

Proceedings

Open Access

Evaluating epistatic interaction signals in complex traits using quantitative traits

Odity Mukherjee*¹, Krishna Rao Sanapala², Padmanabhan Anbazhagana³ and Saurabh Ghosh²

Addresses: ¹National Center for Biological Sciences, Bangalore, India, ²Indian Statistical Institute, Kolkata, India and ³National Institute of Mental Health and Neurosciences, Bangalore, India

E-mail: Odity Mukherjee* - omukherjee@ncbs.res.in; Krishna Rao Sanapala - KrishnaRS@skytechsolutions.co.in; Padmanabhan Anbazhagana - ababubio@gmail.com; Saurabh Ghosh - saurabh@isical.ac.in

*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):S82 doi: 10.1186/1753-6561-3-S7-S82

This article is available from: <http://www.biomedcentral.com/1753-6561/3/S7/S82>

© 2009 Mukherjee et al; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Rheumatoid arthritis (RA) is a complex, chronic inflammatory disease implicated to have several plausible candidate loci; however, these may not account for all the genetic variations underlying RA. Common disorders are hypothesized to be highly complex with interaction among genes and other risk factors playing a major role in the disease process. This complexity is further magnified because such interactions may be with or without a strong independent effect and are thus difficult to detect using traditional statistical methodologies. The main challenge to analyze such gene \times gene and gene \times environment interaction is attributed to a phenomenon referred to as the “curse of dimensionality.” Several combinatorial methodologies have been proposed to tackle this analytical challenge. Because quantitative traits underlie complex phenotypes and contain more information on the trait variation within genotypes than qualitative dichotomy, analyzing quantitative traits correlated with the affection status is a more powerful tool for mapping such trait genes. Recently, a generalized multifactor dimensionality reduction method was proposed that allows for adjustment for discrete and quantitative traits and can be used to analyze qualitative and quantitative phenotypes in a population based study design.

In this report, we evaluate the efficiency of the generalized multifactor dimensionality reduction statistical suite to decipher small interacting factors that contribute to RA disease pathogenesis.

Introduction

Rheumatoid arthritis (RA) is a complex chronic inflammatory disease implicated to have several plausible candidate loci. Many genetic studies have been undertaken

and only two genes, *HLA-DRB1* and *PTPN22*, have been reported to be associated with disease [1-4]. Although these findings are encouraging, they may not account for all the genetic variations in RA because no direct pathogenic role

of these molecules have been established in the development of the disease pathogenesis. Common disorders like RA are hypothesized to be highly complex, with interaction among genes and other risk factors playing a major role in the disease process. This complexity is further magnified because such interactions may be with or without strong independent main effect, and thus difficult to detect using traditional statistical methodologies [5]. The main challenge to analyzing epistatic interactions is attributed to a phenomenon referred to as the "curse of dimensionality," which is a problem caused by the exponential increase in volume associated with adding extra dimensions to a mathematical space. Thus, while analyzing interactions among several loci for a complex phenotype, contingency tables in higher dimensions suffer from the problem of sparse data, leading to unreliable risk estimates. Several combinatorial methodologies have been proposed to overcome this analytical challenge: multifactor dimensionality reduction (MDR) [6]; combinatorial partitioning method (CPM) [7] and restricted partition method (RPM) [8]. Although these methods have been used by several research groups, there exist some limitations in their current form: a) inability to adjust for covariates' MDR, b) inability to use quantitative phenotypes, and c) computationally intense algorithms.

Thus, there is a need to develop and evaluate more powerful statistical methodology so as to decipher small interacting factors that contribute to disease pathogenesis. Because quantitative traits underlie complex phenotypes and contain more information on the trait variation within genotypes than qualitative dichotomy, analyzing quantitative traits correlated with the affection status is a more powerful tool for mapping complex trait genes. Recently, a generalized MDR (GMDR) method was proposed that allows for adjustment for discrete and quantitative traits and can be used to analyze qualitative and quantitative phenotypes in a population based study design [9].

In this report, we use the GMDR statistical suit to evaluate its efficiency to decipher small interacting factors that contribute to RA disease pathogenesis, using the two quantitative traits [anti-CCP (anti-cyclic citrullinated peptide) and RFUW (rheumatoid factor)] as covariates for classifying the data into high and low risk groups.

Data analysis

An initial screen of data for quality control was performed for the markers selected for the current study. Hardy-Weinberg equilibrium (HWE) was estimated in the case, control, and combined groups using the Haploview program (version 3.32). To understand

the degree of correlation between the SNPs, linkage disequilibrium (LD) was estimated using the Haploview program (version 3.32). The D' statistics for the same is presented in Figure 1. Such information is essential when analyzing data employing cross-validation steps because it is possible that the algorithm might identify different SNPs (but in tight LD) for each of its cross-validation intervals [2]. This was followed by the GMDR analysis for detecting epistatic interactions.

Methods

Sample and marker selection

In the current study, we used the Genetic Analysis Workshop 16 (GAW16) RA case-control data set (Problem 1) comprising a total of 2062 sample (case = 868, control = 1194), typed on the 550 k Illumina chip. To evaluate the efficiency of the GMDR algorithm to detect small epistatic interactions involved in RA pathogenesis, analysis was performed on chromosomes 1, 2, 5, and 6, which have shown strong positive association earlier with the phenotype [1-4,10]. Because quantitative trait information was available for only cases, interaction analysis using GMDR was performed on the RA cases ($n = 867$).

Interaction studies

The GMDR is a score-based algorithm based on the MDR framework. Briefly, the MDR uses a novel constructive induction algorithm to facilitate the detection of non-linear interactions among multiple discrete genetic and/or environmental factors that are predictive of a discrete clinical endpoint [11]. Multi-locus genotype combinations are classified as high-risk or low-risk genotype combinations using a threshold that is equal to the ratio of cases and controls. The best model is selected as the combination of marker with maximum cross-validation consistency and minimum prediction error. GMDR works on the same framework of MDR, but is a score-based algorithm. Improving on the original MDR, it can be used on both qualitative and quantitative traits, it allows adjustment for covariates and better handles unbalanced population based data. For the current study, we employed two methods to compute the scores (described below) for GMDR using the QTL information provided in the data set:

1. The GMDR scoring method: The GMDR method uses the original MDR data reduction method, with the ratio of cases to control being replaced by scores in each cell to discriminate between high risk and low risk followed by determining classification accuracy and prediction error. A detailed description of the methodology can be found elsewhere [9]. This generalization of the original MDR algorithm a)

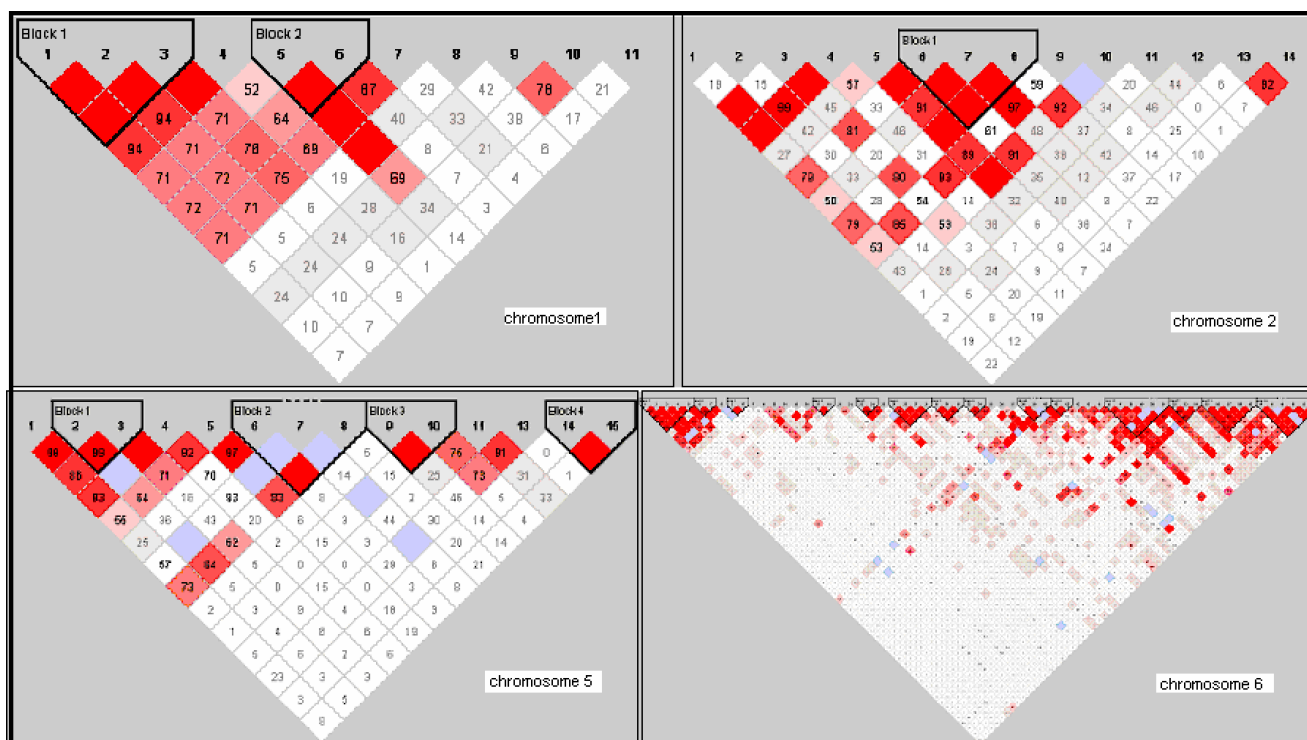


Figure 1
LD block structure across the chromosomal regions used in this study. The figures show the output of Haploview (version 3.32) LD Plot where each square (with D' values written within the box) represents a pair-wise LD relationship between the two SNPs. Red squares indicate statistically significant LD between the pair of SNPs as measured by the D' statistic. Darker colors of red indicate higher values of D', up to a maximum of 1. White squares indicate pair-wise D' values < 1 with no statistically significant evidence of LD.

allows increased flexibility to use covariates, b) is able to handle both dichotomous and continuous phenotypes, c) can be applied to a variety of population-based study designs (e.g., unbalanced case control samples.)

2. We formulated a detailed scoring methodology by using the expression $S = \frac{\exp(\gamma)}{1 + \exp(\gamma)}$, where γ is the standardized quantitative trait. In brief, this was done by computing the mean and standard deviation (SD) of the quantitative trait. Scores were then assigned by subtracting the mean from the individual's quantitative trait value and then dividing it by the SD.

Results

Marker selection

In the current study we used GMDR algorithm to evaluate its efficiency in detecting gene-gene interactions in the complex RA phenotype. For this we used markers information from the GAW16 data set from regions that have been previously implicated in RA. Additional file 1 lists the markers and their chromosomal position used

in this analysis. All the markers selected were in HWE (data not shown). None of the regions selected showed extensive LD between the markers (Figure 1).

Interaction studies

While the MDR software is designed to classify individuals into high risk and low risk groups, GMDR is a score-based method in which the ratio of cases to control is replaced by scores in each cell to discriminate between high risk and low risk cells and then assessing classification accuracy and prediction error. GMDR was performed on the genotype data (cases only) from the GAW16 Problem 1 data set with the computed scores. The phenotype scores used in the analysis were generated using the built-in GMDR scoring method and a detailed scoring method (described above in the Methods section). Analysis was performed individually for the separate chromosomal regions. An exhaustive search was performed to identify all possible one- to five-locus models. We report the prediction accuracy and cross-validation consistency for the most significant models identified by GMDR, the results of which are

Table 1: Summary of the best models obtained using GMDR algorithm for the quantitative trait RFUW (IgM)^a

	No. loci			
	1	2	3	4
SNPs in best model	rs2156875	rs1517352 rs3024896	rs11203368 rs6683201 rs3789607	rs3024912 rs1517352 rs4555370 rs231726
Chromosome	6	6	1	2
Gene	HLA-B	CTLA4	PADI4, PTPN22	STAT4, CTLA4
Predictive accuracy	0.5739	0.5577	0.5069	0.5396
Cross-validation consistency	10	6	5	7
Sign test <i>p</i> -value	0.017	0.001	0.377	0.0547

^aAll models used the GDMR scoring method.

summarized in Table 1. GMDR was able to identify small interacting factors in the regions analyzed in this study. This substantiates the efficiency of the GMDR and the candidate loci for harboring disease-associated markers.

Discussion

Rheumatoid factor (RFUW) has been widely used as a screening test for patients with RA. RFUW is prognostically useful because it correlates with functional and radiographic outcomes in RA [12]. More recently, the anti-cyclic citrullinated peptide (anti-CCP) antibody has been developed, with a sensitivity of ~68% and specificity of 97% [13,14]. Together, these clinical values serve as important indicators of the disease status and are routinely used in clinical setting to aid in diagnosis. Common disorders like RA are hypothesized to be highly complex, with interaction among genes and other risk factors playing a major role in the disease process. Powerful statistical methodology has been developed to overcome these challenges to decipher small epistatic interactions that are characteristic of such phenotypes. Because quantitative traits underlie complex phenotypes and contain more information on the trait variation within genotypes than qualitative dichotomy, we used the anti-CCP value and the RFUW values provided in the GAW16 Problem 1 data set to evaluate the recently developed GMDR algorithm to detect small interacting markers for RA disease status.

In this study we used the GMDR methodology to evaluate its efficiency to detect gene-gene interactions in putative regions for RA using the anti-CCP and RFUW (IgM) values as covariates. Three out of the four models predicted reached statistical significance (Table 1). None of the high-order interactions were between correlated markers, suggesting that there might be more than one signal in these genes. For this study we had used both the anti-CCP and the RFUW values to generate scores for the GMDR analysis. Scoring based on anti-CCP value did not result in significant interaction models. Our results

show that RFUW values are better predictor of high-risk and low-risk classes and further strengthen the role of RFUW (IgM) antibody as a strong prognostic factor. Detailed biological characterization of this quantitative trait are warranted.

List of abbreviations used

anti-CCP: Anti-cyclic citrullinated peptide; CPM: Combinatorial partitioning method; GAW16: Genetic Analysis Workshop 16; GMDR: Generalized multifactor dimensionality reduction; HWE: Hardy-Weinberg equilibrium; IgM: Immunoglobulin M; LD: Linkage disequilibrium; MDR: Multifactor dimensionality reduction; RA: Rheumatoid arthritis; RFUW: Rheumatoid factor; RPM: Restricted partition method; SD: Standard deviation; SNP: Single-nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

OM carried out all statistical analysis, data interpretation, and drafted the manuscript. KRS and PA contributed in data cleaning and analysis. SG contributed in data analysis manuscript writing. All authors have read the paper and approve its contents.

Additional material

Additional file 1
SNPs used in this study.
 Click here for file
[\[http://www.biomedcentral.com/content/supplementary/1753-6561-3-S7-S82-S1.pdf\]](http://www.biomedcentral.com/content/supplementary/1753-6561-3-S7-S82-S1.pdf)

Acknowledgements

The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medicine.

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

- Jawaheer D, Seldin MF, Amos CI, Chen WV, Shigeta R, Etzel C, Damle A, Xiao X, Chen D, Lum RF, Monteiro J, Kern M, Criswell LA, Albani S, Nelson JL, Clegg DO, Pope R, Schroeder HW Jr, Bridges SL Jr, Pisetsky DS, Ward R, Kastner DL, Wilder RL, Pincus T, Callahan LF, Flemming D, Wener MH, Gregersen PK and North American Rheumatoid Arthritis Consortium: **Screening the genome for rheumatoid arthritis susceptibility genes.** *Arthritis Rheum* 2003, **48**:906–916.
- Ritchie MD, Bartlett J, Bush WS, Edward TL, Motsinger AA and Torstenson ES: **Exploring epistasis in candidate genes for rheumatoid arthritis.** *BMC Proc* 2007, **1(suppl 1)**:S70.
- Mei L, Li X, Yang K, Cui J, Fang B, Guo X and Rotter JI: **Evaluating gene × gene and gene × smoking interaction in rheumatoid arthritis using candidate genes in GAW15.** *BMC Proc* 2007, **1(suppl 1)**:S17.
- Pankratz N: **A two stage classification approach identifies seven susceptibility genes for a simulated complex disease.** *BMC Proc* 2007, **1(suppl 1)**:S30.
- Culverhouse R, Suarez BK, Lin J and Reich T: **A perspective on epistasis: limits of models displaying no main effect.** *Am J Hum Genet* 2002, **70**:461–471.
- Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF and Moore JH: **Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.** *Am J Hum Genet* 2001, **69**:138–147.
- Nelson MR, Kardia SL, Ferrell RE and Sing CF: **A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation.** *Genome Res* 2001, **11**:458–470.
- Culverhouse R, Klein T and Shannon W: **Detecting epistatic interactions contributing to quantitative traits.** *Genet Epidemiol* 2004, **27**:141–152.
- Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC and Li MD: **A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence.** *Am J Hum Genet* 2007, **80**:1125–1137.
- Gregersen PK, Silver J and Winchester RJ: **The shared epitope hypothesis: an approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis.** *Arthritis Rheum* 1987, **30**:1205–1213.
- Hahn LW, Ritchie MD and Moore JH: **Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions.** *Bioinformatics* 2003, **19**:376–382.
- Möttönen T, Paimela L, Leirisalo-Repo M, Kautiainen H, Ilonen J and Hannonen P: **Only high disease activity and positive rheumatoid factor indicate poor prognosis in patients with early rheumatoid arthritis treated with “sawtooth” strategy.** *Ann Rheum Dis* 1998, **57**:533–539.
- Schellekens GA, Visser H, de Jong BA, Hoogen van den FH, Hazes JM, Breedveld FC and van Venrooij WJ: **The diagnostic properties of rheumatoid arthritis antibodies recognizing a cyclic citrullinated peptide.** *Arthritis Rheum* 2000, **43**:155–163.
- Bizzaro N, Mazzanti G, Tonutti E, Villalta D and Tozzoli R: **Diagnostic accuracy of anti-citrulline antibody assay for rheumatoid arthritis.** *Clin Chem* 2001, **47**:1089–1093.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

