Mini Review

Open Access

CrossMark

A mini review for aggregating analyses for genomewide association studies (GWAS)

Volume 2 Issue 4 - 2015

Caiyan Li
Takeda Pharmaceuticals, USA

Correspondence: Caiyan Li, Takeda Pharmaceuticals, USA, Email Caiyan07@gmail.com

Received: April 08, 2015 | **Published:** April 29, 2015

Univariate analysis

Genome-wide Association Studies (GWAS) has emerged as a popular tool for identifying common genetic variants for complex disease. A typical case-control GWAS often involves genotyping hundreds of thousands of Single-nucleotide polymorphism (SNPs), in thousands of disease cases and healthy controls, with the goal of identifying individual SNPs that are associated with the disease (or outcome). GWAS usually consists of a discovery phase, in which SNPs are identified followed by a validation phase, in which the SNPs identified in discovery phase are replicated in a separate study cohort. The standard GWAS method is univariate analysis, in which the phenotype is regressed onto each individual SNP to generate p-value for association measurement. The SNPs are then ranked based on their univariate p-values, and a threshold is set such that any SNP with a p-value below the threshold is identified as significant. Multiplicity adjustment methods, such as bonferroni, false discovery rate (FDR) can be used to set the threshold.

Though the use of univariate analysis has helped to identify many disease-susceptibility variants, and it is easy to comprehend and conduct, there are limitations which pose difficulty in some settings. Due to the large number of tested individual SNPs (e.g., 4 million SNPs), the adjusted threshold for genome-wide significance can be extremely low, e.g. $\alpha = 10^{-7}$,[1] which is very difficult to attain. Secondly, univariate-SNP analysis often has poor reproducibility. Many highly-ranked SNPs identified in the discovery phase are false positives due to the low power for detecting SNPs with small effects. Moreover, rare variants (defined as alleles with a frequency of 1%-5%) require a significant large sample size in order to be detected via standard univariate analysis. Since multivariate analysis has the ability to take into account of the correlation between multiple SNPs, as well as help reduce the multiplicity burden (e.g., lower the significance level), it could be advantageous to consider the joint effect of multiple SNPs together. To this end, researchers have established various methodologies to aggregate individual SNPs together as a set and then conduct association analyses, which has aided with the discovery of disease-related variants, especially on rare variants. Among these methods, there are two major categories: burden tests and kernel-based variance component tests. Below, we will review some popular methods within these two categories.

Burden tests

Burden tests assess the cumulative effects of multiple variants within a genomic region by collapsing or summarizing the variants within a region by a single value, which is then used to test for association with the outcome of interest. Some popular burden tests are briefed below:

Cohort allelic sum test (CAST):[2] collapse genotypes across all variants within a specific region into a single dichotomous value for each subject. A subject is coded as 1 if a rare variant is present at any of the variant sites within the region; otherwise as 0. One variation of CAST involves collapsing by counting the number of rare variants within a specific region, instead of dichotomizing.

Combined multivariate and collapsing (CMC):[3] The CMC method is a unified approach that combines collapsing and multivariate tests. SNPs within the region are divided into subgroups based on predefined criteria, such as allele frequencies, and within each subgroup CART was used to collapse genotypes. A multivariate test (e.g., Hotellings's test) is then applied to the subgroups for association analysis.

Weighted sum statistic (WSS):[4] The WSS method specifically considers the case-control setting and collapse SNPs within a set into a single weighted average of the number of rare alleles for each individual, and then apply Wilcoxon rank sum test to compare between case and control. Specifically, the weight for SNP i is defined as $w_i = \frac{1}{\sqrt{n_i q_i (1-q_i)}}$, where $q_i = \frac{m_i^U + 1}{2n_i^U + 2}, m_i^U$ is the number of rare alleles observed for SNP i in the control group, n_i^U is the number of control subjects genotypes for SNP i, and is the total number of subjects genotypes for SNP i. The weight is the inverse of the estimated standard deviation of the total number of rare alleles in the sample (including both cases and controls), under the null hypothesis of no frequency differences. It is used to down-weight rare alleles in constructing the weighted-sum score test.

One limitation for all these burden tests is that they implicitly assume that all rare variants influence the phenotype in the same direction and with the same magnitude of effect (after incorporating weights). However, biological mechanism leads us to believe that most variants within a region have little or no effect on phenotype; in addition some variants are protective and some are deleterious, and the magnitude of effects may vary as well. In these situations, employ the above burden tests may introduce substantial noise and largely reduce the power. Thus, researchers proposed methods that can account for the different direction and magnitude of effects for SNPs within a region.

Sequence kernel association test (SKAT)[5]

Kernel-based test methods, such as SKAT, are non-burden tests. Instead of aggregating variants, SKAT aggregates the associations between variants and the phenotype through a kernel matrix and can allow for SNP-SNP interactions, i.e., epistatic effects. SKAT is a flexible, computationally efficient, regression based approach that tests for association between variants in a region (both common and rare) and a phenotype (dichotomous or continuous) while adjusting for covariates. It borrows information from correlated SNPs that are grouped on the basis of prior biological knowledge (e.g., gene or pathway) and hence produce results with improved reproducibility and power. In this review, we assume a population-based case-control GWAS with n independent genotyped subjects. For a given SNP-set containing p SNPs, let $Z_{i1}, Z_{i2}, Z_{i3}, \ldots\ldots, Z_{ip}$ be the genotype values for subject $i\,(i{=}1,\ldots n)$. Let y_i denote the case-control status for subject i (1 for case and 0 for control); $X_{i1}, X_{i2}, \ldots\ldots, X_{im}$ denote the covariates (e.g., demographic and clinical variables) that we would like to adjust for. The goal is to test the global null of whether any of the p SNPs is related to the outcome while adjusting for covariates. In evaluating the significance of the joint effect of the SNP-set, a logistic kernel-machine regression model is employed as follows:

$$\log it\ P\left(y_i{=}1\right) = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_m x_{im} + h(z_{i1,} z_{i2,} \ldots, z_{ip})$$

The SNPs influence the outcome through the general function h(.), which is an arbitrary function that has a form defined only by a positive semidefinite kernel function K(. , .). For subject i, we can fully define using the kernel function K(. , .) as: $h\left(z_{i1,}z_{i2,\ldots,}z_{ip}\right) = \Sigma_{i=1}^{n} \gamma_{i'} K(Z_i, Z_{i'})$, for some $\gamma_1, \ldots, \gamma_n$. By choosing different kernel functions, we can specify different bases and corresponding models to assess the effects of SNPs. For example, if we define K(. , .) as the linear kernel such that $K(Z_i, Z_{i'}) = \Sigma_{j=1}^{p} z_{ij} z_{i'j}$ then we are assuming the simple linear and logistic model defined by $\log it\ P\left(y_i{=}1\right) = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_m x_{im} + \beta_1 z_{i1} + \ldots + \beta_p z_{ip}$ The outcome depends on the SNPs solely through the function h(.), thus, in order to test whether there is a true SPN-set effect, the following null hypothesis is considered: $H_0 : h(z) = 0$ against the general alternative. A variance-component score test is developed to test the above hypothesis. To learn more about the technical details, please refer to references.

The choice of kernel changes the underlying base for the nonparametric function governing the relationship between the outcome and SNPs in the set. Essentially, K(. , .) is a function that projects the genotype data from the original space to another space and then h(.) is modeled linearly in this new space. More intuitively, $K\left(Z_i, Z_{i'}\right)$ can be viewed as a function that measures the similarity between two individuals with regard to the genotypes. Some specific kernels that SKAT considers are weighted linear kernel $K\left(Z_i, Z_{i'}\right) = \Sigma_{j=1}^{p} w_j z_{ij} z_{i'j}$, which implies that the outcome depends on SNPs in a linear fashion; weighted quadratic kernel $K\left(Z_i, Z_{i'}\right) = (1 + \Sigma_{j=1}^{p} w_j z_{ij} z_{i'j})^2$, which assumes that the model depends on the main effects and the quadratic terms of SNPs and first order of SNP-by-SNP interactions; weighted IBS (identical-by-state) kernel $K\left(Z_i, Z_{i'}\right) = \Sigma_{j=1}^{p} w_j IBS(z_{ij}, z_{i'j})$ which defines similarity between two individuals as the number of alleles that share IBS. The weighted IBS kernel allows for epistatic effects because it does not assume linearity or interactions of any particular order (e.g., second order). Experiences suggest using the IBS kernel when the number of interacting SNPs within the region is modest. In all these kernels, w_j is an allele specific weight that controls the relative importance of the jth SNP. Without prior information, SKAT suggests to use $\sqrt{w_j} = Beta(MAF_j; 1, 25)$, where is the minor allele frequency of the j^{th} SNP. If prior information is available, weight can be selected to increase (or decrease) the weight for likely functionality.

Although SKAT makes few assumptions about rare-variants effects and provides attractive power, it has some limitations in certain settings. For example, in the setting that a large proportion of the rare variants in a given region are truly causal and influence the phenotype in the same direction, SKAT can be less powerful than burden tests. Thus, SKAT-O[6] is proposed, which automatically behaves like the burden test when the burden test is more powerful than SKAT, and behaves like SKAT when SKAT is more powerful than the burden test.

The SKAT research team has developed a powerful R package which implements SKAT and SKAT-O and their Small-Sample Adjustments. The team also developed functions to calculate power and sample size to help design studies to evaluate SNP-set effects. http://www.hsph.harvard.edu/~xlin/software.html

Discussion

In this mini review, we have briefly summarized three-types of methodologies for analyzing SNP data: univariate analysis, burden tests and SKAT (SKAT-O). While univariate analysis lacks of power, it is the classic method that has been used in GWAS to help identify many disease-related SNPs. SKAT has the advantage of taking into account of correlation of a set of SNPs, thus improve power. SKAT-O is a combination of SKAT and burden tests. Personal experiences suggest carrying out both univariate analysis and SKAT (SKAT-O) while dealing with genome-wide association studies. Deep-sequencing will soon generate enormous genetic information for large disease samples and the need for statistical methods to analyze these big data is increasing. More research in this field to search for unified optimal approach which incorporates known biological information is needed.

Acknowledgement

None.

Conflict of interest

The authors declare that they have no financial or non-financial competing interests.

References

1. Wu MC, Kraft P, Epstein MP, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet*. 2010;86(6):929–942.

2. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res*. 2007;615(1–2):28–56.

3. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*. 2009;5(2):e1000384.

4. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 2008;83(3):311–321.

5. Wu MC, Lee S, Cai T, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82–93.

6. Lee S, Emond MJ, Bamshad MJ, et al. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. *Am J Hum Genet*. 2012;91(2):224–237.