Published in final edited form as:
Biometrics. 2012 September ; 68(3): 766–773. doi:10.1111/j.1541-0420.2011.01713.x.

Rapid Testing of SNPs and Gene–Environment Interactions in Case–Parent Trio Data Based on Exact Analytic Parameter Estimation

Holger Schwender[1,2,*], **Margaret A. Taub**[2], **Terri H. Beaty**[3], **Mary L. Marazita**[4], and **Ingo Ruczinski**[2]

[1]Faculty of Statistics, TU Dortmund University, 44221 Dortmund, Germany

[2]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A

[3]Department of Epidemiology, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A

[4]School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15219, U.S.A

Summary

Case–parent trio studies concerned with children affected by a disease and their parents aim to detect single nucleotide polymorphisms (SNPs) showing a preferential transmission of alleles from the parents to their affected offspring. A popular statistical test for detecting such SNPs associated with disease in this study design is the genotypic transmission/disequilibrium test (gTDT) based on a conditional logistic regression model, which usually needs to be fitted by an iterative procedure. In this article, we derive exact closed-form solutions for the parameter estimates of the conditional logistic regression models when testing for an additive, a dominant, or a recessive effect of a SNP, and show that such analytic parameter estimates also exist when considering gene–environment interactions with binary environmental variables. Because the genetic model underlying the association between a SNP and a disease is typically unknown, it might further be beneficial to use the maximum over the gTDT statistics for the possible effects of a SNP as test statistic. We therefore propose a procedure enabling a fast computation of the test statistic and the permutation-based p-value of this MAX gTDT. All these methods are applied to whole-genome scans of the case–parent trios from the International Cleft Consortium. These applications show our procedures dramatically reduce the required computing time compared to the conventional iterative methods allowing, for example, the analysis of hundreds of thousands of SNPs in a few minutes instead of several hours.

Keywords

Conditional logistic regression; Family-based design; Genome-wide association studies; Genotypic transmission/disequilibrium test; International Cleft Consortium; MAX test

*holger.schwender@udo.edu.

8. Supplementary Materials

Web Appendices referenced in Sections 1–6 are available under the Paper Information link at the *Biometrics* website http://www.biometrics.tibs.org/.

1. Introduction

As an alternative to population-based case–control studies considering unrelated individuals, family-based designs are frequently employed to test for association between genetic markers such as single nucleotide polymorphisms (SNPs) and a disease. While family-based studies might be more expensive than comparably powered population-based case–control studies, they have a key advantage in being robust against spurious findings caused by population stratification. In addition, they also enable the simultaneous assessment of association and linkage (Spielman and Ewens, 1996; Gauderman, Witte, and Thomas, 1999; Laird and Lange, 2006).

The transmission/disequilibrium test (TDT) proposed by Spielman, McGinnis, and Ewens (1993) is one of the most popular tests for association in the analysis of case–parent trio data, an often utilized family-based design in which children affected by a disease and their parents are considered. This test (which is usually referred to as the allelic TDT) aims to detect departure from Mendelian expectations, i.e., alleles of an observed marker preferentially transmitted from the parents to the affected offspring reflecting linkage and association between this marker and an unobserved causal gene. It is equivalent to McNemar's test (McNemar, 1947) applied to a 2×2 table summarizing the transmitted and nontransmitted alleles from heterozygous parents.

A prominent alternative procedure to this allelic TDT is the genotypic transmission/disequilibrium test (gTDT), in which both the genotype of the offspring and the Mendelian genotype realizations not transmitted from the parents are considered to test for association, assuming a specific genetic mode of inheritance (e.g., an additive, dominant, or recessive mode). At each marker, one of four possible pairs of parental alleles is transmitted to the affected offspring, and the other three unobserved genotype realizations are used as artificial controls (usually referred to as pseudo-controls). The resulting matching structure can be accounted for using a conditional likelihood (Self et al., 1991; Schaid, 1996). This gTDT can be more powerful than the allelic TDT (Schaid, 1999b). It considers individuals instead of chromosomes as units of the analysis, and enables the direct assessment of the relative risks. Thus, in contrast to the allelic TDT, the gTDT yields parameter estimates, standard errors, and confidence intervals, in addition to p-values. Further, the gTDT can also be used to model specific risk relationships, whereas for the allelic TDT multiplicative effects have to be assumed (Fallin et al., 2002). Another advantage of the gTDT over the allelic TDT is that it allows testing for gene–environment interactions, which makes this test particularly attractive for case–parent trio studies, in which the detection of interaction effects of SNPs with any of several environmental factors may also be of great interest.

The allelic TDT can be applied to hundreds of thousands of SNPs in several minutes, as its test statistic is based on the off-diagonal elements of the 2×2 table summarizing the transmitted and nontransmitted alleles from heterozygous parents. By contrast, genome-wide computations of test statistics such as the gTDT statistic based on the parameters in (conditional) logistic regression models are much more time-consuming, as these parameters usually need to be estimated numerically using an iterative procedure, because in general no closed-form solutions for the parameter estimates exist. One exception to this occurs in an ordinary logistic regression setting with a single binary predictor (e.g., when testing for a recessive or a dominant effect in a population-based case–control study). In this situation, the slope parameter estimate is simply the log-odds ratio, and the conventionally used iteratively reweighted least squares procedure can thus be avoided.

Such an analytic solution does not exist for a conditional logistic regression with a 1:3 matching, i.e., a matching of one case with three (pseudo-)controls, when arbitrary

predictors are considered. However, we show in Section 2 how Mendel's laws impose a structure on the possible genotypes in the child, which allows derivation of closed-form solutions for the parameter estimates when testing for dominant or recessive effects, as well as additive effects.

Such analytic estimates also exist for parameters in conditional logistic regression models when analyzing interactions between SNPs and binary environmental factors, which are the by far most common type of nongenetic variables additionally considered in genetic association studies (see Section 3).

Since the underlying genetic mode of inheritance is usually unknown, it might be beneficial to not just consider one of the three genetic effects of a SNP, as an incorrectly chosen model can lead to a substantial reduction in statistical power (Freidlin et al., 2002). This issue can be addressed by employing the maximum over the test statistics for an additive, a dominant, and a recessive effect of a SNP as test statistic. Thus, for example, Zheng, Freidlin, and Gastwirth (2002) propose such a MAX test based on a TDT-type approach suggested by Schaid and Sommer (1993), while Yan, Zheng, and Li (2008) consider a group sequential MAX test based on the score test statistics of Li, Gastwirth, and Gail (2005) for nuclear families with both affected and unaffected offsprings.

In Section 4, we describe how the MAX test can be adapted for the gTDT. Because the null distribution of the MAX gTDT statistic is unknown and neither the iterative procedure conventionally used for a gTDT nor the closed-form solutions derived in Section 2 would allow in genome-wide association studies a recomputation of the parameter estimates for a sufficiently large number of permutations in any reasonable amount of time, we also propose a procedure to enable a genome-wide determination of permutation-based p-values for this MAX gTDT. This approach essentially builds upon the ideas behind the analytic parameter estimates and the specific permutation scheme used for case–parent trio data.

Afterward, we briefly discuss in Section 5 that—with one exception—no closed-form solutions for the gTDT exist when testing gene–gene interactions. A more comprehensive discussion of gTDTs for testing two-way interactions is presented in Web Appendix C.

All these procedures are applied in Section 6 to the whole-genome SNP data from the case–parent trio study conducted by the International Cleft Consortium (Beaty et al., 2010) to compare the proposed procedures with both the conventional approach to test SNPs with a gTDT and related score tests.

2. Closed-Form Estimates for the Genotypic TDT

In a gTDT for testing whether a particular SNP is associated with disease, the genotypes of the pseudo-controls are derived for each of n case–parent trios as the possible pairs of parental alleles not transmitted to the affected child. A conditional logistic regression is then applied to these strata, each consisting of genotypes of the affected offspring and the three matched pseudo-controls, using the SNP as predictor and the case–pseudocontrol status as response. Denoting genotypes at the SNP for the case ($k = 0$) and the matched pseudo-controls ($k = 1, 2, 3$) in trio $i = 1, \ldots, n$, by x_{ik}, the genotype is either given by the number $m_{ik} \in \{0, 1, 2\}$ of minor alleles present in the respective case or pseudocontrol when testing under an additive model, by $x_{ik} = I(m_{ik} > 0)$ when testing a dominant effect, or by $x_{ik} = I(m_{ik} = 2)$ when testing a recessive effect. The parameter β of this conditional logistic regression model can be estimated by maximizing the likelihood

$$L(\beta)=\prod_{i=1}^{n}\left(\frac{\exp(\beta x_{i0})}{\sum_{k=0}^{3}\exp(\beta x_{ik})}\right) \quad (1)$$

(see Breslow et al., 1978). To test for association, $\hat{\beta}$ is squared and divided by its estimated variance to give the gTDT statistic, which under the null hypothesis is asymptotically χ^2-distributed with one degree of freedom.

Typically, the likelihood (1) needs to be maximized numerically in an iterative procedure. However, the complexity of this maximization problem can be substantially reduced in case–parent trio studies by noticing only 10 possible genotype combinations exist in the 1:3 matching of cases and pseudo-controls (see Table 1) and all case–parent trios with the same combination contribute equally to the maximization of the likelihood (1).

If we test for an additive effect, the corresponding log-likelihood can thus be written as

$$\ell(\beta_{\text{add}})=\log(L(\beta_{\text{add}}))=\sum_{j=1}^{10}a_j\log(w_j(\beta_{\text{add}})), \quad (2)$$

where the numbers a_j, $j = 1, \ldots, 10$, of trios with the same combination of genotypes for the case and the three matched pseudo-controls, and the corresponding weights $w_j(\beta_{\text{add}})$ are summarized in Table 1. Setting the first derivative

$$\frac{\partial\ell(\beta_{\text{add}})}{\partial\beta_{\text{add}}}=(a_2+a_4)-\frac{\exp(\beta_{\text{add}})}{1+\exp(\beta_{\text{add}})}\times\sum_{j=1}^{4}a_j+(a_6+2a_7)-\frac{2\exp(\beta_{\text{add}})}{1+\exp(\beta_{\text{add}})}\times\sum_{j=5}^{7}a_j.$$

of the log-likelihood (2) to zero and solving it for β_{add} yields a closed-form solution for the maximum-likelihood estimate, namely,

$$\widehat{\beta}_{\text{add}}=\text{logit}\left(\frac{a_2+a_4+a_6+2a_7}{a_1+a_2+a_3+a_4+2a_5+2a_6+2a_7}\right) \quad (3)$$

This estimate is thus given by the logit of the ratio of the weighted number $c_{\max}$ of children showing at least as many minor alleles as their parents to the total number p_{het} of heterozygous parents. Hence, the analytic parameter estimate (3) can be computed without specifying pseudo-controls. Using these notations, the variance of (3) can be estimated by plugging $\hat{\beta}_{\text{add}} = \text{logit}(c_{\max}/p_{\text{het}})$ into the negative inverse of the second derivative

$$\frac{\partial^2\ell(\beta_{\text{add}})}{\partial\beta_{\text{add}}^2}=-\frac{\exp(\beta_{\text{add}})}{(1+\exp(\beta_{\text{add}}))^2}p_{\text{het}},$$

leading to the variance estimate

$$\widehat{\text{Var}}(\widehat{\beta}_{\text{add}})=\frac{p_{\text{het}}}{(p_{\text{het}}-c_{\max})c_{\max}}.$$

Thus, the gTDT statistic for testing an additive effect of a SNP is given by

$$g_{\text{add}}=\widehat{\beta}^2_{\text{add}}/\widehat{\text{Var}}(\widehat{\beta}_{\text{add}}).$$

In a similar way, closed-form solutions for parameter estimates in the conditional logistic regression models testing for either a dominant or a recessive effect can be derived. In both models, six case–pseudocontrol combinations exist in case–parent trios, of which only four influence the maximization of the likelihood (see Tables 2 and 3).

With the numbers of trios and weights for the dominant model presented in Table 2, the first derivative of the log-likelihood of this model is given by

$$\frac{\partial \ell(\beta_{\text{dom}})}{\partial \beta_{\text{dom}}}=d_2-(d_2+d_1)\times\frac{\exp(\beta_{\text{dom}})}{\exp(\beta_{\text{dom}})+1}+d_4-(d_4+d_3)\times\frac{\exp(\beta_{\text{dom}})}{\exp(\beta_{\text{dom}})+1/3}.$$

leading to the maximum-likelihood estimate

$$\widehat{\beta}_{\text{dom}}=\log\left(\sqrt{\frac{d_2+d_4}{3(d_1+d_3)}+h^2_{\text{dom}}}-h_{\text{dom}}\right),$$

where

$$h_{\text{dom}}=\frac{1/3d_1-d_2+d_3-1/3d_4}{2(d_1+d_3)},$$

(see Web Appendix A). Analogously, the parameter estimate in a recessive model can be derived as

$$\widehat{\beta}_{\text{rec}}=\log\left(\sqrt{\frac{3(r_2+r_4)}{r_1+r_3}+h^2_{\text{rec}}}-h_{\text{rec}}\right),$$

with

$$h_{\text{rec}}=\frac{3r_1-r_2+r_3-3r_4}{2(r_1+r_3)}.$$

3. Testing Gene–Environment Interactions

Besides testing individual effects of SNPs, it is often of great interest to also test for gene–environment interactions. For this purpose, a conditional logistic regression model

$$\beta_G X + \beta_{GE}(X \times E), \quad (4)$$

can be fitted, where X codes for the considered effect of a SNP (as defined in Section 2), and E is an environmental variable. Because conditioning is based on the family status, the value of the environmental variable is identical for the affected offspring and the three matched pseudo-controls (cf. Maestri et al., 1997; Schaid, 1999a). Thus, the main effect of the environmental variable is not identifiable from the conditional likelihood, and hence, not included in model (4). However, in genetic association studies the effect of an environmental variable is typically only of interest in conjunction with the genotypes, i.e., as gene–environment interaction.

Many of the usually examined environmental variables, for example, gender or maternal smoking, are binary, and can thus be coded with values zero and one. In this situation, it is possible to derive estimates for the parameters β_G and β_{GE} in model (4) based on the parameter estimate from the main effect model discussed in Section 2 that considers the same genetic effect of the SNP. Denoting the estimate of this main effect model by $\hat{\beta}^{(0)}$, when this conditional logistic regression model is fitted only using the case–parent trios with children not exposed to the environmental variable ($e = 0$), and by $\hat{\beta}^{(1)}$, when only the trios with exposed offsprings ($e = 1$) are analyzed, then the maximum-likelihood estimate of β_G and β_{GE} can be determined by

$$\widehat{\beta}_G = \widehat{\beta}^{(0)} \quad \text{and} \quad \widehat{\beta}_{GE} = \widehat{\beta}^{(1)} - \widehat{\beta}^{(0)}.$$

For a derivation of these estimates, see Web Appendix B. Consequently, the gTDT statistic for testing a gene–environment interaction is given by

$$g_{GE} = \frac{\widehat{\beta}^2_{GE}}{\widehat{\mathrm{Var}}(\widehat{\beta}^{(1)}) + \widehat{\mathrm{Var}}(\widehat{\beta}^{(0)})}. \quad (5)$$

Besides this traditional case–parent trio design, there also exist other family-based study designs in which a conditional logistic regression can be used to test gene–environment interactions. Chatterjee et al. (2005), for example, propose a conditional logistic regression model $\beta_G X + \beta_E E + \beta_{GE}(X \times E)$ for a discordant sib-design in which a case is matched with a sibling (or a cousin) not affected by the disease and two pseudo-controls, where they assume genetic susceptibility and environmental exposure are independently distributed within each family. Since this model also employs a 1:3 matching and in particular because Chatterjee et al. (2005) show that this design can lead to an increased power compared to the traditional case–parent trio design, we have investigated whether also closed-form solutions for the estimates of the parameters β_G, β_E, and in particular, β_{GE} exist. As we discuss in detail in Web Appendix D, such analytic solutions cannot be derived, even for a binary environmental variable E, mainly because of the weights in the corresponding log-likelihood that are much more complex than the ones for the case–parent trio design.

4. Permutation-Based p-Values and the MAX gTDT

Instead of considering an additive, a dominant, and a recessive model separately, the maximum g_{max} over the three gTDT statistics g_{add}, g_{dom}, and g_{rec} for these models can be used as test statistic when the underlying genetic model is unknown. The p-value of this MAX gTDT can be computed by a procedure in which the case–pseudocontrol status is

permuted B times. Because this requires accounting for the 1:3 matching, all trios need to be considered separately, and in each permutation the case-status in each trio is randomly assigned to one of the four Mendelian realizations possible given the parents' genotypes.

However, the structure of possible case–pseudocontrol combinations allows us to avoid separate permutations for each trio, and the repeated recomputation of the numbers of trios showing different combinations. If, for example, we consider the $a_1 + a_2$ trios in which one of the parents has no variant allele and the other parent is heterozygous (see Table 1), then under the null hypothesis of no association, the probability that the offspring inherited one variant allele is 0.5. Thus, instead of permuting the case–pseudocontrol status within each of these $a_1 + a_2$ trios separately, we can simply obtain the numbers a_{1b} and a_{2b} of trios showing the corresponding case–pseudocontrol combination in permutation $b = 1, \ldots, B$ by drawing B values a_{2b} from a binomial distribution with $a_1 + a_2$ observations and success probability $p = 0.5$, and by setting $a_{1b} = a_1 + a_2 - a_{2b}$. The same approach applies to the $a_3 + a_4$ trios in which one parent carries one variant allele, and the other parent is homozygous for the variant allele, yielding a_{3b} and a_{4b}. For the $a_5 + a_6 + a_7$ trios in which both parents are heterozygous, permuted numbers can be determined by B random draws from a multinomial distribution with probabilities $p(a_5) = p(a_7) = 0.25$ and $p(a_6) = 0.5$.

Using these numbers, the B permuted values of the parameter estimate (3) and the corresponding gTDT statistic $g^{(b)}_{\text{add}}$ for testing an additive effect can be determined, and only the numerator in the logit term of the estimate (3) needs to be computed in each permutation, as the denominator p_{het} is constant throughout all permutations. The numbers $a_{j\,b}$ can then also be employed to calculate B values of both the gTDT statistic $g^{(b)}_{\text{dom}}$ and $g^{(b)}_{\text{rec}}$ for testing a dominant and a recessive effect of a SNP, respectively (see Tables 2 and 3, respectively) so that the empirical p-value of the MAX gTDT for this SNP can be determined based on the maximum $g^{(b)}_{\text{max}}$ over the values $g^{(b)}_{\text{add}}$, $g^{(b)}_{\text{dom}}$, and $g^{(b)}_{\text{rec}}$, $b = 1, \ldots, B$.

Analogously, the MAX gTDT statistic for testing a gene–environmental interaction can be computed as the maximum over the gTDT statistic (5) determined under each of the three genetic modes of inheritance. Empirical p-values can then be obtained by considering the trios with children not exposed to the binary environmental variable, and trios with exposed children separately when randomly drawing the numbers from the corresponding binomial and multinomial distributions. An extension to gTDTs for testing gene–gene interactions discussed in Section 5 is also possible (see Web Appendix C.3).

5. Testing Gene–Gene Interactions

An analytic solution, which reduces the computing time substantially, would be even more desirable when testing gene–gene interactions than for testing individual SNPs or gene–environment interactions, as there are $m(m-1)/2$ two-way interactions when analyzing m SNPs. Such interactions might be tested either with a simple conditional logistic regression model containing only one term representing the interaction, with a model also including two parameters for the main effects of the two SNPs, or most sophisticatedly with a likelihood ratio test such as the one proposed by Cordell (2002) for epistatic interactions comparing the maximized log-likelihoods of a model composed of factors for the two main effects and a model additionally comprising interaction terms.

As we discuss in detail in Web Appendix C, there do not exist analytic solutions for the parameter estimates in any of these models—with one exception: When considering a conditional logistic regression model $\beta_{\text{GG}} \times (X_s\, X_t)$, where X_s and X_t code for recessive effects of SNPs s and t, then the first derivative of the log-likelihood can be rewritten as a

polynomial of fourth degree, and the roots of this polynomial can be analytically determined by a standard procedure (see Web Appendix G). Because this approach and the closed-form solutions are rather complex, they are not presented here, but in Web Appendix C.2.

Because all other models for testing gene–gene interactions are even (much) more complex, there do not exist analytic solutions for the parameter estimates in these models (for details, see Web Appendix C). However, as we will show in Section 6, rewriting the log-likelihoods analogously to (2) and numerically maximizing these log-likelihoods can also lead to an immense gain in computing time compared to the conventional approach of determining the 15 pseudo-controls per trio (there exist 4×4 pairs of genotype realizations, given the parents' genotypes at the two loci) and fitting a conditional logistic regression model with a 1:15 matching based on the likelihood (1).

6. Application to a Genome-Wide Association Study

To investigate the computational gains in performing a gTDT achieved by the closed-form estimates derived in Sections 2, 3, and 5, as well as the numerical solutions based on the reformulated log-likelihood (2) for testing gene–gene interactions (see Section 5 and Web Appendix C) compared to the conventional approach to the gTDT (i.e., setting up the pseudo-controls for each trio, and then fitting a standard conditional logistic regression model), we reanalyzed the case–parent trio data provided by the International Cleft Consortium (Beaty et al., 2010). DNA samples of children with an oral cleft (cleft lip, cleft lip and palate, or cleft palate) and their parents were genotyped at the Center for Inherited Disease Research using the Illumina Human610-Quadv1_B Beadchip, and processed using the Illumina BeadStudio Genotyping Module. Extensive quality control was carried out, excluding samples and SNPs of low quality (see Beaty et al., 2010, for details).

In our comparison, we considered 569,187 autosomal markers for 1925 case–parent trios from this study in the applications of the gTDT to single SNPs and gene–environment interactions, and a data set consisting of 1000 SNPs randomly drawn from the 569,187 markers when testing all 499,500 two-way interactions comprised by this data set. Both the analytic and the conventional approach to the gTDT were implemented in R (R Development Core Team, 2010) and are available in the R package trio at http://cran.r-project.org. All computations were performed on 2.7 GHz machines.

Additionally, we performed score tests related to the gTDTs. Score tests are often used to screen for interesting SNPs, gene–environment or gene–gene interactions, as they are usually less computationally expensive than Wald tests. These score tests were implemented using our reformulated log-likelihood (2). Implementations based on the logarithm of the original likelihood (1) have an about seven to nine times larger computing time—mainly due to the required generation of pseudo-controls—and are therefore not considered in our comparison.

In our comparison, we focus our interest on the computing times summarized in Table 4. More general results of the analysis of the data from the International Cleft Consortium can be found in Web Appendix E and in Beaty et al. (2010).

In any of our applications of the gTDTs, the analytic estimates and the conventional iterative procedure yielded virtually identical results. When, for example, testing SNPs under an additive model, the maximum difference between the two estimates for β_{add} for any SNP was less than 2×10^{-6}, and the two values of the gTDT statistics differed at most by 8×10^{-7}.

While the median computing time over 10 applications of the iterative fitting procedure to all 569,187 SNPs was 12.03 CPU hours when testing for an additive effect, the gTDT approach based on the closed-form estimates took on average a total of just 8.74 minutes, and was thus 83 times faster than the conventional gTDT procedure. The gain in computing time was even larger when testing dominant and recessive effects. More precisely, testing all SNPs for a dominant or a recessive effect took on average in both cases about 11.87 hours when using the iterative fitting procedure, and about 7.90 minutes, when employing the analytic approach, leading to an about 90 times faster computation (see Table 4).

To further investigate which situations yielded the largest gain in computing time, we applied both approaches to the gTDT to several randomly drawn subsets of these genome-wide data, consisting of between 100 and 20,000 SNPs and between 500 and 1925 trios. In these comparisons, the computing time of the gTDT based on the analytic maximum-likelihood estimates was always 75 times, usually 90 to 100 times, and frequently more than 100 times lower than the time required by the iterative approach for an application to the same data set, with a slight decrease in relative computational gain as the number of trios increased. For details on this simulation study and its results, see Web Appendix F.

For the application of the gTDT to gene–environment interactions, gender—which might be considered as a surrogate for several unmeasured exogenous and endogenous risk factors—was employed as environmental variable. While the conventional procedure required more than 14 hours of CPU time for genome-wide tests of each of the three genetic models, analyzing these gene–environment interactions with the analytic gTDT took on average 9.1 minutes when considering an additive effect and 7.6 minutes under an dominant or recessive model, resulting in a 95 (additive), 114 (dominant), or 117 (recessive) times faster computation (see Table 4).

While under a recessive or dominant model the genome-wide application of the score test was a little faster than applying the analytic gTDT to individual SNPs (in the case of an additive model, the computing times were virtually identical), its application was slightly slower when testing gene–environment interactions. Only in the analysis of gene–gene interactions, the score test showed a moderate gain in computing time compared to the gTDT. Its application was 1.50 – 1.65 times faster when compared with the numerical solution for the gTDT based on the log-likelihood (2), whereas this gain decreased to a factor of 1.18 when the analytic solution of the gTDT for a recessive × recessive model was used.

Employing the numerical solutions for a gTDT for testing gene–gene interactions based on a log-likelihood of the form (2) reduced the computing time of the conventional approach based on the likelihood (1) by a factor of 107 (additive × additive model), 251 (dominant × dominant), or 265 (recessive × recessive). Compared to this reduction, further gain in computing time achieved by using the analytic solution for the recessive × recessive model was (with a factor of 1.4) only relatively small, as computing the different numbers of trios requires most of the computing time and the procedure for determining the maximum-likelihood estimate $\hat{\beta}_{GG}$ analytically is quite complex (see Web Appendices C.2 and G).

Finally, we also applied a MAX gTDT to all autosomal SNPs. Using the analytic parameter estimates, calculation of all 569,187 MAX gTDT statistics required on average 11.5 minutes, which was about 187 times faster than computation with the conventional procedure. This additional gain in computing time is because in the latter approach the three genetic models must be fitted separately, while in the analytic procedure the numbers a_j of trios are computed once and the three gTDT statistics are then determined based on them.

Because one determination of the MAX gTDT statistic for all SNPs in this case–parent trio study took 11.5 minutes, the standard method for computing permutation-based p-values, i.e., shuffing the response B times and recomputing the test statistic for each permutation, would require about 115,000 minutes, i.e., about 80 days, when considering just 10,000 permutations. By contrast, employing the algorithm described in Section 4 took about 185 minutes for 10,000 permutations (additionally to the 11.5 minutes for computing the observed MAX gTDT statistics and thus the numbers a_j), reducing the computing time by a factor of 620 compared to the standard permutation method.

Because actually much more than 10,000 permutations should be used in genome-wide association studies, an approach similar to the one of Sladek et al. (2007) might be employed to further accelerate computation and enable the usage of millions of permutations. Sladek et al. (2007) first determined empirical p-values for the MAX statistic of Freidlin et al. (2002) for all SNPs in a genome-wide population-based association study based on 10,000 permutations, and then computed p-values for the 57 most significant SNPs from the first analysis based on 10 million permutations to obtain better estimates of the empirical p-values.

Here, we followed a similar, but less strict strategy. We started with computing empirical p-values for all 569,187 SNPs based on 10,000 permutations. In the original analysis, Beaty et al. (2010) called a SNP genome-wide significant if its empirical p-value was smaller than 5×10^{-8}. When considering 10^8 permutations, this would mean that a SNP would only be called genome-wide significant, if five or less of the $g_{\max}^{(b)}$ values were larger than or equal to $g_{\max}$. Because we were also interested in getting improved estimates for the p-values of SNPs that show large $g_{\max}$ scores, but are not necessarily genome-wide significant, we chose a less strict cutoff and considered in a second step all SNPs for which 50 or less of the 10,000 $g_{\max}^{(b)}$ values were larger than or equal to the corresponding $g_{\max}$ value. For these 3832 SNPs, empirical p-values based on one million permutations were determined, which took 1.5 hours. In a final step, which was by far the most time-consuming step (requiring about 21 hours), 100 million permutations were used to compute p-values for the 357 SNPs for which 50 or less of the one million $g_{\max}^{(b)}$ values were larger than or equal to the corresponding $g_{\max}$ value in the second step, and to detect the 185 genome-wide significant SNPs among these 357 markers by correcting these p-values for multiple comparisons with a Bonferroni adjustment, i.e., by multiplying these p-values with 569,187, and identifying all SNPs with an adjusted p-value smaller than or equal to 0.05.

7. Discussion

A major goal of SNP association studies in case–parent trio design is to detect alleles preferentially transmitted from parents to an affected offspring. Associations between SNPs and disease are often assessed with the gTDT based on a conditional logistic regression model, which is applied to data from the affected children and their three matched pseudo-controls.

In this article, we have derived closed-form solutions for the parameter estimate in this conditional logistic regression model when testing for an additive, dominant, or recessive effect. These analytic estimates avoid the iterative numerical optimization conventionally used when fitting conditional logistic regression models, and thus, lead to a huge reduction in the required computing times typically seen in genome-wide association studies.

In an application to the case–parent trio data from the International Cleft Consortium (Beaty et al., 2010), we have shown a genome-wide analysis can be carried out in a few minutes, instead of several hours. Using these closed-form estimates, it would thus also be possible to

test the millions of SNPs comprised by the latest types of SNP microarrays in less than an hour, and all SNPs measured in whole-genome sequencing studies in a few hours.

These closed-form solutions become even more beneficial when interactions between SNPs and any of several environmental variables should be analyzed in a genome-wide association study, as they can be adapted to the estimation of the parameters in a conditional logistic regression model testing for interactions of SNPs with binary environmental variables or other binary covariates.

Such analytic estimates would also be of great interest when testing gene–gene interactions. With the exception of the simplest model including only one parameter representing the interaction effect of two SNPs both under a recessive mode of inheritance, closed-form solutions for the gTDT do not exist when testing gene–gene interactions. However, our reformulation (2) of the conditional log-likelihood also allows in these situations a large reduction in computing time compared to the conventional approach.

We have also compared computing times of the gTDTs and related score tests, which—because of their comparably low computational costs—are often used in genome-wide association studies to screen for interesting SNPs. Because our solutions to the gTDT are only a bit more complex to determine than the corresponding score test statistics, both approaches have about the same computing time. Thus, for about the same computational costs, the gTDT not only provides scores and p-values, but—in contrast to the score test—also estimates of genotypic relative risks, standard errors, and confidence intervals. The resulting parameter estimates and their standard errors can be used in ranking procedures such as approximate Bayes factors (Wakefield, 2009) or Bayesian optimal percentiles (Louis and Ruczinski, 2010) to provide a better assessment and ranking of SNPs in association studies than p-values, as these approaches—in contrast to p-values—also take the statistical power into account.

We have also introduced a procedure for computing permutation-based p-values for a MAX test based on the gTDT statistics, which is robust against the underlying genetic mode of inheritance. Our approach makes it possible to calculate genome-wide empirical p-values based on tens of thousands of permutations in a few hours, as compared to months. Along with a simple strategy for selecting potentially interesting SNPs, this procedure was applied with up to 100 million permutations to the case–parent trio data of the International Cleft Consortium in about a day on a single processor. This computing time might be further reduced by choosing more strict selection criteria or parallelizing the computation.

All gTDT procedures (as well as the corresponding score tests) are implemented in the R package trio version 1.2.6 or later, available at http://cran.r-project.org.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Thomas A. Louis for fruitful discussions, and Marvin Newhouse for computational support. We sincerely thank all of the families who participated in the study of the International Cleft Consortium, and we gratefully acknowledge the invaluable assistance of clinical, field, and laboratory staff who contributed to this study. We particularly acknowledge Jeffrey C. Murray (JCM) and Alan F. Scott (AFS) who both played critical roles in bringing together the International Cleft Consortium and in carrying out this genome-wide association study. We also gratefully acknowledge the financial support provided by the Deutsche Forschungsgemeinschaft (SCHW 1508/1-1, SCHW 1508/2-1, and Research Training Group 1032 Statistical Modelling to HS), the National Institute of Health (R01 DK061662 and HL090577 to IR), and a CTSA grant to the Johns Hopkins Medical Institutions.

Funding to support data collection, genotyping, and analysis for the case–parent trio study of the International Cleft Consortium came from several sources, some to individual investigators and some to a consortium supporting the genome-wide study itself. The consortium for GWAS genotyping and analysis was supported by the National Institute for Dental and Craniofacial Research through U01-DE-004425; International Consortium to Identify Genes & Interactions Controlling Oral Clefts, 2007–2009; TH Beaty, PI. Funding for individual investigators include: R01-DE-01458 (THB, AFS), U01-DE-018993 (THB, MLM, IR), R37-DE08559 (JCM, MLM), R01-DE016148 (MLM), P50-DE016215 (JCM, MLM), and R21-DE016930 (MLM). Part of the original recruitment of Norwegian case–parent trios was supported by the Intramural Research Program of the National Institute of Health, National Institute of Environmental Health Sciences.

References

Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, Liang KY, Wu T, Murray T, Fallin MD, Redett RA, Raymond G, Schwender H, Jin SC, Cooper ME, Dunnwald M, Mansilla MA, Leslie E, Bullard S, Lidral AC, Moreno LM, Menezes R, Vieira AR, Petrin A, Wilcox AJ, Lie RT, Jabs EW, Wu-Chou YH, Chen PK, Wang H, Ye X, Huang S, Yeow V, Chong SS, Jee SH, Shi B, Christensen K, Melbye M, Doheny KF, Pugh EW, Ling H, Castilla EE, Czeizel AE, Ma L, Field LL, Brody L, Pangilinan F, Mills JL, Molloy AM, Kirke PN, Scott JM, Arcos-Burgos M, Scott AF. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. Nature Genetics. 2010; 42:525–529. [PubMed: 20436469]

Breslow NE, Day NE, Halvorsen KT, Prentice RL, Sabai C. Estimation of multiple relative risk functions in matched case-control studies. American Journal of Epidemiology. 1978; 108:299–307. [PubMed: 727199]

Chatterjee N, Kalaylioglu Z, Carroll RJ. Exploiting gene-environment independence in family based control studies: Increased power for detecting associations, interactions and joint effects. Genetic Epidemiology. 2005; 28:138–156. [PubMed: 15593088]

Cordell HJ. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics. 2002; 11:2463–2468. [PubMed: 12351582]

Fallin D, Beaty T, Liang KY, Chen W. Power comparisons for genotypic vs. allelic TDT methods with = 2 alleles. Genetic Epidemiology. 2002; 23:458–461. [PubMed: 12432512]

Freidlin B, Zheng G, Li Z, Gastwirth J. Trend tests for case-control studies of genetic markers: Power, sample size and robustness. Human Heredity. 2002; 53:146–152. [PubMed: 12145550]

Gauderman WJ, Witte JS, Thomas DC. Family-based association studies. Journal of the National Cancer Institutes Monograph. 1999; 26:31–37.

Laird NM, Lange C. Family-based designs in the age of large-scale gene-association studies. Nature Reviews Genetics. 2006; 7:385–394.

Li Z, Gastwirth JL, Gail MH. Power and related statistical properties of conditional likelihood score tests for association studies in nuclear families with parental genotypes. Annals of Human Genetics. 2005; 69:296–314. [PubMed: 15845034]

Louis TA, Ruczinski I. Efficient evaluation of ranking procedures when the number of units is large, with application to SNP identification. Biometrical Journal. 2010; 52:34–49. [PubMed: 20131327]

Maestri NE, Beaty TH, Hetmanski J, Smith EA, McIntosh I, Wyszynski DF, Liang KY, Duffy DL, VanderKolk C. Application of transmission disequilibrium tests to non-syndromic oral clefts: Including candidate genes and environmental exposures in the models. American Journal of Medical Genetics. 1997; 73:337–344. [PubMed: 9415696]

McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947; 12:153–157. [PubMed: 20254758]

R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.

Schaid DJ. General score tests for associations of genetic markers with disease using cases and their parents. Genetic Epidemiology. 1996; 13:423–449. [PubMed: 8905391]

Schaid DJ. Case-parents design for gene-environment interaction. Genetic Epidemiology. 1999a; 16:261–273. [PubMed: 10096689]

Schaid DJ. Likelihoods and TDT for the case-parents design. Genetic Epidemiology. 1999b; 16:250–260. [PubMed: 10096688]

Schaid DJ, Sommer SS. Genotype relative risks: Methods for design and analysis of candidate-gene association studies. American Journal of Human Genetics. 1993; 53:1114–1126. [PubMed: 8213835]

Self SG, Longton G, Kopecky KJ, Liang KY. On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics. 1991; 47:53–61. [PubMed: 2049513]

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007; 445:881–885. [PubMed: 17293876]

Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. American Journal of Human Genetics. 1996; 59:983–989. [PubMed: 8900224]

Spielman RS, McGinnis R, Ewens WJ. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). American Journal of Human Genetics. 1993; 52:506–516. [PubMed: 8447318]

Wakefield J. Bayes factors for genome-wide association studies: Comparison with p-values. Genetic Epidemiology. 2009; 33:79–86. [PubMed: 18642345]

Yan LK, Zheng G, Li Z. Two-stage group sequential robust tests in family-based association studies: Controlling type I error. Annals of Human Genetics. 2008; 72:557–565. [PubMed: 18325081]

Zheng G, Freidlin B, Gastwirth JL. Robust TDT-type candidate-gene association tests. Annals of Human Genetics. 2002; 66:145–155. [PubMed: 12174218]

Table 1

The 10 possible combinations of genotypes for an affected child and the three matched pseudo-controls given the parents' genotype, their contribution $w_j(\beta_{add})$ to the likelihood of a conditional logistic regression for testing an additive effect, and the symbols for the numbers of trios exhibiting these genotypes. The genotypes are coded by the numbers of minor alleles. The genotype combinations below the dashed line do not contribute to the likelihood, and thus, do not influence the parameter estimation

Parents	Offspring	Pseudo-controls	Weight $w_j(\beta_{add})$ in likelihood	Number a_j of trios
0, 1	0	0, 1, 1	$\frac{1}{2+2\exp(\beta_{add})}$	a_1
0, 1	1	0, 0, 1	$\frac{\exp(\beta_{add})}{2+2\exp(\beta_{add})}$	a_2
1, 2	1	1, 2, 2	$\frac{\exp(\beta_{add})}{2\exp(\beta_{add})+2\exp(2\beta_{add})}=\frac{1}{2+2\exp(\beta_{add})}$	a_3
1, 2	2	1, 1, 2	$\frac{\exp(2\beta_{add})}{2\exp(\beta_{add})+2\exp(2\beta_{add})}=\frac{\exp(\beta_{add})}{2+2\exp(\beta_{add})}$	a_4
1, 1	0	1, 1, 2	$\frac{1}{1+2\exp(\beta_{add})+\exp(2\beta_{add})}=\frac{1}{(1+\exp(\beta_{add}))^2}$	a_5
1, 1	1	0, 1, 2	$\frac{\exp(\beta_{add})}{1+2\exp(\beta_{add})+\exp(2\beta_{add})}=\frac{\exp(\beta_{add})}{(1+\exp(\beta_{add}))^2}$	a_6
1, 1	2	0, 1, 1	$\frac{\exp(2\beta_{add})}{1+2\exp(\beta_{add})+\exp(2\beta_{add})}=\frac{\exp(2\beta_{add})}{(1+\exp(\beta_{add}))^2}$	a_7
0, 2	1	1, 1, 1	$\frac{\exp(\beta_{add})}{4\exp(\beta_{add})}=\frac{1}{4}$	a_8
2, 2	2	2, 2, 2	$\frac{\exp(2\beta_{add})}{4\exp(2\beta_{add})}=\frac{1}{4}$	a_9
0, 0	0	0, 0, 0	$\frac{1}{4}$	a_{10}

Table 2

The six possible combinations of genotypes that an affected children and the three matched pseudo-controls can show when testing for a dominant effect, their weight in the conditional likelihood, and the symbols representing the numbers of trios having these genotypes, where the symbols in the parenthesis refer to the numbers from Table 1

Affected child	Pseudo-controls	Weight w_j (β_{dom}) in likelihood	Number d_j of trios
0	0, 1, 1	$\frac{1}{2+2\exp(\beta_{dom})}$	d_1 (= a_1)
1	0, 0, 1	$\frac{\exp(\beta_{dom})}{2+2\exp(\beta_{dom})}$	d_2 (= a_2)
0	1, 1, 1	$\frac{1}{1+3\exp(\beta_{dom})}$	d_3 (= a_5)
1	0, 1, 1	$\frac{\exp(\beta_{dom})}{1+3\exp(\beta_{dom})}$	d_4 (= a_6 + a_7)
0	0, 0, 0	0.25	d_5 (= a_{10})
1	1, 1, 1	0.25	d_6 (other a_j)

Table 3

The six possible combinations of genotypes that an affected children and the three matched pseudo-controls can show when testing for a recessive effect, their weight in the conditional likelihood, and the symbols representing the numbers of trios having these genotypes, where the symbols in the parenthesis refer to the numbers from Table 1

Affected child	Pseudo-controls	Weight w_j (β_{rec}) in likelihood	Number r_j of trios
0	0, 1, 1	$\frac{1}{2+2\exp(\beta_{rec})}$	r_1 (= a_3)
1	0, 0, 1	$\frac{\exp(\beta_{rec})}{2+2\exp(\beta_{rec})}$	r_2 (= a_4)
0	0, 0, 1	$\frac{1}{3+\exp(\beta_{rec})}$	r_3 (= a_5 + a_6)
1	0, 0, 0	$\frac{\exp(\beta_{rec})}{3+\exp(\beta_{rec})}$	r_4 (= a_7)
1	1, 1, 1	0.25	r_5 (= a_9)
0	0, 0, 0	0.25	r_6 (other a_j)

Table 4

Average computing times (in seconds) over 10 (in the case of gene–gene interactions, 5) applications of the gTDT based on both analytic (or numerical) parameter estimates using log-likelihoods of the form (2) and the conventional iterative fitting procedure as well as score tests based on (2) to the 569,187 autosomal SNPs from the case–parent trio study of the International Cleft Consortium (Beaty et al., 2010), interactions of each of these SNPs with gender, and all 499,500 interactions between two of 1000 randomly selected SNPs (considering a model consisting of one parameter for the interaction term)

	Individual SNPs			
	Additive	**Dominant**	**Recessive**	**MAX**
Analytic	525	474	472	690
Score	525	437	435	698
Conventional	43,318	42,740	42,726	128,784
	SNP–Gender Interaction			
	Additive	Dominant	Recessive	
Analytic	545	455	454	
Score	556	465	462	
Conventional	51,600	51,834	53,237	
	SNP–SNP Interaction			
	Additive	Dominant	Recessive	
Analytic	–	–	335	
Numerical	1,182	497	469	
Score	785	324	283	
Conventional	126,748	124,753	124,425	