Bioinformatics, 32, 2016, i611–i619
doi: 10.1093/bioinformatics/btw429
ECCB 2016

OXFORD

Gene-set association tests for next-generation sequencing data

Jaehoon Lee[1], Young Jin Kim[2], Juyoung Lee[2], T2D-Genes Consortium, Bong-Jo Kim[2], Seungyeoun Lee[3] and Taesung Park[1,*]

[1]Department of Statistics, Seoul National University, Seoul 151-742, Korea, [2]Division of Structural and Functional Genomics, Korean National Institute of Health, Osong, Chungchungbuk-Do 363-951, Korea and [3]Department of Mathematics and Statistics, Sejong University, Seoul 143-747, Korea

*To whom correspondence should be addressed.

Abstract

Motivation: Recently, many methods have been developed for conducting rare-variant association studies for sequencing data. These methods have primarily been based on gene-level associations but have not been proven to be as effective as expected. Gene-set-level tests have shown great advantages over gene-level tests in terms of power and robustness, because complex diseases are often caused by multiple genes that comprise of biological gene sets.
Results: Here, we propose several novel gene-set tests that employ rapid and efficient dimensionality reduction. The performance of these tests was investigated using extensive simulations and application to 1058 whole-exome sequences from a Korean population. We identified some known pathways and novel pathways whose rare or common variants are associated with elevated liver enzymes and replicated the results in an independent cohort.
Availability and Implementation: Source R code for our algorithm is freely available at http://statgen.snu.ac.kr/software/QTest.
Contact: tspark@stats.snu.ac.kr
Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Genome-wide association studies (GWAS) have focused on the associations between complex diseases and common genetic variants, and have successfully reported an extensive list of single nucleotide polymorphisms (SNPs) associated with complex diseases. However, it has been shown that the associated common variants could explain only a small fraction of the heritability of many common diseases (Bansal *et al.*, 2010). This suggests that other genetic mechanisms such as gene–gene interactions, gene-set-level actions of common variants, or the action of multiple rare variants, could contribute to disease susceptibility. Among these mechanisms, rare variants have become the focus of intense investigation following the development of next-generation sequencing (NGS) technology (Adzhubei *et al.*, 2010; Manolio *et al.*, 2009).

In recent years, many statistical tests have been proposed for detecting the signals of rare variants (Lee *et al.*, 2012; Li and Leal, 2008; Lin and Tang, 2011; Madsen and Browning, 2009; Morris and Zeggini, 2010; Price *et al.*, 2010; Wu *et al.*, 2011). However, current approaches are in the early stages of development, and significant improvements are required, in terms of increasing statistical power and considering the biological context of diseases. Ladouceur et al. (2012) demonstrated that assessing the association between rare variants and complex diseases is still a challenging task and that no single method yields consistently high power, even using large sample sizes.

Complex diseases often result from the combined action of multiple risk factors in a gene or across genes that comprise a gene set or a pathway. Gene-set analysis (GSA) has been widely used for analyses of microarray data (Peng *et al.*, 2010; Subramanian *et al.*, 2005; Wang *et al.*, 2007) and GWAS data (Chai *et al.*, 2009; Goeman et al., 2004), and has played an important role in uncovering the mechanisms of complex diseases. GSA has been shown to possess great advantages over single-gene tests. GSA incorporates related SNPs or genes into a single statistic, and thus requires a small number of tests and yields high power to detect association signals. Even if individual genes of a set have weak or moderate association signals, GSA can combine those into a single strong signal. Incorporation of verified biological knowledge in GSA facilitates interpretation of the underlying genetic background of identified gene sets and reduces false positive results. However, current rare-variant studies have focused mainly on gene-level associations, and GSA of rare variants are scant, as very few pathway-based methods [e.g. smoothed functional principal component analysis (SFPCA; Zhao *et al.*, 2014), for binary traits] have been proposed.

The previously introduced GSA methods can be classified into two types of analysis strategies. The first type of GSA is a one-step method that regards a gene set as a large 'super-gene'. This approach has limited application to high-throughput NGS data in that, as the number of variants within a gene set increases, the power of the analysis decreases. The second type is a two-step method that consists of a gene-level association test followed by a test for association of a gene set with a trait. The two-step method has been widely used in traditional GSA. However, direct application of the traditional two-step GSA used in GWAS to NGS data is not appropriate because the gene-level summarization in the traditional two-step GSA does not collapse or consider rare variants. Although gene-level summarization can be replaced by *P*-values for currently available rare variant association tests, a gene-set-level test that combines *P*-values for multiple genes might be underpowered, due to high degrees of freedom. Combining *P*-values with the assumption of independence could also yield false positive results, if correlations among variants or genes are not taken into account (Price *et al.*, 2010). Therefore, the development of a new GSA is warranted to test gene-set associations by considering the characteristics of rare variants, to increase statistical power by reducing dimensionality, and to account for all possible correlations among variants or genes.

In this article, we propose a powerful gene-set test as well as a single-gene test for use with NGS data. For quantitative traits in particular, we first derive powerful gene-level tests as quadratic forms (QTest) using an eigenvalue decomposition of regression coefficients and applying a dimensionality reduction method. This eigenvalue decomposition step allows our tests to account for correlated variants. Based on the QTest, we then develop the proposed gene-set-level quadratic test (GS.QTest) using an efficient method for reducing degrees of freedom. The proposed tests possess four advantages: (1) QTests provide higher statistical power than existing gene-level tests; (2) GS.QTests are flexible in that they can easily incorporate other existing gene-level tests for rare variants, such as the sequence kernel association test (SKAT); (3) the proposed QTests and GS.QTests cover a broad range of scenarios for joint action of rare variants and common variants; and (4) the tests do not require heavy computational effort because they employ parametric or pre-calculated empirical distributions.

Through extensive simulations, we investigated the performance of the proposed methods by comparing them with other gene-level and gene-set-level association methods. For example, available gene-level association methods include collapsing methods such as GRANVIL (Morris and Zeggini, 2010), weighted sum statistic (MB; Madsen and Browning, 2009) and variable threshold (VT; Price *et al.*, 2010). For the collapsing of rare variants, GRANVIL counts the rare variants; MB aggregates weighted sum based on minor allele frequency (MAF); VT uses the partial sum based on optimal MAF threshold. For bidirectional approaches, we included SKAT (Wu *et al.*, 2011), SKAT-O (Lee *et al.*, 2012), the likelihood ratio test (LRT), and estimated regression coefficients (EREC; Lin and Tang, 2011) in simulation studies. SKAT uses score-based variance component model and SKAT-O is an optimal test combining a burden-type test and SKAT. EREC estimates the regression coefficients and uses these as weights. These bidirectional approaches are useful when deleterious and protective variants are simultaneously present. For the gene-set-level association methods, we included the traditional gene-set-level association methods, GLOSSI (Chai *et al.*, 2009) and GlobalTest (Goeman et al., 2004). GLOSSI combines variant-level *P*-values based on the chi-square statistic, while GlobalTest uses score tests in the framework of generalized linear models.

Finally, we applied the proposed methods to exome sequencing data of 1058 Korean samples (Cho et al., 2009) and some liver enzyme traits and identified some known and novel pathways including the beta-alanine metabolism pathway, lysine degradation pathway which are known to be related to liver cancer. To further validate the potential association of the pathways with elevated liver enzymes, we conducted a replication study in an independent cohort comprising 897 samples; some pathways were successfully replicated (P-value < 0.05).

2 Methods

2.1 Generation of simulated sequencing data

To generate simulation data, we used the software SimRare (Li *et al.*, 2012), which generates sequencing data based on demographic and evolutionary scenarios for the real population. One thousand replicates for the sequence and trait data for 3000 samples were generated for this simulation study. For each individual, we generated a quantitative trait value by adding the effect of multiple causal rare variants and an error term that followed a standard normal distribution. To simulate a gene-level test, we varied the gene size, which is the number of rare variants (10, 20) within a gene, the proportion of causal variants (10, 20, 30, 50 and 70%) and the effect size (β) of causal variants (0.75, 1.0, 1.25 and 1.5).

To check the type 1 error, two million replicates of traits from the null distribution were generated. The type 1 error was defined as the proportion of *P*-values less than various specified significance levels (10^{-3}, 10^{-4}, 10^{-5} and 2.5×10^{-6}) among two million *P*-values. The type 1 errors were calculated for various gene sizes (10, 20 and 50) given the sample size 3000, minimum minor allele count (MAC) 1 and maximum MAF 0.01.

To investigate the performance of the proposed tests for each gene set, another simulation for gene-set analysis was conducted. Based on simulated sequencing data, we assembled a gene set that included 12 genes. In the scenario set for gene-set analysis, trait values were generated under 135 different scenarios by varying the number of causal genes (2, 3 and 4), proportion of causal variants (10, 20, 30, 50 and 70%), effect size (0.75, 1.0 and 1.25) and size of causal gene (10, 30 and 50 variants).

The gene-level power was defined as the proportion of *P*-values less than 2.5×10^{-6} among *P*-values from the simulation under the corresponding scenarios. This cut-off was based on a 5% significance level with Bonferroni correction, under the assumption that 20 000 genes are being tested simultaneously (Lee et al., 2012). Similarly, the gene-set-level power was defined as the proportion of *P*-values less than 2.5×10^{-5} among *P*-values under the assumption that 2000 gene-sets are being tested simultaneously.

2.2 Gene-level association methods based on quadratic tests

As a gene-level association test for rare variants, we introduced a method for combining regression coefficients from a multiple regression framework. Regression coefficients for rare variants can be combined using three different methods according to the relationships among the rare variants. The followings are plausible scenarios for multiple rare variants:

i. All variants within a region have a common effect, either deleterious or protective, or
ii. Some variants are deleterious and others are protective.

QTest$_1$: If there are m rare variants within a region, we can regress a trait, y, on rare variants (S_k's) and covariates. We can first assume that rare variants within a region have common effects, producing either deleterious or protective signals. Given this assumption, the collapsing method, which aggregates effects in only one direction,

provides powerful performance. We derived the $QTest_1$, the inverse variance weighting method for a pooled effect size (β_{pooled}). A chi-square statistic based on pooled effect size is computed as follows:

$$y = \beta_0 + \sum_{k=1}^{m} \beta_k S_k + \gamma Z + \varepsilon, \text{ where } \varepsilon \sim N(0,1)$$

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_k\right)_{m\times 1}, \boldsymbol{\alpha} = (\alpha_k)_{m\times 1}, \text{where } \alpha_k = \frac{1/\text{var}(\widehat{\beta}_k)}{\sum_{k=1}^{m}\left(1/\text{var}(\widehat{\beta}_k)\right)},$$

$$\mathbf{V} = \text{var}(\widehat{\boldsymbol{\beta}}), \mathbf{W} = \begin{pmatrix} w_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_m \end{pmatrix} \text{where } w_k \text{ is a weight for } k\text{th variant}$$

$$\widehat{\beta}_{Pooled} = \boldsymbol{\alpha}^T \mathbf{W}\widehat{\boldsymbol{\beta}} = \sum_{k=1}^{m} \alpha_k w_k \widehat{\beta}_k \sim N(0, \boldsymbol{\alpha}^T \mathbf{WVW}\boldsymbol{\alpha})$$

$$Q_1 = (\boldsymbol{\alpha}^T \mathbf{WVW}\boldsymbol{\alpha})^{-1} \widehat{\beta}^2_{Pooled} \sim \chi^2_1$$

$QTest_2$: We can also assume that some rare variants are deleterious and other rare variants are protective. In this case, the $QTest_1$ shows very poor performance because deleterious and protective signals are offset when combined. To combine association signals with different directions without loss of power, we propose the $QTest_2$.

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_k\right)_{m\times 1}, \text{ for } k = 1, \ldots, m,$$

$$\mathbf{V} = \text{var}(\widehat{\boldsymbol{\beta}}) = \mathbf{U}\Lambda\mathbf{U}^{\mathrm{T}} \text{ where } \Lambda = diag(\lambda_k)$$

where $\mathbf{U}$ consists of eigenvalue vectors of $\mathbf{V}$,

and Λ is the diagonal matrix whose diagonal

elements are eigenvalues of $\mathbf{V}$

$$Q_2^{Wald} = \widehat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{V}^{-1} \widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{U}\Lambda^{-1}\mathbf{U}^{\mathrm{T}} \widehat{\boldsymbol{\beta}} \sim \chi^2_m$$

$$p_j = 2\left(1 - \Phi\left(\mathbf{u}_j^{\mathrm{T}} \widehat{\boldsymbol{\beta}} / \sqrt{\lambda_j}\right)\right) \text{ where } \mathbf{u}_j = j^{th} \text{ column of } \mathbf{U}$$

$$Q_2 = \sum_{j=1}^{m} G^{-1}_{a,1}(1 - p_j) \sim G_{ma,1} = \frac{1}{2}\chi^2_{2ma}$$

G denotes gamma distribution.

$QTest_3$: Although the $QTest_2$ can handle effect sizes that differ in direction, it does not provide better performance than the $QTest_1$ when multiple rare variants have an aggregated effect in the same direction. In order to account for the etiology of rare variants, we should consider all possible cases. Our newly proposed optimal quadratic test ($QTest_3$) statistic is a weighted average of the $QTest_1$ and the $QTest_2$. The $QTest_3$ can operate under assumptions (i) and (ii). The steps for the $QTest_3$ are as follows:

1. Compute $\widehat{\boldsymbol{\beta}}^*$ so that $\widehat{\boldsymbol{\beta}}^* \perp \widehat{\beta}_{Pooled}$

$$\widehat{\boldsymbol{\beta}}^* = \widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}}^* | \widehat{\beta}_{Pooled}) = \widehat{\boldsymbol{\beta}} - \widehat{\beta}_{Pooled}(\boldsymbol{\alpha}^T \mathbf{WVW}\boldsymbol{\alpha})^{-1} \mathbf{WVW}\boldsymbol{\alpha}$$

2. Compute $Q^*_{2|1}$ so that $Q_1 \perp Q^*_{2|1}$

$$Q^*_{2|1} = \widehat{\boldsymbol{\beta}}^* \mathbf{V}^{*-1} \widehat{\boldsymbol{\beta}}^* = \widehat{\boldsymbol{\beta}}^{*\mathrm{T}} (\mathbf{U}^* \Lambda^{*-1} \mathbf{U}^{*\mathrm{T}}) \widehat{\boldsymbol{\beta}}^* \sim \chi^2_{m-1} \text{ where } \mathbf{V}^* = \text{var}(\widehat{\boldsymbol{\beta}}^*)$$

$\mathbf{U}^*$ consists of eigenvalue vectors of $\mathbf{V}^*$

and Λ^* is the diagonal matrix whose diagonal

elements are eigenvalues of $\mathbf{V}^*$

3. Compute $Q_{2|1}$ so that $Q_{2|1}$ follows χ^2_1

$$Q_{2|1} = 2G^{-1}_{0.5,1}(1 - p_{2|1}) \text{ where } p_{2|1} \text{ is obtained from } Q^*_{2|1}$$

4. Compute Q^π_3 for $\pi = 0, 0.1, \ldots, 0.9, 1$

$$Q^\pi_3 = (1-\pi)Q_1 + \pi Q_{2|1} \text{ where } Q_1, Q_{2|1} \sim \text{indep } \chi^2_1$$

From a mixture of two χ^2_1, calculate p value p^π_3

5. Final p value for $Q_3^{\text{optimal } \widehat{\pi}}$ is calculated from empirical distribution

optimal $\widehat{\pi} = \text{Argmin}_\pi\{p^\pi_3;\ \pi = 0,\ 0.1,\ 0.2,\ \ldots,\ 0.9,\ 1\}$

The empirical distribution of $Q_3^{\text{optimal } \widehat{\pi}}$ was calculated by generating a pair of random variables from a chi-square distribution with one degree of freedom, and computing their maximum value of weighted averages over the eleven values of π. In our simulation studies and real data analyses, 10^9 pairs of random variables were generated, and the empirical distribution of was summarized in a cumulative distribution table. Then, the $QTest_3$ used this pre-calculated cumulative distribution table for every gene-level test.

2.3 Gene-set-level analysis for rare and common variants

In order to maximize the genetic variation caused by moderate association, gene-set analysis (GSA) can use prior biological knowledge based on pathway information. We propose two types of gene set tests, the GS_Q.QTest and the GS_B.QTest. The proposed GS_Q.QTest for rare variants requires the following two steps:

Step 1. Compute gene-level *P*-values for rare and common variants using QTests.
Step 2. Combine gene-level chi-square statistics within a predefined gene set.

In **Step 1**, any possible rare variants association test can be used to obtain the gene-level *P*-values, allowing great flexibility. In **Step 2**, given a gene set (*GS*), a gene-set level statistic can be computed based on an inverse gamma transformation (Zhao *et al.*, 2014):

$$Q_{GS} = \sum_{t\in GS} 2G^{-1}_{a,1}(1 - p_t) \sim \chi^2(2ma)$$

p_t = p value for tth gene, m = size of GS

We also proposed another type of gene-set test, the GS_B.QTest. The proposed GS_B.QTest requires the following two steps:

Step 1. Collapse multiple rare variants in each gene within a *GS*.
Step 2. Using the collapsed variants and common variants, conduct QTests.

In **Step 1**, the collapsing method is applied only to rare variants. In **Step 2**, we can use any possible association test, such as the SKAT or the SKAT-O.

3 Results

3.1 Simulation study for gene-level tests

The proposed gene-level test includes three versions that can be applied to various scenarios for rare variants within a gene. The first version ($QTest_1$) is a burden-type test and the second version ($QTest_2$) is a non-burden-type test. The third version ($QTest_3$) is an optimal QTest that combines $QTest_1$ and $QTest_2$ such that it maintains high power regardless of the direction of effects and proportion of causal variants.

The simulation results of type 1 errors of QTests are summarized in Table 1 for several gene sizes and several significance levels. Table 1 shows that the proposed QTests well-preserve the type 1 errors under the null distribution.

We compared the performance of the proposed QTests to existing methods, such as SKAT, SKAT-O, VT, LRT, MB, GRANVIL and EREC under various scenarios. MB, VT and EREC were implemented by SCORE-Seq (Lin and Tang, 2011). All methods were conducted using the default parameter values. MAF threshold value was fixed as 0.01 for rare variant analysis. We generated trait values by varying the causal gene size, the proportion of causal variants,

Table 1. Type 1 error of proposed gene-level QTests

α	Gene size 10			Gene size 50		
	$QTest_1$	$QTest_2$	$QTest_3$	$QTest_1$	$QTest_2$	$QTest_3$
1.0E-03	1.03E-03	1.02E-03	1.06E-03	9.75E-04	1.01E-03	1.05E-03
1.0E-04	9.40E-05	1.06E-04	1.07E-04	1.00E-04	9.39E-04	1.02E-04
1.0E-05	9.00E-05	8.50E-06	8.50E-06	9.95E-06	1.10E-05	9.95E-06
2.5E-06	3.00E-06	1.50E-06	2.50E-06	2.50E-06	2.00E-06	2.50E-06

and the effect size of the causal variants. The relationships between power and effect size are displayed in Figure 1. In general, as the effect size increased, the power increased. Figure 1a shows the result when the proportion of causal variants is low (10 or 20%), and Figure 1b shows the result when the proportion of causal variants is high (50 or 70%). In Figure 1a, $QTest_2$, $QTest_3$, LRT, SKAT and EREC performed much better than burden tests, and $QTest_2$ had the highest power when the effect size was larger than 0.75. As shown in Figure 1b, the optimal tests, such as $QTest_3$ and SKAT-O, perform better than other methods, especially when the effect size is 0.75.

For application to other scenarios, we varied the proportion of protective variants among causal variants within a gene, given a proportion of causal variants 0.5 and effect size 1.0. We also varied the proportions of common variants among causal variants, given a proportion of causal variants 0.5, proportion of common variants 0.5, and effect size 1.0, to observe the joint effect of common and rare variants. In these scenarios, $QTest_2$, $QTest_3$, SKAT, SKAT-O and LRT are robust enough to include protective causal variants or common variants. $QTest_2$, $QTest_3$ and LRT had the largest power among all methods (see Supplementary Figs. S1 and S2). With the existence of non-causal common variants, the burden tests $QTest_1$ and GRANVIL showed poor performance, but other burden tests, MB and VT, showed very good performance, because they focus on rare variants by weighting or thresholding. EREC showed much lower power than $QTest_2$, $QTest_3$, SKAT and SKAT-O, when the proportion of protective variants among causal variants was large; for the cases when the regression coefficients are negative, the weight used in EREC tends to decrease the signal. EREC also showed poor performance when there were non-causal common variants together with causal rare variants (see Supplementary Fig. S2).

For the simulation study of one-directional effects, we varied the effect size of causal variants with the assumption that all variants were deleterious causal variants. The burden tests $QTest_1$ and GRANVIL showed greater power than other tests. The non-burden tests $QTest_2$, SKAT, and LRT showed poor performance, especially when the effect size was not large (see Supplementary Fig. S3).To illustrate a computational burden for QTests, a specific computational time for one simulation setting is summarized in Supplementary Table S1. QTests showed a similar computing time to that of SKAT, and $QTest_3$ was found to be computationally much more efficient than SKAT-O, when implemented in R.

3.2 Simulation study for gene-set-level tests

We developed two types of proposed gene-set-level tests: the GS_Q.QTest and the GS_B.QTest. The GS_Q.QTest calculates gene-level *P*-values by applying the QTest to multiple variants within a gene and then combines the *P*-values using an efficient dimensionality reduction method. The GS_B.QTest employs a burden-type summarization at the gene level and applies the QTest to the collapsed gene-level variants within a set.

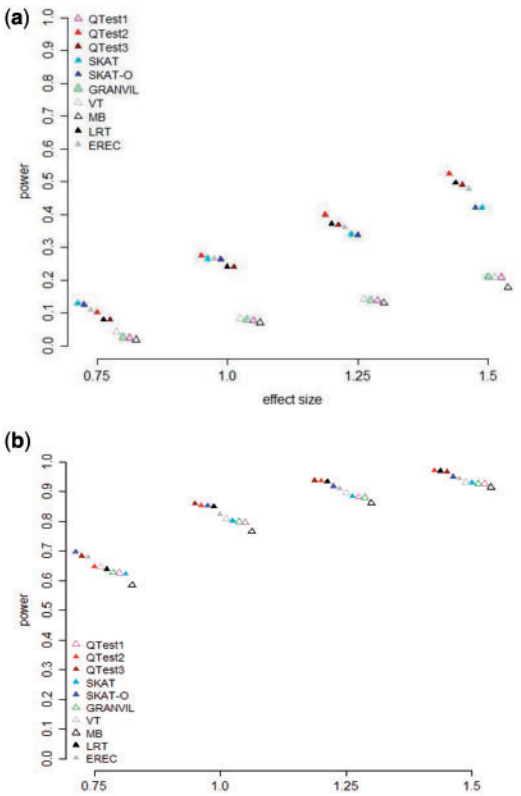

Fig. 1. Comparison of the power of gene-level tests for varying effect sizes; (a) for a low proportion of causal variants; (b) for a high proportion of causal variants

Other rare-variants association methods, such as SKAT and GRANVIL, can easily be incorporated into the proposed gene-set methods. When SKAT was used instead of the QTest in the proposed two gene-set-level tests, the proposed gene set tests are denoted as GS_Q.SKAT and GS_B.SKAT, respectively.

In simulation studies, the following gene-set methods were compared: two traditional gene set analysis methods (GLOSSI, GlobalTest), two QTest-based gene set analysis methods (GS_Q.QTest, GS_B.QTest) and two SKAT-based gene set analysis methods (GS_Q.SKAT, GS_B.SKAT). GLOSSI and GlobalTest are one-step gene set tests that regard all variants from the same gene set as one supergene. The one-step SKAT method, which employs a similar method, was not included in the Figures because the one-step SKAT method produced a very similar pattern to the GlobalTest (data not shown). Under our simulation setting, GS_Q.SKAT was consistently more powerful than the one-step SKAT. In general, considering gene-level summarization makes the gene-set analysis more powerful; however, a few exceptions exist. If causal variants are concentrated in a very few genes, the gene-level summarization may produce a loss of power. Figure 2 shows the relationship between the effect size of causal variants and the power of the gene-set test. Figure 2a shows the result when the proportion of causal variants is low (10 or 20%), whereas Figure 2b shows the result when the proportion of causal variants is high (50 or 70%). As shown in Figure 2a, for a low proportion of causal variants, the GS_Q.SKAT method performs best when the effect size is 0.75, whereas the GS_Q.$QTest_2$ performs best when the effect size is larger than 0.75. In this setting, GS_B-type tests show poorer performance than GS_Q-type tests because collapsing a high proportion of non-causal variants within a gene decreases the power of the analysis.

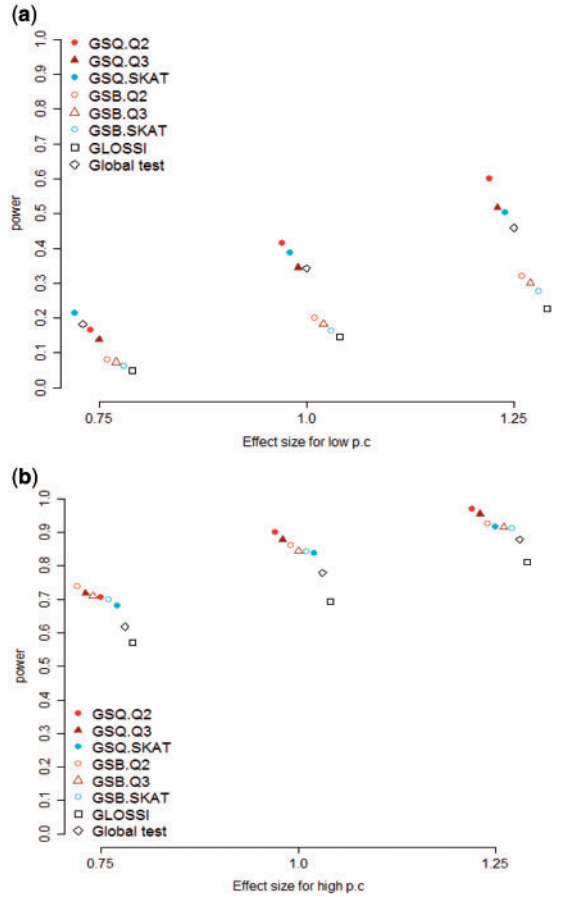

Fig. 2. Comparison of the power of gene-set-level tests for varying effect sizes; (a) for a low proportion of causal variants; (b) for a high proportion of causal variants

For a high proportion of causal variants, performance of GS_B-type tests is similar to or better than GS_Q-type tests when the effect size is 0.75 whereas the power of the $GS_Q.QTest_2$ and the $GS_Q.QTest_3$ exceeds the power of the other tests when effect size is larger than 0.75 (Fig. 2b).

3.3 Application for Korean liver enzymes and exome data

We applied the proposed tests to examine possible associations between liver enzymes and whole-exome sequencing data collected in the Korean Association REsource (KARE; Cho et al., 2009) study. After excluding samples from individuals who were taking medication likely to influence liver enzymes, 1058 samples were used for the proposed tests. For the analysis, 22 654 rare variants and 11 895 common variants of 8322 corresponding genes were included, after excluding low-quality variants and selecting functional variants. In tests for pathways, we used 0.05 or 0.5 as cut-off for maximum MAF and 2 for minimum MAC. We used 1320 canonical pathways from the MSigDB database v.4.0 (Subramanian *et al.*, 2005). The association was declared to be significant if the *P*-value was less than a threshold applying Bonferroni correction over the total number of tests.

For the single rare variants with a MAC > 5, we performed a linear regression in PLINK, and the minimum *P*-values for the three liver enzymes traits [alanine aminotransferase (ALT), aspartate aminotransferase (AST), and gamma glutamyltransferase (GGT)] were 5.18E-06, 5.23E-06 and 1.25E-05 (q-value = 0.31, 0.29, 0.43), respectively. Thus, no rare variant was found to be significantly associated with the three traits, after Bonferroni correction. For the genes with gene size >1, the results of the gene-level test based on rare variants showed that no gene was statistically significant if the Bonferroni correction was applied (See Supplementary Table S2).

In the gene-set level test for rare variants, we discovered that NODAL signaling pathway ($P = 3.4E\text{-}05$; $GS.QTest_3$) significantly associated with AST after Bonferroni correction, and in the gene-set level tests for common and rare variants, we discovered two KEGG pathways significantly associated with GGT after Bonferroni correction; beta-alanine metabolism ($P = 4.5E\text{-}06$; $GS.QTest_2$) and lysine degradation ($P = 4.7E\text{-}06$; $GS.QTest_2$). Table 2 provides the *P*-values from the proposed gene-set-level tests and those from the permutation tests by generating 10^{10} permuted samples. The *P*-values from our approach coincide well with those from the permutation tests for gene-set-level tests.

The results show similar patterns to those of simulation results. For example, the significances of Aurora pathway and ARF6 pathway were identified by the $GS.QTest_1$, since most genes within these pathways showed effects in the same direction (Supplementary Fig. S4). On the other hand, NODAL signaling pathway was significant by the $GS.QTest_2$ and the $GS.QTest_3$, because the genes within this pathway have effects with different directions (Supplementary Fig. S4). The parameter values used in simulation studies such as MAF, effect size and proportion of causal variants were similar to those in these pathways.

These results were applied to the replication study for further validation. Replication analyses for the pathway results were conducted in an independent study population including 897 Korean individuals, a portion of the Cardiovascular Disease Association Study (CAVAS) cohort (Kim et al., 2016a, b). NODAL signaling pathway was not replicated and some pathways including two KEGG pathways were successfully replicated ($P < 0.05$) (Table 3).

We also deconstructed the nature of the signals, which provides the list of key genes involved in the two replicated pathways along with the number of variants in each gene and the gene-level *P*-value (Supplementary Table S3). The gene-level *P*-values from the permutation tests were also computed using the 10^{10} permuted samples. The *P*-values of our gene-set-level tests coincide well with those from the permutation tests. The deconstruction of the two replicated pathways shows that ALDH2 [MIM 100650] is a key gene. The strong effect of ALDH2 and the weak effects of other genes together yielded a strong signal for the two gene sets. In the discovery study, ALDH2 consists of only one common variant, rs671, with a strong statistical significance (MAF = 0.15, single variant *P*-value = 9.58E-08) which has been reported to be associated with GGT (Kamatani *et al.*, 2010). In the replication study, ALDH2 consists of rs671 and two non-significant rare variants.

The two pathways identified here were known to relate to liver functions. Lysine is an essential amino acid that stimulates the biosynthesis of cholesterol in the liver (Schmeisser *et al.*, 1983). Beta-alanine is involved in liver function, and has a protective effect in the presence of toxins (Choi *et al.*, 2009; Lee and Kim, 2007). Graphical representation of genes in the pathway, including information on genetic variants, may help to understand the biological mechanisms underlying the function of the pathways and the accumulated genetic effect of the pathway. For example, Figure 3 shows the genes involved in the pathways and their relationships.

Each gene is colored according to the number of variants used in the association analysis. Red indicates enrichment of variants (≥10) whereas white indicates that no variant of the gene was used in the

Table 2. Gene sets found to be highly associated with liver enzymes in KARE ($P < 1E-03$)

Phenotype	Gene-set	KARE results						
		Maximum MAF	minimum MAC	*P*-value				
				GS. $QTest_1$	GS. $QTest_2$	GS. $QTest_3$	GS. SKAT	GS. SKATO
ALT	PID_AURORA_A_PATHWAY	0.05	2	5.0E-04 (4.2E-04)*	1.2E-01 (1.5E-01)	1.2E-03 (1.2E-03)	1.8E-01 (1.9E-01)	5.9E-03 (6.2E-03)
	BIOCARTA_RELA_PATHWAY	0.05	2	8.4E-04 (8.7E-04)	2.2E-02 (2.7E-02)	2.0E-03 (2.3E-03)	2.4E-01 (2.2E-01)	8.2E-02 (8.1E-02)
	KEGG_FOLATE_BIOSYNTHESIS	all	2	4.6E-02 (4.5E-02)	7.7E-04 (7.9E-04)	2.7E-03 (2.8E-03)	6.9E-01 (6.8E-01)	7.0E-01 (6.8E-01)
AST	REACTOME_SIGNALING_BY_NODAL	0.05	2	9.9E-01 (9.9E-01)	4.2E-05 (4.1E-05)	3.4E-05 (3.4E-05)	1.1E-03 (9.8E-04)	2.3E-03 (2.2E-02)
	REACTOME_GRB2_SOS_PROVIDES_LINKAGE_TO_MAPK_SIGNALING_FOR_INTERGRINS	0.05	2	4.7E-04 (4.5E-04)	4.3E-02 (4.3E-02)	1.1E-03 (1.0E-03)	1.3E-01 (1.4E-01).	1.1E-02 (1.3E-02)
	REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	0.05	2	7.5E-01 (7.5E-01)	7.5E-04 (7.8E-04)	1.2E-03 (1.3E-03)	9.3E-03 (9.5E-03)	1.9E-02 (1.8E-02)
	PID_RETINOIC_ACID_PATHWAY	all	2	9.7E-01 (9.6E-01)	9.6E-04 (9.3E-04)	1.2E-03 (1.0E-03)	4.1E-02 (4.3E-01)	8.3E-02 (8.3E-02)
GGT	PID_ARF6_PATHWAY	0.05	2	4.3E-04 (4.2E-04)	6.9E-02 (6.9E-02)	1.1E-04 (9.9E-05)	1.5E-01 (1.7E-01)	2.4E-03 (2.4E-03)
	KEGG_BETA_ALANINE_METABOLISM	all	2	2.1E-01 (1.9E-01)	4.7E-06 (5.1E-06)	1.1E-05 (1.3E-05)	7.8E-02 (7.8E-02)	8.6E-02 (8.6E-02)
	KEGG_LYSINE_DEGRADATION	all	2	4.0E-02 (4.5E-02)	4.5E-06 (4.7E-06)	1.9E-05 (2.0E-05)	3.0E-03 (2.8E-03)	2.0E-03 (2.0E-03)
	KEGG_BUTANOATE_METABOLISM	all	2	1.6E-01 (1.6E-01)	6.8E-05 (6.6E-05)	2.1E-04 (2.0E-04)	3.8E-02 (3.7E-02)	6.5E-02 (6.3E-02)
	REACTOME_ETHANOL_OXIDATION	all	2	5.0E-04 (5.3E-04)	6.1E-05 (6.1E-05)	2.5E-04 (2.6E-04)	1.7E-07 (1.6E-07)	2.9E-07 (3.1E-07)
	KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	all	2	2.4E-01 (2.7E-01)	8.5E-05 (8.3E-05)	1.0E-03 (1.1E-03)	8.8E-02 (8.6E-02)	1.6E-01 (1.6E-01)
	REACTOME_IL_6_SIGNALING	all	2	2.4E-04 (2.3E-04)	1.8E-03 (1.8E-03)	5.6E-04 (5.9E-04)	4.1E-03 (4.5E-03)	1.8E-03 (1.9E-03)

*Permutation *P*-values are given in parentheses.

Table 3. Results of the replication study conducted in CAVAS

Phenotype	Gene-set	Replication results						
		Maximum MAF	Minimum MAC	*P*-value				
				$GS.QTest_1$	$GS.QTest_2$	$GS.QTest_3$	GS. SKAT	GS. SKATO
GGT	KEGG_BETA_ALANINE_METABOLISM	all	1	3.6E-08	2.8E-01	8.3E-05	1.5E-01	5.9E-02
	KEGG_LYSINE_DEGRADATION	all	1	1.2E-06	5.6E-01	6.9E-05	1.4E-01	1.1E-01

association analysis. Because we performed the association analysis using only functional variants, such as non-synonymous, loss-of-function and splice-site variants, the number of rare functional variants accumulated for each gene can be interpreted as the functional genetic effect on the gene. Several complexes and genes directly related to beta-alanine were shown to be enriched with variants. This indicates that the accumulated functional genetic effects of variants in the beta-alanine metabolic pathway may disrupt the function of the pathway.

Table 2 shows the association results accompanied by their suggestive statistical significance ($P \leq 1E\text{-}03$). Most of the pathways revealed by the test were also biologically plausible in the context of liver function. The overexpression of Aurora-A, which frequently occurs in hepatocellular carcinoma (HCC), causes p53-dependent pre-mitotic arrest during liver regeneration (Li *et al.*, 2009). RelA pathway is associated with the prevention of hepatic apoptosis in mice (Rosenfeld *et al.*, 2000). Folate can affect a decrease of serum ALT level in hypertensive patients (Qin *et al.*, 2012). The NODAL signaling pathway, combined with NANOG, plays a critical role in HCC metastasis (Sun *et al.*, 2013). ARF6 plays an essential role in hepatic cord formation during liver development in mice (Suzuki et al., 2006). Butanoate metabolism is one of the pathways which displayed significant differences during liver development in HCC patients (Diana *et al.*, 2012). Ethanol oxidation was found to occur via hepatic enzymes or liver microsomes (Lieber *et al.*, 1987). IL-6 was upregulated in human HCC and it plays an important role in a tumor development (Ji et al., 2010). Valine, leucine and isoleucine are found to mediate activation of important hepatic metabolic signaling pathways and play a critical role in development of liver disease (Lake et al., 2015).

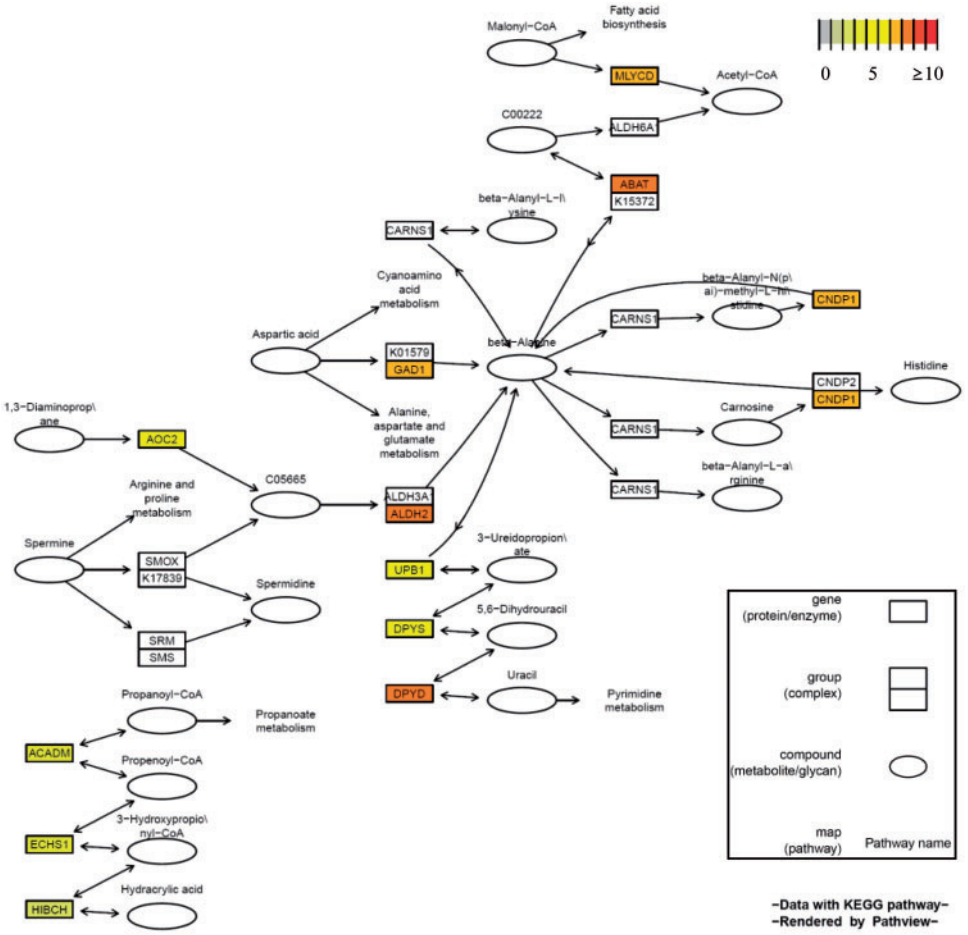

Fig. 3. Beta-alanine metabolism pathway found to be associated with liver enzymes

4 Discussion and conclusions

Many methods focusing on gene-level associations have recently been developed to assess the association between rare variants and complex diseases. However, those methods have some limitations in terms of power because the number of samples of rare variants available is not as large as the number of common variants. Even for large sample sizes, no methods for single-gene analysis have consistently high power across the various rare variant scenarios. In this article, we propose gene-set tests as well as single-gene tests for NGS data, including rare and common variants.

Recent rare-variant association methods usually focus on gene-level analysis. However, the interplay of rare variants affecting complex diseases arises not only at the gene level, but also at the gene-set or pathway level. If multiple mutations in the same functional class can influence a disease or trait, then a gene-set, as well as a gene, can be a key functional class. Here, the quadratic tests for multiple variants within a gene are first defined, and then an efficient method for gene-set-level associations is introduced. The performance of the proposed gene-set tests is demonstrated by comparison with other gene-set-level association methods in various simulation scenarios. This demonstration shows that one-step tests are limited in their applications to gene-set-level analysis, compared to two-step tests. When a gene set contains a large number of variants, one-step tests using SKAT or GlobalTest often lose power because of the many non-causal variants present in a given gene set. Based on our simulation studies, the GS_B.QTest performs best among other gene-set methods in cases containing a large proportion of causal variants with small effects. In many cases, the GS_Q.QTest outperforms other methods.

Our simulation raised some issues that need to be considered for optimization of the performance of rare-variants association methods. In the gene-level analysis, the first issue is that the power of the association tests depends on the proportion and effect sizes of the causal rare variants in a region. In some cases, a small proportion of the causal rare variants affect the disease traits, and in other cases, a large proportion of the causal variants affect them. Non-burden tests, such as the $QTest_2$ and the SKAT, can sensitively detect signals in the former cases. Burden tests, such as the $QTest_1$ and the GRANVIL, performed well in the latter cases, especially when the effect size is small or moderate (e.g. 0.5 or 0.75). However, in our simulation studies, the optimal tests, such as the SKAT-O and the $QTest_3$, yielded higher power than the burden tests except when all the variants were causal and one-directional. In the former case, optimal tests also yielded higher power because they take advantage of the properties of both burden and non-burden tests, and usually maintain consistently high power. In our simulation studies, when the effect of the causal variant was large (e.g. 1.25 or 1.5), the proposed $QTest_2$ and $QTest_3$ were found to detect signals sensitively, regardless of the proportion of causal variants.

The second issue is the direction of the causal variant effects. The burden tests assume that rare variants in a given region have effects in the same direction. However, this is not always the case. When deleterious and protective alleles are combined, the burden tests may lose power because they do not consider opposing effects. On the contrary, the non-burden tests, which do consider effects that operate in different directions, may lose power when the causal variants have signals of the same direction. The $QTest_3$, which combines the burden and non-burden tests, tends to yield fairly consistent power.

The third issue is the existence of common variants within a region. Most association tests for sequencing data focus only on rare variants and give greater weight to rarer variants as a function of MAF. For this reason, the common causal variants in sequencing data would not be focused in these tests. However, it may be better to consider the case when both rare and common variants exist together within a region. QTests can analyze both rare and common variants, because rare and common variants can be included simultaneously as explanatory variables in a multiple regression. QTests are also robust to inclusion of non-causal common variants.

Through extensive simulation studies, we investigated these issues in detail. Our simulation studies for gene-level and gene-set-level analyses show that the proposed QTests maintain consistent performance and high power, regardless of the above issues. The GS_Q.QTests do not lose power, even when faced with a small proportion of causal variants or small effect size, and are robust to inclusion of protective variants with negative effect as well as common variants. In cases with a high proportion of causal variants, the GS_B.QTests increase the power of the gene-set association analysis, but suffer from power loss in the presence of non-causal variants.

To account for heterogeneity between genes, GS_B.QTests can use a function of gene units or of biological importance as a weight. In the case of GS_Q.QTests, the development of a new approach using the weighted sum of gamma random variables is desirable in a future study that considers heterogeneity.

The current versions of the proposed tests can assess only quantitative traits. If the trait in question is binary, then the association statistic (e.g. *P*-value, chi-square statistic) for each rare variant appears not to be stable, in the sense that they can have a very large standard error. However, by collapsing rare variants within a subregion first or by providing a penalty parameter to better estimate the regression coefficients, the proposed QTests could be extended to analysis of binary traits.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant (2012R1A3A2026438), the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the NRF and an intramural grant from the Korea National Institute of Health (2014-NI73001-00), the Republic of Korea. Sequencing data from the T2D-GENES Consortium was supported by NIH/NIDDK U01's DK085501, DK085524, DK085526, DK085545 and DK085584. The exome chip data was supported by the Korean Genome Analysis Project (4845-301), the Korean Genome and Epidemiology Study (4851-302) and the Korea Biobank Project (4851-307) of the Korea Center for Disease Control and Prevention, Republic of Korea.

References

Adzhubei,I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Bansal,V. *et al.* (2010) Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.*, **11**, 773–785.

Chai,H.S. *et al.* (2009) GLOSSI: a method to assess the association of genetic loci-set with complex diseases. *BMC Bioinformatics*, **10**, 102.

Cho,Y.S. *et al.* (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.*, **41**, 527–534.

Choi,D. *et al.* (2009) Taurine depletion by beta-alanine inhibits induction of hepatotoxicity in mice treated acutely with carbon tetrachloride. *Adv. Exp. Med. Biol.*, **643**, 305–311.

Diana,B. *et al.* (2012) Genetic signatures shared in embryonic liver development and liver cancer define prognostically relevant subgroups in HCC. *Mol. Cancer*, **11**, 55. DOI: 10.1186/1476-4598-11-55.

Goeman,J.J. *et al*. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Ji,J. *et al*. (2010) MicroRNA expression, survival, and response to interferon in liver cancer. *N. Engl. J. Med*., **361**, 1437–1447.

Kamatani,Y. *et al*. (2010) Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat. Genet*., **42**, 210–216.

Kim,Y. *et al*. (2016a) Cohort profile: the Korean genome and epidemiology study (KoGES) Consortium. *Int. J. Epidemiol*., **45**, DOI: 10.1093/ije/dyv316.

Kim,Y.K. *et al*. (2016b) Evaluation of pleiotropic effects among common genetic loci identified for cardio-metabolic traits in a Korean population. *Cardiovasc. Diabetol*., **15**, 20.

Ladouceur,M. *et al*. (2012) The empirical power of rare variant association methods: results from Sanger sequencing in 1,998 individuals. *PLoS Genet*., **8**, e1002496.

Lake,A.D. *et al*. (2015) Branched chain amino acid metabolism profiles in progressive human nonalcoholic fatty liver disease. *Amino Acids*, **47**, 603–615.

Lee,S.Y. and Kim,Y.C. (2007) Effect of beta-alanine administration on carbon tetrachloride-induced acute hepatotoxicity. *Amino Acids*, **33**, 543–546.

Lee,S. *et al*. (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, **13**, 1–14.

Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet*., **85**, 311–321.

Li,C.C. *et al*. (2009) Aurora – a overexpression in mouse liver causes p53-dependent premitotic arrest during liver regeneration. *Mol. Can. Res*., **7**, 678–688.

Li,B. *et al*. (2012) SimRare: a program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits. *Bioinformatics*, **28**, 2703–2704.

Lieber,C.S. *et al*. (1987) The microsomal ethanol oxidizing system and its interaction with other drugs, carcinogens, and vitamins. *Ann. N. Y. Acad. Sci*., **492**, 11–24.

Lin,D. and Tang,Z. (2011) A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet*., **89**, 354–367.

Madsen,B.E. and Browning,S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*., **5**, e1000384.

Manolio,T.A. *et al*. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Morris,A.P. and Zeggini,E. (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol*., **34**, 188–193.

Peng,G. *et al*. (2010) Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet*., **18**, 111–117.

Price,A.L. *et al*. (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet*., **86**, 832–838.

Qin,X. *et al*. (2012) Effect of folic acid intervention on ALT concentration in hypertensives without known hepatic disease: a randomized, double-blind, controlled trial. *Eur. J. Clin. Nutr*., **66**, 541–548.

Rosenfeld,M.E. *et al*. (2000) Prevention of hepatic apoptosis and embryonic lethality in RelA/TNFR-1 double knockout mice. *Am. J. Pathol*., **156**, 997–1007.

Schmeisser,D.D. *et al*. (1983) Effect of excess dietary lysine on plasma lipids of the chick. *J. Nutr*., **113**, 1777–1783.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Sun,C. *et al*. (2013) NANOG promotes liver cancer cell invasion by inducing epithelial-mesenchymal transition through NODAL/SMAD3 signaling pathway. *Int. J. Biochem. Cell Biol*., **45**, 1099–1108.

Suzuki,T. *et al*. (2006) Crucial role of the small GTPase ARF6 in hepatic cord formation during liver development. *Mol. Cell Biol*., **26**, 6149–6156.

Wang,K. *et al*. (2007) Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet*., **81**, 1278–1283.

Wu,M.C. *et al*. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet*., **89**, 82–93.

Xb,X. *et al*. (2004) Obstruction of TGF-beta1 signal transduction by anti-Smad4 gene can therapy experimental liver fibrosis in the rat. *Zhonghua Gan Zang Bing Za Zhi*, **12**, 263–266.

Zhao,J. *et al*. (2014) Pathway analysis with next-generation sequencing data. *Eur. J. Hum. Genet*., doi:10.1038/ejhg.2014.121.