European Journal of Human Genetics (2012) 20, 449–456

www.nature.com/ejhg

ARTICLE

A flexible likelihood framework for detecting associations with secondary phenotypes in genetic studies using selected samples: application to sequence data

Dajiang J Liu[1,2] and Suzanne M Leal*[,1,2]

For most complex trait association studies using next-generation sequencing, in addition to the primary phenotype of interest, many clinically important secondary traits are also available, which can be analyzed to map susceptibility genes. Owing to high sequencing costs, most studies use selected samples, and the sampling mechanisms of these studies can be complicated. When the primary and secondary traits are correlated, analyses of secondary phenotypes can cause spurious associations in selected samples and existing methods are inadequate to adjust for them. To address this problem, a likelihood-based method, MULTI-TRAIT-ASSOCIATION (MTA) was developed. MTA is flexible and can be applied to any study with known sampling mechanisms. It also allows efficient inferences of genetic parameters. To investigate the power of MTA and different study designs, extensive simulations were performed under rigorous population genetic and phenotypic models. It is demonstrated that there are great benefits for analyzing secondary phenotypes in selected samples. In particular, using case–control samples and samples with extreme primary phenotypes can be more powerful than analyzing random samples of equivalent size. One major challenge for sequence-based association studies is that most data sets are not of sufficient size to be adequately powered. By applying MTA, data sets ascertained under distinct mechanisms or targeted at different primary traits can be jointly analyzed to map common phenotypes and greatly increase power. The combined analysis can be performed using freely available data sets from public repositories, for example, dbGaP. In conclusion, MTA will have an important role in dissecting the etiology of complex traits.
European Journal of Human Genetics (2012) **20,** 449–456; doi:10.1038/ejhg.2011.211; published online 14 December 2011

Keywords: multiple phenotypes; next-generation sequencing; rare variants; pleiotropy; secondary trait; selective sampling

INTRODUCTION

There is solid evidence that complex traits can be influenced by rare variants.[1–4] The development and large-scale application of next-generation sequencing have already revolutionized genetic studies and enabled detecting associations with complex traits owing to rare variants. In order to design powerful studies, it is necessary to deeply sequence samples from a large number of individuals.[5] However, many existing studies are small- to moderate-sized, owing to the high cost of sequencing or limited availability of samples, and are therefore inadequately powered. It would be advantageous if different studies, which measure the same phenotypes, could be jointly analyzed to increase power. In particular, many clinically important traits, such as body mass index (BMI), systolic (SysBP) and diastolic blood pressure (DiasBP) are often measured in different studies. When combined analysis is performed, in addition to incorporating studies that are targeted at the same primary traits, it is desirable to also analyze data from studies for which the phenotype of interest is measured as an additional outcome. Combined analyses require modeling multiple phenotypes as different studies may sequence selected samples targeted at different primary traits. Similar to the idea of analysis of covariance (ANCOVA), jointly analyzing multiple phenotypes makes it possible to distinguish the phenotype covariance component that is due to gene pleiotropy and the component that is attributable to residual correlations.

Currently, most studies sequence selected samples, for example, case–control samples or individuals with extreme phenotypes.[1,6] Sequencing selected samples reduces sequencing cost and improves power. Owing to sample ascertainment, secondary traits can be associated with the gene region in a selected sample even though they are independent in the general population. For example, consider a gene that is associated with the primary trait, but not with the secondary trait, in the general population (Figure 1). In a sample that consists of individuals with extreme primary trait values, the causative variant frequency will be different between individuals from the upper and lower extremes. The mean value for the secondary trait will also be different owing to phenotypic correlations. Therefore, a spurious association can occur between the gene region and the secondary trait unless the sample ascertainment scheme is correctly modeled. The selection criteria for a sequencing study can be complicated and may involve multiple traits (multiple-trait study) or sub-phenotypes. For instance, it is hypothesized that the etiologies of type-2 diabetes (T2D) are different in obese and non-obese individuals.[7,8] In order to

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA; [2]Department of Statistics, Rice University, Houston, TX, USA
*Correspondence: Dr SM Leal, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. Tel: 71 37 98 4011; Fax: 71 37 98 4012; E-mail: sleal@bcm.edu
Received 17 May 2011; revised 17 October 2011; accepted 20 October 2011; published online 14 December 2011

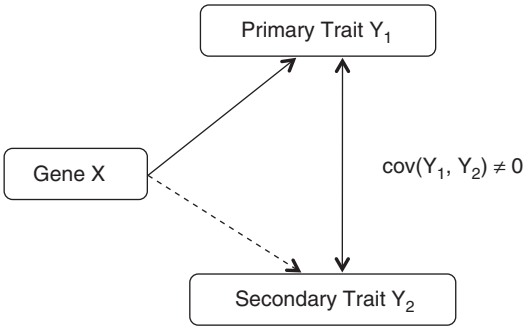

Figure 1 Graphical illustration of gene/multiple phenotypes associations. The gene region is causal for the primary trait Y_1 but not for the secondary trait Y_2. Owing to the correlation between the two traits, a spurious association can be detected between the gene region and the secondary trait if the ascertainment mechanism or phenotypic correlations are not properly modeled.

reduce phenotype heterogeneity and potentially improve power, a study of T2D might be performed using an obese population. There have been methods developed for detecting associations with multiple phenotypes in selected samples.[9,10] However, these methods are limited to case–control studies. They are not applicable to more complicated study designs, for example, studies that sequence individuals with extreme primary traits (extreme-trait study), or studies where secondary phenotypes are also involved in sample selection. In particular, the extreme-trait study design is becoming increasingly popular and widely applied.[11–13] The results for detecting associations with secondary traits can be seriously biased if the secondary traits are not properly analyzed.[9] It is desirable to have a unified approach for analyzing secondary phenotypes from all available data sets.

A flexible likelihood approach, MULTI-TRAIT-ASSOCIATION (MTA), is presented for detecting associations with multiple phenotypes in selected or randomly ascertained samples. This method can be used to detect both common and rare variant/secondary phenotype associations. MTA jointly models multiple phenotypes conditional on the study subjects being ascertained. The sampling mechanisms are incorporated by means of a prospective likelihood approach. The MTA framework is comprehensive and can be used to model multiple continuous or categorical traits. To model traits that are not continuous, a generalized linear model is used. For example, either a probit or logit link function can be applied to model binary traits. In this article, the discussion is focused on using the probit link function and the liability threshold model, which can be justified by the polygenic model of complex traits. It has been suggested that the liability of all complex traits can be considered as 'quantitative'.[14] For complex traits that are not measured in a quantitative scale, there should exist a continuous underlying liability trait, which is due to the aggregated outcome from multiple causative variants with small effects. In this case, a multivariate liability threshold model is naturally used to jointly model multiple phenotypes.

The power of MTA for detecting gene/secondary trait associations is examined in different selective study designs. Three study designs are considered, that is, case–control, extreme-trait and multiple-trait. It is assumed for each of the study designs that the same continuous secondary phenotype T is measured. For comparison purposes, study designs are also evaluated where the quantitative trait T is selected and analyzed as the primary phenotype. Simulation details for each study design can be found in Table 1.

It is very beneficial to be able to use and combine selected samples from existing sequencing-based genetic studies. Through extensive simulation studies, it is shown that the case–control and extreme-trait designs can be more powerful for detecting associations with secondary phenotypes than using a population-based design, where individuals are randomly selected regardless of their phenotypes. The power for detecting associations with secondary phenotypes strongly depends on the aggregation of causative variants in the sample. For study designs that facilitate enrichment of causative variants, power will be increased. In the presence of gene pleiotropy, variants that are associated with both the primary and secondary traits can be enriched through selections on the primary phenotype. When the gene region is only associated with the secondary phenotype, if the primary and secondary traits are correlated, selections on the primary phenotype can also induce selections on the secondary phenotype. In this case, for a sample of equivalent size, the power of rejecting the null hypothesis of no gene/secondary trait association in case–control or extreme-trait studies is still superior or comparable to a population-based study.

The power for detecting associations with secondary phenotypes in selected samples is jointly affected by locus phenotypic effects for both the primary and secondary phenotypes, as well as residual correlations. Concordant with observations from previous studies of multiple-trait linkage/association mapping,[15–17] it is demonstrated that power is maximized when the locus-induced trait correlations are in the opposite direction of the residual correlations. To further demonstrate the utility of MTA in combined analysis, an example is given where samples from a case–control study and a multiple-trait study are jointly analyzed. The power for detecting associations with commonly measured phenotypes can be greatly increased when studies are combined, compared with analyzing each individual study separately.

As an application of MTA, we analyzed the sequence data from the *ANGPTL3*, *ANGPTL4, ANGPTL5* and *ANGPTL6* genes generated by the Dallas Heart Study (DHS). The 3551 study participants of the DHS were phenotyped for multiple metabolism-related traits, including BMI, DiasBP, SysBP, total cholesterol level (TCL), low-density

Table 1 Definitions of selection mechanisms

Study designs	*Definition*
Case–control	Cases and controls are sampled based on the binary primary phenotype A. The trait status is determined by $A=\delta(A^* \geq a^c)$, where a^c is the 90th percentile of the liability trait A^*. A total of 500 cases and 500 controls is sequenced
Extreme-trait	One thousand individuals with quantitative trait B values in the upper and lower 10% were sequenced from a cohort of 5000 individuals
Multiple-trait	The affection status is defined by $C=\delta(C^* \geq c^c)$, where c^c is the 90th percentile for the liability trait C^*. Five hundred affected individuals with trait T-values >65th percentile are sequenced and 500 unaffected individuals are also sequenced regardless of their T-values
Extreme-trait study where T is sampled and analyzed as primary trait	In an extreme-trait study, individuals with extreme T-values in the upper and lower 2, 6 and 10% are sampled, and sequenced from a cohort of 5000 individuals
Population-based study design	A total of 1000, 2000 and 3000 individuals are randomly sampled from the general population regardless of their phenotypes

lipoprotein (LDL), high-density lipoprotein (HDL), triglyceride (TG) and glucose (Gluc). Two primary trait analyses were first performed: (1) analysis of all samples and (2) analysis of selected samples whose quantitative trait values fall within the lower and upper quartiles. Next a secondary phenotype analysis was performed where within each selected sample all other available phenotypes were analyzed as secondary traits. The results from the secondary trait analyses confirmed the primary trait analyses. These results established the importance of analyzing secondary phenotypes and the effectiveness of MTA. They provided solid support to our simulation experiment.

MATERIALS AND METHODS

It is assumed that there are S variant nucleotide sites for a gene locus. The multi-site genotype for individual i is given by $\vec{X}_i = \left(x_i^1, x_i^2, \cdots, x_i^S\right)$, where the genotype at the segregating nucleotide site s is coded by the number of minor alleles, (eg, $x_i^s = 2$ if the individual is homozygous for the minor allele). To detect associations with rare variants, multiple rare variants in a gene locus are usually jointly analyzed.[18–22] The locus genotype coding for an individual i is defined as $X_i = C\left(\vec{X}_i\right)$, where $C(\bullet)$ is the coding function.

Locus multi-site genotype coding schemes

Many statistical methods have been developed for association studies of complex traits owing to rare variants. Existing methods include combined multivariate and collapsing (CMC),[23] test of an aggregated number of rare variants (ANRV),[22] weighted sum statistics (WSS),[20] variable threshold test (VT),[21] kernel-based adaptive cluster (KBAC),[19] the C-alpha test[24] and the RARECOVER (RC) method,[25] and so on. Most of these methods are essentially based on weighting or grouping variants. Among them, CMC and ANRV are regression-based methods, which can be incorporated into MTA through the coding function $C(\bullet)$:

(1) CMC: The coding function is defined as $X_i = C^{CMC}\left(\vec{X}_i\right) = \delta\left(\sum_{s\in RV} x_i^s > 0\right)$, where $\delta(\bullet)$ is an indicator function and $\sum_{s\in RV}$ is a summation over the set of rare variant nucleotide sites RV, which can be determined by a pre-specified frequency cut-off.

(2) ANRV: The coding function belongs to a more general class of weighted sum coding (WSC), which can be defined as $X_i = C^{WSC}\left(\vec{X}_i\right) = \sum_{s\in RV} w^s x_i^s$. In the WSC scheme, the variant from nucleotide site s is assigned weight w^s. The ANRV coding assigns equal weight for all variants, that is, $X_i = C^{ANRV}\left(\vec{X}_i\right) = \sum_{s\in RV} x_i^s$.

A general probability model for multiple phenotypes in selected samples

In order to incorporate the sample ascertainment mechanism and correct for the bias induced by phenotypic residual correlations, multiple phenotypes are jointly modeled. The primary and secondary traits are assumed to follow a multivariate generalized linear model:

$$\begin{cases} F_{Y_1}\left(\vec{\theta}_{Y_1}\right) = \beta_{01} + \beta_{11} X_i + \sum_k \alpha_{k1} W_{ki} \\ F_{Y_2}\left(\vec{\theta}_{Y_2}\right) = \beta_{02} + \beta_{12} X_i + \beta_{Y_1} Y_{1i} + \sum_k \alpha_{k2} W_{ki} \end{cases} \quad (1)$$

$F_{Y_1}\left(\vec{\theta}_{Y_1}\right)$ and $F_{Y_2}\left(\vec{\theta}_{Y_2}\right)$ are link functions, and $\vec{\theta}_{Y_1}$ and $\vec{\theta}_{Y_2}$ are the model parameters related to the primary and secondary traits. This multivariate generalized linear model can be used with any type of link functions, such as probit link function or logit link function.

For selected samples, a conditional likelihood is used, which is similar to Pearson–Aitken correction for ascertainment:[26]

$$L(\beta, \theta; X, \Upsilon) = \prod_{i=1}^{N} \Pr\left(Y_{1i}, Y_{2i} | Z_i = 1, X_i, \{W_{ki}\}_k\right) \quad (2)$$

Z_i is an indicator of individual i being sampled and N is the number of individuals in the sample. Each term $\Pr\left(Y_{1i}, Y_{2i}, | Z_i = 1, X_i, \{W_{ki}\}_k\right)$ in (2) satisfies

$$\Pr\left(Y_{1i}, Y_{2i} | X_i, Z_i = 1, \{W_{ki}\}_k\right) = \frac{\Pr\left(Z_i = 1 | Y_{1i}, Y_{2i}, X_i, \{W_{ki}\}_k\right) \Pr\left(Y_{1i}, Y_{2i} | X_i, \{W_{ki}\}_k\right)}{\int \Pr(Z_i = 1 | y_{1i}, y_{2i}) \Pr(y_{1i}, y_{2i} | X_i) dy_{1i} dy_{2i}} \quad (3)$$

The sampling mechanism is characterized by $\Pr\left(Z_i = 1 | Y_{1i}, Y_{2i}, X_i, \{W_{ki}\}_k\right)$, which can be explicitly calculated for case–control, extreme-trait and multiple-trait studies. The details are shown in Supplementary Material Section 1. When the probit link function is used to model binary phenotypes, the multivariate generalized linear model can be simplified. Computational details can be found in Supplementary Material Section 2.

Association testing

The likelihood-based score statistic can be applied to detect associations with rare variants. Using collapsing coding, P-values for the score statistics can be analytically evaluated. For the WSC, if the weights are only dependent on the multi-site genotypes, the score statistic will asymptotically follow a normal distribution and the P-values can also be analytically evaluated. Permutation procedures cannot be used to analyze secondary phenotypes in selected samples. This is because if the gene region is associated with the primary phenotype, study subjects are not interchangeable under the null hypothesis of no gene/secondary phenotype associations. The analyses in the article were performed using the CMC coding, that is, $X_i = C^{CMC}\left(\vec{X}_i\right)$. The results remain the same when other coding schemes are used.

Combining different cohorts for analyses of secondary phenotypes

Statistical theories for combining multiple studies are well developed.[27] As heterogeneity may exist between different cohorts, meta-analysis methods that combine test statistics should be used.[11,12] For rare variant analysis, multiple rare variants are jointly analyzed and their phenotypic effects are not usually estimated and reported. Therefore, all the joint analyses in this study were performed by combining score statistics from different studies. In the joint analysis, score statistics from different studies are weighted and summed. The weight assigned for each score statistic is proportional to the square root of the sample size according to Skol *et al.*[28]

Generation of genetic and phenotypic data

Following Boyko *et al*,[18] a rigorous population genetic model incorporating demographic change and purifying selections was used to simulate the African variant data. Details of generation of genetic data are given in Supplementary Material Section 3. To generate phenotypes, we assume that the phenotypic effects for causative variants are independent of their fitness. In a case–control study, the augmented phenotype $\left(A_i^*, T_i\right)$ for an individual i with multi-site genotype $\vec{X}_i = \left(x_i^1, x_i^2, \cdots, x_i^S\right)$ follows a bivariate normal distribution $\mathrm{MVN}\left(\vec{\mu}_i^{CC}, \Sigma^{CC}\right)$, with

$$\vec{\mu}_i^{CC} = \left(\tilde{\beta}_{A^*} \sum_{s\in CV_{A^*}} x_i^s, \tilde{\beta}_T \sum_{s\in CV_T} x_i^s\right),$$

and

$$\Sigma^{CC} = \begin{pmatrix} \sigma_{A^*}^2 & \rho_{A^*,T}\sigma_{A^*}\sigma_T \\ \rho_{A^*,T}\sigma_{A^*}\sigma_T & \sigma_T^2 \end{pmatrix} \quad (4)$$

The rare variants sites CV_{A}*and CV_T are randomly chosen to be causative for the traits A^* and T. Either set can be empty if the gene is not associated with the corresponding trait. Variants at sites $CV_{A^*} \cap CV_T$ are pleiotropic and affect both phenotypes. The binary disorder status A_i is determined by $A_i = \delta\left(A_i^* > a^C\right)$. For each scenario, 1000 individuals were simulated. Details for simulating the extreme-trait and multiple-trait study samples can be found in Supplementary Material Section 4.

In order to evaluate type-I errors, phenotype data were generated under the null hypothesis of no gene/secondary trait T associations, that is, β_T=0. Scenarios were considered where (1) the gene region is neither associated with the primary nor the secondary phenotypes and (2) the gene is associated with the primary phenotype but not with the secondary phenotype. Scenarios with a combination of two causative variant primary trait effects $\tilde{\beta}_{A^*} = 0.5\sigma_{A^*}, 0\left(\mathrm{or}\tilde{\beta}_B, \tilde{\beta}_{C^*}\right)$ and four residual correlations $\rho_{A^*,T} = \pm 0.3, \pm 0.6\ \left(\mathrm{or}\rho_{B,T}, \rho_{C^*,T}\right)$ were evaluated.

To compare the power of rejecting the null hypothesis of no gene/secondary trait associations, two causal variant secondary phenotype effects $\beta_T = \pm 0.5\sigma_T$ were used. The power for the three study designs was compared under scenarios with different combinations of genetic parameter values.

Software availability

An R-package implementing MTA is available at http://www.bcm.edu/genetics/leal/software, which is compatible with commonly used operating systems, including Linux, Windows and OS X.

RESULTS

Evaluation of type-I errors

Type-I errors for each study design using MTA were evaluated empirically. Under the null hypothesis of no genetic/secondary phenotype associations, the quantile–quantile (Q–Q) plots of the empirical and theoretical distributions of *P*-values are shown in Figures 2 and 3 for the case–control study design. When the ascertainment mechanism is correctly specified, the type-I errors are controlled. Results are shown in Figure 2 for the scenario where the gene region is not associated with either the primary or the secondary phenotypes, and the scenario where the gene region is only associated with the primary trait. Type-I errors for the extreme-trait and multiple-trait designs were also well controlled (data not shown). The impact of mis-specified sampling mechanisms was investigated. The results are shown in Figure 3 when the prevalence parameter is 10%, but is incorrectly set to be 7% (Figure 3a) or 13% (Figure 3b) in the analyses. The results indicate that mis-specifying prevalence has only a very minimal impact on type-I error rates as can be observed in the Q-Q plot.

In order to illustrate the bias that could be induced by ascertainment, we also analyzed the simulated data using likelihood models without proper ascertainment corrections and the biases in most scenarios can be substantial. The details for the analyses are shown in Supplementary Material Section 5 and Supplementary Figure 1.

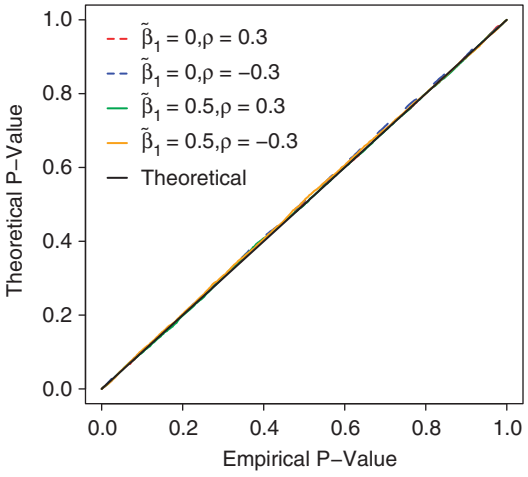

Figure 2 Q-Q plot of *P*-values under the null hypothesis of no gene/ secondary trait (*T*) associations. It is assumed that the disease prevalence (10%) is correctly specified. Scenarios with different combinations of primary trait phenotypic effects $\beta_A{}^*$ and residual correlations $\rho_{A^*,T}$ were investigated. Results are shown where neither the primary nor the secondary traits are associated with the gene region (dashed red and blue lines), and where only the primary but not the secondary trait is associated with the gene region (solid green and brown line). The results were obtained using 10 000 replicates.

Power of detecting secondary phenotype rare variant associations

The efficiency of the three selective sampling designs for detecting secondary trait associations was compared when both the primary and the secondary traits are associated with the same gene (Tables 2). Scenarios were examined where 1000 individuals are sequenced for each study design. There is considerable power for detecting secondary phenotype associations in selected samples. Analyzing secondary phenotypes in a case–control or an extreme-trait study data set can be consistently more powerful than a randomly ascertained population data set of equal size.

When a population-based sample is used where 1000 individuals are randomly selected regardless of their phenotypic values, the power for rejecting the null hypothesis is only 51.7% (Supplementary Table 1). For a case–control sample where the secondary trait T is analyzed, the power can be higher (Table 2). For example, when the primary and secondary trait phenotypic effects, and residual correlation satisfy $\beta_A{}^*=0.5\sigma_A{}^*$, $\beta_T=0.5\sigma_T$ and $\rho_{A^*,T} = -0.6$, the power is 56.5%. It is also comparable to the power (56.6%) when 200 individuals with the most extreme trait T values from a cohort of 5000 are sequenced (Supplementary Table 2).

Compatible with observations from bivariate phenotype association studies,[16] the power for detecting associations with secondary phenotypes is jointly determined by the sizes and directions of the locus phenotypic effects and residual correlations. The power is the highest

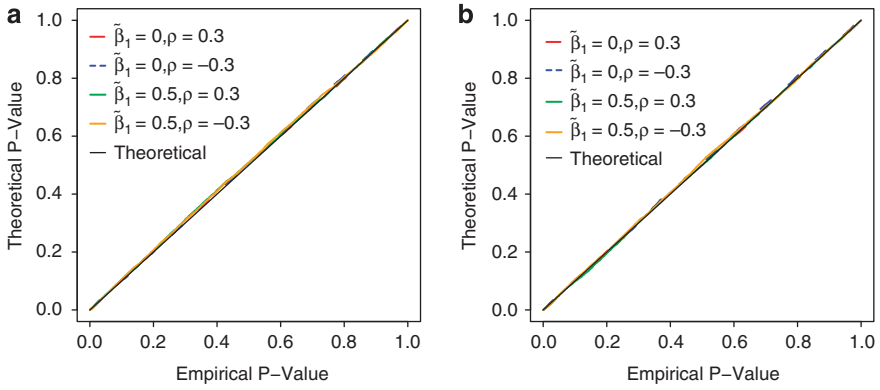

Figure 3 Q-Q plot of *P*-values under the null hypothesis of no gene/ secondary trait (*T*) associations when prevalence is mis-specified. It is assumed that the prevalence (10%) is incorrectly specified as 7% (**a**) or 13% (**b**). Results are shown where neither the primary nor the secondary traits are associated with the gene region (dashed red and blue lines), and where only the primary but not the secondary traits is associated with the gene region (solid green and brown line). The results were obtained using 10 000 replicates.

Table 2 Power to detect secondary trait *T* associations using case–control, extreme-trait and multiple-trait study design

Genetic parameters			Power[d]
$\tilde{\beta}_A$*($\tilde{\beta}_B$, $\tilde{\beta}_C$*)[a]	$\tilde{\beta}_T$[b]	$\rho_{A^*,T}$ ($\rho_{B,T}$, $\rho_{C^*,T}$)[c]	CC[e]/ET[f]/MT[g]
0.5	−0.5	−0.3	0.536/0.562/0.316
0.5	−0.5	0.3	0.548/0.605/0.418
0.5	0.5	−0.3	0.557/0.582/0.448
0.5	0.5	0.3	0.535/0.557/0.506
0.5	−0.5	−0.6	0.533/0.582/0.292
0.5	−0.5	0.6	0.556/0.654/0.471
0.5	0.5	−0.6	0.565/0.667/0.391
0.5	0.5	0.6	0.545/0.589/0.562
0	−0.5	−0.3	0.510/0.555/0.325
0	−0.5	0.3	0.499/0.557/0.412
0	0.5	−0.3	0.508/0.544/0.414
0	0.5	0.3	0.517/0.555/0.497
0	−0.5	−0.6	0.527/0.598/0.315
0	−0.5	0.6	0.513/0.609/0.447
0	0.5	−0.6	0.521/0.606/0.373
0	0.5	0.6	0.531/0.602/0.549

[a]Causal variant phenotypic effect for liability trait *A**, trait *B* and liability trait *C**.
[b]Causal variant effect for secondary trait *T*.
[c]Residual correlation between the primary (liability) trait and secondary trait *T*.
[d]Power was empirically estimated using 5000 replicates under a significance level α=0.05.
[e]Power for case–control study. A case–control study sample consists of 500 cases and 500 controls.
[f]Power for extreme-trait study. An extreme-trait study sample consists of 1000 individuals with extreme trait *B*-values selected from a cohort of 5000.
[g]Power for multiple-trait study. A multiple-trait study sample is obtained based on both trait *C* and trait *T*. The affection status is determined by *C*. Five hundred affected individuals with *T*-values >65th percentile are sequenced, as well as 500 unaffected individuals.

Table 3 Power to detect secondary trait *T* associations for individual studies (case–control and multiple-trait) and the combined analysis

Parameters						Power[g]		
$\tilde{\beta}_{A^*}$[a]	$\tilde{\beta}_T^{CC}$[b]	$\rho_{A^*,T}$[c]	$\tilde{\beta}_{C^*}$[d]	$\tilde{\beta}_T^{MT}$[e]	$\rho_{C^*,T}$[f]	*Case–control*[h]	*Multiple-trait*[i]	*Combined-analysis*
0	−0.5	−0.3	0.5	−0.5	0.3	0.510	0.418	0.690
0	−0.5	0.3	0.5	−0.5	0.3	0.499	0.418	0.680
0	0.5	−0.3	0.5	0.5	0.3	0.508	0.526	0.726
0	0.5	0.3	0.5	0.5	0.3	0.517	0.526	0.732
0	−0.5	−0.6	0.5	−0.5	0.3	0.527	0.418	0.703
0	−0.5	0.6	0.5	−0.5	0.3	0.513	0.418	0.685
0	0.5	−0.6	0.5	0.5	0.3	0.521	0.526	0.731
0	0.5	0.6	0.5	0.5	0.3	0.531	0.526	0.741

[a]Causal variant phenotypic effect for liability trait *A** in the case–control study sample.
[b]Causal variant phenotypic effect for trait *T* in the case–control study sample.
[c]Residual correlations between liability trait *A** and trait *T* in the case–control study sample.
[d]Causal variant phenotypic effect for liability trait *C** in the multiple-trait study sample.
[e]Causal variant phenotypic effect for trait *T* in the multiple-trait study sample.
[f]Residual correlations between liability trait *C** and trait *T* in the multiple-trait study sample.
[g]Power was empirically estimated using 5000 replicates under a significance level α=0.05.
[h]The case–control sample consists of 500 cases and 500 controls.
[i]The multiple-trait data set is obtained based on both trait *C* and trait *T*. The affection status is determined by *C*. Five hundred affected individuals with trait *T*-values >65th percentile are sequenced, as well as 500 unaffected individuals.

when the correlation between the locus phenotypic effects is in the opposite direction of the trait residual correlations. For example, when the locus-induced correlation is positive (ie, $\tilde{\beta}_{A^*} = 0.5\sigma_{A^*}$ and $\tilde{\beta}_T = 0.5\sigma_T$) and the trait residual correlation is negative (ie, $\rho_{A^*,T} = -0.3$), the power is 55.7%. However, if the trait residual correlation is also positive (ie, $\rho_{A^*,T} = 0.3$), the power is 53.5% (Table 2).

Similar patterns of power comparisons are observed for detecting associations with secondary phenotypes T in extreme-trait studies. The power for an extreme-trait study can be substantially higher than that for a population-based study of equivalent size. For example, if the primary and secondary trait effects, and residual correlations are given by $\tilde{\beta}_{C^*} = 0.5\sigma_{C^*}$ $\tilde{\beta}_T = 0.5\sigma_T$ and $\rho_{C^*,T} = -0.6$, the power of rejecting the null hypothesis is 66.7% (Table 2). It is comparable to the power (70.6%) when 600 individuals with the most extreme trait T values from a cohort of 5000 are sequenced (Supplementary Table 2), or the power (66.6%) when 2000 randomly selected samples are sequenced (Supplementary Table 1).

When the gene region is only associated with the secondary trait *T*, using samples ascertained on the primary phenotype will induce selections on the secondary phenotype. For a data set of equivalent size, the power for rejecting the null hypothesis of no gene/secondary trait associations in case–control or extreme-trait samples is still greater than (or comparable to) analyzing the same trait using a randomly ascertained population sample. For example, in an extreme-trait study, which sequences 1000 individuals, when causal variants in the gene affect the secondary trait with effect $\tilde{\beta}_T = 0.5\sigma_T$ and the two traits are positively correlated with correlation coefficient ρ=0.6, the power is 60.2%. If the two traits are negatively correlated with ρ=−0.6, the power is 60.6% (Table 2). The power in these two scenarios is both superior to that of a population-based study (51.6%), which sequences an equivalent number of samples (Supplementary Table 1).

The MTA method can be applied to analyze samples ascertained on multiple phenotypes. In this example of a multiple-trait study, 500 affected individuals with trait *T*-value above the 65th percentile are sequenced and 500 unaffected individuals are also selected regardless of their trait *T*-values (Table 2). Compared with the extreme-trait or case–control study design, the multiple-trait study example that is given is not as powerful. This is because there is not enough phenotypic variability in the sample, as affected individuals are only sampled from the sub-population with trait T above the 65th percentile. However, in some scenarios, there can be considerable power in a multiple-trait study, in particular when sampling on the secondary trait T increases phenotypic variability, for example, affected or unaffected individuals are selected to have secondary T trait values from opposite extreme tails.

MTA allows joint analysis of commonly measured phenotypes in different genetic studies. These studies may be targeted at different primary traits. An example is given where a multiple-trait study is implemented, and the association analysis of the secondary trait T is performed by combining a case–control study data set (Table 3). A wide variety of scenarios were extensively evaluated, and a sizable power increase for the combined analysis is consistently observed.

Applications to the *ANGPTL* family of genes

When each of the eight phenotypes from the DHS was analyzed as primary phenotype using selected samples and the entire sample, four nominally significant associations were found for both types of analyses, that is, *ANGPTL4* with TG (P=0.005), *ANGPTL5* with BMI (P=0.003), *ANGPTL5* with HDL (P=0.024) and *ANGPTL6* with BMI (P=0.022). All of the above significant associations were also successfully detected when TG, BMI and HDL were analyzed as secondary phenotypes. An additional association between *ANGPTL4* and HDL (P=0.018) was identified only when the entire sample was analyzed (Supplementary Table 3).

The association between TG and rare variants in the *ANGPTL4* gene was identified using selected samples where the primary traits are BMI (P=0.025), SysBP (P=0.012) or LDL (P=0.010) (Table 4). These traits are only weakly positively correlated with TG, that is, $\rho_{BMI, TG}$=0.227, $\rho_{LDL, TG}$=0.197 and $\rho_{SysBP, TG}$=0.102 (Supplementary Table 4) The association between *ANGPTL4* and TG is not significant using samples with extreme DiasBP (P=0.137), TCL (P=0.065), Gluc (P=0.117) and HDL (P=0.107) levels.

Although the *ANGPTL4* gene is significantly associated with HDL and the size of the correlation between HDL and TG is larger ($\rho_{HDL, TG}$=−0.374; Supplementary Table 4), the association of TG with *ANGPTL4* gene is not significant when TG is analyzed as a secondary trait using samples with extreme HDL levels. This could have occurred because the locus phenotypic effects for HDL and TG are negatively correlated, and the locus-induced correlation lies in the same direction as the residual correlation, which is shown in our simulations to have reduced power compared with when the locus-induced correlation and trait residual correlations are in opposite directions.

There is one nominally significant association that was only detected in secondary phenotype analyses, that is, the association between Gluc and rare variants in the *ANGPTL3* gene (P=0.024). It was identified when samples with extreme LDL levels were used. But when Gluc was analyzed as primary trait, the association is not significant (P=0.64). This could either be a novel association or a false-positive finding.

Table 4 Results for the secondary phenotype analyses using sequence data from the *ANGPTL3*, *ANGPTL4*, *ANGPTL5* and *ANGPTL6* genes

Primary phenotype	*P-values for analyzing secondary phenotypes*[a]							
	BMI	*DiasBP*	*SysBP*	*TCL*	*LDL*	*HDL*	*TG*	*Gluc*
ANGPTL3								
BMI	—	0.649	0.766	0.429	0.681	0.717	0.121	0.114
DiasBP	0.941	—	0.889	0.580	0.745	0.309	0.441	0.398
SysBP	0.550	0.509	—	0.371	0.223	0.689	0.073	0.222
TCL	0.988	0.955	0.327	—	0.971	0.289	0.163	0.151
LDL	0.871	0.372	0.349	0.114	—	0.116	0.183	0.024*
HDL	0.945	0.616	0.312	0.825	0.668	—	0.561	0.639
TG	0.910	0.883	0.437	0.945	0.418	0.863	—	0.148
Gluc	0.652	0.208	0.351	0.982	0.475	0.692	0.335	—
ANGPTL4								
BMI	—	0.292	0.268	0.733	0.440	0.497	0.025*	0.972
DiasBP	0.965	—	0.380	0.361	0.363	0.121	0.137	0.389
SysBP	0.993	0.551	—	0.728	0.754	0.099	0.012*	0.405
TCL	0.861	0.532	0.571	—	0.052	0.759	0.065	0.933
LDL	0.281	0.894	0.269	0.135	—	0.053	0.010*	0.999
HDL	0.708	0.904	0.286	0.318	0.262	—	0.107	0.874
TG	0.310	0.364	0.584	0.629	0.326	0.784	—	0.845
Gluc	0.824	0.524	0.084	0.848	0.561	0.479	0.118	—
ANGPTL5								
BMI	—	0.920	0.114	0.521	0.233	0.056	0.377	0.797
DiasBP	0.118	—	0.096	0.451	0.803	0.092	0.616	0.367
SysBP	0.203	0.887	—	0.117	0.160	0.304	0.791	0.294
TCL	0.107	0.536	0.923	—	0.399	0.014*	0.221	0.488
LDL	0.084	0.735	0.587	0.202	—	0.002*	0.147	0.458
HDL	0.387	0.866	0.917	0.463	0.991	—	0.569	0.900
TG	0.044*	0.871	0.074	0.296	0.597	0.185	—	0.448
Gluc	0.030*	0.779	0.957	0.546	0.717	0.002*	0.451	—
ANGPTL6								
BMI	—	0.300	1.000	0.606	0.457	0.324	0.401	0.419
DiasBP	0.008*	—	0.385	0.459	0.690	0.478	0.721	0.197
SysBP	0.773	0.816	—	0.622	0.853	0.668	0.338	0.490
TCL	0.024*	0.530	0.992	—	0.823	0.324	0.702	0.940
LDL	0.089	0.383	0.850	0.485	—	0.429	0.801	0.314
HDL	0.034*	0.101	0.873	0.800	0.870	—	0.393	0.215
TG	0.210	0.735	0.974	0.357	0.695	0.561	—	0.811
Gluc	0.153	0.402	0.897	0.340	0.531	0.267	0.905	—

[a]For each phenotype, individuals were selected with trait values in the upper and lower quartiles, and the remaining seven phenotypes were analyzed as secondary traits using MTA.

DISCUSSION

In this article, a flexible likelihood framework MTA is proposed for jointly modeling multiple phenotypes in non-randomly ascertained samples, for example, case–control samples or extreme-trait samples. By coupling multivariate generalized linear models with prospective likelihood, complicated ascertainment mechanisms can be incorporated. The approach is flexible and particularly suitable for analyzing complex traits. It can be applied to any study with known sampling mechanisms. MTA allows efficient statistical inference for the genetic parameters of interest. Although the discussion in this article is focused on analyzing sequence data, MTA can also be applied to analyze genotype data.

The results presented in this article have important implications for the design and analysis of complex traits. Most current studies, owing to their limited sample size, are not adequately powered to detect associations with rare variants. It has been suggested that for an exome study ~10 000 individuals with extreme traits from a cohort of 100 000 need to be sequenced in order to have adequate power.[5] However, the sample size well exceeds the capacity of many existing studies.[5] It is therefore particularly important that combined analysis can be performed using data from multiple studies in order to have sufficient power. Applying MTA, sequencing studies that are targeted at different primary traits can be jointly analyzed for detecting associations with a variety of commonly measured secondary traits.

The power of different selective study designs was investigated. It was shown through extensive simulations that there is considerable power for detecting secondary phenotype associations in selected samples. In particular, when the secondary trait of interest is analyzed in a case–control or an extreme-trait study data set, the power can be greater than analyzing an equivalent sized randomly ascertained sample. Using data-sharing platforms and protocols such as dbGaP,[29] samples from existing studies can be freely obtained and analyzed. The power can be greatly increased when data from multiple studies are jointly analyzed.

Secondary phenotypes not only have their own clinical importance, but they can also be relevant for understanding the primary trait etiologies. For example, among studies of T2D, many are targeted at related quantitative traits, including fasting glucose levels[30] and C-reactive protein.[31] Given that these traits are often available for individuals who participate in T2D case–control studies,[32] MTA can be applied to detect associations with these additional phenotypes.

MTA was also applied to the analysis of sequence data from the DHS. Multiple associations were identified, which confirmed previous data analyses. When the traits were analyzed as secondary phenotypes, although these same set of associations was observed, they were not detected in every selected sample, for example, the association between TG levels and *ANGTPL4* was only detected in secondary trait analyses

using samples with extreme BMI, SysBP and LDL, but not in samples with extreme DiasBP, HDL, TCL and Gluc. This could be affected by the small sample sizes that were analyzed; the moderate effect sizes for variants involved in complex trait etiologies; or the directions and magnitudes of the correlations between the primary and secondary phenotypes. Although these identified associations are only nominally significant, they all have biological support. In fact, the effects of mutant *ANGPTL*-family genes on lipoprotein lipase (LPL) have been investigated through *in vitro* functional studies and *in vivo* mice studies. LPL has been known to affect glucose metabolism,[33] cholesterol level[34] and blood pressure.[35] Additionally, the association between variants in the *ANGPTL4* gene and triglyceride levels has been successfully replicated.[3,36]

Sensitivity of MTA to mis-specified sampling mechanisms was extensively evaluated. When the disease prevalence is reported as an interval of possible values, inferences from MTA can be conducted under different prevalence values from the interval. The results can be integrated using a model averaging procedure. It has been shown that it is an effective approach to further reduce the impact of mis-specified prevalence.[37]

There can be heterogeneities of sequence coverage depth within and between different studies. Coverage depth differences within a single study may cause inflated type-I errors. Possible strategies to reduce the bias include incorporating the mean coverage depth of each individual in the analysis as a covariate.[38] The method can be used with the MTA model. In order to be robust against between-study heterogeneities, a meta-analyses procedure should be implemented for the joint analysis, instead of performing mega-analysis that combines individual participant data.[11,12]

When multiple phenotypes are analyzed, to avoid inflated type-I error owing to testing multiple hypotheses, a stringent significance level must be specified. Owing to phenotypic correlations, Bonferroni corrections for testing multiple genes and phenotypes can be overly conservative. Instead, the spectral decomposition-based method of Nyholt *et al*[39] can be used. In addition to correctly controlling for family-wise error rates, it is important that the findings can be replicated using independent samples.[40]

With large-scale implementation of sequence-based genetic association studies, the capability for mapping complex traits will be further elevated. Detecting associations with rare variants and jointly investigating multiple phenotypes together can be an ambitious and difficult task given the moderate sample sizes of existing studies. Taking advantage of multiple studies and mapping commonly measured phenotypes using MTA is therefore highly beneficial and will greatly accelerate the process of dissecting complex trait genetic etiologies.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This research is supported by National Institutes of Health Grants 1RC4MD005964 and 1RC2HL102926 (SML). DJL is partially supported by a training fellowship from the Keck Center Pharmacoinformatics Training Program of the Gulf Coast Consortia (NIH Grant no. 5 R90 DK071505-04). We thank Drs Jonathan Cohen (JC) and Helen Hobbs for providing us with data from the Dallas Heart Study on the *ANGTPL*-family genes, which was supported by National Institutes of Health Grant RL1HL092550 (JC). Computation for this research was supported in part by the Shared University Grid at Rice funded by NSF under Grant EIA-0216467, and a partnership between Rice University, Sun Microsystems and Sigma Solutions Inc.

1 Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 2004; **305**: 869–872.
2 Ji W, Foo JN, O'Roak BJ *et al*: Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* 2008; **40**: 592–599.
3 Romeo S, Pennacchio LA, Fu Y *et al*: Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat Genet* 2007; **39**: 513–516.
4 Bodmer W, Bonilla C: Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008; **40**: 695–701.
5 Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR: Power of deep, all-exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci USA* 2009; **106**: 3871–3876.
6 Cohen JC, Pertsemlidis A, Fahmi S *et al*: Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc Natl Acad Sci USA* 2006; **103**: 1810–1815.
7 Cauchi S, Nead KT, Choquet H *et al*: The genetic susceptibility to type 2 diabetes may be modulated by obesity status: implications for association studies. *BMC Med Genet* 2008; **9**: 45.
8 Cauchi S, Meyre D, Dina C *et al*: Transcription factor TCF7L2 genetic study in the French population: expression in human beta-cells and adipose tissue and strong association with type 2 diabetes. *Diabetes* 2006; **55**: 2903–2908.
9 Lin DY, Zeng D: Proper analysis of secondary phenotype data in case–control association studies. *Genet Epidemiol* 2009; **33**: 256–265.
10 Richardson DB, Rzehak P, Klenk J, Weiland SK: Analyses of case–control data for additional outcomes. *Epidemiology* 2007; **18**: 441–445.
11 Ioannidis JP, Thomas G, Daly MJ: Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet* 2009; **10**: 318–329.
12 McCarthy MI, Abecasis GR, Cardon LR *et al*: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008; **9**: 356–369.
13 Cirulli ET, Goldstein DB: Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010; **11**: 415–425.
14 Plomin R, Haworth CM, Davis OS: Common disorders are quantitative traits. *Nat Rev Genet* 2009; **10**: 872–878.
15 Lange C, Silverman EK, Xu X, Weiss ST, Laird NM: A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics* 2003; **4**: 195–206.
16 Liu J, Pei Y, Papasian CJ, Deng HW: Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genet Epidemiol* 2009; **33**: 217–227.
17 Allison DB, Thiel B, St Jean P, Elston RC, Infante MC, Schork NJ: Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am J Hum Genet* 1998; **63**: 1190–1201.
18 Boyko AR, Williamson SH, Indap AR *et al*: Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* 2008; **4**: e1000083.
19 Liu DJ, Leal SM: A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet* 2010; **6**: e1001156.
20 Madsen BE, Browning SR: A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009; **5**: e1000384.
21 Price AL, Kryukov GV, de Bakker PI *et al*: Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010; **86**: 832–838.
22 Morris AP, Zeggini E: An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 2009; **34**: 188–193.
23 Li B, Leal SM: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008; **83**: 311–321.
24 Neale BM, Rivas MA, Voight BF *et al*: Testing for an unusual distribution of rare variants. *PLoS Genet* 2010; **7**: e1001322.
25 Bhatia G, Bansal V, Harismendy O *et al*: A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput Biol* 2010; **6**: e1000954.
26 Aitken AC: Notes on selection from a multivariate normal population. *Proc Edin Math Soc* 1934; **4**: 106–110.
27 Munafo MR, Flint J: Meta-analysis of genetic association studies. *Trends Genet* 2004; **20**: 439–444.
28 Skol AD, Scott LJ, Abecasis GR, Boehnke M: Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006; **38**: 209–213.
29 Mailman MD, Feolo M, Jin Y *et al*: The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007; **39**: 1181–1186.
30 Bouatia-Naji N, Rocheleau G, Van Lommel L *et al*: A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels. *Science* 2008; **320**: 1085–1088.
31 Elliott P, Chambers JC, Zhang W *et al*: Genetic loci associated with C-reactive protein levels and risk of coronary heart disease. *JAMA* 2009; **302**: 37–48.
32 Sladek R, Rocheleau G, Rung J *et al*: A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007; **445**: 881–885.
33 Webster RJ, Warrington NM, Weedon MN *et al*: The association of common genetic variants in the APOA5, LPL and GCK genes with longitudinal changes in metabolic and cardiovascular traits. *Diabetologia* 2009; **52**: 106–114.

34 Koster A, Chao YB, Mosior M *et al*: Transgenic angiopoietin-like (angptl)4 overexpression and targeted disruption of angptl4 and angptl3: regulation of triglyceride metabolism. *Endocrinology* 2005; **146**: 4943–4950.
35 Li B, Ge D, Wang Y *et al*: Lipoprotein lipase gene polymorphisms and blood pressure levels in the Northern Chinese Han population. *Hypertens Res* 2004; **27**: 373–378.
36 Romeo S, Yin W, Kozlitina J *et al*: Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J Clin Invest* 2009; **119**: 70–79.
37 Li M, Li C: Assessing departure from Hardy–Weinberg equilibrium in the presence of disease association. *Genet Epidemiol* 2008; **32**: 589–599.
38 Garner C: Confounded by sequencing depth in association studies of rare alleles. *Genet Epidemiol* 2011; **35**: 261–268.
39 Nyholt DR: A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 2004; **74**: 765–769.
40 Liu DJ, Leal SM: Replication strategies for rare variant complex trait association studies via next-generation sequencing. *Am J Hum Genet* 2010; **87**: 790–801.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)