PLOS ONE

RESEARCH ARTICLE

Fast score test with global null estimation regardless of missing genotypes

Shuntaro Sato[1,2☯]*, **Masao Ueki**[3¶], **Alzheimer's Disease Neuroimaging Initiative**[¶]

1 Clinical Research Center, Nagasaki University Hospital, 1-7-1 Sakamoto, Nagasaki, Nagasaki 852-8501, Japan, **2** Biostatistics, Graduate School of Medicine, Kurume University, 67 Asahi-machi, Kurume, Fukuoka, 830-0011, Japan, **3** Statistical Genetics Team, RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan

☯ These authors contributed equally to this work.
¶ Membership of the Alzheimer's Disease Neuroimaging Initiative is provided in the Acknowledgments.
* shuntarosato@nagasaki-u.ac.jp

Citation: Sato S, Ueki M, Alzheimer's Disease Neuroimaging Initiative (2018) Fast score test with global null estimation regardless of missing genotypes. PLoS ONE 13(7): e0199692. https://doi.org/10.1371/journal.pone.0199692

Editor: Xiang Li, Janssen Research and Development, UNITED STATES

Received: August 3, 2017

Accepted: June 12, 2018

Published: July 5, 2018

Data Availability Statement: Relevant data and program code of simulations are within the paper and its Supporting Information files. Data files of "Application to ADNI GWAS Data" are freely and publicly available from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database: http://adni.loni.usc.edu/. Login page: http://adni.loni.usc.edu/data-samples/access-data/.

Funding: This work was supported by Japan Society for the promotion of science (http://www.jsps.go.jp/english/), grant numbers JP16K00064, JP16K08638, JP16H05242, and JP16H01528

Abstract

In genome-wide association studies (GWASs) for binary traits (or case-control samples) in the presence of covariates to be adjusted for, researchers often use a logistic regression model to test variants for disease association. Popular tests include Wald, likelihood ratio, and score tests. For likelihood ratio test and Wald test, maximum likelihood estimation (MLE), which requires iterative procedure, must be computed for each single nucleotide polymorphism (SNP). In contrast, the score test only requires MLE under the null model, being lower in computational cost than other tests. Usually, genotype data include missing genotypes because of assay failures. It loses computational efficiency in the conventional score test (CST), which requires null estimation by excluding individuals with missing genotype for each SNP. In this study, we propose two new score tests, called PM1 and PM2, that use a single global null estimator for all SNPs regardless of missing genotypes, thereby enabling faster computation than CST. We prove that PM2 and CST have an equivalent asymptotic power and that the power of PM1 is asymptotically lower than that of PM2. We evaluate the performance of the proposed methods in terms of type I error rates and power by simulation studies and application to real GWAS data provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI), confirming our theoretical results. ADNI-GWAS application demonstrated that the proposed score tests improve computational speed about 6–18 times faster than the existing tests, CST, Wald tests and likelihood ratio tests. Our score tests are general and applicable to other regression models.

Introduction

Over the last decades, genome-wide association studies (GWASs) have successfully identified many variants that are susceptible to hundreds of human diseases and traits [1, 2]. For discovery of an association between disease and genotypes, researchers often use tests based on a logistic regression model. It can analyze an association between disease (binary trait) and each single nucleotide polymorphism (SNP) while adjusting for the effect of covariates including

(received author is M.U.). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; EliLilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This work was carried out under the ISM General Cooperative Research 1 (2015-ISM-CRP-1013). 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.

age, sex, body mass index, and/or principal components for population stratification [3]. Wald test, likelihood ratio test, and score test are popularly used to examine the effects of each SNP on an outcome and are applied to genome-wide scan. For example, PLINK (http://zzz.bwh.harvard.edu/plink/) [4] and PLINK 1.9 (https://www.cog-genomics.org/plink2/) [5] use the Wald test by default for genome-wide scan in the presence of covariates to be adjusted for. Recently, we need to test over 500,000 loci in SNP-GWAS or tens of millions of loci in the whole-genome sequencing studies. Inclusion of large number of covariates slows the computation further. We often carry out this genome-wide scan for multiple traits. Consequently, computationally efficient method for genome-wide scan for large number of variants is highly desired [6].

In a logistic regression model, an iterative procedure such as Newton–Raphson method is needed to compute maximum likelihood estimator (MLE), which incurs computational burden in application to genome-wide scan. For the Wald test and the likelihood ratio test, MLE under full model for each SNP is required. On the other hand, the score test only requires MLE under null model. Furthermore, since the null model is common for all SNPs in testing association of SNPs (i.e. no SNPs have effect on outcome), if no SNPs have missing genotypes, a single null estimation can be used in score test statistics for all SNPs and computationally demanding iterative optimization process in computing MLE for each SNP is unnecessary. However, genotype data usually include missing genotypes because of assay failures [7]. Then, we still face computational burden even in the score test because missing pattern differs across loci and null estimation by excluding individuals with missing separately for each SNP is necessary. For example, the qtscore function in GenABEL package [8] implements fast genome-wide scan by the score test, where individuals with missing genotpyes are removed for each SNP [9, 10]. The Wald test implemented in PLINK also uses the complete case analysis.

In this study, we propose two fast score tests, called the proposed method 1 (PM1) and the proposed method 2 (PM2), that require only a single global null estimator for all SNPs regardless of missing genotypes unlike the conventional score test (CST) which requires separate null estimations for all SNPs. Fig 1 illustrates our idea. We prove that PM2 and CST have an equivalent asymptotic power and that the power of PM1 is asymptotically lower than that of PM2. We show through simulation studies that our PM1 and PM2 give correct control of the type I error. The simulations also confirm our theoretical results for an equivalent power between PM2 and CST, and the lower power of PM1 although the loss of power seems to be small in a range of practical missing genotype rates (<10%) in current GWAS. Application to real GWAS data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) demonstrates 6–18 times faster computation of the proposed methods than the CST, Wald test, and likelihood ratio test.

Materials and methods

Logistic regression model

We consider a case-control study with total sample size n. For individual i, let $Y_i = 1$ or $Y_i = 0$ be an indicator of disease (case or control), respectively. The probability of being case is $\pi_i = Pr(Y_i = 1)$. Our logistic regression model for a SNP is written as

$$\text{logit}[Pr(Y_i = 1)] = \text{logit}[\pi_i(\beta_0, \beta_e, \beta_g)] = \beta_0 + \beta_e E_i + \beta_g G_i \quad (1)$$

where G_i is some genotype coding, such as an additive coding {0, 1, 2} and E_i is a covariate or an environment factor. Letting $\theta_1 = (\beta_0, \beta_e)^{\mathrm{T}}$, $\theta_2 = \beta_g$, $\theta = (\theta_1^{\mathrm{T}}, \theta_2^{\mathrm{T}})^{\mathrm{T}}$, $X_1 = (1, E)$, and $X_2 = G$, the

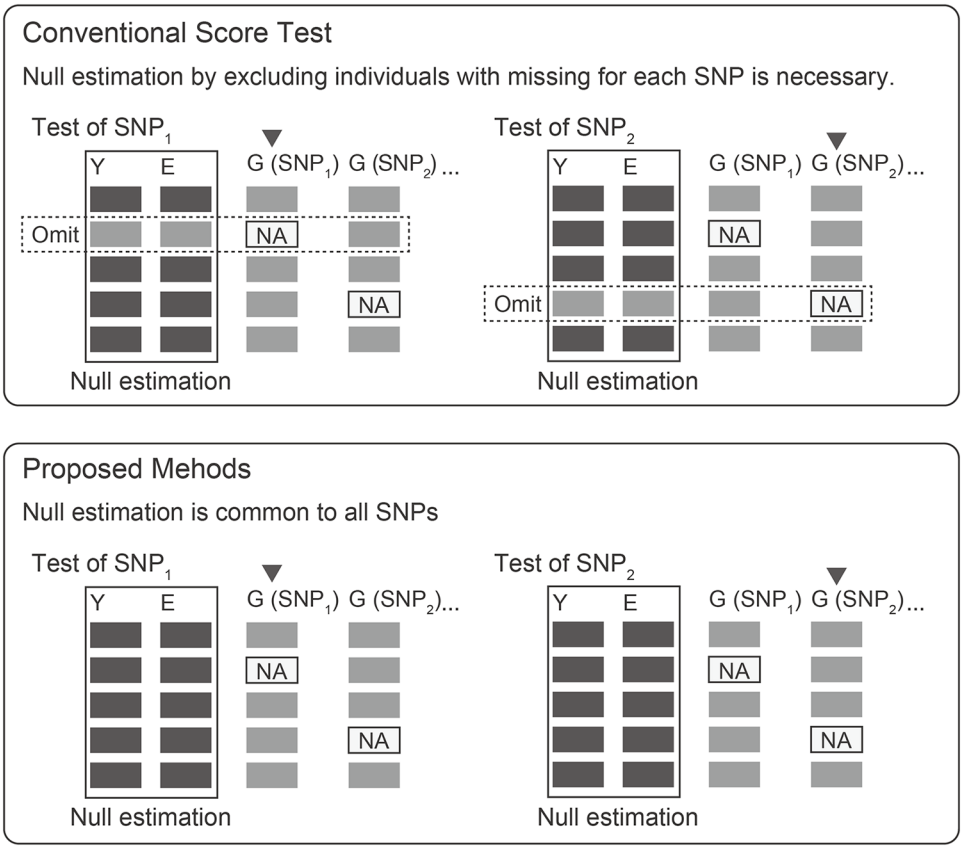

Fig 1. Conceptual difference between the conventional score test and the proposed new score tests. Conceptual difference between the conventional score test (CST) and the proposed new score tests (PM1 and PM2). Outcomes (Y) and covariates (E) are observed in all individuals. NA indicates missing genotype. In CST, null estimation is performed for each SNP excluding individuals with missing genotype. On the other hand, null estimation required in PM1 or PM2 is common in all SNP.

https://doi.org/10.1371/journal.pone.0199692.g001

probability of being case at a full model is

$$\pi_i(\theta) = \pi_i(\theta_1, \theta_2) = \frac{\exp(\beta_0 + \beta_e E_i + \beta_g G_i)}{1 + \exp(\beta_0 + \beta_e E_i + \beta_g G_i)} = \frac{\exp(X_1\theta_1 + X_2\theta_2)}{1 + \exp(X_1\theta_1 + X_2\theta_2)}.$$

Under these setting, the log-likelihood function at the full model is

$$\log f(\theta) = \sum_{i=1}^{n} [Y_i \log \pi_i(\theta) + (1 - Y_i) \log \{1 - \pi_i(\theta)\}].$$

For logistic regression model, no closed-form solution is available for MLE for $(\theta_1^{\mathrm{T}}, \theta_2^{\mathrm{T}})^{\mathrm{T}}$, and iterative procedure such as Newton–Raphson method is required, which causes high computational load in applying to genome-wide scan.

The score function evaluated at the full model is

$$\begin{pmatrix} u_1(\theta) \\ u_2(\theta) \end{pmatrix} = \begin{pmatrix} \partial \log f(\theta)/\partial\theta_1 \\ \partial \log f(\theta)/\partial\theta_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} u_{1i}(\theta) \\ \sum_{i=1}^{n} u_{2i}(\theta) \end{pmatrix}$$

and their covariance matrix at the full model is

$$J_{11}(\theta) = \begin{pmatrix} J_{11}[1,1](\theta) & J_{11}[1,2](\theta) \\ J_{11}[2,1](\theta) & J_{11}[2,2](\theta) \end{pmatrix} = -\frac{1}{n}\begin{pmatrix} \partial u_1(\theta)/\partial\theta_1 & \partial u_1(\theta)/\partial\theta_2 \\ \partial u_2(\theta)/\partial\theta_1 & \partial u_2(\theta)/\partial\theta_2 \end{pmatrix}.$$

Here, the score function and their covariance matrix at a null model ($\theta = (\theta_1^{\mathrm{T}}, 0^{\mathrm{T}})^{\mathrm{T}}$) are

$$\begin{aligned} \begin{pmatrix} u_1(\theta_1) \\ u_2(\theta_1) \end{pmatrix} &= \begin{pmatrix} u_1(\theta) \\ u_2(\theta) \end{pmatrix}\Big|_{\theta=(\theta_1^T,0^T)^T} \\ &= \begin{pmatrix} \partial \log f(\theta)/\partial\theta_1 \\ \partial \log f(\theta)/\partial\theta_2 \end{pmatrix}\Big|_{\theta=(\theta_1^T,0^T)^T} = \begin{pmatrix} \sum_{i=1}^n u_{1i}(\theta) \\ \sum_{i=1}^n u_{2i}(\theta) \end{pmatrix}\Big|_{\theta=(\theta_1^T,0^T)^T} \end{aligned}$$

and

$$\begin{aligned} \begin{pmatrix} J_{11}[1,1](\theta_1) & J_{11}[1,2](\theta_1) \\ J_{11}[2,1](\theta_1) & J_{11}[2,2](\theta_1) \end{pmatrix} &= \begin{pmatrix} J_{11}[1,1](\theta) & J_{11}[1,2](\theta) \\ J_{11}[2,1](\theta) & J_{11}[2,2](\theta) \end{pmatrix}\Big|_{\theta=(\theta_1^T,0^T)^T} \\ &= -\frac{1}{n}\begin{pmatrix} \partial u_1(\theta)/\partial\theta_1 & \partial u_1(\theta)/\partial\theta_2 \\ \partial u_2(\theta)/\partial\theta_1 & \partial u_2(\theta)/\partial\theta_2 \end{pmatrix}\Big|_{\theta=(\theta_1^T,0^T)^T}. \end{aligned}$$

Wald, likelihood ratio, and score tests

In GWAS, the null hypothesis H_0: $\theta_2 = 0$ in the logistic regression model (1) for each SNP is tested using the Wald test, the likelihood ratio test, or the score test. The Wald statistic is

$$W = \hat{\theta}_2^{\mathrm{T}} var(\hat{\theta}_2)^{-1}\hat{\theta}_2,$$

where $\hat{\theta}_2$ is MLE for θ_2 under the full model and $var(\hat{\theta}_2)$ denotes an estimator of (asymptotic) variance of $\hat{\theta}_2$. Wald tests need single optimization for full model MLE for each SNP. The likelihood ratio statistic is

$$LR = -2\{\log f(\check{\theta}_1, 0) - \log f(\hat{\theta}_1, \hat{\theta}_2)\},$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are MLE for θ_1 and θ_2, respectively, under the full model and $\check{\theta}_1$ is MLE for θ_1 under the null model. Likelihood ratio tests need two optimizations for the null model MLE and the full model MLE for each SNP. The score statistic is

$$S = u_2(\check{\theta}_1)^{\mathrm{T}} var\{u_2(\check{\theta}_1)\}^{-1} u_2(\check{\theta}_1),$$

where $var\{u_2(\check{\theta}_1)\}$ denotes an estimator of (asymptotic) variance of $u_2(\check{\theta}_1)$. Score tests require single optimization for the null model MLE for each SNP. If there is no missing genotypes for all SNPs, that is complete data, then null estimation is common for all SNPs. We focus score tests in this study, because parameter estimation which needs iterative optimization can be performed only once. Therefore, the score test with complete data can be computed with much lower computational cost than the Wald and the Likelihood ratio tests.

For a SNP, which we denote the individual i's genotype by G_i, we consider the setting where there are missing genotypes. Under this setting, score functions and their covariance matrix at

the null model are given as follows:

$$\begin{pmatrix} u_1^m(\theta_1) \\ u_2^m(\theta_1) \end{pmatrix} = \begin{pmatrix} u_1^m(\theta) \\ u_2^m(\theta) \end{pmatrix}\Big|_{\theta=(\theta_1^T,0^T)^T} = \begin{pmatrix} \sum_{i=1}^n u_{1i}(\theta)I_i \\ \sum_{i=1}^n u_{2i}(\theta)I_i \end{pmatrix}\Big|_{\theta=(\theta_1^T,0^T)^T}$$

and

$$\begin{aligned} \begin{pmatrix} J_{11}^m[1,1](\theta_1) & J_{11}^m[1,2](\theta_1) \\ J_{11}^m[2,1](\theta_1) & J_{11}^m[2,2](\theta_1) \end{pmatrix} &= \begin{pmatrix} J_{11}^m[1,1](\theta) & J_{11}^m[1,2](\theta) \\ J_{11}^m[2,1](\theta) & J_{11}^m[2,2](\theta) \end{pmatrix}\Big|_{\theta=(\theta_1^T,0^T)^T} \\ &= -\frac{1}{n}\begin{pmatrix} \partial u_1^m(\theta)/\partial\theta_1 & \partial u_1^m(\theta)/\partial\theta_2 \\ \partial u_2^m(\theta)/\partial\theta_1 & \partial u_2^m(\theta)/\partial\theta_2 \end{pmatrix}\Big|_{\theta=(\theta_1^T,0^T)^T}. \end{aligned}$$

where I_i is an indicator defined by

$$I_i = \begin{cases} 1 & \text{if } G_i \text{ is observed} \\ 0 & \text{if } G_i \text{ is missing} \end{cases}$$

For theoretical studies in what follows, we assume missing completely at random (MCAR), that is, I_i independently and identically follows a binomial distribution of size 1, $I_i \sim \text{Bin}(1, 1 - R)$ for $i = 1, \cdots, n$, where R is a probability of random missing, and is independent of u_{1i} and u_{2i}.

Score tests

Conventional score test. The conventional score test (CST) means the score test using null estimator computed by removing individuals having missing genotype to be tested. The score statistic of CST is expressed by the following formula, and asymptotically follows a chi-squared distribution with 1 degree of freedom under the null hypothesis:

$$S_{\text{CST}} = \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n u_{2i}^m(\check{\theta}_1^m)\right\}^{\text{T}} V_m^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n u_{2i}^m(\check{\theta}_1^m)\right\}$$

where $\check{\theta}_1^m$ is MLE for θ_1 under the null model on CST and $V_m = \{-J_{11}^m[2,1](\check{\theta}_1^m)J_{11}^m[1,1](\check{\theta}_1^m)^{-1} J_{11}^m[1,2](\check{\theta}_1^m) + J_{11}^m[2,2](\check{\theta}_1^m)\}/n$. In Appendix (p. 8), it is shown that the convergence rate to chi-square distribution has order $o_p(1)$ as $n \to \infty$.

Proposed methods. We describe two new proposed score tests. The proposed method 1 (PM1) is the score test which uses a single null estimator for all score test statistics regardless of missing genotypes. See Fig 1 for the difference from CST. The score statistic of PM1 is expressed by the following formula, and asymptotically follows a chi-squared distribution with 1 degree of freedom under the null hypothesis:

$$S_{\text{PM1}} = \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n u_{2i}^m(\check{\theta}_1^f)\right\}^{\text{T}} V_f^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n u_{2i}^m(\check{\theta}_1^f)\right\}$$

where $\check{\theta}_1^f$ is MLE for θ_1 under the null model computed independently of genotype data using all individuals, and $V_f = \{-J_{11}^m[2,1](\check{\theta}_1^f)J_{11}[1,1](\check{\theta}_1^f)^{-1}J_{11}^m[1,2](\check{\theta}_1^f) + J_{11}^m[2,2](\check{\theta}_1^f)\}/n$. In Appendix (p. 10), it is shown that the convergence rate to chi-square distribution has order $o_p(1)$ as $n \to \infty$. The MLE $\check{\theta}_1^f$ is a single global null estimator used in all test statistics for

genome-wide scan and does not require re-computation for all SNPs unlike CST, and hence, PM1 achieves lower computational cost than CST. However, we showed that the power of PM1 is asymptotically lower than that of CST. The power of the score test statistic asymptotically increases as the non-centrality parameter increases. In Appendix (p.14–17), we have shown that the mean of CST score function is asymptotically equivalent to that of PM1 score function while the variance of PM1 score function is bigger than the variance of CST score function. That is, the magnitude of non-centrality parameter is dominated only by the magnitude of variance, and the non-centrality parameter of PM1 is smaller than that of CST. Therefore, the power of PM1 is smaller than that of CST.

To improve power, we developed a second test, called the proposed method 2 (PM2). Here, we define the following modified score function:

$$u^*_{2i}(\theta_1) = u^m_{2i}(\theta_1) - J^m_{11}[2,1](\theta_1)J^m_{11}[1,1](\theta_1)^{-1}u^m_{1i}(\theta_1). \quad (2)$$

PM2 uses the above modified score function (2). It is computed without excluding individuals with missing SNP from null model as in PM1. The score statistic of PM2 is expressed by the following formula, and asymptotically follows a chi-squared distribution with 1 degree of freedom under the null hypothesis:

$$S_{\mathrm{PM2}} = \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} u^*_{2i}(\check{\theta}^f_1)\right\}^{\mathrm{T}} V_m^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} u^*_{2i}(\check{\theta}^f_1)\right\}.$$

In Appendix (p.13), it is shown that the convergence rate to chi-square distribution has order $o_p(1)$ as $n \to \infty$. It is shown in Appendix (p.13) that score function $(1/\sqrt{n})\sum_{i=1}^{n} u^*_{2i}(\check{\theta}^f_1)$ is asymptotically equivalent to the score function $(1/\sqrt{n})\sum_{i=1}^{n} u^m_{2i}(\check{\theta}^m_1)$ of CST. Therefore, the power of PM2 is asymptotically equivalent to CST and higher than that of PM1. In PM2, null estimation is common for all SNPs as in PM1. Thus, PM2 can have lower computational costs than CST.

So far, we have considered the test of H_0: $\theta_2 = \beta_g = 0$ under the logistic regression model (1). This framework can be easily extended to other tests. For another application, we consider the following logistic regression model involving gene-environment interaction,

$$\mathrm{logit}[Pr(Y_i = 1)] = \beta_0 + \beta_e E_i + \beta_g G_i + \beta_{ge} G_i E_i. \quad (3)$$

Let $\theta_2 = (\beta_g, \beta_{ge})^{\mathrm{T}}$. We can perform a joint test for combined effect of genetic marginal and of gene-environment interaction [11]. This test constrains $\beta_g = 0$ and $\beta_{ge} = 0$ under the null hypothesis, i.e. the degrees of freedom is two. The joint test is more powerful than the test of interaction alone, which is beneficial, particularly for GWAS where marginal SNP effect is low, and it is applied to real data [12].

More details of this section including formulas, derivations, and additional descriptions are given in S1 Appendix. A program code of simulations are given in S2 Appendix and all data files of "Application to ADNI GWAS Data" are freely and publically available from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database: http://adni.loni.usc.edu/.

Results

Evaluation of proposed methods using simulated data

We performed computer simulations to evaluate the performance (type I error rates and power) of the various test statistics described above. We simulated datasets using R [13] based on two logistic regression models ((1) and (3)) assuming disease prevalence 1%. Case-control

data was generated accoring to the retrospective sampling as described in [14]. The test corresponding to (1) is called 'G test', and the test corresponding to (3) is called 'G-GE test'. We considered binary variables as a covariate for E, e.g. gender, whose population frequency is 50% and set the odds ratio as $OR_e = \exp(\beta_e) = 1.2$. Missing genotypes were generated assuming missing completely at random, in particular, individuals with missing genotype are randomly assigned with a given missing rate.

Type I error rates. We performed 1,000,000 simulation replicates under the null model to estimate type I error rates for a nominal significance threshold of $\alpha = 5 \times 10^{-5}$. We considered a range of missing rates (2%, 5%, 10%), the number of case or control (1,000, 5,000), and minor allele frequencies (MAF) (10%, 30%).

We provided the estimated type I error rates in various settings in Table 1. For 1,000,000 replication, the standard deviation of the estimated type I error rates is $\sqrt{(0.00005 \times 0.99995)/1,000,000} \simeq 0.71 \times 10^{-5}$ and the 95% confidence interval is $(3.6 \times 10^{-5}, 6.4 \times 10^{-5})$ for the nominal significance level of $\alpha = 5 \times 10^{-5}$. From this table, we can see that all of the type I error rates are consistently within the 95% confidence interval, which indicates that the type I error rates are well-controlled at the nominal level.

Next, we constructed quantile-quantile (Q-Q) plots of the distribution of several test settings calculated in the above conditions under the null hypotheses. Fig 2 shows Q-Q plots of G test and G-GE test for missing rate is 5% and MAF is 10% and 30%. We plotted the top 500 score statistics of the CST. Most of the points are distributed around the 45 degree line, which

Table 1. Type I error rates of the conventional score test and the proposed methods.

Test	Missing rate (%)	MAF (%)	#case/control	CST	PM1	PM2
G	2	10	1,000	5.0×10^{-5}	5.5×10^{-5}	5.0×10^{-5}
G	2	10	5,000	3.7×10^{-5}	3.6×10^{-5}	3.7×10^{-5}
G	2	30	1,000	5.6×10^{-5}	4.4×10^{-5}	5.6×10^{-5}
G	2	30	5,000	4.3×10^{-5}	4.6×10^{-5}	4.3×10^{-5}
G	5	10	1,000	4.8×10^{-5}	5.3×10^{-5}	4.8×10^{-5}
G	5	10	5,000	4.2×10^{-5}	3.7×10^{-5}	4.2×10^{-5}
G	5	30	1,000	5.5×10^{-5}	5.8×10^{-5}	5.5×10^{-5}
G	5	30	5,000	4.0×10^{-5}	4.6×10^{-5}	4.0×10^{-5}
G	10	10	1,000	4.6×10^{-5}	4.0×10^{-5}	4.6×10^{-5}
G	10	10	5,000	3.8×10^{-5}	3.6×10^{-5}	3.8×10^{-5}
G	10	30	1,000	5.5×10^{-5}	5.2×10^{-5}	5.5×10^{-5}
G	10	30	5,000	4.9×10^{-5}	4.5×10^{-5}	4.9×10^{-5}
G-GE	2	10	1,000	4.7×10^{-5}	4.8×10^{-5}	4.7×10^{-5}
G-GE	2	10	5,000	4.4×10^{-5}	4.2×10^{-5}	4.4×10^{-5}
G-GE	2	30	1,000	5.7×10^{-5}	6.0×10^{-5}	5.6×10^{-5}
G-GE	2	30	5,000	5.1×10^{-5}	5.0×10^{-5}	5.1×10^{-5}
G-GE	5	10	1,000	4.8×10^{-5}	4.9×10^{-5}	4.8×10^{-5}
G-GE	5	10	5,000	4.9×10^{-5}	4.3×10^{-5}	4.9×10^{-5}
G-GE	5	30	1,000	5.3×10^{-5}	5.4×10^{-5}	5.2×10^{-5}
G-GE	5	30	5,000	5.8×10^{-5}	5.7×10^{-5}	5.8×10^{-5}
G-GE	10	10	1,000	4.6×10^{-5}	4.1×10^{-5}	4.6×10^{-5}
G-GE	10	10	5,000	4.6×10^{-5}	4.1×10^{-5}	4.6×10^{-5}
G-GE	10	30	1,000	5.2×10^{-5}	5.3×10^{-5}	5.0×10^{-5}
G-GE	10	30	5,000	5.8×10^{-5}	4.9×10^{-5}	5.8×10^{-5}

Type I error rates of the conventional score test (CST), the proposed method 1 (PM1), and the proposed method 2 (PM2) at a significance level of $\alpha = 5 \times 10^{-5}$.

https://doi.org/10.1371/journal.pone.0199692.t001

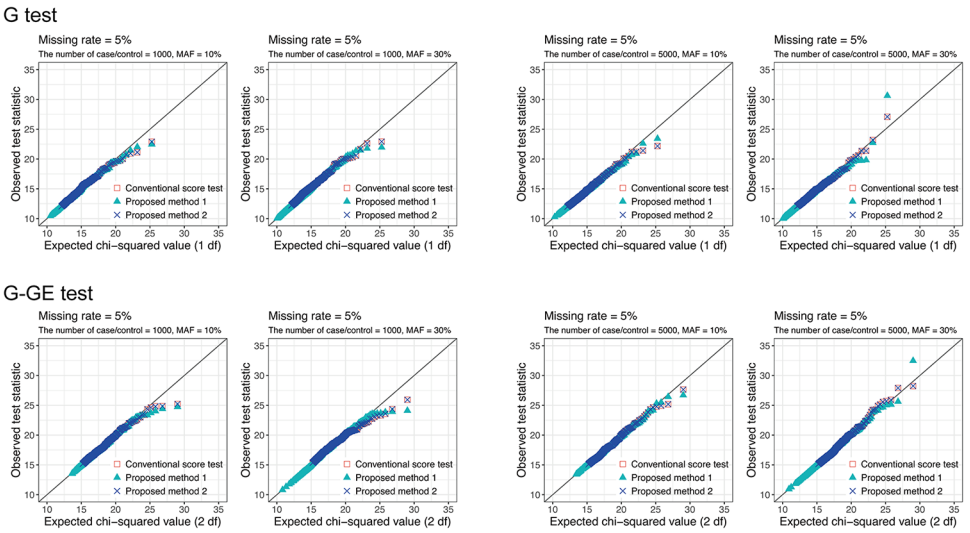

Fig 2. Q-Q plot of the conventional score test and the proposed methods. Chi-squared (1-df or 2-df) Q-Q plot of the top 500 conventional score test, the proposed method 1, and the proposed method 2 score statistics for missing rate is 5% and minor allele frequency (MAF) is 10% and 30% in null simulation.

https://doi.org/10.1371/journal.pone.0199692.g002

implies that χ^2 approximations to the three score statistics are valid. Q-Q plots in other settings are given in S1 and S2 Figs.

Furthermore, we investigated the scenarios with smaller sample size (the number of case or control: 100, 500) and unbalance sample size (the number of case: 1,000 and control: 2,000) in S1 Table. In the scenarios with smaller sample sizes, especially 100, the type I error rates are slightly lower than the nominal level, but the accuracy is improved as sample sizes get large. On the other hand, the type I error with unbalance scenarios are well-controlled at the nominal level.

Power. We performed 1,000 simulation replicates under alternative models to estimate power at $\alpha = 5 \times 10^{-8}$. We considered a range of missing rate (2%, 5%, 10%, 30%), the number of case or control (100, 500, 1,000, 5,000), unbalance sample size (case/control = 1,000/2,000), and MAF (10%, 30%). Although the missing rate of 30% would be unrealistic in practical human GWAS data, it was set to make the difference in power easy to see for confirmation of our theoretical asymptotic results on power. For reference, we also included the method which simply imputes the missing genotypes by their median (called the median imputation).

First, we showed the transition of the power of G tests as the change of OR_g at MAF of 30% in Fig 3. From Fig 3, we can see that score tests of CST and PM2 are more powerful than PM1. Next, we showed the power transition of the G-GE test as the change of OR_{ge} at genetic odds ratios (OR_g = 1.1, 1.2), missing rate (2%, 5%, 10%), the number of case or control (1,000, 5,000), and MAF (30%) in Figs 4 and 5. Similar to G test, G-GE test also has higher power for CST and PM2 than PM1. In G test and G-GE test, the power of median imputation has slightly lower than CST and PM2, and higher than PM1. Even with a small genetic main effect, the joint test can detect the effect of gene-environment interaction [11]. Analogous results were obtained under other settings (see S3–S5 Figs and S2 and S3 Tables). Collectively, the theoretical results shown in the Materials and Methods section can be confirmed by the simulation studies.

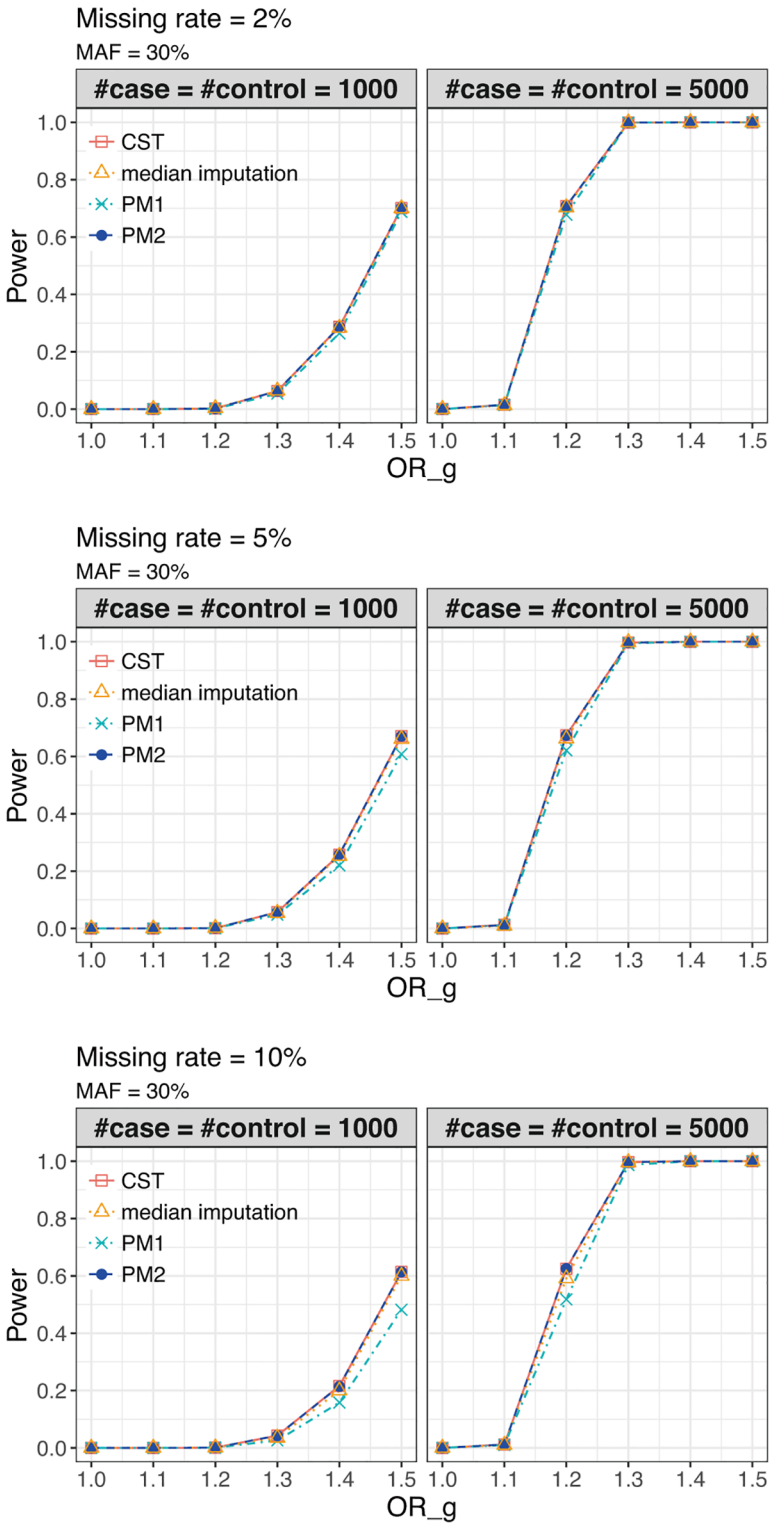

Fig 3. G test Power of the conventional score test and the proposed methods at MAF 30%. G test Power of the conventional score test (CST), the proposed method 1 (PM1), and the proposed method 2 (PM2) under missing rate (2%, 5%, 10%), minor allele frequency (MAF) (30%), and the number of case/control (1,000, 5,000). The x-axis denotes genetic odds ratios ($OR_g = \exp(\beta_g)$). The significance level is 5×10^{-8}.

https://doi.org/10.1371/journal.pone.0199692.g003

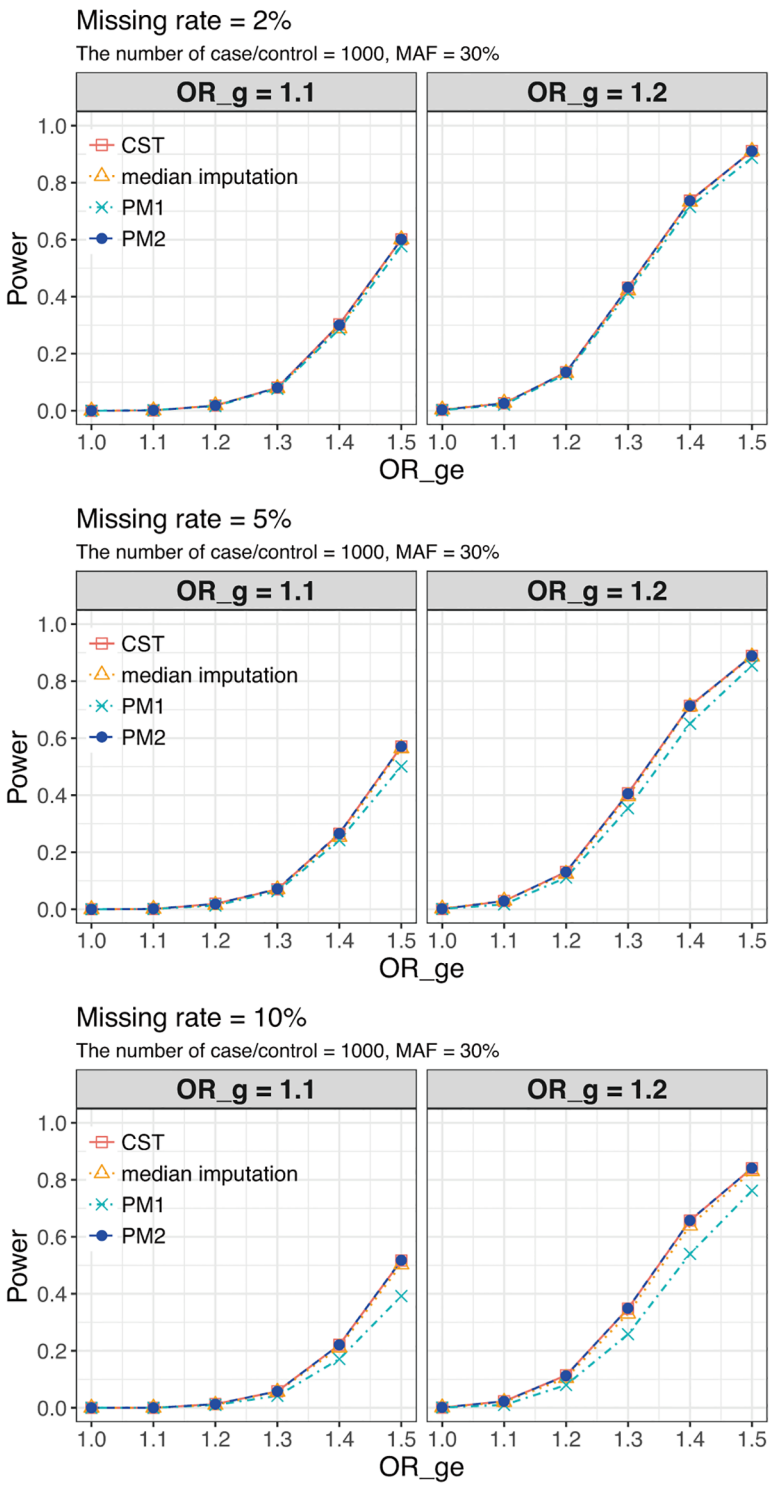

Fig 4. G-GE test Power of the conventional score test and the proposed methods at the number of case/control is 1,000. G-GE test Power of the conventional score test (CST), the proposed method 1 (PM1), and the proposed method 2 (PM2) under genetic odds ratios ($OR_g = \exp(\beta_g) = 1.1, 1.2$), missing rate (2%, 5%, 10%), minor allele frequency (MAF) (30%), and the number of case/control is 1,000. The x-axis denotes gene-environment interaction odds ratios ($OR_{ge} = \exp(\beta_{ge})$). The significance level is 5×10^{-8}.

https://doi.org/10.1371/journal.pone.0199692.g004

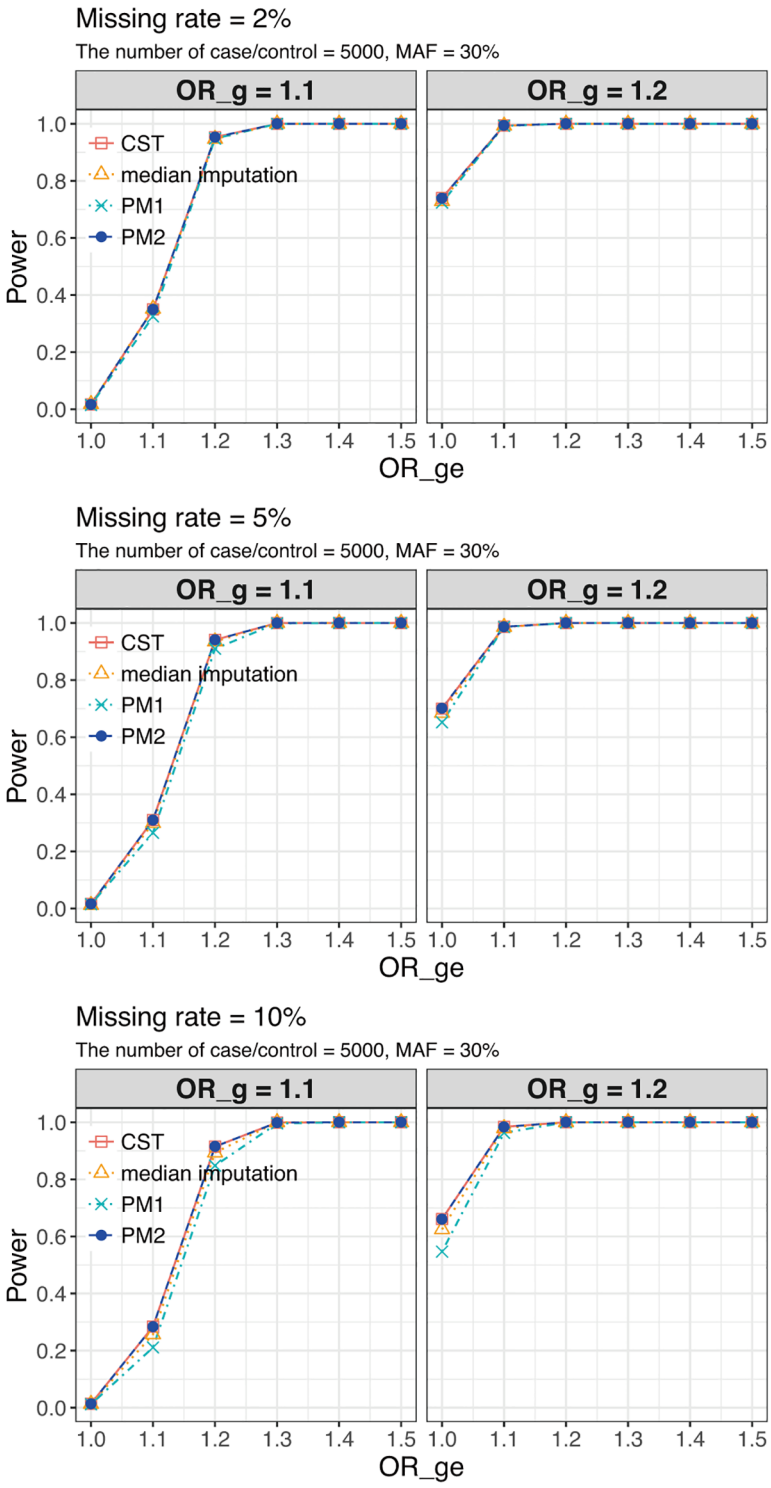

Fig 5. G-GE test Power of the conventional score test and the proposed methods at the number of case/control is 5,000. G-GE test Power of the conventional score test (CST), the proposed method 1 (PM1), and the proposed method 2 (PM2) under genetic odds ratios ($OR_g = \exp(\beta_g) = 1.1, 1.2$), missing rate (2%, 5%, 10%), minor allele frequency (MAF) (30%), and the number of case/control is 5,000. The x-axis denotes gene-environment interaction odds ratios ($OR_{ge} = \exp(\beta_{ge})$). The significance level is 5×10^{-8}.

https://doi.org/10.1371/journal.pone.0199692.g005

Application to ADNI GWAS data

We applied our proposed methods to ADNI-GWAS dataset obtained from the publicly available data of the Alzheimer's Disease Neuroimage Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease. For up-to-date information, see www.adni-info.org. ADNI is an ongoing longitudinal study with the primary purpose of exploring the genetic and neuroimaging information associated with late-onset Alzheimer's disease. The study recruited elderly subjects consisting about 400 subjects with mild cognitive impairment (MCI), about 200 subjects with Alzheimer's disease (AD), and about 200 healthy controls (normal). Each subject was followed for at least 3 years. During the study period, the subjects were assessed with MRI measures and psychiatric evaluation to determine the diagnosis status at each time point.

The ADNI-GWAS data were obtained from 818 DNA samples of ADNI participants using the Illumina Human 610-Quad genotyping assay [15]. The data initially included 620,901 SNPs. We included the *apolipoprotein* E (APOE) SNP rs429358 on chromosome 19 known to affect AD in our analysis. We used data from 684 non-Hispanic Caucasian samples after we excluded one pair showing cryptic relatedness (revealed by the PLINK pairwise $\hat{\pi}$ statistic being greater than 0.125) [4], and we excluded subjects whose reported sex did not match the sex inferred from X-chromosome SNPs. The total number of remaining SNPs was 528,916, and the demographic variables include gender and age. The distribution of missing rate is shown in the S6 Fig. Of the 528,916 SNPs, 45% of them have missing genotypes. In our work, we used 684 subjects: the status at the baseline of normal, MCI, and AD were 192, 329, and 163, respectively. We defined the following phenotype as an outcome: normal (= 0), MCI (= 0), and AD (= 1) as binary traits. We also included the following covariates in a logistic regression model: gender and age. We compared the proposed methods separately for two subsets of SNPs stratified by missing rate, low missing SNPs (43.9%) with 0% $<$ Missing rate $<$ 1%, and high missing SNPs (11.3%) with Missing rate $\geq$ 1%.

Firstly, Fig 6 showed Manhattan plots for all SNPs using CST, PM1, PM2, the Wald test and the likelihood ratio test. All figures were similar in shape, and the APOE SNP was statistically significant in most tests (CST, PM1, and PM2 give P-values of 3.640×10^{-8} because of the absence of missing genotypes at the APOE SNP. The Wald and the likelihood ratio tests give P-values of 7.686×10^{-8} and 5.372×10^{-8}, respectively.). Manhattan plots for two SNPs stratified by missing rate are shown in the S7 and S8 Figs and these figures were also similar in shape.

Secondary, Fig 7 illustrated scatter plots for two SNPs stratified by missing rate comparing top 1,000 P-values of the proposed methods and CST. The Pearson's correlation coefficient in the low missing SNPs between PM1 and CST was 0.9974 (Fig 7A). On the other hand, the correlation between PM2 and CST was 1.0000 (Fig 7B), showing much higher concordance. Similarly, the Pearson's correlation coefficient in the high missing SNPs between PM1 and CST was 0.9837 (Fig 7C), and the correlation between PM2 and CST was 1.0000 (Fig 7D). The above results show that the equivalence between test statistics of PM2 and CST and the difference between PM1 and CST (or PM2) as described in the Material and Methods section.

Finally, we compared the run times from the proposed methods, the CST, the Wald test, and the likelihood ratio test on a personal computer (four CPU cores at 4.0 GHz Intel i7) using all SNPs. We implemented all tests in R without using built-in functions (e.g. glm function in

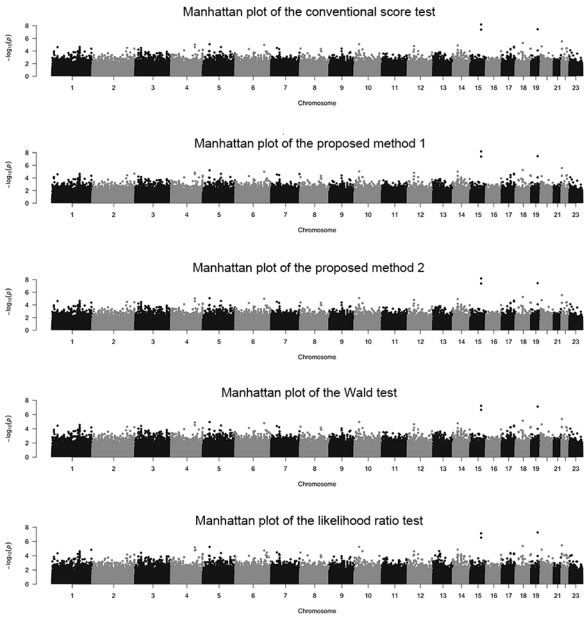

Fig 6. Manhattan plots of each chromosome for ADNI-GWAS dataset with all SNPs. Manhattan plots of each chromosome for ADNI-GWAS dataset. P-values using the conventional score test (CST), the proposed method 1 (PM1) test, the proposed method 2 (PM2) test, the Wald test, and the likelihood ratio test are shown in order from the top. The black and gray points highlight different chromosomes.

https://doi.org/10.1371/journal.pone.0199692.g006

R) for fair comparison of execution speed. Table 2 shows the run times. CST showed similar run times to the Wald test as both tests need a single iterative optimization for MLE for each SNP, i.e. MLE under the null model for CST and MLE under the full model for the Wald test. Likelihood ratio test requires two iterative optimizations for MLE under both null and full models, which make run times about twice longer compared with CST or the Wald test. PM1 and PM2 resulted in about 6–18 times faster than the CST, Wald test and likelihood ratio test. A slightly longer run time was observed for PM2 compared to PM1 because PM2 needs more matrix calculation processes than PM1. Based on these findings, we confirmed that the proposed methods have much lower computational cost than the CST, Wald test, and likelihood ratio test.

Discussion

In this paper, we presented two new fast score tests, PM1 and PM2, that require only a single global null estimator for all SNPs for genome-wide scan when missing genotypes are present. We confirmed that our proposed methods can significantly reduce the computational cost compared to conventional tests for genome-wide scans (e.g. Wald test in PLINK) in an application to ADNI-GWAS data. Run time of PM2 is slightly slower than PM1 because PM2 needs more matrix calculation processes. We theoretically proved that PM2 and CST have an equivalent asymptotic power and that the power of PM1 is lower than that of PM2. Additionally, we evaluated the power of CST, PM1, and PM2 by simulation studies and confirmed theoretical results. Therefore, when even higher power is required in studies, PM2 should be used rather than PM1, although PM2 is slightly slower than PM1 in computation. Our approach can speed up the computation by 6–18 times faster than CST, the Wald test, and the likelihood ratio test

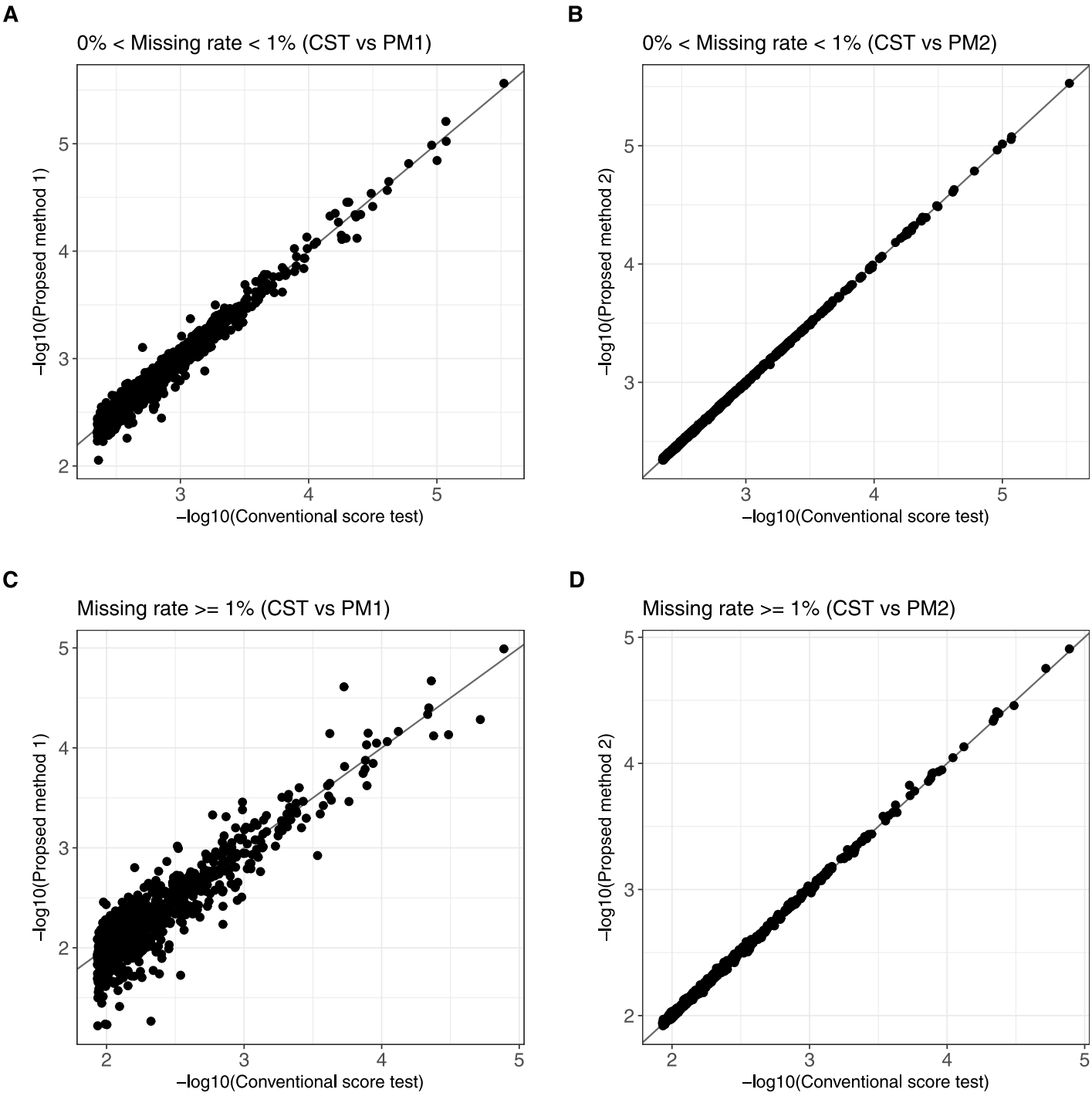

Fig 7. Comparisons of the proposed methods and the conventional score test P-values with the subset SNPs with missing genotypes. Comparisons of the proposed methods (PM1 and PM2) and the conventional score test (CST) P-values that displays only top 1,000. SNPs are stratified by missing rate (low missing SNPs: 0% < Missing rate < 1%, high missing SNPs: Missing rate ≥ 1%).

https://doi.org/10.1371/journal.pone.0199692.g007

for genome-wide scans. The CST, Wald, and likelihood ratio tests require re-computation of null MLE for each SNP because the pattern of missing genotypes differs across loci. Our test statistics only use a single global MLE under the null for all SNPs, which avoids re-computing null MLE for each SNP, and the speed-up is independent of the proportion of missing

Table 2. Run times (CPU sec) from the proposed methods and the conventional tests.

CST	PM1	PM2	Wald	LRT
452.9	49.2	74.6	460.6	874.5

Run times (CPU sec) from the proposed methods (PM1 and PM2), the conventional score test (CST), the Wald test, and the likelihood ratio test (LRT).

https://doi.org/10.1371/journal.pone.0199692.t002

genotypes. The more the number of covariates is, the more computational speed-up is pronounced. Our framework is more valuable for more complicated analyses which require enormous number of hypotheses to be tested such as gene-environment or/and gene-gene interaction analyses.

Missing genotypes may be imputed by the genetic imputation which is a method to predict the genotypes at the SNPs that are not directly assayed in a sample of individuals. It is achieved by using known haplotype reference panel, for example from the HapMap or the 1000 Genomes Project in humans [16, 17]. However, the accuracy of genotype imputation and boosting power of the subsequent association analyses depends on the quality of reference panel. Moreover, genetic imputation requires a lot of computational resources [18–20]. Even if imputation is applied, there usually remain uncertain genotypes that are hard to call, which are often set to missing. Therefore, missing genotype problem is unavoidable even after genetic imputation.

In this study we have assumed MCAR for the proposed methods, which is a reasonable assumption in the case where simply discarding the missing observations (i.e. complete case analysis in the CST) is not too problematic [21]. Although our proposed method worked in real GWAS data from ADNI, there may be a case where missing genotypes cannot be considered as MCAR [22]. Then, simply ignoring missing genotypes from analysis may lead to severe bias [21]. By the same reason, in this case, our theoretical results regarding type I error and power for the proposed methods may not hold. Further work remains to be done in this important topic.

In this paper, we focused only on the logistic regression model for binary traits. However, our framework is general and is extensible to other different score tests, e.g. in survival analysis.

Supporting information

S1 Appendix. Details of the method. More details of the materials and methods section including formulas, derivations, and additional descriptions.
(PDF)

S2 Appendix. Program code. A program code of simulations.
(PDF)

S1 Fig. Q-Q plot of the conventional score test and the proposed methods at missing rate 2%. Chi-squared (1-df or 2-df) Q-Q plot of the top 500 conventional score test, the proposed method 1, and the proposed method 2 score statistics for missing rate is 2% and minor allele frequency (MAF) is 10% and 30% in null simulation.
(EPS)

S2 Fig. Q-Q plot of the conventional score test and the proposed methods at missing rate 10%. Chi-squared (1-df or 2-df) Q-Q plot of the top 500 conventional score test, the proposed

method 1, and the proposed method 2 score statistics for missing rate is 10% and minor allele frequency (MAF) is 10% in null simulation.
(EPS)

S3 Fig. G test Power of the conventional score test and the proposed methods at MAF 10%. G test Power of the conventional score test (CST), the proposed method 1 (PM1), and the proposed method 2 (PM2) under missing rate (2%, 5%, 10%), minor allele frequency (MAF) (30%), and the number of case/control (1,000, 5,000). The x-axis denotes genetic odds ratios ($OR_g = \exp(\beta_g)$). The significance level is 5×10^{-8}.
(EPS)

S4 Fig. G-GE test Power of the conventional score test and the proposed methods at the number of case/control is 1,000 and MAF 10%. G-GE test Power of the conventional score test (CST), the proposed method 1 (PM1), and the proposed method 2 (PM2) under genetic odds ratios ($OR_g = \exp(\beta_g) = 1.1, 1.2$), missing rate (2%, 5%, 10%), minor allele frequency (MAF) (10%), and the number of case/control is 1,000. The x-axis denotes gene-environment interaction odds ratios ($OR_{ge} = \exp(\beta_{ge})$). The significance level is 5×10^{-8}.
(EPS)

S5 Fig. G-GE test Power of the conventional score test and the proposed methods at the number of case/control is 5,000 and MAF 10%. G-GE test Power of the conventional score test (CST), the proposed method 1 (PM1), and the proposed method 2 (PM2) under genetic odds ratios ($OR_g = \exp(\beta_g) = 1.1, 1.2$), missing rate (2%, 5%, 10%), minor allele frequency (MAF) (10%), and the number of case/control is 5,000. The x-axis denotes gene-environment interaction odds ratios ($OR_{ge} = \exp(\beta_{ge})$). The significance level is 5×10^{-8}.
(EPS)

S6 Fig. Missing rate distribution of ADNI. The y-axis denotes the number of SNPs. The x-axis denotes Missing rate.
(EPS)

S7 Fig. Manhattan plots of each chromosome for ADNI-GWAS dataset with the low missing SNPs. The y-axis denotes the number of SNPs. The x-axis denotes Missing rate. Low missing population include SNPs with missing ($0\% <$ Missing rate $< 1\%$).
(PNG)

S8 Fig. Manhattan plots of each chromosome for ADNI-GWAS dataset with the high missing SNPs. The y-axis denotes the number of SNPs. The x-axis denotes Missing rate. High missing population include SNPs with missing (Missing rate $\geq 1\%$).
(PNG)

S1 Table. Type I error rates of the conventional score test and the proposed methods at the scenarios with smaller sample sizes and unbalanced case-control samples. Type I error rates of the conventional score test (CST), the proposed method 1 (PM1), and the proposed method 2 (PM2) at a significance level of $\alpha = 5 \times 10^{-5}$.
(PDF)

S2 Table. G test and G-GE test Power. G test Power of the conventional score test (CST), the proposed method 1 (PM1), and the proposed method 2 (PM2) under missing rate (2%, 5%, 10%, 30%), minor allele frequency (MAF) (10%, 30%), and the number of case/control (1,000, 5,000). The x-axis denotes genetic odds ratios ($OR_g = \exp(\beta_g)$). The significance level is 5×10^{-8}. G-GE test Power of CST, PM1, and PM2 under genetic odds ratios ($OR_g = \exp(\beta_g) = 1.1, 1.2$), missing rate (2%, 5%, 10%, 30%), minor allele frequency (MAF) (10%, 30%), and the

number of case/control is 1,000. The x-axis denotes gene-environment interaction odds ratios ($OR_{ge} = \exp(\beta_{ge})$). The significance level is 5×10^{-8}.
(PDF)

S3 Table. G test and G-GE test Power at the scenarios with smaller sample sizes and unbalance case-control samples. G test Power of the conventional score test (CST), the proposed method 1 (PM1), and the proposed method 2 (PM2) under missing rate (2%, 5%, 10%, 30%), minor allele frequency (MAF) (10%, 30%), and the number of case/control (1,000, 5,000). The x-axis denotes genetic odds ratios ($OR_g = \exp(\beta_g)$). The significance level is 5×10^{-8}. G-GE test Power of CST, PM1, and PM2 under genetic odds ratios ($OR_g = \exp(\beta_g) = 1.1, 1.2$), missing rate (2%, 5%, 10%, 30%), minor allele frequency (MAF) (10%, 30%), and the number of case/control is 1,000. The x-axis denotes gene-environment interaction odds ratios ($OR_{ge} = \exp(\beta_{ge})$). The significance level is 5×10^{-8}.
(PDF)

Acknowledgments

We are very grateful to the reviewers and the editor for their detailed comments and valuable suggestions, which have led to considerable improvement of this paper.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

This work was supported by Japan Society for the promotion of science (http://www.jsps.go.jp/english/), grant numbers JP16K00064, JP16K08638, JP16H05242, and JP16H01528 (received author is M.U.).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; EliLilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This work was carried out under the ISM General Cooperative Research 1 (2015-ISM-CRP-1013).

Author Contributions

Conceptualization: Shuntaro Sato, Masao Ueki.

Data curation: Shuntaro Sato, Masao Ueki.

Formal analysis: Shuntaro Sato.

Funding acquisition: Masao Ueki.

Investigation: Shuntaro Sato, Masao Ueki.

Methodology: Shuntaro Sato, Masao Ueki.

Project administration: Masao Ueki.

Resources: Shuntaro Sato, Masao Ueki.

Software: Shuntaro Sato.

Supervision: Masao Ueki.

Visualization: Shuntaro Sato.

Writing – original draft: Shuntaro Sato, Masao Ueki.

References

1. Manolio TA. Bringing genome-wide association findings into clinical use. Nature Reviews Genetics. 2013; 14(8):549–558. https://doi.org/10.1038/nrg3523 PMID: 23835440
2. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research. 2014; 42(D1):D1001–D1006. https://doi.org/10.1093/nar/gkt1229 PMID: 24316577
3. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006; 38(8):904–909. https://doi.org/10.1038/ng1847 PMID: 16862161
4. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007; 81(3):559–575. https://doi.org/10.1086/519795 PMID: 17701901
5. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4(1):7. https://doi.org/10.1186/s13742-015-0047-8 PMID: 25722852
6. Lee SM, Karrison TG, Cox NJ, Im HK. Quantitative Allelic Test—A Fast Test for Very Large Association Studies. Genetic Epidemiology. 2013; 37(8):831–839. https://doi.org/10.1002/gepi.21768 PMID: 24185610
7. Lin D, Hu Y, Huang B. Simple and efficient analysis of disease association with missing genotype data. The American Journal of Human Genetics. 2008; 82(2):444–452. https://doi.org/10.1016/j.ajhg.2007.11.004 PMID: 18252224
8. Aulchenko YS, Ripke S, Isaacs A, Van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007; 23(10):1294–1296. https://doi.org/10.1093/bioinformatics/btm108 PMID: 17384015
9. Couch FJ, Wang X, McGuffog L, Lee A, Olswold C, Kuchenbaecker KB, et al. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. PLoS genetics. 2013; 9(3):e1003212. https://doi.org/10.1371/journal.pgen.1003212 PMID: 23544013
10. Hicks AA, Pramstaller PP, Johansson Å, Vitart V, Rudan I, Ugocsai P, et al. Genetic determinants of circulating sphingolipid concentrations in European populations. PLoS genetics. 2009; 5(10):e1000672. https://doi.org/10.1371/journal.pgen.1000672 PMID: 19798445
11. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. Human heredity. 2007; 63(2):111–119. https://doi.org/10.1159/000099183 PMID: 17283440
12. Hamza TH, Chen H, Hill-Burns EM, Rhodes SL, Montimurro J, Kay DM, et al. Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via

interaction with coffee. PLoS Genetics. 2011; 7(8):e1002237. https://doi.org/10.1371/journal.pgen.1002237 PMID: 21876681

13. R Core Team. R: A Language and Environment for Statistical Computing; 2017. Available from: https://www.R-project.org/.
14. Ueki M. On the choice of degrees of freedom for testing gene–gene interactions. Statistics in medicine. 2014; 33(28):4934–4948. https://doi.org/10.1002/sim.6264 PMID: 25043617
15. Shen L, Thompson PM, Potkin SG, Bertram L, Farrer LA, Foroud TM, et al. Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. Brain imaging and behavior. 2014; 8(2):183–207. https://doi.org/10.1007/s11682-013-9262-z PMID: 24092460
16. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, et al. The international HapMap project. Nature. 2003; 426(6968):789–796. https://doi.org/10.1038/nature02168
17. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. https://doi.org/10.1038/nature11632
18. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics. 2007; 39(7):906–913. https://doi.org/10.1038/ng2088 PMID: 17572673
19. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature Reviews Genetics. 2010; 11(7):499–511. https://doi.org/10.1038/nrg2796 PMID: 20517342
20. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature genetics. 2012; 44(8):955–959. https://doi.org/10.1038/ng.2354 PMID: 22820512
21. Graffelman J, Nelson S, Gogarten S, Weir B. Exact inference for Hardy-Weinberg proportions with missing genotypes: single and multiple imputation. G3: Genes, Genomes, Genetics. 2015; 5(11):2365–2373. https://doi.org/10.1534/g3.115.022111
22. Graffelman J, Sánchez M, Cook S, Moreno V. Statistical inference for Hardy-Weinberg proportions in the presence of missing genotype information. PLoS One. 2013; 8(12):e83316. https://doi.org/10.1371/journal.pone.0083316 PMID: 24391752