ORIGINAL ARTICLE

Quantitative trait locus analysis for next-generation sequencing with the functional linear models

Li Luo,[1] Yun Zhu,[2] Momiao Xiong[2]

▸ Additional data are published online only. To view these files please visit the journal online (http://dx.doi.org/10.1136/jmedgenet-2012-100798)

[1]Division of Epidemiology, Biostatistics and Preventive Medicine, University of New Mexico, Albuquerque, New Mexico, USA
[2]Department of Biostatistics, Human Genetics Center, The University of Texas School of Public Health, Houston, Texas, USA

Correspondence to
Dr Momiao Xiong, Department of Biostatistics, Human Genetics Center, The University of Texas Health Science Center at Houston, PO Box 20186, Houston, TX 77225, USA; Momiao.Xiong@uth.tmc.edu

Received 31 January 2012
Revised 23 June 2012
Accepted 28 June 2012

ABSTRACT

Background Although in the past few years we have witnessed the rapid development of novel statistical methods for association studies of qualitative traits using next generation sequencing (NGS) data, only a few statistics are proposed for testing the association of rare variants with quantitative traits. The quantitative trait locus (QTL) analysis of rare variants remains challenging. Analysis from low dimensional data to high dimensional genomic data demands changes in statistical methods from multivariate data analysis to functional data analysis.

Methods We propose a functional linear model (FLM) as a general principle for developing novel and powerful QTL analysis methods designed for resequencing data. By simulations we calculated the type I error rates and evaluated the power of the FLM and other eight existing statistical methods, even in the presence of both positive and negative signs of effects.

Results Since the FLM retains all of the genetic information in the data and explores the merits of both variant-by-variant and collective analysis and overcomes their limitation, the FLM has a much higher power than other existing statistics in all the scenarios considered. To evaluate its performance further, the FLM was applied to association analysis of six quantitative traits in the Dallas Heart Study, and RNA-seq eQTL analysis with genetic variation in the low coverage resequencing data of the 1000 Genomes Project. Real data analysis showed that the FLM had much smaller p values to identify significantly associated variants than other existing methods.

Conclusions The FLM is expected to open a new route for QTL analysis.

INTRODUCTION

Rapidly developed next generation sequencing (NGS) technologies with faster, cheaper and more accurate sequencing will soon make whole genomic sequence analysis for a large number of individuals to be a routine analysis in genetic studies of common diseases and will revolutionise biomedical research.[1 2] Resequencing of exomes—and ultimately whole genomes—will generate unprecedented massive, high-dimensional genetic variation data at finer and finer resolution, and provides a powerful tool for detecting the entire allelic spectrum of the causal genetic variations.[3] Despite their promise, NGS platforms also have three specific features: high error rates, enrichment of rare variants, and a large proportion of missing values.[4–8] Available genetic data analysis platforms that are mainly designed for testing association of common variants with the disease provide useful tools for single marker-based genome wide association studies (GWAS), but have limitations in analysing thousands of sequences collected for very large population-based studies of humans.[4] Developing a new approach for the analysis of the massive sequencing data poses novel and great challenges to statistical analysis.[9] Genetic studies of complex diseases are undergoing a paradigm shift from the 'common disease, common variant' hypothesis to the 'common disease, both common and rare variant' hypothesis and from the single market analysis to the joint analysis of multiple variants in a genomic region that can be genes or other functional units.[10–23]

Currently, most association studies of rare variants focus on a qualitative trait.[24] Only two approaches to testing association of rare variants with a quantitative trait have been developed.[9 12 13 24 25] One approach involves group-based simple regression (SRG) methods that combine multiple rare variants into a single variable as a predictor. Another approach uses phenotype extreme selection design that contrasts the frequency of rare variants between individuals in the two tails of the phenotype distribution to transform a quantitative trait association study into a case–control collapsing methods association study. Since statistical methods for phenotype extreme selection design are similar to statistical methods for qualitative trait association analysis, in this paper we only investigate statistical methods that do not belong to the set of analytic tools for case–control association studies.

Although in many cases the regression type collapsing methods that aggregate information across variants for prediction have a higher power than the individual variant tests, they also suffer from limitations in quantitative trait locus (QTL) analysis. First, regression type collapsing methods ignore differences in genetic effects among single nucleotide polymorphism (SNPs) at different genomic locations. Second, collapsing methods do not leverage linkage disequilibrium (LD) in the data. Third, since sequence errors are cumulative when rare variants are grouped, some regression type collapsing methods are sensitive to genotyping errors and missing data. Fourth, collapsing methods often make homogeneity assumptions where putatively functional variants are assumed to have a similar direction of genetic effects. However, in practice, genetic effects are expected to differ in size and direction.[9] Collapsing methods are difficult to deal with size and effect sign (direction) of heterogeneity. Fifth, although some types of multiple regression (MRG) methods such as sequence kernel association test (SKAT)[25] allow different variants with different sign and sizes of genetic effects, MRG methods do not reduce the dimensions of the data and can suffer from multicolinearity.[26]

The purpose of this paper is to develop statistical methods that can utilise the merit of both individual and collapsing methods, but at the same time overcome limitations inherent by single marker test and group test alone. We find that functional data analysis is a powerful tool for sequence-based association studies and can serve this purpose very well. In the past decade, as technology of measurement instruments progresses, we can observe the data sampled over a fine and finer time or space. The observed data points are so close that they can be considered as observations varying over a continuum. Therefore, it is natural to assume that the data are realisation of an underlying stochastic process taking values in a functional space.[27] Functional data analysis is a statistical methodology that takes an entire sequence of observations for an individual as a single functional entity rather than discrete observations.[28] Functional data analysis has received increasing attention.[26 29] It investigates the relationships between functional components, for example, components that can be considered to be (random) functions. Functional data analysis has been applied to many fields including engineering, chemometrics, environmental science, finance, genetics, biology and geophysics.[30–35]

Functional data analysis techniques were recently applied to association studies for next-generation sequencing data where functional component analysis was used to test association of rare variants with a qualitative trait.[23] This paper aims to show that functional linear models (FLMs) are also suited for quantitative trait association analysis with next-generation sequencing data. In the 1000 Genomes Project, more than 36.6 million SNPs have been identified across the genomes in the sequenced 1094 individuals. These densely distributed genetic variants across the genomes in large samples can be viewed as realisations of a Poisson process with its intensity depending on the total mutation rate.[36] The densely typed genetic variants in a genomic region for each individual are so close that these genetic variants can be considered as varying over a continuum and be treated as observed data taken from curves. These data are called functional. Since standard multivariate statistical analyses often fail with functional data,[37] we formulate a test of quantitative trait association with genetic variants in a genomic region as an FLM with scalar response, where the genotype score functions are defined as a function of the genomic position of the genetic variants rather than a set of discrete genotype values and the quantitative trait is predicted by genotype score function. We will show that the FLM with scale response is a natural extension of the multivariate regression for QTL analysis. The FLM can collectively test for the association of genetic variants with the quantitative trait, but can also allow for heterogeneity of genetic effects. The FLM can retain all of the genetic information in the data.

Three types of approaches have been developed for estimating regression coefficient functions in the FLMs.[38] The first approach uses basis function expansion and penalised methods to estimate regression coefficient functions.[26 39–41] The second approach uses functional principal component (PCA) expansions and least square to estimate regression coefficient functions.[42] The third approach uses non-parametric techniques.[27 43] In this paper, we will formulate testing the association of genetic variants including both common and rare variants with a quantitative trait as an FLM and use the basis function expansion and penalised methods to estimate genetic additive effect functions. We use large scale simulations to calculate the type I error rates and evaluate the power of nine alternative statistical methods: FLM, SRG, MRG, regression on PCA, the SKAT,[25] two collapsing-based regression tests (regression on the proportion of rare variants at which an individual carries at least one copy of the minor allele (RVT1), and regression on the presence/absence of at least one minor allele at any rare variant (RVT2)),[24] the weighted sum (WSS)[12] and variable threshold (VT) method[13] under various scenarios of the quantitative trait. To evaluate its performance further, the FLM is applied to ANGPTL4 sequence and six continuous phenotypes data from the Dallas Heart Study[44] and RNA-seq eQTL analysis with genetic variation in the low coverage resequencing data of the 1000 Genomes Project and gene expressions acquired by RNA sequencing (RNA-seq) in lymphoblastoid cell lines (LCLs) from 60 individuals of European origin (CEU).[45] We have found that the FLM substantially outperforms the current statistical methods for quantitative trait association analysis with rare variants in both power analysis and real data applications. A programme for implementing the developed FLMs for quantitative genetic analysis can be downloaded from our website (http://www.sph.uth.tmc.edu/hgc/faculty/xiong/index.htm or http://www.bioconductor.org/).

MODELS AND METHODS

Functional linear model for a quantitative trait

Since it jointly uses multiple marker information, the multiple linear regression models (supplementary material), in general, might have higher power to detect a QTL than the simple linear regression model. However, as the number of markers increases, the degree of freedom of the test statistics also increases, and multicollinearity where the coefficient estimates may change erratically in response to small changes in the model or the data in MRGs are often observed. This will compromise the power of MRGs for identifying QTLs. In addition, when the frequencies of the genetic variant become smaller and smaller, the variances of the estimators of the genetic effects will be larger and larger. To reduce the degrees of freedom in the model, variances of the estimators due to the presence of rare variants in the model and multicolinearity in the data, we consider an FLM for a quantitative trait.

Let t be a genomic position. Define a genotype profile $X_i(t)$ of the i-th individual as

$$X_i(t) = \begin{cases} 2P_m(t), & MM \\ P_m(t) - P_M(t), & Mm \\ -2P_M(t), & mm \end{cases} \tag{1}$$

where M and m are two alleles of the marker at the genomic position t, $P_M(t)$ and $P_m(t)$ are the frequencies of the alleles M and m, respectively. We add a constant $-P_m(t)+P_M(t)$ into $X_i(t)$. Then, the indicator variable will be transformed to

$$X_i(t) = \begin{cases} 1, & MM \\ 0, & Mm \\ -1, & mm \end{cases} \tag{2}$$

which is a widely used indicator variable in quantitative genetic analysis.

Let Y_i be a phenotype value of the i-th individual. An FLM for a quantitative trait is defined as

$$Y_i = \mu + \int_0^T X_i(t)\alpha(t)dt + \varepsilon_i, \tag{3}$$

where ε_i are independent and identically distributed normal variables with a mean of zero and variance σ^2, T is the length of the genome region being considered, and the regression coefficient $\alpha(t)$ is the genetic additive effect of the marker at the genomic position t. The main changes in the FLM is that regression coefficients now become regression coefficient function of genomic position with values $\alpha(t)$. For convenience, the genome region (0, T) is rescaled to (0, 1). If the integrals in equation (12) are discretised, the FLM will be reduced to multiple linear regression models (supplementary material). When the number of variants in the genomic region becomes larger and larger, in the limit, the multiple linear model will converge to the FLM (supplementary material).

We explore the use of restricted basis functions to estimate the additive effect function. Since the genotype functions are non-periodic functions, we use B-spline basis functions to expand the genotype functions and additive effect function. The B-spline can be defined via the recurrence relations (www.cs.unc.edu/~dm/UNC/COMP258/Papers/bsplbasic.pdf, supplementary material).

Let $B_K(t)$ be a B-spline basis function; if we set p to be a specific integer (ie, p=3 yields cubic B-spline basis series). We expand the genotype functions $X_i(t)$ in terms of B-spline basis functions. Let B(t) be a vector of B-spline basis functions of length K_G. Then, we have

$$X_i(t) = \sum_{k=1}^{K_G} u_{ik} B_k(t). \tag{4}$$

The coefficients of the expansion u_{ik} can be obtained by minimising the least square criterion:

$$\sum_{j=1}^{T}\left[X_i(t_j) - \sum_{k=1}^{K_G} B_k(t_j)u_{ik}\right]^2. \tag{5}$$

Let $X_i = [X_i(t_1), \ldots, X_i(t_T)]^T$, $u_i = [u_{i1}, \ldots, u_{iK_G}]^T$,

$$B = \begin{bmatrix} B_1(t_1) & \cdots & B_{K_G}(t_1) \\ \cdots & \cdots & \cdots \\ B_1(t_T) & \cdots & B_{K_G}(t_T) \end{bmatrix}.$$

The least square estimators of the expansion coefficients are then given by

$$\hat{u}_i = (B^TB)^{-1}B^TX_i.$$

The expansion of the genetic effect functions can be similarly written as

$$\alpha(t) = \sum_{k=1}^{K_\beta} \alpha_k B_k = \theta^T\alpha,$$

where $\theta = [B_1(t), \ldots, B_{K_\beta}(t)]^T = [\theta_1(t), \ldots, \theta_{k_\beta}(t)]^T$, $\alpha = [\alpha_1, \ldots, \alpha_{K_\beta}]^T$.

After the genotype functions and genetic effect functions are expanded by B-splines, the FLM can be transformed to ordinary MRGs.

Let $Y = [Y_1, \ldots, Y_n]^T$, $X = [X_1, \ldots, X_n]^T$, $U = [u_1, \ldots, u_n]^T$, $B(t) = [B_1(t), \ldots, B_{KG}(t)]^T$. Then, we have

$$X = UB(t).$$

The FLM (2) can then be written as

$$\begin{aligned} Y &= \mu I + \int_0^T UB(t)\theta^T(t)\alpha\, dt + \varepsilon \\ &= \mu I + U\left[\int_0^T B(t)\theta^T(t)\, dt\right]\alpha + \varepsilon. \end{aligned} \tag{6}$$

Let

$$J_{B\theta} = \begin{bmatrix} \int_0^T B_1(t)\theta_1(t)\, dt & \cdots & \int_0^T B_1(t)\theta_{K_\beta}(t)\, dt \\ \cdots & \cdots & \cdots \\ \int_0^T B_{K_G}(t)\theta_1(t)\, dt & \cdots & \int_0^T B_{K_G}(t)\theta_{K_\beta}(t)\, dt \end{bmatrix}$$

$W = [1, UJ_{B\theta}]$, $\beta = [\mu, \alpha^T]^T$. Then, equation (6) can be rewritten as

$$Y = W\beta + \varepsilon. \tag{7}$$

The lease square estimate of the parameter vector β is given by

$$\hat{\beta} = (W^TW)^{-1}W^TY.$$

The smoothed functional linear models

To avoid excessive local fluctuation in the estimated genetic additive effect function, we incorporate the roughness penalty into the FLMs. We define the penalised residual sum of squares

$$F(\lambda, \beta) = \sum_{i=1}^{n}[y_i - \mu - \int_0^T X_i(t)\alpha(t)dt]^2 + \lambda\int_0^T [D^2\alpha(t)]^2 dt, \tag{8}$$

where $D^2\alpha(t) = (\partial^2\alpha(t))/(\partial t^2)$ is the second order derivative of function α (t).

Define a matrix R as

$$R = \int_0^T [D^2\theta(t)][D^2\theta(t)]^T dt,$$

where $D^2\theta(t) = [D^2\theta_1(t), \ldots, D^2\theta_{k_\beta}(t)]^T$. The penalised residual sum of squares can be expressed as

$$F_p(\lambda, \beta) = ||Y - W\beta||^2 + \lambda\beta^T R_0 \beta, \tag{9}$$

where the penalty matrix R_0 is obtained by augmenting the matrix R via attaching a leading column and row of $k_\beta + 1$ zeros and ||.|| is a L_2-norm of a vector. Minimising values β in F_p (λ, β) yields

$$\hat{\beta} = (W^TW + \lambda R_0)^{-1}W^TY. \tag{10}$$

Assume that the variance–covariance matrix $\Sigma_e = \sigma_e^2 I$, where

$$\hat{\sigma}_e^2 = \frac{1}{N - K_\beta - 1} Y^T [I - 2W(W^T W + \lambda R_0)^{-1} W^T + W(W^T W + \lambda R_0)^{-1} W^T W (W^T W + \lambda R_0)^{-1} W^T] Y.$$

The sampling of variance of the estimate $\hat{\beta}$ is given by

$$\mathrm{Var}(\hat{\beta}) = \hat{\sigma}_e^2 (W^T W + \lambda R_0)^{-1} W^T W (W^T W + \lambda R_0)^{-1}. \quad (11)$$

The smoothing parameter λ can be chosen by cross validation.[26]

Test statistics

An essential problem in genetic studies of the quantitative trait is to test the association of a genomic region with the quantitative trait. Formally, we investigate the problem of testing the following hypothesis:

$$H_0 : \alpha(t) = 0, \quad \forall t \in [0, T] \qquad (12)$$

Against

$$H_a : \alpha(t) \neq 0.$$

If the genetic effect function $\alpha(t)$ is expanded in terms of the basic functions:

$$\alpha(t) = \theta(t)^T \alpha,$$

then, testing the null hypothesis H_0 in equation (12) is equivalent to testing the hypothesis:

$$H_0 : \alpha = 0. \qquad (13)$$

Let Λ be the matrix obtained by removing the first row and the first column of the covariance matrix $\mathrm{Var}(\hat{\beta})$ in equations (10) or (11). Define the test statistic as

$$T_Q = \hat{\alpha}^T \Lambda^{-1} \hat{\alpha}. \qquad (14)$$

Then, under the null hypothesis H_0: α=0, T_Q is asymptotically distributed as a central $\chi^2_{(K_\beta)}$ distribution.

RESULTS

Null distribution of test statistics

In the previous section, we have shown that the test statistics T_Q are asymptotically distributed as a central χ^2 distribution. To examine the validity of this statement, we performed a series of simulation studies to compare their empirical levels with the nominal ones.

We calculated the type I error rates for common, low frequency and rare alleles. We assumed the following model to generate a phenotype.

$$Y_i = \mu + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where ε_i are independent and identically distributed normal variables with mean zero and variance $\sigma^2 = 1$.

We generate 1 000 000 chromosomes by resampling from 2356 individuals with variants in 10 genes selected from the National Heart, Lung, and Blood Institute's Exome Sequencing Project where the description of 10 genes is summarised in supplementary table S1. The number of sampled individuals range from 500 to 5000; 5000 simulations were repeated. Tables 1–4 summarise the average type I error rates of the test statistics for testing the association of common, low frequency, rare variants and all variants over 10 genes, respectively, at the nominal levels α=0.05, α=0.01 and α=0.001. Tables 1–4 showed that the type I error rates of the test statistics in the FLM were not appreciably different from the nominal α levels.

Table 1 Average type 1 error rates of the statistics for testing association of a gene that consists of all common variants with a quantitative trait over 10 genes

Sample size	0.001	0.01	0.05
500	0.0010	0.0096	0.0525
1000	0.0011	0.0102	0.0513
2000	0.0009	0.0107	0.0492
3000	0.0010	0.0097	0.045
4000	0.0011	0.0103	0.048
5000	0.0010	0.0104	0.0502

Power evaluation

To evaluate the performance of the FLMs for testing the association of a genomic region with a quantitative trait, we used simulated data to estimate their power to detect a true association. A true quantitative genetic model is given as follows. Consider L trait ~~loci which are located~~ at the genomic positions $t_1, \ldots, t_L$. Let A_l be a risk allele at the l-th trait locus. The following multiple linear regression is used as an additive genetic model for a quantitative trait:

$$Y_i = \mu_m + \sum_{l=1}^{L} X_{il}\alpha_l + \varepsilon_i,$$

where

$$X_{il} = \begin{cases} 2(1 - P_l), & A_l A_l \\ 1 - 2P_l, & A_l a_l \\ -2P_l, & a_l a_l, \end{cases}$$

$$\alpha_l = P_l G_{11}^l + (1 - 2P_l) G_{12}^l - (1 - P_l) G_{22}^l,$$

G_{11}^l, G_{12}^l and G_{22}^l are genotypic values for the genotypes $A_l A_l$, $A_l a_l$ and $a_l a_l$, respectively, and ε_i is distributed as a standard normal distribution N(0,1).

We considered four disease models: additive, dominant, recessive and multiplicative. The relative risks across all variant sites are assumed to be equal and the variants were assumed to influence the trait independently (ie, no epistasis). Let $f_0 = 1$ be

Table 2 Average type 1 error rates of the statistics for testing association of a gene that consists of all rare variants with a quantitative trait over 10 genes

Sample size	0.0010	0.0100	0.0500
500.0000	0.0008	0.0089	0.0396
1000.0000	0.0007	0.0085	0.0415
2000.0000	0.0008	0.0087	0.0432
3000.0000	0.0008	0.0086	0.0450
4000.0000	0.0009	0.0084	0.0435
5000.0000	0.0008	0.0078	0.0409

Table 3 Average type 1 error rates of the statistics for testing association of a gene that consists of all low frequency variants with a quantitative trait over 10 genes

Sample size	0.001	0.01	0.05
500	0.0008	0.0081	0.0409
1000	0.0007	0.0088	0.0398
2000	0.0008	0.0084	0.0409
3000	0.0008	0.0089	0.0413
4000	0.0009	0.0084	0.0419
5000	0.0008	0.0082	0.0429

a baseline penetrance that is defined as the contribution of the wild genotype to the trait variation and r be a risk parameter. For the dominant model, we assume $G^l_{11} = rf_0$, $G^l_{12} = rf_0$ and $G^l_{22} = f_0$. Thus, the genetic additive effect is defined as $\alpha_l = (1-P_l)(r-1)f_0$. Similarly, the genetic additive effects are defined as $\alpha_l = (r-1)f_0$, $\alpha_l = (rP_l + 1-P_l)(r-1)$ f_0 and $\alpha_l = (r-1)$ P_lf_0 for additive, multiplicative and recessive disease models, respectively.

In our sequenced 1000 individuals with the European origin (data have not been published), the average number of SNPs per kb is eight SNPs. However, the average number of SNPs per kb will rapidly increase as the number of sequenced individuals increases. It was recently reported that the average number of SNPs per kb in 202 drug target genes sequenced in 12 514 European subjects is about 48 SNPs.[46] Due to the high cost of whole genome sequencing, in most genetics studies, only thousands of individuals are often sequenced. In our simulations for power studies, 30 SNPs per kb is assumed. For simplicity, we used 25%, 15% and 60% of the SNPs as common, low frequency and rare SNPs. We consider a 30 kb region (average length of a gene). Since the average number of SNPs per kb is 30, we assume that the number of SNPs in a gene is 900 in the simulations. Therefore, we included 225 common SNPs, 135 low frequency SNPs and 540 rare SNPs in the simulations.

We used the MS software[47] to generate a population of 2 000 000 chromosomes with the above variants. Two haplotypes were randomly sampled from the population and assigned to an individual. We randomly selected 10% of the rare variants as causal variants. A total of 2000 individuals for the dominant, additive and multiplicative models, and 3000 individuals for the recessive model were sampled from the populations. A total of 5000 simulations were repeated for power calculation.

Figures 1 and 2 and supplementary figures S1 and S2 plot the power curves of 10 statistics in the presence of sign homogeneity: the FLM, the smoothed FLM, the SKAT, two collapsing-based regression tests (RVT1 and RVT2),[24] SRG where permutation was used to adjust for multiple testing, MRG, regression on PCAs, WSS and VT for testing association of rare variants in the genomic region under additive, multiplicative, dominant and recessive models, respectively. These power curves are a function of the risk parameter at the significance level α=0.05. We observe that the FLM has the highest power, followed by the smoothed FLM, SKAT, RVT2, WSS, and RVT1, and the SRG has the lowest power. The power curves of the SKAT and RVT2, WSS and RVT1 are quite close. We also observe that when the relative risk parameter r increases the differences in power between the FLM and other collapsing methods roughly increase.

To compare the power of 10 statistics in the presence of sign heterogeneity where both positive and negative signs of effects (increasing or decreasing phenotypes) are present, we plotted figures 3 and 4 and supplementary figures S3 and S4 showing the power of 10 statistics in the presence of 45 risk variants and 45 protective variants under the additive, multiplicative, dominant and recessive models, respectively. These figures clearly demonstrate that the power of the FLM and smoothed FLM are the highest, followed by the SKAT, RVT1, RVT2, WSS, regression on PCAs and MRGs. This demonstrates that these five statistics are less sensitive to the signs of effects. We also observe that the power of two collapsing methods, VT and

Table 4 Average type 1 error rates of the statistics for testing association of a gene that consists of all common, low frequency and rare variants with a quantitative trait over 10 genes

Sample size	0.001	0.01	0.05
500	0.0009	0.0103	0.0533
1000	0.0008	0.0103	0.0477
2000	0.0010	0.0102	0.0521
3000	0.0010	0.0090	0.0508
4000	0.0010	0.0091	0.0508
5000	0.0009	0.0098	0.0514

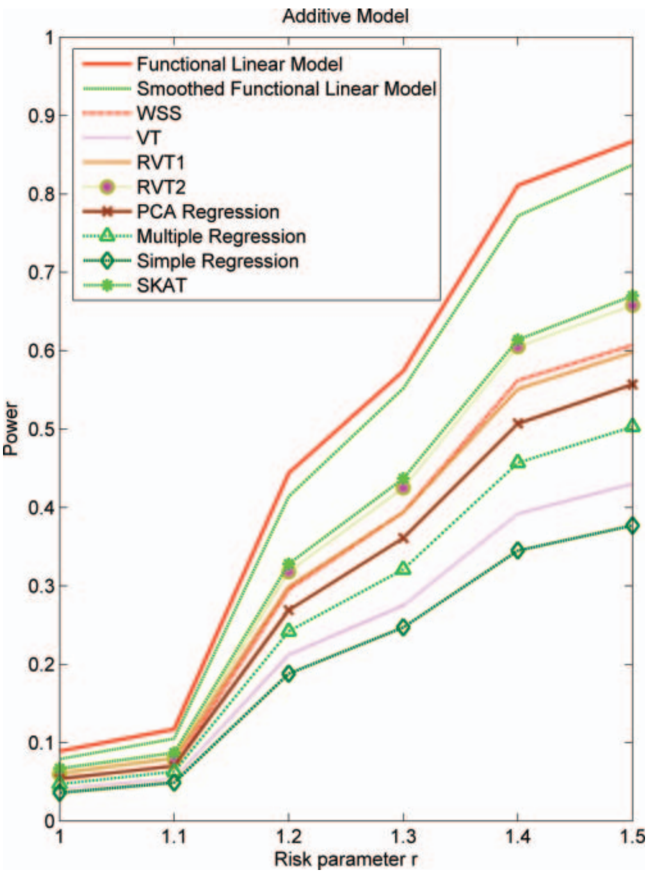

Figure 1 Power curves of 10 statistics: the functional linear model (FLM), the smoothed FLM, the two collapsing-based regression tests RVT1 and RVT2, simple regression where permutation was used to adjust for multiple testing, multiple regression, regression on principal components, sequence kernel association test (SKAT), weighted sum (WSS) and variable threshold (VT) for testing association of multiple rare variants in a genomic region with a quantitative trait as a function of the relative risk parameter r at the significance level α=0.05 under additive models, assuming a baseline penetrance of 1 and sample sizes of 2000.

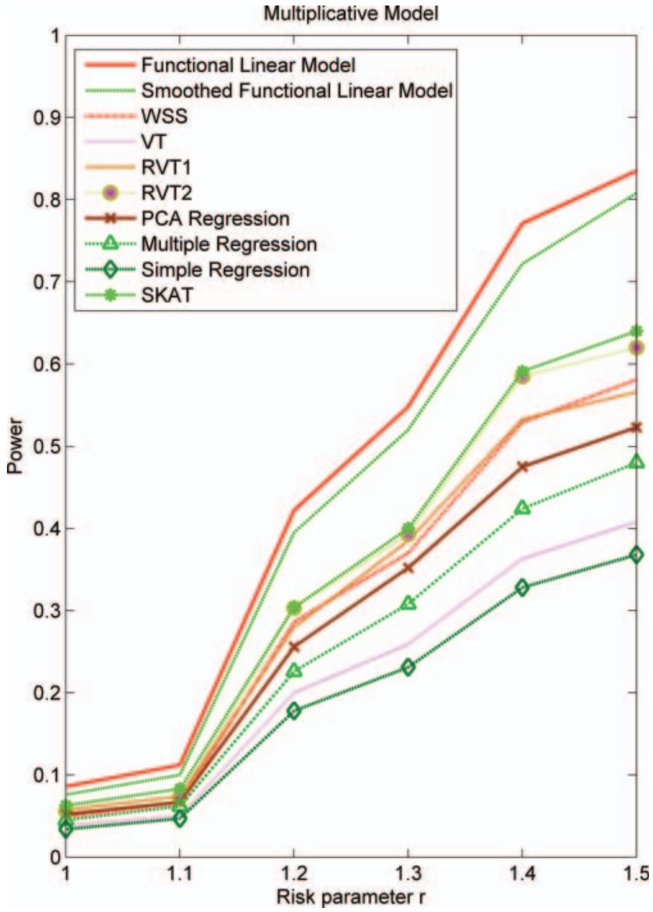

Figure 2 Power curves of 10 statistics: the functional linear model (FLM), the smoothed FLM, the two collapsing-based regression tests RVT1 and RVT2, simple regression, multiple regression, regression on principal components, sequence kernel association test (SKAT), weighted sum (WSS) and variable threshold (VT) for testing association of multiple rare variants in a genomic region with a quantitative trait as a function of the relative risk parameter r at the significance level $\alpha=0.05$ under multiplicative models, assuming a baseline penetrance of 1 and sample sizes of 2000.

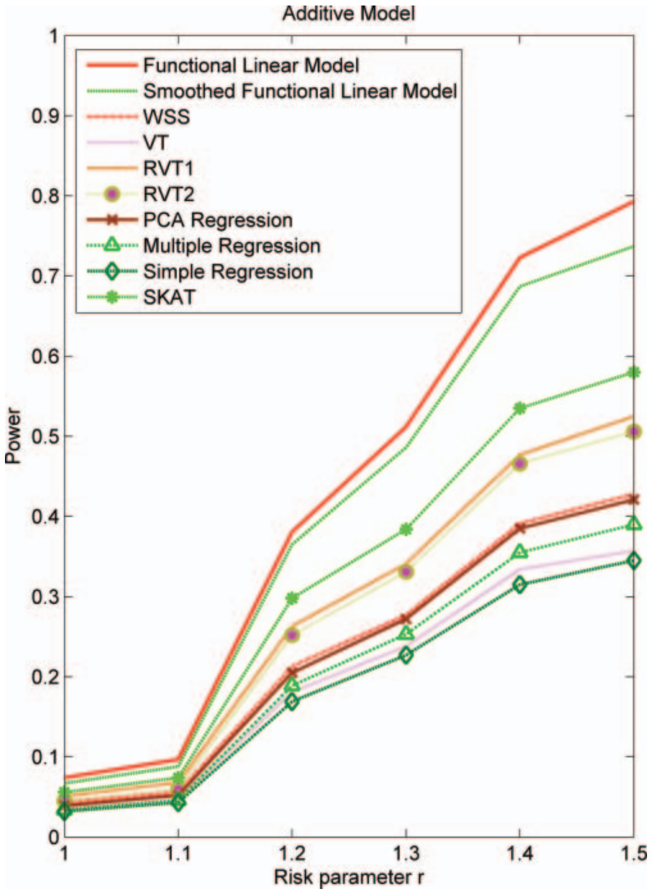

Figure 3 Power curves of 10 statistics: the functional linear model (FLM), the smoothed FLM, the two collapsing-based regression tests RVT1 and RVT2, simple regression, multiple regression, regression on principal components, sequence kernel association test (SKAT), weighted sum (WSS) and variable threshold (VT) for testing association of multiple rare variants in a genomic region with a quantitative trait in the presence of sign heterogeneity as a function of the relative risk parameter r at the significance level $\alpha=0.05$ under additive models, assuming a baseline penetrance of 1 and sample sizes of 2000.

WSS, PCA, MRG and SRG are in the bottom of power curves and they are sensitive to the presence of both risk and protective variants. In the FLM the positive and negative genetic effects are decomposed in the basis function expansion of random genotypic function and genetic effects (regression coefficient) function. The test statistic in the FLM is a quadratic function. Roughly speaking, the statistic in the FLM summarises the squared positive and negative genetic effects where positive and negative effects will not be cancelled in the statistic. The statistic in the WSS and VT is a linear combination of positive and negative effects. These positive and negative effects in the statistic will cancel each other. Therefore, the FLM still has much higher power than the other tests in the presence of both risk and protective variants.

The power of 10 statistics for testing association of 135 low frequency variants with a quantitative trait under the additive and dominant models is plotted in figure 5 and supplementary figure S5, respectively. We observe that the power pattern of 10 statistics is similar to that of 10 statistics for testing association of rare variants.

Now we study the power of 10 statistics for testing association of all variants with a quantitative trait. Figures 6 and 7, and supplementary figures S6 and S7 plot the power of 10 statistics in the presence of sign homogeneity for testing the association of 900 variants in the genomic region as a function of the risk parameter at the significance level $\alpha=0.05$ under the additive, multiplicative, dominant and recessive models, respectively. Again, we observe from these figures that the FLM and smoothed FLM have the highest power, followed by the SKAT, and RVT2. This demonstrates that the FLM can be our best choice in quantitative trait association studies, no matter whether the variants are common or rare.

The pattern of power of the 10 statistics for testing the association of all common, low frequency and rare variants in the genomic region with a quantitative trait in the presence of positive and negative signs of the effects is similar to the power for testing association of rare variants in the presence of positive and negative signs of the effects. To limit the length of this publication, they will not be presented here.

To investigate the impact of the proportion of causal risk variants on the power, figure 8 shows the power curve of 10 statistics for testing association of rare variants at the significance level $\alpha=0.05$ under the additive model as a function of the proportion of risk variants, assuming risk parameter r=1.3. We

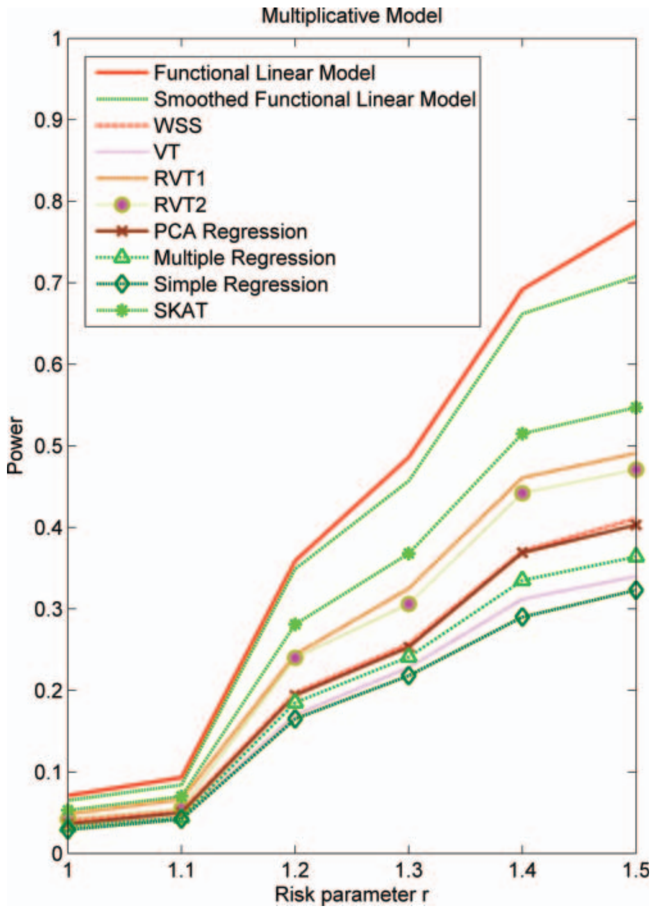

Figure 4 Power curves of 10 statistics: the functional linear model (FLM), the smoothed FLM, the two collapsing-based regression tests RVT1 and RVT2, simple regression, multiple regression, regression on principal components, sequence kernel association test (SKAT), weighted sum (WSS) and variable threshold (VT) for testing association of multiple rare variants in a genomic region with a quantitative trait in the presence of sign heterogeneity as a function of the relative risk parameter r at the significance level $\alpha=0.05$ under multiplicative models, assuming a baseline penetrance of 1 and sample sizes of 2000.

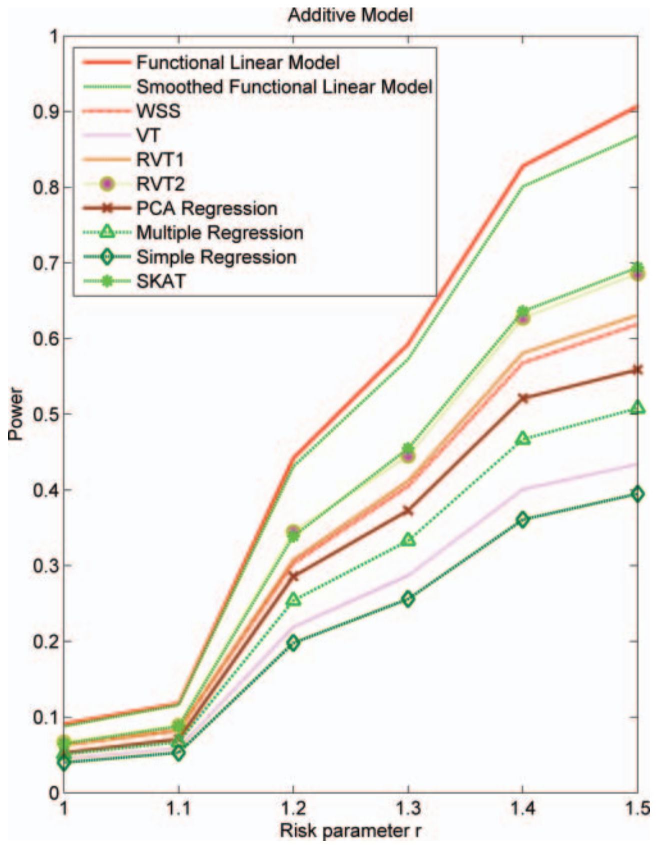

Figure 5 Power curves of 10 statistics: the functional linear model (FLM), the smoothed FLM, the two collapsing-based regression tests RVT1 and RVT2, simple regression, multiple regression, regression on principal components, sequence kernel association test (SKAT), weighted sum (WSS) and variable threshold (VT) for testing association of low frequency variants in a genomic region with a quantitative trait as a function of the relative risk parameter r at the significance level $\alpha=0.05$ under additive models, assuming a baseline penetrance of 1 and sample sizes of 2000.

observe that the FLM methods outperform other methods for every proportion of causal risk variants. This conclusion still holds for other models and types of variants (data not shown).

To investigate the impact of the density of variants on the power, we plot figure 9 and supplementary figure S8. We consider 900 variants in a gene with 30 kb; 60% of them are rare. We assume the risk parameter r=1.4. We also assume that the maximum number of SNPs per kb is 30. In figure 9, we assume sample sizes of 2000 and 10% of causal rare variants which implies that the number of casual rare variants changes as the density of variants changes. Figure 9 shows the power curve of 10 statistics as a function of the number of SNPs per kb under the additive model. We observe that as the density of variants decreases the power of all tests decreases. However, the impact of density of variants on the power of all statistics is not large. The FLM methods still substantially outperform other methods. In supplementary figure S8, we assume that the total number of risk variants is 90 and will not change as the density of variants changes. Other parameters in supplementary figure S8 is the same as that in figure 9. We observe from supplementary figure S8 that the power of all tests will not make large changes as the density of variants changes and is largely determined by the risk variants. The power of the FLM methods is still the highest.

Application to real data examples

To evaluate its performance further, the FLM was first applied to the ANGPTL4 sequence and phenotype data from the Dallas Heart Study.[44] A total of 93 variants were identified from 3553 individuals. The total number of rare variants with a minor allele frequency below 0.03 in the dataset was 71. The study included six quantitative traits: plasma triglyceride, high density lipoprotein cholesterol (HDL), total cholesterol, very low density lipoprotein cholesterol (VLDL), and body mass index (BMI). p Values from the FLM, SKAT, two collapsing-based regression tests, SRG, MRG, regression on PCAs, WSS, and VT for testing the association of rare variants in ANGPTL4 with the six quantitative traits are summarised in table 5 where p values for WSS, VT and SRG were obtained by permutations. Several features emerge from table 5. First, the FLM had the smallest p values to test for the association of the rare variants in ANGPTL4 with all six traits among nine statistics, followed by the PCA, RVT1 and RVT2. Second, the FLM can detect association of the rare variants in ANGPTL4 with all six traits, and the PCA, RT1 and RT2 can detect association with BMI, triglycerides, VLDL and HDL. SRG can detect association

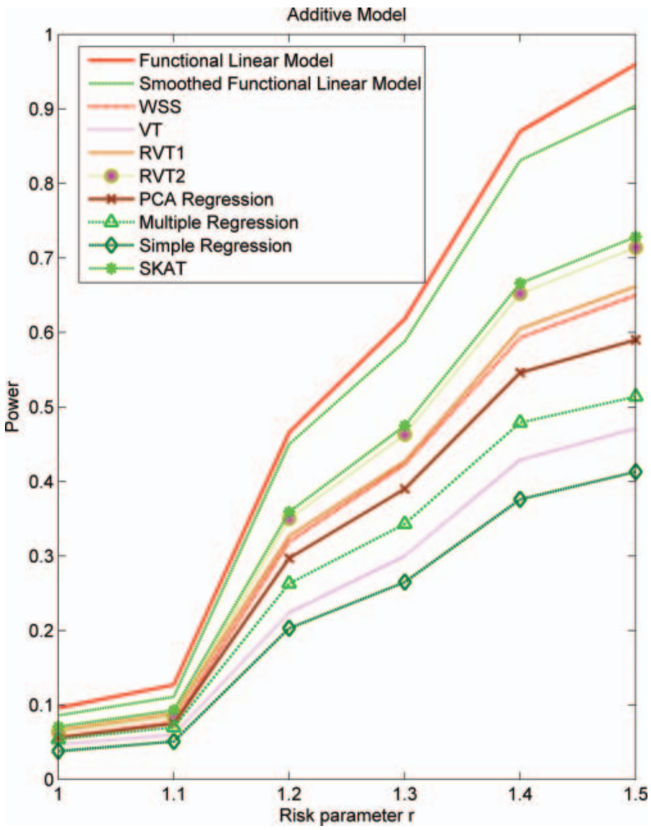

Figure 6 Power curves of 10 statistics: the functional linear model (FLM), the smoothed FLM, the two collapsing-based regression tests RVT1 and RVT2, simple regression, multiple regression, regression on principal components, sequence kernel association test (SKAT), weighted sum (WSS) and variable threshold (VT) for testing association of all common, low frequency and rare variants in a genomic region with a quantitative trait in the presence of sign heterogeneity as a function of the relative risk parameter r at the significance level α=0.05 under additive models, assuming a baseline penetrance of 1 and sample sizes of 2000.

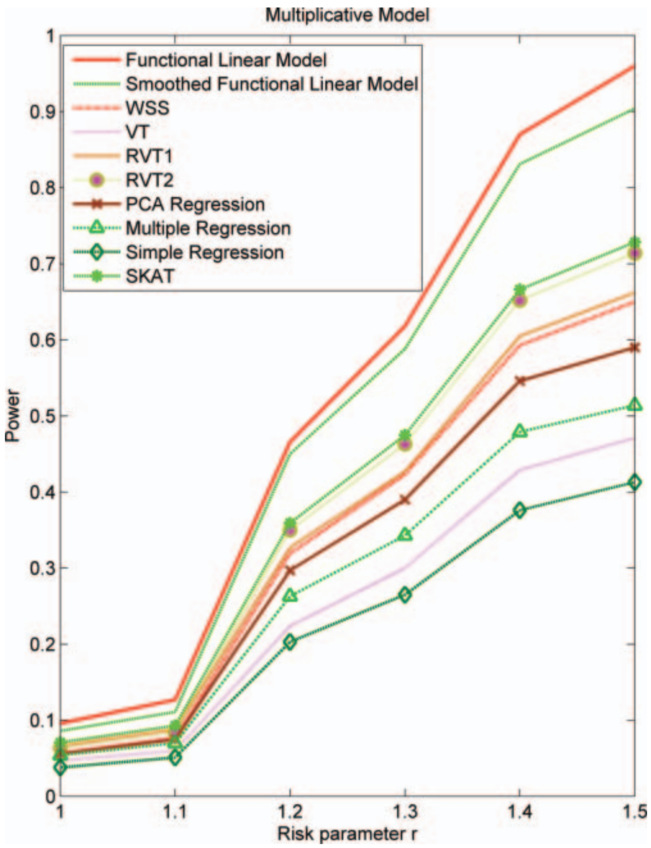

Figure 7 Power curves of 10 statistics: the functional linear model (FLM), the smoothed FLM, the two collapsing-based regression tests RVT1 and RVT2, simple regression, multiple regression, regression on principal components, sequence kernel association test (SKAT), weighted sum (WSS) and variable threshold (VT) for testing association of all common, low frequency and rare variants in a genomic region with a quantitative trait in the presence of sign heterogeneity as a function of the relative risk parameter r at the significance level α=0.05 under multiplicative models, assuming a baseline penetrance of 1 and sample sizes of 2000.

with BMI and HDL, MRGs can detect association with BMI, triglycerides and HDL, SKAT can detect association with BMI, but WSS and VT cannot detect any association of the rare variants in ANGPTL4 with the six traits. Third, we also observed that p values by the FLM ($p<5.48\times10^{-7}$), PCA ($p<2.58\times10^{-6}$) and MRG ($p<0.0032$) for testing association of the rare variants in ANGPTL4 with triglyceride were much smaller than the p value (0.016) in their original studies.[44] Fourth, the p values by the FLM for testing the association of the rare variants in ANGPTL4 with all six traits were much smaller than that by all eight statistics with phenotype extreme selection design (where the individuals whose phenotypic values ≤25th centile were classified as the low quartile group and the individuals whose phenotypic values ≥75th centile were grouped as the high quartile group) in the paper by Luo *et al*.[23]

We also tested the association of all variants in ANGPTL4 with six traits in the Dallas Heart Study. The results are summarised in table 6. Again, we observed that in this scenario, the FLM still outperforms the other statistics.

To investigate the performance of the FLM for the genome-wide quantitative trait association studies, the FLM was then applied to rare expression quantitative trait loci (eQTLs) analysis. We analysed genetic variation in the low coverage resequencing data of the 1000 Genomes Project (released March 2010) and 15 gene expressions acquired by RNA sequencing (RNA-seq) in LCLs from 60 individuals of European origin (CEU).[45] The expression of a gene is measured by a normalised overall expression level of the gene. We take a gene as a unit of rare eQTL association analysis. A total of 2533 genes that consisted of SNPs with minor allele frequency (MAF) <5% were included in the analysis. A p value for declaring significant association after Bonferroni correction for multiple tests was 1.97×10^{-5}. Nine statistics—the FLM, SKAT, two collapsing-based regression tests, SRG, MRG, regression on PCAs, WSS, and VT—were used to test for the association of rare variants in 2533 genes with 15 RNA-seq expressions. The FLM identified 13 genes, SKAT identified three genes, and SRG, MRG, PCA, WSS and VT identified one gene with rare SNPs which were significantly associated with the expressions of 13 genes after Bonferonni correction for multiple tests. It was reported that one SNP (rs7639979) in gene PRSS50 was significantly associated with expression of PRSS46 with a p value $<1.76\times10^{-6}$.[48] The p values of 14 genes calculated by the FLM and other eight statistics for testing their association with the whole-gene expressions are summarised in table 7. Table 7 clearly demonstrates that the FLM substantially outperforms the other seven existing statistics. In addition to 13

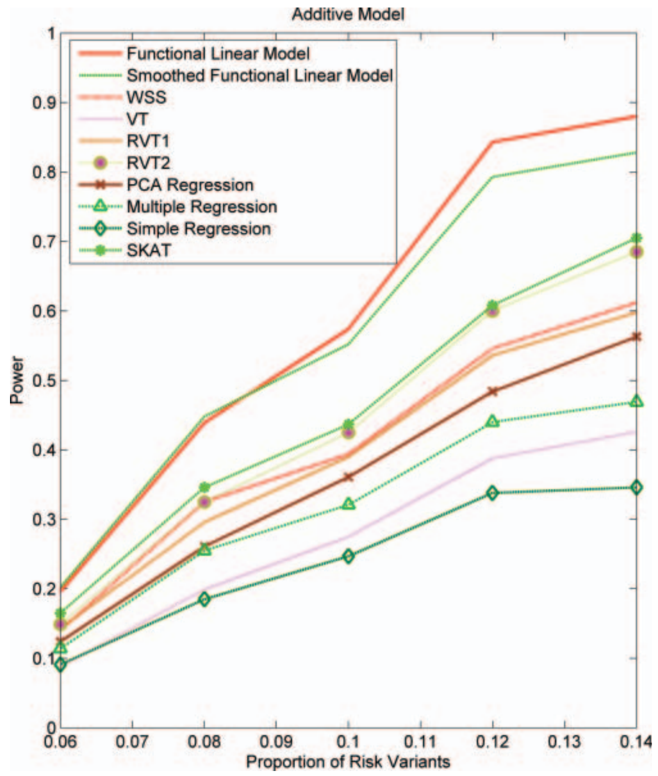

Figure 8 Power curves of 10 statistics: the functional linear model (FLM), the smoothed FLM, the two collapsing-based regression tests RVT1 and RVT2, simple regression, multiple regression, regression on principal components, sequence kernel association test (SKAT), weighted sum (WSS) and variable threshold (VT) for testing association of rare variants in a genomic region with a quantitative trait as a function of the proportion of risk variants at the significance level $\alpha=0.05$ under additive models, assuming a baseline penetrance of 1, risk parameter r=1.3 and sample sizes of 2000.

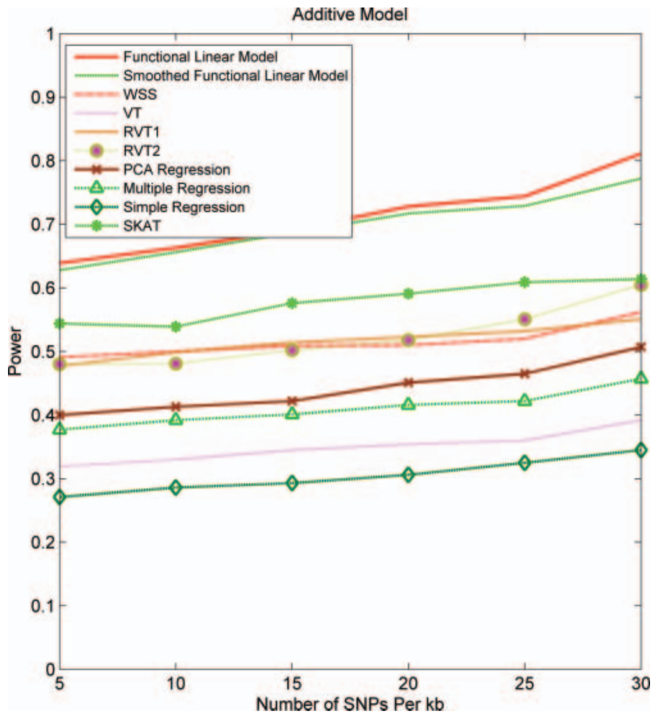

Figure 9 Power curves of 10 statistics: the functional linear model (FLM), the smoothed FLM, the two collapsing-based regression tests RVT1 and RVT2, simple regression, multiple regression, regression on principal components, sequence kernel association test (SKAT), weighted sum (WSS) and variable threshold (VT) for testing association of rare variants in a genomic region with a quantitative trait as a function of the number of single nucleotide polymorphisms (SNPs) per kb at the significance level $\alpha=0.05$ under additive models, assuming a baseline penetrance of 1, risk parameter r=1.4, the maximum number of SNPs per kb is 30, 10% of risk variants, and sample sizes of 2000.

significantly associated genes, table 7 also lists gene OR4Q3 that was not associated with expression of gene MADD by all statistics as a 'control' to show that the FLM will not inflate false positive signal. To illustrate this further, we present quartile-quartile (QQ) plot for p values to test the association of 2533 genes with the MADD expressions by the FLM in supplementary figure S9. It clearly demonstrates that the FLM will not inflate false positive signal in eQTL analysis. Table 8 lists the required computational time of six test statistics for testing the association of a total of 2533 genes with the MADD expressions on intel Core i7-2600 at 3.4 GHz with 16.0 GB memory (Windows 7). The computational time of the FLM is 1 h more than that of the SKAT, but much less than that of MRG and SRG with permutations.

DISCUSSION

NGS can generate several millions of genetic variation data. As a consequence, these genetic variation data are so densely distributed across the genome that the genetic variants can be considered as genomic variation observations varying over a continuum that views a genomic region as a variant site continuously changing interval. As Haldane[49] and Fisher[50] recognised in the last century, the genome can be modelled as a continuum. Specifically, the genome is not purely a collection of independent segregating sites. Rather, the genome is transmitted not in points, but in segments. Instead of modelling the genome as a few separated individual loci, modelling the genome as a continuum where the observed genetic variant function can be viewed as a realisation of the stochastic process in the genome and modelled as a function of genomic location will enrich information on genetic variation across the genome. As Ferraty and Romain pointed out in their book,[37] standard multivariate statistical analysis often fails with functional data. The emergence of NGS demands a paradigm shift in analytic methods for QTL analysis from standard multivariate data analysis to functional

Table 5 p Values of statistics for testing association of rare variants in ANGPTL4 with six traits in the Dallas Heart Study

	Phenotypes					
Statistical methods	BMI	Cholesterol	Triglycerides	VLDL	LDL	HDL
FLM	3.72E-06	4.81E-02	5.48E-07	3.67E-06	1.77E-07	4.27E-06
RVT1	4.18E-06	7.74E-01	2.21E-03	2.13E-03	7.29E-01	2.13E-03
RVT2	6.93E-05	7.18E-01	1.82E-04	1.89E-04	8.93E-01	6.84E-04
PCA	3.93E-05	7.55E-01	2.58E-06	4.59E-06	5.86E-01	4.49E-03
Multiple regression	9.25E-03	1.93E-01	3.19E-03	3.36E-03	2.54E-07	1.02E-02
Simple regression	1.10E-02	7.10E-01	5.50E-02	1.60E-01	1.60E-01	5.00E-03
VT	1.20E-05	2.65E-01	4.95E-01	7.25E-01	3.10E-01	3.80E-01
WSS	5.39E-02	3.55E-01	3.00E-01	8.00E-02	4.60E-01	2.50E-01
SKAT	1.03E-03	7.49E-01	1.57E-01	1.88E-01	2.34E-01	1.05E-01

BMI, body mass index; FLM, functional linear model; HDL, high density lipoprotein; PCA, principal components; SKAT, sequence kernel association test; VLDL, very low density lipoprotein; VT, variable threshold; WSS, weighted sum.

Table 6 p Values of statistics for testing association of all variants in ANGPTL4 with six traits in the Dallas Heart Study

Statistical methods	Phenotypes BMI	Cholesterol	Triglycerides	VLDL	LDL	HDL
FLM	1.35E-08	7.40E-04	1.93E-09	5.54E-10	7.58E-04	8.40E-10
RVT1	3.85E-03	3.26E-03	2.21E-03	2.13E-03	7.29E-01	2.13E-03
RVT2	2.41E-02	7.08E-03	1.82E-04	1.89E-04	8.93E-01	6.84E-04
PCA	3.18E-08	1.36E-03	9.82E-08	1.40E-09	1.05E-02	1.34E-09
Multiple regression	2.37E-04	3.48E-02	1.58E-04	1.48E-02	1.04E-03	2.83E-04
Simple regression	5.37E-05	3.58E-05	6.96E-10	7.59E-06	5.92E-04	3.17E-07
VT	1.10E-02	4.75E-01	9.80E-01	8.42E-01	5.35E-01	6.99E-03
WSS	3.00E-03	8.61E-01	6.34E-01	9.58E-01	8.42E-01	1.80E-02
SKAT	3.74E-07	3.91E-03	4.95E-07	3.40E-08	7.82E-03	6.66E-09

BMI, body mass index; FLM, functional linear model; HDL, high density lipoprotein; PCA, principal components; SKAT, sequence kernel association test; VLDL, very low density lipoprotein cholesterol; VT, variable threshold; WSS, weighted sum.

data analysis. The purpose of this study was to explore existing and newly proposed FLM for QTL analysis with large-scale DNA sequence data and address great challenges arising from the rapid development of NGS.

To utilise the advantages of both single variant analysis and collapsing methods, the highly dense features of NGS genetic variation data distribution across the genome, and address the limitations inherent by single variant analysis and collapsing methods, we view the genome as a continuum and variants in the genome as a realisation of a stochastic process which can be modelled as a random function of genomic position. We then employ FLM to carry out quantitative trait association studies for next-generation sequencing. In this paper, by large-scale simulation and real QTL and eQTL data analysis, we showed that in all scenarios the FLM largely outperform several existing statistical methods for QTL analysis. We also demonstrated that FLM can be used to test the association of the entire allelic spectrum of genetic variation with a quantitative trait and has several remarkable features.

First, the FLM is a natural extension of MRGs. Indeed, replacing summation in equation (9) by an integral will result in the FLM. Second, unlike simple regressions and MRGs discarding a large amount of information due to using limited numbers to summarise the data, the FLM preserves the intrinsic correlation structure in the data and all the positional-level genetic information. Third, the space-ordering of the genetic variation data is a central feature in the FLM. The neighbouring genetic variants are linked. The genotypes at one SNP are dependent on the genotypes at nearby SNPs. This is a key difference between the FLM and other existing methods. The most popular QTL analysis methods will not account for the space-ordering of the data. The FLM simultaneously employs genetic information of the individual variants and correlation information (LD) among all variants. It views the genetic variation across the genomic region as a function of its genomic location and uses intrinsic functional dependence structure of the data and all available genetic information of the variants in the genomic region. Fourth, it can be shown that the test statistic is proportional to the integral of the square of the genetic additive effect function. The square of effects will transform the negative effects to the positive. In the test statistic, the effects with signs in both positive and negative directions will become effects with only one positive direction. The FLM can eliminate the sign heterogeneity in the test. It is also clear that the difference in effect sizes among variants will be taken into account in the integral of the square of the genetic effect functions across the genomic region. Therefore, the size heterogeneity will also be incorporated into the test in the FLM. Fifth, random genetic variant function is flexible. The variable $x_i(t)$ at the individual variant site can take integer values to code alleles or genotypes, or real numbers to represent the number of reads of the sequences, the probability of SNP call, and the probability of the variant being functional or weights at the variant site. The FLM can use various types of genetic variant data and can be extended to testing the

Table 8 The computational times of six statistics for testing the association of a total of 2533 genes with the MADD expressions across the genome

Test statistics	Time
FLM	6 h 47 min 10 s
RVT1	4 min 29.2 s
RVT2	4 min 44.4 s
SKAT	5 h 42 min 38.4 s
Simple regression with permutation	16 h 50 min 32 s
Multiple regression	12 h 39 min 8 s

FLM, functional linear model; SKAT, sequence kernel association test.

Table 7 p Values of 14 genes associated with the expressions of 13 genes

Gene expressions	Gene	FLM	RVT1	RVT2	SRG	MRG	PCA	WSS	VT	SKAT
MADD	PRR19	1.93E-05	4.86E-01	7.95E-01	2.97E-02	1.68E-03	3.66E-04	1.09E-01	6.93E-02	5.73E-05
MADD	OR4Q3	5.83E-01	6.82E-01	6.82E-01	8.01E-02	4.61E-01	5.84E-01	4.46E-01	6.44E-01	1.27E-01
MAST2	C12orf53	9.14E-08	1.28E-01	2.23E-01	5.94E-02	9.33E-05	5.63E-03	3.96E-02	6.93E-02	1.22E-05
PDHB	PLAC2	6.43E-06	3.55E-01	6.62E-01	9.90E-02	4.20E-03	1.92E-01	4.16E-01	4.95E-01	4.62E-04
RP11-364B6.1	KLK11	5.33E-05	4.19E-01	3.69E-01	8.91E-02	4.27E-02	5.46E-02	1.39E-01	3.17E-01	5.97E-04
RP4-773A18.2	NDUFB7	6.69E-07	5.64E-01	2.14E-01	4.65E-01	3.94E-04	5.75E-02	2.57E-01	2.67E-01	1.94E-04
RP4-773A18.2	C22orf27	1.18E-05	9.03E-03	6.70E-03	1.98E-02	6.68E-04	2.56E-02	9.90E-03	3.96E-02	4.04E-04
TRBJ2-1	CHRNE	3.26E-06	9.08E-01	6.04E-01	1.98E-02	2.66E-03	3.31E-03	9.90E-03	9.90E-02	4.59E-05
TRBJ2-1	FUT1	1.64E-05	3.04E-01	6.98E-01	2.97E-02	5.55E-03	8.93E-04	9.90E-03	5.94E-02	1.10E-04
TRBJ2-2	FFAR3	3.72E-06	5.71E-01	3.79E-01	3.27E-01	4.44E-03	2.64E-01	2.08E-01	1.68E-01	8.42E-04
TRBJ2-2P	CCL11	1.52E-05	5.33E-05	5.85E-05	9.90E-03	1.43E-03	3.69E-03	9.90E-03	9.90E-03	1.24E-04
TRBJ2-3	GGT6	1.13E-05	3.94E-01	5.74E-01	8.91E-02	9.88E-04	6.27E-03	5.94E-02	1.39E-01	2.64E-04
TRBJ2-5	C19orf46	6.79E-11	7.62E-02	7.62E-02	2.97E-02	2.55E-04	1.99E-02	2.77E-01	9.90E-03	6.15E-07
PRSS46	PRSS50	8.65E-09	5.55E-05	4.26E-04	1.00E-05	5.68E-06	1.49E-07	1.00E-05	1.00E-05	3.25E-07

FLM, functional linear model; MRG, multiple regression; PCA, principal components; SKAT, sequence kernel association test; SRG, simple regression; VT, variable threshold; WSS, weighted sum.

association of copy number variations (CNVs) with a quantitative trait. They can also incorporate the functional prediction of the variants into the tests. Sixth, genetic variant data in a genomic region which often have strong LD generate multicolinearity and high dimensionality, which other methods are often unable to deal with efficiently. In the FLM, the genetic variant functions and genetic effect functions are expended in terms of orthogonal or closely to orthogonal functions. The component coefficients in the expansion will, in general, not be linearly dependent. Therefore, the multicolinearity problem in the FLM is eliminated. Seventh, we also note that the FLM utilises the data reduction techniques to compress high dimensional data into a few components. Therefore, we can expect that the FLM will have high power to detect association of the genomic regions with a quantitative trait. Through extensive simulations and real data analysis, we clearly demonstrated that in any scenario, the FLM has much higher power to identify the genomic regions associated with the quantitative trait than the current existing tests. Eighth, the FLM use data reduction techniques that compress noised high dimensional genomic data into a few components and smoothing techniques will reduce the effects of noise.

Next-generation sequencing technologies will identify tens of millions of genetic variants across the human genome. Such extremely high-dimensional data that are full of noise and missing data pose fascinating statistical and computational challenges in QTL analysis. Transition of analysis from low dimensional data to extremely high dimensional data demands changes in statistical methods from multivariate data analysis to functional data analysis. In the past decade we have witnessed the emergence of the functional data analysis as an exciting research area of statistics which provides powerful and informative tools for the analysis of various types of high dimensional data including genomic variation. Our theoretic results, preliminary QTL and eQTL real data analysis and simulations showed that the FLM for QTL analysis is able to fully explore all of the information contained in the data, efficiently utilise the merits of both variant-by-variant and collective analyses, while overcoming their limitations. Therefore, the FLM is one of the choices in QTL analysis with next-generation sequencing data. Application of the genomic continuum model and functional data analysis is expected to open a new avenue for quantitative genetic studies. The results in this paper are preliminary. The purpose of this paper is to stimulate further discussions regarding the great challenges we are facing in the quantitative trait studies of high dimensional genomic data produced by next-generation sequencing.

Acknowledgements We thank the anonymous reviewers for their helpful comments.

Contributors MX contributed to the conception and design of the study and wrote the manuscript, LL and YZ performed analysis. All were involved in critical review of the manuscript for intellectual content.

Funding The project described was supported by Grant 1R01AR057120—01, 1R01HL106034-01, P01 AR052915-01A1 and P50 AR054144-01 CORT from National Institutes of Health and NIAMS.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Drmanac R,** Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, Dahl F, Fernandez A, Staker B, Pant KP, Baccash J, Borcherding AP, Brownley A, Cedeno R, Chen L, Chernikoff D, Cheung A, Chirita RCurson B, Ebert JC, Hacker CR, Hartlage R, Hauser B, Huang S, Jiang Y, Karpinchyk V, Koenig M, Kong CLanders T, Le C, Liu J, McBride CE, Morenzoni M, Morey RE, Mutch K, Perazich H, Perry K, Peters BA, Peterson J, Pethiyagoda CL, Pothuraju K, Richter C, Rosenbaum AM, Roy S, Shafto J, Sharanhovich U, Shannon KW, Sheppy CG, Sun M, Thakuria JV, Tran A, Vu D, Zaranek AW, Wu X, Drmanac S, Oliphant AR, Banyai WC, Martin B, Ballinger DG, Church GM, Reid CA. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 2010;**327**:78–81.
2. **Metzker ML.** Sequencing technologies-the next generation. *Nat Rev Genet* 2010;**11**:31–46.
3. **Rakyan VK,** Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nature Rev Gene* 2011;**12**:529–41.
4. **Bansal V,** Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 2010;**11**:773–85.
5. **Chaisson MJ,** Brinza D, Pevzner PA. De novo fragment assembly with short mate-paired reads: does the read length matter? *Genome Res* 2009;**19**:336–46.
6. **Johnson PL,** Slatkin M. Accounting for bias from sequencing error in population genetic estimates. *Mol Biol Evol* 2008;**25**:199–206.
7. **Lynch M.** Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* 2009;**182**:295–301.
8. **Pool JE,** Hellmann I, Jensen JD, Nielsen R. Population genetic inference from genomic sequence variation. *Genome Res* 2010;**20**:291–300.
9. **Bacanu SA,** Nelson MR, Whittaker JC. Comparison of methods and sampling designs to test for association between rare variants and quantitative traits. *Genet Epi* 2011;**35**:226–35.
10. **Shi G,** Rao DC. Optimum designs for next-generation sequencing to discover rare varients for common complex disease. *Genet Epi* 2011;**35**:572–9.
11. **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**:311–21.
12. **Madsen BE,** Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009;**5**:e1000384.
13. **Price AL,** Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010;**86**:832–8.
14. **Li Y,** Byrnes AE, Li M. To identify associations with rare variants, Just WhaIT: weighted haplotype and imputation-based tests. *Am J Hum Genet* 2010;**87**:728–35.
15. **King CR,** Rathouz PJ, Nicolae DL. An evolutionary framework for association testing in resequencing studies. *PLoS Genet* 2010;**6**:e1001202.
16. **Yi N,** Zhi D. Bayesian analysis of rare variants in genetic association studies. *Genet Epidemiol* 2011;**35**:57–69.
17. **Han F,** Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 2010;**70**:42–54.
18. **Neale BM,** Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. *PLoS Genet* 2011;**7**:e1001322.
19. **Ionita-Laza I,** Buxbaum JD, Laird NM, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet* 2011;**7**:e1001289.
20. **Hoffmann TJ,** Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. *PLoS One* 2010;**5**:e13584.
21. **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.
22. **Mukhopadhyay I,** Feingold E, Weeks D, Thalamuthu A. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genetic Epi* 2010;**34**:213–21.
23. **Luo L,** Boerwinkle E, Xiong M. Association studies for next-generation sequencing. *Genome Res* 2011;**21**:1099–108.
24. **Morris AP,** Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epi* 2010;**34**:188–93.
25. **Wu MC,** Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011;**89**:82–93.
26. **Ramsay JO,** Silverman BW. *Functional data analysis*. New York: Springer, 2005.
27. **Baillo A,** Grane A. Local linear regression for functional predictor and scalar response. *J Multivariate Anal* 2009;**100**:102–11.
28. **Coffey N,** Harrison AJ, Donoghue OA, Hayes K. Common functional principal components analysis: a new approach to analyzing human movement data. *Hum Mov Sci* 2011;**30**:1144–66.
29. **Müller HG,** Yao F. *Regressing longitudinal response trajectories on a covariate*. London: Imperial College Press, 2006:305–24.
30. **Horvath L,** Kokoszka P, Reimherr M. Two samples inference in functional linear models. *Can J Stat* 2009;**37**:571–91.
31. **Besse P,** Cardot H, Stephenson D. Autoregressive forecasting of some functional climatic variations. *Scand J Stat* 2000;**27**:673–87.
32. **Ando T,** Imoto S, Miyano S. Functional data analysis of the dynamics of gene regulatory networks. In: 'Knowledge Exploration in Life Science Informatics'. *Lecture notes in computer science*. Vol. 3303. Heidelberg, Berlin: Springer-Verlag, 2004:69–83.

33. **Morris JS,** Arroyo C, Coull BA, Ryan LM, Herrick R, Gortmaker SL. Using wavelet based functional mixed models to characterize population heterogeneityin accelerometer profiles: a case study. *J Am Stat Assoc* 2006;**101**:1352–64.
34. **Müller HG,** Stadtmüller U, Yao F. Functional variance processes. *J Am Stat Assoc* 2006;**101**:1007–18.
35. **Hlubinka D,** Prchal L. Changes in atmospheric radiation from the statistical point of view. *Comput Stat Data Anal* 2007;**51**:4926–41.
36. **Joyce P,** Tavare S. The distribution of rare alleles. *J Math Biol* 1995;**33**:602–18.
37. **Ferraty F,** Romain Y. *The oxford handbook of functional data analysis.* New York: Oxford University Press, 2010.
38. **Febrero-Bande M,** Galeano P, Gonzalez-Manteiga W. Measures of influence for the functional linear model with scalar response. *J Multivariate Anal* 2010;**101**:327–39.
39. **Hastie T,** Mallows C. A discussion of a statistical view of some chemometrics regression tools' by Frank IE and Friedman JH. *Technometrics* 1993;**35**:140–3.
40. **Marx BD,** Eilers PD. Generalized linear regression on sampled signals and curves: a p-spline approach. *Technometrics* 1999;**41**:1–13.
41. **Cardot H,** Ferraty F, Mas A, Sarda P. Testing hypothesis in the functional linear model. *Scand J Stat* 2003;**30**:241–55.
42. **Cardot H,** Ferraty F, Sarda P. Functional linear model. *Stat Prob Lett* 1999;**45**:11–22.
43. **Frank IE,** Friedman JH. A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**:140–3.
44. **Romeo S,** Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, Cohen JC. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat Genet* 2007;**39**:513–16.
45. **Montgomery SB,** Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 2010;**464**:773–7.
46. **Nelson MR,** Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D, Warren L, Aponte J, Zawistowski M, Liu X, Zhang H, Zhang Y, Li J, Li Y, Li L, Woollard P, Topp S, Hall MD, Nangle K, Wang J, Abecasis G, Cardon LR, Zöllner S, Whittaker JC, Chissoe SL, Novembre J, Mooser V. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 2012;**337**:100–4.
47. **Hudson R.** Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 2002;**18**:337–8.
48. **Pickrell JK,** Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010;**464**: 768–72.
49. **Haldane JBS.** The combination of linkage values and the calculation of distance between the loci of linked factors. *J Genet* 1919;**8**:299–309.
50. **Fisher RA.** *The theory of inbreeding*. Edinburgh: Oliver and Boyed, 1949.

JMG

Quantitative trait locus analysis for next-generation sequencing with the functional linear models

Li Luo, Yun Zhu and Momiao Xiong

J Med Genet 2012 49: 513-524
doi: 10.1136/jmedgenet-2012-100798