Quantitative trait analysis in sequencing studies under trait-dependent sampling

Dan-Yu Lin[a,b,1], Donglin Zeng[a,b], and Zheng-Zheng Tang[a,b]

[a]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599; and [b]GO Exome Sequencing Project, National Heart, Lung, and Blood Institute, Bethesda, MD 20892

Edited by David O. Siegmund, Stanford University, Stanford, CA, and approved May 16, 2013 (received for review December 12, 2012)

It is not economically feasible to sequence all study subjects in a large cohort. A cost-effective strategy is to sequence only the subjects with the extreme values of a quantitative trait. In the National Heart, Lung, and Blood Institute Exome Sequencing Project, subjects with the highest or lowest values of body mass index, LDL, or blood pressure were selected for whole-exome sequencing. Failure to account for such trait-dependent sampling can cause severe inflation of type I error and substantial loss of power in quantitative trait analysis, especially when combining results from multiple studies with different selection criteria. We present valid and efficient statistical methods for association analysis of sequencing data under trait-dependent sampling. We pay special attention to gene-based analysis of rare variants. Our methods can be used to perform quantitative trait analysis not only for the trait that is used to select subjects for sequencing but for any other traits that are measured. For a particular trait of interest, our approach properly combines the association results from all studies with measurements of that trait. This meta-analysis is substantially more powerful than the analysis of any single study. By contrast, meta-analysis of standard linear regression results (ignoring trait-dependent sampling) can be less powerful than the analysis of a single study. The advantages of the proposed methods are demonstrated through simulation studies and the National Heart, Lung, and Blood Institute Exome Sequencing Project data. The methods are applicable to other types of genetic association studies and nongenetic studies.

Recent technological advances have made it possible to sequence genomic regions for association studies. At the present time, it is prohibitively expensive to perform large-scale whole-exome sequencing. In the near future, whole-exome sequencing on thousands of subjects will be economically feasible, but not whole-genome sequencing. If a quantitative trait is of primary interest in a large cohort study, a cost-effective strategy is to sequence those subjects with the extreme trait values preferentially. This strategy can substantially increase statistical power (relative to sequencing a random sample with the same number of subjects), as suggested by research in various contexts (1–9). Indeed, such trait-dependent sampling has been adopted in many sequencing projects, including the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) resequencing project. The NHLBI ESP consists of multiple studies, each of which is focused on one trait. For the body mass index (BMI) study, 267 subjects with BMI values >40 and 178 subjects with BMI values <25 were selected for sequencing out of a total of 11,468 subjects from Women's Health Initiative (WHI). Similar designs were used for the LDL and blood pressure (BP) studies, although the sampling was based on residuals (to adjust for age, sex, race, and medication) rather than raw measurements.

Case–control testing is a valid option for comparing the two extremes of a quantitative trait. If the underlying association is quantitative, however, case–control analysis will not be optimal for three major reasons. First, it is less powerful than quantitative trait analysis. Second, it does not estimate the quantitative relationship. Third, its results cannot be efficiently combined with those of other studies with different selection criteria.

In the absence of genetic association, the genetic variant is independent of the quantitative trait in the extremes (Fig. 1, *Upper*); therefore, standard linear regression, which ignores trait-dependent sampling, has correct type I error. In the presence of genetic association, however, standard linear regression will yield biased estimates of genetic effects (Fig. 1, *Lower*). Standard linear regression also yields biased estimates of the effects of confounders, such as ancestry variables for capturing population stratification (whether or not there is genetic association). Consequently, the type I error for testing genetic association will be inflated when there is population stratification (because the effects of ancestry variables are estimated with bias, and thus not correctly adjusted for).

Most sequencing projects, especially those derived from well-designed cohort studies, collect data on a variety of secondary quantitative traits (i.e., quantitative traits other than the one used to select subjects for sequencing). In the NHLBI ESP, a large number of secondary quantitative traits are available in each study. In particular, BMI and BP are available as secondary traits in the LDL study. Association analysis with available data on secondary traits is essentially a "free lunch." By combining the data on a particular trait that is the primary trait in one study and a secondary trait in another, we will have a larger sample size and higher statistical power. This is extremely important because there is little power to detect association with rare variants in small samples.

If the secondary trait is correlated with the primary trait, as is often the case, the genetic effects on the secondary trait may be distorted among the subjects with the extreme values of the primary trait. Thus, standard linear regression may yield biased estimates of the genetic effects on the secondary trait and cause inflation of the type I error. The directions and magnitudes of the bias may be different between the primary and secondary traits. Consequently, combining the results on a particular trait that is the primary trait in one study and a secondary trait in another study may actually reduce instead of increase power (as opposed to analyzing the data on the primary trait alone).

We propose valid and efficient likelihood-based methods for analyzing both primary and secondary quantitative traits and for combining data on a particular trait that is primary in one study and secondary in another under trait-dependent sampling. The newly developed approach, which is referred to as maximum likelihood estimation (MLE), preserves the type I error while achieving the highest power among all valid tests and yields unbiased and efficient estimates of genetic effects. We investigate the theoretical properties of the naive approach, namely, standard linear regression, under trait-dependent sampling. We compare the MLE and naive methods through extensive simulation studies. We provide applications to the NHLBI ESP data.

Author contributions: D.-Y.L. and D.Z. designed research; D.-Y.L., D.Z., and Z.-Z.T. performed research; D.Z. and Z.-Z.T. analyzed data; and D.-Y.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: lin@bios.unc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1221713110/-/DCSupplemental.

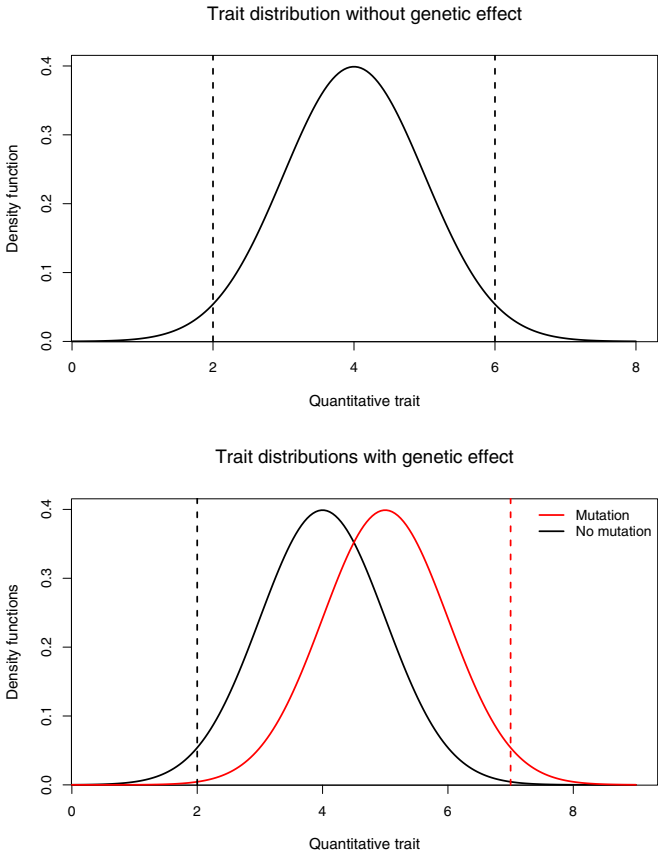

Fig. 1. Density functions of a quantitative trait in the absence (*Upper*) and the presence (*Lower*) of genetic association. In the absence of genetic association, there is no relationship between the trait value and the genetic mutation in each of the two tails. When the mutation tends to increase the trait value, most of the subjects in the right tail have the mutation, whereas most of the subjects in the left tail do not have the mutation; therefore, the difference in the trait distribution between the subjects with and without the mutation is larger in the two tails than in the general population.

Trait-dependent sampling without covariates was previously studied (1–8). It is important to accommodate covariates because population stratification is expected to be a severe issue and can be adjusted for through the use of ancestry covariates. The likelihood function reflecting trait-dependent sampling involves the distribution of covariates, which is high-dimensional in the presence of continuous covariates, and thus entails considerable theoretical and computational challenges. We establish the desired theoretical properties of the MLE through modern asymptotic techniques and develop the corresponding score statistics, which are computationally fast and numerically stable. We develop three types of gene-based tests for rare variants in sequencing studies. In addition, we theoretically investigate the efficiency of trait-dependent sampling and quantify the bias of standard linear regression as a function of the extremity of sampling and as a function of the effects of confounders. No such theoretical results were previously available (even without covariates).

The problem of analyzing secondary traits when the sampling is based on a quantitative trait has not been studied in the literature. We develop statistically efficient and numerically stable methods that properly account for the sampling in the analysis of secondary quantitative traits, paying special attention to testing rare variants in sequencing studies. We theoretically quantify the bias of standard linear regression as a function of the correlation between the primary and secondary traits and as a function of the genetic effect on the primary trait. We provide a meta-analysis method that efficiently combines the results on a quantitative trait that is the primary trait in one study and a secondary trait in another study. We demonstrate that our method is substantially more powerful than the meta-analysis based on standard linear regression.

Methods

We consider the type of design used in the NHLBI ESP, which consists of multiple studies. In each study, a quantitative trait of primary interest (e.g., BMI, LDL, BP) is used to select subjects for sequencing and measurements are available on other (i.e., secondary) quantitative traits. In the association analysis, a particular trait of interest may correspond to the primary trait in one study and to a secondary trait in another. In the NHLBI ESP, BMI is the primary trait in the BMI study and a secondary trait in the LDL and BP studies, whereas LDL is the primary trait in the LDL study and a secondary trait in the BMI and BP studies. In this section, we first show how to analyze primary and secondary quantitative traits in a single study and then show how to combine results on a particular trait that is primary in one study and secondary in another.

For a given study, let Y_1 denote the primary trait and Y_2 denote a secondary trait. (In general, Y_1 and Y_2 stand for different traits in different studies. If sampling is based on residuals rather than raw measurements, Y_1 and Y_2 are defined accordingly.) Also, let G denote the genetic variable of interest and Z denote a set of covariates. The latter may include ancestry variables and demographic/environmental factors. For single-variant analysis, G may denote the number of minor alleles the subject carries at the SNP site or indicate whether the subject carries any minor allele at the SNP site. For gene-based analysis, G may represent the total number of mutations over multiple variant sites within the gene or indicate whether there is any mutation within the gene (10–15).

Suppose that we have a cohort of n subjects, among whom n_1 subjects are selected for sequencing. We assume that the primary trait Y_1 is available on all n cohort members. (If there are missing values on Y_1, we define n as the total number of subjects with available Y_1.) The selection of subjects for sequencing may depend on the values of Y_1 in the entire cohort. By definition, G is available only on the n_1 sequenced subjects. If Z represents ancestry variables (e.g., percentage of African ancestry, principal components for ancestry) constructed from sequencing data, Z is available only on the n_1 sequenced subjects. If Z represents demographic/environmental variables, Z is potentially available on all cohort members. The secondary trait Y_2 is also potentially available on all cohort subjects. For a large cohort (or a study involving multiple large cohorts), however, it is logistically difficult and mathematically unnecessary to retrieve the records on covariates and secondary traits for nonsequenced subjects. Thus, we assume that Z and Y_2 are available only on the n_1 sequenced subjects. The values of Z and Y_2 may be missing among the sequenced subjects. We impute the missing values of Z by their sample means and leave the missing values of Y_2 unchanged.

Based on the above considerations, Y_1 is available on n subjects; (G, Z) is available on a subset of n_1 subjects; and Y_2 is available on a further subset of, say, n_2 subjects. We order the data so that the n_2 subjects with the available Y_2 measurements appear first and the remaining $(n_1 - n_2)$ sequenced subjects appear next. The observed-data likelihood can then be written as

$$\prod_{i=1}^{n_1} P(Y_{1i}|G_i,Z_i)P(G_i,Z_i) \prod_{i=n_1+1}^{n} \sum_{g,z} P(Y_{1i}|g,z)P(g,z) \prod_{i=1}^{n_2} P(Y_{2i}|Y_{1i},G_i,Z_i), \quad [1]$$

where P denotes the density or conditional density function.

It is natural to formulate the joint distribution of Y_1 and Y_2 through the bivariate linear regression model:

$$Y_1 = \beta_1 G + \gamma_1^T Z + \epsilon_1, \quad [2]$$

$$Y_2 = \beta_2 G + \gamma_2^T Z + \epsilon_2, \quad [3]$$

where (ϵ_1, ϵ_2) is bivariate normal with mean zero and covariance matrix $\{\sigma_{kl}; k, l = 1, 2\}$. We absorb the unit component in Z so that the first components of γ_1 and γ_2 pertain to the intercepts. The conditional distribution of Y_2, given (Y_1, G, Z), satisfies the linear model

$$Y_2 = \delta Y_1 + \tilde{\beta}_2 G + \tilde{\gamma}_2^T Z + \tilde{\epsilon}_2, \quad [4]$$

where $\delta = \sigma_{12}/\sigma_{11}$, $\tilde{\beta}_2 = \beta_2 - (\sigma_{12}/\sigma_{11})\beta_1$, $\tilde{\gamma}_2 = \gamma_2 - (\sigma_{12}/\sigma_{11})\gamma_1$, and $\tilde{\epsilon}_2$ is independent of ϵ_1 with mean zero and variance $\tilde{\sigma}_{22} = \sigma_{22} - \sigma_{12}^2/\sigma_{11}$.

In gene-based analysis, G pertains to aggregate information about the mutations within the gene (10–15). If G indicates, by the values 1 vs. 0, whether or not there is any mutation within the gene, β_1 and β_2 have simple

interpretations. If G is the total number of mutations within the gene, there is an implicit assumption that each mutation has the same effect on the quantitative trait. If G is a weighted sum of the mutation counts, β_1 and β_2 can only be interpreted at the aggregate level and the inference is focused on testing rather than estimation.

Note that Expression **1** is a nonparametric likelihood in that the (potentially high-dimensional) distribution of (G,Z) is not parametrized. In *SI Methods*, section A, we describe a computationally efficient and numerically stable expectation-maximization (EM) algorithm for maximizing Expression **1** and show that the resulting maximum likelihood estimators of β_1 and β_2 are approximately unbiased, normally distributed, and statistically efficient. The corresponding test statistics have correct type I error (at least when the sample size is large enough) and are more powerful than any other valid tests.

To make an inference about β_1, the naive method is to perform standard least-squares estimation under model Eq. **2**. This method has correct type I error if and only if there are no confounders. In the presence of genetic association, this method yields biased estimates of genetic effects, whether or not there are confounders, and the degree of bias depends on the extremity of sampling (*SI Methods*, section B).

To make an inference about β_2, the naive approach is to perform standard least-squares estimation under model Eq. **3** or model Eq. **4**. We refer to these two methods as naive-M and naive-C, respectively, where M and C stand for marginal and conditional, respectively. The naive-C method accounts for trait-dependent sampling because Y_1 is included as a covariate; however, $\hat{\beta}_2$ is not equal to β_2 unless $\beta_1=0$ or $\sigma_{12}=0$. Thus, the naive-C method has inflated type I error and biased estimation if the primary and secondary traits are correlated and there is a genetic effect on the primary trait. This conclusion also holds for the naive-M method. In addition, the naive-M method has inflated type I error in the presence of confounders even if there is no genetic effect on the primary trait (*SI Methods*, section B).

We may test the null hypotheses that $\beta_1=0$ and $\beta_2=0$ by using the score, Wald, or likelihood ratio statistics. All the results reported in this paper are based on score statistics, which are statistically more accurate and numerically more stable than Wald and likelihood statistics (13).

When the trait of interest in the association analysis is the primary trait in one study and a secondary trait in another, we perform appropriate analysis on that trait in each study and combine the results through meta-analysis. For example, suppose that we are interested in genetic association with LDL in the NHLBI ESP. In that case, we analyze LDL as the primary trait (i.e., Y_1) in the LDL study by calculating the score statistic for testing $H_0^{(1)}:\beta_1=0$ and analyze it as a secondary trait (i.e., Y_2) in the BMI and BP studies by calculating the score statistics for testing $H_0^{(2)}:\beta_2=0$. (We may perform association analysis on BMI or BP in a similar manner by redefining Y_1 and Y_2 in each study.) We then take the sum of the score statistics on the trait of interest over all the studies to produce an overall test statistic. Likewise, we obtain an overall estimate of the genetic effect on the trait of interest by applying the familiar inverse-variance weighting method to the parameter estimates and variance estimates for that trait from all the studies. This type of meta-analysis is equivalent to the joint analysis of the raw data of all the studies (16). The meta-analysis methods that combine the naive method for the primary trait with the naive-M and naive-C methods for the secondary trait are referred to as the naive-M′ and naive-C′ methods, respectively.

The power of the naive and MLE methods is theoretically investigated in *SI Methods*, section C. For detecting the genetic effect on the primary trait in the absence of a confounder, the naive and MLE methods have similar power. For detecting the genetic effect on the secondary trait when there is neither a confounder nor a genetic effect on the primary trait, the MLE and naive-C methods have similar power and are more powerful than the naive-M method. For combining results on a particular trait that is primary in one study and secondary in another, the MLE method tends to be much more powerful than the naive-M′ and naive-C′ methods. Indeed, the naive-M′ and naive-C′ methods can be less powerful than the analysis of one study only. The loss of power by the naive-M′ and naive-C′ methods is due to the fact that the naive estimates of the genetic effects have different magnitudes (or directions) of bias between the two studies (although the true genetic effects are the same between the two studies).

In gene-based analysis, the variants whose minor allele frequencies (MAFs) exceed a certain threshold may be excluded from the calculation of G. One can choose a fixed threshold, such as 1% or 5%, with the corresponding tests being called T1 and T5. Under the variable-threshold (VT) approach, one calculates the test statistics at all possible thresholds and chooses the threshold that minimizes the P value (12, 13). For the latter approach, it is necessary to account for the multiple testing within the gene. This can be accomplished by using the joint distribution of the test statistics, as described in the last section of *SI Methods*, section A. To detect variants with opposite effects on the trait, we extend the sequence kernel association test (SKAT) (14) to reflect trait-dependent sampling (last section of *SI Methods*, section A).

Results

Simulation Studies. We conducted extensive simulation studies to evaluate the performance of the MLE and naive methods in realistic situations mimicking the NHLBI ESP. We generated two quantitative traits from Eqs. **2** and **3** in which G is the total number of mutations in a gene consisting of 11 variants with MAFs $p_j=0.001j$ $(j=1,\ldots,10)$ and $p_{11}=0.04$ (13), Z is a normally distributed confounder (representing a principal component for ancestry or a different genetically related variable) with mean g conditional on $G=g$ and unit variance, and ϵ_1 and ϵ_2 are potentially correlated standard normal variables. (The variables G and Z have a Pearson correlation coefficient of ~0.38 or R^2 of ~0.14.) We generated a cohort of 10,000 subjects and retained the values of (G,Z,Y_2) for the 250 subjects with the smallest values of Y_1 and the 250 subjects with the largest values of Y_1. We assessed the bias and type I error for the MLE and naive methods. For making inference on β_1, we varied the value of γ_1; for making inference on β_2, we set $\gamma_1=\gamma_2=0$ and varied the values of β_1 and σ_{12}. The parameter values were chosen to reflect the ESP data. We set the nominal significance level α at 10^{-3} and used 1 million replicates.

The results for type I error rates are shown in Fig. 2 (*Left*). MLE has correct control of the type I error. For testing the genetic effect on the primary trait, the naive method has correct type I error in the absence of confounding but has inflated type I error when the effect of the confounder is strong. For testing the genetic effect on the secondary trait, the two naive methods have inflated type I error if the primary and secondary traits are correlated and there is a genetic effect on the primary trait; the inflation is much more severe for the naive-M method than for the naive-C method.

The results for bias are shown in Fig. 2 (*Right*). MLE is unbiased for estimating the genetic effects on both the primary and secondary traits. For estimating the genetic effect on the primary trait, the naive method can be severely biased whether or not the potential confounder has any effect on the trait. The bias is a nonlinear function of the effect of the confounder, which is consistent with the theory of *SI Methods*, section B. For estimating the genetic effect on the secondary trait, the two naive methods are biased when the primary and secondary traits are correlated and there is a genetic effect on the primary trait.

To investigate power, we generated two studies with the above design. The trait of interest was the primary trait in study I and the secondary trait in study II. This setup mimicked the situation where we are interested in BMI, which is the primary trait in the BMI study and a secondary trait in the LDL study. To make fair power comparisons, we considered the setting in which all competing methods have correct type I error. Specifically, we assumed that there is no confounding and that there is no genetic effect on the primary trait in study II, such that the naive methods for testing genetic association with the trait of interest have correct type I error in both study I and study II, and thus also have correct type I error in the meta-analysis. (The MLE methods always have correct type I error.)

Fig. 3 displays the power curves of the MLE and naive methods for detecting the genetic effect on the trait of interest. When the genetic effect is zero, the power (i.e., type I error) of all the methods is indeed near the nominal significance level. When the genetic effect is nonzero, MLE is slightly more powerful than the naive method in study I. In study II, MLE and the naive-C method have similar power, whereas the naive-M method has lower power. In the meta-analysis, MLE is substantially more powerful than both the naive-M′ and naive-C′ methods. Indeed, the two naive meta-analysis methods are less powerful than the naive method for analyzing the primary trait in study I. The loss of power is due to different degrees of bias: The naive estimate for the primary trait in study I is severely biased upward, whereas the two naive estimates, naive-M and naive-C,

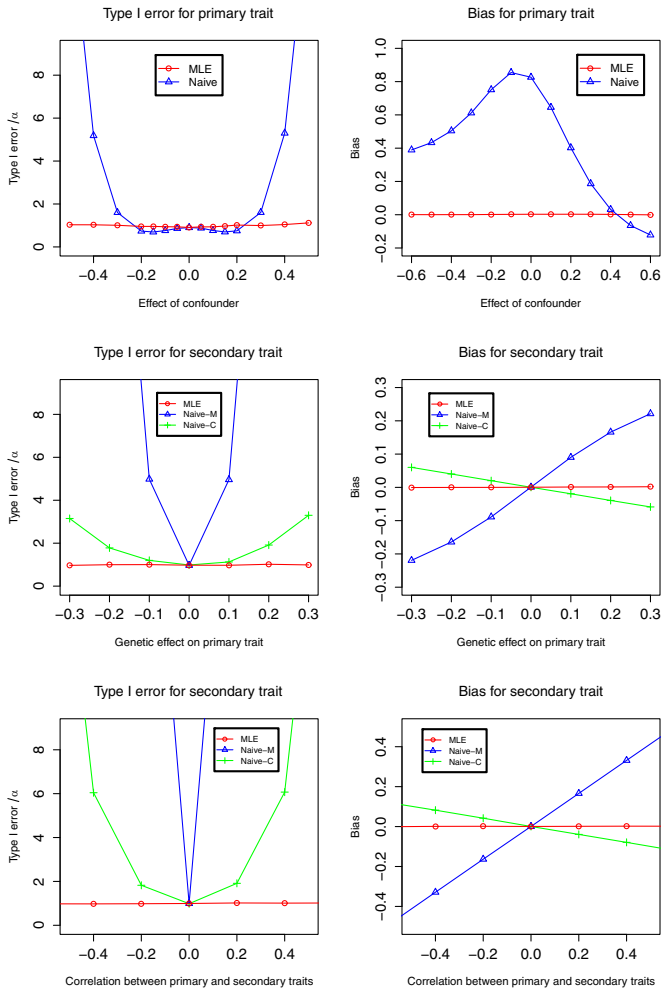

Fig. 2. Type I error (divided by the nominal significance level α) in testing no genetic association and bias in estimating the genetic effect for the MLE and naive methods. (*Top*) Results for the primary trait as a function of the effect of a confounder. (*Middle*) Results for the secondary trait as a function of the genetic effect on the primary trait when the correlation between the primary and secondary traits is 0.2. (*Bottom*) Results for the secondary trait as a function of the correlation between the primary and secondary traits when the genetic effect on the primary trait is 0.2. (*Right*) For the bias estimates, the true effects are 0.2.

for the secondary trait in study II are unbiased (in our simulation setup). For Fig. 3 (*Right*), the bias of the naive estimate for the primary trait in study I is about 0.83, which is more than fourfold the effect size, whereas the naive estimates for the secondary trait in study II are virtually unbiased. As shown in *SI Methods*, section C, meta-analysis of estimates with different degrees of bias can reduce power.

We also evaluated the MLE and naive versions of the VT test in simulation studies. We simulated data in the same manner as before but performed the association test by maximizing the absolute value of the test statistic over the observed MAF thresholds (and accounting for the multiple testing). The MLE approach continues to outperform the naive methods (Fig. S1).

The CHARGE resequencing project adopted a one-tailed sampling design by selecting subjects with the highest values of a quantitative trait, along with a random sample. Our general framework covers this scenario. We conducted a series of simulation studies mimicking the CHARGE design. Specifically, we generated a cohort of 12,000 subjects in the same manner as in the previous simulation studies but selected the 200 subjects with the highest values of Y_1 and a random sample of 1,000 subjects (rather than 250 subjects with the highest values of Y_1 and 250 subjects with the lowest values of Y_1). This sampling is much less extreme than the two-tailed sampling used in the previous simulation studies because only the right tail is preferentially sampled and the sample from the right tail is much smaller than the random sample. The results analogous to Figs. 2 and 3 are displayed in Figs. S2 and S3. The MLE methods continue to perform well. Because the sampling is much less extreme than before (i.e., the two-tailed sampling), the naive methods perform differently: (*i*) the naive methods are less biased than before, especially for the primary trait; (*ii*) the naive-C method is more biased than the naive-M method; and (*iii*) the loss of power for the naive meta-analysis (relative to the MLE meta-analysis) is less severe than before.

To assess the robustness of the methods to the normality assumption, we simulated data in the same manner as in the first series of simulation studies but set ϵ_1 and ϵ_2 to $F^{-1}(\Phi(\epsilon_1^*))$ and $F^{-1}(\Phi(\epsilon_2^*))$, respectively, where $(\epsilon_1^*, \epsilon_2^*)$ is bivariate normal with means 0, variances 1, and correlation 0.2; F is the distribution function of a t random variable; and Φ is the distribution function of a standard normal random variable. The results are shown in Figs. S4 and S5. Both the MLE and naive tests have inflated type I error when the degree of freedom is low and have appropriate type I error when the degree of freedom is high. Even under random sampling, the least-squares methods have inflated

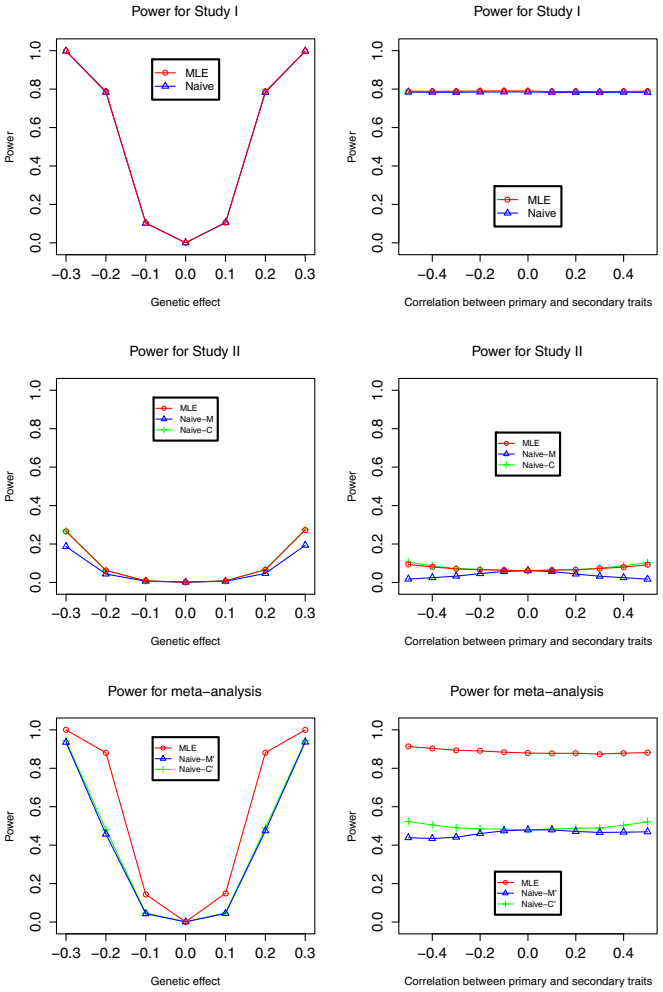

Fig. 3. Power of the MLE and naive methods in detecting genetic association in study I, study II, and the meta-analysis. (*Left*) Power as a function of the genetic effect on the trait of interest when the correlation between the primary and secondary traits in study II is 0.2. (*Right*) Power as a function of the correlation between the primary and secondary traits in study II when the genetic effect on the trait of interest is 0.2, corresponding to R^2 of 6%.

type I error when the degree of freedom is low. In this sense, the MLE tests under trait-dependent sampling are as robust to the normality assumption as the least-squares methods under random sampling. In the presence of genetic effects, the MLE methods yield noticeable bias, especially for the primary trait, but the bias is considerably smaller than that of the naive methods and decreases as the degree of freedom increases.

NHLBI ESP. The NHLBI ESP is a signature project of the National Institutes of Health (NIH) Recovery Act investment. It was designed to identify genetic variants in all protein-coding regions of the human genome that are associated with heart, lung, and blood diseases. The project consists of multiple studies, each of which is focused on one trait. There are three studies that sequenced the subjects with extreme values of a quantitative trait (i.e., BMI, LDL, BP), one case–control study on myocardial infarction (MI), and one case-only study on stroke. As mentioned in the Introduction, 267 subjects with BMI values >40 and 178 subjects with BMI values <25 were selected for sequencing out of a total of 11,468 subjects from the WHI. For the LDL study, 120 subjects with the highest LDL values and 120 subjects with the lowest LDL values were selected out of ∼22,000 European Americans from four cohorts: the Atherosclerosis Risk in Communities (ARIC) study, Cardiovascular Health Study (CHS), Framingham Heart Study (FHS), and Jackson Heart Study (JHS); likewise, 120 subjects with the highest LDL values and 120 subjects with the lowest LDL values were selected out of ∼7,000 African Americans from the same cohorts. The LDL values were adjusted for age, sex, and lipid medication, and the adjusted LDL values for the selected subjects were lower than the first percentile and greater than the 99th percentile for European ancestry and lower than the third percentile and greater than the 97th percentile for African ancestry. For the BP study, 850 subjects were selected from the upper and lower 0.2–1.0% of the BP distribution (adjusted for sex, age, race, BMI, and antihypertensive treatment) out of ∼100,000 European Americans and ∼20,000 African Americans from seven cohorts: the ARIC study, Cardiovascular Risk in Communities Study (CARDIA), CHS, FHS, JHS, Multiethnic Study of Atherosclerosis (MESA), and WHI. The MI study consists of 650 cases with early MI and 650 controls free of MI drawn from the ARIC study, CHS, FHS, JHS, and WHI. The stroke study consists of 600 subjects with ischemic stroke from the ARIC study, CHS, FHS, MESA, and WHI. In addition to the above five studies, there is a random sample of 1,000 European Americans and 500 African Americans from the CARDIA, CHS, FHS, JHS, MESA, and WHI who have a predefined common set of core variables (i.e., phenotypes, traits), which is referred to as a deeply phenotyped reference (DPR). Although each cohort was involved in multiple studies, no subject was selected into more than one study. The ages of the subjects range between 50 and 79 y, 20 and 88 y, 12 and 69 y (<20 y for 4 subjects), 20 and 92 y, 19 and 84 y, and 24 and 87 y in the BMI, LDL, BP, MI, stroke, and DPR studies, respectively. The traits were measured in the same way in all studies. Exome sequencing was performed at the University of Washington and the Broad Institute using the Roche NimbleGen SeqCap EZ or Agilent SureSelect Human All Exon 50 Mb (17).

For illustration, we considered LDL as the trait of interest, which is the primary trait in the LDL study and a secondary trait in the BMI, BP, MI, and stroke studies. We used both the MLE and naive methods to analyze the LDL trait in the LDL, BMI, and BP studies and performed standard linear regression analysis of LDL in the MI study (adjusted for the MI status), stroke study, and DPR. [Both early MI and ischemic stroke are relatively rare. For a case–control or case-only study with rare disease, standard linear regression analysis of secondary quantitative traits conditional on the disease status yields approximately correct results (18).] After restriction to subjects with available sequencing data and exclusion of subjects with sex mismatch or relatedness, there are 412, 489, 656, 839, 428, and 699 subjects in the LDL, BMI, BP, MI, stroke, and DPR studies, respectively. The corresponding numbers of subjects with nonmissing LDL values are 412, 72, 351, 617, 61, and 457. Variants for all studies were called jointly at the University of Michigan. To ensure high-quality genotype calls for the analysis, the individual genotype values were set to missing if the genotype depth was lower than 10. To reduce the number of variant calls due to sequencing and alignment artifacts, a support vector machine was used to separate likely true-positive results from likely artifacts by using a variety of SNP quality metrics, including allelic balance (the proportional representation of each allele in likely heterozygotes), base quality distribution for sites supporting the reference and alternate alleles, and distribution of supporting evidence between strands and sequencing cycle. After implementation of these quality-control filters, there are a total of 115,515 SNPs with call rates >90% and MAFs ≥0.5%. Approximately 60% of the SNPs are in exons. We focused on single-variant analysis under the additive mode of inheritance and included the top five principal components for ancestry, cohorts, and sequencing centers/targets as covariates. The natural logarithm was applied to the LDL and BMI values.

Fig. S6 displays the quantile–quantile plots for the MLE analysis of the LDL study and for the meta-analysis of the six studies based on the MLE, naive-M′, and naive-C′ methods. For the MLE analysis of the LDL study and the MLE meta-analysis, the observed P values agree very well with the global null hypothesis of no association, except at the extreme right tails. By contrast, the observed P values of the naive-M′ and naive-C′ methods deviate substantially from the null distribution, reflecting excessive false-positive results. The inflation is likely due to two sources: (*i*) some SNPs exhibit strong population stratification, and (*ii*) some SNPs are associated with BMI or BP,

Table 1. MLE and naive P-values in the analysis of the LDL study, in the meta-analysis of the other (i.e., BMI, BP, MI, Stroke and DPR) studies, and in the meta-analysis of all six studies for the top 10 SNPs from the LDL study

		MLE			Naive				
					marginal			conditional	
rank of SNP	MAF	LDL	others	all	LDL	others	all	others	all
1	0.053	4.11e-07	1.92e-08	7.18e-13	2.27e-07	2.93e-08	7.77e-11	1.87e-08	4.72e-11
2	0.162	3.06e-06	3.60e-02	5.00e-07	3.13e-06	3.68e-02	4.35e-03	3.74e-02	4.48e-03
3	0.123	9.71e-06	3.36e-01	6.19e-05	1.01e-05	2.78e-01	8.77e-02	3.46e-01	1.17e-01
4	0.123	1.92e-05	3.36e-01	9.86e-05	1.85e-05	2.78e-01	9.11e-02	3.46e-01	1.22e-01
5	0.178	2.93e-05	3.26e-01	1.24e-04	2.32e-05	3.05e-01	1.04e-01	3.26e-01	1.14e-01
6	0.122	3.28e-05	4.23e-01	2.28e-04	3.64e-05	3.74e-01	1.42e-01	4.35e-01	1.74e-01
7	0.132	3.46e-05	8.55e-01	8.72e-04	5.65e-05	7.94e-01	3.93e-01	8.73e-01	4.51e-01
8	0.197	3.51e-05	6.62e-01	2.25e-03	7.38e-05	6.81e-01	8.17e-01	6.64e-01	8.37e-01
9	0.080	3.69e-05	5.73e-01	1.69e-03	3.00e-05	5.39e-01	8.81e-01	5.82e-01	8.34e-01
10	0.210	3.97e-05	7.43e-01	8.79e-04	3.79e-05	7.35e-01	3.64e-01	7.46e-01	3.72e-01

both of which are correlated with LDL. There is good concordance between the MLE analysis of the LDL study and the MLE meta-analysis: The top two SNPs are the same between the two analyses, although the meta-analysis yields more extreme P values, especially for the top SNP, and the third, fourth, and fifth SNPs in the LDL study are the 11th, 15th, and 22th SNPs in the meta-analysis. The results from the naive-M′ and naive-C′ methods are very similar to each other but are quite different from those of the MLE meta-analysis: The lists of the top 20 SNPs based on the naive-M′ and naive-C′ methods have 18 overlaps but only have 5 overlaps with the MLE list.

Table 1 shows the MLE and naive P values in the analysis of the LDL study, in the meta-analysis of the other studies (i.e., BMI, BP, MI, stroke, DPR), and in the meta-analysis of all six studies for the top 10 SNPs from the MLE analysis of the LDL study. For those 10 SNPs, the MLE P values in the meta-analysis of the other studies are similar to their naive counterparts. However, the P values for the MLE meta-analysis of the six studies are much more significant than those of the naive-M′ and naive-C′ methods. Indeed, all the MLE P values are less than 0.01, whereas only the top 2 SNPs using the naive-M′ and naive-C′ methods have P values <0.01.

The forest plots shown in Fig. S7 help to explain the results of Table 1. The MLE estimate for the LDL study is similar to those of the BP, MI, and DPR studies. (There are very few subjects with available LDL measurements in the BMI and stroke studies; thus, the estimates in those two studies are associated with very high variabilities.) Thus, the estimate of the MLE meta-analysis is similar to the MLE estimate of the LDL study but with a smaller SE. The naive estimate for the LDL study is sevenfold larger than the MLE estimate, with a SE that is also sevenfold larger. Due to the extreme trait-dependent sampling, the naive estimate is expected to have this magnitude of bias. Because the naive estimate in the LDL study is severely biased, whereas the estimates in the other studies are roughly unbiased, the naive-M′ and naive-C′ methods yield less significant results than the analysis of the LDL study alone. This phenomenon is consistent with the theoretical analysis (*SI Methods*, section C) and simulation results (Fig. 3).

We also performed gene-based association tests on rare variants. We considered polymorphic variants that are nonsynonymous, stop-gain, stop-loss, or splicing mutations according to the ANNOVAR (functional annotation of genetic variants from high-throughput sequencing data) annotation. We excluded any gene whose total number of mutations is fewer than four and ended up with a total of 16,167 genes. There were a total of 632,003 variants in these genes. For each gene, we defined G as the total number of mutations and applied both the MLE and naive versions of the T1, T5, Madsen–Browning (MB) (11), and VT tests, and the SKAT. For the MB test, each mutation is weighted by the inverse SD of its frequency. The results are displayed in Figs. S8–S12. The conclusions regarding the performance of the MLE and naive tests are similar to those of the single-variant analysis.

Discussion

Trait-dependent sampling provides a cost-effective strategy to conduct sequencing studies of quantitative traits. Failure to account for the biased nature of the sampling can yield gross inflation of type I error and severe loss of power, especially in meta-analysis. Indeed, meta-analysis of standard linear regression results can be less powerful than the analysis of a single study, as shown in our theoretical analysis, simulation studies, and real data. The MLE methods presented in this paper maximize statistical power while preserving type I error. The corresponding numerical algorithms are stable and fast. It took ∼10 s on an IBM HS21 machine to perform one association test for an ESP study.

Case–control sampling is also a form of trait-dependent sampling in that the sampling is based on the disease status. The type of trait-dependent sampling studied in this paper differs from case–control sampling in that the trait is continuous rather than binary. It is well known that case–control sampling can be ignored in the logistic regression analysis of case–control data (19). By contrast, trait-dependent sampling on a quantitative trait cannot be ignored in the linear regression analysis, as demonstrated in this paper, although odds ratio parameters are unaffected (9). There exist MLE methods for analyzing secondary traits in case–control studies (18, 20). If the selection probabilities of cases and controls are known, simple weighting methods (21) can also be used, although they are not as efficient as MLE methods. Weighting methods cannot be applied to the ESP LDL, BMI, or BP study because the subjects with nonextreme trait values had zero probabilities of being selected.

We have focused on secondary quantitative traits. In the NHLBI ESP, investigators are interested in secondary binary traits (e.g., type I diabetes) and longitudinal traits (e.g., diastolic and systolic blood pressures). We are currently extending the MLE methods to such traits.

ACKNOWLEDGMENTS. We thank G. R. Abecasis, P. L. Auer, C. Bizon, N. Franceschini, Y. Hu, E. M. Lange, L. A. Lange, K. North, S. S. Rich, R. Tao, R. P. Tracy, and C. J. Willer for discussions/comments. We thank K.-P. Li for programming assistance and two referees for helpful suggestions. We acknowledge the support of the NHLBI and the contributions of the research institutions, study investigators, field staff, and study participants in creating the ESP data for biomedical research. Additional acknowledgments are provided in *SI Appendix*. This research was supported by the NIH Grants R01 CA082659, R37 GM047845, and P01 CA142538. Funding for ESP was provided by NHLBI Grants RC2 HL-103010 (Heart), RC2 HL-102923 (Lung), and RC2 HL-102924 (WHI Sequencing Project). The exome sequencing was performed through NHLBI Grants RC2 HL-102925 (Broad) and RC2 HL-102926 (Seattle).

1. Allison DB (1997) Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet* 60(3):676–690.
2. Page GP, Amos CI (1999) Comparison of linkage-disequilibrium methods for localization of genes influencing quantitative traits in humans. *Am J Hum Genet* 64(4):1194–1205.
3. Slatkin M (1999) Disequilibrium mapping of a quantitative-trait locus in an expanding population. *Am J Hum Genet* 64(6):1764–1772.
4. Xiong M, Fan R, Jin L (2002) Linkage disequilibrium mapping of quantitative trait loci under truncation selection. *Hum Hered* 53(3):158–172.
5. Chen Z, Zheng G, Ghosh K, Li Z (2005) Linkage disequilibrium mapping of quantitative-trait Loci by selective genotyping. *Am J Hum Genet* 77(4):661–669.
6. Wallace C, Chapman JM, Clayton DG (2006) Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *Am J Hum Genet* 78(3):498–504.
7. Huang BE, Lin DY (2007) Efficient association mapping of quantitative trait loci with selective genotyping. *Am J Hum Genet* 80(3):567–576.
8. Li D, Lewinger JP, Gauderman WJ, Murcray CE, Conti D (2011) Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genet Epidemiol* 35(8):790–799.
9. Chen HY, Li M (2011) Improving power and robustness for detecting genetic association with extreme-value sampling design. *Genet Epidemiol* 35(8):823–830.
10. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am J Hum Genet* 83(3):311–321.
11. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5(2):e1000384.
12. Price AL, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86(6):832–838.
13. Lin DY, Tang ZZ (2011) A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet* 89(3):354–367.
14. Wu MC, et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89(1):82–93.
15. Tzeng JY, et al. (2011) Studying gene and gene-environment effects of uncommon and common variants on continuous traits: A marker-set approach using gene-trait similarity regression. *Am J Hum Genet* 89(2):277–288.
16. Lin DY, Zeng D (2010) On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* 97(2):321–332.
17. Tennessen JA, et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64–69.
18. Lin DY, Zeng D (2009) Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol* 33(3):256–265.
19. Prentice RL, Pyke R (1979) Logistic disease incidence models and case-control studies. *Biometrika* 66(3):403–411.
20. He J, Li H, Edmondson AC, Rader DJ, Li M (2012) A Gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics* 13(3):497–508.
21. Monsees GM, Tamimi RM, Kraft P (2009) Genome-wide association scans for secondary traits using case-control samples. *Genet Epidemiol* 33(8):717–728.