

## Family-based designs for genome-wide association studies

Jurg Ott\*, Yoichiro Kamatani<sup>†</sup> and Mark Lathrop<sup>‡§</sup>

**Abstract** | Association mapping has successfully identified common SNPs associated with many diseases. However, the inability of this class of variation to account for most of the supposed heritability has led to a renewed interest in methods — primarily linkage analysis — to detect rare variants. Family designs allow for control of population stratification, investigations of questions such as parent-of-origin effects and other applications that are imperfectly or not readily addressed in case-control association studies. This article guides readers through the interface between linkage and association analysis, reviews the new methodologies and provides useful guidelines for applications. Just as effective SNP-genotyping tools helped to realize the potential of association studies, next-generation sequencing tools will benefit genetic studies by improving the power of family-based approaches.

### Linkage analysis

A family-based method to search for the chromosomal location of a trait locus by demonstrating co-segregation of the disease with genetic markers of known chromosomal location.

Linkage analysis has a long history<sup>1</sup> and has been hugely successful for mapping Mendelian traits such as familial hypercholesterolaemia<sup>2–4</sup>, Huntington's disease<sup>5</sup> and cystic fibrosis<sup>6</sup>. Linkage analysis has also led to many notable successes in mapping variants that confer susceptibility to common disease, despite the many challenges presented by this application. Among these findings is the identification of the role of the insulin (*INS*) gene in insulin-dependent diabetes mellitus<sup>7,8</sup>, the *BRCA1* and *BRCA2* genes in breast cancer<sup>9</sup>, apolipoprotein E (*APOE*) in Alzheimer's disease<sup>10</sup> and nucleotide-binding oligomerization domain containing 2 (*NOD2*; also known as *CARD15*) in Crohn's disease<sup>11</sup>. Advances in genotyping and sequencing are providing further impetus to linkage studies; for example, by improving the ability to identify mutations responsible for Mendelian diseases.

It was pointed out 15 years ago that for disease genes of weak effect, association studies are sometimes more powerful than linkage studies<sup>12</sup>. Some researchers have spoken out about the distinct advantages of family studies (see, for example, REF. 13). Nevertheless, over the past few years, linkage analysis has lost its predominance in favour of linkage disequilibrium (LD) association mapping, abbreviated to 'association mapping'.

Association studies are now routinely carried out on a genome-wide basis. These genome-wide association studies (GWA studies) are performed by genotyping many hundreds of thousands of SNPs in large case-control

populations, often followed by imputation of genotypes at several million sites<sup>14</sup>. By December 2010, 1,212 GWA studies had reached significance levels of  $P \leq 5 \times 10^{-8}$  for 210 traits<sup>15,16</sup>. However, the inability of the variants detected by GWA studies to account for much of the heritability of most common disorders<sup>17</sup> has led to a renewed interest in linkage and other family-based methods that can be deployed in parallel or in combination with association studies. In particular, present GWA approaches have difficulty detecting rare risk variants through LD with common SNP markers, but such variants can be found by linkage (BOX 1). For example, GWA studies have not identified *BRCA1* and *BRCA2*, which were identified through linkage and are the most predominant of the known breast cancer susceptibility genes. The increasing amount of sequence data that is becoming available from many patients through low-cost exomic or whole-genome resequencing<sup>18</sup> will help to identify rare disease-associated variants that, although individually rare, may be sufficiently clustered within a gene or small genomic region to be identified statistically as a mutational 'hot spot' of variation<sup>19</sup>. However, when disease-related genetic variants occur infrequently within a gene, conclusive evidence of the disease link will require observation of co-inheritance with disease in families.

In general, a combination of linkage and association methodologies should provide the most robust and powerful approach to identify and characterize the full range of disease-susceptibility variants. Family study designs

\*Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, 4A Datun Road, Chaoyang District, Beijing 100101, China.

†Fondation Jean Dausset – Centre d'Etude du Polymorphisme Humain, 27 rue Juliette Dodu, Paris 75012, France.

‡CEA-IC-Centre National de Génotypage, 2 rue Gaston Crémieux, CP 5721, 91057 Evry Cedex, France.

Correspondence to J.O.  
e-mail: [ottjurg@psych.ac.cn](mailto:ottjurg@psych.ac.cn)

doi: 10.1038/nrg2989

Published online 1 June 2011

contribute to this combined approach by providing not only the ability to enrich for genetic loci containing rare variants, as mentioned above, but also by: providing methods to control for heterogeneity and population stratification; allowing direct estimates of the genetic contribution of different loci; making it possible to follow the transmission of variants with phenotypes; revealing the effects of parental origin of alleles<sup>20,21</sup>; and revealing the transmission of phenotypes that modify disease severity<sup>22</sup>, in addition to other applications. This combined use of linkage and association analysis has been particularly successful in genetic studies of Crohn's disease and fetal haemoglobin (HbF), as described later in the article.

The next section briefly discusses the basis for linkage methodology. This is followed by an account of family-based analysis, the development of the well-known transmission disequilibrium test (TDT) and other family-based approaches, and the emergence of combined linkage and association analysis. Linkage strategies are

becoming particularly important now that genome-wide sequencing is being applied to small numbers of individuals. For example, in a recent article, a combination of linkage analysis and exome sequencing of a single offspring led to the identification of a single gene defect leading to T cell deficiency and subsequent Kaposi's sarcoma<sup>23</sup>. Family-based mapping methods have been reviewed previously<sup>24–26</sup>; here we consider their current status, including newer developments<sup>27–29</sup>.

## Linkage-study designs

Because family-based designs enrich for genetic effects, studies based on familial cases have more power than unrelated cases to detect genetic effects given an equivalent number of sampling units<sup>30</sup>. However, family collections are more laborious to acquire than case–control collections. The difficulty of obtaining collections for families depends on which of the many types of pedigree structure is chosen. Typical family designs include:

### Linkage disequilibrium

(LD). The nonrandom combinations of alleles at different loci on a chromosome. LD arises from the evolutionary history of the population and decays across generations at a rate proportional to the degree of linkage between the loci.

### Genome-wide association studies

(GWA studies). An approach that tests the whole genome for a statistical association between a marker and a trait in unrelated cases and controls. It is designed to identify associations with traits, such as common diseases.

### Imputation

The process of inferring genotypes and/or haplotypes at untyped loci based on neighbouring loci in strong linkage disequilibrium with the untyped loci. The assumption is made that no recombination has occurred between these closely spaced loci.

### Heterogeneity

Substructure within a population. This may be due to population admixture (with different allele frequencies occurring in subpopulations), different environmental contributing conditions or different genetic factors leading to disease.

### Population stratification


Heterogeneity within a population.

## Box 1 | Linkage analysis versus association analysis

Although genetic linkage and association mapping are based on the same underlying phenomenon — namely, recombination — in practice these approaches have different characteristics for localizing trait loci, each with particular advantages and disadvantages.

In linkage analysis, recombination events are directly observed (or inferred) in a family pedigree within a limited number of generations, whereas in association analysis we make use of the consequences of non-recombination over many past generations within a short interval surrounding a disease locus. Importantly, the difference between the two approaches often leads to the identification of different classes of disease-related genetic variants. These differences could provide at least a partial explanation for the often poor correspondence between the susceptibility loci identified in genetic linkage and genome-wide association studies.

The table compares linkage analysis, represented here by an affected sibling pair (sib-pair) with one parent heterozygous at a specific marker, and genetic association analysis, represented by a trio family with one affected offspring and a heterozygous parent. In affected sib-pairs, what is estimated is the proportion of pairs sharing an allele that is identical by descent; that is, as copies of one parental allele, where the identity of the shared allele is irrelevant. The sharing proportion is a simple inverse function of the recombination fraction: high sharing proportion corresponds to a small recombination fraction, which in turn points to genetic linkage. In trio families, the estimated quantity is the proportion of affected offspring inheriting a specific marker allele: an excess of over 50% is indicative of association.

		
Property of mapping approach	Linkage analysis	Association analysis
Data type studied	Relatives	Unrelated or related individuals
Relevant parameter	Recombination fraction	Association statistic
Range of effect detected (linkage or association)	Long ( $\leq 5$ Mb)	Short ( $\leq 100$ kb)
Number of markers required for genome-wide coverage	Moderate (500–1,000)	Large ( $> 100,000$ )
Statistics used	Cumbersome (requires tailor-made likelihood methods)	Elegant; can use the range of classical statistical tools
Dealing with correlated markers	Pose problems in presence of ungenotyped individuals	Can be handled efficiently
Biological basis of approach	Observe (or infer) recombination in pedigree data	Exploit unobserved recombination events in past generations
Dealing with allelic heterogeneity	Not a problem	Reduces power
Detecting genotyping errors	Potentially detected as Mendelian inconsistencies	Potentially detected only in family data, but not in case–control data
Most suitable application	Rare, dominant traits	Common traits

# Likelihood methods

Methods that are based on statistical models and that estimate parameters in those models

# Conditioning

In likelihood analysis, conditioning on relatives is based on calculating conditional probabilities when given known or assumed information on relatives of an individual.

# Hidden Markov model

A statistical model that assumes an underlying (unobserved) Markov process; that is, a sequence of events such that a specific event depends only on the one immediately preceding it.

# Markov chain Monte Carlo algorithms

(MCMC algorithms). Computer-based random processes that simulate Markov models.

# Segregation analysis

This refers to estimating segregation ratios (the ratio of presence versus absence of a heritable trait among offspring of a specific set of parents) and testing whether they are equal to some ratios expected under Mendelian laws. For example, when one of two parents is affected with a rare dominant Mendelian disorder and the other is not, the expected segregation ratio among their offspring is 1:1.

# Ascertainment

The process used to find or select subjects for inclusion in a genetic study.

# Penetrances

The conditional probability of a phenotype (specifically, the probability of being affected with disease), given an underlying genotype.

# Recombination mapping

The process that assigns a trait locus to a restricted genomic region by genetic linkage analysis.

# Multiplex families

Families that contain more than one affected individual.

parent–offspring trios; affected sibling pairs (sib-pairs); unselected sib-pairs or related individuals selected from the extremes of a quantitative trait distribution (for example, concordant or discordant sib-pairs); extended pedigrees with multiple affected individuals; consanguineous families; and families obtained from isolated populations. One of these designs, or a combination of them, may be chosen depending on the questions to be investigated. For example, a design involving concordant sib-pairs to map QTLs might be combined with unselected sib-pairs to establish the proportion of the heritability that is accounted for by the mapped loci.

Linkage studies rely on a statistical evaluation of the evidence in favour of the co-segregation of marker loci with a trait in one or more of the family designs outlined above. To do this, multiple techniques adapted to the particular study design or analysis goals may be required.

**Parametric linkage models.** Parametric linkage analysis assumes a particular disease model — for example, recessive or dominant inheritance at a disease locus — and uses likelihood methods to test for the non-independent assortment of markers and disease.

Likelihood calculations can be made for large pedigrees and for a restricted number of markers using ‘peeling’ (REFS 31,32), a method for calculating likelihoods in pedigrees that uses conditioning on relatives. For example, calculations can be carried out on the offspring of two parents at the bottom of a pedigree; results are then carried forward to the sibship containing one of the parents as a sibling, and so on, so that the results for the total pedigree data are obtained in a recursive manner. Likelihood calculations are also possible for smaller pedigrees and for a larger number of markers using a hidden Markov model that uses conditioning on markers<sup>33</sup>; that is, recursing is done over markers rather than over sibships. Although the above methods provide numerically exact results, approximate calculations can be made in arbitrarily large, complex pedigrees and marker data sets using Markov chain Monte Carlo algorithms (MCMC algorithms)<sup>34</sup> which are based on computer simulations. Combined segregation analysis and linkage analysis can be used to estimate the underlying parameters of the disease model from the data, but this requires that ascertainment be taken into account<sup>35</sup>. This is because family pedigrees are usually collected to contain large numbers of individuals that are affected by disease, which potentially distorts estimates of parameters such as allele frequencies.

Recessive or dominant inheritance, affected-only analysis and disease heterogeneity can be taken into account by assuming specific patterns of penetrances at the disease locus. It is often important to evaluate locus heterogeneity in linkage studies, and specific statistical techniques are available for this purpose<sup>36,37</sup>. One of the earliest examples of locus heterogeneity was identified in Charcot–Marie–Tooth neuropathy<sup>38</sup>; multiple genetic loci are known for this condition but, in each family pedigree, the trait is generally due to only one locus. A more recent example is non-syndromic hearing impairment<sup>39</sup>, for which numerous dominant and recessive loci have been identified.

Linkage models can be extended to consider age of onset (and other effects) by allowing penetrance values to depend on age (or on other classification variables). Similarly, different penetrance values can be introduced for male and female meioses to examine parent-of-origin effects<sup>40</sup>. In large family pedigrees with missing genotype data, a methodology for analysing parent-of-origin effects has been developed by Kong *et al.*<sup>21</sup> and applied to the analysis of Icelandic samples; several other methodologies are being actively developed and applied<sup>41,42</sup>. Some disease models specify effects of genes that are unlinked to the marker loci (so-called polygenic background effects); the complex computations that are necessary to account for these effects have been implemented into computer programs<sup>43,44</sup>.

In principle, parametric linkage analysis estimates the recombination fraction between two or more loci, and such estimates tend to be biased if the modes of inheritance have been mis-specified<sup>45</sup>. However, as described in the previous section, parametric linkage analysis is flexible and can be tricked into carrying out analyses for which it was not originally designed; thus, the ‘requirement’ for estimating recombination fractions is not a shortcoming.

Another important feature of parametric linkage analysis is that it permits refined recombination mapping of a locus to a restricted genomic region, a crucial step in the identification of disease-predisposing genes. This technique has been widely used to positionally clone Mendelian disease loci, but it is also important in the analysis of complex disease. A recent example is the identification of a gene involved in neuroblastoma, a rare childhood cancer that sometimes clusters in families. In a linkage study of 18 multiplex families exhibiting dominant patterns of inheritance, refined mapping of recombination events located a likely disease-predisposing gene in a region on chromosome 2p23–p24 spanning 16.2 Mb and containing 104 genes<sup>46</sup>. Sequencing of genes from this region in members of eight of the families revealed mutations in the anaplastic lymphoma kinase (*ALK*) gene. Subsequent resequencing revealed frequent occurrences of somatic mutations of *ALK* in patients with neuroblastoma<sup>47–50</sup>, which demonstrates that this gene is important both in heritable and non-heritable forms of the disease. Diagnostic testing for neuroblastoma based on *ALK* is feasible<sup>51</sup>, and therapeutic targeting of *ALK* fusion genes is emerging as a possibility<sup>52</sup>. The fact that whole-genome association studies have identified other loci potentially implicated in neuroblastoma<sup>46,53,54</sup>, but not *ALK*, is a further illustration of the complementary nature of linkage and association approaches.

**Non-parametric linkage analysis.** Non-parametric (parameter-free) linkage analysis compares identity-by-descent between affected relatives at a specific location in the genome, as estimated from the marker data, to see whether they are significantly different from Mendelian expectations. Nonrandom assortment of the markers and the trait is expected to occur if markers are linked to a disease locus.

Although non-parametric methods are widely assumed to be more robust than parametric methods, complications arise in the former regarding how statistics should be calculated and assessed when multiple related individuals are affected, particularly when combining evidence from families of different sizes. For example, the affected sib-pair method counts the number of sib-pairs inheriting zero, one or two marker alleles as copies of a specific parental allele and checks whether the resulting numbers are in agreement with the 1:2:1 ratio expected under Mendelian inheritance. However, it has been unclear how to optimally extend this approach to more than two affected siblings. One approach for two-point analysis (involving one marker and disease), made formal by Goring and Terwilliger<sup>55</sup>, is to transform the non-parametric test for a binary trait into a model-based test in which the disease status is recoded in families as a marker-like genotype (a so-called 'pseudomarker', see below); linkage of the marker locus is then performed against this pseudomarker. For affected sib-pair designs, this approach is equivalent to analysing the data under a recessive model of inheritance and provides near-optimal statistical performance<sup>56,57</sup>. It is equivalent to the maximum-likelihood binomial method of Abel *et al.*<sup>57</sup> and the technique can also be extended to examine multiple marker loci simultaneously<sup>58,59</sup>.

Early authors were under the impression that linkage to complex traits, for which modes of inheritance are not generally known, should be carried out in a non-parametric fashion. However, as shown above and elsewhere<sup>60</sup>, parametric linkage analysis can be made to handle such situations well.

Homozygosity mapping is another technique that is often considered to be a non-parametric approach to linkage mapping. In fact, this approach assumes a rare disease allele and a recessive mode of inheritance<sup>61</sup>. Such a model leads to a high probability that the genomic region surrounding the disease locus is homozygous in affected individuals owing to the inheritance of the paternal and maternal mutations from a common ancestor. Because this approach relies simply on detecting long stretches of homozygosity in affected individuals, it has the advantage of not requiring any knowledge of the exact pedigree or occurrence of inbreeding for the analysis<sup>62</sup>.

A striking example of homozygosity mapping has recently led to new understanding of the genetics of primary ciliary dyskinesia (PCD), an inherited disorder that is characterized by recurrent infections of the upper and lower respiratory tract. In this example (which was undertaken in dogs), homozygosity mapping in Old English sheepdogs that were found to suffer from a chronic airway inflammation similar to PCD, together with pedigree analysis, demonstrated recessive inheritance of the trait. Genetic studies of five cases and 15 controls identified a 15 Mb segment of homozygosity on dog chromosome 34 that was shared by all cases and contained 151 genes. Systematic analysis identified a mutation in coiled-coil domain containing 39 (*CCDC39*). This led to the examination of the human orthologue in the human disease; remarkably, loss-of-function

mutations in the human orthologue were found to underlie a substantial fraction of PCD cases<sup>63</sup>.

It should be noted, however, that simply searching for long stretches of homozygosity may be unsuccessful because, although this strategy is most suitable for rare marker alleles, it disregards differences in allele frequencies between markers. As an example, a linkage analysis of autosomal-recessive primary congenital glaucoma was successful, whereas the homozygosity-mapping approach failed to identify linked markers<sup>64</sup>.

**Dealing with missing data.** For a given marker, or a set of unlinked markers (multipoint analysis), likelihood analysis can appropriately handle missing marker genotypes. However, most linkage studies are now conducted with dense sets of biallelic SNPs, in which analysis of many markers in close proximity is needed to obtain sufficiently informative linkage information. This means that SNP markers are usually in strong LD and, in this situation, missing parental genotypes have been shown to lead to increased false-positive results<sup>65</sup> because linkage programs generally assume markers to be unlinked. To avoid this problem, linkage calculations are often undertaken with a subset of SNP markers selected for low LD, even though this may result in considerable loss of information. If examination of the pedigrees suggests that the non-genotyped individuals will introduce little or no bias — as would be the case when most parents are genotyped in a nuclear family design — the calculation can be greatly simplified by ignoring the LD between SNPs, at least in a preliminary evaluation. Alternatively, some computer programs, such as MERLIN<sup>66</sup> and MENDEL<sup>67,68</sup>, will allow the user to define blocks of markers that are treated as a single group in the linkage analysis (assuming there is no recombination between markers in the same group). In this case, frequencies of the different combinations of the alleles inherited on the same chromosome (haplotypes) are estimated and used for the linkage analysis.

The determination of haplotypes from genotype data is also of interest in many other situations. High-density SNP arrays include the identification of regions of haplotype identity (homozygosity) for mapping recessive disease traits and for fine-mapping recombination breakpoints in positional gene identification studies, as shown in the example of neuroblastoma above. The availability of family data contributes information on phase, thus removing some of the uncertainty in haplotype determination compared to studies of unrelated individuals at the expense of calculation complexity. Gao *et al.*<sup>69</sup> review statistical issues and computational methods for analysis of haplotype patterns in pedigrees.

## Family-based association studies

**Advantages of using family-based controls.** A major problem in association analyses is the potential for unrecognized population heterogeneity due to, for example, the presence of individuals of different ethnic origin<sup>70</sup>. One of the premises of association studies is that differences in allele frequencies between cases and controls are indicative of disease variants. Thus, other sources

### Identity-by-descent

Two alleles in a genotype (individual) are called identical-by-descent when they are copies of one allele in an ancestor (identical-by-descent is often abbreviated to IBD, but in this article IBD stands for inflammatory bowel disease). By contrast, two alleles that are identical by state just 'look' the same but could have originated from different ancestors.

### Homozygosity mapping

A form of recombination mapping that allows the localization of rare recessive traits by identifying unusually long stretches of homozygosity at consecutive markers.

### Phase

The sequence of alleles at multiple loci inherited from one parent. For an individual who is heterozygous at two loci (Aa and Bb), there are two ways (phases) in which the two alleles, one at each locus, can be inherited from the same parent (BA/ba or Ba/bA).



## Box 2 | Combining linkage and association methodologies: Crohn's disease

The complementary nature of genome-wide association and linkage studies is illustrated by the genetic study of inflammatory bowel disease (IBD). The two major forms of IBD are Crohn's disease and ulcerative colitis. The first susceptibility gene to be identified in Crohn's disease was nucleotide-binding oligomerization domain containing 2 (*NOD2*; also known as *CARD15*). This was based on localization obtained with a non-parametric analysis of multiplex families<sup>118</sup>, followed by the identification of mutations in the colocalized gene<sup>11</sup>. Transmission disequilibrium test (TDT) and pedigree disequilibrium test (PDT) statistics, discussed in the main text, showed a significant (and reproducible<sup>119</sup>) association of Crohn's disease with *NOD2* variants (it should be noted that TDT and PDT are actually tests for linkage, but they are often considered as methods for evaluating association because their power is null if association is absent). Interestingly, the association signals did not explain all the linkage observed in the families, leaving open the possibility of contributions from other unidentified variants at *NOD2* or at another gene in the same region. The issue of deciding whether association with a specific set of genetic variants can account for a linkage signal is of general interest<sup>120–122</sup>.

Genome-wide association (GWA) studies<sup>123–125</sup> have led to the identification of many additional Crohn's disease and ulcerative colitis susceptibility loci. A meta-analysis of three GWA studies and subsequent replication identified 32 susceptibility loci, including *NOD2* (REF. 119). Most of the loci revealed in previous linkage studies of Crohn's disease and ulcerative colitis do not overlap with the regions of association. Thus, these reported linkages are now often attributed to false-positive signals. However, a proportion of the linkage signals could be due to patterns of disease-related genetic variation that is detectable by linkage, but not with current GWA SNP panels (as for *BRCA1* and *BRCA2* in breast cancer). Moreover, the variants identified by association accounted for no more than 20% of the estimated heritability of Crohn's disease.

Recently, linkage has proved fruitful in identifying additional loci and pathways involved in inflammatory bowel disease that were not detected by association. Linkage studies identified mutations in interleukin-10 (*IL10*) receptor genes in severe, early-onset, recessively inherited Crohn's disease<sup>126</sup>. Ulcerative colitis has previously been shown to be associated with markers at *IL10* (REF. 127), and mice that are deficient in IL-10 signalling develop colitis. Taken together, these results suggest a major role for IL-10 signalling in IBD and suggest further screening of related genes. It is plausible that additional genes or pathways could be identified by linkage in clinically defined subsets of IBD.

Many other diseases show little overlap between findings from the meta-analysis of linkage results and findings from the meta-analysis of association results, raising the possibility that disease-implicated genetic variants remain to be identified by systematic sequencing. With the advent of new sequencing technologies, both association and linkage regions will be thoroughly investigated in many diseases to see whether they harbour such rare causative variants (see, for example, recent work on Crohn's disease<sup>127</sup>).

of allele frequencies can greatly increase the rate of false positive results, which is why statisticians have developed analytical methods to handle such heterogeneity<sup>71,72</sup>. This problem is essentially non-existent in linkage analysis, so it has been tempting to carry out genetic association studies by using related individuals, who are necessarily of the same ethnic origin. In other words, family-based controls, as opposed to population-based controls, largely avoid problems of population heterogeneity.

The first such family-based design, the haplotype relative risk (HRR) approach<sup>73</sup>, considered an affected individual and his or her two parents. For a specified marker locus with possibly multiple alleles, it contrasted the two alleles received by the offspring ('case' alleles) with the two alleles not transmitted by the parents ('control' alleles). One may then compute the odds ratio (HRR, approximate relative risk) for a specified allele at this marker. Statistical analysis of this design<sup>74,75</sup> showed that the HRR

is different from 1 only in the presence of LD between disease and marker alleles and when the recombination fraction between the two loci is less than 50%. Several different analysis methods were then developed<sup>76,77</sup> for this 'trio design', notably a McNemar-type statistic that focuses on heterozygous parents<sup>76</sup>.

One of the main differences between family-based data and unrelated case-control individuals is that only the former will potentially exhibit genotyping errors in the form of Mendelian inconsistencies. However, for SNPs in trio families, the error detection rate is in the range of 25–30% so that the true error rate is roughly 3.3–4 times the apparent rate of Mendelian inconsistencies<sup>78</sup>. Particularly in genome-wide scans involving hundreds of thousands of SNPs, the occurrence of Mendelian inconsistencies is almost unavoidable. An extension of the TDT (see below), called TDtae, allows for genotyping errors in the analysis and accommodates various error models<sup>79</sup>.

**The TDT.** The full potential of the trio design described above was realized by formulating a test for 'linkage, given association' called the TDT<sup>80</sup>. It is a McNemar-type test but, importantly, with known parental genotypes and under the null hypothesis of no linkage, offspring genotypes are independent so that the TDT can accommodate multiple affected offspring. For a specific marker, the TDT focuses on heterozygous parents and tests whether a specific marker allele has the same frequency among the alleles inherited and those not inherited by an affected child. As neither genotype nor allele frequencies are required, the TDT is immune to population stratification.

This design was a major step forward in the analysis of family-based data and was pivotal in the search for various DNA variants associated with diseases such as insulin-dependent diabetes mellitus and Crohn's disease<sup>81</sup> (BOX 2). Even though the TDT is a test for linkage, it only has power in the presence of association<sup>74</sup>, and it has been particularly useful in cases when large numbers of associations were found and only a small proportion seemed to be true findings<sup>82–84</sup>; the TDT was then able to either confirm or refute claims of significant results.

Various extensions of the TDT have been developed, notably the family-based association test (FBAT) suite of approaches<sup>24,85</sup>, which is discussed below. A particular variant of the TDT tests has been designed for association that is conditional on a previously identified associated locus<sup>86</sup>. Another extension of the TDT, the entropy-based TDT<sup>87</sup>, was thought to exhibit higher power than the TDT, but this power advantage is purchased entirely by an increased type 1 error and disappears when calibrated to have the same significance level as the TDT<sup>88</sup> (the entropy-based TDT should, therefore, never be used). Based on the same principle of using family-based rather than population-based controls, the affected-family-based controls (AFBAC) method<sup>89</sup> uses data from all parents, rather than from only heterozygous parents as in the TDT, and it was developed as a test for association. However, it is not immune to population stratification<sup>90</sup> because genotype frequencies are relevant in this test.

## McNemar-type statistic

A test that focuses on paired data (either two tests on a specific individual or one test on paired individuals), in which the test has a yes/no outcome. It is of interest to see whether the yes/no distribution is the same for the two members of the pair. For example, two raters may carry out a diagnosis of schizophrenia on a group of probands; the McNemar-type statistic is used to test whether the two raters obtain the same numbers of 'affected' and 'unaffected' individuals.

Table 1 | **A selection of software packages for family-based association analysis**

Program name	Purpose	Website
APL	Association analysis in the presence of linkage	<a href="http://www.chg.duhs.duke.edu/research/aplosa.html">http://www.chg.duhs.duke.edu/research/aplosa.html</a>
FBAT	Family-based association tests	<a href="http://www.biostat.harvard.edu/~fbat/fbat.htm">http://www.biostat.harvard.edu/~fbat/fbat.htm</a>
FBAT-PC, PBAT	Family-based association tests for repeatedly measured quantitative traits	<a href="http://www.biostat.harvard.edu/~clange/default.htm">http://www.biostat.harvard.edu/~clange/default.htm</a>
HAPLOTRY	Enumerates all sets of non-recombinant haplotypes consistent with given genotype data	Available from <a href="mailto:Martin.Farrall@cardiov.ox.ac.uk">Martin.Farrall@cardiov.ox.ac.uk</a> on request
HAPLOVIEW	Haplotype analysis with graphical interface	<a href="http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview">http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview</a>
ILINK	Part of the LINKAGE package for iterative parameter estimation	<a href="http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html">http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html</a>
LAMP	Linkage and association analysis for pedigree data	<a href="http://www.sph.umich.edu/csg/abecasis/LAMP/download/">http://www.sph.umich.edu/csg/abecasis/LAMP/download/</a>
MENDEL	Likelihood analyses (such as linkage or genetic counselling) in pedigree data with small numbers of loci	<a href="http://www.genetics.ucla.edu/software/">http://www.genetics.ucla.edu/software/</a>
MERLIN	Likelihood analyses (such as linkage or variance components) in pedigree data with small numbers of loci; allows for dependent SNPs	<a href="http://www.sph.umich.edu/csg/abecasis/Merlin/">http://www.sph.umich.edu/csg/abecasis/Merlin/</a>
PAP	Likelihood analyses on pedigree data	<a href="http://hasstedt.genetics.utah.edu/">http://hasstedt.genetics.utah.edu/</a>
PDT	Carries out pedigree disequilibrium (association) tests	<a href="http://www.mihg.org/software_download/download_reg.php?software=PDT">http://www.mihg.org/software_download/download_reg.php?software=PDT</a>
PLINK	Multipurpose genome-wide analysis package (can be used for association, linkage and the TDT)	<a href="http://pngu.mgh.harvard.edu/purcell/plink/">http://pngu.mgh.harvard.edu/purcell/plink/</a>
QUANTO	Association analysis, including power and sample size calculations and gene–gene interaction	<a href="http://hydra.usc.edu/GxE">http://hydra.usc.edu/GxE</a>
REGRESS	A version of the LINKAGE programs that incorporates Bonney's regressive models	Available from <a href="mailto:florence.demenais@inserm.fr">florence.demenais@inserm.fr</a> on request
ROADTRIPS	Association testing with partially or completely unknown pedigree structure	<a href="http://www.stat.uchicago.edu/~mcpeek/software/index.html">http://www.stat.uchicago.edu/~mcpeek/software/index.html</a>
SIMWALK	Likelihood approach for generating optimal haplotype configurations in pedigree data	<a href="http://www.genetics.ucla.edu/software/simwalk">http://www.genetics.ucla.edu/software/simwalk</a>
TDTae	TDTs allowing for errors with various error models	<a href="ftp://linkage.rockefeller.edu/software/tdtae2">ftp://linkage.rockefeller.edu/software/tdtae2</a>

Additional program details may be obtained from <http://linkage.rockefeller.edu/soft/>. PAP, Pedigree Analysis Package; PBAT, population-based association tests; PC, principal components; PDT, transmission disequilibrium test.

## Type 1 error

The conditional probability of obtaining a significant result in the absence of any effect tested; it is also called the probability of a false-positive result.

## Genotype relative risk

The ratio of the probability of a trait or disease occurring in an at-risk group to the probability of it occurring in a population that is not considered to be at risk. For example, a risk of 1.2 in heterozygotes that are relative to common homozygotes implies that the heterozygotes are 20% more likely to suffer from the disease.

**Discordant sib-pairs.** For studies of late-onset diseases, parents may no longer be available. It is then tempting to use other unaffected relatives as family-based controls. For example, an unaffected sibling of an affected individual may serve as this person's control<sup>91</sup>. Thus, discordant sib-pairs represent another family-based design. Different test statistics for this design have been proposed<sup>92,93</sup>, but for sibships with one affected and one unaffected sibling, these tests are all equivalent<sup>93</sup>. They differ, however, in the analysis of larger sibships. Although discordant sib-pairs are generally robust to population stratification, they are not as powerful as unrelated case and control data<sup>94</sup> because siblings are genetically similar. Increasing the number of affected siblings per family generally leads to more power<sup>94</sup>.

A multivariate approach has also been developed that can combine data obtained under different designs, including those in which parents are used as controls<sup>95</sup>. Which design is most powerful — those in which the controls are unaffected parents or siblings or those in which they are unrelated individuals — depends on the frequency of the trait, the family configuration (number of affected individuals) and the genotype relative risk<sup>95</sup>.

Discordant sib-pairs have been useful in association analyses of type 1 diabetes<sup>96</sup> and schizophrenia<sup>97</sup>.

**Family-based association tests — FBAT and PDT.** Often, nuclear families with multiple affected and unaffected siblings are collected and used to carry out both genetic linkage and association analysis. A general extension of the TDT has resulted in the FBAT approach, which permits any type of genetic model, affected and unaffected siblings, various phenotypic traits and multiple markers<sup>25,85,98,99</sup>. The statistical principle underlying this approach is complicated and will not be described here in detail. Computer programs are freely available for the FBAT suite of approaches (TABLE 1). The FBAT approach has been successfully applied in a large number of investigations, such as in a recent study of asthma<sup>100</sup>.

The FBAT approach is usually applied to nuclear family data, although the online documentation states that FBAT also uses data from pedigrees. A test that is specifically designed for analysis of LD in general pedigrees is the pedigree disequilibrium test (PDT)<sup>101</sup>. It has been successfully applied in a number of association studies, such as in a recent study on diabetes<sup>102</sup>. As previously mentioned,

for diseases of late onset, parental genotypes are often unavailable. For this situation, a test for association in the presence of linkage (APL) has been developed<sup>103</sup>, which can be more powerful than the PDT.

### Combined linkage and association mapping

**Rationale.** The above considerations motivate the search for a general framework to evaluate linkage and association simultaneously, taking combinations of data from pedigrees with different relationship structures (such as extended pedigrees, sibships and TDT families) and case–control samples. Such an approach is likely to be the most powerful and useful approach for identifying new genetic factors related to trait loci beyond those that can be readily detected by GWA in case–control designs. The detection of new loci will arise principally from interpreting high-resolution genotype and sequence data from studies that combine data from multiple sources. When linkage is detected, researchers may wish to evaluate the extent to which association with a particular variant, or set of variants, accounts for the linkage signal to determine whether these variants encompass all of the genetic factors that can be detected from the region in the available data set. These applications require that family relationships and linkage be appropriately accounted for in the association test. With pedigrees, the power to detect association is maximized with tests that incorporate data on related individuals<sup>104</sup>.

**Implementations.** As discussed above, genetic association between a trait and a marker arises because the marker variant is directly causative or is in strong LD with a causative variant. Parametric linkage analysis allows the LD between a hypothetical disease locus and a marker to be estimated as a parameter in the model. It also provides a statistical test to determine whether these are different from expectations in the absence of LD (no association) or in the presence of complete LD. Including multiple markers in this technique improves the estimation of identity-by-descent and tests whether the linkage signal is completely accounted for by the marker. Such analyses may be carried out manually by using one of the older programs such as ILINK<sup>105</sup>, as has been done, for example, in a successful linkage and association analysis of intervertebral disc disease<sup>106</sup>. A computationally efficient solution for discrete traits has been implemented in the LAMP program (TABLE 1), which is based on approaches described by Li *et al.*<sup>107</sup>. This program uses calculation procedures similar to MERLIN to evaluate identity-by-descent using surrounding marker loci; the likelihood is calculated for the marker data when given the trait information and parameter values for the genetic model.

A general non-parametric framework that can be used to combine data from families and case–control studies for the study of discrete traits has been proposed by Terwilliger and Goring as an extension of their pseudomarker method<sup>55</sup>. With this approach, disease status is recoded as the pseudomarker and the combined linkage and LD analysis is performed against the pseudomarker; the association-test statistic is provided depending on

the linkage in the families. The calculations can be made using special version of ILINK from the FASTLINK 4.1P software package (TABLE 1). An example of the use of this approach for joint linkage and association mapping in extended pedigrees can be found in REF. 108, in which the method was used to identify the variant responsible for lactose intolerance. The pseudomarker approach has more optimal statistical performance than alternatives (J. Terwilliger, personal communication).

Association testing with arbitrary combinations of related and unrelated individuals is also possible by using the ROADTRIPS approach<sup>109</sup>.

**Quantitative traits.** General parametric tests of association (tests of LD) and linkage of markers with a quantitative trait can also be undertaken in families or extended pedigrees<sup>110</sup>. It is generally assumed that the quantitative trait is controlled by an unobserved trait locus with two alleles (high and low) that are linked to and are potentially in LD with marker loci from a specific chromosome region. The test of association is based on estimated trait values for the high and low alleles. Residual familial correlation is included in the model to account for heritability that is not attributed to the specified trait locus, as described above for segregation and linkage analysis. Readers can find implementations of this approach in computer programs such as REGRESS<sup>43</sup> and the Pedigree Analysis Package (PAP)<sup>44</sup>. An example application for mapping of DNA variants controlling plasma angiotensinogen levels is given in REF. 111. Another illuminating example is that of HbF (BOX 3).

**Multiple disease alleles.** A limitation of the methods above is the inability to deal with multiple disease alleles. An alternative approach that both simplifies the calculations and extends the model to multiple alleles assumes that a mapped disease locus or QTL has been completely sequenced. Then we can consider that the variant or variants that are causative of the trait association, or in complete LD with the causative variants, are included in the study. Under this assumption, a measured haplotype analysis is appropriate under which the pedigree information is used to determine probable haplotype configurations for family members, and these are analysed (under the class D regressive model) to estimate trait values for each haplotype. If a single variant is causal, the haplotypes containing the same allele will be equally associated with the trait (the estimated trait values of each haplotype will be statistically indistinguishable). If several different variants that are in complete LD at the locus are causal, then, given a sufficiently large sample size, this can be detected by the pattern of differences between the estimated trait means of the different haplotypes.

An important aspect of the measured haplotype analysis in families is that haplotypes can be determined with less probability of error than for singleton study designs. However, some haplotype uncertainty that is due to lack of complete phase information and/or missing genotype data must generally be taken into account.

**Class D regressive model**  
Regressive models account for major genes and residual familial patterns of dependence, in terms of correlations between relatives, without necessarily introducing a particular scheme of causality for the residual patterns. The class D regressive model incorporates four correlations which may have distinct values to account for residual familial patterns: father–mother, father–offspring, mother–offspring and sibling–sibling. Likelihoods are calculated by successively conditioning on ancestral phenotypes and major genes.

## Lod score

(Base 10 'logarithm of the odds' or 'log-odds'). A statistical estimate of whether two loci are likely to lie near each other on a chromosome and are, therefore, likely to be inherited together. A lod score of three or more is generally taken to indicate that the two loci are close.

## Box 3 | Combining linkage and association methodologies: fetal haemoglobin levels

QTL mapping under a family design has led to the identification of genetic factors controlling individual variation in the residual amounts of fetal haemoglobin (HbF,  $\alpha_2\gamma_2$ ) synthesized in adult life (the major Hb form switches to HbA after the age of 6 months). This is a medically important trait because individuals affected by sickle cell disease or  $\beta$ -thalassaemia have a less severe disease if they have higher HbF. Individual variation of HbF level is under strong genetic control, and mutations involving the  $\beta$ -globin gene (*HBB*) complex play a part. However, family studies show that other major genetic determinants of HbF level segregate independently of *HBB*. Segregation analysis, linkage, combined linkage and association mapping and measured haplotype methods (all described in the main text) have contributed to the identification of these *trans*-acting genetic factors.

Strong evidence for genetic factors outside of *HBB* in controlling HbF was provided by the study of a large Asian-Indian kindred, containing  $\beta$ -thalassaemia patients and more than 163 individuals with measured HbF expression levels. Segregation analysis was applied to obtain a genetic model that was then used in a genome-wide linkage study to identify a QTL in a 1.5 Mb interval on chromosome 6q23, with a lod score of 6.3 (REFS 22, 128). Systematic resequencing followed by combined linkage and association analysis of all SNPs identified multiple markers in the intergenic region between two genes at 6q23 — v-myb myeloblastosis viral oncogene homologue (*MYB*) and HBS1-like (*HBS1L*) — that are strongly associated with the trait, both in this pedigree and in other families ( $P = 10^{-76}$ ) (REF. 129). Measured haplotype analysis led to estimates of 18% of the trait variance attributed to these markers and an additional 12% to markers at *HBB*. As this value was considerably less than the total heritability estimate, a genome-wide association (GWA) study was undertaken in just 179 individuals who were selected from the upper and lower 5% extremes of the distribution of F cells<sup>130</sup> (cells with detectable amounts of HbF) in unaffected individuals from the Caucasian families. This led to the identification of another major QTL at B-cell CLL/lymphoma 11A (*BCL11A*) on chromosome 2p15, which accounts for 15% the trait variance<sup>129</sup>. Further studies of these loci have revealed significant new insights into the regulation of the haemoglobin switch<sup>131</sup> and possible therapeutic approaches<sup>130,132</sup>.

As these loci together account for about 50% of the estimated heritability, other QTLs must also be present. As these were not detected by GWA (an example of the 'missing heritability problem'), rare variants may be implicated. Additional linkage studies in families with substantial residual heritability that was not accounted for by the known loci may be a plausible alternative to identify these variants. Recently, the power of this approach has been demonstrated by applying linkage to a large Maltese pedigree in which high HbF segregated in an apparently autosomal-dominant manner independently of known loci. A novel QTL was located to a 663 kb interval on chromosome 19p13 (lod score of 4.2), and mutations were identified in a colocalized gene, Krüppel-like factor 1 (erythroid) (*KLF1*), a key erythroid transcriptional factor<sup>133</sup>. Diminished *KLF1* activity that was induced by these mutations was shown to lead to decreased expression of *BCL11A* (see above).

This can be achieved by applying a combination of programs (SIMWALK, based on algorithms described in REF. 112, and HAPLOTRY<sup>113</sup>) to prepare input for one of the class D regressive model packages cited above. An application of this approach for high-resolution mapping of the angiotensin converting enzyme (ACE)-linked QTL, which influences circulating ACE activity, is given in REF. 113.

TABLE 1 contains a list of publicly available computer programs for interested readers who want to implement analyses based on these methods.

## Conclusions

Current methodologies and implementations of family-based association analysis have reached a high level of sophistication and exploit much of the linkage and association information about the localization of trait

variants. Statistically, the main advantage of family-based association analysis (that is, combined linkage and association analysis) is that two almost-independent contributions to the genetic mapping of trait variants are united for a joint assessment of the localization of such variants.

In future, other sources of information could be tapped, such as copy number variants (CNVs) and expression levels. Currently, such additional information must be incorporated by statistical ad hoc methods, but they might eventually be combined into one comprehensive approach. Family-based association tests for CNVs have already been developed<sup>114,115</sup>. Finally, for many complex traits, such as heart disease and some forms of diabetes, non-genetic risk factors tend to be more important determinants than genetic variants<sup>116,117</sup> and should be included in genetic studies as much as possible.

- Ott, J. *Analysis of Human Genetic Linkage* (Johns Hopkins Univ. Press, Baltimore, USA, 1999). **This book provides the basics of linkage analysis.**
- Ott, J. *et al.* Linkage studies in a large kindred with familial hypercholesterolemia. *Am. J. Hum. Genet.* **26**, 598–603 (1974).
- Elston, R. C., Namboodiri, K. K., Go, R. C., Siervogel, R. M. & Glueck, C. J. Probable linkage between essential familial hypercholesterolemia and third complement component (C3). *Cytogenet. Cell Genet.* **16**, 294–297 (1976).
- Berg, K. & Heiberg, A. Linkage between familial hypercholesterolemia with xanthomatosis and the C3 polymorphism confirmed. *Cytogenet. Cell Genet.* **22**, 621–623 (1978).
- Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).
- Tsui, L. C. *et al.* Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* **230**, 1054–1057 (1985).
- Bell, G. I., Horita, S. & Karam, J. H. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* **33**, 176–183 (1984).
- Julier, C. *et al.* Insulin-IGF2 region on chromosome 11p encodes a gene implicated in HLA-DR4-dependent diabetes susceptibility. *Nature* **354**, 155–159 (1991).
- Walsh, T. & King, M. C. Ten genes for inherited breast cancer. *Cancer Cell* **11**, 103–105 (2007).
- Coon, K. D. *et al.* A high-density whole-genome association study reveals that *APOE* is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J. Clin. Psychiatry* **68**, 613–618 (2007).
- Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
- Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
- Clerget-Darpoux, F. & Elston, R. C. Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. *Hum. Hered.* **64**, 91–96 (2007).



14. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9**, 356–369 (2008).
15. Hindorf, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
16. Hindorf, L. A., Junkins, H. A., Hall, P. N., Mehta, J. P. & Manolio, T. A. A catalog of published genome-wide association studies. *National Human Genome Research Institute* [online], <http://www.genome.gov/gwastudies/> (2010).
17. Maher, B. Personal genomes: the case of the missing heritability. *Nature* **456**, 18–21 (2008).
18. Teer, J. K. & Mullikin, J. C. Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.* **19**, R145–R151 (2010).
19. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nature Rev. Genet.* **11**, 773–785 (2010).
20. Paterson, A. D., Naimark, D. M. & Petronis, A. The analysis of parental origin of alleles may detect susceptibility loci for complex disorders. *Hum. Hered.* **49**, 197–204 (1999).
21. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).
22. Thein, S. L. *et al.* Detection of a major gene for heterocellular hereditary persistence of fetal hemoglobin after accounting for genetic modifiers. *Am. J. Hum. Genet.* **54**, 214–228 (1994).
23. Byun, M. *et al.* Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. *J. Exp. Med.* **207**, 2307–2312 (2010).
24. Laird, N. M. & Lange, C. Family-based designs in the age of large-scale gene-association studies. *Nature Rev. Genet.* **7**, 385–394 (2006).
25. Laird, N. M. & Lange, C. Family-based methods for linkage and association analysis. *Adv. Genet.* **60**, 219–252 (2008).
26. Zhang, K. & Zhao, H. in *Handbook on Analyzing Human Genetic Data: Computational Approaches and Software* (eds Lin, S. & Zhao, H.) 191–240 (Springer, Berlin, 2010).
27. Thornton, T. & McPeck, M. S. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am. J. Hum. Genet.* **81**, 321–337 (2007).
28. Rakovski, C. S., Weiss, S. T., Laird, N. M. & Lange, C. FBAT-SNP-PC: an approach for multiple markers and single trait in family-based association tests. *Hum. Hered.* **66**, 122–126 (2008).
29. Hoffmann, T. J. *et al.* Parsing the effects of individual SNPs in candidate genes with family data. *Hum. Hered.* **69**, 91–103 (2010).
30. Li, M., Boehnke, M. & Abecasis, G. R. Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *Am. J. Hum. Genet.* **78**, 778–792 (2006).
31. Elston, R. C. & Stewart, J. A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**, 523–542 (1971).  
**This study provides the theoretical foundation for peeling algorithms in likelihood calculations for large family pedigrees.**
32. Elston, R. C., George, V. T. & Severtson, F. The Elston-Stewart algorithm for continuous genotypes and environmental factors. *Hum. Hered.* **42**, 16–27 (1992).
33. Lander, E. S. & Green, P. Construction of multilocus genetic linkage maps in humans. *Proc. Natl Acad. Sci. USA* **84**, 2363–2367 (1987).  
**This paper provides the theoretical foundation for peeling algorithms, which are used on family data in likelihood calculations involving large numbers of markers.**
34. Heath, S. C., Snow, G. L., Thompson, E. A., Tseng, C. & Wijsman, E. M. MCMC segregation and linkage analysis. *Genet. Epidemiol.* **14**, 1011–1016 (1997).
35. Ma, J., Amos, C. I. & Warwick Daw, E. Ascertainment correction for Markov chain Monte Carlo segregation and linkage analysis of a quantitative trait. *Genet. Epidemiol.* **31**, 594–604 (2007).
36. Ott, J. Linkage analysis and family classification under heterogeneity. *Ann. Hum. Genet.* **47**, 311–320 (1983).
37. Hodge, S. E., Vieland, V. J. & Greenberg, D. A. HLODs remain powerful tools for detection of linkage in the presence of genetic heterogeneity. *Am. J. Hum. Genet.* **70**, 556–559 (2002).
38. Bird, T. D., Ott, J. & Giblett, E. R. Evidence for linkage of Charcot-Marie-Tooth neuropathy to the Duffy locus on chromosome 1. *Am. J. Hum. Genet.* **34**, 388–394 (1982).
39. Schraders, M. *et al.* Homozygosity mapping reveals mutations of GRXCR1 as a cause of autosomal-recessive nonsyndromic hearing impairment. *Am. J. Hum. Genet.* **86**, 138–147 (2010).
40. Strauch, K. *et al.* Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. *Am. J. Hum. Genet.* **66**, 1945–1957 (2000).
41. Shete, S., Elston, R. C. & Lu, Y. A novel approach to detect parent-of-origin effects from pedigree data with application to Beckwith-Wiedemann syndrome. *Ann. Hum. Genet.* **71**, 804–814 (2007).
42. Greenberg, D. A., Monti, M. C., Feenstra, B., Zhang, J. & Hodge, S. E. The essence of linkage-based imprinting detection: comparing power, type 1 error, and the effects of confounders in two different analysis approaches. *Ann. Hum. Genet.* **74**, 248–262 (2010).
43. Bonney, G. E., Lathrop, G. M. & Lalouel, J. M. Combined linkage and segregation analysis using regressive models. *Am. J. Hum. Genet.* **43**, 29–37 (1988).
44. Hasstedt, S. J. Variance components/major locus likelihood approximation for quantitative, polytomous, and multivariate data. *Genet. Epidemiol.* **10**, 145–158 (1993).
45. Risch, N. & Giuffra, L. Model misspecification and multipoint linkage analysis. *Hum. Hered.* **42**, 77–92 (1992).
46. Maris, J. M. *et al.* Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N. Engl. J. Med.* **358**, 2585–2593 (2008).
47. Mosse, Y. P. *et al.* Identification of *ALK* as a major familial neuroblastoma predisposition gene. *Nature* **455**, 930–935 (2008).
48. Janoueix-Lerosey, I. *et al.* Somatic and germline activating mutations of the *ALK* kinase receptor in neuroblastoma. *Nature* **455**, 967–970 (2008).
49. George, R. E. *et al.* Activating mutations in *ALK* provide a therapeutic target in neuroblastoma. *Nature* **455**, 975–978 (2008).
50. Chen, Y. *et al.* Oncogenic mutations of *ALK* kinase in neuroblastoma. *Nature* **455**, 971–974 (2008).
51. Eng, C. Cancer: a ringleader identified. *Nature* **455**, 883–884 (2008).
52. Garber, K. *et al.* *ALK*, lung cancer, and personalized therapy: portent of the future? *J. Natl Cancer Inst.* **102**, 672–675 (2010).
53. Capasso, M. *et al.* Common variations in *BARD1* influence susceptibility to high-risk neuroblastoma. *Nature Genet.* **41**, 718–723 (2009).
54. Diskin, S. J. *et al.* Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* **459**, 987–991 (2009).
55. Goring, H. H. & Terwilliger, J. D. 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**, 1310–1327 (2000).
56. Davis, S. & Weeks, D. E. Comparison of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation. *Am. J. Hum. Genet.* **61**, 1431–1444 (1997).
57. Abel, L., Alcais, A. & Mallet, A. Comparison of four sib-pair linkage methods for analyzing sibships with more than two affecteds: interest of the binomial maximum likelihood approach. *Genet. Epidemiol.* **15**, 371–390 (1998).
58. Abel, L. & Muller-Myhsok, B. Robustness and power of the maximum-likelihood-binomial and maximum-likelihood-score methods, in multipoint linkage analysis of affected-sibship data. *Am. J. Hum. Genet.* **63**, 638–647 (1998).
59. Goring, H. H. & Terwilliger, J. D. Linkage analysis in the presence of errors I: complex-valued recombination fractions and complex phenotypes. *Am. J. Hum. Genet.* **66**, 1095–1106 (2000).
60. Abreu, P. C., Greenberg, D. A. & Hodge, S. E. Direct power comparisons between simple LOD scores and NPL scores for linkage analysis in complex diseases. *Am. J. Hum. Genet.* **65**, 847–857 (1999).
61. Smith, C. A. B. The detection of linkage in human genetics. *J. R. Stat. Soc. B* **15**, 153–192 (1953).
62. Lander, E. S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567–1570 (1987).
63. Merveille, A. C. *et al.* CCDC39 is required for assembly of inner dynein arms and the dynein regulatory complex and for normal ciliary motility in humans and dogs. *Nature Genet.* **43**, 72–78 (2011).
64. Plasilova, M. *et al.* Linkage of autosomal recessive primary congenital glaucoma to the *GLC3A* locus in Roms (Gypsies) from Slovakia. *Hum. Hered.* **48**, 30–33 (1998).
65. Huang, Q., Shete, S. & Amos, C. I. Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *Am. J. Hum. Genet.* **75**, 1106–1112 (2004).
66. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin — rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.* **30**, 97–101 (2002).  
**This paper describes the methodology and software for carrying out general linkage analysis in small families.**
67. Lange, K., Weeks, D. & Boehnke, M. Programs for pedigree analysis: MENDEL, FISHER, and dGENE. *Genet. Epidemiol.* **5**, 471–472 (1988).
68. Lange, K. *et al.* 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** (Suppl.), 504 (2001).
69. Gao, G., Allison, D. B. & Hoeseche, I. Haplotyping methods for pedigrees. *Hum. Hered.* **67**, 248–266 (2009).
70. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
71. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).  
**This paper presents the first approach to control for the negative effects of population stratification.**
72. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904–909 (2006).
73. Falk, C. T. & Rubinstein, P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* **51**, 227–233 (1987).
74. Ott, J. Statistical properties of the haplotype relative risk. *Genet. Epidemiol.* **6**, 127–130 (1989).  
**This study describes the statistical basis for the HRR approach that underlies development of the TDT.**
75. Knapp, M., Seuchter, S. A. & Baur, M. P. The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. *Am. J. Hum. Genet.* **52**, 1085–1093 (1993).
76. Terwilliger, J. D. & Ott, J. A haplotype-based ‘haplotype relative risk’ approach to detecting allelic associations. *Hum. Hered.* **42**, 337–346 (1992).
77. Seuchter, S. A., Knapp, M. & Baur, M. P. in *Recent Progress in the Genetic Epidemiology of Cancer* Vol. 1 (eds Lynch, H. T. & Tautu, P.) 89–94 (Springer, Berlin, 1991).
78. Gordon, D., Heath, S. C. & Ott, J. True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Hum. Hered.* **49**, 65–70 (1999).  
**This provides the first evaluation of true and apparent error rates in trio families.**
79. Gordon, D. *et al.* A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. *Eur. J. Hum. Genet.* **12**, 752–761 (2004).
80. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516 (1993).  
**This paper discusses the development of the TDT, the first method for carrying out family-based association mapping with multiple affected offspring.**
81. Rioux, J. D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genet.* **29**, 223–228 (2001).
82. Lernmark, A. & Ott, J. Sometimes it's hot, sometimes it's not. *Nature Genet.* **19**, 213–214 (1998).
83. Ott, J. Association of genetic loci: replication or not, that is the question. *Neurology* **63**, 955–958 (2004).
84. Thomas, D. C. & Clayton, D. G. Betting odds and genetic associations. *J. Natl Cancer Inst.* **96**, 421–423 (2004).

85. Rabinowitz, D. & Laird, N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.* **50**, 211–223 (2000).  
**This study provides the initial theoretical foundation for the development of the FBAT approach.**
86. Koeleman, B. P., Dudbridge, F., Cordell, H. J. & Todd, J. A. Adaptation of the extended transmission/disequilibrium test to distinguish disease associations of multiple loci: the conditional extended transmission/disequilibrium test. *Ann. Hum. Genet.* **64**, 207–213 (2000).
87. Zhao, J., Boerwinkle, E. & Xiong, M. An entropy-based genome-wide transmission/disequilibrium test. *Hum. Genet.* **121**, 357–367 (2007).
88. Ewens, W. & Li, M. Comments on the entropy-based transmission/disequilibrium test. *Hum. Genet.* **123**, 97–100 (2008).
89. Thomson, G. Mapping disease genes: family-based association studies. *Am. J. Hum. Genet.* **57**, 487–498 (1995).
90. Spielman, R. S. & Ewens, W. J. The TDT and other family-based tests for linkage disequilibrium and association. *Am. J. Hum. Genet.* **59**, 983–989 (1996).  
**This paper describes the use of family-based tests that also account for association.**
91. Boehnke, M. & Langefeld, C. D. Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *Am. J. Hum. Genet.* **62**, 950–961 (1998).
92. Spielman, R. S. & Ewens, W. J. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.* **62**, 450–458 (1998).
93. Teng, J. & Risch, N. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res.* **9**, 234–241 (1999).
94. Risch, N. & Teng, J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling. *Genome Res.* **8**, 1273–1288 (1998).
95. Schaid, D. J. & Rowland, C. Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. *Am. J. Hum. Genet.* **63**, 1492–1506 (1998).
96. He, C., Hamon, S., Li, D., Barral-Rodriguez, S. & Ott, J. MHC fine mapping of human type 1 diabetes using the T1DGC data. *Diabetes Obes. Metab.* **11** (Suppl. 1), 53–59 (2009).
97. Schwab, S. G. *et al.* Genome-wide scan in 124 Indonesian sib-pair families with schizophrenia reveals genome-wide significant linkage to a locus on chromosome 3p26–21. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **147B**, 1245–1252 (2008).
98. Lange, C. *et al.* A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Stat. Appl. Genet. Mol. Biol.* **3**, Article 17 (2004).
99. Won, S. *et al.* On the analysis of genome-wide association studies in family-based designs: a universal, robust analysis approach and an application to four genome-wide association studies. *PLoS Genet.* **5**, e1000741 (2009).
100. Smit, L. A. *et al.* CD14 and toll-like receptor gene polymorphisms, country living, and asthma in adults. *Am. J. Respir. Crit. Care Med.* **179**, 363–368 (2009).
101. Martin, E. R., Monks, S. A., Warren, L. L. & Kaplan, N. L. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am. J. Hum. Genet.* **67**, 146–154 (2000).
102. Bronson, P. G., Ramsay, P. P., Thomson, G. & Barcellos, L. F. Analysis of maternal-offspring HLA compatibility, parent-of-origin and non-inherited maternal effects for the classical HLA loci in type 1 diabetes. *Diabetes Obes. Metab.* **11** (Suppl. 1), 74–83 (2009).
103. Martin, E. R., Bass, M. P., Hauser, E. R. & Kaplan, N. L. Accounting for linkage in family-based tests of association with missing parental genotypes. *Am. J. Hum. Genet.* **73**, 1016–1026 (2003).
104. Sahana, G., Guldbrandtsen, B., Janss, L. & Lund, M. S. Comparison of association mapping methods in a complex pedigree population. *Genet. Epidemiol.* **34**, 455–462 (2010).
105. Lathrop, G. M., Lalouel, J. M., Julier, C. & Ott, J. Strategies for multilocus linkage analysis in humans. *Proc. Natl Acad. Sci. USA* **81**, 3443–3446 (1984).
106. Annunen, S. *et al.* An allele of COL9A2 associated with intervertebral disc disease. *Science* **285**, 409–412 (1999).
107. Li, M., Boehnke, M. & Abecasis, G. R. Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. *Am. J. Hum. Genet.* **76**, 934–949 (2005).
108. Enattah, N. S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nature Genet.* **30**, 223–237 (2002).
109. Thornton, T. & McPeck, M. S. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am. J. Hum. Genet.* **86**, 172–184 (2010).
110. Ewens, W. J., Li, M. & Spielman, R. S. A review of family-based tests for linkage disequilibrium between a quantitative trait and a genetic marker. *PLoS Genet.* **4**, e1000180 (2008).
111. Brand, E. *et al.* Detection of putative functional angiotensinogen (AGT) gene variants controlling plasma AGT levels by combined segregation-linkage analysis. *Eur. J. Hum. Genet.* **10**, 715–723 (2002).
112. Sobel, E. & Lange, K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am. J. Hum. Genet.* **58**, 1323–1337 (1996).
113. Soubrier, F. *et al.* High-resolution genetic mapping of the ACE-linked QTL influencing circulating ACE activity. *Eur. J. Hum. Genet.* **10**, 553–561 (2002).
114. Ionita-Laza, I. *et al.* On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. *Genet. Epidemiol.* **32**, 273–284 (2008).
115. Murphy, A. *et al.* On the genome-wide analysis of copy number variants in family-based designs: methods for combining family-based and population-based information for testing dichotomous or quantitative traits, or completely ascertained samples. *Genet. Epidemiol.* **34**, 582–590 (2010).
116. Willett, W. C. Balancing life-style and genomics research for disease prevention. *Science* **296**, 695–698 (2002).
117. Yusuf, S. *et al.* Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* **364**, 937–952 (2004).
118. Hugot, J. P. *et al.* Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* **379**, 821–823 (1996).
119. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genet.* **40**, 955–962 (2008).
120. Sun, L., Cox, N. J. & McPeck, M. S. A statistical method for identification of polymorphisms that explain a linkage result. *Am. J. Hum. Genet.* **70**, 399–411 (2002).
121. Chen, M. H. *et al.* Evaluation of approaches to identify associated SNPs that explain the linkage evidence in nuclear families with affected siblings. *Hum. Hered.* **69**, 104–119 (2010).
122. Chen, M. H. *et al.* Joint modeling of linkage and association using affected sib-pair data. *BMC Proc.* **1**, S38 (2007).
123. Duerr, R. H. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).
124. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* **3**, e58 (2007).
125. Rioux, J. D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nature Genet.* **39**, 596–604 (2007).
126. Glocker, E. O. *et al.* Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. *N. Engl. J. Med.* **361**, 2033–2045 (2009).
127. Franke, A. *et al.* Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nature Genet.* **40**, 1319–1323 (2008).
128. Craig, J. E. *et al.* Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach. *Nature Genet.* **12**, 58–64 (1996).
129. Thein, S. L. *et al.* Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc. Natl Acad. Sci. USA* **104**, 11346–11351 (2007).
130. Thein, S. L., Menzel, S., Lathrop, M. & Garner, C. Control of fetal hemoglobin: new insights emerging from genomics and clinical implications. *Hum. Mol. Genet.* **18**, R216–R223 (2009).
131. Borg, J. *et al.* Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nature Genet.* **42**, 801–805 (2010).
132. Lettre, G. *et al.* DNA polymorphisms at the BCL11A, HBS11L-MYB, and  $\beta$ -globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc. Natl Acad. Sci. USA* **105**, 11869–11874 (2008).
133. Sankaran, V. G. *et al.* Developmental and species-divergent globin switching are driven by BCL11A. *Nature* **460**, 1093–1097 (2009).

## Acknowledgements

We are grateful to J. Terwilliger for providing us with unpublished results on his pseudomarker method. Our work was partially supported by the National Natural Science Foundation of China (NSFC) grant no. 30730057 to J.O.

## Competing interests statement

The authors declare no competing financial interests.

## FURTHER INFORMATION

Jurg Ott's homepage: [http://english.psych.cas.cn/ge/gel/2011102/t20110215\\_65089.html](http://english.psych.cas.cn/ge/gel/2011102/t20110215_65089.html)  
 Nature Reviews Genetics series on Genome-Wide Association Studies: <http://www.nature.com/nrg/series/gwas/index.html>  
 Nature Reviews Genetics series on Study Designs: <http://www.nature.com/nrg/series/studydesigns/index.html>  
**ALL LINKS ARE ACTIVE IN THE ONLINE PDF**