NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

RESEARCH

Tero Hiekkalinna

On the superior power of likelihood-based linkage disequilibrium mapping in large multiplex families compared to population based case-control designs

88

RESEARCH 88/2012

Tero Hiekkalinna

On the superior power of likelihood-based linkage disequilibrium mapping in large multiplex families compared to population based case-control designs

ACADEMIC DISSERTATION

To be presented with the permission of the Faculty of Medicine,
University of Helsinki, for public examination in the lecture hall 1,
Haartman Institute, on October 5th, 2012, at 12 noon

Department of Medical Genetics, Haartman Institute, University of Helsinki,
Helsinki, Finland
and
Institute for Molecular Medicine Finland FIMM, Helsinki, Finland
and
Public Health Genomics Unit, National Institute for Health and Welfare,
Helsinki, Finland

Helsinki 2012

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

Cover graphic: The great Andromeda galaxy (M31) and satellite galaxies M32 and M110. Photograph by Jussi Kantola. Total exposure time eight hours. Printed with permission.

ISBN 978-952-245-712-7 (print)
ISSN 1798-0054 (print)
ISBN 978-952-245-713-4 (pdf)
ISSN 1798-0062 (pdf)
URN:ISBN:978-952-245-713-4
http://urn.fi/URN:ISBN: 978-952-245-713-4

Juvenes print – Tampereen Yliopistopaino Oy
Tampere, Finland 2012

YMPÄRISTÖMERKKI
MILJÖMÄRKT
441 729
Painotuote

Supervised by

Academician of Science, Professor Leena Peltonen-Palotie †
Wellcome Trust Sanger Institute, Cambridge, United Kingdom
National Institute for Health and Welfare, Unit of Public Health Genomics and
University of Helsinki, Institute for Molecular Medicine Helsinki, Finland

Associate Professor Joseph D Terwilliger
University of Helsinki, Institute for Molecular Medicine Helsinki, Finland,
Columbia University, Department of Psychiatry, Department of Genetics and
Development, Columbia Genome Center and
New York State Psychiatric Institute Division of Medical Genetics,
New York, USA

Research Professor Markus Perola
National Institute for Health and Welfare, Unit of Public Health Genomics and
University of Helsinki, Institute for Molecular Medicine
Helsinki, Finland

Reviewed by

Professor Daniel E. Weeks
Departments of Human Genetics and Biostatistics
Graduate School of Public Health
University of Pittsburgh
Pittsburgh, Pennsylvania, USA

Professor Hannes Lohi
Department of Veterinary Biosciences, Faculty of Veterinary Medicine,
Research Programs Unit, Molecular Medicine, Faculty of Medicine,
University of Helsinki and
Folkhälsan Research Center, Helsinki, Finland

Opponent

John Blangero, Ph.D., Director
AT&T Genomics Computing Center
Texas Biomedical Research Institute, Department of Genetics,
San Antonio, Texas, USA

† *Academican, Professor Leena Peltonen-Palotie deceased in March 2010*

"Joulu Joulu, maha kurnuttaa!"
-Oiva 5-v.

”Isi, miksi viet kaikki loitsukirjat työpaikalle?”
-Ukko 6-v, kun vein Mathematical Methods in Physical Sciences-kirjan työpaikalle

”Ootsä oikeesti niin huono kirjoittaja, että sä kirjoitit siihen väitöskirjaan vääriä kirjaimia?”
-Ukko 9-v.

To my family

Abstract

Tero Hiekkalinna. On the superior power of likelihood-based linkage disequilibrium mapping in large multiplex families compared to population based case-control designs. National Institute for Health and Welfare (THL). Research 88/2012. 173 pages. Helsinki, Finland 2012.
ISBN 978-952-245-712-7 (printed); ISBN 978-952-245-713-4 (pdf)

In this thesis, we developed software for automated genome-wide linkage and linkage disequilibrium analysis based on common gene mapping methods for qualitative and quantitative phenotypes. We further developed likelihood-based software for joint linkage and/or linkage disequilibrium (LD) analysis in general pedigrees based on a novel variation of the classical lod score approach, the so-called pseudomarker method, and evaluated its statistical properties as compared with the existing family-based association methods. This was done using real-life migraine and schizophrenia pedigree structures from Finland. In addition, we compared various study designs for association analysis and investigated statistical properties of the likelihood ratio test for conditional analysis of LD given linkage.

First, we automated the laborious process of running a variety of genome-wide linkage and linkage disequilibrium analysis software packages, including ANALYZE, MERLIN, GENEHUNTER, and SOLAR. With this software tool, data file format conversion, and running of the analyses are completely automated. This tool has been applied to many large genome-wide mapping studies.

Second, we developed user-friendly PSEUDOMARKER software, which performs likelihood-based linkage and/or linkage disequilibrium analysis in general pedigrees. This software allows for joint analysis of heterogeneous relationship structures, such as singletons (i.e. cases and controls), triads, sibships, and large multigenerational pedigrees. The performance of this software was evaluated in comparison to the existing repertoire of other family-based association methods.

Third, we performed an extensive simulation study to investigate the statistical properties (i.e. type-I error and power) of PSEUDOMARKER and other commonly used family-based association methods. Our results demonstrate that many widely-used methods are not valid for testing LD in the presence of linkage, and likelihood-based methods which can properly account for missing data and individual relationships in pedigrees, such as PSEUDOMARKER, outperform the other approaches over a wide variety of etiological models. We also demonstrated that association mapping in families is far more powerful than in population-based samples.

Fourth, we investigated the statistical properties of the likelihood ratio test for association conditional on linkage when inaccurate parametric models were used. Our results showed that while under most situations they perform appropriately despite the parametric model being improperly specified, under certain conditions, when there is complete linkage between disease and marker loci, overly-deterministic dominant analysis models can lead to false inferences of LD in the presence of linkage when the true etiological model is recessive in character.

In this study, we have developed powerful and easy to use tools for analysis of linkage and LD in general pedigrees and unrelated individuals jointly, and have demonstrated the superiority of such methods in the general case. Our results provide important information for the human genetics community about optimal ways to collect and analyze data.

Keywords: likelihood-based methods, genome-wide linkage analysis, linkage disequilibrium, family-based association, genome-wide association studies, bias, type-I error, power, computer software, study design

Tiivistelmä

Tero Hiekkalinna. On the superior power of likelihood-based linkage disequilibrium mapping in large multiplex families compared to population based case-control designs [Uskottavuus-pohjaisen kytkentäepätasapainokartoituksen suuresta voimasta perherakenteissa verrattuna populaatiopohjaiseen tapaus-verrokki asetelmaan]. Terveyden ja hyvinvoinnin laitos (THL). Tutkimus 88/2012. 173 sivua. Helsinki, Finland 2012.
ISBN 978-952-245-712-7 (painettu); ISBN 978-952-245-713-4 (pdf)

Tässä väitöskirjatutkimuksessa kehitettiin menetelmä koko perimänlaajuiseen automaattiseen kytkentä- ja kytkentäepätasapainoanalyysiin kvalitatiivisilla ja kvantitatiivisilla ominaisuuksilla käyttäen yleisiä geenikartoitusmenetelmiä. Lisäksi tässä työssä kehitettiin uskottavuuteen perustuva tietokoneohjelma yhdistettyyn kytkentä- ja kytkentäepätasapainoanalyysiin perheaineistossa, joka perustuu muunnelmaan klassisesta lod score kytkentäanalyysimenetelmästä. Tämän ns. pseudomarker-menetelmän tilastotieteellisiä ominaisuuksia verrattiin simulaatiotutkimuksella muihin yleisesti käytettyihin perhepohjaisiin assosiaatiomenetelmiin käyttäen suomalaisia skitsofrenia ja migreeni perherakenteita. Lisäksi vertailtiin erilaisia tutkimusasetelmia assosiaatioanalyysissä ja tutkittiin ehdollisen uskottavuusosamäärätestin tilastotieteellisiä ominaisuuksia, kun testataan assosiaatiota kytkennän vallitessa

Ensimmäisessä osatyössä automatisoitiin koko perimänlaajuinen kytkentä- ja kytkentäepätasapainoanalyysi käyttäen ANALYZE, MERLIN, GENEHUNTER ja SOLAR geenikartoitusohjelmia. Tämän tietokoneohjelman avulla aineiston käsittely geenikartoitusmenetelmän vaatimaan muotoon ja analyysi on täysin automatisoitu. Lisäksi se on mahdollistanut useita automatisoituja laajoja koko perimänlaajuisia geenikartoitusanalyysejä.

Toisessa osatyössä kehitettiin käyttäjäystävällinen PSEUDOMARKER-tietokoneohjelma, joka on uskottavuus-pohjainen yhdistetty kytkentä- ja kytkentäepätasapainoanalyysi perheissä. Tämä PSEUDOMARKER-ohjelma mahdollistaa erilaisten aineistojen yhdistämisen yhteen ja samaan analyysiin, kuten tapauksia, verrokkeja, trioja, ydinperheitä ja suuria usean sukupolven perheitä. Lisäksi pseudomarker-menetelmää verrattiin muihin perhepohjaisiin assosiaatiomenetelmiin simulaatiotutkimuksella.

Kolmannessa osatyössä vertailtiin laajasti simulaatiotutkimuksella PSEUDOMARKER-ohjelman ja muiden yleisesti käytettyjen perhepohjaisten assosiaatio-ohjelmien tyypin I virhettä ja tilastotieteellistä voimaa. Tässä tutkimuksessa havaittiin, että jotkin menetelmät eivät sovellu assosiaatiotestaukseen kytkennän vallitessa ja uskottavuus-pohjaiset assosiaatiomenetelmät, jotka pystyvät ottamaan huomioon puuttuvaa genotyyppi- tai fenotyyppidataa ja pystyvät analysoimaan suuria perherakenteita, kuten PSEUDOMARKER-ohjelma, ovat voimakkaimpia erilaisten tautimallien vallitessa. Lisäksi tulostemme perusteella assosiaatioanalyysi perheaineistossa on voimakkaampaa kuin tapaus-verrokki aineistossa.

Neljännessä osatyössä tutkittiin uskottavuusosamäärätestin tilastotieteellisiä ominaisuuksia, kun testataan assosiaatiota kytkennän vallitessa, käyttäen epätarkkoja parametrista malleja. Tutkimuksessa havaitsimme, että suurimmassa osassa tilanteita epätarkat mallit toimivat moitteettomasti, mutta tietyissä tilanteissa, kun täysi kytkentä vallitsee tautilokuksen ja geenimerkin välillä, niin deterministisen vallitsevan analyysimallin käyttö ehdollisessa uskottavuusosamäärätestissä (olettaen kytkentä) saattaa johtaa väärää johtopäätökseen assosiaatiosta, kun todellinen malli on peittyvästi periytyvä.

Tässä väitöskirjatyössä olemme kehittäneet voimakkaan ja helppokäyttöisen ohjelman kytkentä- ja kytkentäepätasapainoanalyysin perheissä ja tapaus-verrokki aineistossa. Lisäksi olemme osoittaneet vastaavien menetelmien ylivoimaisuuden. Tutkimuksemme tulokset antavat tärkeää tietoa ihmisgenetiikan tutkijoille parhaista tavoista kerätä ja analysoida aineistoa.

Avainsanat: uskottavuus-pohjaiset menetelmät, koko perimän laajuinen kytkentäanalyysi, kytkentäepätasapaino, perhe-pohjainen assosiaatio, koko perimän laajuinen assosiaatiotutkimus, vinouma, tyypin I virhe, voima, tietokoneohjelma, tutkimusasetelma

Contents

Abbreviations

DNA	Deoxyribonucleic acid
GWAS	Genome-wide association study
HWE	Hardy-Weinberg equilibrium
HRR	Haplotype relative risk
HHRR	Haplotype-based haplotype relative risk
IBD	Identity-by-descent
i.i.d.	independent and identically distributed
LD	Linkage disequilibrium
LR	Likelihood ratio
LRT	Likelihood ratio test
MLE	Maximum likelihood estimate
RR	Relative risk
SNP	Single nucleotide polymorphism
TDT	Transmission/Disequilibrium test

Symbols

ϕ	Disease prevalence
p_D	Disease allele frequency
δ	Gametic linkage disequilibrium coefficient
D’	Lewontin’s D’ (D-prime)
Λ	Pseudomarker linkage test statistic
Ψ	Pseudomarker LD given linkage test statistic
Υ	Pseudomarker LD given no linkage test statistic
ζ	Pseudomarker linkage given LD test statistic
Ξ	Pseudomarker joint linkage and LD test statistic
θ	Recombination fraction
r^2	Squared correlation coefficient
α	Significance level

List of original publications

This thesis is based on the following original articles referred to in the text by their Roman numerals:

I **Hiekkalinna T**, Terwilliger JD, Sammalisto S, Peltonen L, Perola M: AUTOGSCAN: Powerful Tools for Automated Genome-wide Linkage and Linkage Disequilibrium Analysis. Twin Res Hum Genet. 2005 Feb;8(1):16-21.

II **Hiekkalinna T**, Schäffer AA, Lambert BW, Norrgrann P, Göring HHH, Terwilliger JD: PSEUDOMARKER: A Powerful Program for Joint Linkage and/or Linkage Disequilibrium Analysis on Mixtures of Singletons and Related Individuals. Hum Hered 2011;71:256-266.

III **Hiekkalinna T**, Göring HHH, Lambert BW, Weiss KM, Norrgrann P, Schäffer AA, Terwilliger JD: On the statistical properties of family-based association tests in datasets containing both pedigrees and unrelated case-control samples. European Journal of Human Genetics (2012) 20, 217-223.

IV **Hiekkalinna T**, Göring HHH, Terwilliger JD: On the Validity of the Likelihood Ratio Test and Consistency of Resulting Parameter Estimates in Joint Linkage and Linkage Disequilibrium Analysis under Improperly Specified Parametric Models. Annals of Human Genetics (2012) 76, 63-73.

These articles are reproduced with the kind permission of their copyright holders.

Also, unpublished data from one other study is presented.

1 Introduction

Over the last few decades human geneticists have characterized thousands of genetic variants which influence human traits, ranging from severe Mendelian diseases to normal quantitative variation in traits such as the serum levels of different proteins (Comuzzie, Hixson et al. 1997). Classical genetic studies investigated highly penetrant Mendelian traits in large pedigrees, often ascertained from isolated populations, in which the inheritance of traits across generations was studied. However, in the last few years, some focus has shifted from Mendelian traits to complex traits, and large pedigrees studies have given way to cross sectional studies. A complex trait, by definition, has unknown etiology, but is often presumed to be influenced by a large number of genetic variants with small effects and a large number of environmental factors, often related to nutrition and lifestyle. Common complex traits, such as diabetes, cardiovascular disease, and obesity have a major impact on public health, and therefore there is great interest in unravelling their architecture. Nevertheless, cross sectional studies, even with thousands of individuals and hundreds of thousands of genotyped markers have proven to be relatively uninformative and even such enormous studies are therefore often euphemistically referred to as “underpowered” implying the solution lies in even larger samples (Spencer, Su et al. 2009).

Technological advances in molecular genetics have made it possible to generate large amounts of genotype and DNA sequence data, and parallel developments in microprocessor technology have completely revolutionized the possibilities for statistical computing. Statistical theory in genetics, as developed in the early 20th century, is implemented today in computer programs developed by scientists all over the world. Popular program packages, which are used for analyzing population-based samples can mostly analyze cross sectional (i.i.d.) data only. Furthermore, as the cost of whole genome sequencing becomes affordable, and the error-rates become tolerably low, studies are beginning to sequence related individuals in families, a trend that is likely to continue in populations with good genealogical records like Finland. This will require additional analytical expertise from young researchers, as many students and postdocs today have only been trained to analyze unrelated samples, rather than individuals in families. Eventually, all individuals in the population are likely to be sequenced for screening of Mendelian risk variants, at which time the family information from our population registers can be used to trace the segregation of shared genomic segments through the population. This can be linked with available phenotypic data in Finland, where numerous health-related databases of phenotypic information and family relationships are available, such as FINRISK (Vartiainen, Jousilahti et al. 2000), the Northern Finland Birth Cohorts

(NFBC66 and NFBC86) (Rantakallio 1969; Jarvelin, Hartikainen-Sorri et al. 1993), HEALTH 2000 (Aromaa, Koskinen et al. 2004), the Finnish Twin Registry (Kaprio, Sarna et al. 1978), COROGENE: The Genetic Predisposition of Coronary Heart Disease in Patients Verified with Coronary Angiogram (Vaara, Nieminen et al. 2011), and The Cardiovascular Risk in Young Finns Study (Raitakari, Juonala et al. 2008). These enormous amounts of data will require intelligent and automated approaches to analysis.

Currently, there are hundreds of statistical analysis programs available for geneticists, which implement a wide variety of mapping methods. In a typical mapping study, scientists apply several programs, because none of them can perform all the desired analyses. In this study, we have automated the process of using various packages in large genome-wide scans, developed software for linkage and LD-based mapping in pedigrees, evaluated the statistical properties of commonly used family-based methods, and investigated the relative efficiency of a variety of mapping strategies.

2 Review of the literature

2.1 The human genome

The human genome is composed of a sequence of approximately 3.17 billion base pairs (The Genome Reference Consortium, assembly GRCh37.p7) (The Genome Reference Consortium 2012) of double stranded deoxyribonucleic acid (DNA), organized into 23 pairs of chromosomes. Normally, each individual receives one copy of each of the 22 autosomes from each parent along with a sex chromosome. There are two sex chromosomes; X and Y, such that an individual with two X chromosomes would be female and an individual with one X and one Y chromosome would be male. Only a small fraction, 1.5%, of the genome is known to be protein coding sequence, while the function of the rest is not well understood (International Human Genome Sequencing Consortium 2004), however there are known functional non-protein-coding ribonucleic acids (ncRNAs), which are supporting protein translation (Birney, Stamatoyannopoulos et al. 2007).

2.1.1 Genetic linkage

Genetic linkage is the result of a physical phenomenon, where large chromosomal segments are inherited intact from parents to offspring, disrupted occasionally by crossovers that occur in meiosis. A crossover refers to an exchange of DNA segments between an individual's pair of homologous chromosomes in meiosis (Figure 1). The genetic map distance between two loci on the same chromosome (syntenic loci) is measured in terms of the expected number of crossovers between them per meiosis (measured in Morgans). This *genetic* distance between loci cannot be directly predicted from the *physical* distance between them, because the frequency of crossovers along each chromosome varies tremendously from region to region. Furthermore, the relationship between physical distance and frequency of recombination also varies with both age and sex of the parent in which the meiosis occurred. On average there are about 27 crossovers per meiosis in spermatogenesis and about 42 in oogenesis (Chowdhury, Bois et al. 2009). Alleles of loci which are closely linked on the same chromosome are more likely transmitted together in meiosis, than those which are on opposite ends of the same chromosome or on different chromosomes (non-syntenic loci) (Sham 1998). This property forms the basis of genetic linkage analysis, in which segregation patterns of known marker loci are compared with inferred segregation patterns of putative genetic variants influencing the trait of interest.

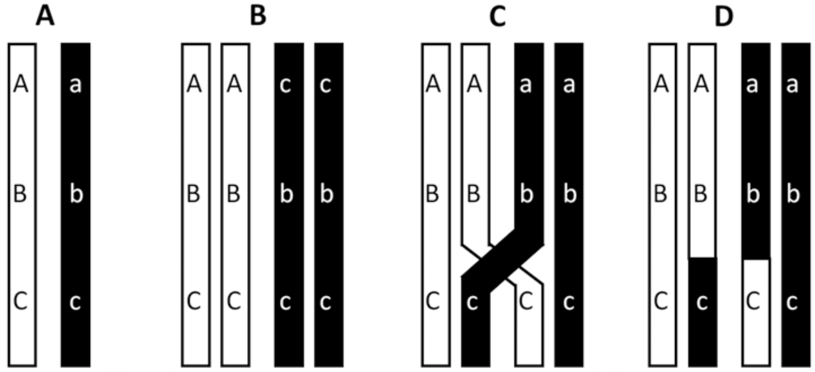

Figure 1: Crossing over in meiosis. (A) A pair of homologous chromosomes (paternally and maternally derived), with three polymorphic loci shown with haplotypes ABC and abc. (B) Chromosome duplication. (C) Crossing over between homologous chromosomes. (D) Recombinant chromosomes are formed with haplotypes ABc and abC.

2.1.2 Genome-wide linkage scan

In a classical hypothesis-free genome-wide linkage scan, the segregation of a large number of polymorphic marker loci spanning the entire length of the human genome is investigated in families. In such genome-wide screens, the genomic location of a putative trait locus is not known and co-segregation of each marker locus with the disease locus is tested, where disease locus genotypes are inferred from phenotypes using some sort of probabilistic inheritance model. These models may be dominant or recessive, for example. In a dominant model, one copy of the disease allele can increase disease risk by itself, where under a recessive model two disease alleles would be required to increase risk of disease. A typical linkage study can contain a few large pedigrees or hundreds of smaller nuclear pedigrees, in which multiple related individuals share some phenotype of interest. Phenotypes can be qualitative traits (affected or unaffected with some disease) or they can be quantitative traits related to normal variation such as bone mineral density (Styrkarsdottir, Cazier et al. 2003), serum levels of different proteins, or total fat mass (Comuzzie, Hixson et al. 1997). For example, linkage analysis has been a powerful method for localizing disease alleles with large effects such as cystic fibrosis (Tsui, Buchwald et al. 1985), Huntington's disease (Gusella, Wexler et al. 1983), Duchenne muscular dystrophy (Murray, Davies et al. 1982), and many rare diseases in Finland (Peltonen, Jalanko et al. 1999).

2.1.3 Linkage disequilibrium

When a mutation occurs in the DNA sequence, this newly created sequence variant (or allele) is only found on the single haplotype where it arose. As the allele segregates through the population, the length of the shared haplotype decays slowly over time due to recombination. This allele, therefore, will not occur independently of the alleles at very tightly-linked loci, as they co-segregate in the population over many generations. This nonindependence of alleles occurring on haplotypes is referred to as linkage disequilibrium (LD), which decays over time as rare recombination events occur (Ott 1999).

2.1.4 Genome-wide association scan

Recent technological achievements resulting in a dramatically reduced cost of large-scale genotyping have made possible genome-wide association studies (GWAS) with hundreds of thousands or even millions of single nucleotide polymorphisms (SNPs) spanning the genome. GWAS was motivated by the 'common disease-common variant' (CD/CV) hypothesis (Reich and Lander 2001), where common disease variants could have a role in common diseases. In such studies, these SNP markers are examined in a large (typically population-based) sample of thousands of individuals, for example see the Wellcome Trust Case Control Consortium (WTCCC 2007). The method is based on the idea that, when large numbers of SNP markers are analyzed over the genome, one or more of these SNPs might be in LD with any disease-predisposing allele. This is likely because most of sequence variants were once created by mutations occurring on single haplotypes deep in evolutionary history. Therefore, if recombination has occurred only rarely between adjacent loci around such a variant, then significant LD will remain over a small region (Hartl and Clark 2007). Although LD mapping in unrelated individuals is currently in fashion, association analysis is typically more powerful in families, though they can be more difficult to ascertain in outbred populations. So far these large cross-sectional GWA studies have only found a relatively small number of variants which explain a fairly large percentage of the genetic contribution to these common complex phenotypes (Visscher, Brown et al. 2012).

2.2 Statistical testing in linkage and association analysis

In human genetics, gene mapping is almost exclusively based on statistical inference, which allows us to evaluate if the disease locus and the marker locus are correlated. One basis for developing statistical tests is to use likelihood functions, which are introduced below.

2.2.1 Likelihood and likelihood ratio test

The likelihood of a set of data under a given hypothesis is $L(\mathbf{H})$, which is proportional to the probability of the data given parameters in $\mathbf{H}$:

$$L(\mathbf{H}) \propto P(data; \mathbf{H}).$$

The method of maximum likelihood tries to find numerical values for the parameters in $\mathbf{H}$ which maximize the likelihood function $L(\mathbf{H})$ (Fisher 1922), the so-called maximum likelihood estimates (MLE) of those parameters.

The exact value of the likelihood function is not meaningful by itself, but rather what is of interest is the ratio of likelihoods under null and alternative hypotheses (i.e. how many times more likely is the data under the alternative hypothesis than under the null). Therefore we evaluate the likelihood ratio as

$$LR = \frac{P(data;\ H_A)}{P(data;\ H_0)} = \frac{L(H_A)}{L(H_0)},$$

where H_0 is the null hypothesis, and the alternative hypothesis is H_A. This ratio forms the basis for the likelihood ratio test (LRT). LRTs have been increasingly popular in the era of computing because according to the Neymann-Pearson lemma (Neymann and Pearson 1933), if there exists a most-powerful test of a given hypothesis, it would be of the form of a likelihood ratio test Asymptotically, under certain regularity conditions, the statistical distribution of the LRT is well-known. Wilks' theorem states that, if sample size approaches ∞, then $2\ln(LR)$ is asymptotically distributed according to a χ^2 distribution with k degrees of freedom (where k is the difference in the number of free parameters between H_A and H_0) (Wilks 1938).

When the regularity conditions of Wilks' theorem are not met, the distribution of the test statistic can be rather complicated (Davies 1977). Such problems arise, for example, when the likelihood is a function of more parameters under the alternative hypothesis than under the null, or when the null hypothesis occurs at a boundary value for a parameter. In such cases, the statistical distribution of the LRT is often approximated by a mixture of χ^2 distributions and sometimes a point-mass at 0 (examples of this will be pointed out later).

Likelihoods can be used for evaluating a probability model, which describes known correlations among individuals in a data set. Then the likelihood can be maximized over admissible values of the parameters of the probability model. One attractive feature of maximum likelihood estimation is that some of the parameters, such as allele frequencies, can be treated as *nuisance parameters*, which are estimated separately under both the null and alternative hypothesis, without changing the degrees of freedom.

2.2.2 The lod score

If two loci are very closely linked on the same chromosome, it is highly unlikely that recombination will occur between them in any given meiosis. The probability of recombination in a given meiosis is called the recombination fraction, θ, which ranges on 0 ≤ θ ≤ 0.5 in humans (some polyploid species can have $\theta > 0.5$ (Wright, Johnson et al. 1983)). The recombination fraction between two unlinked or non-syntenic loci is θ=0.5. The goals of linkage analysis are to estimate the recombination fraction between two loci and test the null hypothesis that $\theta = 0.5$.

Historically, the idea of testing for linkage by using the likelihood ratio was proposed by Haldane and Smith (Haldane and Smith 1947), and the lod ("logarithm-of-odds") score for statistical inference was introduced by Barnard (Barnard 1949). The linkage test statistic, the lod score, was formulated in terms of the common logarithm ($\log_{10}$) of the likelihood ratio. The lod score statistic was introduced in genetic analysis by Smith (Smith 1953) and Morton (Morton 1955) as

$$Z(\theta) = \log_{10}(LR) = \log_{10}\left[\frac{P(data;\theta)}{P(data;\theta = 0.5)}\right] = \log_{10}\left[\frac{L(\theta)}{L(\theta = 0.5)}\right].$$

The traditional threshold for significance testing was proposed by Morton (Morton 1955) to be Z > 3, corresponding to a pointwise p-value of 0.0001 asymptotically (p-value is discussed in section 2.2.3). When the lod score is maximized over the recombination fraction (Z$_{\max}$) on the range $\theta \in [0, 0.5]$, then $2\ln(10)Z_{\max} = 2\ln\left[\max_{\theta} L(\theta)/L(\theta = 0.5)\right]$ is asymptotically distributed as a 50:50

mixture of a point mass at 0 and χ_1^2, because the test is performed in a one-sided manner (θ cannot exceed 0.5). In a genome-wide linkage scan with an infinitely dense marker map, a lod score threshold of 3.3 has been shown to correspond roughly to a genome-wide false positive rate of 5%, (i.e. corrected for multiple testing (Lander and Kruglyak 1995)).

Linkage and LD are correlations between alleles of linked loci and have nothing to do with trait phenotypes. The relationship between the trait phenotypes and the putative underlying (unknown) risk genotypes is always modeled probabilistically. In most cases, one models disease loci as having two alleles: a disease-predisposing allele D and a wild-type allele +, because for rare Mendelian disorders, it is highly unlikely that multiple alleles would be segregating in any single pedigree. However, there are almost always multiple variants in every disease locus in the population (Terwilliger and Weiss 1998). The genotype → phenotype relationship is modeled then in terms of the conditional probabilities of having the disease phenotype given each possible disease locus genotype (i.e. penetrance functions).

2.2.3 The p-value

The significance of a statistical test is typically measured with a p-value, which is a probability of obtaining the same or more extreme value for the test statistic when H_0 is true. The p-value can be determined from the sampling distribution of the test statistic under H_0. A type-I error occurs when the test rejects H_0 although H_o is true, and a type-II error occurs when the test fails to reject H_0 when H_o is false. The power of a test is a probability of rejecting H_0 when H_0 is not true (Table 1). The choice of the significance level (denoted with α) for any given test is fairly arbitrary, based on some a priori decision about how many false positive results we would be willing to accept. For example, if we use α=0.0001 (one out of every ten thousand tests would give a positive result even if there were no true signal), and our test statistic give a p-value ≤ α, then we would declare our test results to be statistically significant and reject H_0.

Table 1. Type-I, type-II, and power, where H_0 is the null hypothesis.

Statistical decision	**True state of nature**	
	H_0 true	H_0 false
Reject H_0	Type-I error, False positive	Power, True positive
Do not reject H_0	True negative	Type-II error, False negative

2.3 Linkage analysis of qualitative traits

The purpose of linkage analysis is to test whether two loci co-segregate in meiosis more often than expected by chance. In such an analysis, one tries to determine whether offspring have inherited non-recombinant or recombinant gametes from their parents. In Figure 2, a three-generation pedigree is shown with known genotypes of two loci for each individual. The vertical line between the genotypes indicates the phase of the alleles, meaning which alleles that individual received in each parental gamete. Four children in this pedigree received non-recombinant gametes from their father (denoted with N), while one child received a recombinant gamete from him (denoted with R). This can be determined with certainty because it is clear that the father received alleles 2 and D from his father and alleles + and 1 from his mother, but he has transmitted alleles + and 2 to the 5th child, which he himself had received from different parents (hence, a "re-"combination). If the loci were unlinked, then in expectation, 50% of gametes would be non-recombinant and 50% recombinant. However, if two loci are linked, non-recombinant gametes should be more frequent.

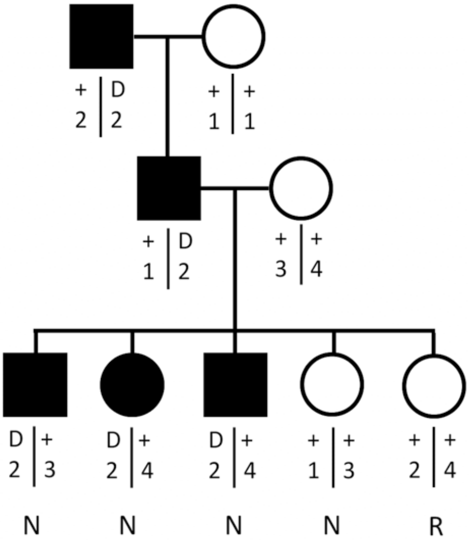

Figure 2. Three-generation pedigree illustrating co-segregation and recombination. Each individual has alleles of two loci shown where the first locus has alleles D and +, and the second locus has alleles 1, 2, 3 and 4. A vertical line between the loci indicates that the phase is known, i.e. father with genotypes +/D and 1/2 has phase <u>+ 1</u> and <u>D 2</u> because he received alleles + and 1 from his mother and alleles 2 and D from his father. In the third generation each recombinant meiosis from the father is indicated with R and each non-recombinant meiosis with N. The mother is homozygous at the first locus and therefore the recombination status of her gametes cannot be classified.

If recombinant and non-recombinant offspring can be counted as in Figure 2, then the likelihood of the data would be

$$L(\theta) = K\theta^{R}(1-\theta)^{N} \,,$$

where the constant *K* contains the segregation probabilities, genotype probabilities for the (genotyped) founders, and combinatoric terms (i.e. binomial coefficients). Because this constant *K* is identical in both numerator and denominator of the likelihood ratio, it can be factored out, and thus its value is irrelevant and need not be computed. The number of recombinant offspring, *R=1,* and the number of non-recombinant offspring, *N=4*. The lod score statistic for the pedigree in Figure 1 can be computed as

$$\begin{aligned} Z(\theta) = \log_{10}\left[\frac{P(data;\theta)}{P(data;\theta=0.5)}\right] &= \log_{10}\left[\frac{L(\theta)}{L(\theta=0.5)}\right] \\ = \log_{10}\left[\frac{\theta^{R}(1-\theta)^{N}}{(0.5)^{R}(1-0.5)^{N}}\right] &= \log_{10}\left[\frac{\theta^{R}(1-\theta)^{N}}{(0.5)^{R+N}}\right] \\ = \log_{10}\left[\frac{\theta^{1}(1-\theta)^{4}}{(0.5)^{1+4}}\right] &= \log_{10}\left[\frac{\theta(1-\theta)^{4}}{(0.5)^{5}}\right]. \end{aligned}$$

Now, the lod score can be maximized over recombination fraction and it turns out that $Z_{\max}$ is obtained when $\hat{\theta} = R/(R+N)$ = 1/5 = 0.2. The lod scores can be summed over all pedigrees in an analysis for each fixed set of parameter values (including the model parameters and the recombination fraction). Linkage analysis can be performed one marker at time (two-point linkage), or with multiple markers jointly (multipoint linkage), as first automated by Lathrop et al. (Lathrop, Lalouel et al. 1984).

2.4 Association analysis of qualitative traits

Linkage analysis can be only performed in families, because the method tests whether the alleles of two or more loci co-segregate more often than expected by chance. In association analysis using a case-control design, study subjects are assumed to be 'unrelated'- that is to say their genotypes are assumed to be independent and identically distributed (i.i.d.). However, the power of the method is reliant on the assumption that affected individuals are distant relatives, who are connected to each other by an unknown number of historical meioses, and share clonal risk alleles identical by descent (IBD) through a single historical lineage. In Figure 3 this unknown pedigree structure is indicated by dotted lines. Linkage analysis tests for co-segregation in directly observed meioses within families, as indicated by solid lines, while LD analysis tests for co-segregation in unobserved historical meioses on the assumption that a single lineage for the risk alleles exists although its specific gene genealogy is unknown (Sham 1998; Terwilliger and Göring 2000).

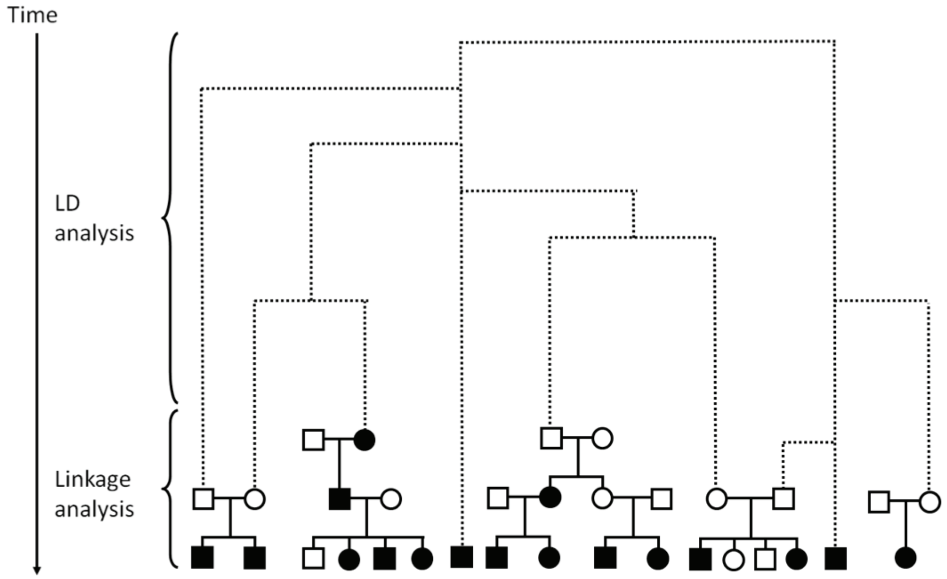

Figure 3. Linkage analysis uses families where the relationships between individuals are known (solid lines), while LD analysis is essentially linkage analysis in the enormous pedigree of unknown structure connecting these individuals together historically (dotted lines), resulting in shared haplotypes across observed pedigrees. The dotted lines connect the individuals we implicitly assume are distant relatives, with an unknown number of meioses connecting them.

2.4.1 Definition of linkage disequilibrium

Let us assume that we have two loci with alleles A and a at the first locus and alleles B and b at the second, allele frequencies P(A), P(a) = 1 - P(A), P(B), and P(b) = 1 – P(B), and haplotype frequencies P($\underline{\text{A B}}$), P($\underline{\text{A b}}$), P($\underline{\text{a B}}$), and P($\underline{\text{a b}}$) = 1 – P($\underline{\text{A B}}$) - P($\underline{\text{A b}}$) - P($\underline{\text{a B}}$). Alleles of the two loci are said to be in LD if the haplotype frequency deviates significantly from the product of the allele frequencies, i.e. P($\underline{\text{A}}$

B)≠P(A)P(B). We can quantify this nonindependence as P(A B)=P(A)P(B) + δ, where δ is the gametic linkage disequilibrium coefficient (Lewontin and Kojima 1960). Because the range of δ is dependent on allele frequencies, its numerical value is not particularly meaningful. For this reason, geneticists prefer to work with more easily-interpretable measures of LD, whose meanings are not allele-frequency dependent. One such measure for LD is Lewonti's D' (Lewontin 1964), which is defined as

$$D' = \frac{\delta}{\delta_{MAX}} = \begin{cases} \dfrac{\delta}{\min\{P(a)P(B), P(A)P(b)\}}, \delta \geq 0 \\ \dfrac{\delta}{\min\{P(a)P(b), P(A)P(B)\}}, \delta < 0 \end{cases}.$$

The range of D' is -1 ≤ D' ≤ 1, but usually the absolute value of D' is used, i.e. |D'|, because the sign depends only on the arbitrary labeling of the alleles. If two loci are in linkage equilibrium (LE), then D'=0, while when a new allele first arises in the population by mutation, it is always on a single haplotype such that |D'| = 1 with every other locus in the genome, after which, in a large expanding population it decays each generation according to the equation $\delta_k = (1-\theta)^k \delta$, where k is the number of generations (Lewontin and Kojima 1960).

Another increasingly popular measure of LD is the squared correlation coefficient, $r^2 = \delta^2 / \left[P(A)P(a)P(B)P(b)\right]$, which measures how well the alleles of either locus can be predicted from the other. This is often used in the GWAS era to measure power, in that in order to detect association between marker and disease loci with $r^2 < 1$, one would need to increase the sample size by roughly a factor of $1/r^2$ (Gabriel, Schaffner et al. 2002). Therefore, r^2 and D' do not measure the same thing. For example, |D'|=1 in the case of a very recent mutation, but the value of r^2 is typically close to 0, because although the presence of the new allele predicts well what allele is present at the second locus, the allelic state of the second locus has almost no predictive value about the presence of the new allele. Alleles of two loci must be of similar frequency in order to have high r^2 (Hartl and Clark 2007), and therefore r^2 is of less utility in population genetics than it is in epidemiology as its value is uninformative about population history.

2.4.2 Cases and controls (singletons)

The simplest design to study allelic association is to collect a random sample of affected cases and unaffected controls from some population. If a locus has an allele (or alleles) which have an effect on the disease outcome, then the allele frequencies would be expected to be different between affected and unaffected individuals in the population (Sham 1998).

Let *Ph* represent the observed phenotype of a given individual, G_M - the observed marker genotype (with alleles 1 and 2), and G_D - the underlying unobserved disease locus genotype (with alleles D and +). The conditional probability of the observed phenotype given each possible disease locus genotype are called penetrances, $P(Ph \mid G_D)$, and the disease locus genotype frequencies are $P(G_D)$. The conditional probability of the marker genotypes given each possible disease locus genotype is $P(G_M \mid G_D)$, which is a function of the marker locus genotype frequencies in the population and linkage disequilibrium between the disease and marker loci. That is to say, for a random individual $P(G_M = 1/1 \mid G_D = +/+) = \left[P(1 \mid +)\right]^2$ and $P(G_M = 1/1 \mid G_D = D/+) = P(1 \mid D)P(1 \mid +)$, assuming the Hardy-Weinberg equilibrium (HWE) (Hardy 1908; Weinberg 1908) at the marker locus, where $P(1 \mid D)P(D) = P(\underline{1\,D})$, etc. The likelihood of the data can thus be computed as a function of linkage disequilibrium between marker and disease locus genotypes as

$$L \propto P(Ph, G_M) = \sum_{G_D} P(Ph, G_M \mid G_D)P(G_D) = \sum_{G_D} P(Ph \mid G_D)P(G_M \mid G_D)P(G_D).$$

The likelihood under the null hypothesis of no linkage disequilibrium would be computed assuming $P(1 \mid +) = P(1 \mid D) = P(1)$ and $P(2 \mid +) = P(2 \mid D) = P(2)$, while under the alternative hypothesis the conditional allele frequencies would be estimated freely.

2.4.3 Family-based association

While sampling unrelated singletons from the population is quite simple, it can be sensitive to various sampling biases, such as population stratification. One approach to avoid this is to use non-transmitted alleles from the parents of an affected child as control alleles, because non-transmitted alleles are well matched, as they come from the homologous chromosomes of the same parent. A method based on this idea is the (genotype-based) haplotype relative risk (HRR) (Rubinstein, Walker et al. 1981). An HRR sampling unit is a triad, consisting of two parents and one affected offspring. In this approach, the affected child contributes one case genotype and the two non-transmitted alleles from the parents contribute one "cohort" genotype (Ahsan, Hodge et al. 2002). A more powerful (in the presence of HWE) version of this is the haplotype-based haplotype relative risk (HHRR) (Terwilliger and Ott 1992), which compares alleles rather than genotypes. The HHRR is likewise a test of linkage disequilibrium, which is only powerful in the presence of linkage.

The HRR can also be formulated as a paired-sampling McNemar test (McNemar 1947), in which case only transmission from heterozygous parents would be analyzed. This censoring of information from homozygous parents provides protection from false positives when HWE does not hold at the marker locus, at a nontrivial cost, however, in terms of power if HWE does, in fact, hold (Terwilliger and Ott 1992). A test of the same algebraic form, based on paired sampling was introduced as the transmission/disequilibrium test (TDT) of linkage by Spielman et al (Spielman, McGinnis et al. 1993), which is powerful only in the presence of LD. The TDT test was designed to be applied to alleles transmitted from all heterozygous parents to all affected offspring, treating all such transmissions, even to siblings from the same parent, as independent observations. In this case, the TDT is a valid test of linkage, which is more powerful in the presence of linkage disequilibrium, but a significant TDT does not imply that there must be LD (Spielman and Ewens 1996; Göring and Terwilliger 2000).

While HHRR and TDT only count non-transmitted and transmitted alleles in triads, we can also compute the full likelihood in general family data as

$$L \propto P(\mathbf{G_M}, \mathbf{Ph}) = \sum_{\mathbf{G_D}} P(\mathbf{Ph}, \mathbf{G_M} \mid \mathbf{G_D}) P(\mathbf{G_D}) = \sum_{\mathbf{G_D}} P(\mathbf{Ph} \mid \mathbf{G_D}) P(\mathbf{G_M} \mid \mathbf{G_D}) P(\mathbf{G_D}),$$

except that now $\mathbf{G_D}$ is a vector of phased disease locus genotypes, $\mathbf{G_M}$ is a vector of phased marker locus genotypes, and **Ph** is a vector of observed phenotypes for all individuals in a pedigree jointly, and P($\mathbf{G_D}$) is a function of disease locus genotype frequencies for founders, and Mendelian transmission probabilities for all other individuals, conditional on their parent's genotypes. $P(\mathbf{G_M} \mid \mathbf{G_D})$ is a function of linkage disequilibrium parameters (i.e. conditional allele frequencies) in the unrelated pedigree founders, and is a function of the segregation probabilities and recombination fraction between disease and marker loci for all other individuals in the pedigree, conditional on disease locus alleles in vector $\mathbf{G_D}$. The "case" and "control" alleles in such family-based analysis retain the desirable property of the HHRR design that they derive from the same set of individuals who are clearly part of the same "breeding population".

In Figure 4, one affected sib-pair is shown with each individual genotyped at one marker locus. In order to compute the likelihood of the data, observed phenotypes must be probabilistically related to disease locus genotypes using some inheritance model. If we assume the model $\{P(D) = p_D = 0.001$, $P(Affected \mid D/D) = 1$, $P(Affected \mid D/+ \, or + /+) = 0.\}$, then both affected offspring must have genotype D/D and both unaffected parents must have genotype D/+ (Figure 4B).

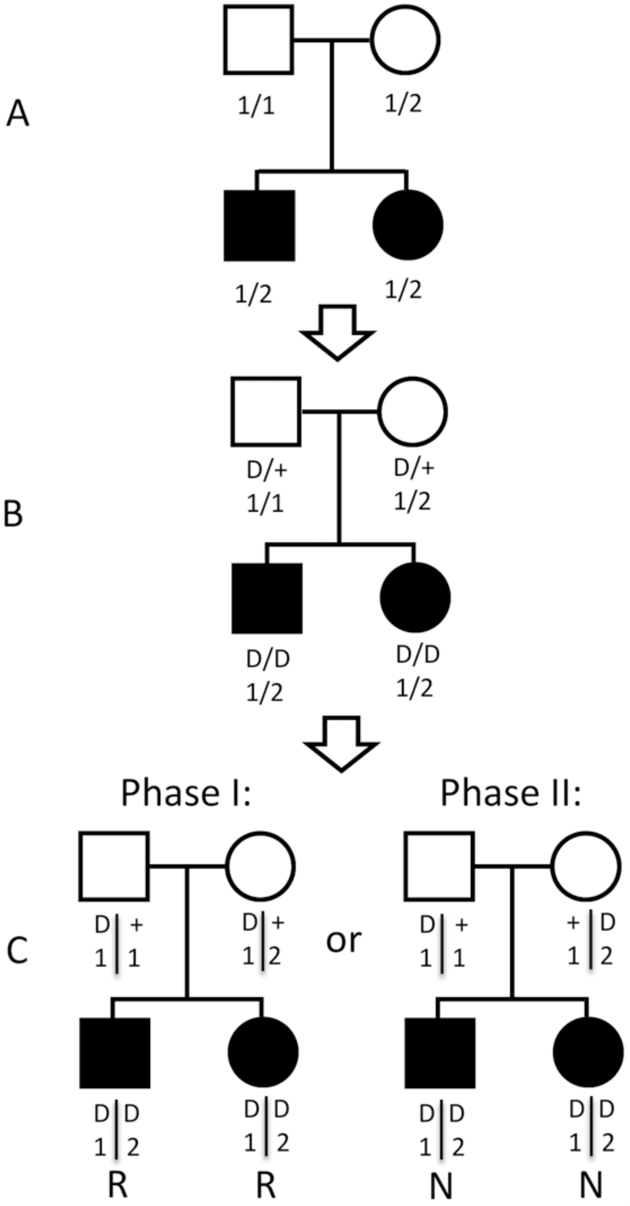

Figure 4: (A) An affected sib-pair with genotypes of one marker locus is shown. (B) Disease locus genotypes uniquely inferred from inheritance model { $p_D = 0.001$, $P(Affected \mid D/D) = 1$, $P(Affected \mid D/+ \; or + /+) = 0.$ } (C) Two possible phases for the mother, where phase I is D 1/+ 2 (both offspring recombinant) and phase II is + 1/D 2 (both offspring non-recombinant).

However, we do not know the phase of these genotypes in the mother, since she is heterozygous at both loci, and therefore we must analyze the data allowing for the possibility of each possible phased genotypes for her (Figure 4C), i.e. phase I: $\underline{\text{D 1}}/\underline{\text{+ 2}}$ or phase II: $\underline{\text{+ 1}}/\underline{\text{D 2}}$, such that the likelihood of this pedigree is

$$\begin{aligned} L &\propto P(\mathbf{Ph},\mathbf{G_M},\mathbf{G_D},\text{Phase}) \\ &= \sum\nolimits_{\mathbf{G_D}} P(\mathbf{Ph}\mid\mathbf{G_M},\mathbf{G_D},\text{Phase})P(\mathbf{G_M},\mathbf{G_D},\text{Phase}) \\ &= \sum\nolimits_{\mathbf{G_D}} P(\mathbf{Ph}\mid\mathbf{G_D})P(\mathbf{G_M},\mathbf{G_D},\text{Phase}) \\ &= \sum\nolimits_{\mathbf{G_D}} P(\mathbf{Ph}\mid\mathbf{G_D})P(\mathbf{G_D})P(\mathbf{G_M},\text{Phase}\mid\mathbf{G_D}), \end{aligned}$$

where under both phases $P(\mathbf{G_D}) = [2p_D(1-p_D)][2p_D(1-p_D)](0.25)(0.25)$, and $(0.25)'s$ are the segregation probabilities for a child having disease locus genotype D/D given that both parents are D/+. $P(\mathbf{Ph}\mid\mathbf{G_D}) = (1)(1)(1)(1) = 1$, because for the two affected children penetrances are P(Affected|D/D)=1and for the unaffected parents, the penetrances are P(Unaffected|D/+)=1-P(Affected|D/+)=1. The probability of the marker genotypes conditional on the inferred disease genotypes under phase I, where the mother transmits a recombinant gamete, $\underline{\text{D 2}}$, to both of her offspring, is

$$P(\mathbf{G_M},\text{Phase I}\mid\mathbf{G_D}) = p_{1|D}p_{1|+}p_{1|D}p_{2|+}(\theta)(\theta),$$

where, for example, $P(1\mid D) = p_{1|D}$, etc (where $p_{1|D} = p_{\underline{1D}}/p_D$), and θ is the recombination fraction. However, under phase II, the $\underline{\text{D 2}}$ gamete would be non-recombinant, such that

$$P(\mathbf{G_M},\text{Phase II}\mid\mathbf{G_D}) = p_{1|D}p_{1|+}p_{1|+}p_{2|D}(1-\theta)(1-\theta).$$

Now, the total likelihood is

$$\begin{aligned} L &\propto \sum\nolimits_{\mathbf{G_D}} P(\mathbf{Ph}\mid\mathbf{G_D})P(\mathbf{G_D})P(\mathbf{G_M},\text{Phase}\mid\mathbf{G_D}), \\ &= [2p_D(1-p_D)][2p_D(1-p_D)]\big(0.25\big)\big(0.25\big)\times \\ &\quad \big[p_{1|D}p_{1|+}p_{1|D}p_{2|+}(\theta)(\theta) + p_{1|D}p_{1|+}p_{1|+}p_{2|D}(1-\theta)(1-\theta)\big] \\ &= 0.25[p_D(1-p_D)]^2 p_{1|D}p_{1|+}\left[p_{1|D}p_{2|+}(\theta)^2 + p_{1|+}p_{2|D}(1-\theta)^2\right], \end{aligned}$$

which is evidently a function of both linkage disequilibrium between the alleles of the loci, and the recombination fraction between marker and disease loci.

2.4.4 Family-based association programs

There have been a great number of statistical methods and associated software packages developed for family-based association analyses (cf. (Ott, Kamatani et al. 2011)). Most algorithms can only efficiently utilize homogenous relationship structures, such as triads, sib-pairs or nuclear families, and only a few of these methods can jointly test for both linkage and LD. None of those software packages is fully satisfactory when it comes to: 1) combining various relationship structures (singletons, triads, sib pairs, large nuclear families, and extended pedigrees) into one analysis; 2) using all the information from extended pedigrees; 3) allowing for missing genotype data; and 4) testing for LD conditional on linkage. Some of the most commonly used programs are briefly introduced below.

2.4.4.1 GENEHUNTER TDT

GENEHUNTER (Kruglyak, Daly et al. 1996; Kruglyak and Lander 1998) is a multipurpose program for parametric and non-parametric linkage analysis. It also implements the classic TDT (Spielman, McGinnis et al. 1993) test of linkage, which is more powerful in the presence of LD.

2.4.4.2 PLINK

PLINK (Purcell, Neale et al. 2007) is a feature rich program for genome-wide association analysis of population-based samples. In addition, it implements rudimentary methods for family-based association studies, including the classic TDT and a variant of this test called *parentTDT*, which incorporates parental phenotype information. In this method, the alleles in affected vs. unaffected parents are counted, treating parents of a nuclear family as a matched pair. These counts are then combined with the classic TDT’s transmitted and non-transmitted allele counts.

2.4.4.3 QTDT

QTDT (Abecasis, Cardon et al. 2000; Abecasis, Cookson et al. 2000) is a program which implements a TDT-type test for general pedigrees. The main difference between this and the classic TDT is that it accounts for the kinship coefficient between family members. The kinship coefficient is the probability that (under the null hypothesis of no linkage and no LD) a random allele drawn at random from each of two individuals at the same locus will be IBD (Wright 1922). In QTDT, the measure of allelic transmission is the difference between the observed genotype score and the expected genotype score, where the kinship coefficient is used in computing the expected genotype score. If there are multiple affected offspring in a family, transmission from parents to all offspring are treated as a single unit.

2.4.4.4 FBAT

The FBAT (Laird, Horvath et al. 2000; Rabinowitz and Laird 2000) program implements a TDT-type test for nuclear families, which allows association testing in the presence of linkage by using minimal sufficient statistics (Lake, Blacker et al. 2000). If pedigree data contains multigenerational families, those are decomposed to nuclear families which are treated as if they were independent in the analysis.

2.4.4.5 TRANSMIT

The TRANSMIT (Clayton 1999) program implements a TDT-type test for nuclear families. According to the author, it allows for multiple affected offspring, missing genotype data and can be used for association testing in the presence of linkage by using a robust variance estimator. Like FBAT, it decomposes multigenerational families into nuclear families and treats them as if they were independent in the analysis and ignores parental phenotypes. However, unlike FBAT, when there is missing parental genotype data, TRANSMIT attempts to infer all compatible genotypes for parents if there are additional unaffected or unaffected siblings in the family and averages over all compatible configurations of the data.

2.4.4.6 UNPHASED

UNPHASED (Dudbridge 2008) is similar to TRANSMIT in how it treats multigenerational families, it allows testing association in the presence of linkage, and allows for missing data. In contrast to TRANSMIT, UNPHASED can incorporate unrelated subjects in the analysis. However, unrelated individuals and nuclear families are treated as separate samples in the likelihood computation, such that HWE is assumed for the singletons.

2.4.4.7 MENDEL

MENDEL (Lange, Cantor et al. 2001) is a comprehensive software package for linkage and association analysis of qualitative and quantitative traits. Cantor et al. (Cantor, Chen et al. 2005) introduced a parametric maximum-likelihood-based option in MENDEL to model joint linkage and association analysis in full pedigrees. In this option, MENDEL estimates the recombination fraction and the conditional frequencies of the disease allele given each marker allele $P(D|1,2,...,n)$, such that the constraint $p_D = \sum_{i=1}^{n} p_i p_{D|i}$ is forced and the population allele frequency p_i of each marker allele i is fixed. These marker allele frequencies can be estimated from the data using other options of MENDEL. MENDEL also implements a so-called "gamete-competition" model (Sinsheimer, Blangero et al. 2000), which is a TDT-type test of linkage in pedigrees, whose power increases with increasing LD.

2.4.4.8 LAMP

LAMP (Li, Boehnke et al. 2005; Li, Boehnke et al. 2006) is a maximum-likelihood-based program for joint linkage and association analysis in general pedigrees, which allows association testing conditional on linkage. The method uses the Lander-Green-algorithm (Lander and Green 1987) for the likelihood computation and therefore can only analyze relatively small pedigrees in a reasonable time. LAMP assumes complete linkage between disease and marker loci, and estimates marker allele frequencies, haplotype frequencies and disease model parameters jointly, conditionally on assumed disease prevalence ϕ. Estimation of such disease model parameters can be constrained to recessive, dominant, additive or multiplicative models, or penetrances can be estimated freely, with the constraint that prevalence in the population would be

$$\begin{aligned}\phi = {p_D}^2 P(Disease \mid D/D) \; + \; 2p_D(1-p_D)P(Disease \mid D/+) \\ + \; (1-p_D)^2 P(Disease \mid +/+).\end{aligned}$$

However, such models estimated from nonrandomly ascertained data are typically biased, so caution should always be used.

2.4.4.9 PSEUDOMARKER

Likelihood-based joint linkage and LD analysis had already been performed almost two decades ago (Tienari, Terwilliger et al. 1994; Trembath, Clough et al. 1997; Kainulainen, Perola et al. 1999; Enattah, Sahi et al. 2002). For example, the linkage analysis software package LINKAGE (Lathrop and Lalouel 1984) and the segregation analysis package PAP (Hasstedt 1982) have included the capability to jointly model linkage (recombination fraction), and LD (with disease-marker haplotype frequencies) since the 1980's. Göring and Terwilliger (Göring and Terwilliger 2000) introduced a unified theoretical model for joint linkage and/or linkage disequilibrium analysis, which allows joint analysis of singletons, triads, nuclear families and extended pedigrees, which is stochastically equivalent to the classic model-free methods for LD and linkage analysis (Göring and Terwilliger 2000).

In this method, each individual is assigned an artificial ‘pseudomarker’ locus genotype with alleles D (disease allele) and + (wild-type allele), based on the observed disease phenotypes. The idea of this assignment is to make all meioses which connect affected individuals together informative for linkage, such that they share as many pseudomarker alleles IBD as possible. Furthermore, if large pedigrees with multiple affected individuals are ascertained, this would be because of an implicit assumption that affected individuals share a common genetic risk factor affecting the disease phenotype. Therefore, co-segregation of these pseudomarker-genotypes with some genomic region would be consistent with linkage under this hypothesis. Errors in the assumption that all affecteds share a risk allele IBD could be modeled via the recombination fraction parameter. An example of this can be seen in Figure 5, a recessive pseudomarker locus assignment in an affected sib-pair, where the father and mother are forced to be informative for linkage by assigning a pseudomarker locus genotype D/+, with affected children receiving genotype D/D.

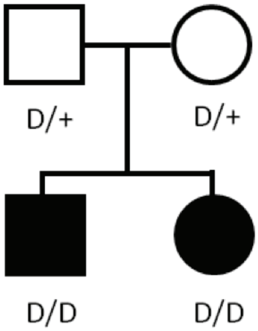

Figure 5. A recessive pseudomarker locus assignment on an affected sib-pair pedigree.

It has been shown that performing linkage analysis by using recessive pseudomarker locus structure is stochastically equivalent to the traditional affected sib-pair mean test on affected sib-pairs (Knapp, Seuchter et al. 1994; Kuokkanen, Sundvall et al. 1996; Satsangi, Parkes et al. 1996). This recessive assignment is completely identical to recessive linkage analysis as in Figure 4. Similarly, applying pseudomarker locus assignment on unrelated singletons and various-sized nuclear pedigrees, yields likelihood-ratio tests of linkage and/or LD that are stochastically equivalent to the classical case-control analyses, the haplotype-based haplotype relative risk (HHRR), and the transmission disequilibrium test (TDT).

2.5 Detectance

The detectance is the inverse of the penetrance (Figure 6) and is defined as $P(\mathbf{G}_\mathbf{D} \mid \mathbf{Ph}; \mathit{ascertainment})$ (Weiss 1993). The power of a study depends on both the strength of linkage and/or LD between marker and trait loci *and* how well the observed phenotype predicts the underlying unobserved trait locus genotypes: the detectance (Weiss and Terwilliger 2000). Technological advances in genotyping and sequencing can only influence the strength of linkage and LD, because in a very large set of markers over the genome, at least one marker is likely to be in linkage and LD with any potential functional variant. However, technology cannot influence the detectance, which can be only modified by altering the ascertainment scheme, as penetrance is an inherent biological property. The ascertainment of large multigenerational pedigrees with multiple affected individuals (who share some trait of interest) from small isolated populations increases the predictive value of phenotype on risk genotype, because affected family members are more likely to share genetic risk factors which have influence on the phenotype.

In Table 2 are listed some simple examples to demonstrate the lack of equivalence between detectance of risk genotypes and penetrance in random samples; a) A complex multifactorial disease, such as stroke, is likely influenced by multiple variants of minimal individual risk (low penetrance) together with environmental factors (Dichgans 2007), and therefore may require enormous sample sizes to detect any functional variant (low detectance), b) The relationship between sex and prostate cancer is a trivial example of a genetic risk factor (sex) which has low penetrance (most men do not have prostate cancer) but high detectance (most prostate cancer patients are male) (cf. American Cancer Society (2012)), c) Retinitis pigmentosa (RP) can be caused by multiple alleles at multiple loci with high penetrance, but given a patient with RP, one cannot predict which of the risk genotypes at which of the risk loci he carries (low detectance) (cf. (Hartong, Berson et al. 2006)), and d) Adult-type hypolactasia in Northern Europe is almost exclusively caused by a single genotype of a single SNP such that individuals with this risk genotype are all unable to digest lactase (high penetrance), and all individuals unable to digest lactase share this same genotype (high detectance) (Enattah, Sahi et al. 2002).

Table 2. Examples of high/low detectance and high/low penetrance phenotypes.

		Detectance	
		Low	**High**
Penetrance	**Low**	a) Rare variants, weak effects (common multifactorial stroke)	b) Common variant, very weak effect (prostate cancer vs sex)
	High	c) Multiple rare variants, each high effect (Retinitis pigmentosa)	d) Single variant, large effect (adult-type hypolactasia)

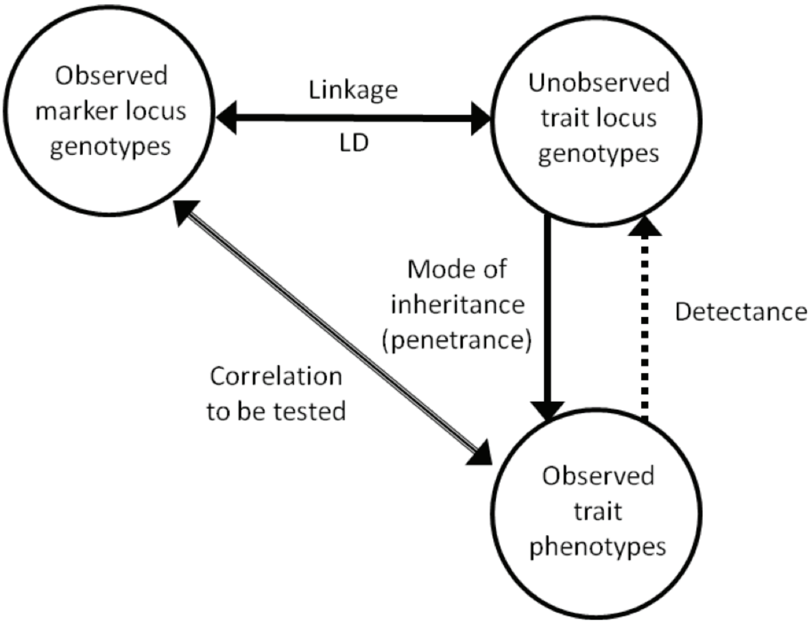
Observed marker locus genotypes
Linkage
LD
Unobserved trait locus genotypes
Mode of inheritance (penetrance)
Detectance
Correlation to be tested
Observed trait phenotypes

Figure 6. In linkage and association studies one tests for correlations between observed marker genotypes and observed trait phenotypes. The power of a study is dependent on the detectance, which is the predictive value of the observed phenotype on the unobserved disease locus genotype, in addition to linkage and/or LD between the disease locus genotypes and marker locus genotypes which have been characterized. Figure modified from Terwilliger and Göring (Terwilliger and Göring 2000).

3 Aims of the study

The aim of this study were to automate genome-wide linkage and association analysis, to develop family-based association software, to evaluate statistical properties of commonly used family-based association methods, and to investigate the validity of likelihood ratio tests for joint and conditional linkage and LD analysis by addressing the following specific aims:

1) To develop software for automating large-scale genome-wide linkage and association analysis (I, III).

2) To develop family-based association software for a combined analysis of singletons, triads, sib-pairs, larger sibships and multigenerational pedigrees (II).

3) To evaluate statistical properties of commonly used family-based association tests in a sample of singletons and related individuals and to compare various study designs (II, III).

4) To investigate properties of the likelihood ratio test in joint and conditional linkage and LD analysis (IV).

4 Materials and methods

4.1 AUTOGSCAN software (I)

In practice, a large genome-wide scan requires a large amount of file formatting when a variety of software packages are used for data analysis. It is typical that each software package has their own file formats for describing pedigrees, trait phenotypes, marker genotypes, marker allele frequencies, marker maps, disease locus parameters, etc. A genome-wide analysis also requires the user to repeat the same basic steps multiple times for each chromosome. This process can be extremely time-consuming and prone to error. Thus skilled statistical geneticists were overloaded by 'simple' analysis help requests from students and post-docs, even though their expertise would be much more valuable in interpreting the results rather than performing the analyses. There was an obvious need for automating these processes for researchers.

There exist at least one program tool for automated file formatting and analysis such as MEGA2 (Mukhopadhyay, Almasy et al. 2005). MEGA2 is a menu-driven (can be run in batch-mode) program which can create input files for more than 30 different analysis programs, and allows flexible selection of phenotypes and markers for the analysis. Additionally, MEGA2 includes the capability to create graphical plots from the analysis results, because most of the statistical software packages create only plain text output files.

In Study I, the automation of genome-wide linkage and linkage disequilibrium analysis with several widely-used programs was implemented in the AUTOGSCAN program. The input files for AUTOGSCAN software are a LINKAGE format pedigree file (Lathrop and Lalouel 1984) describing family relationships, gender, phenotype, and marker genotypes (each chromosome in a separate file), MEGA2 format marker map file (Mukhopadhyay, Almasy et al. 2005) describing the chromosome, a unique name, and a genomic position for each of the marker loci, a phenotype file, a parametric model file, and a control file (for specific analysis options). A separate phenotype file allows for easy analysis of multiple phenotypes, because most of the programs require phenotype in the pedigree file.

This program contains several scripts and auxiliary programs which automatically handle file formatting (including combining pedigree and phenotype files), checking for Mendelian inconsistencies with PedCheck (O'Connell and Weeks 1998), computing pedigree, phenotype, and marker statistics using PedStats (Wigginton and Abecasis 2005), and estimating of the marker allele frequencies from the data by simple counting (if required). After successful file formatting and if Mendelian inconsistencies are not found, automatic running of the statistical analyses are performed using ANALYZE, GENEHUNTER (Kruglyak, Daly et al. 1996), MERLIN (Abecasis, Cherny et al. 2002), or SOLAR (Almasy and Blangero 1998). The ANALYZE software package provides qualitative trait two-point analysis; parametric linkage analysis, affected sib-pair analysis, and haplotype-based haplotype relative risk analysis. Multipoint linkage analysis is performed with GENEHUNTER or MERLIN. Quantitative trait linkage analysis (2-point or multipoint) is performed with MERLIN or SOLAR.

Each chromosome is analysed under a separate chromosomal subfolder, because this allows for simultaneous analysis of all chromosomes, thus avoiding overwriting of intermediate files. After all the analyses are completed, summary files are created and program-specific output files are copied in the same results folder. Furthermore, all converted input files and intermediate files are kept in subfolders for viewing detail analysis results or if manual re-analysis is required. For example, SOLAR writes computed IBD matrices into files and those can be reloaded (i.e. there is no need to compute IBD matrices if the genotype data does not change) when additional phenotypes are analyzed.

The main difference between MEGA2 and AUTOGSCAN is that the former is focused on providing support for multiple file formats and more flexible analysis options, while the latter is focused on completely automated genome-wide scan using fixed options with few commonly used software packages. AUTOGSCAN is intended for simple first pass analyses, and all assumed options may not be optimal for all data sets. However, AUTOGSCAN allows for easy initial genome-wide screens with multiple parametric models and multiple phenotypes. AUTOGSCAN accepts several command line options, and if multiple models or phenotypes are have to be used, this only requires one or two command line options before re-analysis. The full list of command line options is listed in the AUTOGSCAN documentation. More detailed analysis of specific genomic locations typically requires manual usage of additional programs which may not be automated by AUTOGSCAN. A detailed description of AUTOGSCAN can be found in the original publication (I).

Documentation, Unix-shell scripts, C/C++ source codes, and usage examples for the AUTOGSCAN can be found at the webpage:
http://www.helsinki.fi/~tsjuntun/autogscan/index.html.

4.2 PSEUDOMARKER software (II)

In study II, a likelihood-based method for joint linkage and/or linkage disequilibrium analysis (see Section 2.4.4.9), was implemented in the PSEUDOMARKER program. A key feature of the PSEUDOMARKER program is that it can combine singletons and pedigrees of varying structure into a single unified analysis. Other features are that it estimates a) marker allele frequencies, or b) marker allele frequencies conditionally on disease locus alleles, sometimes jointly with the recombination fractions when maximizing the likelihood of the data under a variety of hypotheses (Table 3). The likelihood calculating engine used by the PSEUDOMARKER program is a specially modified version of the ILINK program (Lathrop and Lalouel 1984; Lathrop, Lalouel et al. 1984; Lathrop, Lalouel et al. 1986) from the FASTLINK4.1P package (Cottingham, Idury et al. 1993; Schäffer, Gupta et al. 1994). FASTLINK uses the Elston-Stewart-algorithm (Elston and Stewart 1971) for traversing pedigrees, and therefore all family relationships are used correctly, marriage and consanguinity loops and missing phenotype or genotype data can be handled, and *theoretically* any size pedigree could be analyzed. Unrelated individuals are included in the analysis by creating pedigrees in which cases and controls are unrelated founders, with offspring of unknown phenotype and genotype.

Table 3. Likelihoods. Table modified from Study II.

Hypothesis	**Linkage**	**LD**	**Likelihood**	**Estimated parameters from the data**
H_0	no	no	$L_0 = L(\theta = 0.5, \mathbf{\delta} = 0)$	Marker allele frequencies
H_1	yes	no	$L_1 = \max_{\theta} L(\theta, \mathbf{\delta} = 0)$	Marker allele frequencies and recombination fraction
H_2	no	yes	$L_2 = \max_{\mathbf{\delta}} L(\theta = 0.5, \mathbf{\delta})$	Conditional marker allele frequencies
H_3	yes	yes	$L_3 = \max_{\theta, \mathbf{\delta}} L(\theta, \mathbf{\delta})$	Conditional marker allele frequencies and recombination fraction

The modified version of ILINK uses a direct-search optimization method (Torczon 1991) instead of the default GEMINI procedure (Lalouel 1979). The major benefit of using direct-search over GEMINI is that when estimation is performed over multiple dimensions jointly (i.e. conditional allele frequencies and recombination fraction jointly), GEMINI can get trapped in a local optimum when some parameter values are close to their boundary values. A more detailed description of the PSEUDOMARKER program work flow and of the modifications to the ILINK optimization procedure can be found in the original publication (II) and the references therein.

4.2.1 Pseudomarker locus assignment with penetrances

A recessive pseudomarker trait locus genotype assignment was shown in Figure 5. Equivalently, recessive pseudomarker analysis can be performed by using a rather extreme penetrance model: P(D)=0.00001 (or any other very small number), P(Affected | D/D)=0.00001, and P(Affected | D/+ or +/+) = 0. The disease locus genotypes inferred from this "affecteds-only" model are virtually identical to recessive pseudomarker genotype assignment (Terwilliger and Ott 1994; Göring and Terwilliger 2000), and we therefore use this approach in general, for simplicity, and to avoid inducing errors.

4.2.2 Statistical tests

After likelihoods are maximized under the four different hypotheses from Table 3, it is possible to perform several likelihood ratio tests (LRT). In Table 4, all the tests which the PSEUDOMARKER program performs are listed along with their analogous model-free tests. The linkage statistic (Λ), a test of LD allowing for linkage (Ψ), a test of LD assuming the absence of linkage (Υ), a test of linkage allowing for LD (ζ), and a joint test of linkage and LD (Ξ). A more detailed discussion of these tests can be found in Göring and Terwilliger (Göring and Terwilliger 2000). In general, the test of LD without linkage Υ, is not particularly meaningful in practice. Of particular interest is the statistic Ψ, with which one can test for LD in the presence of linkage after a significant linkage signal has been found.

Table 4. Likelihood test for linkage and/or LD can be performed with PSEUDOMARKER software.

Test statistics	Application	Analogous model-free test
$\Lambda = 2\ln\frac{L_1}{L_0} = 2\ln\frac{\max_{\theta} L(\theta,\delta=0)}{L(\theta=0.5,\delta=0)}$ *	Test of linkage (without LD)	Affected sib-pair tests, affected relative-pair tests
$\Psi = 2\ln\frac{L_3}{L_1} = 2\ln\frac{\max_{\theta,\delta} L(\theta,\delta)}{L(\theta,\delta=0)} \sim \chi^2_{n-1}$	Test of LD allowing for linkage	Haplotype relative risk (HRR) tests, case-control tests
$\Upsilon = 2\ln\frac{L_2}{L_0} = 2\ln\frac{\max_{\delta} L(\theta=0.5,\delta)}{L(\theta=0.5,\delta=0)} \sim \chi^2_{n-1}$	Test of LD (without linkage)	Haplotype independence test (HIND) **
$\zeta = 2\ln\frac{L_3}{L_2} = 2\ln\frac{\max_{\theta,\delta} L(\theta,\delta)}{L(\theta=0.5,\delta)} \sim \chi^2_1$	Test of linkage allowing for LD	Transmission/ disequilibrium tests (TDT)
$\Xi = \Lambda + \Psi = 2\ln\frac{L_3}{L_0} = 2\ln\frac{\max_{\theta,\delta} L(\theta,\delta)}{L(\theta=0.5,\delta=0)}$ ***	Joint test of linkage and LD	-

The n is the number of alleles at the marker locus.

* The linkage statistic Λ is distributed as 50-50 mixture of a point mass at 0 and $\chi_1{}^2$, because of the one-sided nature of the test (θ cannot be > 0.5).

** The haplotype independence test (HIND) was discussed by Terwilliger and Ott (Terwilliger and Ott 1992).

*** The joint test statistic Ξ is distributed as a 50-50 mixture of χ^2_n and χ^2_{n-1}, because for biological reasons, θ is restricted to be less than 0.5. Since $\Xi = \Lambda + \Psi$, this mixture distribution approximately applies.

4.2.3 Documentation

A web-based documentation, C/C++ source codes, and usage examples for the PSEUDOMARKER can be found at the webpage: http://www.helsinki.fi/~tsjuntun/pseudomarker/index.html.

4.3 Simulations (III, IV)

There are several recent published studies comparing family-based association methods (Howson, Barratt et al. 2005; Millstein, Siegmund et al. 2005; Jonasdottir, Humphreys et al. 2007; Nicodemus, Luna et al. 2007; Glaser and Holmans 2009; Infante-Rivard, Mirea et al. 2009; Callegaro, Lebrec et al. 2010), but they have exclusively focused on TDT-based methods and have not included the generally more powerful likelihood-based methods such as those implemented in LAMP, MENDEL, or PSEUDOMARKER.

In Study III, we evaluated the statistical properties of the most commonly used family-based association tests (see section 2.4.4). A total of 10 programs and 42 analysis options used in the simulation are enumerated in Table 5. The programs test types and relationship structures they can utilize can be found from the original publication (II) Tables 1 and 2. The simulation was performed under the following conditions: 1) no linkage and no association, 2) linkage and no association, and 3) linkage and association. Genotypes were simulated with (Fast)SLINK (Ott 1989; Weeks, Ott et al. 1990), in which the random number generator was modified to use three seeds instead of one. In addition, we generated data with the phenogenetic evolutionary simulator, ForSim (Lambert, Terwilliger et al. 2008), under an oligogenic model with multiple risk alleles per locus having varying effects on the phenotype.

Table 5. The programs and analysis options used in the simulations. Table modified from Study III.

Program	**Analysis option(s) used (with program specific abbreviations)**
FBAT	Robust variance estimator with dominant, recessive and additive models
TRANSMIT	One affected individual per nuclear family (one) • with robust estimator (one, ro) • with bootstrapping (one, bs) • without robust estimator or bootstrapping* (one) One nuclear family (nonuc) • with robust estimator (nonuc, ro) • with bootstrapping (nonuc, bs) • without robust estimator or bootstrapping* (nonuc) Multiple nuclear pedigrees (mf) • with robust estimator (mf, ro) • with bootstrapping (mf, bs) • without robust estimator or bootstrapping* (mf)
UNPHASED	Nelder & Mead's downhill-simplex method for likelihood maximization (-neldermead) • plain (no additional options) • with missing (missing) • with missing and parentrisk (missing,pr) • plain with controls (cc) • with missing and controls (missing,cc) • with missing, parentrisk, and controls (missing,pr,cc)
GENEHUNTER	Transmission/Disequilibrium Test (TDT)
PLINK	• Transmission/Disequilibrium Test (TDT) • TDT, parent of origin analysis (tdt,poo) • The parental discordance test (parentdt1) • The combined TDT and parental discordance test (parentdt2) • The sib-tdt (sibtdt)
QTDT	No additional options other than TDT test for the qualitative phenotype
HHRR**	Genotype and allele based HHRR • Genotype based, theoretical (GT) • Genotype based, randomized (GR) • Allele based, theoretical (AT) • Allele based, randomized (AR)
MENDEL	Gamete competition (A generalized version of TDT with full

	pedigrees) • Model 1 (gc1), where allele frequencies are fixed to some value (In these simulation analysis to obtain true generating values) • Model 2 (gc2), where allele frequencies are estimated from the data internally by MENDEL Association given linkage • Analysis model M_{Dom*} and the marker locus allele frequencies estimated from the data using MENDEL’s option 6 model 1 (M_{Dom*},fixed1)*** • Analysis model M_{Dom*} and the marker locus allele frequencies estimated from the data using MENDEL’s option 6 model 2 (M_{Dom*},fixed2) *** • Analysis model M_{Rec*} and the marker locus allele frequencies estimated from the data using MENDEL’s option 6 model 1 (M_{Rec*},fixed1) *** • Analysis model M_{Rec*} and the marker locus allele frequencies estimated from the data using MENDEL’s option 6 model 2 (M_{Rec*},fixed2) ***
LAMP	Dominant, recessive, additive, multiplicative and free. The ‘maxbits’ was set to 10[****].
PSEUDOMARKER	Dominant (M_{Dom}) and recessive (M_{Rec}).

[*] Bootstrapping was done with 100000 samples.

[**] We have used the likelihood based HHRR program

[***] Parametric models used were M_{Dom*}={P(D)=0.01; P(Aff|D/D)= P(Aff|D/+)=0.01; P(Aff|+/+)=0}, and M_{Rec*}={ P(D)=0.01; P(Aff|D/D)= 0.01;P(Aff|D/+)=P(Aff|++)=0}. NOTE: MENDEL had underflow problems when we used the M_{Dom} and M_{Rec} models, so we had to increase the penetrances and allele frequencies to 0.01. Option 6 model 1 estimates allele frequencies while preserving pedigree relationships, while model 2 treats all individuals as unrelated.

[****] The maxbits option was used to disregard large pedigrees from the analysis, because without this limitation it would have taken an average of one week to analyze one simulated replicate of our schizophrenia (or migraine) pedigrees. The bit complexity of a pedigree, which refers to the depth of the ‘gene-flow-trees’ (Abecasis, Cherny et al. 2002), is calculated by using the formula *2n-3f*, where *n* is the total number of individuals and *f* is the number of founders.

4.3.1 Pedigree data

The pedigree structures used in Studies II and III were taken from ongoing schizophrenia (Ekelund, Hovatta et al. 2001) and migraine (Wessman, Kallela et al. 2002; Kaunisto, Tikka et al. 2005) studies from Finland (Table 6). The phenotype in the migraine families was 'Migraine with Aura' and in the schizophrenia families was DSM-IV diagnosis. All other individuals without 'Migraine with Aura' or DSM-IV diagnosis were set as unknown. The availability of each individual for genotyping was determined based on a sample marker from each study. This information was used in simulations to allow for a realistic amount of missing genotype data. The schizophrenia families were mostly nuclear families with multiple affected offspring, while the migraine families were mostly multigenerational pedigrees with affected individuals across generations.

Table 6. Migraine and schizophrenia pedigree characteristics used in Studies II and III. Pedigree statistics were computed with PedStats (Wigginton and Abecasis 2005). Table modified from Study III.

		Migraine	Schizophrenia
Pedigrees		84	438
Individuals		1099	2535
Founders		366	914
Average pedigree size		13.08 (4 to 47)	5.79 (3 to 14)
Generations	**2**	10 (11.9%)	436 (99.5%)
	3	47 (56.0%)	2 (0.5%)
	4	27 (32.1%)	-
Affected individuals	**All**	398 (36.2%)	918 (36.2%)
	Founders	26 (7.1%)	60 (6.6%)
Genotyped individuals	**All**	810 (73.3%)	1906 (75.2%)
	Founders	147 (40.2%)	442 (48.4%)
Additional singletons	**Cases**	270	-
	Controls	884	199

4.3.2 No linkage and no association simulation

In the double null simulation, empirical type-I error rates were estimated for each program and analysis option when there was no linkage ($\theta=0.5$) and no association ($\delta=0$). The marker locus was assumed to be diallelic, with alleles 1 and 2, with minor allele frequency $p_1 = 0.1$. Empirical null distributions were estimated from 1,000 replicates of the schizophrenia and migraine datasets.

4.3.3 Linkage and no association simulation

Empirical type-I error rates were estimated when there was complete linkage (θ=0) and no association (δ=0). In order to simulate with linkage we had to assume a model for the trait locus, which in this case was chosen as follows: diallelic trait locus, with etiological model M_{Rec} (Table 7) in the schizophrenia families and M_{Dom} (Table 7) in the migraine families, and again a marker with a minor allele frequency of 0.1 as above. These extreme models were used to maximize the effects of linkage on the test statistics. Empirical null distributions (for tests of LD allowing for linkage) were estimated from 1,000 replicates of the schizophrenia and migraine dataset.

Table 7. Etiological models M_{Rec} and M_{Dom} used in type-I error simulations (θ=0 and δ=0).

Model	p_D	Penetrance		
		P(Affected|D/D)=$f_{D/D}$	P(Affected|D/+)=$f_{D/+}$	P(Affected|+/+)=$f_{+/+}$
M_{Rec}	0.00001	0.00001	0	0
M_{Dom}	0.00001	0.00001	0.00001	0

4.3.4 Linkage and no association simulation when parental genotypes are missing

In addition, the empirical type-I error rates for the tests of LD allowing for linkage were estimated as a function of the family size and number of missing genotypes in the parents. The pedigree structures in this analysis were: 200 triads (one affected child) and 100 sibships (two, three or four affected children). The number of sibships were reduced from 200 to 100 to save computational time, because it would have taken several months to compute empirical type-I error rates for the evaluated programs especially in case of a large sibship with unknown parent genotypes. Type-I error rates were estimated when parental genotypes were known, one parental genotype unknown, and both parental genotypes unknown. An additional 200 population controls were included in all simulations. Complete linkage was assumed between disease and marker loci (SNP marker with minor allele frequency of 0.1). The true generating model was M_{Rec}.

4.3.5 Power simulations

The power was estimated in both schizophrenia and migraine datasets, where complete linkage was assumed (θ=0) and where LD and etiological parameters were varied. We assumed diallelic disease and marker loci, with alleles (D, +) and (1, 2), respectively. The allele frequencies were $p_D = 0.1$ $(p_+ = 0.9)$ and $p_1 = 0.1$ $(p_2 = 0.9)$. In the first simulation, the genotype relative risk was varied when there was complete linkage and complete LD (D'=1) between disease and marker loci. The genotype relative risks are the ratio of the penetrances

$$RR_{D/D} = \frac{f_{D/D}}{f_{+/+}} \; ; \; RR_{D/+} = \frac{f_{D/+}}{f_{+/+}},$$

such that $f_{D/D} = RR_{D/D} f_{+/+}$ and $f_{D/+} = RR_{D/+} f_{+/+}$, where, assuming disease prevalence ϕ, the penetrance f_{++} is determined as follows (assuming HWE)

$$\phi = f_{D/D} p_D{}^2 + f_{D/+} 2p_D(1-p_D) + f_{+/+}(1-p_D)^2 \Rightarrow$$

$$\phi = RR_{D/D} f_{+/+} p_D{}^2 + RR_{D/+} f_{+/+} 2p_D(1-p_D) + f_{+/+}(1-p_D)^2 \Rightarrow$$

$$\phi = f_{+/+}\left[RR_{D/D} p_D{}^2 + RR_{D/+} 2p_D(1-p_D) + (1-p_D)^2\right] \Rightarrow$$

$$f_{+/+} = \frac{\phi}{RR_{D/D} p_D{}^2 + RR_{D/+} 2p_D(1-p_D) + (1-p_D)^2}.$$

In the second simulation, the LD parameter D' was varied between 0 ≤ D' ≤ 1, while the genotype relative risk was fixed (see below) and complete linkage between the disease and marker loci was assumed. The conditional allele frequency $p_{1|D}$ was varied from 0.1 (D'=0) to 1.0 (D'=1). Then, the conditional allele frequency $p_{1|+}$ was set to

$$p_{1|+} = \frac{p_1 - p_{1|D} p_D}{1 - p_D}.$$

Haplotype frequencies can be computed from the conditional allele frequencies as

$$\begin{aligned} p_{\underline{1D}} &= p_{1|D} p_D, \\ p_{\underline{1+}} &= p_{1|+} p_+, \\ p_{\underline{2D}} &= p_{2|D} p_D = (1 - p_{1|D}) p_D, \\ p_{\underline{2+}} &= p_{2|+} p_+ = (1 - p_{1|+}) p_+. \end{aligned}$$

The corresponding D' values were computed using the equations in section 2.4.1.

4.3.6 Genotype relative risk scan

In the genotype relative risk scan the disease and marker loci were assumed to be in complete linkage (θ=0) and in complete LD (D'=1 and $r^2 = 1$) (i.e. the marker is the functional variant). The genotype relative risk was varied from 1 to 6 under a recessive model in the schizophrenia pedigrees and from 1 to 2 under a dominant model in the migraine pedigrees. The rationale for selecting a recessive model for schizophrenia and a dominant model for migraine was based on pedigree structures and observed phenotypes, which suggested such models of inheritance.

4.3.7 D' scan

In the D' scan the disease locus and the marker locus were assumed to be in complete linkage (θ=0), while D' was varied as described above. In the schizophrenia dataset the genotype relative risk was fixed to 6 under a recessive model and in the migraine dataset it was fixed to 2 under a dominant model.

4.3.8 Complex multifactorial trait simulation with ForSim

ForSim (Lambert, Terwilliger et al. 2008) was used to simulate a complex multifactorial trait in an entire population, over an evolutionary timeframe. In the simulation there were five chromosomes, each containing three etiological loci, where one locus on each chromosome had multiple variants contributing to the disease phenotype additively. The population was simulated over 10,000 generations, where hundreds of variants emerged by mutation and were subjected to natural selection on the resulting phenotype. The disease prevalence in the last generation was 9%. This simulation generated 10,000 multigenerational pedigrees and 1,000 random control individuals with thousands of SNP markers. One functional SNP was selected for the analysis, which showed strongest linkage and association in the population. The power of each analysis method was estimated using a set of randomly sub-sampled pedigrees. A detailed description of the simulation can be found in the original publication (III).

4.3.9 Comparison of different ascertainment strategies

In Study III we compared four different ascertainment schemas with fixed sample size; 2000 cases and 2000 controls, 1000 triads and 1000 controls, 800 sib pairs and 800 controls, and 667 sib-trios and 665 controls. Each sample contained a total of

4000 individuals. The hypothesis was: complete linkage, complete LD, recessive genotype relative risk of 4, disease prevalence of 10%, and $p_D = p_1 = 0.1$. PSEUDOMARKER's recessive LD given linkage test was used to estimate power over all sampling schemas.

4.3.10 Additional controls

We investigated the effect of adding population controls to the LD analysis conditional on linkage in the schizophrenia and migraine datasets. We assumed complete linkage, complete LD, and fixed relative risks as in the D' scan (See Section 4.3.7). The allele frequencies were as described in section 4.3.5. Various numbers of population controls were added and the power to detect association conditional on linkage was compared.

4.3.11 Parametric linkage analysis under true and inaccurate models

In Study IV we compared the power of parametric linkage analysis under the true generating model with M_{rec} in 800 fully genotyped affected sib pairs (with unaffected parents). We assumed complete linkage (θ=0), marker minor allele frequency of 10%, disease allele frequency of 10%, and disease prevalence of 5%. The same comparison was performed using the schizophrenia dataset, except the disease prevalence was 1%.

We computed the expected maximum lod score, $E[Z_{\max}]$, under true model and M_{Rec} models in both datasets with MLINK (Lathrop and Lalouel 1984; Lathrop, Lalouel et al. 1985; Lathrop, Lalouel et al. 1986) program from the FASTLINK 4.1P package (Cottingham, Idury et al. 1993; Schäffer, Gupta et al. 1994) for each replicate, and the expected maximum lod score was computed as

$$E\left[Z_{\max}\right] = \frac{1}{N}\sum_{i=1}^{N} \max_{\theta} Z_i(\theta),$$

where N is number of replicates. Then we compared the expected maximum lod scores by using their ratio of $E_{M_{\mathrm{Rec}}}\left[Z_{\max}\right] / E_{True}\left[Z_{\max}\right]$, which measures the relative gain in the expected lod score statistics, when M_{Rec} model is used, compared to the true generating model is used.

4.3.12 Two-point linkage analysis vs. multipoint linkage analysis

We used ForSim-generated data to compare the power of parametric two-point linkage analysis and multipoint linkage analysis when a set of loci were in both linkage and LD with a functional variant. The pedigree sample for linkage analysis was sampled without replacement from the set of pedigrees we previously simulated with ForSim (see above). All individuals were genotyped at all marker loci. We selected 11 SNPs from an 89kb region containing functional variants simulated in 722 individuals from 79 three-generational pedigrees. The functional variant had relatively large effect, genotype relative risk of 10. The minor allele frequency of the functional variant in the sample was 0.2, and the population prevalence of the disease was 0.212.

In addition, we performed a similar comparison using the migraine dataset, in the presence of incomplete LD between the functional variant and a marker locus. A dominant model for the disease and parameters of the SNP marker were as described in Section 4.3.7. We also simulated a highly polymorphic microsatellite marker with 10 alleles, each having allele frequency of 0.1, at the same genomic location (θ=0) with the SNP marker.

4.4 Maximum likelihood estimates of parameters in a test of LD conditional on linkage (IV)

In Study IV, we investigated statistical properties of likelihood ratio tests of LD conditional on linkage under various true (M_{True}) and analysis ($M_{Analysis}$) model combinations. To examine the maximum likelihood estimates (MLEs) of parameters under each model, the BIAS program was written, which computes the expected log-likelihood of the data as a function of θ, $p_{1|D}$, $p_{1|+}$, and the true and analysis models as follows

$$E\left[\ln L\left(\theta, p_{1|D}, p_{1|+}\right)\right] = \sum_{\mathbf{G_D}} \sum_{\mathbf{G_M}} P(\mathbf{G_M}, \mathbf{G_D}; M_{True}) \ln L\left(\theta, p_{1|D}, p_{1|+}; M_{Analysis}\right).$$

The BIAS program uses a downhill simplex method (Nelder and Mead 1965) for maximization of the likelihood.

In order to evaluate the properties of the conditional test of LD given linkage, we assumed complete linkage and no LD under some true model with parameters θ=0 and $p_{1|D} = p_{1|+} = 0.1$ (i.e. marker allele frequency of 0.1 and no LD). The true and analysis disease models all permutations of the models M_{Rec} and M_{Dom} from Table 7.

The expected profile log-likelihoods were computed in a single sib pair and a single sib-trio treating the recombination fraction as a nuisance parameter as

$$E\left[\ln L\left(p_{1|D}, p_{1|+}\right)\right] = \max_\theta E\left[\ln L\left(\theta, p_{1|D}, p_{1|+}\right)\right].$$

The profile log-likelihoods were computed as a function of $p_{1|D}$ and $p_{1|+}$, where $0 < \left(p_{1|D}, p_{1|+}\right) < 1$.

Additionally, we investigated the properties of the MLEs of the parameters under the analysis model M$_{\mathrm{Dom}}$, when the M$_{\mathrm{True}}$ models were dominant $(RR_{D/+} = RR_{D/D})$, recessive $(RR_{D/+} = 1)$, additive $(RR_{D/+} = 0.5 \times RR_{D/D})$ and multiplicative $(RR_{D/+} = \sqrt{RR_{D/D}})$, assuming the following parameters: $p_D = 0.05$, θ=0, and $\phi = 0.01$. The genotype relative risk, $RR_{D/D}$, varied between 1 and 50.

5 Results and their evaluation

5.1 Applications of AUTOGSCAN software (I)

Before AUTOGSCAN was published, it was extensively tested by our research scientists at the National Institute of Public Health (KTL), Helsinki, Finland, later, the National Institute for Health and Welfare, Helsinki, Finland. Automated analysis has enabled efficient data analysis of numerous large genome-wide scans by students and researchers themselves. Furthermore, the valuable time of trained statistical geneticists can now be used for interpretation of results, rather than running repetitive analyses.

For example, an early version of AUTOGSCAN was used by Paunio et al. (Paunio, Tuulio-Henriksson et al. 2004), in which SOLAR was used for variance component linkage analysis. Since AUTOGSCAN was published it has been used in many published genome-wide scans. Automated ANALYZE was utilized in (Al-Yahyaee, Al-Gazali et al. 2006; Rehnstrom, Ylisaukko-oja et al. 2006; Ylisaukko-oja, Alarcon et al. 2006; Turunen, Rehnstrom et al. 2008; Wider, Melquist et al. 2008; Tikka-Kleemola, Artto et al. 2010; Polvi, Siren et al. 2012).

Automated MERLIN was utilized in (Sammalisto, Hiekkalinna et al. 2005; Knaapila, Keskitalo et al. 2007; Perola, Sammalisto et al. 2007; Magnusson, Boman et al. 2008; Wedenoja, Loukola et al. 2008; Kettunen, Perola et al. 2009; Sammalisto, Hiekkalinna et al. 2009; Haataja, Karjalainen et al. 2011; Kantojarvi, Kotala et al. 2011), and both automated ANALYZE and automated MERLIN were utilized in (Loukola, Broms et al. 2008; Wessman, Forsblom et al. 2011).

Even though AUTOGSCAN is intended for performing simple quick-and-dirty genome-wide analysis, it allows for a flexible selection of subsets of chromosomes to be analyzed. This option was applied in a large genome-wide linkage scan by Sammalisto et al. (Sammalisto, Hiekkalinna et al. 2005), where the resulting analysis was performed in a 'parallel' sense, with each chromosome was simultaneously analyzed on a separate individual processor to dramatically reduce the analysis time. While today, such approaches are in common practice, at that time it was novel. In conclusion, AUTOGSCAN has proven to be very useful tool for research scientists.

5.2 Empirical type I error rates (III)

The empirical type-I error rates were estimated for 0.01 and 0.05 significance levels in both schizophrenia and migraine data sets using 1,000 replicates. The selection of the α-level is in general more or less arbitrary, however estimation of the much smaller type-I error rates normally used in gene mapping would have required an extremely large number of replicates. Furthermore, in our case, several months of computing time would have been needed to evaluate all the programs and options we considered. As our goal was to identify analysis methods that were grossly invalid, we applied these less-stringent criteria, and already identified several program/analysis option combinations which gave grossly anticonservative results, obviating the need to explore deeper into the tail of their distributions. The results under the hypothesis of no linkage and no association, and for the hypothesis of complete linkage and no association are given in the original publication (III) Supplementary Tables 2-5. Figure 7 shows the empirical type-I error rates when no linkage and no association are assumed and Figure 8 show the empirical type-I error rates for the schizophrenia and migraine data sets, when each program's author's recommended options were used in testing for association in the presence of linkage when there is missing parental genotype data. When the program allowed for parametric analysis, the simulation model-type was used (recessive or dominant).

5.2.1 No linkage and no association

Under the null hypothesis of no linkage and no association all the programs provided valid tests except LAMP (See the original publication (III) Supplementary Tables 2-3), which consistently showed significantly inflated type-I error rates (Figure 7). LAMP's analysis models were all anticonservative with the most anticonservative option being 'free', which had a type-I error rate of 0.17 on schizophrenia pedigrees and 0.19 on migraine pedigrees at the 0.05 significance level.

LAMP estimates parameters of the etiological model under constraints concerning disease prevalence and model type (recessive, dominant, multiplicative, etc.) jointly with linkage and LD, and it assumes complete linkage ($\theta=0$) between trait and marker loci. This leaves many nonorthogonal nuisance parameters to be estimated both under the null (penetrances and allele frequencies of disease and marker loci) and alternative (penetrances and haplotype frequencies of disease and marker loci) hypotheses. Furthermore, due to the complexity of the parameter constraints, and the nonorthogonality of the several parameters being estimated, the regularity conditions of Wilks' theorem do not apply, such that significance testing in LAMP is based on a simulated null distribution assuming no association, rather than some approximate mixture of χ^2 distributions. The null distribution of the test statistic given complete linkage and no association is simulated in LAMP by using the MLEs of the parameters under the hypothesis of complete linkage and no association. Then the observed test statistic is compared with the simulated null distribution (Li, Boehnke et al. 2006). However, these estimated parameters are not the "true state of nature", and estimates of the same parameters under the alternative hypothesis can be dramatically different. For all of these reasons, the anticonservative nature of the test statistic in LAMP is not particularly surprising.

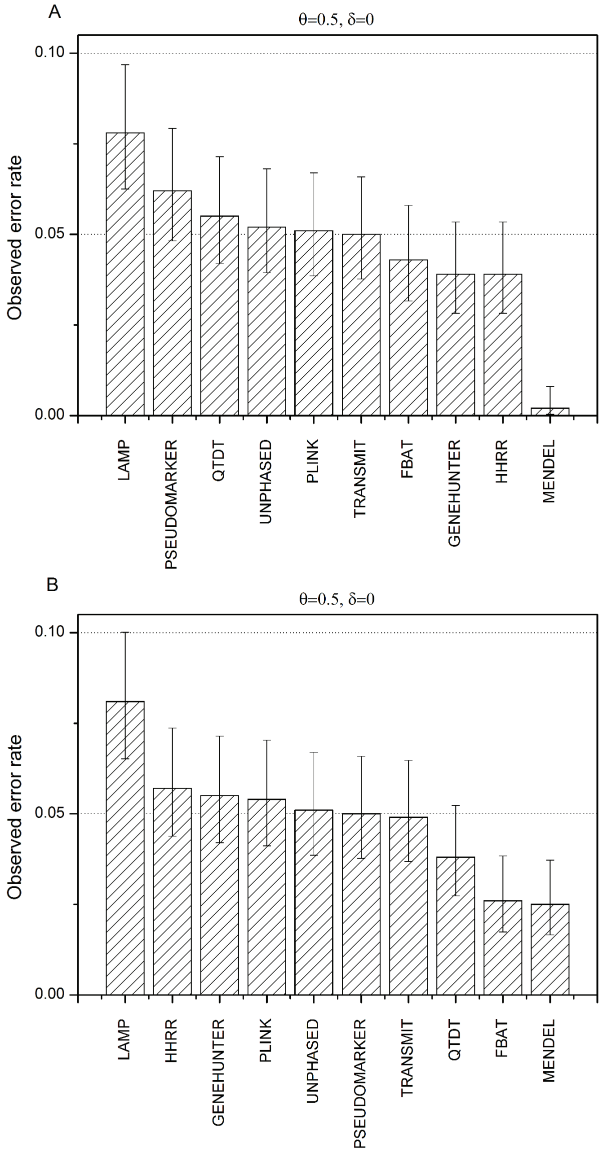

Figure 7. Empirical error rates from the no linkage no LD simulation, (A) the schizophrenia data set, and (B) the migraine data set at 0.05 significance level with 95% confidence intervals. The following analysis options were used: FBAT [(A) recessive and (B) dominant], PSEUDOMARKER [(A) recessive and (B) dominant, LD given linkage), GENEHUNTER (TDT), PLINK (sib-TDT), HHRR (AR), MENDEL [(A) $M_{Rec*,fixed1}$, (B) $M_{Dom*,fixed1}$, association given linkage), QTDT , TRANSMIT (mf, ro), LAMP [(A) recessive (B) dominant], and UNPHASED (missing, pr, cc). The results are based on 1,000 replicates.

5.2.2 Complete linkage and no association

Under the null hypothesis of complete linkage and no association, GENEHUNTER TDT, PLINK (TDT-based options), MENDEL (gamete competition option) had some power to detect linkage, (See the original publication (III) Supplementary Tables 4-5), even though there was no association, consistent with the null hypothesis of those tests being that of no linkage, not of no association. Furthermore, this means that significant test statistics from those methods imply nothing about the existence of any association.

In Figure 8 the empirical type-I error rates are presented for programs which allow testing for LD in the presence of linkage. The TDT-type methods which claim to test for LD in the presence of linkage, such as TRANSMIT ('robust estimator'-option) and UNPHASED ('parentrisk'-option), showed inflated type-I error rates, both for schizophrenia (Figure 8A) and migraine (Figure 8B) data sets, however these error rates were somewhat lower on the multigenerational migraine dataset than the nuclear family dataset with schizophrenia. UNPHASED had an enormous error rate of 0.498 on the schizophrenia dataset (Figure 8A), even when options for missing data and inclusion of controls were used, and the presence of linkage was allowed for in the analysis, as per the author's instructions (Dudbridge, personal communication). FBAT's test for LD in the presence of linkage (robust variance estimator) and HHRR were valid in all tests, as were the likelihood-based methods in PSEUDOMARKER and MENDEL (association given linkage).

Our results confirm the obvious, that it is not suitable to use TDT-based methods [GENEHUNTET TDT, PLINK (TDT-based options), MENDEL (gamete competition), and some options of TRANSMIT and UNPHASED] for testing the null hypothesis of no association in datasets where there are families with multiple affected individuals, unless one subsamples a single affected individual and his parents and censors the rest.

Type-I error rates were elevated, even when UNPHASED's ('parentrisk'-option) and TRANSMIT's ('robust estimator'-option) options controlling for linkage were used. The reason for UNPHASED's and TRANSMIT's relatively less anti-conservative performance in the migraine families are because there were relatively more nuclear families with genotyped parents (in large pedigrees), than in the schizophrenia dataset, which consisted solely of nuclear families. FBAT was the only exception among the set of TDT-type methods we compared, with type-I error rates at the expected level, meaning that the correction for linkage implemented in

FBAT provides a valid test, as advertised. Perhaps the most interesting result was with UNPHASED, when additional controls were added to the analysis. It appears that UNPHASED treats affected individuals with two missing parents (phenotypes and genotypes unknown) differently from random affected individuals drawn from the population in the analysis, even though those are fundamentally identical sampling units, because everyone has parents regardless of whether they have been genotyped or not. LAMP's elevated type-I error rates show that estimation of the disease model conditional on the disease prevalence and the significance testing based on simulated null distributions is not really adequate.

PSEUDOMARKER (LD given linkage) and MENDEL (association given linkage) were valid in both simulations and in both datasets, because those methods use all available data and can account for linkage in full pedigrees. Although MENDEL had lower type-I error rates than expected, it may be a result of how the conditional allele frequencies are estimated (i.e. MENDEL adds more constraints on the values of the nuisance parameters than PSEUDOMARKER). The HHRR was valid, because when there is no LD, "case" and non-transmitted "control" genotypes are independent (Terwilliger and Ott 1992).

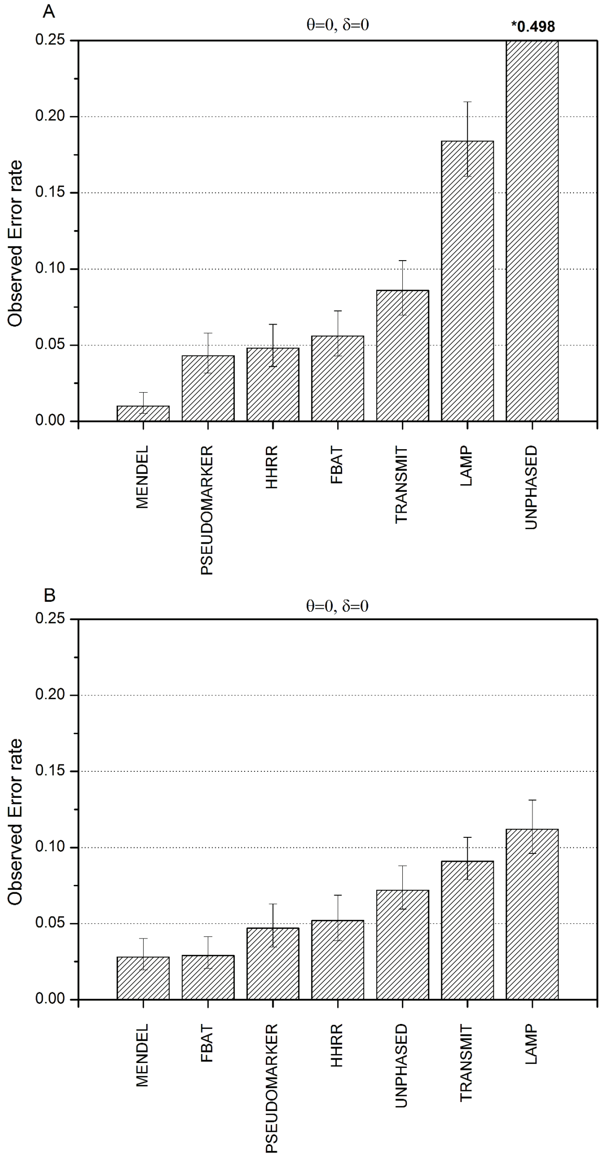

Figure 8. Empirical error rates from the complete linkage no LD simulation, (A) the schizophrenia data set, and (B) the migraine data set at 0.05 significance level with 95% confidence intervals. The following analysis options (LD tests conditional on linkage) were used: FBAT [(A) recessive and (B) dominant], PSEUDOMARKER [(A) recessive and (B) dominant, LD given linkage), HHRR (AR), MENDEL [(A) $M_{Rec^*,fixed1}$, (B) $M_{Dom^*,fixed1}$, association given linkage), TRANSMIT (mf, ro), LAMP [(A) recessive (B) dominant], and UNPHASED (missing, pr, cc). The results are based on 1,000 replicates.

5.2.3 Complete linkage and no association: effect of missing parental genotypes

In an additional set of simulations of complete linkage and no LD, we considered nuclear families (triads, sibships and additional controls), with various proportions of missing parental genotypes. In the presence of such missing data, several programs for testing LD in the presence of linkage showed excessive type-I error rates and some programs could not analyze the data at all when parental genotypes were missing (See Supplementary Table 6 of the original publication (III)).

The TDT-type test is a test of linkage which is only powerful in the presence of LD, such as GENEHUNTER TDT, MENDEL (gamete competition), TRANSMIT (without ‘robust estimator’-option), UNPHASED (without ‘parentrisk’-option), and PLINK (TDT-options) did show some power to detect linkage when all individuals were genotyped and sibship size was increasing. Furthermore, interestingly, power to detect linkage increased as missing data increased (one parent not genotyped and both parents not genotyped) for MENDEL (gamete competition), TRANSMIT (one affected per nuclear pedigree) and UNPHASED (‘missing’-option and controls), indicating potential problems with the algorithms for handling missing data.

TRANSMIT’s and UNPHASED’s options for testing LD in the presence of linkage showed highly elevated type-I error rates. For example, in affected sibships with both parental genotypes missing TRANSMIT (‘robust estimator’-option), and UNPHASED (‘parent risk’-option) had type-I error rates of 100% at the 0.05 significance level (See Supplementary Table 6 of the original publication (III)). FBAT, HHRR, LAMP, and MENDEL (association given linkage) were valid in all tests. PSEUDOMARKER’s test “recessive LD given linkage” showed an elevated type-I error rate of 0.1 on triads (parental genotypes known). This elevated type-I error rate arose from the fact that the null hypothesis likelihood is not a function of the recombination fraction while under the alternative hypothesis, it is leading to a well-known violation of the regularity conditions of Wilks’ theorem (see (Davies 1977)). This problem and its solutions are discussed in detail in Section 5.4 below.

When there is no parental genotype information available, there is no data for methods like the classic TDT (such as GENEHUNTER TDT and PLINK’s TDT-options) to analyze. TRANSMIT and UNPHASED overcome the missing genotype problem by integrating over all possible complete data configurations consistent with the known genotypes, and type-I error-rates from those programs were at expected levels when the dataset consisted solely of triads and singletons. MENDEL (gamete competition) was also valid for triad data. However, when there were multiple affected offspring and parental genotypes unknown, those tests are no longer valid, as shown above in the complete linkage and no association simulations. Interestingly LAMP was valid for all tests and the reason might be related to the specific relationship structures being analyzed. The main difference between sibships and the schizophrenia dataset are that in the schizophrenia dataset we only have phenotypic information for definitively affected individuals (36.2%) and all others are phenotypically unknown. Additionally, there is a mixture of family structures, and a mixed assortment of missing data, where in these analyses the entire dataset consisted of independent and identically distributed family structures. The LAMP association test (conditional on linkage) was valid when the data is 'idealistic', where all individuals are phenotyped, however that is never the case in real-life data sets.

5.3 Power (III)

Only those programs and options that provided valid type-I error rates (at α=0.05 significance level) for tests of LD given linkage were included in the power analyses. Therefore classic TDT-based methods GENEHUNTER TDT and PLINK (TDT-options) were omitted from further consideration, along with LAMP. Some options of TRANSMIT and UNPHASED programs had valid type-I error rates under the hypothesis of complete linkage and no association, so those options which had the least anticonservative empirical type-I error rates were used in the power analyses (See Supplementary Tables 4-6 from the original publication (III)).

In the schizophrenia dataset, TRANSMIT's option for selecting one triad (one) from each pedigree was used (the empirical type-I error rate was 0.057) and UNPHASED's option including controls (plain,cc) (the empirical type-I error rate was 0.056) but not the option allowing for missing data was used, because the missing data option had an anticonservative type-I error rate of 0.29. In the migraine dataset, TRANSMIT's option for selecting one nuclear family from a pedigree (nonuc) using the robust variance estimator (ro) was used (the empirical type-I error rate was 0.067) and UNPHASED's option including controls (plain,cc) without allowing for missing data (the empirical type-I error rate was 0.066). Power was estimated at the α=0.0001 significance level, because in genome-wide linkage analysis, a pointwise significance level of 0.0001 (which correspond to a lod score of 3) is approximately equivalent to a genome-wide significance level of 0.05. The choice of α–level should be even lower for genome-wide association scans, where large numbers of marker are studied.

5.3.1 Genotype relative risk and D' scans

In the genotype relative risk scan using the schizophrenia dataset, PSEUDOMARKER was the most powerful followed by FBAT, HHRR, MENDEL (association given linkage), TRANSMIT, QTDT, and UNPHASED (Figure 9). In the multigenerational migraine dataset (Figure 10), TDT-based tests (UNPHASED, TRANSMIT, FBAT, and QTDT) had a striking reduction in power, with FBAT and QTDT having almost no power at all. These TDT-based methods are unable to utilize multigenerational pedigrees correctly, and in the migraine dataset there are fewer informative nuclear families (both parents genotyped) than in the schizophrenia dataset. Furthermore, FBAT and QTDT require both parents to be genotyped and heterozygous, and those programs do not handle missing data at all. The most powerful approach, once again, was PSEUDOMARKER, followed by HHRR and MENDEL (association given linkage).

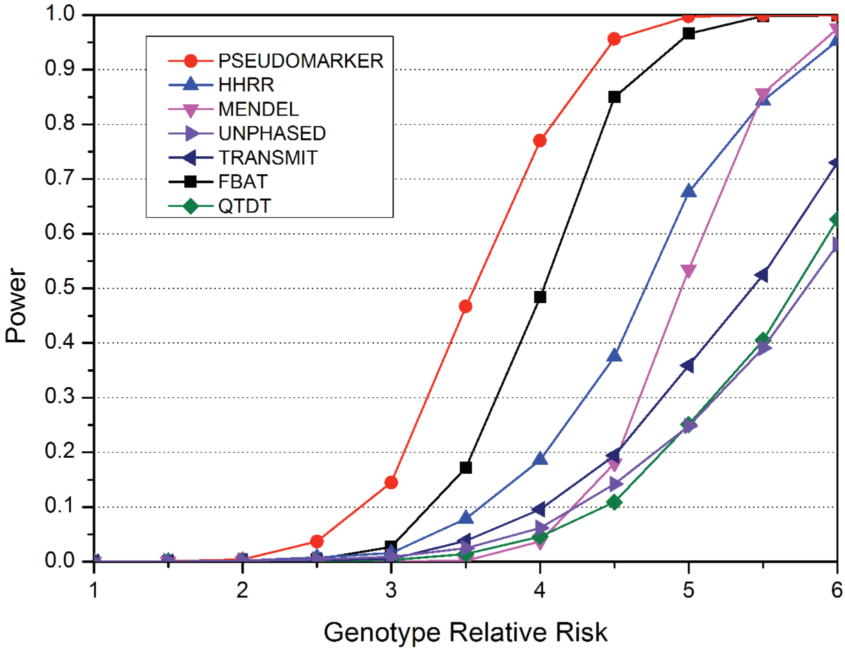

Figure 9. Power of each program to detect allelic association in the presence of linkage at significance level α=0.0001 in the schizophrenia dataset, as a function of the genotype relative risk. The disease prevalence was 1%, θ=0, p_D=0.1, and the risk allele itself was genotyped. The following analysis options were used: FBAT (recessive model), PSEUDOMARKER (recessive LD given linkage), HHRR (AR), MENDEL ($M_{Rec^*,fixed1}$, association given linkage), QTDT, TRANSMIT (one), and UNPHASED (plain, cc). The results are based on 1,000 replicates. The gigure modified from Study III.

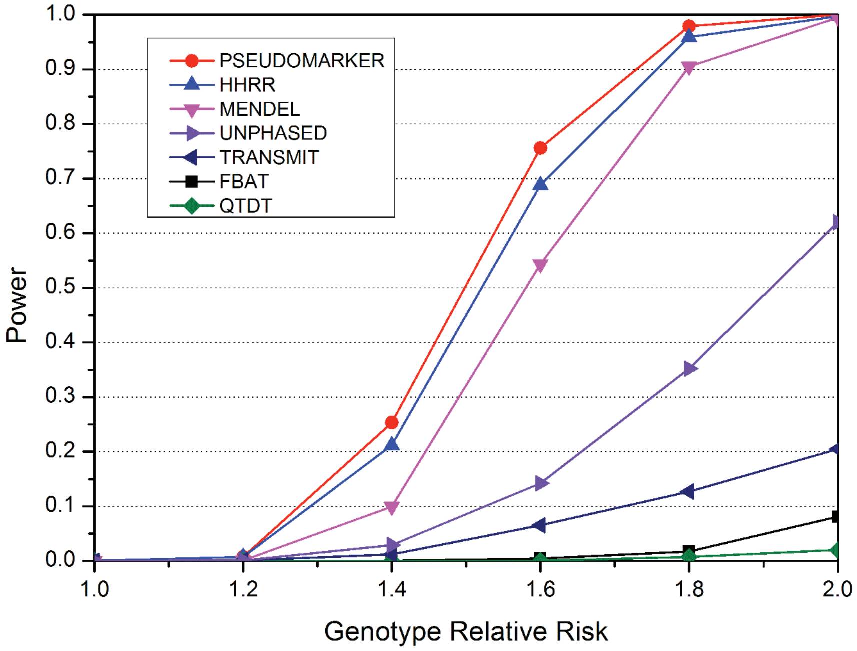

Figure 10. Power of each program to detect allelic association in the presence of linkage at significance level α=0.0001 in the migraine dataset, as a function of the genotype relative risk. The disease prevalence was 10%, θ=0, p_D=0.1, and the risk allele itself was genotyped. The following analysis options were used: FBAT (dominant model), PSEUDOMARKER (dominant LD given linkage), HHRR (AR), MENDEL ($M_{Dom^*,\ fixed1}$, association given linkage), QTDT, TRANSMIT (nonuc,ro), and UNPHASED (plain,cc). The results are based on 1,000 replicates. The figure modified from Study III.

In the D' scan (Figures 11 and 12), a similar trend was observed in relative power across programs. HHRR was surprisingly powerful because it can; a) include singletons in the analysis, because for example, unaffected healthy control individuals provide information about the allele frequencies, and b) use homozygous parents (all the data is used for estimating the allele frequencies). Thus, when there is complete linkage and complete LD and all individuals in a pedigree are genotyped, genotyping an additional affected non-founder individual does not add significant new information about association, since under a dominant model the disease allele most likely enters each pedigree only once. Therefore, most of the information can be obtained from a single triad, making HHRR surprisingly powerful.

There are several reasons why PSEUDOMARKER is the most powerful: a) all available data are used, including unrelated cases and controls, b) pedigree relationships in families are modelled correctly, c) allele frequencies, conditional allele frequencies, and recombination fractions are estimated directly from the data, and d) missing genotype data can be integrated over accurately. The tests based on TDT were less powerful and possible reasons for that are a) large extended pedigrees are decomposed into nuclear families and treated incorrectly as independent, b) they cannot adjust properly for missing data or cannot use families at all, when parents are missing (such as classic TDT).

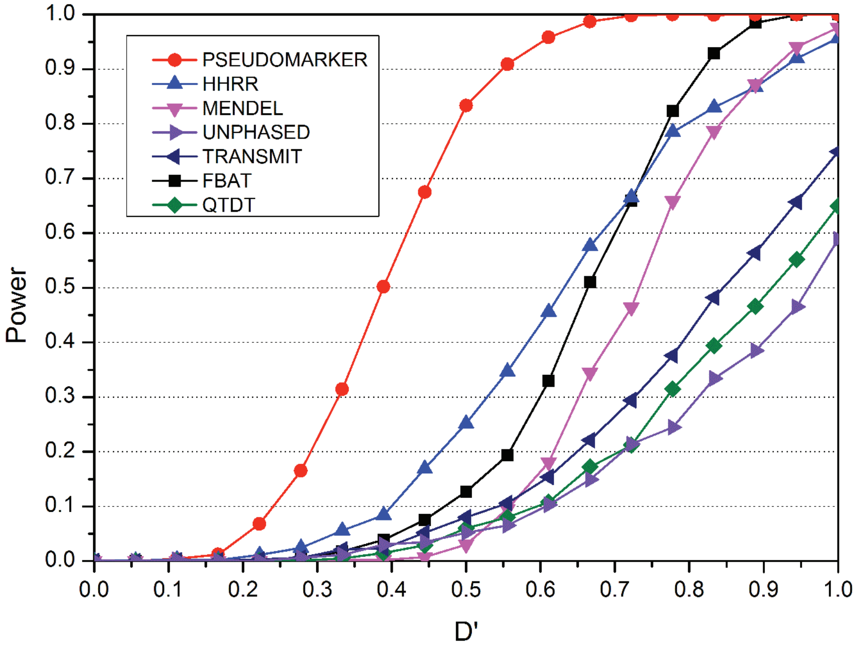

Figure 11. Power of each program to detect allelic association in the presence of linkage at significance level α=0.0001 in the schizophrenia dataset, as a function of the strength of LD between marker and disease loci. The disease prevalence was 1%, θ=0, recessive genotype relative risk was fixed to 6, and $p_D = p_1 = 0.1$. The following analysis options were used: FBAT (recessive model), PSEUDOMARKER (recessive LD given linkage), HHRR (allele-based randomized), MENDEL ($M_{Rec^*,fixed1}$, association given linkage), QTDT, TRANSMIT (nonuc,ro), and UNPHASED (plain, cc). The results are based on 1,000 replicates. Figure modified from the Study III.

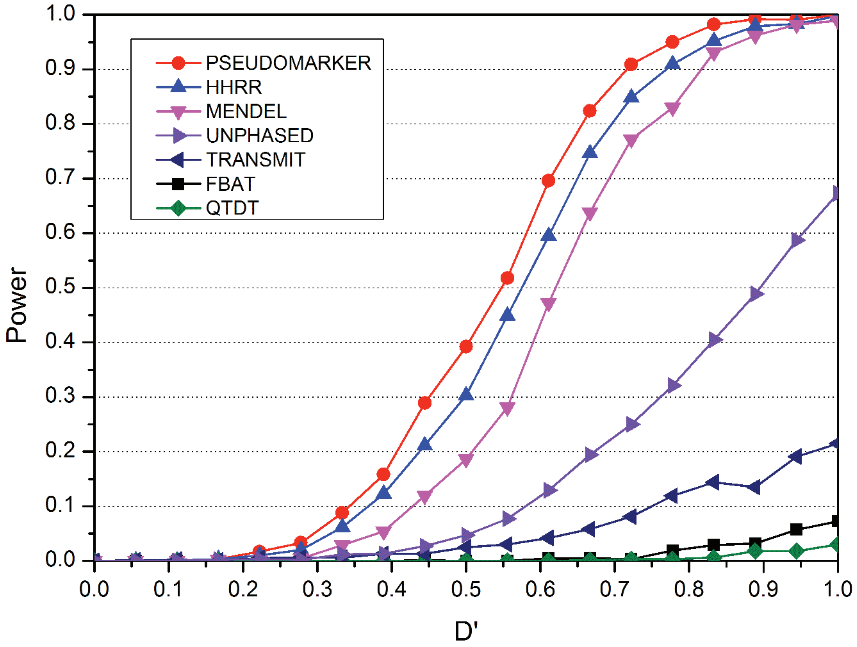

Figure 12. Power of each program to detect allelic association in the presence of linkage at significance level α=0.0001 in the migraine dataset, as a function of the strength of LD between marker and disease loci. The disease prevalence was 10%, θ=0, dominant genotype relative risk was fixed to 2, and $p_D = p_1 = 0.1$. The following analysis options were used: FBAT (dominant model), PSEUDOMARKER (dominant LD given linkage), HHRR (AR), MENDEL ($M_{Dom*, fixed1}$, association given linkage), QTDT, TRANSMIT (one), and UNPHASED (plain,cc). The results are based on 1,000 replicates. The figure modified from Study III.

5.3.2 Complex multifactorial trait simulation with ForSim

The power analysis results in the ForSim-generated data (see Section 4.3.8) were similar to those from the migraine pedigrees (See Figure 3 from the original publication (III)). PSEUDOMARKER was the most powerful followed by HHRR, MENDEL (association given linkage), TRANSMIT, FBAT, QTDT, and UNPHASED.

This shows that our observations made under a single functional variant model on migraine data hold also under more complex models with multiple functional variants in multiple genes influencing the trait. It is likely that real-life diseases are caused by multiple functional variants across many different genes with a variety of effect sizes. ForSim enabled us to simulate phenotype-based natural selection in an entire population over many thousands of generations, in which multiple etiological loci with hundreds of functional variants arise by mutation, and were subjected to natural selection on the resulting phenotypes in each generation. In this model, all variants were assumed to influence the trait in an additive fashion. Pedigrees for the analysis were selected from the last generations of this population-based simulation. This is oversimplified, but yet much more realistic than the single functional variant models, the fact that the results were essentially the same reassures us that our conclusions are generalizable.

5.3.3 Comparison of different sampling schemas

In a comparison of four different ascertainment schemes (case-control, triads, affected sib pairs, and affected sib-trios) (see Section 4.3.9) with fixed number of genotyped samples, the lowest power per genotype came from the sample of triads as expected (Figure 13), because one triad (father, mother and affected offspring, total of three genotypes) contributes one "case" genotype (i.e. transmitted) and one "cohort" genotype (i.e. non-transmitted) to the analysis. The power of a case-control study was higher, because twice as many affected individuals were included in the study, with 50% more of the "controls" having been screened for not having the disease compared with non-transmitted alleles in the triad design. The change in the power from a case-control design to an affected sib pair and to an affected sib-trio design was enormous. On average, offspring in a sib pair share 50% of their genomes IBD, increasing the chance that they are sharing a phenotype because of shared genetic exposures rather than environmental ones (i.e. the detectance at the disease locus genotype is increased). With three affected sibs in a single family, the detectance is increased even more, as three individuals sharing a rare disease phenotype by chance in the same family is very small. In affected sib-trios, therefore, even more genetic risk factors are likely to have been transmitted from parents to offspring. The key element behind increased power is this increased detectance for the disease locus genotype (D/D). This is discussed in detail in Section 5.8 below.

Similar results have been previously reported by Risch and Merikangas (Risch and Merikangas 1996), who demonstrated that association mapping in sib pairs requires far fewer genotyped individuals than in triads. However, the paper by Risch and Merikangas is often misinterpreted as having argued for GWAS in unrelated individuals as optimal, while they really pointed out that this was not true and furthermore, rather argued that the increased sample size required when not analyzing multiplex families was not that large, and therefore GWAS might be feasible.

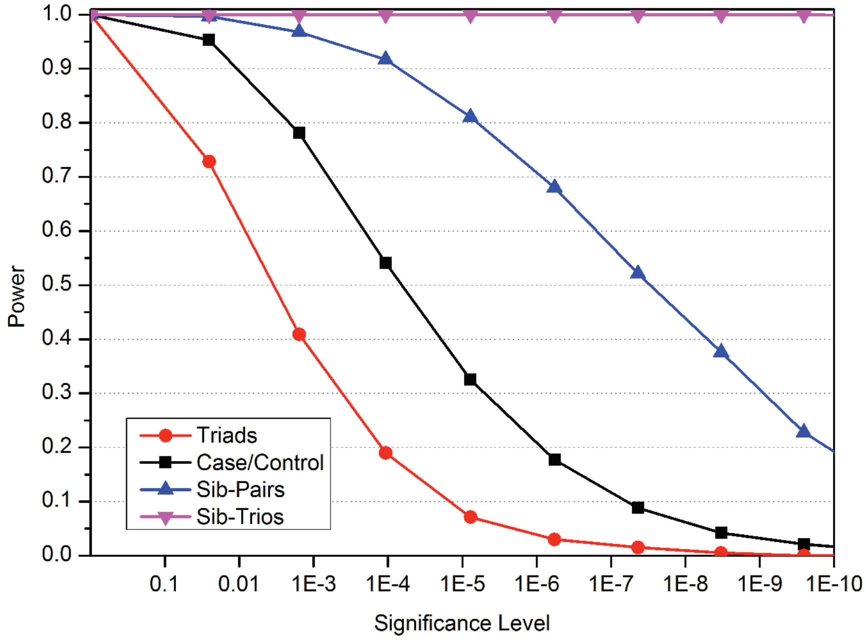

Figure 13. The power increase when sampling family material. There were total of 4000 genotyped individuals; a) 2000 cases and 2000 controls, b) 1000 triads and 1000 controls, c) 800 sib pairs and 800 controls, and d) 667 sib-trios and 665 controls. The generating model parameters were: a recessive genotype relative risk of 4, disease prevalence of 10%, disease allele frequency of 10%, complete linkage, and complete LD. The test statistic was PSEUDOMARKER recessive LD given linkage. The results are based on 1,000 replicates. The figure modified from Study III.

5.3.4 Additional controls in family-based association analysis

The increase in power, when adding random population controls to the schizophrenia dataset, is shown in Figure 14 (Hiekkalinna et al., unpublished results). The test statistic used here was again PSEUDOMARKER recessive LD given linkage. When there are missing data, which is always the case in real-life studies, genotyping additional random population controls improves the estimates of the population frequency of a risk allele in unaffected individuals, and therefore increases power.

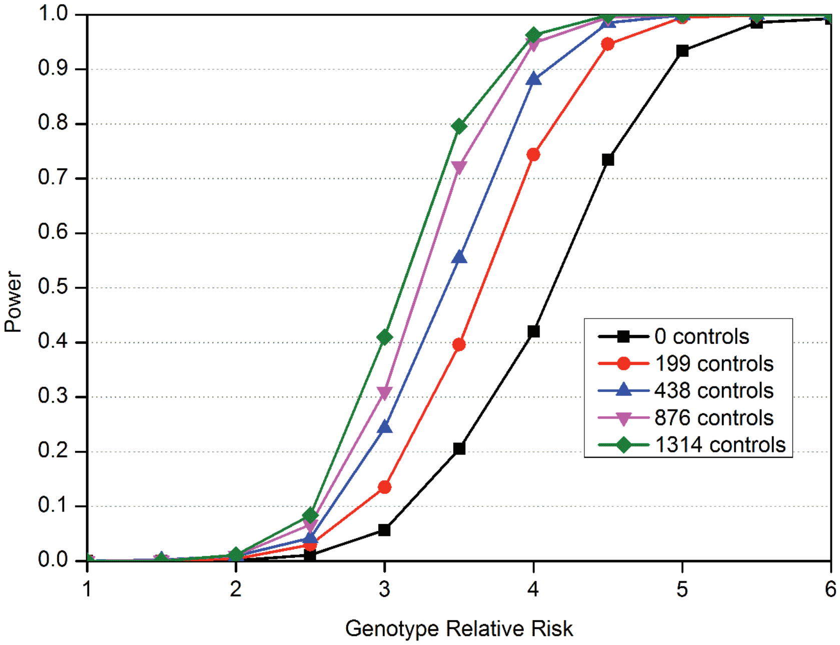

Figure 14. Power in the presence of increasing number of controls as a function of the genotype relative risk. The disease was simulated in the schizophrenia dataset assuming complete linkage, complete LD, and a recessive mode of inheritance with disease allele frequency of 10% and prevalence of the disease of 1%. The test statistic used was recessive PSEUDOMARKER LD given linkage. The number of controls 199, 438, 876, and 1314 corresponds to a) controls available in original study, b) one control per family, c) two controls per family, and d) three controls per family, respectively. The results are based on 1,000 replicates.

One could argue that power was higher when adding controls simply because the overall sample size grew. To address this issue we computed the expectation of the LRT statistic, Ψ , divided by the number of genotyped individuals in the study to get the expected information per genotype. Figure 15 shows that there is an increase in $E\left[\Psi\right]$ per genotype whenever the genotype relative risk is more than 2. Whether one adds one, two, or three controls per family the LRT per genotyped individual remains roughly the same in the schizophrenia dataset. This implies that after a certain number of additional controls there is no significant further gain in power, and that therefore it is not cost effective to genotype too many controls per founder.

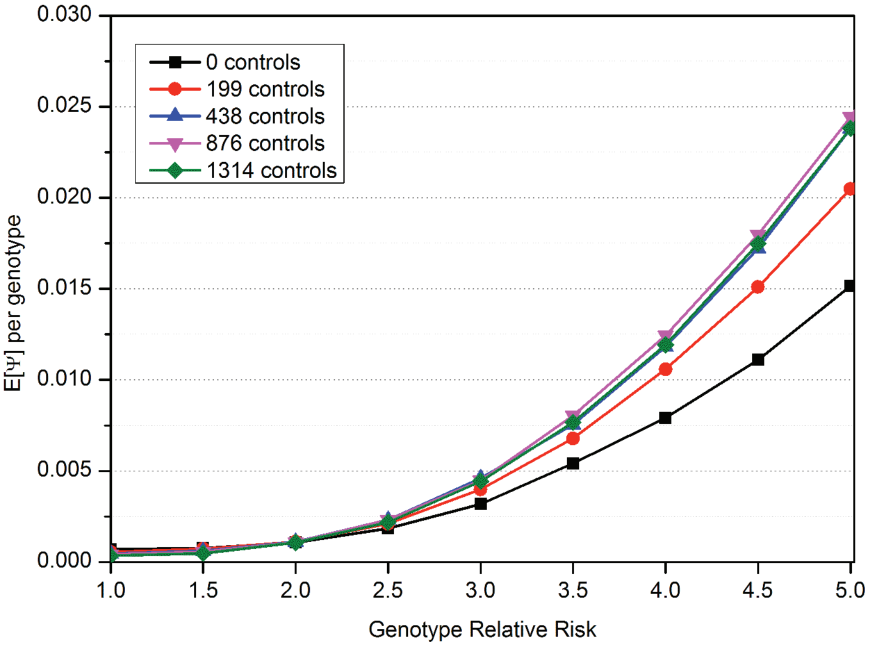

Figure 15. E[Ψ] per genotype as a function of the genotype relative risk in the presence of varying numbers of controls added to the analysis. Family relationship structures were taken from the schizophrenia dataset and the marker was simulated assuming complete linkage, complete LD, and a recessive mode of inheritance with disease allele frequency of 10% and prevalence of the disease of 1%. The test statistic used was recessive PSEUDOMARKER LD given linkage. The number of controls 199, 438, 876, and 1314 corresponds to a) controls available in original study, b) one control per family, c) two controls per family, and d) three controls per family, respectively. The results are based on 1,000 replicates.

In the migraine dataset, where the disease was dominant and the prevalence was much higher (10%), the ratio of $E\left[\Psi\right]/E\left[\Psi \mid \text{no controls}\right]$ was investigated, which measures the relative gain in the expected LRT statistic when adding controls, compared to the expected LRT statistic with no controls. Figure 16 shows the degree of this increase in $E\left[\Psi\right]/E\left[\Psi \mid \text{no controls}\right]$ when more controls are added. This figure shows that after the genotype relative risk exceeds 1.2, the ratio of the expected LRTs plateaus. This means that the ratio of the expected LRTs is independent of the genotype relative risk.

Our results show that including unrelated controls to the family-based association increases the power for detecting LD. Similar results was obtained by Lasky-Su et al

(Lasky-Su, Won et al. 2010). Furthermore, it is important that unrelated controls (and cases) are from the same genetic population to avoid population stratification.

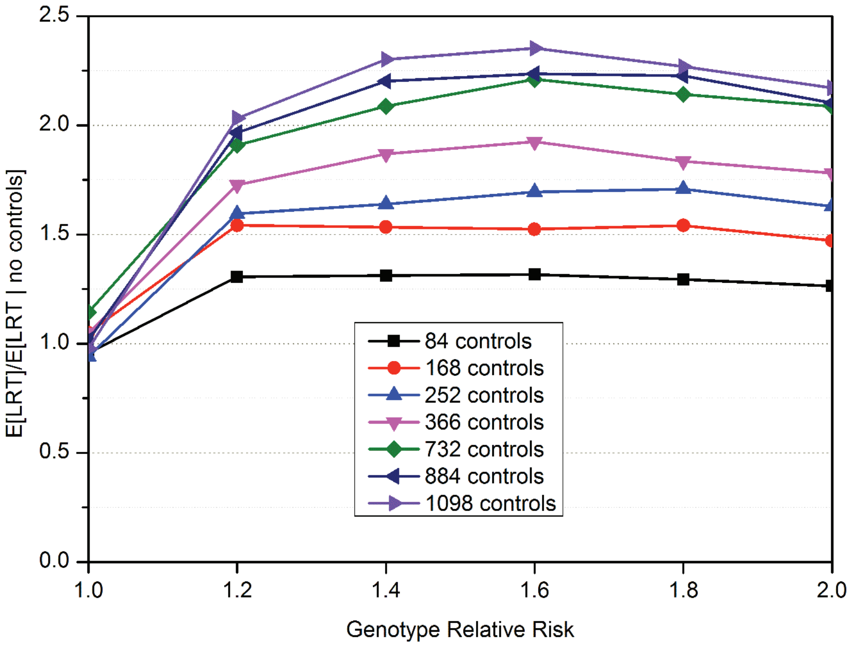

Figure 16. The ratio of E[Ψ]/ E[Ψ | no controls] as a function of the genotype relative risk (fix x-axis label) with varying numbers of controls added to the dataset. A marker was simulated in the migraine dataset assuming complete linkage, complete LD, and a dominant mode of inheritance with disease allele frequency of 10% and prevalence of the disease of 10%. The test statistic used was PSEUDOMARKER dominant LD given linkage. The number of controls 84, 168, 252, 366, 732, 884, and 1098 corresponds to a) one control per family, b) two controls per family, c) three controls per family, d) one control per founder, e) two controls per founder, f) controls available in the original study, and g) three controls per founder, respectively. The results are based on 1,000 replicates.

5.4 On the validity of the test of association conditional on linkage (III)

PSEUDOMARKER's recessive LD given linkage test showed an elevated type-I error rate of 0.1 at the α=0.05 level when the dataset consisted of 200 fully genotyped triads and 200 controls, but not in larger families (See Supplementary Table 6 in the original publication (III)). This likelihood ratio test Ψ assumes that the dataset contains information about the recombination fraction under both null and alternative hypothesis (See Table 4). However, the recombination fraction parameter does not exist under the null, creating a well-known pathology in the distribution of the LRT, as the conditions of Wilks' theorem are violated (Davies 1977). When additional affected sib pairs were added to the dataset, type-I error rates dropped to the appropriate levels when there as few as 10 informative sib-pairs (See Table 2 of the original publication (III)).

Whenever performing likelihood ratio tests with nuisance parameters, it is important that the likelihood be a function of those parameters under both null and alternative hypotheses. When testing LD conditional on linkage that means that the likelihood should be a function of the recombination fraction in the absence of LD (i.e. under the null hypothesis). In practice, this means one should have sufficient numbers of informative meioses for the lod score to have a nondegenerate distribution. Supplementary Figure 1 from the original study (III) shows that, when there are at least 20 informative meioses (10 fully informative sib-pairs), the exact p-value of a lod score of 3 is approximately at its asymptotic value of 0.0001.

5.5 Bias in parameters estimates on likelihood-based conditional tests (IV)

In order to examine the MLEs of parameters in LD analysis conditional on linkage, we computed profile log-likelihoods for all permutations of models M_{Rec} and M_{Dom} as true and analysis models shown in Table 7 (see section 4.3.3). Under the null hypothesis of complete linkage but no LD, the profile likelihoods maximize at the true parameter values, $p_{1|D} = p_{1|+} = 0.1$ (See Figures 1A, 1B and 1C in the original publication (IV)) for all combinations of true and analysis models with the single exception of when the true model was M_{Rec} and the analysis model was M_{Dom}. In this case, the estimated parameter values were $\hat{p}_{1|D} = 0.147$ and $\hat{p}_{1|+} = 0.084$ (Figure 17). The bias in the parameter estimates becomes even worse when one or both parental genotypes were unknown (See Figures 2E and 2F in the original publication (IV)).

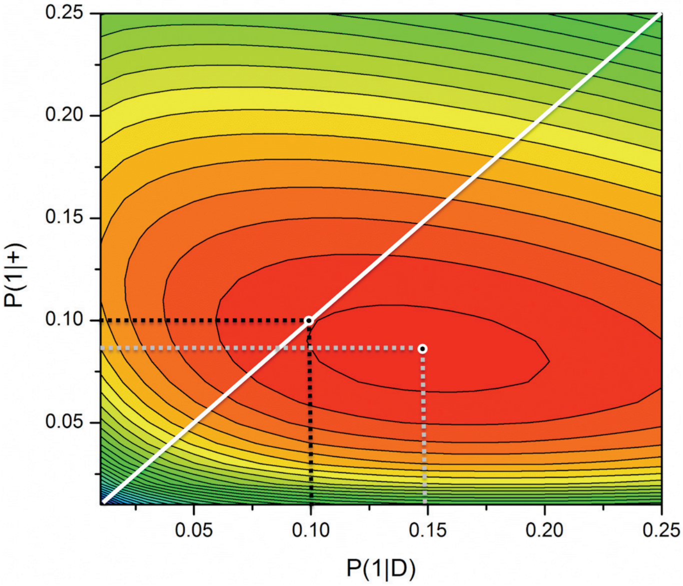

Figure 17. The expected profile log-likelihood surface for a single sib pair. True etiological model was M_{Rec}, while the inaccurate analysis model was M_{Dom}. The MLEs for conditional allele frequencies were $\hat{p}_{1|D} = 0.147$ and $\hat{p}_{1|+} = 0.084$. The figure is modified from Study IV.

When additional control individuals were added to the analysis, the bias in $\hat{p}_{1|D}$ was not removed, even with an infinite number of them. However, the bias in $\hat{p}_{1|+}$ almost vanished when adding controls (See Figure 3 in the original publication (IV)), because under these analysis models controls would all be inferred to have two + alleles. The bias exists only when the true mode of inheritance is recessive and the genotype relative risk is more than 15, and became worse when parental genotypes were missing (See Figure 4 in the original publication (IV)).

In order to uncover the source of this bias in the parameter estimates, we computed the detectance of the marker genotypes, P($\mathbf{G_M}$|**Ph**), in both members of an affected sib pair under models M_{Dom} and M_{Rec}. Both parents were assumed to be unaffected, complete linkage (θ=0) between diallelic marker and disease loci, marker allele 1 frequency of 0.1 $(p_2 = 0.9)$, and no LD. In Table 8, the detectance is shown for each possible marker genotype vector, $\mathbf{G_M}$, under both models, where children with different genotypes are indicated with shaded rows.

Table 8. Detectances for marker (diallelic marker with alleles 1 and 2) genotype vectors G_M, assuming models M_{Rec} and M_{Dom}. Shaded rows indicate children with different marker genotypes. The table is modified from Study IV.

G_M (Parents)	P(G_M-Parents)	G_M (Sibs)	P(G_M|M_{Rec})	P(G_M|M_{Dom})
1/1×1/1	0.0001	1/1-1/1	0.0001	0.0001
1/1×1/2	0.0036	1/1-1/1	0.0018	0.00135
		1/1-1/2	0.00000036	0.0009
		1/2-1/2	0.0018	0.00135
1/1×2/2	0.0162	1/2-1/2	0.0162	0.0162
1/2×1/2	0.0324	1/1-1/1	0.0081	0.00405
		1/1-1/2	0.00000324	0.0081
		1/1-2/2	0.00000000000162	0.000000162
		1/2-1/2	0.0162	0.0081
		1/2-2/2	0.000000324	0.00809998
		2/2-2/2	0.0081	0.00405
1/2×2/2	0.2916	1/2-1/2	0.146	0.10935
		1/2-2/2	0.00000292	0.0729027
		2/2-2/2	0.146	0.10935
2/2×2/2	0.6561	2/2-2/2	0.6561	0.6561

In Table 8, the mating type 1/2×2/2 with affected children with genotypes 1/2-2/2 has a great difference in the detectance. In this mating under model M_{Rec}, the proportion of families with both affected children having marker genotype 1/2 would be 50% and both with genotype 2/2 would be 50%. Under model M_{Dom}, if the parent with genotype 1/2 were carrying the disease allele, both children would have genotype 1/2 half of the time and genotype 2/2 the other half of the time. However, if the parent with genotype 2/2 were the carrier of the disease allele, then 25% of the time both children would have genotype 1/2, 25% of the time both would be 2/2, and 50% of the time there would be a 1/2 child and a 2/2 child.

Under analysis model M_{Dom}, it is more likely the parent having marker genotype 1/2 is the carrier of the disease allele than the parent with marker genotype 2/2, because it would appear that the sibs share marker allele 1 IBD from that parent in every sib pair in the study sample. This creates the illusion of LD between the disease allele and marker allele 1. If the segregation were random, and M_{Dom} were true, one would expect many more 1/2-2/2 sib pairs in the absence of LD. Thus, when analyzing recessive data under a dominant model, too many of the 1/2x2/2 matings are inferred to have the disease allele with the 1 allele, creating the false positive evidence of LD given linkage.

In real life, the true mode of inheritance is never known. If the dataset consists of predominantly nuclear families, one should be wary of applying dominant analysis models to test for LD conditional on linkage. The recessive analysis model did not suffer from this bias, and is therefore recommend because of its increased robustness and generally higher power as well, unless there is a clear evidence of intergenerational transmission of disease in multigenerational families.

5.6 Parametric linkage analysis using incorrect model and true model (IV)

In order to argue for highly inaccurate analysis models, we compared parametric linkage analysis under a very inaccurate etiological model assumption, M_{Rec}, with the same sort of analysis under the perfectly accurate simulated generating models. This comparison showed uniformly higher power when M_{Rec} was used for linkage analysis, rather than the true etiological model, in sharp contrast to the implications of earlier claims by Hodge and Elston (Hodge and Elston 1994; Greenberg, Abreu et al. 1998; Abreu, Greenberg et al. 1999), Greenberg et al (Greenberg, Abreu et al. 1998) and Abreu et al (Abreu, Greenberg et al. 1999).

In Figure 5 of the original publication (IV) the ratio of expected maximum lod scores under these models is shown for a dataset consisting of 800 affected sib pairs with unaffected parents. The M_{Rec} analysis model was always more powerful than the true parametric model; especially when the genotype relative risk was small, which is the situation presumed to apply in most complex multifactorial trait mapping studies. In Figure 18, results from an analogous simulation in the schizophrenia dataset are presented (Hiekkalinna et al., unpublished results). In this figure, three curves are graphed, with varying amounts of LD between marker and disease loci, D'=0 (no LD), D'=0.5 (intermediate LD), and D'=1 (complete LD). The fact that these lines are superimposed and indistinguishable confirms that this gain in power from the application of improperly specified analysis models is effectively independent of LD between disease and marker loci (Terwilliger 2001).

Two decades ago, when the scientific community began the shift from studying Mendelian traits to mapping of loci underlying complex traits, there was a great debate whether one should use parametric (“model-based”) or non-parametric linkage (“model-free”, i.e. no need to specify penetrance model) analysis methods (Hodge and Elston 1994; Kruglyak 1997). One argument favoring model-free methods was that for complex diseases it was impossible to determine the mode of inheritance. However, it was soon recognized that there are similarities between “model-based” and “model-free” methods (Knapp, Seuchter et al. 1994), such that “model-free” analyses are a special case of “model-based” analyses with overly deterministic etiological models (i.e. models with high detectance, such as M_{Rec} and M_{Dom} (Göring and Terwilliger 2000)). Because lod score analysis has the advantage of modeling all genetic relationships as they really are, applying these methods tends to lead to generally more powerful tests (Davis and Weeks 1997), consistent with the predictions of the Neymann-Pearson lemma. As we have demonstrated here, linkage analysis using the pseudomarker model is generally more powerful than the true correct model when the genotype relative risk is not enormous and the relationship structures consist of predominantly nuclear families.

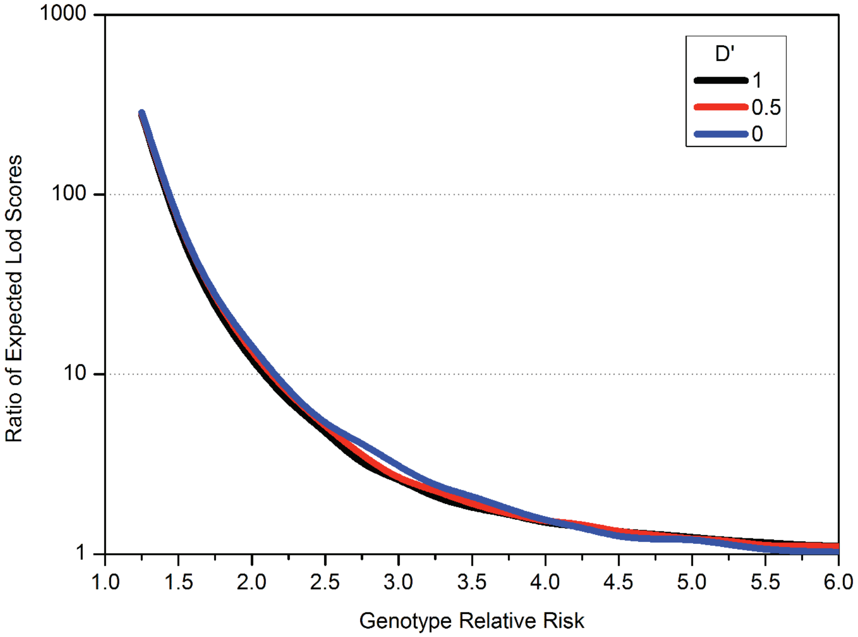

Figure 18. The ratio of M_{Rec} and true model-based two-point linkage analyses under the true etiological model using the Finnish schizophrenia dataset. Separate curves are presented as a function of LD between marker and trait loci with D’ set to 0, 0.5, and 1. The marker and disease allele frequency was 10% and disease prevalence was 1%. The results are based on 1,000 replicates.

5.7 Two-point vs. multipoint linkage analysis

To show importance of performing two-point linkage analysis in addition to multipoint linkage analysis, we compared the power of two-point and multipoint linkage analysis in ForSim generated data (see Section 4.3.12) (Hiekkalinna et al., unpublished results). Figure 19 shows a graph of the parametric two-point and multipoint lod scores and the LD plot among 11 markers from HaploView (Barrett, Fry et al. 2005). The lod scores were computed with MERLIN (Abecasis, Cherny et al. 2002) using the option for clustering markers to control for LD among them (Abecasis and Wigginton 2005). The etiological model for the functional variant in this gene was $p_D = 0.2$, $P(\mathit{Affected} \mid D/D \text{ or } D/+) = 0.5$ and $P(\mathit{Affected} \mid +/+) = 0.05$. The functional variant in this simulation, SNP5, showed a maximum two-point lod score of 3.5, while the multipoint lod score was only 1.5 (solid line). In Table 9, detectances are shown for each SNP in affected and unaffected individuals, where SNP5 shows the greatest difference in the detectance distributions as well.

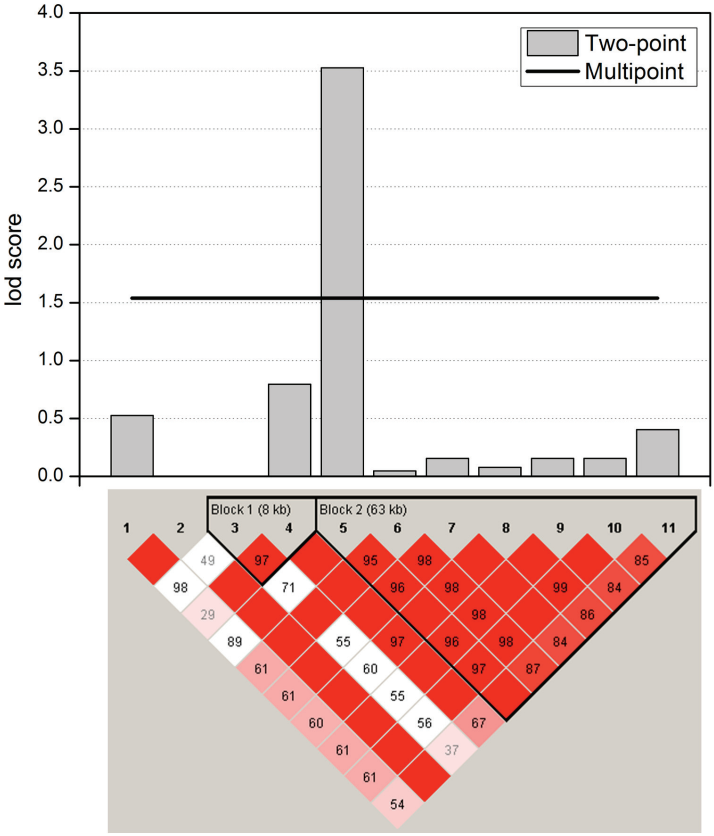

Figure 19. Multipoint and two-point linkage analysis of 11 tightly linked SNPs covering an area of ~89k base pairs. The pedigrees for linkage analysis were sampled w/o replacement from generated simulated population of 10,000 pedigrees using ForSim. This analysis sample contained 722 individuals from 79 pedigrees. All the pedigrees had three generations and individuals were all genotyped at all marker loci. The etiological model simulated for SNP5 was $P(Affected \mid D/D \text{ or } D/+) = 0.5$ **,** $P(Affected \mid +/+) = 0.05$ **and** $p_D = 0.2$.

This is possible, because SNP5 is the functional variant and co-segregates in every meiosis and is uninformative for linkage in meioses where it is not segregating. Other SNPs co-segregate with this locus when both are heterozygous for the same meioses, but segregate randomly to the offspring independent of affection status for all meioses in which SNP5 was homozygous and the SNP being analyzed was heterozygous. Because multipoint lod scores look at allele-sharing among affected offspring in all meioses, not just the ones in which the risk allele is actually segregating, multipoint lod scores tend to be smaller than two-point lod scores with either the functional variant itself or markers in high LD with it. Essentially, the flanking SNP markers are just adding noise to the analysis (Terwilliger and Göring 2000).

Table 9. The detectances for affecteds and unaffecteds for each SNP on ForSim simulated data.

Marker	Affected			Unaffected		
	1/1	1/2	2/2	1/1	1/2	2/2
SNP1	0.889	0.106	0.005	0.842	0.154	0.004
SNP2	0.000	0.134	0.866	0.004	0.162	0.834
SNP3	0.014	0.111	0.875	0.006	0.209	0.785
SNP4	0.380	0.546	0.074	0.279	0.455	0.267
SNP5	**0.181**	**0.787**	**0.032**	**0.731**	**0.253**	**0.016**
SNP6	0.005	0.171	0.824	0.008	0.231	0.761
SNP7	0.824	0.171	0.005	0.747	0.245	0.008
SNP8	0.815	0.181	0.005	0.737	0.255	0.008
SNP9	0.005	0.171	0.824	0.008	0.245	0.747
SNP10	0.005	0.171	0.824	0.008	0.245	0.747
SNP11	0.028	0.287	0.685	0.045	0.352	0.603

In Figure 20 the ratio of the expected maximum lod scores (computed as in Section 4.3.11) of two-point linkage analysis and multipoint linkage analysis in migraine dataset is graphed as a function of the strength of LD between the disease locus and SNP marker. The ratio of lod scores are shown for three genotype relative risks; 4, 7, and 10, to demonstrate that the general gain in power of two-point linkage analysis over multipoint is virtually independent of the true mode of inheritance. Two-point and multipoint lod scores are almost identical when D'=0.5 (intermediate LD) and when D'=1, two-point linkage is over three-times more powerful than multipoint linkage.

When using GWAS chips with hundreds of thousands of SNPs in families, it is likely that one of the SNPs would be in LD to some degree with any putative functional variant. Therefore is would be wise to perform simple and fast two-point linkage analysis in addition to multipoint analysis, because one could easily miss true positives. However, with two point analysis there are many more independent tests, and multiple test corrections would need to be applied (Bonferroni 1935).

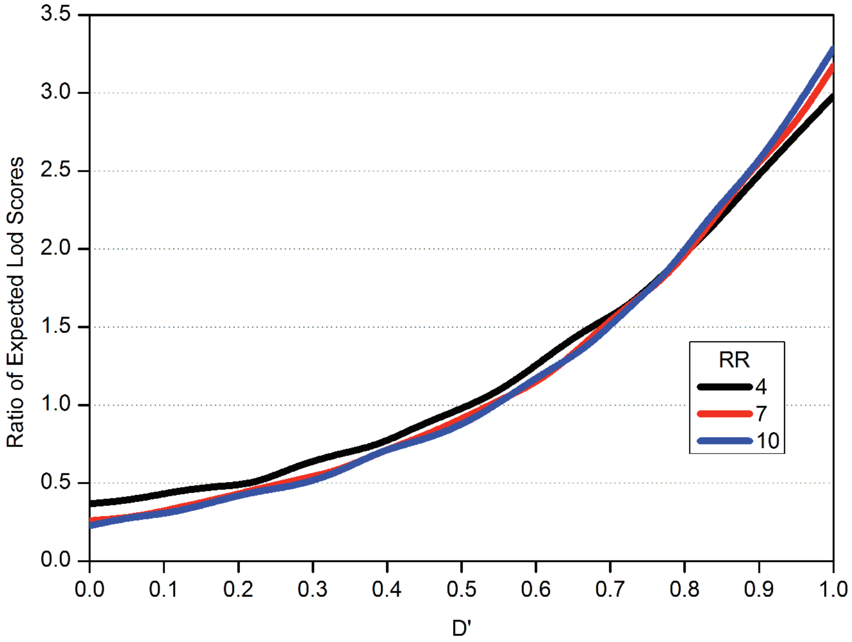

Figure 20. The ratio of expected two-point lod score to expected multipoint lod scores using the migraine dataset. The disease model was dominant, where $p_D = 0.1$, $\phi = 0.1$ and $\theta = 0$. The analysis model was M$_{\mathbf{Rom}}$. Power of two-point linkage analysis (assuming no LD in the analysis) increases when D' between the disease locus and SNP marker is higher. Curves are shown for genotype relative risks of 4, 7, and 10. In two-point linkage analysis, the SNP marker minor allele frequency was 10%. The results are based on 1,000 replicates.

5.8 Ascertainment bias and detectance

In order to emphasize the benefits of ascertainment bias in linkage and linkage disequilibrium analysis, detectances, $P(\mathbf{G_D} \mid \mathbf{Ph}; \mathit{ascertainment})$, for the risk locus genotype (D/D) were computed for cases and controls, and probands from triads (affected child with unaffected parents), sib pairs (two affected offspring with unaffected parents), and sib-trios (three affected offspring with unaffected parents) using our DETECTANCE software (Hiekkalinna et al., unpublished results). In Figure 21 detectances are shown for an affected child randomly chosen from each sampling unit as a function of genotype relative risk, where the highest detectance is for the proband in an affected sib-trio, followed by (in order) sib-pair, case, affected child in triad, non-transmitted genotype in triad, and control. A similar pattern was shown in Figure 13, where the power to detect LD given linkage was compared in these same relationship structures.

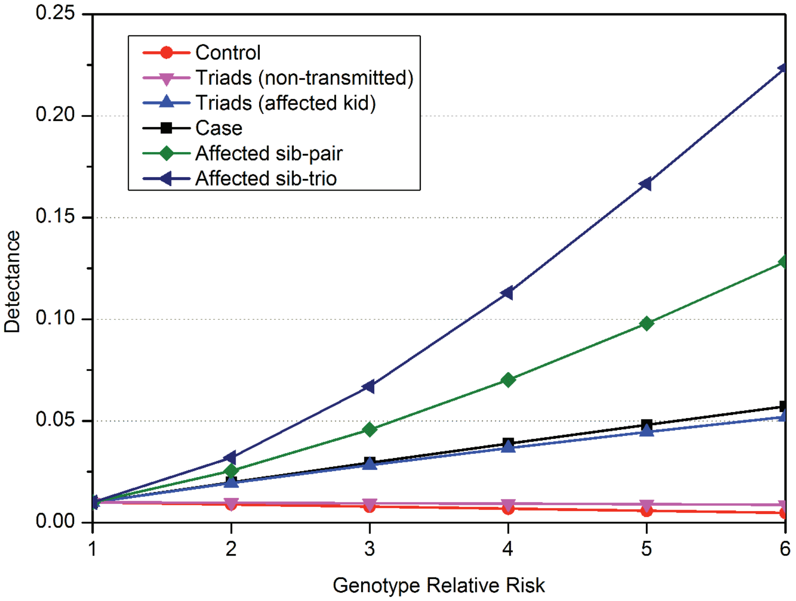

Figure 21. The detectance of disease locus genotype D/D for a proband selected under various study designs. The disease model was recessive with disease allele frequency of 10% and disease prevalence of 10%.

The control (screened for the phenotype) has the lowest detectance for genotype D/D. This is as expected, because under the recessive disease model, controls are most likely to have genotypes D/+ or +/+. The non-transmitted alleles in a triad had higher detectance to have genotype D/D than the controls, because essentially non-transmitted genotypes are random genotypes from the population (not screened for disease), as discussed in Section 2.4.3. The transmitted genotype in a triad had lower detectance than random cases, because we assumed parents in the triad were unaffected (if they were unknown it should be identical to random cases, of course). There is huge difference in the detectance between a case (from the population) and a case chosen from an affected sibship. Affected individuals in a sibship are more likely to share the disease for genetic reasons and thus share alleles, leading to higher detectance than for a random case from the population. The detectance is even higher in an affected sib-trio, because adding more affected relatives increases the probability that there are more risk alleles segregating in that family. From this result is clear to see that in general sampling families with multiple affected individuals increases the detectance of the underlying disease locus genotypes, and therefore increases the power of any study to detect linkage and LD.

6 General discussion

6.1 Monogenic trait mapping

Human geneticists have been quite successful in their search for the genetic determinants of monogenic traits in humans. For example, as of March 16, 2012, Online Mendelian Inheritance in Man (OMIM) reported a total of 21,134 entries, covering 3,441 monogenic traits with phenotype descriptions for which the molecular basis is known (Online Mendelian Inheritance in Man, OMIM®). The identification of the allelic variation underlying these phenotypes has been possible, because such traits run in families, due to highly-penetrant functional alleles. The latest data (February 10, 2012) from the Human Gene Mutation Database (HGMD) public release (Stenson, Ball et al. 2009), where one of the requirements is "*Novel appearance and subsequent co-segregation of the lesion and disease phenotype through the family pedigree*" (among other firm criteria), indicates that a total of 85,840 functional variants from 3,253 genes have been firmly associated with some clinical phenotype. The number of genes listed in HGMD is smaller, one of the reasons being that they do not include somatic mutations and mutations in the mitochondrial genome.

6.2 Complex trait mapping

A complex trait is by definition, complex, where the phenotype is a result of multiple factors, such as genetic factors (typically alleles of individually small effect at a large number of independent loci), environmental factors, and cultural factors. After years of successful identification of alleles of individually large effect (often themselves being sufficient to cause some disease outright), researchers began to shift their focus to more common multifactorial conditions, such as cardiovascular disease, and mental illnesses. For example, cardiovascular disease alone is a significant cause of premature death (WHO 2011) and treatment of chronic diseases uses a very large proportion of the available health care resources. Therefore, there was a great public health interest in the search for genetic risk factors related to such diseases. However, the search for genetic determinants for complex diseases with small effects seems to require enormous sample sizes and therefore ascertainment was focused on a large number of smaller family structures or large population-based samples, many of which were already collected and available.

Complex trait mapping has been boosted by technological advances in the last two decades. The first draft of the human genome was sequenced in 2001 (Lander, Linton et al. 2001; Venter, Adams et al. 2001; International Human Genome Sequencing Consortium 2004), followed by the International HapMap Project, which developed a dense genome-wide map of SNPs (The International HapMap Consortium 2003). Methods for rapid analysis of such SNPs has been developed (Purcell, Neale et al. 2007; Kang, Sul et al. 2010; Yang, Lee et al. 2011) and it is likely that within a decade even complete genome sequence data will become available at affordable costs for all individuals in any study sample. At the moment, GWAS is routinely done with roughly 1,000,000 genotyped SNPs.

In Figure 22 shows a cumulative graph of the number of identified highly-penetrant causal variants related to monogenic diseases (HGDM, http://www.hgmd.org) (Stenson, Ball et al. 2009) and the number of loci (reported SNP associations, rather than actual functional variants) which have been correlated to the multifactorial traits through GWAS (A Catalog of GWAS studies, http://www.genome.gov/gwastudies). Significantly, while the functional variants in monogenic disease are themselves causative of some disorder, the GWAS loci may not be themselves functional and typically do not even identify which specific gene harbors the putative functional effect. The GWAS data obtained from the aforementioned website was filtered to include only genotyped SNPs associated to some phenotype with a p-value $< 5\times10^{-8}$, such that the number of SNPs (over all phenotypes, disease and otherwise) which met this criteria was 1749. The total number of SNPs in the GWAS Catalog was 5864, most of which did not meet the required significance threshold. This graph shows that identification of Mendelian functional variants has been remarkably more successful than GWAS.

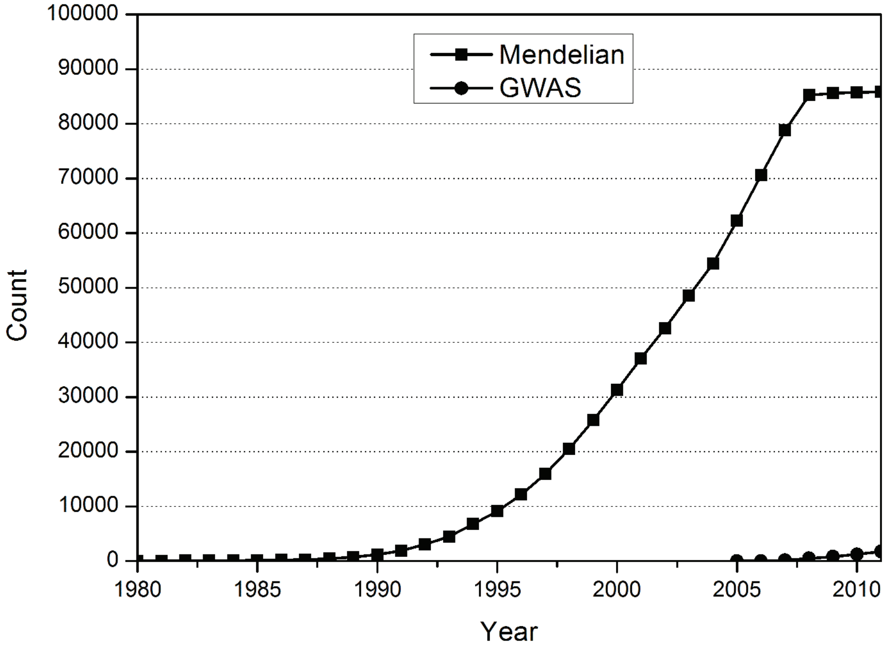

Figure 22. A cumulative number of published Mendelian functional mutations (The Human Gene Mutation Database, public release) and reported GWAS loci (A Catalog of published GWAS studies, http://www.genome.gov/gwastudies) between 1980 and 2011. The GWAS data was filtered and contains only genotyped SNPs with p-values < 5×10^{-8}. The data was obtained from the web sites on February 10, 2012.

6.3 Singletons

In the last few years, large population based studies have dominated the mapping of complex traits. These initial GWAS pursuits were motivated by the common-variant/common-disease hypothesis (Risch and Merikangas 1996), under which common variants of high attributable fraction were assumed to play the major role in the etiology of common diseases such as heart disease, diabetes, or obesity. Epidemiologists have for a long time pursued cross-sectional studies of large numbers of unrelated individuals, often saving blood samples for future analysis. For this reason, it was quick and easy to apply biotechnological advances to these readily available samples of deeply-phenotyped cohorts. International collaboration has enabled large consortium studies, for example of blood pressure and cardiovascular risk factors containing hundreds of thousands of individuals (Ehret, Munroe et al. 2011). However, if identification of a risk variant for any given trait requires such enormous sample sizes, it is doubtful that such variants could possibly have clinical relevance or significant public health impact (Weiss and Terwilliger 2000).

As shown in this study, case-control sampling is not very powerful for finding such risk variants, because the detectance is low compared to sampling of related individuals in families, which are more likely to share causal risk factors. Cross-sectional studies can be useful for estimating the prevalence of a disease, and for estimation of the actual effect size of known risk factors. Repeated longitudinal follow-up studies in these large surveys, such as the Finnish FINRISK study for cardiovascular risk factors (Vartiainen, Jousilahti et al. 2000), may provide useful information about health changes in the population.

6.4 Triads

The use of triads (affected offspring and parents) was proposed to avoid false-positive results due to population stratification (Rubinstein, Walker et al. 1981). Methods to analyze triad data are simple to implement, which led to rapid development of multiple program packages implementing variations on the HRR and TDT. Another attractive feature of such computational simplicity is that analysis of large amounts of data is very fast, because such methods only count transmitted and non-transmitted alleles.

Our results confirmed earlier findings that the triad-based design is the least powerful design for testing allelic association, because it is required to genotype three individuals to get one “case” genotype and one “control” genotype (constructed from non-transmitted alleles). Therefore, 50% more genotyping is needed for the same information as a case-control study. Furthermore, non-transmitted genotypes are essentially random genotypes from the population not screened for the disease (Ahsan, Hodge et al. 2002), rather than phenotypically healthy controls, making it more analogous to a less-powerful case-cohort design rather than case-control. This result is consistent with earlier reports (Terwilliger and Ott 1992; Risch and Merikangas 1996).

6.5 Sibships

It is surprising how often professional human genetic researchers erroneously believe that association analysis cannot be performed in families. Historically speaking, one or two decades ago, association analysis was done in many studies, when it was common to do large genome-wide linkage analysis on sets of families and then after significant results were found, follow-up studies were done to screen for LD in the area of significant linkage, using the same family material. Although for a long time software tools have been available for joint linkage and LD analysis in large pedigrees, they were either very hard to use or required extensive computing resources to run, which were not often accessible to biologists.

We have shown that association analysis in affected sib pairs (or sibships) is more powerful than in singletons or triads, consistent with the well-known results of previous decades (Risch and Merikangas 1996; Göring and Terwilliger 2000; Terwilliger and Göring 2000). This is obvious, because sampling affected relatives enriches the sample for the genetic portion of the "multifactorial" etiology. Furthermore, affected relatives are likely to share the same risk alleles, and families with more affected relatives are more likely to be segregating more genetic risk factors. The use of affected sib pairs or larger families likewise allows for the analysis of linkage as well as linkage disequilibrium as we have shown, meaning that there are more analysis options when pedigrees are collected.

6.6 Complex pedigrees

Multigenerational families have been used for localization and identification of a large number of highly penetrant risk factors for inherited diseases using linkage and LD analysis strategies for the past several decades. Many attempts were made to pursue the same strategy for complex traits as well, with limited success. This paucity of results resulted because the sample size requirements are much larger than people expected, because they grossly overestimated the marginal effect size of the genetic risk factors underlying such traits, when the effect sizes of the inherited risk factors are small. One recent example of a successful joint linkage and LD analysis in a set of complex pedigrees was the identification of the variant responsible for adult-type hypolactasia in Northern Europeans (Enattah, Sahi et al. 2002). In that study, LD analysis conditional on linkage was used to pinpoint the genomic region associated with lactase non-persistence. Although multigenerational pedigrees are difficult to ascertain, they can be the most informative sampling units in the search for functional variants which have an effect on some phenotype.

6.7 Joint analysis of singletons and families

We have shown the benefits of joint analysis of singletons, triads, sib pairs, larger sibships, and multigenerational families using full likelihood-based methods. Sib pairs (and larger sibships) and multigenerational pedigrees provide information about linkage, while singletons provide important information about the allele frequencies. This is essential, because it is quite common to have missing genotype information in families, especially in families with multiple generations. As we have shown, adding controls to the analysis leads to more powerful and more accurate results. It is important to note that joint linkage and LD analysis requires much more computing time than traditional linkage analysis or simple family-based association analysis approaches. When millions of SNP markers are genotyped, it is prudent to first perform simple and fast two-point linkage analysis, and LD analysis with the computationally trivial HHRR for all markers (this statistic is recommended based on its performance in our simulation studies compared with the other computationally simple alternatives). If some region shows evidence of linkage and LD with some markers in this initial scan, then a full joint linkage and LD analysis should be performed with PSEUDOMARKER.

The renaissance of family-based studies is at our doorstep, especially in Finland, where we have computerized population registers of the pedigree relationships among all individuals. GWAS on cross sectional data has not been as successful as predicted (Reich and Lander 2001) and researchers are starting to re-think family-based strategies again. However, only a small percentage of people who work in genetic epidemiology today have had any experience or training in the concepts and practice of linkage analysis. This creates a great demand for automated easy to use linkage and LD analysis packages like our own.

6.8 Study design

We have shown that association mapping based on samples from large multiplex families is significantly more powerful than studies based on cross-sectional data. Ascertaining families with multiple affected individuals increases the probability (the detectance) that affected individuals are affected because of shared genetic factors (as opposed to environmental factors).

The phenotype (qualitative or quantitative) is critical for any given mapping study, because it is essential to have a well characterized biological phenotype that has a strong correlation with some underlying trait locus genotypes (i.e. with high detectance). If the phenotype is poorly defined or is biologically too complex and multifactorial, there is little hope of finding any important functional variants. The success of a gene mapping study is also dependent of other factors shown in Figure 23, where good quality sequencing and genotyping is important in addition to powerful analysis methods.

In human genetics we cannot do experiments like we can with yeast and mice, nor do we have extremely complex and large families like in the canine breeding pedigrees used in gene mapping, and therefore we must always look for natural experiments in the human population. For example, we may look for isolated sub-populations or families harboring unusual and interesting phenotypes. The Finnish population is an example of such a natural experiment which was extremely useful for identifying the causes of many rare monogenic diseases, because of Finland's isolation in northern Europe, small founder population size, minimal immigration and rapid population growth (Norio, Nevanlinna et al. 1973; Peltonen, Jalanko et al. 1999). However, for complex common traits, these advantages may be largely attenuated because of the complexity of the etiology and high frequency of such phenotypes even in Finland. Therefore, one might need to seek out other sorts of natural experiments when dealing with less rare and less deterministic conditions. For example, populations where first-cousin marriages are the rule rather than the exception could prove interesting.

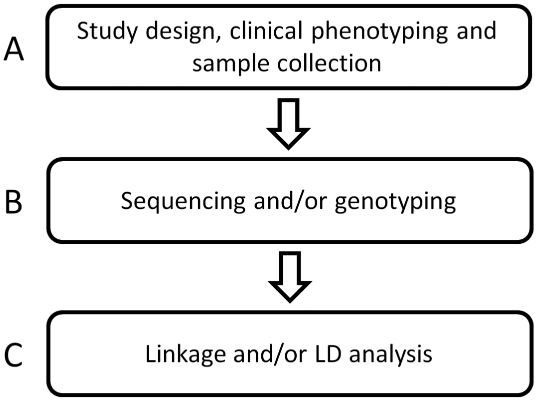
A
Study design, clinical phenotyping and sample collection
B
Sequencing and/or genotyping
C
Linkage and/or LD analysis

Figure 23. An oversimplified flow chart of a gene mapping study. (A) Study design, clinical phenotyping and sample collection, where ascertainment of multiplex families is preferred. (B) Genome sequencing and/or genotyping of the sample. (C) Joint linkage and/or LD analysis of families and unrelated individuals. This study was focused on developing tools for easy and powerful linkage and/or LD analysis of the data, but if the study is poorly designed, the phenotype is poorly defined or sequencing and genotyping quality is low, there is no linkage or association method which can rescue the study.

6.9 Future prospects

The technological advances of the last few decades have made it possible to look at the complete sequence of the human genome. However, it is not current practice to sequence every sampled individual in a study, but it is likely to become the reality in the near future, as a by-product of screening for the tens of thousands of highly penetrant risk factors known to be related to monogenic diseases. This would mean that each individual would have almost over three billion data points (i.e. base pairs). This will be a nightmare for IT-professionals and data managers handling servers and databases, because the amount of genetic data that sequencing will produce is astronomical (although it will not compete with the data created by CERN's Large Hadron Collider in the search for the Higgs boson!). Besides that, this large amount of data points means one will face the curse of dimensionality in statistical hypothesis testing; testing of over three billion sites would require ridiculously small p-values to compensate for multiple testing. Furthermore, when family registers are linked with the sequence data, the computational burden will become enormous.

In the end, because of the curse of dimensionality, it is not unlikely that we will be back to classical positional cloning, where investigators will start by selecting a sparse map of markers for linkage analysis, followed by a slightly more dense HapMap style set of markers for GWAS, with interesting regions followed up by joint linkage and LD analysis and sequence comparisons. The advantage of cheap and affordable sequencing, especially when it is being performed for health-screening purposes, is that the time and cost of obtaining this molecular information will no longer be a factor in the selection of experimental designs. The era when "*...study designs for gene mapping be compared under the assumption that we have the complete sequence of the entire genome of every samples individual, as a means of focusing the discussion on the more relevant issue of whether or not there are detectable genetic risk factors in a given data set under the best of circumstances.*" (Terwilliger and Göring 2000) may be upon us sooner than we think.

7 Conclusions

This study would not have been possible without extensive national and international collaboration between physicians, geneticists, statistical geneticists, evolutionary biologists, and last but not least, computer scientists. Experts from different fields of specialization are crucial for any successful study.

It seems to be that the era of large GWAS studies population-based samples for common diseases is coming to an end as we have seen in the past few years. It is reality that family-based studies are coming back from the geneticist's "dusty drawers", although they were never completely forgotten.

To this end, we should start to prepare our analysis methods, biotechnology centers, and IT-infrastructure for the full genomic sequence of thousands or even millions of individuals. It is likely that in ten or twenty years, when sequencing technology is affordable and accurate, all individuals in a population will be sequenced by default.

This would be a highly valuable resource for the researchers in the future when individuals and their relatives could be just directly obtained from databases based on some clinical information. However, even with complete sequence data, we should still begin our gene hunting quest with classical positional cloning.

8 Acknowledgments

This study was carried out in the Department of Molecular Medicine, Finland's National Public Health Institute, Helsinki, Finland, later the National Institute of Health and Welfare, Helsinki, Finland, and Institute of Molecular Medicine Finland FIMM. I would like to thank the director general of the National Institute of Health and Welfare, Professor Pekka Puska, director of Institute of Molecular Medicine Finland, Professor Olli Kallioniemi and head of Public Health Genomics Unit, Adjunct professor Anu Jalanko for providing the excellent research facilities.

I greatly acknowledge the financial support from the Helsingin Sanomat Foundation, the Emil Aaltonen Foundation, the Finnish Cultural Foundation, Academy of Finland, the Biomedicum Helsinki Foundation, the Otto A. Malm Foundation, the Jenny ja Antti Wihuri Foundation, and the Paulo Foundation. The support for high performance computing from CSC - IT Center for Science is highly appreciated.

There are no words enough in this universe to express my thankfulness for the late Academician of Science Leena Peltonen-Palotie. From the day one I felt warmly welcomed into her group at KTL, the atmosphere in her group was amazing. That time in her group changed my life. She took a boy from Paltamo - Kainuu, to Los Angeles - USA, to build the Department of Human Genetics computer infrastructure from the scratch, and I'm truly grateful for that. I could never even dream of that kind of opportunity in my life. My switch from IT-support to scientific research would not have been possible without Leena's encouragement and support. On my first scientific abstract is Leena's handwritten comment:"*Tämä on TOSI hyvä!!, Leena (Kukka)*", I still have that paper in safe place. She was truly inspirational and exceptional scientist.

My utmost gratitude goes to Professor Joseph Terwilliger who is the visionary and the mastermind behind my thesis work. I truly admire Joe's broad knowledge of statistics and genetics (and life in general!). I'm extremely lucky to have him supervise my thesis. We met first time at the KTL many years before he became my thesis supervisor and along these years we have become really good friends. I have had incredible journey with him; our Schlaab N' Hammer Tours has taken us all over the world (Tartu, Tallinn, Riga, Uppsala, Stockholm, Seoul, Tampa Bay, Penn State, New York City, Washington DC, Montreal, and the list goes on…). I'm proud to say that it's amazing to have a friend like him.

I wish to express my huge gratitude to Research Professor Markus Perola with who I have shared office for almost 12 awesome years. During these years we have become really good friends and I'm extremely lucky to have friend and supervisor like him. Markus has helped me since beginning when I started my research and his support has been very crucial during these years. It's incredible to have a friend like him.

I express my gratitude for Professor Tiina Paunio, Olli Pietiläinen and Adjunct Professor Teppo Varilo for providing the Schizophrenia pedigree structures and Professor Aarno Palotie, Adjunct Professor Maija Wessman, Dr. Mari Kaunisto, Dr. Verneri Anttila for providing the Migraine pedigree structures for the simulation studies.

I am extremely grateful to thank Dr. John Blangero accepting the role as Opponent in my thesis defense. I am very grateful to Professors Hannes Lohi and Daniel E. Weeks for reviewing my thesis – their comment were extremely valuable for improving the manuscript. Donald Smart is greatly acknowledged for language revision of the manuscript.

This thesis would not have been possible without the great international collaboration with Dr. Alejandro S. Schäffer, Dr. Harald H.H. Göring, Associate Professor Joseph H. Lee, Brian W. Lambert, and Professor Kenneth M. Weiss. I'm extremely lucky that I have been able to collaborate with you and I sincerely hope that our collaboration will continue in the future. Very special thanks go to Alejandro, whose patience for getting our results published has been extremely long. Special thanks go to Harald whose always positive attitude is just truly amazing and friendship during off-work hours have been so much fun. Special thanks go to Joe Lee and Ken with who brainstorming sessions in New York City and in Penn State have been very inspirational and fun.

I wish to thank many people in the National Public Health Institute, Department of Molecular Medicine at Mannerheimintie 166, where I got my first glimpse into genetics: Ann-Christine Syvänen, Leena Karttunen, Iiris Hovatta, Johanna Aaltonen, Petra Pekkarinen, Satu Kuokkanen, Tomi Pastinen, Jesper Ekelund, Lasse Lönnqvist, Juha Isosomppi, Päivi Pajukanta, Teppo Varilo, Tiina Paunio, Miina Öhman, Maria Halonen, Ilona Visapää, Ville Holmberg, Pekka Ellonen, Heidi Lilja, Jani Saarela, Nabil Enattah, Aki Suomalainen, Teemu Perheentupa and many others. The list is incomplete, my deepest apologies.

I would also like to thank many wonderful people at the University of California, Los Angeles, Department of Human Genetics, where I was able to hangout and work almost three years with Jaana & Mikko, Tutsa & Vesku, Maikki, Markus & Virpi, Pia & Iikka M, Mira, Jenny, Tuula, Lennu, Juha P, Heidi, Mira, Elina, Aino, Niklas, Jouni, Hilde, Joni, Päivi, Janna, Kirsi, Kismat & Aimee, Anthony 'Tony'

Metdizis and many others. Very special thanks go to Tony, we had so good times. Special thanks go to Professor Aarno Palotie, for all his help and support during LA years. Also I would like to thank senior scientist in the Department of Human Genetics: Rita Cantor, Janet Sinsheimer, Jeanine Papp, Eric Sobel and Kenneth Lange. It was honor to work with you and start my career as 'wannabe-scientist' in UCLA.

I wish many thanks to 'the next generation of scientists' at KTL/THL/FIMM: William, Elina, Tero Y, Mira, Kaisu, Mikko, Niklas, Jenny, Joni, Kirsi, Kaisu, Jussi, Kati, Annina, Annika, Sampo, Mervi Antti, Juha K, Carina, Jonna, Karola, Suvi, Riikka, Heli, Liisa, Juho, Marika, Johannes, Anu, Heidi, Annu, Tony, Kismat, Aimee, Minttu, Outi, Päivi, Pekka, Markus L, Hanna, Antti, Olli, Mikko M, Taru, Ansku, Mari, Tiia, Virpi, Ida, Emmi, Tea, PP, Jarkko, and Verneri. I have surely forgotten someone from the list, and therefore I express my deepest apologies. You are all awesome. I would like thank Tony, Jari, Juri, Hannu, Timo, Tomi and Teemu for their excellent computer support. Special thanks go to Juri, Tony and Teemu whose company at the office made everyday hilarious. Petri Norrgrann is thanked for his programming support. Special thanks go to Teppo, who has helped me countless of times during these years. I wish to give very warm thanks to our group, former and present members: Outi, Kirsi, Mervi, Sampo, Johannes, Juha K, Kaisu, Antti, Katja, Emma, Jasmine, Marjis, Anni, Perttu, Niina and Emilia. You have been excellent company at the office and in the conferences all over the world. Special thanks go to Teppo, Perttu, Sampo, Johannes and Markus, who had made work and off-work atmosphere truly pleasant and enjoyable!

I would like to thank our senior scientist in our department and unit for the help along these several years: Ann-Christine Syvänen, Ismo "Iski" Ulmanen, Marjo Kestilä, Tiina Paunio, Anu Jalanko, Kaisa Silander, Teppo Varilo, Iiris Hovatta, Janna Saarela, Juha Saharinen, Vesa Olkkonen, Samuli Ripatti, Aija Kyttälä, Anu Loukola and Matti Jauhiainen

I would like to acknowledge our wonderful secretaries and administration who have helped me during these years: Tuija Koski, Liisa Penttilä, Sari Kivikko, Shen Huei-Yi, Sari Mustala and Susanna Rosas.

My friends are specially thanked and just to list few of them: Jani & Satu, Samuli & Hellevi, Simo & Kati, Elsa & Lauri, Kovis & Jatta, Antti & Grace and Marja. I wish to thank my dear friends from Paltamo (The Kings from Paltamo): Jute, Jouni, Jukka, Iikka and Jorkki. Special thanks go to Jute who has helped me to explore the universe beyond our Solar System. Very special thanks go to Metallica and their superior heavy metal music, from which I absorb energy to my everyday life.

I wish to warmly thank my dear Mom and uncle Oiva for the love and care when I needed it. Tommi & Anne are thanked for the care and company.

My warmest and deepest thanks and love goes to my lovely wife Paukka and my two wonderful sons Ukko and Oiva. Without your love and support this would not have been possible. Nothing else matters.

Helsinki, August 10th, 2012

Tero Hiekkalinna

9 References

(2012). American Cancer Society: Cancer Facts and Figures 2012 http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-031941.pdf.

(International Human Genome Sequencing Consortium 2004). "Finishing the euchromatic sequence of the human genome." Nature **431**(7011): 931-45.

(Online Mendelian Inheritance in Man, OMIM®). "Online Mendelian Inheritance in Man, OMIM®." from http://omim.org/.

(The Genome Reference Consortium 2012). "The Genome Reference Consortium." from http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/index.shtml.

(The International HapMap Consortium 2003). "The International HapMap Project." Nature **426**(6968): 789-96.

(WTCCC 2007). "The Wellcome Trust Case Control Consortium Study: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-78.

Abecasis, G. R., L. R. Cardon and W. O. Cookson (2000). "A general test of association for quantitative traits in nuclear families." Am. J. Hum. Genet. **66**(1): 279-92.

Abecasis, G. R., S. S. Cherny, W. O. Cookson and L. R. Cardon (2002). "Merlin--rapid analysis of dense genetic maps using sparse gene flow trees." Nat. Genet. **30**(1): 97-101.

Abecasis, G. R., W. O. Cookson and L. R. Cardon (2000). "Pedigree tests of transmission disequilibrium." Eur. J. Hum. Genet. **8**(7): 545-51.

Abecasis, G. R. and J. E. Wigginton (2005). "Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers." Am. J. Hum. Genet. **77**(5): 754-67.

Abreu, P. C., D. A. Greenberg and S. E. Hodge (1999). "Direct power comparisons between simple LOD scores and NPL scores for linkage analysis in complex diseases." Am J Hum Genet **65**(3): 847-57.

Ahsan, H., S. E. Hodge, G. A. Heiman, M. D. Begg and E. S. Susser (2002). "Relative risk for genetic associations: the case-parent triad as a variant of case-cohort design." Int J Epidemiol **31**(3): 669-78.

l-Yahyaee, S., L. I. Al-Gazali, P. De Jonghe, H. Al-Barwany, M. Al-Kindi, E. De Vriendt, P. Chand, R. Koul, P. C. Jacob, A. Gururaj, L. Sztriha, A. Parrado, C. Van Broeckhoven and R. A. Bayoumi (2006). "A novel locus for hereditary spastic paraplegia with thin corpus callosum and epilepsy." Neurology **66**(8): 1230-4.

Almasy, L. and J. Blangero (1998). "Multipoint quantitative-trait linkage analysis in general pedigrees." Am. J. Hum. Genet. **62**(5): 1198-211.

Aromaa, A., S. Koskinen and eds. (2004). Health and functional capacity in Finland. Baseline results of the Health 2000 health examination survey. Helsinki, National Public Health Institute (http://www.terveys2000.fi/julkaisut/baseline.pdf).

Barnard, G. A. (1949). "Statistical Inference." Journal of the Royal Statistical Society. Series B **11**: 115-149.

Barrett, J. C., B. Fry, J. Maller and M. J. Daly (2005). "Haploview: analysis and visualization of LD and haplotype maps." Bioinformatics **21**(2): 263-265.

Birney, E., J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H.

Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W. K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C. L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. Karaoz, A. Siepel, J. Taylor, L. A. Liefer, K. A. Wetterstrand, P. J. Good, E. A. Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Loytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. A. Stone, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King, A. Ameur, S. Enroth, M. C. Bieda, J. Kim, A. A. Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega, C. W. Lee, P. Ng, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. A. Singer, T. A. Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. A. Nix, G. Euskirchen, S. Hartman, A. E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T. H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C. K. Glass, M. G. Rosenfeld, S. F. Aldred, S. J. Cooper, A. Halees, J. M. Lin, H. P. Shulha, M. Xu, J. N. Haidar, Y. Yu, V. R. Iyer, R. D. Green, C. Wadelius, P. J. Farnham, B. Ren, R. A. Harte, A. S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A. S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R. M. Kuhn, D. Karolchik, L. Armengol, C. P. Bird, P. I. de Bakker, A. D. Kern, N. Lopez-Bigas, J. D. Martin, B. E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyras, I. B. Hallgrimsdottir, J. Huppert, M. C. Zody, G. R. Abecasis, X. Estivill, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. A. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. A. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu and P. J. de Jong (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Bonferroni, C. E. (1935). "Il calcolo delle assicurazioni su gruppi di teste." Studi in Onore del Professore Salvatore Ortu Carboni: 13-60.

Callegaro, A., J. J. Lebrec and J. J. Houwing-Duistermaat (2010). "Testing for genetic association in the presence of linkage and gene-covariate interactions." Biom. J. **52**(1): 22-33.

Cantor, R. M., G. K. Chen, P. Pajukanta and K. Lange (2005). "Association testing in a linked region using large pedigrees." Am. J. Hum. Genet. **76**(3): 538-42.

Chowdhury, R., P. R. Bois, E. Feingold, S. L. Sherman and V. G. Cheung (2009). "Genetic analysis of variation in human meiotic recombination." PLoS Genet **5**(9): e1000648.

Clayton, D. (1999). "A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission." Am. J. Hum. Genet. **65**(4): 1170-7.

Comuzzie, A. G., J. E. Hixson, L. Almasy, B. D. Mitchell, M. C. Mahaney, T. D. Dyer, M. P. Stern, J. W. MacCluer and J. Blangero (1997). "A major quantitative trait locus determining serum leptin levels and fat mass is located on human chromosome 2." Nat Genet **15**(3): 273-6.

Cottingham, R. W., Jr., R. M. Idury and A. A. Schäffer (1993). "Faster sequential genetic linkage computations." Am. J. Hum. Genet. **53**(1): 252-63.

Davies, R. B. (1977). "Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative." Biometrika **64**(2): 247-254.

Davis, S. and D. E. Weeks (1997). "Comparison of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation." Am. J. Hum. Genet. **61**(6): 1431-44.

Dichgans, M. (2007). "Genetics of ischaemic stroke." Lancet Neurol **6**(2): 149-61.

Dudbridge, F. (2008). "Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data." Hum. Hered. **66**(2): 87-98.

Ehret, G. B., P. B. Munroe, K. M. Rice, M. Bochud, A. D. Johnson, D. I. Chasman, A. V. Smith, M. D. Tobin, G. C. Verwoert, S. J. Hwang, V. Pihur, P. Vollenweider, P. F. O'Reilly, N. Amin, J. L. Bragg-Gresham, A. Teumer, N. L. Glazer, L. Launer, J. H. Zhao, Y. Aulchenko, S. Heath, S. Sober, A. Parsa, J. Luan, P. Arora, A. Dehghan, F. Zhang, G. Lucas, A. A. Hicks, A. U. Jackson, J. F. Peden, T. Tanaka, S. H. Wild, I. Rudan, W. Igl, Y. Milaneschi, A. N. Parker, C. Fava, J. C. Chambers, E. R. Fox, M. Kumari, M. J. Go, P. van der Harst, W. H. Kao, M. Sjogren, D. G. Vinay, M. Alexander, Y. Tabara, S. Shaw-Hawkins, P. H. Whincup, Y. Liu, G. Shi, J. Kuusisto, B. Tayo, M. Seielstad, X. Sim, K. D. Nguyen, T. Lehtimaki, G. Matullo, Y. Wu, T. R. Gaunt, N. C. Onland-Moret, M. N. Cooper, C. G. Platou, E. Org, R. Hardy, S. Dahgam, J. Palmen, V. Vitart, P. S. Braund, T. Kuznetsova, C. S. Uiterwaal, A. Adeyemo, W. Palmas, H. Campbell, B. Ludwig, M. Tomaszewski, I. Tzoulaki, N. D. Palmer, T. Aspelund, M. Garcia, Y. P. Chang, J. R. O'Connell, N. I. Steinle, D. E. Grobbee, D. E. Arking, S. L. Kardia, A. C. Morrison, D. Hernandez, S. Najjar, W. L. McArdle, D. Hadley, M. J. Brown, J. M. Connell, A. D. Hingorani, I. N. Day, D. A. Lawlor, J. P. Beilby, R. W. Lawrence, R. Clarke, J. C. Hopewell, H. Ongen, A. W. Dreisbach, Y. Li, J. H. Young, J. C. Bis, M. Kahonen, J. Viikari, L. S. Adair, N. R. Lee, M. H. Chen, M. Olden, C. Pattaro, J. A. Bolton, A. Kottgen, S. Bergmann, V. Mooser, N. Chaturvedi, T. M. Frayling, M. Islam, T. H. Jafar, J. Erdmann, S. R. Kulkarni, S. R. Bornstein, J. Grassler, L. Groop, B. F. Voight, J. Kettunen, P. Howard, A. Taylor, S. Guarrera, F. Ricceri, V. Emilsson, A. Plump, I. Barroso, K. T. Khaw, A. B. Weder, S. C. Hunt, Y. V. Sun, R. N. Bergman, F. S. Collins, L. L. Bonnycastle, L. J. Scott, H. M. Stringham, L. Peltonen, M. Perola, E. Vartiainen, S. M. Brand, J. A. Staessen, T. J. Wang, P. R. Burton, M. S. Artigas, Y. Dong, H. Snieder, X. Wang, H. Zhu, K. K. Lohman, M. E. Rudock, S. R. Heckbert, N. L. Smith, K. L. Wiggins, A. Doumatey, D. Shriner, G. Veldre, M. Viigimaa, S. Kinra, D. Prabhakaran, V. Tripathy, C. D. Langefeld, A. Rosengren, D. S. Thelle, A. M. Corsi, A. Singleton, T. Forrester, G. Hilton, C. A. McKenzie, T. Salako, N. Iwai, Y. Kita, T. Ogihara, T. Ohkubo, T. Okamura, H. Ueshima, S. Umemura, S. Eyheramendy, T. Meitinger, H. E. Wichmann, Y. S. Cho, H. L. Kim, J. Y. Lee, J. Scott, J. S. Sehmi, W. Zhang, B. Hedblad, P. Nilsson, G. D. Smith, A. Wong, N. Narisu, A. Stancakova, L. J. Raffel, J. Yao, S. Kathiresan, C. J. O'Donnell, S. M. Schwartz, M. A. Ikram, W. T. Longstreth, Jr., T. H. Mosley, S. Seshadri, N. R. Shrine, L. V. Wain, M. A. Morken, A. J. Swift, J. Laitinen, I. Prokopenko, P. Zitting, J. A. Cooper, S. E. Humphries, J. Danesh, A. Rasheed, A. Goel, A. Hamsten, H. Watkins, S. J. Bakker, W. H. van Gilst, C. S. Janipalli, K. R. Mani, C. S. Yajnik, A. Hofman, F. U. Mattace-Raso, B. A. Oostra, A. Demirkan, A. Isaacs, F. Rivadeneira, E. G. Lakatta, M. Orru, A. Scuteri, M. Ala-Korpela, A. J. Kangas, L. P. Lyytikainen, P. Soininen, T. Tukiainen, P. Wurtz, R. T. Ong, M. Dorr, H. K.

Kroemer, U. Volker, H. Volzke, P. Galan, S. Hercberg, M. Lathrop, D. Zelenika, P. Deloukas, M. Mangino, T. D. Spector, G. Zhai, J. F. Meschia, M. A. Nalls, P. Sharma, J. Terzic, M. V. Kumar, M. Denniff, E. Zukowska-Szczechowska, L. E. Wagenknecht, F. G. Fowkes, F. J. Charchar, P. E. Schwarz, C. Hayward, X. Guo, C. Rotimi, M. L. Bots, E. Brand, N. J. Samani, O. Polasek, P. J. Talmud, F. Nyberg, D. Kuh, M. Laan, K. Hveem, L. J. Palmer, Y. T. van der Schouw, J. P. Casas, K. L. Mohlke, P. Vineis, O. Raitakari, S. K. Ganesh, T. Y. Wong, E. S. Tai, R. S. Cooper, M. Laakso, D. C. Rao, T. B. Harris, R. W. Morris, A. F. Dominiczak, M. Kivimaki, M. G. Marmot, T. Miki, D. Saleheen, G. R. Chandak, J. Coresh, G. Navis, V. Salomaa, B. G. Han, X. Zhu, J. S. Kooner, O. Melander, P. M. Ridker, S. Bandinelli, U. B. Gyllensten, A. F. Wright, J. F. Wilson, L. Ferrucci, M. Farrall, J. Tuomilehto, P. P. Pramstaller, R. Elosua, N. Soranzo, E. J. Sijbrands, D. Altshuler, R. J. Loos, A. R. Shuldiner, C. Gieger, P. Meneton, A. G. Uitterlinden, N. J. Wareham, V. Gudnason, J. I. Rotter, R. Rettig, M. Uda, D. P. Strachan, J. C. Witteman, A. L. Hartikainen, J. S. Beckmann, E. Boerwinkle, R. S. Vasan, M. Boehnke, M. G. Larson, M. R. Jarvelin, B. M. Psaty, G. R. Abecasis, A. Chakravarti, P. Elliott, C. M. van Duijn, C. Newton-Cheh, D. Levy, M. J. Caulfield, T. Johnson, H. Tang, J. Knowles, M. Hlatky, S. Fortmann, T. L. Assimes, T. Quertermous, A. Go, C. Iribarren, D. Absher, N. Risch, R. Myers, S. Sidney, A. Ziegler, A. Schillert, C. Bickel, C. Sinning, H. J. Rupprecht, K. Lackner, P. Wild, R. Schnabel, S. Blankenberg, T. Zeller, T. Munzel, C. Perret, F. Cambien, L. Tiret, V. Nicaud, C. Proust, A. Uitterlinden, C. van Duijn, J. Whitteman, L. A. Cupples, S. Demissie-Banjaw, V. Ramachandran, A. Smith, A. Folsom, A. Morrison, I. Y. Chen, J. Bis, K. Volcik, K. Rice, K. D. Taylor, K. Marciante, N. Smith, N. Glazer, S. Heckbert, T. Harris, T. Lumley, A. Kong, G. Thorleifsson, G. Thorgeirsson, H. Holm, J. R. Gulcher, K. Stefansson, K. Andersen, S. Gretarsdottir, U. Thorsteinsdottir, M. Preuss, S. Schreiber, I. R. Konig, W. Lieb, C. Hengstenberg, H. Schunkert, M. Fischer, A. Grosshennig, A. Medack, K. Stark, P. Linsel-Nitschke, P. Bruse, Z. Aherrahrou, A. Peters, C. Loley, C. Willenborg, J. Nahrstedt, J. Freyer, S. Gulde, A. Doering, C. Meisinger, N. Klopp, T. Illig, A. Meinitzer, A. Tomaschitz, E. Halperin, H. Dobnig, H. Scharnagl, M. Kleber, R. Laaksonen, S. Pilz, T. B. Grammer, T. Stojakovic, W. Renner, W. Marz, B. O. Bohm, B. R. Winkelmann, K. Winkler, M. M. Hoffmann, D. S. Siscovick, K. Musunuru, M. Barbalic, C. Guiducci, N. Burtt, S. B. Gabriel, A. F. Stewart, G. A. Wells, L. Chen, O. Jarinova, R. Roberts, R. McPherson, S. Dandona, A. D. Pichard, D. J. Rader, J. Devaney, J. M. Lindsay, K. M. Kent, L. Qu, L. Satler, M. S. Burnett, M. Li, M. P. Reilly, R. Wilensky, R. Waksman, S. Epstein, W. Matthai, C. W. Knouff, D. M. Waterworth, H. H. Hakonarson, M. C. Walker, A. S. Hall, A. J. Balmforth, B. J. Wright, C. Nelson, J. R. Thompson, S. G. Ball, J. F. Felix, S. Demissie, L. R. Loehr, W. D. Rosamond, A. R. Folsom, E. Benjamin, Y. S. Aulchenko, T. Haritunians, D. Couper, J. Murabito, Y. A. Wang, B. H. Stricker, J. S. Gottdiener, P. P. Chang, J. T. Willerson, C. A. Boger, C. Fuchsberger, X. Gao, Q. Yang, H. Schmidt, S. Ketkar, G. Pare, E. J. Atkinson, K. Lohman, M. C. Cornelis, N. M. Probst-Hensch, F. Kronenberg, A. Tonjes, G. Eiriksdottir, L. J. Launer, E. Rampersaud, B. D. Mitchell, M. Struchalin, M. Cavalieri, F. Giallauria, J. Metter, J. de Boer, D. Siscovick, M. C. Zillikens, M. Feitosa, M. Province, M. de Andrade, S. T. Turner, P. S. Wild, R. B. Schnabel, S. Wilde, T. F. Munzel, T. S. Leak, W. Koenig, L. Zgaga, T. Zemunik, I. Kolcic, C. Minelli, F. B. Hu, A. Johansson, G. Zaboli, D. Ellinghaus, M. Imboden, D. Nitsch, A. Brandstatter, B. Kollerits, L. Kedenko, R. Magi, M. Stumvoll, P. Kovacs, M. Boban, S. Campbell, K. Endlich, M. Nauck, S. Badola, G. C. Curhan, A. Franke, T. Rochat, B. Paulweber, W. Wang, R. Schmidt, M. G. Shlipak, I. Borecki, B. K. Kramer, U. Gyllensten, N. Hastie, I. M. Heid, C. S. Fox, S. B. Felix, N. Watzinger, G. Homuth, J. Aragam, R. Zweiker, L. Lind, R. J. Rodeheffer, K. H. Greiser, J. W. Deckers, J. Stritzke, K. J. Lackner, E. Ingelsson, I. Kullo, J. Haerting, T. Reffelmann, M. M. Redfield, K. Werdan, G. F. Mitchell, D. K. Arnett, M. Blettner, N. Friedrich, E. J. Benjamin, G. M. Lord, D. P. Gale, M. N. Wass, K. R. Ahmadi, J.

Beckmann, H. J. Bilo, H. T. Cook, I. Cotlarciuc, G. Davey Smith, R. de Silva, G. Deng, O. Devuyst, L. D. Dikkeschei, N. Dimkovic, M. Dockrell, A. Dominiczak, S. Ebrahim, T. Eggermann, J. Floege, N. G. Forouhi, R. T. Gansevoort, X. Han, J. J. Homan van der Heide, B. G. Hepkema, M. Hernandez-Fuentes, E. Hypponen, P. E. de Jong, N. Kleefstra, V. Lagou, M. Lapsley, K. Luttropp, C. Marechal, L. Nordfors, B. W. Penninx, E. Perucha, A. Pouta, P. J. Roderick, A. Ruokonen, S. Sanna, M. Schalling, D. Schlessinger, G. Schlieper, M. A. Seelen, J. H. Smit, P. Stenvinkel, M. J. Sternberg, R. Swaminathan, L. J. Ubink-Veltmaat, C. Wallace, D. Waterworth, K. Zerres, G. Waeber, P. H. Maxwell, M. I. McCarthy and L. Lightstone (2011). "Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk." Nature **478**(7367): 103-9.

Ekelund, J., I. Hovatta, A. Parker, T. Paunio, T. Varilo, R. Martin, J. Suhonen, P. Ellonen, G. Chan, J. S. Sinsheimer, E. Sobel, H. Juvonen, R. Arajärvi, T. Partonen, J. Suvisaari, J. Lönnqvist, J. Meyer and L. Peltonen (2001). "Chromosome 1 loci in Finnish schizophrenia families." Hum. Mol. Genet. **10**(15): 1611-7.

Elston, R. C. and J. Stewart (1971). "A general model for the genetic analysis of pedigree data." Hum. Hered. **21**(6): 523-42.

Enattah, N. S., T. Sahi, E. Savilahti, J. D. Terwilliger, L. Peltonen and I. Järvelä (2002). "Identification of a variant associated with adult-type hypolactasia." Nat. Genet. **30**(2): 233-7.

Fisher, R. A. (1922). "On the Mathematical Foundations of Theoretical Statistics." Philosophical Transactions of the Royal Society of London. Series A **222**: 309-368.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly and D. Altshuler (2002). "The structure of haplotype blocks in the human genome." Science **296**(5576): 2225-9.

Glaser, B. and P. Holmans (2009). "Comparison of methods for combining case-control and family-based association studies." Hum. Hered. **68**(2): 106-16.

Greenberg, D. A., P. Abreu and S. E. Hodge (1998). "The power to detect linkage in complex disease by means of simple LOD-score analyses." Am J Hum Genet **63**(3): 870-9.

Gusella, J. F., N. S. Wexler, P. M. Conneally, S. L. Naylor, M. A. Anderson, R. E. Tanzi, P. C. Watkins, K. Ottina, M. R. Wallace, A. Y. Sakaguchi and et al. (1983). "A polymorphic DNA marker genetically linked to Huntington's disease." Nature **306**(5940): 234-8.

Göring, H. H. and J. D. Terwilliger (2000). "Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified." Am. J. Hum. Genet. **66**(4): 1310-27.

Haataja, R., M. K. Karjalainen, A. Luukkonen, K. Teramo, H. Puttonen, M. Ojaniemi, T. Varilo, B. P. Chaudhari, J. Plunkett, J. C. Murray, S. A. McCarroll, L. Peltonen, L. J. Muglia, A. Palotie and M. Hallman (2011). "Mapping a new spontaneous preterm birth susceptibility gene, IGF1R, using linkage, haplotype sharing, and association analysis." PLoS Genet **7**(2): e1001293.

Haldane, J. B. and C. A. Smith (1947). "A new estimate of the linkage between the genes for colourblindness and haemophilia in man." Ann Eugen **14**(pt 1): 10-31.

Hardy, G. H. (1908). "Mendelian proportions in a mixed population." Science **28**: 49-50.

Hartl, D. L. and A. G. Clark (2007). Principles of Population Genetics, Sinauer Associates, Inc.

Hartong, D. T., E. L. Berson and T. P. Dryja (2006). "Retinitis pigmentosa." Lancet **368**(9549): 1795-809.

Hasstedt, S. J. (1982). "A mixed-model likelihood approximation on large pedigrees." Comput. Biomed. Res. **15**(3): 295-307.

Hodge, S. E. and R. C. Elston (1994). "Lods, wrods, and mods: the interpretation of lod scores calculated under different models." Genet Epidemiol **11**(4): 329-42.

Howson, J. M., B. J. Barratt, J. A. Todd and H. J. Cordell (2005). "Comparison of population- and family-based methods for genetic association analysis in the presence of interacting loci." Genet. Epidemiol. **29**(1): 51-67.

Infante-Rivard, C., L. Mirea and S. B. Bull (2009). "Combining case-control and case-trio data from the same population in genetic association analyses: overview of approaches and illustration with a candidate gene study." Am. J. Epidemiol. **170**(5): 657-64.

Jarvelin, M. R., A. L. Hartikainen-Sorri and P. Rantakallio (1993). "Labour induction policy in hospitals of different levels of specialisation." Br J Obstet Gynaecol **100**(4): 310-5.

Jonasdottir, G., K. Humphreys and J. Palmgren (2007). "Testing association in the presence of linkage--a powerful score for binary traits." Genet. Epidemiol. **31**(6): 528-40.

Kainulainen, K., M. Perola, J. Terwilliger, J. Kaprio, M. Koskenvuo, A. C. Syvänen, E. Vartiainen, L. Peltonen and K. Kontula (1999). "Evidence for involvement of the type 1 angiotensin II receptor locus in essential hypertension." Hypertension **33**(3): 844-9.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti and E. Eskin (2010). "Variance component model to account for sample structure in genome-wide association studies." Nat Genet **42**(4): 348-54.

Kantojarvi, K., I. Kotala, K. Rehnstrom, T. Ylisaukko-Oja, R. Vanhala, T. N. von Wendt, L. von Wendt and I. Jarvela (2011). "Fine mapping of Xq11.1-q21.33 and mutation screening of RPS6KA6, ZNF711, ACSL4, DLG3, and IL1RAPL2 for autism spectrum disorders (ASD)." Autism Res **4**(3): 228-33.

Kaprio, J., S. Sarna, M. Koskenvuo and I. Rantasalo (1978). "The Finnish Twin Registry: formation and compilation, questionnaire study, zygosity determination procedures, and research program." Prog Clin Biol Res **24 Pt B**: 179-84.

Kaunisto, M. A., P. J. Tikka, M. Kallela, S. M. Leal, J. C. Papp, A. Korhonen, E. Hämäläinen, H. Harno, H. Havanka, M. Nissilä, E. Säkö, M. Ilmavirta, J. Kaprio, M. Färkkilä, R. A. Ophoff, A. Palotie and M. Wessman (2005). "Chromosome 19p13 loci in Finnish migraine with aura families." Am. J. Med. Genet. B Neuropsychiatr. Genet. **132**(1): 85-9.

Kettunen, J., M. Perola, N. G. Martin, B. K. Cornes, S. G. Wilson, G. W. Montgomery, B. Benyamin, J. R. Harris, D. Boomsma, G. Willemsen, J. J. Hottenga, P. E. Slagboom, K. Christensen, K. O. Kyvik, T. I. Sorensen, N. L. Pedersen, P. K. Magnusson, T. Andrew, T. D. Spector, E. Widen, K. Silventoinen, J. Kaprio, A. Palotie and L. Peltonen (2009). "Multicenter dizygotic twin cohort study confirms two linkage susceptibility loci for body mass index at 3q29 and 7q36 and identifies three further potential novel loci." Int J Obes (Lond) **33**(11): 1235-42.

Knaapila, A., K. Keskitalo, M. Kallela, M. Wessman, S. Sammalisto, T. Hiekkalinna, A. Palotie, L. Peltonen, H. Tuorila and M. Perola (2007). "Genetic component of identification, intensity and pleasantness of odours: a Finnish family study." Eur J Hum Genet **15**(5): 596-602.

Knapp, M., S. A. Seuchter and M. P. Baur (1994). "Linkage analysis in nuclear families. 2: Relationship between affected sib-pair tests and lod score analysis." Hum. Hered. **44**(1): 44-51.

Kruglyak, L. (1997). "Nonparametric linkage tests are model free." Am. J. Hum. Genet. **61**(1): 254-5.

Kruglyak, L., M. J. Daly, M. P. Reeve-Daly and E. S. Lander (1996). "Parametric and nonparametric linkage analysis: a unified multipoint approach." Am. J. Hum. Genet. **58**(6): 1347-63.

Kruglyak, L. and E. S. Lander (1998). "Faster multipoint linkage analysis using Fourier transforms." J. Comput. Biol. **5**(1): 1-7.

Kuokkanen, S., M. Sundvall, J. D. Terwilliger, P. J. Tienari, J. Wikström, R. Holmdahl, U. Pettersson and L. Peltonen (1996). "A putative vulnerability locus to multiple sclerosis maps to 5p14-p12 in a region syntenic to the murine locus Eae2." Nat. Genet. **13**(4): 477-80.

Laird, N. M., S. Horvath and X. Xu (2000). "Implementing a unified approach to family-based tests of association." Genet. Epidemiol. **19 Suppl 1**: S36-42.

Lake, S. L., D. Blacker and N. M. Laird (2000). "Family-based tests of association in the presence of linkage." Am J Hum Genet **67**(6): 1515-25.

Lalouel, J. M. (1979). GEMINI - a computer program for optimization of a nonlinear function, Department of Medical Biophysics and Computing, University of Utah, Salt Lake City.

Lambert, B. W., J. D. Terwilliger and K. M. Weiss (2008). "ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth." Bioinformatics **24**(16): 1821-1822.

Lander, E. and L. Kruglyak (1995). "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results." Nat Genet **11**(3): 241-7.

Lander, E. S. and P. Green (1987). "Construction of multilocus genetic linkage maps in humans." Proc. Natl. Acad. Sci. U. S. A. **84**(8): 2363-7.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi and Y. J. Chen (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lange, K., R. Cantor, S. Horvath, M. Perola, C. Sabatti, J. Sinsheimer and E. Sobel (2001). "Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets." Am. J. Hum. Genet. **69(Supp):504**.

Lasky-Su, J., S. Won, E. Mick, R. J. Anney, B. Franke, B. Neale, J. Biederman, S. L. Smalley, S. K. Loo, A. Todorov, S. V. Faraone, S. T. Weiss and C. Lange (2010). "On genome-wide association studies for family-based designs: an integrative analysis approach combining ascertained family samples with unselected controls." Am J Hum Genet **86**(4): 573-80.

Lathrop, G. M. and J. M. Lalouel (1984). "Easy calculations of lod scores and genetic risks on small computers." Am. J. Hum. Genet. **36**(2): 460-5.

Lathrop, G. M., J. M. Lalouel, C. Julier and J. Ott (1984). "Strategies for multilocus linkage analysis in humans." Proc. Natl. Acad. Sci. U. S. A. **81**(11): 3443-6.

Lathrop, G. M., J. M. Lalouel, C. Julier and J. Ott (1985). "Multilocus linkage analysis in humans: detection of linkage and estimation of recombination." Am. J. Hum. Genet. **37**(3): 482-98.

Lathrop, G. M., J. M. Lalouel and R. L. White (1986). "Construction of human linkage maps: likelihood calculations for multilocus linkage analysis." Genet. Epidemiol. **3**(1): 39-52.

Lewontin, R. C. (1964). "The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models." Genetics **49**(1): 49-67.

Lewontin, R. C. and K. Kojima (1960). "The evolutionary dynamics of complex polymorphisms." Evolution(14): 458-472.

Li, M., M. Boehnke and G. R. Abecasis (2005). "Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal." Am. J. Hum. Genet. **76**(6): 934-49.

Li, M., M. Boehnke and G. R. Abecasis (2006). "Efficient study designs for test of genetic association using sibship data and unrelated cases and controls." Am. J. Hum. Genet. **78**(5): 778-92.

Loukola, A., U. Broms, H. Maunu, E. Widen, K. Heikkila, M. Siivola, A. Salo, M. L. Pergadia, E. Nyman, S. Sammalisto, M. Perola, A. Agrawal, A. C. Heath, N. G. Martin, P. A. Madden, L. Peltonen and J. Kaprio (2008). "Linkage of nicotine dependence and smoking behavior on 10q, 7q and 11p in twins with homogeneous genetic background." Pharmacogenomics J **8**(3): 209-19.

Magnusson, P. K., M. Boman, U. de Faire, M. Perola, L. Peltonen and N. L. Pedersen (2008). "Genome-wide search for QTLs for apolipoprotein A-I level in elderly Swedish DZ twins: evidence of female-specific locus on 15q11-13." Eur J Hum Genet **16**(9): 1103-10.

McNemar, Q. (1947). "Note on the sampling error of the difference between correlated proportions or percentages." Psychometrika(12): 153-157.

Millstein, J., K. D. Siegmund, D. V. Conti and W. J. Gauderman (2005). "Testing association and linkage using affected-sib-parent study designs." Genet. Epidemiol. **29**(3): 225-33.

Morton, N. E. (1955). "Sequential tests for the detection of linkage." Am. J. Hum. Genet. **7**(3): 277-318.

Mukhopadhyay, N., L. Almasy, M. Schroeder, W. P. Mulvihill and D. E. Weeks (2005). "Mega2: data-handling for facilitating genetic linkage and association analyses." Bioinformatics **21**(10): 2556-7.

Murray, J. M., K. E. Davies, P. S. Harper, L. Meredith, C. R. Mueller and R. Williamson (1982). "Linkage relationship of a cloned DNA sequence on the short arm of the X chromosome to Duchenne muscular dystrophy." Nature **300**(5887): 69-71.

Nelder, J. A. and R. Mead (1965). "A Simplex Method for Function Minimization." Computer J. **7**(4): 6.

Neymann, J. and E. Pearson (1933). "On the Problem of the Most Efficient Tests of Statistical Hypotheses." Philosophical Transactions of the Royal Society of London. Series A(231): 289–337.

Nicodemus, K. K., A. Luna and Y. Y. Shugart (2007). "An evaluation of power and type I error of single-nucleotide polymorphism transmission/disequilibrium-based statistical methods under different family structures, missing parental data, and population stratification." Am. J. Hum. Genet. **80**(1): 178-85.

Norio, R., H. R. Nevanlinna and J. Perheentupa (1973). "Hereditary diseases in Finland; rare flora in rare soul." Ann Clin Res **5**(3): 109-41.

O'Connell, J. R. and D. E. Weeks (1998). "PedCheck: a program for identification of genotype incompatibilities in linkage analysis." Am. J. Hum. Genet. **63**(1): 259-66.

Ott, J. (1989). "Computer-simulation methods in human linkage analysis." Proc. Natl. Acad. Sci. U. S. A. **86**(11): 4175-8.

Ott, J. (1999). Analysis of Human Genetic Linkage, Johns Hopkins University Press, Baltimore.

Ott, J., Y. Kamatani and M. Lathrop (2011). "Family-based designs for genome-wide association studies." Nat Rev Genet **12**(7): 465-74.

Paunio, T., A. Tuulio-Henriksson, T. Hiekkalinna, M. Perola, T. Varilo, T. Partonen, T. D. Cannon, J. Lonnqvist and L. Peltonen (2004). "Search for cognitive trait components of schizophrenia reveals a locus for verbal learning and memory on 4q and for visual working memory on 2q." Hum Mol Genet **13**(16): 1693-702.

Peltonen, L., A. Jalanko and T. Varilo (1999). "Molecular genetics of the Finnish disease heritage." Hum Mol Genet **8**(10): 1913-23.

Perola, M., S. Sammalisto, T. Hiekkalinna, N. G. Martin, P. M. Visscher, G. W. Montgomery, B. Benyamin, J. R. Harris, D. Boomsma, G. Willemsen, J. J. Hottenga, K. Christensen, K. O. Kyvik, T. I. Sorensen, N. L. Pedersen, P. K. Magnusson, T. D. Spector, E. Widen, K. Silventoinen, J. Kaprio, A. Palotie and L. Peltonen (2007). "Combined genome scans for body stature in 6,602 European twins: evidence for common Caucasian loci." PLoS Genet **3**(6): e97.

Polvi, A., A. Siren, M. Kallela, H. Rantala, V. Artto, E. M. Sobel, A. Palotie, A. E. Lehesjoki and M. Wessman (2012). "Shared loci for migraine and epilepsy on chromosomes 14q12-q23 and 12q24.2-q24.3." Neurology **78**(3): 202-9.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly and P. C. Sham (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am. J. Hum. Genet. **81**(3): 559-75.

Rabinowitz, D. and N. Laird (2000). "A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information." Hum. Hered. **50**(4): 211-23.

Raitakari, O. T., M. Juonala, T. Ronnemaa, L. Keltikangas-Jarvinen, L. Rasanen, M. Pietikainen, N. Hutri-Kahonen, L. Taittonen, E. Jokinen, J. Marniemi, A. Jula, R. Telama, M. Kahonen, T. Lehtimaki, H. K. Akerblom and J. S. Viikari (2008). "Cohort profile: the cardiovascular risk in Young Finns Study." Int J Epidemiol **37**(6): 1220-6.

Rantakallio, P. (1969). "Groups at risk in low birth weight infants and perinatal mortality." Acta Paediatr Scand **193**: Suppl 193:1+.

Rehnstrom, K., T. Ylisaukko-oja, T. Nieminen-von Wendt, S. Sarenius, T. Kallman, E. Kempas, L. von Wendt, L. Peltonen and I. Jarvela (2006). "Independent replication and initial fine mapping of 3p21-24 in Asperger syndrome." J Med Genet **43**(2): e6.

Reich, D. E. and E. S. Lander (2001). "On the allelic spectrum of human disease." Trends Genet. **17**(9): 502-10.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science **273**(5281): 1516-7.

Rubinstein, P., M. Walker, C. Carpenter, C. Carrier, J. Krassner, C. T. Falk and F. Ginsberg (1981). "Genetics of HLA disease associations: The use of the haplotype relative risk (HRR) and the "haplo-delta" (Dh) estimates in juvenile diabetes from three racial groups." Hum. Immunol. **3**: 384.

Sammalisto, S., T. Hiekkalinna, K. Schwander, S. Kardia, A. B. Weder, B. L. Rodriguez, A. Doria, J. A. Kelly, G. R. Bruner, J. B. Harley, S. Redline, E. K. Larkin, S. R. Patel, A. J. Ewan, J. L. Weber, M. Perola and L. Peltonen (2009). "Genome-wide linkage screen for stature and body mass index in 3.032 families: evidence for sex- and population-specific genetic effects." Eur J Hum Genet **17**(2): 258-66.

Sammalisto, S., T. Hiekkalinna, E. Suviolahti, K. Sood, A. Metzidis, P. Pajukanta, H. E. Lilja, A. Soro-Paavonen, M. R. Taskinen, T. Tuomi, P. Almgren, M. Orho-Melander, L. Groop, L. Peltonen and M. Perola (2005). "A male-specific quantitative trait locus on 1p21 controlling human stature." J Med Genet **42**(12): 932-9.

Satsangi, J., M. Parkes, E. Louis, L. Hashimoto, N. Kato, K. Welsh, J. D. Terwilliger, G. M. Lathrop, J. I. Bell and D. P. Jewell (1996). "Two stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12." Nat. Genet. **14**(2): 199-202.

Schäffer, A. A., S. K. Gupta, K. Shriram and R. W. Cottingham, Jr. (1994). "Avoiding recomputation in linkage analysis." Hum. Hered. **44**(4): 225-37.

Sham, P. (1998). "Statistics in Human Genetics." Arnold (Hodder Headline Group), London.

Sinsheimer, J. S., J. Blangero and K. Lange (2000). "Gamete-competition models." Am J Hum Genet **66**(3): 1168-72.

Smith, C. A. B. (1953). "Detection of linkage in human genetics." Journal of the Royal Statistical Society. Series B **15**: 153-192.

Spencer, C. C., Z. Su, P. Donnelly and J. Marchini (2009). "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip." PLoS Genet **5**(5): e1000477.

Spielman, R. S. and W. J. Ewens (1996). "The TDT and other family-based tests for linkage disequilibrium and association." Am. J. Hum. Genet. **59**(5): 983-9.

Spielman, R. S., R. E. McGinnis and W. J. Ewens (1993). "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)." Am. J. Hum. Genet. **52**(3): 506-16.

Stenson, P. D., E. V. Ball, K. Howells, A. D. Phillips, M. Mort and D. N. Cooper (2009). "The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics." Hum Genomics **4**(2): 69-72.

Styrkarsdottir, U., J. B. Cazier, A. Kong, O. Rolfsson, H. Larsen, E. Bjarnadottir, V. D. Johannsdottir, M. S. Sigurdardottir, Y. Bagger, C. Christiansen, I. Reynisdottir, S. F. Grant, K. Jonasson, M. L. Frigge, J. R. Gulcher, G. Sigurdsson and K. Stefansson (2003). "Linkage of osteoporosis to chromosome 20p12 and association to BMP2." PLoS Biol **1**(3): E69.

Terwilliger, J. D. (2001). "On the resolution and feasibility of genome scanning approaches." Adv. Genet. **42**: 351-91.

Terwilliger, J. D. and H. H. Göring (2000). "Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design." Hum. Biol. **72**(1): 63-132.

Terwilliger, J. D. and J. Ott (1992). "A haplotype-based 'haplotype relative risk' approach to detecting allelic associations." Hum. Hered. **42**(6): 337-46.

Terwilliger, J. D. and J. Ott (1994). Handbook of Human Genetic Linkage, Johns Hopkins University Press, Baltimore.

Terwilliger, J. D. and K. M. Weiss (1998). "Linkage disequilibrium mapping of complex disease: fantasy or reality?" Curr. Opin. Biotechnol. **9**(6): 578-94.

Tienari, P. J., J. D. Terwilliger, J. Ott, J. Palo and L. Peltonen (1994). "Two-locus linkage analysis in multiple sclerosis (MS)." Genomics **19**(2): 320-5.

Tikka-Kleemola, P., V. Artto, S. Vepsalainen, E. M. Sobel, S. Raty, M. A. Kaunisto, V. Anttila, E. Hamalainen, M. L. Sumelahti, M. Ilmavirta, M. Farkkila, M. Kallela, A. Palotie and M. Wessman (2010). "A visual migraine aura locus maps to 9q21-q22." Neurology **74**(15): 1171-7.

Torczon, V. (1991). "On the convergence of the multidirectional search algorithm." SIAM J. Optim. **1**(1): 123-145.

Trembath, R. C., R. L. Clough, J. L. Rosbotham, A. B. Jones, R. D. Camp, A. Frodsham, J. Browne, R. Barber, J. Terwilliger, G. M. Lathrop and J. N. Barker (1997). "Identification of a major susceptibility locus on chromosome 6p and evidence for further disease loci revealed by a two stage genome-wide search in psoriasis." Hum Mol Genet **6**(5): 813-20.

Tsui, L. C., M. Buchwald, D. Barker, J. C. Braman, R. Knowlton, J. W. Schumm, H. Eiberg, J. Mohr, D. Kennedy, N. Plavsic and et al. (1985). "Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker." Science **230**(4729): 1054-7.

Turunen, J. A., K. Rehnstrom, H. Kilpinen, M. Kuokkanen, E. Kempas and T. Ylisaukko-Oja (2008). "Mitochondrial aspartate/glutamate carrier SLC25A12 gene is associated with autism." Autism Res **1**(3): 189-92.

Vaara, S., M. S. Nieminen, M. L. Lokki, M. Perola, P. J. Pussinen, J. Allonen, O. Parkkonen and J. Sinisalo (2011). "Cohort Profile: The Corogene study." Int J Epidemiol.

Vartiainen, E., P. Jousilahti, G. Alfthan, J. Sundvall, P. Pietinen and P. Puska (2000). "Cardiovascular risk factor changes in Finland, 1972-1997." Int J Epidemiol **29**(1): 49-56.

Wedenoja, J., A. Loukola, A. Tuulio-Henriksson, T. Paunio, J. Ekelund, K. Silander, T. Varilo, K. Heikkila, J. Suvisaari, T. Partonen, J. Lonnqvist and L. Peltonen (2008). "Replication of linkage on chromosome 7q22 and association of the regional Reelin gene with working memory in schizophrenia families." Mol Psychiatry **13**(7): 673-84.

Weeks, D. E., J. Ott and G. M. Lathrop (1990). "SLINK: a general simulation program for linkage analysis." Am. J. Hum. Genet. **A204**(47:A204).

Weinberg, W. (1908). "Über den Nachweis der Vererbung beim Menschen. ." Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg **64**: 368-382.

Weiss, K. M. (1993). Genetic Variation and Human Disease: Principles and evolutionary approaches. Cambridge, Cambridge University Press.

Weiss, K. M. and J. D. Terwilliger (2000). "How many diseases does it take to map a gene with SNPs?" Nat. Genet. **26**(2): 151-7.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L.

Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Wessman, M., C. Forsblom, M. A. Kaunisto, J. Soderlund, J. Ilonen, R. Sallinen, T. Hiekkalinna, M. Parkkonen, A. P. Maxwell, L. Tarnow, H. H. Parving, S. Hadjadj, M. Marre, L. Peltonen and P. H. Groop (2011). "Novel susceptibility locus at 22q11 for diabetic nephropathy in type 1 diabetes." PLoS One **6**(9): e24053.

Wessman, M., M. Kallela, M. A. Kaunisto, P. Marttila, E. Sobel, J. Hartiala, G. Oswell, S. M. Leal, J. C. Papp, E. Hämäläinen, P. Broas, G. Joslyn, I. Hovatta, T. Hiekkalinna, J. Kaprio, J. Ott, R. M. Cantor, J. A. Zwart, M. Ilmavirta, H. Havanka, M. Färkkilä, L. Peltonen and A. Palotie (2002). "A susceptibility locus for migraine with aura, on chromosome 4q24." Am. J. Hum. Genet. **70**(3): 652-62.

WHO (2011). Global Atlas on cardiovascular disease prevention and control.

Wider, C., S. Melquist, M. Hauf, A. Solida, S. A. Cobb, J. M. Kachergus, J. Gass, K. D. Coon, M. Baker, A. Cannon, D. A. Stephan, D. F. Schorderet, J. Ghika, P. R. Burkhard, G. Kapatos, M. Hutton, M. J. Farrer, Z. K. Wszolek and F. J. Vingerhoets (2008). "Study of a Swiss dopa-responsive dystonia family with a deletion in GCH1: redefining DYT14 as DYT5." Neurology **70**(16 Pt 2): 1377-83.

Wigginton, J. E. and G. R. Abecasis (2005). "PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data." Bioinformatics **21**(16): 3445-7.

Wilks, S. S. (1938). "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." The Annals of Mathematical Statistics **9**: 60-62.

Visscher, P. M., M. A. Brown, M. I. McCarthy and J. Yang (2012). "Five Years of GWAS Discovery." Am J Hum Genet **90**(1): 7-24.

Wright, J. E., Jr., K. Johnson, A. Hollister and B. May (1983). "Meiotic models to explain classical linkage, pseudolinkage, and chromosome pairing in tetraploid derivative salmonid genomes." Isozymes Curr Top Biol Med Res **10**: 239-60.

Wright, S. (1922). "Coefficients of inbreeding and relationship." American Naturalist **56**: 330-338.

Yang, J., S. H. Lee, M. E. Goddard and P. M. Visscher (2011). "GCTA: a tool for genome-wide complex trait analysis." Am J Hum Genet **88**(1): 76-82.

Ylisaukko-oja, T., M. Alarcon, R. M. Cantor, M. Auranen, R. Vanhala, E. Kempas, L. von Wendt, I. Järvelä, D. H. Geschwind and L. Peltonen (2006). "Search for autism loci by combined analysis of Autism Genetic Resource Exchange and Finnish families." Ann. Neurol. **59**(1): 145-55.