Multi-locus Test Conditional on Confirmed Effects Leads to Increased Power in Genome-wide Association Studies

Li Ma^{1,2,3*}, Shizhong Han^{3,9}, Jing Yang¹, Yang Da^{1*}

1 Department of Animal Science, University of Minnesota, Saint Paul, Minnesota, United States of America, **2** Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, United States of America, **3** Department of Psychiatry, Division of Human Genetics, School of Medicine, Yale University, New Haven, Connecticut, United States of America

Abstract

Complex diseases or phenotypes may involve multiple genetic variants and interactions between genetic, environmental and other factors. Current genome-wide association studies (GWAS) mostly used single-locus analysis and had identified genetic effects with multiple confirmations. Such confirmed single-nucleotide polymorphism (SNP) effects were likely to be true genetic effects and ignoring this information in testing new effects of the same phenotype results in decreased statistical power due to increased residual variance that has a component of the omitted effects. In this study, a multi-locus association test (MLT) was proposed for GWAS analysis conditional on SNPs with confirmed effects to improve statistical power. Analytical formulae for statistical power were derived and were verified by simulation for MLT accounting for confirmed SNPs and for single-locus test (SLT) without accounting for confirmed SNPs. Statistical power of the two methods was compared by case studies with simulated and the Framingham Heart Study (FHS) GWAS data. Results showed that the MLT method had increased statistical power over SLT. In the GWAS case study on four cholesterol phenotypes and serum metabolites, the MLT method improved statistical power by 5% to 38% depending on the number and effect sizes of the conditional SNPs. For the analysis of HDL cholesterol (HDL-C) and total cholesterol (TC) of the FHS data, the MLT method conditional on confirmed SNPs from GWAS catalog and NCBI had considerably more significant results than SLT.

Citation: Ma L, Han S, Yang J, Da Y (2010) Multi-locus Test Conditional on Confirmed Effects Leads to Increased Power in Genome-wide Association Studies. PLoS ONE 5(11): e15006. doi:10.1371/journal.pone.0015006

Editor: Amanda Ewart Toland, Ohio State University Medical Center, United States of America

Received: August 14, 2010; **Accepted:** October 7, 2010; **Published:** November 16, 2010

Copyright: © 2010 Ma et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This research was supported in part by project MN-16-043 of the Agricultural Experiment Station at the University of Minnesota. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the NHLBI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: maxxx131@umn.edu (LM); yda@umn.edu (YD)

⁹ These authors contributed equally to this work

Introduction

Genome-wide association studies (GWAS) have identified genetic variants associated with a number of complex diseases or phenotypes [1], [2], [3] and some of these variants had confirmations from several studies [1]. Published GWAS studies typically used a single-locus test (SLT), in which each variant is tested individually for association with a specific phenotype. Single-locus analysis may not be the best approach in the presence of confirmed SNP effects, because confirmed effects become a component of random residuals and decrease statistical power for detecting new effects if those true effects are omitted in the analysis. Controlling genetic backgrounds in single-marker tests was well explored in the area of QTL mapping including Zeng's CIM [4] and Jansen's multiple regression method [5]. Roeder et al [6] proposed to use linkage genome scan results in GWAS to achieve higher power. In a recent meta-analysis on smoking behavior, a novel SNP was identified with genome-wide significance related to smoking conditional on a known SNP [7]. Conditional analysis was also used in other meta-analysis in GWAS [8]. In this study, we propose a multi-locus test (MLT) that tests each candidate SNP conditional on confirmed SNPs for GWAS analysis to increase the statistical power for detecting new

SNP effects, and we demonstrate the MLT method had increased statistical power relative to SLT using analytical formulae derived in this study and using simulation, case studies, and the Framingham Heart Study (FHS) data.

Methods

Predicted Statistical Power of MLT and SLT

The multiple linear regression model for the MLT method can be expressed as:

$$Y_i = \mu + Z_i \beta_Z + G_i \beta_G + G_{i,s} \beta_s + e_i \quad (1)$$

where μ = the population mean of the phenotypic values, $Z_i = 1 \times p$ vector of the p covariates for subject i ($i = 1, \dots, N$) to account for environment, population stratification, and other factors; $\beta_Z = p \times 1$ vector of the partial regression coefficients of the covariates Z_i ; $G_i = (G_{i,1}, G_{i,2}, \dots, G_{i,s-1}) = 1 \times (S-1)$ vector of the $S-1$ SNP genotypes that were confirmed to be associated with the phenotype for subject i , with $G_{i,j}$ taking values of 0, 1 or 2 according to the number of copies of the minor allele for subject i at SNP j ; $\beta_G = (S-1) \times 1$ vector of partial regression coefficients of $S-1$ conditional SNPs; $G_{i,s}$ = the genotype value of the candidate

SNP; β_s is the partial regression coefficient of the candidate SNP, and e_i = random residual that is assumed to follow $N(0, \sigma^2)$ normal distribution. The residual variance of Equation 1 is

$$\sigma_{e_i}^2 = \text{var}(y_i | G_{i,s}, G_i, Z_i) = \text{var}(e_i) = \sigma^2 \quad (2)$$

A standard t-test can be used for testing the significance of the candidate SNP based on testing the following hypotheses, $H_0: \beta_s = 0$, where β_s is the regression coefficient of the candidate SNP and is the M th element of $\beta = (\mu, \beta_Z, \beta_G, \beta_s)$ and $M = 1 + p + S$. One-sided statistical power was derived in this study for simplicity (two-sided t-test statistical power can be obtained similarly but is not considered here). Using a multiple linear regression framework, the power of the one-sided t-test can be formulated as:

$$P_I = \int_{t_\alpha}^{+\infty} f(t, \lambda, N - M) dt \quad (3)$$

where t_α denotes the $\alpha\%$ quantile value for t distribution and f denotes the non-central t-distribution with $N - M$ degrees of freedom and non-centrality parameter:

$$\lambda = \frac{\hat{\beta}_s}{\sqrt{\text{var}(\hat{\beta}_s)}} = \frac{\hat{\beta}_s}{\sqrt{\hat{\Sigma}_{MM}}} \quad (4)$$

where $\hat{\Sigma}_{MM}$ = the element at the M th row and M th column of variance-covariance matrix $\hat{\Sigma}$, and $\hat{\Sigma}$ is a $M \times M$ variance-covariance matrix of $\hat{\beta}_s$ and can be estimated as:

$$\hat{\Sigma} = \text{var}(\hat{\beta}) = (X'X)^{-1} \sigma^2 \quad (5)$$

where X is the design matrix in Equation 1.

For SLT, the statistical model is the same as Equation 1 except that the residual term is now a summation of confirmed effects and random residuals. The residual variance for the SLT model is no longer σ^2 and has the following mathematical expression:

$$\sigma_{e_i}^2 = \text{var}(y_i | Z_i, G_{i,s}) = \text{var}(G_i \beta_G + e_i) = \beta_G^T \text{var}(G_i) \beta_G + \sigma^2 \quad (6)$$

where $\text{var}(G_i)$ is calculated based on the G_i values defined in Equation 1. Equation 6 shows that the residual variance of the SLT model is inflated over the MLT model of Equation 1 due to the fact that confirmed SNP effects are now in the residual term of Equation 6. Using this inflated estimated residual variance, the variance of $\hat{\beta}_s$ can be estimated by:

$$\text{var}(\hat{\beta}_s) = c \sigma_{e_i}^2 \quad (7)$$

where c is the element at the $(p+2)$ th row and $(p+2)$ th column of matrix $(X'X)^{-1}$ and X is the design matrix for the regression model of SLT. Therefore, the t-test statistic for SLT does not have a t-distribution but a t-distribution divided by a constant

$$\tau = \sqrt{\frac{c \sigma_{e_i}^2}{\hat{\Sigma}_{MM}}}$$

Similar to Equations 3, the power of the one-sided t-test can be formulated as:

$$P_I = \int_{t_\alpha}^{+\infty} f(t, \lambda, N - p - 2) dt \quad (8)$$

where t_α denotes the $\alpha\%$ quantile value for t-distribution with $N - p - 2$ degrees of freedom and f denotes the non-central t-distribution with $N - p - 2$ degrees of freedom and non-central parameter λ .

Results and Discussion

Evaluation of Statistical Power of MLT and SLT Using Simulation and Real Data

The analytical formulae for statistical power for MLT accounting for confirmed effects and for SLT without accounting for confirmed effects (Equation 3 and Equation 8) were validated by simulated data of 2000 subjects for various effect sizes of the candidate SNP and confirmed SNPs with 10,000 repeats. The phenotypic values were simulated by the summation of a population mean, three additive SNP effects and a random error which followed a standard normal distribution. The three SNPs were simulated under Hardy-Weinberg equilibrium and linkage equilibrium with allele frequencies, 0.3, 0.4 and 0.2. The first two SNPs were assumed to have confirmed effects and the last SNP was assumed to be the candidate SNP. The candidate SNP was tested by the MLT and SLT methods in each simulation. Empirical power was calculated as the proportion of significant results from all 10,000 simulation results. We fixed the effects of the two confirmed SNPs as 0.3 and 0.2 standard deviation of residuals (SD) and varied the effect of the candidate SNP from 0.04 to 0.2 SD. Simulated statistical power were nearly identical to the predicted power for MLT and SLT based on different candidate SNP effect sizes (Table S1) and on different population sizes (Table S2). With this knowledge of the power formulae being correct, predicted statistical power for MLT and SLT were calculated for various effect sizes of the confirmed SNPs (Table 1).

Table 1. Power comparison between MLT (Power I) and SLT (Power II) with various conditional SNP effect sizes and constant effect size of 0.1 SD for the candidate SNP.

Effect Size	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Power I	0.713	0.713	0.713	0.713	0.713	0.713	0.713	0.713	0.713	0.713
Power II	0.710	0.701	0.686	0.665	0.638	0.606	0.568	0.525	0.479	0.429
Improvement	0.003	0.012	0.027	0.048	0.075	0.108	0.146	0.188	0.235	0.284

doi:10.1371/journal.pone.0015006.t001

We further evaluated predicted statistical power using reported effect sizes for some confirmed SNP effects. We collected all reported SNP effects for HDL cholesterol (HDL-C), LDL cholesterol (LDL-C), triglycerides (TG), total cholesterol (TC), and serum metabolites (SM) from the GWAS catalog [1]. The effect sizes and risk allele frequencies of those SNPs were extracted and utilized for the power calculations. After filtering out SNPs in high linkage disequilibrium (LD) by only keeping one SNP with the largest effect size in each high LD region, the final selection of confirmed SNP markers included 22 relatively independent SNPs with effect sizes of 0.07–0.24 SD for HDL-C, 24 SNPs with effect sizes of 0.07–0.35 SD for LDL-C, 13 SNPs with effect sizes of 0.06–0.42 SD for TG, and 19 SNPs with effect sizes of 0.06–0.24 SD for TC. For SM, we extracted five SNPs which explained 5.6 to 36.3 percent of the total phenotypic variation [1], [2]. Conditional on those known significant SNP effects, statistical power of MLT increased over that of SLT by about 4–5% for HDL-C, LDL-C, TG and TC. The pattern of the heatmap for statistical power was similar for these four traits and the heatmap for HDL-C is shown in Figure 1A. Largest improvements were in the region where candidate SNPs had small effect sizes and large allele frequencies or candidate SNPs had medium effect sizes and relatively low allele frequencies (0.1–0.2). The increase in statistical power of MLT over SLT was much larger for SM, varying from 10% to 30%, because of the large effect sizes of the known SNPs (Figure 1B).

For GWAS analysis using real data, true statistical power is not observable but MLT is expected to have more significant results than SLT. To compare observed statistical significance of MLT and SLT, we used the FHS GWAS data (version 2) that had 6575 individuals with SNP genotypes of the 500k SNP panel from dbGAP [9]. Of the 6575 individuals, 6078 individuals had observations on HDL-C and 6431 individuals had observations on TC. From the 500k SNP panel, 432,096 SNP markers with

known locations and minor allele frequencies 0.01 or greater were selected and tested. The original cholesterol measures deviated from normality and had outliers. The Box-Cox transformation analysis [10] implemented by the R package [11] showed that the log-transformation was approximately the best transformation to achieve normality for HDL-C and TC. Age, age-squared, cholesterol treatment, blood sugar, body mass index, smoking status, number of cigars smoked, alcohol consumption and sex were adjusted for transformed HDL-C. Blood sugar, body mass index, smoking status, and sex were adjusted for transformed TC. The testing of SNP effects used the generalized least squares version [12] of epiSNP [13]. From the GWAS catalog [1] and NCBI (<http://www.ncbi.nlm.nih.gov>), we selected six SNP markers (Table 2) with multiple confirmations. These six SNP markers were independent of each other because pairwise correlations measured by R-square among these SNPs were nearly zero. Results showed that MLT had more significant results than SLT for both HDL-C and TC (Table 2). The first two markers in Table 2 had the largest improvement in observed significance (reduced P-value), while the remaining markers only had minor improvement.

In our analysis, we did not impute genotypic data to fill in missing genotypes so that the MLT test had smaller sample size than SLT. The observed significance should have been larger than observed in Table 2 if the missing genotypes were filled by genotype imputing using software like MACH [14] and BEAGLE [15]. Although improvement in statistical significance could be small in some cases, such improvement could be easily achieved using GWAS analysis software like PLINK [16] and epiSNP [13] that provide options to incorporate covariates. Due to incorporating confirmed effects, MLT has smaller degrees of freedom for residuals and can be less robust than SLT. Fortunately, sample size in typical GWAS studies (thousands) is large enough for incorporating relatively small number of confirmed effects (tens).

Figure 1. Power improvement of the MLT method conditional on confirmed SNP effects over the SLT method. a, HDL Cholesterol. b, Serum metabolites.

doi:10.1371/journal.pone.0015006.g001

Table 2. Comparison of P-values of MLT and SLT using the Framingham Heart Study data.

SNP	P-value		Phenotype	Conditional SNPs
	SLT	MLT		
rs1800775	7.06×10^{-34}	4.85×10^{-35}	HDL-C	rs765547, rs6507945, rs17259942
rs765547	2.19×10^{-12}	9.53×10^{-13}	HDL-C	rs1800775, rs6507945, rs17259942
rs6507945	9.77×10^{-10}	9.51×10^{-10}	HDL-C	rs1800775, rs765547, rs17259942
rs17259942	2.77×10^{-09}	1.52×10^{-09}	HDL-C	rs1800775, rs765547, rs6507945
rs599839	1.23×10^{-15}	9.77×10^{-16}	TC	rs4245791
rs4245791	3.32×10^{-09}	3.21×10^{-09}	TC	rs599839

doi:10.1371/journal.pone.0015006.t002

With more and more confirmed effects reported in the literature, MLT is useful for identifying new genetic variants with smaller effects.

Supporting Information

Table S1 Power comparison between analytic formulas and simulation for MLT (Power I) and SLT (Power II) with varied candidate SNP effect sizes, constant confirmed effect sizes (0.2 and 0.3 SDs for two confirmed SNPs) and constant sample size (2000). (DOC)

Table S2 Power comparison between analytic formulas and simulation for MLT (Power I) and SLT (Power II) with varied sample sizes, constant candidate SNP effect size (0.1 SD) and

confirmed effect sizes (0.2 and 0.3 SDs for two confirmed SNPs). (DOC)

Acknowledgments

We are thankful to Dr. Andrew Clark and Dr. Alon Keinan for helpful discussions on this manuscript. We also thank the reviewers for their comments that improved the presentation of the manuscript.

Author Contributions

Conceived and designed the experiments: LM SH YD. Performed the experiments: LM JY. Analyzed the data: LM JY. Wrote the paper: LM SH YD.

References

- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 106: 9362–9367.
- Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, et al. (2010) A genome-wide perspective of genetic variation in human metabolism. *Nat Genet* 42(2): 137–41.
- Sandhu MS, Waterworth DM, Debenham SL, Wheeler E, Papadakis K, et al. (2008) LDL-cholesterol concentrations: a genome-wide association study. *Lancet* 371(9611): 483–491.
- Zeng Z-B (1994) Precision mapping of quantitative trait loci. *Genetics* 136: 1457–1468.
- JANSEN RC, STAM P (1994) High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* 136: 1447–1455.
- Roeder K, Bacanu SA, Wasserman L, Devlin B (2006) Using linkage genome scans to improve power of association in genome scans. *Am J Hum Genet* 78: 243–252.
- Saccone NL, Culverhouse RC, Schwantes-An T-H, Cannon DS, Chen X, et al. (2010) Multiple Independent Loci at Chromosome 15q25.1 Affect Smoking Quantity: a Meta-Analysis and Comparison with Lung Cancer and COPD. *PLoS Genet* 6(8): e1001053. doi:10.1371/journal.pgen.1001053.
- Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, et al. (2010) Biological, clinical, and population relevance of 95 loci for blood lipids. *Nature* 466: 707–713.
- Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, et al. (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 39: 1007–1181.
- Box GEP, Cox DR (1964) An analysis of transformations (with discussion). *J Roy Stat Soc B* 26: 211–252.
- R Development Core Team. R (2008) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Ma L, Amos CI, Da Y (2008a) Accounting for correlations among individuals for testing SNP single-locus and epistasis effects in Genome-wide association analysis (Abstract). *Plant & Animal Genomes XVI Conference*, January 12–16, San Diego, CA. http://www.intl-pag.org/16/abstracts/PAG16_P11_903.html
- Ma L, Runesha HB, Dvorkin D, Garbe JR, Da Y (2008b) Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies. *BMC Bioinformatics* 9: 315.
- Li Y, Abecasis GR (2006) Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference. *Am J Hum Genet* S79: 2290.
- Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81: 1084–1097.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* 81: 559–575.