

GEENIKARTOITUSOPAS

VESA OLLIKAINEN JA PEKKA UIMARI
CSC

Geenikartoitusopas

Vesa Ollikainen ja Pekka Uimari

CSC – Tieteellinen laskenta Oy

Tämän teoksen tekijänoikeudet kuuluvat CSC – Tieteellinen laskenta Oy:lle. Teoksen tai osia siitä voi kopioida ja tulostaa vapaasti henkilökohtaiseen käyttöön sekä Suomen yliopistojen ja korkeakoulujen kurssikäyttöön edellyttäen, että kopioon tai tulosteeseen liitetään tämä ilmoitus teoksen tekijästä ja tekijänoikeuksista. Teosta ei saa myydä tai sisällyttää osaksi muita teoksia ilman CSC:n lupaa.

© Kirjoittajat ja
CSC – Tieteellinen laskenta Oy
2001–2004

ISBN 952-5520-00-5

<http://www.csc.fi/oppaat/geenikartoitus/>

Esipuhe

Tämä on uudistettu versio CSC:n vuonna 2001 sähköisesti julkaisemasta Geenikartoitusoppaasta.

Tässä oppaassa esitellään geenikartoituksen keskeisiä menetelmiä sekä opastetaan niiden soveltamiseen CSC:n laskentaympäristössä. Painopiste on ihmisen monitekijäisten sairauksien taustalla olevien alttiusgeenien kartoittamisessa. Oppaassa keskitytään sekä kytkentä- että assosiaatioanalyysin menetelmiin ja siinä tarkastellaan erityisesti tapauksia, joissa kiinnostuksen kohteena oleva fenotyyppi (sairauden aste) voidaan ilmaista kaksiarvoisella muuttujalla.

Opas on tarkoitettu henkilöille, jotka tarvitsevat geenikartoituksen laskentamenetelmiä työssään. Teksti alkaa alkeista, joten syvälinen tietämys aihepiiristä ei ole tarpeen.

Teoksen sisältöön ja ulkoasuun liittyvät kommentit ja ehdotukset voi toimittaa CSC:lle sähköpostiosoitteeseen vesa.ollikainen@csc.fi.

Haluamme kiittää Jarno Tuimalaa arvokkaista kommentista sekä oppaan kieliasun tarkistamisesta.

Espoossa 10.12.2004

Tekijät

Sisältö

Esipuhe	5
1 Johdanto	8
2 Geenikartoituksen lähestymistavat	10
2.1 Monogeeniset ja monitekijäiset sairaudet	10
2.2 Ihmisen genomi, markerit ja fenotyypit	11
2.3 Sukupuut ja periytyminen	12
2.4 Lähestymistavat ja geenikartoitusohjelmistot	15
3 Aineistojen esitys tietokoneella	17
3.1 Linkage-tiedostomuoto	17
3.1.1 Linkage-sukupuutiedoston rakenne	18
3.1.2 Sukupuutiedoston tarkistaminen Pedcheck-ohjelmalla	20
3.1.3 Linkage-parametritiedosto ja sen luonti	22
3.2 Tiedostomuotojen muunnokset: Mega2	27
4 Tilastollisia näkökulmia	31
4.1 Geenikartoituksen tavoitteet ja hypoteesit	31
4.2 Tilastollinen merkitsevyys	32
4.3 Tilastollinen voima	32
4.4 Moninkertaisen testauksen ongelma	33
5 Parametrinen kytkentäanalyysi	35
5.1 Taustaa	35
5.1.1 Hypoteesien määrittäminen	36
5.1.2 Uskottavuusfunktion muodostaminen	37
5.1.3 Parametrien estimointi	39
5.1.4 Hypoteesien testaus	39
5.1.5 Tilastollisten johtopäätösten muodostaminen	40
5.2 Analyysin nopeudesta	40
5.3 Analyysi Genehunter-ohjelmalla	41
5.4 Analyysi Linkage-ohjelmistolla	43
5.4.1 Sukupuutiedoston muuntaminen Makeped-ohjelmalla	44
5.4.2 Kaksipisteanalyysi	45
5.4.3 Monipisteanalyysi	52
6 Ei-parametrinen kytkentäanalyysi	54
6.1 Taustaa	54

6.1.1	Sairaiden sisarusparien menetelmä	54
6.1.2	Menetelmän laajennukset	56
6.2	Genehunter-ohjelman käyttö	58
6.2.1	Kytkentäanalyysi Genehunter-ohjelmassa	58
6.2.2	Genehunterin NPL-testit	60
6.2.3	Tiedostojen rakenteet	60
6.2.4	Genehunter-ohjelmiston käskyt	60
6.2.5	Komennot ASP-menetelmälle	69
6.3	Analyysi Merlin-ohjelmalla	70
6.3.1	Syötetiedostojen tarkastelu	70
6.3.2	Kytkentäanalyysin käynnistys	71
6.4	Suurten sukupuiden analysointi: Simwalk	73
6.4.1	Simwalk-analyysit Mega2-ohjelman avulla	73
6.4.2	Analyysin ohjaaminen manuaalisesti	77
7	Kytkentäepätasapaino ja assosiaatioanalyysi	79
7.1	Taustaa	79
7.1.1	Assosiaatioanalyysi yksilötason χ^2 -testillä	81
7.2	Periytyksen epätasapainotesti (TDT)	82
7.2.1	Binomijakaumaan perustuva eksakti testi	84
7.2.2	TDT-testisuureen approksimointi	85
7.2.3	Analyysi Genehunter-ohjelmalla	86
7.3	Assosiaatioanalyysin tilastollisen voiman arviointi	89
8	Haplotyyppien rakentaminen ja analyysi	93
8.1	Periytyksen epätasapainotesti haplotyypeille	93
8.2	Perhepohjainen haplotyyppaus	95
8.2.1	Haplotyyppaus Merlin-ohjelmalla	96
8.2.2	Tapaus-verrokkitiedoston rakentaminen	97
8.3	Assosioituvien haplotyyppien etsintä	97
8.4	Väestöpohjainen haplotyyppaus	100
8.5	Suuret SNP-aineistot ja haplotyyppiblokit	103
8.5.1	HapMap-projekti	104
8.5.2	Haploview-ohjelma	105
9	Analyyysin automatisointi Unix-ympäristössä	111
9.1	Tiedostojen automaattinen syöttäminen	112
9.2	Tulosten järjestäminen	112
9.3	Empiirisen p-arvon määrittäminen kytkentätulokselle	114
	Liitteet	117
A	Kirjoittajien yhteystiedot	118
B	Tietoa CSC:stä	119
	Kirjallisuutta	120
	Hakemisto	122

1 Johdanto

Geenikartoituksen tavoitteena on löytää tilastollisia yhteyksiä yksilöiden perimässä olevien varianttien ja näiden yksilöiden ilmiäsuun välillä. Geenikartoitus on haastava tilastollinen ja tietojenkäsittelytieteellinen ongelma.

Tässä oppaassa tarkastellaan geenikartoitusta sovellettuna ihmisen perinnöllisten sairauksien alttiusgeenien löytämiseen. Painopiste on alusta loppuun sellaisissa sairauksissa, joissa tarkasteltava ilmiäsuu voidaan esittää kaksiarvoisen muuttujan avulla. Tällöin jokaisesta tutkimukseen osallistuvasta ihmisestä voidaan periaatteessa sanoa, sairastaako hän tarkasteltavaa sairautta vai ei. Käytännössä tällainen lähtökohta valitaan useimmissa ihmisen sairausgeenien kartoitusprojekteissa, sillä sairauden asteen tarkempi mittaaminen on usein paitsi vaikeaa myös moniselitteistä.

Ihmisen sairausgeenien kartoitus on paitsi laskennallisesti haastava myös palkitseva ongelma, jolla on tärkeitä lääketieteellisiä sovelluksia: sairauksien taustalla olevien geenien tunteminen edistää niiden diagnostiikkaa ja hoitoa. Kuvattavat menetelmät ovat toki sovellettavissa ja laajennettavissa muihinkin organismeihin ja piirteisiin. Erityisesti kasvi- ja eläinjalostuksessa painopiste kohdistuu kuitenkin jatkuva-arvoisiin ominaisuuksiin, jotka jätetään tässä oppaassa tarkastelun ulkopuolelle. Ihmisen sairausgeenien kartoituksen sekä kasvi- ja eläingenetiikan välillä on toki muitakin merkittäviä eroja, jotka liittyvät koeasetelmiin: siinä missä vaikkapa kasvinjalostuksessa voidaan luoda räätelöityjä risteytysasetelmia, on ihmisen kohdalla tyytyminen havaittuihin, todellisista sukupuista ja väestöistä kerättäviin aineistoihin.

Ongelman haasteellisuutta lisää, että pääpaino perinnöllisten sairauksien geenikartoituksessa on siirtynyt *monitekijäisiin* sairauksiin, joissa mikään geneettinen tekijä yksin ei riitä selittämään sairauden syntyä. Tällaisten sairauksien alttiusgeenien etsiminen on vaikeaa: se vaatii sairauden ja perimän välisen syy-seuraussuhteen analysointia, innovatiivista koesuunnittelua, suuria aineistoja ja usein laskentaintensiivistä, eksploratiivista lähestymistapaa. Koeasetelmissa painopiste on hitaasti siirtymässä perheaineistojen analysoinnista kohti populaatiopohjaisia lähestymistapoja. Myös laboratoriotekniikan kehittyminen on muuttanut koeasetelmia: ihmisen sekvensoinnin tuloksena on löydetty suuri määrä yhden emäksen mutaatioita, niin kutsuttuja SNP-kohtia (*single nucleotide polymorphisms*), joiden avulla yksittäisen ihmisen DNA:sta saatavan datan määrä voi olla huomattavan suuri. Eräs keskeisistä kysymyksistä on, kuinka tästä suuresta informaatiomäärästä saadaan eristet-

tyä tutkijoita kiinnostavaa tietoa.

Tämän oppaan näkökulma on käytännönläheinen: keskeisen teorian ohella tutustutaan geenikartoitusaineistojen käsittelyyn CSC:n laskentaympäristössä ja perehdytään tavallisimpien geenikartoitusohjelmistojen käyttöön. Näkökulma on tarkoituksella sovelluspainotteinen, mikä johtuu alan ohjelmistokehityksen luonteesta. Koska huomattava osa alan ohjelmistoista on nopeasti etenevän akateemisen kehitystyön tuloksia, on loppukäyttäjälle tarjolla usein vain komentopohjaisia käyttöliittymiä, jotka vaativat aluksi opettelua, mutta ovat osaavissa käsissä toisaalta joustavia ja tehokkaita.

Tämä opas alkaa katsauksella geenikartoituksen keskeisiin käsitteisiin ja lähestymistapoihin (luku 2). Seuraavaksi tarkastelemme geenikartoituksen tarvittavan informaation esittämistä tietokoneella (luku 3) ja tutustumme geenikartoitusongelman taustalla oleviin keskeisiin tilastollisiin käsitteisiin (luku 4). Sen jälkeen tarkastelemme kahta perheineistojen lähestymistapaa, parametrissa (luku 5) ja ei-parametrissa kytkentäanalyysiä (luku 6). Luvussa 7 tutustumme kytkentäepätasapainon käsitteeseen, joka luo perustan väestöpohjaiselle geenikartoitukselle. samassa yhteydessä esitellään väestöpohjaista analysointia käytännössä. Luku 8 perehdyttää haplotyyppipohjaisiin assosiaatiomenetelmiin, joissa hyödynnetään samanaikaisesti informaatiota useista lähekkäisistä perimän kohdista. Oppaan viimeisessä luvussa tarjotaan suuntaviivoja analyysivaiheiden automatisoinniksi suurissa geenikartoitusprojekteissa.

2 Geenikartoituksen lähestymistavat

Tässä luvussa määritellään keskeisiä käsitteitä, joiden päälle geenikartoituksen teoria voidaan rakentaa. Luvun lopussa tutustutaan lyhyesti kahteen geenikartoituksen lähestymistapaan: perheaineistoihin perustuvaan kytkentäanalyysiin ja väestöaineistoja hyödyntävään assosiaatioanalyysiin.

2.1 Monogeeniset ja monitekijäiset sairaudet

Viime vuosina ihmisen geenikartoituksen painopiste on siirtynyt yhden geenin *monogeenisistä* sairauksista *monitekijäisiin* sairauksiin, joissa yksilön sairastumisen ajatellaan olevan seurausta geneettisten tekijöiden ja ympäristötekijöiden yhteisvaikutuksesta, joka voi olla hyvinkin monimutkainen. Monitekijäisen sairauden ollessa kyseessä mikään geeni yksinään ei siis riitä selittämään yksilön sairastumista, ja geenien sijainnin lisäksi monet analyysin kannalta keskeiset parametrit ovat tuntemattomia. Tällaisia parametreja ovat esimerkiksi vaikuttavien geenien lukumäärä, yksittäisten geenien vaikutusten voimakkuus, ympäristötekijöiden vaikutus ja eri tekijöiden yhteisvaikutukset.

Erityisesti monitekijäisten sairauksien alttiusgeenien paikantamiseksi on toistaiseksi vähän menetelmiä. Niinpä lähtökohtana on etsiä aluksi yhtä alttiusgeeniä, jolla toivotaan olevan riittävän selkeä vaikutus ilmiäsuun.

Kun geneettistä sairautta kuvataan formaalisti, puhutaan usein *sairausmallista*, joka määritetään *penetranssivektorin* sekä riskialleelin populaatiofrekvenssin avulla. Penetranssivektori määrittää ehdolliset todennäköisyydet, joilla erilaisia genotyyppettä alttiuslokuksessa kantavat henkilöt sairastuvat. Esimerkiksi penetranssivektori

$$\Pi = \begin{bmatrix} \pi_{DD} \\ \pi_{D+} \\ \pi_{++} \end{bmatrix} = \begin{bmatrix} 0.99 \\ 0.8 \\ 0.001 \end{bmatrix}$$

määrittää riskialleelin suhteen homotsygootin sairastumistodennäköisyydeksi 99%, heterotsygootin sairastumistodennäköisyydeksi 80% ja riskialleelia

kantamattoman henkilön sairastumistodennäköisyydeksi 0.1%. Edellä määritelty sairausmalli kuvaa monitekijäisen sairauden tapauksessa sitä vaikutusta, jonka yksittäinen alttiustekijä aiheuttaa. Muiden tekijöiden vaikutus on tällöin projisoitu näkymättömiin.

Yksinkertaisimmillaan geneettiset sairaudet ovat *dominantteja* tai *resessiivisiä*. Dominantti ja resessiivinen sairaus ovat erikoistapauksia sairausmalleista: edellisen penentranssivektori on $[1, 1, 0]^T$ ja jälkimmäisen $[1, 0, 0]^T$.

Sairauden *prevalenssi* kuvaa sen yleisyyttä väestössä. Prevalenssi voidaan helposti laskea sairausmallin avulla, mutta käytännössä tilanne on usein käänteinen: prevalenssi tunnetaan, mutta penentranssivektori ja riskialleelin populaatiofrekvenssi ovat hämärän peitossa. Minkä tahansa prevalenssiarvon kanssa yhteensopivia yhden geenin sairausmalleja on ääretön määrä. Geenikartoitusprojekteissa sairausmallia joudutaan näin ollen yleensä pitämään tuntemattomana.

Kuten todettiin, monitekijäisten sairauksien yhteydessä mikään geneettinen tekijä ei riitä yksin selittämään sairauden syntyä. Tilannetta, jossa useat geenit vaikuttavat sairauden syntyyn, kutsutaan *lokusheterogeniaksi*. Tilannetta, jossa saman geenin erilaiset variantit aiheuttavat (saman tai samankaltaisen) sairauden, kutsutaan *alleeliheterogeniaksi*. Näitä kahta heterogenian tyyppiä kutsutaan yhteisnimellä *geneettinen heterogenia*.

Geneettisen tekijän voimakkuutta kuvataan toisinaan myös *relatiivisen riskin* avulla. Relatiivinen riski kertoo, kuinka moninkertainen sairastumistodennäköisyys riskialleelin kantajalla on verrattuna riskialleelia kantamattomaan henkilöön.

2.2 Ihmisen genomi, markkerit ja fenotyypit

Ihmisen genomi on organisoitunut 22 autosomaaliseksi kromosomipariksi sekä kahdeksi sukupuolikromosomiksi.

Geenikartoitus perustuu *markkereiden* käyttöön. Markkerilla tarkoitetaan sellaista perimän kohtaa, jossa eri yksilöiden välillä on geneettistä vaihtelua. Eri yksilöiden variantteja noissa perimän kohdissa kutsutaan *alleleiksi*. Geenikartoitusprojektissa kunkin yksilön DNA-näytteestä selvitetään, mitä kahta alleelia hän kussakin markkerilokuksessa kantaa. Tuota alleeliparia kutsutaan *genotyypiksi*.

Suurin osa markkereista sijaitsee perimän ei-koodaavissa kohdissa, joissa eri alleleihin ei kohdistu minkäänlaista valintapainetta. Yleisimmin käytetyt markkerit ovat joko kaksialleelisiä SNP-markkereita tai monialleelisiä *mikrosatelliittimarkkereita*, joissa eri alleelit vastaavat eri mittaisia toistojaksoja. Ihmisen genomien kartoituksen myötä saatavilla olevien SNP-markkereiden määrä on kasvanut erittäin suureksi ja kasvaa edelleen.

Markkerien eri alleelien yleisyyksiä kutsutaan *alleelifrekvensseiksi*. Alleelifrekvenssit vaihtelevat populaatiosta toiseen väestöhistoriallisista eroista joh-

tuen. Useat geenikartoitusmentelmät toimivat paremmin, jos väestötason alleelifrekvensseistä on saatavilla mahdollisimman oikeaa tietoa. Käytännössä alleelifrekvenssit joudutaan usein arvioimaan varsinaista tutkimusta varten kerätystä otoksesta.

Markkerin *informatiivisuus* kuvaa sen käyttökelpoisuutta geenikartoituksessa. Informatiivisuus on sitä korkeampi, mitä suurempi on markkerin alleelien lukumäärä ja mitä tasaisempi alleelifrekvenssien jakauma on. Informatiivisuudelle on kehitetty mittoja, joista tunnetuimmat ovat:

1. Heterotsygotia-aste (*heterozygosity*):

$$H = 1 - \sum_{i=1}^n p_i^2$$

2. polymorfismin informaatioisältö (*polymorphism information content, PIC*):

$$PIC = 1 - \sum_{i=1}^n p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^n 2p_i^2 p_j^2$$

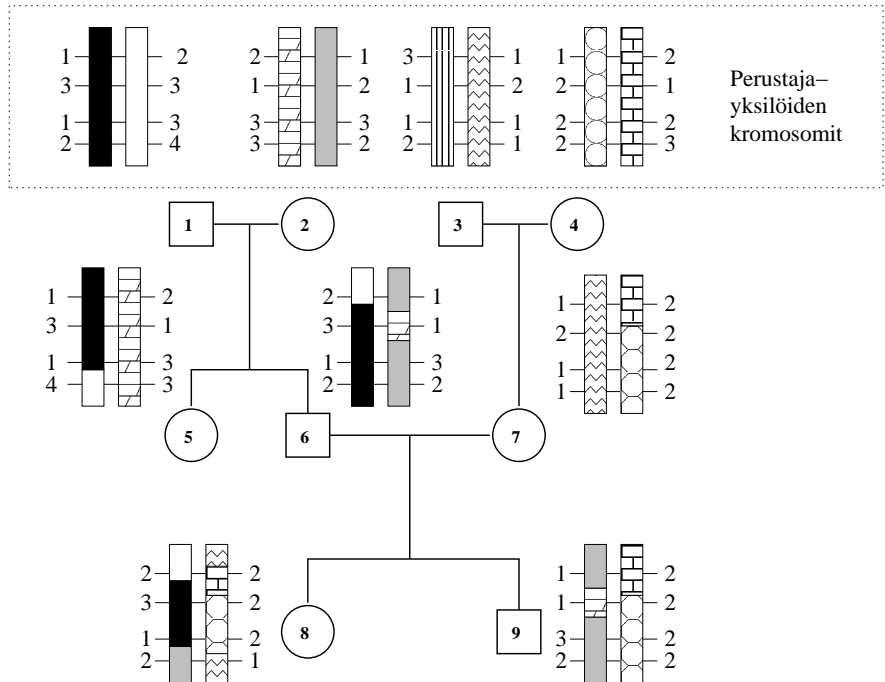
Edellä merkintä p_i tarkoittaa alleelin i frekvenssiä. Esimerkiksi kaksialleelisen SNP-markkerin, jonka alleelifrekvenssit ovat 0.6 ja 0.4, heterotsygotia-asteeksi saadaan 0.48 ja PIC-arvoksi 0.3648. Mitan arvot vaihtelevat nollan ja yhden välillä.

Fenotyyppiä tarkoitetaan yksilön ilmiänsua eli sitä piirrettä, josta tutkija on kiinnostunut. Sairausgeenien kartoitusprojekteissa fenotyyppi voi olla kaksiarvoinen, jolloin yksilö on sairas tai terve, tai jatkuva-arvoinen, jolloin sairauden aste kuvataan mitattavana arvona. Fenotyyppiä käytetään usein kriteerinä, jonka perusteella otos poimitaan tutkimusta varten.

2.3 Sukupuut ja periytyminen

Tarkastellaan yhden autosomaalisen kromosomiparin periytymistä sukupuussa. Kuva 2.1 esittää sukupuuta, johon kuuluu yhdeksän henkilöä. Naispuoliset henkilöt on merkitty ympyröin ja miespuoliset neliöin. Sukupuun neljää ylintä henkilöä (numerot 1–4), joiden vanhempia ei ole sukupuussa mukana, kutsutaan *perustajayksilöiksi*. Kullekin neljästä perustajayksilöstä on kuvaan piirretty kaksi kromosomia, jotka on merkitty erilaisin rasteroinnein. Kussakin kromosomiparissa isältä peritty kromosomi on piirretty vasemmanpuoleiseksi ja äidin oikeanpuoleiseksi.

Näemme, että kussakin periytymistapahtumassa lapsi perii toisen kromosomin isältään ja toisen äidiltään. Vanhemmalta peritty kromosomi ei kuitenkaan aina periä sellaisenaan, vaan se voi olla yhdistelmä kahden vanhemmalla esiintyvän kromosomin perimäaineksestä. Esimerkiksi tyttö 5 on pe-



Kuva 2.1: Esimerkki kromosomisegmenttien ja niiden alueella olevien markkerien periytymisestä sukupuussa.

rinyt yläosan kromosomistaan isän isänpuoleisesta ja alaosan isän äidinpuoleisesta kromosomista. Tällöin isän *meioosissa* (tapahtumassa, jossa sukusolut muodostuvat) isän kahden vastinkromosomin välillä on tapahtunut *tekijäinvaihtoa* (engl. *crossing-over*). Kaikki geenikartoitusmenetelmät perustuvat tavalla tai toisella tekijäinvaihdon laskennalliseen mallintamiseen.

Se, montako tekijäinvaihtokohtaa lapselle periytyvään kromosomiin on muodostunut, riippuu karkeasti muun muassa kromosomin fyysikaalisesta pituudesta. Odotusarvo kromosomissa yhden meiosisin aikana muodostuville tekijäinvaihtokohdille määrittelee kromosomin *geneettisen pituuden*. Yhden *morganin* – eli 100 *senttimorganin* – mittaiselle kromosomialueelle muodostuu odotusarvoisesti yksi tekijäinvaihtokohta. Ihmisen kromosomien pituudet vaihtelevat muutamasta kymmenestä muutamaa sataan senttimorganiin. Laskennallisissa analyysissä tehdään useimmiten oletus, että tekijäinvaihtokohtien lukumäärä noudattaa Poisson-jakaumaa, jonka parametrina on kromosomin geneettinen pituus morganeina. Tekijäinvaihtokohtien oletetaan yleensä jakautuvan satunnaisesti ja toisistaan riippumattomasti tarkasteltavalle kromosomialueelle.

Kahden pisteen välillä sanotaan tapahtuneen *rekombinaatio*, jos havaitaan, että meiosisin seurauksena pisteet ovat lähtöisin vanhemman eri vastinkromosomeista. Rekombinaatio voidaan havaita, jos pisteiden välillä on ollut pariton määrä tekijäinvaihtokohtia. Todennäköisyyttä, että genomissa kah-

den pisteen välillä on tapahtunut rekombinaatio, kutsutaan *rekombinaatiofraktioksi*. Rekombinaatiofraktion arvo vaihtelee nollan ja puolikkaan välillä; erityisesti arvo on nolla, jos lokukset ovat päällekkäin. Kahden samassa kromosomissa olevan lokuksen arvo kasvaa nolasta ja lähestyy puolikasta, kun lokukset siirtyvät kauemmaksi toisistaan. Eri kromosomeissa olevien lokusten välinen rekombinaatiofraktio on aina 0.5 johtuen siitä, että eri kromosomit periytyvät toisistaan riippumattomasti. Lokusten sanotaan olevan sitä vahvemmin *kytkeytyneitä*, mitä lähempänä nollaa niiden välinen rekombinaatiofraktio on.

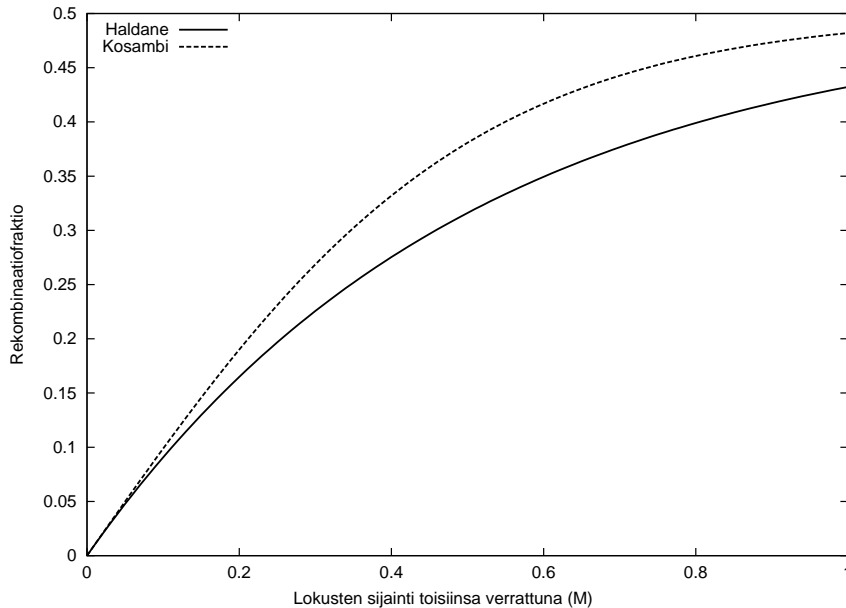
Geneettiset etäisyydet (m) voidaan muuttaa rekombinaatiofraktioiksi (θ) *karttafunktioiden* avulla. Näistä tunnetuimpia ovat Haldanen karttafunktio

$$\theta = \frac{1 - e^{-2m}}{2}$$

sekä Kosambin karttafunktio

$$\theta = \frac{1 - e^{-4m}}{2(1 + e^{-4m})}$$

Myös käännteisten kaavojen johtaminen on helppoa. Geneettisten etäisyyksien ja rekombinaatiofraktion välistä suhdetta on havainnollistettu kuvassa 2.2.



Kuva 2.2: Geneettisen etäisyyden ja rekombinaatiofraktion suhde Haldanen ja Kosambin karttafunktioita soveltaen.

Lokusten välistä geneettistä kytkeytystä voidaan hyödyntää sen selvittämisessä, miten kromosomien kohdat ovat periytyneet. Kuvaan 2.1 on merkitty neljä markkeria ja niissä esiintyvät alleelit. Markkerilokusten alleelit periytyvät

lapsille vastaavien kromosomisegmenttien mukana. Sovittamalla tarkasteltavalle kromosomialueelle riittävän tiheä markkerikartta ja määrittämällä yksilöiden genotyypit markkerikohdissa voidaan arvioida sitä, miten kromosomialueet periytyvät sukupuissa ja mitä kromosomialueita tietyn fenotyypin mukaiset yksilöt jakavat. Tämä on geenikartoituksen taustalla oleva ajatus.

2.4 Lähestymistavat ja geenikartoitusohjelmistot

Geenikartoitukseen on olemassa kaksi lähestymistapaa: *kytkentäanalyysi* ja *assosiaatioanalyysi*. KytKentäanalyysissä tarkastellaan sairauden periytymistä sukupuaineistoissa, kun taas kytKentäepätasapainoon perustuvassa assosiaatioanalyysissä aineisto kerätään väestöstä tai väestöisolaatista, jossa sairautta esiintyy. KytKentäanalyysiä esitellään luvuissa 5 ja 6, ja assosiaatioanalyysiin palataan luvussa 7.

Genominlaajuiset geenikartoitusprojektit ovat perinteisesti edenneet siten, että genomista etsitään aluksi kytKentäanalyysin keinoin lupaavia alueita käyttäen suhteellisen harvaa markkerikarttaa. Tämän jälkeen näiden alueiden informatiivisuutta parannetaan genotyypaamalla lisää markkereita ja ottamalla mahdollisuuksien mukaan uusia perheitä analysoitavaksi. KytKentämenetelmän tarkkuusrajojen tullessa vastaan aluetta pyritään kaventamaan edelleen kytKentäepätasapainoon perustuvien menetelmin (assosiaatioanalyysi), ja ihannetapauksessa tässä vaiheessa voidaankin hyödyntää samaa, geneettisestä isolaatista kerättyä aineistoa kuin kytKentäanalyysissä. Monitekijäisten sairauksien ja tiheiden SNP-karttojen myötä geenikartoituksen painopiste on siirtymässä assosiaatiomenetelmien suuntaan, ja jopa puhtaasti kytKentäepätasapainoon perustuvat geenikartoitusprojektit ovat yleistymässä. Toisaalta kytKentäanalyysi säilyttäneenä asemansa geenikartoituksen keskeisenä työkaluna vielä pitkään: esimerkiksi eri puolilla maailmaa kerättyjen aineistojen meta-analyysi kytKentämenetelmin tarjoaa monia mahdollisuuksia.

Taulukossa 2.1 esitetään tässä oppassa käytettävät ohjelmistot, jotka on asennettu CSC:n Corona-sovelluspalvelimelle. Corona-palvelin otetaan käyttöön alkuvuodesta 2005, johon saakka sovelluspalvelimena toimii Cedar-palvelin. Tämän oppaan esimerkeissä komentokehoteen konenimenä esiintyy corona. Ohjelmien käyttö Cedar-palvelimelta on konenimeä lukuunottamatta samanlaista.

CSC:lle asennetut geenikartoitusohjelmistot alustetaan ennen käyttöä komennolla:

```
corona% use genemap
```

Alustus päivittää eräitä ympäristömuuttujien arvoja siten, että ohjelmia voi

Taulukko 2.1: *Sovelluspalvelimelle asennettuja geenikartoitusohjelmistoja*

Aihe	Ohjelmistot
Parametrinen kytkentäanalyysi	Genehunter, Linkage
Ei-parametrinen kytkentäanalyysi	Merlin, Genehunter, Simwalk
Periytymisen epätasapainotesti (TDT)	Genehunter
Linkage-parametritiedoston luonti	Linkagepar
Tiedostomuotojen muunnokset	Mega2
Sukupuutiedoston tarkistaminen	Pedcheck
Assosiaatioanalyysi	Haplo-assoc
Periytymisen simulointi	Merlin

sen jälkeen käyttää istunnon aikana viittaamalla niihin pelkällä ohjelman nimellä täydellisen polun sijasta.

3 Aineistojen esitys tietokoneella

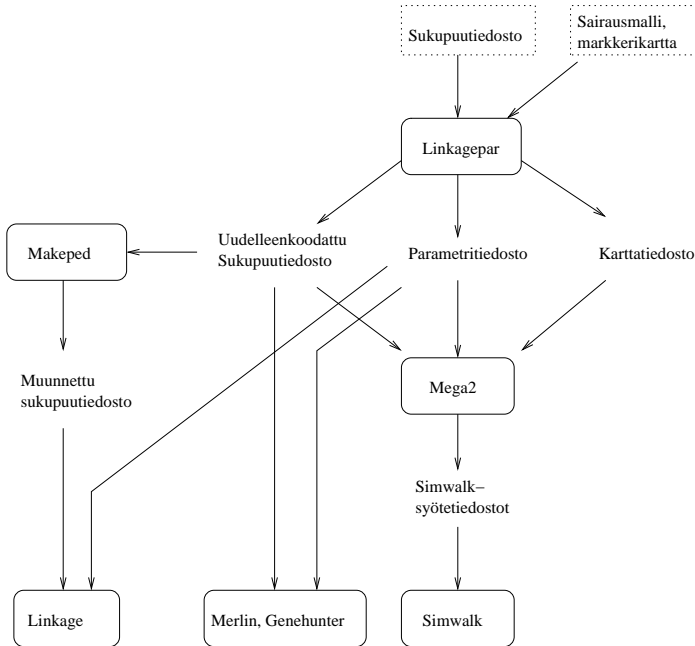
Geenikartoitusprojektissa analysoitava informaatio koostuu henkilöiden sukulaisuussuhteista, fenotyypeistä, genotyypeistä ja toisinaan ympäristötekijöiden arvoista. Jotta geneettistä aineistoa voitaisiin käsitellä tietokoneella, on tämä informaatio sekä sukupuiden rakenne kuvattava formaalissa, tietokoneen ymmärtämässä muodossa.

Eräänlaisen standardin aseman on saavuttanut Linkage-esitysmuoto, joka koostuu sukupuu- ja parametritiedostosta. Linkage-esitysmuoto otetaan tässä oppaassa aineistojen esityksen lähtökohdaksi. Aineistojen rakentaminen ja esikäsittely CSC:n laskentaympäristössä yleisimpiä kytkentäanalyysiohjelmia varten on esitetty kaaviona kuvassa 3.1. Yleisimmin käytetyt kytkentäanalyysiohjelmat, Merlin[[ACCC02](#)] ja Genehunter[[KDRDL96](#)], saavat syötteenään Linkage-muotoisen sukupuu- ja parametritiedoston sellaisenaan. Näistä sukupuutiedosto rakennetaan esimerkiksi tekstieditorilla tai (tuotantoympäristössä) automaattisesti perhe- ja genotyypiaineiston perusteella, kun taas parametritiedoston luontiin on olemassa apuohjelmia (tässä esitelty Linkagepar sekä Makedata, Downfreq ja Preplink). Kolmas yleinen kytkentäanalyysiohjelma, Simwalk (eli Simwalk2)[[SL96](#)], edellyttää lisäksi tiedostomuodon muuttamista omaan formaattiinsa; tähän voidaan käyttää esimerkiksi Mega2-apuohjelmaa[[MAS+99](#)]. Myös Linkage-kytkentäanalyysiohjelmaa[[LLJO84](#), [LL84](#), [LLW86](#)] varten sukupuutiedosto on käsiteltävä apuohjelmalla (Makeped). Syötetiedostojen rakentamiseen ja tiedostomuotojen muunnoksiin tutustutaan tässä luvussa yksityiskohtaisesti.

3.1 Linkage-tiedostomuoto

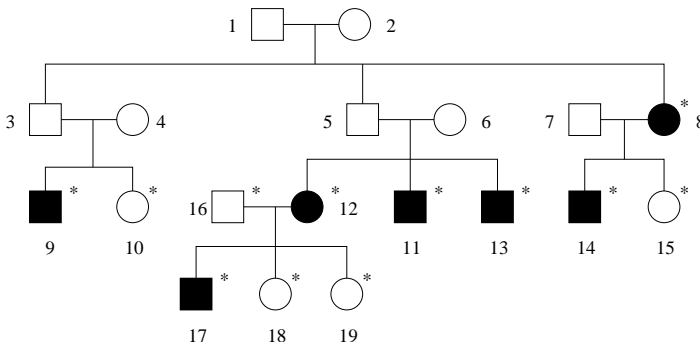
Tarkastellaan aluksi tiedostojen esittämistä Linkage-muodossa, joka on yleisimmin käytetty geneettisten aineistojen esitysmuoto.

Linkage-muotoinen aineiston kuvaus koostuu kahdesta tiedostosta: sukupuutiedostosta ja parametritiedostosta. Muodostamme tällaiset tiedostot kuvan 3.2 esittämälle sukupuulle, jonka tähdellä merkityistä henkilöistä on genotyyppattu kahdeksan mikrosatelliittimarkkeria. Seitsemän ensimmäisen henkilön sairausstatuksesta ei ole varmaa tietoa, joten status merkitään näiden



Kuva 3.1: Sukupuu- ja parametritiedostojen luonti ja esikäsittely kytkentäanalyysiä varten.

yksilöiden osalta tuntemattomaksi.



Kuva 3.2: Esimerkkitiedostoa vastaava sukupuu. Sairaats yksilöt on merkitty täytetyllä mustalla symbolilla. Genotyypit on kerätty asteriskilla merkityiltä yksilöiltä, ja heistä tiedetään myös sairausstatus.

3.1.1 Linkage-sukupuutiedoston rakenne

Linkage-muotoisessa sukupuutiedostossa kuvataan perherakenne, fenotyyppi sekä markerilokusten genotyypit. Tiedosto voidaan laatia käsin tekstieditorilla tai se voidaan – automatisoidussa tuotantoympäristössä – generoida tietokannassa olevan genotyypidatan avulla. Kun analysoitavana on yksi binäärinen fenotyyppi, on sukupuutiedoston rakenne seuraava:

Niinpä esimerkiksi yksilöistä 14 ja 15 on genotyyppatyn äidin (yksilö 8) lisäksi sukupuutiedostossa esitettävä genotyypaamattoman ja sukupuuhun muuten kuulumattoman isän tiedot. Tämä on tehty siten, että tiedostoon on luotu ylimääräinen yksilö 7, jonka fenotyyppi- ja genotyypitiedot on merkitty tuntemattomiksi.

Osa kytkentäanalyysiohjelmista sallii perhe- ja yksilötunnisteiksi mielivaltaisia numeroita tai merkkijonoja. Jotta tiedosto olisi yhteensopiva kaikkien ohjelmistojen kanssa, kannattaa tunnisteina käyttää pieniä kokonaislukuja.

Alleelit koodataan kokonaisluvuin. Koska alleelifrekvenssit on parametritiedostossa (ks. luku 3.1.3) lueteltava kaikille alleeleille ykkösestä alkaen suurimpaan alleelinumeroon asti, on alleelitunnisteinakin järkevää käyttää pieniä kokonaislukuja.

Edellä olevassa esimerkkitiedostossa näytteet on kerätty 12 henkilöltä (yksilöt 8–19). Yksittäiset markkerit ovat olleet toimimattomia joillakin yksilöillä; tällöin vastaava genotyyppi on merkitty tiedostossa "0 0". Yksilöiltä, jotka kuuluvat sukupuuhun, mutta joilta DNA-näytettä ei ole saatu, merkitään kaikki genotyypit puuttuviksi.

Linkage-tiedostomuoto sallii yksilöille valinnaisen alttiusluokan kuvaamisen. Alttiusluokalla tarkoitetaan aineiston alaryhmää, jolle määritetään parametrista kytkentäanalyysiä varten oma penetranssivektorinsa. Penetranssivektorit voivat riippua esimerkiksi iästä tai sukupuolesta. Jos aineisto halutaan jakaa esimerkiksi kolmeen alttiusluokkaan (alle 30-vuotiaat, 30-60-vuotiaat ja yli 60-vuotiaat), merkitään nuorimpaan ryhmään kuuluville ihmisille ylimääräiseen, seitsemänteen sarakkeeseen (sairausstatuksen ja genotyyppien väliin) luku 1. Vastaavasti keskimmäiseen ikäryhmään kuuluville merkitään kyseiseen sarakkeeseen luku 2 ja vanhimpaan ikäryhmään kuuluville luku 3. Varsinainen penetranssivektorien sisältö kuvataan erillisessä parametritiedostossa (ks. luku 3.1.3). Edellä esitetyssä esimerkissä alttiusluokat eivät ole käytössä.

3.1.2 Sukupuutiedoston tarkistaminen Pedcheck-ohjelmalla

Sukupuuaineiston tekninen virheettömyys on hyvä tarkistaa aina ennen tilastollisia analyysejä. Tyypillisiä virheitä ovat yksinkertaiset laboratorio- ja koodausvirheet. Nämä johtavat siihen, että lapsilla saattaa esiintyä alleeleja, joita kummallakaan vanhemmista ei ole. Jos virheitä ilmenee useassa lokuksessa samalla yksilöllä, voi tämä selittyä sillä, että sukulaisuussuhteet (erityisesti biologinen isyys) poikkeavat sukupuussa kuvatusta.

Aineiston oikeellisuus voidaan selvittää Pedcheck-nimisen [OW98] ohjelman avulla. Pedcheck-ohjelmalle annettavan sukupuutiedoston tulee olla Linkage-muodossa. Ohjelmalle annetaan sukupuutiedoston ohella joko Linkage-muotoinen parametritiedosto tai erityinen markkerinimitiedosto. Ohjelman keskeiset komentoriviparametrit ovat:

- p** *sukupuutiedosto*: määrittää sukupuutiedoston nimen
- d** *parametritiedosto*: määrittää parametritiedoston nimen
- m**: ilmoittaa, että käytetään sukupuutiedostoa, joka ei ole Makeped-apuohjelman luoma
- n** *markkerinimitiedosto*: määrittää markkerinimitiedoston nimen. Tiedosto on annettava silloin, jos parametritiedostoa ei anneta syötteenä.
- o** *tulostiedosto*: määrittää tulostiedoston nimen. Oletusnimenä on `pedcheck.err`.
- 2**: kertoo, että käyttäjä haluaa ns. toisen tason tarkistuksen. Tämä tehdään, jos ns. ensimmäisen tason tarkistuksessa ei löytynyt virheitä. Jos sukupuutiedosto läpäisee toisen tason tarkistuksen, on aineisto varmasti Mendelin periytymissäntöjen mukainen.

Seuraavaksi laadimme Pedcheck-ohjelmaa varten markkerinimitiedoston. Markkerinimitiedostossa merkitään X- tai x-symbolein sellaiset lokukset, joita ei haluta tarkistaa. Muille lokuksille annetaan nimi. Seuraavassa on tästä esimerkki (tiedostonimenä `list.markers`):

```
X
M1
M2
M3
M4
M5
M6
M7
M8
```

Sukupuutiedostossa esiintyvä ensimmäinen lokus kuvaa sairausstatusta, joten sille ei haluta tehdä Mendel-tarkistusta.

Testaamme ohjelmaa edellä esitetylle, nimellä `demoped.pre` tallennetulle sukupuutiedostolle, johon luomme ensin keinotekoisen Mendel-virheen. Muutetaan tiedostoa `demoped.pre` siten, että toiseksiviimeisen yksilön 18 viimeisen lokuksen toinen alleli on virheellinen:

```
1 18 16 12 2 1 6 5 5 5 1 5 5 4 2 7 2 3 6 6 4 1
```

Käynnistämme tarkistuksen komennolla:

```
corona% pedcheck -p demoped.pre -m -n list.markers
```

Tämän jälkeen katsomme tiedostoa `pedcheck.err`, johon virheet on kirjoitettu. Esimerkkiaineistosta löytyi yksi virhe:

```
##### GENOTYPE ERROR: Pedigree 1 Locus 8 Name M8 #####
```

```
ERROR: Child 18 is consistent with each parent separately,  
but not as a pair.
```

```
ORIGINAL SCORING:
```

```
Father 16: 2/4 Mother 12: 4/5
```

```
Child 17: 5/2
```

```
Child 18: 1/4
```

```
Child 19: 0/0
```

Virheiden korjaamisen jälkeen edellinen komento on annettava uudestaan. Kun kaikki virheet on korjattu, ohjelma antaa raportin, jossa lukee

```
PedCheck has found NO LEVEL 0 inconsistencies in the pedigree  
data.
```

```
PedCheck has found NO LEVEL 1 inconsistencies in the pedigree  
data.
```

```
RERUN USING LEVEL 2.
```

Tämän jälkeen teemme ns. toisen tason tarkistuksen antamalla komennon, johon on lisätty optio `-2`:

```
corona% pedcheck -p demoped.pre -m -n list.markers -2
```

Koska ohjelma ei havaitse uusia virheitä, aineisto noudattaa mendelistisiä pe-rietymissääntöjä. Jos ensimmäisen tason virheitä ei löydy, voidaan Pedcheck-ohjelmalla myös etsiä todennäköisimpiä virhelähteitä (optio `-3`) korjaus-ehdotuksineen (optio `-4`). Vaikka datan korjaaminen tällaisten todennäköisyys-laskelmien perusteella on uhkarohkeaa, voi tulosteista olla apua genotyyp-pausvirheiden jäljityksessä.

3.1.3 Linkage-parametritiedosto ja sen luonti

Edellä käsiteltiin Linkage-muotoisen sukupuutiedoston luontia ja tarkistus-ta. Sukupuutiedostossa kuvattiin yksilökohtaiset tiedot: perherakenteet, sai-rausstatukset ja henkilöiden genotyypit. Linkage-muotoisessa parametritie-dostossa puolestaan kuvataan kaikille henkilöille yhteiset parametrit: sairaus-malli, käytetyt fenotyypit ja markkerit sekä markkereiden alleelifrekvenssit.

Historiallisista syistä tiedosto sisältää myös kenttiä, joilla välitetään parametreja analyysiohjelmalle; useat ohjelmistot jättävät nämä kentät huomiotta.

Parametritiedosto on kätevintä luoda jollakin tarkoitukseen kehitetyllä apuohjelmalla. Esimerkiksi Linkagepar-nimisellä ohjelmalla parametritiedosto voidaan helposti luoda sukupuutiedostosta lähtien. CSC:llä kehitetty ohjelma käynnistetään antamalla parametrina sukupuutiedoston nimi:

```
corona% linkagepar demoped.pre
```

Ohjelma käynnistyy tulostamalla yhteenvedon aineistosta:

```
Pedigree statistics:  
19 individuals.  
8 markers.  
  
Press Enter to continue.
```

Painamalla Enter-näppäintä ruudulle ilmestyy päävalikko:

```
Linkagepar - a CSC tool for constructing Linkage parameter files  
Version 20-Feb-04. Requests/comments to vesa.ollikainen@csc.fi  
  
(1.) Penetrance vector 1      : 0.01 0.99 0.99  
(2.) Disease allele frequency : 0.001  
(3.) Intermarker distances   : 0.1 0.1 0.1 0.1 0.1 0.1 0.1  
(4.) Allele frequencies      : Estimate from all individuals.  
(5.) Downcode alleles       : yes -- to file: out.pre  
(6.) Output parameter file   : out.dat  
(7.) Output map file        : out.map  
  
(9.) Generate files and exit.  
(X ) Exit without generating files.
```

Ohjelma näyttää muokattavissa olevat parametrit (1-7) sekä kunkin parametrin voimassa olevat arvot. Arvoja voidaan muuttaa syöttämällä vaihtoehdon numero ja vastaamalla ohjelman esittämiin kysymyksiin. Valikossa voidaan muuttaa seuraavia arvoja:

Penetranssivektori. Penetranssit eli sairauden todennäköisyydet annetaan kunkin genotyypille edustajille: riskialleelia kantamatomille henkilöille (++) , heterotsygooteille riskialleelin haltijoille (D+) ja riskialleelin

suhteen homotsygooteille henkilöille (*DD*). Jos aineistossa on määritetty alttiusluokat, ohjelma kysyy pentetranssivektorin arvot erikseen kullekin alttiusluokalle.

Sairausalleelin frekvenssi. Riskialleelin frekvenssi populaatiotasolla.

Markkerien etäisyydet. Markkerien etäisyydet syötetään joko rekombinaatiofraktioina tai senttimorganeina. Jos käytetään senttimorganeita, käyttäjä valitsee joko Haldanen tai Kosambin karttafunktion. Päävalikossa etäisyydet näkyvät aina rekombinaatiofraktioina riippumatta siitä, missä muodossa ne on syötetty.

Alleelifrekvenssit. Ohjelmalle kerrotaan tapa, jolla alleelifrekvenssit arvioidaan: arviointi voidaan tehdä perustajayksilöiden perusteella ("estimate from founders"), kaikkien yksilöiden perusteella ("Estimate from all individuals") tai alleelifrekvenssit voidaan asettaa yhtä suuriksi ("Use equal allele frequencies"). Käytetyillä alleelifrekvensseillä on vaikutusta kytkeäntäilyanalyysin tuloksiin, sillä ohjelmistot laskevat niiden perusteella tuntemattomien perustajagenotyyppien todennäköisyyksiä. Jos populaatiotason alleelifrekvenssejä ei ole saatavilla, joudutaan ne arvioimaan aineistosta. Tällöin pelkkien perustajayksilöiden käyttämisessä on se etu, että ne ovat lähtökohtaisesti riippumattomia toisistaan ja heijastavat paremmin populaatiotason alleelifrekvenssejä. Toisaalta perustajayksilöitä on yleensä vähän, jolloin frekvenssit ovat herkkiä satunnaisvaihtelulle. Jos populaatiotason alleelifrekvenssit markkereille ovat tiedossa esimerkiksi aiemmista tutkimuksista, kannattaa niitä käyttää. Tällöin frekvenssit voidaan syöttää jälkikäteen ohjelman generoimaan parametritiedostoon.

Alleelien uudelleenkodeausmenetelmä. Alleelit voidaan tarvittaessa uudelleennumeroida ykkösestä alkaen. Jos uudelleennumerointi valitaan, joudutaan myös sukupuutiedosto kirjoittamaan uudelleen. Tällöin ohjelma kysyy käyttäjältä tiedostonimen uudelle sukupuutiedostolle.

Tulostettava parametritiedosto. Tulostettavan parametritiedoston nimi.

Tulostettava karttatiedosto. Ohjelma tulostaa myös karttatiedoston Mega2-tiedostomuunnosohjelmaa varten. Käyttäjä voi muuttaa karttatiedoston nimen.

Ohjelmasta poistutaan valinnalla 9. Poistuttaessa ohjelma kirjoittaa parametritiedoston ja – jos sukupuoli haluttiin uudelleennumeroida – uuden sukupuutiedoston. Valinnalla X ohjelman suoritus voidaan keskeyttää generoimatta tulostiedostoja.

Muutamme esimerkkitiedostomme karttaetäisyyksiksi 5 cM valitsemalla päävalikosta valinnan 5. Seuraavaksi ohjelmalle kerrotaan, että etäisyydet syötetään senttimorganeina käyttäen Haldanen karttafunktiota. Tämän jälkeen markkerien väliset etäisyydet annetaan yksi kerrallaan. Kussakin vaiheessa

ohjelma näyttää valintaa vastaavan oletusarvon hakasulkeissa. Oletusarvo hyväksytään painamalla pelkästään Enter-näppäintä. Jos oletusarvoa ei hyväksytä, syötetään uusi arvo.

```
How do you want to enter intermarker distances?
(1.) Recombination fractions.
(2.) Haldane centiMorgans.
(3.) Kosambi centiMorgans.
[1] 2

Enter 7 intermarker distances one at a time.
Use Haldane centiMorgans.

Markers 1 - 2 : [11.1572] 5
Markers 2 - 3 : [11.1572] 5
Markers 3 - 4 : [11.1572] 5
Markers 4 - 5 : [11.1572] 5
Markers 5 - 6 : [11.1572] 5
Markers 6 - 7 : [11.1572] 5
Markers 7 - 8 : [11.1572] 5
```

Esimerkissämme arvioimme alleelifrekvenssit kaikkien sukupuun jäsenten perusteella. Yhdestä sukupuusta koostuva aineisto olisi todellisuudessa liian pieni alleelifrekvenssien luotettavaan arviointiin. Alleelifrekvenssien arviointi tehdään tässä vain demonstrointitarkoituksessa.

Ohjelman avulla generoidaan kolme tulostiedostoa: uudelleenkoodattu sukupuutiedosto, parametritiedosto ja karttatiedosto. Viimeksimainittua tiedostoa tarvitaan, jos tiedostot halutaan muuttaa Linkage-tiedostomuodosta johonkin toiseen muotoon Mega2-muunnostyökalun avulla (ks. luku 3.2). Ennen generointia vaihdamme käyttöliittymässä vielä tulostettavien sukupuun-, parametri- ja karttatiedostojen nimiksi ex-1.pre (valinta 5), ex-1.dat (valinta 6) ja ex-1.map (valinta 7):

```
(1.) Penetrance vector 1 : 0.01 0.99 0.99
(2.) Disease allele frequency : 0.001
(3.) Intermarker distances : 0.0475813 0.0475813 0.047...
(4.) Allele frequencies : Estimate from all indiv.
(5.) Downcode alleles : yes -- to file: ex-1.pre
(6.) Output parameter file : ex-1.dat
(7.) Output map file : ex-1.map
```

- (9.) Generate files and exit.
 (X) Exit without generating files.

Valinnalla 9 poistutaan ohjelmistosta. Generoitu parametritiedosto ex-1.dat näyttää seuraavanlaiselta (kenoviivamerkki kuvaa rivin jatkumista seuraavalta riviltä. Parametritiedostossa rivit on kirjoitettava yhteen ilman rivinvaihtoa tai kyseistä merkkiä):

```

9 0 0 5
0 0.0 0.0 0
1 2 3 4 5 6 7 8 9
1 2 # TRAIT
0.999 0.001
1
0.01 0.99 0.99
3 5 # M1
0.0833333 0.125 0.333333 0.125 0.333333
3 3 # M2
0.277778 0.444444 0.277778
3 5 # M3
0.136364 0.318182 0.136364 0.363636 0.0454545
3 3 # M4
0.3125 0.25 0.4375
3 4 # M5
0.458333 0.0833333 0.0833333 0.375
3 3 # M6
0.181818 0.590909 0.227273
3 3 # M7
0.25 0.125 0.625
3 4 # M8
0.272727 0.136364 0.318182 0.272727
0 0
0.5 0.0475813 0.0475813 0.0475813 0.0475813 0.0475813\
0.0475813 0.0475813
1 0.1 0.45

```

Parametritiedoston ensimmäisellä rivillä kerrotaan kuinka monta lokusta analyysissä on mukana (9 = sairauslokus + 8 markkerilokusta). Kolmannella rivillä on lueteltu lokukset järjestyksessä. Neljännellä rivillä on ensimmäinen lokus määritelty sairauslokukseksi (1) ja sen alleelien määräksi 2. Rivillä viisi on esitetty sairausalleelien frekvenssit. Rivillä kuusi kerrotaan alttiusluokkien lukumäärä (1) ja sitä seuraavalla rivillä ensimmäisen ja ainoan alttiusluokan penetranssit riskilokuksen eri genotyypeille. Jos alttiusluokkia olisi

useampi kuin yksi, olisi parametritiedostossa yksi rivi kutakin alttiusluokkaa kohden siten, että ensimmäinen rivi määrittää ensimmäisen alttiusluokan penetransivektorin, toinen toisen jne. Rivillä kahdeksan määritellään seuraava lokus markkeriksi (3) ja annetaan alleelien lukumäärä ja rivillä yhdeksän vastaavat alleelifrekvenssit. Seuraavat rivit vastaavat muita markkerilokuksia. Toiseksi viimeisellä rivillä annetaan lokusten väliset etäisyydet rekombinaatiofraktion avulla siten, että ensimmäinen luku vastaa sairausalleelin ja ensimmäisen markkerilokuksen välistä etäisyyttä (tätä ei luonnollisestikaan tiedetä, ja annettu arvo 0.5 vastaa alkuarvoa) ja tämän jälkeiset luvut markkerilokusten välisiä tunnettuja etäisyyksiä (ensin markkerien 1 ja 2 etäisyys toisistaan, sitten markkerien 2 ja 3 jne.). Loput tiedostossa annettavat arvot eivät ole käyttäjän kannalta merkityksellisiä. Lisää tietoa parametritiedoston sisällöstä löytyy mm. Terwilligerin ja Ottin teoksesta [TO94].

3.2 Tiedostomuotojen muunnokset: Mega2

Mega2-ohjelmalla[MAS⁺99] voidaan muuntaa syötetiedostoja eri geenikartoitusohjelmien tiedostomuotojen välillä. Ohjelmiston avulla voidaan myös rakentaa Unix-komentotiedosto, joka ajaa halutun analyysin.

Tarkastellaan aluksi Simwalk-syötetiedostojen tekoa esimerkisukupuutamme vastaavista Linkage-tiedostoista. Simwalk-ohjelman käyttöä esitellään luvussa 6.4. Tässä yhteydessä keskitymme ainoastaan syötetiedostojen muuntamiseen.

Mega2-ohjelmisto vaatii syötteekseen kolme pakollista tiedostoa: sukupuutiedoston, parametritiedoston ja karttatiedoston. Nämä syötetiedostot voidaan rakentaa Linkage-muotoisista tiedostoista erittäin helposti:

Sukupuutiedostoksi kelpaa Linkage-muotoinen sukupuutiedosto sellaisenaan. Tiedosto saa olla käsitelty Makeped-ohjelmalla, mutta se ei ole välttämätöntä.

Parametritiedostoksi käy Linkage-muotoinen parametritiedosto, jossa kaikkien lokusten on oltava nimettyjä. Kukin lokus nimetään rivillä, jossa esitellään lokuksen tyyppi. Nimeä edeltää risuaitamerkki (#).

Karttatiedostoksi laaditaan tiedosto, jossa on yksi rivi kutakin kartassa esiintyvää lokusta kohden. Kullakin rivillä esitellään kromosomin numero, sijainti senttimorganeina ja lokuksen nimi. Tiedoston alussa on otsikoriivi, josta Mega2 pääättelee käytettävän karttafunktion. Jos toisena sanana esiintyy merkkijono KOSAMBI, soveltaa ohjelma Kosambin karttafunktiota geneettisten etäisyyksien muuttamisessa rekombinaatiofraktioiksi, muussa tapauksessa Haldanen.

Linkagepar-apuohjelma (ks. luku 3.1.3) tuottaa kaikki kolme Mega2-ohjelman vaatimaa syötetiedostoa käyttäjän antamasta sukupuutiedostosta lähtien käyt-

täjän antamien parametriarvojen perusteella. Linkagepar-ohjelman tuottama karttatiedosto `ex-1.map` näyttää esimerkkiaineistollemme seuraavanlaiselta:

CHROMOSOME	KOSAMBI	NAME
1	0.0000	M1
1	4.7726	M2
1	9.5451	M3
1	14.3177	M4
1	19.0903	M5
1	23.8629	M6
1	28.6354	M7
1	33.4080	M8

Karttatiedostossa kerrotaan, että kaikki markkerit sijaitsevat kromosomissa 1, ne sijaitsevat 4.77 senttimorganin välein (vastaa viittä Haldanen senttimorgania) ja ovat nimeltään $M1, \dots, M8$. Kartan ensimmäisen markkerin sijainniksi voidaan asettaa nolla, jolloin loput sijainnit ilmaistaan suhteessa tähän ensimmäiseen markkeriin. Kromosominumerolla 1 ei tässä yhteydessä ole merkitystä. Numerolla on käyttöä vain siinä tapauksessa, jos halutaan käsitellä samaan aikaan useita kromosomeja, jolloin niille voidaan luoda yhteinen karttatiedosto.

Ohjelmisto käynnistetään komennolla:

```
corona% mega2
```

Ohjelman käynnistyttyä näytölle ilmestyy ensimmäinen valikko, jossa annamme syötetiedostojen nimet:

```
=====
Mega2 input menu:
=====
0) Done with this menu - please proceed
1) Chromosome number:          1
2) Input file extension:       01
3) Locus datafile:             ex-1.dat
4) Pedigree datafile:          ex-1.pre
5) Map datafile:                ex-1.map
6) Omit datafile (optional):   _
7) Include all pedigrees whether typed or not
8) Set up for simulating genotyping errors: [no ]
```

Aloitusvalikossa annetaan kromosomin numero (valinta 1). Ohjelma etsii automaattisesti työskentelyhakemistossa olevia kyseiseen kromosomiin liittyviä syötetiedostoja. Näiden syötetiedostojen kromosomikohtainen pääte voidaan antaa valinnalla 2. Esimerkiksi kromosomille 1 ohjelma etsisi automaattisesti syötetiedostoja nimeltä data.in.01, ped.in.01 ja map.01. Syötetiedostojen täydelliset nimet voi vaihtoehtoisesti antaa käsin – kuten yllä olevassa esimerkissä – valitsemalla vuoron perään kohdat 3, 4 ja 5 ja kirjoittamalla analysoitavien tiedostojen nimet. Tällöin tiedostonimille ei ole mitään rajoitteita.

Valitsemalla 0 päästään seuraavaan valikkoon, jossa valitaan haluttu tiedostomuoto:

ANALYSIS MENU	
1 SimWalk2-format	15 TDTMax analyses
2 MENDEL-format	16 SOLAR-format
3 ASPEX-format	17 Vitesse-format
4 GeneHunter-Plus format	18 Linkage-format
5 GeneHunter format	19 Test loci for HWE
6 APM-format	20 Allegro-format
7 APM MULT format	21 MLBQTL format
8 Create nuclear families	22 SAGE 4.0 format
9 SLINK-format	23 Pre-makeped format
10 SPLINK-format	24 Merlin/SimWalk2-NPL
11 Homogeneity analyses	25 PREST format
12 SIMULATE-format	26 PAP format
13 Create summary files	27 Merlin format
14 SAGE-format	28 Loki format
Select an option between 1-28 >	

Mega2-ohjelma tukee useita tiedostomuotoja, ja vain osa yllä luetelluista ohjelmista on asennettu CSC:n laskentaympäristöön. Näistä keskeisiä ovat:

- 1. SimWalk2-format:** Ohjelma tekee Mendel-muotoiset syötetiedostot sekä ajotiedoston Simwalk-ohjelmaa varten.
- 5. Genehunter format:** Ohjelma kirjoittaa Linkage-muotoiset syötetiedostot sekä ajotiedoston Genehunter-ohjelmaa varten.
- 8. Create nuclear families:** Ohjelma pilkkoo suuret sukupuut trioiksi, joissa on mukana isä, äiti ja lapsi. Trioaineistoja voidaan käyttää esimerkiksi periytymisen epätasapainotestin (TDT) tekemiseksi. Ohjelmalla luodut trioaineistot ovat Linkage-formaatissa.

- 9. SLINK-format:** Ohjelma tekee syötetiedostot SLINK-simulaatio-ohjelmalle[WOL90]. SLINK-ohjelman avulla voidaan simuloida genotyypien periytymistä ehdollistaen havaituille sairausstatuksille. Ohjelmaa käytetään kytkentäanalyysin voimalaskelmissa yleensä yhdessä Linkage-ohjelmistopakettin kanssa.
- 18. Linkage-format:** Ohjelma tekee Linkage-muotoiset syötetiedostot siten, että sukupuutiedosto on post-makeped-formaatissa (vastaa Makeped-ohjelmalla käsiteltyä sukupuutiedostoa).
- 23. Pre-makeped format:** Ohjelma tekee Linkage-muotoiset syötetiedostot, joita ei ole käsitelty Makeped-ohjelmalla.
- 24. Merlin/Simwalk2-NPL:** Ohjelma tekee syötetiedostot Merlin- ja Simwalk-ohjelmia varten samoin kuin komentotiedoston, jonka avulla analyysi voidaan automaattisesti jakaa kahden ohjelmiston välille. Tällöin pienet sukupuut analysoidaan Merlin-ohjelmalla, ja saadut tulokset syötetään simwalk-ohjelmalle, joka yhdistää ne suurista sukupuista saatuihin tuloksiin.
- 27. Merlin format:** Ohjelma tekee ns. QTDT-muotoiset syötetiedostot Merlin-ohjelmaa varten. Merlin-ohjelman syöte voidaan tosin antaa myös Linkage-muodossa.

Kun haluttu analyysivaihtoehto on valittu, Mega2-ohjelma kysyy tarvittaessa kyseiseen muunnosvaihtoehtoon liittyviä lisätietoja. Tästä eteenpäin ohjelman suoritus etenee valitun vaihtoehdon mukaisesti. Esimerkiksi aineistojen muuntaminen Simwalk-muotoon (valinta 1) esitellään luvussa 6.4.

4 Tilastollisia näkökulmia

Geenikartoituksen ja genetiikan juuret ovat syvällä tilastotieteessä. Jo laskennallisen genetiikan isän Gregor Mendelin (1822-1884) tulokset perustuvat puhtaasti tilastotieteen soveltamiseen, ja tilastotieteen uranuurtaja R.A. Fisher (1890-1962) oli taustaltaan geneetikko. Tässä luvussa tarkastelemme geenikartoitusta tilastollisena ongelmana ja pohdimme tilastollisen merkitsevyyden ja voiman käsitteitä, jotka ovat geenikartoituksessa keskeisiä.

4.1 Geenikartoituksen tavoitteet ja hypoteesit

Geenikartoitusprojektin tavoitteena voi olla genomien läpikäynti, geenin sijainnin tilastollinen osoittaminen tai sen tilastollinen poissulkeminen. Tavoitteesta riippuen geenikartoitusprojektin hypoteesit tulee muotoilla eri tavoin.

Genominlaajuisessa geenikartoituksena tavoitteena on yleensä etsiä genomista lupaavia alueita, joihin tarkempi mielenkiinto kohdennetaan. Tällöin voidaan valita kustannusteoreettinen lähestymistapa, jossa perimästä etsitään laskettavan tunnusluvun tai -lukujen perusteella osajoukko, eli kokoelma alueita eri kromosomeista, johon tarkempi mielenkiinto kohdistetaan. Kyse on tällöin eksploraatiivisesta analyysistä. Tavoitteena ei ole kytkennän tai assosiaation tilastollisesti pitävä osoittaminen, vaan halutaan vain asettaa genomien alueet paremmuusjärjestykseen.

Geneettisen kytkennän tilastollinen osoittaminen perustuu kytkentä- tai assosiaatioanalyysiin. Lähtökohtaisesti kussakin analysoitavassa pisteessä tehdään tilastollinen testi, jossa nollahypoteesi asetetaan vastaamaan konservatiivista tilannetta, vallitsevaa käsitystä:

H_0 : tarkasteltava kohta (markkeri) ei ole geneettisesti kytkeytynyt fenotyyppiin.

Vaihtoehtoinen hypoteesi määritellään siten, että se poissulkee nollahypoteesin. Jos nollahypoteesi hylätään, jää vaihtoehtoinen hypoteesi voimaan:

H_A : tarkasteltava kohta (markkeri) on geneettisesti kytkeytynyt fenotyyppiin.

Siinä missä testauksen tavoitteena on yleensä geneettisen kytkennän osoittaminen, pyritään joskus myös sulkemaan pois genomien kohtia. Periaatteessa geneettinen kytkentä voidaan katsoa suljetuksi pois silloin, jos nollahypoteesi kykenee selittämään aineistosta lasketut tunnusluvun arvot huomattavasti paremmin kuin vaihtoehtoinen hypoteesi. Tällainen poissuluku on kuitenkin tilastollisesti pätevä vain, jos testauksen taustalla olevat oletukset sairauden ja kohdan (markkerin) välisestä suhteesta pätevät. Tästä suhteesta voidaan yleensä esittää vain valistuneita arvauksia.

Vaikka geenikartoitus pystytään pukemaan tilastollisiksi testeiksi, muodostaa testien suuri määrä ongelman. Kun testattavia kohtia (markkereita) on suuri määrä, kasvaa todennäköisyys, että puhdas sattuma tuottaa yhden tai useamman tapauksen, jossa nollahypoteesi hylättäisiin. Tätä kutsutaan moninkertaisen testauksen ongelmaksi, ja siihen palataan kappaleessa 4.4.

4.2 Tilastollinen merkitsevyys

Löydetyt tulokset *tilastollinen merkitsevyys* kertoo siitä, kuinka voimakkaasti tulokseen on syytä uskoa. Tilastollista merkitsevyyttä kuvataan *p-arvolla*, joka määrittää todennäköisyytenä, jolla pelkkä sattuma johtaisi nollahypoteesin hylkäämiseen. Mitä pienempi *p*-arvo on, sitä huonommin sattuma kelpaa tuloksen selittäjäksi ja sitä luotettavampi nollahypoteesin hylkäämiseen perustuva johtopäätös on.

Formaalisti tilastollista merkitsevyyttä kuvaava *p*-arvo voidaan kirjoittaa:

$$p = P(H_0 \text{ hylätään} \mid H_0 \text{ pätee}).$$

Tilastollista merkitsevyyttä kuvaava *p*-arvo tulkitaan toisinaan virheellisesti todennäköisyydeksi, että nollahypoteesi on tosi.

4.3 Tilastollinen voima

Tutkimuksen *tilastollisella voimalla* tarkoitetaan todennäköisyyttä, että koeasetelman avulla kyetään löytämään se aineistosta löydettävissä oleva tulos tai ilmiö, josta ollaan kiinnostuneita. Tilastollinen voima määrittää todennäköisyytenä, että nollahypoteesi hylätään silloin, kun se kuuluukin hylätä. Formaalisti tilastollinen voima *F* voidaan kirjoittaa:

$$F = P(H_0 \text{ hylätään} \mid H_A \text{ pätee})$$

Geenikartoituksessa tutkimuksen tilastollinen voima on ymmärrettävästi sitä parempi, mitä suurempi on otoskoko ja mitä selkeämpi taustalla oleva muuttujien (fenotyypin ja markkerien) riippuvuus on. Tilastollisen voiman suuruus riippuu kriteeristä, jolla nollahypoteesi hylätään. Jos hylkäys halutaan

tehdä vain erittäin varmoissa tapauksissa, asetetaan hylkäyksen kriteeriksi asetettu p -arvoraja, ns. *merkitsevyystaso*, alhaiseksi ja kääntäen.

Geenikartoituksessa voimalaskelmia tehdään ainakin kahdesta syystä:

1. Halutaan arvioida ennen näytteiden keräämistä ja/tai genotyypaamista, minkälaiset onnistumisen mahdollisuudet aiotulla koeasetelmalla on.
2. Jos tutkimuksesta saadaan negatiivinen tulos (alttiusgeeniä ei löydy), voidaan vakuuttua siitä, että koeasetelma olisi paljastanut geenin, jos sellainen olisi tarkasteltavalla alueella ollut edellyttäen, että sen vaikutus fenotyyppiin on oletetun kaltainen. Voimalaskelmalla pyritään tällöin osoittamaan, että negatiivinen tulos ei johdu liian pienestä tai vääränlaisesta otoksesta tai liian harvasta markkerikartasta.

Nykyään yhä useammat tieteelliset julkaisut arvostavat tai jopa edellyttävät voimalaskelmia julkaistavien tulosten yhteydessä.

4.4 Moninkertaisen testauksen ongelma

Moninkertaisen testauksen ongelma aiheutuu siitä, että koesarjassa tehdään useita tilastollisia testejä, jolloin todennäköisyys, että jokin niistä antaa tilastollisesti merkitsevän tuloksen, kasvaa. Geenikartoituksessa ongelmaa aiheuttavia tekijöitä on useita:

- Tilastollinen testi toistetaan useille markkereille tai markkeriryhmille.
- Testi toistetaan markkerin tai markkeriryhmän eri alleeleille.
- Testi toistetaan käyttäen eri fenotyyppiä.
- Testi toistetaan käyttäen erilaista oletusta fenotyypin ja genotyypin suhteesta.
- Testi toistetaan muuttamalla otokseen kuuluvia henkilöitä ja/tai markkereita.

Tunnetuin menetelmä moninkertaisen testauksen ongelman korjaamiseksi on *Bonferroni-korjaus*, jossa yksittäiset testit oletetaan toisistaan riippumattomiksi. Tällöin korjattu p -arvo kuvaa todennäköisyyttä, että ainakin yksi näistä riippumattomista testeistä antaa tilastollisesti merkitsevän tuloksen. Tarkastellaan tilannetta, jossa yksittäisessä testissä havaittu kiinnostavin p -arvo on p_0 . Tällöin todennäköisyys, että yksittäinen testi tuottaa tätä huonomman p -arvon, on $1 - p_0$, ja todennäköisyys, että toistettaessa n testiä kaikki niistä tuottavat arvoa p_0 huonomman p -arvon, on $(1 - p_0)^n$. Todennäköisyys p^* , että jokin testeistä tuottaa arvoa p_0 paremman p -arvon, on tällöin:

$$p^* = 1 - (1 - p_0)^n$$

Tämä on koko testisarjan Bonferroni-korjattu p -arvo.

Edellä todettiin, että Bonferroni-korjaus soveltuu tilanteisiin, jossa tilastolliset testit ovat toisistaan riippumattomia. Geenikartoituksessa tämä ei yleensä ole tilanne, sillä geneettisesti kytkeytyneille markkereille tehdyt testit riippuvat toisistaan. Tämä riippuvuusrakente on usein monimutkainen, ja sen analyttinen käsittely on vaikeaa. Tällöin eräs mahdollisuus on muodostaa simuloimalla nollahypoteesijakauma parhaalle yksittäisen testin antamalle p -arvolle (tai sen perustana olevalle tunnusluvulle), ja verrata todellisesta aineistosta saatua arvoa tähän jakaumaan. Esimerkiksi kytkentäanalyyysissä tarkasteltavaksi tunnusluvuksi voidaan ottaa korkeimman kytkentähuipun arvo tarkasteltavalla alueella tai koko genomissa. Nollahypoteesin mukaiset simuloituidut aineistot voidaan generoida siten, että kussakin aineistossa perustajayksilöille arvotaan alleelit niiden frekvenssijakauman perusteella, ja alleelien annetaan periytyä sukupuissa siten, että simuloidaan kromosomialueelle osuvia tekijänvaihtokohtia. Tällaisen *alleelienpudotussimulaatioiden* tuloksena saadaan keinotekoisia aineistoja, joista kukin voidaan analysoida täsmälleen samoin kuin todellinen aineisto. Kunkin analyysin tuloksena kirjataan muistiin voimakkain kytkentätunnusluvun arvo, jolloin tälle arvolle saadaan empiirinen, nollahypoteesin mukainen jakauma. Lopuksi todellisesta aineistosta saatua arvoa verrataan tähän jakaumaan. Jos todellisessa aineistossa havaittiin esimerkiksi korkein kytkentätunnusluvun arvo 2.8, ja tuhannesta simuloitusta aineistosta 13:ssa havaitaan jossain kohdassa arvo, joka on vähintään yhtäsuuri kuin 2.8, saadaan empiiriseksi, moninkertaisen testauksen suhteen korjatuksi p -arvoksi

$$p = 13/1000 = 0.013 \approx 0.01$$

Tapaus-verrokkiasetelmiin perustuvassa assosiaatioanalyyysissä korjatut p -arvot perustuvat usein niin sanottuun *permutaatiotestaukseen*, jossa sairausstatusta ilmaisevan sarakkeen arvot sekoitetaan kussakin aineistossa korttipakan tavoin. Sekoituksen tuloksena fenotyypin ja genotyyppien välinen yhteys rikkoutuu, ja aineistosta laskettavat tunnusluvut vastaavat nollahypoteesitilannetta, jossa fenotyypin ja genotyypin välinen yhteys on puhtaasti satunnainen.

Raportoidut p -arvot voidaan jättää myös korjaamatta moninkertaisen testauksen suhteen. Tällöin koeasetelma ja testausproseduuri on kuvattava erityisen tarkasti, ja lukijalle on tehtävä selväksi, että kyse on *nominaalisista* p -arvoista.

5 Parametrinen kytkentäanalyysi

5.1 Taustaa

Kytkeäanalyysin tavoitteena on kartoittaa alttiusgeenin todennäköinen sijaintipaikka tarkastelemalla fenotyyppien ja alleelien esiintymistä ja periytymistä sukupuissa.

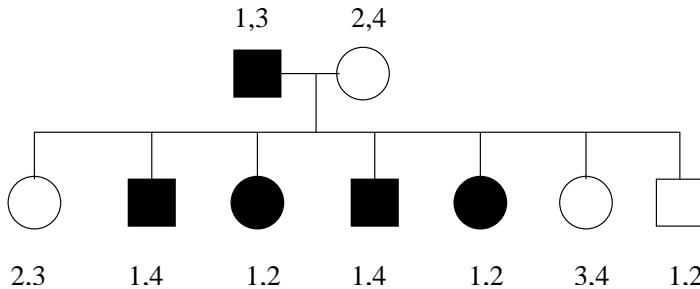
Parametrisessa kytkentäanalyysissä sairauden taustalla oleva geneettinen malli – eli penetranssivektori ja riskialleelin populaatiofrekvenssi – oletetaan tunnetuksi. Vaikka tämä vaatimus on monitekijäisille sairauksille haastava, voidaan tarkastelemalla ilmiön periytymistä sukupuissa (ns. *segregaatioanalyysi*) toisinaan silti saada karkea arvio sairauden periytymismallista tarkasteltavassa aineistossa. On myös mahdollista, että monitekijäisen sairauden ollessa kyseessä löytyy potilaiden alaryhmä, joiden suvuissa sairausfenotyyppi periytyy mendelistisesti dominantin tai resessiivisen sairausmallin mukaisesti. Jos sairausmalli voidaan päätellä, parametrinen kytkentäanalyysi näyttää kyntensä: sen tilastollinen voima on ei-parametrisia analyysimenetelmiä parempi, kun sairausmalli on määritelty oikein.

Monien muiden tilastollisten estimointitehtävien tavoin kytkentäanalyysi perustuu ns. suurimman uskottavuuden (*maximum likelihood*) menetelmään, jossa pyritään löytämään kaikkein todennäköisin arvo tuntemattomalle parametrille, tässä tapauksessa sairausgeenin sijainnille. Tällöin kytkentäanalyysin ajatellaan koostuvan seuraavista vaiheista:

1. Hypoteesien määrittäminen
2. Uskottavuusfunktion muodostaminen
3. Parametrien estimointi
4. Hypoteesien testaus
5. Tilastollisten johtopäätösten muodostaminen

Seuraavaksi tarkastelemme näitä vaiheita esimerkin avulla. Analysoimme kuvassa 5.1 esitettyä sukupuuta yhden mikrosatelliittimarkkerin suhteen. Ole-

tamme, että kyseessä on dominantti, täysin penetrantti sairaus, jossa ei esiinny fenokopioita haettavan sairauslokuksen suhteen. Alttiusgeenin ja fenotyypin välillä on siis täydellinen vastaavuus.



Kuva 5.1: Esimerkkisukupuu kytentäanalyysiä varten. Kaikkien sukupuun henkilöiden DNA-näytteistä on genotyyppattu yksi mikrosatelliittimarkkeri.

Teemme sukupuusta ja annetuista oletuksista aluksi muutaman havainnon, jotka yksinkertaistavat laskennallista käsittelyä:

1. Koska kyseessä on dominantti, täysin penetrantti sairaus, on perheen isällä riskialleeli, mutta äidillä ei sitä ole.
2. Koska perheessä on terveitä jälkeläisiä, on isä riskialleelin suhteen heterotsygootti.
3. Edellä olevan perusteella kullakin sairaalla jälkeläisellä on yksi kappaale riskialleeleita ja terveillä yksilöillä kaksi kappaletta normaaleja alleleleita.
4. Jokaisen markkeri- ja sairauslokuksen alleelin periytyminen voidaan päätellä sukupuussa yksikäsitteisesti (lukuunottamatta sitä, kummasta isän vastinkromosomista riskialleeli on lähtöisin).

5.1.1 Hypoteesien määrittäminen

Suurimman uskottavuuden menetelmässä kyse on hypoteesien testaamisesta. Kytentäanalyysissä nollahypoteesi tarkoittaa, että etsittävä sairauslokuks ei ole kytkeytynyt tarkasteltavan markkerilokuksen kanssa (kaksipisteanalyysi) tai ei ole kytkeytynyt tarkasteltavaan kytentäryhmään (monipisteanalyysi). Vastaavasti vaihtoehtoinen hypoteesi tarkoittaa, että sairauslokuks on kytkeytynyt tutkittavan markkerin tai kytentäryhmän kanssa.

Esimerkissämme on kyse kaksipisteanalyysistä. Nollahypoteesiksi määritetään nyt konservatiivinen tilanne:

$$H_0: \theta = 0.5, \quad \text{ominaisuuden vaikuttava lokus ei ole kytkeytynyt markkerilokuksen kanssa}$$

Vaihtoehtoinen hypoteesi on tällöin:

$$H_A: \theta < 0.5, \quad \text{ominaisuuteen vaikuttava lokus on kytkeyty-} \\ \text{nyt markkerilokuksen kanssa}$$

5.1.2 Uskottavuusfunktion muodostaminen

Uskottavuusfunktio määrittää todennäköisyyden havaitulle aineistolle estimoitavan parametrin funktiona. Ajatuksena on, että mitä paremmin jokin parametrin arvo kykenee selittämään havainnot, sitä luultavammin tuo parametrin arvo vastaa todellisuutta. Formaalisti uskottavuusfunktio voidaan kirjoittaa:

$$L_{D=d}(\theta) = P(d|\theta)$$

Edellä merkintä d kuvaa havaittua aineistoa, joka ajatellaan satunnaismuuttujan D havaittuna arvona, ja θ on estimoitava parametri. Erityisesti kaksipisteanalyysissä parametri θ kuvaa sairaus- ja markkerilokuksen välisen rekombinaatiofraktion arvoa.

Esimerkkimme perheelle (kuva 5.1) uskottavuusfunktio kokonaisuudessaan saa seuraavan muodon:

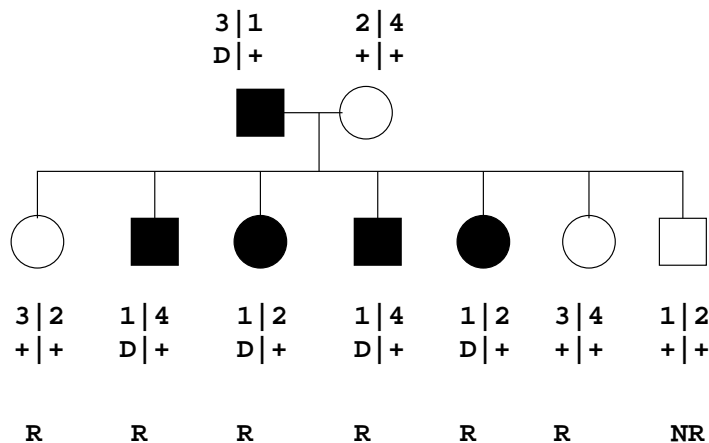
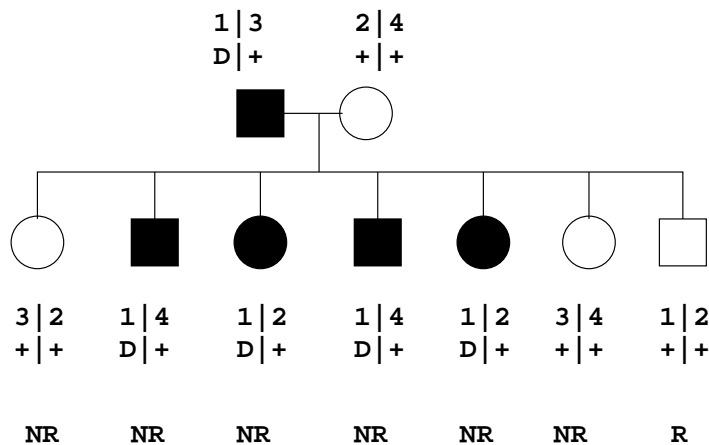
$$L(\theta) = \sum_{G_M} \sum_{G_F} \sum_{G_{C1}} \dots \sum_{G_{C7}} P(G_F)P(Y_F|G_F)P(G_M)P(Y_M|G_M) \\ \prod_{i=1}^7 P(G_{Ci}|G_F, G_M)P(Y_{Ci}|G_{Ci})$$

Edellä olevassa uskottavuusfunktion lausekkeessa merkinnät G_F ja G_M kuvaavat vanhempien mahdollisia haplotyyppikonfiguraatioita ja merkinnät G_{C1}, \dots, G_{C7} vastaavasti lasten $1, \dots, 7$ mahdollisia haplotyyppikonfiguraatioita. Haplotyyppikonfiguraatiolla tarkoitetaan tässä tapaa, jolla sairaus- ja markkerilokusten alleelit järjestäytyvät vastinkromosomeihin. Esimerkiksi sukupuun (kuva 5.1) isällä on kaksi mahdollista haplotyyppikonfiguraatiota. Markkerilokuksen alleeli 1 voi sijaita samassa vastinkromosomissa sairauslokusten riskialleelin kanssa, tai se voi sijaita samassa vastinkromosomissa sairauslokusten normaalin alleelin kanssa.

Uskottavuusfunktion lausekkeessa esiintyy todennäköisyys kullekin haplotyyppikonfiguraatiolle (esimerkiksi isälle $P(G_F)$). Isän tapauksessa kumpikin kahdesta mahdollisesta haplotyyppikonfiguraatiosta on lähtökohtaisesti yhtä todennäköinen. Fenotyyppien todennäköisyydet ehdolla haplotyyppikonfiguraatio (esimerkiksi isälle $P(G_F)P(Y_F|G_F)$) määräytyvät puolestaan suoraan annettujen penetranssien perusteella.

Lasten haplotyyppikonfiguraatioiden todennäköisyydet ehdolla vanhempien vastaavat konfiguraatiot (edellä kuvatuin merkinnöin $P(G_{Ci}|G_F, G_M)$) riip-

puvat lokusten välisestä rekombinaatiofraktiosta. Todennäköisyys, että lokusten välillä on tapahtunut rekombinaatio yhdessä meioosissa, on suuruudeltaan θ . Vastaavasti rekombinaatiota ei ole tapahtunut todennäköisyydellä $1 - \theta$. Nuo todennäköisyydet esitetään kuvassa 5.2. Kuten edellä todettiin, mahdollisia isän haplotyyppikonfiguraatiota on kaksi kappaletta. Ylempi kuva esittää toista näistä mahdollisuuksista ja alempi toista.



Kuva 5.2: Kaksi mahdollista isän haplotyyppikonfiguraatiota esimerkksikukupuuksa. Kunkin lapsen alle on merkitty, onko meioosissa täytynyt tapahtua rekombinaatio (R) vai ei (NR).

Ylemmän kuvan tilanteessa lasten haplotyyppikonfiguraatioiden selittämiseksi tarvitaan yksi rekombinaatio ja alemman kuvan tilanteessa kuusi rekombinaatiota. Lapsia — ja meiooseja — on yhteensä seitsemän. Uskottavuusfunktion laskennassa hyödynnetään sitä, että periytyminen eri lapsille on toisistaan riippumatonta. Vastaavat todennäköisyydet voidaan tällöin ker-

toa keskenään.

Ylemmän kuvan tapauksessa lasten haplotyyppikonfiguraatioiden todennäköisyys sievenee muotoon

$$\theta \cdot (1 - \theta)^6$$

ja alemman kuvan tapauksessa muotoon

$$\theta^6 \cdot (1 - \theta)$$

Uskottavuusfunktio voidaan nyt kirjoittaa muodossa

$$L(\theta) = 1/2 \cdot \theta \cdot (1 - \theta)^6 + 1/2 \cdot \theta^6 \cdot (1 - \theta)$$

Huomaa, että kaikkiin muihin vanhempien ja lasten haplotyyppikonfiguraatioihin liittyvät termit on voitu sieventää pois, koska ne ovat havaitun datan kanssa ristiriidassa. Tällöin ne kerrottaisiin nollan suuruisilla fenotyyppitodennäköisyyksillä.

Esimerkissämme uskottavuusfunktio kirjoitettiin kahdesta sukupolvesta muodostuvalle perheelle. Se voidaan yleistää mielivaltaiselle sukupuulle. Kun perheen koko kasvaa, muuttuu funktio nopeasti rakenteeltaan monimutkaisemmaksi, ja uskottavuusfunktion arvojen laskenta tehdään käytännössä numeerisesti.

5.1.3 Parametrien estimointi

Parametrien estimoinnin tavoitteena on löytää sellainen parametrin arvo, joka maksimoi uskottavuusfunktion arvon.

Yksinkertaisessa esimerkkitapauksessakin uskottavuusfunktio on seitsemännen asteen funktio. Se voitaisiin periaatteessa maksimoida johtamalla derivaattafunktio ja määrittämällä sen nollakohdat. Käytännössä estimointi tehdään numeerisesti. Tässä tapauksessa rekombinaatiofraktiolle saadaan estimaatti $\hat{\theta} = 0.143$. Tätä vastaa uskottavuusfunktion arvo $L(\hat{\theta}) = 0.02833$.

5.1.4 Hypoteesien testaus

Hypoteesien testaamisella haetaan vastausta siihen, kuinka luotettavana pidämme estimoitua parametrin arvoa. Tällöin vertamme estimaattiin liittyvää uskottavuusfunktion arvoa nollahypoteesin mukaiseen arvoon, ja katsomme, poikkeavatko ne toisistaan enemmän, kuin pelkkä sattuma antaisi aiheen olettaa.

Esimerkissämme nollahypoteesia ($\theta = 1/2$) vastaa uskottavuusfunktion arvo 0.0078125. Kytkentäanalyysissä hypoteesien testaaminen perustuu ns. *LOD scoreen* (Z_{max}) [Mor55], joka määritetään kymmenkantaisena logaritmina uskottavuusfunktioiden arvojen suhteesta:

$$Z_{max} = \log_{10} \frac{L(\hat{\theta})}{L(1/2)}$$

Tässä tapauksessa LOD score on 0.559. LOD scoreit ovat additiivisia perheiden yli, joten jos analyysi olisi toistettu samaa sairausmallia ja markkerikarttaa käyttäen usealle perheelle, voitaisiin arvot summata tässä vaiheessa.

KytKentäanalyyseissä arvon 3 ylittävää LOD scorea pidetään perinteisesti osoituksena tilastollisesti merkitsevää kytkennästä. Tämä vastaa tilastotieteessä yleisemmin käytetyn uskottavuusosamäärän testisuureen arvoa 13.8. Siitä johdettu p-arvo on suuruudeltaan 0.0002. Koska kytkentäanalyysi toistetaan yleensä suurelle määrälle markkereita, on LOD scorerajan 3 valinnalla haluttu saavuttaa tilanne, jossa nollahypoteesin hylkäämisen todennäköisyys olisi riittävän pieni myös genomilaajuisessa haussa. Esimerkkimme tapauksessa LOD score jää reilusti tuon rajan alle, ja nollahypoteesi jää voimaan.

5.1.5 Tilastollisten johtopäätösten muodostaminen

Kun hypoteesit on testattu, jää tutkijan viimeiseksi tehtäväksi tilastollisten johtopäätösten muodostaminen. Käytännössä kytkentäanalyysin tuloksen perusteella ei yleensä voida julistaa geenin olemassaoloa, vaan lopullinen verifiointi on tehtävä kaventamalla aluetta edelleen esimerkiksi assosiaatiomenetelmin ja hakemalla alueella sijaitseva vaikuttava geeni ja siinä oleva mutaatio. Tämä on yleensä huomattavan työläistä.

Vastaavasti negatiivinen tulos ei yleensä ole pitävä osoitus siitä, ettei tarkastelluilla alueilla ole alttiusgeenejä. Syynä voi olla, että tarkastellut alueet eivät ole tässä otoksessa kytkettyneet fenotyyppiin, tai otoskoko voi olla riittämätön kytkennän osoittamiseen.

5.2 Analyysin nopeudesta

KytKentäanalyyseissä tietokoneressurssien tarve on otettava huomioon, sillä suuren sukupuun analyysi tietokoneella voi kestää useita tunteja tai jopa vuorokausia. Toisaalta laskennan vaatiman muistin määrä voi pahimmillaan olla niin suuri, että tarpeet täyttävää tietokonetta ei maailmasta löydy.

Kaikkein tärkein laskenta-aikaan vaikuttava tekijä on laskenta-algoritmi. KytKentäanalyyseihin on kaksi tunnettua, tarkkaa laskenta-algoritmia: Elston-Stewart- ja Lander-Green-algoritmi. Elston-Stewart-algoritmi [ES71] mahdollistaa nopean laskennan suurille sukupuille, jos markkereita on vähän. Algoritmin tehokkuus heikkenee ratkaisevasti, jos halutaan käyttää useampaa kuin 4-5 markkeria samanaikaisesti. Elston-Stewart-algoritmi on käytössä Linkage-ohjelmistossa. Lander-Green -algoritmi [LG87], mahdollistaa nopean laskennan monipisteanalyyseissä, jossa hyödynnetään useita markkereita samanaikaisesti. Algoritmi on tehokas vain pienille sukupuille, joissa on korkeintaan 30-40 yksilöä. Tarkka laskenta-aika riippuu sukupuun rakenteesta. Algoritmi on toteutettu mm. Merlin- ja Genehunter-ohjelmistoissa. Edellisessä algoritmi on optimoitu tehokkaammin, jolloin laskenta-aika on pienempi.

Uusien tiheiden markkerikarttojen myötä Lander-Green-algoritmi on muodostunut suosittumaksi valinnaksi. Algoritmin tila- ja aikavaativuus kasvaa kuitenkin eksponentiaalisesti sukupuun koon funktiona. Tämä tarkoittaa, että tietokoneiden tehon kasvamisesta ei ole kuin marginaalista apua: todella suurten sukupuiden analysointi ei tule olemaan algoritmillä mahdollista nyt eikä tulevaisuudessa.

Jotta myös suuria sukupuita voitaisiin analysoida monen pisteen kytkentä-analyysillä, on kehitetty approksimatiivisia menetelmiä, jotka perustuvat ns. Markovin ketjuihin ja Monte Carlo -otantaan. Tähän perustuu mm. Simwalk-ohjelma. Heuristisilla menetelmillä on kuitenkin kääntöpuolensa: se mikä voitetaan tilantarpeessa ja laskenta-ajassa, menetetään tuloksen tarkkuudessa ja luotettavuudessa.

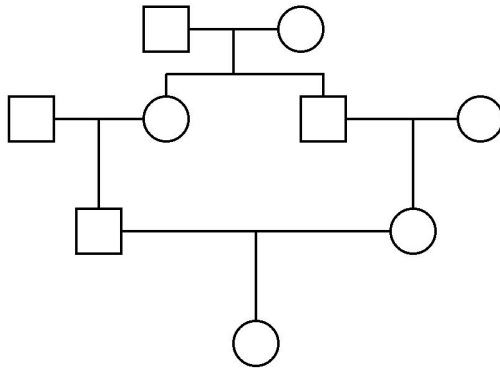
Algoritmin ohella laskenta-aika riippuu tietenkin analysoitavasta aineistosta: kuinka monta yksilöä ja sukupuuta aineisto sisältää, ovatko sukupuut suuria vai pieniä, ja ovatko perherakenteet yksinkertaisia vai onko niissä silmukoita (ks. kuva 5.2). Myös analysoitavien markkerien määrä aineistossa vaikuttaa laskentanopeuteen: kuinka monta markkeria halutaan testata yhtä aikaa, eli halutaanko käyttää kaksipiste- vai monipisteanalyysiä ja kuinka monelta yksilöltä puuttuu markkerihavainnot. Yleensä suuret sukupuut, joissa markkeriaineistoa on vain parin viimeisen sukupolven yksilöistä, ovat laskennallisesti raskaita, koska analyysialgoritmin on käytävä läpi kaikki markkerigenotyypimahdollisuudet, joita puuttuvan havainnon haltijoilla voisi olla.

Myös tietokoneen suorittimen nopeus vaikuttaa laskenta-aikaan, mutta vain vakiokertoimen verran. Muistin määrä voi sen sijaan muodostua pullonkaulaksi. Esimerkiksi Lander-Green-algoritmin vaatima muistitila kasvaa laskenta-ajan tavoin eksponentiaalisesti sukupuun koon kasvaessa. Jos muisti loppuu kesken, laskenta epäonnistuu, tai sukupuusta joudutaan karsimaan pois vähiten informatiivisia yksilöitä.

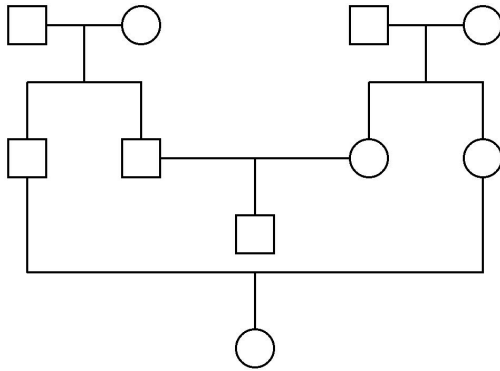
5.3 Analyysi Genehunter-ohjelmalla

Genehunter-ohjelma soveltuu parametriseen kytkentäanalyysiin tilanteissa, joissa sukupuiden koko ei ole erityisen suuri. Toisaalta ohjelman käyttämän Lander-Green-algoritmin ansiosta laskenta-aika kasvaa ainoastaan lineaarisesti markkerien määrän funktiona, joten ohjelma soveltuu tilanteisiin, joissa analysoitavia markkereita on paljon, jopa satoja tai tuhansia. Koska sekä parametrinen että ei-parametrinen kytkentäanalyysi tehdään Genehunter-ohjelmalla samalla tavalla, käsittelemme ohjelman käyttöä tarkemmin luvussa 6.2 ei-parametriseen kytkentäanalyysin yhteydessä. Tutustumme seuraavaksi Linkage-ohjelmistoon, joka on geenikartoitusohjelmista tunnetuin ja yhä yleisesti käytössä.

Sukulaisuussilmukka (*inbreeding loop, consanguinity loop*):



Aviosilmukka (*marriage loop*):



Kuva 5.3: Aineistossa olevat silmukat.

5.4 Analyysi Linkage-ohjelmistolla

Linkage-ohjelma[LLJO84, LL84, LLW86] soveltuu parametriseen kytkentä-analyysiin. Se käyttää laskennassa Elston-Stewart -algoritmia, jossa laskenta-aika kasvaa lineaarisesti suhteessa sukupuiden kokoon mutta eksponentiaalisesti suhteessa markkereiden määrään. Niinpä ohjelmalla voidaan analysoida suuriakin sukupuita, mutta markkereita voidaan ottaa mukaan analyysiin vain muutama kerrallaan. Ohjelmaa kannattaa siis käyttää tilanteissa, joissa sukupuut ovat suuria, mutta halutaan analysoida vain yhtä tai korkeintaan muutamaa markkeria kerrallaan. CSC:lle on asennettu ohjelmiston perusalgoritmeja korvaava, jonkin verran nopeampi Fastlink-toteutus[JIS93, SGSJ94].

Linkage-ohjelmisto apuohjelmineen muodostuu joukosta eri tarkoituksiin kehitettyjä ohjelmia, joita kutsutaan peräkkäisessä järjestyksessä. Näitä ohjelmia ovat (katso kuvaa 5.4):

Makeped: muuntaa sukupuuaineiston Linkage-ohjelmalle sopivaan muotoon

Unknown: määrittää mahdolliset genotyypiyhdistelmät yksilöille, joilta puuttuu markkereita

Mlink: laskee LOD scoreit käyttäjän määrittämällä rekombinaatiofraktion arvoilla

Ilink: etsii sen rekombinaatiofraktion arvon, joka maksimoi uskottavuusfunktion arvon

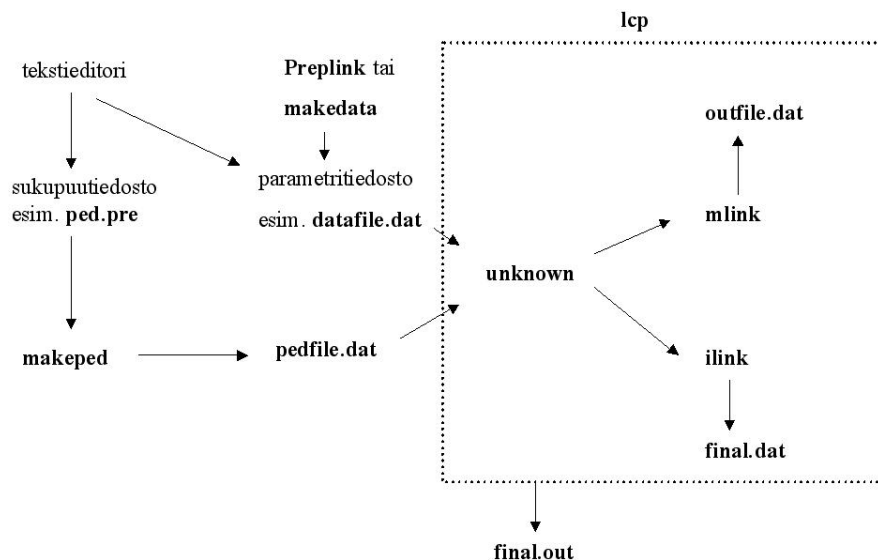
Linkmap: tekee monipisteanalyysin tapauksessa, jossa lokusta verrataan usean markkerin muodostamaan kiinteään karttaan

Lcp: apuohjelma, jonka avulla voidaan käynnistää analyysitehtäviä

Lrp: ohjelma, joka tulostaa kytkentätuloksesta havainnollisemman raportin

Tyypillisimmillään kytkentäanalyysin teko Linkage-ohjelmistolla koostuu CSC:n laskentaympäristössä seuraavista työvaiheista:

1. Sukupuutiedosto laaditaan tekstieditorilla. Tiedosto voidaan myös tuoda tekstimuodossa vaikkapa Excel-taulukosta. Valmis tiedosto on syytä tarkastaa Pedcheck-apuohjelmalla (ks. luku 3.1.2).
2. Sukupuutiedosto muunnetaan makeped-ohjelmalla analyysiohjelmien ymmärtämään muotoon.
3. Parametritiedosto tehdään joko tekstieditorilla tai esim. Linkagepar-ohjelmalla (ks. luku 3.1.3). Ohjelman avulla kuvataan käytettävä sairasmalli ja markkerikartta.
4. Halutut analyysit määritellään Lcp-ohjelmalla. Ohjelma tuottaa tulosteena komentotiedoston.



Kuva 5.4: Kaavio Linkage-ohjelmistoon kuuluvista ohjelmista ja apuohjelmista sekä syöte- ja välitiedostoista.

5. Analyysit suoritetaan ajamalla komentotiedosto, joka muuntaa tarpeen mukaan syötetiedostoja ja kutsuu sukupuun tarkistusohjelmaa unknown sekä kytkentäanalyysiohjelmia Mlink ja Ilink.
6. Tulostiedostoa tarkastellaan Lrp-ohjelman avulla.

Kohtien 4 ja 5 sijasta voidaan toimia myös siten, että Unknown-, Mlink- ja Ilink-ohjelmia kutsutaan käsin. Tällöin tehtävä analyysi määritellään parametritiedoston yhteydessä.

Tutustumme seuraavaksi Makeped-muunnosohjelman käyttöön. Ohjelmaa tarvitaan sukupuutiedoston muuntamiseksi Linkage-ohjelmien ymmärtämään muotoon.

5.4.1 Sukupuutiedoston muuntaminen Makeped-ohjelmalla

Makeped-apuohjelmaa käytetään sukupuutiedoston muuntamiseen ns. post-makeped -tiedostomuotoon. Esimerkiksi Linkage-kytkentäanalyysiohjelmisto odottaa saavansa syötteen tässä muodossa. Käytettäessä pelkästään Genehunter- ja Merlin-ohjelmia voidaan tämä vaihe ohittaa.

Sukupuutiedoston käsitellään Makeped-ohjelmalla näin:

```
corona% makeped ex-1.pre ex-1.ped n
```


1. Hankitaan Mlink-ohjelman avulla yleiskuva siitä, miten uskottavuusfunktion arvo muuttuu rekombinaatiofraktion funktiona.
2. Etsitään uskottavuusfunktion maksimiarvo sekä sitä vastaavat rekombinaatiofraktion arvo ja LOD score Ilink-ohjelmalla.

Lcp-ohjelman avulla määritellään, minkälaisia analyysejä aineistolle halutaan tehdä. Käynnistetään Lcp-ohjelma:

```
corona% lcp
```

Aluksi avautuu päävalikko:

COMMAND file name [pedin] :	pedin
LOG file name [final.out] :	final.out
STREAM file name [stream.out] :	stream.out
PEDIGREE file name [pedin.dat] :	ex-1.dat
PARAMETER file name [datain.dat] :	ex-1.ped
Secondary PEDIGREE file name [] :	
Secondary PARAMETER file name [] :	

Käyttäjä määrittelee tässä valikossa ohjelman syöte- ja tulostiedostot. Ensimmäisellä rivillä määritellään sen komentotiedoston nimi, johon Lcp-ohjelma tallentaa analyysiohjelmien kutsusarjan. Kun käyttäjä poistuu Lcp-ohjelmasta, hän kutsuu tätä komentotiedostoa, joka suorittaa analyysin eri osat alusta loppuun.

Ylös ja alas voi siirtyä kohdistinnäppäimillä. Valikosta toiseen pääsee PageDown-näppäimellä tai näppäinyhdistelmällä Ctrl-n. Päävalikosta eteenpäin siirryttäessä ohjelma tarkastaa, että sukupu- ja parametritiedostot ovat olemassa.

Seuraavaksi ilmestyy valikko Pedigree options:

General pedigrees :	<-
Three-generation pedigrees :	
Experimental cross pedigrees :	

Tässä valitaan sukupuiden tyyppi. Ylin valinta sopii rakenteeltaan mielivaltaisille sukupuille.

Valikko General pedigree analysis options sisältää nyt eri analyysivaihtoehdot:

LODScore :
ILINK :
LINKMAP :
MLINK : <-

Käytössä on neljä ohjelmaa:

1. **Lodscore**-ohjelma laskee parittaiset LOD scoren -arvot kahden lokusjoukon kaikkien lokusten välillä. Ohjelmaa ei käsitellä tässä esimerkissä.
2. **Ilink** etsii iteratiivisesti sellaisen rekombinaatiofraktion arvon, joka tuottaa mahdollisimman korkean LOD scoren. Saatava rekombinaation arvo on siis suurimman uskottavuuden arvio (maximum likelihood estimate, MLE) rekombinaatiofraktiolle.
3. **Linkmap** on ohjelma monipistekytkenälyysin tekemiseksi. Monipisteanalyysissä sairauslokusta ei verrata ainoastaan yhteen vaan useampaan markkeriin samanaikaisesti. Menetelmällä saavutetaan hyötyä, jos yksittäisten markkerien informatiivisuus on puutteellinen.
4. **Mlink** laskee kiinteisiin rekombinaatiofraktion arvoihin liittyvät uskottavuusfunktion arvot ja LOD scoren. Käyttäjä määrittelee Mlink-ohjelmalle rekombinaatiofraktion alkuarvon, loppuarvon ja askeleen suuruuden. Lcp-ohjelman avulla Mlink-ohjelmaa voi kutsua myös aivan vapaavalintaisella joukolla rekombinaatiofraktion arvoja.

Koska haluamme aluksi laskea tiettyjä, kiinteitä rekombinaatiofraktion arvoja vastaavat LOD scoren, valitsemme Mlink-ohjelman. Tämän jälkeen ohjelma kysyy, millä tavoin rekombinaatiofraktion arvot valitaan.

Specific evaluation : <-
Lod score table :
Multiple pairwise Lod table :

Vaihtoehtoista ylin, **Specific evaluation**, vastaa tilannetta, jossa Mlink-ohjelmaa kutsuttaisiin suoraan komentoriviltä. Käyttäjä määrittelee tällöin rekombinaatiofraktion alkuarvon, loppuarvon ja askelvälin. Toinen vaihtoehto, **Lod score table**, sallii käyttäjän luetella vapaavalintaisen joukon rekombinaatiofraktion arvoja.

Valitsemme aluksi ylimmän vaihtoehdon. Seuraavasta valikosta valitsemme ainoan vaihtoehdon:

No sex difference : <-

Tämän jälkeen kerromme ohjelmalle, mitä lokuksia verrataan toisiinsa, ja mitä rekombinaatiofraktion arvoja käytetään:

Locus order []	: 1 2
Recombination fractions [.1]	: .0
Recombination varied [1]	: 1
Increment value [.1]	: .1
Stop value [.5]	: .5

Kohtien merkitys on seuraava:

- Kohta `Locus order` määrittää testattavat lokukset. Esimerkissämme valinta `1 2` kertoo, että sairauslokusta (1) halutaan verrata ensimmäiseen markkerilokukseen (2).
- Kohdassa `Recombination fractions` määritetään alaraja rekombinaatiofraktiolle.
- Kohdassa `Recombination varied` arvo `1` kertoo, että rekombinaatiofraktion arvoa halutaan vaihtaa askeltaen.
- Kohdassa `Increment value` määritetään rekombinaatiofraktion askelväli.
- Kohdassa `Stop value` määritetään suurin testattava arvo rekombinaatiofraktiolle. Ohjelma laskee joka tapauksessa rekombinaatiofraktion arvoa `0.5` vastaavan uskottavuusfunktion arvon, joten tässä kohdassa voisi olla pienempikin arvo.

Ylläolevassa esimerkissä käytetään rekombinaatiofraktion arvoja `0`, `0.1`, `0.2`, `0.3`, `0.4`, ja `0.5`. Kun arvot ovat kohdallaan, käyttäjä painaa `PageDown`-näppäintä, jolloin ohjelma lisää analyysin tiedot komentotiedostoon. Tässä kohdin ohjelmisto on suunniteltu hieman heikosti, sillä käyttäjä ei saa painalluksesta minkäänlaista palautetta.

Haluamme tehdä kaksipisteanalyysin `Mlink`-ohjelmalla myös sairauslokuksen ja muiden markkereiden suhteen. Tätä varten vaihdetaan `Locus order`-kohtaan arvoksi `1 3` ja painetaan `PageDown`-näppäintä. Seuraavaksi annetaan arvoksi `1 4` ja painetaan `PageDown`, minkä jälkeen käydään läpi muut markkerit siten, että viimeisenä arvona annetaan `1 9`. Näin olemme määritelleet kahdeksan erillistä kaksipisteanalyysiä, jotka suoritetaan peräjälkeen.

Lopuksi `Lcp`-ohjelmasta poistutaan näppäinyhdistelmällä `Ctrl-z`. `Lcp`-ohjelma on nyt kirjoittanut analyysikomennot tiedostoon nimeltä `pedi.n`. Analyysi käynnistetään ajamalla kyseinen komentotiedosto:

```
corona% pedi.n
```

Analyyysin tulokset tulostuvat tiedostoon final.out. Tulokset alkavat seuraavasti:

```

...
Locus Order          : 1 2
...
LINKAGE (V5.1) WITH 2-POINT AUTOSOMAL DATA
ORDER OF LOCI:      1 2
-----
THETAS 0.500
-----
PEDIGREE | LN LIKE | LOG 10 LIKE
-----
          1 -41.387416 -17.974288 LOD=      0.000000
-----
TOTALS      -41.387416 -17.974288
-2 LN(LIKE) = 8.27748e+01 LOD SCORE =      0.000000
-----
THETAS 0.000
-----
PEDIGREE | LN LIKE | LOG 10 LIKE
-----
          1 -38.032712 -16.517362 LOD=      1.456926
-----
TOTALS      -38.032712 -16.517362
-2 LN(LIKE) = 7.60654e+01 LOD SCORE =      1.456926
-----
THETAS 0.100
-----
PEDIGREE | LN LIKE | LOG 10 LIKE
-----
          1 -38.786277 -16.844630 LOD=      1.129658
-----
TOTALS      -38.786277 -16.844630
-2 LN(LIKE) = 7.75726e+01 LOD SCORE =      1.129658
-----

```

Tulosten alussa kohdassa Order of loci nähdään, että analyysi on tehty aluksi sairauslokuksen (1) ja ensimmäisen markkerilokuksen (2) välillä. Ohjelma laskee aluksi rekombinaatiofraktion arvoa 0.5 vastaavan uskottavuusfunktion arvon, jota tarvitaan LOD scoren laskennassa. Tämän jälkeen rekombinaatiofraktion arvoa liu'utetaan ja kutakin arvoa kohti tulostetaan uskottavuusfunktion arvo sekä LOD score. Esimerkiksi rekombinaatiofraktion arvoa 0 vastaa LOD score 1.46.

Tuloksista saadaan kauniimpi raportti Lrp-ohjelman avulla. Raportin tuottamiseen tarvittavat tiedot ovat tiedostossa stream.out, joka syntyi ajetun analyysin aikana. Lrp-ohjelman käyttö on helppoa. Ohjelma käynnistetään komentoriviltä:

```
corona% lrp
```

Raportille annetaan haluttu otsikko:

```
STREAM file name [stream.out] : stream.out
```

```
REPORT title [] : Kaksipisteanalyysi
```

Eteenpäin siirrytään — kuten Lcp-ohjelmassakin — näppäimellä [PageDown](#). Seuraavaksi kerrotaan, että kyseessä olivat tavalliset sukupuut:

```
General pedigree reports : <-
```

```
Three-generation pedigree reports :
```

```
Experimental cross pedigree reports :
```

```
Full stream file reports :
```

Haluamme LOD score -taulukon Mlink-ohjelman tulosteista:

```
Two-point lodscore report (LODScore) :
```

```
Multi-point order report (ILINK) :
```

```
Location score report (LINKMAP) :
```

```
Lod table report (MLINK) : <-
```

Seuraavaksi voidaan määritellä tulostusasua. Haluamme tulosteen taulukko-
na kuvaruudulle kymmenkantaiseen log-uskottavuusfunktioon perustuen
(LOD score). Koska mukana on vain yksi sukupuoli, ei tuloksia tässä tapauk-
sessa tarvitse eritellä sukupuoli sukupuulta.

```
Table format : <-
```

```
Full format :
```

```
Display LOG 10 results [Yes] : Yes
```

```
Display LOG e results [No] : No
```

```
Include Pedigrees [No] : No
```

```
Output report to the screen : <-
```

```
Output report to a file :
```

Nyt LOD scoreit nähdään muotoiltuna taulukkona:

L O D T A B L E R E P O R T							
File: stream.out	Kaksipisteanalyysi					Screen: 1 of 1	
Order	0.0	0.1	0.2	0.3	0.4	0.5	
1=2	1.46	1.13	0.79	0.45	0.15	0.00	a
1=3	1.15	0.87	0.59	0.32	0.09	0.00	a
1=4	1.40	1.08	0.75	0.41	0.13	0.00	a
1=5	0.86	0.62	0.40	0.20	0.06	0.00	a
1=6	-0.04	-0.02	-0.01	-0.00	0.00	0.00	a
1=7	1.39	1.05	0.71	0.38	0.11	0.00	a
1=8	1.67	1.32	0.96	0.58	0.21	0.00	a
1=9	1.16	0.93	0.66	0.38	0.13	0.00	a
= = Test Interval							
a = LOD Scores				(LOG 10)			
b = LOG 10 Likelihoods				(LOG 10)			

Tulosteesta nähdään, että käytetyllä rekombinaatiofraktion askelvälillä sairauslokukseen (1) ja toiseksi viimeisen markkerin (8) välinen LOD scoren maksimi-arvo on 1.67, jota vastaa rekombinaatiofraktion arvo 0.

Seuraavaksi tarkennamme rekombinaatiofraktion estimaattia käyttämällä ohjelmaa Ilink, jolloin uskottavuusfunktion arvoa ei lasketa ennalta määrättyissä pisteissä, vaan ohjelma etsii sen pisteen, jossa maksimi-arvo saavutetaan. Tuo piste on rekombinaatiofraktion suurimman uskottavuuden arvio.

Jotta analyysien teko olisi helpompaa, käytämme jälleen apuna Lcp-ohjelmaa. Käynnistämme Lcp-ohjelman ja syötämme vastaavat tiedot kuin edelläkin, mutta valitsemme käytettäväksi ohjelmaksi tällä kertaa Ilink-ohjelman:

LODScore :
ILINK : <-
LINKMAP :
MLINK :

Viimeisellä sivulla, jossa lokukset määritellään, ohjelma kysyykin vain yhtä rekombinaatiofraktion arvoa. Tähän ohjelmalle annetaan rekombinaatiofraktion alkuarvo, josta maksimointi aloitetaan:

Locus order [] : 1 2
Recombination fractions [.1] : .1

Analyysin tiedot kirjoitetaan jälleen komentotiedostoon painamalla näppäintä [PageDown](#). Tämän jälkeen vaihdetaan testattaviksi lokuksiksi 1 3, sen

jälkeen 1 4 ja näin jatkaen arvoon 1 9 asti. Tämän jälkeen Lcp-ohjelmasta poistutaan näppäinyhdistelmällä [Ctrl-z]. Komentotiedosto ajetaan jälleen komennolla:

```
corona% pedin
```

Esimerkiksi sairauslokuksen (1) ja toiseksi viimeisen markkerilokuksen osalta tulokset (tiedostossa final.out) näyttävät tältä:

```

Locus Order           : 1 8
Male Recomb. Fractions : 0.1000000
*****
CHROMOSOME ORDER OF LOCI :
 1 2
*****
THETAS:
 0.001
*****
-2 LN(LIKE) = 6.45798e+01
LOD SCORE = 1.66509e+00
NUMBER OF ITERATIONS = 3
NUMBER OF FUNCTION EVALUATIONS = 9
PTG = -4.20175e+01
*****
*****

```

Tulosteesta selviää, että suurimman uskottavuuden arvio rekombinaatiofraktion arvolle on 0.001, ja maksimoitu LOD score on suuruudeltaan 1.67. Myös Ilink-ohjelmalla tehtyjen analyysien tuloksia voidaan katsella raportointiohjelmalla Lrp.

5.4.3 Monipisteanalyysi

Monipisteanalyysissä sairauslokusta verrataan usean markkerin muodostamaan kytkentäryhmään. Laskenta on tällöin hitaampaa, mutta markkereiden informaatiota käytetään tehokkaammin hyväksi, mikä saa usein aikaan paremman paikannustuloksen. Monipisteanalyysi voidaan kaksipisteanalyysin tapaan määritellä Lcp-apuohjelmalla. Tällöin käytettäväksi ohjelmaksi määritetään Linkmap. Koska Elston-Stewart-algoritmi hidastuu analysoitaessa useita markkereita samanaikaisesti, otetaan analyysiin mukaan vain muutama markkeri kerrallaan. Taulukko 5.1 näyttää, miten sairauslokusta (X) voidaan liu'uttaa esimerkkitapauksessamme yhdeksästä markkerista (M1-M9) koostuvan kartan yli siten, että analysoitavana on kussakin vaiheessa neljä

vierekkäistä markkeria. Taulukon osoittamat lokus- ja intervallinumerot syötetään Lcp-apuohjelmassa seuraavasti (esimerkki vastaa taulukon 5.1 ylintä riviä):

Test loci [] : 1
Order of fixed loci [] : 2 3 4 5
Recombination fractions [.1] : .0476 .0476 .0476
Test interval [0] : 0
Number of evaluations in interval [5] : 5

Esimerkissä kahden vierekkäisen markkerin väliseksi rekombinaatiofraktioksi on syötetty 0.0476 (vastaa n. 5 cM geneettistä etäisyyttä). Arvoista ensimmäinen vastaa kahden vasemmanpuoleisen ikkunaan kuuluvan markkerin välistä rekombinaatiofraktiota. Neljää markkeria vastaa näin ollen kolme rekombinaatiofraktion arvoa. Viimeinen kohta (Number of evaluations in interval) kertoo, monessako askeleessa sairauslokusta liu'utetaan kunkin intervallin sisällä.

Kunkin analyysimäärityksen jälkeen painetaan **PageDown**-näppäintä ja syötetään seuraava analyysivaihe (taulukon 5.1 seuraava rivi).

Taulukko 5.1: *Esimerkki monipisteanalyysistä Linkmap-ohjelmalla. Sairauslokusta (X) liu'utetaan yli yhdeksästä markkerista (M1-M9) koostuvan kartan siten, että analyysissä on samanaikaisesti mukana neljä vierekkäistä markkeria. Kullakin rivillä esitetään myös analyysivaihetta vastaavat Lcp-apuohjelman kenttiin syötettävät arvot.*

Analysoitavat lokukset	Test loci	Order of fixed loci	Test interval
X-M1---M2---M3---M4--	1	2 3 4 5	0
--M1-X-M2---M3---M4--	1	2 3 4 5	1
--M1---M2-X-M3---M4--	1	2 3 4 5	2
--M2---M3-X-M4---M5--	1	2 3 4 5	2
--M3---M4-X-M5---M6--	1	3 4 5 6	2
--M4---M5-X-M6---M7--	1	4 5 6 7	2
--M5---M6-X-M7---M8--	1	5 6 7 8	2
--M6---M7-X-M8---M9--	1	6 7 8 9	2
--M6---M7---M8-X-M9--	1	6 7 8 9	3
--M6---M7---M8---M9-X	1	6 7 8 9	4

Ohjelma tulostaa kullekin sijainnille uskottavuusosamäärään perustuvan location scoren . Location score muutetaan LOD scoreksi jakamalla se vakiolla $2 \ln 10 \approx 4.605$. Tulosten taulukoimiseksi voidaan käyttää Lrp-apuohjelmaa.

6 Ei-parametrinen kytkentäanalyysi

6.1 Taustaa

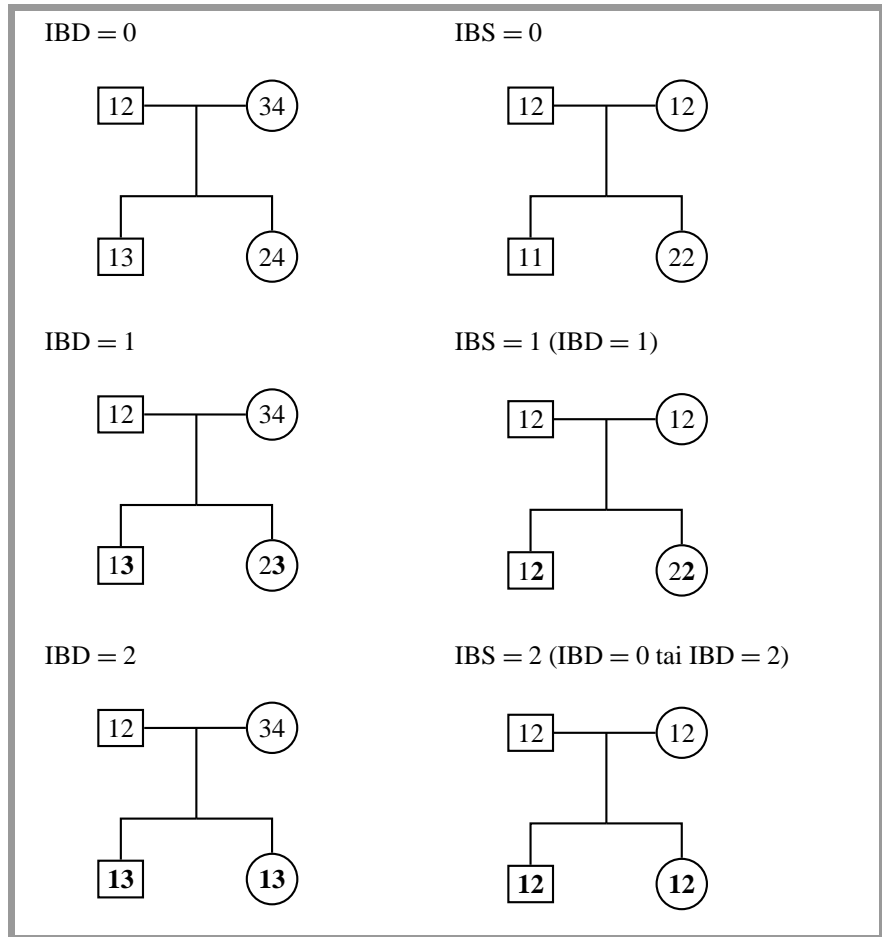
Parametrisessa kytkentäanalyysissä määritellään sairauden periytymismalli, joka koostuu penentransseista ja riskialleelien frekvensseistä populaatiossa. Ei-parametrisessa kytkentäanalyysissä (*non-parametric linkage analysis, model free linkage analysis*) periytymismallia ei yksiselitteisesti määritetä, vaan tutkitaan, missä kohdissa saman fenotyypin omaavilla sukulaisilla on yhteistä alkuperää olevia genomien kohtia ja keillä sukulaisilla nuo yhteiset kohdat esiintyvät.

Kun tarkastellaan yksilöille yhteisiä genomien kohtia, puhutaan ns. IBS-jakamisesta (*sharing identity-by-state*) ja IBD-jakamisesta (*sharing identity-by-descent*). IBS-jakamisella tarkoitetaan sitä, että yksilöillä on samassa markerikohdassa sama (samannumeroinen) alleeli. IBD-jakamisessa vaaditaan lisäksi, että tuo alleeli on yhteistä esivanhempaisalkuperää. IBD-jakamisen ehtona on siis, että yksilöillä on yhteinen esivanhempi, jonka samasta vastinkromosomista molemmat ovat tuon yhteisen alleelin perineet.

Yksinkertaisin ei-parametrinen kytkentäanalyysi perustuu sairaiden sisarusparien analysointiin [Pen35]. Tällöin arvioidaan, jakavatko sairastuneet sisarusparit odotettua enemmän alleeleita IBD jossakin markerilokuksessa. Tarkastelun kohteena ovat IBD-statukset siitä syystä, että niistä saadaan enemmän informaatiota kuin IBS-statuksista. Kuvassa 6.1 on esitetty eroavaisuuksia IBD- ja IBS-statusten välillä. Näkyvin ero on viimeisessä tapauksessa, jossa kummatkin vanhemmat sekä lapset ovat heterotsygootteja samojen alleelien suhteen. Tässä tapauksessa ei tiedetä, onko sisarusten alleelit 1 peritty samalta vanhemmalta vai onko toinen sisaruksesta perinyt alleelin 1 isältään ja toinen äidiltään.

6.1.1 Sairaiden sisarusparien menetelmä

Sairaiden sisarusparien menetelmä (*affected sib pair method, ASP*) on monen muun menetelmän tavoin tehokkaimmillaan resessiivisesti periytyvien sairauksien geenien paikallistamisessa. Sitä käytetään myös monitekijäisten



Kuva 6.1: Lasten IBD- ja IBS-statusia nelihenkisissä perheissä.

sairauksien geenikartoituksessa, koska se ei vaadi periytymismallin määrittämistä. Monet ei-parametriset menetelmät ovatkin tavalla tai toisella ASP-menetelmän laajennuksia. ASP-testissä käytettävä aineisto koostuu joukosta perheitä, joissa on kaksi sairasta sisarusta sekä heidän yhteiset vanhempansa. Tällaisia nelikkoja kerätään suuri joukko. Yleisesti myös oletetaan näiden nelikoiden olevan toisistaan riippumattomia, jolloin ne eivät ole välitöntä sukua keskenään. Jotta menetelmä olisi mahdollisimman tehokas, saa puuttuvia genotyyppisiä olla vain vähän. Tehokkuutta lisää myös, jos vanhemmat ovat heterotsygootteja markkereiden suhteen siten, että lasten välinen IBD voidaan selvittää.

Testisuureen laskeminen alkaa IBD-statusen määrittämisestä kaikille sisaruspareille; tällöin tutkitaan, jakavatko sisarusparit 0, 1, vai 2 alleelia IBD. Tämän jälkeen merkitään niiden sisarusparien, joissa IBD-status on 0, lukumäärää n_0 :lla, ja vastaavasti n_1 :llä ja n_2 :llä niiden sisarusparien lukumäärä, jotka jakavat 1 tai 2 alleelia IBD. Näiden lukumäärien avulla lasketaan ASP-

testisuureen arvo S :

$$S = \sum_{i=0}^2 \frac{(n_i - e_i)^2}{e_i} = \frac{(n_0 - e_0)^2}{e_0} + \frac{(n_1 - e_1)^2}{e_1} + \frac{(n_2 - e_2)^2}{e_2},$$

missä $e_0 = 0.25n$, $e_1 = 0.5n$ ja $e_2 = 0.25n$. Testisuure S noudattaa nollahypoteesitilanteessa χ^2 -jakaumaa kahdella vapausasteella eli $S \sim \chi^2_2$.

Kuvassa 6.2 havainnollistetaan, mistä testisuureen lukumäärät tulevat nollahypoteesin vallitessa ($e_0 = 0.25n$, $e_1 = 0.5n$ ja $e_2 = 0.25n$, missä n on havaintojen kokonaismäärä eli $n = n_0 + n_1 + n_2$). Oletetaan, että isän alleelit ovat 1 ja 2 ja äidin alleelit ovat 3 ja 4, jolloin voidaan yksiselitteisesti määrittää sisarususten välinen IBD-status. Tällöin kummallakin sisaruksella on neljä erilaista genotyypivaihtoehtoa: 13, 14, 23 ja 24. Näistä saadaan 16 erilaista sisarusgenotyypikombinaatiota esim. 13 ja 13, 14 ja 24 jne.

Kun kaikille näille kombinaatioille määritetään IBD-status, nähdään, että 25 % sisarususten genotyypiyhdisteistä antaa IBD-statuksiksi 0. Samoin nähdään, että 50 % antaa IBD-statuksiksi 1 ja 25 % statuksen 2. Toisaalta voidaan katsoa myös, kuinka monta alleeliparia on IBD eli peritty samalta vanhemmalta. Näiden osuus on 50 % ($(0.5 \times 1 + 0.25 \times 2)/2$).

χ^2 -testin lisäksi käytetään ns. keskiarvotestiä (*mean test*), jossa testataan onko IBD-alleeliparien lukumäärä suurempi kuin puolet kaikista alleelipareista. Testisuure tässä tapauksessa on:

$$S = \frac{(0.5n_1 + n_2) - 0.5n}{\sqrt{n/8}},$$

missä n on sisarusparien kokonaislukumäärä, $0.5n$ on lausekkeen $(0.5n_1 + n_2)$ odotusarvo ja $n/8$ sen varianssi nollahypoteesin vallitessa. Testisuure noudattaa standardia normaalijakaumaa. Koska testataan $(0.5n_1 + n_2) > 0.5n$, testi on yksisuuntainen.

6.1.2 Menetelmän laajennukset

Menetelmää voidaan laajentaa siten, että puuttuvien havaintojen (markerialleelitiedot puuttuvat esim. toiselta vanhemmalta) aiheuttamaa informaatiohukkaa korvataan laskemalla todennäköisyydet eri IBD-statuksille käyttäen hyväksi suvun muiden henkilöiden (kuten sisarususten ja isovanhempien) genotyypitietoja.

Toinen tapa laajentaa testiä on huomioida kaikki perheen sairaat sisarusukset kahden sisarusksen sijasta. Testiä voidaan edelleen yleistää ottamalla huomioon sisarususten sijasta sukuun kaikki sairaat yksilöt. Tällöin puhutaan ns. *extended relative pair analysis (ERPA)* -menetelmästä.

Genehunter- ja Merlin-ohjelmissa menetelmää on laajennettu usealla tavalla: ne sallivat puuttuvia havaintoja ja tarkastelevat samaan aikaan mielivaltaisen monen sairaan sukulaisen IBD-jakamista. Lisäksi ne yhdistävät eri markkereista saatavan informaation (monipisteanalyysi).

Tautilokus ei ole kytkeytynyt markkerilokuksen kanssa, joten aineistossa on odotusarvoisesti 25 % sisaruspareja, jotka eivät jaa yhtään alleelia keskenään, 50 %, jotka jakavat yhden alleelin keskenään ja 25 %, jotka jakavat kaksi alleelia IBD keskenään.

		Sisarus 1		Sisarus 2	
		Isän gameetit		Isän gameetit	
		1	2	1	2
Äidin gameetit	3	1 3	2 3	1 3	2 3
	4	1 4	2 4	1 4	2 4

Sisarusyhdistelmät:

			IBD
1 3	1 3	2	
1 3	2 3	1	
1 3	1 4	1	
1 3	2 4	0	
2 3	1 3	1	
2 3	2 3	2	
2 3	1 4	0	
2 3	2 4	1	
1 4	1 3	1	
1 4	2 3	0	
1 4	1 4	2	
1 4	2 4	1	
2 4	1 3	0	
2 4	2 3	1	
2 4	1 4	1	
1 4	2 4	2	

IBD=0 : $4/16 = 0.25$

IBD=1 : $8/16 = 0.5$

IBD=2 : $4/16 = 0.25$

Kaikkiaan alleelipareja 32. Näistä
16 on IBD eli 50 %.

Kuva 6.2: Nollahypoteesi ASP-testissä.

6.2 Genehunter-ohjelman käyttö

Genehunter-ohjelma [KDRDL96] on yleisimmin käytettyjä kytkentäanalysohjelmistoja. Sitä voi käyttää seuraaviin tarkoituksiin:

1. parametrinen kaksi- ja monipistekytkenäanalyysi
2. ei-parametrinen kaksi- ja monipistekytkenäanalyysi
3. markkereiden informatiivisuuden laskeminen
4. haplotyyppien määrittäminen
5. Sairaiden sisarusparien analyysi
6. QTL-varianssikomponenttianalyysi
7. Periytymisen epätasapainotesti (TDT).

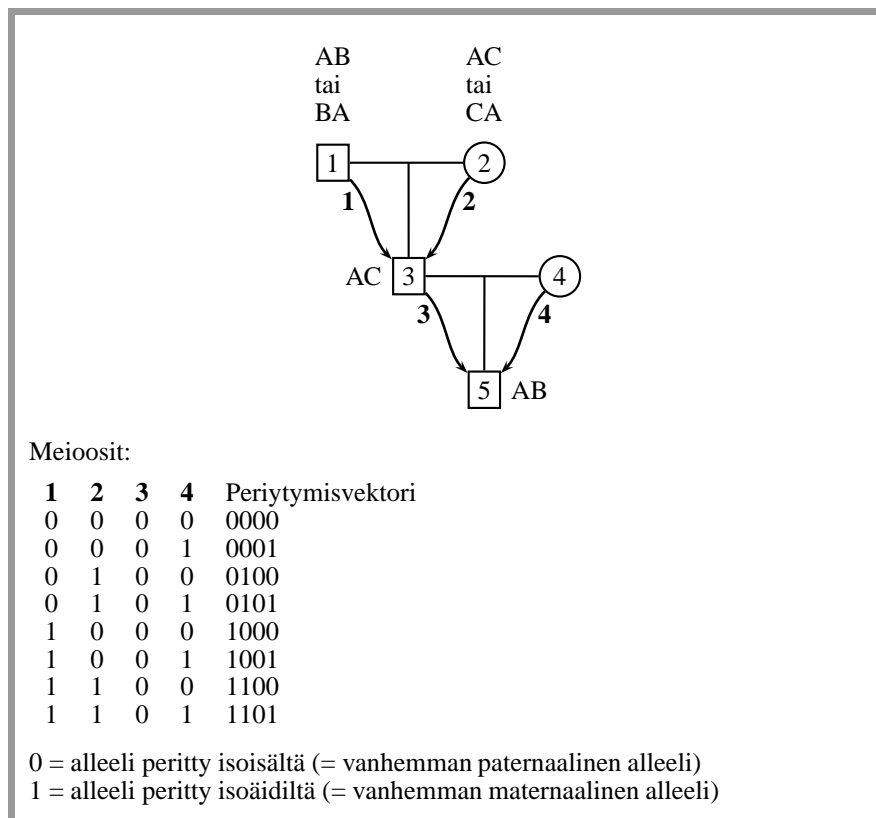
Kohdat 1–6 perustuvat periytymisvektoreihin. Genehunter-ohjelman käyttäjästä Lander-Green-algoritmista johtuen ohjelmalla voidaan analysoida vain kohtuullisen pieniä sukupuita, joissa tapahtuu noin 20–25 meioosia. Toisaalta laskenta-algoritmi mahdollistaa usean markkerin samanaikaisen kytkentäanalyysin eli monipisteanalyysin.

6.2.1 Kytkentäanalyysi Genehunter-ohjelmassa

Genehunter-ohjelman avulla voidaan samalle aineistolle suorittaa samanaikaisesti sekä parametrinen että ei-parametrinen kytkentäanalyysi. Ohjelma käyttää Lander-Green -algoritmia [LG87] uskottavuusfunktion laskemisessa. Laskenta-algoritmi perustuu ns. *periytymisvektoreiden (inheritance vector)* määrittämiseen ja niiden todennäköisyyksien laskemiseen.

Kuvassa 6.3 on esimerkki periytymisvektoreiden muodostamisesta yhdelle lokukselle ja yhdelle sukupuulle. Esimerkissä oletetaan, että isän isän genotyyppi on AB ja isän äidin genotyyppi on AC. Isän genotyyppi on AC ja lapsen AB, mutta äidin genotyyppiä ei tunneta. Näiden tietojen valossa tiedetään, että isä on perinyt A-alleelin isältään ja C-alleelin äidiltään. Tiedetään myös, että perheen pojan alleeli A on peritty isältä, ja se on alunperin tullut isän isältä eli tiedetään tämän alleelin isovanhempaisalkuperä. Tiedetään myös, että poika on perinyt alleelin B äidiltään. Tämä jättää joukon avoimia kysymyksiä vastaamatta eli isän alleelien isovanhempaisalkuperät sekä pojan äidiltään perimän B-alleelin isovanhempaisalkuperä (eli kummasta vastinkromosomista alleelit on peritty).

Kysymystä voidaan lähteä ratkaisemaan numeroimalla kaikki sukupuussa tapahtuneet meioosit. Tässä sukupuussa niitä on neljä: (1: isän isältä isälle, 2: isän äidiltä isälle, 3: isältä pojalle ja 4: äidiltä pojalle). Ainoastaan meioosi 3 voidaan määrittää suoraan, eli se on peritty isoisältä. Merkitään tapahtumaa koodilla 0 (koodi 1 tarkoittaa, että alleeli on peritty isoisäidiltä eli on



Kuva 6.3: Periytymismahdollisuudet meiooseissa ja niitä vastaavat periytymisvektorit.

maternaalista alkuperää). Muidenkin meioosien kohdalla on aina kaksi vaihtoehtoa; peritty alleeli on joko paternaalista tai maternaalista alkuperää eli 0 tai 1. Kaikki mahdollisuudet (8 kappaletta) on kirjoitettu taulukkoon, joten tälle sukupuulle on olemassa 8 erilaista periytymisvektoria. Näille kaikille lasketaan todennäköisyydet. Genehunter käsittelee kaikkia markkerilokuksia yhdessä, joten periytymisvektoreista muodostuu matriisi, jossa esitetään kaikkien lokusten yli määritellyt periytymisvektorit perhekohtaisesti. Tämän matriisin eri periytymisvektoreiden mahdollisuuksille lasketaan todennäköisyydet, jotka riippuvat markkereiden välisistä etäisyyksistä ja sukupuunformaatiosta.

Määriteltyjen periytymisvektoreiden avulla voidaan lokuskohtaisesti määrittää eri yksilöiden väliset IBD-statusmahdollisuudet ja niiden todennäköisyydet. Samoin voidaan laskea markkereiden välillä oleville alueille IBD-statusien todennäköisyydet. Periytymisvektoreiden todennäköisyyksissä esiintyvä vaihtelu on myös lähtökohtana, kun lasketaan informatiivisuusastetta kromosomin eri kohdissa.

6.2.2 Genehunterin NPL-testit

Genehunter-ohjelma sisältää kaksi erilaista IBD-statusiin perustuvaa ei-parametrista kytkentäanalyysitestistä: S_{pairs} ja S_{all} . Näistä S_{all} on yleensä voimakkaampi testi, mutta voima riippuu aineistosta ja sairauden periytymismekanismista.

S_{pairs} -testissä haetaan kaikki sukupuusta löytyvät sairaut yksilöt, ja lasketaan IBD-status näistä muodostettujen sukulaisparien kesken. Esimerkiksi silloin, jos sukupuussa on kaksi sairasta lasta sekä heidän sairas isoäitinsä, lasketaan IBD-statukset sisarusten kesken, isoäidin ja ensimmäisen sisaruksen kesken sekä isoäidin ja toisen sisaruksen kesken. Tämän jälkeen lasketaan, kuinka monta alleeliparia sairaut sukulaisparit jakavat keskenään IBD. Lisätietoa testistä löytyy artikkelista [WH94].

Toinen testi, S_{all} , perustuu samoin IBD-statusten laskemiseen kaikkien sairaiden sukulaisten kesken. Nyt kuitenkin katsotaan koko sairaiden yksilöiden joukkoa samanaikaisesti. Perusajatuksena on, että jos useat sairaut sukulaiset jakavat saman alleelin IBD, on se arvokkaampi osoitus lokuksen kytkeyttämisestä sairauslokuksen kanssa kuin jos useat sukulaisparit jakaisivat eri alleelin keskenään. Tästäkin testistä on enemmän Whitemoren ja Halpernin artikkelissa [WH94].

Arvojen S_{pairs} ja S_{all} avulla lasketaan testisuure, joka noudattaa *standardoituua normaalijakaumaa* nollahypoteesin H_0 vallitessa (H_0 = sairauslokuks ja markkerilokus eivät ole kytkettyneitä toisiinsa).

6.2.3 Tiedostojen rakenteet

Genehunter tarvitsee kaksi Linkage-muotoista syötetiedostoa: sukupuutiedoston ja parametritiedoston. Linkage-tiedostomuotoon tutustuttiin luvussa 3.

6.2.4 Genehunter-ohjelmiston käskyt

Genehunter käynnistetään komennolla:

```
corona% gh
```

Ohjelma avaa komentorivikäyttöliittymän, johon käyttäjä voi kirjoittaa komentoja yksi kerrallaan:

```
np1 : 1>
```

Load

Ohjelmalle annetaan ensin parametritiedosto komennolla `load parametritiedoston nimi`. Seuraavassa esimerkki:

```
np1:1> load ex-1.dat
Parsing Linkage marker data file...
8 markers read (last one = loc8)
```

Ohjelma varoittaa tässä vaiheessa, jos jonkin markkerin alleelifrekvenssit eivät summaudu ykköseen tai jos tiedostossa on jokin muu syntaksivirhe.

Scan

Aineisto voidaan yksinkertaisimmillaan analysoida komennolla `scan sukuuutiedoston nimi`:

```
np1:3> scan ex-1.pre
analyzing pedigree 1...
WARNING: due to computation time and memory constraints,
individuals 18 19 have been dropped from the analysis.
using non-originals: 8 3 9 5 13 14 15 12 17 10 11
position  LOD score    NPL score  p-value    information
  0.00    2.004752    1.999322  0.046875    0.733700
  1.00    2.010418    2.010278  0.046875    0.727282
  2.00    2.016873    2.023015  0.046875    0.724886
  3.00    2.024115    2.037563  0.046875    0.725526
  4.00    2.032142    2.053953  0.046875    0.729032
  5.00    2.040952    2.072213  0.027344    0.735605
  6.00    2.044646    2.053064  0.046875    0.731756
  7.00    2.049122    2.035540  0.046875    0.736611
  8.00    2.054378    2.019608  0.046875    0.750260
  9.00    2.060415    2.005234  0.046875    0.774546
 10.00    2.067232    1.992381  0.046875    0.820090
 11.00    2.068469    1.986718  0.046875    0.812122
 12.00    2.070484    1.982549  0.046875    0.807882
 13.00    2.073277    1.979868  0.046875    0.806287
 14.00    2.076850    1.978668  0.046875    0.807147
 15.00    2.081202    1.978941  0.046875    0.810636
 16.00    2.081269    1.970490  0.046875    0.809112
 17.00    2.082125    1.963587  0.046875    0.810561
 18.00    2.083770    1.958225  0.046875    0.814636
```

19.00	2.086206	1.954393	0.046875	0.821531
20.00	2.089432	1.952080	0.046875	0.832524
21.00	2.090444	1.941886	0.046875	0.816253
22.00	2.092249	1.933288	0.046875	0.806123
23.00	2.094847	1.926272	0.046875	0.799660
24.00	2.098241	1.920817	0.046875	0.796395
25.00	2.102432	1.916905	0.046875	0.796516
26.00	2.091690	1.859900	0.062500	0.777785
27.00	2.081484	1.805289	0.062500	0.764689
28.00	2.071811	1.752970	0.078125	0.756358
29.00	2.062669	1.702843	0.078125	0.752696
30.00	2.054055	1.654807	0.078125	0.755033
31.00	2.022823	1.554892	0.078125	0.718040
32.00	1.991166	1.458684	0.078125	0.696745
33.00	1.958963	1.365983	0.078125	0.686667
34.00	1.926073	1.276590	0.078125	0.688905
35.00	1.892327	1.190310	0.078125	0.714431

Ohjelma tulostaa seuraavat tiedot:

position: sairauslokuksen etäisyys ensimmäisestä markkerista senttimorganeina

LOD score: parametrinen monipisteanalyysin tuottama LOD score

NPL score: ei-parametrisesti testatun kytkennän voimakkuutta kuvaavan tunnusluvun arvo

p-value: NPL scorea vastaava tilastollinen merkitsevyys (p-arvo), joka perustuu oletukseen äärimmäisen informatiivisesta markkerikartasta

information: markkeriaineiston informatiivisuus, jonka yksikkönä käytetään entropiamittaa. Mitan arvot vaihtelevat nollan (ei lainkaan informaatiota) ja ykkösen (täydellinen informaatio) välillä.

Tulosteen alussa Genehunter esitti varoituksen siitä, että kaikkia yksilöitä ei muistirajan vuoksi otettu mukaan analyysiin. Muistirajan kasvattamista esitellään komennon `max bits` yhteydessä.

Use

Markkerikartan markkereiden väliset etäisyydet voi antaa joko suoraan parametritiedostossa tai erikseen komennolla

```
use 1_markkerin_nimi etäisyys 2_markkerin_nimi
```

Esimerkki:

```

np1:4> use 1 5.2 2 5.2 3 5.2 4 5.2 5 5.2 6 5.2 7 5.2 8
Current map (8 markers):
loc1 5.2000 loc2 5.2000 loc3 5.2000 loc4 5.2000 loc5 5.2000
loc6 5.2000 loc7 5.2000 loc8

```

Ohjelma ilmoittaa käytetyn kartan senttimorganasteikolla.

Total stats

Komento `total stats` summaa LOD scoreit ja NPL scoreit yli perheiden eli antaa koko aineistoa vastaavat arvot. Komento voidaan antaa sen jälkeen, kun aineisto on analysoitu Scan-komennolla. LOD scoreit lasketaan suoraan yhteen. NPL scoreit yhdistetään summaamalla ne ja jakamalla summa sukupuiden lukumäärän neliöjuurella.

Max bits

Komento `max bits` määrittää, kuinka suuria sukupuita ohjelmalla voidaan analysoida. Parametrina annettava luku kertoo, montako bittiä periytymisvektorissa saa enintään olla. Periytymisvektorin bittien määrä voidaan laskea sukupuusta kaavalla $2n - f$, missä n on sukupuun jälkeläisten ja f perustajajksilöiden määrä. Jos jokin sukupuuvu vaatii enemmän bittejä kuin raja sallii, karsii Genehunter automaattisesti sukupuusta vähiten informatiivisia yksilöitä.

Esimerkkiaineistollemme oletusarvona ollut raja oli liian pieni, koska kaksi yksilöä jouduttiin karsimaan. Kasvatamme rajan 20 bittiin:

```

np1:2> max bits 20
Currently analyzing a maximum of 20 bits per pedigree

```

Single point

Komennolla `single point` on voidaan kytkeä päälle kaksipisteanalyysi. Esimerkkiaineiston kaksipisteanalyysi (käyttäen edellä korotettua max bits -arvoa) näyttää tältä:

```

np1:3> single point on
Single point mode is now 'on'

np1:4> scan ex-1.pre

analyzing pedigree 1...
using non-originals:  8 3 9 5 13 14 15 12 17 10 11 18 19

```

position	LOD score	NPL score	p-value	information
M1	1.404340	0.699873	0.105469	0.575417
M2	1.105625	0.512174	0.132812	0.215846
M3	1.346509	0.542153	0.132812	0.341879
M4	0.836817	0.019589	0.347656	0.167874
M5	-0.035105	0.377323	0.144531	0.400945
M6	1.334552	0.409835	0.144531	0.260851
M7	1.607164	0.554887	0.132812	0.524790
M8	1.109399	0.130700	0.265625	0.422012

Ohjelma antaa samat LOD scoreit kuin Linkage-ohjelmisto.

Count recs

Komento `count recs` on laskee havaitut rekombinaatiot ja vertaa havaittuihin rekombinaatioihin perustuvaa karttaa (`observed map`) käyttäjän antamaan karttaan (`real map`). Komento annetaan ennen `scan`-komennolla tehtävää monipisteanalyysiä. Esimerkkiaineistollemme tuloste on seuraavanlainen:

OBSERVED RECOMBINATION:		
interval	theta	observed
1-2	0.0476	0.0228 (0.5023/22)
2-3	0.0476	0.0426 (0.9373/22)
3-4	0.0476	0.0124 (0.2736/22)
4-5	0.0476	0.0134 (0.2955/22)
5-6	0.0476	0.0182 (0.3995/22)
6-7	0.0476	0.0287 (0.6315/22)
7-8	0.0476	0.0542 (1.1931/22)
real map = 35.00, observed map = 19.96, total recs = 4.23		

Jos kartat eroavat toisistaan huomattavasti, syynä voivat olla virheet markkeriaineistossa tai käytetyssä geneettisessä kartassa. Yllä olevassa esimerkissä meioosien lukumäärä oli vain 20, joten havaintoaineisto oli pieni, ja havaittujen rekombinaatioiden määrässä on paljon satunnaisvaihtelua. Jos ero olisi havaittu suuremmassa aineistossa, olisi syytä tarkistaa virhemahdollisuudet.

Haplotype

Komento `haplotype` on tulostaa kaikkein todennäköisimmät haplotyyppit aineiston kaikille yksilöille. Haplotyyppillä tarkoitetaan saman kromosomin markkerilokuksissa esiintyvien alleelien jonoa. Haplotyyppi kuvaa siis sitä,

mitkä alleelit kullakin yksilöllä on toisaalta isältä ja toisaalta äidiltä perityssä kromosomissa. Koska Genehunter antaa vain todennäköisimmän haplotyyppikonfiguraation, muutkin konfiguraatiot ovat mahdollisia ja joskus jopa yhtä todennäköisiä, kuin ohjelman antama paras vaihtoehto. Mahdollisuuksien lukumäärä yleensä kasvaa sitä mukaa, mitä enemmän puuttuvia havaintoja on ja mitä harvempaa karttaa käytetään.

Sukupuun perustajayksilöiden osalta Genehunter tulostaa haplotyyppit ruudulle:

HAPLOTYPES OF ORIGINAL INDIVIDUALS:								
indiv 1:	3	3	2	1	1	3	3	1
	0	0	0	0	0	0	0	0
indiv 2:	5	2	4	3	1	2	1	4
	0	0	0	0	0	0	0	0
indiv 4:	3	1	2	3	4	1	3	3
	0	0	0	0	0	0	0	1
indiv 6:	4	2	4	2	4	2	3	3
	3	0	0	0	3	0	2	3
indiv 7:	0	0	1	1	4	2	2	4
	1	1	5	1	3	2	2	2
indiv 16:	2	1	3	2	4	2	3	2
	5	2	1	3	2	1	3	3

Esimerkiksi yksilön 16 haplotyyppit ovat 2-1-3-2-4-2-3-2 ja 5-2-1-3-2-1-3-3. Nolla kuvaa puuttuvaa informaatiota; näissä kohdissa haplotyyppien määrittäminen ei ollut mahdollista puuttuvien genotyyppien vuoksi.

Kaikkien yksilöiden haplotyyppit löytyvät tiedostosta haplo.dump. Tiedosto näyttää seuraavalta:

*****	1	2.072									
1	0	0	0	3	3	2	1	1	3	3	1
				0	0	0	0	0	0	0	0
2	0	0	0	5	2	4	3	1	2	1	4
				0	0	0	0	0	0	0	0
4	0	0	0	3	1	2	3	4	1	3	3
				0	0	0	0	0	0	0	1
6	0	0	0	4	2	4	2	4	2	3	3
				3	0	0	0	3	0	2	3
7	0	0	0	0	0	1	1	4	2	2	4
				1	1	5	1	3	2	2	2
16	0	0	1	2	1	3	2	4	2	3	2
				5	2	1	3	2	1	3	3
3	1	2	0	3	3	2	1	1	3	3	1
				5	2	4	3	1	2	1	4

5	1	2	0	3	3	2	1	1	3	3	1
				5	2	4	3	1	2	1	4
8	1	2	2	3	3	2	1	1	3	3	1
				5	2	4	3	1	2	1	4
9	3	4	2	3	3	2	1	1	3	3	1
				3	1	2	3	4	1	3	3
10	3	4	1	5	2	4	3	1	2	1	4
				3	1	2	3	4	1	3	1
11	5	6	1	5	2	4	3	1	2	1	4
				3	0	0	0	3	0	2	3
12	5	6	1	5	2	4	3	1	2	1	4
				4	2	4	2	4	2	3	3
13	5	6	2	3	3	2	1	1	3	3	1
				4	2	4	2	4	2	3	3
14	7	8	2	1	1	1	1	4	2	2	4
				3	3	2	1	1	3	3	1
15	7	8	2	1	1	5	1	3	2	2	2
				3	3	2	1	1	3	3	1
17	16	12	2	2	1	3	2	4	2	3	2
				5	2	4	3	1	2	1	4

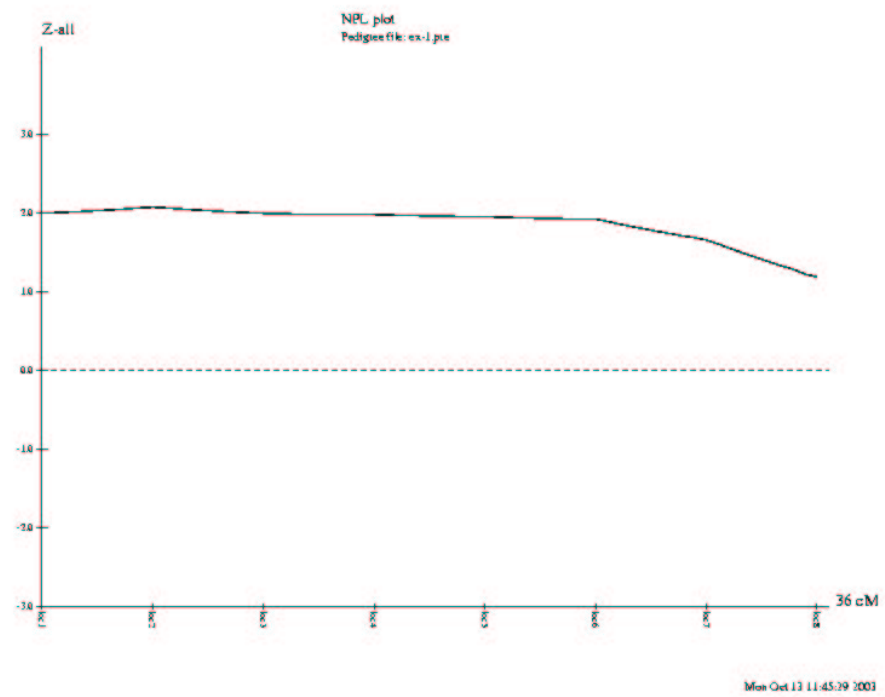
Jokaista yksilöä kohti tulostetaan kaksi riviä. Ensimmäisellä rivillä ovat yksilön tunniste, isän tunniste, äidin tunniste, sairausstatus ja näiden jälkeen isältä perityt alleelit. Toinen rivi sisältää äidiltä perityt alleelit.

Postscript output

Komento `postscript output` on (lyhyemmin `ps on`) tuottaa PostScript-tiedoston (`.ps`), joka sisältää LOD-, NPL- ja markkereiden informatiivisuusarvot graafisessa muodossa. Jos `haplotype on` -komento on annettu, `ps on` -komento tuottaa myös postscript-tiedoston sukupuista, joissa näkyvät yksilöiden haplotyytit sekä x-symbolein merkityt rekombinaatiokohtat. Esimerkki NPL-kuvaajasta on kuvassa 6.4 ja haplotyyppitulosteesta kuvassa 6.5.

Increment

Laskupisteiden määrää voidaan muuttaa käskyillä `increment distance etäisyys` tai `increment step askel`. Esimerkiksi `increment distance 2` laskee LOD- ja NPL-arvot kahden senttimorganin välein ja `increment step 10` jakaa markkerivälit kymmeneen osaan.



Kuva 6.4: Genehunter-ohjelman tuottama NPL-kuvaaja.

Off end

Analyysin aloituskohtaa voidaan muuttaa käskyllä `off end etäisyys`. Esimerkiksi `off end 10` aloittaa analyysin 10 cM ennen ensimmäistä markkeria ja lopettaa sen 10 cM viimeisen markkerin jälkeen.

Score

Komennolla `score pairs` tai `score all` voidaan valita, lasketaanko S_{pairs} - vai S_{all} -tunnuslukujen arvot. Oletusarvona on jälkimmäinen.

Map function

Käytetyn karttafunktion voi valita komennolla `map function haldane` tai `kosambi`. Kaikissa kytkentäanalyseissä, joissa kartoitetaan sairauslokuksen paikka kromosomissa, suositellaan käytettäväksi Haldanen karttafunktiota (`map function haldane`).

Kun kaikki tarpeelliset komennot on annettu, komento `scan sukupuutiedosto` on annettava uudestaan.

Date run: Mon Oct 13 11:45:22 2003

Filename: ex-1.ped

Pedigree: 1

Males: 1 2 3 4 5 6 7 8

A: 4 6 2 3 1 4 6 1

a: - - - - -

B: 6 5 5 1 3 2 5

b: - - - - -

C: 4 1 2 5 7 2 6 4

c: - - - - -

D: 5 5 5 4 7 3 6 4

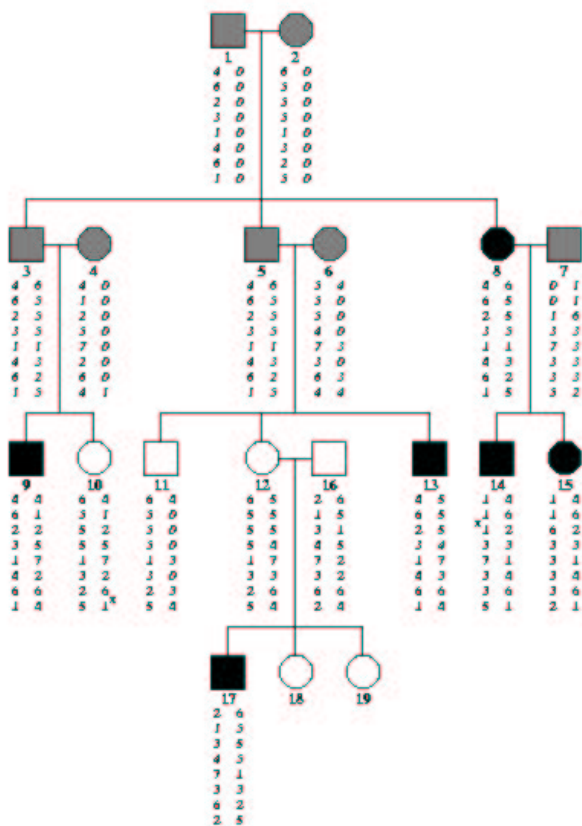
d: 4 - - 3 - 3 4

E: - - 1 3 7 3 3 5

e: 1 1 6 3 3 3 3 2

F: 2 1 3 4 7 3 6 2

f: 6 5 1 5 2 2 6 4



Kuva 6.5: Genehunter-ohjelman tuottama haplotyyppituloste.

Photo

Genehunter-istunnon tapahtumat voi tallentaa komennolla `photo file name`. Esimerkiksi `photo sessio.out` tallentaa kaikki päätteelle annetut käskyt ja ohjelman tekemät tulostukset tästä eteenpäin `sessio.out`-tiedostoon. Tallennuksen voi kytkeä pois päältä komennolla `photo off`.

Quit

Komento lopettaa Genehunter-istunnon.

Esimerkki tyypillisestä Genehunter-istunnosta

Alla esitetään tyypillisen Genehunter-istunnon kulku. Oletetaan, että Linkage-sukupuutiedoston nimi on `example.ped` ja parametritiedoston nimi on `example.dat`. Käyttäjä haluaa vedostaa istunnon kulun tiedostoon `session.out`:

```
photo session.out
load example.dat
count recs on
ps on
haplotype on
scan example.ped
total stat
quit
```

6.2.5 Komennot ASP-menetelmälle

Estimate

Tavallisen sairaiden sisarusparien analyysin voi suorittaa Genehunter-ohjelman komennolla estimate.

```
np1:11> estimate
analyze under the assumption of no dominance variance?
  y/n [n]: y
file to store ML estimation [mls.out]:
analysis complete
text output file successfully written
```

Ohjelmassa voi suorittaa uskottavuusfunktion maksimoinnin kahdella eri tavalla vastaamalla kysymykseen

```
analyze under the assumption of no dominance variance? y/n [n]:
```

joko y tai n.

Vastaamalla y oletetaan, että dominanssivarianssia ei ole, jolloin z_1 kiinnitetään arvoon 0.5 ja vain z_2 estimoidaan ($z_0 = 1 - z_1 - z_2$). Vastaamalla n sekä z_1 että z_2 estimoidaan.

Tässä tapauksessa analyysin tulokset tallentuivat tiedostoon `mls.out`. Käyttäjä voi suorittaa Genehunter-ohjelman sisältä Unix-komentoja liittämällä eteen avainsanan `system`. Esimerkiksi ohjelman tuottamaa tiedostoa `mls.out` voi katsella komennolla:

```
np1:12> system more mls.out
```

Pairs used

Oletusarvoisesti estimate-komento tekee ASP-testin käyttäen vain yhtä sisarusparia jokaisesta perheestä. Tämän voi muuttaa komennolla `pairs used`.

```

npl:12> pairs used

the current pair setting is: *first affected/phenotyped
  sibpair only*

Possible pair options:
  1. First pair of affected/phenotyped sibs
  2. All independent pairs of affected/phenotyped sibs
  3. All pairs of affected/phenotyped sibs
  4. All pairs of affected/phenotyped sibs - UNWEIGHTED
Enter the index of the analysis you want to use [1]: 2

```

6.3 Analyysi Merlin-ohjelmalla

Merlin [[ACCC02](#)] on uusi kytkentäanalyysiohjelmisto, joka on nopeasti saatuttanut vankan jalansijan. Ohjelmalla voidaan tehdä ei-parametrinen kytkentäanalyysiä, muodostaa haplotyyppisiä, arvioida henkilöiden sukulaisuuden astetta ja etsiä genotyypausvirheitä. Ohjelmaa voidaan käyttää myös varianssikomponenttianalyysiin.

Merlin perustuu Lander-Green-algoritmiin ja käyttää ns. harvoja puita periytymisvektorien tallentamiseksi. Tämän ansiosta Merlin ei joudu – toisin kuin Genehunter – käymään läpi kaikkia mahdollisia periytymisvektoreita. Merlin muodostaa periytymisvektoreista puurakenteen, jonka ansiosta se voi välttää sellaisten periytymisvektoreiden tarkastelun, jotka ovat mahdottomia aiemmin kerätyn informaation perusteella. Lisäksi Merlin hyödyntää eräitä periytymisvektoreiden symmetriaominaisuuksia hakuvaruuden rajoittamiseksi.

Ohjelman käyttämät tietorakenteet ja algoritmit nopeuttavat tuntuvasti pienten ja keskisuurien perheiden kytkentäanalyysiä. Annetulle sukupuulle Merlin vaatiikin vain murto-osan Genehunter-ohjelman laskenta-ajasta. Jos laskenta-aika puolestaan pidetään vakiona, sallii Merlin hieman suuremman sukupuun analysoinnin kuin Genehunter.

Merlin-ohjelman voi asentaa myös omaan Linux-koneeseen, mutta suuren muistin tarpeen vuoksi ajo CSC:n koneilla voi olla hyvä vaihtoehto.

6.3.1 Syötetiedostojen tarkastelu

Merlin-pakettiin kuuluvat ohjelmat käynnistetään suoraan Unix-komentokäyttöliittymästä. Samalla annetaan ajettavan ohjelman tarvitsemat parametrit. Alo-

tamme esimerkkisukupuumme (kuva 3.2) analysoinnin keräämällä pakettiin kuuluvalla Pedstats-ohjelmalla koostetietoa tarkasteltavasta aineistosta.

Merlin ymmärtää Linkage-muotoisia tiedostoja, joten voimme käyttää aiemmin laadittuja syötetiedostoja `ex-1.pre` ja `ex-1.dat`. Ohjelma tukee myös niin sanottua QTDT-tiedostomuotoa, jota ei tässä oppaassa käsitellä.

Koostetiedot pyydetään komennolla

```
corona% pedstats -p ex-1.pre -d ex-1.dat
```

Ohjelma tuottaa koosteen fenotyypin ja genotyyppien jakaumasta. Jos sukupuussa on Mendel-virheitä, ne paljastuisivat tässä vaiheessa. Jos esimerkiksi vaihdamme viimeisen yksilön 19 ensimmäisen markkerilokuksen genotyypin "2/6"muotoon "2/2", tuottaisi ohjelma virheilmoituksen:

```
M1 - Fam 1: Child 19 [2/2] has parents [2/6]*[5/6]
```

```
FATAL ERROR -
```

```
Mendelian inheritance errors detected
```

Pedstats-ohjelmalla voidaan paitsi etsiä Mendel-virheitä myös tuottaa graafisia kuvaajia syötetiedoston ominaisuuksista kuten sukupuiden rakenteesta ja alleelifrekvenssijakaumista. Tämä onnistuu valitsimella `--pdf`.

6.3.2 Kytkentäanalyysin käynnistys

Käynnistämme kytkentäanalyysin antamalla seuraavan komennon:

```
corona% merlin -p ex-1.pre -d ex-1.dat --np1
```

Valitsinta `-p` seuraa sukupuutiedoston ja valitsinta `-d` parametritiedoston nimi. Ohjelma tunnistaa automaattisesti, että kyseessä on Linkage-muotoinen syöte. Valitsin `--np1` kertoo, että aineistolle halutaan ajaa ei-parametrinen kytkentäanalyysi.

Ohjelma näyttää sille annetut parametrit sekä tuottaa kytkentäanalyysituloksen:

```
Phenotype: LOCUS1 [ALL] (1 family)
```

	Pos	Zmean	pvalue	delta	LOD	pvalue
min		-0.87	0.8	-0.121	-0.04	0.7
max		8.24	0.00000	1.155	1.02	0.02
	0.000	2.00	0.02	1.155	0.52	0.06
	5.000	2.07	0.02	1.155	0.53	0.06
	10.000	1.99	0.02	1.155	0.52	0.06
	15.000	1.98	0.02	1.155	0.52	0.06
	20.000	1.95	0.03	1.155	0.51	0.06
	25.000	1.92	0.03	1.155	0.51	0.06
	30.000	1.65	0.05	1.155	0.46	0.07
	35.000	1.19	0.12	1.155	0.38	0.09

Tulosteen sarake Zmean vastaa Genehunter-ohjelman tuottamaa NPL scorea. Tätä seuraa normaaliapproksimaatiolla saatu p-arvo tunnusluvulle. Normaaliapproksimaatio perustuu tunnusluvun oletettuun jakaumaan ideaalitalanteessa, jossa oletetaan täysin informatiivinen markkerikartta.

Kolme viimeistä saraketta perustuvat Kongin ja Coxin menetelmään [KC97], jossa IBD-jakamisen määrää arvioidaan yhdellä parametrilla (delta). Menetelmä on tuttu Genehunter Plus -ohjelmistosta, ja se hyödyntää kehittäjien mukaan paremmin informaatiota markkerien välisillä alueilla.

Ei-parametrissa kytkeäntäanalyysiä voidaan ohjata erilaisin komentorivillä annettavin asetuksin, joista keskeisimpiä esitetään seuraavassa:

- bits:n:** Kertoo, kuinka pitkille periytymisvektoreille muistia varataan. Korottamalla tämän parametrin arvoa voidaan analysoida suurempia sukupuita. Vakio n asettaa maksimiarvon lausekkeelle $2k - f$, missä k on sukupuun yksilöiden lukumäärä ja f perustajayksilöiden lukumäärä.
- perFamily:** Erittelee ei-parametrisen kytkeäntäanalyysin tulokset tiedostoon perheittäin.
- pdf:** Luo analyysin tuloksista pdf-muotoisen tiedoston, jota voi katsella Adobe Acrobat -ohjelmalla.
- information:** Tulostaa markkerikartan informatiivisuuden karttasi-jainnin funktiona. Informaation mittana Merlin käyttää Genehunter-ohjelman tapaan entropiaa.
- error:** Kirjoittaa tiedostoon merlin.err luettelon genotyypeistä, jotka ovat epätodennäköisiä mutta eivät silti aiheuta Mendel-virheitä. Tällainen tilanne voi aiheutua esimerkiksi sellaisesta genotyypausvirheestä, joka voitaisiin periaatteessa selittää kahdella hyvin lähellä toisiaan olevalla rekombinaatiolla.

- simulate:** Suorittaa ns. alleelienpudotussimulaation, jolloin Merlin simuloi kromosomin tai kromosomialueiden periytymistä kussakin sukupuussa. Tämän jälkeen se arpoo perustajayksilöille markkerikohdissa alleelit kyseisen alleelifrekvenssijakauman mukaisesti ja päättelee niiden sekä periytyneiden kromosomisegmenttien perusteella kaikkien yksilöiden genotyypit. Lopuksi Merlin merkitsee samat genotyypit puuttuviksi kuin alkuperäisessä aineistossa. Näin muodostettu keinotekoinen aineisto on nollahypoteesin (ei kytkentää) mukainen, ja se voidaan analysoida samoin kuin todellinen aineisto, esimerkiksi valitsimella `--np1`. Toistamalla simulointi useita kertoja saadaan nollahypoteesijakauma sille tunnusluvulle, josta ollaan kiinnostuneita. Lisätietoja alleelienpudotussimulaatioista on luvussa 9. Simulaatioita toistettaessa on lisäksi annettava valitsin `-r`, jota seuraa satunnaisluku-generaattorin siemenluku. Siemenluvun on vaihduttava toistokerrasta toiseen.
- ibd:** Tulostaa kullekin yksilöparille ja kullekin kohdalle todennäköisyysjakauman sille, montako alleelia (0, 1 vai 2) yksilöt jakavat tuossa kohdassa IBD. Jakauma on seurausta sukulaisuuden asteesta ja markkerien informatiivisuudesta.

6.4 Suurten sukupuiden analysointi: Simwalk

Edellä käsitellyt Genehunter- ja Merlin-ohjelmat käyttävät Lander-Green-algoritmia sukupuiden läpikäymiseen. Algoritmin antama tulos on tarkka, mutta laskenta suurille sukupuille kestää kauan ja vie paljon muistia. Suurten sukupuiden analysoimiseksi on käytettävä heuristista menetelmää, joka antaa riittävän hyvän vastauksen suurella todennäköisyydellä. Sallimalla pieni epävarmuus saavutetaan suuri säästö laskenta-ajassa. Simwalk-ohjelma käyttää kahta heuristista lähestymistapaa: Markov Chain Monte Carlo -simulaatioita sekä simuloitua jäädytystä. Tarkka kuvaus ohjelman käyttämistä algoritmeista on artikkeleissa [\[SL96\]](#), [\[ESW01\]](#) ja [\[SPL02\]](#).

Tässä luvussa tarkastellaan ei-parametrista kytkentäanalyysiä Simwalk-ohjelmalla. Vaikka Simwalk-ohjelma ei hyväksy syötteekseen Linkage-muotoisia tiedostoja, voidaan tiedostot muuttaa helposti oikeaan formaattiin Mega2-tiedostomuunnostyökalun [\[MAS⁺99\]](#) avulla.

6.4.1 Simwalk-analyysit Mega2-ohjelman avulla

Mega2-ohjelma ja sen käyttömahdollisuudet esiteltiin luvussa 3.2. Tutuimme Mega2-ohjelman syötetiedostoihin ja syötimme ohjelmalle esimerkkiaineistoomme liittyvät sukupuu-, parametri- ja karttatiedostot. Etenimme

tuolloin pisteeseen, jossa Mega2-ohjelma kysyy valittavaa analyysiohjelmistoa:

ANALYSIS MENU	
1 SimWalk2-format	15 TDTMax analyses
2 MENDEL-format	16 SOLAR-format
3 ASPEX-format	17 Vitesse-format
4 GeneHunter-Plus format	18 Linkage-format
5 GeneHunter format	19 Test loci for HWE
6 APM-format	20 Allegro-format
7 APM MULT format	21 MLBQTL format
8 Create nuclear families	22 SAGE 4.0 format
9 SLINK-format	23 Pre-makeped format
10 SPLINK-format	24 Merlin/SimWalk2-NPL
11 Homogeneity analyses	25 PREST format
12 SIMULATE-format	26 PAP format
13 Create summary files	27 Merlin format
14 SAGE-format	28 Loki format

Select an option between 1-28 > 1

Koska haluamme luoda syöte- ja ajotiedostot erityisesti Simwalk-ohjelmaa varten, valitsemme vaihtoehdon 1.

Simwalk-ohjelmalla voidaan tehdä muun muassa haplotyyppianalyysiä sekä parametrista ja ei-parametrista kytkentäanalyysiä. Valitsemme tässä viimeksi mainitun.

SimWalk2 program options:
1) Haplotype analysis
2) Parametric Linkage analysis
3) Non-Parametric Linkage analysis
4) IBD analysis
5) Mistyping analysis
Enter selection 1-5 > 3

Tässä vaiheessa ohjelma tuottaa esimerkkiaineistollamme varoituksia siitä, että osa alleelifrekvensseistä on nolli. Tämä on seurausta siitä, että alleelifrekvenssit arvioitiin kovin pienestä aineistosta, jossa kaikkia alleeleja ei esiintynyt.

Seuraavaksi ohjelma tarjoaa mahdollisuuden valita kartasta yksittäisiä markkereita; valitsemme kuitenkin kartan kaikki markkerit siinä järjestyksessä, jossa ne on karttatiedostossa määritelty:

```

LOCUS REORDERING MENU
1) Select all loci in map order on chromosome: 1
2) Select by locus number.
3) Select one chromosome and select loci on it.
4) Select one or more chromosomes and select loci on them.
Enter 1, 2, 3 or 4 > 1

```

Lopuksi ohjelma kysyy tulostiedostojen nimiä. Ohjelma tarjoaa myös mahdollisuuden tuottaa analyysin tuloksesta graafisia kuvia R-tilasto-ohjelmiston avulla (valinta 8). Tämä toiminto kannattaa kytkeä päälle.

```

0) Done with this menu - please proceed
1) Locus file name :                LOCUS.01                [new]
2) Pedigree filename :              PEDIGREE.01            [new]
3) Penetrance file name :           PEN.01                 [new]
4) Batch file name :                BATCH2.01              [new]
5) C-shell file name :              npl.01.sh             [new]
6) Person id in output pedigree file: Individual id
7) Pedigree identifier in output
   pedigree file: Premakeped pedigree number.
8) Generate R-plots for SimWalk2-NPL: [yes]

Select options 0-7 to enter new values,
8 to toggle > 0

```

Tämän jälkeen ohjelma tiedustele, minkä tunnuslukujen arvoista tulostetaan graafiset käyrät.

```

R plot statistic selection menu:
=====
1) BLOCKS
2) MAX-TREE
3) ENTROPY
4) NPL_PAIRS
5) NPL_ALL
=====
Enter string of statistic numbers ('e' to terminate) > 4 5 e

```

Valittavia tunnuslukuja on viisi erilaista, ja ne ovat:

BLOCKS: Niiden perustajyksilöiden alleelien lukumäärä, joita esiintyy sairailta yksilöillä.

MAX-TREE: Suurin kopiomäärä, jona jokin perustajyksilön alleeli on levinnyt sairaille yksilöille.

ENTROPY: Sairailta yksilöillä havaittavien alleelien entropia. Entropia on suurin silloin, kun eri alleelit ovat jakautuneet yksilöille mahdollisimman tasaisesti.

NPL_PAIRS: IBD-sukulaisuuskerroin, joka kuvaa alleelien jakamisen astetta sairaiden sukulaisparien välillä.

NPL_ALL: Genehunter-ohjelman käyttämää NPL-all-testisuureta vastaava tunnusluku, joka kuvaa jakamista sairaiden sukulaisparien välillä.

Kolmen ensimmäisen tunnusluvun pieni arvo kuvaa kytkentää. Muilla tunnusluvuilla suuri arvo osoittaa kytkennän.

Yllä olevassa esimerkissä valitsimme tulosteeseen tunnusluvut NPL_PAIRS ja NPL_ALL. Tunnusluvut syötetään antamalla niitä vastaavat numerot, joita seuraa e-kirjain. Tämän jälkeen Mega2-ohjelma tiedustelee graafisen tuloksen ulkoasuparametreja:

```
R plot parameter selection menu:
```

```
0) Done with this menu - please proceed
```

```
1) Postscript output file name      SW2NPL.01.ps [overwrite]
```

```
2) Minimum Y-axis value             0.00
```

```
3) Maximum Y-axis value             3.00
```

```
4) Horizontal cut-off line at       2.00
```

```
5) Postscript plot orientation      [portrait]
```

```
Select from options 0-5 (5 to toggle, 2-4 to change values) > 0
```

Oletusarvot voidaan valita syöttämällä 0.

Esimerkkisukupuun muunnoksen tuloksena Mega2-ohjelma tuottaa kromosomikohtaiset syötetiedostot nimeltä PEDIGREE.01, LOCUS.01, PEN.01 ja BATCH2.01. Syötetiedostojen sisältöön palataan hieman myöhemmin.

Ohjelma tuottaa myös analysointia helpottavan ajokelpoisen komentotiedoston nimeltä np1.01.sh. Komentotiedosto ajetaan — ja kytkentäanalyysi käynnistetään — antamalla komentotiedoston nimi komentorivillä:

```
corona% np1.01.sh
```

Analyysi ja tulosten koostaminen tapahtuvat tämän jälkeen automaattisesti, mutta katsotaan silti, mitä komentotiedosto tekee: se käsittelee yhden kromosomin kerrallaan siten, että vuorossa olevat kromosomikohtaiset syötetiedostot kopioidaan nimille PEDIGREE.DAT, LOCUS.DAT, PEN.DAT sekä BATCH2.DAT. Seuraavaksi komentotiedosto ajaa Simwalk-ohjelman, joka lukee nämä syötetiedostot ja tekee kytkentäanalyysin. Lopuksi komentotiedosto kutsuu toista, juuri generoitua komentotiedostoa nimeltä Rsimwalk.01.sh. Tämä käynnistää R-tilasto-ohjelmiston, joka puolestaan tuottaa NPL-tunnuslukua vastaavan käyrän tiedostoon nimeltä SW2NPL.01.ps.

Tuloksena saatavaa kytkentäkäyrää voidaan katsella esimerkiksi Ghostview-ohjelmalla:

```
corona% ghostview SW2NPL.01.ps
```

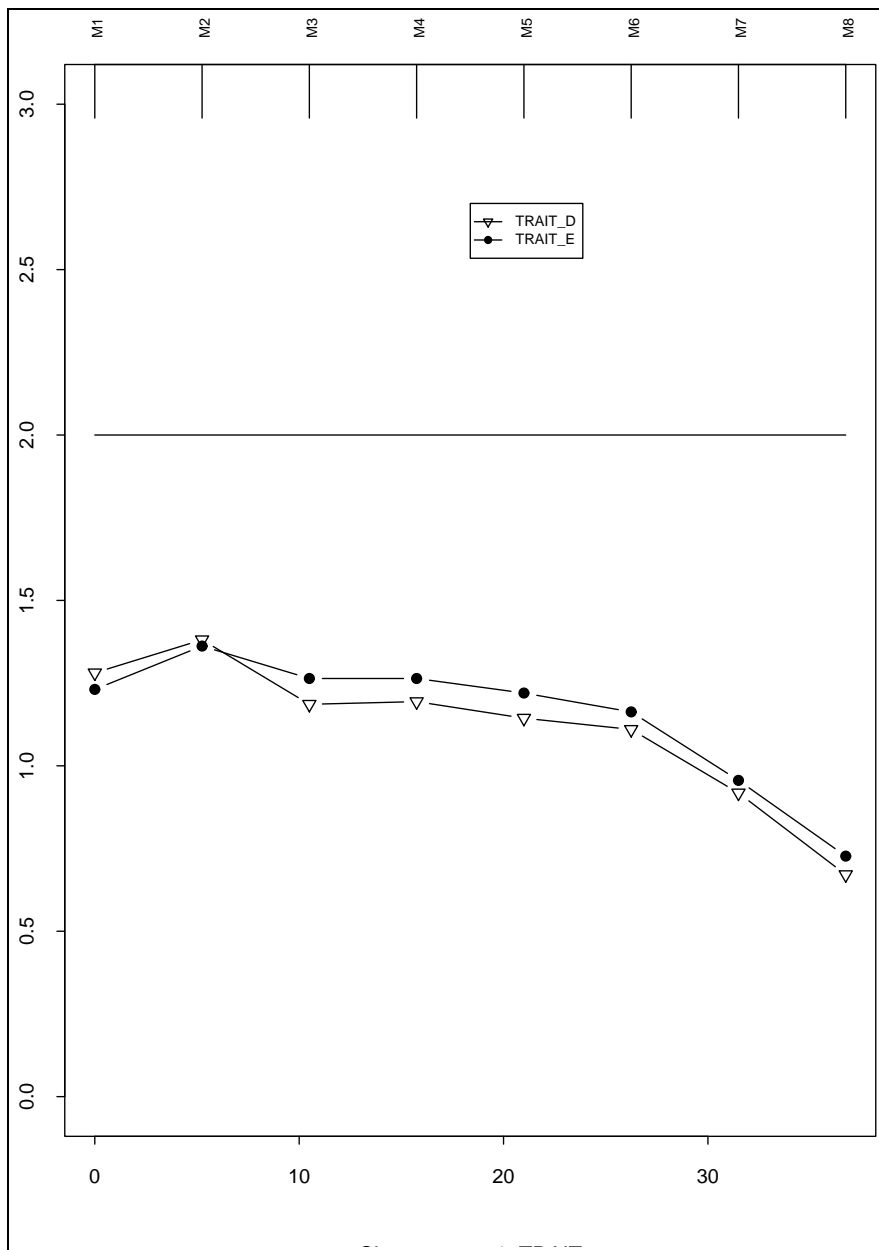
Tuloksena saatavat kuvaajat on esitetty kuvassa 6.6. Kuvaajaan on merkitty empiirisistä p -arvoista laskettujen kymmenkantaisten logaritmien vastaluvut ($-\log P$), joten asteikko ei ole suoraan yhtenevä Genehunter-ohjelman NPL-tunnuslukujen kanssa.

6.4.2 Analyysin ohjaaminen manuaalisesti

Vaikka syötetiedostot voidaan rakentaa Mega2-apuohjelman avulla, on niiden rakenteen tuntemisesta hyötyä, jos halutaan muuttaa manuaalisesti ohjelman parametreja.

Kun Simwalk2 käynnistetään komentoriviltä komennolla `simwalk` (ilman Mega2-ohjelman generoimaa ajotiedostoa), etsii ohjelma syötetiedostoja nimeltä PEDIGREE.DAT, LOCUS.DAT, PEN.DAT ja BATCH2.DAT. Näistä sukupuutiedosto (PEDIGREE.DAT) sisältää tiedot yksilöistä, heidän sukulaisuussuhteistaan ja genotyypeistään. Lokustiedosto (LOCUS.DAT) määrittelee käytettävien markkereiden nimet, alleelifrekvenssit ja sijainnit geneettisellä kartalla. Penetranssitudosto (PEN.DAT) määrittää sairausmallin mahdollista parametrissa kytkentäanalyysiä varten. Ei-parametrisessa analyysissä penetranssitudoston arvoilla ei ole merkitystä. Ohjaustiedosto (BATCH2.DAT) sisältää suuren joukon parametreja, joiden avulla analyysin kulkua voidaan säädellä. Käyttäjä voi tehdä näihin syötetiedostoihin haluamansa muutokset ja ajaa analyysin komennolla:

```
corona% simwalk
```



Kuva 6.6: Simwalk-analyysin tuloksena saatava PostScript-muotoinen kytkentäkäyrä kahdelle kytkentää kuvaavalle tunnusluvulle.

7 KytKentäepätasapaino ja assosiaatioanalyysi

7.1 Taustaa

Assosiaatioanalyysin tavoitteena on löytää viitteitä siitä, esiintyykö sairausstatus odotettua useammin yhdessä jonkin tietyn alleelin tai haplotyyppin kanssa. Lähestymistavassa hyödynnetään populaatioissa esiintyvää sairaus- ja markkerilokusten alleelien välistä *kytKentäepätasapainoa*, jolla tarkoitetaan lähekkäisten lokusten alleelien esiintymistä yhdessä odotettua useammin. Sairausgeenien paikantamisessa hyödynnettävä kytKentäepätasapaino on usein seurausta yhteisen väestöhistorian pullonkaulasta, jolloin *geneettinen ajautuminen (drift)* on karsinut suuren osan kromosomeista. Jäljelle jääneet kromosomit ovat tällöin monistuneet väestöhistorian kuluessa, eivätkä lyhyen, pullonkaulan jälkeisen historian aikana tapahtuneet rekombinaatiot vielä ole tuhonneet väestöön tulneiden riskialleelien ja läheisten markkerialleelien välistä yhteyttä. Kuva 7.1 esittää kytKentäepätasapainoon perustuvan geenikaritoituksen taustalla olevan ajatuksen.

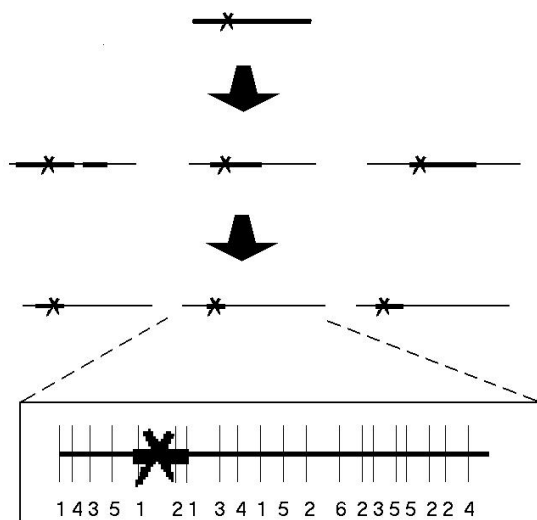
Kahden lokuksen alleelien välistä kytKentäepätasapainoa kuvataan erilaisin tunnusluvuin. Näistä yleisimmin käytettyjä ovat D -mitta, skaalattu kytKentäepätasapainomitta D' sekä korrelaatiokertoimen neliö r^2 . Tarkastellaan kaksialleelisiä loksia ja merkitään kahden lokuksen haplotyyppien 1-1, 1-2, 2-1 ja 2-2 todennäköisyyksiä p_{11} , p_{12} , p_{21} ja p_{22} . Ensimmäisen lokuksen alleelifrekvenssejä kuvataan lisäksi muuttujilla p_{1+} sekä p_{2+} , ja toisen lokuksen p_{+1} sekä p_{+2} . Tällöin kytKentäepätasapainomitat määritetään seuraavasti:

$$D = p_{11}p_{22} - p_{12}p_{21}$$

$$D' = \frac{D}{D_{max}}, \text{ missä } D_{max} = \begin{cases} \min(p_{1+}p_{+2}, p_{+1}p_{2+}) & \text{kun } D \geq 0 \\ \min(p_{1+}p_{+1}, p_{+2}p_{2+}) & \text{kun } D < 0 \end{cases}$$

$$r^2 = \frac{D^2}{p_{1+}p_{2+}p_{+1}p_{+2}}$$

KytKentäepätasapainon suuruus riippuu väestöhistoriasta. Kahden lokuksen alleelien välinen kytKentäepätasapaino heikkenee populaation iän kasvaessa, mutta toisaalta suurempi rekombinaatioiden määrä voi sallia sairausmutaation yhä tarkemman paikantamisen. Populaatiohistorialla on muutenkin ratkaiseva merkitys assosiaatioanalyysin onnistumiselle, mistä kuva 7.2 antaa



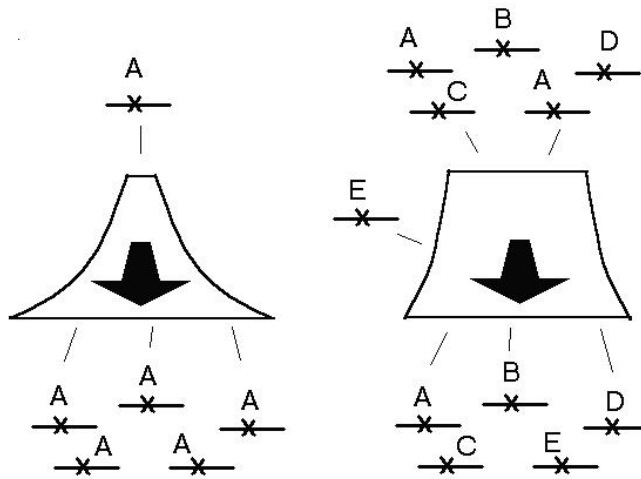
Kuva 7.1: Kytentäpätasapaino ja sen hyödyntäminen geenikartoituksessa ihannelitilanteessa. Kuvan yläreunassa olevan esivanhemman paksulla viivalla merkitty kromosomi sisältää sairautta aiheuttavan mutaation (X). Sukupolvien kuluessa jälkeläisille periytynyt kromosomisegmentti pienenee. Kun nykypopulaatiosta kerätään sairaita mutaation kantajia, ovat he perineet riskialleelin (X) lisäksi vaihtelevanmittaisen kromosomisegmentin tuon kohdan ympäriltä. Kuvan alareunassa yksi kromosomeista on suurennettu: havaitaan, että riskialleelin (X) lisäksi kyseinen yksilö oli perinyt myös läheisten markerilokusten yli ulottuvan haplotyyypin 1-2-1.

esimerkin. Vasemmalla on pienen alkuperäisväestön asuttama perustajapopulaatio, joka on kasvanut voimakkaasti historian saatossa. Oikeanpuoleisen kuvan populaatio on sen sijaan vanhempi. Tarkastelun alkuhetkellä populaatio on ollut paljon suurempi eikä voimakasta kasvua ole esiintynyt. Lisäksi populaatioon on kohdistunut runsaasti muuttoliikettä myöhemmissä vaiheissa.

Koska vasemmanpuoleisen populaation perustajaväestö on pieni, on luultavaa, että sinne on saapunut vain pieni määrä eri sairausmutaatioita, kuvan tapauksessa yksi (A). Oikeanpuoleisen kuvan tilanteessa on luultavaa, että sairait perustajayksilöt ovat suuren määränsä vuoksi geneettisesti heterogeenisempiä ja mutaatiosta on useampia erilaisia kopioita (A–D). Lisäksi muuttoliike on tuonut populaatioon uusia mutaatioita (E).

Seurauksena tästä vasemman vasemmanpuoleisen kuvan tilanteesta myös nykyväestön sairait yksilöt kantavat enimmäkseen samaa tai samoja mutaatioita. Oikeanpuoleisen kuvan tilanteesta sairaita yksilöistä otettu otos sisältää suuren joukon erilaisia mutaatioita.

Vasemmanpuoleisen kuvan tilanteesta mutaation paikantaminen on helpompaa, koska suuri osa sairaita kantaa samaa, samoihin alleeleihin assosioitunutta mutaatiota.



Kuva 7.2: Kaksi erilaista populaatiohistoriaa. Vasemmalla on voimakkaasti laajentunut pullonkaulapopulaatio ja oikealla hitaasti laajentunut populaatio.

7.1.1 Assosiaatioanalyysi yksilötason χ^2 -testillä

Yksinkertaisin tapa mitata sairausstatuksen ja alleelien assosiaation voimakkuutta on tehdä ns. χ^2 -testi 2×2 -kontingenssitaululle, joka sisältää havaitut frekvenssit sairaiden ja verrokkien otoksissa. Tarkastellaan esimerkiksi assosiaation voimakkuuden testaamisesta (taulukko 7.1). Eräessä markkerilokuksessa sairailta yksilöillä esiintyy 79 kertaa alleeli 1 ja 21 kertaa alleeli 2. Terveiden yksilöiden otoksessa vastaavat lukumäärät ovat 46 ja 54.

Nollahypoteesitilanteessa, jossa alleelien ja sairausstatuksen välinen assosiaatio on puhtaasti satunnaista, saadaan odotetut frekvenssit laskettua kertomalla vastaavat rivi- ja sarakesummat ja jakamalla tulo havaintojen kokonaismäärällä. Esimerkiksi odotettu frekvenssi alleelin 1 esiintymien määrälle sairaiden yksilöiden otoksessa on tällöin $(100 \times 125)/200 = 62.5$.

Testisuureen χ^2 -arvo voidaan nyt laskea havaittujen (o_{ij}) ja odotettujen (e_{ij}) frekvenssien avulla:

$$\begin{aligned}\chi^2 &= \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \\ &= \frac{(79 - 62.5)^2}{62.5} + \frac{(46 - 62.5)^2}{62.5} + \frac{(21 - 37.5)^2}{37.5} + \frac{(54 - 37.5)^2}{37.5} = 23.23.\end{aligned}$$

Nollahypoteesin vallitessa testisuure jakautuu χ^2 -jakauman mukaisesti yhdellä vapausasteella. Kertymäfunktion jakaumasta nähdään, että testisuureen arvoon liittyvä p-arvo on jopa alle 0.000001, joten tulos on tilastollisesti erittäin merkitsevä.

Taulukko 7.1: Havaitut ja odotetut alleelien 1 ja 2 esiintymiskerrat terveiden ja sairaiden yksilöiden otoksissa.

Havaitut	Sairas	Terve	Yht.
Alleeli 1	79	46	125
Alleeli 2	21	54	75
Yht.	100	100	200

Odotetut	Sairas	Terve	Yht.
Alleeli 1	62.5	62.5	125
Alleeli 2	37.5	37.5	75
Yht.	100	100	200

7.2 Periytymisen epätasapainotesti (TDT)

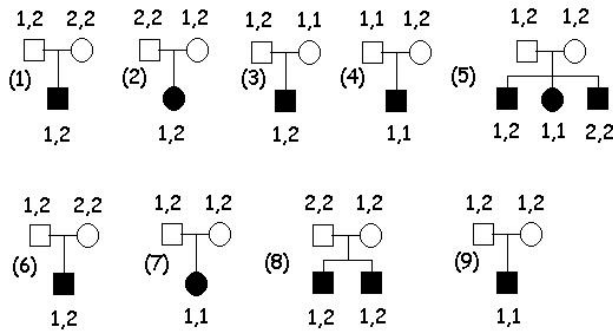
Periytymisen epätasapainotesti (*Transmission/Disequilibrium Test, TDT*) yhdistää kytkentäanalyysin ja assosiaatioanalyysin hyviä puolia. Testi perustuu periytymisen tarkastelemiseen ydinperheiden sisällä siten, että samasta perheestä ei tarvita kuin yksi sairaskäyttäjä.

Tässä oppaassa tarkastellaan testin perusversiota [SME93] sekä sen approksimointia χ^2 -jakaumaan perustuen. Periytymisen epätasapainotestin keskeisiä ominaisuuksia ovat:

- TDT perustuu ydinperheiden analysointiin. Testissä tarkastellaan alleelien siirtymistä heterotsygootilta vanhemmalta sairaille jälkeläisille. Kussakin ydinperheessä voi olla yksi tai useampi sairaskäyttäjä.
- Kyseessä on geneettisen kytkennän testi. Lokusten välisen kytkentäepätasapainon olemassaolo on ehdoton edellytys TDT:n menestykselle käytölle.
- Testi ei edellytä oletuksia periytymismallista. Testin tilastollinen voima riippuu kuitenkin periytymismallista ja on suurimmillaan resessiivisen sairauden tapauksessa.
- Testi on laskennallisesti varsin yksinkertainen ja intuitiivinen.

TDT-testissä otos muodostetaan valitsemalla satunnaisia ydinperheitä, joissa on ainakin yksi sairaskäyttäjä. Esimerkki tällaisesta aineistosta on kuvassa 7.3.

TDT-testin taustalla oleva ajatus on seuraava: Kussakin ydinperheessä heterotsygootilla vanhemmalla on markkerilokuksessa alleelit 1 ja 2. Jos lokus ei ole kytketty sairausstatukseen, siirtyy alleeli 1 jälkeläiselle samalla todennäköisyydellä kuin alleeli 2. Tällöin em. ydinperheiden otokseen on tullut sairaita alleelin 1 ja alleelin 2 kantajia odotusarvoisesti sama määrä, ja



Kuva 7.3: Pieni esimerkkiaineisto periytymisen epätasapainotestiä varten. Genotyypit on merkitty yhden markkerin suhteen.

otoksen sisällä kumpikin alleeleista 1 ja 2 näyttäisi siirtyneen sairaalle jälkeläiselle samassa suhteessa. Jos sen sijaan esimerkiksi alleeli 1 olisi assosioitunut sairausstatukseen, olisi valtaosalla sairaista yksilöistä markkerilokuksessa yksi tai kaksi kappaletta alleelia 1. Alleelin 1 kantajat ovat tällöin yliedustettuja otoksessa, ja havaitaan, että otoksen sisällä alleeli 1 on siirtynyt heterotsygoottivanhemmalta sairaalle jälkeläiselle suuremmalla todennäköisyydellä kuin alleeli 2.

Tarkastellaan aluksi kahden alleelin tapausta (tulokset yleistyvät myös markkerilokuksille, joissa on useampia alleeleita). KytKentää testattaessa nollahypoteesin vallitessa sairausstatus periytyy markkerilokuksesta riippumatta. Määrittelemme aineistosta laskettavan TDT-testisuureen, jonka teoreettinen nollahypoteesin mukainen jakauma tunnetaan. Kun todellisesta aineistosta laskettua testisuuren arvoa verrataan tähän jakaumaan, voidaan esittää arvio siitä, onko aineisto todennäköisesti syntynyt nollahypoteesin mukaisesta jakaumasta. Jos tämä vaikuttaa riittävän epätodennäköiseltä, nollahypoteesi hylätään.

Tarkastellaan joukkoa ydinperheitä, jossa on n sairasta lasta. Markkerilokuksessa M esiintyy kahta alleelia, M_1 ja M_2 . Vanhempien genotyypit voidaan esittää taulukkona seuraavasti:

	M_1 ei siirtynyt	M_2 ei siirtynyt	Yhteensä
M_1 siirtyi	t_{11}	t_{12}	$t_{11} + t_{12}$
M_2 siirtyi	t_{21}	t_{22}	$t_{21} + t_{22}$
Yhteensä	$t_{11} + t_{21}$	$t_{12} + t_{22}$	$2n$

Nelikentän luvut t_{12} ja t_{21} kuvaavat heterotsygootteja vanhempia. Luvut t_{11} ja t_{22} liittyvät homotsygootteihin, eivätkä ne sisällä lainkaan transmissioinformaatiota.

Havaitut arvot sekä nollahypoteesitilanteen mukaiset odotetut arvot voidaan kirjoittaa taulukkona:

<i>Tilanne</i>	<i>Havaittu</i>	<i>Odotettu</i>
M_1 siirtynyt, M_2 ei siirtynyt	t_{12}	$(t_{12} + t_{21})/2$
M_1 ei siirtynyt, M_2 siirtynyt	t_{21}	$(t_{12} + t_{21})/2$
Yhteensä	$t_{12} + t_{21}$	$t_{12} + t_{21}$

7.2.1 Binomijakaumaan perustuva eksakti testi

Seuraavassa esimerkissä tehdään ensin binomijakaumaan perustuva eksakti testi. Sen jälkeen tarkastellaan sen erästä approksimaatiota, jossa käytetään χ^2 -testisuuretta.

Tarkastellaan kuvan 7.3 yhdeksän ydinperheen joukkoa. Kuvaan 7.3 on merkitty sukupuut sekä kunkin jäsenen genotyyppi markkerilokuksessa.

Alleelien siirtymätiedot voidaan esittää taulukkona:

Perhe	Isä	Äiti	Sairas lapsi	Heterotsygootit vanhemmat	(1) Siirt.	(2) Siirt.
1.	1,2	2,2	1,2	1	1	
2.	2,2	1,2	1,2	1	1	
3.	1,2	1,1	1,2	1		1
4.	1,1	1,2	1,1	1	1	
5.	1,2	1,2	1,2	2	1	1
	1,2	1,2	1,1	2	2	
	1,2	1,2	2,2	2		2
6.	1,2	2,2	1,2	1	1	
7.	1,2	1,2	1,1	2	2	
8.	2,2	1,2	1,2	1	1	
	2,2	1,2	1,2	1	1	
9.	1,2	1,2	1,1	2	2	
Yhteensä				17	13	4

Nollahypoteesitilanteessa kukin 17 heterotsygootivanhemmasta siirtää yhtä suurella todennäköisyydellä ykkös- kuin kakkosalleelin sairaalle jälkeläiselle, joten testisuureemme, siirtyneiden ykkösalleelien (vastaavasti kakkosalleelien) lukumäärä, noudattaa nollahypoteesitapauksessa binomijakaumaa. Tilannetta voi verrata kolikon heittoon: kun rahaa heitetään 17 kertaa, on luultavaa, että kruunia ja klaavoja saadaan likimain sama määrä. Jos jompia-kumpia tulee selvästi enemmän, on aihetta epäillä kolikkoa painotetuksi.

Laskemme binomitodennäköisyyden kaavan avulla p-arvon eli todennäköisyyden sille, että havaintojen jakauma koetta toistettaessa on vähintään yhtä epätodennäköinen kuin ylläolevassa aineistossa.

Kuvatkoön satunnaismuuttuja X niiden kertojen lukumäärää, jolloin alleeli 1 oli siirtynyt.

Tällöin

$$P(X \geq 3) = P(X = 13) + P(X = 14) + P(X = 15) + P(X = 16) + P(X = 17).$$

Tässä lausekkeessa saadaan esimerkiksi

$$P(X = 13) = \binom{17}{13} \times 0.5^{13} \times 0.5^4 = 0.018158$$

ja vastaavasti:

$$P(X = 14) = 0.005188, P(X = 15) = 0.001038,$$

$$P(X = 16) = 0.000130 \text{ ja } P(X = 17) = 0.000008.$$

Koko summan arvoksi saadaan $P(X \geq 13) \approx 0.0245$

Testaamme tässä kuitenkin kaksisuuntaista hypoteesia; emmehan tiedä ennalta, kumpi alleeleista, 1 vai 2, on assosioitunut sairausstatukseen. Yhtä poikkeavia kuin tapaukset $X \geq 13$ ovat kuitenkin myös tapaukset $X \leq 4$. Koska binomijakauma on tässä tapauksessa symmetrinen, lopulliseksi kaksisuuntaiseksi p-arvoksi saadaan $2 \times 0.0245 = 0.049$. Saatu tulos antaa näin ollen lievää todistetta nollihypoteesia vastaan.

7.2.2 TDT-testisuureen approksimointi

Edellä käytetyn binomijakaumaan perustuvan eksaktin testin ohella käytetään laskennallisesti helpompaa χ^2 -testisuureta, joka tunnetaan myös alkuperäisenä TDT-testisuurena ja McNemarin testisuurena:

$$\begin{aligned} \chi^2 &= \sum \frac{(O-E)^2}{E} = \frac{(t_{12}-0.5(t_{12}+t_{21}))^2}{0.5(t_{12}+t_{21})} + \frac{(t_{21}-0.5(t_{12}+t_{21}))^2}{0.5(t_{12}+t_{21})} \\ &= \frac{(t_{21}-t_{12})^2}{t_{12}+t_{21}} \end{aligned}$$

Testisuureen arvo jakautuu asymptoottisesti χ^2 -jakauman mukaisesti. Tällöin esitetään peukalosääntönä, että testiä voi käyttää luotettavasti, kun kukin odotetuista alleelifrekvensseistä on vähintään viisi. Tämä edellyttäisi noin kymmentä transmissiotapahtumaa.

Testille valittuun merkitsevyytasoon liittyvät tunnusluvun kriittiset rajat saadaan jakauman kertymäfunktio-aulukosta:

p	0.100	0.050	0.025	0.010	0.005
d.f. = 1	2.71	3.84	5.02	6.63	7.88

Esimerkki [SME93]:

Tarkasteltavana on yksittäinen polymorfismi ja sen yhteys sairauteen. Otokseen kerätyissä ydinperheissä oli mukana 57 vanhempaa, jotka ovat heterotyygootteja alleeleille 1 ja X. Lapsille transmittoituneita alleeleja oli 124 kappaletta, joista 78 kappaletta 1-alleeleita ja 46 kappaletta X-alleeleita.

Lasketaan arvo TDT-testisuurelle: $\chi^2 = \frac{(78-46)^2}{78+46} \approx 8.26$.

Verrataan arvoa nollahypoteesitilanteen mukaiseen χ^2 -jakaumaan vapausasteilla yksi. Koska testisuureen arvoa vastaava p-arvo on alhainen ($p < 0.005$), antaa aineisto todistetta nollahypoteesia vastaan.

7.2.3 Analyysi Genehunter-ohjelmalla

Genehunter-ohjelma tekee Spielmanin kaksialleelisen TDT-testin kunkin markkerilokuksen jokaiselle alleelille erikseen niputtaen yhteen kaikki muut paitsi tarkasteltavan alleelin. SNP-markkereita käytettäessä niputtamista ei tietenkään tarvitse tehdä.

Genehunter-ohjelma saa syötteenään kaksi Linkage-muotoista tiedostoa: sukupuutiedoston ja parametritiedoston (ks. luku 3). Alla on hieman alkua 20 perhettä käsittävästä esimerkkitiedostosta `trio1.pre`. Huomaa, että kaikki perheet ovat trioja (isä, äiti ja lapsi). Isän ja äidin sairausstatus on tiedostossa merkitty tuntemattomaksi, mutta lapsista tiedetään, että he ovat sairaita.

1	1	0	0	1	0	1	1	1	1	2	1	1	2	1	2	1	1
1	2	0	0	2	0	1	1	1	1	1	2	2	1	2	2	1	2
1	3	1	2	2	2	1	1	1	1	2	1	1	2	1	2	1	1
2	1	0	0	1	0	1	2	1	1	1	2	1	1	1	1	2	1
2	2	0	0	2	0	2	1	1	2	2	2	2	1	1	2	1	1
2	3	1	2	2	2	1	2	1	1	1	2	1	2	1	1	2	1
3	1	0	0	1	0	2	1	1	1	1	2	1	2	1	2	2	2
3	2	0	0	2	0	1	2	1	1	1	2	2	2	1	1	1	1
3	3	1	2	1	2	2	1	1	1	1	1	1	2	1	1	2	1
4	1	0	0	1	0	2	2	1	2	1	1	1	2	1	1	2	2
4	2	0	0	2	0	1	2	1	1	2	2	2	2	1	2	2	2
4	3	1	2	2	2	2	1	1	1	1	2	1	2	1	1	2	2

Luomme seuraavaksi parametritiedoston `trio1.dat`. Parametritiedoston alleelifrekvensseillä ja sairausalleelin yleisyydellä ei ole merkitystä; edelliset on tässä esimerkissä arvioitu aineiston perusteella.

7	0	0	5			
0	0.0	0.0	0			
1	2	3	4	5	6	7
1	2					

```

0.99 0.01
1
0.001000 0.999000 0.999000
3 2 # M15
0.632 0.368
3 2 # M16
0.078 0.922
3 2 # M17
0.809 0.191
3 2 # M18
0.575 0.425
3 2 # M19
0.73 0.27
3 2 # M20
0.52 0.48
0 0
0.10 0.02 0.02 0.02 0.02 0.02
1 0.1 0.45

```

Käynnistämme Genehunter-ohjelman ja ajamme analyysin komennoilla

```

np1:1> load markers trio1.dat
Parsing Linkage marker data file...
6 markers read (last one = M20)

np1:2> tdt trio1.pre

```

Saamme tulosteen alleelikohtaisista TDT-testeistä:

```

TDT Summary - (0 non-original affecteds):
Marker M15          trans untrans  Chi2  p-val
M15      - Allele 1    10    10    0.00  1.000000
M15      - Allele 2    10    10    0.00  1.000000
Marker M16          trans untrans  Chi2  p-val
M16      - Allele 1    10     3    3.77  0.052204
M16      - Allele 2     3    10    3.77  0.052204
Marker M17          trans untrans  Chi2  p-val
M17      - Allele 1    17     7    4.17  0.041227  +
M17      - Allele 2     7    17    4.17  0.041227  -
Marker M18          trans untrans  Chi2  p-val
M18      - Allele 1    19     5    8.17  0.004267  ++
M18      - Allele 2     5    19    8.17  0.004267  --
Marker M19          trans untrans  Chi2  p-val

```

M19	- Allele 1	17	3	9.80	0.001745	++
M19	- Allele 2	3	17	9.80	0.001745	--
Marker M20		trans	untrans	Chi2	p-val	
M20	- Allele 1	2	7	2.78	0.095581	
M20	- Allele 2	7	2	2.78	0.095581	

Genehunter merkitsee plus-merkein (miinus-merkein) rivit, joihin liittyy merkitsevä p-arvo :

- yksi plus-merkki: tilastollisesti melkein merkitsevä tulos, $p < 0.05$
- kaksi plus-merkkiä: tilastollisesti merkitsevä tulos, $p < 0.01$
- kolme plus-merkkiä: tilastollisesti erittäin merkitsevä tulos, $p < 0.001$

Näemme, että esimerkiksi lokuksen M19 alleeli 1 oli siirtynyt heterotsygootivanhemmalta 17 kertaa 20 mahdollisesta kerrasta. Vaikka tulos näyttää erittäin lupaavalta, on syytä ottaa huomioon moninkertaisen testauksen ongelma: kun testataan riittävän innokkaasti, saadaan joskus myös sattuman vaikutuksesta merkitsevältä näyttäviä arvoja. Niinpä saatua p-arvoa on korjattava ylöspäin esimerkiksi permutaatiotestin avulla.

Permutaatiotestaus

Edellisessä esimerkissä tehtiin joukko testejä ja valittiin tuloksista se, joka on merkitsevin. Permutaatiotestauksen tavoitteena on selvittää kokeellisesti, miten korkein havaittu testisuureen arvo jakautuu nollahypoteesin (ei kytkentää) mukaisesti generoiduissa aineistoissa. Vertaamalla havaittua korkeinta arvoa tähän korkeimpien arvojen jakaumaan saamme korjatun p-arvon, joka ottaa huomioon tekemämme moninkertaisen testauksen.

Permutaatiotestaus on Genehunter-ohjelmiston avulla hyvin helppoa. Se käynnistetään komennolla `perm1` sen jälkeen, kun varsinainen TDT-analyysi on jo tehty.

```
np1:3> perm1 10000
```

Parametrina annettu luku (10000) kertoo iteraatioiden lukumäärän. Mitä suurempaa lukua käytetään, sitä tarkempi empiirinen p-arvo saadaan, mutta sitä kauemmin myös laskenta kestää. Permutaatiotestauksessa Genehunter muodostaa jokaisella kierroksella keinotekoisien aineiston, jossa kumpikin vanhemmilla olevista kahdesta alleelistä on siirtynyt yhtä suurella todennäköisyydellä (0.5) jälkeläiselle. Tästä keinotekoisesta aineistosta ohjelma tekee TDT:n normaaliin tapaan ja poimii parhaan havaitun testisuureen arvon. Aineiston muodostaminen ja analyysi toistetaan useita (tässä 10000) kertoja,

jolloin saadaan empiirinen jakauma parhaalle χ^2 -testisuureen arvolle. Empiirisen p-arvon saamiseksi lasketaan, kuinka monta kertaa havaittu testisuureen maksimi-arvo ylittyi keinotekoisissa nollahypoteesin mukaisissa aineistoissa. Näiden kertojen lukumäärä jaetaan lopuksi iteraatioiden kokonaismäärällä.

Esimerkkitapauksemme tuloste on seuraavanlainen:

PERMUTATION SUMMARY:
38 of 10000 had a larger maximum value than the real best (9.80)
65 of 10000 had as many tests (2) exceeding p=.01
80 of 10000 had as many tests (1) exceeding p=.001
10000 of 10000 had as many tests (0) exceeding p=.0001

Havaittu χ^2 -tunnusluvun maksimi-arvomme ylittyi 38 keinotekoisessa aineistossa. Tästä saadaan empiiriseksi p-arvoksi $38/10000 \approx 0.004$, eli tulosta voi moninkertaisesta testauksesta huolimatta pitää tilastollisesti merkitsevänä.

Tuloksen viimeisillä riveillä kerrotaan, kuinka usein havaittujen merkitsevien testien lukumäärä ylittyi simuloituissa aineistoissa. Näin saatujen empiiristen p-arvojen laskennassa on siis käytetty tunnuslukuna merkitsevien testien lukumäärää; näin laskettujen p-arvojen tulkinta on epäselvempi.

7.3 Assosiaatioanalyysin tilastollisen voiman arvointi

Assosiaatioanalyysin voimalaskelmien tavoitteena on selvittää, onko suunniteltu koeasetelma (otos, markkerikartta, analyysimenetelmät) sellainen, että sen avulla on riittävät mahdollisuudet löytää haettavan alttiusgeenin sijainti. Laskelmien lähtökohdaksi joudutaan aina tekemään oletus tai oletuksia siitä, minkälaista mallia sairaus noudattaa ja mikä on geenien ja sairausfenotyypin välinen suhde.

Assosiaatioanalyysin voimalaskelmiin soveltuu Internetin kautta käytettävä GPC (Genetic Power Calculator) -ohjelmisto [PCS03]. Ohjelmiston avulla voidaan muun muassa arvioida voimaa ja otoskokoa tilanteissa, joissa etsitään χ^2 -testillä assosiaatiota markkerilokuksessa olevan yhden alleelin ja sen kanssa kytKentäepätasapainossa olevan sairauslokuksen riskialleelin kanssa. Haplotyyppipohjaisissa menetelmissä todellinen voima saattaa olla suurempi, jolloin ohjelmiston tuottama voima-arvio on konservatiivinen.

Esimerkki GPC-ohjelmiston käytöstä assosiaatioanalyysin voiman arvioinniksi (päävalikon vaihtoehto *Case-control for discrete traits*) syötteistä on kuvassa 7.4. Syötelomakkeella käyttäjä määrittää testattavan koeasetelman. Esimerkin tapauksessa laskelma tehdään tilanteelle, jossa tarkasteltava sairaus on varsin yleinen (prevalenssi 4%), ja jossa oletettu riskialleeli on varsin yleinen väestössä (frekvenssi 10%). Riskialleelin kantajuuden oletetaan nostavan sairastumistodennäköisyyden kaksinkertaiseksi (heterotsygootin Aa ja homotsygootin AA relativiset riskit). Kyseessä on tällöin monitekijäinen sai-

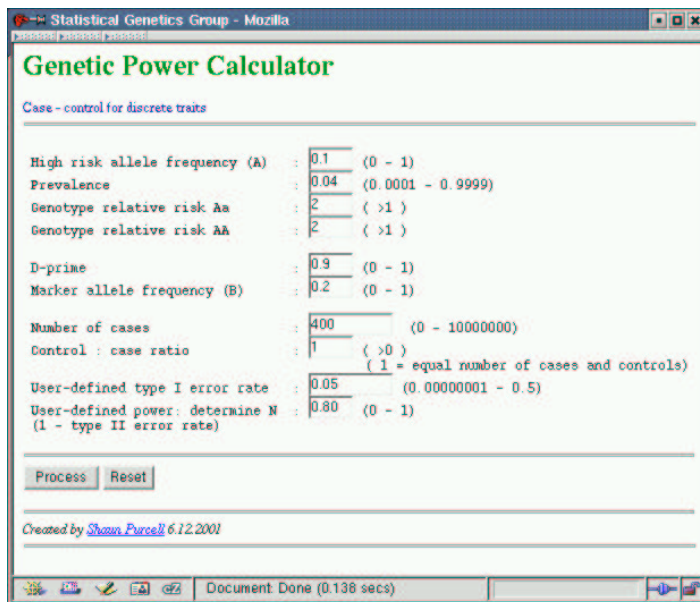
raus, jossa yleinen riskialleeli nostaa sairastumistodennäköisyyttä, mutta ei yksin kykene selittämään sairauden yleisyyttä väestössä.

Lomakkeessa kuvataan seuraavaksi tarkasteltavan markkerin tiedot. Markkereita oletetaan olevan käytössä vain yksi, ja yhden sen alleeleista oletetaan assosioituvan sairausstatukseen. Sairauslokuksen ja markkerilokuksen välisen kytkentäepätasapainon voimakkuudeksi D' -mitalla mitattuna arvioidaan 0.9 (arvo 1 vastaa täydellistä kytkentäepätasapainoa), ja assosioituvan markkerialleelin frekvenssiksi 0.2. Edellä kuvatut arvot riippuvat väestöhistoriasta ja käytetystä markkerikartasta, ja ne voidaan arvioida esimerkiksi analyysin kannalta huonoimman vaihtoehdon mukaan.

Ohjelmalla saadaan laskettua sekä annettua otoskokoa vastaava tilastollinen voima että annettuun tilastolliseen voimaan vaadittava otoskoko.

Esimerkissä tapausyksilöiden määräksi on asetettu 400, ja verrokkiyksilöitä on sama määrä. Tämä kuvataan asettamalla verrokkien ja tapauksen suhdeluvuksi 1. Jos verrokkeja olisi vaikkapa 600, asetettaisiin suhdeluvuksi 1.5.

Syötelomakkeen lopussa annetaan koeasetelmassa sallittava väärän positiivisen tuloksen todennäköisyys (esimerkissä 0.05) sekä tavoiteltava voimataso. Jälkimmäinen saa esimerkiksi arvon 0.8, mikä tarkoittaa, että markkerille tehdyn testin pitäisi paljastaa kuvatun kaltainen alttiusgeeni 80% todennäköisyydellä. Mitä suuremmaksi voimavaatimus asetetaan, sitä suurempi otoskoko tarvitaan.



Kuva 7.4: Syötelomake assosiaatioanalyysin voiman arvioinniksi Genetic Power Calculator (GPC) -työkalulla.

Esimerkin syötteistä laskettu tuloste on kuvassa 7.5. Tulosteen yläosassa on koottuna sairausmalliin liittyvät parametrit ja niistä laskettu riskialleelia kan-

tamattoman homotsygootin penetranssi, joka on esimerkissämme 0.03361. Kantajilla tämä riski nousee siis kaksinkertaiseksi.

Seuravaassa laatikossa esitetään oletetun markkerin lähtötiedot sekä niistä lasketut penetranssit kolmelle markkerilokuksen genotyypille (bb, Bb ja BB). Niitä seuraavat vedonlyöntisuhteet (*odds ratio*) kuvaavat genotyypin kantajien osuuksien suhdetta kahdessa otoksessa. Esimerkiksi heterotsygootin Bb vedonlyöntisuhde 1.459 tarkoittaa, että heterotsygoottien (Bb) suhde eikantajiin (bb) on sairaiden otoksessa 1.459-kertainen terveiden otoksesta laskettuun vastaavaan suhteeseen verrattuna.

Tulosten kohdessa *Expected allele frequencies* on laskettu alleelifrekvenssit kummassakin otoksessa, ja seuraavassa laatikossa esitetään niistä lasketut genotyypifrekvenssit. Näitä seuraa Hardy-Weinbergin tasapainoa testaavan testisuureen arvo (*H-W test NCP*), ja todennäköisyys, että testisuure antaa tilastollisesti merkitsevän tuloksen. Voimaltaan heikohkon testin taustalla on ajatus, että poikkeama Hardy-Weinbergin tasapainosta on suurempaa sairaiden joukossa, jossa riski- ja sen seurauksena markkerialleelin suhteen homotsygootit yksilöt ovat yliedustettuina.

Kaikkein kiinnostavimmat tulokset sijaitsevat taulukon lopussa: kutakin väärin positiivisten todennäköisyyttä (*Alpha*) kohti on esitetty saavutettava tilastollinen voima (*Power*). Oikeanpuolinen sarake (*N cases for 80% power*) puolestaan kuvaa, kuinka suuri otos kyseisellä väärin positiivisten todennäköisyydellä tarvitaan, jotta saavutetaan käyttäjän antama 80 prosentin tilastollinen voima.

GPC-ohjelmiston verkkosivu on <http://statgen.iop.kcl.ac.uk/gpc/>. Ohjelmistoa voidaan käyttää myös periytymisen epätasapainotestin (TDT) voiman arvioimiseksi; käyttö on varsin analogista edellä esitetyn tapaus-verrokiasetelmaa koskevan laskennan kanssa.

Käytännössä assosiaatioanalyysin tilastollisen voiman arviointi on huomattavan vaikeaa, sillä syötteenä olevien parametrien tarkka arviointi on usein mahdotonta. Tilastollinen voima riippuu ratkaisevasti haettavan riskialleelin oletetusta efektistä. Tämä ei yleensä ole tiedossa, ja riskialleelitkin voivat vaihdella populaatiosta toiseen. Niinpä voiman arvioinnissa joudutaan turvautumaan lukuisiin entä-jos-tyyppisiin laskelmiin erilaisille geneettisille efekteille.

Genetic Power Calculator - Mozilla

Genetic Power Calculator

Case-control for discrete traits

Case-control parameters

Number of cases	400
Number of controls	400
High risk allele frequency (A)	0.1
Prevalence	0.04
Genotypic relative risk Aa	2
Genotypic relative risk AA	2
Genotypic risk for aa (baseline)	0.03361

Marker locus B

High risk allele frequency (B)	0.2
Linkage disequilibrium (D')	0.9
Penetrance at marker genotype bb	0.03428
Penetrance at marker genotype Bb	0.04926
Penetrance at marker genotype BB	0.05743
Genotypic odds ratio Bb	1.459
Genotypic odds ratio BB	1.716

Expected allele frequencies

	Case	Control
B	0.2545	0.1977
b	0.7455	0.8023

Expected genotype frequencies

	Case	Control
BB	0.05743	0.03927
Bb	0.3941	0.3169
bb	0.5485	0.6438
H-W test NCP	0.5958	0.0004944
Power (alpha=0.05)	0.1205	0.05006

Case-control statistics

Sample NCP = 7.355

Alpha	Power	N cases for 80% power
0.1	0.8571	336.2
0.05	0.774	426.8
0.01	0.5542	635.1
0.001	0.2815	928.6
0.05	0.774	426.8

Document: Done (0...)

Kuva 7.5: Esimerkki Genetic Power Calculator (GPC) -työkalulla saadusta assosiaatioanalyysin voima-arviosta.

8 Haplotyyppien rakentaminen ja analyysi

Edellä on tarkasteltu yksittäisissä markkerilokuksissa esiintyvien alleelien assosioitumista yksilön sairausstatukseen. Jos käytössä on tiheä markkerikartta, voidaan analyysin voimaa usein parantaa määrittämällä kumpaan vastinkromosomiin – isältä vai äidiltä perittyyn – kukin alleeli kuuluu ja testaamalla peräkkäisten markkerilokusten yli ulottuvien, samassa vastinkromosomissa olevien alleelien kombinaatioiden – haplotyyppien – assosioitumista sairausstatukseen. Haplotyyppianalyysin merkitys korostuu tiheitä SNP-karttoja käytettäessä, sillä useiden SNP-markkereiden samanaikaisella tarkastelulla voidaan kompensoida yksittäisen markkerin puutteellista informatiivisuutta. Tässä luvussa käsittelemme haplotyyppipohjaisia assosiaatiotestejä, haplotyyppien rakentamista sekä väestöissä esiintyvien odotettua useammin kokonaisina periytyvien alueiden eli niin sanottujen haplotyyppi-blokkien etsintää.

8.1 Periytymisen epätasapainotesti haplotyypeille

Edellisessä luvussa tarkasteltiin periytymisen epätasapainotestin teoriaa ja käytännön laskentaa. Periytymisen epätasapainotestin voiman paraneminen haplotyypeihin siirryttäessä riippuu mm. markkerien välisestä kytkentäepätasapainosta, populaatiohistoriasta ja lokuksissa tapahtuneista mutaatioista. Tarkastelemme seuraavaksi, miten periytymisen epätasapainotesti toteutetaan haplotyypeille Genehunter-ohjelmalla.

Kahden lokuksen muodostaman haplotyyppin TDT-analyysi voidaan suorittaa komennolla `tdt2` ilman tiedoston nimeä. Tätä ennen on kuitenkin annettava komennot `load` ja `tdt` (ks. luku 7.2.3). Valinnaisena voidaan antaa parametri, joka määrää kuinka mones haplotyyppin toinen lokus on ensimmäisestä lokuksesta: esim. komento `tdt2 2` testaa haplotyyppit lokusten 1-3, 2-4, 3-5 jne. välillä. Vastaava komento permutaatiotestille on `perm2`. Seuraavassa testamme kahden markkerin mittaiset esimerkkiaineistossa esiintyvät haplotyyppit ja arvioimme tuloksen tilastollista merkitsevyyttä permutaatiotestillä:

```
np1:4> tdt2
```

```
Markers 1 and 2      trans untrans  Chi2  p-val (M15 - M16)
```

Haplotype:	1	1	8	4	1.33	0.248213	
Haplotype:	1	2	2	6	2.00	0.157299	
Haplotype:	2	1	9	6	0.60	0.438578	
Haplotype:	2	2	0	3	3.00	0.083264	
Markers 2 and 3			trans	untrans	Chi2	p-val	(M16 - M17)
Haplotype:	1	1	16	4	7.20	0.007290	++
Haplotype:	1	2	4	9	1.92	0.165518	
Haplotype:	2	1	1	3	1.00	0.317311	
Haplotype:	2	2	1	6	3.57	0.058782	
Markers 3 and 4			trans	untrans	Chi2	p-val	(M17 - M18)
Haplotype:	1	1	15	0	15.00	0.000108	+++
Haplotype:	1	2	5	9	1.14	0.285049	
Haplotype:	2	1	0	3	3.00	0.083264	
Haplotype:	2	2	3	11	4.57	0.032509	-
Markers 4 and 5			trans	untrans	Chi2	p-val	(M18 - M19)
Haplotype:	1	1	12	0	12.00	0.000532	+++
Haplotype:	1	2	1	2	0.33	0.563703	
Haplotype:	2	1	6	8	0.29	0.592980	
Haplotype:	2	2	0	9	9.00	0.002700	--
Markers 5 and 6			trans	untrans	Chi2	p-val	(M19 - M20)
Haplotype:	1	1	3	5	0.50	0.479500	
Haplotype:	1	2	16	0	16.00	0.000063	++++
Haplotype:	2	1	2	4	0.67	0.414216	
Haplotype:	2	2	0	12	12.00	0.000532	---
np1:5> perm2 10000							
PERMUTATION SUMMARY:							
0 of 10000 had a larger maximum value than the real best (16.00)							
0 of 10000 had as many tests (7) exceeding p=.01							
0 of 10000 had as many tests (6) exceeding p=.001							
0 of 10000 had as many tests (4) exceeding p=.0001							

Tuloksista huomataan, että lokusten M19 ja M20 muodostama haplotyyppi 1-2 antaa pienemmän empiirisen p-arvon kuin kumpikaan lokus erikseen, vaikka informatiivisia havaintoja oli vain 16 kpl.

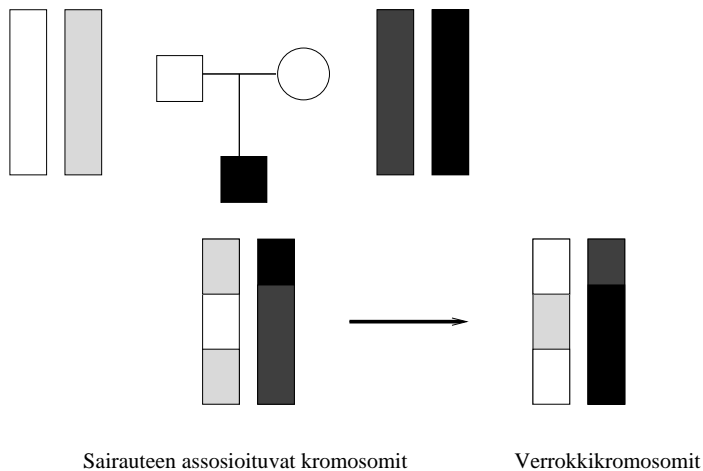
Haplotyyppianalyysiä voidaan laajentaa vielä kolmen ja neljän lokuksen analyysiin komennoilla tdt3 ja tdt4, mutta niille ei voi tehdä enää permutaatiotestiä.

8.2 Perhepohjainen haplotyyppaus

Assosiaatioanalyysin voimaa pyritään usein parantamaan hyödyntämällä otoksessa olevista yksilöistä saatavaa haplotyyppi-informaatiota. Haplotyyppien määrittämiseen on kaksi tapaa: perhepohjainen ja väestöpohjainen haplotyyppaus. Tarkastelemme seuraavassa perhepohjaista haplotyyppausta; väestöpohjaiseen haplotyyppaukseen palataan luvussa 8.4.

Perhepohjaisessa haplotyyppauksessa genotyypit hankitaan paitsi mukaan otettavilta sairailta yksilöiltä myös heidän perheenjäseniltään, yleensä vanhemmilta. Tällä tavoin kerättävä aineisto koostuu siis trioista eli ydinperheistä, joissa on isä, äiti ja sairas lapsi. Vertaamalla isän, äidin ja lapsen genotyyppiä kussakin markkerilokuksessa voidaan yleensä päätellä, kumpi alleeleista on peritty isältä ja kumpi äidiltä. Isältä perityt alleelit muodostavat tällöin yhden haplotyyppin ja äidiltä perityt toisen. Jos mukana on useampia sisarusksia, voidaan mahdolliset rekombinaatiokohdat usein tunnistaa.

Keräämällä ydinperheitä assosiaatioanalyysiä varten saavutetaan toinenkin etu: vanhempien genotyyppi-informaation avulla voidaan luoda verokkiaaineisto. Tällöin kummankin vanhemman niistä alleeleista, jotka eivät siirtyneet jälkeläiselle, voidaan muodostaa keinotekoinen *pseudoverrokkikromosomi*. Yhdestä triosta saadaan tällä tavoin kaksi sairauteen assosioituvaa ja kaksi pseudoverrokkikromosomia (kuva 8.1).



Kuva 8.1: Ydinperheestä muodostettavat sairauteen assosioituvat ja pseudoverrokkikromosomit. Sairaalan lapsen molemmat kromosomit merkitään sairauteen assosioituviksi. Pseudoverrokkikromosomit koostetaan niistä kromosomisegmenteistä, jotka eivät periytyneet sairaalle lapselle. Kromosomien rakentaminen tehdään käytännössä markeridatan perusteella.

Assosiaatioanalyysi perhepohjaisesti muodostetuille haplotyypeille on analoginen periytymisen epätasapainotestin (TDT) kanssa sillä erotuksella, että jälkimmäisessä tarkastellaan vain heterotsygootteja vanhempia. Toisaalta perhepohjaisesti muodostettuja haplotyyppiä voidaan analysoida myös

muilla assosiaatiomenetelmillä.

Perhepohjainen haplotyyppaus voidaan tehdä kytkentäanalyysiohjelmistoilla kuten Merlin ja Genehunter. Merlin-ohjelmalla voidaan muodostaa todennäköisimmät (genotyyppiaineistoon parhaiten soveltuvat) haplotyyppit, tai haplotyyppikonfiguraatio voidaan valita satunnaisotannalla siten, että otannassa kutakin konfiguraatiota painotetaan sitä vastaavalla uskottavuusfunktion arvolla. Genehunter tuottaa aina todennäköisimmän haplotyyppikonfiguraation.

Seuraavassa esimerkissä haplotyyppaamme Merlin-ohjelman avulla esimerkkiaineistomme, jota käytimme TDT-testin demonstroimiseen. Tämän jälkeen rakennamme tulosteesta aineiston, joka koostuu sairauteen assosioituvista ja pseudoverrokkikromosomeista. Lopuksi testaamme löytyviä alleeli- ja haplotyyppiassosiaatioita.

8.2.1 Haplotyyppaus Merlin-ohjelmalla

Haplotyyppien muodostaminen Merlin-ohjelmistolla käy yhtä yksinkertaisesti kuin kytkentäanalyysikin. Komentorivillä valitaan haplotyyppien poimintavaihtoehdoksi joko `--best` (todennäköisin haplotyyppikonfiguraatio), `--sample` (satunnaisotantaan perustuva datan kanssa yhteensopiva haplotyyppikonfiguraatio) tai `--zero` (haplotyyppit, joissa ei rekombinaatioita):

```
corona% merlin -d tdt.snp.dat -p tdt.snp.pre --best
--horizontal
```

Merlin tuottaa oletusarvoisesti tulosteen haplotyypeistä siten, että kukin kromosomeista on kuvattu pystysuunnassa merkkigrafiikalla. Esitysmuoto on hyvä visuaalisia tarkasteluja varten, mutta se soveltuu huonosti koneelliseen jatkokäsittelyyn. Määreellä `horizontal` ilmoitetaan, että haplotyyppit kuvataan vaakasuuntaisena tulosteena, jossa on yksi rivi yhtä yksilöä kohden.

Haplotyyppauksen tulos ilmestyy tiedostoon `merlin.chr`, ja se näyttää ensimmäisen ydinperheen osalta tältä (välilyöntejä on karsittu tulosteen kaventamiseksi):

FAMILY 1 [Uninformative]														
1	(FOUNDER)	2	2	1	1	1	2	1	2	2	1	1	1	A 1,2 A 1,2 1 1 1 2 2 1 1
1	(FOUNDER)	1	2	1	2	1	2	2	2	1	1	A 2,1 A 2,1 2 1 2 1 1 1 2		
2	(FOUNDER)	2	2	1	1	2	2	1	2	2	1	1	A 2,1 A 2,1 2 1 1 1 1 1 1	
2	(FOUNDER)	1	2	1	1	2	2	1	1	2	2	1	A 1,2 A 1,2 2 2 1 1 1 2 2	
3	(MATERNAL)	2	2	1	1	2	2	1	2	2	1	1	A 2,1 A 2,1 2 1 1 1 1 1 1	
3	(PATERNAL)	2	2	1	1	1	2	1	2	2	1	1	A 1,2 A 1,2 1 1 1 2 2 1 1	

Tulosteesta nähdään, että haplotyyppaus on onnistunut yksikäsitteisesti lukuunottamatta kahta markkeria, joiden kohdalla on A-kirjain ja kaksi yhtä todennäköistä vaihtoehtoa pilkulla erotettuna. Näissä kohdissa kukin kolmesta yksilöstä on ollut samanlainen heterotsygootti 1,2.

8.2.2 Tapaus-verrokkitiedoston rakentaminen

Merlin-ohjelman tuloste voidaan muuttaa tapaus-verrokkiaineistoksi CSC:llä laaditulla `merlin2assoc`-työkalulla. Työkalulle annetaan parametreina alkuperäinen sukupuutiedosto sekä Merlin-ohjelman tuottama haplotyyppitiedosto `merlin.chr`. Ohjaamme tuloksen tiedostoon `assoc.data`:

```
corona% merlin2assoc tdt.snp.pre merlin.chr > assoc.data
```

Tuloste alkaa ensimmäisestä perheestä muodostetuilla haplotyypeillä:

```
a 2 2 1 1 2 2 1 2 2 2 1 1 0 0 2 1 1 1 1 1 1
c 1 2 1 1 2 2 1 1 2 2 1 1 0 0 2 2 1 1 1 2 2
a 2 2 1 1 1 2 1 2 2 1 1 1 0 0 1 1 1 2 2 1 1
c 1 2 1 2 1 2 1 2 2 2 1 1 0 0 2 1 2 1 1 1 2
```

Ensimmäinen sarake kertoo kunkin kromosomin statuksen: sairauteen assosioitunut (a) tai verrokki (c). Verrokkikromosomit on rakennettu siten, että vanhempien genotyypeistä on valittu jälkeläiselle siirtymätön alleeli. Alleelit on merkitty puuttuviksi niissä kohdissa, joissa haplotyyppaus ei onnistunut yksikäsitteisesti.

8.3 Assosioituvien haplotyyppien etsintä

Kytkenäepätasapainoon perustuvaan geenikartoitukseen on olemassa joukko menetelmiä ja valmisohjelmistoja mutta monet menetelmistä ovat varsin hitaita. Alleeli- ja haplotyyppiassosiaatioiden automaattiseksi testaamiseksi tiedonlouhintatyyppisesti voimme käyttää sovelluspalvelimeen asennettua ja CSC:llä kehitettyä Haplo-assoc-työkalua. Kyseessä on C-kielinen ohjelma, joka etsii aineistosta kaikki esiintyvät haplotyyppit ja testaa niiden tilastollisen merkitsevyyden χ^2 -testillä.

Ohjelmalle voidaan antaa kolme parametria:

χ^2 -raja (valitsin -c): Raja sille, mikä χ^2 -testisuureen arvon on vähintään oltava, jotta testi otetaan mukaan tulosteeseen. Asettamalla raja korkeaksi voidaan tulostetta lyhentää karsimalla heikkoja, usein satunnai-

47	0	124	124	57.990	4	-	-	-	-	-	-	-	1	1	1	-	-	-	-
48	1	126	126	55.963	3	-	-	-	-	-	-	-	1	1	1	-	-	-	-
42	0	116	116	51.284	4	-	-	-	-	-	-	-	1	1	1	2	-	-	-
41	0	114	114	49.989	5	-	-	-	-	-	-	-	1	1	1	2	-	-	-
41	0	116	116	49.801	5	-	-	-	-	-	-	-	1	1	1	2	1	-	-
40	0	114	114	48.511	6	-	-	-	-	-	-	-	1	1	1	2	1	-	-
64	13	136	136	47.118	4	-	-	-	-	-	-	-	1	2	1	2	-	-	-
41	1	114	114	46.697	5	-	-	-	-	-	-	-	1	1	2	1	2	-	-
125	63	168	168	46.420	1	-	-	-	-	-	-	-	1	-	-	-	-	-	-
32	1	114	114	34.049	5	-	-	-	-	-	-	2	2	1	1	1	-	-	-
34	2	124	124	33.275	4	-	-	-	-	-	-	2	1	1	1	-	-	-	-
28	0	110	110	32.083	5	-	-	-	-	-	-	2	1	1	1	1	-	-	-
27	0	104	104	31.028	6	-	-	-	-	-	-	-	1	1	1	2	1	2	-
26	0	100	100	29.885	6	-	-	-	-	-	-	2	2	1	1	1	1	-	-
26	0	102	102	29.798	7	-	-	-	-	-	-	-	1	1	1	2	1	2	-
93	45	164	164	28.822	3	-	-	-	-	-	-	-	-	-	-	2	1	2	-
28	2	100	100	26.510	6	-	-	-	-	-	-	-	1	2	1	2	2	1	-
23	0	100	100	25.989	6	-	-	-	-	-	-	2	1	1	1	2	-	-	-
141	94	186	186	25.524	2	-	-	-	-	-	-	-	-	-	-	2	1	-	-
22	0	100	100	24.719	7	-	-	-	-	-	-	2	1	1	1	2	1	-	-
21	0	90	90	23.774	7	-	-	-	-	-	-	2	2	1	1	1	2	-	-
20	0	84	84	22.703	7	-	-	-	-	-	-	-	1	1	2	1	2	2	1
33	6	112	112	22.633	5	-	-	-	-	-	-	-	1	2	1	2	2	-	-
20	0	90	90	22.500	8	-	-	-	-	-	-	2	2	1	1	1	2	1	-
148	106	186	186	21.894	1	-	-	-	-	-	-	-	2	-	-	-	-	-	-
44	13	124	124	21.891	4	-	-	-	-	-	-	-	-	1	2	2	1	-	-
22	1	94	94	21.847	6	-	-	-	-	-	-	-	1	1	2	1	2	2	-

Alttiusgeenin sijainti markkerin 14 tuntumassa näkyy selvästi. Ensimmäiseltä tulosriviltä nähdään, että vahvimmin assosioituva haplotyyppi 1-1 esiintyy 71 kertaa sairautteen assosioituvassa ja vain 5 kertaa verrokkihaplotyyppissä. Vastaavasti jokin muu alleeli esiintyi sairautteen assosioituvassa kromosomissa 140 kertaa ja verrokkikromosomissa yhtä usein. Saatu χ^2 -testisuureen arvo haplotyypille on 78.669, joka on kiistatta tilastollisesti erittäin merkitsevä. Todellisuudessa signaali on harvoin näin selkeä.

Tuloksia tulkittaessa on huomioitava, että mukana ovat olleet vain sellaiset haplotyyppit, jotka on pystytty yksikäsitteisesti määrittämään. Tämän on osoitettu [DKTC00] aiheuttavan harhaa perhepohjaisessa assosiaatiotestauksessa, ja harha pahenee haplotyyppien pidentyessä. Harhan ansiosta χ^2 -testien tarkkaa tilastollista merkitsevyyttä ei voi arvioida suoraviivaisesti. Eräänä ratkaisuna esitetään [DKTC00] p-arvojen määrittäminen empiirisesti. Tämä voidaan tehdä permutaatiotestauksen avulla. Tällöin haplotyyppatun aineiston statussarake sekoitetaan ennen analyysiä. Sekoitettu aineisto analysoidaan samalla tavoin kuin alkuperäinen aineisto, ja sekoitetun aineiston paras χ^2 -testisuureen arvo talletetaan. Tätä toistetaan silmukassa useita kertoja, ja lopuksi havaitusta aineistosta laskettuja arvoja voidaan verrata tähän empiiriseen jakaumaan.

8.4 Väestöpohjainen haplotyyppaus

Assosiaatioanalyysin lähtökohdaksi voidaan ottaa populaatiopohjainen otanta: tällöin väestöstä kerätään sekä sairaita yksilöitä että terveitä verrokkiyksilöitä, jolloin vanhempia ei tarvitse kerätä eikä genotyypata. Väestöpohjaisen aineiston kerääminen on yleensä helpompaa ja halvempaa kuin vastaavan kokosen perhepohjaisen aineiston. Toisaalta otantatapa tuottaa assosiaatioanalyysissä helposti vääriä positiivisia tuloksia, jos sairaat ja verokkikromosomit eroavat geneettiseltä taustaltaan toisistaan.

Irrallisista yksilöistä koostuvaan aineistoon voidaan tehdä alleeliassosiaatio-testejä, mutta haplotyyppi-informaatiota ei lähtökohtaisesti ole käytössä. Jos yksilöt on kuitenkin kerätty samasta väestöstä, joissa esiintyy rajallinen määrä eri alkuperää olevia kromosomeja, voidaan väestötason haplotyyppijakau-mia kuitenkin hyödyntää yksilön haplotyyppien muodostamisessa. Toisaalta perhepohjaisen haplotyyppauksen lopputulosta voidaan täydentää väestöpohjaisin menetelmin.

Populaatiopohjaiseen haplotyyppaukseen on erilaisia algoritmisia lähestymistapoja kuten Clarkin algoritmi (jossa rakennetaan iteratiivisesti sellaista haplotyyppien joukkoa, joka selittää genotyyppidatan), EM-algoritmi (*Expectation maximization algorithm*), Markov Chain Monte Carlo -simulaatiot ja geneettiset algoritmit. Tarkastellaan tässä EM-algoritmiin perustuvan SNP-HAP-ohjelmiston käyttöä. Ohjelmisto on kuvattu osoitteessa

<http://www-gene.cimr.cam.ac.uk/clayton/software/>

ja se on asennettu CSC:n Cedar-koneeseen.

EM-algoritmi koostuu kahdesta askeleesta, joita toistetaan niin kauan, kunnes haettava tulos ei enää oleellisesti muutu:

E-askeleessa algoritmi olettaa, että eri haplotyyppien globaalit todennäköisyydet ovat kiinteitä. Tämän perusteella se laskee – kunkin yksilön genotyyppien perusteella – yksilön eri haplotyyppikonfiguraatioiden todennäköisyydet. Algoritmi skaalaa kunkin yksilön haplotyyppitodennäköisyydet niin, että ne summautuvat ykköseksi.

M-askeleessa algoritmi laskee uudet globaalit haplotyyppitodennäköisyydet edellisen askeleen yksilökohtaisten haplotyyppitodennäköisyyksien perusteella. Kunkin haplotyyppin yksilökohtaiset todennäköisyydet lasketaan yhteen ja skaalataan siten, että ne summautuvat ykköseksi. Tällä tavoin päivitettyjä globaaleja haplotyyppitodennäköisyyksiä käytetään jälleen seuraavassa E-askeleessa.

SNP-HAP-ohjelma toteuttaa EM-algoritmin siten, että se lähtee liikkeelle lyhyistä haplotyypeistä ja pidentää niitä laskennan edetessä. Hakuvaiheen aikana ohjelma pienentää etsintäavaruutta karsimalla heuristisesti epätodennäköisiä haplotyyppisiä, joten algoritmi ei tuota kaikissa tapauksissa optimaalista ratkaisua.

Tarkastellaan SNP-HAP-ohjelman käyttöä pienelle esimerkkiaineistolle, joka koostuu 100 yksilöstä. Selkeyden vuoksi SNP-markkereita on genotyypattu vain viisi kappaletta; todellisuudessa niitä voisi olla moninkertainen määrä. Nimellä `snpgeno.txt` tallennettavan aineiston 10 ensimmäistä riviä on esitelty alla. Kukin rivi alkaa yksilötunnisteella, joka voi koostua numeroista ja kirjaimista; siinä ei kuitenkaan saa olla välilyöntejä. Tämän jälkeen luetellaan yksilön genotyypit karttajärjestyksessä; kukin kahden numeron pari vastaa yhtä markkeria. Alleelit koodataan käyttäen kokonaislukuja 1 ja 2, jolloin puuttuva data ilmaistaan luvulla 0. Vaihtoehtoisesti voitaisiin käyttää nukleotidikoodausta (A, C, G, T), jolloin puuttuvaa dataa kuvaa mikä tahansa muu merkki.

1	2	2	2	2	1	1	1	1	1	2
2	1	1	1	2	1	1	1	2	1	2
3	1	2	2	2	1	1	1	2	1	2
4	1	2	2	2	1	1	2	2	1	2
5	1	2	1	1	1	1	1	2	1	2
6	1	2	2	2	1	2	2	2	1	1
7	1	2	1	2	1	1	1	1	2	2
8	2	2	2	2	1	1	1	1	2	2
9	1	1	2	2	1	2	1	2	1	2
10	2	2	2	2	1	1	1	1	1	2

SNP-HAP-ohjelma voidaan käynnistää komennolla:

```
corona% snphap snpgeno.txt out.1 out.2
```

Ensimmäinen argumentti (`snpgeno.txt`) kertoo syötetiedoston nimen ja muut kaksi argumenttia määrittävät tulostiedostojen nimet (`out.1` ja `out.2`).

Ensimmäinen tulostiedosto (tässä `out.1`) tuottaa haplotyyppien arvioidut todennäköisyydet sekä kumulatiivisen todennäköisyysjakauman:

.....	Probability	Cumulative
22121	0.172676	0.172676
12121	0.114922	0.287597
22111	0.108045	0.395642
12111	0.104856	0.500498
22112	0.086869	0.587367
12212	0.053497	0.640864
22211	0.048899	0.689763
12112	0.045975	0.735738

12211	0.040734	0.776472
11121	0.038527	0.814999
12221	0.037010	0.852009
12122	0.036657	0.888666
11122	0.035348	0.924014
22221	0.021660	0.945674
11111	0.016717	0.962390
11112	0.013447	0.975837
21111	0.010956	0.986794
22222	0.010891	0.997684
12222	0.002310	0.999995
21112	0.000005	1.000000
21121	0.000000	1.000000
22212	0.000000	1.000000
22122	0.000000	1.000000
21122	0.000000	1.000000

Toinen tulostiedosto (out. 2) näyttää kunkin yksilön todennäköisimmän haplotyyppikonfiguraation. Kutakin yksilöä vastaa tulosteessa kaksi riviä.

.....	Probability
1 1 22111	1.0000000
1 2 22112	1.0000000
2 1 11122	0.4855658
2 2 12111	0.4855658
3 1 12121	0.4558502
3 2 22112	0.4558502
4 1 12122	1.0000000
4 2 22121	1.0000000
5 1 11122	0.9992769
5 2 21111	0.9992769
6 1 12221	0.7210019
6 2 22121	0.7210019
7 1 11112	0.9997141
7 2 22112	0.9997141
8 1 22112	1.0000000
8 2 22112	1.0000000
9 1 12121	0.6400092
9 2 12212	0.6400092
10 1 22111	1.0000000
10 2 22112	1.0000000

Tulosteesta nähdään, että esimerkiksi toisen yksilön haplotyyppi oli erittäin epävarma.

SNPHAP-ajoa ja tulostusta voidaan säädellä joukolla parametreja, joista keskeisiä esitellään seuraavassa:

- th**: Valitsimella säädellään sitä, kuinka paljon vaihtoehtoisia haplotyyppi-konfiguraatioita otetaan mukaan yksilökohtaisiin tuloksiin. Jos arvoksi annetaan 1 (oletusarvo), tulostetaan ainoastaan kunkin yksilön todennäköisimmät haplotyyppit. Jos arvoksi annetaan esimerkiksi 0.6, tulostetaan sellaiset haplotyyppit, joiden todennäköisyys on vähintään 60 prosenttia todennäköisimmän haplotyyppin todennäköisyydestä.
- mi**: Valitsimen avulla voidaan luoda aineistoja, jotka on haplotyyppattu siten, että kunkin yksilön haplotyyppit on poimittu satunnaisotannalla yksilökohtaisiin haplotyyppijakaumiin perustuen. Valitsimen perään annetaan luotavien aineistojen lukumäärä. Jos tietyllä yksilöllä haplotyyppin A todennäköisyys on 80 prosenttia ja haplotyyppin B 20 prosenttia, valitaan kuhunkin tulosteeseen haplotyyppi A 80 prosentin ja haplotyyppi B 20 prosentin todennäköisyydellä.
- mm**: Valitsimen avulla toistetaan haplotyyppaus useita kertoja siten, että kohta, josta haplotyyppettä aletaan laajentaa, valitaan joka kerralla satunnaisesti. Valitsimen perään annetaan toistojen lukumäärä. Ohjelma tuottaa sen tulosteen, jonka uskottavuus on suurin.
- ss**: Tuottaa tabuloinnein erotetun tulosteen, jonka jatkokäsittely esimerkiksi taulukkolaskentaohjelmistolla on helpompaa.

Populaatiopohjainen haplotyyppaus on yleensä varsin virheeltistä, ja aineistosta saattaa löytyä useita erilaisia mutta lähes yhtä luotettavia haplotyyppi-konfiguraatioita. Useat menetelmät eivät toimi välttämättä tyydyttävästi, jos haplotyyppattava alue on niin laaja, että rekombinaatioilla on ollut suuri merkitys havaittujen haplotyyppien muodostumiseen. Haplotyyppausta helpottaa, jos henkilöt ovat väestöhistoriallisesti samankaltaisia ja markkerikartta on mahdollisimman tiheä.

8.5 Suuret SNP-aineistot ja haplotyyppiblokit

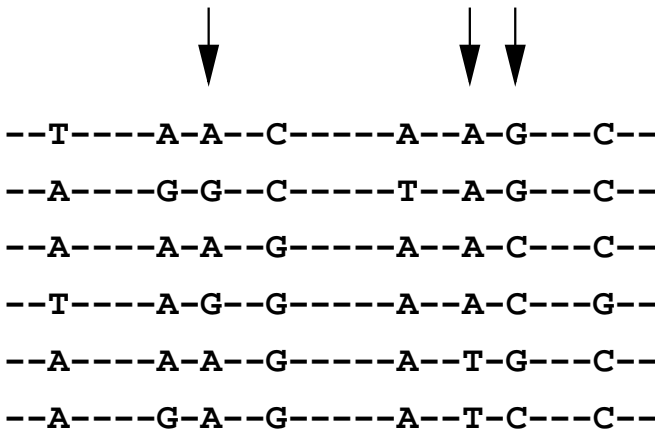
Ihmisen genomien läpikartoituksen myötä on löytynyt miljoonia SNP-markkereita, joita siruteknologian ansiosta voidaan hyödyntää geenikartoituksessa. Lähestymistapaa sovellettaen pian myös kytkentäepätasapainoon perustuvaan genomilaajuiseen geenikartoitukseen. Tällaisissa kartoitusprojekteissa arvioidaan tarvittavan joitain satoja tuhansia markkereita.

Tiheiden SNP-karttojen käytön myötä geenikartoitusprojektien tiedonhallinnalliset tarpeet kasvavat. Suurten tietomäärien analysointi ja hallinnointi edellyttävät, että genotyypit ja muu data tallennetaan hyvin organisoituihin teksti- tai relaatiotietokantoihin. Tarkoitukseen on kehitetty ja kehitetään – pitkälti kaupallisia – integroituja käyttöympäristöjä, joissa tietokantojen ja analyyttialgoritmien sujuva yhteiskäyttö helpottuu.

8.5.1 HapMap-projekti

Koska tunnettujen SNP-markkereiden määrä on suurempi kuin geenikartoituksessa realistisesti käytettävä markkereiden määrä, herää kysymys siitä, miten tarjolla olevista SNP-markkereista voidaan valita mahdollisimman kustannustehokas osajoukko. HapMap-projekti tuo valaistusta tähän kysymykseen.

Ihmisten yhteisestä väestöhistoriasta johtuen eri SNP-markkereiden alleelit eivät esiinny yksilöissä toisistaan riippumatta: genomista on löydetty *haplotyyppiblokkeja* (*haplotype blocks*), joissa yhdessä tai useammassa SNP-lokuksessa esiintyvien alleelien perusteella voidaan luotettavasti päätellä, mitä alleeleita henkilö kantaa toisissa saman blokin sisällä olevissa SNP-lokuksissa. Tällöin geenikartoituksessa käytettäviin markkerikarttoihin riittää poimia mukaan tällaiset *tunniste-SNP-markkerit* (*tag-SNPs*) kustakin blokista. Näin genotyypattavien SNP-markkerien kokonaismäärä laskee. Kuva 8.2 esittää tunniste-SNP-markkerien taustalla olevan ajatuksen: esimerkissä haplotyyppiblokki koostuu kahdeksasta markkerista, ja yksilöllä esiintyy kuutta eri haplotyyppiä. Kuitenkin kolme nuolella merkittyä markkeria riittää tunnistamaan haplotyyppit täydellisesti, joten nämä kolme markkeria voidaan valita tunniste-SNP-markkereiksi, ja niiden mukaanotto markkerikarttaan riittää periaatteessa kattamaan koko blokin. Rajoittamalla testattavien markkerien määrää saavutetaan myös se hyöty, että tehtävien tilastollisten testien määrä — ja samalla väärin positiivisten tulosten lukumäärä — vähenee.



Kuva 8.2: Esimerkki haplotyyppiblokista ja sen alueelta haetuista tunniste-SNP-markkereista. Kuvan haplotyyppiblokki sisältää kahdeksan SNP-markkeria, joista kolmen nuolella merkityn markkerin alleelit riittävät tunnistamaan, mikä kuudesta haplotyypistä on kyseessä.

Tietämys haplotyyppiblokkien synnystä ja historiasta on toistaiseksi rajallista. Aiheeseen pureutuu kansainvälinen HapMap-projekti. HapMap-projektin tavoitteena on ihmisen haplotyyppikartan luominen ja haplotyyppiblokkien sekä tunniste-SNP-markkerien etsiminen. Projektin etenemistä voi seurata verkkosivulla <http://www.hapmap.org>. Sivulla voi tutustua löytyneisiin

SNP-markkereihin ja haplotyyppiblokkeihin sekä ladata taustalla olevia genotyypaineistoja omalle koneelle.

8.5.2 Haploview-ohjelma

SNPHAP-projektin aineistojen haplotyyppiblokkeja ja aineistoissa olevaa kytKentäepätasapainoa voi tarkastella Haploview-ohjelmalla. Ohjelma soveltuu hyvin myös omien SNP-aineistojen tarkasteluun. Haploview-ohjelmalla voidaan laskea markkeriparien välisen kytKentäepätasapainon suuruus, löytää konservoituneita haplotyyppiblokkeja sekä laskea markkerikohtaisia tunnuslukuja kuten Hardy-Weinbergin tasapainomitan arvot. Ohjelman avulla voidaan tehdä myös yksinkertainen alleeliassoiaatiotestaus.

Geenikartoitusprojektissa markkerien välisen kytKentäepätasapainon laskennan tavoitteena voi olla yleiskäsityksen saaminen kartan soveltuvuudesta alttiusgeenin löytämiseen. Analyysi voidaan aloittaa laskemalla kytKentäepätasapainomittojen arvoja, jolloin saadaan arvio taustalla esiintyvän kytKentäepätasapainon (*background linkage disequilibrium*) voimakkuudesta. Mitä matalampi taustalla esiintyvä kytKentäepätasapaino on, sitä tiheämpi markkerikartta tarvitaan geenikartoituksen onnistumiseksi. Jos markkerien välillä ei havaita selkeää kytKentäepätasapainoa, on vaarana, että markkerien välinen sairauslokuks jää assosiaatioanalyysissä havaitsematta.

Haploview-ohjelma on toteutettu Java-kielellä, ja sen voi joko asentaa omalle koneelle, tai ohjelmaa voi käyttää CSC:n sovelluspalvelimelta¹. Asennusohjeet omalle koneelle löytyvät Haploview-ohjelman verkkosivulta

<http://www.broad.mit.edu/personal/jcbarret/haploview/>.

Graafinen käyttöliittymä

CSC:n laskenta-ympäristössä Haploview-ohjelman graafinen käyttöliittymä käynnistetään komennolla

```
corona% haploview
```

Jos grafiikkaa ei voida tai haluta näyttää (esimerkiksi pääteyhteyden vuoksi), annetaan komentorivillä valitsin `-n`, jolloin ohjelma ajetaan komentorivillä annetuin argumentein ja tulokset kirjoitetaan tiedostoon. Komentorivillä annettavat argumentit määrittävät, minkälainen tuloste halutaan; niihin palataan tuonnempana.

¹Uuden sovelluspalvelimen käyttöönottoon asti Ibm-sc-koneelta.

Tarkastellaan seuraavaksi graafisen käyttöliittymän toimintoja. Kun ohjelma käynnistetään, se kysyy syötetiedoston muodon (kuva 8.3, vasen). Vaihtoehdot ovat:

Load genotypes (Linkage format): Tavallinen Linkage-muotoinen sukupuutiedosto sekä lokustiedosto, joka sisältää markkerien nimet ja niiden absoluuttiset sijainnit. Sukupuutiedosto voi sisältää myös pelkkää tapaus-verrokkidataa; tällöin jokainen yksilö koodataan omalla perhenumrollaan, ja vanhemmat merkitään puuttuviksi.

Load phased haplotypes: Tiedosto, joka sisältää yksilöiden valmiiksi lasketut haplotyyppit. Lisäksi syötetään lokustiedosto.

Browse HapMap data from DCC: HapMap-projektin verkkosivulta tallennettu, haluttuun kromosomiin tai sen osaan liittyvä genotyypitiedosto.

Alla on esimerkki pienestä Linkage-muotoisesta syötetiedostosta, joka sisältää tapaus-verrokkiaineiston, jossa faasia (tietoa siitä, kumpi alleeli on peritty isältä ja kumpi äidiltä) ei ole määrätty. Tällöin jokainen yksilö koodataan omalla perhenumrollaan ja vanhemmat merkitään puuttuviksi. Puuttuva genotyyppi merkitään tavalliseen tapaan kahdella nollalla (0 0).

1	1	0	0	1	1	1	1	1	1	2	1	1	
2	1	0	0	1	1	2	2	2	2	2	2	2	
3	1	0	0	2	1	1	2	1	2	0	0	2	2
4	1	0	0	1	1	1	1	1	2	1	2	2	2
5	1	0	0	2	1	1	2	1	1	1	2	1	2
6	1	0	0	1	1	1	1	1	1	1	2	1	2

Vaihtoehtoisesti ohjelmalle voitaisiin syöttää trioaineistoja (lapsi ja vanhemmat).

Jos faasi on tunnettu, annetaan syötetiedosto alla olevan kaltaisessa muodossa. Kukin rivi alkaa perhetunnisteella, jota seuraa yksilötunniste ja kromosomiin liittyvät haplotyyppit. Jokaisesta yksilöstä muodostetaan kaksi riviä. Esimerkissä ensimmäisellä yksilöllä on toisessa vastinkromosomissa haplotyyppi 1-1-1-1 ja toisessa vastinkromosomissa haplotyyppi 1-1-2-1. Kolmannen yksilön genotyyppi kolmannessa markkerissa on tuntematon (0), ja kuudennen yksilön kolmannessa markkerissa yksilö on heterotsygootti, jonka haplotyyppi ei ollut yksikäsitteinen (h).

1	1	1	1	1	1
1	1	1	1	2	1
2	1	2	2	2	2
2	1	2	2	2	2

3	1	2	2	0	2
3	1	1	1	0	2
4	1	1	1	2	2
4	1	1	2	1	2
5	1	2	1	2	1
5	1	1	1	1	2
6	1	1	1	h	2
6	2	1	1	h	1

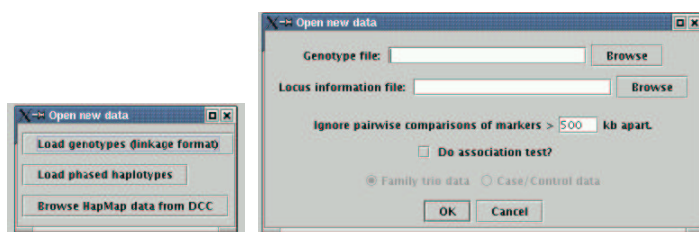
Markkeritiedosto sisältää markkerien nimet sekä absoluuttiset fysikaaliset etäisyydet. Jos fysikaaliset etäisyydet eivät ole tiedossa, voidaan LD-laskelmat toki suorittaa muuttamalla geneettiset etäisyydet fysikaaliseksi jollain yksinkertaisella skaalauksella. Tällöin etäisyydet eivät ole todellisia, mutta ne eivät vaikuta D' -mitan arvoihin. Esimerkki markkeritiedoston (4 markkerille) sisällöstä on alla:

M1	0
M2	4182
M3	11300
M4	14728

Tarkastelemme tässä Linkage-muotoisten tapaus-verrokkiaineistojen analysointia. Muista analyysivaihtoehdoista löytyy lisätietoa Haploview-ohjelman verkkosivulta. Kun Linkage-muoto on valittu, avautuu ikkuna (kuva 8.3, oikea), jossa käyttäjä ilmoittaa syötetiedostojen nimet sekä antaa tarkentavia lisämääreitä: käyttäjä voi asettaa rajan, jota etäämpänä toisestaan sijaitseville markkereille kytkentäepätasapainoa ei lasketa. Valitsemalla kohdan *Do association test* ohjelmaa pyydetään tekemään myös tapaus-verrokkiassosiaatiotestit kunkin markkerin ja sairausstatuksen välille. Jos kohta valitaan, määrittää käyttäjä myös sen, onko kyseessä trioaineisto (jolloin lapsen alleeleista tehdään sairauteen assosioituvia ja vanhempien periytymättömistä alleeleista verrokkeja) vai valmis tapaus-verrokkiaineisto (jolloin sairausstatuksia käytetään sellaisinaan).

Kun syötetiedot on luettu, saadaan valitsemalla välilehti *LD plot* graafinen tuloste ennustetuista haplotyyppiblokeista sekä markkerien välisestä kytkentäepätasapainosta (kuva 8.4). Ennustetut haplotyyppiblokot on merkitty kolmiomaisiin kehyksiin. Taulukkoon on merkitty kunkin markkeriparin välisen kytkentäepätasapainon suuruus D' -mitalla (ks. s. 79) mitattuna. Haluttua markkeriparia vastaava ruutu löydetään etsimällä kyseisistä markkereista lähtevien diagonaalien leikkauspiste. Käyttäjän näpäyttäessä halutussa ruudussa hiiren oikeanpuoleista näppäintä ilmestyy laatikko, joka sisältää kyseisen markkeriparin D' - ja LOD-arvot, korrelaatiokertoimen neliön sekä 95% luottamusvälin kytkentäepätasapainomitan D' arvoille. LOD-arvo kuvaa kytkentäepätasapainon tilastollista merkitsevyyttä. Se on usein samansuuntainen

D' -mitan arvojen kanssa, mutta esimerkiksi tilanteissa, joissa toisen markkerin informatiivisuus on hyvin heikko, voi otoksesta laskettu kytkentäepätasapaino olla hyvin suuri, vaikka tulos ei olisi tilastollisesti merkitsevä. D' - ja LOD-arvoa on käytetty perusteena ruutujen väriytykselle: esimerkiksi punaiseksi väritetyissä ruuduissa on korkea D' -arvo (punavärin voimakkuus riippuu arvosta) ja tulos on tilastollisesti merkitsevä (LOD vähintään 2).



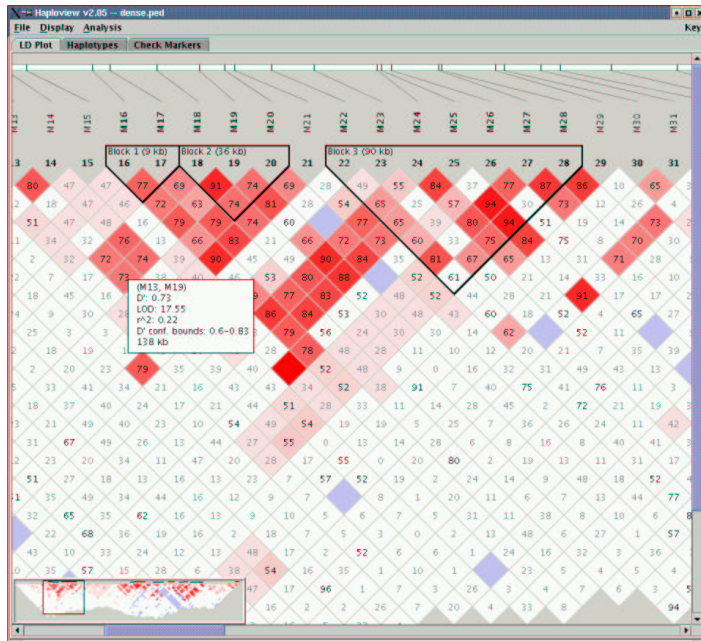
Kuva 8.3: Syötetiedostojen valinta Haploview-ohjelmassa. Vasemmalla syötteen muodon valintaikkuna. Oikealla genotyypitiedostojen määritysikkuna.

Ohjelman ennustamia haplotyyppiblokkeja on mahdollista muuttaa kehystämällä halutut markkerit hiiren avulla. Muodostettuja blokkeja voi tarkastella välilehdellä *Haplotypes* (kuva 8.5). Ennustettujen haplotyyppiblokkien päälle on merkitty pieniin kolmioihin blokin yksikäsitteisesti tunnistavat SNP-markkerit eli ns. *tag-SNP:t*. Blokkien ennustusalgoritmia voi muuttaa valinnalla *Analysis/Define blocks*. Algoritmit esitellään lyhyesti jäljempänä kommentorivikäyttöliittymän yhteydessä. Lisätietoa algoritmeista löytyy Haploview-ohjelman kotisivulta.

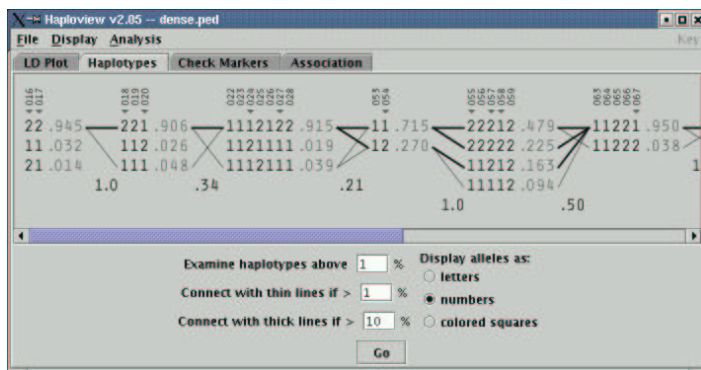
Välilehdellä *Check markers* saa yhteenvetotietoa käytetyistä markkereista. Jos assosiaatiotestit valittiin Linkage-muotoisen datan lukemisen yhteydessä, välilehti *Association* sisältää alleliassosiaatiotestien tulokset kullekin markkerille.

Komentorivikäyttöliittymä

Haploview-ohjelmaa voidaan käyttää myös komentorivikäyttöliittymän kautta. Tällöin komennon nimen perään annetaan parametri *-n*, joka kertoo, että ohjelmaa käytetään komentorivipohjaisesti. Parametrin *-n* jälkeen annetaan muut parametrit, jotka määrittävät analyysin tarkemmin. Näitä parametreja ovat:



Kuva 8.4: SNP-markkerien välisen kytentäepätasapainon laskentaa Haploview-ohjelmalla.



Kuva 8.5: Haploview-ohjelmalla etsittyjä haplotyyppiblokkeja.

- h:** tulostaa opastustekstin, joka sisältää täydellisen luettelon komentoriviparametreista.
- p tiedosto:** ilmoittaa sukupuutiedoston nimen
- i tiedosto:** ilmoittaa markkeritiedoston nimen
- d:** pyytää ohjelmaa tulostamaan markkerien välisen kytke-
täepäätasapainon suuruuden tiedostoon. Tiedosto ohjautuu
samaa hakemistoon, jossa syötetiedosto sijaitsi, ja se saa
päätteekseen merkkijonon .DPRIME.
- c:** tulostaa markkerikohtaiset tunnusluvut tiedostoon. Tiedosto
ohjautuu samaan hakemistoon, jossa syötetiedosto sijaitsi,
ja se saa päätteekseen merkkijonon .CHECK.
- o algoritmi:** tulostaa haplotyyppiblokki tiedostoon valittua algoritmia
käyttäen. Algoritmi voi saada arvokseen GAB, GAM, SPI
tai ALL. Valittaessa GAB sovelletaan Gabrielin algoritmia,
jossa blokkien rakentaminen perustuu markkerien välisen
kytkentäepäätasapainon suuruuteen. Valinta GAM ottaa käyt-
töön Wangin menetelmän[AZCJ02], jossa lasketaan kulle-
kin vierekkäisten markkerien parille haplotyyppifrekvens-
sit, ja jos niissä esiintyy korkeintaan kolmea haplotyyppiä,
katsotaan, että kohta kuuluu haplotyyppiblokkiin. Kolmas
algoritmi, SPI, on tekijöiden oma algoritmi, jossa vaaditaan,
että etsittävän blokin ensimmäinen ja viimeinen markkeri
ovat vahvassa kytkeäepäätasapainossa kaikkien välillä ole-
vien markkerien kanssa. Valinta ALL ottaa käyttöön kaikki
kolme algoritmia.
- m kb:** määrää maksimietäisyyden markkerien väliselle etäisyydel-
le. Kytkeäepäätasapainomitan D' arvo lasketaan vain niille
markkeripareille, joissa markkereiden etäisyys on tuota ra-
jaa pienempi. Raja annetaan kiloemäksinä.

Esimerkiksi Linkage-muotoinen genotyyppitiedosto haplo.ped ja markke-
ritiedosto haplo.info voitaisiin analysoida komennolla:

```
corona% haploview -n -p haplo.ped -i haplo.info -d -c -o  
GAM
```

Komento tuottaa kytkeäepäätasapainotulosteen haplo.ped.DPRIME,
markkeritulosteen haplo.ped.CHECK sekä blokkitulosteen
haplo.ped.4GAMblocks. Viimeksimainitun laskennassa on käytetty Wangin
algoritmia, joka perustuu kahden markkerin haplotyyppien lukumääriin.

9 Analyysin automatisointi Unix-ympäristössä

Tässä oppaassa esitetyillä ohjelmistoilla voidaan tyypillisesti analysoida geneettisiä aineistoja yksi kromosomi tai kandidaattialue kerrallaan. Tällöin kustakin kromosomista saadaan oma tulosteensa, ja haluttu tulos on etsittävä niistä manuaalisesti.

Se, että geenikartoitusohjelmistot on toteutettu Unix-käyttäjärjestelmässä komentorivipohjaisina ohjelmina, tarjoaa mahdollisuuden laajentaa niiden toiminnallisuutta. Perehtymällä hieman Unix-työkaluohjelmiin voidaan vähällä vaivalla kirjoittaa komentotiedostoja, jotka vähentävät käsityötä ja toistuvia työvaiheita: komentotiedoston avulla voidaan esi- ja jälkikäsitellä geenikartoitusohjelmistojen syötteitä ja tulosteita, kutsua ohjelmia toistuvasti ja suorittaa erilaisia simulaatioita.

Lähestytään aihetta kolmen havainnollisen esimerkin kautta. Oletetaan, että käyttäjällä on syötetiedostot täydellisestä genomilaajuisesta hausta, joka koostuu 22 kromosomista. Aineistojen analysointia voidaan nyt automatisoida esimerkiksi seuraavin tavoin:

- Luodaan komentotiedosto, joka syöttää aineistoja analysoitavaksi kromosomi kerrallaan. Näin vältetään 22 eri analyysivaiheen manuaaliselta toistamiselta (luku 9.1).
- Laajennetaan komentotiedostoa siten, että se etsii kustakin kromosomista kaikki kohdat, jossa NPL score ylittää käyttäjän antaman rajan ja tulostaa nuo kohdat paremmuusjärjestyksessä (luku 9.2).
- Laaditaan komentotiedosto, jonka avulla haetaan löytyneen korkeimman kytkentähuipun tilastollinen merkitsevyys, joka on korjattu moninkertaisen testauksen suhteen. Tällöin hyödynämme Merlin-ohjelman toimintoa, jolla voidaan simuloida periytymistä alleeleita pudottaen (*allele dropping simulations*, luku 9.3).

9.1 Tiedostojen automaattinen syöttäminen

Oletetaan, että analysoitavana on 22 eri kromosomiin liittyvät Linkage-muotoiset aineistot, jotka on nimetty seuraavasti:

1.link.dat	13.link.dat	17.link.dat	20.link.dat	4.link.dat	8.link.dat
1.link.pre	13.link.pre	17.link.pre	20.link.pre	4.link.pre	8.link.pre
10.link.dat	14.link.dat	18.link.dat	21.link.dat	5.link.dat	9.link.dat
10.link.pre	14.link.pre	18.link.pre	21.link.pre	5.link.pre	9.link.pre
11.link.dat	15.link.dat	19.link.dat	22.link.dat	6.link.dat	
11.link.pre	15.link.pre	19.link.pre	22.link.pre	6.link.pre	
12.link.dat	16.link.dat	2.link.dat	3.link.dat	7.link.dat	
12.link.pre	16.link.pre	2.link.pre	3.link.pre	7.link.pre	

Tiedostot voidaan syöttää CSC:n laskentaympäristössä helposti Merlin-geenikartoitusohjelmistolle laatimalla tekstieditorilla seuraavan kaltainen komentotiedosto:

```
#!/bin/bash
for C in 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 \
20 21 22 ; do
echo "Kromosomi " $C
merlin -d $C.link.dat -p $C.link.pre --np1
done
```

Komentotiedosto tallennetaan levyllä (tässä nimellä `skripti-1`), minkä jälkeen ohjelmalle annetaan suoritusoikeudet, alustetaan tarvittavat ympäristömuuttujat ja analyysi käynnistetään:

```
corona% chmod 700 skripti-1
corona% use genemap
corona% skripti-1 > res.merlin
```

Ohjelma antaa syötetiedostoja Merlin-kytkentäanalyysiohjelmistolle kromosomi kerrallaan ja kokoaa tulokset yhteen tiedostoon nimeltä `res.merlin`.

9.2 Tulosten järjestäminen

Edellisessä luvussa kirjoitimme komentotiedoston `skripti-1`, joka syöttää genomilaajuisen geenikartoitusprojektin kaikkia kromosomeja vastaavat tiedostot Merlin-ohjelmalla kromosomi kerrallaan ja luo yhteisen tulostiedoston

nimeltä `res.merlin`. Kirjoitamme seuraavaksi toisen komentotiedoston, joka jatkaa tästä: se lukee tulostiedoston sekä tulostaa kaikki ne genomien kohdat, joissa käyttäjän komentoriviparametrina antama NPL scoren raja ylittyy.

Komentotiedosto voidaan kirjoittaa esimerkiksi seuraavasti:

```
#!/bin/bash
gawk -v Raja=$1 '$1=="Kromosomi" {
    print;
}
NF==6 && substr($1,length($1)-3,1)=="." && $2>Raja {
    print;
}
}'
```

Komentotiedosto hyödyntää Unix-työkaluohjelmaa `gawk`, jota voidaan käyttää tekstimuotoisten tiedostojen käsittelyyn.

Annetaan tallennetulle tiedostolle suoritusoikeudet ja ajetaan se käyttäen NPL scoren raja-arvoa 1.5:

```
corona% chmod 700 skripti-2
corona% skripti-2 1.5 < res.merlin
```

Komentotiedosto tuottaa seuraavan tulosteen:

Kromosomi	1						
Kromosomi	2						
Kromosomi	3	133.886	1.58	0.06	0.765	0.70	0.04
Kromosomi	4						
Kromosomi	5	200.829	1.74	0.04	0.765	0.75	0.03
Kromosomi	6						
Kromosomi	7						
Kromosomi	8	22.314	1.89	0.03	0.459	0.42	0.08
		33.472	1.93	0.03	0.765	0.69	0.04
		44.629	1.82	0.03	0.765	0.77	0.03
		100.415	1.57	0.06	0.765	0.73	0.03
Kromosomi	9						
Kromosomi	10						
Kromosomi	11	122.729	2.13	0.02	0.765	0.88	0.02

	133.886	1.89	0.03	0.765	0.78	0.03
	145.043	1.53	0.06	0.462	0.35	0.10
Kromosomi	12					
	100.415	1.88	0.03	0.617	0.55	0.06
Kromosomi	13					
	55.786	2.23	0.013	0.765	0.88	0.02
	66.943	3.70	0.00011	0.765	1.66	0.003
	78.100	3.83	0.00006	0.765	1.71	0.003
	89.257	2.10	0.02	0.765	1.11	0.012
	111.572	1.65	0.05	0.765	0.86	0.02
Kromosomi	14					
	66.943	1.92	0.03	0.765	0.88	0.02
Kromosomi	15					
	33.472	2.42	0.008	0.765	1.29	0.007
	44.629	3.69	0.00011	0.765	1.69	0.003
	55.786	3.81	0.00007	0.765	1.72	0.002
	66.943	3.61	0.0002	0.765	1.64	0.003
	78.100	2.79	0.003	0.765	1.04	0.014
	89.257	2.24	0.012	0.598	0.54	0.06
	100.415	2.13	0.02	0.765	0.74	0.03
Kromosomi	16					
Kromosomi	17					
	78.100	2.20	0.014	0.612	0.66	0.04
Kromosomi	18					
	22.314	1.76	0.04	0.317	0.25	0.14
	66.943	1.57	0.06	0.590	0.39	0.09
Kromosomi	19					
Kromosomi	20					
Kromosomi	21					
Kromosomi	22					

9.3 Empiirisen p-arvon määrittäminen kytkentätulokselle

Seuraavaksi kirjoitamme komentotiedoston, jonka avulla saamme laskettua nollahypoteesijakauman genominlaajuudessa geenikartoituksessa havaittavalle korkeimmalle NPL scorelle. Vertaamalla todellisesta aineistosta laskettua korkeimman huipun NPL scorea tähän jakaumaan saamme empiirisen p-arvon, joka kuvaa tuloksen tilastollista merkitsevyyttä. Tällä tavoin laskettu arvio on korjattu moninkertaisen testauksen suhteen, ja se ottaa huomioon aineiston markkerikartan rajallisen informatiivisuuden toisin kuin Merlinin NPL scoren perusteella laskema, normaalijakaumaoletukseen perustuva p-arvo.

Genominlaajuisen p-arvon määrittäminen tällä tavoin simuloimalla on laskennallisesti erittäin raskasta. Kytkeä tuloksia julkaisevat lehdet kuitenkin arvostavat simulaatiotuloksia, sillä niiden avulla lukija saa varsin hyvän käsityksen siitä, kuinka merkitseviä tulokset ovat.

Edellä kuvattu tilastollisen merkitsevyyden arviointi voidaan CSC:n ympäristössä tehdä esimerkkinä 22 kromosomin syötetiedoille vaikkapa seuraavanlaisen lyhyehkön komentotiedoston avulla. Komentotiedosto hyödyntää toistuvasti Merlin-ohjelman alleelienpudotussimulaatiotoimintoa.

```
#!/bin/bash
echo "" > res.merlin
R=0
while [ $R -lt $1 ]; do
  let R=R+1
  echo "Iteraatio " $R >> res.merlin
  for C in 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 \
    20 21 22 ; do
    echo "Kromosomi " $C >> res.merlin
    merlin -d $C.link.dat -p $C.link.pre --npl --simulate \
      -r $R >> res.merlin
  done
done

gawk '$1=="Iteraatio" {
  if ($2>1)
    print korkein;
  korkein = -999;
}'

NF==6 && substr($1,length($1)-3,1)=="." {
  if ($2>korkein)
    korkein = $2;
} END {
  print korkein;
}' < res.merlin | sort -rn |
gawk '{print ++n " " $1}'
```

Jos ajamme simulaation esimerkiksi kymmenellä toistolla, tapahtuu se antamalla komennot:

```
corona% chmod 700 skripti-3  
corona% use genemap  
corona% skripti-3 10
```

Jos komento `use genemap` on annettu aiemmin saman istunnon aikana, ei sitä tarvitse antaa uudestaan. Laskenta kestää kymmeniä minuutteja ja tuottaa seuraavan kaltaisen tulosteen:

1	3.83
2	2.88
3	2.70
4	2.56
5	2.47
6	2.44
7	2.37
8	2.27
9	1.78
10	1.54

Tämän pienen simulaation tuloksena havaitsemme, että todellisesta aineistosta saatu NPL score 3.83 saavutettiin tai ylitettiin yhdellä toistolla kymmenestä. Tästä saamme karkean arvion empiiriselle p-arvolle: $p = 1/10 = 0.1$. Käytännössä toistoja on syytä olla moninkertainen määrä, vähintään sata, ja jotta pienikin p-arvo voitaisiin arvioida tarkasti, mielellään tuhat.

Koska laskenta kestää kauan, on simulointityö yleensä syytä ajaa eräajona. Eräajotyö asetetaan jonoon, josta se ohjautuu automaattisesti suorittimelle, jolla on vapaata laskentakapasiteettia. Tällöin laskentaresurssien käyttö on tehokkaampaa. Eräajotöistä löytyy lisätietoa CSC:n kotisivuilta löytyvistä koneoppaista.

CSC:n asiantuntijat neuvovat ja avustavat mielellään koneiden ja eräajojärjestelmän käytössä.

Liitteet

A Kirjoittajien yhteystiedot

VESA OLLIKAINEN

CSC – Tieteellinen laskenta Oy
Tekniikantie 15 a D (PL 405)
02101 Espoo

puh. (09) 457 2282
vesa.ollikainen@csc.fi

PEKKA UIMARI

Jurilab Ltd.
Microkatu 1
70210 Kuopio

puh. (017) 467 8027
pekka.uimari@jurilab.com

B Tietoa CSC:stä

CSC on opetusministeriön hallinnoima tieteen tietotekniikan keskus. CSC tarjoaa korkeakouluille ja tutkimuslaitoksille tietoteknistä tukea ja resursseja: mallinnus- ja laskentapalveluja sekä informaatiopalveluja. Tutkijat voivat käyttää CSC:ssä sijaitsevaa Suomen laajinta tieteellisten ohjelmistojen ja tieteen tietokantojen valikoimaa sekä Suomen tehokkainta superlaskentaympäristöä Funet-tietoliikenneyhteyksien kautta.

Lisätietoja CSC:stä saa www-osoitteesta <http://www.csc.fi/>.

Kirjallisuutta

- [ACCC02] G. Abecasis, S. Cherny, W. Cookson ja L. Cardon, Merlin – rapid analysis of dense genetic maps using sparse flow gene trees, *Nat. Genet.*, 2002, Volume 30, sivut 97–101.
- [AZCJ02] N. Wang J.M. Akey, K. Zhang, R. Chakraborty ja L. Jin, Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation, *Am. J. Hum. Genet.*, 2002, Volume 71, sivut 1227–1234.
- [DKTC00] F. Dudbridge, B.P. Koeleman, J.A. Todd ja D.G. Clayton, Unbiased application of the transmission/disequilibrium test to multilocus haplotypes, *Am. J. Hum. Genet.*, 2000, Volume 66, sivut 2009–2012.
- [ES71] R.C. Elston ja J. Stewart, A general model for the genetic analysis of pedigree data, *Hum. Hered.*, 1971, Volume 21, sivut 523–542.
- [ESW01] E. Sobel E, H. Sengul ja D.E. Weeks, Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees, *Hum. Hered.*, 2001, Volume 52, sivut 121–131.
- [JIS93] R.W. Cottingham Jr., R.M. Idury ja A.A. Schäffer, Faster sequential genetic linkage computations, *Am. J. Hum. Genet.*, 1993, Volume 53, sivut 252–263.
- [KC97] A. Kong ja N.J. Cox, Allele-sharing models: Lod score and accurate linkage tests, *Am. J. Hum. Genet.*, 1997, Volume 61, sivut 1179–1188.
- [KDRDL96] L. Kruglyak, M.J. Daly, M.P. Reeve-Daly ja E.S. Lander, Parametric and nonparametric linkage analysis: a unified multipoint approach, *Am. J. Hum. Genet.*, 1996, Volume 58, sivut 1347–1363.
- [LG87] E.S. Lander ja P. Green, Construction of multilocus genetics maps in humans, *Proc. Natn. Acad. Sci. U.S.A.*, 1987, Volume 84, sivut 2363–2367.
- [LL84] G.M. Lathrop ja J.M. Lalouel, Easy calculations of lod scores and genetic risks on small computers, *Am. J. Hum. Genet.*, 1984, Volume 36, sivut 460–465.
- [LLJO84] G.M. Lathrop, J.M. Lalouel, C. Julier ja J. Ott, Strategies for multilocus linkage analysis in humans, *Proc. Natn. Acad. Sci. U.S.A.*, 1984, Volume 81, sivut 3443–3446.
- [LLW86] G.M. Lathrop, J.M. Lalouel ja R.L. White, Construction of human linkage maps: likelihood calculations for multilocus linkage analysis, *Gen. Epi.*, 1986, Volume 3, sivut 39–52.

- [MAS⁺99] N. Mukhopadhyay, L. Almasy, M. Schroeder, W.P. Mulvihill ja D.E. Weeks, Mega2, a data-handling program for facilitating genetic linkage and association analyses, *Am. J. Hum. Genet.*, 1999, Volume 65, sivu A436.
- [Mor55] N.E. Morton, Sequential tests for the detection of linkage, *Am. J. Hum. Genet.*, 1955, Volume 3, sivut 277–318.
- [OW98] J.R. O’Connell ja D.E. Weeks, Pedcheck: A program for identifying genotype incompatibilities in linkage analysis, *Am. J. Hum. Genet.*, 1998, Volume 63, sivut 259–266.
- [PCS03] S. Purcell, S.S. Cherny ja P.C. Sham, Genetic power calculator: design of linkage and association genetic mapping studies of complex traits, *Bioinformatics*, 2003, Volume 19, sivut 149–150.
- [Pen35] L.S. Penrose, The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage, *Annals of Eugenics*, 1935, Volume 6, sivut 133–138.
- [SGSJ94] A.A. Schäffer, S.K. Gupta, K. Shriram ja R.W. Cottingham Jr., Avoiding recomputation in linkage analysis, *Hum. Hered.*, 1994, Volume 44, sivut 225–237.
- [SL96] E. Sobel ja K. Lange, Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics, *Am. J. Hum. Genet.*, 1996, Volume 58, sivut 1323–1337.
- [SME93] R.S. Spielman, R.E. McGinnis ja W.J. Ewens, Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM), *Am. J. Hum. Genet.*, 1993, Volume 52, sivut 506–516.
- [SPL02] E. Sobel, J.C. Papp ja K. Lange, Detection and integration of genotyping errors in statistical genetics, *Am. J. Hum. Genet.*, 2002, Volume 70, sivut 496–508.
- [TO94] J. Terwilliger ja J. Ott. *Handbook of Human Genetic Linkage*. Johns Hopkins University Press, 1994.
- [WH94] A. Whittemore ja J. Halpern, A class of tests for linkage using affected pedigree members, *Biometrics*, 1994, Volume 50, sivut 109–117.
- [WOL90] D.E. Weeks, J. Ott ja G.M. Lathrop, Slink: a general simulation program for linkage analysis, *Am. J. Hum. Genet.*, 1990, Volume 47, sivu A204.

Hakemisto

A

alleeli, 11, 14
 alleelienpudotussimulaatio, 34, 111, 115
 Merlin-ohjelmassa, 73
 alleelifrekvenssit, 11, 74
 estimointi, 24
 alleeliheterogenia, 11
 alttiusluokka, 26
 ASP, affected sib pair method, 54, 58, 69
 keskiarvotesti, 56
 nollahypoteesi, 57
 ASP-testisuure, 56
 assosiaatioanalyysi, 10, 15, 79, 81, 97

B

Bonferroni-korjaus, 33

C

crossing-over, 13
 CSC – Tieteellinen laskenta Oy, 119

D

dominanssivarianssi, 69
 Downfreq, 17

E

ei-parametrinen kytkentäanalyysi, 54, 58
 eksploratiivinen analyysi, 31
 EM-algoritmi, 100
 eräajo, 116
 ERPA, extended relative pair analysis, 56
 esitysmuodot, 17

F

fenotyyppi, 12, 15

G

geenikartoitus, 8, 10
 geneettinen ajautuminen, 79
 geneettinen algoritmi, 100

geneettinen etäisyys, 14
 geneettinen heterogenia, 11
 geneettinen kytkentä, 14
 poissulku, 32
 Genehunter, 58
 ASP-analyysi, 69
 istunnon kulku, 68
 käskyt, 60
 kytkentäanalyysi, 58
 NPL, 58
 NPL-analyysi, 60
 parametrinen analyysi, 41
 periytymisvektori, 59
 TDT, 86
 TDT haplotyypeille, 93
 tiedostot, 60
 Genehunter Plus, 72
 genomi, 11
 genomilaajuinen geenikartoitus, 15, 31, 111
 genotyyppi, 11, 15
 Ghostview, 77
 GPC, Genetic Power Calculator, 89

H

Haplo-assoc, 97
 haplotyyppi, 58, 64, 93
 assosiaatioanalyysi, 97
 perhepohjainen, 95, 96
 väestöpohjainen, 100
 haplotyyppiblokki, 103
 haplotyyppikonfiguraatio, 37
 Haploview, 105
 HapMap-projekti, 104
 harvat puut, 70
 heterogenia, 11
 heterotsygotia-aste, 12
 hypoteesi, 36
 testaus, 39

I

IBD, 54, 55
 IBD-status, 55
 IBS, 54, 55

identity-by-descent, 54
identity-by-state, 54
informatiivisuus, 12, 58, 59, 62
isovanhempaisalkuperä, 58

K

kaksiarvoinen muuttuja, 8
karttafunktio
 Haldane, 14, 67
 Kosambi, 14, 67
karttatiedosto, 25, 28
Kongin ja Coxin menetelmä, 72
kontingenssitaulu, 81
kromosomi, 11
 sairauteen assosioitunut, 97
 status, 97
 verrokki, 97
kromosomipari, 11
kromosomisegmentti, 13, 15
kytkentäanalyysi, 10, 15, 35
 ei-parametrinen, 54, 58
 kaksipiste, 41, 45, 63
 monipiste, 41, 52
 parametrinen, 35, 58
kytkentäepätasapaino, 15, 79, 82, 93
 mitat, 79
kytkeytyneet lokukset, 14

L

lääketieteelliset sovellukset, 8
laskennallinen vaativuus, 40
laskenta-algoritmi, 40
 Elston-Stewart, 40
 heuristinen, 41
 Lander-Green, 40, 58, 70
Linkage, 40, 43, 51, 64
 analyysivaiheet, 43
 Ilink, 44, 47, 51
 Lcp, 43, 45
 Linkmap, 47, 52
 Lodscore, 47
 Lrp, 44, 50
 Mlink, 44, 47, 50
 ohjelmat, 43
 Unknown, 44
Linkage-muoto, 17, 86, 106
 parametritiedosto, 17, 22, 26, 86
 sukupuutiedosto, 17, 18
Linkagepar, 17, 23, 27
location score, 53
LOD score, 62
lokusheterogenia, 11

M

Makedata, 17
Makeped, 17, 43, 44

markkeri, 11, 14
 virheet, 20
markkerikartta, 15, 64
MCMC-simulaatiot, 41, 73, 100
McNemarin testi, 85
Mega2, 17, 25, 27, 73
meioosi, 13
Mendel-virhe, 71
merkitsevyytaso, 33
Merlin, 70
 käskyt, 71, 72
 tiedostot, 70
Merlin2assoc, 97
meta-analyysi, 15
mikrosatelliittimarkkeri, 11
moninkertaisen testauksen ongelma,
 32, 33
monitekijäinen sairaus, 8, 10, 55
monogeeninen sairaus, 10
morgan, 13

N

nollahypoteesi, 31, 81
NPL score, 62

P

p-arvo, 32, 88
 empiirinen, 34, 88, 114, 116
 korjattu, 88
 nominaalinen, 34
parametrien estimointi, 39
parametrinen kytkentäanalyysi, 35
pedcheck, 20
Pedstats, 71
penetranssivektori, 10
periytymisen epätasapainotesti, 82, 93
periytymismalli, 35, 54
 resessiivinen, 54, 82
periytymisvektori, 58, 59
permutaatiotestaus, 34, 88, 99
perustajajyksilö, 12
PIC, 12
populaatiofrekvenssi, 10
Prelink, 17
prevalenssi, 11

R

R, 77
rekombinaatio, 13
rekombinaatiofraktio, 14
relatiivinen riski, 11

S

sairausmalli, 10
 dominantti, 11
 resessiivinen, 11

S_{all}, 60
segregaatioanalyysi, 35
senttimorgan, 13
silmutka, sukupuussa, 41, 42
simuloitu jäähtytys, 73
simuloitu nollahypoteesijakauma, 34
Simwalk, 17, 41, 73
SNP, 8, 11, 15, 86
SNPHAP, 100
 komennot, 101, 103
sovelluspalvelin, 15
S_{pairs}, 60
sukupu, 12
sukupuutiedoston tarkistaminen, 20
suuret sukupuut, 73
suurimman uskottavuuden menetelmä,
 35

T

tapaus-verrokkitiedosto, 97
TDT, 58, 82
 alleelikohtainen, 87
 approksimaatio, 85
 binomijakaumaan perustuva, 84
 Genehunter-ohjelmalla, 86
 permutaatiotestaus, 88
tekijäinvaihto, 13
tiedostot
 sukupuutiedosto, 20
tilastollinen merkitsevyys, 32
tilastollinen testi, 31
tilastollinen voima, 32
 assosiaatioanalyysin, 89
tunniste-SNP-markkeri, 104

U

Unix, 111
uskottavuusfunktio, 37

V

väestöhistoria, 79
varianssikomponentti, 58
voimalaskelma, 33



2004

<http://www.csc.fi/oppaat/geenikartoitus>
ISBN 952-5520-00-5