Bioinformatics and Biology Insights

METHODOLOGY

A Novel Approach for the Simultaneous Analysis of Common and Rare Variants in Complex Traits

Ao Yuan[1], Guanjie Chen[2], Yanxun Zhou[3], Amy Bentley[2] and Charles Rotimi[2]

[1]National Human Genome Center, Howard University, Washington DC, USA. [2]Center for Research on Genomics and Global Health, NHGRI, NIH, Bethesda, Maryland. [3]Suizhou Central Hospital, Sui zhou, P.R. China, 441300.
Corresponding authors email: ayuan@howard.edu; rotimic@mail.nih.gov

Abstract: Genome-wide association studies (GWAS) have been successful in detecting common genetic variants underlying common traits and diseases. Despite the GWAS success stories, the percent trait variance explained by GWAS signals, the so called "missing heritability" has been, at best, modest. Also, the predictive power of common variants identified by GWAS has not been encouraging. Given these observations along with the fact that the effects of rare variants are often, by design, unaccounted for by GWAS and the availability of sequence data, there is a growing need for robust analytic approaches to evaluate the contribution of rare variants to common complex diseases. Here we propose a new method that enables the simultaneous analysis of the association between rare and common variants in disease etiology. We refer to this method as SCARVA (simultaneous common and rare variants analysis). SCARVA is simple to use and is efficient. We used SCARVA to analyze two independent real datasets to identify rare and common variants underlying variation in obesity among participants in the Africa America Diabetes Mellitus (AADM) study and plasma triglyceride levels in the Dallas Heart Study (DHS). We found common and rare variants associated with both traits, consistent with published results.

Keywords: association, common variant, haplotype, rare variant

Bioinformatics and Biology Insights 2012:6 1–9

doi: 10.4137/BBI.S8852

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

Introduction

Genome-wide association studies (GWAS) have proved to be an important tool for the identification of common genetic variants associated with many complex diseases and traits.[1,2] Notably, however, the collection of variants identified so far through GWAS explain only a small fraction of the heritability estimated from family studies for any particular disease or trait.[3,4] It has been suggested that this "missing heritability" is due to the collective effects of rare variants which are usually unaccounted for in GWAS. The "rare variant hypothesis"[5–7] proposes that a significant proportion of the inherited susceptibility to relatively common human chronic diseases may be due to the cumulative effects of a series of low frequency dominantly and independently acting variants at different genes, each conferring a moderate, but detectable, increase in relative risk. It is believed that such rare variants will mostly be population-specific, because of founder effects resulting from genetic drift. Data from published results show that the effects of rare variants tend to be larger than those of common variants. For example, only a handful of risk estimates for common variants (ie, frequency ≥5%) exceeds 2 with the majority falling between 1.1 and 1.4.[8,9] In contrast, rare variants tend to have risk estimates that are larger than 2. Moreover, it is believed that associated rare variants are more likely to be causal.[7,9] A comprehensive review of current understanding of the allelic complexity of human disease genes is provided by Smith and Lusis.[10] In addition, Bodmer and Bonilla[7] provided a historical review of the search for genetic variants influencing susceptibility of an individual to a chronic disease, from R.A. Fisher's seminal work to the current progress of whole-genome association studies.

The current thinking about the contribution of rare variants to complex diseases and traits has motivated the development of new analytic tools. Li and Leal[11] developed combined multivariate and collapsing and kernel based adaptive cluster methods to test for rare variant associations with complex traits. Price et al[12] considered a method for the analysis of rare variants. Other approaches have been proposed by Grady et al,[13] Morris and Zeggini[14] and Zhu et al.[15] McClellan et al[16] summarized evidence for rare alleles responsible for Schizophrenia, Shental et al[17] proposed a method based on compressed sequences. Notably, all of these approaches are based on the separate analysis of common and rare variants. However, we believe that the most efficient strategy to localize disease/trait variants will involve approaches that can identify both common and rare variants in the same model. Also, the method should distinguish between significant rare variants that increase risk and those that are protective. We present such an approach in this study.

The Method

Our method uses quantitative trait data with typed haplotypes and covariates from unrelated individuals. The term "rare variant" seems to lack a common definition; some define it as a variant with a minor allele frequency less than 1%, but with non-negligible effect, residing in a functional unit, such as a gene.[18] Here we define a rare variant as a haplotype with population frequency less than 1%. In this study, genomic loci (eg, genes or chromosomes) are first partitioned into haplotypes, defined as a consecutive strings of SNPs transmitted together from parents to offspring, using existing methods (for example, HapLink, the HapMap website.[19–21] The association of common haplotypes are modeled separately, while the combined association of all rare haplotypes is modeled, to overcome the problem of a low number of observations. The proposed method is a joint regression model with common and rare alleles as covariates, along with other covariates.

We refer to this method as SCARVA (simultaneous common and rare variants analysis).

Let $Y = (y_1, \ldots, y_n)'$ be the quantitative traits of n unrelated individuals, with covariates $X = (X_1, \ldots, X_n)'$, where each X_i is a row vector of covariates, and $H = (h_1, \ldots, h_n)$ is the observed haplotypes for the n individuals at a given genomic loci. Suppose that in the population under consideration, there are $m+1$ different haplotypes (or alleles, in a simpler terminology) out of which $h_1^c, \ldots, h_m^c$ are common and $h_1^r, \ldots, h_l^r$ rare. Each of the observedc or 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 $p = (p_1, \ldots, p_m)$ and $q = (q_l, \ldots, q_l)$ be the population frequencies of H^c and H^r respectively (note $\sum_{i=1}^{m} p_i + \sum_{j=1}^{l} q_j = 1$). Since haplotypes can contain several SNPs, the computational burden using haplotypes instead of individual SNPs is expected to be several fold less; however, the results from SNP based analyses are likely to have

higher resolution. Also, as SNPs in a given haplotype tend to be transmitted together from parents to offspring, they are usually highly correlated to each other. Thus by searching for risk variants in a haplotype instead of individual SNPs, the proposed method (SCARVA) significantly reduces computational burden several fold depending on the size of the analysis dataset.

A standard method of analyzing quantitative phenotype in the presence of covariates is regression. First, we describe a regression model in which the effects of all rare alleles are modeled by a single parameter. Due to the expectation that some rare variants will be positively associated, while others will be inversely associated, we first identified the direction of the association (Step III below) in a single effect model and then modeled the positive and negative associations using different parameters. This modeling strategy minimizes the loss of power that is likely to result from the single effect model and simultaneously analyzes rare variants that are positively and negatively associated with the underlying trait(s). Also, the proposed stepwise regression approach effectively addresses a major limitation of most existing rare variants analysis, which is the combined analysis of non-functional and functional variants with the resulting loss of statistical power.

Let I be the indicator function, the saturated model would be

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

where α_j is the effect of j-th common allele h_j^c; λ is the cumulative effect of the rare alleles (haplotypes); $\beta = (\beta_1, \ldots, \beta_k)'$ is the effect of the covariates, and the $\in_i$'s are i.i.d. with $E(\in_i) = 0$ and $Var(\in_i) = \sigma^2$, which is unknown and is estimated. Due to the fact that observed haplotypes are likely to be individually rare, we model the effects of all the rare variants with a single parameter λ (as in Morris and Zeggini).[14] λ is approximately the mean effect of the individual effects λ_j's

$$\lambda \approx \sum_{j=1}^{l} \lambda_j q_j / q,$$

where $q = \sum_{j=1}^{l} q_j$.

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 $U = (u_1, \ldots, u_n)'$, $V = (v_{ij}) 1 \le i \le n; 1 \le j \le l-1$ (note for each fixed i, we have $\sum_{j=1}^{l} v_{ij} = 1$, so we only use the first $l - 1$status v_{ij}'s, otherwise the matrix $Z'Z$ will be singular), $X = (X_1', \ldots, X_n')'$, with

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

Then (1) is re-written as

$$Y = Z\theta + \in,\ E(\in) = 0_n,\ Var(\in) = \sigma^2 I_n \qquad (2)$$

So the proposed approach for the identification of common and rare variants that are associated with the trait of interest consists of several steps as discribed below.

Step I. Fit the saturated model (2)

The least squares estimate $\hat{\theta}$ of θ under model (2) is

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

and the estimated variance is

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

where $z_i = (1, u_i, v_i, X_i')$ is the i-th row of Z, and v_i is the i-th row of V.

Step II. Analysis of common risk allele(s)

Here we test the significance of the coefficient α_j (j = 1, …, m) of each allele 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 $\phi(\cdot)$. Denote the hypothesis of no effect of the j-th common haplotype as H_j: $\alpha_j = 0$. Let $Z_{-j} = (1_n, U, V_{-j}, X) := (z_{-j,1}, \ldots, z_{-j,n})'$, where V_{-j} is V with the j-th column removed, and let $\hat{\theta}_{-j} = (\hat{\mu}_{-j}, \hat{\lambda}_{-j}, \hat{\alpha}_{-j}, \hat{\beta}_{-j}) = (Z'_{-j} Z_{-j})^{-1} Z'_{-j} Y$ 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 $\hat{\sigma}^2_{-j} = 1/n - 2\sum_{i=1}^{n}(y_i - \hat{y}_{-j,i})^2$, $\hat{y}_{-j,i} = z_{-j,i}\hat{\theta}_{-j}$.

Let χ^2_1 be the centered chisquared distribution with 1 degree of freedom. If H_j true, then approximately

$$2(l(\hat{\theta}, \hat{\sigma}^2) - l(\hat{\theta}_{-j}, \hat{\sigma}^2_{-j})) \sim \chi^2_1.$$

Given a significance level of δ (often $\delta = 0.01$, 0.02 or 0.05), if

$$\Lambda_j := 2(l(\hat{\theta}, \hat{\sigma}^2) - l(\hat{\theta}_{-j}, \hat{\sigma}^2_{-j})) > \chi^2_1(1-\delta),$$

we reject H_j. Where $\chi^2_1(1 - \delta)$ is the $(1 - \delta)$-th upper quantile of χ^2_1, we accept H_j.

After testing all the α_j's ($j = 1, \ldots, m$), remove all the non-significant components of α (it is still denoted it as a to simplify notation). Let H^c be the collection of all the risk common haplotypes, 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 the new estimate of θ (still denoted as $\hat{\theta} = (Z'Z)^{-1}Z'Y$).

Step III. Analysis of rare allele(s)

The risk rare alleles are of two types: alleles that are positively associated with the trait of interest (ie, contributes positively to the effect λ), and alleles that are negatively associated with the trait (ie, contributes negatively to the effect of λ). Let R^+ and R^- denote the collection of these two types of rare variants in a given haplotype. We propose modeling the effects of the positive and negative rare variants using different coefficients. First, we identify the rare variants in R^+ and R^- respectively. To achieve this goal, we test the significance of each rare allele h^r_j and its effect based on the Z estimated from Step II. Let H'_j be the hypothesis: h^r_j is non-risk. Similarly, let $Z_{-j} = (1_n, U_{-j}, V, X)$, where $U_{-j} = (u_{-j,i}, \ldots, u_{-j,n})' u_{-j,i} = \sum_{k=1,k\neq j}^{l} I(h_i = h^r_k)$ $(i = 1, \ldots, n)$. 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, and the variance under H'_j can be 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 II). In this step, however, the hypothesis H'_j is not nested within the full model, hence we cannot use the chisquare test as in step II. Instead we use a version of the Bayesian information criterion (BIC).[22] BIC and the related AIC criteria have been used extensively in statistical applications. Let m_j be the number of parameters under H'_j, by this criterion. Model under H'_j is preferred if $l(\tilde{\theta}_{-j}, \tilde{\sigma}^2_{-j}) - m_j/2)\log(n)$ is largest among all $j = 1, \ldots, l$. Here m_j is the same for all j, thus, we pick the rare alleles h^r_j's as risk for those j's where $l(\tilde{\theta}_{-j}, \tilde{\sigma}^2_{-j})$ is larger than the others. Let $\delta_j = |l(\tilde{\theta}, \tilde{\sigma}^2) - l(\tilde{\theta}_{-j}, \tilde{\sigma}^2_{-j})|$ $(j = 1, \ldots, m)$, 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 our simulation studies and with the assumption that risk rear variants generally account for no more than 30% of all real variants), we recommend the following values for γ: $\gamma = 1.1$, 1.3 and 1.5 to represent somewhat significant, significant and very significant.

If h^r_j is significant by the above method, and $\hat{\lambda}_{-j} < \hat{\lambda}$ then removing h^r_j resulted in underestimate of the total effect; thus we can deduce $h^r_j \in R^+$. If h^r_j is not significant, then $h^r_j \in R^-$.

Thus, we can identify all the positively and negatively associated rare variants in a given haplotype. Now let $U = (U^+, U^-)'$, with $U^+ = (u^+_1, \ldots, u^+_n)$ as $u^+_i = \sum_{h^r_j \in R^+} I(h_i = h^r_j)$, $U^- = (u^-_1, \ldots, u^-_n)$ as $u^-_i \sum_{h^r_j \in R^-} I(h_i = h^r_j)$ V as after Step II, $Z = (1_n, U, V, X)'$, $\lambda = \lambda^+$, λ^- and θ be the corresponding components for Z.

Note: if an analysis locus has only 1 rare allele, it may not be meaningful to analyze it because the corresponding number of observations will be too few to make reliable conclusion.

Step IV. Fit the final model

Now with Z from Step III, we re-fit the following model

$$Y = Z\theta + \in$$

The least squares estimates of the parameter $\theta, \hat{\theta} = (Z'Z)^{-1}Z'Y$, is the final characterization of the associations of all the risk rare and common variants.

Simulation Study

We simulated a range of datasets with varied parameter values and different numbers of variants, and used SCARVA to analyze the generated data. In this simulation we sampled data sets based on a set of 4,000 observed quantitative traits, covariates, and corresponding alleles within a given haplotype region; for brevity, we present the results from one of these simulation exercises. We simulated observed haplotypes directly, without simulating genotypes and constructing haplotypes by existing methods. The simulated haplotype region contained 20 alleles, with the first 10 designated as common and the last 10 as rare, with frequencies $(p;\ q) = (p_1, \ldots, p_{10};\ q_1,\ldots,q_{10}) = (0.075, 0.115, 0.130, 0.060, 0.220, 0.085, 0.105, 0.050, 0.015, 0.095;\ 0.003, 0.007, 0.006, 0.004, 0.005, 0.004, 0.006, 0.003, 0.005, 0.007)$. Among the rare variants, we define $H^r = R^+ \cup R^-$, with $R^+ = \{h_2^r, h_3^r, h_{10}^r\}$ representing all positively associated rare alleles, with effects $(\lambda_2^+, \lambda_3^+, \lambda_{10}^+) = (0.45, 0.50, 0.48)$; and $R^- = \{h_5^r, h_8^r\}$ representing all negatively associated rare alleles, with effects $(\lambda_5^-, \lambda_8^-) = (-0.53, -0.49)$. Thus the overall effect of rare variants is $\lambda^+ = \lambda_2^+ q2/q + \lambda_3^+ q3/q + \lambda_{10}^+ q10/q = 0.4755$ and $\lambda^- = \lambda_5^- q5/q + \lambda_8^- q8/q$; let $\lambda = (\lambda^+, \lambda^-)$. Among the common variants, we define the collection of risk variants as $H^c = \{h_3^c\}$, with effects $\alpha_3 = 0.37$, thus $\alpha_j = 0$ $(j \neq 3)$. The covariates are $X = (x_1, x_2, x_3) =$ (gender, age, body mass index [BMI]), where gender takes values 0 or 1 with probability 0.5 each, age in years is uniformly distributed [10,70], and BMI values are uniformly distributed (12,42). The effects of covariates are $\beta = (\beta_1, \beta_2, \beta_3) = (0.0167, 0.008, 0.120)$. Given the haplotype and the covariates, the quantitative trait follows the normal $N(1.5, 2)$ distribution.

For each of the individual observations $(n = 2000$ y_i's), we generated haplotypes using the probabilities and covariates as described above. Using the simulated data sets, we generated y_i from (1) with $\mu = 1.5$ and $\sigma^2 = 2$. We then use the algorithms described in Steps I–IV to detect rare and common risk haplotypes. The results of these analyses for both the common and rare allele are displayed below (Table 1). As all the 10 common alleles satisfy a linear constraint under the model, we only show the results for the first 9 alleles. For each common allele j, the Λ_j values and the corresponding chisquare P-values in Step II are displayed to show how some common alleles are removed from the model. For each rare allele, the ratio $\delta_j / \bar{\delta}$ in Step III are given to show how some rare alleles are removed from the model. Estimates of the regression parameters in the final model, as in Step IV, (standard errors in brackets) are displayed in Table 1.

We correctly identified the common risk allele 3 with a P-value of almost zero. All other common alleles, simulated to be low-risk, are rejected.

Table 1. Λ/P-values, $\delta_j / \bar{\delta}$ for each allele for simulated data.

Common allele	**1**	**2**	**3**	**4**	**5**
Λ_j	0.80392	0.27248	1256.665	0.00985	0.78023
P-value	0.370	0.602	0.000	0.921	0.377
	6	**7**	**8**	**9**	**10**
	0.81300	0.33284	0.52457	0.12862	
	0.367	0.564	0.469	0.720	
Rare allele	**1**	**2**	**3**	**4**	**5**
$\delta_j / \bar{\delta}$	0.297	1.302	1.428	0.299	2.182
	6	**7**	**8**	**9**	**10**
	0.029	1.025	1.318	0.108	2.010
Parameter	**intercept**	α_3	λ^+	λ^-	**bmi**
real		0.800			0.017
estimates	1.505	0.806	0.545	–0.500	0.0173
(sd)	(0.00033)	(0.00015)	(0.00034)	(0.00057)	(0.00001)
	age	**gender**			
	0.008	0.120			
	0.0077	0.128			
	(2.896E-6)	(0.0001)			

For the rare alleles, the ratios $\delta_j / \bar{\delta}$'s of alleles 2, 3, 5, 8 and 10 are bigger than the critical value $\gamma = 1.3$, suggesting rare alleles 2, 3, 5, 8 and 10 are likely risk alleles. With the deletion of alleles 2,3, and 10, the estimates of λ were smaller, suggesting that these alleles are positively associated with the trait (ie, they belong to R^+). Similarly, deleting alleles 5 and 8 resulted in larger estimates of λ, suggesting that these two alleles are negatively associated with the trait and thus belong to R^-. These results are consistent with the 'truth' as implemented in our simulated dataset. Finally, we refitted the model with only the associated (risk) common and rare alleles, as in the last three rows of Table 1, which also show the effects of the risk alleles with covariates included. Analyses of simulated data under different conditions also reached the correct conclusions and are not displayed.

Real Data Analysis

We used SCARVA to analyze two independent real datasets to identify rare and common variants underlying variation in obesity among participants in the Africa America Diabetes Mellitus (AADM) study and plasma triglyceride levels in the Dallas Heart Study (DHS).[23] The software PHASE,[24] was used to construct haplotypes. For both traits, our results were consistent with published results.

First real dataset

The AADM dataset included 141 unrelated individuals from West Africa who were part of a linkage and association study of type 2 diabetes (T2D) and associated risk factors, including BMI, a commonly used measure of the degree of adiposity. The AADM protocol was approved by the institutional review board of Howard University and the respective institutions in West Africa. Written informed consent was obtained from each participant.

For this study we focused on the linkage and association signal observed in a 19cM region on chromosome 5. After evidence for strong linkage in this region (125906 bp to 125960 bp) on chromosome 5, we conducted fine-mapping using experimentally and imputed SNPs genotypes for an average map density of less than 1 kb. The results of the fine-mapping (manuscript in preparation) identified a very strong candidate gene for obesity and this gene was subsequently sequenced using Sanger technology. It is this sequence data that was analyzed using SCARVA. Using an established method,[25] we identified 9 haplotype blocks in this gene. The results of the analyses of the haplotypes within these blocks using SCARVA were similar to those obtained using traditional methods, like logistic regression. Some numerical details of these results, including values of the corresponding j's, the log-likelihood, P-values and regression parameters (with standard errors) are displayed for the first haplotype in Table 2.

As shown in Table 2, we observed a significant ($P = 0.023$, from Step II) association between common allele 1 with BMI; other common alleles were not significant at $P = 0.05$. The $\delta_j / \bar{\delta}$ value (from Step III) for rare allele 3 was 4.205, which is much larger than the suggested highly significant critical value of $\gamma = 2.5$, indicating that rare allele 3 is strongly associated with obesity as measured by BMI. These results show that both common allele 1 and rare allele 3 are strongly associated with obesity.

The overall results for all the nine haplotypes are summarized in Table 3 below. Displayed in the table are the number of common and rare haplotypes, the significant common allele with the corresponding P-value, from Step II, in bracket, and the significant

Table 2. Λ/P-values $\delta_j / \bar{\delta}$ for each allele in block 1.

Common allele	**1**	**2**	**3**	**4**	**5**	**6**
Λ_j	5.12324	0.70894	0.00744	0.82887	0.02079	0.02010
P-value	0.02361	0.39980	0.93128	0.36260	0.88534	0.88725
Rare allele	**1**	**2**	**3**	**4**	**5**	
$\delta_j / \bar{\delta}$	0.194	0.001	4.205	0.388	0.212	
Parameter	**intercept**	α_1	λ	**age**	**gender**	**T2D**
estimates	36.948	5.466	7.795	0.033	–1.860	–3.224
(sd)	(0.167)	(0.138)	(0.217)	(0.004)	(0.038)	(0.042)

Table 3. Summary results for the AADM data.

Block	No. of comm. hap.	Sig. *P*-value	No. of rare hap.	Sig. ratio
1	7	1 (0.023)	5	3 (4.205)
2	2		2	
3	7	4 (0.027)	2	2 (1.996)
4	2		2	
5	8		3	
6	2		2	
7	3		1	
8	2		1	
9	2		1	

rare allele with the corresponding $\delta_j / \bar{\delta}$ value, from Step III, in bracket, also by haplotype,

In addition to the results described above for haplotype 1, we observed that common allele 4 (*P*-value = 0.027) and rare allele 2 ($2(\delta_2 / \bar{\delta} = 1.996) = 1.996$) are associated with obesity.

To evaluate our real dataset and compare the results to those obtained from SCARVA, we used QuTie[26] approach (the Rare Variant Analysis Tool for Quantitative Trait). Notably, the QuTie method is designed to detect association of rare allele(s) only. It pools the low frequency/rare variants within defined regions and treats them as a single super locus, with analysis by linear regression and student's *t*-test. Using haplotypes to pool rare variants, we observed a BMI value of 29.51 for QT + RV (the quantitative traits for individuals with at least one rare variant minor allele) compared to a mean of 23.71 for QT-RV (quantitative traits for individuals with no rare variant minor allele; *P*-value = 0.004, beta = −5.79, and std error = 1.98). These results indicate that rare variants are associated with obesity, although the specific rare variant(s) responsible for the association is not necessarily identified. In contrast, SCARVA identified rare allele 3 and common allele 1 as the likely risk alleles.

Second real dataset

The aim of the Dallas Heart Study[23] was to use a reverse genetic strategy to test the hypothesis that 4 angiopoietin-like proteins (*ANGPTL* 3, 4, 5 and 6) play key roles in triglyceride (TG) metabolism in humans by re-sequencing the coding regions of the genes encoding these proteins. Analyses of the sequence data identified multiple rare nonsynonymous (NS) sequence variants that were associated with low plasma TG level but not with other metabolic phenotypes. Functional studies revealed that all mutant alleles of *ANGPTL* 3 and *ANGPTL* 4 that were associated with low plasma TG level interfered either with the synthesis or secretion of the protein to inhibit lipoprotein lipase (LPL). Interestingly, a total of 1% of the DHS population and 4% of these participants with a plasma TG in the lowest quartile had a rare loss-of-function mutation in *ANGPTL* 3, *ANGPTL* 4, or *ANGPTL* 5. Hence the investigators concluded that *ANGPTL*) 3, *ANGPTL* 4 and *ANGPTL* 5, but not ANGPTL 6, play non-redundant roles in TG metabolism, and that multiple alleles at these loci cumulatively contribute to variability in plasma TG levels in human populations.

We reanalyzed the DHS sequencing data for the three genes (*ANGPTL* 3, 4, 5) using SCARVA. The results of the significant common and/or rare variants are displayed in Table 4. The displayed results include *P*-values, the $\delta_j / \bar{\delta}$ values for the significant rare variant and the number of total common variants and rare variants along with the coefficients of the significant common/rare variants (with standard error).

Briefly, the results of our reanalysis using SCARVA are as follows: we observed 2 common and 7 rare variants in *ANGPTL* 3 gene. The 2nd common variant is significant (*P* = 0.0461), the corresponding regression coeffcient is 0.027 with standard error 0.00017. The 2nd and 4th rare variants are significant with ratio $\delta_j / \bar{\delta}$ values 1.916 and 1.376 respectively, and positive association of 0.218 (se 0.0009). The 6th rare variant with ratio value 2.7, is negatively associated with coeffcient −0.111 (SE 0.00017). We observed 2 common and 27 rare variants in the *ANGPTL* 4 gene. None of the common variants was significantly associated with the trait. In contrast, rare variants 1, 13, 14, 17 and 24 were negatively associated with the trait, with significant ratio values of 1.951, 2.260, 9.236, 1.500, 2.670 respectively, and estimated effect of −0.0023 (0.000047). In the *ANGPTL* 5 gene, we observed 2 common and 19 rare variants. None of the common variants was significantly associated with the trait. Rare variants 1 and 4 are positively associated with the trait (ratio values 1.377 and 1.298 respectively, and coeffcient 0.0960, se = 0.00039). Rare variants 7, 18 and 19 were negatively associated with the trait (ratio values of 1.961, 1.748 and 1.777, and with coeffcient −0.1376, se = 0.00057).

Table 4. Summary results for the ANGPTL data.

ANGPTL	No. of comm. hap.	sig. comm (*P*-value)	Coef. (se)
3	2	v_2 (0.0461)	0.027 (0.00017)
4	2		
5	2		
ANGPTL	**No. of rare hap.**	**sig. rare (ratio)**	**Coef. (se)**
3	7	u_2+ (1.916)	0.218 (0.0009)
		u_4+ (1.376)	
		u_6– (2.700)	–0.111 (0.00017)
4	27	u_1– (1.951)	–0.0023 (0.000047)
		u_{13}– (2.260)	
		u_{14}– (9.236)	
		u_{17}– (1.500)	
		u_{24}– (2.670)	
5	19	u_1+ (1.377)	0.0960 (0.00039)
		u_4+ (1.298)	
		u_7– (1.961)	–0.1376 (0.00057)
		u_{18}–(1.748)	
		u_{19}–(1.777)	

Discussion

We proposed a novel approach (SCARVA) for the combined association analysis of common and rare variants in disease and non-disease trait research. SCARVA is a regression-based strategy that uses quantitative trait and haplotype data together with covariates. The common alleles analysis implemented in SCARVA is a straightforward linear regression. However, to avoid the problem of dimensionality (ie, large number of parameters with very small dataset), SCARVA models the effect of rare alleles using a single parameter with the well-developed approach of identifying variants that show positive as well as negative associations. Furthermore, we implemented the BIC and the AIC as test statistics, because the modeling of rare alleles is partly non-nested, the classical chi-squared approach is not appropriate. In this regard, the rare variants analysis in SCARVA is less 'quantitative' than that of the common alleles. We note that, as implemented, SCARVA addresses a major limitation (ie, dilution of power due to the combined analysis of functional and non-functional variants) of current rare variants analysis software packages. Finally, we showed that the method is simple to use and computationally effcient. Simulation studies showed that the method works well and can accurately identify both the common and rare risk alleles defined as those variants with at least moderate effects on the trait.

In principle Step 2 and 3 can be done iteratively, but we prefer the current order of Step 2 then followed by Step 3, as data on common alleles have more observations, and results inferred from them are more reliable than those from rare alleles. Thus we use the common alleles to guide the regressor selection in the model.

SCARVA uses haplotype information instead of individual SNPs, which lowers the computational burden of the analysis. However, this computational advantage is at the cost of lower resolution. As part of future efforts in our lab, we are actively exploring how to extend SCARVA to accommodate the analysis of both haplotypes and individual SNPs. In this case synthetic association[27,28] can be considered.

Acknowledgement

This work is supported in part by the National Center for Research Resources at NIH grant 2G12RR003048, and by the Center for Research on Genomics and Global Health (CRGGH) at NHGRI/NIH.

Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects.

If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

References

1. Coetezee V, Barrett L, Greeff JM, Henzi SP, Perrett DI, et al. Common HLA alleles associated with health, but not with facial attractiveness. *PLoS ONE*. 2007;2(7):e640. (doi:10.1371/journal.pone.0000640).
2. Hindorff LA, Junkins HA, Hall PN, Mehta JP, Manolio TA. A Catalog of Published Genome-Wide Association Studies. 2011 http://www.genome.gov/gwastudies. [cited May 1, 2011]; Available from: www.genome.gov/gwastudies.
3. Maher B. Personal genomes: The case of the missing heritability. *Nature*. 2008;456(7218):18–21.
4. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747–53. PMCID: 2831613.
5. Frayling I, et al. The APC variants I1307K and E1317Q are associated with colorectal tumors, but not always with a family history. *Proc Natl Acad Sci U S A*. 1998;95:10722–7.
6. Bodmer W. Familial adenomatous polyposis (FAP) and its gene, APC. Cyto-genet. *Cell Genet*. 1999;86:99–104.
7. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*. 2008;40(6):695–701.
8. Fearnhead N, Winney B, Bodmer W. Rare variant hypothesis for multifactorial inheritance: susceptibility to colorectal adenomas as a model. *Cell Cycle*. 2005;4:521–5.
9. Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. *Annu Rev Genet*. 2010;44:293–308.
10. Smith DJ, Lusis AJ. The allelic structure of common disease. *Human Molecular Genetics*. 2002;11(20):2455–61.
11. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics*. 2008;83(3):311–21.
12. Price AL, Kryukov GV, de Bakker P, et al. Pooled association tests for rare variants in exon-resequening studies. *American Journal of Human Genetics*. 2010;86:832–8.
13. Grady DL, Chi HC, Ding YC, et al. High prevalence of rare dopamine receptor D4 alleles in children diagnosed with attention-deficit hyperactivity disorder. *Molecular Psychiatry*. 2003;8:536–45.
14. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*. 2010;34:188–93.
15. Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. *Genetic Epidemiology*. 2010;34:171–87.
16. McClellan JM, Susser E, King MC. Schizophrenia: a common disease caused by multiple rare alleles. *British Journal of Psychiatry*. 2007;190:194–9.
17. Shental N, Amir A, Zuk O. Identification of rare alleles and their carriers using compressed se(que)nsing. *Nucleic Acids Research*. 2010;38(22). doi:10.1093/nar/gkq675.
18. Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. *American Journal of Human Genetics*. 2011;89:354–67.
19. Daly M, Rioux J, Schaffner S, Hudson T, Lander E. High-resolution haplotype structure in the human genome. *Nature Genetics*. 2010;29:229–32.
20. Eskin E, Halperin E, Karp R. Large scale reconstruction of haplotypes from genotype data. Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB-2003) San Diego, CA, March 27–31, 2004.
21. Kimmel G, Sharan R, Shamir R. Computational problems in noisy SNP and haplotype analysis: block scores, block identification and population stratification. Manuscript. 2004.
22. Schwarz G. Estimating the dimension of a model. *Annals of Statistics*. 1978;6:461–4.
23. Romeo S, Yin W, Kozlitina J, et al. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *Journal of Clinical Investigation*. 2009;119:70–9.
24. Stephens M, Smith N, Donnelly P. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*. 2001;68:978–89.
25. Purcell S, Neale B, Todd-Brown K, et al. A toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*. 2007;81:559–75.
26. Lawrence R, Day-Williams AG, Elliot KS, Morris AP, Zeggini E. CCRaVAT and QuTie—enabling analysis of rare variants in large-scale case control and quantitative trait association studies. *BMC Bioinformatics*. 2010;11:527.
27. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLOS Biology*. 2010;8(1):e1000294. doi:10.1371/journal.pbio.1000294.
28. Wray NR, Purcell SM, Visscher PM. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS biology*. 2011;9(1):e1000579.