

Proceedings

Open Access

## Armitage's trend test for genome-wide association analysis: one-sided or two-sided?

Yixin Fang\*<sup>1</sup>, Yuanjia Wang<sup>2</sup> and Nanshi Sha<sup>2</sup>

Addresses: <sup>1</sup>Department of Mathematics and Statistics, Georgia State University, 750 COE, 7<sup>th</sup> Floor, 30 Pryor Street, Atlanta, Georgia 30303, USA and <sup>2</sup>Department of Biostatistics, Columbia University, 722 West 168 Street, Floor 6, New York 10032, USA

E-mail: Yixin Fang\* - matyxf@langate.gsu.edu; Yuanjia Wang - yw2016@columbia.edu; Nanshi Sha - ns2397@columbia.edu

\*Corresponding author

from Genetic Analysis Workshop 16  
St Louis, MO, USA 17-20 September 2009

Published: 15 December 2009

BMC Proceedings 2009, 3(Suppl 7):S37 doi: 10.1186/1753-6561-3-S7-S37

This article is available from: <http://www.biomedcentral.com/1753-6561/3/S7/S37>

© 2009 Fang et al; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

The importance of considering confounding due to population stratification in genome-wide association analysis using case-control designs has been a source of debate. Armitage's trend test, together with some other methods developed from it, can correct for population stratification to some extent. However, there is a question whether the one-sided or the two-sided alternative hypothesis is appropriate, or to put it another way, whether examining both the one-sided and the two-sided alternative hypotheses can give more information. The dataset for Problem 1 of Genetic Analysis Workshop 16 provides us with a chance to address this question. Because it is a part of a combined sample from the North American Rheumatoid Arthritis Consortium (NARAC) and the Swedish Epidemiological Investigation of Rheumatoid Arthritis (EIRA), the results from the combined sample can be used as references. To test this aim, the last 10,000 single-nucleotide polymorphisms (SNPs) on chromosome 9, which contain the common genetic variant at the *TRAF1-C5* locus, were examined by conducting Armitage's trend tests. Examining the two-sided alternative hypothesis shows that SNPs *rs12380341* ( $p = 9.7 \times 10^{-11}$ ) and *rs872863* ( $p = 1.7 \times 10^{-15}$ ), along with six SNPs across the *TRAF1-C5* locus, *rs1953126*, *rs10985073*, *rs881375*, *rs3761847*, *rs10760130*, and *rs2900180* ( $p \sim 1 \times 10^{-7}$ ), are significantly associated with anti-cyclic citrullinated peptide-positive rheumatoid arthritis. But examining the one-sided alternative hypothesis that the minor allele is positively associated with the disease shows that only those six SNPs across the *TRAF1-C5* locus are significantly associated with the disease ( $p \sim 1 \times 10^{-8}$ ), which is consistent with the results from the combined sample of the NARAC and the EIRA.

### Background

The Genetic Analysis Workshop 16 (GAW16) rheumatoid arthritis (RA) dataset is the initial batch of whole genome-wide association study (GWAS) data for the

North American Rheumatoid Arthritis Consortium (NARAC) cases ( $N_1 = 868$ ) and controls ( $N_0 = 1194$ ) after removing duplicated and contaminated samples [1]. The high-throughput genotyping technology [ $\sim 550$  k

single-nucleotide polymorphisms (SNPs)] in the NARAC data makes it a challenge to interpret this GWAS.

One of the disadvantages of the case-control GWAS is that they are prone to a number of biases including population stratification [2]. The importance of considering confounding due to population stratification in GWAS using case-control designs [3,4] has been a source of debate. The Armitage’s trend tests can correct for population stratification to some extent [5-7]; some other methods based on the Armitage’s trend tests have also been developed, such as genomic control approach [8,9]. However, there is still a question as to whether the one-sided or the two-sided alternative hypothesis is appropriate, or put it another way, whether examining both the one-sided and the two-sided alternative hypotheses can give more information. The dataset for the Problem 1 of GAW16 provides us with a chance to address this question. Because it is a part of a combined sample from the NARAC and the Epidemiological Investigation of Rheumatoid Arthritis (EIRA), the results from the combined sample can be used as references.

To this aim, the last 10,000 SNPs on chromosome 9, which contains the common genetic variant at the *TRAF1-C5* locus, were examined by conducting Armitage’s trend tests. Two alternative hypotheses, the two-sided alternative hypothesis that the genotypes at a locus are associated with the disease and the one-sided alternative hypothesis that the minor allele at a locus is positively associated with the disease, were considered. Three types of scores, co-dominant score, dominant score, and recessive score, were chosen to construct the Armitage’s trend tests.

**Methods**

At any SNP, the data can be summarized in a contingency table as in Table 1. Always assume that “M” is the major allele and “m” is the minor allele. Scores  $x_0$ ,  $x_1$ , and  $x_2$ , are chosen to construct Armitage’s trend test. The Armitage’s trend test statistic is defined as [5,6].

**Table 1: Contingency table at any SNP (M is major allele and m is minor allele)**

	Genotype			Total
	MM	Mm	Mm	
Case	$n_{10}$	$n_{11}$	$n_{12}$	$N_1$
Control	$n_{00}$	$n_{01}$	$n_{02}$	$N_0$
Total	$N_{+0}$	$N_{+1}$	$N_{+2}$	$N$
Score	$x_0$	$x_1$	$x_2$	

$$X_A^2 = \frac{N(N\sum n_{1j}x_j - N_1\sum N_{+j}x_j)^2}{N_1N_0\{N\sum N_{+j}x_j^2 - (\sum N_{+j}x_j)^2\}} \tag{1}$$

Under the null hypothesis, it is approximately distributed with  $\chi_1^2$ . This test statistic is suitable for the two-sided alternative hypothesis that the genotypes at a SNP are associated with the disease of interest. As discussed in Armitage [5], whatever the scoring system chosen, the validity of the test  $X_A^2$  is not affected, but the choice of scoring system affects the power of the test. There are three common choices of scoring system: 1) co-dominant score:  $x_0 = 0$ ,  $x_1 = 1$ , and  $x_2 = 2$ ; 2) dominant score:  $x_0 = 0$ ,  $x_1 = 1$ , and  $x_2 = 1$ ; 3) recessive score:  $x_0 = 0$ ,  $x_1 = 0$ , and  $x_2 = 1$ . Here, the names of scoring systems are in favor of the minor allele “m”.

From the rationale of the genetic association analysis (see, for example, Risch and Merikangas [10]), it is more informative to look at two one-sided alternative hypotheses, i) the alternative that the minor allele is positively associated with the disease and ii) the alternative that the major allele is positively associated with the disease. Furthermore, because the disease of interest is rare, it is more reasonable to concentrate on the first alternative, despite that in practice we would do better to consider both alternatives if no prior information is available on which allele is positively associated with the disease. Another reason is that it can reduce the false-positive rate.

Hereafter, we concentrate on the alternative hypothesis that the minor allele is positively associated with the disease. To this aim, one-sided can be defined as

$$Z_A = \frac{\sqrt{N}(N\sum n_{1j}x_j - N_1\sum N_{+j}x_j)}{\sqrt{N_1N_0\{N\sum N_{+j}x_j^2 - (\sum N_{+j}x_j)^2\}}} \tag{2}$$

Under the null hypothesis, it is approximately distributed with  $N(0,1)$ . Similarly, those three scoring systems can also be used here. It is shown in Knapp [11] that if the co-dominant scoring system is chosen, then  $Z_A = Z / \sqrt{1 + F}$ , where  $F$  is the Wright’s coefficient of inbreeding, and  $Z$  is the test statistic simply comparing the frequencies of minor allele “m” in the case and control groups. Here the value of  $F$  automatically corrects the population stratification to some extent.

**Results**

For simplicity of interpretation, we only consider the last 10,000 SNPs on chromosome 9, which contain the common genetic variant at the *TRAF1-C5* locus. The same analysis can be extended to the whole genome of approximately 550,000 SNPs.

**Table 2: LOD values for the two-sided alternative**

SNP <sup>a</sup>	Z <sup>2</sup> <sup>b</sup>	X <sub>A1</sub> <sup>2</sup> <sup>c</sup>	X <sub>A2</sub> <sup>2</sup>	X <sub>A3</sub> <sup>2</sup>	F <sup>d</sup>
rs4078292	6.14	5.23	<b>0.91<sup>d</sup></b>	6.96	0.1958
rs12380341	11.61	10.01	<b>0.36</b>	13.45	0.1722
rs16929545	7.22	6.71	<b>1.01</b>	7.62	0.0850
*rs1953126	7.56	7.53	5.05	5.44	<i>0.0037<sup>f</sup></i>
*rs10985073	6.98	6.87	5.63	4.19	<i>0.0173</i>
*rs881375	7.64	7.63	4.81	5.71	<i>0.0020</i>
*rs3761847	7.91	7.75	5.92	5.03	<i>0.0230</i>
*rs10760130	7.42	7.30	6.03	4.40	<i>0.0190</i>
*rs2900180	8.21	8.19	5.20	6.09	<i>0.0022</i>
rs872863	15.65	14.78	<b>1.51</b>	15.11	0.0617
rs888229	6.17	5.27	<b>1.09</b>	6.10	0.1914
rs1185665	7.54	6.48	<b>0.32</b>	9.68	0.1817
rs11792145	8.58	6.53	<b>0.04</b>	12.24	0.3488

<sup>a</sup>Asterisks indicates SNPs are located on *TRAF1-C5*.

<sup>b</sup>Z<sup>2</sup> is the Chi-square test comparing the frequencies of the minor allele in the two groups.

<sup>c</sup>Subscripts A1, A2 and A3 denote test (1) with score systems 1, 2, and 3, respectively.

<sup>d</sup>F is the Wright's coefficient of inbreeding.

<sup>e</sup>Bold font indicates X<sub>A2</sub><sup>2</sup> is significantly smaller than X<sub>A1</sub><sup>2</sup> and X<sub>A3</sub><sup>2</sup>.

<sup>f</sup>Italic font indicates F value is smaller than 0.03.

For the two-sided alternative that the genotypes at a SNP are associated with the disease, Table 2 summarizes the LOD scores (-log<sub>10</sub> p) of the test Z<sup>2</sup>, which simply compares the frequencies of the minor allele in both groups, the Armitage's tests X<sub>A1</sub><sup>2</sup> with co-dominant score, X<sub>A2</sub><sup>2</sup> with dominant score, X<sub>A3</sub><sup>2</sup> with recessive score, and the Wright's coefficient of inbreeding F; only those SNPs with LOD > 6 are reported. The SNPs across the *TRAF1-C5* locus are marked with asterisks.

In Table 2, those six SNPs marked with asterisks have small F (<0.03), and this explains why their X<sub>A1</sub><sup>2</sup> values in the third column, which correct for population stratification, are almost the same as Z<sup>2</sup> in the second column. Also, for these six SNPs, X<sub>A1</sub><sup>2</sup> is a bit more significant than X<sub>A2</sub><sup>2</sup> and X<sub>A3</sub><sup>2</sup>, and the latter two are close to each other, which means that these SNPs are very likely co-dominant. For the other seven SNPs, X<sub>A3</sub><sup>2</sup> is a bit more significant than X<sub>A1</sub><sup>2</sup>, but X<sub>A2</sub><sup>2</sup> is not significant at all. This shows that these SNPs are very likely recessive.

Another thing learned from Table 2 is that two SNPs, rs12380341 and rs872863, have extreme large LOD scores for Z<sup>2</sup>, X<sub>A1</sub><sup>2</sup>, and X<sub>A3</sub><sup>2</sup>, but surprisingly they were not reported by Plenge et al. [1], which was based on the combined sample from the NARAC and the EIRA. Are these two SNPs truly associated with the disease, or are they just false positives? Table 3 summarizes the LOD values for the one-sided alternative that the minor allele at a SNP is positively associated with the disease. Similarly, Z<sub>A1</sub> is the statistic Z<sub>A</sub> with co-dominant score, Z<sub>A2</sub> dominant score, and Z<sub>A3</sub> recessive score.

**Table 3: LOD values for the one-sided alternative of the minor allele**

SNP <sup>a</sup>	Z <sup>b</sup>	Z <sub>A1</sub> <sup>c</sup>	Z <sub>A2</sub>	Z <sub>A3</sub>
rs4078292	0.00	0.00	0.03	0.00
rs12380341	0.00	0.00	0.11	0.00
rs16929545	0.00	0.00	0.02	0.00
*rs1953126	7.86	7.83	5.35	5.74
*rs10985073	7.28	7.17	5.93	4.49
*rs881375	7.95	7.93	5.12	6.01
*rs3761847	8.21	8.05	6.22	5.33
*rs10760130	7.72	7.60	6.33	4.70
*rs2900180	8.51	8.49	5.50	6.39
rs872863	0.00	0.00	0.01	0.00
rs888229	0.00	0.00	0.02	0.00
rs1185665	0.00	0.00	0.63	0.00
rs11792145	0.00	0.00	0.27	0.00

<sup>a</sup>Asterisks indicates SNPs are located on *TRAF1-C5*.

<sup>b</sup>Z is the z-test comparing the frequencies of the minor allele in the two groups.

<sup>c</sup>Subscripts A1, A2 and A3 denote the test (2) with score systems 1, 2, and 3, respectively.

From Table 3, only those six SNPs marked with asterisks are significant for the one-sided alternative that the minor allele is positively associated with the disease. These results are completely consistent with the ones in Plenge et al. [1]. By consider the other one-sided alternative that the major allele is positively associated with the disease, the other seven SNPs are significant. Therefore, as discussed in the preceding section, and particularly for this dataset, it seems that it is more reasonable to consider the one-sided alternative that the minor allele is positively associated with the disease.

### Discussion

The question of whether the two-sided alternative or the one-sided alternatives should be considered is intractable, but this manuscript attempts to raise the question and address it to some extent. Table 3 shows that if we concentrate on the one-sided alternative that the minor allele is positively associated with the disease, we get exactly the same results as Plenge et al. [1]. For rare diseases, and we have reason to believe that the alleles positively associated with them have low frequencies in a general population. Based on this belief (or alternative hypothesis), it seems that those SNPs without asterisks are false positives under the two-sided alternative.

But if we do not want to believe that the minor allele is positively associated with the disease and do not want to miss any SNPs related to the disease, we had better consider the two-sided alternative.

### Conclusion

More information can be gained from GWAS by using multiple scoring systems in the Armitage's trend tests

and examining both the one-sided and the two-sided alternative hypotheses.

### List of abbreviations used

EIRA: Epidemiological Investigation of Rheumatoid Arthritis; GAW16: Genetic Analysis Workshop 16; GWAS: Genome-wide association; NARAC: North American Rheumatoid Arthritis Consortium; RA: Rheumatoid arthritis; SNP: Single-nucleotide polymorphism(s).

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

YF conceived the idea, performed the statistical analysis, and drafted the manuscript. YW and NS helped to perform the statistical analysis and draft the manuscript. All authors read and approved the final manuscript.

### Acknowledgements

The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences.

This article has been published as part of *BMC Proceedings* Volume 3 Supplement 7, 2009: Genetic Analysis Workshop 16. The full contents of the supplement are available online at <http://www.biomedcentral.com/1753-6561/3?issue=S7>.

### References

1. Plenge R, Seielstad M and Padyukov L, et al: **TRAF1-C5 as a risk locus for rheumatoid arthritis—a genome-wide study.** *N Engl J Med* 2007, **357**:1199–1209.
2. Pearson TA and Manolio TA: **How to interpret a genome-wide association study.** *JAMA* 2008, **299**:1335–1345.
3. Thomas DC and Witte JS: **Point: population stratification: a problem for case-control studies of candidate-gene associations?.** *Cancer Epidemiol Biomarkers Prev* 2002, **11**:505–512.
4. Wacholder S, Rothman N and 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.
5. Armitage P: **Tests for linear trends in proportions and frequencies.** *Biometrics* 1955, **11**:375–386.
6. Sasieni PD: **From genotypes to genes: doubling the sample size.** *Biometrics* 1997, **53**:1253–1261.
7. Schaid DJ and Jacobsen SJ: **Biased tests of association: comparisons of allele frequencies when departing from Hardy-Weinberg proportions.** *Am J Epidemiol* 1999, **149**:706–711.
8. Devlin B and Roeder K: **Genomic control for association studies.** *Biometrics* 1999, **55**:997–1004.
9. Reich D and Goldstein D: **Detecting association in a case-control study while correcting for population stratification.** *Genet Epidemiol* 2001, **20**:4–16.
10. Risch N and Merikangas K: **The future of genetic studies of complex human disease.** *Science* 1996, **273**:1516–1517.
11. Knapp M: **Re: “Biased tests of association: comparisons of allele frequencies when departing from Hardy-Weinberg proportions”.** *Am J Epidemiol* 2001, **154**:287–288.

Publish with **BioMed Central** and every scientist can read your work free of charge

*"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."*

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

