EVIDENCE OF SNP EFFECT ON THE RISK OF RHEUMATOID ARTHRITIS: EFFECTS OF COVARIATE ADJUSTMENT UPON ASSOCIATION RESULTS

Suely Ruiz GIOLO[1]
Júlia Maria Pavan SOLER[2]
Maria Jacqueline BATISTA[3]
Márcio Augusto Afonso de ALMEIDA[4]
Alexandre Costa PEREIRA[4]

▪ *ABSTRACT: In this paper, an association analysis approach through covariate adjustment is proposed to classify Single Nucleotide Polymorfism (SNP) effects associated with the risk of rheumatoid arthritis (RA). Initially, the marginal effect of each SNP is evaluated by considering a single locus logistic regression. This effect is then evaluated adjusted by a covariate of known biological effect (HLA-DRB alleles) on the RA risk. To take into account possible influences of population stratification, the first ten axes of variation, resulting from a principal components analysis of the SNPs, are incorporated into the analysis as covariates. For comparison purposes, analysis without these axes is also performed. SNPs from all autosomal chromosomes of the RA data from the Genetic Analysis Workshop 16 are used in the analysis. From comparisons carried out with regard to the SNP effects obtained without and with adjustment by a covariate of known biological effect, classification of these effects is suggested in terms of direct and indirect influence, as well as influence through epistatic mechanisms. The proposed analytical procedure was shown to be useful overall, not only to disclose potentially important higher-order effects, but also to provide a classification of the SNPs.*

[1]Federal University of Paraná – UFPR, Department of Statistics, CEP: 81531-990, Curitiba, PR, Brazil. E-mail: *giolo@ufpr.br*
[2]University of São Paulo – USP, Department of Statistics, CEP: 05315-970, São Paulo, SP, Brazil. E-mail: *pavan@ime.usp.br*
[3]Federal University of Ceará – UFCE, Department of Statistics and Applied Mathematics, CEP: 60455-760, Fortaleza, CE, Brazil. E-mail: *mjb@ime.usp.br*
[4]Medical School of University of São Paulo, Heart Institute, Laboratory of Genetics and Molecular Cardiology, CEP: 05403-000, São Paulo, SP, Brazil. E-mail: *marcio.almeida@incor.usp.br* / *lbmpereira@incor.usp.br*

▪ *KEYWORDS: Association studies; human genetics; logistic regression; molecular markers; population-stratification.*

1 Introduction

The application of statistical methodologies are particularly useful in finding genetic variations that contribute to complex diseases such as asthma, cancer, diabetes and heart disease. The advent of cost-effective high-throughput genotyping technologies has greatly helped the identification of genetic variants underlying human diseases. High-throughput means that large numbers of genetic markers can be quickly assayed in a large number of DNA samples for a small cost per assay. A genetic marker, for instance, may be a short DNA sequence, such as a sequence surrounding a single base-pair change like single nucleotide polymorphisms (SNPs) which are the most common type of genetic variation and represent over 80% of the genetic variation between individuals (Ziegler and Konig, 2006). SNPs are ideal candidates for research correlating phenotype and genotype. Since some SNPs predispose individuals to a certain disease or a trait or cause an altered reaction to a drug, they are proving to be highly useful in diagnostics and drug development. There are tens of millions of SNPs present in the genome of a typical organism. However, usually only a very small subset of these will be developed into genetic markers (SNP markers). The main advantages of SNPs are: (1) they are so common and evenly-distributed in the genome, and (2) methods of detecting (or assaying) SNPs can be easily automated. This ease of automation is what makes SNPs high-throughput markers. The main disadvantage of SNPs is the small number of alleles typically present. Although, in theory, each SNP marker can have up to four possible alleles (A, C, G, and T), in practice, only two alleles usually are present at any given SNP (e.g., C or T). However, the large number of available SNPs and their low assay costs overcome the disadvantage of their low variability per marker. Affymetrix and Illumina are two companies that offer GeneChips that allow the genotyping of a large number of human SNPs. The Affymetrix SNP Array 6.0, for instance, enables genotyping more than 906,600 single nucleotide polymorphisms (SNPs). Similarly, the Illumina 550K chip enables genotyping about 550,000 SNPs while the Illumina Human 1M-Duo, over than 1 million SNPs markers.

In this context, a genome-wide association study (GWAS) is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. The simplest way to analyze data generated by genome-wide association studies is to carry out an association test for each SNP ascertained in the chip. Nevertheless, this approach is severely affected by the multiple tests involved in this strategy, as well as by possible population-stratification effects and association between loci due to, for instance, linkage (i.e., association of two or more loci on a chromosome with limited

recombination between them), linkage disequilibrium (LD) which is the non-random association of alleles at two or more loci, not necessarily on the same chromosome, and epistasis (interaction between two or more genes). For complex traits, it is widely accepted that disease-associated individual SNPs have a minor effect overall; it is only when these effects are combined amongst them or with additional loci that they may strongly provide more consistent predictive information.

In order to contribute to genome-wide association analysis, we propose in this paper a procedure that compares the results of two different models (with and without adjustment for covariates of known biological effect), in order to classify SNP effects associated with the risk of rheumatoid arthritis (RA). Case-control data from the Genetic Analysis Workshop 16 is used in the analyzes performed.

2 The data

Rheumatoid arthritis (RA) is a common and complex inflammatory human disorder (Plenge *et al.*, 2005; Begovich *et al.*, 2004; du Montcel *et al.*, 2005) involving genetic determinants and environmental factors in its development. Numerous studies have suggested that the HLA region in 6p21 is associated to RA, and there is consistent evidence that HLA-DRB alleles contribute to increase the risk of this disease. Moreover, there is evidence that shared epitope alleles (SENum) are also associated with RA risk, and that their effect is modulated by alleles in the HLA region (Irigoyen *et al.*, 2005).

In this paper are assessed data from 2,062 individuals (72.4% of them females), being 868 cases (affected by RA) and 1,194 controls. This data is the initial batch of whole genome association data for the North American Rheumatoid Arthritis Consortium (NARAC). For each individual it was recorded the number of shared-epitope alleles (NN = 0, SN = 1, SS = 2) and the HLA-DRB1 allele 1 and the HLA-DRB1 allele 2. The alleles that confer increased risk for rheumatoid arthritis include DRB1 0101, 0102, 0104, 0105, **0401, 0404, 0405, 0408, 0409**, 1001, 1402 and 1406, with highest risk alleles being bolded (Newton *et al.*, 2004). For our analyzes, a covariate denoted by DRB was defined by a joint codification of the HLA-DRB1 allele 1 and HLA-DRB1 allele 2, where scores 0, 1 and 2 were assumed for DRB genotypes according to the number of high risk alleles. For all individuals in this case-control study, information from 545,080 SNP-genotype from the Illumina 550K chip are also available. The genotypes are in the format X_X, where X is a base (A,C,G,T). Each record contains the following: SNP name, chromosome (0 = mitochondrial, 1 = chromosome 1, 2 = chromosome 2, ..., 22 = chromosome 22, 23 = X, 24 = Y, 25 = pseudo-autosomal), and SNP position in basepairs.

3 Statistical methods

Using a logistic regression approach (Hosmer e Lemeshow, 2000; Stokes et al, 2000), the procedure considered in our analysis consists of an unadjusted and an adjusted genome scan for SNPs additive effects upon the RA risk. First, the p-value

associated with the effect of each SNP is calculated according to the model without adjustment for covariates of known biological effect. Next, the corresponding p-value, now associated with the effect adjusted by the covariate (HLA-DRB), is calculated and compared with the former. Possible population-stratification influences on the SNP effects are also taken into account by means of a methodology based on principal components analysis (Price *et al.*, 2006).

Comparison of the significance level of the SNP effect obtained through these two different models may reveal a substantial change in the significance level of a particular SNP. In particular, changes in the AIC statistic as well as in the Wald statistic are evaluated to reveal SNPs that suggest an indirect influence with RA, not independently of previous knowledge; that are independently and suggest a direct influence; or that suggest an influence through epistatic mechanisms (or through effect modification). Next, each step of the analysis performed is described in more details.

3.1 Data set cleaning procedure

To remove potential genotype errors, we first identified the minor allele frequency (MAF) for each SNP and codified it as 1. SNP genotypes took the scores of 0, 1 and 2. We then conducted for all SNPs a Hardy-Weinberg equilibrium test calculation by using the Chi-square test available in the genetics library of the R package (R Development Core Team 2010). The significance level used in this step was 10^{-4}. SNPs showing evidence of disequilibrium and those with MAF lower than 1% were removed from the analysis. Any imputation procedure was considered to impute those few missing genotypes remained in the data set.

3.2 Analysis without adjustment by a covariate of known biological effect

The first model considered is based on a single locus genome scan described by a logistic model at the genotype level, expressed as:

$$\text{logit}[P(Y = 1|G_j)] = \beta_{0j} + \beta_{G_j} G_j \tag{1}$$

where Y is the RA affection status, G_j is a variable associated with the j-th SNP effect, and β_{0j} and β_{G_j} are regression coefficients. For the j-th SNP genotype, it is assumed that $G_j = 0$, 1, 2 corresponding to the number of alleles of lowest frequency. To test association between each SNP and the disease, the Wald statistic is used.

3.3 Analysis with adjustment by a covariate of known biological effect

The second model extends the first by including a covariate of known biological effect in the RA risk, and an interaction term. This model is given by:

$$\text{logit}[P(Y = 1|X, G_j)] = \beta_{0j} + \beta_X X + \beta_{G_j} G_j + \beta_{Ij} X * G_j \tag{2}$$

where X and G_j are covariates, β_{Ij} is the interaction effect between the j-th SNP and the covariate X, and β_X and β_{G_j} are regression coefficients. Considering the variables X and G_j, the SNP effect adjusted by the covariate X may be substantially different from the unadjusted effect, allowing for the classification of SNP effects into different classes. The parameter β_{Ij} is tested by means of a statistic that asymptotically has a chi-square distribution with the degree of freedom of 1. If there is evidence of an interaction effect between the SNP and the covariate X, it is possible to find SNPs not identified by model (1). Otherwise, a reduced model is fitted and it is given by:

$$\text{logit}[P(Y = 1|X, G_j)] = \beta_{0j} + \beta_X X + \beta_{G_j} G_j. \quad (3)$$

Comparisons between models (1) and (2) or (1) and (3) may be carried out computing differences in the Wald statistic (or in their p-values), as well as in the Akaike Information Criterion statistic (denoted as AIC). Changes in these statistics may reveal more consistent signals of the SNP effects than those obtained directly from model (1).

3.4 Correcting for population-stratification

In order to prevent possible population-stratification influences upon the SNP effects, a principal components analysis is performed following the methodology proposed by Price *et al.* (2006). According to this methodology, some axes of variation are extracted from the spectral decomposition of the covariance matrix of the individuals and are then used as covariates in models (1), (2) and (3). If k axes are used, models (1), (2) and (3) become:

$$\text{logit}[P(Y = 1|G_j)] = \beta_{0j} + \beta_{A_1} A_1 + ... + \beta_{A_k} A_k + \beta_{G_j} G_j \quad (4)$$

$$\text{logit}[P(Y = 1|X, G_j)] = \beta_{0j} + \beta_{A_1} A_1 + ... + \beta_{A_k} A_k + \beta_X X + \beta_{G_j} G_j + \beta_{Ij} X * G_j \quad (5)$$

$$\text{logit}[P(Y = 1|X, G_j)] = \beta_{0j} + \beta_{A_1} A_1 + ... + \beta_{A_k} A_k + \beta_X X + \beta_{G_j} G_j, \quad (6)$$

respectively, where A_m denotes the m-th axis selected ($m = 1, \ldots, k$) and β_{A_k} its corresponding effect. Although it is advisable to use a subset of SNPs to obtain these axes, we used all SNPs that passed the cleaning procedure, given that another study performed with this data provided similar results for different subsets of SNPs (Peloso *et al.*, 2008). An empirical evaluation of the cases and controls plotted against the first principal component is used to select the number of axes to be considered in the analysis. Only those axes showing a non-random pattern should be selected. The flowchart shown in Figure 1 summarizes the steps used in the statistical analyses performed.

4 Results

Using the data cleaning process described previously, a total of 43,616 SNPs were removed with 501,464 remaining for the analyses. From the spectral

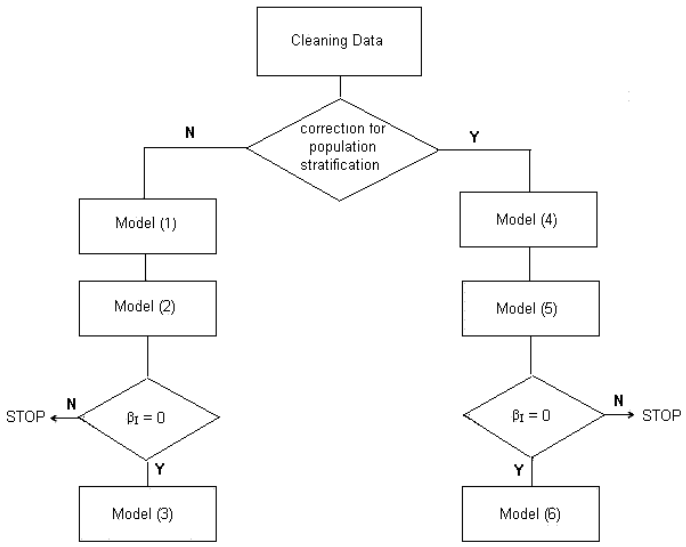

Figure 1 - Flowchart displaying the steps and models used in the analyses.

decomposition of the covariance matrix between individuals, we selected the first ten principal axes of variation to be used as covariates in models (4), (5) and (6), in order to prevent possible population-stratification influences. Figure 2 displays six of these ten axes where non-random patterns can be observed. DRB and SENum were used separately as biological covariates in models (2), (3), (5) and (6). As mentioned, the covariate DRB was defined by a joint codification of the HLA-DRB1 allele 1 and HLA-DRB1 allele 2, where scores 0, 1 and 2 were assumed for DRB genotypes according to the number of high risk alleles. Since the results that took into account these covariates were similar, only those for DRB are presented.

Analyses were performed with and without correction for population-stratification. Thus, model (1) was compared with model (2) or (3) and model (4) with model (5) or (6) regarding their AIC statistic profiles and AIC differences (denoted as ΔAIC). For models with correction, Figure 3 shows these statistics and differences for all SNPs in the 22 autosomal chromosomes. In general, a drop in the AIC values for the adjusted model ((5) or (6)) can be seen, with the highest changes concentrated in the HLA region of the chromosome 6.

Assuming 10^{-5} as the p-value cut-off (Duggal *et al.*, 2008), Tables 1 and 2 show the distribution of SNPs for models with and without correction for population-stratification, respectively. When correction is considered, 89 + 57 = 146 SNPs showed significant effects in model (4) but not in model (5) or (6), which is suggestive of an indirect influence with RA, depending on the biological covariate; from these, 117 are in the HLA region (Figure 4a). Furthermore, 20 SNPs had significant

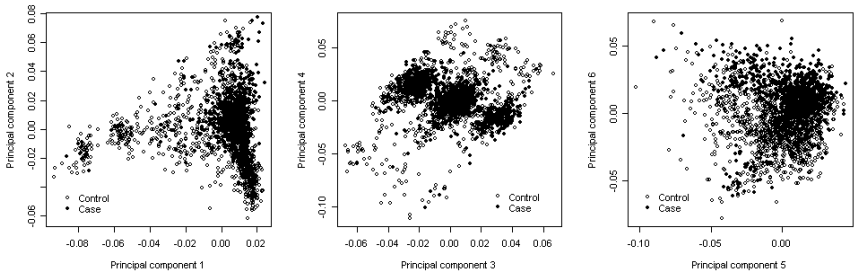

Figure 2 - First six axes of variation obtained from a principal components analysis performed with the 501,464 SNPs from the 22 autosomal chromosomes that passed the cleaning procedure.

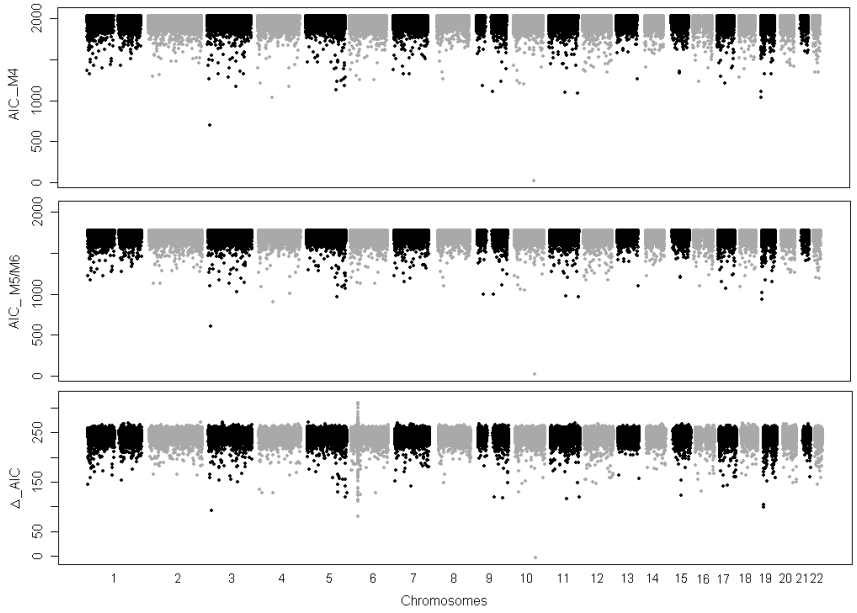

Figure 3 - AIC statistics for unadjusted and adjusted models under correction for population-stratification where M4 = model (4), M5 = model (5) and M6 = model (6).

effects in models (4) and (5), suggesting influence through epistatic mechanisms (or effect modification), and 71 in models (4) and (6), suggesting an independent direct influence. From these 91 SNPs, 34 are in the HLA region (Figure 4b). In addition, 9 SNPs were not significant in model (4) but significant in (5), also suggesting influence through epistatic mechanisms (or effect modification), and 36 were not significant in model (4) but significant in (6), suggesting independent direct influence with increased precision. From these 45 SNPs, 22 are in the HLA region (Figure 4c). Analogous results are shown in Table 2 regarding the analysis conducted without correction for population-stratification.

Table 1 - SNPs distribution for models with correction for population-stratification

	Model 5 (or 6)		
Model 4	$p \leq 10^{-5}$	$p > 10^{-5}$	Total
$p \leq 10^{-5}$	20 (71)	89 (57)	237
$p > 10^{-5}$	9 (36)	28466 (472716)	501227
Total	29 (107)	28555 (472773)	501464

Table 2 - SNPs distribution for models without correction for population-stratification

	Model 2 (or 3)		
Model 1	$p \leq 10^{-5}$	$p > 10^{-5}$	Total
$p \leq 10^{-5}$	26 (164)	160 (432)	782
$p > 10^{-5}$	8 (69)	28174 (472431)	500682
Total	34 (233)	28334 (472863)	501464

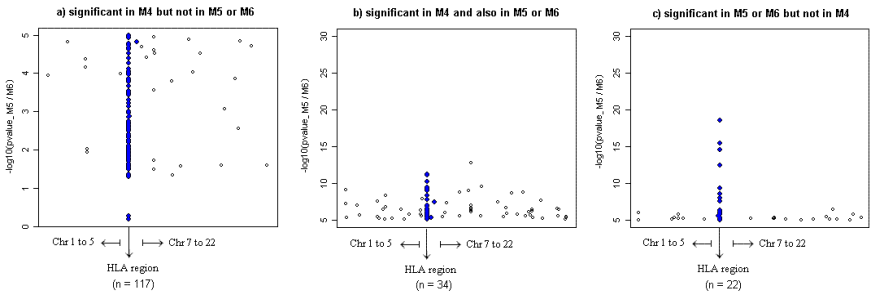

Figure 4 - SNPs position for unadjusted and adjusted association analysis under correction for population-stratification where M4 = model 4, M5 = model 5 and M6 = model 6.

As expected, a higher number of SNPs with potentially significant effects was observed in the analysis without correction for population-stratification, either under model (1) or under model (2) (or (3)). Considering the model without adjustment by a covariate of known biological effect, Figure 5a shows that 216 SNPs had significant effects in both analyses without and with correction whereas 21 SNPs had significant effects with correction for population-stratification but not without it. In addition, when the model with adjustment by a covariate of known biological effect is considered, Figure 5b shows 104 SNPs classified as significant either in the analysis conducted without or with correction for population-stratification.

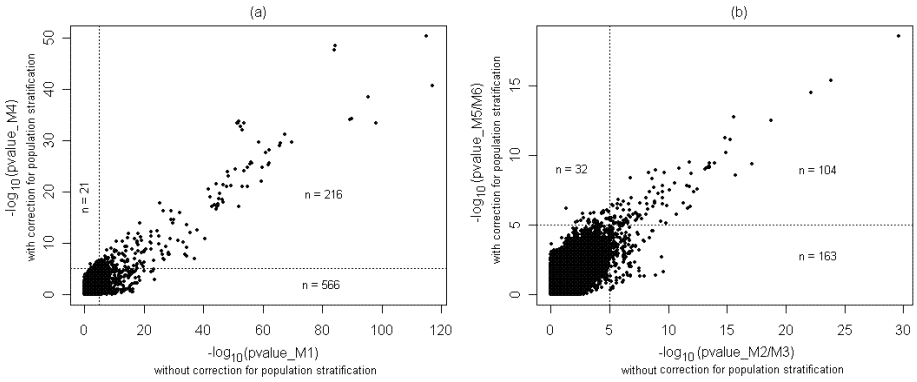

Figure 5 - P-values of SNP effects for models unadjusted and adjusted by a covariate of known biological effect without and with correction for population-stratification.

5 Conclusion and discussion

The Genome Wide Association Studies (GWAS) approach has become a widely used tool for mapping genes associated with complex diseases. Despite the benefits provided by this drastically increased genomic information, challenges are still at play regarding the control of both false positive and false negative results, due to the multiple tests nature of this approach, the population-stratification effects, and the possible associations among loci. In addition, the impact of higher-order effects is not being addressed in the current analytical approaches. Given this context, the main objective of our analyses was to find more consistent evidence of SNP effects upon the RA risk by comparing different models based on association analysis. By evaluating a single SNP effect without and with adjustment for covariates of known biological effect, it is possible not only to disclose potentially important higher-order effects, but also to provide a classification of SNPs that may be relevant in the decision process of validation and replication efforts.

In this paper, results for SNPs in all 22 autosomal chromosomes were presented. The analysis was performed taking into account the properties of the multiple regression analysis in terms of partial regression coefficients. From comparisons among the results of the unadjusted and adjusted models, and assuming a p-value cut-off of 10^{-5}, it was possible to classify SNP effects in terms of suggestive direct and indirect influence, as well as influence through epistatic mechanisms. Based on our results, we were able to find evidence of SNPs that were significant in the models without adjustment by a covariate of known biological effect, but not in the models with such adjustment, many of them being located in the HLA region, probably in LD with the DRB alleles. In addition, we were able to find SNPs with a non significant effect in the models without the adjustment mentioned, but a significant effect outside the HLA region in the models with the adjustment. In this particular case, the evidence for new SNPs is possibly due to interaction effects between SNPs and DRB alleles or to improved precision due to covariate adjustment. Additionally, several significant SNPs under both models (unadjusted and adjusted) were found outside of the HLA region.

The interpretation of the SNP effects on chromosome 6 deserves especial attention since DRB alleles are indeed variables genetically determined and are located in the human chromosome 6. Accordingly, SNP effects associated with RA and located near these loci may denote one or more of several options: LD between genetic markers also located in the HLA region, allelic association, or possible interaction effects. One important insight from our analytical approach is that it enables the discrimination of different classes of SNP effects according to the biological covariate contribution for the adjustment. Indeed, the classification of a SNP previously classified as significant as not significant can be extremely useful when one has to decide between hundreds of *associated SNPs* in a confirmatory study.

In the different possible scenarios, covariate adjustment may be interesting to disclose markers that suggest indirect influence, depending on the biological covariate (significant in model (4) but not in (5) or (6)); that suggest independent and direct influence (significant in models (4) and (6) or not significant in model (4) but significant in (6)); or that suggest influence through epistatic mechanisms or effect modification (significant in model (4) and model (5) or not significant in model (4) but significant in model (5)). Here, one should not limit the proposed approach to markers near the adjusted covariate, since this would limit the potential informativeness primarily to the first scenario. Furthermore, the search for SNPs associated with DRB through epistatic effects would be severely limited, since this effect is not restricted to closed loci.

Additionally, our results showed changes in the analyses with and without correction for population-stratification. However, we do not currently know whether the axes of variation selected represent ancestry measures or are informative about some other stratification pattern. Besides, our analysis assumed a significance level of 10^{-5} to correct for multiple loci tests, but we did not use any correction procedure for the different tests performed within each locus. Bonferroni's criterion could be

used for this purpose. Further insight into the classification of the SNP effect may result from tests of the LD (or allelic association) pattern with DRB or SENum, or by taking into account the inclusion of other available covariates, although this was not considered in our analysis. Nevertheless, our results are a remainder that just the simple inclusion of a covariate with known biological effect can significantly alter overall results. This information may be useful for more efficient planning of follow-up studies of potentially significant markers within the context of a complex phenotype.

Acknowledgments

This work is based on data that was gathered with the suppport of grants from the National Institutes of Health (NO1-AR-2-2263 and RO1-AR-44422), and the National Arthritis Foundation. The Genetic Analysis Workshop was supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. Authors were partially supported by FAPESP (SP, Brazil) and CNPq (Brazil, Grant 150653/2008-5).

GIOLO, S. R.; SOLER, J. M. P.; BATISTA, M. J.; ALMEIDA, M. A. A.; PEREIRA, A. C. Evidências do efeito de SNP sobre o risco de artrite reumatóide: Efeitos do ajuste de covariáveis nos resultados de associação. *Rev. Bras. Biom.*, São Paulo, v.29, n.1, p.47-59, 2011.

- RESUMO: Nesse artigo, uma abordagem de análise de associação por meio do ajuste de covariáveis é proposta para classificar o efeito de polimorfismos de um único nucleotídeo (SNP) associados com o risco de artrite reumatóide (RA). Inicialmente, o efeito marginal de cada SNP é avaliado considerando-se uma regressão logística para cada loco. Esse efeito é, então, avaliado ajustado por uma covariável de efeito biológico conhecido (alelos HLA-DRB) sobre o risco de RA. Para levar em consideração possíveis influências de estratificação da população, os primeiros dez eixos de variação, resultantes da análise de componentes principais dos SNPs, são incorporados na análise como covariáveis. Para fins de comparação, análise sem esses eixos também é realizada. SNPs de todos os cromossomos autossômicos dos dados de artrite reumatóide do Genetic Analysis Workshop 16 são utilizados nas análises. A partir das comparações realizadas com respeito aos efeitos dos SNPs obtidos sem e com o ajustamento para a covariável de efeito biológico conhecido, classificação desses efeitos é sugerida em termos de influência direta e indireta, bem como influência devido à epistasia. O procedimento analítico proposto mostrou ser útil não somente para mostrar efeitos de ordem superior potencialmente importantes, mas também para fornecer uma classificação dos SNPs.

- PALAVRAS-CHAVE: Estudos de associação; genética humana; regressão logística; marcadores moleculares; estratificação da população.

References

BEGOVICH, A. B.; CARLTON, V. E. H.; HONIGBERG, L. A.; SCHRODI, S. J.; CHONOKKALINGAM, A. P.; ALEXANDER, H. C.; ARDLIE, K. G.; HUANAG, Q.; SMITH, A. M.; SPOERKE, J. M.; CONN, M. T.; CHANG, M.; CHANG, S-YP.; SAIKI, R. K.; CATANESE, J. J.; LEONG, D.; GARCIA, V. E.; McALLISTER, L. B.; JEFFERY, D. A.; LEE, A. T.; BATLIWALLA, F.; REMMERS, E.; CRISWELL, L. A.; SELDIN, M. F.; KASTER, D. L.; AMOS, C. I.; SNINSKY, J. J.; GREGERSEN, P. K. A missense SNP in the protein tyrosine phosphatase PTPN22 is associated with rheumatoid arthritis. *Am. J. Hum. Genet.*, v.75, p.330-337, 2004.

du MONTCEL, S. T.; MICHOU, L.; PETIT-TEIXEIRA, E.; OSORIO, J.; LEMAIRE, I.; LASBLEIZ, S.; PIERLOT, C.; QUILLET, P.; BARDIN, T., PRUM, B.; CORNELIS, F.; CLERGET-DARPOUX, F. New classification of HLA-DRB1 alleles supports the shared epitope hypothesis of rheumatoid arthritis susceptibility. *Arthritis Rheum.*, Atlanta, v.52, p.1063-1068, 2005.

DUGGAL, P.; GILANDERS, E.; HOLMES, T. N.; BILEY-WILSON, J. E. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics*, London, v.9, p.516, 2008.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression.* New York: John Wiley & Sons, 2000. 390p.

IRIGOYEN, P.; LEE, A. T.; WENER, M. H.; LI, W.; KERN, M.; BATLIWALLA, F.; LUM, R. F.; MASSAROTTI, E.; WEISMAN, M.; BOMBARDIER, C. Regulation of anti-cyclic citrullinated peptide antibodies in rheumatoid arthritis: contrasting effects of HLA-DR3 and the shared epitope alleles. *Arthritis Rheum.*, Atlanta, v.52, p.3813-3818, 2005.

NEWTON, J. L.; HARNEY, S. M.; WORDSWORTH, B. P.; BROWN, M. A. A review of the MHC genetics of rheumatoid arthritis. *Genes Immun.*, London, v.5, n.3, p.151-157, 2004.

PELOSO, G. M.; TIMOFEEV, N.; LUNETT, K. L. Detecting population stratification with different subsets of SNPs and subsequent adjustment in the NARAC data. In: GENETICS ANALYSIS WORKSHOP, 16., 2008, St Louis. *Proceedings...* USA: St Louis, 2008. p.16-20.

PLENGE, R. M.; PADYUKOV, L.; REMMERS, E. F.; PURCELL, S.; LEE, A. T.; KARLSON, E. W.; WOLFE, F.; KASTNER, D. L.; ALFREDSSON, L.; ALTSHULER, D. Replication of putative candidate-gene association with rheumatoid arthritis in 4,000 samples from North American and Sweden: association of susceptibility with PTPN22, CTLA4 and PADI4. *Am. J. Hum. Genet.*, Cambridge, v.77, n.6, p.1044-1060, 2005.

PRICE, A. L.; PATTERSON, N. J.; PLENGE, R. M.; WEINBLATT, M. E.; SHADICK, N. A.; REICH, D. Principal components analysis corrects for

stratification in genome-wide association studies. *Nat. Genet.*, New York, v.38, p.904-909, 2006.

R Development Core Team. R: A language and environment for statistical computing. R: Foundation for Statistical Computing. Vienna, Austria, ISBN 3-900051-07-0, 2010. <http://www.R-project.org>.

STOKES, M. E.; DAVIS, C. S.; KOCH, G. G. *Categorical data analysis using the SAS system*. 2.ed. Cary, USA: SAS Institute Inc., 2000. 623p.

ZIEGLER, A.; KONIG, I. R. *A statistical approach to genetic epidemiology*. Weinheim: Wiley-VCH, 2006. 350p.

Received in 01.11.2010.

Approved after revised in 08.04.2011.