Biostatistics (2018) **00**, 00, *pp.* 1–17
doi:10.1093/biostatistics/kxy030

STEPS: an efficient prospective likelihood approach to genetic association analyses of secondary traits in extreme phenotype sequencing

WENJIAN BI

Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

YUN LI

Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA, and Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599, USA

MATTHEW P. SMELTZER

Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN 38152, USA

GUIMIN GAO

Department of Public Health Sciences, University of Chicago, Chicago, IL 60637, USA

SHENGLI ZHAO

School of Statistics, Qufu Normal University, Qufu 273165, PR China

GUOLIAN KANG*

Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
Guolian.Kang@stjude.org

Summary

It has been well acknowledged that methods for secondary trait (ST) association analyses under a case–control design (ST_{CC}) should carefully consider the sampling process to avoid biased risk estimates. A similar situation also exists in the extreme phenotype sequencing (EPS) designs, which is to select subjects with extreme values of continuous primary phenotype for sequencing. EPS designs are commonly used in modern epidemiological and clinical studies such as the well-known National Heart, Lung, and Blood Institute Exome Sequencing Project. Although naïve generalized regression or ST_{CC} method could be applied, their validity is questionable due to difference in statistical designs. Herein, we propose a general prospective likelihood framework to perform association testing for binary and continuous STs under EPS designs (STEPS), which can also incorporate covariates and interaction terms. We provide a computationally efficient and robust algorithm to obtain the maximum likelihood estimates. We also

*To whom correspondence should be addressed.

present two empirical mathematical formulas for power/sample size calculations to facilitate planning of binary/continuous STs association analyses under EPS designs. Extensive simulations and application to a genome-wide association study of benign ethnic neutropenia under an EPS design demonstrate the superiority of STEPS over all its alternatives above.

Keywords: Extreme phenotype sequencing; Genome-wide association studies; Maximum likelihood estimate; Next generation sequencing studies; Secondary trait analysis.

1. INTRODUCTION

Genome-wide association studies (GWAS) and next generation sequencing (NGS) studies have successfully detected thousands of genetic variations associated with a wide variety of traits (Klein *and others*, 2005; Sanna *and others*, 2008; Solovieff *and others*, 2010; Sanders *and others*, 2012). Besides the specific primary trait that GWAS/NGS is designed for, many secondary traits (STs) are also worthy of investigation to further decipher the disease etiology or pathology. For example, studies designed for neutropenia typically measure the number of white blood cells (WBCs) as a primary trait, and additional blood test results (such as platelet counts) could be recorded and retained for secondary objectives (such as analysis of thrombocytopenia, Bunimov *and others*, 2013). Another common case is meta-analyses where the trait of interest (such as height) is not usually the primary trait of the single studies (Speliotes *and others*, 2010).

In this study, we focus on ST genetic association analyses under the extreme phenotype sequencing (EPS) designs. Due to the high genotyping or sequencing cost, the EPS design only selects and sequences the subjects with extremely large and small primary continuous trait values from the whole cohort. The EPS designs are widely adopted in many GWAS/NGS because it can obtain greater statistical power compared with randomly sequencing the same number of subjects (Kang *and others*, 2012). For example, National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (phs000400.v5.p1) included five subgroups for five primary phenotypes, of which three subgroups used the EPS design to select participants. A GWAS in benign ethnic neutropenia (BEN) selected and genotyped subjects with leukocyte counts at the lowest 1–7th percentile and at the 85th to 95th percentile. These projects also have a number of outcomes including both categorical and continuous traits, which could be used for ST analysis.

ST analyses have to consider both the sampling scheme and the correlation between primary trait and ST. Otherwise highly biased parameter estimates would occur. This property has been well studied for case–control designs (Lin and Zeng, 2009; Monsees *and others*, 2009; Wang and Shete, 2011; Ghosh *and others*, 2013; Kang *and others*, 2017) but has not received enough attention for EPS. A simple simulated example was used to show that the EPS design may significantly alter the correlation between genotype and ST (Figure S1, simulation details are in Appendix A1 of Supplementary Materials available at *Biostatistics* online). Suppose in the population, ST is positively correlated with primary trait (Figures S1A, S1E of Supplementary Materials available at *Biostatistics* online) and is not correlated with genotype (Figures S1B, S1F of Supplementary Materials available at *Biostatistics* online). If subjects with extreme large or extreme small primary trait are selected to sequence from the study cohort (EPS design, Figures S1C, S1G of Supplementary Materials available at *Biostatistics* online), then the ST is statistically significantly correlated with genotype (Figure S1D of Supplementary Materials available at *Biostatistics* online: analysis of variance P-value is 2.95×10^{-62}; Figure S1H of Supplementary Materials available at *Biostatistics* online: contingency table χ^2 test P-value is 8.29×10^{-37}). Even if we incorporate primary trait as a covariate, the false positive associations still exist. This example clearly shows that ignoring the EPS design would generate highly biased results for associating genotype and STs.

One typical method for ST analyses in EPS designs is to consider subjects in two extreme regions as "cases" and "controls" so that methods for a case–control design (ST_{CC}) can be applied (Lin and Zeng,

2009; Monsees *and others*, 2009; Wang and Shete, 2011; Ghosh *and others*, 2013; He *and others*, 2012; Kang *and others*, 2017). However, the transformation process of the primary trait from continuous type to binary type could result in huge loss of useful information. More importantly, the generated "case–control" data is not actually derived from a case–control design, which could result in biased estimates (as shown in simulation of Section 3.2). Thus, novel valid statistical methods for ST analyses under EPS designs are urgently needed.

To the best of our knowledge, only a few methods have considered the ST analysis under EPS. Lin *and others* (2013) proposed a nonparametric likelihood-based method to analyze continuous STs under trait-dependent sampling through a bivariate linear regression model. The method is able to adjust for covariates and has correct type I error control at a significance level of 10^{-3} in some situations. A free command-line program named SEQTDS can be easily accessed at http://dlin.web.unc.edu/software/score-seqtds/. However, some properties limit its applications. First, the method only considered continuous ST but cannot be used to analyze binary ST. Second, complete primary traits data in the original whole study cohort is required to assess the sampling scheme, which may not be directly available in some cases. Although we could still apply SEQTDS by treating binary ST as a continuous one and imputing the primary traits based on some marginal distribution assumptions, the SEQTDS method cannot control type I error rate at $\alpha = 0.05$ given some specific parameter settings (as shown in simulation of Section 3.2).

We propose a set-valued model to jointly characterize the relationship among genotype, primary trait, and ST, which can be continuous or binary. Then, we propose a novel ST association analysis method under EPS designs [we call it STs under EPS designs (STEPS) for short throughout the article]. We first use a prospective likelihood function to estimate model parameters. Then, we give a closed form of the Fisher information matrix to conduct the Wald test for associating genotype and ST. The model and the estimation approach can easily incorporate covariates and allow for environmental factors, genetic principle components, or interactions between genotype and environmental factors. We performed extensive simulations to compare STEPS with existing methods including straightforward linear/logistic regression, ST_{CC}, and SEQTDS method proposed by Lin *and others* (2013). Simulations and application to a GWAS of BEN all validated the super advantage of our new method. In addition, we also conducted simulations to evaluate STEPS under the polygenic architecture, which is the first time that polygenic effect, a well-known phenomenon in GWAS/NGS, is fully considered in ST genetic association analysis.

The remainder of the article is organized as follows. In Section 2, we introduce the models and propose the STEPS method for associating ST and genotype. In Section 3, we conduct extensive simulations to evaluate the properties of the proposed STEPS method. In Section 4, we apply the STEPS method to a real data example. Finally, Section 5 gives a brief summary.

2. Methods

In this section, we first briefly review three common approaches currently used in ST genetic association analyses under EPS designs. Then, we propose a joint set-valued model to characterize the relationships among primary and STs, genotype and covariates. Next, we use a prospective likelihood function to estimate model parameters and to construct a Wald test statistic for associating genotype and ST.

2.1. *Three commonly used approaches*

For ST genetic association analysis under EPS designs, three categories of approaches are widely used. The first one is the naïve linear/logistic regression (we call it LR for short throughout the article) that directly models the relationship between genotype and ST disregarding the primary trait and its corresponding

EPS design. The second one is to apply ST_{CC} to ST analysis under EPS designs. In the simulations below, we chose one of ST_{CC}, SPREG method (Lin and Zeng, 2009, http://dlin.web.unc.edu/software/spreg-2/) to show its property. The SPREG method employed a logistic regression model and retrospective likelihood conditional on disease status to handle the case–control sampling in the analysis of ST. It controls type I error rate at a liberal significance level of 0.05 but not at more stringent significance level such as 10^{-5} in some situations such as common disease and rare variants (RVs). Through a profile likelihood approach, environmental covariates can also be incorporated into model as a high-dimensional nuisance parameter. We did not consider the set-valued method proposed by Kang *and others* (2017) because it cannot incorporate covariates into the model, although the method provides more accurate type I error control and greater power, especially under stringent significance levels. The third one is SEQTDS proposed by Lin *and others* (2013) whose properties and limitations have been described in the introduction section.

2.2. *Joint modeling of the primary and secondary traits*

Suppose a cohort of N subjects are randomly selected from a general population. For the ith subject ($i \leq N$), let Y_i denote a continuous primary trait, let $\boldsymbol{X_i} = (X_{i1}, X_{i2}, \ldots, X_{im})$ denote m covariates, which might include age, gender, genetic ancestry scores, and so on. Then, we select n subjects with extreme large or extreme small primary traits from the N subjects for genotyping or sequencing. Let G_i denote the genotype for a specific single nucleotide polymorphism (SNP) locus, $i = 1, \ldots, n$. Let S_i be a selection indicator, which equals one if the ith subject is selected and equals zero otherwise. The indices of the N subjects are re-ordered so that the first n subjects are selected. We assume that both the primary trait and ST could be affected by genotype and covariates, and that primary trait could be affected by the ST. This assumption is widely adopted for ST analysis in case–control study design (Lin and Zeng, 2009; Kang *and others*, 2017).

If a ST is continuous, let Z_i denote the ST, which is a linear combination of G_i and $\boldsymbol{X_i}$. And let primary trait Y_i be a linear combination of G_i, $\boldsymbol{X_i}$ and Z_i. Four cut-offs of $y_0 > y_1 > y_2 > y_3$ are used to select subjects to genotype or sequence. Borrowing the idea of the set-valued model proposed for associating the primary binary trait and genotype (Kang *and others*, 2014), for the ith subject ($i \leq N$), the set-valued model under EPS is as follows.

$$\begin{cases} Z_i &= \gamma_0 + \gamma_1 G_i + \boldsymbol{\gamma_2 X_i} + e_{i,1} \\ Y_i &= \beta_0 + \beta_1 G_i + \boldsymbol{\beta_2 X_i} + \beta_3 Z_i + e_{i,2} \\ S_i &= I(y_0 > Y_i > y_1 \text{ or } y_2 > Y_i > y_3) \end{cases} \tag{2.1}$$

where γ_0 and β_0 are intercept terms, γ_1 and β_1 are regression coefficients for the SNP locus, $\boldsymbol{\gamma_2} = (\gamma_{21}, \gamma_{22}, \ldots, \gamma_{2m})^T$ and $\boldsymbol{\beta_2} = (\beta_{21}, \beta_{22}, \ldots, \beta_{2m})^T$ are vectors of regression coefficients for the m covariates. Coefficient β_3 represents the effect size of ST on primary trait. Error terms $e_{i,1}$ and $e_{i,2}$ are assumed to be independent and identically distributed with a normal distribution with a mean of 0 and a variance of σ_1^2 and σ_2^2, respectively. The cut-offs of y_1, y_2 are used to define the extreme large and small primary traits and the cut-offs of y_0, y_3 are used to define the outlier subjects with unreasonably large and small primary traits. If the study design does not exclude the outliers, then $y_0 = \text{Inf}$ and $y_3 = -\text{Inf}$, that is, $S_i = I(Y_i > y_1 \text{ or } y_2 > Y_i)$. After substituting Z_i into the formula of Y_i, we can derive that model (2.1) above is exactly the same as bivariate models (2) and (3) in Lin *and others*, 2013. More details are in Appendix A2 of supplementary material available at *Biostatistics* online.

If ST is dichotomous (e.g. binary variable of 1 or 0), we let Z_i denote a latent continuous variable and let D_i denote the binary ST. If the latent variable Z_i is greater than the cut-off z_0, then D_i is 1, otherwise,

D_i is 0. Similar to equation (2.1), the set-valued model is as follows.

$$\begin{cases} Z_i &= \gamma_0 + \gamma_1 G_i + \boldsymbol{\gamma_2 X_i} + e_{i,1} \\ D_i &= I(Z_i > z_0) \\ Y_i &= \beta_0 + \beta_1 G_i + \boldsymbol{\beta_2 X_i} + \beta_3 D_i + e_{i,2} \\ S_i &= I(y_0 > Y_i > y_1 \text{ or } y_2 > Y_i > y_3) \end{cases} \tag{2.2}$$

where model parameters $(\gamma_0, \gamma_1, \boldsymbol{\gamma_2}, \beta_0, \beta_1, \boldsymbol{\beta_2}, \beta_3, y_0, y_1, y_2, y_3)$ are similar with model (2.1). Error terms $e_{i,1}$ and $e_{i,2}$ also follow an independent normal distribution with a mean of 0 and a variance of σ_1^2 and σ_2^2, respectively.

2.3. *Maximum likelihood estimate (MLE) and Wald statistics*

We propose a maximum likelihood estimation method based on a prospective likelihood function. If the ST is continuous, the likelihood function is

$$L = \prod_{i=1}^{n} \Pr(Z_i = z_i, Y_i = y_i | G_i, \boldsymbol{X_i}, S_i = 1) \tag{2.3}$$

where z_i and y_i are secondary and primary trait values of ith subject, respectively. If the ST is dichotomous, the likelihood function is

$$L = \prod_{i=1}^{n} \Pr(D_i = d_i, Y_i = y_i | G_i, \boldsymbol{X_i}, S_i = 1) \tag{2.4}$$

where d_i and y_i are secondary and primary trait values of ith subject, respectively. The detailed derivation process of the probabilities can be seen in Appendix A3 of supplementary material available at *Biostatistics* online.

We employ a quasi-Newton algorithm to optimize the likelihood function, which is implemented with the R function optim() method "BFGS". For model (2.1), we estimate the parameters $(\gamma_0, \gamma_1, \boldsymbol{\gamma_2}, \beta_0, \beta_1, \boldsymbol{\beta_2}, \beta_3, \sigma_1, \sigma_2)$ that maximize the likelihood function (2.3). And for model (2.2), we fix $z_0 = 0, \sigma_1 = 1$ and estimate parameters $(\gamma_0, \gamma_1, \boldsymbol{\gamma_2}, \beta_0, \beta_1, \boldsymbol{\beta_2}, \beta_3, \sigma_2)$ that maximize the likelihood function (2.4). In the R package, we provide a simple method to estimate cutoffs y_1 and y_2 in case the information is unknown (Appendix A4 of supplementary material available at *Biostatistics* online).

The null hypothesis to test the association between genotype and ST is $H_0 : \gamma_1 = 0$, and the alternative hypothesis is $H_1 : \gamma_1 \neq 0$. Here, we propose a Wald test statistic, $\hat{\gamma}_1^2 / \hat{\text{var}}(\hat{\gamma}_1)$, which should follow χ^2 distribution with 1 degree of freedom under the regularity conditions under H_0. Here $\hat{\gamma}_1$ is obtained by the MLE shown above and $\hat{\text{var}}(\hat{\gamma}_1)$ is obtained by Fisher information matrix. The closed form of Fisher information matrix and the related proof about the regularity conditions can be seen in Appendices A3 and A5 of supplementary material available at *Biostatistics* online.

3. Simulations

We conducted extensive simulations in three parts. In part 1, we first compared STEPS with three methods of LR, SPREG, and SEQTDS in terms of type I error control and power at a liberal significance level $\alpha = 0.05$. Then, in part 2, we only evaluated STEPS in terms of type I error control and power under

more comprehensive parameter settings at more stringent significance levels $\alpha = 0.01$ and 10^{-5} because simulations from part 1 show that the other three methods cannot control type I error in some situations. In GWAS/NGS, the polygenicity is a well-known phenomenon that a large proportion of weak effects collectively contribute to the trait. As for ST analysis, it is even more complex since the polygenic architecture could affect both primary and STs. To the best of our knowledge, the polygenic architecture effect on ST analysis has not been discussed comprehensively. Thus, in part 3, we lastly evaluated STEPS under polygenic architecture in terms of type I error control and power at significance levels $\alpha = 0.01$, 0.001, and 10^{-5}. For SEQTDS, we adopted two methods to impute primary traits for the subjects without available genotype or STs, one is based on the true primary trait and the other one is imputed based on the marginal normal distribuiton (details can be seen in Appendix A6 of supplementary material available at *Biostatistics* online).

3.1. *Simulation process*

For each replication, we first generated N genotypes following Hardy–Weinberg equilibrium given minor allele frequency (MAF) of the tested SNP. Then, we simulated covariates and model error terms $\{(\boldsymbol{X_i}, e_{i,1}, e_{i,2}), i \leq N\}$ following independent standard normal distribution. Next, primary and STs were simulated based on model (2.1) or (2.2), depending on the type of ST. In this section, we simulated $m = 1$ covariate and fixed parameters $\gamma_0 = 0, \gamma_2 = 0.4, \beta_0 = 0, \beta_2 = 0.4, z_0 = 0, y_0 = \text{Inf}, y_3 = -\text{Inf}$. The upper ρ quantile and the lower ρ quantile of N primary traits were selected as cutoffs y_1 and y_2, so that $n = N \times 2\rho$ subjects in the cohort were retained based on EPS as the study sample.

For comparisons of STEPS with LR, SPREG, and SEQTDS, we fixed the sample size $n = 1000$, MAF $= 0.3, \rho = 0.2, \beta_1 = -0.4/0.4$ and increased β_3 from -0.7 to 0.7 in increments of 0.1. We considered H_0 with $\gamma_1 = 0$ and H_1 with $\gamma_1 = 0.15$ or -0.15, for which the heritability of the ST (h^2, i.e. the proportion of phenotypic variation of Z_i attributing to G_i) is 0.8%. For each parameter setting, 10 000 replications were simulated to assess the type I error rate and power at a liberal significance level $\alpha = 0.05$, parameter estimation, mean squared error, and coverage probability of 95% Wald-type confidence intervals for the genetic effect on ST.

For examination of STEPS at more stringent significance levels $\alpha = 0.01$ and 10^{-5} under H_0 with $\gamma_1 = 0$, we considered coefficient β_1 of 0.4 and -0.4 to simulate different effects of genotype on primary trait, coefficient β_3 of -0.7, -0.4, 0, 0.4, and 0.7 to simulate different effect sizes and directions of ST on primary trait, and $\rho = 0.2, 0.1, 0.05$, and 0.01 to simulate different EPS designs. Three MAFs of 0.3, 0.05, and 0.005 were used to simulate common variants, less common variants (LCV), and RVs, respectively. For continuous (binary) ST, we fixed sample sizes $n = 1000$ (2000). For each parameter setting, we evaluated type I error rates with 10^7 replications. We also designed three simulation scenarios to evaluate power under different parameter settings at stringent significance levels $\alpha = 10^{-5}$ based on 10 000 replications (details in Appendix A7 of supplementary material available at *Biostatistics* online).

For assessment of the effect of the polygenic architecture on STEPS, the primary trait is assumed to be affected by 100 causal SNPs in four regions (25 SNPs/region) each with a different linkage disequilibrium (LD) structure of no, weak, moderate or strong LD, respectively. This means that the genetic effect on primary trait $\beta_1 G$ in models (2.1) and (2.2) is replaced by $\sum_{j=1}^{100} \beta_{1j} G_j$, where $\{G_j, j = 1, \ldots, 100\}$ are genotypes of 100 causal SNPs affecting the primary trait with effect sizes of $\{\beta_{1j}, j = 1, \ldots, 100\}$. The genotypes of 100 causal SNPs among four regions were simulated independently based on R code of simRareSNP.R (http://www.biostat.umn.edu/∼weip/prog/BasuPanGE11/simRareSNP.R, Basu and Pan, 2011; Wang, 2016) with a fixed MAF of 0.3 and parameter rho = 0, 0.3, 0.6 and 0.9 to simulate 4 LD regions, respectively. For ST, we considered two scenarios: (i) no associations between all of these 100 SNPs and ST, $\gamma_1 = 0$. We randomly selected 4 SNPs with one from each of four regions for their association

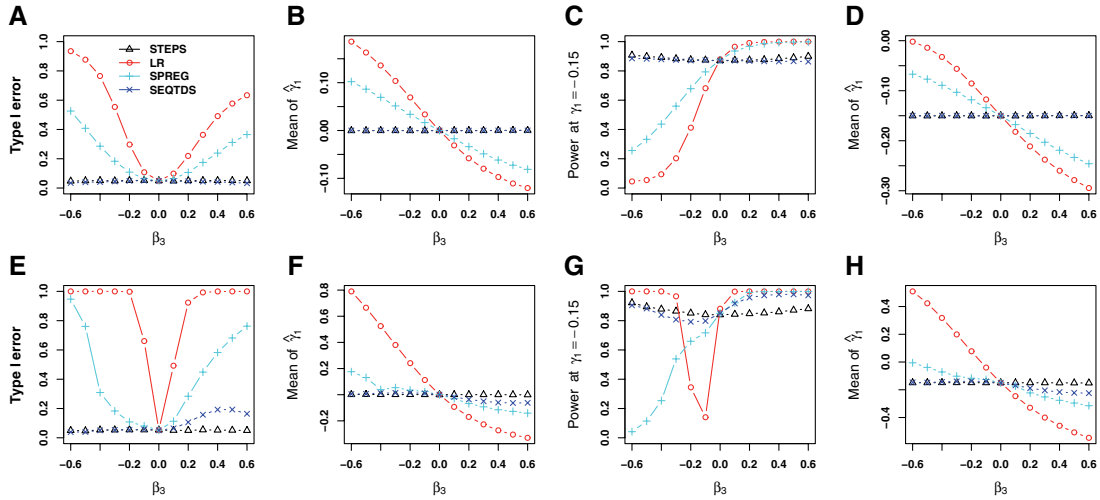

Fig. 1. Continuous STs: comparisons of STEPS, LR, SPREG, and SEQTDS methods at a significance level of 0.05 based on 10 000 replications. $\beta_1 = -0.4$, MAF $= 0.3$, sample size $n = 1000$. (A–D) $\rho = 0.2$; (E–H) $\rho = 0.01$.

testing with ST; (ii) four SNPs with one randomly selected from each of the four regions are associated with ST as four causal SNPs of ST but all others are not associated with ST. This means that the genetic effect on ST $\gamma_1 G_i$ in models (2.1) and (2.2) is replaced by $\sum_{k=1}^{4} \gamma_{1k} G_{ik}$ where $\{G_{ik}, k = 1, 2, 3, 4\}$ are genotypes of four selected causal SNPs affecting ST with an effect size of γ_{1k}. We considered $\gamma_{1k} = -0.15$ and 0.15 for the four selected causal SNPs of ST which represents its overall heritability 3.64% (0.91% for each causal SNP). For both scenarios, we let $\beta_1 = -0.04/0.04$ to simulate weak polygenic effects with a heritability of the primary trait per each causal SNP less than 0.1% (h^2 for 100 SNPs are from 21.6% to 36.4%). Five β_3 values ranging from -0.7 to 0.7 were considered.

We tested the associations between ST and each of the four selected tested SNPs and estimated the type I error rate and power of STEPS for scenarios 1 and 2 as the proportions of replicates with P-values $< \alpha = 0.01, 0.001$ and 10^{-5}. For Scenario 2, besides these four causal SNPs of ST, we also randomly selected another four non-causal SNPs of ST with one from each of 4 LD regions, among which three non-causal SNPs are in LD with three causal SNPs of ST in three LD regions. We tested their associations with ST and reported the proportions of replicates with P-values $< \alpha$ for each SNP. Here, 5×10^6 and 50 000 replicated datasets were simulated for scenarios 1 and 2, respectively, with a given sample size of 1000. We also randomly generated MAFs for 100 SNPs following a uniform distribution of U(0.05,0.5) instead of a fixed constant MAF of 0.3 across all 100 SNPs with all the other parameters exactly same and conclusions are similar (data not shown).

3.2. *Comparison of STEPS, LR, SPREG, and SEQTDS methods*

Figures 1 and 2 show the simulation results for continuous and binary STs with two EPS designs $\rho = 0.2$ (Figures 1A–D and 2A–D) and $\rho = 0.01$ (Figures 1E–H and 2E–H) given $\beta_1 = -0.4$. As SPREG cannot output P-values for many replications when $|\beta_3| = 0.7$ and $\rho = 0.01$, we only showed results for $|\beta_3| \leq 0.6$ in Figures 1 and 2. We can see that no matter ST is continuous (Figures 1A–B and 1E–F) or binary (Figures 2A–B and 2E–F), STEPS always gave accurate parameter estimation (Table S1 of Supplementary Materials available at *Biostatistics* online), which leads to correct type I error control at $\alpha = 0.05$ regardless of β_3. However, as expected, LR and SPREG were invalid unless primary trait is

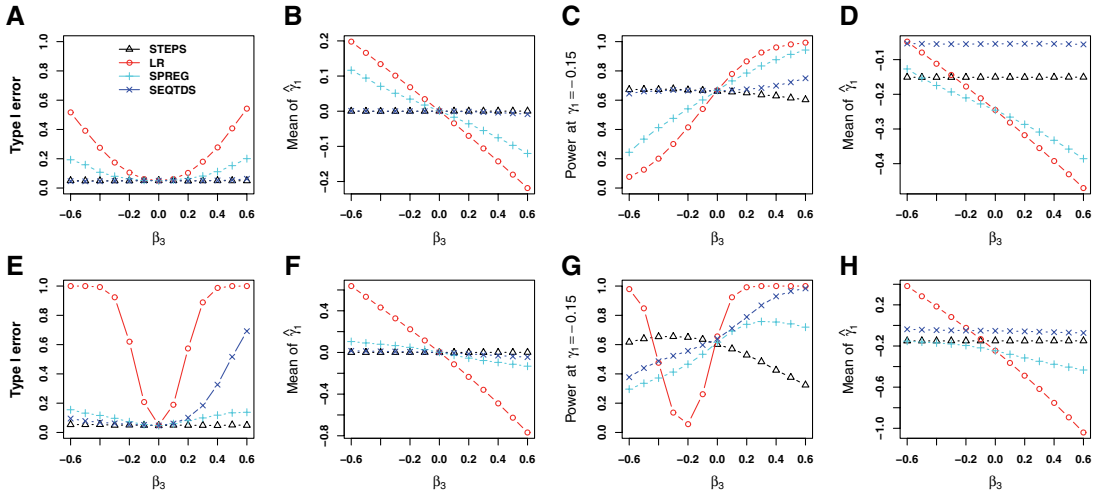

Fig. 2. Binary STs: comparisons of STEPS, LR, SPREG, and SEQTDS methods at a significance level of 0.05 based on 10 000 replications. $\beta_1 = -0.4$, MAF $= 0.3$, sample size $n = 1000$. (A–D) $\rho = 0.2$; (E–H) $\rho = 0.01$.

not correlated with ST, i.e., $\beta_3 = 0$ because of their biased parameter estimations due to disregard or inappropriate consideration of EPS. SPREG performs better than LR for either binary or continuous ST, which indicates that considering EPS as the "case–control" study could help the parameter estimation and type I error control to some extent. Under EPS design $\rho = 0.2$, SEQTDS generally performed similar to STEPS. Only when $|\beta_3|$ is greater than 0.4, the $\hat{\gamma}_1$ by SEQTDS is a little biased, which leads to inflated type I error rate (Table S6 of Supplementary Materials available at *Biostatistics* online). Under more extreme EPS design $\rho = 0.01$, SEQTDS could not control type I error for both continuous and binary ST if $\beta_3 > 0$. The inflated type I error is also due to the biased estimate $\hat{\gamma}_1$. For example, when $\beta_3 = 0.7$, SEQTDS gave $\hat{\gamma}_1$ of $-0.05(-0.06)$ and type I error rates of 0.82 (0.14) for binary (continuous) ST at $\alpha = 0.05$ (Table S6 of Supplementary Materials available at *Biostatistics* online).

Under $H_1(\gamma_1 \neq 0)$, no matter ST is continuous (Figures 1C–D and 1G–H) or binary (Figures 1C–D and 1G–H), STEPS always gave accurate parameter estimates and power at $\alpha = 0.05$ regardless of β_3. However, the parameter estimates $\hat{\gamma}_1$ with LR and SPREG changed a lot and increased with increase in β_3 and approached true γ_1 when $\beta_3 = 0$ (the trend is similar to that under H_0). We also performed similar simulations with $\beta_1 = 0.4$ (Figures S2 and S3 of Supplementary Materials available at *Biostatistics* online). Under sampling design $\rho = 0.2$, SEQTDS generally performed similar to STEPS if ST is continuous. Only when $|\beta_3| > 0.4$, SEQTDS slightly lost power compared with STEPS (Table S6 of Supplementary Materials available at *Biostatistics* online). When ST is binary, power of SEQTDS would be greater than that of STEPS if true $\gamma_1 < 0$ and $\beta_3 > 0$ ($\text{sign}(\gamma_1) \times \text{sign}(\beta_3) \times \text{sign}(\beta_1) = 1$), and would be less than that of STEPS if true $\gamma_1 < 0$ and $\beta_3 < 0$ ($\text{sign}(\gamma_1) \times \text{sign}(\beta_3) \times \text{sign}(\beta_1) = -1$). However, the greater power of LR, SPREG and SEQTDS here should be interpreted cautiously due to their uncontrolled type I error rates.

SEQTDS shows similar performance as STEPS in some cases. While for very EPS design (very small ρ) or binary ST, its parameter estimate is biased and type I error rate is inflated. Simulations also show that performances of SEQTDS with two imputing methods are almost the same, which indicates that the primary trait could be reasonably imputed if the primary traits truly approximately follow normal distribution. SEQTDS is not designed for bianry STs analysis, so that the unstable performance for binary ST is expected. While more importantly, the simulations showed that SEQTDS does not perform

as stable as STEPS even for continuous ST when $\rho = 0.01$. The finding is striking since the model in Lin *and others* (2013) is actually same as model (2.1) in this article after a transformation (see Section 2.2). To confirm the consistence between models, we also conducted simulations following the model in Lin *and others* (2013) and validated that SEQTDS cannot control type I error rate at $\alpha = 0.05$ given some specific parameter settings even under its own model (Appendix A2 of supplementary material available at *Biostatistics* online). Although both SEQTDS and STEPS assume the same model, SEQTDS is based on a nonparametric likelihood function and STEPS is based on a parametric likelihood function, which should be the main reason that the two methods perform differently.

3.3. *Type I error of STEPS*

STEPS could control the type I error rate at $\alpha = 0.01$ and 10^{-5} across all parameter settings for both continuous (Table 1) and binary STs (Tables S3 of Supplementary Materials available at *Biostatistics* online). This is resulted from the fact that the mean of estimated parameter $\hat{\gamma}_1$ is very close to 0, and the empirical standard deviation $\text{sd}(\hat{\gamma}_1)$ is very close to the mean of the estimated standard error $\widehat{\text{se}}(\hat{\gamma}_1)$ (Tables S2 and S4 of Supplementary Materials available at *Biostatistics* online). These ensure that the Wald statistic could truly follow chi-square distribution so that its type I error could be controlled. It is striking that STEPS can control type I error rates even for RVs (MAF = 0.005) under a significance level as stringent as $\alpha = 10^{-5}$.

3.4. *Power of STEPS*

For continuous STs, the power of STEPS would increase with increase in effect size γ_1 (Figures 3A–C) or in sample size n (Figures 1D–F). Interestingly, the smaller ρ gives greater power conditional on the same sample size, especially for LCV with MAF of 0.05 and RV with MAF of 0.005 because of more enriched minor alleles in selected samples. For example, for a RV with MAF of 0.005, $\beta_1 = -0.4$, and $\beta_3 = -0.7$, if the top and bottom of $\rho = 0.2$ ($\rho = 0.01$) from a cohort of 2500 (50000) individuals were selected under EPS, i.e., $n = 1000$, then the mean numbers of minor allele counts captured in the study samples are 13.98 (37.34), which leads to their respective power of 0.256 (0.779) (Figure 1F). This strongly supports the assumption that rare causal variants are likely to be enriched in samples with more extreme phenotypes so that EPS designs can capture these causal RVs for primary and STs with higher probability and have greater statistical power to detect them. However, this superiority does not hold when $\beta_1 = \beta_3 = 0$, i.e. different ρ correspond to similar number of minor allele counts and power (Figures S4 and S5 of Supplementary Materials available at *Biostatistics* online). In sharp contrast, as (β_1, β_3) diverge from $(0, 0)$, MAFs estimated in selected samples increases and the corresponding power also increases, especially for LCV and RV. The patterns of MAFs changes and of power changes are similar (Figure S4 of Supplementary Materials available at *Biostatistics* online). Hence, the power change under EPS are partially explained by the number of MA counts in the selected study samples. In addition, Table S2 of Supplementary Materials available at *Biostatistics* online shows as the decrease in MAF, $\widehat{\text{se}}(\hat{\gamma}_1)$ increases, which indicates the power loss. Similarly, as the decrease in ρ, $\text{sd}(\hat{\gamma}_1)$ decreases, especially for LCV and RV, which explains why smaller ρ corresponds to greater power given n.

3.5. *Polygenic architecture effect on STEPS*

To the best of our knowledge, it is the first time that polygenic architecture is comprehensively considered in the context of ST association testing. The results can be seen in Table S5 of Supplementary Materials available at *Biostatistics* online. Strikingly, under polygenic architecture, STEPS can still control type I error rates at $\alpha = 0.01$ and 10^{-5} for each of the selected four tested SNPs in four regions for scenario 1,

Table 1. Ratio of the empirical type I error rates to significance levels of α based on 10^7 replications given $n = 1000$

ρ	MAF	$\alpha = 0.01$					$\alpha = 10^{-5}$				
		$\beta_3 = -0.7$	$\beta_3 = -0.4$	$\beta_3 = 0$	$\beta_3 = 0.4$	$\beta_3 = 0.7$	$\beta_3 = -0.7$	$\beta_3 = -0.4$	$\beta_3 = 0$	$\beta_3 = 0.4$	$\beta_3 = 0.7$
$\beta_1 = -0.4$											
0.2	0.3	1.026	1.027	1.022	1.029	1.022	1.110	1.190	1.260	1.140	1.160
	0.05	1.030	1.029	1.042	1.029	1.028	1.250	0.930	1.200	1.190	1.090
	0.005	1.045	1.083	1.038	1.059	1.039	0.920	1.190	1.230	1.090	1.160
0.1	0.3	1.023	1.034	1.028	1.030	1.027	1.040	1.180	1.070	1.190	1.280
	0.05	1.022	1.029	1.036	1.036	1.029	0.960	1.180	1.070	1.280	1.060
	0.005	1.021	1.066	1.038	1.045	1.032	0.990	1.130	1.420	1.200	0.990
0.05	0.3	1.024	1.028	1.027	1.027	1.034	1.040	1.130	1.090	1.030	1.110
	0.05	1.028	1.027	1.032	1.026	1.028	1.250	1.410	1.120	1.060	1.250
	0.005	1.015	1.069	1.034	1.035	1.006	0.960	1.160	1.080	0.870	1.200
0.01	0.3	1.026	1.028	1.030	1.033	1.033	1.170	1.190	1.230	1.310	1.010
	0.05	1.027	1.028	1.033	1.029	1.026	1.160	1.220	1.000	1.350	1.150
	0.005	0.959	1.055	1.040	1.001	0.967	1.240	0.950	1.060	1.160	1.290
$\beta_1 = 0.4$											
0.2	0.3	1.028	1.028	1.021	1.022	1.023	1.050	1.060	1.070	1.250	0.980
	0.05	1.027	1.033	1.041	1.030	1.030	1.290	1.210	1.310	1.030	1.010
	0.005	1.046	1.078	1.032	1.057	1.041	0.820	1.230	1.330	1.160	0.930
0.1	0.3	1.026	1.031	1.031	1.031	1.031	1.350	1.240	1.110	1.200	1.010
	0.05	1.023	1.027	1.030	1.028	1.027	1.090	1.230	0.970	1.110	1.110
	0.005	1.027	1.072	1.030	1.038	1.027	0.850	1.000	1.380	0.940	0.760
0.05	0.3	1.028	1.028	1.032	1.027	1.028	1.320	1.360	1.300	1.300	1.110
	0.05	1.030	1.032	1.031	1.030	1.027	1.090	1.180	1.130	1.400	1.020
	0.005	1.015	1.071	1.045	1.038	1.006	0.810	1.040	1.250	1.100	0.930
0.01	0.3	1.023	1.029	1.028	1.030	1.035	1.050	1.170	1.090	1.170	1.110
	0.05	1.021	1.031	1.032	1.030	1.033	0.900	1.200	1.040	1.060	1.170
	0.005	0.958	1.049	1.041	1.007	0.965	0.910	1.130	1.190	1.030	1.28

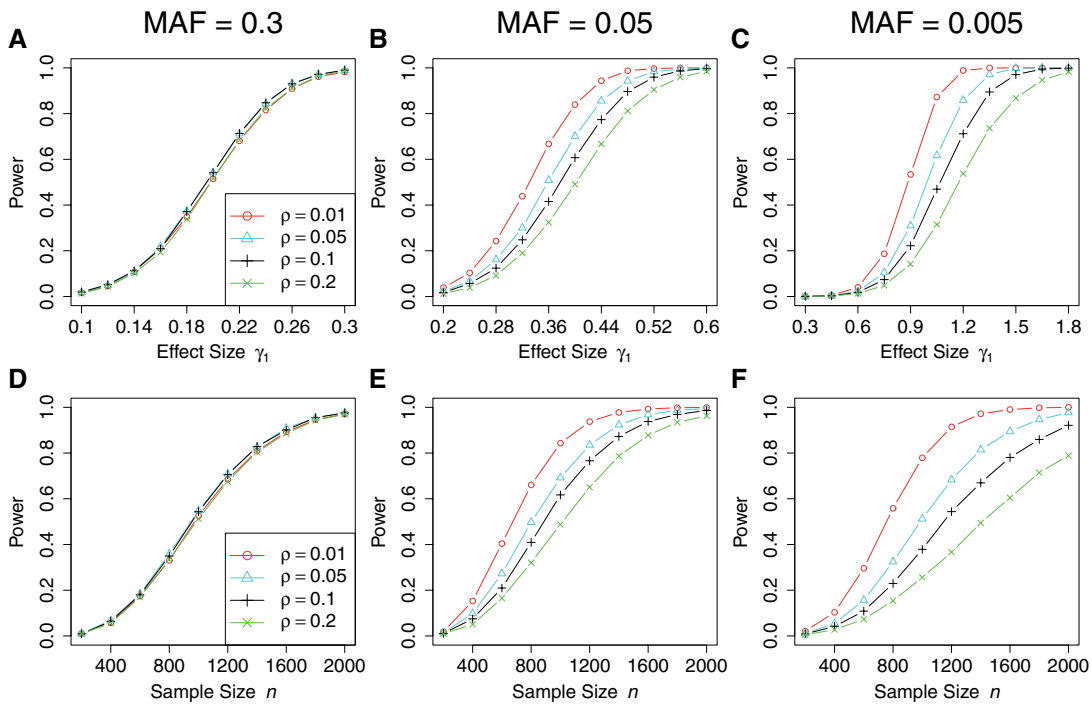

Fig. 3. Power of STEPS as a function of effect size and sample size. For (A–C), sample size is fixed at $n = 1{,}000$; For (D–F), effect size is fixed at $\rho_1 = 0.2, 0.4$, and 1 respectively. The results are the empirical power at a significance level of 10^{-5} based on 10 000 replications. $\beta_1 = -0.4, \beta_3 = -0.7$.

no matter the tested SNP is in LD or no LD with the other causal SNPs of the primary trait. This is resulted from the facts that (i) all the estimates of the coefficients related to ST are unbiased; (ii) the effects of the other causal SNPs on the primary traits can be absorbed into the error term and the effect of the tested SNP on the primary trait with decomposition levels depending on the LD structure between the tested SNP and the other causal SNPs of the primary trait showed by simulations (Table S10 of Supplementary Materials available at *Biostatistics* online). For scenario 2, for the non-causal SNP selected in the no LD region, the proportion of replicates for which the test is rejected was very close to α. This means that the type I error rate could also be correctly controlled for non-causal SNPs of ST which are causal SNPs of the primary trait and are in linkage equilibrium with causal SNPs of ST due the reason that $\sum_{k=1}^{4} \gamma_{1k} G_k$, the effect of the four causal SNPs of ST can be absorbed in the error term.

The power of STEPS for each of four causal SNPs under different LD scenarios were also quite similar given the same effect size (same h^2). Interestingly, for the other three non-causal SNPs selected in small, moderate, and strong LD regions under H_1, the proportion of replicates for which it is rejected was largest for the non-causal SNP in strong LD region and was smallest for the non-causal SNP in small LD region with one in the moderate region in-between. For example, given $\beta_1 = -0.04, \beta_3 = -0.7$, the power for testing four causal SNPs of binary ST in four LD regions was 0.42 to detect an $h^2 = 0.91\%$ with a sample size of 1000. In sharp contrast, the proportions of replicates for which the non-causal SNP are rejected was 0.18, 0.05, and 0.02 if non-causal SNP is in strong, moderate and small LD with the causal SNP, respectively.

4. APPLICATION TO A GWAS OF BEN

BEN is a clinical condition characterized by a relative reduction in neutrophil count. Hence, WBC is a common continuous index to indicate the BEN. Here, we applied four methods above to a GWAS of BEN in which around 1000 samples were selected from a large national cohort study including over 14 000 African-Americans with low WBC (at the lowest 1–7th percentile) and high WBC (at the 85th to 95th percentile). More description about the dataset can be seen in dbGaP (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000507.v1.p1).

We considerd 7 STs including C-reactive protein (CRP), triglycerides (TL), platelet count (PC), high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol (TC), and Albumin serum (ALS) and eight covariates of age, smoking status, gender and top five principle components. We used log-transformed CRP, TC, TL, PC, and square-root-transformed HDL, LDL and un-transformed ALS (Ma *and others*, 2010; Bryant *and others*, 2014; Oh *and others*, 2014; Ligthart *and others*, 2016; Prins *and others*, 2017; Zhu *and others*, 2017). After removing subjects whose WBC or any one covariate is missing, we retained 980 genetically independent subjects with 677,755 SNPs after removing SNPs whose MAFs are less than 0.005. Of the seven STs, PC, TL, and CRP are positively correlated with WBC and HDL is negatively correlated with WBC in the study sample (Table S7 of Supplementary Materials available at *Biostatistics* online). For STEPS, cutoffs of $y_0 = 8.92, y_1 = 7.37, y_2 = 3.35, y_3 = 1.5$ were given based on the distribution of WBC in the study sample.

As a ST analysis method designed for case–control study, SPREG requires a critical pre-given parameter of population prevalence. Although we can simply treat subjects with low WBC as cases and subjects with high WBC as controls, it is not intuitive how to give the prevalence parameter since any one is biased compared with true sampling process. Table S8 of Supplementary Materials available at *Biostatistics* online showed that when a prevalence of 0.07 was used to match the lowest 1–7th percentile as "cases", SPREG has an inflation factor of greater than 2 for CRP; when a prevalence of 0.5 was used, almost all *P*-values are 1; but a prevalence of 0.25 gave the most reasonable results in terms of QQ-plot and inflation factor. The Manhattan and QQ plots are shown in Figure S6 of Supplementary Materials available at *Biostatistics* online. All four methods gave reasonable QQ plots and had inflation factors between 0.99 and 1.08 (Table S8 of Supplementary Materials available at *Biostatistics* online).

To demonstrate the effectiveness of STEPS in analyzing real BEN data and to demonstrate its potential usefulness in GWAS/NGS, we summarized all significant SNPs defined as p-values $< 10^{-6}$ into three groups based on the existing results on GWAS catalog (https://www.ebi.ac.uk/gwas/) in Table 2. If a significant SNP is within a reported gene or on an intergenic region adjacent to a reported gene, the association is considered as highly possible positive. If a significant SNP is within a gene whose adjacent gene has been reported, the association is considered as medium possible positive. Otherwise, the association is considered as lowly possible positive. For the three STs of ALS, TC, and LDL not correlated with WBC, as expected, all four methods gave similar results in terms of QQ plots and identified significant SNPs. This is consistent with the simulation results that LR, SPREG, and SEQTDS can be used to analyze STs under EPS designs when primary trait is not associated with ST. However, for the four STs (PC, TL, CRP, and HDL) correlated with WBC, STEPS identified more significant SNPs which are highly/medium possible positive than but similar number of lowly possible positive SNPs to the other three methods. Furthermore, for these SNPs, 7 SNPs corresponds to $\text{sign}(\gamma_1)\times\text{sign}(\beta_3)\times\text{sign}(\beta_1) = -1$. Their p-values by STEPS are smallest among all four methods. This is perfectly consistent with the simulation results that LR, SPREG, and SEQTDS are less powerful to detect associations of STs with SNPs if $\text{sign}(\gamma_1)\times\text{sign}(\beta_3)\times\text{sign}(\beta_1) = -1$. For the other five SNPs, *P*-values by STEPS are still smaller than those by SEQTDS although not by LR and/or SPREG.

Ridker *and others*, 2008 reported the association between CRP and SNP rs3091244 that locates on the upstream of gene CRP. Only STEPS identified their association at a significance level of 10^{-6}, while

Table 2. Comparison of the analysis results of seven STs in a GWAS of BEN data

ST	Number of highly possible positive SNPs					Number of medium possible positive SNPs					Number of lowly possible positive SNPs			
	Methods				Reported	Methods				Adjacent	Methods			
	LR	SEQTDS	SPREG	STEPS	Gene	LR	SEQTDS	SPREG	STEPS	Gene	LR	SEQTDS	SPREG	STEPS
ST is not correlated with WBC														
LDL	1	1	1	1	APOE	0	0	0	0		0	0	0	0
ALS	0	0	0	0		0	0	0	0		0	0	0	0
TC	0	0	0	0		0	0	0	0		0	0	0	0
ST is positive correlated with WBC														
PC	0	0	0	0		0	0	0	1	EHD3	0	0	0	0
TL	2	2	2	2	MIR148A	0	0	0	0		1	0	0	0
CRP	1	3	2	4	CRP	0	0	0	0		1	0	1	1
ST is negative correlated with WBC														
HDL	0	0	0	0		0	0	0	1	AMPD3	0	0	0	0

ST, secondary trait; CRP, C-reactive protein; TL, triglycerides; PC, platelet count; HDL, high-density lipoprotein; LDL, low-density lipoprotein; TC, total cholesterol; ALS, Albumin serum.

P-values of SPREG, LR and SEQTDS were greater than the cutoff (Table S9 of Supplementary Materials available at *Biostatistics* online). STEPS has also uniquely identified two novel SNPs locating on known regions for HDL and PC. As for HDL, STEPS identified SNP rs1035691 which locates in the intron region of gene MRVI1. The gene is on cytoband of 11p15.4 and is adjacent to gene AMPD3 in which several SNPs have been reported to be associated with HDL (Teslovich *and others*, 2010; Willer *and others*, 2013; Spracklen *and others*, 2017). In addition, Webb *and others*, 2017 also reported the association between MRVI1 and coronary artery disease. As for PC, STEPS identified SNP rs207444 which locates in the intron of gene XDH. The gene is on cytoband of 2p23.1 and is adjacent to gene EHD3 in which several SNPs have been reported to be associated with PC (Astle *and others*, 2016). And O'Byrne *and others*, 2000 also reported a potential relationship among platelet, Xanthine Oxidoreductase (XO) and Xanthine DeHydrogenase (XDH). Furthermore, for some reported SNPs such as rs726640, although four methods all identified its association with CRP, STEPS gives the smallest *P*-value (Table S9 of Supplementary Materials available at *Biostatistics* online). All these evidences strongly indicate that the new STEPS method could be more effective and powerful to identify SNPs truly associated with STs under EPS than the other three methods.

5. DISCUSSION

We have proposed a novel STEPS method to test for association between binary or continuous STs and genetic variants under different EPS designs. To the best of our knowledge, there is no statistical method that appropriately takes into account the EPS designs when only study data is available, although EPS designs are widely adopted in many GWAS or NGS projects. Currently, to test associations between STs and genetic variants, naïve generalized regression or STs association analyses methods implemented for case–control designs are often used, which have been proven invalid both theoretically and empirically if both traits are correlated. Nonparametric likelihood method (SEQTDS) (Lin *and others*, 2013) can be used to analyze ST under EPS but cannot control type I error in some situations and could have smaller power than STEPS. Compared with the existing methods, STEPS takes account of the EPS designs more appropriately and therefore generates unbiased parameter estimations and better type I error control at both liberal (0.05) and stringent (10^{-5}) significance levels. In addition, in some situations, STEPS is more powerful than the existing methods, while the latter could not control type I error rates.

Most complex traits have polygenic architecture. That is, the primary trait can be affected by hundreds or thousands of causal SNPs each with weak effect. As a consequence, the second equation in Equation (2.1) and the third equation in Equation (2.2) are no longer valid. Under this situation, strikingly, the new proposed STEPS is still valid by simulations. This is intuitively understandable because the effect of the other causal SNPs on the primary trait could be absorbed into the error term and/or the effect of the tested SNP on the primary trait depending on the LD structure between the tested SNP and the causal SNPs of the primary trait and would not modify the effect of the tested SNP on ST. This is the first time to show that polygenetic architecture would not affect the ST genetic association analysis if appropriate statistical method is employed.

Simulations show that BFGS algorithms usually need less than 20 iterations to find the MLE, which makes STEPS computationally efficient. Although only demonstrated for a single SNP analysis in this study, STEPS can readily be applied to analyses of any form of predictor variables such as environmental exposure variables, gene expression, or other genomic features. In addition, the method can easily incorporate covariates such as age, gender, genetic ancestry estimates, or gene-environment interactions.

As a single-variant analysis method, STEPS is also underpowered to identify RVs, although the type I error rate for RVs could be well controlled. For RVs analysis, the standard method is to aggregate a set of variants as a genomic region and to perform region-based analysis. For example, burden-based and SNP-set Kernel Association Tests are two main categories of region-based methods and have been generalized

in many fields. Although Liu and Leal, 2012 proposed a framework to analyze RVs with selected samples, we still believe that is an important area for further investigation using the set-valued model.

In summary, the power of STEPS is a complicated function of the SNP MAF, cohort sample size, the proportion of extremes selected, effect size of SNP, and the correlations among primary trait STs and genotype. Via extensive simulation studies ($\sim$11 000 parameter combinations), we have quantified the relationship between association parameters and the power of STEPS for binary and continuous STs at four different significance levels $\alpha = 10^{-5}, 10^{-6}, 10^{-7}$, and 10^{-8} and have included them as a R function in STEPS software. These formulas are very crucial and can be easily and readily used to calculate power given sample size and all the other parameters in the planning stage of new ST-association study under EPS designs.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgements

We thank the editors and two anonymous reviewers for their insightful and helpful comments which have significantly improved the manuscript. This research is supported by the American Lebanese and Syrian Associated Charities (ALSAC). We acknowledge dbGAP for approval of our use of benign ethnic neutropenia data. The data were obtained from Matthew Hsieh's ancillary proposal to the Reasons of Geographic and Racial Differences in Stroke (REGARDS) study. Matthew Hsieh is supported by the intramural research program of NHLBI and NIDDK at NIH. Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096C and HHSN268201100011I. We acknowledge the High Performance Computing Facility (HPCF) at SJCRH for providing shared HPC resources that have contributed to the research results reported within this article. *Conflict of Interest*: None declared.

References

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M. A. *and others*. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429.

Basu, S. and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology* **35**, 606–619.

Bryant, E. K., Dressen, A. S., Bunker, C. H., Hokanson, J. E., Hamman, R. F., Kamboh, M. I. and Demirci, F. Y. (2014). A multiethnic replication study of plasma lipoprotein levels-associated snps identified in recent GWAS. *PLoS One* **8**, e63469.

Bunimov, N., Fuller, N. and Hayward, C. P. M. (2013). Genetic loci associated with platelet traits and platelet disorders. *Semin Thromb Hemost* **3**, 291–305.

Ghosh, A., Wright, F. A. and Zou, F. (2013). Unified analysis of secondary traits in case–control association studies. *Journal of the American Statistical Association* **108**, 566–576.

He, J., Li, H., Edmondson, A. C., Rader, D. J. and Li, M. (2012). A gaussian copula approach for the analysis of secondary phenotypes in case–control genetic association studies. *Biostatistics* **13**, 497–508.

Kang, G., Bi, W., Zhang, H., Pounds, S., Cheng, C., Shete, S., Zou, F., Zhao, Y., Zhang, J. F. and Yue, W. (2017). A robust and powerful set-valued approach to rare variant association analyses of secondary traits in case-control sequencing studies. *Genetics* **205**, 1049–1062.

KANG, G., BI, W., ZHAO, Y., ZHANG, J. F., YANG, J. J., XU, H., LOH, M. L., HUNGER, S. P., RELLING, M. V., POUNDS, S. *and others*. (2014). A new system identification approach to identify genetic variants in sequencing studies for a binary phenotype. *Human Heredity* **78**, 104–116.

KANG, G., LIN, D., HAKONARSON, H. AND CHEN, J. (2012). Two-stage extreme phenotype sequencing design for discovering and testing common and rare genetic variants: efficiency and power. *Human Heredity* **73**, 139–147.

KLEIN, R. J., ZEISS, C., CHEW, E. Y., TSAI, J. Y., SACKLER, R. S., HAYNES, C., HENNING, A. K., SANGIOVANNI, J. P., MANE, S. M., MAYNE, S. T. *and others*. (2005). Complement factor h polymorphism in age-related macular degeneration. *Science* **308**, 385–389.

LIGTHART, S., VAEZ, A., HSU, Y. H., STOLK, R., UITTERLINDEN, A. G., HOFMAN, A., ALIZADEH, B. Z., FRANCO, O. H. AND DEHGHAN, A . (2016). Bivariate genome-wide association study identifies novel pleiotropic loci for lipids and inflammation. *BMC Genomics* **17**, 443.

LIN, D. Y. AND ZENG, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology* **33**, 256–265.

LIN, D. Y., ZENG, D. AND TANG, Z. Z. (2013). Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 12247–12252.

LIU, D. J. AND LEAL, S. M. (2012). A unified method for detecting secondary trait associations with rare variants: application to sequence data. *PLoS Genetics* **8**, e1003075.

MA, L., YANG, J., BIRALI, R. H., TANAKA, T., FERRUCCI, L., BANDINELLI S., AND DA, Y. (2010). Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the framingham heart study data. *BMC Medical Genetics* **11**, 55.

MONSEES, G. M., TAMIMI, R. M. AND KRAFT, P. (2009). Genome-wide association scans for secondary traits using case-control samples. *Genetic Epidemiology* **33**(8), 717–728.

O'BYRNE, S., SHIRODARIA, C., MILLAR, T., STEVENS, C., BLAKE D. AND BENJAMIN N. (2000). Inhibition of platelet aggregation with glyceryl trinitrate and xanthine oxidoreductase. *Journal of Pharmacology and Experimental Therapeutics* **292**, 326–330.

OH, J. H., KIM, Y. K., MOON, S., KIM, Y. J. AND KIM, B. J. (2014). Genome-wide association study identifies candidate loci associated with platelet count in koreans. *Genomics & Informatics* **12**, 225–230.

PRINS, B. P., KUCHENBAECKER, K. B., BAO, Y., SMART, M., ZABANEH, D., FATEMIFAR, G., LUAN, J., WAREHAM, N. J., SCOTT, R. A., PERRY, J. R. B. *and others*. (2017). Genome-wide analysis of health-related biomarkers in the uk household longitudinal study reveals novel associations. *Scientific Reports* **7**, 11008.

RIDKER, P. M., PARE, G., PARKER, A., ZEE, R. Y., DANIK, J. S., BURING, J. E., KWIATKOWSKI, D., COOK, N. R., MILETICH, J. P., AND CHASMAN, D. I. (2008). Loci related to metabolic-syndrome pathways including lepr, hnf1a, il6r, and gckr associate with plasma C-reactive protein: the women's genome health study. *The American Journal of Human Genetics* **82**, 1185–1192.

SANDERS, S. J., MURTHA, M. T., GUPTA, A. R., MURDOCH, J. D., RAUBESON, M. J., WILLSEY, A. J., ERCAN-SENCICEK, A. G., DILULLO, N. M., PARIKSHAK, N. N., STEIN, J. L. *and others*. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237.

SANNA, S., JACKSON, A. U., NAGARAJA, R., WILLER, C. J., CHEN, W. M., BONNYCASTLE, L. L., SHEN, H., TIMPSON, N., LETTRE, G., USALA, G. *and others*. (2008). Common variants in the gdf5-uqcc region are associated with variation in human height. *Nature Genetics* **40**, 198.

SOLOVIEFF, N., MILTON, J. N., HARTLEY, S. W., SHERVA, R., SEBASTIANI, P., DWORKIS, D. A., KLINGS, E. S., FARRER, L. A., GARRETT, M. E., ASHLEY-KOCH, A. *and others*. (2010). Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5' olfactory receptor gene cluster. *Blood* **115**, 1815–1822.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Lango Allen, H., Lindgren, C. M., Luan, J., Mägi, R. *and others*. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42**, 937.

Spracklen, C. N., Chen, P., Kim, Y. J., Wang, X., Cai, H., Li, S., Long, J., Wu, Y., Xing Wang, Y., Takeuchi, F. *and others*. (2017). Association analyses of east asian individuals and trans-ancestry analyses with european individuals reveal new loci associated with cholesterol and triglyceride levels. *Human Molecular Genetics* **26**, 1770–1784.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello J. P., Ripatti, S., Chasman, D. I., Willer, C. J. *and others*. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707.

Wang, J. and Shete, S. (2011). Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genetic Epidemiology* **35**, 190–200.

Wang, K. (2016). Boosting the power of the sequence kernel association test by properly estimating its null distribution. *The American Journal of Human Genetics* **99**, 104–114.

Webb, T. R., Erdmann, J., Stirrups, K. E., Stitziel, N. O., Masca, N. G., Jansen, H., Kanoni S., Nelson, C. P., Ferrario, P. G., König, I. R. *and others*. (2017). Systematic evaluation of pleiotropy identifies 6 further loci associated with coronary artery disease. *Journal of the American College of Cardiology* **69**, 823–836.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S. *and others*. (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45**(11), 1274.

Zhu, Y., Zhang, D., Zhou, D., Li, Z., Li, Z., Fang, L., Yang, M., Shan, Z., Li, H., Chen, J. *and others*. (2017). Susceptibility loci for metabolic syndrome and metabolic components identified in han chinese: a multi-stage genome-wide association study. *Journal of Cellular and Molecular Medicine* **21**, 1106–1116.

[*Received 20 October 2017; revised 16 May 2018; accepted for publication 2 June 2018*]