Bioinformatics and Biology Insights

OPEN ACCESS
Full open access to this and thousands of other papers at http://www.la-press.com.

METHODOLOGY

Simultaneous Analysis of Common and Rare Variants in Complex Traits: Application to SNPs (SCARVAsnp)

Guanjie Chen[1], Ao Yuan[2], Yanxun Zhou[3], Amy R. Bentley[1], Jie Zhou[1], Weiping Chen[4], Daniel Shriner[1], Adebowale Adeyemo[1] and Charles N. Rotimi[1]

[1]Center for Research on Genomics and Global Health, NHGRI, NIH, Bethesda, Maryland, USA. [2]National Human Genome Center, Howard University, Washington DC, USA. [3]Sui Zhou Central Hospital, Sui Zhou, China. [4]National Institute of Diabetes and Digestive and Kidney Diseases, NIH, Bethesda, Maryland, USA.
Corresponding authors email: chengu@mail.nih.gov; rotimic@mail.nih.gov

Abstract: Advances in technology and reduced costs are facilitating large-scale sequencing of genes and exomes as well as entire genomes. Recently, we described an approach based on haplotypes called SCARVA[1] that enables the simultaneous analysis of the association between rare and common variants in disease etiology. Here, we describe an extension of SCARVA that evaluates individual markers instead of haplotypes. This modified method (SCARVAsnp) is implemented in four stages. First, all common variants in a pre-specified region (eg, gene) are evaluated individually. Second, a union procedure is used to combined all rare variants (RVs) in the index region, and the ratio of the log likelihood with one RV excluded to the log likelihood of a model with all the collapsed RVs is calculated. On the basis of previously-reported simulation studies,[1] a likelihood ratio ≥ 1.3 is considered statistically significant. Third, the direction of the association of the removed RV is determined by evaluating the change in λ values with the inclusion and exclusion of that RV. Lastly, significant common and rare variants, along with covariates, are included in a final regression model to evaluate the association between the trait and variants in that region. We apply simulated and real data sets to show that the method is simple to use, computationally effcient, and that it can accurately identify both common and rare risk variants. This method overcomes several limitations of existing methods. For example, SCARVAsnp limits loss of statistical power by not including variants that are not associated with the trait of interest in the final model. Also, SCARVAsnp takes into consideration the direction of association by effectively modelling positively and negatively associated variants.

Keywords: complex traits, rare and common variants

Bioinformatics and Biology Insights 2012:6 177–185

doi: 10.4137/BBI.S9966

This article is available from http://www.la-press.com.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article. Unrestricted non-commercial use is permitted provided the original work is properly cited.

Introduction

Biological and empirical evidence suggests that rare variants (RVs) may account for a significant proportion of the genetic component of several disorders including common complex diseases.[2] It is also believed that better insight into the role of RVs will directly inform our understanding of disease pathophysiology, and that if RVs display sufficiently high penetrance RVs may then have predictive value.[3] However, association analyses of RVs present many challenges, including diminished power, potential biases, and the need for a large sample size. Several of the currently available rare variant analysis methods collapse or group RVs.[4–7] While this approach may help alleviate the problem of small numbers, it may dilute or mask the direction of association by including variants with no effect or that have an effect in a different direction. Recently, Lin and Tang[2] proposed a general framework for detecting disease associations with rare variants in sequencing studies. They employed a score-type statistic (hereafter referred to as the SCORE-TEST). The SCORE-TEST is more powerful and efficient than other methods, but it could not completely resolve the problem of analyzing variants with no effect. Additionally, this method could not simultaneously consider the combined effect of common variants (CVs) and RVs on a trait in a given genomic region. The Sequence Kernel Association Test (SKAT),[8] in contrast, can simultaneously evaluate CVs and RVs. It is a score-based variance-component test that employs a regression method to test for association of variants within a specified region. As with the SCORE-TEST, SKAT could not account for the direction of association of individual variants or exclude variants that are not associated with the trait of interest. As such, implementation of this method produces a single *P*-value for all variants in the region, making it uninformative regarding the specific variants or sets of variants within that region are responsible for observed associations. In this article, we present a modified version of the SCARVA method, in which rare variants are analyzed for association in the context of common variants that influence the trait of interest. The new method (SCARVAsnp), based on the analysis of individual markers instead of haplotypes, is applied to simulated and real datasets. Compared to existing methods, SCARVAsnp has the following advantages: (1) common variants are not ignored; (2) rare variants are sequentially removed from a collapsed rare variant term to maximize the sample size retained in the analysis and avoid biases associated with small numbers; (3) RVs with different directions of association are collapsed separately; (4) RVs that are not associated with the index trait are not included in the final model; and (5) different modes of inheritance can be modeled. Although this method is not directly comparable to existing methods because of differences in definining the analytic unit (ie, the number and type of variants collapsed in the analysis), results from analysis of the same regions using existing methods (SCORE-TEST and SKAT) are provided for context.

The Method

This method modifies and extends the previously-described SCARVA technique, which is haplotyped-based, for use with marker-level data in a given region (which could be defined by a gene, target sequence, window size, or pathway involving multiple genes). A variant is considered rare with a minor allele frequency (MAF) $<$ 5% and with an allele count $\geq$ 5 (RVs with allele acounts $<$ 5 are removed from the analysis). Common variants (CVs) are modeled separately to determine the association of each with the phenotype. RVs, however, are combined using a union method, and the combined effect of all RVs are modeled to overcome RV-associated diminished power. After eliminating variants of no effect, a final regression model is constructed with collapsed positively-associated RVs, collapsed negatively-associated RVs, and covariates (including CVs).

Let $Y = (y_1, \ldots, y_n)'$ be the quantitative trait outcome of n unrelated individuals, with covariates $X = (X_1, \ldots, X_n)'$, where each X_i is a row vector of covariates. $H = (h_1, \ldots, h_n)$ is the observed genotypes for n individuals and $h_j = (h_{j1}, \ldots, h_{jk})$ is the genotype of the j-th individual at the k-th loci. Each variant is assumed to be biallelic. Suppose m of the k genotypes $h_1^c, \ldots, h_m^c$ are common and l of them $h_1^r, \ldots, h_l^r$ are rare ($m + l = k$). Each of the observed h_i is one of the $(h_1^c, \ldots, h_m^c) := H^c$ or one of $(h_1^r, \ldots, h_l^r) := H^r$ $(i = 1, \ldots, n)$.

Let I be the indicator function. The saturated model would be

$$y_i = \mu + \lambda \sum_{j=1}^{l} I\left(h_i = h_j^r\right) + \sum_{j=1}^{m} \alpha_j I\left(h_i = h_j^c\right) + X_i\beta + \in_i, \qquad \left(i = 1, \ldots, n\right) \tag{1}$$

where α_j is the effect of j-th associated $CV\ h_j^c; \lambda$ is the accumulated effects of the associated RVs; $\beta = (\beta_1, \ldots \beta_k)'$ is the effects of all the covariates, and the $\in_i$'s are i.i.d. with $E(\in_i) = 0$ and $\mathrm{Var}(\in_i) = \sigma^2$, which is unknown and is estimated. We model the effects of all the rare variants with a single parameter λ, instead of modeling each of the individual effects with the λ_j's as suggested by others.[9] To simplify notations, let $\alpha = (\alpha_1, \ldots, \alpha_m)$, $\theta = (\mu, \lambda, \alpha, \beta')'$, $1_n = (1, \ldots, 1)'$ of length n, $0_n = (0, \ldots, 0)'$ of length n, I_n be the identity matrix of dimension n, $Z = (1_n, U, V, X)$, and $\in = (\in_1, \ldots, \in_n)'$, where $U = (u_1, \ldots, u_n)'$, $V = (v_{ij})_{1 \le i \le n; 1 \le j \le l}$, $X = (X_1', \ldots, X_n')'$, with

$$u_i = \sum_{j=1}^{l} I\left(h_i = h_j^r\right) \quad \left(i = 1, \ldots, n\right) \text{ and } v_{ij} = I(h_i = h_j^c).$$

Then (1) is re-written as

$$Y = Z\theta + \in, \quad E(\in) = 0_n, \quad Var(\in) = \sigma^2 I_n. \tag{2}$$

The proposed approach for the identification of CVs and RVs that are associated with the trait of interest uses the same basic technique as the SCARVA method.[1] To fit the saturated model (2), the least squares estimate $\hat{\theta}$ of θ under model (2) is

$$\hat{\theta} = \left(\hat{\mu}, \hat{\lambda}, \hat{\alpha}', \hat{\beta}'\right)' = (Z'Z)^{-1} Z'Y$$

and the estimated variance is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2, \quad \hat{y}_i = z_i \hat{\beta},$$

where $z_i = \left(1, u_i, v_i, X_i'\right)$ is the i-th row of Z, and v_i is the i-th row of V. The algorithm contains the following steps.

Step 1: Analysis of CVs

Here we test the significance of the coefficient α_j (j = 1, ..., m) of each CV separately. Of note, the least squares estimate is equivalent to the maximum likelihood estimate under the normal model. Let $\phi(\cdot)$ be the density function of the standard normal distribution, and $l(\theta)$ be the log-likelihood of the data under $\phi(\cdot)$. The hypothesis that the j-th CV is not associated with the trait is represented by H_j: $\alpha_j = 0$. Let $z_{-j} = \left(1_n, U, V_{-j}, X\right) := \left(z_{-j,1}, \ldots, z_{-j,n}\right)'$, where V_{-j} is V with the j-th column removed, and let $\hat{\theta}_{-j} = (\hat{\mu}_{-j}, \hat{\lambda}_{-j} \hat{\alpha}_{-j}, \hat{\beta}_{-j}) = (Z_{-j}' Z_{-j})\, Z_{-j}' Y$ be the least squares estimate of $\theta_{-j} = (\mu, \lambda, \alpha_{-j}, \beta')'$ under H_j, where α_{-j} is α with the j-th component removed, and the estimation of variance under H_j is $\hat{\sigma}_{-j}^2 = (1/(n-2)) \sum_{i=1}^{n} \left(y_i - \hat{y}_{-j,i}\right)^2$, $\hat{y}_{-j,i} = z_{-j,i} \hat{\theta}_{-j}$.

Let χ_1^2 be the centered chi-squared distribution with 1 degree of freedom. If H_j is true, then, approximately,

$$2(l(\hat{\theta}, \hat{\sigma}^2) - l(\hat{\theta}_{-j}, \hat{\sigma}_{-j}^2)) \sim \chi_1^2.$$

Given a significance level of δ, if

$$\Lambda_j := 2(l(\hat{\theta}, \hat{\sigma}^2) - l(\hat{\theta}_{-j}, \hat{\sigma}_{-j}^2)) > \chi_1^2(1-\delta),$$

we reject H_j. When $\chi_1^2(1-\delta)$ is the $(1-\delta)$-th upper quantile of χ_1^2, we accept H_j. After testing all the α_j's ($j = 1, \ldots, m$), remove all the non-significant components of α (the reduced term will still be denoted by α). Let H^c be the collection of all the significantly associated CVs, and let V and Z denote their counterparts with the corresponding columns removed. Re-fit the model in equation (2) with the current Z to get a new estimate of $\theta\,(\hat{\theta} = (Z'Z)^{-1} Z'Y)$. Linear regression models are used to evaluate CVs (MAF $\ge$ 0.05) within a given region. Those CVs with $P < 0.05$ after Bonferroni correction are added to the covariate matrix.

Step 2: Genetic coding of RVs

RVs (those with MAF $<$ 0.05 and observed minor allele counts $\ge$ 5) are coded as dominant and recessive variables. For each coding, collapsed RV parameters are made from the union of all RVs (ie, all recessively-coded RVs are collapsed, all dominantly-coded RVs are collapsed, and the two codings are summed to create a set of additively-coded RVs, then collapsed).

Step 3: Analysis of RVs

Associated RVs can either be positively or negatively associated with the trait; let R^+ and R^- denote collections of these two types of RVs. The association of the positively- and negatively-associated RVs are modeled using different coefficients. First, we test the statistical significance of each RV h_j^r and its effect based on the Z estimated in Step 2. Let H_j' be the hypothesis that h_j^r is not associated with the outcome. Similarly, let $Z_{-j} = (1_n, U_{-j}, V, X)$, where $U_{-j} = \left(u_{-j,1}, \ldots u_{-j,n}\right)', u_{-j,i} = \sum_{k=1,k\neq j}^{l} I(h_i = h_k^r)(i = 1, \ldots, n)$. Let $\hat{\theta}_{-j} = (\hat{\mu}_{-j}, \hat{\lambda}_{-j}, \hat{\alpha}_{-j}, \hat{\beta}_{-j}) = (Z'_{-j}Z_{-j})^{-1} Z'_{-j}Y$ be the least squares estimate of θ under H_j'. The variance under H_j' is estimated as $\hat{\sigma}^2_{-j} = (1/(n-2)) \sum_{i=1}^{n} (y_i - y_{-j,i})^2$, $y_{-,i} = z_{-j,i}\hat{\theta}_{-j}$ (the same notation was used in Step 2). The hypothesis H_j' is not nested within the full model, hence we cannot use the chi-square test (as in Step 2). Instead a version of the Bayesian information criterion (BIC)[10] is used. Let m_j be the number of associated parameters under H_j', using this criterion. The model under H_j' is preferred if $l(\tilde{\theta}_{-j}, \tilde{\sigma}^2_{-j}) - (m_j/2)\log(n)$ is the largest among all $j = 1, \ldots, l$. Here m_j is the same for all values of j, thus, we pick the RVs h_j^r's as associated for those j's where $l(\tilde{\theta}_{-j}, \tilde{\sigma}^2_{-j})$ is larger than the others. Let $\delta_j = | l(\hat{\theta}, \hat{\sigma}^2) - l(\tilde{\theta}_{-j}, \tilde{\sigma}^2_{-j}) | (j = 1, \ldots m)$ and $\bar{\delta} = m^{-1} \sum_{j=1}^{m} \delta_j$. We reject H_j' if there is a big relative increase in δ_j, ie, if

$$\frac{\delta_j}{\bar{\delta}} > \gamma.$$

Based on Yuan et al,[1] the following values for γ: $\gamma = 1.3$, and 1.5 to represent significant and very significant, respectively.

If h_j^r is significant by the above method, and $\hat{\lambda}_{-j} < \hat{\lambda}$ then removing h_j^r resulted in underestimate of the total effect; thus we can deduce $h_j^r \in R^+$.

Thus, we can identify all the positively and negatively associated rare variants. Now let $U = (U^+, U^-)'$, with $U^+ = (u_1^+, \ldots, u_n^+)$ as $u_i^+ = \sum_{h_j^r \in R+} I(h_i = h_j^r), U^- = (u_1^-, \ldots, u_n^-)$ as $u_i^- = \sum_{h_j^r \in R^-} I(h_i = h_j^r)$, V as after Step 2, $Z = (1_n, U, V, X)'$, $\lambda = (\lambda^+, \lambda^-)$ and θ be the corresponding components for Z.

Briefly, first, run regression models of the combined RVs. Second, a reduced model of the RV is implemented by removing one RV at a time and noting the λ and log likelihood values for the reduced models. Then, calculate the λ difference and ratio of the likelihood of full model to that of the reduced model with one of the RV removed at a time. If the ratio of the log likelihood of the full to the reduced models is ≥ 1.3[1], the RV is considered to be associated with the trait. If the λ difference between the full and the reduced model is positive, the excluded RV is negatively-associated; otherwise the RV is positively-associated.

Step 4: Combining associated RVs

Combine all positively-associated RVs into one group, with coefficient λ^+, and H^+, and λ^- and H^- for all the negatively-associated RVs. H^c is the set of all significantly-associated CVs. The final model is

$$y_i = \mu + \lambda^+ \sum_{h_j^r \in H^+} I(h_i = h_j^r) + \lambda^- \sum_{h_j^r \in H^-} I(h_i = h_j^r) + \sum_{h_j^c \in H^c \in} \alpha_j I(h_i = h_j^c) + X_i\beta + \in_i, \quad (i = 1, \ldots, n) \tag{3}$$

Step 5: Final model

Fit the final linear regression model (3), and evaluate the positively- and negatively-associated RVs, along with the significantly-associated CVs. Let $\mathbb{H}^+ : \lambda^+ = 0$; $\mathbb{H}^- : \lambda^- = 0$, and $\mathbb{H}_j : \alpha_j = 0$ for $h_j^c \in H^c$. Chi-squared test statistics can be used to find and report the P-values of the null hypotheses.

Type I Error and Power

Let $H_0 : \lambda = \alpha_j = 0, (j = 1, \ldots, m)$ be the null hypothesis that there is no association of the variants. Let $\hat{\theta}$ be the MLE of θ under the full model, $\hat{\sigma}^2$ be the corresponding variance estimator. $\hat{\theta}_0$ be the MLE of θ under H_0, and $\hat{\sigma}^2_0$ be the corresponding variance estimator. When H_0 is true, we have, asymptotically,

$$\Lambda_0 := 2(l(\hat{\theta}, \hat{\sigma}^2) - l(\hat{\theta}_0, \hat{\sigma}_0^2)) \sim \chi^2_{m+1},$$

where χ^2_{m+1} is the chi-squared distribution with $(m + 1)$ degrees of freedom. Thus, assuming the data is generated under H_0, for a given significance level δ, the type I error is approximated by

$$\gamma = \gamma(\delta) = P_{H_0}(\Lambda_0 \geq \chi^2_{m+1}(1-\delta)) \approx \delta.$$

Thus, the test is asymptotically unbiased.

The power will depend on the magnitude of the effect size. Let $\eta = (\lambda^2 + \alpha_1^2 + \cdots + \alpha_m^2)/\hat{\sigma}$, and Λ_0 as given above. Assuming the data are generated not from H_0, but from H_1, then asymptotically,

$$\Lambda_0 \sim \chi^2_{m+1,\eta},$$

where $\chi^2_{m+1,\eta}$ is the chi-squared distribution with $(m + 1)$ degrees of freedom (where m is number of associated CVs) and non-centrality parameter η. The power for a given δ at η is approximated by

$$\rho = \rho(\delta,\eta) = P_{H_1}(\Lambda_0 \geq \chi^2_{m+1}(1-\delta))$$
$$\approx P(\chi^2_{m+1,\eta} \geq \chi^2_{m+1}(1-\delta)).$$

For a given δ and η, $\rho(\delta, \eta)$ can be determined using a table of the non-central chi-squared distribution. Figure 1 gives the $\rho(\delta, \eta)$ values when η goes from 1 to 20, $\delta = 0.05$, 0.025 and 0.01, and $m = 2$, 4, 6, 8, and 10.

Simulation Study

We conducted a range of simulations based on a set of 4,000 observed quantitative traits, covariates, and corresponding alleles within a given region; for brevity, we present the results from one of these simulation exercises. Ten CVs and 10 RVs were simulated, with frequencies of $(p; q) = (p_1, \ldots, p_{10}; q_1, \ldots, q_{10}) = (0.075, 0.115, 0.130, 0.060, 0.220, 0.085, 0.105, 0.050, 0.015, 0.095; 0.008, 0.007, 0.006, 0.005, 0.005, 0.008, 0.009, 0.007, 0.008, 0.009)$, where $p_{1,\ldots,10}$ denote the frequencies of $\text{CVs}_{1 \text{ to } 10}$, and $q_{1,\ldots,10}$ denote the frequencies of $\text{RVs}_{1 \text{ to } 10}$. For the RVs, let $H^r = R^+ \cup R^-$, with $R^+ = \{h_2^r, h_3^r, h_{10}^r\}$ (representing all positively-associated RVs) and effect sizes $(\lambda_2^+, \lambda_3^+, \lambda_{10}^+) = (0.42, 0.52, 0.62)$, and with $R^- = \{h_5^r, h_8^r\}$ (representing all negatively-associated RVs) with effect sizes $(\lambda_5^-, \lambda_8^-) = (-0.53, -0.49)$. For the CVs, we define the collection of associated variants as $H^c = \{h_3^c, h_7^c\}$, with effects $\alpha_3 = 0.45$, and $\alpha_7 = 0.50$, thus $\alpha_j = 0$ $(j \neq 3, 7)$. Covariates are contained in $X = (x_1, x_2, x_3)$ = (gender, age, Body Mass Index (BMI)), where gender has the value of 0 or 1 with 0.5 probability of each, age (years) is uniformly distributed [10,70], and BMI values are uniformly distributed [12,42]. The effect sizes of covariates are $\beta = (\beta_1, \beta_2, \beta_3) = (0.0167, 0.008, 0.120)$. Given the genotypes and covariates, the quantitative trait follows the normal $N(1.5, 2)$ distribution.

Using SCARVAsnp, the joint analysis detected significant associations for two CVs (v_3 and v_7, $P < 0.0001$ for both), positively-associated RVs (u_2, u_3, u_{10}, $P < 0.0001$) and negatively-associated RVs (u_5, u_8, $P < 0.0001$) as displayed in Table 1.

Data Analysis for DHS Application

We used the SCARVAsnp method to analyze the association of genotypes to plasma triglyceride (TG) in the Dallas Heart Study (DHS).[11] In the DHS, angtiopoietin-like (*ANGPTL*) genes 3, 4, and 5 were sequenced in 3551 participants. Multiple rare non-synonymous (NS) sequence variants in these genes have been reported to be associated with lower plasma TG levels[11] based on a Wilcoxon rank-sum test (with adjustments for age and gender). Using SCARVAsnp, the joint analysis detected significant associations were displayed in Table 2. For comparison, we analyzed the

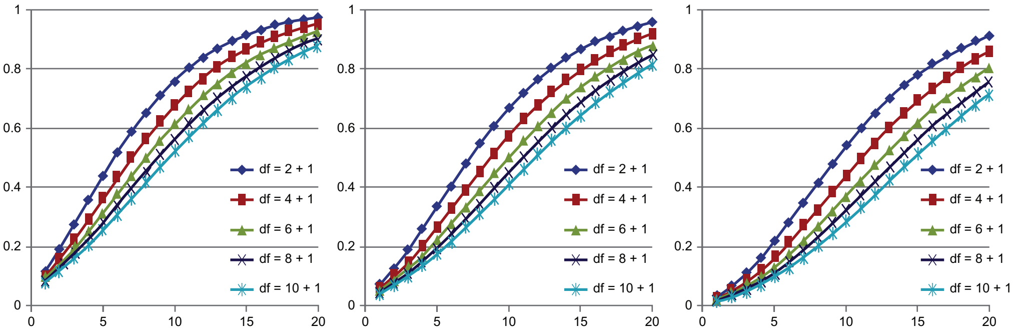

Figure 1. Power (Y-axis) and non-centrality parameters (X-axis) for different values of δ (from left to right, panels represent δ of 0.05, 0.025, and 0.01).
Note: The degrees of freedom (df) are indicated by the color of the lines.

Table 1. Association analysis of the simulated sequence data set using SCARVAsnp.

Type of variant	Sig. CVs/RVs	Single	Joint analysis	*P*-value
		P/ratio	β or λ (SE)	
v*($n = 10$)	v_3	<0.0001	0.46 (0.02)	<0.0001
	v_7	<0.0001	0.47 (0.02)	<0.0001
u**($n = 10$)	Rare (+)			
	u_2	1.38		
	u_3	1.30		
	u_{10}	2.36	0.61 (0.04)	<0.0001
	Rare (–)			
	u_5	1.92		
	u_8	1.69	–0.57 (0.05)	<0.0001

Notes: Sig. CVs/RVs—statistically significant common and rare variants. *Total number of CVs analyzed; **total number of RVs analyzed. Rare (+): Positively-associated RVs. Rare (–): Negatively-associated RVs. β: regression coefficients for CVs. λ: regression coefficients for collapsed RV terms.

DHS data set using the recently published SCORE-TEST method[2] as well as SKAT.[8] We implemented the T_5, F_p, V_T, and T_{max} variations of the SCORE-TEST for the analysis of the simulated and real datasets, but for brevity, we present only the results of the T_5. The results of these analyses are displayed in Figure 2 (SCORE-TEST) and Table 3.

Three CVs and 15 RVs in *ANGPTL3* were analyzed (Table 2). CV ANG3-008357 (MAF = 0.40) was associated with TG levels ($P = 2.17 \times 10^{-7}$ in the joint analysis). There were 3 RVs with ratio values ≥ 1.3: ANG3-005308 and ANG3-005424 were negatively associated with TG ($P = 2.86 \times 10^{-7}$) and ANG3-004520 was positively associated with TG levels ($P = 1.99 \times 10^{-2}$). RVs ANG3-005308, ANG3-005424, and ANG3-004520 also had the top scores (4.61, 2.73 and 2.10, respectively) using the SCORE-TEST method, and variants in this gene were associated with TG using the SKAT method as well ($P = 3.28 \times 10^{-7}$).

Four CVs and 28 RVs were included in the analysis of *ANGPTL4* (Table 2). Two CVs (ANG4-010707 and ANG4-009155) were associated with TG levels ($P = 2.75 \times 10^{-5}$ and 1.54×10^{-2}, respectively). Five RVs (ANG4-006052, ANG4-009191, ANG4-001175, ANG4-010620, and ANG4-006175) had high ratio values (consistent with the top scores from SCORE-TEST, Figure 2). All 5 RVs were negatively associated with TG levels (P-value $= 1.00 \times 10^{-20}$, Table 2). An association with this gene was also identified using the SKAT method ($P = 3.78 \times 10^{-23}$).

Four CVs and 30 RVs were included in the analysis of *ANGPTL5*. No CVs were associated with TG levels. Four RVs (ANG5-014661, ANG5-011617, ANG5-022751, and ANG5-012530) were negatively

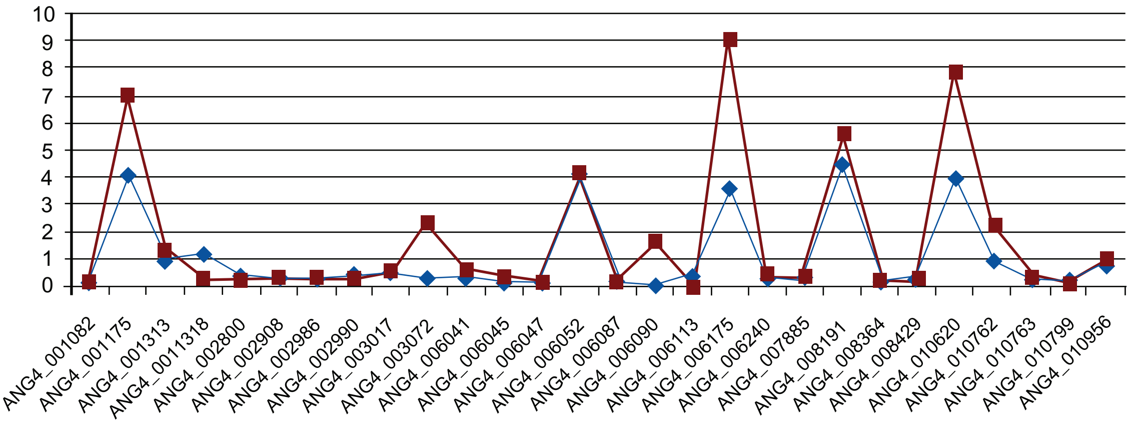

Figure 2. The distribution of the ratio of the log likelihood values from SCARVAsnp (blue line), and the distribution of scores from SCORE-TEST (red line) for rare variants in *ANGPT4* in the Dallas Heart Study (DHS).

Table 2. Results of the DHS sequence data for 3 lipid genes using SCARVAsnp.

Genes	SNPs	MAF	Sig. CVs/RVs	*P*/ratio	Joint model
ANGPTL3	ANG3-008357	0.40	Common	5.44×10^{-10}	2.17×10^{-7}
($v^* = 3$, $u^{**} = 15$)	ANG3-005424	0.01	Rare (–)	2.56	
	ANG3-005308	0.02	Rare (–)	5.21	2.86×10^{-7}
	ANG3-004520	0.01	Rare (+)	1.86	1.99×10^{-2}
ANGPTL4	ANG4-010707	0.06	Common	1.50×10^{-7}	2.75×10^{-5}
($v = 4$, $u = 28$)	ANG4-009155	0.28	Common	5.09×10^{-3}	1.54×10^{-2}
	ANG4-006052	0.03	Rare (–)	3.88	
	ANG4 009191	0.03	Rare (–)	4.44	
	ANG4-001175	0.04	Rare (–)	4.08	
	ANG4-010620	0.04	Rare (–)	3.98	
	ANG4-006175	0.04	Rare (–)	3.61	1.00×10^{-20}
ANGPTL5	ANG5-014661	0.01	Rare (–)	1.69	
($v = 4$, $u = 30$)	ANG5-011617	0.02	Rare (–)	2.67	
	ANG5-022751	0.02	Rare (–)	1.58	
	ANG5-012530	0.04	Rare (–)	1.32	2.20×10^{-4}
	ANG5-026244	0.01	Rare (+)	3.22	
	ANG5-012581	0.01	Rare (+)	2.59	
	ANG5-017106	0.03	Rare (+)	5.58	6.66×10^{-6}

Notes: *Total number of CVs analyzed; **total number of RVs analyzed. Rare (+): Positively-associated RVs. Rare (–): Negatively-associated RVs.
Abbreviations: DHS, Dallas Heart Study; Sig. CVs/RVs, statistically significant common and rare variants; *P*/ratio, *P* values for CVs or ratio values for RVs.

associated ($P = 2.24 \times 10^{-4}$), and 3 RVs (ANG5-026244, ANG5-012581, and ANG5-017106) were positively associated with TG levels ($P = 6.66 \times 10^{-6}$; Table 2). The 7 RVs also received the top scores using the SCORE-TEST method, and this gene was also associated with TG when the SKAT method was used ($P = 2.01 \times 10^{-7}$).

While all of these methods were identified an association between each of these genes and TG, there are notable differences in the inferences possible given the output given from each method. SCARVAsnp provides a separate *P*-value for the union of all positively-associated RVs, the union of all negatively-associated RVs, and for each CV (Table 3). In contrast, the SCORE-TEST provides one *P*-value for all RVs evaluated in the region, and SKAT produces a single *P*-value all CVs and RVs together.[8] Thus, SCARVAsnp supports conclusions

Table 3. Results of simulated and DHS data comparing SCARVAsnp, SCORE-TEST, and SKAT.

Data sets	Total # of SNPs	SCARVAsnp	SCORE-test (T_5***)	SKAT****
Simulated	$v^* = 10$	$v_3 < 0.0001$		2.02×10^{-66}
		$v_7 < 0.0001$		
	$u^{**} = 10$	Rare (+) < 0.0001	0.000044	
		Rare (–) < 0.0001		
ANGPTL3	$v = 3$	ANG3_008357 = 2.17×10^{-7}		3.28×10^{-7}
	$u = 15$	Rare (+) = 1.99×10^{-2}	0.008470	
		Rare (–) = 2.86×10^{-7}		
ANGPTL4	$v = 4$	ANG4_010707 = 2.75×10^{-5}		3.78×10^{-23}
		ANG4_009155 = 1.54×10^{-2}		
	$u = 28$	Rare (+) = N/A	0.000001	
		Rare (–) = 1.00×10^{-20}		
ANGPTL5	$v = 4$	*P*-value > 0.05		2.01×10^{-7}
	$u = 30$	Rare (+) = 6.66×10^{-6}	0.015066	
		Rare (–) = 2.20×10^{-4}		

Notes: *Total number of CVs analyzed; **total number of RVs analyzed; ****P*-value for the set of RVs with MAF < 5%; *****P*-value for the set of CVs and RVs.
Abbreviation: DHS, Dallas Heart Study.

about associations at a finer level of detail than existing methods, and avoids the potential pitfalls of collapsing variants that differ in their level of association (ie, including variants of no effect) or in the direction of that assocation.

Discussion

We have extended the haplotype-based simultaneous common and rare variant analysis (SCARVA) method to be used with marker-level data (SCARVAsnp). The method is easy to use and could be computationally efficient, as illustrated in the analysis of simulated and real datasets.

In general, RV analysis is challenged by a low number of observations with the relevant genotype, and the resulting decreased power to detect associations. To solve this problem, grouping strategies have been used, which increase the number of observations with relevant genotypes and reduce the number of tests needed by evaluating collapsed markers simultaneously. It is known that not all RVs within a given region have an effect, grouping all RVs regardless of effect may produce an unsatisfactory signal-to-noise ratio. Incorporating information from prior studies or background population variation to grouping strategies has been suggested.[12] In practice, it is unlikely to find this information for a particular disease in a specific population, of significance given that rare variation is less shared across populations than is common variation.[3,13] The present method is implemented in three stages to deal with issues above. First, RVs are collapsed and the log-likelihood of a model with this collapsed term is compared to the log likelihood of a model with a collapsed term that excludes one of the RVs. After each of the RVs has been evaluated in this manner, the RVs for which the reduced model does not significantly differ from the full model are excluded from the analysis as having no effect. Secondly, RVs are grouped according to the direction of association, with separate groups for positively- and negatively-associated RVs. Thirdly, the group of positively-associated RVs, the group of negatively-associated RVs, associated CVs, and covariates are analyzed together in a joint analysis. This strategy significantly reduces the number of tests, resulting in a large number of degrees of freedom. Our method identified the same RVs as were identified using the SCORE-TEST method for a simulated and a real data set, indicating that the method is as accurate as the SCORE-TEST method for RVs.

In our simulation, CVs (v_3 *and* v_7) explained 24% of the total variance, while positive and negative RVs explained only 6% of the total variance alone (combined, CVs and RVs explained 29% of total variance). These findings suggest that the combined analysis of RVs and CVs may be important in explaining the so-called missing heritability from GWAS analyses. A benefit of SCARVAsnp for RV identification is that it treats CVs as covariates, acknowledging the potential contribution of both types of variation in the outcome. Future work will focus on the interaction between CVs and RVs at different levels of linkage disequilibrium.

Conclusions

The modified SCARVA method (SCARVAsnp) combines rare variants (RVs) for association analysis to avoid the problem of low number of observations with relevant genotypes. It eliminates variants with no effect within a given region, separately analyzes positively- and negatively-associated RVs, and allows the adjustment for common variants and covariates. SCARVAsnp was used to analyze simulated and well known data sets. In all cases, it identified the RVs that produced the highest scores in the SCORE-TEST, while eliminating the costly effect of analyzing markers with no clear association with the trait of interest. Notably, and in contrast to the SCORE-TEST and SKAT, SCARVAsnp provides *P*-values for the union of all positively- and all negatively-associated RVs separately.

Acknowledgement

This work is supported in part by the National Center for Research Resources at NIH grant 2G12RR003048.

Author Contributions

Conceived and designed the experiments: GC, AY, AA, CR. Analysed the data: JZ, GC, AY. Wrote the first draft of the manuscript: GC, AY, WP, YZ. Contributed to the writing of the manuscript: AB, DS, AA, RC. Agree with manuscript results and conclusions: GC, AY, YZ, AB, JZ, WC, DS, AA, CR. Jointly developed the structure and arguments for the paper: GC, AY, AB, AA, DS, CR. Made critical revisions and approved final version: GC, AY, AB, CR. All authors reviewed and approved of the final manuscript.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Yuan A, Chen G, Zhou Y, Bentley A, Rotimi C. A novel approach for the simultaneous analysis of common and rare variants in complex traits. *Bioinform Biol Insights*. 2012;6:1–9.
2. Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet*. 2011;89: 354–67.
3. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome wide association. *PLoS Biol*. 2000;8:e1000294.
4. 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;615:28–56.
5. Li B, Leal SM. Methods for detecting associations with rare variants for common disease: application to analysis of sequence data. *Am J Hum Genet*. 2008;83:311–21.
6. Madsen BE, Browning SR. A groupwide association test for rare mutations using a weighted sum statistic. *PLoS Genet*. 2009;5:e1000384.
7. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, et al. Pooled Association Test for Rare Variants in Exon-Resequencing Studies. *Am J Hum Genet*. 2010;86:832–8.
8. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare variant association testing for sequencing data using the sequence Kernel association test (SKAT). *Am J Hum Genet*. 2011;89:82–93.
9. Morries AP, Zeggini E. An Evaluation of Statistical Approaches to Rare Variant Analysis in Genetic Association Studies. *Genet Epidemiol*. 2010;34:188–93.
10. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6:461–4.
11. Romeo S, Yin W, Kozlitina J, Pennacchio AL, Boerwinkle E, et al. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J Clin Invest*. 2009;119:70–9.
12. Zhang L, Pei YF, Li J, Papasian JC, Deng HW. Improved detection of rare genetic variants for disease. *PLoS One*. 2010:e13857.
13. The international HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010;467:52–8.