Gao et al. BMC Proceedings 2018, 12(Suppl 9):49
https://doi.org/10.1186/s12919-018-0136-7

BMC Proceedings

PROCEEDINGS Open Access

Methods to evaluate rare variants gene-age interaction for triglycerides

Tony Huayang Gao[1], Jianjun Zhang[2], Diaz Medina Miguelangel[3] and Xuexia Wang[2*]

From Genetic Analysis Workshop 20
San Diego, CA, USA. 4 - 8 March 2017

Abstract

Triglycerides are an important measure of heart health. Although more than 90 genes have been found to be associated to lipids, they only explain 12 to 15% of the variance in lipid levels. Evidence suggests that age may interact with the genetic effect on lipid levels. Existing methods to detect the main effect of rare variants cannot be readily applied for testing the gene environment interaction effect of rare variants, as those methods either have unstable results or inflated Type I error rates when the main effect exists. To overcome these difficulties, we developed two statistical methods: testing of optimally weighted combination of single-nucleotide polymorphism (SNP) environment interaction (TOW-SE) and a variable weight TOW-SE (VW-TOW-SE) to test the gene environment interaction effect of rare variants by grouping SNPs into biologically meaningful SNP-sets (SNPs in a gene or pathway) to improve power and interpretability. The proposed methods can be applied to either continuous or binary environmental variables, and to either continuous or binary outcomes. Simulation studies show that Type I error rates of the proposed methods are under control. Comparing the two methods with the existing interaction sequence kernel association test (iSKAT), the VW-TOW-SE is the most powerful test and the TOW-SE is the second most powerful test when gene environment interaction effect exists for both rare and common variants. The three tests were applied to the GAW20 simulated data, among the five regions in which the main effect of common SNPs was simulated and the gene–age interaction effect was not included. As expected, none of the tests indicated positive results.

Background

Highly heritable triglycerides (TG) [1] are an important measure of heart health. Having excess levels of TG can increase the risk of heart disease. Identified common variants only explain 12% to approximately 15% of the variance in lipid levels [2]. A substantial proportion of lipid heritability is unexplained [3]. This suggests that rare (minor allele frequency [MAF] < 1%) or intermediate variants (0.01 < MAF < 0.05) with potentially larger effect sizes or other mechanisms, such as gene–environment interactions, may play a role in explaining the substantially missing heritability.

Clear evidence shows that lipids vary by age. A handful of lipid loci with age-dependent effects were identified from candidate gene studies and genome-wide association study (GWAS) [4, 5]. However, few of these explored the role of gene–age interaction for rare and intermediate variants in lipid levels. More than 74.6% of variants are rare and intermediate variants [6], which may have a larger effect size than common variants and explain substantial proportions of lipid variance. In this GAW20 study, we attempted to detect the effect of gene–age interactions on TG for rare and intermediate variants with novel statistical methods.

Due to the allelic heterogeneity and the extreme rarity of individual variants [7], most existing methods focus on improving the power of detecting gene–environment (G × E) interactions only for individual markers, especially for common variants, and are not optimal for detecting rare variants. Although there has been interest in multiple-marker analysis by grouping single-nucleotide polymorphisms (SNPs) into biologically meaningful SNP-sets (eg, SNPs in a gene or pathway) to improve power and interpretability, the existing SNP set analysis has focused on testing for the marginal effect of a SNP set [8, 9]. Limited work has been done on testing the interactions between a SNP set and an

* Correspondence: Xuexia.Wang@unt.edu
[2]Department of Mathematics, University of North Texas, 1155 Union Circle #311430, Denton, TX 76203, USA
Full list of author information is available at the end of the article

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

environmental variable, especially as it pertains to rare variants. Although the SNP-set–based interaction sequence kernel association test (iSKAT) [10] can be applied to detect G × E interactions in rare variants, its power is very restricted and lacks robustness to the shape of the data in many circumstances. Motivated by the need for powerful methods to test G × E interactions for rare variants, we developed two novel methods: testing of optimally weighted combination of SNP environment interaction (TOW-SE) and a variable weight TOW-SE (VW-TOW-SE) to identify G × E interactions for SNP sets of common and/or rare variants in GWAS, exome, or next-generation sequencing data. Our simulation studies show that the Type I error rates of the proposed methods are under control. Comparing the two methods with the iSKAT, and the VW-TOW-SE is the most powerful test, TOW-SE is the second most powerful test when G × E interaction effect exists for both rare and common variants.

The Genetic Analysis Workshop (GAW) 20 "SimulationBestOneRepresentative" data includes TG levels before and after treatment with fenofibrate and genotyped genome-wide SNPs from the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study [11]. We imputed chromosomes 1, 6, 8, 9, 10, and 17, where five major main-effect causal SNPs of TG reside. We applied the proposed methods (TOW-SE and VW-TOW-SE) and iSKAT to the unrelated individuals (sample size n = 246) to test TG susceptible to gene–age interactions on the five imputed genes. The main effect of common SNP was simulated in the five regions. However, the gene–age interaction effect was not included. As expected, using the proposed methods, none of the regions indicated significantly gene–age interaction effects on TG.

Methods

Consider a sample of n individuals. Each individual has been genotyped at M variants in a genomic region (a gene or a pathway). For the i^{th} individual, denote y_i as the trait value (continuous or binary); E_i as the environmental variable (continuous or binary); $G_i = (g_{i1}, \cdots, g_{iM})$ as the genotypic scores at M variants, where $g_{im} \in \{0, 1, 2\}$ is the number of minor alleles the i^{th} individual has at the m^{th} variant. Z_i denotes the potential confounder covariates.

We use the generalized linear model (GLM):

$$f\left(E(y_i|G_i, E_i, Z_i)\right) = \alpha_0 + Z_i\alpha + E_iG_i\beta + G_i\zeta + \eta E_i \tag{1}$$

to model the relationship between trait values and G × E interactions E_iG_i, where $f(\cdot)$ is a monotone "link" function. For a quantitative trait, $f(\cdot)$ will be an identity link function. For a binary trait, a logit link function will be used. Coefficients of each term in Eq. (1) are denoted by α_0, a, β, ζ, and η, respectively. To test for the G × E interaction for a SNP set of M SNPs is equivalent to testing the null hypothesis $H_0 : \beta = 0$ in Eq. (1).

To test $H_0 : \beta = 0$ in Eq. (1), we developed a score test by treating α_0, a, ζ, and η as nuisance parameters. First, we adjusted both trait value y_i and G × E interaction E_iG_i for the covariates Z_i, the genotypic score G_i and the environmental variable E_i by applying linear regression and obtaining residuals. Denote $\tilde{y}_i$ as the residual of y_i and $\tilde{X}_i = (\tilde{x}_{i1}, \cdots, \tilde{x}_{iM})$ as the residual of E_iG_i. Then, the relationship between $\tilde{y}_i$ and $\tilde{X}_i = (\tilde{x}_{i1}, \cdots, \tilde{x}_{iM})$ can be modeled by the GLM:

$$f\left(E(\tilde{y}_i|\tilde{X}_i)\right) = \beta_0{}^* + \tilde{X}_i\beta^* \tag{2}$$

To test $H_o : \beta = 0$ in equation (1) is equivalent to test $H_o : \beta_* = 0$ in equation (2). Sha et al. [12] proposed a score test to test $H_o : \beta^* = 0$ in GLM. However, for rare variants' SNP–environment interactions, the score test may lose power as a consequence of the sparse data and a large degree of freedom. To increase power by effectively using information from data, we proposed to test the G × E interactions by testing the effect of a weighted combination of SNP–environment interactions, $\tilde{x}_i = \sum_{m=1}^{M} w_m\tilde{x}_{im}$.

To test $\tilde{x}_i = \sum_{m=1}^{M} w_m\tilde{x}_{im}$, the score test is:

$$\begin{aligned} S(w_1, \cdots, w_M) &= n\frac{\left(\sum_{i=1}^{n}(\tilde{y}_i-\bar{\tilde{y}})(\tilde{x}_i-\bar{\tilde{x}})\right)^2}{\sum_{i=1}^{n}(\tilde{y}_i-\bar{\tilde{y}})^2\sum_{i=1}^{n}(\tilde{x}_i-\bar{\tilde{x}})^2} \\ &= n\frac{\left(\sum_{m=1}^{M}w_m\sum_{i=1}^{n}(\tilde{y}_i-\bar{\tilde{y}})(\tilde{x}_{im}-\bar{\tilde{x}}_m)\right)^2}{\sum_{i=1}^{n}(\tilde{y}_i-\bar{\tilde{y}})^2\sum_{i=1}^{n}(\tilde{x}_i-\bar{\tilde{x}})^2} \end{aligned} \tag{3}$$

It reaches its maximum $S_o(w_1^0, \cdots, w_M^0) = n\sum_{i=1}^{n}(\tilde{y}_i-\bar{\tilde{y}})(\tilde{x}_i^0-\bar{\tilde{x}}^0)/\sum_{i=1}^{n}(\tilde{y}_i-\bar{\tilde{y}})^2$ when $w_m^0 = \frac{\sum_{i=1}^{n}(\tilde{y}_i-\bar{\tilde{y}})(\tilde{x}_{im}-\bar{\tilde{x}}_m)}{\sum_{i=1}^{n}(\tilde{x}_{im}-\bar{\tilde{x}}_m)^2}$; $\tilde{x}_i^0 = \sum_{m=1}^{M} w_m^0\tilde{x}_{im}$, as rare variants are essentially independent. Thus, w_m^0 is the optimal weight. We define $T_{T-SE} = \sum_{i=1}^{n}(\tilde{y}_i-\bar{\tilde{y}})(\tilde{x}_i^0-\bar{\tilde{x}}^0)$ as the statistic to Test the effect of the Optimally Weighted combination of SNP-Environment interactions (TOW-SE), which is equivalent to $S_o(w_1^0, \cdots, w_M^0)$ when we use a permutation test to evaluate p-values.

We analytically derive optimal weights for TOW-SE. The optimal weight w_m^0 will put a big weight to SNP–environment interactions that have strong association with

the trait of interest and also adjust the direction of the association. Moreover, it will put big weights to SNP–environment interactions with small variations that are often rare variants. iSKAT assigns weights to variants based on the MAFs via a beta function. It put decent nonzero weights for variants with MAF in (0.01, 0.05). If MAFs of causal variants are not in the range of (0.01, 0.05), iSKAT will be less powerful than TOW-SE. TOW-SE targets rare variants and may lose power when testing G × E effects of both rare and common variants.

To test for the G × E interaction of both rare and common variants, we propose variable weight TOW-SE (VW-TOW-SE). We divide variants into rare and common. Let T_r and T_c denote the test statistic of TOW-SE for rare and common variants, respectively. Let $T_\lambda = \lambda \frac{T_r}{\sqrt{var(T_r)}} + (1-\lambda)\frac{T_c}{\sqrt{var(T_c)}}$. Denote p_λ as the p-value of T_λ. The test statistic of VW-TOW-SE is defined as $T_{VW-T-SE} = min_{0 \le \lambda \le 1} p_\lambda$. We will use permutations to evaluate p-values of both T_{T-SE} and $T_{VW-T-SE}$.

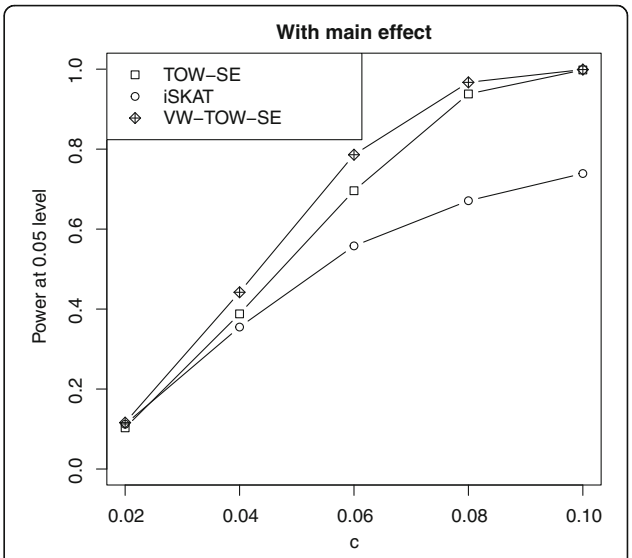

Fig. 1 Power comparisons of three tests (TOW-SE, iSKAT, VW-TOW-SE) for $n = 2000$ at $\alpha = 0.05$ level of significance for testing both rare and common variant G × E interaction effects on a continuous outcome with main effect

Simulations

Following the simulation setting in Lin et al. [10], we conducted simulation studies using the GAW17 empirical mini-exome sequenced data. The data set contains genotypes of 697 unrelated individuals on 3205 genes. Research shows that SNP rs11583200 on gene *ELAVL4* is associated with body mass index [13] and rare variants on gene *ELAVL4* are associated with the quantitative trait Q1 in the GAW17 data [14]. Therefore, we chose gene *ELAVL4* in our simulation study. There are 10 variants on gene *ELAVL4* of which 8 are rare variants and 2 are common variants. The rare variants threshold was chosen as 0.01. We use the program *fastPHASE* [15] to infer haplotypic phase for the 697 individuals and calculate haplotype frequencies. To generate the genotype of an individual, we generate 2 haplotypes according to the haplotype frequencies. The quantitative trait was generated using the following model:

$$Y = 0.5Z_1 + 0.5Z_2 + E\alpha_1 + \boldsymbol{G}^T\alpha_2 + \boldsymbol{EG}^T\beta + \boldsymbol{EG_c}B^c + \epsilon \quad (4)$$

where $Z_1 \sim N(0,1)$; $Z_2 \sim$ Binomial$(1, 0.5)$ and $\epsilon \sim N(0,1)$. The environmental variable E is assumed to be continuous following standard normal distribution and we set $\alpha_1 = 0.015$; $\boldsymbol{EG}$ is the rare variants G × E interaction and $\boldsymbol{EG_c}$ is one common variant G × E interaction.

Table 1 Type I error rates for both rare and common variants in the presence of main effects (*top panel*) and in the absence of main effects (*bottom panel*) for $n = 2000$

	α-level	TOW-SE	iSKAT	VW-TOW-SE
With main effect				
$n = 2000$	0.050	0.050	0.055	0.059
	0.010	0.011	0.012	0.015
	0.001	0.000	0.001	0.000
Without main effect				
$n = 2000$	0.05	0.051	0.061	0.056
	0.01	0.011	0.013	0.009
	0.001	0.000	0.003	0.001

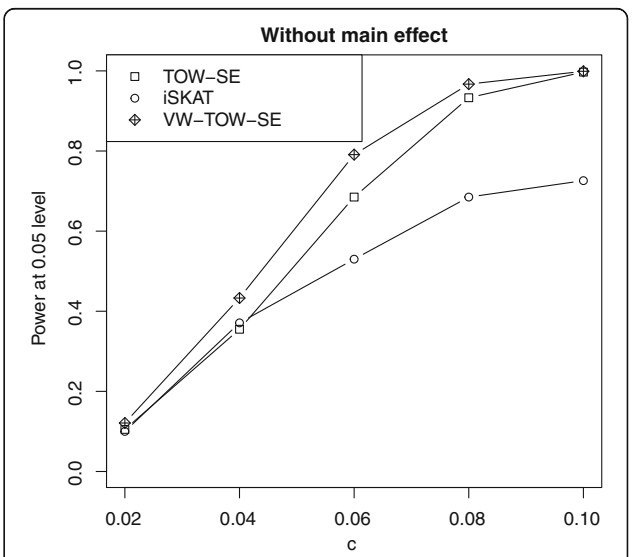

Fig. 2 Power comparisons of three tests (TOW-SE, iSKAT, VW-TOW-SE) for $n = 2000$ at $\alpha = 0.05$ level of significance for testing both rare and common variant G × E interaction effects on a continuous outcome without main effect

Table 2 Results of testing G × E interaction in the five causal regions using the three methods

Region name (SNP set)	Median of MAF (range)	$P_{TOW\text{-}SE}$	P_{iSKAT}	$P_{VW\text{-}TOW\text{-}SE}$
rs736004	0.013 (0.001–0.387)	0.020	0.022	0.019
rs1012116	0.015 (0.002–0.167)	0.024	0.030	0.021
rs4399565	0.006 (0.001–0.449)	0.022	0.023	0.019
rs9551059	0.007 (0.001–0.275)	0.031	0.030	0.037
rs10828412	0.003 (0.001–0.394)	0.100	0.099	0.095

We consider two scenarios: (a) with main effect and (b) without main effect in the model (4). When there are no main effects, we set the magnitudes of vector $\alpha_2 = 0.3$ for each element and their signs are randomly sampling from (−1, 1). When there are no main effects, we set $\alpha_2 = 0$. To evaluate the Type I error, we set β and β^c all to 0. To evaluate power, we vary the number of non-zero elements β_j in β. We set the magnitude of the nonzero β_j as $|\beta_j| = c$, and increase c from 0.02 to 0.1. Of the β_j, 50% are positive. β^c is positive and twice the magnitude of β_j. The sample size is 2000 for each scenario. P values are estimated by 10,000 permutations. The Type I error rates and power are evaluated using 1000 replicated samples.

GAW20 data analysis

We applied TOW-SE, VW-TOW-SE, and iSKAT to the GAW20 "SimulationBestOneRepresentative" data, which includes TG levels before and after treatment with fenofibrate and genotyped genome-wide SNPs from the GOLDN project [11]. We imputed chromosomes 1, 6, 8, 9, 10, and 17 with *minimac2* software [16]. Five major main effect causal SNPs (rs9661059, rs736004 [*LYRM4*], rs1012116, rs10828412, and rs4399565 [*HS3ST3A1*]) of TG reside on chromosomes 1, 6, 8, 9, 10, and 17, respectively. The 1000 Genomes project haplotypes integrated phase I served as the reference panel. It includes 1092 individuals. Our analysis was based on 246 unrelated individuals. Both pre-treatment and post-treatment TG values were provided for two visits. We used the log ratio of the pre-treatment mean and the post-treatment mean as the phenotype trait in our G × E interaction analysis. For individuals who did not have two visits, we just used the existing value. The median age is 64 years (range: 28–83 years). We used the centered age in our analysis. The median of the TG ratio is 1.54 (range: 0.72–4.24). We excluded 8 individuals from our analysis because of completely missing post-treatment TG values.

We evaluated the performance of TOW-SE, VW-TOW-SE, and iSKAT by testing G × E interaction effect on TG for the aforementioned 5 regions. Each region consists of 10 SNPs, and the fifth SNP in each region is the major main effect causal SNP of TG. Additionally, we assessed the effect of region size on the power of the tests using the region of rs4399565.

Results

Table 1 shows that the Type I error rates of all the three methods are under control. Power comparisons of the three tests (VW-TOW-SE, TOW-SE, and iSKAT) for different values of G × E effect for rare and common variants are given in Fig. 1 (with main effect) and Fig. 2 (without main effect). The power of the three tests increases as the effect size increases. When there is a G × E interaction effect for both rare and common variants, VW-TOW-SE is the most powerful test and TOW-SE is the second most powerful test. Table 2 shows that the median of the MAF ranges from 0.003 to 0.013 in the 5 regions. When we apply the three tests to test G × E interaction effect, under Bonferroni correction, none of the regions are significantly associated with TG. Table 3 suggests that all of the three methods perform better when the region size is larger.

Discussion

The computation time for TOW-SE and VW-TOW-SE using 10,000 permutations for analyzing 1000 individuals in a region that includes 50 SNPs is 9 s and 20 s, respectively. Suppose a whole genome sequencing data with 13,498,188 SNPs, TOW-SE will take 674 h (28 days) to conduct a whole genome analysis. The effect analysis of the region size suggests that the three methods will perform better when the region size is larger. However, the larger the region size, the higher chance for collinearity to appear in the region, which makes the computation more complex. To minimize the problem of collinearity, we recommend a region size between 10 SNPs and 30 SNPs when we apply the proposed methods to a genome-wide scan.

Conclusions

In summary, we developed two novel statistical methods: TOW-SE and VW-TOW-SE by grouping SNPs into

Table 3 Region size effect analysis for the three methods based on the region of rs4399565

Region size	Median of MAF (range)	$P_{TOW\text{-}SE}$	P_{iSKAT}	$P_{VW\text{-}TOW\text{-}SE}$
10 SNPs	0.006 (0.001–0.449)	0.022	0.023	0.019
20 SNPs	0.008 (0.001–0.449)	0.014	0.018	0.015
30 SNPs	0.008 (0.001–0.449)	0.013	0.014	0.012

biologically meaningful SNP-sets, which improved power and interpretability. Simulation studies show that the proposed methods yielded well-controlled Type I error rates under all study conditions. When gene environment interaction effect exists for both rare and common variants, VW-TOW-SE is the most powerful test, TOW-SE is the second most powerful test, and iSKAT is the least powerful test.

Funding
Publication of this article was supported by NIH R01 GM031575.

Availability of data and materials
The data that support the findings of this study are available from the Genetic Analysis Workshop (GAW), but restrictions apply to the availability of these data, which were used under license for the current study. Qualified researchers may request these data directly from GAW.

About this supplement
This article has been published as part of BMC Proceedings Volume 12 Supplement 9, 2018: Genetic Analysis Workshop 20: envisioning the future of statistical genetics by exploring methods for epigenetic and pharmacogenomic data. The full contents of the supplement are available online at https://bmcproc.biomedcentral.com/articles/supplements/volume-12-supplement-9.

Authors' contributions
XW designed the overall study. JZ, MDM, and TG conducted statistical analyses. TG, JZ, MDM, and XW drafted the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

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

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Texas Academy of Mathematics & Science, University of North Texas, 1155 Union Circle #311430, Denton, TX 76203, USA. [2]Department of Mathematics, University of North Texas, 1155 Union Circle #311430, Denton, TX 76203, USA. [3]Brady Corporation, 6555 W Good Hope Rd, Milwaukee, WI 53223, USA.

Published: 17 September 2018

References

1. Coram MA, Duan Q, Homann TJ, Thornton T, Knowles JW, Johnson NA, Robinson JG. Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. Am J Hum Genet. 2013;92(6):904–16.
2. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, Mora S, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45(11):1274–83.
3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindor LA, Hunter DJ, Cho JH. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53.
4. Snieder H, Van Doornen LJ, Boomsma DI. The age dependency of gene expression for plasma lipids, lipoproteins, and apolipoproteins. Am J Hum Genet. 1997;60(3):638.
5. Carvalho-Wells AL, Jackson KG, Gill R, Olano-Martin E, Lovegrove JA, Williams CM, Minihane AM. Interactions between age and apoE genotype on fasting and postprandial triglycerides levels. Atherosclerosis. 2010;212(2):481–7.
6. Morrison AC, Voorman A, Johnson AD, Liu X, Yu J, Li A, Muzny D, Yu F, Rice K, Zhu C, et al. Whole-genome sequence-based analysis of a model complex trait, high density lipoprotein cholesterol. Nat Genet. 2013;45(8):899–901.
7. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83(3):311–21.
8. Tzeng JY, Zhang D, Chang SM, Thomas DC, Davidian M. Gene-trait similarity regression for multimarker-based association analysis. Biometrics. 2009;65(3):822–32.
9. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010;86(6):929–42.
10. Lin X, Lee S, Wu MC, Wang C, Chen H, Li Z, Lin X. Test for rare variants by environment interactions in sequencing association studies. Biometrics. 2016;72(1):156–64.
11. Irvin MR, Kabagambe EK, Tiwari HK, Parnell LD, Straka RJ, Tsai M, Ordovas JM, Arnett DK. Apolipoprotein E polymorphisms and postprandial triglyceridemia before and after fenofibrate treatment in the genetics of lipid lowering and diet network (GOLDN) study. Circ Cardiovasc Genet. 2010;3(5):462–7.
12. Sha Q, Wang X, Wang X, Zhang S. Detecting association of rare and common variants by testing an optimally weighted combination of variants. Genet Epidemiol. 2012;36(6):561–71.
13. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, Croteau-Chonka DC. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):179–206.
14. Shi G, Gu CC, Kraja AT, Arnett DK, Myers RH, Pankow JS, Rao DC. Genetic effect on blood pressure is modulated by age. Hypertension. 2009;53(1):35–41.
15. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006;78(4):629–44.
16. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. Bioinformatics. 2015;31(5):782–4.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC