BMC Proceedings

BioMed Central

Proceedings

Open Access

An integrated genome-wide association analysis on rheumatoid arthritis data

Jun Zhang*[1], Xiaofeng Zhu[2] and Richard S Cooper[3]

Address: [1]Department of Statistics, University of Chicago, 5734 South University Avenue, Chicago, Illinois 60637 USA, [2]Department of Epidemiology and Biostatistics, Case Western Reserve University, 2103 Cornell Road, Cleveland, Ohio 44106 USA and [3]Department of Preventive Medicine and Epidemiology, Loyola University, 2160 South First Avenue, Maywood, Illinois 60153 USA

Email: Jun Zhang* - junzhang@galton.uchicago.edu; Xiaofeng Zhu - xzhu1@darwin.epbi.cwru.edu; Richard S Cooper - rcooper@lumc.edu

* Corresponding author

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

Published: 18 December 2007

BMC Proceedings 2007, **1**(Suppl 1):S35

This article is available from: http://www.biomedcentral.com/1753-6561/1/S1/S35

Abstract

We propose a nonparametric association analysis combining both family and unrelated case-control genotype data. Under the assumption of Hardy-Weinberg equilibrium, we formed an affected group to compare with a group of unaffecteds.

Comparison with traditional case-control chi-square test and transmission-disequilibrium test shows that this new approach has noticeably improved power. All analysis was based on the simulated rheumatoid arthritis data provided by Genetic Analysis Workshop 15. In the situation of population stratification, we also suggest an approach to update the genotype data using principal components. However, the Genetic Analysis Workshop 15 simulation data does not simulate population stratification. All analysis was done without knowledge of the answers.

Background

Traditional linkage analysis has achieved great success in the genetic dissection of mendelian diseases caused by a single gene with large effect. However, it is well known that association analysis has more power than linkage analysis for complex diseases such as rheumatoid arthritis (RA) [1]. Nowadays genome-wide association studies have been widely planned and carried out due to biotechnical improvements and decreasing experimental costs. Traditional approaches to association study designs are either family-based or unrelated case-control subjects based. Here we demonstrate an integrated association analysis using both family and unrelated simulation data on RA from Genetic Analysis Workshop 15 (GAW15).

Methods

Simulated data without population stratification

The RA data set was simulated according to familial patterns and other environmental effects. Each of the 100 replicates has 1500 nuclear families consisting of one affected sibling pair (ASP) and their parents, and 2000 unrelated unaffected individuals as controls. Markers include 730 microsatellite markers, 9187 evenly distributed SNPs on 22 autosomal chromosomes, and 17,820 dense SNPs on chromosome 6. In the analysis, we used

the first 200 families and the first 200 people of the 2000 controls. To include unrelated cases in the analysis, we randomly picked one of the two affected siblings from the next 200 families. Our final data set includes 200 families, 200 unrelated cases, and 200 controls. Among the 200 selected families, there were 56 families with a single parent and two families with both parents affected. In the most general setting, we form one group of all affected individuals consisting of affected siblings, affected parents, and unrelated cases, which was compared with a group of all unaffected individuals consisting of unaffected siblings, unaffected parents, and unrelated controls. Depending on the number of affected parents, there are three possible groupings for a family with r affected siblings with genotype $x_1,..., x_r$; s unaffected siblings with genotype $y_1,..., y_s$, and parents with genotype x_m and x_f Here, genotypes x and y denote the number of a particular allele whose allele frequency is p. Suppose in the data there are l families with both unaffected parents, m families with one affected parent (say the mother), n families with both affected parents, and additionally unrelated cases wi, $i = 1,..., u$, and v controls z_i, $i = 1,..., v$. The allele frequencies of the two groups are given by:

$$p_a = \frac{\sum_{i=1}^{l}(x_1^{(i)}+...+x_r^{(i)})+\sum_{j=1}^{m}(x_1^{(j)}+...+x_r^{(j)}+x_m^{(j)})+\sum_{k=1}^{n}(x_1^{(k)}+...+x_r^{(k)}+x_m^{(k)}+x_f^{(k)})+\sum w_i}{2[lr+m(r+1)+n(r+2)+u]}$$

$$p_u = \frac{\sum_{i=1}^{l}(y_1^{(i)}+...+y_s^{(i)}+x_m^{(i)}+x_f^{(i)})+\sum_{j=1}^{m}(y_1^{(j)}+...+y_s^{(j)}+x_f^{(j)})+\sum_{k=1}^{n}(y_1^{(k)}+...+y_s^{(k)})+\sum z_i}{2[l(s+2)+m(s+1)+ns+v]}.$$

We then use a normal test statistic $z = \frac{p_a - p_u}{\sqrt{Var(p_a - p_u)}}$, which is a generalization of Risch and Teng's result [2]. In particular, $Var(p_a - p_u) = Var(p_a) + Var(p_u) - 2Cov(p_a, p_u)$. Assuming Hardy-Weinberg equilibrium, each term is given below:

$$Var(p_a) = \frac{l(r^2+r)+m(r^2+3r+2)+2n(r^2+2r+2)+2u}{4[lr+m(r+1)+n(r+2)+u]^2}p(1-p)$$

$$Var(p_u) = \frac{2l(s^2+2s+2)+m(s^2+3s+2)+n(s^2+s)+2v}{4[l(s+2)+m(s+1)+ns+v]^2}p(1-p)$$

$$Cov(p_a, p_u) = \frac{lr(s+2)+m(r+1)(s+1)+ns(r+2)}{[lr+m(r+1)+n(r+2)+u][l(s+2)+m(s+1)+ns+v]}p(1-p),$$

And p is the estimated average allele frequency of all subjects in the data. For our final data, $r = 2$; $s = 0$; $l = 140$; $m = 56$; $n = 2$, and $u = v = 200$.

In the presence of population stratification

In the situation of population stratification, we suggest an approach to adjust the genotype data using principal components before the above procedures are applied. Unfortunately, the RA data was simulated without a population stratification effect, therefore we only give brief idea of this method here. The rationale of this approach is that across the genome there should be a consistent pattern among allele frequency differences, and that pattern is summarized by principal components to which many markers contribute. We sketch the procedures below. Details may be found in Price et al. [3]. First, pick founders from each family and all unrelated case-controls. Denote the genotype at the i^{th} locus for j^{th} individual by g_{ij}, $i = 1,..., M$ and $j = 1,..., N$. Let $u_i = \frac{1}{N}\sum_{j=1}^{N} g_{ij}$ be the sample mean for i^{th} locus and $X = (x_{ij})$ the matrix normalized by subtracting u_i from each row and dividing by $\sqrt{\frac{1}{2}u_i(1-u_i)}$. Second, compute the estimated covariance matrix of all markers $\psi_{M\times M} = \frac{1}{N-1}XX^T$, and list the first k largest eigenvalues $\lambda_1,..., \lambda_k$ with corresponding eigenvectors $v_1,..., v_k$ The l^{th} eigenvector $v_l = (v_{l1},..., v_{lM})$ gives the l^{th} principal component as $v_l \cdot g = (v_{l1},..., v_{lM}) \cdot (g_1,..., g_M) = \sum_{i=1}^{M} v_{li}g_i$. Finally, regress genotypes on the markers by $g_{ij,update} = g_{ij} - \sum_{l=1}^{k} v_{li}\sum_{s=1}^{M} v_{ls}g_{sj}$, where $\sum_{s=1}^{M} v_{ls}g_{sj}$ is the regression coefficient for l^{th} marker and j^{th} individual.

Results

Because population stratification was not simulated in GAW15, we did not adjust the genotype data using principal component procedures. We directly applied the test to the 9187 SNPs, and identified four SNPs whose p-values are far less than the Bonferroni corrected p-value $0.05/9187 = 5.44 \times 10^{-6}$. We used the software Haploview [4] to test the linkage disequilibrium pattern among them. The D' scores among SNP6-152, SNP6-153, and SNP6-154 are above 0.93, suggesting strong LD, and the D' between SNP6-155 and the rest was less than 0.38. Next, we applied a case-control chi-square test to the unrelated 200 cases and controls, and a family-based test (transmission-disequilibrium test, or TDT) to the family data. As a comparison, we also applied our test *zfam* only to the family data. All the test results were consistent, and are sum-

Table 1: The most significant SNPs out of the total 9187 markers and their test values with associated *p*-values before Bonferroni correction.

SNP	Location (cM)	test z	p_z	χ^2	p_{x2}	family z_{fam}	p_{fam}	TDT	p_{TDT}
SNP6-152	49.4300	16.10	0	92.93	0	12.31	0	7.57	1.87×10^{-14}
SNP6-153	49.4606	24.07	0	276.53	0	16.49	0	10.52	0
SNP6-154	49.4662	23.26	0	225.80	0	16.68	0	10.06	0
SNP6-155	49.6216	10.30	0	57.68	3.09×10^{-14}	6.60	2.06×10^{-11}	3.88	5.22×10^{-5}

marized in Table 1. The squares of the new test value z are strictly larger than the square sum of the corresponding chi-square test and TDT. For the family data, the value of our statistic *zfam* is also bigger than the value of TDT test statistic. These suggest that the proposed combined test has improved power. Also, as expected, the values of test statistic z are much larger than the test statistic *zfam*, which is restricted only to families, because more information from the unrelated case-control sample is used.

The type I errors of the proposed test are reasonable and comparable to the other two tests, which are listed in Table 2. At the significance level $\alpha = 0.05$, we observed 483 SNPs with p-values less than 0.05, giving a slightly higher type I error rate of 0.0525, which might be caused by correlation with disease loci. Thus, we excluded all the 674 SNPs on chromosome 6, and then observed 433 SNPs with p-value less than 0.05, with a corresponding type I error of 0.0508 (Table 2). Next, we applied our test to the dense map of chromosome 6, and got 56 significant SNPs whose p-values are less than the Bonferroni corrected p-value $0.05/(17820 + 9187) = 1.85 \times 10^{-6}$. In particular, the markers 3439, 3442, 3437, 3436, 3440, 3430, and 3426 have the largest test value. Together with the LD patterns from Haploview, we conclude that the most likely interval for a major gene is between 49.4262 cM and 49.5184 cM on chromosome 6.

Discussion

Under the assumption of Hardy-Weinberg equilibrium, the proposed approach has improved power by combining families of different structures with unrelated subjects, and it also give a potential way to resolve the issue of population stratification. Compared with the traditional TDT test, the proposed test can combine all the available families and may have better power than the TDT because the TDT excludes a certain proportion of families. Under the assumption of no population stratification and low disease prevalence in parents, another simpler test that Risch and Teng describe is to regard all parents from families as unaffected, with the remainder of this test being the same as ours [2]. However, when we carried out this test on the RA data, it led to an inflated type I error rate. At the significance level $\alpha = 0.05$, the type I error rate reached 0.055. On the other hand, our new proposed test might lose power without the random mating assumption.

Table 2: Type I error rates of different tests for all markers except those on chromosome 6.

Test	Type I error
Combined z test	0.0508
Family z_{fam}	0.0507
Case/control χ^2	0.0509
TDT	0.0506

Recently Epstein et al. [5] described a likelihood-based approach for combining triads and unrelated subjects, but it requires further work to combine families of different structures. Li et al. [6] also published another likelihood-based approach using hidden Markov model of affected sibling pairs. However, their approaches can not deal with the issue of population stratification. We proposed a principal-component based approach to resolve this, and will test the performance of adjusting population stratification procedure elsewhere.

Competing interests

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

Acknowledgements

The authors are very grateful to the reviewers for their numerous suggestions for improving the format and content of this paper. This work was supported by a grant from National Human Genome Research Institute (R01 HG003054) to XZ.

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. Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273:**1516-1517.
2. 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.
3. 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.

4. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21:**263-265.
5. Epstein MP, Veal CD, Trembath RC, Barker JN, Li C, Satten GA: **Genetic association analysis using data from triads and unrelated subjects.** *Am J Hum Genet* 2005, **76:**592-608.
6. Li M, Boehnke M, Abecasis G: **Efficient study for test of genetic association analysis using sibship data and unrelated cases and controls.** *Am J Hum Genet* 2006, **78:**778-792.