BMC Proceedings

Proceedings

# Case-control genome-wide association study of rheumatoid arthritis from Genetic Analysis Workshop 16 using penalized orthogonal-components regression-linear discriminant analysis

Min Zhang[1], Yanzhu Lin[1], Libo Wang[1], Vitara Pungpapong[1], James C Fleet[2] and Dabao Zhang*[1]

Addresses: [1]Department of Statistics, Purdue University, 150 North University Street, West Lafayette, IN 47907, USA and [2]Department of Foods and Nutrition, Purdue University, 700 West State Street, West Lafayette, IN 47907, USA

E-mail: Min Zhang - minzhang@purdue.edu; Yanzhu Lin - lin43@purdue.edu; Libo Wang - wang220@purdue.edu;
Vitara Pungpapong - vpungpap@purdue.edu; James C Fleet - fleet@purdue.edu; Dabao Zhang* - zhangdb@purdue.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):S17 doi: 10.1186/1753-6561-3-S7-S17

This article is available from: http://www.biomedcentral.com/1753-6561/3/S7/S17

## Abstract

Currently, genome-wide association studies (GWAS) are conducted by collecting a massive number of SNPs (i.e., large *p*) for a relatively small number of individuals (i.e., small *n*) and associations are made between clinical phenotypes and genetic variation one single-nucleotide polymorphism (SNP) at a time. Univariate association approaches like this ignore the linkage disequilibrium between SNPs in regions of low recombination. This results in a low reliability of candidate gene identification. Here we propose to improve the case-control GWAS approach by implementing linear discriminant analysis (LDA) through a penalized orthogonal-components regression (POCRE), a newly developed variable selection method for large *p* small *n* data. The proposed POCRE-LDA method was applied to the Genetic Analysis Workshop 16 case-control data for rheumatoid arthritis (RA). In addition to the two regions on chromosomes 6 and 9 previously associated with RA by GWAS, we identified SNPs on chromosomes 10 and 18 as potential candidates for further investigation.

## Background

Genome-wide association studies (GWAS) are challenged by the "curse of dimensionality", i.e., a large number of single-nucleotide polymorphisms (SNPs) are genotyped (i.e., large *p*) from a small number of biological samples (i.e., small *n*). Because of this, in practice, only one SNP is evaluated for association at a time [1]. However, such univariate approaches ignore the high correlation between SNPs in certain regions of the genome due to linkage disequilibrium (LD) [2]. Recently, Zhang et al. [3] developed a penalized orthogonal-components regression (POCRE) method for efficiently selecting variables in large *p* small *n* settings. Here we propose to implement linear discriminant analysis (LDA) combined with POCRE, and apply the so-called POCRE-LDA to a case-control GWAS dataset.

## Methods

### *POCRE*

POCRE works well to fit a large $p$ small $n$ regression model [3],

$$\mathbf{Y} = \mu + \sum_{j=1}^{p} \beta_j \mathbf{X}_j + \varepsilon, \tag{1}$$

where the sample ($\mathbf{Y}$, $\mathbf{X_1}$,..., $\mathbf{X_p}$) is of size $n$. Let $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_p^T)$, and further assume both $\mathbf{Y}$ and $\mathbf{X}$ are centralized ($\mu$ = 0 in the above model). Starting with $\tilde{\mathbf{X}}_1 = \mathbf{X}$, POCRE sequentially constructs components $\tilde{\mathbf{X}}_k \omega_k$ such that $\tilde{X}_k$ is orthogonal to $\{\tilde{\mathbf{X}}_i \omega_i, \ldots, \tilde{\mathbf{X}}_{k-1} \omega_{k-1}\}$, and the loading $\omega_k$ = $\gamma/||\gamma||$ with $\gamma$ minimizing

$$-2\gamma^T \tilde{\mathbf{X}}_k^T \mathbf{Y}\mathbf{Y}^T \tilde{\mathbf{X}}_k \alpha + ||\gamma||^2 + g_\lambda(\gamma), \quad \text{subject to } ||\alpha|| = 1. \tag{2}$$

Here $g_\lambda(\gamma)$, is a penalty function with tuning parameter $\lambda$, which Zhang et al. [3] implemented with empirical Bayes thresholding methods proposed by Johnstone and Silverman [4]. Such implementation introduces a proper regularization on $\gamma$, and provides adaptively sparse loadings of orthogonal components.

When the optimal $\gamma$ solving Eq. (2) is zero, we stop the sequential construction because the constructed orthogonal components $\{\tilde{\mathbf{X}}_1 \omega_1, \tilde{\mathbf{X}}_2 \omega_2, \ldots\}$ account for almost all contributions of $\mathbf{X}$ to the variation in $\mathbf{Y}$. An estimate of $\beta_1$,..., $\beta_p$ in Eq. (1) can be derived by regressing $\mathbf{Y}$on these orthogonal components. Resultant estimates of $\beta_1$,..., $\beta_p$ are mostly zero due to the sparse loadings in $\omega_j$, $j$ = 1, 2,.... This algorithm is computationally efficient as it only involves constructing penalized leading principal components.

### *POCRE-LDA*

POCRE can efficiently construct orthogonal components by excluding insignificant SNPs, and therefore simultaneously identify significant SNPs for GWAS [5]. In a case-control GWAS, we can define the response variable using the group membership, i.e., $y_i$ = 1 if individual $i$ is from the case population, and $y_i$ = -1 otherwise. Then, regressing $\mathbf{Y} = (y_1, \ldots, y_n)^T$ on $\mathbf{X}$ using POCRE implements LDA with threshold $c$ = 0. Indeed, the resultant $\sum_{j=1}^{p} b_j \mathbf{X_j}$ is a penalized version of Fisher's LDA direction [6], with $b_j$ estimating $\beta_j$. We therefore call it POCRE-LDA, with the tuning parameter $\lambda$ elicited by employing a 10-fold cross-validation and considering candidates $\lambda \in \{0.8, 0.82, 0.84, 0.86, 0.88, 0.9, 0.92, 0.94, 0.96, 0.98, 1\}$.

We applied POCRE-LDA to the rheumatoid arthritis (RA) case-control data in Genetic Analysis Workshop (GAW) 16. Of the 545,080 SNPs, 490,613 (90.2%) SNPs and all 2,062 individuals (868 cases and 1,194 controls) were kept for our analysis after using PLINK [7] to preprocess the data and control the data quality. To control the underlying population structure, EIGENSTRAT [8] was used to derive the first 10 principal components of the genome-wide genotype data. Then POCRE-LDA was applied separately to each chromosome. The effects of the 10 principal components constructed by EIGENSTRAT were controlled, where, for each chromosome, only the first several principal components were identified to be associated with the case/control status (results not shown).

## Results

The results of our analysis are shown in Figure 1, where the estimated effect size of each SNP is plotted against the physical location of the SNP. Several clusters of nonzero effects appear on chromosomes 6, 9, 10, and 18. The cluster on chromosome 6 covers a wide genomic region ranging from 6p22.1 to 6p21.32 and includes many genes related to the immune system. For example, this region contains the human leukocyte antigen (HLA) genes that encode the major histocompatibility complex (MHC) proteins necessary for antigen presentation and the *TAP2* gene that encodes a membrane-associated ATP-binding cassette peptide transporter necessary for delivering antigens to MHC class I proteins. Because there are many SNPs with nonzero effects in each of these clusters, Table 1 reports only the most significant SNPs within each region. The gene information corresponding to this genetic location was obtained from the Ensemble database http://www.ensembl.org. Several of the genes on chromosome 6 listed in Table 1 have previously been shown to be associated with RA, i.e., *MICB* [9], *BAT1* [9], *TAP2* [10,11], and *BTNL2* [12]. The most significant SNP on chromosome 9 (rs2900180), together with another significant SNP in that region (rs3761847), are in LD with the TNF receptor associated factor 1 (*TRAF1*) gene as well as the *C5* gene. Polymorphisms in these two genes were previously associated with RA [13,14]. Our results suggest weak evidence of association for SNPs on chromosomes 10 and 18 with RA (i.e., only a few SNPs with nonzero effects occur there). Neither of these regions has previously been associated with RA. Our analysis also reveals a large number of individual SNPs with nonzero effects (Figure 1). These may also reflect genetic variation controlling the risk for RA. For example, rs2476601 on chromosome 1 has a nonzero

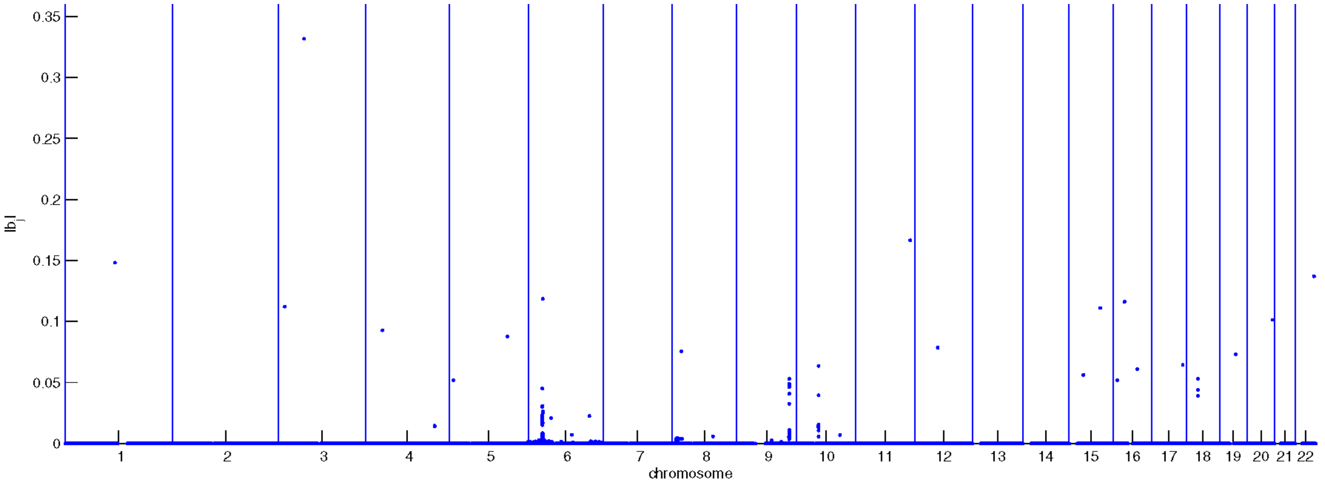

**Figure 1**

**SNPs identified using POCRE-LDA**. The *x*-axis indicates the physical location of each SNP on the chromosome, and the *y*-axis represents the absolute value of the estimated coefficient, i.e., $|b_j|$. Genetic regions with multiple SNPs are identified in chromosomes 6, 8, 9, 10, and 18.

**Table 1: Genomic regions and candidate genes identified for case-control study of rheumatoid arthritis in GAW16**

| Chromosome | Genomic region (Mb) | SNP with the largest effect | Number of genes | Candidate genes |
|---|---|---|---|---|
| 6p21.33 | 31.55-31.62 | rs2523647 | 4 | *MICB-001*; *MICB-002*; *MCCD1-001*; *BAT1* |
| 6p21.32 | 32.33-32.41 | rs10484560 | 1 | *C6orf10* |
| 6p21.32 | 32.47-32.69 | rs3135363 | 6 | *BTNL2*; *HLA-DRA*; *AL662796.6*; *HLA-DRB9*; *HLA-DRB5*; *HLA-DRB1* |
| 6p21.32 | 37.74-32.79 | rs9275601 | 1 | *HLA-DQB1* |
| 6p21.32 | 32.87-32.97 | rs9380326 | 5 | *HLA-DOB*; *TAP2*; *PSMB8*; *PSMB9*; *TAP1* |
| 6p21.32 | 33.21-33.29 | rs3130237<br>rs6901221 | 5 | *COL11A2*; *RXRB*; *SLC39A7*; *HSD17B8*; *RING1* |
| 9q33.1 | 12.05-12.12 | rs2900180 | 3 | *DBC1*; *TRAF1*[a]; *C5*[a] |
| 10q11.22 | 49.64-49.79 | rs2671692 | 2 | *C10orf64*; *LRRC18* |
| 18q12.1 | 26.82-26.88 | rs2852003 | 1 | *DSC3* |

[a] The identified SNPs are in linkage disequilibrium with these genes.

effect and is about 71 kb upstream of the *PTPN22* gene identified by Plenge et al. [14] as associated with RA.

## Discussion

For general settings with large *p* small *n* data, the superior performance of POCRE over existing methods such as LASSO and ridge regression were presented in Zhang et al. [3]. The results of our analysis compare favorably with the earlier GWAS conducted by Plenge et al. [14]. This is in spite of the fact that we conducted only a stage I analysis (i.e., a full GWAS in a single population) rather than the two-stage approach reported by Plenge et al. [14] (i.e., follow-up analysis of a sub-set of highly significant SNPs identified from stage I using a second, unrelated population). Thus, our new analytical procedure appears to be more sensitive and less open to false positives compared with the traditional univariate approach used by Plenge et al. [14]. In addition to the multi-SNP approach we used, another difference between our approach and the method used by Plenge et al. [14] is that we used the first 10 principal components for population stratification in our analysis, whereas Plenge et al. [14] used only the first principal component. It should be noted that our analysis did not find an association between the *STAT4* gene polymorphism and RA that was previously reported by others [15]. However, this earlier analysis was conducted in a case-control association study using only 13 candidate genes selected from within the long (q) arm of chromosome 2 that was previously shown to be in linkage with RA in 642 families of European ancestry [15]. Our data showing a lack of association between the *STAT4* polymorphism and RA is consistent with the previous GWAS by Plenge et al. [14].

## Conclusion

Combination of the novel method POCRE with LDA allows us to identify genomic regions (chromosomes 6, 9, 10, and 18) harboring genes associated with the susceptibility to RA. In addition, we identified several single SNPs that are in LD with genes that have previously been associated with RA.

## List of abbreviations used

GAW: Genetic Analysis Workshop; GWAS: Genome-wide association studies; HLA: Human leukocyte antigen; LD: Linkage disequilibrium; LDA: Linear discriminant analysis; MHC: Major histocompatibility complex; POCRE: Penalized orthogonal-components regression; RA: Rheumatoid arthritis; SNP: Single-nucleotide polymorphism.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

MZ and DZ both conceived the study and drafted the manuscript. MZ and YL designed the study and performed statistical analysis. LW and VP participated in the design of the study and preprocessing of the data. JCF participated in interpreting the statistical analysis results, reviewing and editing 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. Support from the CCE project at the Oncological Sciences Center of Purdue University is gratefully acknowledged. This research was also partially supported by NSF grant IIS-0844945 to DZ, Purdue Research grants to MZ, and Purdue Discovery Park seed grant to MZ, JCF, and DZ.

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. Balding DJ: **A tutorial on statistical methods for population association studies.** *Nat Rev Genet* 2006, **7:**781–791.
2. Waldron ERB, Whittaker JC and Balding DJ: **Fine mapping of disease genes via haplotype clustering.** *Genet Epidemiol* 2006, **30:**170–179.
3. Zhang D, Lin Y and Zhang M: **Penalized orthogonal-components regression for large *p* small *n* data.** *Electron J Stat* 2009, **3:**781–796.
4. Johnstone IM and Silverman BW: **Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences.** *Ann Statist* 2004, **32:**1594–1649.
5. Lin Y, Zhang M, Wang L, Pungpapong V, Fleet JC and Zhang D: **Simultaneous genome-wide association studies of anti-cyclic citrullinated peptide in rheumatoid arthritis using penalized orthogonal-components regression.** *BMC Proceedings* 2009, **3(suppl 7):**S20.
6. Bartlett MS: **Further aspects of the theory of multiple regression.** *Proc Camb Phil Soc* 1938, **34:**33–40.
7. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Baker PIW, Daly MJ and Sham PC: **PLINK: A tool set for whole-genome association and population-based linkage analysis.** *Am J Hum Genet* 2007, **81:**559–575.
8. 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.
9. Ota M, Katsuyama Y, Kimura A, Tsuchiya K, Kondo M, Naruse T, Mizuki N, Itoh K, Sasazuki T and Inoko H: **A second susceptibility gene for developing rheumatoid arthritis in the human MHC is localized within a 70-kb interval telomeric of the TNF genes in the HLA class III region.** *Genomics* 2001, **71:**263–270.
10. Zhang SL, Chabod J, Penfornis A, Reviron D, Tiberghien P, Wendling D and Toussirot E: **TAP1 and TAP2 gene polymorphism in rheumatoid arthritis in a population in eastern France.** *Eur J Immunogenet* 2002, **29:**241–249.
11. Yu MC, Huang CM, Wu MC, Wu JY and Tsai FJ: **Association of TAP2 gene polymorphism in Chinese patients with rheumatoid arthritis.** *Clin Rheumatol* 2004, **23:**35–39.
12. 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 sysmtematic lupus erythematosus.** *Hum Immunol* 2005, **66:**1235–1241.
13. Kurreeman FA, Padyukov L, Marques RB, Schrodi SJ, Seddighzadeh M, Stoeken-Rijsbergen G, Helm-van Mil van der AH, Allaart CF, Verduyn W, Houwing-Duistermaat J, Alfredsson L, Begovich AB, Klareskog L, Huizinga TW and Toes RE: **A candidate gene approach identifies the TRAF1/C5 region as a risk factor for rheumatoid arthritis.** *PLoS Med* 2007, **4:**e278.
14. 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.
15. 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.