BMC Proceedings

BioMed Central

Proceedings

A new gene-based association test for genome-wide association studies

Alfonso Buil*[1], Angel Martinez-Perez[1], Alexandre Perera-Lluna[2], Leonor Rib[1], Pere Caminal[2] and Jose Manuel Soria[1]

Addresses: [1]Unitat de Genomica de Malalties Complexes, Institut de Recerca de l'Hospital de la Santa Creu i Sant Pau, Barcelona, 08025, Spain and [2]Centre de Recerca en Enginyeria Biomedica, Departament d'Enginyeria Sistemes i Automatica Industrial, Universitat Politecnica de Catalunya, Pau Gargallo, 5, 08028 Barcelona, Spain

E-mail: Alfonso Buil* - abuil@santpau.cat; Angel Martinez-Perez - amartinezp@santpau.cat;
Alexandre Perera-Lluna - alexandre.perera@upc.edu; Leonor Rib - mrib@santpau.cat; Pere Caminal - pere.caminal@upc.edu;
Jose Manuel Soria - jsoria@santpau.cat
*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):S130 doi: 10.1186/1753-6561-3-S7-S130

This article is available from: http://www.biomedcentral.com/1753-6561/3/S7/S130

Abstract

Genome-wide association studies are widely used today to discover genetic factors that modify the risk of complex diseases. Usually, these methods work in a SNP-by-SNP fashion. We present a gene-based test that can be applied in the context of genome-wide association studies. We compare both strategies, SNP-based and gene-based, in a sample of cases and controls for rheumatoid arthritis.

We obtained different results using each strategy. The SNP-based test found the *PTPN22* gene while the gene-based test found the *PHF19-TRAF1-C5* region. That suggests that no single strategy performs better than another in all cases and that a certain underlying genetic architecture can be delineated more easily with one strategy rather than with another.

Introduction

The last three years have witnessed an enormous increase in genome-wide association studies (GWAS). The original idea behind these studies is to genotype some hundreds of thousands of single-nucleotide polymorphisms (SNP) in a sample of cases and controls for a given disease, and then look for association between every SNP and the disease. Thus, the researcher has to perform as many statistical tests as SNPs that he has. In other words, the unit of association is the SNP. In this paper, we propose a statistical test for GWAS in which the basic unit of association is the gene–that is, the researcher performs one statistical test per gene. The idea is to combine the genetic information given by all the SNPs in a gene to obtain a more informative result.

Our test presents two advantages compared with the classical SNP-based test. On the one hand, it suffers much less from the problem of multiple testing because we are doing around 20,000 tests instead of half a million. On the other hand, we expect that in those genes with multiple functional variants, the gene-based test will be more powerful than the SNP-based test [1]. However, the gene-based test presents some drawbacks. Because genes do not cover the whole genome, the gene-based test does not use all of the SNPs available. Moreover, the gene-based test is not as simple as the SNP-based test and therefore it requires more computational resources. Our goal was to compare the results using a SNP-based test to those obtained using a new gene-based test for GWAS.

To perform the comparison, we used data from individuals with and without rheumatoid arthritis (RA)–a systemic autoimmune disease characterized by inflammation of the synovial tissue and local articular damage. Studies of RA heritability in two European populations reported that 60% of the disease variance can be attributed to genetic factors [2]. Linkage and association studies have demonstrated that alleles at the human leukocyte antigen (HLA) class II gene *DRB1* have a strong effect on the risk of RA [3]. However, these variants do not explain all of the heritability. It is possible that loci not linked to the HLA region play an important role in RA susceptibility. Several studies have reported genes or genomic regions related to RA susceptibility outside the HLA region [4-8].

Methods

The data

We used the subset of the data from the North American Rheumatoid Arthritis Consortium (NARAC) study provided for Genetic Analysis Workshop 16 [9]. It consisted of 868 cases and 1194 controls. The Ilumina 550 k chip (545,080 SNPs) was genotyped in the whole sample. Before the analysis, we cleaned the data. For example, we excluded SNPs with a call rate smaller than 0.95 (18,627 SNPs), minor allele frequency smaller than 0.01 (23,047 SNPs), or with a *p*-value for the Hardy-Weingberg equilibrium test smaller than 1×10^{-5} (1,342 SNPs). We also excluded six individuals with sex genotype inconsistencies. All of the individuals had a call rate greater than 0.95. All of our analyses were done with 'affection status' as the trait of interest and 'sex' as a covariate.

SNP-based test

We estimated the association between the trait RA and a SNP using logistic regression assuming a codominant model for the genetic effect of every SNP. We used the *p*-value of the test as a measure of statistical significance. We used the GenABEL software package [10].

Gene-based test

To perform a gene-based test, the first problem is to define the genes and to assign SNPs to the genes. Also, it is important to be sure that the physical positions of the genes and SNPs refer to the same annotation release. We used the NCBI build 129 release 36.3 for both genes and SNPs. We accepted that a SNP was in a gene if it was inside the gene plus or minus 5,000 base pairs. We analyzed 21,672 genes that contained 272,604 SNPs. That means that around half of the available SNPs were not assigned to genes and, thus, they were left out of the gene-based analysis.

The paradigm for the proposed gene-based test has three steps [11]:

1. Estimate the genetic similarity among individuals based on the genotypes of the SNPs in a given gene.

2. Cluster the individuals in groups by genetic similarity.

3. Test the association between the groups of individuals and the trait of interest.

In the first step, we used the Gower distance. Also known as Gower's coefficient [12], it is a measure of the similarity between two individuals based on the information given by a set of quantitative or qualitative variables. We realized that, in the special case of SNP genotypes, Gower distance is the same as the identity-by-state (IBS) multilocus measure. IBS allele sharing is a measure of genetic similarity between two individuals. Given the genotypes of two individuals at a given SNP, the IBS between them is 0, 1, or 2 depending on whether they share 0, 1, or 2 alleles at that SNP. This measure can be extended to several SNPs by adding the IBS for each locus and dividing by twice the number of loci [13]:

$$mIBS_{ij} = \frac{\sum_{l=1}^{L} IBS_{ij}^{l}(g_i^l, g_j^l)}{2L},$$

where L is the number of loci considered in the calculation; $g^l{}_i$ and $g^l{}_j$ are the genotypes of individuals *i* and *j*, respectively, at the l^{th} locus (l = 1, ..., L); and $\text{IBS}^l{}_{ij}$ is the IBS between *i* and *j* at locus *l*. We estimated this similarity measure for every pair of individuals at every gene. Thus, the result of the first step is a distance matrix among individuals in a given gene.

In the second step, the distance matrix is used for finding groups of individuals with similar genotypic distribution in the given gene. This clustering is performed in a hierarchical procedure by means of a complete linkage agglomerative algorithm. Complete linkage evaluates distances between two groups as the distance of their most distant pair of individuals. We divided the individuals in three groups of similarity. To test the effect of the chosen cluster algorithm on the results, we repeated the analysis in chromosomes 1, 6, and 9 using two other cluster algorithms: hierarchical average linkage agglomerative algorithm and spectral clustering. Average linkage evaluates distances between two groups as the mean distance between individuals of each cluster.

Spectral clustering is a method that defines *k* clusters on a set of *n* data points representing arbitrary objects. It is based on the spectral decomposition of the normalized Laplacian graph defined from a similarity matrix among the objects [14]. In our case, the objects are the individuals and the similarity matrix is the genetic similarity matrix defined above.

Finally, in the third step, association between the groups of individuals and the phenotype of interest was estimated using logistic regression with the group as a factor.

Results

SNP-based association

With 545,080 SNPs, the strict Bonferroni *p*-value for a genome-wide significance of 0.05 for the SNP-based test is $p = 9.17 \times 10^{-8}$. However, because there is some amount of linkage disequilibrium among them, the effective number of tests is smaller. A recent study has estimated that the number of effective independent tests done with the Illumina 550 k chip is 324,559 [15]. That means that the adjusted critical *p*-value is 1.5×10^{-7}. As expected, there are many SNPs showing statistically significant association with RA in chromosome 6. Specifically, there are 213 statistically significant SNPs, that cover 74 genes, all of them in the HLA region, in 6p21. In addition, the SNP rs2476601 in chromosome 1 has a *p*-value of 2.04×10^{-8}. This SNP lies on the *PTPN22* gene, which has been previously associated with RA [4].

Gene-based association

With 21,672 genes analyzed, the strict Bonferroni *p*-value for a genome-wide significance of 0.05 for the gene-based test is $p = 2.3 \times 10^{-6}$. The gene-based association test presents 60 statistically significant hits, with *p*-values as low as 1×10^{-74}, in 6p21, as expected. The number of SNPs in these genes ranges from 2 to 37, with an average of 7.7. Table 1 shows the top gene-based association results outside of chromosome 6.

The *PHF19* gene lies in 9q33.2–an area that has been associated recently with RA [5,8], and shows a clear statistical significance with the gene-based test, but not with the SNP-based test. Three other areas that show suggestive association with RA are 16p13.12, 12q12, and 3p25. These areas contain previously described quantitative-trait loci for RA, which can be found at the UCSC genome browser database [16].

To test the sensitivity of the new method to the cluster algorithm, we repeated the analysis of chromosomes 1, 6, and 9 using two other cluster algorithms: hierarchical agglomerative average algorithm and spectral clustering. In the HLA region, the average clustering found 42 of the 60 genes found by the original cluster method and 12 genes not detected by the original. On the other hand, the spectral clustering found 54 of the 60 genes found also by the original cluster method and 28 not detected by the original. Outside of the HLA region, the gene *PHF19* on chromosome 9 was found by the spectral, but not by the hierarchical average algorithm. None of the cluster algorithms found the *PTPN22* gene that was found by the SNP-based test.

A possible question associated with the new method is whether it is sensitive to the number of SNPs in the gene. We discretized the number of SNPs in quartiles and tested the association between this variable and the $-\log_{10}$ of the *p*-value by means of a one-way ANOVA. We did not find an association between the two variables ($p = 0.14$). We repeated the analysis for each chromosome separately and none of them gave a statistically significant association.

Discussion

We have compared the results of a new gene-based test with the results using a standard SNP-based test in a sample of cases and controls for RA in a GWAS. The procedure of the test was to group the individuals by

Table 1: Gene-based association results out of chromosome 6

Symbol	Map location	No. SNPs analyzed on gene	*p*-Value of the gene-based test for the gene	Minimum *p*-value obtained for any SNP on the gene using the SNP-based test
PTPN22	1p13	6	0,11	2.04×10^{-8}
PHF19	9q33.2	4	9.61×10^{-7}	7.07×10^{-6}
FLJ11151	16p13.12	40	3.98×10^{-6}	2.69×10^{-4}
GLT8D3	12q12	6	8.37×10^{-6}	4.92×10^{-2}
LSM3	3p25.1	5	1.24×10^{-5}	2.31×10^{-4}

genetic similarity in a given gene and test whether the distribution of groups is different between the cases and the controls. In some ways, this approach is related to the association with haplotypes, because both approaches combine genetic information from several SNPs. However, haplotype-based methods suffer from a statistical problem–that is, the abundance of rare haplotypes. To solve this problem, different strategies that reduce the number of haplotypes have been proposed, i.e., haplotype clustering [17] or inference of ancestral haplotypes [18]. Our approach shares this idea, but is much simpler because it works directly with genotypes instead of haplotypes. In fact, in a previous study, we used this approach with data of a complete resequenced gene and we were able to group the individuals in the same way that when using the ancestral haplotypes [11].

As expected, both gene-based and SNP-based tests gave many statistically significant results in the HLA region. However, both tests gave different results outside of the HLA region. The SNP-based test found the *PTPN22* gene while the gene-based test found the *PHF19* gene. A possible explanation is that different underlying genetic architectures can be more easily detected with one strategy but not with the other.

The gene-based test has some limitations. One of these limitations is that, as defined here, it uses only half of the genetic information available. That is a consequence of how we assign SNPs to a gene: a SNPs belongs to a gene if it is inside the gene plus or minus 5,000 base pairs. Some researchers solve this problem assigning every SNP to the nearest gene, no matter how far it is from the gene. We think that a better strategy is to change the gene-based test to a region-based test. In our test, we used the gene as the unit of the analysis, but a genomic region of a fixed size could be used instead. Then, the test could be applied with a 'sliding window' strategy and utilize all the available genetic information.

Other limitations are related to the clustering algorithm. On the one hand, different algorithms can give slightly different results and, on the other hand, the test does not take into account the uncertainty of the clustering process. To measure and, eventually, correct this bias, we are planning to evaluate the effects of using different cluster algorithms and to use re-sampling techniques, as bootstrap or permutations, to estimate the statistical evidence of the test.

Conclusion

We performed a GWAS for RA affected status first with a standard SNP-based association test and second with a new gene-based association test. Both strategies gave a large amount of statistical evidence of association in 6p21 and little evidence outside of this region. With the SNP-based test we found a SNP in the *PTPN22* gene that was not found with the gene-based approach. On the other hand, with the gene-based test we found an area in 9q33.2 that was not detected by the SNP-based test. We do not think that one test is better than the other. They simply use the genetic information in a different way. We consider both tests as complementary.

List of abbreviations used

GWAS: Genome-wide Association Studies; HLA: Human leukocyte antigen; IBS: Identity by state; RA: Rheumatoid arthritis; SNP: Single-nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AB designed the study, participated in the statistical analysis, and drafted the manuscript. AM-P carried out most of the statistical analyses. AP-L carried out the spectral clustering analysis and participated in the writing of the manuscript. LR participated in the statistical analysis. PC and JMS participated in the design of the study and reviewed the manuscript.

Acknowledgements

The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. We acknowledge the advice and helpful discussions of Professor W.H. Stone. This study was supported partially by grants 2 R01 HL070751-05 from the USA NIH, PI-05/1361, PI-05/1382. J.M. Soria was supported by "Programa d'Estabilització d'Investigadors de la Direcció d'Estrategia i Coordinació del Departament de Salut" (Generalitat de Catalunya). This work was supported by the Spanish Ministerio de Educacion y Ciencia under the Ramon y Cajal Program and TEC2007-63637/TCM and the CIBER-Centro De Investigacion Biomedica en Red en Bioingenieria, Biomateriales y Nanomedicina. We also thank two anonymous reviewers for their useful comments.

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. Neale BM and Sham PC: **The future of association studies: gene-based analysis and replication.** *Am J Hum Genet* 2004, **75:**353–362.
2. MacGregor AJ, Snieder H, Rigby AS, Koskenvuo M, Kaprio J, Aho K and Silman AJ: **Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins.** *Arthritis Rheum* 2000, **43:**30–37.
3. Hasstedt SJ, Clegg DO, Ingles L and Ward RH: **HLA-linked rheumatoid arthritis.** *Am J Hum Genet* 1994, **55:**738–746.
4. Begovich AB, Carlton VE, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC, Ardlie KG, Huang Q, Smith AM, Spoerke JM, Conn MT, Chang M, Chang SY, Saiki RK, Catanese JJ, Leong DU, Garcia VE, McAllister LB, Jeffery DA, Lee AT, Batliwalla F,

Remmers E, Criswell LA, Seldin MF, Kastner DL, Amos CI, Sninsky JJ and Gregersen PK: **A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis.** *Am J Hum Genet* 2004, **75:**330–337.

5. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LR, Li W, Tan AK, Bonnard C, Ong RT, Thalamuthu A, Pettersson S, Liu C, Tian C, Chen WV, Carulli JP, Beckman EM, Altshuler D, Alfredsson L, Criswell LA, Amos CI, Seldin MF, Kastner DL, Klareskog L and Gregersen PK: **TRAF1-C5 as a risk locus for rheumatoid arthritis–a genomewide study.** *N Engl J Med* 2007, **357:**1199–1209.
6. Remmers EF, Plenge RM, Lee AT, Graham RR, Hom G, Behrens TW, de Bakker PI, Le JM, Lee HS, Batliwalla F, Li W, Masters SL, Booty MG, Carulli JP, Padyukov L, Alfredsson L, Klareskog L, Chen WV, Amos CI, Criswell LA, Seldin MF, Kastner DL and Gregersen PK: **STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus.** *N Engl J Med* 2007, **357:**977–986.
7. Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, Burtt NP, Gianniny L, Korman BD, Padyukov L, Kurreeman FA, Chang M, Catanese JJ, Ding B, Wong S, Helm-van Mil van der AH, Neale BM, Coblyn J, Cui J, Tak PP, Wolbink GJ, Crusius JB, Horst-Bruinsma van der IE, Criswell LA, Amos CI, Seldin MF, Kastner DL, Ardlie KG, Alfredsson L, Costenbader KH, Altshuler D, Huizinga TW, Shadick NA, Weinblatt ME, de Vries N, Worthington J, Seielstad M, Toes RE, Karlson EW, Begovich AB, Klareskog L, Gregersen PK, Daly MJ and Plenge RM: **Common variants at CD40 and other loci confer risk of rheumatoid arthritis.** *Nat Genet* 2008, **40:**1216–1223.
8. Chang M, Rowland CM, Garcia VE, Schrodi SJ, Catanese JJ, Helm-van Mil van der AH, Ardlie KG, Amos CI, Criswell LA, Kastner DL, Gregersen PK, Kurreeman FA, Toes RE, Huizinga TW, Seldin MF and Begovich AB: **A large-scale rheumatoid arthritis genetic study identifies association at chromosome 9q33.2.** *PLoS Genet* 2008, **4:**e1000107.
9. Amos CI, Chen WV, Seldin MF, Remmers E, Taylor KE, Criswell LA, Lee AT, Plenge RM, Kastner DL and Gregersen PK: **Data for Genetic Analysis Workshop 16 Problem 1, association analysis of rheumatoid arthritis data.** *BMC Proc* 2009, **3 (suppl 7):**S2.
10. Aulchenko YS, Ripke S, Isaacs A and van Duijn CM: **GenABEL: an R library for genome-wide association analysis.** *Bioinformatics* 2007, **23:**1294–1296.
11. Perera A, Buil A, Di Bernardo MC, Sabater-Lleal M, Pattini L, Souto JC, Fontcuberta J, Vallverdu M, Soria JM and Caminal P: **Clustering of individuals given SNPs similarity based on normalized mutual information: F7 SNPs in the GAIT sample.** *Conf Proc IEEE Eng Med Biol Soc* 2007, **2007:**123–126.
12. Gower JC: **A general coefficient of similarity and some of its properties.** *Biometrics* 1971, **27:**623–637.
13. Wessel J and Schork NJ: **Generalized genomic distance-based regression methodology for multilocus association analysis.** *Am J Hum Genet* 2006, **79:**792–806.
14. Shi J and Malik J: **Normalized cuts and image segmentation.** *IEEE Trans Pattern Anal Mach Intell* 2000, **22:**888–905.
15. Lewinger JP, Duggan DJ, Tavera DM, Gauderman WJ, Stram DO and Thomas DC: **Multiple testing in genomewide association studies: how bad really is the Bonferroni correction? [abstract].** *Proceedings of the International Genetic Epidemiology Society Meeting: 2008 September 15-16; St. Louis. Houston: International Genetic Epidemiology Society* 2008 http://www.geneticepi.org/meetings/2008/files/2008Abstracts.pdf.
16. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D and Kent WJ: **The UCSC Genome Browser Database: 2008 update.** *Nucleic Acids Res* 2008, **36:**D773–D779.
17. Waldron ER, Whittaker JC and Balding DJ: **Fine mapping of disease genes via haplotype clustering.** *Genet Epidemiol* 2006, **30:**170–179.
18. Su SY, Balding DJ and Coin LJ: **Disease association tests by inferring ancestral haplotypes using a hidden markov model.** *Bioinformatics* 2008, **24:**972–978.