BMC Proceedings

Genome-wide association study of rheumatoid arthritis by a score test based on wavelet transformation

Renfang Jiang, Jianping Dong* and Yilin Dai

Address: Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA

E-mail: Renfang Jiang - rjiang@mtu.edu; Jianping Dong* - jdong@mtu.edu; Yilin Dai - ydai@mtu.edu
*Corresponding author

Abstract

Background: We have conducted a genome-wide association study on the Genetic Analysis Workshop (GAW) 16 rheumatoid arthritis data using a multilocus score test based on wavelet transform proposed recently by the authors. The wavelet-based test automatically adjusts for the amount of noise suppressed from the data. The power of the test is also increased by using the genetic information contained in the spatial ordering of single-nucleotide polymorphisms on a chromosome.

Results: After adjusting for the effect of population stratification, the test identified some previously discovered rheumatoid arthritis susceptibility loci (*HLA-DRB1* and rs3761847) as well as some loci (rs2076530 and rs3130340) known to have association with sarcoidosis and bone mineral density. It was previously reported that patients with rheumatoid arthritis have elevated prevalence of sarcoidosis and have reduced bone mass.

Conclusion: This new test provides a useful tool in genome-wide association studies.

Background

In genome-wide association studies, the first choice of tests is usually a single-marker test. If a single-nucleotide polymorphism (SNP) has a strong association with a disease, single-marker tests should have higher power than multilocus tests. Multilocus tests can achieve higher power if several SNPs are associated with the disease. However, the potential high power of multilocus tests could be diminished as an increased number of markers results in an increased number of degrees of freedom.

Therefore, reducing the number of degrees of freedom is essential to increasing the power of multilocus tests. Different strategies were introduced to reduce the number of degrees of freedom. Tests based on haplotype sharing (the longest continuous interval of matching alleles between haplotypes) effectively reduce the number of degrees of freedom [1,2]. Another popular approach is to apply principal components analysis (PCReg) [3]. A score test based on the Fourier transform [4] is another attempt to reduce the degree of freedom,

thereby increasing power. Instead of using the genotype data, weighted Fourier transform coefficients of the genotype data are used to form a score test.

Many multilocus association tests are not affected by permuting spatial order of SNPs; thus, they do not use the information contained in the ordering of SNPs. For example, the results of logistic regression will not change if the order of SNPs are permuted. The same is true for the test obtained by fitting a regression function with one SNP followed by Bonferroni correction to find the global *p*-value. PCReg is also invariant under the permutation of SNPs. We provide a proof of the above claims as follows (See Wang and Abbott [3] for notations in the following proof.). Let G be a matrix of coded genotypes and suppose it is centered. The variance-covariance matrix of genotypes is $A = G^{T}G/(n - 1) = VDV^{-1}$, where D is a diagonal matrix with eigenvalues of A as diagonal entries, and the columns of V are the eigenvectors. The regression model is $y = GV_1b + \varepsilon$, where V_1 is the first several columns of V. After permutation of SNPs, the columns of G are also permuted, and G becomes GP, where P is a permutation matrix. Since $(GP)^{T} (GP) = (P^{T}V) D(P^{T}V)^{-1}$, the eigenvalues are not changed by the permutation, and the matrix of eigenvectors becomes $P^{T}V$. The regression model $y = GPP^{T}V_1b + \varepsilon = GV_1b + \varepsilon$ remains unchanged. For the tests based on shared haplotype length, changing the ordering of SNPs could mean a shortened shared length, and the association between the disease and the SNPs could disappear. Therefore, the ordering of SNPs contains important genetic information, and ignoring it could lower the power of association tests.

We recently proposed a score test based on wavelet transform [5,6], which is used in this report. This test has three advantages. First, it uses the wavelet transformation of genotype data instead of the genotype data. The wavelet transform is designed to handle unsmooth noisy signals. Genetic data are usually unsmooth and can be dealt with by the wavelet transform naturally. Second, it uses an empirical Bayes thresholding [7]. It was proved by Johnston and Silverman that this thresholding effectively suppresses noise from data [7]. Therefore, it increases the power of the test. The exact amount of noise being reduced depends on the specific problem. Third, because our method views multilocus genotypes of an individual as a discretized function, the spatial ordering of the SNPs on a chromosome is taken into consideration, which increases the power of the test.

Methods
There is a total of 2,062 individuals consisting of 868 cases and 1,194 controls in the North American Rheumatoid Arthritis Consortium (NARAC) data for Genetic Analysis Workshop (GAW) 16. These individuals were genotyped on the 550 k Illumina SNP chip. We analyzed 22 autosomal chromosomes in this report. SNPs satisfying one of the following criterion were excluded: missing genotype rate >0.05, or minor allele frequency < 0.05, or having *p*-values < 0.00001 in the Hardy-Weinberg equilibrium test. About 12.8% of SNPs were removed. Missing genotypes were imputed by fastPHASE [8]. We adopted a strategy of a moving window of eight SNPs without overlapping (the first window contains SNPs 1-8, the next window contains SNPs 9-16, etc.) on the chromosome. We have tried windows of sizes 16 and 32 SNPs and there was not much difference. Overlapping windows were also tried and again, the results were similar. Computational burden increases when large windows or overlapping windows were used. The kth window is from the $(8(k -1) + 1)$th SNP to the $8k$th SNPs. Let G_k be the corresponding eight columns of the genotype matrix G. Different ways of coding SNPs do make a difference. The SNPs in G_k were recoded to maximize the number of positive pairwise correlations. This removes any ambiguity in the coding of genotypes. In our simulation studies, the recoding increased the power of the test while keeping the type I error rate in check.

We applied EIGENSTRAT [9] to remove the effects of population stratification. The top three eigenvectors obtained by EIGENSTRAT were used to adjust genotypes and phenotypes as suggested by Price et al. [9]. Let G be the genotype matrix, C be the matrix of the principal component vectors, and Y be the vector of phenotypes. The adjusted genotype matrix is $\tilde{G} = G - CC^{T}G$ and the adjusted phenotype vector is $\tilde{Y} = Y - CC^{T}Y$. Before the adjustment, the inflation factor [10] $\lambda_{GC} = 1.399$, which reduced to 1.025 after the adjustment. To avoid a possible bias, all chromosomes except 6 and 8 were included to produce the eigenvectors, which were used to adjust genotypes on chromosomes other than 6. Chromosome 6 was used to produce eigenvectors to adjust genotypes on chromosome 6.

Consider the kth window and the corresponding matrix of adjusted genotype $\tilde{G}_k$. Applying wavelet transform, thresholding, and inverse wavelet transform consecutively on the ith row of $\tilde{G}_k$ resulted in the modified genotype $(x_{i1}, x_{i2}, ..., x_{im})$ of the ith individual. Subtract the mean of the jth column of $X = (x_{ij})$ from x_{ij} such that the mean of each column of $X = 0$. Let $\tilde{Y} = (y_1, y_2, ..., y_n)$ be the adjusted phenotype of n individuals. Assume a generalized linear model [11] $f(E(y_i)) = \alpha + X_i\beta$ between X and Y, where f is a link function. Let $U_j = \sum_{i=1}^{n} y_i x_{ij}$ be the score statistic for the markers. The variance of U_j

under the null hypothesis can be estimated by $V_j = (1/(n-1)) \sum \left(\gamma_i - \overline{Y} \right)^2 \sum x_{ij}^2$ The global score statistic of this window is $T = \sum U_j / \sqrt{\sum V_j}$.

The global *p*-values of the wavelet-based test were obtained as follows. We randomly permuted the phenotypes (case-control status) 5,000 times, and received 5,000 sets of permuted phenotypes. Adjust each set with the principal component vectors, and they are still called permuted phenotypes. At each window of eight SNPs, the value of *T* was calculated 5,001 times, using the adjusted phenotypes and 5,000 sets of permuted phenotypes, respectively. In *M* windows along the chromosomes, we have 5000 *M* values of *T* with permuted phenotypes, denoted by a 5,000 × *M* matrix (T_{ij}), where T_{ij} is the absolute value of *T* at the *j*th window using the *i*th set of permuted phenotypes. Let $m_i = \max_j T_{ij}$ be the maximum absolute value of *T* at *M* windows using the *i*th set of permuted phenotypes. Let T_j be the absolute value of *T* at the *j*th window using the adjusted phenotypes. The global *p*-value of the wavelet-based test at the *j*th window is the proportion of $m_i > T_j$: global *p*-value of $T_j = \#\{m_i | m_i > T_j\}/5,000$.

Results and discussion

After correcting for population stratification, significant signals were only found on chromosomes 6 and 9. Four windows on chromosome 6 attracted our attention. The first window (rs9268005, rs3130340, rs3115553, rs9268132, rs926070, rs6935269, rs7775397, rs17422797) contains rs3130340, which was identified to have association with bone mineral density and fractures [12]. It has been reported that a large proportion of men with rheumatoid arthritis (RA) had reduced bone mass [13]. The *p*-value at this window is 0.0074. The second window (rs4424066, rs12529049, rs3117099, rs3117098, rs3817973, rs1980493, rs2076530, rs4248166) contains rs2076530, which is associated with sarcoidosis [14]. Increased prevalence of sarcoidosis among RA patients has been reported [15]. The association between sarcoidosis, RA, and rs2076530 is an interesting phenomenon. The association between rs2076530 and RA was investigated before, but whether it is a causal SNP for RA or its effect is merely a carryover effect of nearby haplotypes [16] merits further investigation. The third window (rs2395182, rs3129890, rs9268832, rs6903608, rs2395185, rs477515, rs2516049, rs2858870) contains *HLA-DRB1*, which has long been identified as a major genetic risk contributor to RA [17]. The last window (rs9275224, rs5000634, rs6457617, rs2647012, rs9357152, rs10484561, rs9275313, rs1794282) contains rs6457617, which has also been reported as being associated with RA [18]. The

p-values of the test at the above three windows are less than 0.0002 (it is 0 after 5,000 permutations). The most significant window on chromosome 9 (rs1953126, rs10985073, rs3761847, rs10985095, rs10985097, rs2900180, rs12235400, rs10985112) which contains rs3761847 and it was identified as a risk locus for RA [19], has a *p*-value 0.53.

We applied the wavelet-based test on a moving window of eight SNPs with overlapping (the first window contains SNPs 1-8, the second window contains SNPs 2-9, etc.) on a 550-kb region of chromosome 6 for fine mapping. The results are shown in Figure 1. We have tried other window sizes (16 and 32) and found similar results (not shown). Comparisons of the wavelet-based test using overlapping and non-overlapping windows showed no significant differences. In Figure 2, we compared the wavelet-based test with Armitage χ^2 test. If *p*-value of the wavelet-based test is 0 after 5,000 permutations, we set the *p*-value as 0.0002 in order to calculate log(*p*-value) in Figures 1 and 2, which makes 0.0002 the smallest possible *p*-value in this study. Some significant loci (*HLA-DRB1*, rs2076530, rs6457617) were identified by both tests. However, the Armitage χ^2 test was not significant around rs3130340 (*p*-value after Bonferroni correction was 1), while the *p*-value of the wavelet-based test at the window containing the SNP was 0.0074, after correcting for multiple testing.

Conclusion

A wavelet-based multilocus score test was applied in a genome-wide association study on RA data followed by fine mapping of regions identified in our genome-wide association study. Several statistically significant risk loci for RA were identified after adjustment for population stratification. Some windows contain genes and/or SNPs (*HLA-DRB1*, rs3761847, and rs6457617) previously associated with RA. Some windows contain SNPs (rs2076530, rs3130340) previously associated with sarcoidosis [14] or bone mineral density and fractures [12], all of which have been reported as being related with RA [13,15].

List of abbreviations used

NARAC: North American Rheumatoid Arthritis Consortium; PCReg: Principal components analysis; RA: Rheumatoid arthritis; SNP: Single-nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.
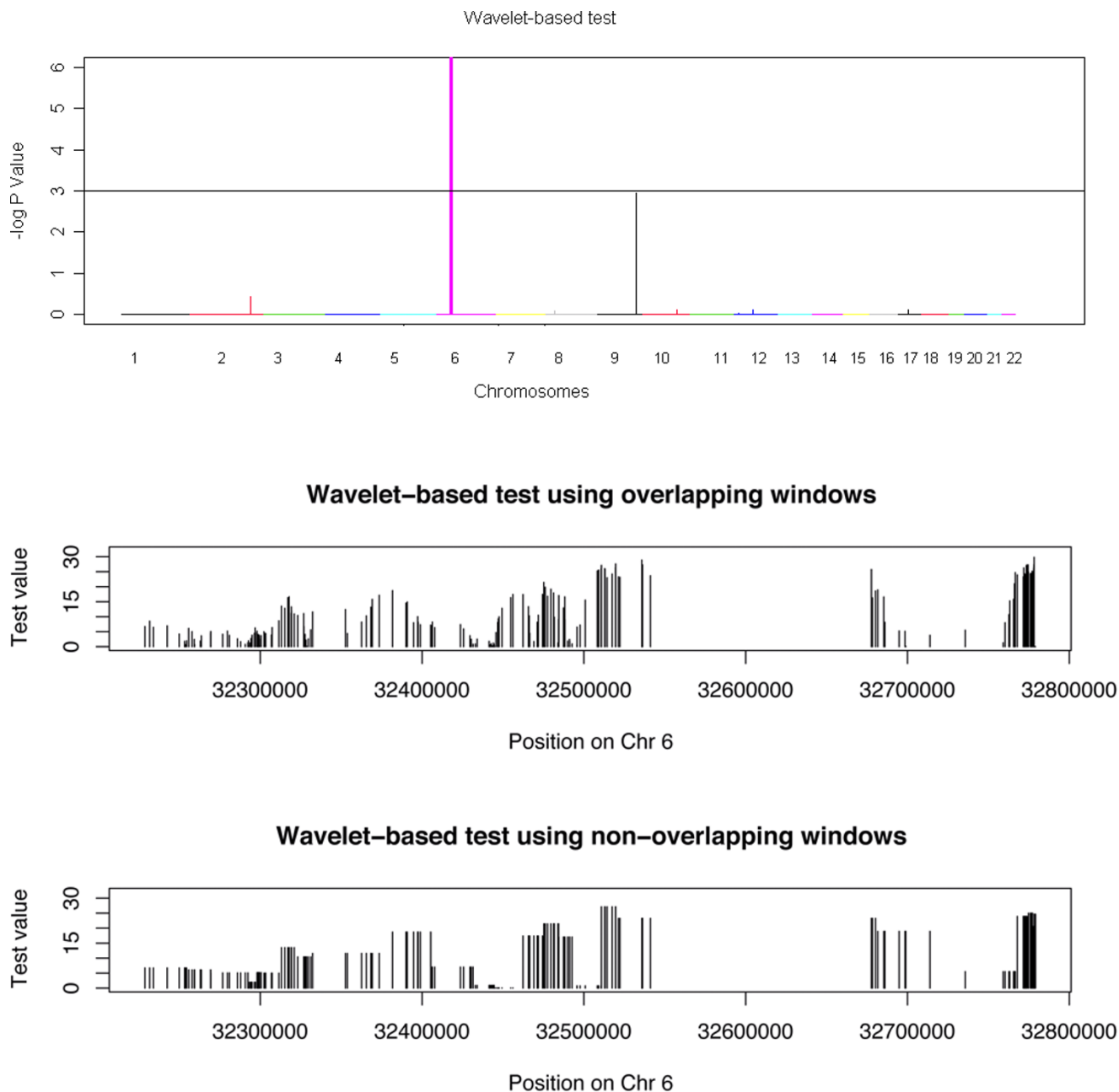
Figure 1
_p_-Values on 22 chromosomes, and comparison of windows. The plot on the left contains the _p_-values of
the wavelet-based test on 22 chromosomes. The horizontal line indicates 5% significance level. The plots on the right
are a comparison of overlapping windows (top) and non-overlapping windows (bottom) for a 550-kb region
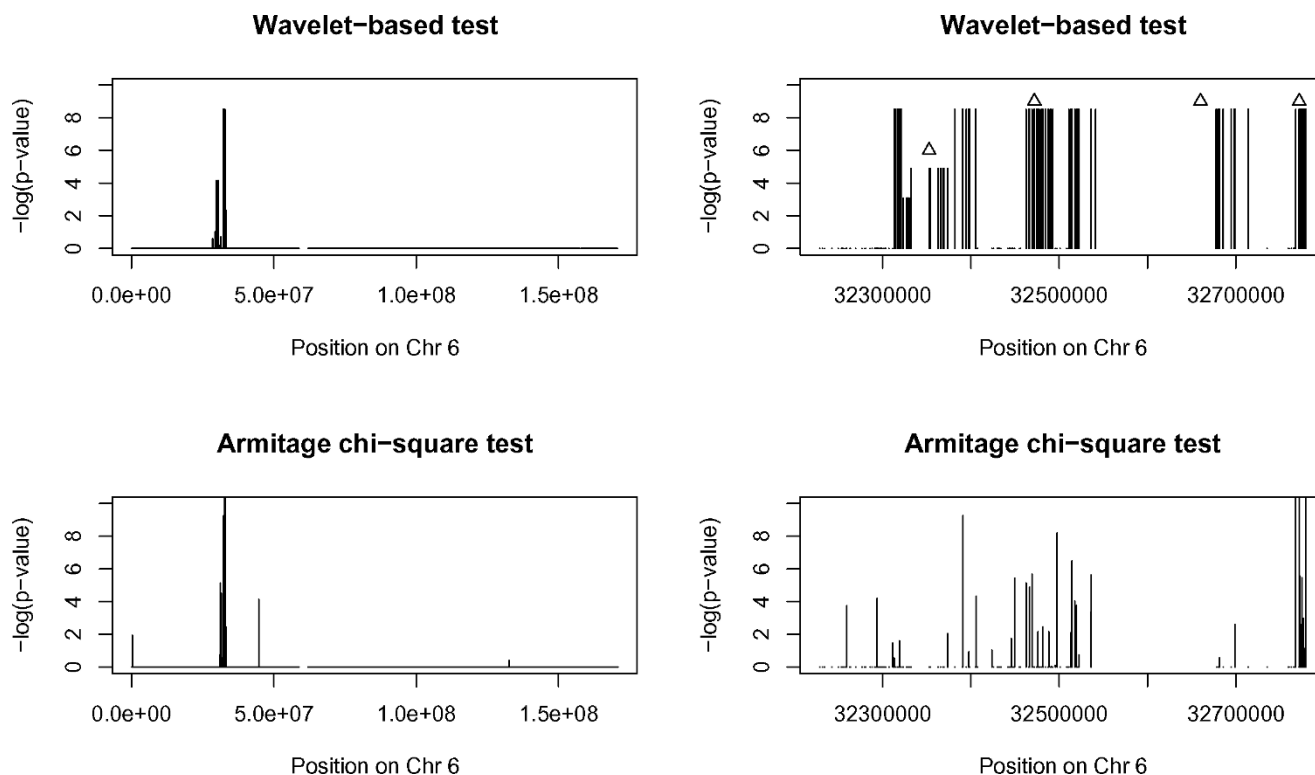on chromosome 6.

Authors' contributions

RJ and JD both contributed in development of the statistical
test, provided simulation strategies, and drafted the manu-
script. RJ also participated and guided the numerical
calculations. YD carried out part of the programming
work. All authors read and approved the manuscript.

Acknowledgements

Figure 2
Comparison with single marker test. Comparisons of the wavelet-based test and the Armitage χ^2 test on chromosome 6. The plots on the left are for the whole chromosome 6, and the plots on the right are fine mapping results for a 550-kb region of chromosome 6. The triangles indicate the positions (from left to right) of rs3130340 (associated with bone mineral density and RA), rs2076530 (associated with sarcoidosis and RA), HLA-DRB1 (Note: no data from 32.54 Mb to 32.67 Mb), and rs6457617 (associated with RA).

the supplement are available online at http://www.biomedcentral.com/1753-6561/3?issue=S7.

References

1. Tzeng JY, Devlin B, Roeder K and Wasserman I: **On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit.** *Am J Hum Genet* 2003, **72:**891–902.
2. Zhang S, Sha Q, Chen HS, Dong J and Jiang R: **Transmission/disequilibrium test based on haplotype sharing for tightly linked markers.** *Am J Hum Genet* 2003, **73:**566–579.
3. Wang K and Abbott D: **A principal components regression approach to multilocus genetic association studies.** *Genet Epidemiol* 2008, **32:**108–118.
4. Wang T and Elston RC: **Improved power by use of a weighted score test for linkage disequilibrium mapping.** *Am J Hum Genet* 2007, **80:**353–360.
5. Jiang R, Dong J and Dai Y: **A wavelet based method in association [abstract 96].** *Genet Epidemiol* 2008, **32:**697.
6. **Improving Power in Genetic-association Studies Via Wavelet Transformation.** *BMC Genetics* 2009, **10:**53.
7. Johnston IM and Silverman BW: **Empirical Bayes selection of wavelet thresholds.** *Ann Stat* 2005, **33:**1700–1752.
8. Scheet P and Stephens M: **A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.** *Am J Hum Genet* 2006, **78:**629–644.
9. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA and Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38:**904–909.
10. Devlin B and Roeder K: **Genomic control for association studies.** *Biometrics* 1999, **55:**997–1004.
11. Schaid DJ, Rowland CM, Tines DE, Jacbson RM and Poland GA: **Score tests for association between traits and haplotypes when linkage is ambiguous.** *Am J Hum Genet* 2002, **70:**425–434.
12. Styrkarsdottir U, Halldorsson BV, Gretarsdottir S, Gudbjartsson DF, Walters GB, Ingvarsson T, Jonsdottir T, Saemundsdottir J, Center JR, Nguyen TV, Bagger Y, Gulcher JR, Eisman JA, Christiansen C, Sigurdsson G, Kong A, Thorsteinsdottir U and Stefansson K: **Multiple genetic loci for bone mineral density and fractures.** *N Engl J Med* 2008, **358:**2355–2365.
13. Tengstrand B and Hafstrom I: **Bone mineral density in men with rheumatoid arthritis is associated with erosive disease and sulfasalazine treatment but not with sex hormones.** *J Rheumatol* 2002, **29:**2299–2305.
14. Valentonyte R, Hampe J, Huse K, Rosenstiel P, Albrecht M, Stenzel A, Nagy M, Gaede KI, Franke A, Haesler R, Koch A, Lengauer T, Seegert D, Reiling N, Ehlers S, Schwinger E, Platzer M, Krawczak M, Muller-Quernheim J, Schurmann M and Schreiber S: **Sarcoidosis is associated with a truncating spliice site mutation in BTNL2.** *Nat Genet* 2005, **37:**357–364.
15. Kucera RF: **A possible association of rheumatoid arthritis and sarcoidosis.** *Chest* 1989, **95:**604–606.
16. Orozco G, Eerligh P, Sánchez E, Zhernakova S, Roep BO, González-Gay MA, López-Nevot MA, Callejas JL, Hidalgo C, Pascual-Salcedo D, Balsa A, González-Escribano MF, Koeleman BP and Martín J: **Analysis of a functional BTNL2 polymorphism in type 1 diabetes, rheumatoid arthritis, and systemic lupus erythematosus.** *Hum Immunol* 2005, **66:**1235–1241.
17. Stastny P: **Association of the B-cell alloantigen DRw4 with rheumatoid arthritis.** *N Engl J Med* 1978, **298:**869–871.

18. Julià A, Ballina J, Cañete JD, Balsa A, Tornero-Molina J, Naranjo A, Alperi-López M, Erra A, Pascual-Salcedo D, Barceló P, Camps J and Marsal S: **Genome-wide association study rheumatoid arthritis in the Spanish population: KLF12 as a risk locus for rheumatoid arthritis susceptibility.** *Arthritis Rheum* 2008, **58:**2275–2286.
19. 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.