A Comprehensive Analysis of Association between Several Candidate Genes and Rheumatoid Arthritis in Samples from NARAC

Lei Cong1*, Tian Zheng1*, Iuliana Ionita-Laza2, Yuejing Ding1 and Shaw-Hwa Lo1

1Department of Statistics, Columbia University, New York, NY 10027; 2Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115.

Objective

Comprehensive study on a set of genes of Rheumatoid Arthritis, with emphasis on joint effects of multiple genes.

Data (Problem 2)

- 839 cases and 855 controls from the North American Rheumatoid Arthritis Consortium (NARAC);
- 20 SNPs in 14 candidate genes, previously studied by Plenge et al. (2005)

Methods

1. Data processing

17 SNPs have missing genotypes under 3%. SNP rs2240340 (PADI4) have about 65% missing. SNPs rs1061622 and 5509_5511delCAA have missing level around 15%.

The fastPHASE program (Scheet and Stephens, 2006) were used to impute missing genotypes.

2. Multi-maker association information measure

Markers: M1, ..., Mk
Genotypes: G1, G2, ..., G 3^k
ni cases and nu controls
Disease status

Genotypes G1, G2, ..., Gv
n1 cases and n2 with genotype Go,
i=1, ..., 3^k

- Genotype-Trait Distortion (GTD) measures genotype i’s sample relative frequency among the cases with that of the controls.

\[ \text{GTD} = \frac{c(n)}{\sum_{i=1}^{3^k} (n_i^d/n_i - n_i^u/n_u)^2} \]

(Zheng et al., 2006)

- The expectation of the GTD score is

\[ E(\text{GTD}(k)) = c(n) [\sum (p_i^d - p_i^u)^2 + O(1/n)] \]

\( p_i^d \) and \( p_i^u \) are genotype i’s population relative frequencies among cases and controls.

- Using the backward greedy screening developed in Zheng et al. (2006), we screen the 20 SNPs to identify local optimal clusters (subsets) with high GTD scores.

3. Selection threshold and evaluation of significance

Since we are searching for optimal SNP clusters of different sizes in the greedy algorithm, we identify significant SNP clusters as follows.

- For the original data: compute GTD and its standard deviation on all possible subsets.
- Permute the disease status labels (case or control) randomly. Obtain 100 such permuted data sets.
- For each of the permutation set, compute GTD and its standard deviation on all possible subsets.
- Selection threshold: for a cluster (subset) size, select the local optimal clusters identified with GTD score higher than the maximum GTD score obtain in the permutations.

Using the permutation, we are controlling the family-wise type I error at 1% level.

We show in Figure 1 the results from real data and permuted data. The signal in the real data is much higher than that observed in permutations, even after corrected using the Bonferroni procedure.

1. Significance of the association information (Figure 1)

- Red solid line: at each subset size, the highest GTD score for real data with a 95% confidence interval corrected using the Bonferroni procedure.
- Black dotted line at 1: unexpected value under the null hypothesis of no association.
- Black solid line: at each subset size, the median maximum GTD scores and its 95% confidence interval out of 100 permutations are plotted.

For each permutation, we also calculated Bonferroni-corrected 95% confidence intervals for the maximum GTD scores.

Blue shading: the coverage of these 100 confidence intervals at each subset size, with the darkest being 0.9-1 (or 90%-100%) and the lightest being 0.0-0.1 (or 0%-10%).

2. Association network construction

- 8 SNP subsets on 7 SNPs identified.
- Marginally significant (or one-SNP subsets): rs2476601 (PTPN22), rs237025 and rs577001 (both at SUMO4).
- Using subsets with more than one SNP, a graphical network is constructed using GUESS (Adar 2006).

Conclusions

- This a comprehensive analysis of the candidate genes for Rheumatoid Arthritis.
- For case-control data, the unphased genotypes on multiple markers were used.
- We studied the association between the RA phenotype with multiple genetic loci without assuming any disease model.
- Interesting significant results are identified even after corrected for multiple comparison based on permutations.

Acknowledgment

The research was supported by NIH grant R01 GM070789.

* Correspondences: cong1@stat.columbia.edu or tzheng@stat.columbia.edu

---

Table 1: 20 SNPs genotyped on the loci of 14 putative RA candidate genes.