Published in final edited form as:

Genet Epidemiol. 2013 December ; 37(8): 831–839. doi:10.1002/gepi.21768.

Quantitative Allelic Test - a Fast Test for Very Large Association Studies

Sang Mee Lee¹, Theodore G Karrison¹, Nancy J Cox², and Hae Kyung Im¹

¹Department of Health Studies, University of Chicago, Chicago, Illinois

²Department of Medicine, University of Chicago, Chicago, Illinois

Abstract

Advances in high throughput technology have enabled the generation of unprecedented amounts of genomic data (e.g., next generation sequence data, transcriptomics, metabolomics, and proteomics), which promises to unravel the genetic architecture of complex traits. These discoveries may lead to novel therapeutic targets, guide disease prevention, and enable personalized medicine. However, the pace of data generation surpasses the ability to process and analyze the vast amounts of data. For example, in a typical study of transcription regulation, the relationship between more than 1 million genetic variants and ten thousand transcript levels are explored, requiring tens of billions of tests. In order to address this problem, we propose a fast, accurate, and robust method that can assess the significance of associations between quantitative phenotypes and genotypes. The method is an extension of the allelic test commonly used in case-control studies for the analysis of quantitative traits. We show the asymptotic equivalence of the proposed test to linear regression results. We also reduce a generalized linear regression problem to the comparison of two groups, which can handle non-normal and survival time phenotypes.

Keywords

GWAS; quantitative traits; allelic methods

Introduction

Over the past few years Genome Wide Association Studies (GWAS) have been extremely successful in discovering reproducibly associated genetic variants that affect disease susceptibility and other complex traits [Visscher et al., 2012]. However, the biological mechanisms behind these discoveries are still lacking for a large number of variants. To address this problem, an enormous amount of data on intermediate phenotypes such as gene expression, metabolite, and protein abundance are being mapped to genetic variants. Huge numbers of tests are computed in order to sift through these relationships and understand the biological machinery that determines the observed complex traits. Thus, fast and robust methods to compute these tests are highly desired.

Correspondence to : Hae Kyung Im, Department of Health Studies, University of Chicago, 5841 S. Maryland Ave., Chicago, IL 60637, USA, haky@uchicago.edu.

The authors declare no conflict of interests.

Here we propose an allele-based method for association studies that maintains estimation accuracy while reducing computational costs. Our approach is motivated by an allele-based test commonly used in case-control studies. For binary outcomes, it was shown that the allelic test based on allele frequencies is asymptotically equivalent to the genotype-based trend test under Hardy-Weinberg Equilibrium [Sasieni, 1997; Guedj et al., 2006; Guedj et al., 2008; Zheng, 2008]. Similarly, we develop a method based on the decomposition of genotypes into alleles for evaluating the association between quantitative traits and genetic variants.

In this article, we explore properties of estimators using our new method and compare them to the genotype-based method, i.e. linear regression estimators. Compared with Matrix eQTL [Shabalin, 2012], our method can be implemented with the same level of computational complexity for normally distributed traits. The main advantage of our method is that it can also be applied to non-normal and survival time phenotypes.

Method

The allelic test is an approach to evaluate the association between disease status and genotype in case-control studies. Table 1 shows a typical genotypic contingency table and the corresponding allelic contingency table. In the genotypic contingency table, each individual is a statistical unit and the number of individuals in each combination of disease status and genotype (aa, aA, AA) is tabulated. In the allelic contingency table, each allele (every individual has two alleles, one coming from the mother and one from the father) is a statistical unit and the number of alleles in each combination of disease status and allele (a, A) is tabulated.

We extend the idea behind the allelic test to quantitative traits and show that the regression problem in the genotypic formulation can be reduced to the comparison of two groups. Let us illustrate the method with a simple example with three individuals. These individuals have phenotypic values y_1, y_2 , and y_3 and genotypic values aa, aA, and AA, respectively. We transform the data such that each allele becomes the statistical unit instead of an individual. Thus, the first individual's data becomes two data points with phenotypic values y_1 and y_1 and allelic values a and a . We proceed similarly for the remaining two individuals as illustrated in Table 2. The association problem can be reduced to the comparison of two groups a and A with phenotypic values $\{y_1, y_1, y_2\}$ and $\{y_2, y_3, y_3\}$, respectively. As a result, this comparison can be readily performed using a two sample t-test or a non-parametric Wilcoxon rank sum test. Obviously, the main problem with this approach is that we would be ignoring the induced correlation in the data by the fact that we are, in some sense, "double counting" the observations. We will evaluate the consequences of ignoring this fact.

Comparison of the genotypic vs. allelic methods

Let us assume the following model for the phenotype genotype relationship:

$$Y = X\beta + \varepsilon \quad (1)$$

where X is a vector with genotypes taking values 0, 1, and 2 before centering and $-2MAF$, $1 - 2MAF$, and $2 - 2MAF$ after centering (MAF =minor allele frequency). Y is the phenotype of interest, which has also been centered at its mean. β is the effect of the genotype on phenotype. ε is the vector of error terms, assumed to be i.i.d and normally distributed with mean zero and variance σ^2 . We will refer to equation (1) as the genotypic model.

We now represent the genotypic model in (1) in a seemingly cumbersome way in terms of paternal and maternal alleles. (The method does not need to distinguish which one is which.)

$$\begin{pmatrix} Y \\ Y \\ \mathbb{Y} \end{pmatrix} = \begin{pmatrix} F \\ M \\ \mathbb{X} \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_F \\ \varepsilon_M \\ \mathbb{E} \end{pmatrix} \quad (2)$$

where $\mathbb{Y} = [Y' \ Y']'$, $\mathbb{X} = [F' \ M']'$ and $\mathbb{E} = [\varepsilon_F' \ \varepsilon_M']'$ with $\varepsilon_F = M\beta + \varepsilon$ and $\varepsilon_M = F\beta + \varepsilon$. F and M are vectors taking values $-MAF$ and $1 - MAF$ so that they are centered at zero.

The allelic model in (2) is equivalent to the genotypic model as long as we take into account the fact that the error terms have the following covariance matrix.

$$Var(\mathbb{E}) = \sigma^2 \begin{pmatrix} I & I \\ I & I \end{pmatrix}$$

where I is a $N \times N$ identity matrix and we have used the fact that $\varepsilon_M = F\beta + \varepsilon$ and $\varepsilon_F = M\beta + \varepsilon$. More specifically, model (2) is repeating model (1) twice and defining ε_F and ε_M so that the two models are exactly the same. Then we compute the variance of $\mathbb{E} = [\varepsilon_F' \ \varepsilon_M']'$ so that everything is consistent.

Let us consider the following estimator for β :

$$\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y} \quad (3)$$

This is actually the linear regression estimate of β for model (2) if we ignore the correlation structure in the data. Assuming Hardy-Weinberg equilibrium, we can show that the following properties hold for the estimator (2) under the true distribution of the data, i.e. having accounted for the true correlation structure.

$$E(\hat{\beta}) = \beta + O\left(\frac{1}{\sqrt{N}}\right) \quad (4)$$

$$Var(\hat{\beta}) = \sigma^2(X'X)^{-1} \left(1 + O\left(\frac{1}{\sqrt{N}}\right)\right) \quad (5)$$

(see Appendix for more details). Note that $Var(\hat{\beta})$ converges to zero as N goes to infinity since $X'X$ is $O(N)$. In other words, the estimator (3) has the same asymptotic properties as the linear regression estimator of model (1), $\tilde{\beta} = (X'X)^{-1} X'Y$.

We are interested in the properties of the estimator in (3) because when we reduce the linear regression problem to the comparison of two groups we will be ignoring the correlation structure in the data and this will be the estimator to be used. We are also using the fact that a t-test is equivalent to a univariate linear regression where the covariate has two levels.

Next, we evaluate the properties of $\hat{\beta}$ if we ignore the correlation structure. The expected value of $\hat{\beta}$ is not changed. We show that the estimated variance ignoring correlation is larger than the estimated variance accounting for correlation. This means that, even though the true mean and variance of the estimator is asymptotically equivalent to the linear regression estimator, the estimated variance is inflated when correlation is neglected. As a consequence, the estimated association strength will be conservative unless we adjust it using R^2 values. Since the effects of genetic variants on the phenotype rarely exceeds 10% in real GWAS data, this approximation should be more than sufficient in most cases. In cases where there is more inflation of the variance of $\hat{\beta}$, the effect is so large that the association would not be missed even with the slightly lower power. The inflation of the estimated variance when ignoring correlation is given by

$$\frac{Var(\hat{\beta}|ignoring\ correlation)}{Var(\hat{\beta}|accounting\ for\ correlation)} = \left(1 + \frac{R^2}{2(1 - R^2)}\right) + O_p\left(\frac{1}{\sqrt{N}}\right) \quad (6)$$

The properties of estimator (3) compared to the one from (1) are derived in the Appendix.

Extension to non-normal phenotypes

A linear regression model is typically used to assess the effects of genotype on different types of quantitative traits with the underlying assumption that the error term is normally distributed. In practice, this assumption may not always hold. The effects of violations of the normal assumption severely affect the power and type I error in finite samples [Feingold, 2002]. One strategy is to transform the data to achieve approximate normality using an inverse normal, logarithm, Box-Cox, or rank-based transformation method [Goh and Yap, 2009]. Alternatively, non parametric methods can be used but these tend to be more computationally expensive. For non-normal traits, Cuzick's nonparametric trend test [Cuzick, 1985] is available to calculate the association between genotypes and phenotype. With our allelic method, the comparison of two groups can be carried out with the Wilcoxon rank sum test. For survival outcomes subject to censoring, the Cox proportional hazards models is standard. Again, with our allelic approach the Cox regression can be replaced by a log-rank test. In the following section, through simulation studies and real data analyses, allele-based and genotype-based methods will be compared.

Results

Simulation Study

Normally distributed phenotypes—To assess the performance of the allelic and the genotypic method, we simulated phenotypes and genotypes for different values of minor allele frequency (MAF), total phenotypic variance explained by the genotype, and sample size. The phenotype was generated from the genotype using the model $Y = \beta X + \varepsilon$. X was simulated as the sum of two independent Bernoulli random variables with probability given by the MAF. The Hardy-Weinberg equilibrium assumption implies the independence of the maternal and paternal alleles. ε was generated as i.i.d normal variate with mean 0 and variance 1. All the results from the allelic method shown below are computed ignoring correlation. This is the more relevant comparison since we are interested in knowing the performance of the two group comparison approach assuming independence of the error terms.

In Figures 1 and 2 we show the comparison of our allelic method to linear regression (genotypic) with 1000 simulations using MAF = 5% and sample size 1000. Figure 1 shows results for $R^2 = 10\%$ and Figure 2 shows results for $R^2 = 1\%$. Corresponding β 's are 1.03 and 0.32, respectively. The figures show the $-\log_{10} p$ values, estimated regression coefficients, the R^2 , and the variance of the regression coefficient. The vertical axis shows the results using the quantitative allelic test (by allele) and the horizontal axis shows the results using regular linear regression (by genotype). The regression coefficient estimated using the allelic approach is very close to the one estimated using the genotypic approach, whereas the R^2 of the genotypic method is approximately twice that of the allelic model. This is expected since each allele should explain about half of what the sum of the two would. For relatively common SNPs the difference in performance depends mostly on the proportion of phenotypic variability explained by the genotype.

These results are consistent with our theoretical results, i.e. that the estimator in Eq.(3) is unbiased and with a slightly inflated variance. Nevertheless, even with a large effect size ($R^2 = 10\%$ rarely seen in GWAS studies) the regression coefficients, p values, and standard errors are quite similar.

Non-normal phenotypes—We next considered a non-normal phenotype and compared the allelic approach using Wilcoxon rank sum test and the genotypic approach implemented using Cuzick's trend test. We generated 1000 non-normal phenotypes based on a similar model as the previous simulation where the error term was generated using a t-distribution with two degrees of freedom. Figure 3 shows in $\log_{10}$ scale p values computed using the two methods. Thus, we find that also for the nonparametric test the allelic model provides similar results to the genotypic model, with slightly larger differences when the effect size is larger, as expected.

Survival phenotypes—We simulated survival times from a Cox regression model with a specified baseline hazard function. We considered two values for the log hazard ratios, $\beta = 0.8$ and 0.2 , representing strong and weak effects of genetic variation on survival. The baseline hazard function was assumed to have an exponential distribution with the hazard

rate of 0.005. Censoring times were generated from an exponential distribution with a parameter λ , which was determined by the censoring fraction. Here we set the censoring fraction as 20% and MAF as 5%. Figure 4 shows $\log_{10} p$ values for both methods. Similar to normal and non-normal phenotype simulations, the log-rank test and Cox model produce very comparable test results while the larger genetic effect ($\beta = 0.8$) shows more variability in the derived p values than the smaller effect ($\beta = 0.2$).

Application to GoKinD data

We applied both genotypic and allelic methods to GWAS data from the GoKinD (Genetics of Kidney Disease) study [Manolio et al., 2007; Pluzhnikov et al., 2010]. The dataset consisted of more than 1800 probands with long-standing type I diabetes. Over 300 dichotomous and quantitative phenotypes were genotyped on the Affymetrix Genome-Wide Human SNP Array 5.0 platform. Using Plink, we excluded SNPs with Hardy-Weinberg equilibrium $p < 0.001$, a minor allele frequency of less than 0.05 or genotyping call-rates of less than 0.97. We analyzed a subset of 1644 individuals reported to be Caucasian. We present results for body mass index (BMI) phenotype. A random subset of 25,000 SNP results are shown. Figure 5 shows that the genotypic and allelic method resulted in almost identical values in $-\log_{10} p$, estimated regression coefficients and variances. The R^2 of the genotypic method is roughly twice that of the allelic method. The quantile-quantile (QQ) plot (Figure 6) reveals no SNP was significantly associated with BMI in both approaches.

Computational efficiency

Although t-tests are computationally cheaper than running linear regressions, correlation statistics can be implemented in a very efficient manner as shown in the R package, matrix eQTL [Shabalin, 2012]. The t statistic can be implemented with similar computational cost to the correlation statistics (see Appendix). The main advantage of our approach is that the method can be used for phenotypes that need not be normally distributed as well as for survival phenotypes with censored data. By reducing the problem into a comparison of two groups we can use the non parametric Wilcoxon test for non normal variables and log-rank test for survival outcomes. The gain in computational efficiency is most notable for survival outcomes, due to the fact that the Cox procedure is an iterative method whereas the log-rank test is not.

Discussion

In this paper we proposed a fast and robust approach to assess the significance of association between genotype and quantitative traits. The method is an extension of the Allelic Test, which uses alleles as statistical units as opposed to a linear regression approach, where each individual is a statistical unit.

We showed that the allelic approach and genotypic approach yield estimates of β that are asymptotically equivalent when the correlation in the data is properly accounted for. However, when we ignore the correlation, the estimator is still unbiased but the estimated variance of the estimator is slightly inflated. When we reduce the problem to the comparison

of two groups we are, in fact, ignoring the correlation, and the estimated association strengths become conservative. We provide an asymptotic estimate of the inflation.

Our simulations results and application to GoKinD data are consistent with our theoretical findings. There is a small loss of power when the effect sizes are large, but of course, in these cases, we can afford to be conservative since the significance will be large regardless.

We have focused on the association test between quantitative traits and genotypes without consideration of other factors, such as confounding or environmental variables, which may influence phenotypes. However, it is possible to adjust for them prior to the genome wide analysis. It is not uncommon to preprocess the data by regressing the outcome on relevant covariates and then performing GWAS on the residuals without covariate adjustment. The allelic version of the test would be run similarly on the residuals, hence resulting in computational speed-up. Note that the preprocessing step is performed only once. This workaround may not perform as well when adjusting for population structure in this manner (i.e., first regressing the outcome on principal components). However, our method is still relevant since, in many instances, GWAS is performed on relatively homogeneous populations and cross population results are obtained by meta analysis. The allelic test would improve the computational efficiency of the within-population analyses. Thus, our method should be useful in many instances. Since the p values from the genotypic approach and allelic approach do not differ in any substantive way, we expect the genomic control λ values to be very close.

We recognize that when performing regression for a handful of phenotypes computational time is not an issue. The need to increase computational speed is more pressing in the context of large eQTL mapping studies where tens of thousands of regressions need to be performed. In comparison with matrix eQTL [Shabalin, 2012], our method can be implemented with the same computational efficiency for normally distributed outcomes.

Furthermore, when permutation based p values need to be computed, computational speed becomes essential. Another context where our method is relevant is the case where the outcome is survival with censored data. The gain in speed is substantial when using a log-rank test instead of fitting a Cox regression model, since the former requires only one-step computation while the latter requires iteration to maximize the (partial) likelihood.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The Genetics of Kidneys in Diabetes (GoKinD) Study was conducted by the GoKinD investigators and supported by the Juvenile Diabetes Research Foundation, the CDC, and the Special Statutory Funding Program for Type 1 Diabetes Research administered by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). This manuscript was not prepared in collaboration with investigators of the GoKinD study and does not necessarily reflect the opinions or views of the GoKinD study or the NIDDK.

References

- Cuzick J. A Wilcoxon-type test for trend. *Statistics in Medicine*. 1985; 4:87–90. [PubMed: 3992076]
- Feingold E. Regression-based quantitative-traitlocus mapping in the 21st century. *The American Journal of Human Genetics*. 2002; 71:217–222.
- Goh L, Yap VB. Effects of normalization on quantitative traits in association test. *BMC Bioinformatics*. 2009; 10:415. [PubMed: 20003414]
- Guedj M, Nuel G, Prum B. A note on allelic tests in case-control association studies. *Annals of Human Genetics*. 2008; 72:407–409. [PubMed: 18355390]
- Guedj M, Wojcik J, Della-Chiesa E, Nuel G, Forner K. A fast, unbiased and exact allelic test for case-control association Studies. *Human Heredity*. 2006; 61:210–221. [PubMed: 16877868]
- Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*. 2007; 39:1181–1186. [PubMed: 17898773]
- Manolio TA, Rodriguez LL, Brooks L, Abecasis G. Group, T. G. C. R. New models of collaboration in genome-wide association studies: the genetic association information network. *Nature Genetics*. 2007; 39:1045–1051. [PubMed: 17728769]
- Pluzhnikov A, Below J, Konkashbaev A, Tikhomirov A, Kistner-Griffin E, Roe C, Nicolae D, Cox N. Spoiling the whole bunch: quality control aimed at preserving the integrity of high-throughput genotyping. *The American Journal of Human Genetics*. 2010; 87:123–128.
- Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics*. 1997; 53:1253–1261. [PubMed: 9423247]
- Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 2012; 28:1353–1358. [PubMed: 22492648]
- Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *American Journal of Human Genetics*. 2012; 90:7–24. [PubMed: 22243964]
- Zheng G. Can the allelic test be retired from analysis of case-control association studies? *Annals of Human Genetics*. 2008; 72:848–851. [PubMed: 18652599]

Figure 1. Comparison of results using allelic vs genotypic methods

This figure shows the results of 1000 simulations with genotype explaining 10% of the total variability ($R^2 = 10\%$). The top left panel shows $-\log_{10} p$ values, the bottom left panel shows the estimated regression coefficients (effect size), the top right panel shows the computed R^2 , and the bottom right panel shows the variance of the regression coefficient. The vertical axis represents the results for the allelic method and the horizontal axis represents the results from the genotype based method.

Figure 2. Comparison of simulation results using allelic vs genotypic methods

This figure shows the results of 1000 simulations with the genetic marker explaining 1% of the total variability ($R^2 = 1\%$). The top left panel shows $-\log_{10} p$ values, the bottom left panel shows the estimated regression coefficients (effect size), the top right panel shows the computed R^2 , and the bottom right panel shows the variance of the regression coefficient. The vertical axis represents the results for the allelic method and the horizontal axis represents the results from the genotype based method.

Figure 3. Comparison of simulation results using Wilcoxon rank sum test vs. Cuzick's trend test
 The non-normal phenotypes were generated using a t-distribution with two degrees of freedom. The left panel shows $-\log_{10} p$ values from 1000 simulations with the genetic marker explaining 10% of the total variability ($R^2 = 10\%$) and the right panel shows the results of simulations with the genetic marker explaining 1% of the total variability ($R^2 = 1\%$)

Figure 4. Comparison of simulation results using log-rank test vs. Cox regression model

The left panel shows $-\log_{10} p$ values from 1000 simulations for survival time phenotypes with log hazard ratio 0.8 ($\beta = 0.8$) and the right panel shows the results of simulations with log hazard ratio 0.2 ($\beta = 0.2$).

Figure 5. Results of allelic and genotypic approaches for the GoKinD study

The top left panel shows $-\log_{10} p$ values, the bottom left panel shows the estimated regression coefficients (effect size), the top right panel shows the computed R^2 , and the bottom right panel shows the variance of the regression coefficient. The vertical axis represents the results for the allelic method and the horizontal axis represents the results from the genotype based method.

Figure 6. Quantile-Quantile plots for BMI association

The left panel shows the quantile-quantile plots from genotypic method while the right panel shows allelic method.

Table 1
Genotypic and allelic contingency tables

Lower case a represents the minor allele and capital A represents the major allele. The genotype contingency table shows the number of individuals with a given combination of disease status and genotype (aa,aA,AA). The allelic contingency table shows the number of alleles (a, A) with a given combination of disease status and allele type (a,A).

Genotypic contingency table				
	aa	aA	AA	Total
Cases	D_0	D_1	D_2	N_D
Controls	H_0	H_1	H_2	N_H
Total	N_0	N_1	N_2	N

Allelic contingency table			
	a	A	Total
Cases	$2D_0 + D_1$	$2D_2 + D_1$	$2N_D$
Controls	$2H_0 + H_1$	$2H_2 + H_1$	$2N_H$
Total	$2N_0 + N_1$	$2N_2 + N_1$	$2N$

Table 2
Genotypic and allelic data tables

Left side shows genotypic data and the right hand side shows the allelic data. The unit becomes the allele instead of the individual. The right hand side classifies the alleles into two group a and A

phenotype	genotype		phenotype	allele
y_1	aa		y_1	a
			y_1	a
y_2	aA	$\rightarrow$	y_2	a
			y_2	A
y_3	AA		y_3	A
			y_3	A