Gene-based analysis of rare and common variants to determine association with blood pressure

Xiaofeng Liu, Joseph Beyene*

From Genetic Analysis Workshop 18
Stevenson, WA, USA. 13-17 October 2012

Abstract

Systolic blood pressure and diastolic blood pressure are known risk factors for cardiovascular diseases and understanding their genetic basis will have important public health implications. For rare variants, it is extremely challenging to make statistical inference for single-maker tests. Therefore, joint analysis of a set of variants has been proposed. In this paper, we applied recently proposed methods "test for testing the effect of an optimally weighted combination of variants" and "variable weight-TOW" to determine genetic regions that are associated with blood pressure. Then least absolute shrinkage and selection operator, as well as sparse partial least square methods, were used to identify significant markers within a gene or in intergenic regions. We investigated the effect of rare variants and common variants, and their combined effect.

Background

It is well known that high blood pressure is an important risk factor for cardiovascular diseases. Elevated blood pressure is a complicated trait that affects more than 30% of the adult population [1,2]. An increase in systolic and diastolic blood pressure has a continuous impact on the risk of cardiovascular diseases. Globally, every year, high blood pressure contributes to approximately 13.5% of premature deaths, 54% of stroke, and 47% of ischemic heart disease [1,3]. Genetic heritance is one of the major risk factors for hypertension. For complex diseases, the common disease-common variant (CD-CV) hypothesis that underpins genome-wide association studies (GWAS) has led to the identification of several novel susceptibility loci. However, a majority of the heritability is unexplained. It has been pointed out that the GWAS-identified variants can only explain a small portion of the heritability; therefore, exploration is still needed to unveil the undiscovered variants [12]. Recently, arguments have been put forward against CD-CV, and common disease-rare variants (CD-RV) as an alternative has been proposed. It is based on the assumption that the etiology for common diseases is caused by the cumulative effect of multiple rare variants [4,5]. Nevertheless, another merging hypothesis states that common diseases are caused by the combination of common and rare variants [6-8].

In this paper, we focused on identifying whether a gene is associated with blood pressure. We applied recently proposed tests called “test for testing the effect of an optimally weighted combination of variants (TOW)” and “variable weight-TOW (VW-TOW)” [9] to determine significant genetic regions. Our interest also lies on identifying the associated variants for regions that are found significantly associated by applying sparse methods Lasso and SPLS [10,11].

Methods

Data
Both the real and simulated data that were made available for Genetic Analysis Workshop 18 (GAW18) were used. We focused on the genotype data on chromosome 3 for unrelated individuals. The baseline data for the covariates and the phenotypes were considered. We considered the
first time point of systolic blood pressure (SBP) and diastolic blood pressure (DBP) as the traits. We also used a composite of the 2 phenotypes called the mean arterial pressure, which is defined as (2/3)*DBP + (1/3)*SBP. For the genotype data, we mapped single-nucleotide polymorphisms (SNPs) to the genes; the remaining SNPs that do not belong to any genes, were grouped as intergenic regions. A total of 2286 regions (consisting of 1224 genes and 1062 intergenic regions) that include all the SNPs were defined. The regions were further divided into “rare” or “common” based on minor allele frequency (MAF) threshold of 0.01.

**Association tests**

TOW and VW-TOW are recently proposed methods that allow covariates and account for direction effects for causal variants. Let \( Z_i = (z_{i1}, \ldots, z_{ip})^T \), \( X_i = (x_{i1}, \ldots, x_{iM})^T \) and \( Y_i \) be the covariates, genotype (coded 0, 1, 2) and phenotype for the \( i^{th} \) individual, respectively. The effects of the covariates on \( Y_i \) and \( X_{im} \) are adjusted by the residuals of the following linear models

\[
y_i = \alpha_0 + \alpha_1 z_{i1} + \ldots + \alpha_p z_{ip} + \varepsilon_i
\]

and

\[
x_{im} = \alpha_{0m} + \alpha_{1m} z_{i1} + \ldots + \alpha_{pm} z_{ip} + \tau_{im}.
\]

The methods are based on the optimal weighting scheme, which is defined as \( w_m^p = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_{im} - \bar{x}_m)^2}{\sum_{i=1}^n (\bar{x}_m - \bar{x}_m)^2} \), where \( \bar{y} \) and \( \bar{x}_m \) denote the residuals from equations (1) and (2) for the \( i^{th} \) individual respectively. Let \( x_i^m = \sum_{m=1}^M w_m^p x_{im} \). The test statistics for TOW is defined as \( T_{\text{TOW}} = \sum_{i=1}^n (y_i - \bar{y}) (x_i^m - x^m)^T \). For VW-TOW, let \( T_r \) and \( T_c \) denote the test statistics of TOW for rare and common variants, \( T_r = \lambda \frac{T_r}{\sqrt{\text{var}(T_r)}} + (1 - \lambda) \frac{T_c}{\sqrt{\text{var}(T_c)}} \) and \( \rho_\lambda \) be the \( p \) value of \( T_r \). The test statistics for VW-TOW is defined as \( T_{\text{VW-TOW}} = \min_{0 \leq k < K} \rho_k = \min_{0 \leq k < K} \rho_{k\lambda} \), where \( \lambda_k = k/K \) for \( k = 0, 1, \ldots, K \). The \( p \) values are evaluated by permutation.

After identifying the significant genomic regions, we further investigated the SNPs that have important contribution to the phenotypes for the significant regions by variable selection methods Lasso and SPLS, which are available in the R package: “RV tests.” Because this package does not allow covariates, we adjusted the effect of environmental factors using the linear model shown in equation (1). Instead of the observed trait, the residuals from the linear model are treated as the phenotype.

**A summary of the steps we followed for real data**

- **Step 1:** Map the SNPs to gene and intergenic regions based on the annotation file refGene.txt.gz (available from http://hgdownload.cse.ucsc.edu/). Then the genes or intergenic regions were further divided into subregions (“rare” vs. “common”) based on a threshold of MAF = 0.01.
- **Step 2:** Extract the genotype, phenotype (baseline measures) and covariates (baseline measures) data for the unrelated individuals. Remove the participants that have missing variables in phenotype or covariates data.
- **Step 3:** TOW and VW-TOW are applied to identify the regions that are associated with the traits.
- **Step 4:** Apply Lasso and SPLS to the regions to discriminate the associated variants from noise (using the R package “RV tests”).

**Results**

**Real data**

The sample used in our analysis is made up of 142 independent individuals. After removing missing variables, 129 subjects were analyzed. There are, in total, 1,215,296 markers on chromosome 3; approximately one-sixth of the markers were removed as a result of zero variation across the 129 independent samples.

The association tests (TOW and VW-TOW) were applied to each genetic region for SBP, DBP, and mean arterial pressure (MAP) on chromosome 3. Both tests produce an empirical \( p \) value, based on 10,000 permutations for each region. Figure 1 displays the \( p \) value plot for DBP, where the x-axis denotes the position of the genes in original order on chromosome 3. The \( p \) values for intergenic regions are not included in Figure 1. By parallel comparison, we can see that effects of the genes are caused by the rare variants or the common variants. We note that there is a small cluster of genes that appear highly significant around the 440th region on the upper and lower plots.

After obtaining all the \( p \) values, regions that have strong association with the traits are picked according to the ranking of the \( p \) values. We decided to set the significant level threshold to be 0.001, so as to be more selective. Genes are only selected if they satisfy this criterion for the trait using both TOW and VW-TOW. Table 1 lists the regions that appear to be potentially important. For SBP, there are 3 genes; 2 genes with common variants only and 1 gene with rare variants only are highly associated with the trait. For MAP there are 4 genes when a combined analysis of “rare” and “common” variants is done, and 2 genes are significant with common variants only. For DBP, the number of significant regions is greater than the other 2 traits. For this trait, not only variants that belong to genes, but
also variants in intergenic regions exhibit strong association.

As mentioned earlier, there is a cluster of regions (shown in Figure 1) that show strong significance for DBP. The region names are: TWF2, PPM1M, region between PPM1M and WDR82, WDR82, region between WDR82 and GLYC7K, GLYC7K, region between GLYC7K and DNAH1, BAP1, PHF7, SEMA3G, and TNNC1. The above regions all fall inside the physical location range of (52262625, 52488057).

Then variable selection methods Lasso and SPLS are applied to the regions that are picked at the gene (or region) level. Table 1 also summarizes the numbers of significant markers that were selected using these sparse methods. The number of selected markers can be varied with different choice of penalty parameter.

Simulated data

In stage I, we focused on the top significant genes on chromosome 3, which are MAP4, FLNB, and ABTB1, with common and rare variants combined. We analyzed all 200 replicates with the target genes to assess the power of TOW and VW-TOW. MAP4 has large effect on both SBP and DBP, whereas FLNB and ABTB1 have small effects on SBP only. We adjusted the phenotypes by all the covariates at baseline. Table 2 reports the results. We can see that both methods have very poor power when the variants are all rare in the genes. Table 2 does show, however, that TOW has better performance than VW-TOW in most cases. MAP shows better power than the other 2 phenotypes. In the cases of small effect size, the power is very low for both TOW and VW-TOW.

In stage II, we assessed the performance of Lasso and SPLS by analyzing all 200 replicates on MAP4 with all the variants. There are 6 target SNPs in MAP4, but 1 of the SNPs is removed because of monomorphism. The location numbers of the 5 SNPs are 48040283, 47957996, 47956424, 48040284, and 47913455. Both Lasso and SPLS are variable selection methods. With the careful selection of the penalty parameters for both methods, on average approximately 5 variants are selected with every replicate. Table 3 shows the results. We can see that using MAP as phenotype demonstrates higher power than using SBP or DBP. Lasso and SPLS have very poor power to detect 47956424 and 47913455.

Discussion

Most recently proposed methods assign large weights to rare variants and small weights to common variants, resulting in low power. On the other hand, TOW and
VW-TOW assign corresponding weight, which can account for the direction effect, to individual variants. The methods outperform some currently popular methods, such as Combined Multivariate and Collapsing (CMC) and sequence kernel association test (SKAT), in various scenarios [9]. In addition, both TOW and VW-TOW can be modified to account for population stratification using principal component approach.

Overall, we were able to detect some significant genes based on association tests (TOW and VW-TOW) with

Table 1 Common, rare and total number of variants identified by LASSO, SPLS and by both methods.

| Trait | Region | Common variants | Rare variants | Combined variants |
|-------|--------|-----------------|---------------|------------------|
|       |        | A    | B    | C    | D    | Region | A    | B    | C    | D    | Region | A    | B    | C    | D    |
| SBP   | LAMP3  | 167  | 28   | 21   | 2    | TP63   | 485  | 109  | 94   | 31   | LAMP3  | 267  | 18   | 21   | 2    |
|       | LIMD1  | 371  | 20   | 39   | 15   |        |      |      |      |      |        |      |      |      |      |
| DBP   | Inter-530 | 315 | 72   | 104  | 29   | BAP1   | 24   | 7    | 8    | 7    | BAP1   | 30   | 8    | 9    | 8    |
|       | CTDSP1 | 277  | 20   | 19   | 14   | Inter-378 | 42  | 10   | 10   | 7    | CCCDC66 | 373  | 42   | 57   | 20   |
|       | SLC25A36 | 88   | 7    | 6    | 6    | Inter-896 | 23  | 4    | 8    | 4    | Inter-376 | 10   | 1    | 2    | 1    |
|       |        |      |      |      |      | Inter-898 | 55  | 8    | 14   | 8    |        | 41   | 5    | 8    | 4    |
|       |        |      |      |      |      | CP81   | 52   | 3    | 7    | 2    | Inter-378 | 75   | 11   | 10   | 7    |
|       | DNAH1  | 211  | 40   | 41   | 16   | Inter-896 | 47   | 4    | 8    | 4    |        |      |      |      |      |
|       | E1F5A2 | 34   | 11   | 5    | 5    | Inter-982 | 206  | 27   | 2    | 2    |        |      |      |      |      |
|       | GLYCTK | 13   | 4    | 5    | 4    | CP81   | 164  | 18   | 9    | 5    |        |      |      |      |      |
|       | NCBP2  | 24   | 2    | 2    | 2    | DNAH1  | 326  | 114  | 43   | 31   |        |      |      |      |      |
|       | SEMA3G | 30   | 7    | 8    | 5    | E1F5A2 | 77   | 19   | 5    | 5    |        |      |      |      |      |
|       | TNNC1  | 7    | 3    | 4    | 3    | GLYCTK | 25   | 4    | 5    | 4    |        |      |      |      |      |
|       | WDR82  | 36   | 11   | 12   | 10   | PHF7   | 40   | 3    | 14   | 3    |        |      |      |      |      |
| MAP   | LAMP3  | 167  | 20   | 2    | 2    |        |      |      |      |      |        |      |      |      |      |
|       | LIMD1  | 371  | 10   | 51   | 5    |        |      |      |      |      |        |      |      |      |      |
|       |        |      |      |      |      |        |      |      |      |      |        |      |      |      |      |

A, number of total variants; B, number of variants selected by LASSO; C, number of variants selected by SPLS; D, number of variants selected by both LASSO and SPLS; Inter-376, region between PPMIM & WDR82; Inter-377, region between WDR82 & GLYCTK; Inter-378, region between GLYCTK & DNAH1; Inter-530, region between NXPE3 & LOC152225; Inter-896, region between RPL22L1 & E1F5A2; Inter-898, region between SLC2A2 & TNIK; Inter-982, region between ST6GAL1 & RPL39L

VW-TOW assign corresponding weight, which can account for the direction effect, to individual variants.

Table 2 Power of TOW and VW-TOW to detect MAP4, FLNB, and ABTB1

| Traits     | Power (MAP4) | Power (FLNB) | Power (ABTB1) |
|------------|--------------|--------------|---------------|
|            | TOW          | VW-TOW       | TOW           | VW-TOW       | TOW          | VW-TOW       |
| Gene (combined) |              |              |               |               |              |              |
| SBP        | 0.32         | 0.265        | 0.02          | 0.015         | 0.095        | 0.07         |
| DBP        | 0.325        | 0.26         |               |               |              |              |
| MAP        | 0.435        | 0.35         |               |               |              |              |
| Gene (rare) |              |              |               |               |              |              |
| SBP        | 0.035        | 0.03         | 0.005         | 0.01          | 0.095        | 0.075        |
| DBP        | 0.06         | 0.06         |               |               |              |              |
| MAP        | 0.05         | 0.045        |               |               |              |              |
| Gene (common) |             |              |               |               |              |              |
| SBP        | 0.335        | 0.29         | 0.03          | 0.045         | 0.08         | 0.075        |
| DBP        | 0.345        | 0.245        |               |               |              |              |
| MAP        | 0.445        | 0.33         |               |               |              |              |
SBP, DBP, and MAP. Although we used Lasso and SPLS only as variant selection methods, they can also be used to do the association test for genotype with complex traits. However, both Lasso and SPLS are very computationally intensive. In addition, our analysis is focused on the independent subjects only, which limits our sample size. For future study, it is essential to incorporate family structure that not only increases the size of the sample available for analysis, but also the number of variants. Because SBP and DBP are correlated, it is deficient to analyze them separately. MAP, which is a combination of SBP and DBP, has better power than SBP and DBP separately.

**Competing interests**
The authors declare that they have no competing interests.

**Authors’ contributions**
XFL conducted statistical analyses and drafted the manuscript. JB conceived the study and assisted in drafting the manuscript. All authors read and approved the final manuscript.

**Acknowledgements**
JB would like to acknowledge Discovery Grant funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) (grant number 293295-2009) and Canadian Institutes of Health Research (CIHR) (grant number 84392). JB holds the John D. Cameron Endowed Chair in the Department of Clinical Epidemiology and Biostatistics, McMaster University. The GAW18 whole genome sequence data were provided by the T2D-GENES Consortium, which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW18 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH grants U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW18 were provided by the T2D-GENES Consortium, which is supported by NIH grants U01 DK085526, and U01 DK085545. The Genetic Analysis Workshop is supported by NIH grant R01 GM031575. We would like to thank two anonymous reviewers and the editor for insightful comments that improved the presentation and clarity of our manuscript.

This article has been published as part of BMC Proceedings Volume 8 Supplement 1, 2014: Genetic Analysis Workshop 18. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcproc/supplements/8/S1. Publication charges for this supplement were funded by the Texas Biomedical Research Institute.

**Published:** 17 June 2014

**References**
1. Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, Dehghan A, Glazer NL, Morrison AC, Johnson AD, Aspelund T, et al.: Genome-wide association study of blood pressure and hypertension. Nat Genet 2009, 41:677-687.

2. Fields LE, Burt VL, Cutler JA, Hughes J, Roccella EJ, Sorlie P: The burden of adult hypertension in the United States 1999-2000: a rising tide. Hypertension 2004, 44:398-404.

3. Lawes CM, Vander Hoorn S, Rodgers A: International Society of Hypertension: Global burden of blood-pressure-related disease, 2001. Lancet 2008, 371:1513-1518.

4. Jyengar SK, Elston RC. The genetic basis of complex traits: rare variants or "common gene, common disease"? Methods Mol Biol 2007, 376:71-84.

5. Smith DJ, Luis AJ. The allelic structure of common disease. Hum Mol Genet 2002, 11:2455-2461.

6. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 2008, 40:695-701.

7. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 2001, 69:124-137.

8. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet 2002, 11:2417-2423.

9. Sha Q, Wang X, Wang X, Zhang S. Detecting association of rare and common variants by testing an optimally weighted combination of variants. Genet Epidemiol 2012, 36:561-571.

10. Xu C, Ladouceur M, Dastani Z, Richards JB, Ciampi A, Greenwood C, Yu Z: Multiple regression methods show great potential for rare variant association tests. PLoS One 2012, 7:e41694.

11. Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B 2000, 58:267-288.

12. McCarthy M, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008, 9:356-369.