Computational Biology and Chemistry 62 (2016) 1–11

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Using propensity score adjustment method in genetic association studies

Amrita Sengupta Chattopadhyay[a,b,c,1], Ying-Chao Lin[d,1], Ai-Ru Hsieh[e,1], Chien-Ching Chang[d], Ie-Bin Lian[f,*], Cathy S.J. Fann[a,d,*]

[a] Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan
[b] Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan
[c] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[d] Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan
[e] Graduate Institute of Biostatistics, China Medical University, Taichung, Taiwan
[f] Department of Mathematics, National Changhua University of Education, Changhua, Taiwan

ARTICLE INFO

Article history:
Received 8 January 2016
Received in revised form 7 February 2016
Accepted 17 February 2016
Available online 3 March 2016

Keywords:
Propensity score
Logistic regression
Single-point association test
Gene–gene interaction
Rheumatoid arthritis
Single nucleotide polymorphism

ABSTRACT

Background: The statistical tests for single locus disease association are mostly under-powered. If a disease associated causal single nucleotide polymorphism (SNP) operates essentially through a complex mechanism that involves multiple SNPs or possible environmental factors, its effect might be missed if the causal SNP is studied in isolation without accounting for these unknown genetic influences. In this study, we attempt to address the issue of reduced power that is inherent in single point association studies by accounting for genetic influences that negatively impact the detection of causal variant in single point association analysis. In our method we use propensity score (PS) to adjust for the effect of SNPs that influence the marginal association of a candidate marker. These SNPs might be in linkage disequilibrium (LD) and/or epistatic with the target-SNP and have a joint interactive influence on the disease under study. We therefore propose a propensity score adjustment method (PSAM) as a tool for dimension reduction to improve the power for single locus studies through an estimated PS to adjust for influence from these SNPs while regressing disease status on the target-genetic locus. The degree of freedom of such a test is therefore always restricted to 1.

Results: We assess PSAM under the null hypothesis of no disease association to affirm that it correctly controls for the type-I-error rate (<0.05). PSAM displays reasonable power (>70%) and shows an average of 15% improvement in power as compared with commonly-used logistic regression method and PLINK under most simulated scenarios. Using the open-access multifactor dimensionality reduction dataset, PSAM displays improved significance for all disease loci. Through a whole genome study, PSAM was able to identify 21 SNPs from the GAW16 NARAC dataset by reducing their original trend-test *p*-values from within 0.001 and 0.05 to *p*-values less than 0.0009, and among which 6 SNPs were further found to be associated with immunity and inflammation.

Conclusions: PSAM improves the significance of single-locus association of causal SNPs which have had marginal single point association by adjusting for influence from other SNPs in a dataset. This would explain part of the missing heritability without increasing the complexity of the model due to huge multiple testing scenarios. The newly reported SNPs from GAW16 data would provide evidences for further research to elucidate the etiology of rheumatoid arthritis. PSAM is proposed as an exploratory tool that would be complementary to other existing methods. A downloadable user friendly program, PSAM, written in SAS, is available for public use.

Abbreviations: PS, propensity Score; PSAM, propensity score adjustment method; ULRM, univariate logistic regression method; S-MLRM, stepwise-multivariate logistic regressioin method; GWAS, Genome Wide Association studies; GAW, genetic analysis workshop; NARAC, North American Rheumatoid Arthritis Consortium; SNP, single nucleotide polymorphism; CPU, Central Processing Unit; LD, linkage disequilibrium; Dom, dominant; Rec, recessive; MDR, multifactor dimensionality reduction; ATM, Ataxia telangiectasia mutated; PTPN22, protein tyrosine phosphatase, non-receptor type 22 lymphoid; IPA, ingenuity pathway analysis; DNA, deoxyribo-nucleic acid; QC, quality control; PSM, propensity score matching; SPS, stratified propensity score; PC, principle component; PCA, principle component analysis.

* Corresponding authors.
E-mail addresses: maiblian@cc.ncue.edu.tw (I.-B. Lian), csjfann@ibms.sinica.edu.tw (C.S.J. Fann).

[1] These authors contributed equally to the manuscript.

http://dx.doi.org/10.1016/j.compbiolchem.2016.02.017
1476-9271/

1. Introduction

Identifying disease susceptibility loci for complex diseases through genetic association studies assume that finding highly significant statistical differences between marker allele frequencies in case and control populations would help to discover genes influencing disease expression (Corso and Greenberg, 2014). The statistical tests used to detect causal variant using single locus association analyses are mostly under-powered. One could conceivably focus on the complex relationship between multiple genetic variants to account for the loss of heritability in such single locus association strategies (Zuk et al., 2012). If a disease associated causal single nucleotide polymorphism (SNP) operates essentially through a complex mechanism that involve multiple SNPs or environmental factors, the effect of the causal SNP might be missed if it is studied in isolation without accounting for these unknown genetic influences. Hence, methods which focus on pair-wise statistical interaction between SNPs, in search for disease association have been developed (Sengupta Chattopadhyay et al., 2014; Hahn et al., 2003; Purcell et al., 2007). Multifactor dimensionality reduction method (MDR) is one such popular method, which uses a non-parametric technique via exhaustive search strategy for higher order epistasis analysis (Hahn et al., 2003). The epistasis module of PLINK is another popular state of the art procedure which uses a parametric technique to conduct gene-gene interaction analysis (Purcell et al., 2007). However, there are several issues that pose challenge for such interaction studies. These include the problem of handling higher order interactions (>2-way) where the number of possible marker combinations is manifold for a large number of SNPs, hence, analysis involving large samples are required to obtain a significant p-value. The amount of time needed to perform such analyses even with advanced CPU power and space is unaccountable for large-scale association data. Moreover, presence of allelic heterogeneity, varying amounts of linkage disequilibrium (LD) between disease alleles and marker SNP, and differing interactions of disease allele with alleles at other loci add to the challenge of identifying disease association with interacting loci.

In single locus analysis strategies, statistical power to detect the causal locus is likely to be reduced as the effect of the causal locus is masked by the effects of genetic variants at other loci (definition of epistasis) thereby compromising the inference of detecting causal variant (Cordell, 2002). This masking effect is essentially a definition of epistatic effect of other SNPs on the causal SNP as stated by Cordell (2002). Accounting for such masking/confounding effects would explain part of the missing heritability in single locus analysis strategies.

Propensity score (PS) approach is popular to epidemiologists for mitigating confounding effects in observational studies (Austin, 2011). It was first proposed by Rosenbaum and Rubin to infer cause and effect from observational studies (Rosenbaum and Rubin, 1983). PS is estimated through a conditional probability where a particular treatment is assigned given a vector of observed covariates and adjustment for this PS is sufficient to remove bias due to all observed covariates (Rosenbaum and Rubin, 1983). An observational study attempts to estimate the effects of a treatment or an exposure by comparing outcomes for subjects who were not assigned at random to treatment or control (Jepsen et al., 2004). PS reduces or adjusts for the effects of confounders to estimate treatment effects, and it balances covariates (such as age, gender or population principle components) so that treated and control groups are comparable (Austin, 2011). Most often a logistic regression model is used to estimate the true PS by regressing the treatment status on observed covariates (McCullagh and Nelder, 1989). Several ways to adjust for covariates when estimating the effects of treatment on outcomes using PS include matching, stratification (or subclassification), covariate adjustment and inverse probability treatment weighing (IPTW) (Pourhoseingholi et al., 2012; Pirracchio et al., 2015).

In genetic association studies, to infer the direct causal effect of a genetic variant, Vansteelandt et al. (2009) have proposed to use PS to control for the effect of biological phenotypes to adjust for confounding. Jiang and Zhang (2011) have also suggested using non-parametric techniques to obtain PS while adjusting for covariates like population stratification or environmental factors for SNPs of interest to identify disease association. Instead of directly testing epistatic effects from numerous combinations of SNPs, in this paper, we propose using PS as a dimension-reduction tool to improve the marginal single-point association result for each SNP by accounting for loss of heritability. The effects include SNPs that are epistatic and/or in LD (correlated) with the target-SNP. Epistatic SNPs have a joint interactive influence on the disease under study. Correlated SNPs i.e., those in LD with the target-SNP have an effect on the disease. The underlying model $logit(Y) = \beta_0 + \Sigma_{k=1\sim n}\beta_k \cdot \mathrm{SNP_k} + g\ (\mathrm{SNP}_i, \mathrm{SNP}_j)$, consists of both main effects for each SNP ($\Sigma_{k=1\sim n}\beta_k \cdot \mathrm{SNP_k}$), and epistasis effects g (SNP_i,SNP_j) caused by disease-related SNP_i and SNP_j for specific i and j. The epistatic function g can only be expressed in a form of two-way matrix (Table 1). The goal of this study is to detect single point disease association of SNP_k ($k = 1 \sim n$). When SNP_i and SNP_j for unknown i and j have relatively weak main effects (β_i,β_j)but strong epistatic effect g, models including either SNP_i or SNP_j usually have low power due to the weak main effects, and their effects are likely to be ignored unless both are present in the fitted model. However it is not plausible to include either all SNPs (main effects) in one model or to try models of all possible interacting SNP combinations (for all i and j) when n is large. A conventional method, stepwise-multiple logistic regression model (S-MLRM), is to fit Y on the *SNP*s using stepwise selection procedure (Cordell and Clayton, 2002). However this method have low power when the target marker SNP_i and SNP_j have weak main effects. Another naïve choice, is univariate logistic regression model (ULRM), where univariate SNPs (SNP_k for each $k = 1 \sim n$) are fitted individually through a logistic regression model and significant SNP_ks' are picked, may also result in low power due to the similar reason.

The proposed propensity score adjustment method (PSAM) tests single locus association through an estimated PS to adjust for influence from correlated and/or epistatic genetic factors. Such genetic factors (SNPs) could be termed as covariate SNPs as they could possibly be predictive of the disease under study or may be in association with the SNP under study (Clarke et al., 2011). The PS of a target-SNP is expected to summarize the necessary information from the numerous other SNPs which may be potentially epistatic or correlated with the disease locus, and therefore reduces the dimensionality of statistical testing. The original application of PS was to balance the covariates' distributions for a binary target-explanatory variable (exposure and non-exposure). However, for continuous or multi-level target-variable, it had been shown that the direct adjustment by incorporating PS into the model can also efficiently reduce the bias due to the imbalance, whereas in this case the PS was estimated by linear regression (Lian, 2003). In this study, for simplification, we illustrate the effect of the PS for a binary target-SNP i.e., the SNP under study, by assuming the target-SNP as dominant or recessive. We propose to identify candidate covariate SNPs, prioritize them using a stepwise regression method and integrate them into the PS based covariate adjustment model (Schneeweiss et al., 2009). The target-SNP is regressed over such covariate SNPs to obtain the predicted PS. The systematic differences in these SNPs are taken care of between cases and control subjects while estimating the adjusted effect of the