BMC Proceedings

BioMed Central

Proceedings

Open Access

Combining multiple family-based association studies

Hua Tang*[1], Jie Peng†[2], Pei Wang†[3], Marc Coram and Li Hsu[3]

Address: [1]Department of Genetics, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, California 94305, USA, [2]Department of Statistics, University of California, Davis, California 95616, USA and [3]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA

Email: Hua Tang* - huatang@fhcrc.org; Jie Peng - jie@wald.ucdavis.edu; Pei Wang - pwang@fhcrc.org; Marc Coram - mcoram@stanford.edu; Li Hsu - lih@fhcrc.org

* Corresponding author †Equal contributors

from Genetic Analysis Workshop 15
St. Pete Beach, Florida, USA. 11–15 November 2006

Published: 18 December 2007

BMC Proceedings 2007, **1**(Suppl 1):S162

This article is available from: http://www.biomedcentral.com/1753-6561/1/S1/S162

Abstract

While high-throughput genotyping technologies are becoming readily available, the merit of using these technologies to perform genome-wide association studies has not been established. One major concern is that for studies of complex diseases and traits, the whole-genome approach requires such large sample sizes that both recruitment and genotyping pose considerable challenge. Here we propose a novel statistical method that boosts the effective sample size by combining data obtained from several studies. Specifically, we consider a situation in which various studies have genotyped non-overlapping subjects at largely non-overlapping sets of markers. Our approach, which exploits the local linkage disequilibrium structure without assuming an explicit population model, opens up the possibility of improving statistical power by incorporating existing data into future association studies.

Background

The past few years have seen a rapid development in high-throughput genotyping platforms, making genome-wide association studies technologically feasible. Yet the merit of using these technologies to perform genome-wide association studies has not been established [1]. Insufficient sample size is a particular concern in genome-wide association studies for several reasons. First, the genotyping cost, even though decreasing, nonetheless puts a pressure to reduce the number of individuals assayed. Second, the sample size required to declare statistical significance is increased as a result of the large number of hypotheses tested. Third, a majority of these studies investigate complex diseases, in which each disease variant confers a moderate risk. Thus, assuming an odds ratio of 1.5 and an allele frequency of 0.2, more than 1000 cases and controls are required for a statistical power of 80% and a significance level of $p < 10^{-6}$ [2].

Several approaches are being developed to improve the efficiency of genome-wide association studies. One attractive solution is a multi-stage design, in which only a subset of individuals are genotyped at the full set of markers in the initial phase. In subsequent stages, additional individuals are genotyped at increasingly restricted marker sets [3]. In the context of family-based association testing, methods have also been developed that use the same data for genomic screening and replication [4]. Both of these methods focus on reducing the number of hypotheses tested.

As an alternative, we note that association studies of some diseases, such as rheumatoid arthritis (RA), are being performed by more than one group of investigators, giving rise to multiple data sets [5,6]. One major challenge is that subjects from different studies, or within a study over the study period, may be genotyped by different platforms; therefore, different sets of single-nucleotide polymorphisms (SNPs) are assayed. In linkage analysis, it is relatively straightforward to combine families genotyped at different marker panels: a LOD score at an unobserved marker may be evaluated using neighboring observed markers through multipoint interval mapping techniques [7]; LOD scores at corresponding genome locations can then be added across families. In contrast, to our knowledge, a multipoint interval mapping approach has not been developed in the context of linkage disequilibrium (LD) association studies. This is because without a priori knowledge of LD patterns between markers, it is difficult to infer LD between disease and an unobserved marker based on neighboring observed markers. Furthermore, haplotypes constructed on each study cannot be easily combined. On the other hand, there is evidence that strong LD exists among tightly linked markers, and therefore we expect the test statistics at these markers to be correlated [8].

In this paper, we propose a novel approach that allows us to increase the effective sample size by combining data obtained from several studies. Specifically, we consider a situation in which both the subjects and marker panels are non-overlapping among the studies. In this situation, it is not possible to evaluate LD between markers from different panels. Our approach, which exploits the local LD structure without assuming an explicit population model, opens up the possibility of improving statistical power by incorporating existing data into future association studies. We illustrate our method by analyzing the simulated RA data. We had no knowledge of the "answers" at the time of analysis.

Methods

We have previously described a multipoint transmission-disequilibrium test (TDT) method that is based on local smoothing [8]. Our study demonstrated that a) TDT statistics at tightly linked markers are correlated, and b) when tightly linked markers are genotyped, the smoothed TDT statistics can achieve a greater statistical power compared with the non-smoothed version. These findings suggest that TDT statistics can be combined, even though the different studies have genotyped non-overlapping set of markers.

Combining data sets with non-overlapping markers and individuals

In this section, we outline our statistical methodology in a simple setting: two studies have genotyped non-overlapping sets of markers on independent sets of individuals in a common genomic region, and both studies have used a case-parents trio design. In each study, the TDT statistics, TDT^A or TDT^B, can be computed at the genotyped markers [9].

To motivate our test statistic, we first consider a marker that has been genotyped in both studies. In Study A, let b_1 denote the number of informative transmissions, in which **A** alleles are transmitted but **a** alleles are not transmitted, and let c_1 denote the converse (i.e., **a** alleles but not **A** alleles are transmitted). Likewise, let b_2 and c_2 denote the corresponding numbers of informative transmissions in Study B. With complete genotype data, we would compute the TDT by pooling the data:

$$\begin{aligned} T_{pool} &= \frac{(b_1+b_2-c_1-c_2)^2}{b_1+b_2+c_1+c_2} \\ &= \frac{(b_1-c_1)^2}{b_1+c_1}\frac{b_1+c_1}{b_1+b_2+c_1+c_2} + \frac{(b_2-c_2)^2}{b_2+c_2}\frac{b_2+c_2}{b_1+b_2+c_1+c_2} + \frac{2(b_1-c_1)(b_2-c_2)}{b_1+b_2+c_1+c_2} \\ &= \frac{n_1}{n_1+n_2}\mathrm{TDT}^A + \frac{n_2}{n_1+n_2}\mathrm{TDT}^B + \frac{2(b_1-c_1)(b_2-c_2)}{b_1+b_2+c_1+c_2}. \end{aligned}$$

We next show that, under the null hypothesis, the last term in Eq. (1) has an expectation of 0. Let $R_1 = B_1 + C_1 > 0$ (the capital letters denote the random variables), and $R_2 = B_2 + C_2 > 0$. Under the null hypothesis, $L(B_1 \mid R_1) \sim \mathrm{Binom}(R_1, 0.5)$, $L(B_2 \mid R_2) \sim \mathrm{Binom}(R_2, 0.5)$, and B_1 and B_2 are independent. We then have:

$$\begin{aligned} E\left(\frac{(B_1-C_1)(B_2-C_2)}{B_1+B_2+C_1+C_2}\right) &= E\left[E\left(\frac{(B_1-C_1)(B_2-C_2)}{B_1+B_2+C_1+C_2}\right) \mid R_1, R_2\right] \\ &= E\left[\frac{1}{R_1+R_2}E(2B_1-R_1 \mid R_1)E(2B_2-R_2 \mid R_2)\right] = 0. \end{aligned}$$

Therefore, under the null hypothesis, the pooled TDT statistic is nearly a weighted average of the corresponding TDT statistics in the respective studies. The weights are proportional to the number of informative parents ($n_i = b_i + c_i$). Assuming that the two studies sampled comparable populations (e.g., allele frequencies are similar at all loci), we approximate these weights by the number of trios.

For a marker that is not genotyped in one study (but is genotyped in the other), we try to impute the TDT statistic using neighboring markers. We then add the observed and imputed TDT scores from the two studies. Suppose M markers have been genotyped by either Study A or Study

B. Denote the physical locations of these markers by $\{t_1,...,t_M\}$. Let $V^A = [v_1^A,\ldots,v_M^A]$ be a vector indicating whether marker *m* is genotyped in Study A. Denote the TDT statistics using each of the two samples as TDT^A and TDT^B, respectively. Let $f^A(t)$ be the results of applying a local linear regression fitting to (t_i, TDT^A) data. At each marker, we compute:

$$T^A(t_i) = v_i^A \mathrm{TDT}^A(t_i) + (1 - v_i^A) f^A(t_i).$$

In other words, if a marker is genotyped in Study A (v^A = 1), we simply take the TDT statistic; if a marker t_i is not gentoyped in Study A, we impute its expected TDT statistic using the predicted value. Similarly, we compute $T^B(t)$ using data from Study B. The combined test statistic at each marker is simply:

$$\mathrm{TDT}_{\mathrm{comb}}(t_i) = \frac{n_1}{n_1+n_2} T^A(t_i) + \frac{n_2}{n_1+n_2} T^B(t_i).$$

We implement the imputation step using the loess function in R. The choice of the smoothing parameter depends on many factors such as the age of the disease mutation, the population under study, and the marker density. While an optimal window size is difficult to define, an examination of inter-marker LD guides our choice: we seek a region within which the genotyped markers are in high LD. Roughly speaking, we are faced with a trade-off between bias and variance: smoothing over a wide region tends to reduce variance of the imputed statistics at the cost of an increased bias. Therefore, an alternative to loess with pre-specified bandwidth is a smoothing spline with the degree of freedom chosen by cross-validation. To properly account for the imputation, and to correct for multiple comparison, we perform a simulation-based test: conditioning on the parents' genotype, we generate the transmitted and the non-transmitted haplotypes under the null hypothesis, re-impute the TDT statistics, and compute $\mathrm{TDT}_{\mathrm{comb}}$ on the simulated data. The observed $\max_i \mathrm{TDT}_{\mathrm{comb}}(t_i)$ is compared with the null distribution of the corresponding maxima in the simulated data.

Data set example

To illustrate our proposed method, we analyze Replicate 1 of the simulated RA data. This data set consists of 1500 nuclear families, each of which has both parents and two affected children genotyped. It is known that there is a strong effect of DR type at the HLA locus on chromosome 6. A simple TDT analysis using all 1500 families unambiguously demonstrates preferential transmission of DR-2 or DR-3 alleles to the affected individuals. However, is the DR allele the sole variant affecting the disease in the region? To address this question, we examine the transmission from parents who are homozygous 1/1 at the DR locus. If the DR locus explains the entire association in the region, conditioning on parents being 1/1, there should not be preferential transmission at any markers nearby. Among 1500 mothers, 70 have genotype 1/1. Our analyses highlight a practical difficulty: performing stratified analysis on a subset of samples further reduces the sample size; thus, stratified analyses are particularly likely to suffer from small sample size even when the main study has good power.

On chromosome 6, we restrict ourselves to the 293 SNP markers falling within 1.5×10^6 bp around the DR locus. We consider a situation in which each third of the families are genotyped on a different platform. The 293 SNPs are randomly divided into three sets, and there is no overlap in the three sets of markers or individuals. Because the risk of RA is much higher among women, we hypothesize that there may be gene × sex interaction. Furthermore, there has been ambiguous evidence regarding maternally transmitted risk elements [10]. Therefore, we looked at four types of transmission: father to son, father to daughter, mother to son, and mother to daughter. Because the phase is known for all the affected children, the four types of transmission can be examined independently. For each type of transmission, we perform a TDT analysis on each of the three subsets of families. Because the diagnosis of RA is often ambiguous, we hypothesized that the more severe cases are more likely to carry the genetic risk factor. Therefore, a severity measure, on the scale of 1 to 5, is used as a relative weight. We then combine the three sets of TDT scores to compute $\mathrm{TDT}_{\mathrm{comb}}$, with a bandwidth approximately 15 markers.

Results

The results of various TDT test for the mother-daughter transmission in the 3-cM region are shown in Figure 1. Figure 1a displays the TDT scores using each subset of one-third of the families. Because the transmissions from a mother to two daughters are independent, a meaningful measure of sample size is the mother-daughter pairs. In our data, the numbers of mother-daughter transmissions in the three subsets are 46, 39, and 23, respectively. For all TDT tests, we use 10,000 permutations to establish the null distribution and significance level. The *p*-values of TDT on the three subsets are 0.0011, 0.02, and 0.61, respectively. In Figure 1b, the points represent the TDT scores from subset A (families 1–500) and the solid line represents the loess prediction, $\mathrm{TDT}_{\mathrm{comb}}(t_i) = \frac{n_1}{n_1+n_2} T^A(t_i) + \frac{n_2}{n_1+n_2} T^B(t_i).$. Figure 1c compares the TDT scores when all markers are genotyped in every individual ($\mathrm{TDT}_{\mathrm{all}}$, open square) versus $\mathrm{TDT}_{\mathrm{comb}}$

(filled points). While the maximum value achieved by TDT_{comb} appears substantially lower than the corresponding value by TDT_{all}, the same is true under the null hypothesis, because the imputed TDT statistics tend to be smoother than observed ones. As a result, at a specific significance level (say, 0.99), the critical value for TDT_{all} is 17.38, while the corresponding critical value for TDT_{comb} is 6.87. We perform a quantile transformation based on the null distribution, and Figure 1d compares TDT_{all} with transformed TDT_{comb}. It indicates that, upon suitable transformation, TDT_{comb} can achieve similar significance level as if TDT_{all}. However, the location of the peak shifts slightly: while the marker with highest TDT_{all} lies to the right of the DR locus, that with the highest TDT_{comb} lies to the left of DR locus. Another consequence of smoothing and imputing TDT scores is that the "peak" of TDT_{comb} appears somewhat narrower than TDT_{all}. In a similar fashion, we analyzed the other three types of transmission. The results, summarized in Table 1, suggest the existence of another variant that influences the disease risk. Interestingly, transmission is distorted in mother-daughter and father-daughter transmission, but not transmissions to sons. This suggests possible gene × sex interaction. Finally, while we set out to use severity as a relative weight for each individual, retrospective comparison indicates that the weight makes little difference.

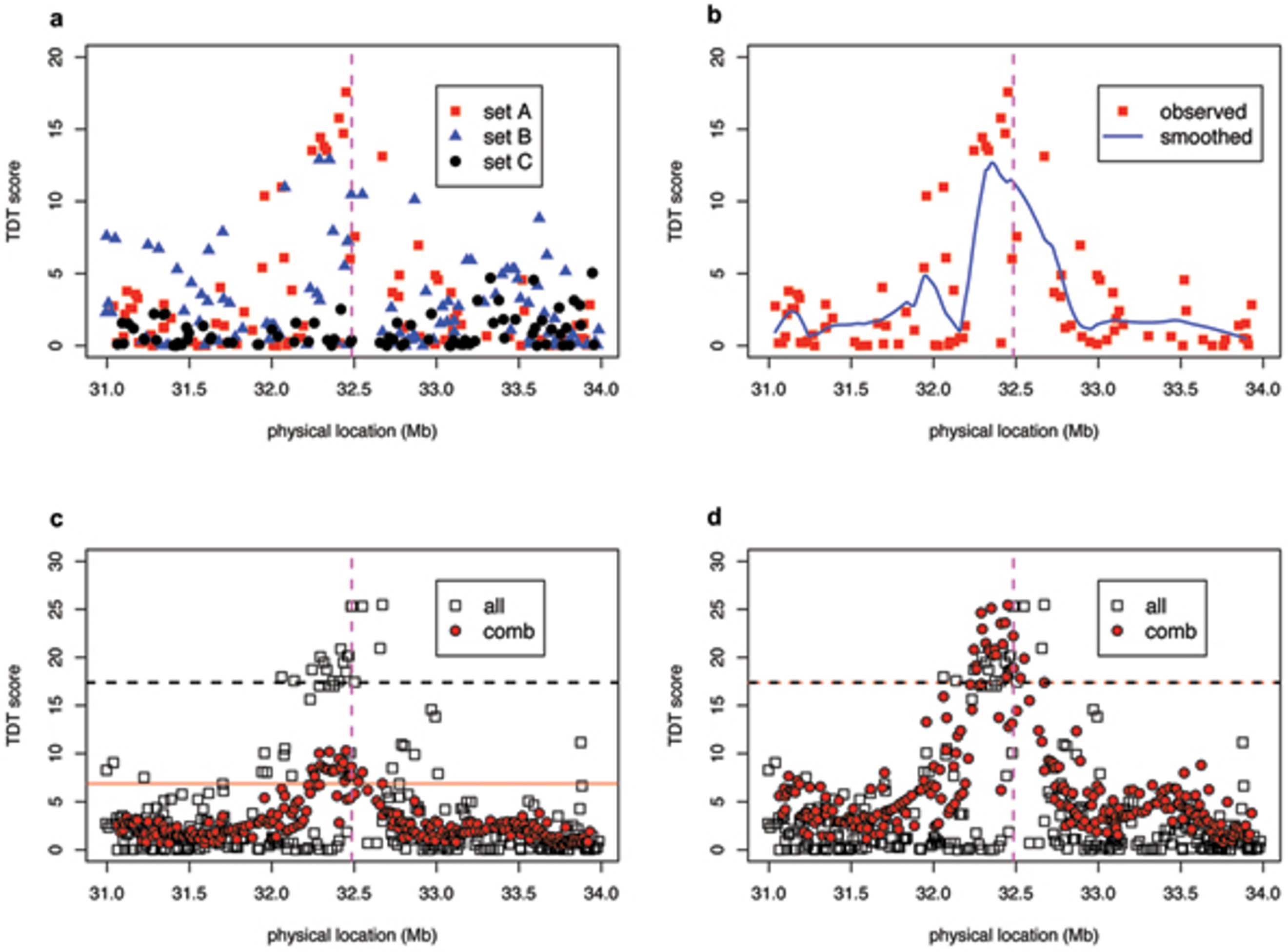

Figure 1

TDT tests for mother-daughter transmissions, restricted to mothers with DR genotype 1/1. a, TDT for three subsets separately. Dotted vertical line indicates the location of DR locus. b, TDT scores using sub-sample A (square) versus imputed scores (line). c, TDT scores assuming all markers are genotyped in each individual (TDT_{all}), open square) versus TDT_{comb}. Dotted line indicates 0.01 critical value for TDT_{all}, and solid line represents the corresponding critical value for TDT_{comb}. d, Comparison of TDT_{all} and TDT_{comb} after a quantile transformation.

Table 1: Results of TDT tests for sex- and parent-specific transmissions

Transmission	TDT^A	TDT^B	TDT^C	TDT_{all}	TDT_{comb}
Mother-daughter	0.0011	0.019	0.60	**$<10^{-4}$**	**$<10^{-4}$**
Father-daughter	6×10^{-4}	**$<10^{-4}$**[a]	0.0023	**$<10^{-4}$**	**$<10^{-4}$**
Mother-son	0.056	0.241	0.77	0.159	0.0896
Father-son	0.22	0.6634	0.4440	0.68	0.62

[a] Bold indicates statistical significance at $p = 0.05$ level.

Discussion

We have performed additional simulations to assess the gain in power of our combined test (not shown). In a wide range of disease models and parameters, the combined test achieves a greater power than the probability that at least one of the two studies achieved statistical significance, even though the latter procedure has an inflated type I error. Furthermore, the gain in power of our approach depends on marker density. In the current setting, the entire SNP set on chromosome 6 has a density comparable to a genome-wide set of 300 K, and hence each of the three subsets we analyzed has a density comparable to the 100 K SNP array. When each subset of individuals has been genotyped on a denser marker set, the imputed TDT scores tend to be more accurate. Thus, had subsets of individuals been genotyped on different platforms of 500 K SNP arrays, we would expect the imputed TDT scores to be more accurate, and therefore combining across studies will achieve even greater power. On the other hand, compared to the pooled TDT with complete marker data, the combined test incurs a loss of power. This loss of power is to be expected for two reasons. First, the smoothing process introduces uncertainties both under the null and alternative hypotheses. Second, if a locus increases the disease risk in both studies and the high risk alleles are the same allele, the last term in Eq. (1) has an expectation greater than 0; hence, $E(TDT_{pool}) > E(TDT_{comb})$.

We have developed and evaluated our approach in the context of a family-based study using the TDT design. The approach can be generalized to case-control design when participants in all samples represent a relatively homogeneous population and all studies use the exact same phenotype definition. Combining samples in the presence of population stratification requires extensions to our method, including modification on the weight and the smoothing parameters. For example, if LD among markers decays slower in one population than in the other, it maybe desirable to use a wider smoothing parameter in the former population. These issues should be examined more thoroughly in the future. On the other hand, as long as the cases and controls are matched within each study, the combined test offers greater protection against population stratification than a test on the pooled genotypes. Finally, if different genetic factors underlie the etiology of each study, combing these studies will not improve power. Therefore, an important issue to address in the future is how to decide whether different studies can be combined.

Conclusion

Genome-wide association studies often suffer from insufficient sample size. The problem is exacerbated by the need to restrict the analysis to a subset of individuals. Our approach, which exploits the local LD structure without assuming an explicit population model, opens up the possibility of improving statistical power by incorporating existing data into future association studies.

Competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

This work was supported by NIH GM073059, CA125489, and CA053996.

This article has been published as part of *BMC Proceedings* Volume 1 Supplement 1, 2007: Genetic Analysis Workshop 15: Gene Expression Analysis and Approaches to Detecting Multiple Functional Loci. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/1?issue=S1.

References

1. Hirschhorn JN, Daly MJ: **Genome-wide association studies for common diseases and complex traits.** *Nat Rev Genet* 2005, **6:**95-108.
2. Wang WY, Barratt BJ, Clayton DG, Todd JA: **Genome-wide association studies: theoretical and practical concerns.** *Nat Rev Genet* 2005, **6:**109-118.
3. van den Oord EJ, Sullivan PF: **False discoveries and models for gene discovery.** *Trends Genet* 2003, **19:**537-542.
4. Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, Demeo DL, Murphy A, Su J, Datta S, Rosenow C, Christman M, Silverman EK, Laird NM, Weiss ST, Lange C: **Genomic screening and replication using the same data set in family-based association testing.** *Nat Genet* 2005, **37:**683-691.
5. Etzel CJ, Chen WV, Shepard N, Jawaheer D, Cornelis F, Seldin MF, Gregersen PK, Amos CI: **Genome-wide meta-analysis for rheumatoid arthritis.** *Hum Genet* 2006, **119:**634-641.
6. Radstake TR, Barrera P, Albers MJ, Swinkels HL, van de Putte LB, van Riel PL, European Consortium on Rheumatoid Arthritis Families: **Genetic anticipation in rheumatoid arthritis in Europe.** *J Rheumatol* 2001, **28:**962-967.
7. Fulker DW, Cherny SS, Cardon LR: **Multipoint interval mapping of quantitative trait loci, using sib pairs.** *Am J Hum Genet* 1995, **56:**1224-1233.

8. Hsu L, Yu X, Houwing-Duistermaat JJ, Uh HW, El Galta R, Lebrec JJ, Tang H: **Locally weighted transmission/disequilibrium test for genetic association analysis.** *BMC Genet* 2005, **6(Suppl 1):**S60.
9. 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.
10. Deighton C, Criswell LA, Lum RF, Silman A: **Ages of onset suggestive of genetic anticipation in rheumatoid arthritis multicase sibships can be explained by observational bias.** *Rheumatology* 2007, **46:**120-123.