$\mathrm{E}[L] = \sum_{l=1}^{m} \Pr(L \geq l) = a + a \sum_{l=a+1}^{m} \frac{1}{l} \approx a\,(1 + \ln m - \ln a)\,,$$

missä on viimeisessä vaiheessa käytetty tietoa $\sum_{i=1}^{n} 1/i \approx \ln n$. Sijoittamalla tähän a:n paikalle αm saadaan haluttu tulos. □

Jos esimerkiksi $m = 10^5$ ja kynnysarvo $\alpha = 10^{-4}$, tarvitaan näytteitä keskimäärin noin 100. Parannus on huomattava, mutta esimerkiksi myöhemmin osiossa 3.2 esiteltävässä FDR-valintametodissa tarkka kynnysarvo riippuu havaituista p-arvoista, eikä sitä voida siten tietää etukäteen. Luonnollisesti selkeä nopeutus saataisiin jo asettamalla kynnysarvolle ennalta karkea yläraja, jonka ylittävät p-arvot voidaan yllä mainitulla tavalla karsia pois laskuista jo ennen lopullista merkitsevien p-arvojen valintaa.

P-arvon adaptiivinen approksimointi

Kuten totesimme, tietoa kynnysarvosta ei välttämättä ole etukäteen saatavilla eikä edellä esitelty adaptiivinen näytteenotto sellaisenaan onnistu. Koska pienet p-arvot

ovat kuitenkin suuria kiinnostavampia, haluaisimme arvioida niitä tarkemmin, kun taas suurille p-arvoille riittäisi karkea approksimaatio. Besagin ja Cliffordin [BC91] kuvailemassa menetelmässä pyritään rajoittamaan estimaatin $\hat{\mathrm{p}}$ keskihajontaa alle kynnysarvon $c\mathrm{p}$, missä $c > 0$ on ennalta valittu vakio. Näytteitä tuotetaan yksi kerrallaan, ja aina uuden näytteen jälkeen arvioidaan p-arvoa $\hat{\mathrm{p}}_k = h_k/k$, missä k on näytteiden lukumäärä siihen mennessä. Tarkoituksena on jatkaa näytteenottoa, kunnes lopetusehto $\mathrm{Var}[\hat{\mathrm{p}}_k] \leq (c\mathrm{p})^2$ täyttyy.

Koska h_k noudattaa binomijakaumaa, saadaan $\mathrm{Var}[\hat{\mathrm{p}}_k] = \mathrm{Var}[h_k]/k^2 = \mathrm{p}(1-\mathrm{p})/k$. Nyt sijoittamalla arvio $\hat{\mathrm{p}}_k$ aidon p-arvon paikalle saadaan lopetusehdoksi

$$\frac{h_k}{k^2}\left(1 - \frac{h_k}{k}\right) \leq \left(c\frac{h_k}{k}\right)^2,$$

ja tästä edelleen

$$h_k \geq \frac{1}{k^{-1} + c^2}.$$

Lopetukseen vaadittavien t:tä suurempien näytteiden lukumäärä siis lähestyy alhaalta päin rajaa c^{-2} sitä mukaa, kun näytteiden lukumäärää lisätään. Jos p-arvo on erityisen pieni, kasvaa tähän vaadittavien otosten määrä turhan suureksi. Tästä syystä on järkevää jälleen asettaa jokin yläraja m näytteiden määrälle. Näin saatu menetelmä muistuttaa suuresti edellä esiteltyä adaptiivista näytteenottoa arvolla $\alpha = c^{-2}$, tosin alussa lopettamiskriteeri on löysempi.

Vaikka esitetyt kaksi adaptiivista menetelmää voivat olla merkittävästi nopeampia kuin naiivi toteutus, tämäkään ei välttämättä riitä. Jos vaadittu tarkkuus on suuri, voi jo yhden nollaa riittävän lähellä olevan p-arvon laskeminen olla liian työlästä.

P-arvon jakauman arviointi

Eräs vaihtoehto vähentää tarvittavien otosten määrää on pyrkiä aluksi arvioimaan p-arvoja jollakin yksinkertaisella (esimerkiksi heuristisella) tavalla ja tarkentaa arvioita vain tarvittaessa. Kustra ja kumppanit [KSM+08] ovat kehittäneet tämän tyyppisen menetelmän. Se on tarkoitettu nimenomaisesti tilanteisiin, joissa p-arvoja täytyy laskea suuri määrä. Menetelmä palauttaa kutakin p-arvoa kohden todennäköisyysjakauman, joka toivon mukaan on keskittynyt tiukasti todellisen p-arvon ympärille.

Olkoon M assosiaatiotestien lukumäärä. Menetelmä jakautuu kolmeen vaiheeseen: Ensimmäisessä vaiheessa valitaan pieni osa testeistä opetusdataksi ja rakennetaan näiden avulla ennustaja, jonka avulla puolestaan toisessa vaiheessa muodostetaan

kunkin testin $i = 1, \ldots, M$ p-arvolle priorijakauma. Lopuksi kolmannessa vaiheessa saatuja jakaumia päivitetään näytteenotolla, kunnes saavutetaan haluttu tarkkuustaso. Seuraavaksi käymme hieman yksityiskohtaisemmin kunkin vaiheen läpi.

Ensimmäisessä vaiheessa pyritään oppimaan säännöt, joilla ristiintaulun reunafrekvenssien perusteella saadaan ennustettua tunnusluvun T_i nollajakauman 95. prosenttipiste, siis kohta x, jossa $\Pr(T_i < x) = 0.95$. Algoritmi on seuraava:

Algoritmi 1 Opi ennustaja tunnusluvuilta 95. prosenttipisteille

1: Valitse opetusdata $L \subset \{1, \ldots, M\}$.
2: **for** $i \in L$ **do**
3: Laske nollajakauma tunnusluvulle T_i käyttämällä näytteenottoa.
4: Laske jakauman 95. prosenttipiste.
5: Opi ennustaja testien suuruusjärjestykseen järjestetyiltä reunafrekvensseiltä lasketuille prosenttipisteille.

Opetusdataksi L valitaan satunnaisesti pieni osajoukko kaikista testeistä. Kustra ja kumppanit käyttivät esimerkiksi artikkelinsa kokeissa opetusdatana 3000 satunnaisesti valittua testiä, kun yhteensä testejä oli noin 400 000. Ennustajan oppimiseen Kustra ja kumppanit käyttivät satunnaismetsämenetelmää. Tehtävään voisi luonnollisesti soveltaa muitakin koneoppimisalgoritmeja.

Menetelmän toisessa vaiheessa muodostetaan ennustajien avulla arvioidut todennäköisyysjakaumat kunkin tunnusluvun T_i todelliselle p-arvolle p_i. Tämä tehdään seuraavasti:

Algoritmi 2 Laske todennäköisyysjakaumat p-arvoille

1: **for** $i \leftarrow 1$ **to** M **do**
2: Ennusta opitulla ennustajalla tunnusluvun T_i jakauman 95. prosenttipiste reunafrekvenssien perusteella.
3: Valitse opetusdatasta h nollajakaumaa, joiden 95. prosenttipisteet ovat lähimpänä. Laske näiden avulla h arviota t_i:n p-arvolle.
4: Laske arvioiden geometrinen keskiarvo $\hat{\mathrm{p}}_i$.
5: Muodosta $\hat{\mathrm{p}}_i$:n perusteella priorijakauma p-arvolle p_i.

Keskiarvo $\hat{\mathrm{p}}_i$ on ennustajien avulla laskettu arvio tunnusluvun todellisesta p-arvosta. Priorijakauma muodostetaan siten, että se antaa suuren todennäköisyyden $\hat{\mathrm{p}}_i$:n lähistöllä sijaitseville p-arvoille. Priorijakaumana käytetään Kustran ja kumppaneiden

artikkelissa kahden $\hat{\mathrm{p}}_i$:n ympärille keskittyneen sopivasti valitun beeta-jakauman sekoitetta. Vaiheen 3 vakiona h Kustra ja kumppanit käyttivät arvoa 10.

Kun kullekin p-arvolle on muodostettu jakauma, halutaan kolmannessa vaiheessa tarkentaa näitä halutulle tasolle. Tätä varten asetetaan raja q, jota pienemmistä p-arvoista olemme erityisen kiinnostuneita. Kuten aiemmin tämä voisi siis olla valittu merkitsevyysraja, jos se on tiedossa etukäteen. Lisäksi asetetaan yläraja m testiä kohden otettavien näytteiden määrälle. Käytetään p-arvoja p_i vastaavista satunnaismuuttujista merkintää P_i. Nyt $\Pr(\mathrm{P}_i < q)$ eli todennäköisyys, että P_i alittaa kynnysrajan, voidaan laskea priorijakaumasta. Päämääränä on laskea maksimitarkkuudella p-arvot, joilla tämä todennäköisyys on suuri, ja tarkentaa loppuja p-arvoja vain välttämätön määrä. Rajan q alittavien, alle m:llä näytteellä tarkennettujen p-arvojen lukumäärän odotusarvo on

$$\sum_{i:k_i<m} \Pr(\mathrm{P}_i < q),$$

missä k_i on tunnusluvusta T_i otettujen näytteiden lukumäärä. Algoritmi pyrkii pienentämään ahneesti tätä odotusarvoa arpomalla näytteitä sopivasti valituista tunnusluvuista ja päivittämällä p-arvojen jakaumia näiden avulla. Jos priorijakaumana käytetään beeta-sekoitejakaumaa, kuuluu päivittämisen tuloksena saatava posteriorijakauma samaan perheeseen. Kun odotusarvo on laskenut ennalta määritellyn rajan alle, lopettaa algoritmi päivittämisen.

Kustran ja kumppaneiden kuvaaman menetelmän lopputuloksena saadaan kullekin p-arvolle todennäköisyysjakauma. Näitä jakaumia voidaan edelleen käyttää suoraan merkisevien assosiaatioiden valintaan. Vaihtoehtoisesti voitaisiin arvioida p-arvoja jakaumien perusteella lasketuilla odotusarvoilla. Heuristisen priorijakauman ansiosta suuret p-arvot voidaan tiputtaa entistä nopeammin pois laskuista. Pienille p-arvoille menetelmä tarvitsee silti edelleen suuren määrän näytteitä, joten tässä suhteessa se ei eroa aiemmin esitellyistä adaptiivisista menetelmistä.

Satunnaisten otosten tuottaminen

Kaikissa tässä osiossa kuvatuissa menetelmissä on ollut tärkeässä osassa riippumattomien nollahypoteesin mukaisten otosten arpominen tunnusluvusta. Yleinen tapa otoksen tuottamiseen on sekoittaa toisen testattavan muuttujan havaitut arvot satunnaiseen järjestykseen. Permutoinnin seurauksena muuttujien väliset mahdolliset assosiaatiot häviävät. Sekoitetuille muuttujille voidaan tämän jälkeen muodostaa

ristiintaulut ja laskea tunnusluvun arvo normaaliin tapaan. Permutointi säilyttää ristiintaulun reunafrekvenssit, ja se noudattaa siten kiinnitettyjen reunojen mallia. Se onkin näytteenotosta ilman takaisinpanoa erikoistapaus, jossa otettava näyte on koko aineiston kokoinen.

Vapaiden reunojen mallin mukaisten näytteiden arpominen ei onnistu samaan tapaan, sillä tarvittavat parametrit ovat tuntemattomia. Käyttämällä n:n näytteen otantaa takaisinpanolla kummallekin muuttujalle saataisiin ristiintaulun arvoille kyllä tulomultinomijakauma, jonka parametrit ovat suurimman uskottavuuden estimaatteja todellisille parametreille.

Sekä permutoinnissa että otannassa takaisinpanolla otoksen muodostamiseen kuluva aika on lineaarinen näytteiden määrään n sekä ristiintaulun kokoon nähden – siis $O(n + rc)$, missä r ja c ovat muuttujien X ja Y erilaisten arvojen määrä. Jos n on suuri, olisi houkuttelevaa keksiä arpomiseen nopeampi menetelmä. Jos tunnusluvun jakauma olisi tiedossa, voitaisiin toisaalta p-arvot laskea suoraan, eikä näytteenottoa tarvittaisi lainkaan.

Välimaastosta oleva keino olisi arpoa ristiintauluja. Kahden yksittäisen markkerin välistä assosiaatiota mitattaessa sisältää ristiintaulu korkeintaan yhdeksän arvoa, joten parannus saattaisi olla useita suuruusluokkia. Toisaalta jos osapuolina ovat useamman markkerin ryhmät, kasvavat ristiintaulut helposti niin isoiksi, että etu voitaisiin menettää. Vapaiden reunojen mallissa ristiintaulujen arpominen olisi suhteellisen helppoa – ainakin, jos tulomultinomijakauman parametrit tunnettaisiin. Jos reunafrekvenssit on kiinnitetty, ei ristiintaulujen arpominen halutusta jakaumasta kuitenkaan ole helppoa. Itse asiassa annetuilla reunafrekvensseillä pelkästään erilaisten mahdollisten taulujen lukumäärän selvittäminen on erittäin vaikeaa. Joitakin Markovin ketjuihin perustuvia menetelmiä taulujen arpomiseen on kehitetty. Esimerkiksi Raymondin ja Roussetin [RR95] algoritmilla uusi ristiintaulunäyte voidaan arpoa vakioajassa. Useissa tapauksissa myös tunnusluvun laskeminen arvotulle näytteelle voidaan sisällyttää tähän vakioaikaan, jolloin saavutettu nopeusetu permutointiin nähden on teoriassa varsin merkittävä. Lyhyiden testien perusteella algoritmin käyttämä Markovin ketju sekoittuu kuitenkin niin hitaasti, että otoksia tarvitaan moninkertainen määrä permutoinnilla saataviin riippumaattomiin otoksiin verrattuna. Tämän seurauksena tietyn tarkkuuden saavuttamiseen kuluu käytännössä suunnilleen saman verran laskenta-aikaa kuin permutoinnillakin.

Johtopäätöksenä satunnaistestaukseen liittyen voimme todeta, etteivät esitetyt satunnaistestausmenetelmät vaikuta olevan riittävän nopeita assosiaatioiden etsimi-

seen koko kromosomistosta. Tästä syystä emme käytä niitä myöhemmissä kokeissa. Seuraavaksi tutustumme muutamaan tapaan laskea p-arvoja tarkasti sopivasti valituille tunnusluvuille.

2.4 Tarkka maksimipoikkeamatesti

Ehkä helpoiten mieleen juolahtava tapa tutkia ristiintaulun poikkeavuutta olisi mitata arvojen n_{ij} poikkeamia odotusarvoistaan ja valita suurin näistä poikkeamista. Tämän innoittamana määrittelemme maksimipoikkeama-tunnusluvun

$$T_{\dot{\epsilon}} = \max_{ij} |N_{ij} - e_{ij}|,$$

missä $e_{ij} = np_{i*}p_{*j}$ on N_{ij}:n odotusarvo. Merkinnässä $T_{\dot{\epsilon}}$, symboli ϵ viittaa havaitun arvon virheeseen odotusarvoon nähden ja $\dot{\epsilon}$ sen maksimiin. Khii-toiseen-testistä poiketen minkäänlaista normalisointia ei tehdä ja summaamisen tilalla on maksimin ottaminen. Voidaan ajatella, että normalisoinnin puutteesta johtuen $T_{\dot{\epsilon}}$ painottaa enemmän absoluuttisesti suuria poikkeamia. Tavoitteista riippuen tämä voidaan nähdä hyvänä tai huonona asiana.

Kuten olemme aiemmin todenneet, odotusarvot e_{ij} valitettavasti tunnetaan yleisesti vain kiinnitettyjen reunojen mallissa. Oletetaan kuitenkin, että ne olisivat tiedossa myös vapaiden reunojen tapauksessa. Oletetaan lisäksi, että reunafrekvenssejä ei ole lukittu. Tällöin khii-toiseen-testin ja G-testin tunnuslukuihin verrattuna maksimipoikkeamalla on eräs selkeä etu: maksimipoikkeaman tarkka p-arvo voidaan laskea dynaamisella ohjelmoinnilla käyttäen tapaa, jonka kuvaamme seuraavaksi.

Tarkastellaan havaittuja ristiintaulun ruutujen arvoja. Merkintöjen helpottamiseksi järjestetään ruutujen indeksit $11, 12, \ldots, 1c, 21, 22, \ldots, rc$ mielivaltaiseen järjestykseen, ja nimetään ne yksittäisillä luvuilla $1, 2, \ldots, m$, missä $m = rc$ on ristiintaulun ruutujen määrä. Havaittuihin ruutujen arvoihin viitataan nyt siis muuttujilla $n_1, n_2, \ldots, n_m$. Merkitään vastaavia satunnaismuuttujia $N_1, N_2, \ldots, N_m$. Havainnon $t_{\dot{\epsilon}}$ p-arvo on nyt

$$\begin{aligned} \mathrm{p} &= \Pr(\max_k |N_k - e_k| \geq t_{\dot{\epsilon}}) \\ &= 1 - \Pr\left(|N_k - e_k| < t_{\dot{\epsilon}} \ \forall\ k\right). \end{aligned}$$

Kuten olemme aiemmin reunafrekvenssivapaasta mallista todenneet, ruutujen arvot noudattavat siinä multinomijakaumaa parametrein $p_1, p_2, \ldots, p_m$ ja n, missä p_i on ruutua i vastaava todennäköisyys. Jos arvot $N_1, N_2, \ldots, N_m$ olisivat keskenään

riippumattomia, voitaisiin todennäköisyys $\Pr\left(|N_k - e_k| < t_{\dot{\epsilon}} \ \forall \ k\right)$ hajottaa helposti yksittäisiä muuttujia koskevien todennäköisyyksien tuloiksi. Multinomijakauman tapauksessa näin ei kuitenkaan ole, joten joudumme ottamaan riippuvuudet jotenkin huomioon.

Riippuvuuksien hallitsemiseksi esitämme multinomijakauman toisessa muodossa samaan tapaan kuin Levin [Lev81] johtaessaan tapaa laskea multinomijakauman kertymäfunktio. Tiedetään yleisesti, että multinomijakauma voidaan esittää riippumattomien Poisson-jakautuneiden satunnaismuuttujien avulla

$$\Pr(N_1 = n_1, \ldots, N_m = n_m) = \Pr\left(X_1 = n_1, \ldots, X_m = n_m \;\middle|\; \sum_k X_k = n\right),$$

missä $X_k \sim Poisson(e_k)$ ja $e_k = np_k$. Siis kun Poisson-jakautuneiden satunnaismuuttujien summa kiinnitetään arvoon n, saadaan tuloksena haluttu multinomijakauma. P-arvolle saadaan siten lauseke

$$\mathrm{p} = 1 - \Pr\left(|X_k - e_k| < t_{\dot{\epsilon}} \ \forall \ k \;\middle|\; \sum_k X_k = n\right).$$

Merkitsemällä $E_k = |X_k - e_k|$ ja käyttämällä ehdollisen todennäköisyyden määritelmää saadaan tämä edelleen muotoon

$$\mathrm{p} = 1 - \frac{\Pr\left((E_k < t_{\dot{\epsilon}} \ \forall \ k) \cap (\sum_k X_k = n)\right)}{\Pr\left(\sum_k X_k = n\right)}.$$

Saadussa lausekkeessa on kaksi todennäköisyyttä, jotka haluamme ratkaista. Tarkastellaan ensin nimittäjää. Riippumattomien Poisson-satunnaismuuttujien summa on tunnetusti Poisson-jakautunut; siis $\sum_k X_k \sim Poisson(\sum_k e_k) \sim Poisson(n)$. Siispä nimittäjä saa arvon $n^n/n! \cdot \exp(-n)$. Saadaan siis

$$\mathrm{p} = 1 - \frac{n!}{n^n} \exp(n) \cdot \Pr\left((E_k < t_{\dot{\epsilon}} \ \forall \ k) \cap \left(\sum_k X_k = n\right)\right).$$

Jäljellä olevan todennäköisyyden laskeminen osoittautuu hieman monimutkaisemmaksi. Olkoot $l \in \{0, \ldots, m\}$ ja $h \in \{0, \ldots, n\}$. Merkitään

$$P(l, h) = \exp\left(\sum_{k=1}^{l} e_k\right) \cdot \Pr\left(\left(\bigcap_{k=1}^{l} (E_k < t_{\dot{\epsilon}})\right) \cap \left(\sum_{k=1}^{l} X_k = h\right)\right).$$

Nyt haluttu p-arvo on siis $p = 1 - P(m, n) \cdot n!/n^n$. Muodostamme seuraavaksi dynaamista ohjelmointia hyödyntävän algoritmin arvojen $P(l, h)$ laskemiseen.

Tarkastellaan ensin tapausta $l = 0$. Tällöin yllä olevassa lausekkeessa konjunktion vasen puoli muodostaa tyhjän konjunktion ja siten aina tosi. Oikealle puolella tyhjä summa saa arvon nolla, joten tapahtumaksi tulee $0 = h$. Näin ollen $P(0,0) = \exp(0) \cdot \Pr(0 = 0) = 1$ ja $P(0,h) = \exp(0) \cdot \Pr(0 = h) = 0$, kun $h > 0$.

Olkoon nyt $l \in [1,n]$. Osoitamme seuraavaksi, että arvot $P(l,0), \ldots, P(l,n)$ saadaan laskettua arvojen $P(l-1,0), \ldots, P(l-1,n)$ avulla. Huomataan, että konjunktion oikealla puolella oleva tapahtuma $\sum_{k=1}^{l} X_k = h$ voidaan osittaa erillisiin tapauksiin $i = 0, \ldots, h$, joissa kussakin aina $X_l = i$ ja $\sum_{k=1}^{l-1} X_k = h - i$. Todennäköisyys voidaan siten hajottaa näiden erillisten tapausten todennäköisyyksien summaksi. Lisäksi konjunktion vasemman puolen konjunktioketjusta voidaan erottaa tapahtuma $E_k < t_\epsilon$ erikseen. Saadaan siis

$$\begin{aligned}
P(l,h) &= \exp\left(\sum_{k=1}^{l} e_k\right) \cdot \sum_{i=0}^{h} \Pr\left(\left(\bigcap_{k=1}^{l-1}(E_k < t_\epsilon)\right) \cap (E_l < t_\epsilon) \cap \right. \\
&\qquad\qquad \left. \left(\sum_{k=1}^{l-1} X_k = h - i\right) \cap (X_l = i)\right) \\
&= \sum_{i=0}^{h} \exp(e_l) \Pr((E_l < t_\epsilon) \cap (X_l = i)) \cdot \\
&\qquad \exp\left(\sum_{k=1}^{l-1} e_k\right) \Pr\left(\left(\bigcap_{k=1}^{l-1}(E_k < t_\epsilon)\right) \cap \left(\sum_{k=1}^{l-1} X_k = h - i\right)\right) \\
&= \sum_{i=0}^{h} \exp(e_l) \Pr((E_l < t_\epsilon) \cap (X_l = i)) \cdot P(l-1, h-i).
\end{aligned}$$

Toisessa yhtäsuuruudessa on hyödynnetty oletusta, jonka mukaan muuttuja X_l (ja siten myös E_l) on riippumaton edeltävistä muuttujista $X_1, \ldots, X_{l-1}$ (ja siten myös muuttujista $E_1, \ldots, E_{l-1}$).

Tapahtuma $E_l < t_\epsilon$ toteutuu, jos $X_l \in (e_l - t_\epsilon, e_l + t_\epsilon)$. Siis $\Pr((E_l < t_\epsilon) \cap (X_l = i))$ saa arvon $\Pr(X_l = i) = (e_l)^i \cdot \exp(-e_l)/(i!)$, jos $i \in (e_l - t_\epsilon, e_l + t_\epsilon)$, ja arvon nolla muuten. Näin ollen summaaminen voidaan rajoittaa kyseiselle välille, jolloin saadaan

$$P(l,h) = \sum_{i=a_l}^{b_l} \frac{(e_l)^i}{i!} P(l-1, h-i),$$

missä $a_l = \max(0, \lfloor e_l - t_\epsilon \rfloor + 1)$ ja $b_l = \min(h, \lceil e_l + t_\epsilon \rceil - 1)$. Arvojen $P(l,h)$ laskemiseksi muodostetaan nyt seuraava algoritmi.

Algoritmi 3 Laske taulukko $P(0 \dots m, 0 \dots n)$

1: $P(0,0) \leftarrow 1$ ja $P(0,h) \leftarrow 0$ kaikilla $h = 1, \dots, n$.
2: **for** $l \leftarrow 1$ **to** m **do**
3: **for** $h \leftarrow 0$ **to** n **do**
4: $P(l,h) \leftarrow 0$
5: **for** $i \leftarrow \max(0, \lfloor e_l - t_\epsilon \rfloor + 1)$ **to** $\min(h, \lceil e_l + t_\epsilon \rceil - 1)$ **do**
6: $P(l,h) \leftarrow P(l,h) + P(l-1, h-i) \cdot (e_l)^i/(i!)$

Rivillä 6 tarvittavat kertoimet $(e_l)^i/(i!)$ voidaan laskea etukäteen kaikille mahdollisille arvoille $l = 1, \dots, m$ ja $i = 0, \dots, n$ ajassa $O(mn)$. Siten yllä oleva taulukon P laskeva algoritmi 3 vaatii selvästi ajan $O(mn \cdot t_\epsilon)$. Koska $t_\epsilon \leq n$, on aikavaativuus erityisesti luokkaa $O(mn^2)$.

On huomattava, että mainitut aikavaativuudet pätevät vain, jos tarvittavat yksittäiset peruslaskutoimitukset voidaan suorittaa vakioajassa. Jos laskuissa käytetään vakiotarkkuuden liukulukuja, ovat tarvittavat laskut vakioaikaisia. Toisaalta, mikäli laskut halutaan suorittaa tarkasti suurilla rationaaliluvuilla, ei tämä yleisesti ole mahdollista ja aikavaativuus kasvaa jonkin verran. Esimerkiksi pelkästään lukujen $i!$ esittämiseen voidaan tarvita $\Omega(\log(n!)) = \Omega(n \log n)$ bittiä, joten aikavaativuuskin kasvaa vähintään vastaavalla kertoimella. Lisäksi murtolukuversion edellytyksenä on, että odotusarvot e_l voidaan esittää rationaalilukuina tarpeeksi pienessä tilassa.

Saimme siis muodostettua ajassa $O(mn \cdot t_\epsilon)$ toimivan algoritmin p-arvon laskemiseen havaitulle maksimipoikkeamalle t_ϵ. Menetelmä voidaan myös varsin suoraviivaisesti yleistää normalisoidulle maksimipoikkeamalle, jossa $E_k = |N_k - e_k|$ korvataan normalisoidulla versiolla $E_k = |N_k - e_k|/f_k$. Normalisointivakioiksi f_k voidaan valita vaikkapa arvojen N_k keskihajonta. Lopullisessa yleistetyssä algoritmissa ainoastaan taulukon P laskennan sisimmässä silmukassa täytyy läpikäytävien muuttujan i arvojen rajat määritellä uudelleen muotoon $\max(0, \lfloor e_l - f_l t_\epsilon \rfloor + 1)$ ja $\min(h, \lceil e_l + f_l t_\epsilon \rceil - 1)$. Muilta osin yleistetty algoritmi vastaa täysin esitettyä.

Esitelty p-arvon laskeva algoritmi sekä itse asiassa jo pelkän maksimipoikkeamatunnusluvun laskeminen vaativat, että ristiintaulun ruutujen odotusarvot e_{ij} tunnetaan. Tämä saattaa joissain tapauksissa olla mahdollista, mutta yleisesti näin ei ole. Oikeiden odotusarvojen tilalla voidaan käyttää luvussa 2.2 esiteteltyä arvioita $\hat{e}_{ij} = n_{i*} n_{*j}/n$. Jos n on suuri, on näiden suhteellinen virhe todellisesta odotusarvosta todennäköisesti pieni. Tunnusluvun jakauma on tällöin sama kuin jakauma, joka saataisiin näytteenotolla takaisinpanolla.

Jos laskettavat p-arvot ovat kovin pieniä, muodostuu ongelmaksi myös rajallinen liukulukujen tarkkuus ja siihen liittyen myös jatkuvasta summaamisesta mahdollisesti kerääntyvät pyöristysvirheet. Erityisesti koska p-arvo lasketaan komplementtitodennäköisyyden $1-p$ kautta, ei tarkkuus pääse kasvamaan p-arvojen pienentymisen myötä. Tarkoilla rationaaliluvuilla laskettaessa tätä ongelmaa ei luonnollisesti ole. Kolmas merkittävä ongelma maksimipoikkeamatestillä on hitaus. Vaikka saavutettu aikavaativuus on huomattavasti parempi, kuin mitä saavutettaisiin kaikki mahdolliset ristiintaulut luettelemalla, on jo liukulukuversiokin silti käytännössä liian hidas koko kromosomiston kaikkien markkeriparien läpikäymiseen. Emme siis käytä sitä myöhemmissä kokeissa.

2.5 Genotyyppikohtainen tarkka testi

Tähän saakka olemme puhuneet markkerien välisistä assosiaatioista. Vuorovaikutuksen voidaan kuitenkin ajatella tapahtuvan myös yksittäisten genotyyppien välillä. Koko ristiintaulua vastaavan p-arvon sijaan voimmekin laskea kukin taulun ruudulle erikseen merkitsevyyttä kuvaavan p-arvon. Siis markkeriparin XY sijaan tarkastellaan erikseen kaikkia sen saamia arvoja eli genotyyppipareja $\frac{AB}{AB}$, $\frac{AB}{Ab}$, $\frac{Ab}{Ab}$ jne. Lähestymistavalla on sekä hyviä että huonoja puolia. Tärkein saavutettava etu lienee p-arvojen laskemisen helpottuminen. Lisäksi saatavat tulokset kertovat tarkemmin siitä, miten markkeriparin välinen assosiaatio rakentuu.

Olkoon n_{ij} havaittu arvo yksittäisessä ristiintaulun ruudussa ij. Käsitellään mallia, jossa reunafrekvenssit on lukittu. Tällöin vastaava satunnaismuuttuja N_{ij} noudattaa nollahypoteesin vallitessa hypergeometrista jakaumaa [Zar99]:

$$N_{ij} \sim HypGeom(n, n_{i*}, n_{*j}).$$

Vapaiden reunafrekvenssien mallissa noudattaisi N_{ij} vastaavasti binomijakaumaa parametrein p_{ij} sekä n, ja jatko menisi täysin samaan tapaan. Käsittelemme tässä kuitenkin vain lukittua mallia, jossa kaikki jakauman parametrit ovat tunnettuja.

Koska assosiaatiot voivat ilmetä sekä liian suurina että pieninä ruudun arvoina, tulee merkitsevyyden testaukseen käyttää kaksipuoleista p-arvoa. Tähän on useita hiukan poikkeavia lähestymistapoja [Agr02], joista käytämme yksipuoleisen p-arvon kaksinkertaistamista. Merkitään tavanomaisia vasemman- ja oikeanpuoleisia p-arvoja vastaavasti $p_{\leq} = \Pr(N_{ij} \leq n_{ij})$ ja $p_{\geq} = \Pr(N_{ij} \geq n_{ij})$, missä n_{ij} on ruudun havaittu arvo ja N_{ij} on sitä vastaava satunnaismuuttuja. Nyt määrittelemme kaksipuoleisen

p-arvon näiden avulla

$$\mathrm{p} = 2\min(p_{\leq}, p_{\geq}).$$

Vaihtoehtoisesti voitaisiin summata p-arvot niistä tapauksista, joissa N_{ij} poikkeaa odotusarvostaan e_{ij} vähintään havaitun verran. Yleinen tapa on myös laskea yhteen kaikkien korkeintaan yhtä todennäköisten tapausten todennäköisyydet. Valitussa määritelmässä on kuitenkin etuna, että oikeanpuoleinen p-arvo voidaan laskea helposti, jos vasemmanpuoleinen p-arvo on tiedossa: koska N_{ij} voi saada vain kokonaislukuarvoja, on $p_{\geq} = 1 - p_{\leq} + \Pr(N_{ij} = n_{ij})$. Vasemmanpuoleinen p-arvo saadaan suoraan määritelmän mukaan laskettua summaamalla $n_{ij} + 1$ pistetodennäköisyyttä

$$p_{\leq} = \sum_{k=0}^{n_{ij}} \Pr(N_{ij} = k).$$

Hypergeometrisen jakauman pistetodennäköisyydet saadaan tunnetusti laskettua kaavalla $\Pr(N_{ij} = k) = \binom{n_{i*}}{k}\binom{n-n_{i*}}{n_{*j}-k}/\binom{n}{n_{*j}}$ [Agr02]. Binomikertoimien laskemisessa tarvittavat kertomat puolestaan voidaan laskea etukäteen kaikille mahdollisille arvoille $0, 1, \ldots, n$ ajassa $O(n)$, joten pistetodennäköisyydet saadaan itse algoritmissa laskettua vakioajassa. Nyt koska $\sum_{ij} n_{ij} = n$, voidaan kaikkien ristiintaulun ruutujen p-arvot laskea ajassa $\sum_{ij} O(n_{ij} + 1) = O(n + rc)$, missä r ja c ovat taulun rivien ja sarakkeiden määrät. Ristiintaulun muodostaminen aineiston perusteella vaatii jo itsessään saman ajan, joten kokonaisaikavaativuus ei kasva esimerkiksi aiemmin esiteltyihin khii-toiseen-testiin ja G-testiin verrattuna. Kuten maksimipoikkeamatestin tapauksessa, myös tässä mainitun aikavaativuuden saavuttaminen edellyttää lukuarvojen käsittelemistä vakioajassa – siis käytännössä liukulukujen käyttämistä. Tarkassa murtolukuversiossa aikavaativuuteen vaikuttaisivat myös käsiteltävien lukuarvojen suuruudet.

Käytännössä liukulukujen rajallinen tarkkuus saattaa rajoittaa tilanteita, joissa yllä kuvattua laskentatapaa voidaan käyttää. Ongelmia tarkkuuden kanssa syntyy, jos havaitut arvot ovat merkitsevästi odotusarvoaan suurempia. Tällöin $p_{\leq}$ on lähellä ykköstä, ja pyöristysvirheet muodostuvat siten suuriksi. Ratkaisuna on näissä tilanteissa laskea aluksi vasemmanpuoleisen p-arvon sijaan oikeanpuoleinen $p_{\geq}$ summaamalla n_{ij}:tä suurempien arvojen pistetodennäköisyydet. Vasen häntätodennäköisyys $p_{\leq}$ saadaan analogisesti laskettua oikeanpuoleisen avulla. Käytännössä tarvittava suoritusaika pysyy suunnilleen samana kuin aina vasemmalta puolelta summattaessa.

Genotyyppikohtainen tarkka testi (tarvittaessa käytämme testistä jatkossa lyhennettä E-testi) on käytännössä suora yleistys Fisherin [Fis22] tarkalle testille 2×2-

ristiintauluille: Kun ij on testattava genotyyppiyhdistelmä, niin voidaan ajatella, että kummastakin markkerista loput genotyypit $i^c = \{i' = 1, \ldots, r \mid i' \neq i\}$ ja $j^c = \{j' = 1, \ldots, c \mid j' \neq j\}$ yhdistetään. Näin saadaan 2×2-ristiintaulu arvoilla n_{ij}, n_{ij^c}, n_{i^cj} ja $n_{i^cj^c}$. Yksittäinen genotyyppikohtainen tarkka testi vastaa saadulle taululle tehtyä Fisherin tarkkaa testiä.

Genotyyppikohtainen tarkka testi vaikuttaa etenkin yksinkertaisuutensa ja nopeutensa ansiosta houkuttelevalta. Siinä on kuitenkin pari huonoa puolta: Ensinnäkin suoritettavien testien määrä moninkertaistuu. Yksittäisten markkerien välisissä assosiaatioiden mittauksissa käsitellään tyypillisesti 3x3-ristiintauluja, ja testejä tulee siten yhdeksänkertainen määrä. Kahden markkerin ryhmien välillä puolestaan ristiintaulut voivat teoriassa sisältää jopa $9 \cdot 9 = 81$ ruutua. Todellisuudessa kaikkia mahdollisia reunayhdistelmiä ei esiinny, mutta testejä tulee tavallisesti joka tapauksessa useita kymmeniä taulua kohden. Testien määrän lisääntyminen puolestaan hankaloittaa merkitsevien p-arvojen havaitsemista kohinan seasta.

Toinen lievempi ongelma on, että saman markkeriparin genotyyppiyhdistelmien frekvenssit riippuvat selkeästi toisistaan. Riippuvuuksien syntyminen vaikeuttaa teoriassa merkitsevien assosiaatioiden valintaa, jota käsitellään luvussa 3. Toisaalta koska LD-ilmiön ansiosta korrelaatiota esiintyy myös läheisten markkeriparien kesken, ei markkerien tarkasteleminen aiempaan tapaan kokonaisuuksina hävitä riippuvuusongelmaa.

3 Merkitsevien assosiaatioiden valinta

Edellä olemme käyneet läpi erilaisia menetelmiä yksittäisen markkeriparin assosiaation merkitsevyyden mittaamiseen. Yksittäisissä merkitsevyystesteissä käytetään tavallisesti ennalta sovittua merkitsevyysrajaa, esimerkiksi 0.05 tai 0.01 halutusta varmuudesta riippuen. Jos saatu p-arvo on alle valitun kynnysarvon, hylätään nollahypoteesi – tapauksessamme siis oletus testatun markkeriparin riippumattomuudesta. Muussa tapauksessa todistusaineiston ei katsota riittävän nollahypoteesin hylkäämiseen.

Tavoitteenamme on löytää mahdollisimman monta aineistossa esiintyvää todellista assosiaatiota, mutta välttää silti valitsemasta mukaan liian paljon sellaisia riippumattomia markkeripareja, jotka vain sattuvat näyttämään assosioituneilta. Valinnassa tehdyt virheet voidaan jakaa kahteen luokkaan:

- *Hylkäysvirhe* (tyypin I virhe) tarkoittaa havaintoa, joka tuomittiin merkitseväksi (nollahypoteesi hylättiin) vaikka todellista riippuvuutta ei ole. Jos hylkäysvirheiden määrä on pieni, sanotaan valintaprosessia (ja testiä) *tarkaksi*.
- *Hyväksymisvirhe* (tyypin II virhe) puolestaan on ei-merkitseväksi tuomittu pari, joka sisältää todellisen riippuvuuden. Valintaprosessi on *voimakas*, jos hyväksymisvirheitä on vähän eli suuri osa riippuvuuksista löydetään.

P-arvo mittaa suoraan määritelmän mukaan hylkäysvirheen todennäköisyyttä yksittäisen testin tapauksessa. Merkitsevyysrajaa muuttamalla voidaan siis suoraan vaikuttaa hylkäysvirheiden määrään ja siten valinnan tarkkuuteen. Tyypillisesti tämä muuttaa hyväksymisvirheiden määrää päinvastaiseen suuntaan. Jos vaihtoehtoinen hypoteesi on riittävän hyvin rajattu, hyväksymisvirheisiin voidaan merkitsevyysrajan lisäksi pyrkiä vaikuttamaan esimerkiksi sopivan tunnusluvun valinnalla: tunnusluku valittaisiin niin, että nimenomaan vaihtoehtoisen hypoteesiin mukaiset tapaukset saisivat mahdollisimman pieniä p-arvoja. Tämä on kuitenkin ongelmallista, jos vaihtoehtoinen hypoteesi on liian laaja. Esimerkiksi tapauksessamme, jos assosiaatioiden etsinnässä mahdollisten riippuvuuksien tyypistä ei etukäteen tiedetä mitään, täytyy tunnusluvun valinta tehdä muilla perusteilla. Oletamme jatkossa, että tunnusluku ja p-arvon laskutapa on etukäteen valittu, ja keskitymme hylkäysvirheiden rajoittamiseen.

P-arvolla voidaan siis hallita yksittäisen testin hylkäysvirhetodennäköisyyttä. Kun sama testi suoritetaan kaikille halutuille markkeripareille, saadaan kuitenkin suuri määrä p-arvoja. Merkitään näitä $\mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{p}_m$, missä m on suoritettujen testien määrä. Nyt tavallinen tapa on valita jokin merkitsevyysraja α, jonka alapuolelle jäävät p-arvot tuomitaan merkitseviksi. Kuitenkin jos merkitsevyysraja valitaan liian suureksi, aiheuttaa todennäköisesti jo pelkkä satunnaisheilahtelu lukuisia näennäisiä löydöksiä. Ajatellaan esimerkiksi tilannetta, että testattavia markkeripareja olisi 10^{10} kappaletta ja merkitsevyysrajaksi valittaisiin $\alpha = 0.01$. Tällöin, vaikka kaikki parit olisivat riippumattomia, alittaisi p-arvon tasaisesta jakautumisesta johtuen odotusarvoisesti $10^{10} \cdot 0.01 = 10^8$ paria merkitsevyyskynnyksen. Tämä ei selvästikään ole toivottavaa.

Merkitsevien tulosten valintaan monien testien joukosta on kehitetty lukuisia erilaisia menetelmiä. Tyypillisesti nämä valitsevat jollain perusteella sopivan kynnysarvon ja toteavat sen alapuolelle jäävät p-arvot tilastollisesti merkitseviksi. Yksittäistä piirrettä (tai markkeria) koskevia koko genomin laajuisia assosiaatiotestejä varten on myös yritetty löytää yleisesti käypää merkitsevyyskynnystä [DG08, HCDI+08].

Taulukko 4: Kukin riippumattomuustesti voidaan sijoittaa yhteen neljästä luokasta sen mukaan, onko parin välillä riippuvuutta ja tuomittiinko havainto merkitseväksi. Taulukon kirjaimet merkitsevät kunkin tulostyypin kokonaismäärää. Isolla kirjoitetut ovat satunnaismuuttujia, pienellä kirjoitetut puolestaan vakioita. Testien kokonaismäärä m on luonnollisesti tiedossa, mutta tosien nollahypoteesien määrä m_0 on tuntematon. Lisäksi kun valinta on suoritettu, tunnetaan R, mutta ei muiden satunnaismuuttujien arvoja.

	tuomittu epämerkitseväksi	tuomittu merkitseväksi	yhteensä
ei riippuvuutta	U	V	m_0
riippuvuus	T	S	$m - m_0$
yhteensä	$m - R$	R	m

Tyypillisesti näissä on päädytty suuruusluokkaa 10^{-8} tai 10^{-7} olevaan kynnysarvoon. Tämä ei luonnollisesti suoraan käy markkeriparien assosiaation testaamiseen, sillä tapauksessamme testien määrä on moninkertainen. Olemmekin keskittyneet yleisiin merkitsevien p-arvojen valintamenetelmiin.

Kun merkitsevien tulosten valinta on tehty, voidaan testit jakaa taulukon 4 mukaisesti neljään kategoriaan. Taulukon mukaisesti hylkäysvirheiden määrää merkitsemme V:llä. Tutustumme seuraavaksi muutamaan eri V:tä rajoittavaan valintamenetelmään, jotka perustuvat kahden hylkäysvirheisiin liittyvän suureen – koekohtaisen virhetodennäköisyyden ja hylkäysvirheasteen – pitämiseen halutuissa rajoissa.

P-arvojen riippuvuus

Kuten kohta näemme, merkitsevien p-arvojen valinta hylkäysvirheiden määrää halutulla tavalla kontrolloiden on suhteellisen yksinkertaista, jos nollahypoteesin mukaiset testit (ja niistä saadut p-arvot) ovat keskenään riippumattomia. Koska geneettisen kytkennän ansiosta vierekkäisten markkerien arvot ovat vahvasti assosioituneita, eivät testit kuitenkaan todellisuudessa ole riippumattomia. Jos esimerkiksi eri kromosomeissa sijaitsevien markkereiden A ja B välillä näyttäisi olevan jonkinasteinen riippuvuus, on todennäköistä, että A näyttää olevan assosioitunut myös B:n vieressä sijaitsevan markkerin C kanssa. Vastaavasti jos A ja B ovat täysin riippumattomia, ei A:n ja C:n välilläkään luultavasti ole havaittavaa riippuvuutta. Lähekkäisten testien p-arvot ovat siis positiivisesti korreloituneita. Käytännössä positiivinen korrelaatio

vähentää testien efektiivistä määrää; äärimmäisen vahvan kytkennän tapauksessa kaikki kromosomin markkerit olisivat kopioita toisistaan, jolloin kaikki testit antaisivat saman tuloksen ja testien efektiivinen määrä olisi siten yksi.

Samankaltainen riippuvuusilmiö voidaan havaita pienoiskoossa genotyyppikohtaista tarkasta testistä saaduille saman markkeriparin eri genotyyppien p-arvoille: Jos yksi ristiintaulun ruutu poikkeaa selvästi odotusarvostaan, täytyy poikkeama kompensoitua jotenkin muissa ruuduissa. Myös tässä tapauksessa p-arvot näyttäisivät siis olevan positiivisesti korreloituneita.

3.1 FWER-kontrollointi

Perinteinen tapa suorittaa valinta moninkertaisen testauksen yhteydessä on rajoittaa *koekohtaista virhetodennäköisyyttä*, josta käytetään lyhennettä *FWER* (engl. familywise error rate). FWER määritellään todennäköisyytenä, että merkitseviksi valitut testit sisältävät vähintään yhden hylkäysvirheen:

$$\mathrm{FWER} = \Pr(V \geq 1).$$

Tarkoituksena on rajoittaa FWER halutun kynnysarvon α alapuolelle. Tähän on kehitetty lukuisia menetelmiä, joista käsittelemme lyhyesti muutamaa. Yksinkertaisin tapa on käyttää *Bonferroni-korjausta* [Sha95], eli valita merkitseviksi ne testit i, joilla $\mathrm{p}_i < \alpha/m$. Bonferroni-korjatuilla p-arvoilla viitataan arvoihin mp_i, joita voidaan suoraan verrata kynnysarvoon α. On helppo nähdä, että Bonferroni-korjauksen käyttö rajoittaa FWER:ää halutulla tavalla: Olkoon $T_0 \subset \{1, 2, \ldots, m\}$ niiden testien joukko, joilla nollahypoteesi on tosi. Tällöin koekohtainen virhetodennäköisyys on $\mathrm{FWER} = \Pr(\cup_{i \in T_0}(\mathrm{p}_i < \alpha/m))$. Arvioimalla yhdisteen todennäköisyyttä ylöspäin todennäköisyyksien summalla saadaan edelleen

$$\mathrm{FWER} \leq \sum_{i \in T_0} \Pr(\mathrm{p}_i < \alpha/m) = |T_0|\frac{\alpha}{m} \leq \alpha.$$

Holmin [Hol79] esittelemässä versiossa p-arvot järjestetään aluksi pienimmästä suurimpaan. Olkoot $p_1, p_2, \ldots, p_m$ nämä järjestetyt p-arvot. Tämän jälkeen ensimmäinen näistä valitaan merkitseväksi, mikäli $\mathrm{p}_1 < \alpha/m$. Jos p_1 ei ollut merkitsevä, lopetetaan. Muussa tapauksessa valitaan seuraava merkitseväksi, jos $\mathrm{p}_2 < \alpha/(m-1)$. Jos myös p_2 oli merkitsevä, jatketaan vertailulla $\mathrm{p}_3 < \alpha/(m-2)$. Näin jatketaan, kunnes kohdataan ensimmäinen p-arvo, jota ei todeta merkitseväksi. Holm osoitti,

että myös tämä menettelytapa rajoittaa koekohtaisen virhetodennäköisyyden kynnysarvon α:n alapuolelle.

Esitetyt perusversiot Bonferroni-korjauksesta ja Holmin proseduurista eivät tee mitään oletuksia p-arvojen välisistä suhteista. Jos p-arvojen tiedetään olevan riippumattomia, voidaan kummassakin tapauksessa ehtoja hieman löysentää korvaamalla α/k arvolla $1-(1-\alpha)^{1/k}$ kaikilla $k = 1, 2, \ldots, m$. Kuten aiemmin olemme todenneet, riippumattomuus ei kuitenkaan tarkalleen ottaen päde markkerien välisissä assosiaatiomittauksissa. Lisäksi pienillä p-arvoilla löysennyksestä saatava parannus on käytännössä merkityksetön.

3.2 FDR-kontrollointi

Edellisen luvun menetelmissä merkitsevien p-arvojen valinta tehdään siten, että FWER pysyy tarpeeksi pienenä. Tämä on ehkä liiankin tiukka rajoitus, jonka takia valintaprosessin voimakkuus kärsii turhaan. Jos merkitseviä p-arvoja saataisiin löydettyä paljon, ei muutaman hylkäysvirheen eksyminen valittujen joukkoon välttämättä olisi kovinkaan haitallista. Merkitään Q:lla virheellisten valintojen määrän suhdetta kaikkien valintojen määrään. Siis $Q = V/R$. Jos yhtään valintaa ei tehdä, on $R = 0$. Tällöin määritellään erikseen hylkäysvirheiden osuudeksi $Q = 0$. Koska V:n arvoa ei tiedetä, on myös Q satunnaismuuttuja, jonka arvoa ei tunneta eikä voida suoraan kontrolloida. Benjaminin ja Hochbergin [BH95] esittelemä *hylkäysvirheaste* eli *FDR* (engl. false discover rate) on odotusarvoinen hylkäysvirheiden osuus kaikista merkitseväksi valituista:

$$\mathrm{FDR} = \mathrm{E}\left[Q\right] = \Pr(R > 0)\,\mathrm{E}\left[\frac{V}{R} \mid R > 0\right].$$

Jos $m_0 = m$ eli nollahypoteesi pätee kaikilla testeillä, on $V = R$. Tässä tapauksessa siis $\mathrm{E}[Q] = \Pr(V > 0)$, eli FDR:n kontrollointi kontrolloi suoraan myös yllä esiteltyä koekohtaista virhetodennäköisyyttä. Toisaalta jos osalla testeistä nollahypoteesi ei päde, on FDR pienempi kuin FWER, ja sen kontrollointi on siten helpompaa.

Riippumattomien p-arvojen tapaus

Olkoot p-arvot järjestetty pienimmästä suurimpaan $p_1, p_2, \ldots, p_m$. Nyt hylkäysvirheastetta voidaan kontrolloida valitsemalla merkitsevät p-arvot seuraavasti: Olkoon q ennalta valittu kynnysarvo ja k suurin $i \in \{1, 2, \ldots, m\}$, jolle pätee

$$p_i \leq \frac{i}{m} q. \tag{2}$$

Merkitseväksi valitaan k pienintä p-arvoa, eli arvot $p_1, p_2, \ldots, p_k$. Jos tällaista k ei ole, ei yhtään p-arvoa valita merkitseväksi. Kutsumme jatkossa tätä menetelmää *FDR-valintamenetelmäksi*.

Benjamini ja Hochberg osoittivat, että jos p-arvot ovat nollahypoteesin vallitessa keskenään riippumattomia, pyrkii FDR-valintamenetelmä maksimoimaan valittujen lukumäärän $R = k$ pitäen silti FDR:n annetun kynnysarvon alapuolella:

$$E[Q] \leq \frac{m_0}{m} q \leq q,$$

missä m_0 on tosien nollahypoteesien lukumäärä. Kuten yllä olevasta epäyhtälöstä voidaan huomata, on kontrollointi itse asiassa hieman konservatiivinen. Kuitenkin jos suurimmalla osalla testeistä nollahypoteesi pätee, on raja varsin tiukka. Huomioitavaa on, että riippumattomuusvaatimus koskee ainoastaan nollahypoteesin mukaisten testien p-arvoja eikä ota kantaa p-arvoihin, joille vaihtoehtoinen hypoteesi pätee.

FDR:n toimintaa voidaan helposti havainnollistaan seuraavan erityistapauksen avulla: Ajatellaan tilannetta, jossa suoritetaan yhteensä m testiä ja näistä m_0 noudattaa nollahypoteesia. Oletetaan, että lopuilla $m_1 = m - m_0$ vaihtoehtoista hypoteesia noudattavalla testillä saatu p-arvo on todella pieni (niin pieni, että FDR-tulee sen varmasti valitsemaan). Merkitään näitä vaihtoehtoiseen hypoteesiin liittyviä p-arvoja suuruusjärjestyksessä $\epsilon_1, \epsilon_2, \ldots, \epsilon_{m_1}$. Tehdään lisäksi oletus, että nollahypoteesin mukaisten testien p-arvot muodostavat järjestettynä jonon $\frac{1}{m_0}, \frac{2}{m_0}, \ldots, \frac{m_0}{m_0}$ sijoittuen siten tasaisin välein yksikkövälille $[0, 1]$. Koska nollahypoteesin vallitessa p-arvo on tasaisesti jakautunut, voidaan kyseistä sijoittelua ajatella edustavana esimerkkinä. Kaikki p-arvot muodostavat siis järjestämisen jälkeen jonon $(p_1, p_2, \ldots, p_m) = (\epsilon_1, \epsilon_2, \ldots, \epsilon_{m_1}, \frac{1}{m_0}, \frac{2}{m_0}, \ldots, \frac{m_0}{m_0})$. Merkitään FDR:n valitsemien p-arvojen lukumäärää $R = m_1 + j$. Nyt koska $p_{m_1+j} = j/m_0$, seuraa FDR:n kynnysehdosta (2), että

$$\frac{j}{m_0} \leq \frac{m_1 + j}{m} q.$$

Vaihtamalla tekijöitä sopivasti ristiin saadaan ehto muotoon

$$\frac{j}{m_1 + j} \leq \frac{m_0}{m} q \leq q.$$

Koska valittujen kokonaismäärä on $R = m_1 + j$ ja tästä $V = j$ on hylkäysvirheiden määrä, on hylkäysvirheiden osuus $Q = j/(m_1 + j) \leq q$. Saimme siis tässä erikoistapauksessa Benjaminin ja Hochbergin osoittamaa yleistä odotusarvoista tulosta vastaavan ylärajan hylkäysvirheiden osuudelle.

Riippuvuuden huomioiminen

Alun perin Benjamini ja Hochberg osoittivat FDR-valintamenetelmän kontrolloivan hylkäysvirheastetta vain tapauksessa, jossa nollahypoteesin mukaiset p-arvot (tai tunnusluvut) ovat riippumattomia. Myöhemmin Benjamini ja Yekutieli [BY01] kuitenkin todistivat menetelmän toimivan myös tietynlaisen positiivisen riippuvuuden vallitessa tunnuslukujen välillä. Tarkempaa luonnehdintaa varten määrittelemme *PRDS-ominaisuuden* (engl. positive regression dependency on a subset) suoritettujen testien mielivaltaiselle osajoukolle $I_0 \subset \{1, 2, \ldots, m\}$.

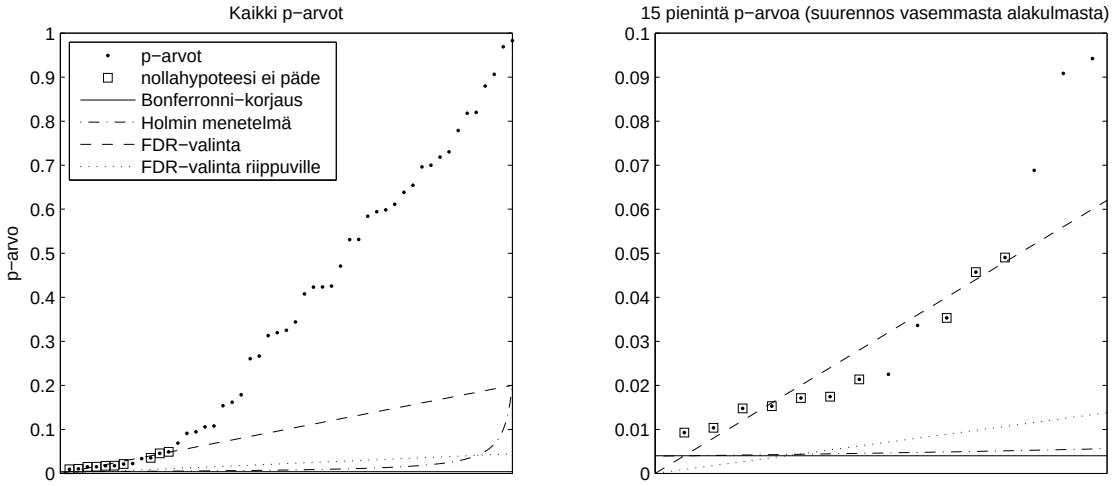
Määritelmä 1. Joukolla I_0 on PRDS-ominaisuus, jos mille tahansa kasvavalle joukolle D ja kaikille $i \in I_0$ pätee, että $\Pr(\mathbf{X} \in D \mid X_i = x)$ on ei-vähenevä x:n suhteen.

Yllä olevassa määritelmässä $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ koostuu testien tunnuslukuja vastaavista satunnaismuuttujista, ja joukon D alkiot ovat siten mahdollisia tunnuslukujen arvoasetuksia. Joukon D kasvavuusvaatimus tarkoittaa, että jos $\mathbf{x} \in D$ ja $\mathbf{y} \geq \mathbf{x}$, niin myös $\mathbf{y} \in D$.

Oletetaan, että tunnusluvuille suoritettava merkitsevyystestaus on yksipuoleinen (muuten tunnuslukujen riippuvuuksien jaottelu positiivisiin ja negatiivisiin ei olisi edes mielekästä). Benjamini ja Yekutieli osoittivat, että jos testien tunnuslukujen yhteisjakauma noudattaa PRDS-ominaisuutta tosia nollahypoteeseja vastaavalla testien osajoukolla I_0, rajoittaa yllä esitelty FDR-valintamenetelmä hylkäysvirheasteen edelleen saman rajan $m_0/m \cdot q$ alapuolelle. Tunnuslukujen jakauman sijaan PRDS-ominaisuus voidaan vaatia suoraan myös p-arvoilta. On helppo nähdä, että tässä tapauksessa alkuperäisen tunnusluvun ei tarvitse enää olla yksipuoleinen, sillä p-arvo on itsessään eräs yksipuoleinen tunnusluku. Markkeripareille tehtyjen assosiaatiotestien välisten riippuvuuksien luonteesta on vaikea sanoa mitään täysin varmaa. Kuitenkin intuitiivisesti ajatellen PRDS-ominaisuuden voisi olettaa ainakin pääpiirteittäin pätevän näille testeille.

FDR-valintamenetelmää voidaan helposti muokata toimimaan myös tapauksissa, joissa PRDS-ominaisuus ei päde. Benjamini ja Yekutieli osoittivat, että jos valintamenetelmässä kynnysarvo q korvataan arvolla $q/(\sum_{i=1}^{m} 1/i)$, on hylkäysvirheaste aina korkeintaan $m_0/m \cdot q$. Koska $\sum_{i=1}^{m} 1/i \approx \ln m$, on muunnetun valintaprosessin voimakkuus selvästi aiempaa heikompi. Toisaalta, jos valintoja tulee enemmän kuin $\ln m$ kappaletta, on menetelmä silti voimakkaampi kuin Bonferroni-korjaus.

Kuvassa 3 on havainnollistettu FDR-valintamenetelmien sekä Bonferroni-korjauksen



Kuva 3: Kuvassa on havainnollistettu eri valintamenetelmien toimintaa 50 keinotekoisesti tuotetulla p-arvolla. P-arvoista 40 on arvottu nollahypoteesin mukaisesti tasaisesta $[0, 1]$-jakaumasta, ja loput 10 vaihtoehtoista hypoteesia edustavat p-arvot puolestaan on arvottu beeta-jakaumasta parametrein 2 ja 100. Tämän jälkeen p-arvot on järjestetty pienimmästä suurimpaan. Jokaista valintamenetelmää edustaa kuvassa oma viiva tai käyrä. Kynnysarvoiksi on valittu $\alpha = q = 0.2$. Bonferroni-korjauksen tapauksessa merkitseviksi valitaan kaikki viivan alapuolelle jäävät p-arvot. Holmin menetelmä etsii vasemmalta alkaen ensimmäisen p-arvon, joka on käyrän yläpuolella, ja valitsee kaikki sen vasemmalle puolelle jäävät p-arvot. FDR-valintamenetelmät etsivät viimeisen viivan alapuolelle jäävän p-arvon ja valitsevat merkitseviksi sen sekä kaikki siitä vasemmalle. Vasemmanpuoleinen kuva sisältää kaikki 50 p-arvoa ja oikeanpuoleisessa on kuvattu suurennettuna näistä 15 pienintä. Esimerkin tapauksessa huomataan, että ainoa vähintään yhden valinnan tekevä menetelmä on riippumattomille ja positiivisesti korreloituneille p-arvoille tarkoitettu FDR-valintamenetelmä, joka poimii kymmenen ensimmäistä p-arvoa. Näistä kaksi kappaletta on todellisuudessa nollahypoteesin mukaisia. Hylkäysvirheiden osuus on siis 20 %, joka tässä tapauksessa vastaa täsmälleen asetettua kynnysarvoa. Hyväksymisvirheitä on myös tässä tapauksessa kaksi.

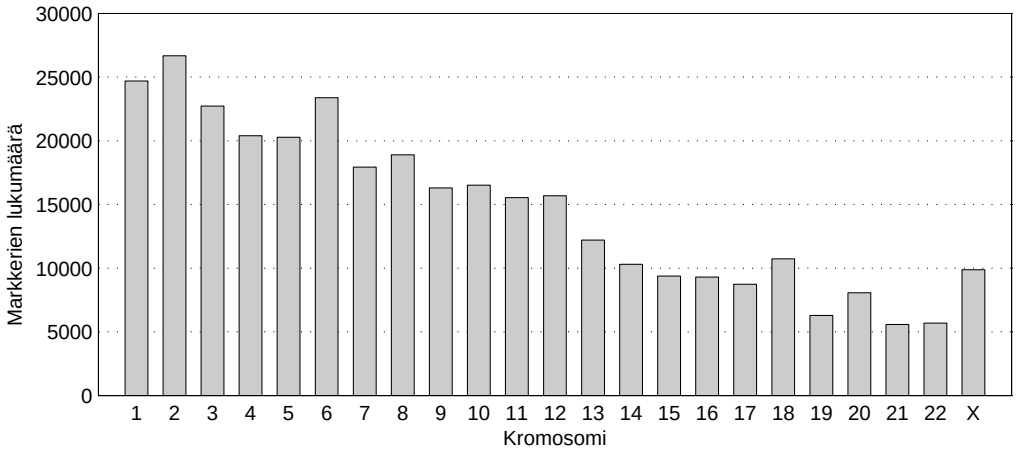
ja Holmin menetelmän toimintaa keinotekoisilla p-arvoilla. Selvästi on nähtävissä, että Holmin menetelmällä saavutettu voimakkuuden kasvu verrattuna Bonferroni-korjaukseen on mitätön, ellei merkitseviä p-arvojen osuus ole todella suuri. FDR-menetelmistä saavutetaan huomattavaa etua verrattuna FWER:n rajoittamiseen erityisesti tilanteissa, joissa merkitseviä p-arvoja on useita ja ne ovat samaa suuruusluokkaa kuin pienimmät ei-merkitsevät p-arvot.

Toisaalta, jos vaihtoehtoisesta hypoteesista saatavat p-arvot ovat lähes aina kertaluokkia pienempiä pienimmät nollahypoteesista saatavat p-arvot, löytyvät ne myös Bonferroni-korjausta käyttämällä. Tällöin, mikäli nollahypoteesin mukaiset p-arvot ovat tasaisesti jakautuneet, valitsee FDR oikeasti merkitsevien p-arvojen lisäksi mukaan myös suurin piirtein kynnysarvoa q vastaavan osuuden ”turhia”, todellisuudessa ei-merkitseviä p-arvoja.

Muunnelmat

FDR-valintamenetelmästä on kehitetty lukuisia erilaisia muunnelmia. Esitetty valintamenetelmä pitää hylkäysvirheasteen kertoimen m_0/m verran haluttua pienempänä. Korvaamalla valinnassa kynnysarvo q arvolla $q^* = qm/m_0$ päästäisiin tiukkaan rajaan. Adaptiivisissa FDR-menetelmissä [BH00, BKY06] valintaprosessi suoritetaan tyypillisesti kaksi kertaa: ensimmäisen kerran tuloksen perusteella muodostetaan arvio tosien nollahypoteesien määrästä m_0, ja arviota käytetään hyväksi toisella valintakierroksella. Adaptiivisuudella saavutetaan merkittävää hyötyä vain, jos tosien nollahypoteesien määrä m_0 on paljon pienempi kuin kaikkien testien määrä m. Tämä tuskin pitää kromosomien välisillä genotyyppipareilla paikkansa: oletettavasti kromosomien pitäisi olla ainakin pääosin täysin riippumattomia.

Jos nollahypoteesin mukaiset p-arvot ovat keskenään vahvasti riippuvia, saattaa tavanomainen FDR-menetelmä toimia tarpeettoman konservatiivisesti. Yekutieli ja Benjamini [YB99] ovat ehdottaneet näytteenottoon perustuvaa FDR:n arviointia ratkaisuna ratkaisuna tämänkaltaisiin tilanteisiin. Luonnollisesti näytteenoton haittapuolena on sen suhteellinen raskaus verrattuna tavanomaiseen FDR-valintaan. FDR:n muunnelmia on käsitelty hieman tarkemmin ja vertailtu FWER:n kanssa esimerkiksi Reinerin ja kumppaneiden [RYB03] sekä Verhoevenin ja kumppaneiden [VSM05] artikkeleissa.



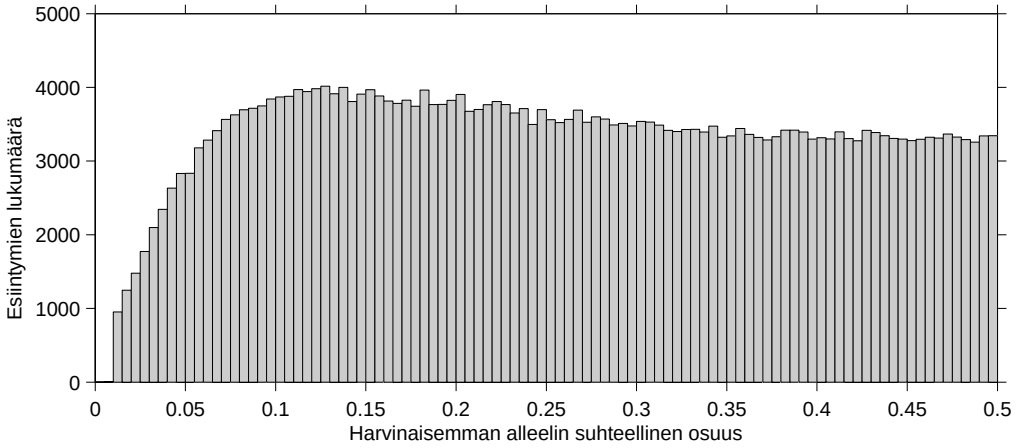
Kuva 4: Markkerien jakautuminen eri kromosomeihin NFBC-aineistossa.

4 Koetuloksia

Tässä osiossa tavoitteenamme on testata edellä esiteltyjä menetelmiä oikealla markkeriaineistolla. Tätä ennen suoritamme kuitenkin muutamia kokeita keinotekoisilla aineistokokoelmilla ja näytämme, että testit toimivat näissä tapauksissa oletetulla tavalla. Sen jälkeen haemme assosiaatioita käytössä olevasta NFBC-aineistokokoelmasta sekä vertailun vuoksi myös muutamasta pienemmästä HapMap-projektin aineistokokoelmasta. Lopuksi vertailemme löytyneitä assosiaatioita ja pohdimme niiden syitä.

Kuten johdannossa totesimme, käytämme kokeissa pääasiassa reilun 5300 näytteen NFBC-aineistoa. Tämän lisäksi vertailukohtana on neljä erillistä HapMap-projektin aineistoa, CHB, MEX, MKK sekä YRI, joiden koot vaihtelevat vajaasta sadasta noin kahteensataan yksilöön. NFBC sisältää kaikkiaan noin 340 000 markkeria. HapMap-aineistot ovat alkujaan suurempia, yli miljoona markkeria, mutta näistä käytetään vain noin 260 000 markkeria, jotka sisältyvät kaikkiin viiteen testiaineistoon. Jatkossa HapMap-aineistoista puhuttaessa viittaamme karsittuun nämä 260 000 yhteistä markkeria sisältäviin aineistoihin. NFBC-aineiston tapauksessa kokeet koskevat kuitenkin pääosin kaikkia 340 000 markkeria poikkeuksena luonnollisesti vertailut HapMap-aineistosta saatuihin tuloksiin.

Kaikki käytetyt aineistot sisältävät likimäärin yhtä paljon näytteitä kummaltakin sukupuolelta. Suurin suhteellinen ero on CHB-aineistossa, joka sisältää 53 miestä ja 84 naista. Kaikissa aineistoissa markkerit ovat jakautuneet eri kromosomeihin suurin



Kuva 5: Kuvassa on NFBC-aineiston harvinaisemman alleelin suhteellisien osuuksien jakauma. Kaikilla markkereilla kumpaakin alleelia esiintyy vähintään 80 kertaa, ja harvinaisemman alleelin osuus on siten kaikissa tapauksissa ainakin 0.77 %.

piirtein kromosomien kokojen mukaisessa suhteessa. Kromosomijako on koko NFBC-aineiston osalta esitetty kuvassa 4. Rajoittuminen yhteisiin markkereihin ei juuri vaikuta kromosomien suhteellisiin osuuksiin. Poikkeuksena on X-kromosomi, jonka suhteellinen osuus markkereista laskee noin kolmasosaan. X-kromosomista on hyvä huomata myös, että miespuolisilla näytteillä ainoastaan kaksi genotyyppiä ovat mahdollisia. Testiaineistoissa näitä on merkitty heterotsygooteilla genotyypeillä ($\binom{A}{A}$ ja $\binom{a}{a}$). Tästä aiheutuu luonnollisesti selkeä ero mies- ja naispuolisten näytteiden genotyyppijakaumien välille. Koska ero esiintyy vain X-kromosomin markkereilla, ei sen kuitenkaan pitäisi aiheuttaa ylimääräisiä kromosomien välisiä assosiaatioita. Poikkeus tähän luonnollisesti olisi, jos markkerin arvo jossain muussa kromosomissa olisi assosioitunut sukupuolen kanssa, mikä itsessään olisi mielenkiintoinen löytö. Kokeissa on pidetty X-kromosomi mukana ilman minkäänlaista erikoiskohtelua.

Kuvassa 5 oleva histogrammi näyttää harvinaisemman alleelin suhteellisen osuuden jakautumisen NFBC-aineistossa. Pienimpien osuuksien puuttumista lukuun ottamatta jakauma on suhteellisen tasainen. Huomattavaa on, että toisin kuin NFBC-aineistossa, kaikissa kolmessa HapMap-aineistossa pienimpiä alleelifrekvenssejä ei ole leikattu pois, ja jopa täsmälleen yhden kerran esiintyviä alleeleja löytyy näistä aineistoista.

Kolmen mahdollisen genotyypin lisäksi aineistoissa esiintyy myös puuttuvia arvoja. Tyypillisesti genotyyppi puuttuu muutamalta yksilöltä markkeria kohden, mutta joillakin markkereilla arvoista saattaa puuttua noin 5 %. Markkeriparin välistä as-

sosiaatiota testattaessa on hyödynnetty näytteistä vain sitä osaa, jolla kummankin markkerin arvo on saatavilla. Vastaavasti markkeriryhmien välisen assosiaation tapauksessa on poistettu näytteet, joilla puuttuu yksikin arvo jommastakummasta testin kohteena olevasta markkeriryhmästä. Puuttuvien arvojen suhteellisen pienestä määrästä johtuen tämän ei pitäisi vaikuttaa merkittävästi testien voimakkuuteen. Toinen vaihtoehto olisi täyttää puuttuvat arvot jotenkin, esimerkiksi arpomalla näihin genotyypit markkerissa esiintyvän genotyyppijakauman mukaisesti. Suurta hyötyä tästä ei kuitenkaan saataisi ja keinotekoisesti täydennetyt arvot voisivat pahimmassa tapauksessa vääristää saatavia tuloksia.

Kokeissa on käytetty pääasiassa genotyyppikohtaista tarkkaa testiä, mutta tuloksia on vertailtu myös khii-toiseen-testin ja G-testin kanssa. Merkitsevien tulosten valintaan on käytetty tavanomaista Bonferroni-korjausta sekä FDR-valintamenetelmää riippumattomille ja positiivisesti riippuville testeille. Genotyyppikohtaisen testin tapauksessa valintaprosessi on suoritettu kaikille genotyyppien p-arvoille ja mahdollinen saman markkeriparin päällekkäiset valinnat on yhdistetty jälkikäteen.

Kokeet suoritettiin 64-bittisessä Linux-ympäristössä. Suurelta osin kokeissa käytettiin MATLAB-ohjelmistoa, mutta suorituskyvyn kannalta kriittiset ohjelmaosat kirjoitettiin C-kiellä. Käytettyjen koneiden ominaisuuksista on kerrottu tarkemmin ajankulutuksesta kertovien mainintojen yhteydessä.

4.1 Kokeet keinotekoisella aineistolla

Keinotekoisella aineistolla suoritettujen kokeiden tarkoituksena on osoittaa kontrolloiduissa olosuhteissa, että testit toimivat halutulla tavalla. Käytössä oli kahdenlaisia aineistoja: täysin satunnaisesti tuotettuja sekä aidosta `NFBC`-genotyyppidatasta muokattuja otoksia. Kaikissa kokeissa käytettiin kahta samankokoista (esimerkiksi sadasta markkerista koottua) kromosomia, joiden väliltä laskettiin p-arvot kaikkien mahdollisten markkeriparien välisille riippumattomuuksille. Keinotekoisissa testiaineistoissa näytekokona käytettiin tuhatta yksilöä, joka on sopivasti `NFBC`-aineiston reilun 5000 näytteen ja HapMap-aineistojen parin sadan näytteen välimaastossa.

Täysin riippumaton aineisto

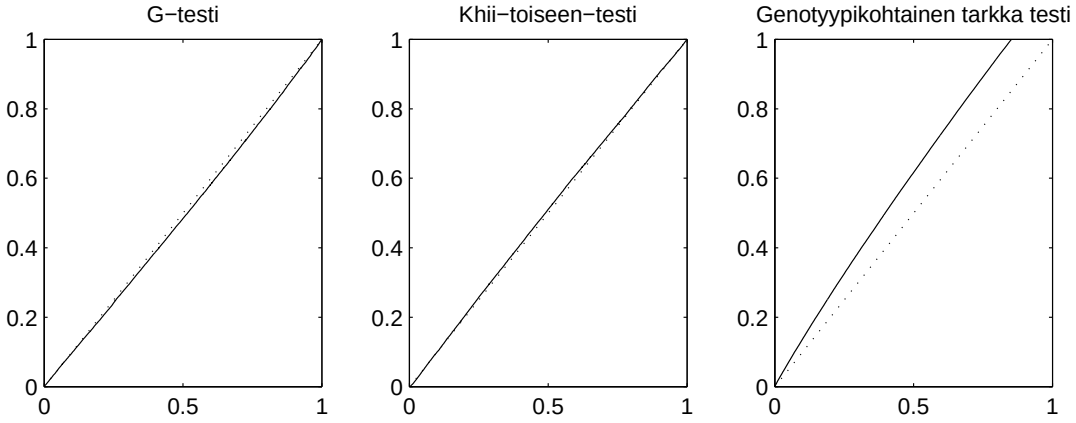
Aluksi testejä kokeiltiin aineistolla, jossa minkäänlaisia assosiaatioita ei esiinny. Täysin ideaalista riippumatonta genotyyppiaineistoa tuotettiin seuraavasti. Ensin arvottiin kullekin markkerille alleelifrekvenssit (p_A ja $p_a = 1 - p_A$) riippumattomasti ta-

saisesta jakaumasta väliltä $[0, 1]$. Alleelijakaumista laskettiin genotyyppien jakaumat Hardy–Weinberg-tasapainoehdon mukaisesti: $p_A^A = (p_A)^2$, $p_a^A = 2p_Ap_a$ ja $p_a^a = (p_a)^2$. Tämän jälkeen kullekin yksilölle arvottiin genotyypit riippumattomasti näistä jakaumista.

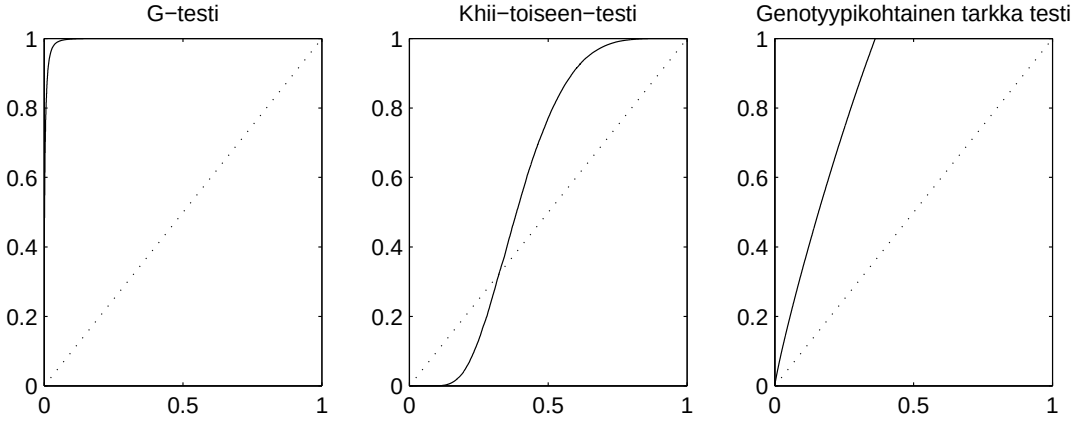
Tällä tavalla muodostettiin kaksi sadan markkerin kromosomia, joiden väliltä mitattiin assosioitumista ristiin kaikkien 10 000 markkeriparien välillä. Yksittäisten markkerien parien lisäksi kokeiltiin myös kolmen peräkkäisen markkerin ryhmien välisten assosiaatioiden mittaamista. Kolmen markkerin ryhmistä kromosomien välisiä markkeripareja saadaan muodostettua yhteensä $98^2 = 9604$ kappaletta. Kokeessa käytettiin kolmea eri testimenetelmää: G-testiä, khii-toiseen testiä ja genotyyppikohtaista tarkkaa testiä. Kuvassa 6 on esitetty tuloksena saatujen p-arvojen jakautuminen eri testeillä. Yksittäisten testien välisissä assosiaatioissa G-testin ja khii-toiseen testin asymptotiikat toimivat varsin hyvin, mutta suuremmat markkeriryhmät aiheuttavat vakavia ongelmia. Genotyyppikohtaisella testillä p-arvot ovat kummassakin tapauksessa selvästi konservatiivisia, mutta markkerien ryhmittelystä seuraavat ongelmat eivät ole yhtä pahoja kuin kahdella muulla testillä.

Kullakin testillä saaduista p-arvoista valittiin merkitsevät käyttäen sekä tavallista Bonferroni-korjausta että normaalia FDR-valintamenetelmää. Valintakynnyksenä kummallakin valintamenetelmällä käytettiin arvoa $\alpha = q = 0.10$. Taulukossa 5a on luetteloitu kullakin testaus- ja valintamenetelmällä saavutettavat keskimääräinen merkitseväksi valittujen p-arvojen lukumäärä sekä hylkäysvirheaste, kun assosiaatioita mitattiin yksittäisten markkerien välillä. Lukemat on saatu toistamalla kukin koe 500 kertaa ja laskemalla näistä saatujen tulosten keskiarvot. Taulukko 5b sisältää vastaavat lukemat kolmen markkerin ryhmien välisille assosiaatioille. Koska aineisto ei sisällä yhtään todellista assosiaatiota, ovat kaikki valinnat hylkäysvirheitä. Näin ollen FDR on tässä erikoistapauksessa sama kuin FWER ja kertoo suoraan, kuinka suuressa osassa tapauksia vähintään yksi p-arvo tuli valittua merkitseväksi.

Taulukoista voidaan havaita, että yksittäisten markkerien tapauksessa G-testi toimii hyvin halutulla tavalla ja kummatkin valintamenetelmät saavuttavat suunnilleen kynnysarvoa vastaavan FDR:n ja FWER:n. Genotyyppikohtaisella testillä valinnat toimivat jossain määrin konservatiivisesti rajoittaen FDR:n alle puoleen kynnysarvosta. Khii-toiseen-testillä sen sijaan käytännössä aina tuli vähintään yksi p-arvo tuomittua merkitseväksi ja valintamenetelmästä riippuen valituksi tuli keskimäärin muutama promille kaikista pareista. Ongelmat johtuvat pienistä genotyyppifrekvensseistä. Kolmen markkerin ryhmillä G-testi on selvästi liian konservatiivinen ja myös



(a) Assosiaatiot yksittäisten markkerien välillä.



(b) Assosiaatiot kolmen markkerin ryhmien välillä.

Kuva 6: P-arvojen jakautuminen eri testeillä, kun käytetty riippumatonta täysin keinotekoista testiaineistoa. Arvot on piirretty vasemmalta oikealle pienimmästä suurimpaan, jolloin muodostuu p-arvojen jakauman kertymäfunktion käänteisfunktio. Ideaalitapauksessa jakauman tulisi olla tasainen ja arvojen kulkea diagonaalia pitkin. Tätä on merkitty pisteviivalla. G-testillä ja khii-toiseen testillä on etenkin suuremmilla ryhmillä selvästi havaittavissa asymptoottisen jakauman väärän muodon ja sijainnin aiheuttamat ongelmat. G-testillä p-arvoista tulee järjestään liian suuria ja khii-toiseen-testillä puolestaan suurelta osin liian pieniä. Khii-toiseen-testillä muodostuu vasempaan alakulmaan myös yksittäisillä markkereilla vastaavanlainen p-arvojen notkahdus kuin alemmassakin kuvassa. Tämä on kuitenkin liian pieni näkyäkseen kuvassa. Genotyyppikohtaisella tarkalla testillä on huomattavissa diskreeteistä tapauksista aiheutuva p-arvojen selkeä arvioiminen ylöpäin.

Taulukko 5: Taulukoissa on tarkasteltu merkitsevyystestien ja valintamenetelmien toimivuutta tilanteissa, joissa kromosomien välillä ei ole riippuvuuksia. Kullekin testimenetelmä–valintamenetelmä-parille on listattu keskimääräinen valittujen p-arvojen lukumäärä sekä hylkäysvirheaste. Valintakynnyksenä on sekä FDR- että Bonferroni-menetelmällä käytetty arvoa $\alpha = q = 0.1$.

(a) Riippumaton keinotekoinen aineisto, yksittäiset markkerit

	FDR-valinta			Bonferroni-valinta		
	G-testi	χ^2-testi	E-testi	G-testi	χ^2-testi	E-testi
Hylkäysvirheitä	0.104	19.250	0.058	0.092	9.334	0.046
Hylkäysvirheaste (FDR)	0.090	0.998	0.042	0.088	0.996	0.040

(b) Riippumaton keinotekoinen aineisto, kolmen markkerin ryhmät

	FDR-valinta			Bonferroni-valinta		
	G-testi	χ^2-testi	E-testi	G-testi	χ^2-testi	E-testi
Hylkäysvirheitä	0.000	1579.410	0.012	0.000	744.052	0.012
Hylkäysvirheaste (FDR)	0.000	1.000	0.012	0.000	1.000	0.012

(c) Permutoitu otos `NFBC`-aineistosta, yksittäiset markkerit

	FDR-valinta			Bonferroni-valinta		
	G-testi	χ^2-testi	E-testi	G-testi	χ^2-testi	E-testi
Hylkäysvirheitä	0.208	3.196	0.224	0.090	1.946	0.066
Hylkäysvirheaste (FDR)	0.078	0.768	0.056	0.058	0.742	0.046

(d) Permutoitu otos `NFBC`-aineistosta, kolmen markkerin ryhmät

	FDR-valinta			Bonferroni-valinta		
	G-testi	χ^2-testi	E-testi	G-testi	χ^2-testi	E-testi
Hylkäysvirheitä	0.006	855.710	0.046	0.002	414.190	0.022
Hylkäysvirheaste (FDR)	0.004	1.000	0.018	0.002	1.000	0.016

genotyyppikohtaisella testillä konservatiivisuus lisääntyy jonkin verran. Khii-toiseen-testillä valittujen p-arvojen määrä pyörii kymmenen prosentin kummallakin puolella, mitä ei voi pitää mitenkään hyväksyttävänä. Kaiken kaikkiaan G-testi näyttäisi suoriutuvan erinomaisesti yksittäisten markkerien tapauksessa ja genotyyppikohtainen testi suhteellisen hyvin markkeriryhmien koosta riippumatta.

Täysin keinotekoisen aineiston lisäksi toistettiin kokeet myös muunnetulla otoksella `NFBC`-aineistosta. Testiaineisto muodostettiin valitsemalla 100 peräkkäistä markkeria kromosomeista 1 (markkerit 1001–1100) ja 2 (markkerit 1001–1100). Tämän jälkeen näytteet järjestettiin uudelleen satunnaisesti toisessa kromosomissa hävittäen siten mahdolliset kromosomien väliset riippuvuudet. Saatu testiaineisto erosi täysin keinotekoisesta pääasiassa neljällä tavalla: Ensinnäkin näytteiden lukumäärä oli noin viisinkertainen. Toiseksi aineisto sisälsi puuttuvia arvoja. Kolmas, ehkä tärkein eroavaisuus oli geneettisen kytkennän säilyminen läheisten markkerien välillä. Neljänneksi alleelifrekvensseillä oli aiemmin mainittu alaraja (kuva 5). Tulokset on esitetty taulukoissa 5c ja 5d. Saadut arvot muistuttavat suurelta osin keinotekoisella aineistolla saatuja lukuja. Hylkäysvirheaste pysyy G-testillä ja genotyyppikohtaisella testillä edelleen kynnysarvon määräämissä rajoissa, mutta etenkin valittujen lukumäärä on hieman aiempaa korkeampi. Tämä on luultavasti suoraa seurausta kytkennästä aiheutuvasta testien riippuvuudesta. Khii-toiseen-testin tulokset puolestaan olivat selvästi parempi kuin täysin keinotekoisella aineistolla, mutta eivät edelleenkään pysyneet halutuissa rajoissa. Toisena havaintona muutokset ovat pienempiä siirryttäessä kolmen markkerin ryhmiin yksittäisten markkerien sijaan. Myös tästä on kiittäminen kytkentää, jonka ansiosta erilaisia genotyyppiyhdistelmiä esiintyy tyypillisesti vähemmän ja frekvenssit ovat siten suurempia.

Riippuvuuksien simuloiminen

Riippumattoman datan lisäksi kokeiltiin testi- ja valintamenetelmiä myös aineistolla, johon oli lisätty keinotekoisia assosiaatioita. Tässä käytettiin jälleen muokattua otosta `NFBC`-aineistosta. Testiaineisto muodostettiin seuraavalla tavalla: Aluksi `NFBC`-aineistosta poimittiin sama $100 + 100$ markkerin otos kuin riippumattoman aineiston tapauksessakin ja sekoitettiin toisen kromosomin yksilöt vastaavasti. Tämän jälkeen kromosomeista arvottiin 10 satunnaista markkeriparia. Kutakin arvottua markkeriparia kohden valittiin satunnainen määrä muutettavia yksilöitä sekä genotyyppi, jonka osuutta haluttiin muuttaa. Muutettavien yksilöiden osuus kaikista yksilöistä arvottiin beeta-jakaumasta parametrein (2, 8). Muutettavaksi genotyy-

Taulukko 6: Merkitsevyystestien ja valintamenetelmien toimivuus tilanteissa, joissa kromosomien välille on lisätty keinotekoisia riippuvuuksia. Kullekin testimenetelmä–valintamenetelmä-parille on listattu keskimääräiset oikein ja väärin valittujen p-arvojen lukumäärät sekä keskimääräinen hylkäysvirheaste. Genotyyppikohtaiselle tarkalle testille (E-testi) listatut pääasialliset tulokset koskevat tilannetta, jossa jokaisen genotyypin p-arvoa on käsitelty erillisenä ja valittu genotyyppi luetaan oikeaksi valinnaksi, vaikka juuri kyseisen genotyypin todennäköisyys ei olisikaan aineistoa tuotettaessa muunnettu. Suluissa on lueteltu vastaavat lukemat, kun samaan markkeripariin liittyvät moninkertaiset valinnat on yhdistetty.

	FDR-valinta			Bonferroni-valinta		
	G-testi	χ^2-testi	E-testi	G-testi	χ^2-testi	E-testi
Oikeita hylkäyksiä	8.692	8.748	48.870 (8.798)	8.550	8.612	43.730 (8.554)
Hylkäysvirheitä	1.128	4.950	4.064 (3.256)	0.128	1.928	0.062 (0.050)
Hylkäysvirheaste	0.092	0.310	0.071 (0.231)	0.011	0.160	0.001 (0.005)

piksi poimittiin satunnaisesti yksi kaikista yhdeksästä mahdollisesta vaihtoehdosta. Tämän jälkeen kunkin markkeriparin genotyyppien jakaumassa muutettiin valitun genotyypin todennäköisyyttä satunnaisesti pienemmäksi tai suuremmaksi, ja valituille yksilöille arvottiin uudet genotyypit saadun muunnetun jakauman mukaan. Muunnettu jakauma muodostettiin kertomalla valitun genotyypin todennäköisyys Gamma$(4, 1/4)$-jakaumasta ja normalisoimalla todennäköisyydet todennäköisyysjakaumaksi.

Kuten aiemmin, laskettiin jälleen kaikkien markkeriparien väliset p-arvot G-testillä, khii-toiseen-testillä sekä genotyyppikohtaisella testillä, ja valittiin kullakin testillä saaduista p-arvoista merkitsevät käyttäen sekä Bonferroni-korjausta että FDR-valintamenetelmää ja kynnysarvoa 0.10. Taulukossa 6 on listattu keskimääräiset oikeiden valintojen ja virheellisten valintojen määrät sekä näistä laskettu hylkäysvirheaste. Jälleen G-testillä ja genotyyppikohtaisella testillä FDR:n kontrollointi toimii halutulla tavalla. Jälkimmäisen kohdalla on kuitenkin huomattava, että koska valinnat koskevat yksittäisiä genotyyppejä, voi sama markkeripari tulla valituksi useamman kerran. Tämä myös näkyy selvästi suurempana valintojen kokonaismääränä. Jos samaan markkeripariin osuvat valinnat yhdistetään, kasvaa FDR jonkin verran eikä pysy enää kynnysarvon alapuolella. Tämä johtuu siitä, että todelliset assosiaatiot ovat muodostetussa aineistossa tyypillisesti vahvoja ja näkyvät herkästi useassa genotyypissä.

4.2 Kokeet aidolla aineistolla

Markkerien välisiä assosiaatioita etsittiin koko NFBC-aineistosta käyttäen p-arvojen laskemiseen genotyyppikohtaista tarkkaa testiä. Testinmenetelmän valintaan on useita syitä: Ensinnäkin se on tarpeeksi nopea koko perimäaineiston läpikäymiseen. Toiseksi saadut tulokset ovat tarkkoja, sillä menetelmään ei sisälly jakauman approksimointia. Tämän ansiosta testi toimii oikein kaikissa tilanteissa, myös useamman markkerin ryhmien välisen assosiaation mittaamisessa. Kolmantena perusteena voidaan pitää hienoista epätavallisuutta; lähestymistapa on tavallisesta poikkeava verrattuna esimerkiksi yleisesti käytettyihin khii-toiseen-testiin ja G-testiin.

Merkitsevien p-arvojen valintaan käytettiin normaalia FDR-valintamenetelmää, joka aiempien kokeiden perusteella näyttäisi toimivan hyvin myös geneettisen kytkennän aiheuttamien riippuvuuksien kanssa. Kynnysarvoksi asetettiin $q = 0.01$, jonka pitäisi riittää takaamaan korkeintaan muutaman prosentin hylkäysvirheaste vielä saman markkeriparin eri genotyyppejä koskevien p-arvojen yhdistämisen jälkeenkin.

NFBC-aineistosta haettiin sekä yksittäisten markkerien välisiä assosiaatioita (1–1), että korkeintaan kolmen markkerin kokoisten ryhmien välisiä assosiaatioita. Jälkimmäisessä tapauksessa siis kokeiltiin kaikkia mahdollisia ryhmäkokojen yhdistelmiä (1–1, 1–2, 1–3, 2–1, . . . , 3–3). Koska kaikkien tuloksena saatujen p-arvojen tallentaminen olisi vienyt liikaa tilaa (teratavuluokkaa jo pelkästään yksittäisten markkerien tapauksessa), tallennettiin vain p-arvot, jotka olivat pienempiä kuin 10^{-5}. Jos oletetaan, että lopullinen merkitsevyysraja on tätä pienempi, ei suurempien p-arvojen unohtaminen vaikuta valintaprosessien toimintaan.

Kokeet suoritettiin koneilla, jotka oli varustettu kahdella 2.83 GHz Intel Xeon neliydinsuorittimella. Tärkeimmät osat laskennasta oli kirjoitettu C-kielellä ja säikeistetty hyödyntämään kaikkia kahdeksaa ydintä. Yksittäisten markkerien välinen testaus vei aikaa noin 20 konetyöpäivää (5 päivää neljällä koneella) ja ryhmien välinen testaus noin 270 konetyöpäivää (reilut 12 päivää 22 koneella).

Yksittäisten markkerien väliset assosiaatiot

Yksittäisten markkerien välisillä assosiaatioilla suoritettujen testien kokonaismäärä oli $4.7 \cdot 10^{11}$. FDR-valinta kynnysarvolla 0.01 johti merkitsevyysrajaan $2.5 \cdot 10^{-9}$, jolla yhteensä 121 873 p-arvoa tuomittiin merkitseväksi. Erillisiä markkeripareja nämä muodostavat 36 143 kappaletta. Koska merkitsevä makkeripari sisältää keskimäärin noin 3.4 merkitsevää p-arvoa, pitäisi todellisen hylkäysvirheasteen siten olla

korkeintaan luokkaa 0.034.

Kuvassa 7 on havainnollistettu löydettyjen assosiaatioiden jakautumista kromosomien välille. Voidaan huomata, että assosiaatioita näyttäisi löytyvän melko tasaisesti kaikkien kromosomien väliltä. Silmiinpistävästi erottuva piirre on useimpien assosiaatioiden ryhmittyminen selkeiksi rivi- ja sarakelinjastoiksi. Jotkin markkerit näyttäisivät siten olevan vahvasti assosioituneita todella monien muiden eri kromosomeissa sijaitsevien markkerien kanssa. Toisaalta myös rivi-sarakekuvion väleissä sijaitsee useita yksinäisiä assosiaatioita. Seuraavaksi pyrimme luokittelemaan löydettyjä assosiaatioita.

Alipopulaatiot

Kuvitellaan tilanne, jossa käytettävissä oleva aineisto koostuisi kahdesta eri populaatiosta poimituista näytteistä. Nyt, jos näissä populaatioissa joidenkin markkerien alleelijakaumat ja/tai genotyyppijakaumat poikkeaisivat merkittävästi toisistaan, olisivat kyseiset markkerit kaikki keskenään assosioituneita koko aineistossa. Kuvan 7 kaltaisessa kartassa nämä assosiaatiot muodostaisivat ruudukkomaisen kuvion. Kyseisessä kuvassa näyttäisi käyvän juuri tällä tavalla – suurin osa assosiaatioista on sijoittunut yhteisille vaaka- ja pystysuuntaisille linjoille.

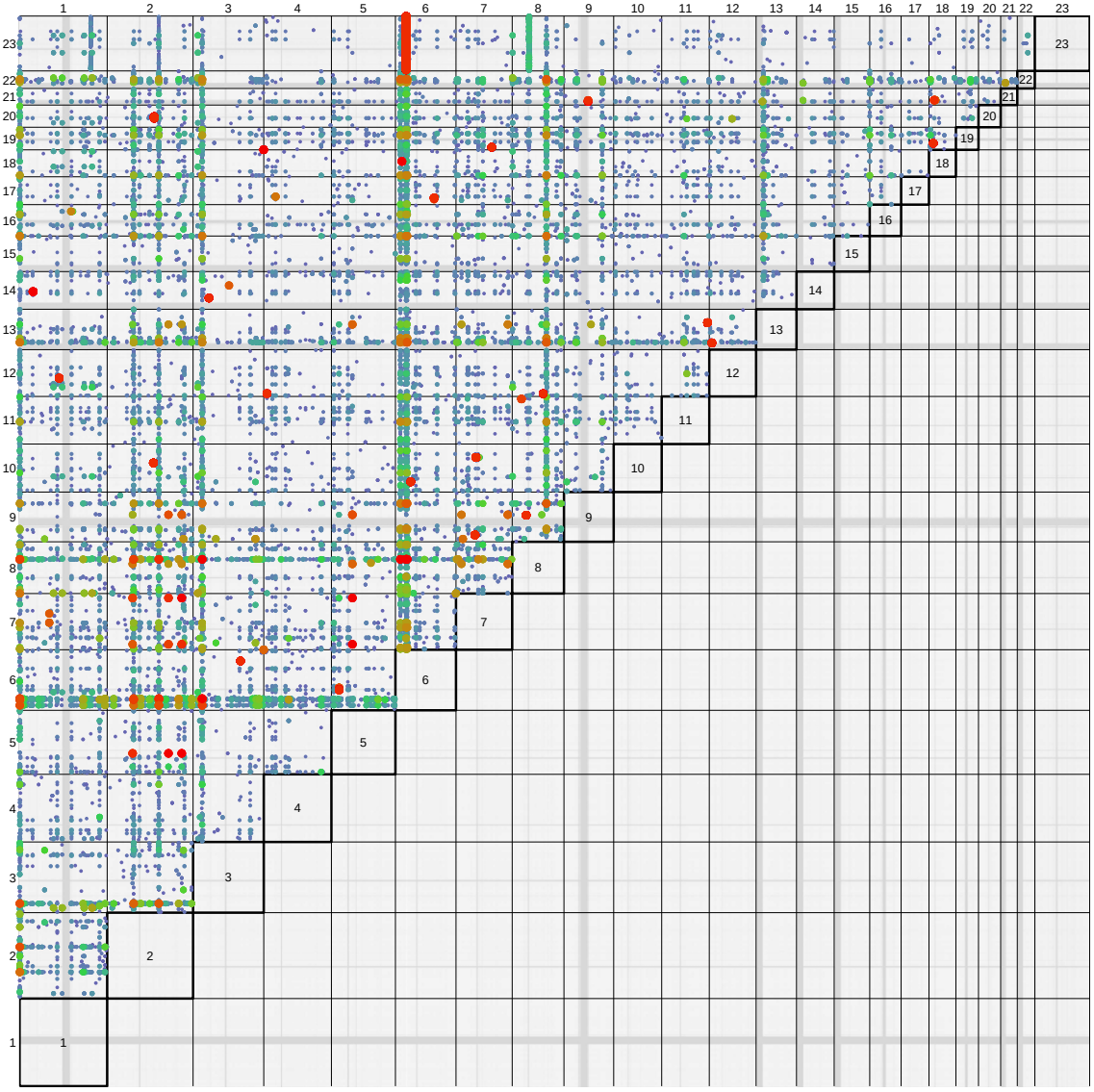
Suurin osa kuvan 7 linjamuodostelmista voidaan todella selittää mainitun kaltaisilla alipopulaatioilla. Nämä alipopulaatiot voidaan erottaa seuraavanlaisella karkealla menetelmällä:

1. Yhdistetään lähekkäin sijaitsevat merkitsevät assosiaatiot.

2. Olkoon k sellaisten merkitsevien genotyyppiassosiaatioiden määrä, joilla genotyypin frekvenssi on liian suuri, ja olkoon m näytteiden määrä. Luodaan $m \times k$ -binäärimatriisi, joka kohdassa (i, j) sisältää arvon 1, jos näytteellä i on assosiaatiota j vastaava genotyyppi.

3. Järjestetään muutamaan (esim. 10) kertaan vuorotellen matriisin rivit ja sarakkeet leksikografiseen järjestykseen suurimmasta pienimpään.

4. Matriisista on erotettavissa sarakeryhmiä, joissa tietyillä näytteillä esiintyy lähes yksinomaan ykköstä ja muilla lähes yksinomaan nollaa. Valitaan tällainen ryhmä sarakkeita. (Lopetetaan, jos selkeää sarakeryhmää ei voida erottaa.)



Kuva 7: Kuvaan on merkitty E-testillä ja FDR-valinnalla `NFBC`-aineistosta löydetyt merkitsevät yksittäisten markkerien väliset assosiaatiot. Diagonaalilla olevat neliöt ovat eri kromosomeja; X-kromosomia on merkitty numerolla 23. Kukin piste tarkoittaa vastaavalla kohdalla vaaka- ja pystysuunnassa sijaitsevista kromosomeista löytyvien markkerien välistä merkitsevää assosiaatiota. Pisteen suhteellinen sijainti kromosomien sisällä vastaa assosioituneiden markkereiden fyysistä sijoittumista kromosomeissa. Mitä suurempi ja punaisempi piste on, sitä vahvempi on kyseinen havaittu assosiaatio. Kartan taustalla oleva harmaan sävy kertoo kromosomien markkeritiheyden eri kohdissa. Tummanharmaat raidat vastaavat käytännössä sentromeereja, joiden alueelta markkereita ei ole juuri saatavilla.

5. Valitulle sarakeryhmälle etsitään ne näytteet, joilla ykkösten lukumäärä ylittää sopivasti valitun kynnysarvon. Nämä näytteet muodostavat alipopulaation.

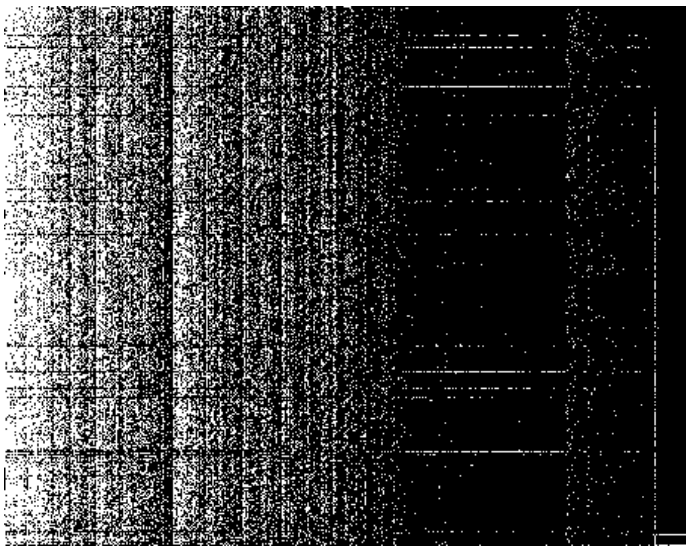
6. Käydään koko genotyyppiaineiston läpi, ja etsitään kaikki markkerit, joilla genotyyppien frekvenssit poikkeavat tilastollisesti merkitsevästi muodostetun alipopulaation ja muiden näytteiden välillä.

7. Poistetaan löydetyistä assosiaatioista ne, joilla ainakin toinen osapuoli kuuluu näihin alipopulaation kanssa assosioituneisiin markkereihin.

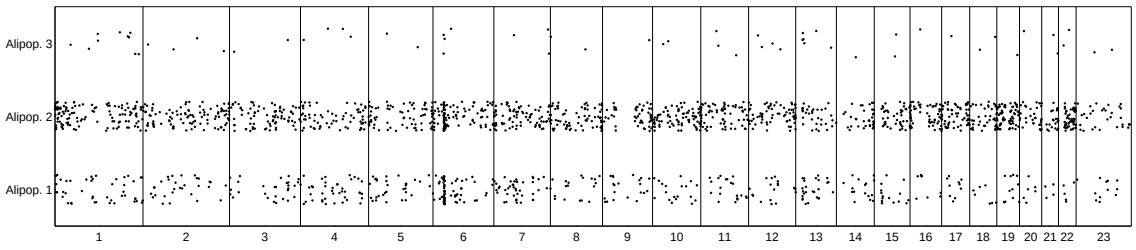
8. Jatketaan vaiheesta 2 seuraavan alipopulaation etsimiseksi.

Ensimmäisessä vaiheessa yhdistettiin aluksi kaksi alle kymmenen markkerin etäisyydellä toisistaan olevaa markkeria samaksi, jos ne kummatkin ovat osallisena vähintään yhdessä merkitsevässä assosiaatiossa. Tämän jälkeen kaikista markkeriryhmäparin assosiaatioista valittiin aina merkitsevimmän edustamaan koko assosiaatiojoukkoa. Päämääränä ensimmäisessä vaiheessa on assosiaatioiden määrää vähentämällä helpottaa seuraavia vaiheita ja estää pelkästään näennäisten alipopulaatioiden muodostaminen. Jos menetelmällä löydettävissä olevia selkeitä alipopulaatioita esiintyy, niin vaiheen 3 tuloksena saadaan kuvan 8 kaltainen matriisi. Esimerkkikuvan tapauksessa matriisin oikean puoliskon keskellä on suuri joukko peräkkäisiä sarakkeita, jotka muutaman näytteen kohdalla sisältävät pääosin arvoa 1 ja muilla näytteillä pääosin arvoa 0. Siis on olemassa muutama yksilö, joilla on (lähes) näitä kaikkia sarakkeita vastaavissa markkeripareissa tietty genotyyppiyhdistelmä, kun taas muilla yksilöillä vastaavissa markkeripareissa (lähes) aina jokin muu genotyyppiyhdistelmä. Vaiheessa 6 käytettiin genotyyppikohtaista tarkkaa testiä assosiaatioiden etsimiseen löydetyn alipopulaation ja aineiston markkerien välillä. Merkitsevät assosiaatiot valittiin FDR-menetelmällä käyttäen kynnysarvoa $q = 0.05$.

Yllä kuvatulla tavalla saatiin erotettua kolme selkeää alipopulaatiota, kooltaan 7, 82 ja 76 näytettä. Kuvaan 9 on merkitty näiden kanssa merkitsevästi assosioituneiden markkerien sijainnit kromosomeissa. Merkitsevästi assosioituneita markkereita alipopulaatioille löytyi 505, 1426 ja 62 kappaletta. Kun löytyneistä assosiaatioista poistetaan sellaiset, joilla ainakin toinen kuuluu näihin vajaaseen 2000 alipopulaatioiden kanssa assosioituneiden markkereiden joukkoon, ja FDR-valinta suoritetaan uudelleen, jää edelleen jäljelle 21 475 assosioitunutta markkeriparia. Lähes puolet aiemmin valituista assosiaatiosta liittyi siis kolmeen havaittuun osapopulaatioon. Kuvaan 12a on piirretty jäljelle jääneet riippuvuudet.



Kuva 8: Esimerkki tekstissä esitetyn, alipopulaatioiden erottamiseen käytetyn menetelmän vaiheessa 3 saatavasta genotyyppimatriisista, jonka rivit ja sarakkeet on järjestetty uudelleen. Valkoinen piste tarkoittaa arvoa 1, musta arvoa 0. Piste on siis valkoinen, mikäli sen riviä vastaavalla yksilöllä on sen saraketta vastaava genotyyppi. Matriisin oikean puoliskon keskellä on joukko genotyyppejä (sarakkeita), joita esiintyy vain muutamalla tietyllä yksilöllä. Kuva on skaalattu pienemmäksi valitsemalla matriisista mukaan vain joka viidestoista rivi ja sarake.



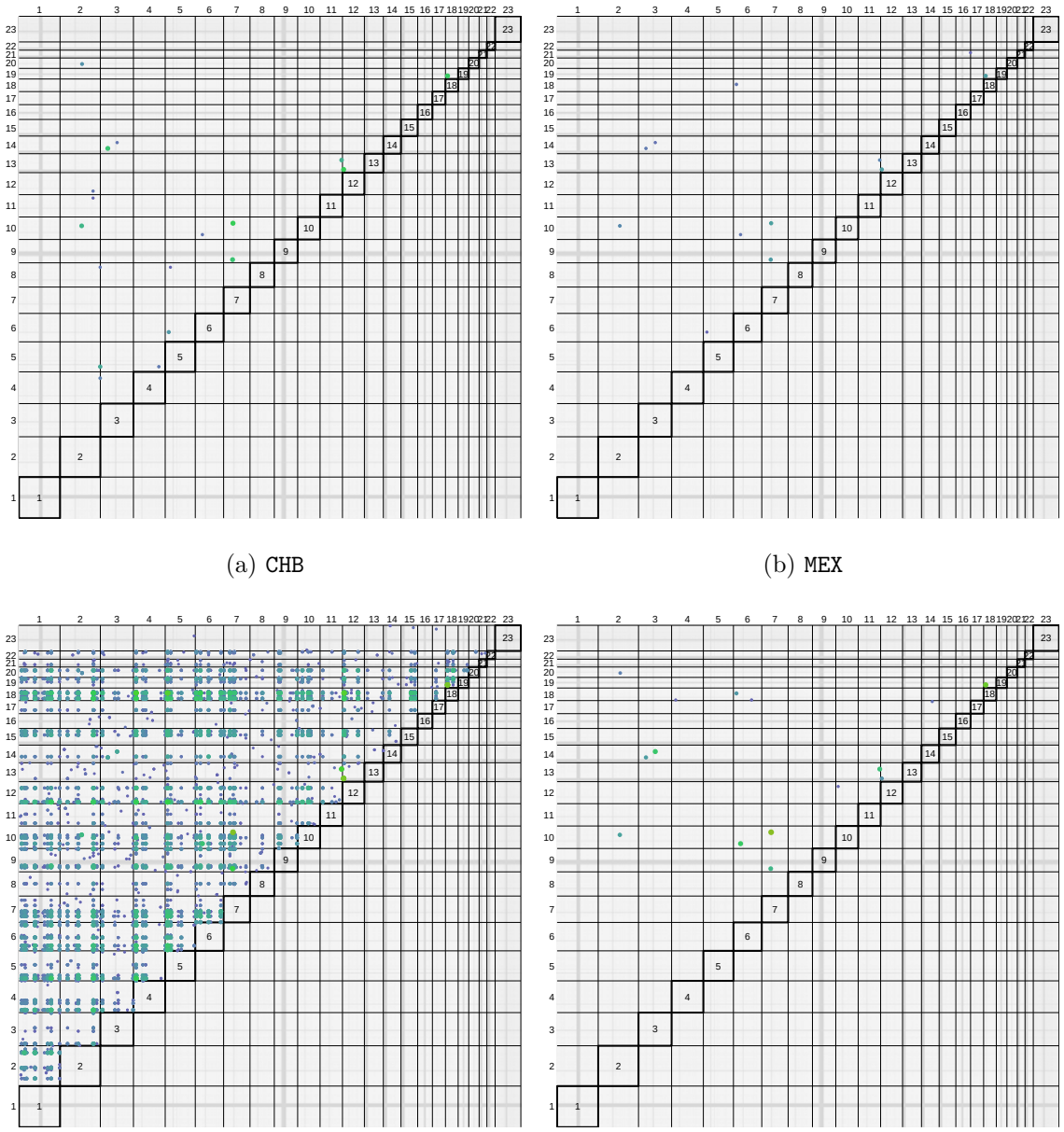
Kuva 9: Kolmen löydetyn alipopulaation kanssa assosioituneiden markkerien sijainnit eri kromosomeissa. X-akselilla on fyysinen sijainti perimässä, pystyviivat ovat kromosomien välisiä rajoja. Ryhmien sisällä Y-koordinaatilla ei ole merkitystä, satunnainen hajauttaminen on pelkästään esitystekninen keino. Kromosomin 6 alkupäässä olevat tihentymät sijaitsevat niin kutsutulla HLA-alueella, jossa aineiston markkeritiheys oli moninkertainen muihin paikkoihin verrattuna.

Tapamme poistaa alipopulaatioihin liittyvät assosiaatiot on jossain määrin ylimalkainen, ja se saattaa teoriassa poistaa myös alipopulaatioihin liittymättömiä havaintoja. Oikeaoppisempi tapa olisi poistaa vain sellaiset assosiaatiot, joiden kumpikin osapuolimarkkeri on assosioitunut saman alipopulaation kanssa. Koska poistettujen markkereiden osuus kaikista on vain noin 0.6 %, on satunnaisesti valitun markkeriparin todennäköisyys tulla poistetuksi alle 1.2 %. Ylimalkaisuudesta aiheutuva haitta on siten varsin pieni.

Alipopulaatioselitys on oikeastaan vain yksi tapa tarkastella havaittua ilmiötä. Vastaavasti voitaisiin myös väittää, ettei oikeastaan mitään alipopulaatiorakennetta ole olemassa, ja havaitut markkerit vain ovat jostain muusta syystä kaikki keskenään assosioituneita. Koska tapauksessamme tällaisia poikkeavia markkereita on suuri määrä – kymmeniä tai satoja – vaikuttavat alipopulaatiot kuitenkin luontevammalta selitykseltä. Alipopulaatioita voisivat muodostaa esimerkiksi sukulaiset tai muualta saapuneet ihmisryhmät.

Assosiaatiot HapMap-aineistossa

Yksittäisten kromosomien välisiä assosiaatioita etsittiin `NFBC`-aineiston lisäksi samalla menetelmällä myös mainituista neljästä HapMap-aineistosta. Kuvassa 10 on esitetty vastaavat löydettyjen assosiaatioiden sijoittumista kuvaavat kartat kullekin neljälle aineistolle. Kartoista voidaan tehdä muutamia mielenkiintoisia havaintoja. Ensinnäkin `MKK`-aineistolla ilmenee saman tyyppistä ruudukkokuviota kuin `NFBC`-aineistollakin, joten sekin sisältää luultavasti jonkinlaista alipopulaatiojakoa. Jostain syystä ilmiö ei näytä koskevan X-kromosomia. Kolmella muulla aineistolla puolestaan havaittujen assosiaatioiden lukumäärä on varsin pieni ja niiden sijainnit näyttävät toisistaan riippumattomilta. Huomattavaa on kuitenkin se, että suuri osa assosiaatioista näyttäisi olevan samoja kaikilla kolmella aineistolla. Testatuista `NFBC`- ja HapMap-aineistoista voidaankin mittaustavasta riippuen löytää noin 30–40 kaikille viidelle aineistolle yhteistä markkeriparien välistä assosiaatiota, tai noin 10, jos lähekkäin sijaitsevat assosiaatiot lasketaan yhdeksi. Havainto vaikuttaa erittäin mielenkiintoiselta. Valitettavasti nämä kaikki yhteiset assosiaatiot johtuvat kuitenkin melko varmasti markkereista, jotka on jostakin syystä sijoitettu väärään kromosomiin. Tätä virhesijoittelua käsittelemme seuraavaksi.



(a) CHB (b) MEX

(c) MKK (d) YRI

Kuva 10: Vastaavat assosiaatiokartat yksittäisten markkerien välisille assosiaatioille HapMap-aineistoissa (vertaa kuvaan 7).

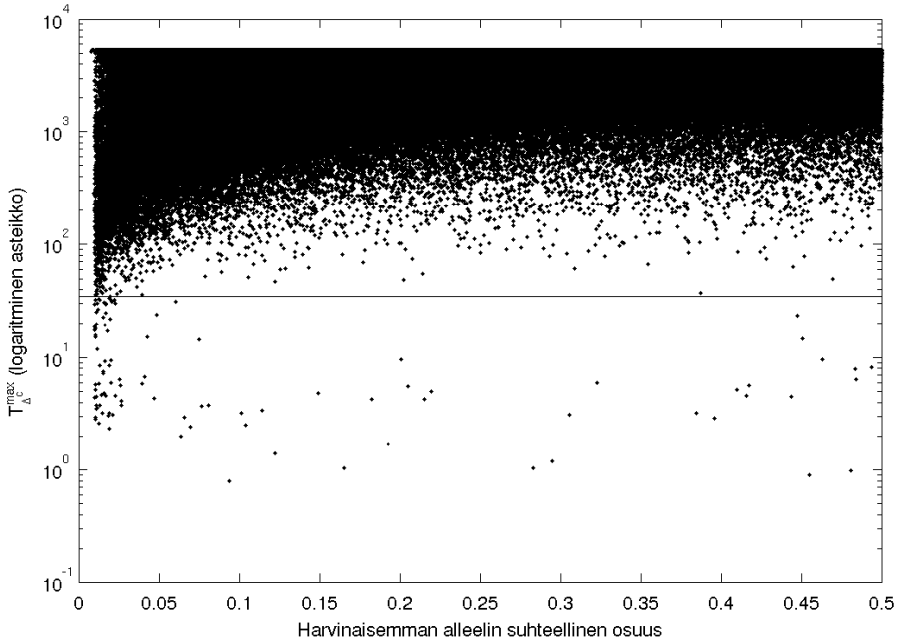
Virheellisesti sijoitetut markkerit

Kuten olemme todenneet, geneettinen kytkentää aiheuttaa tyypillisesti vahvan riippuvuuden kromosomissa lähekkäin sijaitsevien markkereiden välille. Mahdollisten omituisuuksien löytämiseksi mitattiin kullekin aineiston markkerille naapurikytkennän voimakkuutta laskemalla komposiittikorrelaation tunnusluvun T_{Δ^c} arvon kummaltakin puolelta kymmenen lähimmän markkerin kanssa ja valitsemalla saaduista lukemista suurimman. Merkitään tätä $T_{\Delta^c}^{\max}$. Koska T_{Δ^c} mittaa alleeliassosiaation merkitsevyyttä, on $T_{\Delta^c}^{\max}$ sitä suurempi, mitä merkitsevämpi riippuvuus markkerin ja sen naapuruston väliltä löytyy.

Kuvaan 11 on piirretty pisteinä kaikki `NFBC`-aineiston markkerit. Pystyakselille on sijoitettu $T_{\Delta^c}^{\max}$ ja vaaka-akselille harvinaisemman alleelin suhteellinen frekvenssi. Jos alleelijako on erittäin epätasapainoinen, ei merkittävää assosioitumista tapahdu yhtä helposti. Tämä näkyy vasemman reunan pienempinä $T_{\Delta^c}^{\max}$-arvoina. Suurimmalla osalla markkereista kytkentä naapureiden kanssa on kuitenkin suhteellisen selkeää. Muutamat markkerit sijoittuvat kuvassa epätavallisen alas, mikä tarkoittaa, että kytkentää naapureiden kanssa ei ilmene. Kytkennän kynnysarvoksi valittiin 35 (kuvassa vaakaviiva), mikä vastaa korjaamatonta p-arvoa $2.5 \cdot 10^{-10}$, kun nollahypoteesina on riippumattomuus naapureiden kanssa. Koska markkereita testattiin noin 340 000 ja kutakin markkeria kohden 20 naapuria, on vastaava Bonferroni-korjattu p-arvo $3.3 \cdot 10^{-9} \cdot 340000 \cdot 20 = 0.022$.

Yhteensä 103 markkerin maksiminaapuriassosiaatio jää valitun kynnysarvon alapuolelle. Erityisesti aiemmin havaituilla kaikissa viidessä aineistossa assosioituneilla markkeripareilla jokaisella toinen toinen markkeri sijoittuu kynnysviivan alapuolelle. Näille markkereille siis assosioituminen naapurimarkkereiden kanssa on olematonta ja assosioituminen jonkin muun sijainnin kanssa epätavallisen voimakasta. Näyttäisi selvältä, että markkerit on jostain syystä sijoitettu väärään paikkaan. Päätelmää vahvistaa haku NCBI:n ylläpitämästä SNP-tietokannasta[3]: suunnilleen puolet todennäköisesti väärin sijoitelluista markkereista on tietokannassa sijoitettu paikkaan, jonka kanssa selkeää riippuvuutta on havaittavissa. Oletettavasti myös loput ovat väärin sijoiteltuja ja tieto sijainnista tietokannassa ei pidä paikkansa.

[3]Yhdysvaltojen NCBI:n (The National Center for Biotechnology Information) ylläpitämään dbSNP-tietokantaan voi tehdä kyselyjä web-käyttöliittymän kautta osoitteessa `http://www.ncbi.nlm.nih.gov/projects/SNP/`.



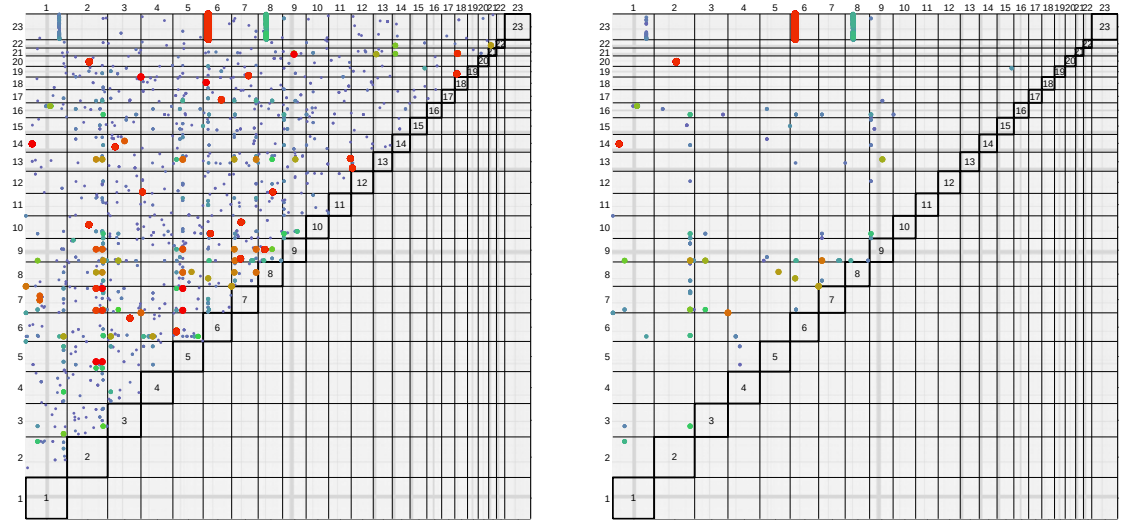
Kuva 11: Naapurikytkennän merkitsevyys markkereilla. Mitä korkeampana piste sijaitsee, sitä merkitsevämpi kytkentä markkerilla löytyy lähimpien naapureiden kanssa. Kuvaan on piirretty valittu merkitsevyysraja kohtaan $T_{\Delta^c}^{\max} = 35$.

Muita syitä

Varsin pieni osa havaituista assosiaatioista – muutama kymmenen kappaletta — liittyy todennäköisesti väärin sijoiteltuihin markkereihin. Vaikka alipopulaatioihin liittyvien assosiaatioiden lisäksi poistetaan joukosta sellaiset markkeriparien assosiaatiot, joiden toinen markkeri näyttäisi olevan väärässä paikassa, jää assosiaatioita edelleen suuri suuri määrä jäljelle.

Suurin osa, yli 80 %, jäljelle jääneistä liittyy kromosomiin X. Nämä assosiaatiot muodostavat kuvassa 12a kolme X-kromosomin läpi ulottuvaa pystysuuntaista raitaa. Kromosomeissa 1, 6 ja 8 näyttäisi siten kussakin olevan kohta, joka on merkitsevästi assosioitunut koko X-kromosomin kanssa. Tämä viittaisi markkerin arvon riippuvuuteen sukupuolesta. Voidaan havaita, että kaikkien kolmen kohdalta asia myös on näin; markkerit rs3767423 (kromosomissa 1), rs6917603 (kromosomissa 6) ja rs6989593 (kromosomissa 8) ovat selvästi assosioituneita sukupuolen kanssa. Taulukoissa 7a ja 7b on näistä kahdelle ensimmäiselle muodostettu ristiintaulut sukupuolen kanssa.

Sukupuoleen liittyvä suuri assosiaatiomäärä aiheuttaa FDR-valinnan kanssa ikä-



(a) Alipopulaatioihin liittyvät assosiaatiot poistettu ja merkitsevien valinta tehty FDR-menetelmällä.

(b) Alipopulaatioihin sekä väärin sijoiteltuihin markkereihin liittyvät assosiaatiot poistettu ja merkitsevien valinta tehty Bonferroni-menetelmällä.

Kuva 12: Karsitut assosiaatiokartat `NFBC`-aineistolle.

vän sivuilmiön ja vetää mukanaan joukon luultavasti todellisuudessa riippumattomia markkeripareja. Ongelman kiertämiseksi kuvassa 12b on käytetty Bonferroni-valintaa kynnysarvolla $\alpha = 0.05$ testituloksiin, joista on karsittu alipopulaatioihin ja väärin sijoiteltuihin markkereihin liittyvät assosiaatiot. Jäljelle jää kromosomiin X liittyvien assosiaatioiden lisäksi 526 assosioitunutta markkeriparia. Jos alle 100 markkerin päässä olevat assosiaatioiden osapuolet yhdistetään, jää jäljelle 74 erillistä riippuvuutta. Kuvasta nähdään, että edelleen osa assosiaatioista muodostaa pysty- ja vaakasuuntaisia linjoja ja niiden perusteella voisi varmastikin muodostaa ainakin yhden alipopulaation lisää. Toisaalta, jos ryhmän genotyyppifrekvenssit poikkeavat valtavirrasta vain muutaman markkerin kohdalla, voi alipopulaatiosta puhumisen mielekkyyden kyseenalaistaa. Joka tapauksessa myös linjojen ulkopuolelle jää muutamia assosiaatiopisteitä. Taulukoissa 7c ja 7d on kaksi esimerkkiä tällaisesta assosiaatiosta kromosomien 5 ja 8 sekä vastaavasti kromosomien 2 ja 20 välillä. Kumpaankin esimerkkiin on valittu markkeripari, joka tuottaa vahvimman assosiaation kyseisten kromosomien välillä.

Taulukkoon 8 on koottu yhteenvetona assosioituneiden markkeriparien lukumäärien jakautuminen eri selitysten kesken. Toisin kuin aiemmin tekstissä, taulukon arvoissa FDR-valintaa ei ole suoritettu uudelleen aina selityksien eliminoimisen jälkeen.

Taulukko 7: Esimerkkejä jäljelle jäävistä assosiaatioista. Markkerin tunnuksen jälkeen on suluissa kromosomin numero.

(a) sukupuoli ja rs3767423 (1)

sukupuoli	rs3767423			
	$\frac{B}{B}$	$\frac{B}{b}$	$\frac{b}{b}$	$*$
mies	0	787	1799	2586
nainen	103	781	1893	2777
$*$	103	1568	3692	5363

(b) sukupuoli ja rs6917603 (6)

sukupuoli	rs6917603			
	$\frac{B}{B}$	$\frac{B}{b}$	$\frac{b}{b}$	$*$
mies	210	2317	3	2530
nainen	176	1	2592	2769
$*$	386	2318	2595	5299

(c) rs30341 (5) ja rs4150895 (8)

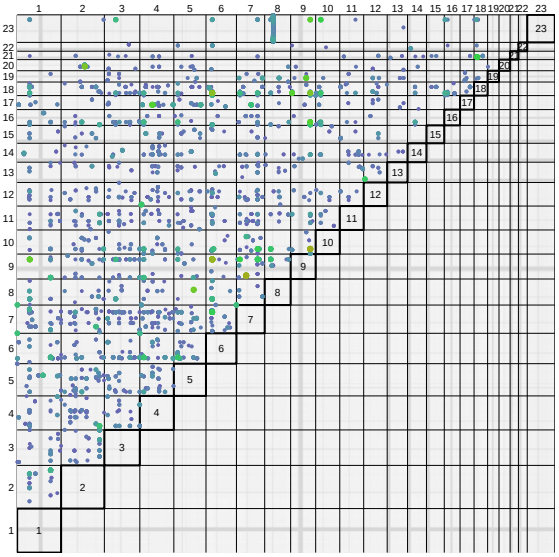
rs30341	rs4150895			
	$\frac{B}{B}$	$\frac{B}{b}$	$\frac{b}{b}$	$*$
$\frac{A}{A}$	0	3	1118	1121
$\frac{A}{a}$	0	117	2410	2527
$\frac{a}{a}$	6	317	1303	1626
$*$	6	437	4831	5274

(d) rs4082157 (2) ja rs6115178 (20)

rs4082157	rs6115178			
	$\frac{B}{B}$	$\frac{B}{b}$	$\frac{b}{b}$	$*$
$\frac{A}{A}$	33	31	7	71
$\frac{A}{a}$	38	803	274	1115
$\frac{a}{a}$	7	348	3818	4173
$*$	78	1182	4099	5359

Taulukko 8: Löydettyjen assosioituneiden markkeriparien lukumäärät eri selityksien mukaan jaoteltuna. FDR-valinnassa on käytetty kynnysarvoa $q = 0.01$ ja Bonferroni-valinnassa vastaavasti kynnysarvoa $\alpha = 0.05$.

Selitys	FDR-valinta	Bonferroni-valinta
Alipopulaatiot	14235	5708
Väärin sijoitettu markkeri	2589	1612
Sukupuoleen liittyvä	17635	14699
Ei selitystä	1684	526
Yhteensä	36143	22545



Kuva 13: Assosiaatiokartta korkeintaan yhden, kahden ja kolmen markkerin ryhmille. Kaikki yksittäisten markkerien väliset assosiaatiot sekä niiden päälle osuvat ryhmien assosiaatio poistettu.

Markkeriryhmien väliset assosiaatiot

Markkeriryhmien välisiä assosiaatioita haettiin yhden, kahden ja kolmen markkerin ryhmillä. Eriasteisesta ryhmittelystä aiheutuu luonnollisesti huomattavasti päällekkäisyyttä testien kesken. Tämä päällekkäisyys aiheuttaa entistä enemmän ongelmia merkitsevyysvalinnan kannalta. Tästä syystä päädyttiinkin käyttämään konservatiivisesti Bonferroni-korjausta kynnysarvolla $\alpha = 0.05$, jotta väärien valintojen riski olisi mahdollisimman alhainen.

Markkeriryhmien tapauksessa kiinnostavinta on nähdä, löytyykö niiden avulla assosiaatioita, joita yksittäisten markkerien tapauksessa ei havaita. Merkitseviksi valittujen joukosta poistettiin siis kaikki aiemmin FDR-menetelmällä valitut yksittäisten markkereiden väliset assosiaatiot sekä näiden kanssa päällekkäin menevät markkeriryhmien assosiaatiot. Jäljelle jää suuri joukko valintoja, jotka ovat näkyvissä kuvassa 13. Jonkin asteista linjoihin järjestäytymistä on edelleen havaittavissa. Joka tapauksessa markkereita ryhmittelemällä on selkeästi mahdollista löytää assosiaatioita, jotka eivät erotu tarpeeksi voimakkaasti yksittäisten markkereiden välisissä testeissä.

5 Yhteenveto

Kromosomissa lähekkäin sijaitsevien markkereiden välisen sekä markkereiden ja piirteiden välisen tilastollisen riippuvuuden voimakkuutta on mitattu lukuisissa eri tutkimuksissa. Tyypillisesti näissä on keskitytty alleeliassosiaatioon, mikä usein vaatii, että näytteiden haplotyypit ovat tiedossa. Koska useimmiten saatavilla olevat markkeriaineistot sisältävät genotyyppejä, täytyisi haplotyypit yrittää päätellä tilastollisesti. Tässä tutkielmassa keskityttiin assosiaation mittaamiseen suoraan genotyypeistä. Tällaisessa testauksessa pitäisi näkyä paitsi alleelien väliset assosiaatiot, myös puhtaasti genotyyppeihin liittyvät assosiaatiot, jos sellaisia on olemassa. Suhteellisen hyvin tunnetun, geneettisestä kytkennästä johtuvan toisiaan lähellä olevien alleelien assosioitumisen sijaan mittasimme riippuvuutta eri kromosomeissa olevien markkereiden välillä. Koska kromosomit periytyvät riippumattomasti, teoriassa minkäänlaista assosiaatiota niiden välillä ei pitäisi ilmetä.

Alleeliassosiaation mittaamiseen kehitettyjä lukuisia testejä ja testisuureita ei yleensä voi helposti yleistää genotyypeille. Tästä syystä vertailimme pääasiassa yleisiä kahden kategorisen muuttujan välisiä riippuvuustestejä. Valintamme mahdollistaa samojen testien käyttämisen yksittäisten markkerien sijaan suoraan markkeriryhmien välisten assosiaatioiden etsimiseen. Vertaillut testimenetelmät voidaan jakaa kolmeen ryhmään: tarkan p-arvon kertovat testit, tunnettuun testisuureen asymptoottiseen jakaumaan perustuvat testit sekä satunnaista näytteenottoa käyttävät testit. Näytteenotto on näistä joustavin ja mahdollistaa mielivaltaisen testisuureen valinnan, mutta testiaineiston suuresta koosta ja p-arvoilta vaaditusta suuresta tarkkuudesta johtuen siihen perustuvat menetelmät ovat tapauksessamme liian hitaita. Pienemmällä aineistolla ne ovat kuitenkin varteenotettava vaihtoehto. Asymptoottiset khii-toiseen-testi ja G-testi ovat yleisesti käytettyjä tilastollisen riippuvuuden mittareita, jotka on erittäin nopea laskea. Harvat ristiintaulut kuitenkin aiheuttavat tunnetusti niille ongelmia, ja etenkään markkeriryhmien tapaukseen ne eivät kokeiden mukaan sellaisenaan sovellu. Ainakin G-testi vaikuttaisi kylläkin toimivan suhteellisen hyvin yksittäisten markkerien tapauksessa käytössä olleen aineiston suhteellisen suurella reilun 5000 henkilön näytekoolla. P-arvojen tarkka laskeminen vaatii yleisesti eksponentiaalisen ajan, mutta sopivalla tunnusluvun valinnalla myös siitä voidaan saada nopeaa. Esittelimme kaksi tarkkaa testiä: maksimipoikkeamatestin ja genotyyppikohtaisen tarkan testin. Koko käytössä olleen `NFBC`-aineiston laajuiseen assosiaatioiden hakuun valittiin genotyyppikohtaisen tarkka testi, joka on samaa nopeusluokkaa khii-toiseen ja G-testien kanssa, mutta toimii suhteellisen hy-

vin myös harvoilla ristiintauluilla. Tilastollisesti merkitsevien p-arvojen valintaan sovellettiin suosittua FDR-valintamenetelmää sekä joissakin tapauksissa myös perinteistä Bonferroni-korjausta.

Kokeissa käytetty testimenetelmä löysi suuren määrän eri kromosomeissa sijaitsevien markkereiden välisiä assosiaatioita. Löydöksistä vajaa puolet oli selitettävissä aineiston sisäisellä alipopulaatiorakenteella, ja yli puolet muutamalla sukupuolen – ja siten käytännössä koko X-kromosomin – kanssa assosioituneella markkerilla. Näiden lisäksi jäljelle jäi pienempi joukko assosiaatioita, joista osa vaikuttaisi olevan seurausta aineistossa väärään paikkaan sijoitetuista markkereista. Joillekin assosiaatioille ei minkäänlaista selitystä löytynyt. Mahdollisia muita selityksiä näille voisivat olla virheet genotyyppauksessa eli genotyyppien lukemissa, aineiston sisäiset sukulaisuussuhteet, luonnonvalinta sekä jokin tuntematon biologinen mekanismi. Alipopulaatioihin sekä sukupuoleen liittyvien assosiaatioiden selityksetkin ovat vain osittaisia. Avoin kysymys onkin myös, mistä näiden kaltainen rakenteisuus perimmiltään johtuu.

Lähteet

Agr02 Agresti, A. *Categorical Data Analysis*, luku 3 Inference for Contingency Tables. Wiley series in probability and statistics. John Wiley & Sons, toinen painos, 2002.

AWB79 Agresti, A., Wackerly, D. ja Boyett, J. M., Exact conditional tests for cross-classifications: Approximation of attained significance levels. *Psychometrika*, 44,1(1979), sivut 75–83.

BB07 Browning, B. L. ja Browning, S. R., Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet. Epidemiol.*, 31,5(2007), sivut 365–375.

BC91 Besag, J. ja Clifford, P., Sequential Monte Carlo p-values. *Biometrika*, 78,2(1991), sivut 301–304.

BH95 Benjamini, Y. ja Hochberg, Y., Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the*

Royal Statistical Society. Series B (Methodological), 57,1(1995), sivut 289–300.

BH00 Benjamini, Y. ja Hochberg, Y., On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25,1(2000), sivut 60–83.

BKY06 Benjamini, Y., Krieger, A. M. ja Yekutieli, D., Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93,3(2006), sivut 491–507.

BY01 Benjamini, Y. ja Yekutieli, D., The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29,4(2001), sivut 1165–1188.

CM98 Collins, A. ja Morton, N. E., Mapping a disease locus by allelic association. *Proc. Natl. Acad. Sci. U.S.A.*, 95,4(1998), sivut 1741–1745.

Col07 Collins, A. R., Linkage disequilibrium and association mapping: An introduction. Teoksessa *Linkage Disequilibrium and Association Mapping*, Collins, A. R., toimittaja, osa 376 sarjasta *Methods in Molecular Biology*, Humana Press, 2007, sivut 1–15.

Con03 Consortium, T. I. H., The international HapMap project. *Nature*, 426,6968(2003), sivut 789–796.

DG08 Dudbridge, F. ja Gusnanto, A., Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.*, 32,3(2008), sivut 227–234.

DH97 Davison, A. C. ja Hinkley, D. V. *Bootstrap methods and their applications*, luku 4 Tests, sivut 136–190. Cambridge University Press, 1997.

DR95 Devlin, B. ja Risch, N., A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29, sivut 311–322.

EMMC06 Evans, D. M., Marchini, J., Morris, A. P. ja Cardon, L. R., Two-stage two-locus models in genome-wide association. *PLoS Genetics*, 2,9(2006), sivut 1424–1432.

Fis22 Fisher, R. A., On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85,1(1922), sivut 87–94.

HCDI+08 Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C. ja Balding, D. J., Genome-wide significance for dense SNP and resequencing data. *Genet. Epidemiol.*, 32,2(2008), sivut 179–185.

Hed05 Hedrick, P. W., *Genetics of populations*. Jones & Bartlett Publishers, kolmas painos, 2005.

Hol79 Holm, S., A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6,2(1979), sivut 65–70.

KFZ08 Kim, Y., Feng, S. ja Zeng, Z.-B., Measuring and partitioning the high-order linkage disequilibrium by multiple order Markov chains. *Genet. Epidemiol.*, 32,4(2008), sivut 301–312.

KL08 Kooperberg, C. ja Leblanc, M., Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet. Epidemiol.*, 32,3(2008), sivut 255–263.

KRB+89 Kerem, B., Rommens, J., Buchanan, J., Markiewicz, D., Cox, T., Chakravarti, A., Buchwald, M. ja Tsui, L., Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245,4922(1989), sivut 1073–1080.

KSM+08 Kustra, R., Shi, X., Murdoch, D. J., Greenwood, C. M. T. ja Rangrej, J., Efficient p-value estimation in massively parallel testing problems. *Biostatistics (Oxford, England)*, 9,4(2008), sivut 601–612.

Lev81 Levin, B., A representation for multinomial cumulative distribution functions. *The Annals of Statistics*, 9,5(1981), sivut 1123–1126.

MCP+06 Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z. S., Munro, H. M., Abecasis, G. R. ja Donnelly, P., A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.*, 78,3(2006), sivut 437–450.

MDC05 Marchini, J., Donnelly, P. ja Cardon, L. R., Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, 37,4(2005), sivut 413–417.

Mor07 Morton, N. E., A history of association mapping. Teoksessa *Linkage Disequilibrium and Association Mapping*, osa 376, 2007, sivut 17–21.

MSL+07 Musani, S. K., Shriner, D., Liu, N., Feng, R., Coffey, C. S., Yi, N., Tiwari, H. K. ja Allison, D. B., Detection of gene x gene interactions in genome-wide association studies of human population data. *Hum. Hered.*, 63,2(2007), sivut 67–84.

NFR02 Nothnagel, M., Fürst, R. ja Rohde, K., Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum. Hered.*, 54,4(2002), sivut 186–198.

PGC+05 Petkov, P. M., Graber, J. H., Churchill, G. A., DiPetrillo, K., King, B. L. ja Paigen, K., Evidence of a large-scale functional organization of mammalian chromosomes. *PLoS Genetics*, 1,3(2005), sivut 0312–0322.

RR95 Raymond, M. ja Rousset, F., An exact test for population differentiation. *Evolution*, 49,6(1995), sivut 1280–1283.

RYB03 Reiner, A., Yekutieli, D. ja Benjamini, Y., Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19,3(2003), sivut 368–375.

SD03 Stephens, M. ja Donnelly, P., A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, 73,5(2003), sivut 1162–1169.

Sev04 Sevon, P., *Algorithms for Association-Based Gene Mapping*. Väitöskirja, Helsingin yliopisto, Tietojenkäsittelytieteen laitos, 2004. Raportti A-2004-4.

Sha95 Shaffer, J. P., Multiple hypothesis testing. *Annual Review of Psychology*, 46, sivut 561–584.

SR96 Strachan, T. ja Read, A. P., *Human molecular genetics*. BIOS Scientific Publishers, 1996.

TNH+07 Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E. ja Visscher, P. M., Recent human effective population size estimated from linkage disequilibrium. *Genome Res.*, 17,4(2007), sivut 520–526.

VSM05 Verhoeven, K. J., Simonsen, K. L. ja McIntyre, L. M., Implementing false discovery rate control: increasing your power. *Oikos*, 108,3(2005), sivut 643–647.

Wei79 Weir, B., Inferences about linkage disequilibrium. *Biometrics*, 35,1(1979), sivut 235–254.

Wil35 Wilks, S. S., The likelihood test of independence in contingency tables. *The Annals of Mathematical Statistics*, 6,4(1935), sivut 190–196.

YB99 Yekutieli, D. ja Benjamini, Y., Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82,1-2(1999), sivut 171–196.

Zar99 Zar, J. H. *Biostatistical analysis*, luku 23 Contingency Tables. Prentice Hall, neljäs painos, 1999.

Zay04 Zaykin, D. V., Bounds and normalization of the composite linkage disequilibrium coefficient. *Genet. Epidemiol.*, 27,3(2004), sivut 252–257.

ZPW08 Zaykin, D. V., Pudovkin, A. ja Weir, B. S., Correlation-based inference for linkage disequilibrium with multiple alleles. *Genetics*, 180,1(2008), sivut 533–545.