Original Paper

Human Heredity

Hum Hered 2008;66:67–86
DOI: 10.1159/000119107

Published online: March 31, 2008

# Review and Evaluation of Methods Correcting for Population Stratification with a Focus on Underlying Statistical Principles

Hemant K. Tiwari[a] Jill Barnholtz-Sloan[c] Nathan Wineinger[a] Miguel A. Padilla[a]
Laura K. Vaughan[a] David B. Allison[a, b]

[a]Department of Biostatistics, Section on Statistical Genetics, and [b]Clinical Nutrition Research Center, University of Alabama at Birmingham, Birmingham, Ala., [c]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, Ohio, USA

**Key Words**
Admixture · Ancestry · Association · Covariance-based tests · Genomic control · Linkage · Marginal-based tests · QTL · RAM · Randomization · SAT · Structure · Sufficient statistics · TDT

**Abstract**
When two or more populations have been separated by geographic or cultural boundaries for many generations, drift, spontaneous mutations, differential selection pressures and other factors may lead to allele frequency differences among populations. If these 'parental' populations subsequently come together and begin inter-mating, disequilibrium among linked markers may span a greater genetic distance than it typically does among populations under *panmixia* [see glossary]. This extended disequilibrium can make association studies highly effective and more economical than disequilibrium mapping in panmictic populations since less marker loci are needed to detect regions of the genome that harbor phenotype-influencing loci. However, under some circumstances, this process of intermating (as well as other processes) can produce disequilibrium between pairs of unlinked loci and thus create the possibility of confounding or spurious associations due to this *population stratification*. Accordingly, researchers are advised to employ valid statistical tests for linkage disequilibrium mapping allowing conduct of genetic association studies that control for such confounding. Many recent papers have addressed this need. We provide a comprehensive review of advances made in recent years in correcting for population stratification and then evaluate and synthesize these methods based on statistical principles such as (1) randomization, (2) conditioning on sufficient statistics, and (3) identifying whether the method is based on testing the genotype-phenotype covariance (conditional upon familial information) and/or testing departures of the marginal distribution from the expected genotypic frequencies.



## Introduction

Theoretical developments, computer simulations, and empirical evidence from population studies continue to indicate that population stratification due to genetic admixture, as well as other departures from random mating, can confound genetic association studies and produce false positive results [1–4]. Population admixture, however, can also 'mask' true genotype-phenotype associations and produce false negative results. In either case, departures from non-random matings can result in biased estimates and faulty conclusions. This form of population heterogeneity is often regarded as an impediment to genetic association studies given its potential to con-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com


0001–5652/08/0662–0067$24.50/0

Accessible online at:
www.karger.com/hhe

Hemant K. Tiwari, PhD
Department of Biostatistics, Section on Statistical Genetics
Ryals Public Health Building, 420D, University of Alabama at Birmingham
1665 University Blvd., Birmingham, AL 35294 (USA)
Tel. +1 205 934 4907, Fax +1 205 975 2541, E-Mail htiwari@uab.edu

found statistical analyses and induce spurious genotype-phenotype associations.

Experimentally controlling mating type in plant and animal studies is the most extreme way to control for this confounding effect and is accomplished with the use of recombinant inbred strains. However, this is not necessarily feasible for all plant and animal studies and is impossible in human genetic research. Concerns about the effects of population stratification led to the recommendation of using familial data and to the development of the seminal paper on the transmission-disequilibrium test (TDT) [5], based on a related idea proposed by Rubinstein et al. [6] and later by Falk and Rubinstein [7]. The TDT is a family-based association test designed for testing linkage disequilibrium by comparing the proportion of alleles transmitted versus the proportion not transmitted from informative parental matings (i.e., matings with at least one heterozygous parent) to affected offspring. By focusing on affected offspring (i.e., case-only), the TDT assesses whether the distribution of alleles among affected children conditional on parental genotypes differs from what is expected under the null hypothesis of no linkage and/or no association.

Although effective at eliminating false positives due to stratification and genetic admixture, TDT type designs may result in substantially lower power relative to other types of association studies since they utilize only those individuals who are informative for allelic transmission and exclude all others. Population-based association studies (e.g. the case-control study design) usually have greater power than family-based and case-only designs as long as correction for population stratification is properly modeled. Recently, significant advances have been made in statistical methodology to control for the potential confounding effects of population admixture via use of measures of 'individual admixture' and related techniques [8–16]. We refer to such methods as *structured association testing* (SAT). Of equal interest are exciting new developments in the use of individual admixture estimates for what we call *regional admixture mapping* (RAM) [16–21]. In principle, these methods allow researchers to localize genomic regions containing trait-influencing genes in samples of unrelated individuals.

With novel procedures being proposed at such a rapid pace, it is difficult for investigators to keep abreast of the latest methods and their utility. Thus, here we review many of the statistical procedures which aim to create valid test statistics for linkage and disequilibrium mapping studies that control for confounding due to population stratification.

## Review of TDT – Association Testing Methods for Family Data

In the late 1980s and early 1990s, several approaches were proposed to identify disease genes that combined the advantages of linkage and population association approaches [6, 7, 22–24]. These methods typically compared alleles transmitted from parents to affected offspring against alleles that were not transmitted, considering the parental alleles that were not transmitted as 'pseudo controls'. For example, Rubinstein et al. [6] and later Falk and Rubinstein [7] proposed a method for calculating the odds ratio of transmitted vs. non-transmitted alleles to offspring from parents. They termed this 'Haplotype Relative Risk' (HRR) because they were investigating HLA haplotypes, and an odds ratio is similar to relative risk if the disease prevalence is low. It is an unmatched case-control design comparing frequencies of transmitted alleles vs. non-transmitted alleles from parents. A similar method was proposed by Terwilliger and Ott [24]. Ott [23], who studied the properties of HRR and theoretically derived the expected frequencies of transmitted and non-transmitted alleles assuming a recessive disease. Although the test proposed by Falk and Rubinstein [7] was not a valid test for linkage [5], Spielman et al. [5] proposed a valid test for linkage in the presence of association[1] based on the idea of Falk and Rubinstein [7]. The transmission disequilibrium test or TDT is a McNemar [25] test for a matched case-control design that compares transmitted alleles from heterozygous parents to an affected offspring with the expected non-transmitted alleles, assuming there is no transmission distortion. Here, transmitted and non-transmitted alleles from heterozygous parents are considered both as cases and controls, creating a matched case-control design. Tiwari et al. [26] noted that the informative families used in TDT designs can be viewed as a mixture of experimental backcrosses (one heterozygous parent) and F2 intercrosses (two heterozygous parents) as an analogy to experimental crosses.

The original TDT design requires the collection of family trios that include two parents and an affected offspring and is limited to di-allelic marker loci, and dichotomous traits. Although the TDT method is a valid test for

[1] By valid test of linkage in the presence of association, we mean a test that (A) yields p values less than or equal to no more than 100 * $\alpha$% of the time when the marker is either unlinked to or not associated with a locus causing variation in the phenotype; and (B) yields p values less than or equal to $\alpha$ more than 100 * $\alpha$% of the time when the marker is both linked to and associated with a locus causing variation in the phenotype.

linkage, it only has power in the presence of population association and is robust against population admixture [27]. There are more than two hundred publications describing extensions and variations of the original TDT. Figure 1 shows the distribution of 223 published extensions and variations of the TDT from 1993 to 2007. In supplemental table 1 (www.karger.com/doi/10.1159/000119107), we summarize some (but not all) of the extensions or variations of the TDT type procedures.

The extensions to the TDT fall mainly in four categories: (1) relaxing the requirement of only two alleles at the marker locus; (2) relaxing the requirement of the trait to be dichotomous; (3) relaxing the requirement of a parent/offspring trio design, and (4) extension to using genotype information from the X chromosome (X-linked TDT). Other extensions to the TDT include multiple loci, Bayesian TDT, multiple phenotypes, parent of origin/imprinting effects, inbreeding, TDT for haplotypes, censored data, simultaneous and separately modeling of the linkage and association parameters, and other variations to increase power; we choose to focus this review mostly on the four main categories listed above with some discussion of the other extensions.

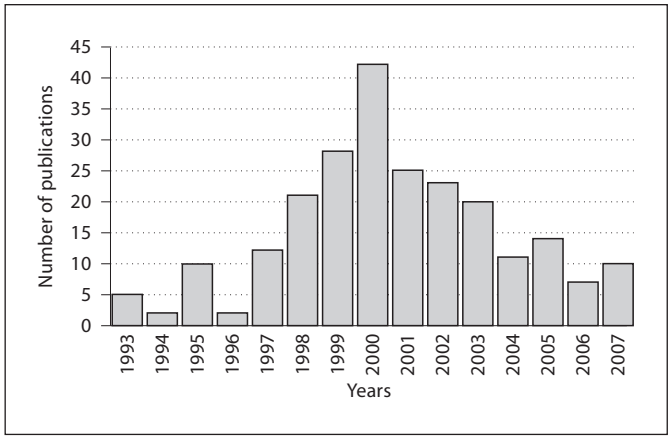


**Fig. 1.** Distribution of TDT type methods manuscripts published since 1993.

### Relaxing the Requirement of Two Alleles at the Marker Locus

Several extensions to the TDT have been proposed to allow for multiple alleles at the marker locus. Bickeboller and Clerget-Darpoux [28] extended the TDT for multiallelic markers by comparing the genotypes formed by the two transmitted alleles (genotype of index) and the genotypes formed by the two nontransmitted alleles (internal control genotype) similar to Terwilliger and Ott [24], thus using the information on both parents simultaneously. This test of transmission patterns of genotypes ($T_g$) was based on the homogeneity test for contingency table of genotype frequencies. Bickeboller and Clerget-Darpoux [28] also proposed an allelic test ($T_c$) based on testing the complete symmetry of the contingency table of allele frequencies. In addition, Rice et al. [29] proposed an extension of the TDT that allows analysis with multiallelic markers, and at about the same time Sham and Curtis [30] introduced an extended TDT (ETDT) based on a logistic regression procedure. The advantage of the ETDT is that it can be easily programmed in any standard statistical software. Other adaptations have followed: Morris et al. [31] used a likelihood ratio test. Spielman and Ewens [32] proposed an alternative test of marginal homogeneity ($T_{mhet}$) that is similar to Biekeboller and Clerget-Darpoux [28], allowing for multi-allelic markers. Kaplan et al. [33] used a Monte Carlo approach, called the MC-$T_m$ statistic, and showed that MC-$T_m$ is more powerful than $T_{mhet}$ and ETDT. Cleves et al. [34] proposed an exact test which is implemented using an exact algorithm and Monte Carlo-Markov chain (MCMC) simulation. Finally, Schaid [35] proposed using each allele separately and then using the maximal TDT as the test statistic to infer linkage. He also proposed a class of model-based approaches using conditional likelihood analyzing all alleles simultaneously under specific genetic models [35]. The maximal TDT statistic, however, does not follow a chi-square distribution. Bentensky and Rabinowitz [36] provided a refinement to Bonferroni's correction for multiple testing based on maximal spanning trees to calculate accurate upper bounds for type 1 error and p values for the maximal TDT.

### Relaxing the Requirement of the Trait to Be Dichotomous

Extensions of the original TDT test of a dichotomous trait to quantitative traits are mainly based on regression framework where covariates can be easily modeled. Allison [37] proposed five TDTs for quantitative traits sequentially called TDTQ1 to TDTQ5. The first four versions of these TDTQs were based on extreme-threshold sampling, and TDTQ5 uses the full distribution of a quantitative trait. TDTQ5 is the most flexible in the sense that it can be easily extended to multiple alleles, multiple

loci, gene-environment interaction, etc., and it is also most powerful of the five. TDTQ5 requires family trios consisting of at least one heterozygous parent and one child. In TDTQ5, the quantitative trait is regressed on offspring genotypes while controlling for parental mating types defined by their genotypes. The test statistic for TDTQ5 is an F ratio that compares the fit of two models with or without the genetic effect in a regression framework that includes the offspring's genotype and parental mating type. Xiong et al. [38] developed a similar approach that allows for more than one child per family.

A non-parametric TDT for quantitative traits was introduced independently by Rabinowitz [39]. The advantage of this test lies in its flexibility in modeling multiple alleles at the marker locus, inclusion of other siblings, and incorporation of covariates. Sun et al. [40] extended Rabinowitz's [39] approach to include families with only one parent available. All these tests assume that model residuals are independent, and therefore they are applicable, as a test for linkage, only for nuclear family data.

George et al. [41] proposed a regression-based TDT for linkage between a marker locus and a quantitative trait locus, treating the trait as the dependent variable and transmission status along with other predictors and confounders, as independent variables. This method does not require independence of observations, thus allowing for analysis of extended pedigree data as well, and modeling any number of covariates. Zhu and Elston [42] proposed conditional likelihood-ratio test statistics that allow multi-generational data as well as a test either for linkage in the presence of allelic association or for allelic association in the presence of linkage. Abecasis et al. [43, 44] proposed a general test of association for quantitative traits in nuclear families (QTDT) based on Fulker et al.'s [45] variance components approach. Monks and Kaplan [46] introduced three extensions to the TDT for quantitative traits: (1) $T_{QP}$ statistic uses genotype information for parents and their children; (2) $T_{QS}$ uses genotypes for at least two siblings having different genotypes in the absence of parental genotypes, and (3) $T_{QPS}$ which was a combination of $T_{QP}$ and $T_{QS}$. Note that the $T_{QP}$ statistic is similar to the statistic proposed by Rabinowitz [39]. Waldman et al. [47] proposed a logistic regression framework instead of the ordinary linear regression for continuous and categorical data. This framework can be easily extended to include multiple phenotypes by simply including phenotypes as predictors in the regression model, and it can easily accommodate multiple offspring per nuclear family. No phenotype distributional assumptions are required with this approach. Lastly, it does not require stand alone software and any standard statistical software such as SAS or SPSS can be used for the analysis.

Liu et al. [48] offered a unified framework for TDT analysis for discrete and continuous traits based on a conditional score test that maximizes power to detect small effects for any distribution in the exponential family, regardless of skewness or kurtosis. Kistner and Weinberg [49] proposed quantitative trait extension of their log-linear approach for qualitative traits [50]. Like the log-linear approach for quantitative traits their quantitative trait extension allows for population admixture by conditioning on parental genotypes.

### Relaxing the Requirement of a Parent/Offspring Trio Design

Parental genotype data are often difficult or impossible to obtain when studying diseases with adulthood or late in life onset. Several approaches have been developed to alleviate the problems that arise from missing and incomplete parental genotypic data.

*Using Information from Unaffected Siblings*

When unaffected siblings are available for the study, their genotype information can be used in tests for allelic transmission. Curtis [51] proposed an extension to the TDT utilizing only discordant sibling pairs for both phenotype and genotype. S-TDT, a similar approach developed by Spielman and Ewens [52], requires (1) that at least one affected and one unaffected sibling, and (2) that all members of the sibship do not have the same genotype at the marker locus. With these requirements met, the S-TDT can be used to analyze linkage disequilibrium between a marker allele and a putative disease allele without reconstructing parental genotypes and without relying on allele frequency estimates. Statistically, the S-TDT tests for significant marker allele frequency differences in affected offspring compared to their unaffected siblings [52]. Generally, the S-TDT is less powerful than the TDT when parental genotypes are available because data on the preferential transmission of parental alleles is more informative. In fact, the S-TDT can be used jointly with the TDT to construct a combined test (C-TDT) using nuclear families, trios, and discordant siblings. Schaid and Rowland [53] showed that the S-TDT is equivalent to the conditional likelihood with the log-additive effects of the marker alleles.

The sibling TDT method by Curtis [51] requires randomly selecting one affected sibling and then selecting

one unaffected sibling whose marker genotype is different from that of affected sibling. To include all available siblings from the same family, Horvath and Laird [54] proposed a sibling disequilibrium test (SDT) based on a standard nonparametric sign test. The SDT is effective in cases where parental information is not available. The data design requirement is the same as S-TDT, with the only difference being that the SDT is a non-parametric test. In 1998, Boehnke and Langefeld [55] introduced seven association tests for multi-allelic markers which they represent using a $2 \times k$ contingency table (k is the number of alleles at the marker locus). The rows represent the disease status and columns represent marker alleles. In some cases these discordant-alleles tests (DATs), ($AC_1$, $AC_2$, and $AC_{ws}$) are identical to each other as well as equivalent to S-TDT but the $AC_2$ statistics have the best power overall. Boehnke and Langefeld [55] proposed to get p values for these DATs by a permutation procedure involving randomly permuting affection status of the siblings. Risch and Teng [56, 57] noted that one can derive additional information from the sample by analyzing the relative frequency of different sibship genotype configurations. This information can then be used to estimate the proportion of mating type frequencies for a di-allelic marker. Weinberg [58] proposed a likelihood approach for families with incomplete parental data. Schaid and Rowland [59] proposed a score test statistic using parents as controls, siblings as controls, or unrelated individuals as controls. Note that their method generalizes the S-TDT and the DAT. In 2000, Siegmund et al. [60] introduced a test of association in the presence of linkage using multivariate regression for correlated outcome data to analyze sibship data.

*Using Information from Nuclear Families in the Absence of Unaffected Siblings and Only One Parent Available*

Bias can arise in the TDT statistic when information is only available from one heterozygous parent, leading to higher false positive rates [30]. Sun et al. [61] introduced 1-TDT to detect linkage between candidate locus and a disease locus using genotypes of affected individuals and only one available parent of the affected individual. The 1-TDT is a valid test of the null hypothesis of no linkage or association. In 2000, Wang and Sun derived the sample size needed to detect linkage disequilibrium for S-TDT and 1-TDT, finding that the required sample size is roughly the same as for the S-TDT with one affected and one unaffected sibling, and is about twice the sample size needed for the original TDT [62]. Clayton [63], Weinberg [58], and Cervino and Hill [64], also provided extensions to TDT when one parent is missing. Allen et al. [65] extended parental controlled association tests for a di-allelic marker and disease that are valid when parental genotype data are informatively missing (i.e. when the missing genotype of parent influences the probability of the parent's genotype data being observed). Also, Allen et al. [66] proposed a multi-allelic extension of their missingness model [65] which also incorporated a bootstrap calibration of missing at random (MAR) procedures to account for informative missingness.

*Using Sibship Data Only and Reconstructing Missing Parental Genotypes*

For some families it might be possible to reconstruct the genotypes of missing parents. However, Curtis [51], Spielman and Ewens [67] and Knapp [68] pointed out that reconstructing genotypes to achieve more power for the TDT procedure can introduce bias. Knapp [68] proposed a statistical procedure to overcome the potential bias induced by the parental genotype reconstruction. Knapp [68] incorporated a reconstruction approach that corrects for bias into C-TDT and called the resulting procedure the reconstruction combined TDT (RC-TDT). Comparisons showed that RC-TDT is more powerful than the S-TDT.

*Using Information on Non-Informative Mating Types*

Because no inference on linkage disequilibrium can be obtained from homozygous parents or other cases of non-informative transmissions, these types of nuclear families are not included in the classical TDT analysis. This problem is often encountered when using binary markers, such as single-nucleotide polymorphisms (SNPs), which are highly abundant throughout the genome and cost effective. The maximum frequency of heterozygotes at a binary marker locus in Hardy-Weinberg equilibrium is 0.5. In this scenario, at least half of the parents would be non-informative in a traditional TDT. Analyzing marker haplotypes is a relatively straightforward solution. However, the haplotype phase is often uncertain, and restricting analyses to pedigrees where the phase is known may lead to bias. As a result, Clayton [63] proposed a new approach to TDT methods using tests based upon score vectors which are averaged over all possible parental haplotypes and transmissions consistent with the observed data (TRANSMIT 2.5.4 documentation: www-gene.cimr.cam.ac.uk/clayton/software). At its implementation, this approach possessed three distinct advantages over earlier TDT methods: (1) it could use any

available parental data; (2) it could use multiple affected offspring in the analysis, and (3) it was the only approach that could adequately deal with phase uncertainty in multilocus haplotypes [63]. The TRANSMIT program also implements Allen et al. [66] bootstrap calibration of missing at random procedures to account for informative missingness.

*Using the Sibling-Based TDT for Quantitative Traits*

The premise behind sibling-based quantitative traits in a regression framework is that any test of association between a genetic marker and a phenotype is also a valid test of linkage if one conditions on parental genotypes since full siblings are nested within parental genotypes. Allison et al. [69] proposed two sibling-based tests of linkage and association for quantitative traits. One is a mixed model, in which the genotype is modeled as a fixed effect and the sibship as a random effect. This test is extremely flexible and can be implemented in standard statistical software. It allowed for multiple alleles at the marker locus, sibships of any size, multiple loci, gene-gene interaction, gene-environment interaction and additional covariates of any number. The second procedure of Allison et al. is a permutation test. Schaid and Rowland [59] proposed another TDT for quantitative traits that allows for missing parental data. Van den Oord [70], Whitmore and Tu [71], Rabinowitz and Laird [72], and Horvath et al. [73] have all offered methods incorporating missing data.

*Using Information from Extended Families*

Traditional TDT-type tests in trios or discordant sibships assume that observations are independent, an assumption violated when trios or discordant sibships from the same extended family are used. Thus, when larger pedigrees are investigated using these methods, only one unit from the pedigree is analyzed and the rest of the information from the pedigree is discarded. As a result, the pedigree disequilibrium test (PDT) was developed [74]. Using this method, the average disequilibrium for each general pedigree is treated as an individual observation. Martin et al. [75] proposed two alternatives to the PDT which correct for bias when multiple generations are contributing to the disequilibrium, the genetic effect due to the locus is strong, and marker-allele frequencies are uneven. The PDT-avg averages all phenotypically informative units regardless of heterozygosity in trios or informative discordant sibships. Meanwhile the PDT-sum method removes the within-family LD average from the original PDT. The PDT-avg gives equal weight to all families, whereas larger families are more heavily weighted in the PDT-sum. The geno-PDT was later developed to test genotype-specific association in general pedigrees [76, 77].

Testing non-random transmission of an allele from parent to affected offspring follows similar statistical methodology regardless of locus, pedigree, or trait characteristic. Developing the FBAT statistic, Laird et al. [78] took advantage of this trend which generalized some of the more specific TDT-type tests including the original TDT, the S-TDT, and the RC-TDT. At the time of implementation, the FBAT statistic could be manipulated by a set of user-defined codings to analyze data from di-allelic or multi-allelic loci and dichotomous, quantitative, or censored traits [73]. At the present time, the FBAT software [78] is able to accommodate a wide variety of pedigree structures, genetic models, and trait characteristics as well as perform haplotype analysis and test multiple markers simultaneously.

Schaid and Sommer [22, 79] developed a likelihood procedure for trios by modeling the probability of an affected offspring's genotype conditional on parental genotypes as a function of the genotype relative risks of the offspring. In 2000, Whittemore and Tu [71] developed a class of likelihood-based score tests for arbitrary family structure and incomplete data extending the work of Schaid and colleagues [22, 35, 79, 80]. The score statistic comprises of two components, namely, a *non-founder statistic* (NFS) and a *founder statistic* (FS). The *non-founder statistic* evaluates transmission disequilibrium from parents to offspring and is based on the conditional distribution of the offspring genotypes given the observed or inferred genotypes of their parents. The *non-founder statistic* is a direct extension of the transmission disequilibrium test (TDT). The *founder statistic* compares marker genotypes in the family founders with those expected under the null hypothesis. In companion paper [81] they examined these two statistics using nuclear family data. Shih and Whittemore [82] and further extended previous work of Whittemore and others [63, 71, 79, 81] to accommodate affected and unaffected offspring, missing parental genotypes, and to include other phenotypes such as censored survival data and quantitative traits. These algorithms are implemented in software named Family Genotype Analysis Program (FGAP). Whittemore and Halpern [83] compared FGAP with FBAT [72] and another alternative association test proposed by Rabinowitz [84]. They observed that FBAT procedures tended to have less power than the other two tests, particularly when applied to families in whom all offspring were affected. The Rabi-

nowitz test and the tests implemented in FGAP performed equally well with respect to overall statistical power.

Methodology is still being developed to improve the power and robustness of TDT approaches to the various forms of ascertainment, genotype, and phenotype characteristics. The informative-transmission disequilibrium test (i-TDT) improves on the design of extended pedigree analysis first addresses by Martin et al. [85]. The i-TDT is a valid joint test of linkage and association that is more powerful than its alternative approach in FBAT because it also incorporates transmission information for heterozygous parents to unaffected offspring [86]. A recent study has expanded the robustness of QTDT methods that rely upon a normality assumption [37] by developing methodology to adequately analyze linkage disequilibrium when traits are not normally distributed [87]. These discoveries show that TDT methods can still be extended further.

### Extension to Using Genotype Information from the X Chromosome (X-Linked TDT)

Horvath et al. [88] modified S-TDT [89] for X-linked diseases (XS-TDT). In addition, they extended RC-TDT [68] to the X-linked reconstruction-combination TDT (XRC-TDT). These tests make no assumption about the mode of disease inheritance or the ascertainment of the sample, and they protect against spurious association due to population stratification similar to S-TDT and RC-TDT. The X-linked RC-TDT employs parental-genotype reconstruction by combining data from families in which parental genotypes are available with data from families in which genotypes of unaffected siblings are available but parental marker information is incomplete, and corrects for the biases resulting from the reconstruction. It does not depend on population allele frequencies, and it outperforms X-linked S-TDT with respect to power. Also, a freely available SAS implementation of these tests allows for the calculation of exact p values. Ho and Bailey-Wilson [90] independently extended the TDT, S-TDT, and the C-TDT [5] for X-linked loci, terming them X-linked TDT, X-linked S-TDT, and X-linked C-TDT.

### Other Notable Extensions

*Multiple Loci and Haplotype Sharing Statistics*

Recently, several approaches to association/linkage mapping have been proposed that utilize the data from multiple loci simultaneously. All these methods are based on the assumption that cases are expected to share not only the disease allele, but also haplotype flanking markers containing the disease allele. This led van der Muelen and te Meerman [91] to propose a haplotype sharing statistic comparing the extent of similarity between transmitted and un-transmitted haplotypes. Wilson [92] extended the TDT to include information on two linked multi-allelic markers instead of one marker only following the likelihood ratio test proposed by Sham and Curtis's [30] ETDT. She also described how the contribution from each locus could be evaluated, both separately and jointly. Collins and Morton [93] proposed a likelihood-based procedure for haplotype sharing in the case-control study design setting. Clayton and Jones [94] offered a haplotype TDT for both qualitative and quantitative traits. However, Wilson [92] and Clayton and Jones [94] assume that the haplotypes of the parents are known. Thus, their methods are not applicable to haplotype phase-unknown data. McPeek and Strahs [95] introduced the decay of haplotype-sharing procedure by modeling the decay of sharing of the ancestral haplotype by descendents, where the number of haplotypes with common ancestral DNA decreases with increasing genetic distance from the variant. Zhao et al. [96] proposed variations of the TDT using multiple tightly linked markers based on phase-known or phase-unknown haplotypes of parental data. In 2000, MacLean [97] proposed the trimmed-haplotype test for linkage disequilibrium applied to both parent-offspring trios and multiplex pedigrees. There are multiple other extensions that also allow for multiple linked markers [98–110]. Furthermore multiple marker loci can be accommodated when considering quantitative traits in a regression framework. Using this methodology, epistatic effects can be investigated and haplotypes of linked loci can be treated as multi-allelic markers.

However, there are two major issues that must be considered more carefully about these haplotype sharing procedures. Firstly, in general haplotype sharing procedures often assume the haplotype phase is known or inferred accurately. However, the misspecification of the distribution of parental haplotypes can lead to substantial bias in parameter estimates even when complete genotype information is available. To resolve this problem, Allen et al. [111] proposed a geometric approach to estimation in the presence of nuisance parameters and derived locally efficient tests and estimators of haplotype effects that are robust to misspecification of the haplotype frequency distribution. Allen and Satten [112] generalized a previous result of Allen et al. [111], allowing for missing genotype data and haplotype × environment in-

teractions. Secondly, Allen and Satten [113] pointed out that variance estimation of haplotype sharing statistic is either very complex [100, 110] or requires the use of permutation testing [91, 101, 102, 104–106, 108]. Permutation testing can be computationally prohibitive in case of genome-wide association studies. Also, permutation variances may be invalid if the model used for the reconstruction of haplotypes is invalid (i.e. Hardy-Weinberg Equilibrium is not met in the data). Therefore, Allen and Satten [113] proposed a simple framework for a class of haplotype sharing statistics for association testing in case-parent trio data by providing a simple variance estimator for haplotype sharing statistics.

*Parent-of-Origin/Imprinting Effects*

Genomic imprinting, also known as 'parent-of-origin effect' is an epigenetic product. A natural way to identify parent-of-origin effects is to stratify transmission/non-transmission allele counts of parental origin and test for their symmetry. To date, there are more than 1700 mutations with parent-of-origin effects catalogued in the University of Otago database [114] (http://igc.otago.ac.nz/home.html). Wilcox et al. [115] proposed a simple method to analyze case-parent trios in effort to detect maternal genetic risk and estimate relative risks associated with both the mother's and the offspring's genotype. Weinberg et al. [50] further extended the TDT to detect parent-of-origin effects based upon a log-linear likelihood approach. In 1999, Weinberg provided a new test of imprinting which resolved deficiencies in her previous test [116]. Recently, Zhou et al. [117] and Hu et al. [118, 119] proposed several methods to detect imprinting in a TDT-type framework. Van den Oord [70] used mixture models to perform a test of parent-of-origin effects for quantitative traits. Furthermore, imprinting can be tested in quantitative traits in a regression framework by coding an additional dummy variable to indicate which parent is heterozygous in a heterozygote-homozygote mating. A significant interaction between this variable and offspring's genotype would indicate imprinting.

*Gene-Environment (G × E) Effect*

The genetic architecture of complex traits may involve multiple genetic or environmental factors and interactions between them. Multiple alleles at a marker locus associated with disease susceptibility may differ in their sensitivity to certain environmental exposures. Most methods developed for G × E interaction using trios implicitly assume that an individual's environmental exposure status is independent of their genotype at the candidate locus of their parent's genotype. In 1999, Schaid [120] proposed likelihood-based methods to assess interaction. A similar test was proposed by Umbach and Weinberg [121] based upon likelihood ratio tests. Alternatively, Eaves and Sullivan [122] used a logistic regression approach, extending the original tests of main effects proposed by Sham and Curtis [30]. This method provided separate tests of the main effects and the interaction effects. Lunetta et al. [123] proposed family-based tests for association and linkage by constructing a score statistic based upon the likelihood of the phenotypic distribution, given individual genotype. Their method is available in the FBAT software.

*Modeling Multiple Phenotypes*

Several authors have pointed out that effective use of multivariate phenotypes can potentially enhance the power of linkage analysis [124–134]. Analyzing each phenotype separately requires correction for multiple testing. Multiple phenotypes can be treated as predictor variables in a TDT-based logistic regression framework by treating transmission status as the dependent variable, using multivariate analysis of variance (MANOVA) [132].

*Inbreeding*

In a number of human populations, inbreeding is common and even encouraged. Bennett and Curnow [135] and Génin et al. [136] first investigated the consequences and benefits of using related parents on the TDT. The TDT remains a valid test of linkage in the presence of inbreeding but is not a valid test of association. However, when inbreeding is taken into account and no recombination exists between the disease susceptibility locus and the marker locus ($\theta = 0$), power to detect linkage is gained under certain genetic models: (1) recessive mode of inheritance and the frequency of the disease allele $<0.5$, (2) multiplicative or additive models, and (3) dominant mode of inheritance. Meanwhile, power increases with inbreeding coefficient but is considerably reduced when linkage disequilibrium between the marker and disease susceptibility locus is decreased [136].

*Bayesian TDT*

Information may be lost when known prior trait information (i.e. mode of inheritance, penetrance, etc.) is not incorporated into TDT-type analysis methods. In these circumstances Bayesian methods are an excellent alternative to more common frequentist approaches. For example, when mode of inheritance is known, incorporat-

ing this information results in an increase in power [137]. Additionally, joint and marginal posterior distributions of the recombination fraction and disequilibrium coefficient may be attained. However, the Bayes factor, or a measure of disagreement between two competing models, is not designed for error control; it is a measure of difference between prior beliefs only.

*Joint Testing of Linkage and Association*

The joint test approach was applied to TDT designs in order to capitalize on the information available in both covariance-based and marginal-based tests of linkage. The results showed that a multinomial joint test provides the highest overall power irrespective of allele frequency or mode of inheritance [26].

*Combining Case-Parent Trios and Unrelated Subjects*

In 2004, Nagelkerke et al. [138] proposed a likelihood-based association analysis of the data comprising of trio data, with unrelated controls, and possibly some unrelated cases. Nagelkerke et al. [138] provided ad hoc procedures to determine whether trios and unrelated data can be safely combined. Epstein et al. [139] modified the Nagelkerke et al. approach and provided formal statistical procedures to determine when it is appropriate to combine trios, unrelated controls, and unrelated cases together in a combined association analysis.

## Association Testing Methods for Population-Based Data Using Unrelated Individuals

Successfully identifying genes by linkage and association analyses using family-based designs can be difficult because the sample size required to achieve adequate power is often not attainable [140]. Hence, many researchers have turned to population-based association studies as a powerful tool for identifying these variants that underlie complex disease risk [141].

Genetic association studies aim to correlate differences in disease frequency with differences in allele frequencies at a particular genetic locus, where a specific allelic variant is either a direct disease-causing variant or is in linkage disequilibrium (LD) with the disease-causing variant. The most commonly used study design in population-based association studies is the case-control design. As with any study design, the case-control design assumes that the differences in allele frequencies between cases and controls relates directly to the trait of interest; in other words, there are no confounding effects [142].

Allele frequencies, however, are known to vary widely within and between populations, and these differences are widespread throughout the genome [143, 144]. When cases and controls have different allele frequencies attributable to variation in genetic ancestry within or between race/ethnicity groups, population stratification (PS) is said to be present, and ancestry becomes a confounding variable leading to spurious associations in the analysis. Redden and Allison [145] have shown that, contrary to popular conceptions, admixture-like patterns and spurious associations can occur in the presence of non-random mating patterns that would traditionally not be considered admixture, and they have evaluated the extent to which genomic control [146] and structured association testing [12, 13] can manage this potential confounding. PS is not only present in recently admixed populations like African Americans and Latinos [147–149], but also in European-American populations [150–153] and historically isolated populations including Tibeto-Burmans and Icelanders [154, 155].

As previously discussed, a consequence of PS in association studies is the potential for bias in the estimate of allelic associations due to deviations from the Hardy-Weinberg equilibrium and the induction of linkage disequilibrium [156, 157]. In order for bias due to PS to exist, both the frequency of the marker variant of interest and the background disease prevalence must vary significantly by race/ethnicity [158, 159]. If either of these conditions is not fulfilled, bias due to PS cannot occur. Bias due to PS can induce both false positive [1, 3, 4, 8] and false negative associations [160]. Controlling for self-reported race has generally been thought to suffice [161], but recent data shows that matching on ancestry is more robust; however, in many populations, whether recently admixed or not, individuals are not aware of their precise ancestry [4, 162].

No true consensus has been reached on how to test and/or adjust for population stratification [158, 159, 163], although many methods have been developed [11–13, 146, 164, 165]. Here, we provide short descriptions of association methods designed to identify genetic variants predisposing to the trait while simultaneously controlling for stratification in population-based data. These methods can be grouped into three categories: Genomic Control (GC), Structured Association Testing (SAT), and Regional Admixture Mapping (RAM).

*Genomic Control*

One of the early methods developed to control for PS or admixture induced confounding was genomic control.

It is based on the idea that the false positive rate (Type I error) increases in the presence of PS. The GC technique uses a set of non-candidate, random markers (sometimes called null markers) to estimate an inflation factor, λ; λ is equal to 1 if there is no population stratification present. Estimates for λ have been suggested for additive genetic models [146, 166] and for dominant/recessive models [167]. This inflation is assumed to be caused by population stratification and the GC method corrects the standard $\chi^2$ association test by this factor λ, where the new $\chi^2/\lambda$ test statistic still has $\chi^2$ distribution. Therefore, GC performs uniform adjustment to all association tests assuming the same inflation factor. One of the main assumptions of this method is that if the study population comes from larger population made up of a mixture of subpopulations with different disease prevalences and disease allele frequencies, then the $\chi^2$ association test statistic follows a non-central $\chi^2$ distribution [11]. If the non-central parameter is truly small, then adjusting by the estimated inflation factor λ is a good approximation to the distribution, however, if the non-centrality parameter is truly large then adjusting for the estimated inflation factor λ will not be sufficient to prevent false positive associations and loss of statistical power [168]. If AIMs are used instead of random markers, more false positive associations will result simply because the AIMs show large population differences in allele frequencies and there will be a tendency towards over-correction [168]. However, GC is a relatively computationally easy method to implement and interpret. In addition, Bacanu et al. [166] have shown that the GC approach is more powerful than TDT barring substantial population stratification.

*Structured Association Testing*

Some structured association methods utilize Bayesian techniques to assign individuals to 'clusters' or subpopulation classes using information from a set of non-candidate, unlinked loci and then tests for association within each 'cluster' or subpopulation class [2, 8, 12, 13, 19, 169]. The clusters are determined using individual ancestry estimates or principal component analysis (PCA). To estimate the ancestry of each individual, the genetic markers with different allele frequencies in the founding populations (called ancestry informative markers, or AIMs) are used to estimate the proportion of an individual's genome derived from each founding population. These proportions are then used to cluster individuals into subpopulations and to control for population structure during association testing. Alternatively, one can use PCA to estimate a genetic background score for each individual based on the AIMs, and control for stratification by accounting for variation in the data associated with the differences in allele frequencies [10, 15, 18, 170]. Satten et al. [14] proposed a latent class logistic regression procedure that simultaneously estimates PS and tests for association between a marker allele and a binary phenotype, assuming the marker loci are unrelated to disease and in linkage equilibrium with a putative disease gene in the same subpopulation. The advantages of this model are that it offers a unified treatment of both association and PS, using straightforward likelihood estimation accounting for substructure differences that will occur between cases and controls, which is ignored by Pritchard et al. [12, 13]. However, it has a significant drawback in that there is no software available for the analysis, which is complex given the number of nuisance parameters it involves.

*Regional Admixture Mapping*

RAM methodology builds upon the following premises: (1) Disease mutation occurred in one population and propagated into another through inter-mating (i.e. prevalence of the disease in the donor population is higher than in the recipient population); (2) the process of recent admixture creates disequilibrium among linked loci that tends to extend over longer genetic distances in the admixed population compared to the non-admixed population, and (3) the degree of individual regional admixture will vary with disease-predisposing loci and also in disequilibrium with loci even after appropriately adjusting for the degree of individual ancestry. Rife [171] was the first to point out that hybrid populations can provide useful information regarding linkage. RAM is also known as admixture mapping [17, 19–21, 169, 172], mapping by admixture disequilibrium (MALD) [173], and marker location-specific ancestry mapping [18]. RAM is a form of association testing in which genome-wide ancestry estimates and region-specific ancestry estimates are used to identify specific regions of the genome potentially harboring loci predisposing the disease or trait.

McKeigue may have been the first to introduce RAM based on sound statistical principles to control for spurious associations induced by variations in ancestry [174]. McKeigue noted that, if one conditioned on the admixture of the individuals' parents, linkage could be detected by testing for the association of a phenotype with the ancestry of alleles at a marker locus in an admixed sample. Using a combination of Bayesian and frequentist approaches, he employed a Hidden Markov Model (HMM) to generate the posterior distribution of individual admixture in the population and utilized likelihood-based

**Table 1.** Summary and comparisons of individual admixture estimation methods

| | Structure [11–13, 169] | AdmixMap [2, 8] | FRAPPE [172]/ PSMIX [180] | AncestryMap [19] |
|---|---|---|---|---|
| Methodology | Bayesian | Bayesian | ML | Bayesian |
| Software readily available? | yes | yes | yes | yes |
| Allows for family data? | no | no | no | no |
| Allows for more than two parental populations? | yes | yes | yes | no |
| Requires members of founding populations in analysis? | yes | no | yes | no |
| Requires estimates of allele frequencies in founding populations? | no | yes | no | yes |
| Provides prior estimates of founding allele frequencies? | yes | yes | yes | yes |
| Requires ancestry informative markers? | no | yes | yes | yes |
| Designed for linked markers? | yes | yes | no | yes |
| Gives region-specific estimates? | yes | yes | no | yes |
| Provides credible or confidence intervals? | yes | no | yes | no |
| Speed of analysis | very slow | fast | very fast | fast |
| Provides for multiple models of admixture | yes | no | no | no |
| Ease of use/user friendly | very easy | difficult | easy | difficult |
| Allows for missing data? | yes | yes | no | yes |
| Accommodates measurement error in ancestry? | no | yes | no | no |
| Studied through simulations? | yes | yes – limited | yes – limited | yes – limited |
| Tested with 'real' data? | yes | yes | yes | yes |

score statistics for linkage testing [174–177]. In the RAM method of Patterson et al. [19], a HMM is used to scan the genome to identify regions associated with a particular trait or phenotype by simultaneously estimating individual admixture proportion and individual region-specific admixture proportions, and testing for linkage of specific genomic regions to specific phenotypes. Patterson et al. [19] used a likelihood ratio test for the case-only design but offered a simple t test for use in a case-control design; these algorithms are available in the software *AncestryMap*.

Zhu et al. [18] developed a method that has several advantages and is intended to be an extension to McKeigue [174, 176]. In practice, assumptions made by McKeigue [174, 176] about admixture patterns are unlikely to hold for natural populations, resulting in an inflation of the type I error rate when testing for linkage by the McKeigue method. They generalized McKeigue's approach to allow for two different admixture models: (1) hybrid isolation admixture, and (2) a continuous gene flow model. Zhu et al.'s method is very similar to Patterson et al.'s [19] approach for case-only testing but uses a two-stage rather than simultaneous estimation and testing procedure.

Montana and Pritchard [17] also introduced a test of whether cases' region-specific admixture values are significantly different from their genome-wide admixture value, as implemented in the *MALDsoft* software. They recommend that the following should be collected for proper analysis: (a) a sample of affected individuals from the admixed population; (b) a sample of unaffected or random control individuals, also from the admixed population, and (c) 'learning samples' that consist of random individuals from each of the ancestral populations (or a close approximation thereof) that can be used to estimate the ancestral allele frequencies. However, they note that 'it is preferable but not required to have both controls and learning samples.' Montana and Pritchard also use a HMM to estimate the admixture proportion in the first stage and then conduct simple t tests in a second stage. In a highly similar paper, Zhang et al. [20] develop a simultaneous HMM framework for estimating region-specific and genome-wide individual admixture values and then incorporate these values in a logistic, regression-like framework to model case-control data. They present simulation results suggesting that their method works well for both hybrid-isolation and continuous gene-flow models. Redden et al. [16] developed a method based on a generalized linear model that can accommodate both SAT and RAM tests and can be used in standard statistical software, such as SAS. Recently, Clarke and Whittemore [178] proposed an admixture mapping test for a case only study design. The test compares the case's ancestry as inferred from his/her marker genotypes to the ancestry inferred from information from their family [176].

*Software*

Several software packages are available to analyze recently admixed populations. These methods generally fall into one of two categories: Bayesian or Maximum likelihood. There are advantages and disadvantages of both the Bayesian (*AdmixMap*, *AncestryMap,* and *Structure* [2, 8, 19, 169]) and the ML methods (*FRAPPE, IBGA*, and *PSMIX* [172, 179, 180]). Table 1 gives a listing of these available programs and compares important features. *PSMIX* was chosen as a representative method for the very similar ML methods. *Structure* is by far the most popular of all of these programs, and has been used by a number of investigators for a wide variety of populations and complex phenotypes.

A concern with all of the methods (except some of the ML methods) is the limited, or restricted, amount of testing that has been conducted. Several authors reported correlations of their estimates with *Structure* estimates but not of their estimates with true admixture, as determined by simulation. Tang et al. provide the only direct evaluation of individual admixture estimates [172]. They show via simulation that, with informative markers and well represented parental populations, both *Structure* and *FRAPPE* estimations work well. With less informative markers, or only a few members of the parental populations, the *FRAPPE* estimation is unbiased while *Structure* estimates can be highly biased.

## Conceptual Frameworks: Seeking the Connections

Exploring how the various methods are connected can help to identify why these methods work, what their underlying assumptions are, which are simply special cases of and redundant with another, and potentially point the way to gaps where new methods may be needed or how existing methods might be improved. Here we describe two frameworks with which one can conceptualize the extant methods. We then categorize methods via these frameworks as summarized in table 2.

### A Conceptual Framework for Joint Tests of Linkage and Association Test Methods Based on Underlying Statistical Principles

In tests of association in the presence of linkage (TALs) (aka, tests of linkage in the presence of association; joint tests of linkage and association), we wish to identify situations in which (A) genotypes at a marker locus (either directly or indirectly through intermediary phenotypes) cause variations in a phenotype; or (B) the marker locus is in linkage disequilibrium with another locus at which genotypes cause variations in the phenotype; and to distinguish those situations from (C) situations in which genotypic variation at the marker locus is correlated with (but not linked to) some other inherited factor that causes variation in the phenotype. Although less commonly discussed, we also wish to (D) identify marker loci that are both linked to loci that (either directly or indirectly through intermediary phenotypes) cause variations in a phenotype and, when certain other variables are conditioned on, also associated with loci that (either directly or indirectly through intermediary phenotypes) cause variations in a phenotype; even in situations where (E) genotypic variation at the marker locus is not associated with variations in the phenotype in the absence of conditioning on those other variables (i.e., when the association is *masked* [160] or *suppressed* [181]).

Everything in the preceding paragraph is just another way of saying we need to control for potential *confounding* so that we may infer a causal influence of the marker locus itself or something in linkage disequilibrium with it on the phenotype. The ultimate source of the potential confounding that we wish to control for in TDT-type tests in *non-linkage disequilibrium* (NLD), i.e., correlation or disequilibrium among unlinked loci. NLD can result from many sources including selection [182], assortative mating [145], and the admixture process [16].

There is a rich literature on detecting causal effects in scientific research. To the extent that causation can ever be determined, most methodologists concur that we can have no stronger basis than a randomized experiment [183]. This is because the act of randomization assures that, in the hypothetical population to which we wish to make inferences (not the specific sample in hand), there can be no association between the independent variable to which we assign subjects and *any* variable that existed prior to randomization. Therefore, randomization is the only method that controls for both known and unknown sources of confounding. Thus, in an ideal world, we would randomly assign individuals to genotypes at marker loci and then do our tests without any concern of confounding. Of course, in reality, this is not possible. So what is the next best thing?

We can find the root of the next best thing in the work of Gregor Mendel's second law of genetics – the law of independent assortment. Mendel [184] wrote *'All constant combinations which in peas are possible by the combination of the said 7 differentiating characters were actu-*

**Table 2.** Comparison of statistical methods based on sound statistical principles

| Design | Methodology | Reference | Principles/properties | | |
|---|---|---|---|---|---|
| | | | randomization (selection/assignment) | conditioning on sufficient statistics | marginal/ covariance-based test |
| Family-based | TDT-type | Spielman et al. [5]: TDT | yes | yes | yes/no |
| | | Sham and Curtis [30]: ETDT | yes | no | yes/no |
| | | Cleves et al. [34]: TDT-EX | yes | yes | yes/no |
| | | Allison [37]: TDTQ5 | yes | yes | no/yes |
| | | Rabinowitz [39]: $S_{max}$ | yes | yes | no/yes |
| | | George et al. [41]: SAGE | no | yes | no/yes |
| | | Abecasis et al. [43, 44]: QTDT | yes | yes | no/yes |
| | | Spielman and Ewens [52]: S-TDT | yes (if only one affected and one unaffected sibs are used) | yes (conditioned on sibships) | yes/no |
| | | Sun et al. [61]: 1-TDT | no | no | yes/no |
| | | Knapp [68]: RC-TDT | yes | yes (conditioned on reconstructed parental genotypes) | yes/no |
| | | Clayton and Jones [94]: TRANSMIT | yes | yes | no/yes |
| | | Martin et al. [75]: PDT | yes | yes | yes/no |
| | | Rabinowitz and Laird [72]: FBAT | yes (no sampling strategy required/any pedigree) | yes | yes/yes |
| | | Guo et al. [86]: i-TDT | yes | yes | yes/no |
| | | Horvath et al. [88, 192]: XS-TDT; XRC-TDT | yes | yes (conditional on reconstructed parental genotypes) | yes/no |
| | | Weinberg et al. [50]: LRT | affected sibs trios | yes | no/yes |
| | | Hu et al. [118, 119] | yes | yes | yes/no |
| | | Umbach and Weinberg [121]: LRT | yes (affected/unaffected trios) | yes | no/yes |
| | | George and Laud [137]: Bayesian TDT | yes | yes | yes/no |
| | | Tiwari et al. [26] | yes | yes | yes/yes |
| | | Lunetta et al. [123] | yes | yes | no/yes |
| Population-based | GC | Devlin and Roeder [146] | yes | no | yes/no |
| | SAT | Pritchard and Rosenberg, Pritchard et al. [11–13] | yes | yes | yes/no |
| | RAM | Patterson et al. [19] | no | yes | yes/no |
| | | Zhu et al. [18] | no | yes | yes/no |

*ally obtained by repeated crossing. Their number is given by* $2^7 = 128$*. Thereby is simultaneously given the practical proof that the constant characters which appear in the several varieties of a group of plants may be obtained in all the associations which are possible according to the laws of combination, by means of repeated artificial fertilization*.'[2] A more formal statement of the law of independent assortment is '*When gametes are formed the alleles for one trait segregate independently of the alleles of a gene for another trait*.' In other words, Mendel believed that genes for different traits segregate independently. We now know that this is only true for genes at unlinked loci. Nevertheless, Mendel's second law implies that every act of meiosis is an act of randomization in which parents randomly assign alleles to the gametes they form from their available alleles. This further implies that, *conditional upon parents' genotypes*, all individuals have equal probability of inheriting (i.e., being assigned to) any particular genotype. Thus, *conditional upon parents' genotypes*, individuals are essentially randomized to genotypes. The only caveat (which often works in our favor in genetic research) is that the genotypes to which individuals are randomly assigned at one locus will be correlated with the genotypes to which they are assigned at other loci, but only when the loci in question are physically linked. Hence, conditioning on parents' genotypes offers us a natural randomized experiment that eliminates the possibility of confounding by NLD. It does not eliminate potential confounding by LD, but this 'confounding' by LD is actually just what we are counting on to help us identify genes in many association studies (especially genome-wide association studies).

How can we condition on parents' genotypes? There are several ways in which this can be achieved. The first

[2] From: http://www.mendelweb.org/Mendel.html.

and most straightforward way would be to begin with two individuals that are of the opposite sex and, at every locus, are homozygous. However, at many loci the two individuals will be different from each other. If such individuals produce a large number of offspring, these offspring will all be genetically identical and heterozygous at every locus at which the two parents differed. These offspring can then be intermated to produce another generation. In the second generation that descends from the original set of parents (conventionally denoted the F2 generation) every individual would have an equal probability of being assigned to each genotype compared with every other individual. Thus, individuals are essentially randomized to genotypes and we have the equivalent of a true experiment with randomization. This is essentially a description of a F2 cross among inbred lines that is classically used to map genes for complex traits in animals such as mice and flies. It is noteworthy that the individuals comprising the F2 population are admixed. And yet there is no concern about confounding due to admixture because all individuals have the same ancestry. As pointed out by Redden et al. [16] this indicates that it is variation in ancestry and not variations in admixture per se that can cause confounding by NLD. Thus, the F2 cross among inbred lines can be seen as the geneticist's experiment in which meiosis is used to enact the process of randomization. It can also be seen as a precursor to the TDT.

Of course, we cannot set up inbred lines and do controlled breeding in humans. How then can we achieve similar objectives? We can do so by recognizing that in order for individuals to be assigned essentially at random (i.e. with equal probability across individuals) to genotypes *at the marker locus*, it is only necessary that their parents have the same genotypes at the marker locus, not that their parents are genetically identical with every other set of parents at *all* loci. Hence, we should select only individuals whose parents all had a common genotype at the marker locus. For example, if we had a locus that was di-allelic with alleles *A* and *a*, we could select only individuals in which one parent was *AA* and the other parent was *Aa*. In the offspring, we could then assess whether individuals who 'randomly' receive an '*a*' allele from one of their parents tend to be phenotypically different than individuals that receive no '*a*' alleles from their parents. Such a design would, at the locus in question, essentially recapitulate a backcross among an experimental population such as mice in which heterozygotes at the F1 generation are backcrossed to one of the parental strains. A design in which we only selected individuals whose parents had the genotypes *Aa* and *Aa* would essentially recapitulate an F2 cross at that locus.

In practice of course, the approach described in the preceding paragraph would be infeasible. Instead, rather than selecting individuals who only have parents with particular genotypes, we can statistically control for (i.e., condition on) the two parental genotypes (which we often denote mating types). This yields equivalent control because conditional upon the parental genotypes, the assignment to offspring genotypes is essentially random. Thus, our second way of achieving the benefits of randomization of allowing strong causal inferences and eliminating confounding by NLD is to statistically control for parental genotypes by directly observing them and including them in the statistical models. This is the basis of several TDTs [e.g. Allison, 37]. Using a similar argument Tiwari et al. [26] and Beasley et al. [185] apply the rules of randomization by conditioning on parental genotypes.

Of course, one may not be able to or wish to observe the genotypes of the parents themselves. One can then recognize that full siblings (by definition) share the same parents. Therefore, if one controls for sibship using studies of multiple siblings, one has effectively controlled for parents' genotypes because all siblings have the parents with the same genotypes offering yet another way to effectively condition upon parental genotypes and enjoy the inferential strength that randomization offers.

As this discussion indicates, there are multiple variables one could control for that may yield valid inference in this context allowing the randomization by meiosis to eliminate confounding by NLD. Rabinowitz and colleagues [39, 72, 84, 186, 187] have extended this idea to talk about conditioning on sufficient statistics. They seek to identify statistics that are 'sufficient' in the sense that if conditioned upon they would eliminate confounding by NLD. At root, this is still the same concept but expressed in a different form. This different expression of the concept is the basis for several other TDT type approaches [37, 69, 78]. Horvath et al. [188] express succinctly the importance of conditioning on sufficient statistics: 'The general principle is to evaluate the distribution of test statistics using the conditional distribution of offspring genotypes under the null hypothesis, where the conditioning is on the sufficient statistics for any nuisance parameters in the model [72]. The potential nuisance parameters for nuclear families include the distribution of the phenotypes, the parental allele frequencies, and the model for ascertainment. By conditioning the offspring genotype distribution on the phenotypes, one eliminates sensitivity of the tests to misspecification of

the phenotype distribution and to ascertainment conditions that depend on the phenotypes. Conditioning on the parental genotypes eliminates sensitivity to population admixture, when parents' genotypes are unknown. The procedures in Allison's TDTs [37], George et al. [41], Allison et al. [69], and FBAT [73, 78] all correct for association by conditioning on the parental genotype or transmission status of the individual.

Finally, a new class of tests known as structured association tests [16] attempt to use the rest of the genome to derive, via various machinations, a variable that, if conditioned upon, would control for or eliminate NLD as a confounder. In the original formulations of such approaches, the variable one sought to control for was an index of genetic admixture under some assumption of a particular population dynamic including population admixture [2, 8, 11–13, 19, 169, 172]. More recently, these approaches are being extended to allow for other background genetic factors [10, 15]. It is important to note that unlike family-based TDT-type approaches which strictly eliminate confounding by NLD, such approaches as structured association testing only do so to the extent that one has effectively captured the important background covariates for inclusion in the model and modeled them successful [16]. Expressed in this way, one can see structured association testing as essentially trying to achieve the same goals that propensity score analysis attempts to achieve in more general epidemiologic studies [189, 190]. Indeed, our group is currently working on formalizing the propensity score analysis approach to structured association testing. Note that the genomic control method [191] achieves valid inference by correcting the variance inflation factor rather than conditioning on sufficient statistics.

## Conclusion

In conclusion, we have reviewed family- and population-based designs that have been in the literature proposed for eliminating or controlling for confounding due to population stratification in order to draw valid inference in the context of genetic linkage and association studies. Also, we described how and why these methods of linkage and association testing follow general statistical principles such as (1) randomization, (2) conditioning on sufficient statistics, and (3) identifying whether the method is based on testing the genotype-phenotype covariance (conditional upon familial information) and/or testing departures of the marginal distribution from the expected genotypic frequencies. Because of the vast number of options available, the reader is cautioned to take care when applying these methods in terms of meeting the required assumptions and assuring that the method is testing the hypothesis that the reader intends to test.


## Acknowledgements

This study is supported in part by R21LM008791, T32AR007450, R01DK52431, U01HL072510-02, P20RR016430, P30DK56336, R01RR017009, R01DK56366, R01ES09912, P01AR049084, U54CA100949, R01GM077490, R01AR052658, and 3R01AR052658-03S1, 2R01HL055673-11A1, R01GM074913-01A1.


## Glossary

*Individual admixture:* The degree to which the genome of an individual in the population that results from the intermating process is composed of DNA segments descended from one particular parental population relative to others.

*Individual ancestry:* Proportion of the ancestors of an individual from the resulting population who were from one particular parental population in a generation before intermating.

*Assortative mating:* Assortative mating can refer to any nonrandom mating patterns, but most often refers to the common observation that individuals are more likely to mate with phenotypically similar individuals.

*Confounding:* It is a type of bias when an association between a risk factor and disease can be explained by a factor associated with both disease and risk factor.

*Disequilibrium:* The dependence between any two loci is called disequilibrium irrespective of the locations of the loci.

*Linkage:* Two loci are linked when they are close to each other on the same chromosome.

*Linkage disequilibrium:* The dependence between two loci is called linkage disequilibrium if both loci are linked, i.e., marker locus is close to disease locus on the same chromosome.

*Masking:* See *suppression* below.

*Non-linkage disequilibrium* (NLD)*:* We use the term NLD to refer to situations in which two loci are in disequilibrium, but not linked. NLD is the ultimate source of potential confounding that we wish to control for in TDT-type tests.

*Panmixia:* Panmixia is used as synonym for random mating within a breeding population.

*Stratification:* Stratification exists when the population has been formed by admixture of diversified subpopulations with different allele frequencies.

*Structure:* Subpopulations in the study population.

*Sufficient statistic:* A statistic is sufficient for a parameter if it gives as much information about the parameter as do the full data.

*Suppression/suppressor variable:* A suppressor variable has zero (or close to zero) correlation with the independent variable but is correlated with one or more of the predictor variables, and therefore, it will suppress irrelevant variance of independent variables.

## References


1 Halder I, Shriver M: Measuring and using admixture to study the genetics of complex diseases. Hum Genet 2003;1:52–62.
2 Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM: Design and analysis of admixture mapping studies. Am J Hum Genet 2004;74:965–978.
3 Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. Nat Genet 2004;36:512–517.
4 Ziv E, Burchard EG: Human population structure and genetic association studies. Pharmacogenomics 2003;4:431–441.
5 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–516.
6 Rubinstein P, Walker M, Carpenter C, Carrier C, Krassner J, Falk C, Ginsberg F: Genetics of HLA disease and associations: The use of the haplotype relative risk (HRR) and the 'haplo-delta' (Dh) estimates in juvenile diabetes from three racial groups. Hum Immunol 1981;3:384.
7 Falk CT, Rubinstein P: Haplotype relative risks: An easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 1987;51:227–233.
8 Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM: Control of confounding of genetic associations in stratified populations. Am J Hum Genet 2003;72:1492–1504.
9 Purcell S: Sample selection and complex effects in quantitative trait loci analysis. Ph.D. Dissertation. University of London, 2003.
10 Chen HS, Zhu X, Zhao H, Zhang S: Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. Ann Hum Genet 2003;67:250–264.
11 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999;65:220–228.
12 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: Association mapping in structured populations. Am J Hum Genet 2000;67:170–181.
13 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. Genetics 2000;155:945–959.
14 Satten GA, Flanders WD, Yang QH: Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 2001;68:466–477.
15 Zhang SL, Zhu XF, Zhao HY: On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genet Epidemiol 2003;24:44–56.
16 Redden D, Divers J, Vaughan L, Tiwari H, Beasley T, Fernandez J, Kimberly R, Feng R, Padilla M, Lui N, Miller M, Allison D: Regional admixture mapping and structured association testing: Conceptual unification and an extensible general linear model. PLoS Genet 2006;2:1254–1264.
17 Montana G, Pritchard JK: Statistical tests for admixture mapping with case-control and cases-only data. Am J Hum Genet 2004;75:771–789.
18 Zhu X, Cooper RS, Elston RC: Linkage analysis of a complex disease through use of admixed populations. Am J Hum Genet 2004;74:1136–1153.
19 Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly MJ, Reich D: Methods for high-density admixture mapping of disease genes. Am J Hum Genet 2004;74:979–1000.
20 Zhang WH, Collins A, Gibson J, Tapper WJ, Hunt S, Deloukas P, Bentley DR, Morton NE: Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. Proc Natl Acad Sci USA 2004;101:18075–18080.
21 Freeman AR, Meghen CM, MacHugh DE, Loftus RT, Achukwi MD, Bado A, Sauveroche B, Bradley DG: Admixture and diversity in West African cattle populations. Mol Ecol 2004;13:3477–3487.
22 Schaid DJ, Sommer SS: Genotype relative risks: Methods for design and analysis of candidate-gene association studies. Am J Hum Genet 1993;53:1114–1126.
23 Ott J: Statistical properties of the haplotype relative risk. Genet Epidemiol 1989;6:127–130.
24 Terwilliger JD, Ott J: A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Hum Hered 1992;42:337–346.
25 McNemar Q: Note on the sampling error of the differences between correlated proportions of percentages. Psychometrika 1947;12:153–157.
26 Tiwari HK, Holt J, George V, Beasley TM, Amos CI, Allison DB: New joint covariance- and marginal-based tests for association and linkage for quantitative traits for random and non-random sampling. Genet Epidemiol 2005;28:48–57.
27 Ewens WJ, Spielman RS: The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 1995;57:455–464.
28 Bickeboller H, Clerget-Darpoux F: Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. Genet Epidemiol 1995;12:865–870.
29 Rice JP, Neuman RJ, Hoshaw SL, Daw EW, Gu C: TDT with covariates and genomic screens with mod scores: Their behavior on simulated data. Genet Epidemiol 1995;12:659–664.
30 Sham PC, Curtis D: An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. Ann Hum Genet 1995;59:323–336.
31 Morris AP, Whittaker JC, Curnow RN: A likelihood ratio test for detecting patterns of disease-marker association. Ann Hum Genet 1997;61:335–350.
32 Spielman RS, Ewens WJ: The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 1996;59:983–989.
33 Kaplan NL, Martin ER, Weir BS: Power studies for the transmission/disequilibrium tests with multiple alleles. Am J Hum Genet 1997;60:691–702.
34 Cleves MA, Olson JM, Jacobs KB: Exact transmission-disequilibrium tests with multiallelic markers. Genet Epidemiol 1997;14:337–347.
35 Schaid DJ: General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 1996;13:423–449.
36 Betensky RA, Rabinowitz D: Simple approximations for the maximal transmission/disequilibrium test with a multi-allelic marker. Ann Hum Genet 2000;64:567–574.
37 Allison DB: Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet 1997;60:676–690.
38 Xiong MM, Krushkal J, Boerwinkle E: TDT statistics for mapping quantitative trait loci. Ann Hum Genet 1998;62:431–452.
39 Rabinowitz D: A transmission disequilibrium test for quantitative trait loci. Hum Hered 1997;47:342–350.
40 Sun FZ, Flanders WD, Yang QH, Zhao HY: Transmission/disequilibrium tests for quantitative traits. Ann Hum Genet 2000;64:555–565.
41 George V, Tiwari HK, Zhu X, Elston RC: A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. Am J Hum Genet 1999;65:236–245.
42 Zhu X, Elston RC: Transmission/disequilibrium tests for quantitative traits. Genet Epidemiol 2001;20:57–74.
43 Abecasis GR, Cookson WOC, Cardon LR: Pedigree tests of transmission disequilibrium. Eur J Hum Genet 2000;8:545–551.
44 Abecasis GR, Cardon LR, Cookson WOC: A general test of association for quantitative traits in nuclear families. Am J Hum Genet 2000;66:279–292.
45 Fulker DW, Cherny SS, Sham PC, Hewitt JK: Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 1999;64:259–267.

46 Monks SA, Kaplan NL: Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus. Am J Hum Genet 2000;66:576–592.
47 Waldman ID, Robinson BF, Rowe DC: A logistic regression based extension of the TDT for continuous and categorical traits. Ann Hum Genet 1999;63:329–340.
48 Liu Y, Tritchler D, Bull SB: A unified framework for transmission-disequilibrium test analysis of discrete and continuous traits. Genet Epidemiol 2002;22:26–40.
49 Kistner EO, Weinberg CR: Method for using complete and incomplete trios to identify genes related to a quantitative trait. Genet Epidemiol 2004;27:33–42.
50 Weinberg CR, Wilcox AJ, Lie RT: A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet 1998;62:969–978.
51 Curtis D: Use of siblings as controls in case-control association studies. Ann Hum Genet 1997;61:319–333.
52 Spielman RS, Ewens WJ: A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 1998;62:450–458.
53 Schaid DJ, Rowland C: Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. Am J Hum Genet 1998;63:1492–1506.
54 Horvath S, Laird NM: A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 1998;63: 1886–1897.
55 Boehnke M, Langefeld CD: Genetic association mapping based on discordant sib pairs: The discordant-alleles test. Am J Hum Genet 1998;62:950–961.
56 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 1999;9:234–241.
57 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 1998; 8:1273–1288.
58 Weinberg CR: Allowing for missing parents in genetic studies of case-parent triads. Am J Hum Genet 1999;64:1186–1193.
59 Schaid DJ, Rowland CM: Quantitative trait transmission disequilibrium test: allowance for missing parents. Genet Epidemiol 1999; 17(suppl 1):S307–S312.
60 Siegmund KD, Langholz B, Kraft P, Thomas DC: Testing linkage disequilibrium in sibships. Am J Hum Genet 2000;67:244–248.
61 Sun F, Flanders WD, Yang Q, Khoury MJ: Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. Am J Epidemiol 1999;150:97–104.
62 Wang D, Sun F: Sample sizes for the transmission diequilibrium tests: TDT, D-TDT and 1-TDT. Theory Methods 2002;29:1129–1140.
63 Clayton D: A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 1999; 65:1170–1177.
64 Cervino AC, Hill AV: Comparison of tests for association and linkage in incomplete families. Am J Hum Genet 2000;67:120–132.
65 Allen AS, Rathouz PJ, Satten GA: Informative missingness in genetic association studies: Case-parent designs. Am J Hum Genet 2003;72:671–680.
66 Allen AS, Collins JS, Rathouz PJ, Selander CL, Satten GA: Bootstrap calibration of TRANSMIT for informative missingness of parental genotype. Bmc Genetics 2003;4: (suppl 2):S39.
67 Spielman RS, Ewens WJ: TDT clarification. Am J Hum Genet 1999;64:668.
68 Knapp M: The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. Am J Hum Genet 1999; 64:861–870.
69 Allison DB, Heo M, Kaplan N, Martin ER: Sibling-based tests of linkage and association for quantitative traits. Am J Hum Genet 1999;64:1754–1763.
70 van den Oord EJ: The use of mixture models to perform quantitative tests for linkage disequilibrium, maternal effects, and parent-of-origin effects with incomplete subject-parent triads. Behav Genet 2000;30:335–343.
71 Whittemore AS, Tu IP: Detection of disease genes by use of family data. I. Likelihood-based theory. Am J Hum Genet 2000;66: 1328–1340.
72 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 2000;50:211–223.
73 Horvath S, Xu X, Laird NM: The family based association test method: Strategies for studying general genotype-phenotype associations. Eur J Hum Genet 2001;9:301–306.
74 Martin ER, Monks SA, Warren LL, Kaplan NL: A test for linkage and association in general pedigrees: The pedigree disequilibrium test. Am J Hum Genet 2000;67:146–154.
75 Martin ER, Bass MP, Kaplan NL: Correcting for a potential bias in the pedigree disequilibrium test. Am J Hum Genet 2001;68:1065–1067.
76 Martin ER, Bass MP, Hauser ER, Kaplan NL: Accounting for linkage in family-based tests of association with missing parental genotypes. Am J Hum Genet 2003;73:1016–1026.
77 Martin ER, Bass MP, Gilbert JR, Pericak-Vance MA, Hauser ER: Genotype-based association test for general pedigrees: The genotype-PDT. Genet Epidemiol 2003;25: 203–213.
78 Laird NM, Horvath S, Xu X: Implementing a unified approach to family-based tests of association. Genet Epidemiol 2000;19:S36–S42.
79 Schaid DJ, Sommer SS: Comparison of statistics for candidate-gene association studies using cases and parents. Am J Hum Genet 1994;55:402–409.
80 Schaid DJ, Li H: Genotype relative-risks and association tests for nuclear families with missing parental data. Genet Epidemiol 1997;14:1113–1118.
81 Tu IP, Balise RR, Whittemore AS: Detection of disease genes by use of family data. II. Application to nuclear families. Am J Hum Genet 2000;66:1341–1350.
82 Shih MC, Whittemore AS: Tests for genetic association using family data. Genet Epidemiol 2002;22:128–145.
83 Whittemore AS, Halpern J: Genetic association tests for family data with missing parental genotypes: A comparison. Genet Epidemiol 2003;25:80–91.
84 Rabinowitz D: Adjusting for population heterogeneity and misspecified haplotype frequencies when testing nonparametric null hypotheses in statistical genetics. J Am Stat Assoc 2002;97:742–751.
85 Martin ER, Kaplan NL, Weir BS: Tests for linkage and association in nuclear families. Am J Hum Genet 1997;61:439–448.
86 Guo CY, Lunetta KL, DeStefano AL, Ordovas JM, Cupples LA: Informative-transmission disequilibrium test (i-TDT): Combined linkage and association mapping that includes unaffected offspring as well as affected offspring. Genet Epidemiol 2007;31:115–133.
87 Diao G, Lin DY: Improving the power of association tests for quantitative traits in family studies. Genet Epidemiol 2006;30:301–313.
88 Horvath S, Laird NM, Knapp M: The transmission/disequilibrium test and parental-genotype reconstruction for X-chromosomal markers. Am J Hum Genet 2000;66: 1161–1167.
89 Horvath S, Laird NM: A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 1998;63: 1886–1897.
90 Ho GY, Bailey-Wilson JE: The transmission/disequilibrium test for linkage on the X chromosome. Am J Hum Genet 2000;66: 1158–1160.
91 Van der Meulen MA, te Meerman GJ: Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. Genet Epidemiol 1997;14: 915–920.

92 Wilson SR: On extending the transmission/disequilibrium test (TDT). Ann Hum Genet 1997;61:151–161.
93 Collins A, Morton NE: Mapping a disease locus by allelic association. Proc Natl Acad Sci USA 1998;95:1741–1745.
94 Clayton D, Jones H: Transmission/disequilibrium tests for extended marker haplotypes. Am J Hum Genet 1999;65:1161–1169.
95 McPeek MS, Strahs A: Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet 1999;65:858–875.
96 Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK: Transmission/disequilibrium tests using multiple tightly linked markers. Am J Hum Genet 2000;67:936–946.
97 MacLean CJ, Martin RB, Sham PC, Wang H, Straub RE, Kendler KS: The trimmed-haplotype test for linkage disequilibrium. Am J Hum Genet 2000;66:1062–1075.
98 Seltman H, Roeder K, Devlin B: Transmission/disequilibrium test meets measured haplotype analysis: Family-based association analysis guided by evolution of haplotypes. Am J Hum Genet 2001;68:1250–1263.
99 Yu K, Zhang S, Borecki IB, Kraja A, Xiong C, Myers R, Province MA: A haplotype similarity based transmission/disequilibrium test under founder heterogeneity. Ann Hum Genet 2005;69:455–467.
100 Tzeng JY, Devlin B, Wasserman L, Roeder K: On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet 2003;72:891–902.
101 Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F: Search for multifactorial disease susceptibility genes in founder populations. Ann Hum Genet 2000;64:255–265.
102 Zhang S, Sha Q, Chen HS, Dong J, Jiang R: Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. Am J Hum Genet 2003;73:566–579.
103 Sha Q, Dong J, Jiang R, Chen HS, Zhang S: Haplotype sharing transmission/disequilibrium tests that allow for genotyping errors. Genet Epidemiol 2005;288:341–351.
104 Levinson DF, Kirby A, Slepner S, Nolte I, Spijker GT, te Meerman GJ: Simulation studies of detection of a complex disease in a partially isolated population. Am J Med Genet. 2001;105:65–70.
105 Beckmann L, Thomas DC, Fischer C, Chang-Claude J: Haplotype sharing analysis using mantel statistics. Hum Hered 2005;59:67–78.
106 Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN: Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet 2005;76:780–793.
107 Yu K, Xu J, Rao DC, Province M: Using tree-based recursive partitioning methods to group haplotypes for increased power in association studies. Ann Hum Genet 2005;69:577–589.
108 Beckmann L, Fischer C, Obreiter M, Rabes M, Chang-Claude J: Haplotype-sharing analysis using Mantel statistics for combined genetic effects. BMC Genet 2005;6:S70.
109 Dudbridge F, Koeleman BP, Todd JA, Clayton DG: Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. Am J Hum Genet 2000;66:2009–2012.
110 Nolte IM, de Vires AR, Spijker GT, Jansen RC, Brinja D, Zelikovshy A, te Meerman GJ: Whole genome association analysis by haplotype sharing length based methods. BMC Genomics 2007, in press.
111 Allen AS, Satten GA, Tsiatis AA: Locally-efficient robust estimation of haplotype-disease association in family-based studies. Biometrika 2005;92:559–571.
112 Allen AS, Satten GA: Inference on haplotype/disease association using parent-affected-child data: the projection conditional on parental haplotypes method. Genet Epidemiol 2007;31:211–223.
113 Allen AS, Satten GA: Statistical models for haplotype sharing in case-parent trio data. Hum Hered 2007;64:35–44.
114 Glaser RL, Ramsay JP, Morison IM: The imprinted gene and parent-of-origin effect database now includes parental origin of de novo mutations. Nucleic Acids Res 2006;34:D29–D31.
115 Wilcox AJ, Weinberg CR, Lie RT: Distinguishing the effects of maternal and offspring genes through studies of 'case-parent triads'. Am J Epidemiol 1998;148:893–901.
116 Weinberg CR: Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. Am J Hum Genet 1999;65:229–235.
117 Zhou JY, Hu YQ, Fung WK: A simple method for detection of imprinting effects based on case-parents trios. Heredity 2007;98:85–91.
118 Hu YQ, Zhou JY, Sun F, Fung WK: The transmission disequilibrium test and imprinting effects test based on case-parent pairs. Genet Epidemiol 2007;31:273–287.
119 Hu YQ, Zhou JY, Fung WK: An extension of the transmission disequilibrium test incorporating imprinting. Genetics 2007;175:1489–1504.
120 Schaid DJ: Likelihoods and TDT for the case-parents design. Genet Epidemiol 1999;16:250–260.
121 Umbach DM, Weinberg CR: The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet 2000;66:251–261.
122 Eaves LJ, Sullivan P: Genotype-environment interaction in transmission disequilibrium tests. Adv Genet 2001;42:223–240.
123 Lunetta KL, Faraone SV, Biederman J, Laird NM: Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. Am J Hum Genet 2000;66:605–614.
124 Amos CI, Elston RC, Bonney GE, Keats BJ, Berenson GS: A multivariate method for detecting genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype. Am J Hum Genet 1990;47:247–254.
125 Elston RC: Genetic analysis of multivariate traits. Epilepsy Res Suppl 1991;4:161–171.
126 Amos CI, Laing AE: A comparison of univariate and multivariate tests for genetic linkage. Genet Epidemiol 1993;10:671–676.
127 Schork NJ: Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. Am J Hum Genet 1993;53:1306–1319.
128 Markel PD, Corley RP: A multivariate analysis of repeated measures: linkage of the albinism gene (Tyr) to a QTL influencing ethanol-induced anesthesia in laboratory mice. Psychiatr Genet 1994;4:205–210.
129 Jiang C, Zeng ZB: Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 1995;140:1111–1127.
130 Korol AB, Ronin YI, Kirzhner VM: Interval mapping of quantitative trait loci employing correlated trait complexes. Genetics 1995;140:1137–1147.
131 Boomsma DI: Using multivariate genetic modeling to detect pleiotropic quantitative trait loci. Behav Genet 1996;26:161–166.
132 Allison DB, Neale MC: Joint tests of linkage and association for quantitative traits. Theor Popul Biol 2001;60:239–251.
133 Allison DB, Thiel B, St Jean P, Elston RC, Infante MC, Schork NJ: Multiple phenotype modeling in gene-mapping studies of quantitative traits: Power advantages. Am J Hum Genet 1998;63:1190–1201.
134 Blangero J, Williams-Blangero S, Mahaney MC: Multivariate genetic analysis of apo AI concentration and HDL subfractions: Evidence for major locus pleiotropy. Genet Epidemiol 1993;10:617–622.
135 Bennett S, Curnow RN: Consanguinity and the transmission/disequilibrium test for allelic association. Genet Epidemiol 2001;21:68–77.
136 Genin E, Todorov AA, Clerget-Darpoux F: Properties of the transmission-disequilibrium test in the presence of inbreeding. Genet Epidemiol 2002;22:116–127.
137 George V, Laud PW: A Bayesian approach to the transmission/disequilibrium test for binary traits. Genet Epidemiol 2002;22:41–51.

138 Nagelkerke NJ, Hoebee B, Teunis P, Kimman TG: Combining the transmission disequilibrium test and case-control methadology using generalized logistic regression. Eur J Hum Genet 2004;12:964–970.
139 Epstein MP, Veal CD, Trembath RC, Barker JNWN, Li C, Satten GA: Genetic association analysis using data from triads and unrelated subjects. Am J Hum Genet 2005;76: 592–608.
140 Risch N, Merikangas K: The future of genetic studies of complex human diseases. Science 1996;273:1516–1523.
141 Risch NJ: Searching for genetic determinants in the new millennium. Nature 2000; 405:847–856.
142 Rothman K, Greenland S: Modern Epidemiology. Philadelphia, Lippincott Williams & Wilkins, 1998.
143 Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang RH, Messer CJ, Chew A, Han JH, Duan JC, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu CB, Kalbfleisch TS, Shaner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Ruano G, Vovis GF: Haplotype variation and linkage disequilibrium in 313 human genes. Science 2001;293:489–493.
144 Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P: A haplotype map of the human genome. Nature 2005;437:1299–1320.
145 Redden DT, Allison DB: The effect of assortative mating upon genetic association studies: Spurious associations and population substructure in the absence of admixture. Behav Genet 2006;36:678–686.
146 Devlin B, Roeder K: Genomic control for association studies. Biometrics 1999;55: 997–1004.
147 Salari K, Choudhry S, Tang H, Naqvi M, Lind D, Avila PC, Coyle NE, Ung N, Nazario S, Casal J, Torres-Palacios A, Clark S, Phong A, Gomez I, Matallana H, Perez-Stable EJ, Shriver MD, Kwok PY, Sheppard D, Rodriguez-Cintron W, Risch NJ, Burchard EG, Ziv E: Genetic admixture and asthma-related phenotypes in Mexican American and Puerto Rican asthmatics. Genet Epidemiol 2005;29:76–86.
148 Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, Tsai HJ, Naqvi M, Phong A, Ung N, Matallana H, Avila PC, Casal J, Torres A, Nazario S, Castro R, Battle NC, Perez-Stable EJ, Kwok PY, Sheppard D, Shriver MD, Rodriguez-Cintron W, Risch N, Ziv E, Burchard EG: Population stratification confounds genetic association studies among Latinos. Hum Genet 2006;118: 652–664.
149 Hanis CL, Chakraborty R, Ferrell RE, Schull WJ: Individual admixture estimates – disease associations and individual risk of diabetes and gallbladder-disease among Mexican-Americans in Starr County, Texas. Am J Phys Anthropol 1986;70: 433–441.
150 Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, Pfaff C, Jones C, Massac A, Cameron N, Baron A, Jackson T, Argyropoulos G, Jin L, Hoggart CJ, McKeigue PM, Kittles RA: Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 2003;112:387–399.
151 Campbell C, Ogburn E, Lunetta K, Lyon H, Freedman ML, Groop L, Altshuler D, Ardlie K, Hirschhorn JN: Demonstrating stratification in a European American population. Nat Genet 2005;37:868–872.
152 Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, Silva G, Belmont JW, Klareskog L, Gregersen PK: European population substructure: clustering of northern and southern populations. PLoS Genet 2006;2: e143.
153 Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD: Measuring European population stratification with microarray genotype data. Am J Hum Genet 2007;80:948–956.
154 Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K: An Icelandic example of the impact of population structure on association studies. Nat Genet 2005;37: 90–95.
155 Wen B, Xie XH, Gao S, Li H, Shi H, Song XF, Qian TZ, Xiao CJ, Jin JZ, Su B, Lu D, Chakraborty R, Jin L: Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet 2004; 74:856–865.
156 Chakraborty R, Weiss KM: Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci USA 1988;85: 9119–9123.
157 Chakraborty R, Smouse PE: Recombination of haplotypes leads to biased estimates of admixture proportions in human-populations. Proc Natl Acad Sci USA 1988;85: 3071–3074.
158 Wacholder S, Rothman N, Caporaso N: Population stratification in epidemiologic studies of common genetic variants and cancer: Quantification of bias. J Natl Cancer Inst 2000;92:1151–1158.
159 Wacholder S, Rothman N, Caporaso N: Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. Cancer Epidemiol Biomarkers Prev 2002; 11:513–520.
160 Deng HW: Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. Genetics 2001;159:1319–1323.
161 Dean M: Approaches to identify genes for complex human diseases: Lessons from Mendelian disorders. Hum Mutat 2003;22: 261–274.
162 Burnett MS, Strain KJ, Lesnick TG, de Andrade M, Rocca WA, Maraganore DM: Reliability of Self-reported Ancestry among Siblings: Implications for Genetic Association Studies. Am J Epidemiol 2006;163: 486–492.
163 Thomas DC, Witte JS: Point: Population stratification: A problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prevention 2002;11:505–512.
164 Pritchard JK, Donnelly P: Case-control studies of association in structured or admixed populations. Theor Popul Biol 2001; 60:227–237.
165 Devlin B, Roeder K, Wasserman L: Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 2001; 60:155–166.
166 Bacanu SA, Devlin B, Roeder K: The power of genomic control. Am J Hum Genet 2000; 66:1933–1944.
167 Zheng G, Freidlin B, Li Z, Gastwirth JL: Genomic control for association studies under various genetic models. Biometrics 2005; 61:186–192.
168 Chen HS, Zhu X, Zhao H, Zhang S: Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. Ann Hum Genet 2003;67:250–264.
169 Falush D, Stephens M, Pritchard JK: Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 2003;164: 1567–1587.
170 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–909.
171 Rife D: Populations of hybrid origin as source material for the detection of linkage. Am J Hum Genet 1954;6:26–33.
172 Tang H, Peng J, Wang P, Risch NJ: Estimation of individual admixture: Analytical and study design considerations. Genet Epidemiol 2005;28:289–301.
173 Stephens JC, Briscoe D, Obrien SJ: Mapping by admixture linkage disequilibrium in human-populations – limits and guidelines. Am J Hum Genet 1994;55:809–824.
174 McKeigue PM: Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admired populations. Am J Hum Genet 1997;60:188–196.

175 McKeigue PM: Multipoint admixture mapping. Genet Epidemiol 2000;19:464–465.
176 McKeigue PM: Mapping genes that underlie ethnic differences in disease risk: Methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am J Hum Genet 1998;63: 241–251.
177 McKeigue PM, Carpenter JR, Parra EJ, Shriver MD: Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. Ann Hum Genet 2000;64:171–186.
178 Clarke G, Whittemore AS: Comparison of admixture and association mapping in admixed families. Genet Epidemiol 2007;31: 763–775.
179 Bonilla C, Parra EJ, Pfaff CL, Dios S, Marshall JA, Hamman RF, Ferrell RE, Hoggart CL, McKeigue PM, Shriver MD: Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping. Ann Hum Genet 2004; 68:139–153.
180 Wu B, Liu N, Zhao H: PSMIX: an R package for population structure inference via maximum likelihood method. BMC Bioinformatics 2006;7:317.
181 Tzelgov J, Stern I: Relationships between variables in three variable linear regression and the concept of supressor. Educ Psychol Measurement 1978;38:325–335.
182 Zhang XS, Hill WG: Predictions of patterns of response to artificial selection in lines derived from natural populations. Genetics 2005;169:411–425.
183 Fisher R: Statistical methods for research workers. Edinburgh, Oliver and Boyd, 1925.
184 Mendel G: Versuche über Pflanzen-Hybriden. Verh Naturforsch Ver Brünn 1866; 4:3–47.
185 Beasley TM, Yang DY, Yi NJ, Bullard DC, Travis EL, Amos CI, Xu SZ, Allison DB: Joint tests for quantitative trait loci in experimental crosses. Genet Select Evol 2004; 36:601–619.
186 Rabinowitz D: Adjusting for population heterogeneity: A framework for characterizing statistical information and developing efficient test statistics. Genet Epidemiol 2003;24:284–290.
187 Rabinowitz D: Unbiased discordant sibpair tests when parental genotypes are missing. Am J Med Genet 2001;105:57–59.
188 Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM: Family-based tests for associating haplotypes with general phenotype data: Application to asthma genetics. Genet Epidemiol 2004;26:61–69.
189 Rosenbaum P, Rubin D: The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70: 41–55.
190 Rosenbaum P, Rubin D: Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 1985;39:33–38.
191 Devlin B, Roeder K: Genomic control for association studies. Am J Hum Genet 1999; 65:A83–A83.
192 Horvath S, Windemuth C, Knapp M: The disequilibrium maximum-likelihood-binomial test does not replace the transmission/disequilibrium test. Am J Hum Genet 2000;67:531–534.