Bioinformatics Advance Access published October 12, 2015

Genome analysis

Unified tests for fine scale mapping and identifying sparse high-dimensional sequence associations

Shaolong Cao[1,2], Huaizhen Qin[2,3], Alexej Gossmann[2,4], Hong-Wen Deng[2,3] and Yu-Ping Wang[1,2,3*]

[1]Department of Biomedical Engineering, [2]Center for Bioinformatics and Genomics, [3]Department of Biostatistics and Bioinformatics, and [4]Department of Mathematics, Tulane University, New Orleans, LA, USA

Associate Editor: Dr. John Hancock

ABSTRACT

Motivation: In searching for genetic variants for complex diseases with deep sequencing data, genomic marker sets of high-dimensional genotypic data and sparse functional variants are quite common. Existing sequence association tests are incapable of identifying such marker sets and individual causal loci, although they appeared powerful to identify small marker sets with dense functional variants. In sequence association studies of admixed individuals, cryptic relatedness and population structure are known to confound the association analyses.

Method: We here propose a unified marker wise test (uFineMap) to accurately localize causal loci and a unified high-dimensional set based test (uHDSet) for identifying high-dimensional sparse associations in deep sequencing genomic data of multi-ethnic individuals with random relatedness. These novel tests are based on scaled sparse linear mixed regressions with L_p ($0<p<1$) norm regularization. They jointly adjust for cryptic relatedness, population structure and other confounders to prevent false discoveries and improve statistical power for identifying promising individual markers and marker sets that harbor functional genetic variants of a complex trait.

Results: With large scale simulation data and real data analysis, the proposed tests appropriately controlled Type I error rate and appeared more powerful than several prominent methods. We illustrated their practical utilities by the applications to DNA sequence data of Framingham Heart Study for osteoporosis. The proposed tests identified 11 novel significant genes that were missed by the prominent famSKAT and GEMMA. In particular, four out of six most significant pathways identified by the uHDSet but failed by famSKAT have been reported to be related to BMD or osteoporosis in the literature.

Availability: The computational toolkit is available for academic use: https://sites.google.com/site/shaolongscode/home/uhdset

Contact: wyp@tulane.edu

1. INTRODUCTION

Deep sequencing technologies have been generating huge amounts of data of rare and common DNA sequence variants. A number of sequence association tests have been developed for identifying marker sets (e.g., a group of SNPs or CNVs) that contain functional genetic variants. Most of them, however, do not jointly model cryptic relatedness, population structure and other covariates. With the growing demand of analyzing next generation sequencing data of multi-ethnic individuals, linear mixed models have become popular because of their demonstrated effectiveness in accounting for sample relatedness (Amos, 1994) and population structure which occurs when there are large-scale systematic differences in genetic ancestry among individuals in a sample. Typical examples include individuals with various levels of immigrant ancestry and more recent shared ancestors than one would expect in a homogenies population. Cryptic relatedness, refers to the presence of relatives in a sample of ostensibly unrelated individuals, could pose more serious confounding than population structure (Devlin and Roeder, 1999), especially for samples from small and isolated populations (Voight and Pritchard, 2005). Accounting for population structure is more challenging when family structure or cryptic relatedness is also present (Price, et al., 2010). We paved the way to correct for the effects of both confounders jointly.

Within the framework of linear mixed models, famSKAT (Chen, et al., 2013) and GEMMA (Zhou and Stephens, 2012) appeared as two powerful sequence association tests for identifying small marker sets that harbor dense functional genetic variants. FamSKAT is a set based test which is an extension of SKAT to be applicable to family data. GEMMA is a computationally efficient method for fitting multivariate linear mixed models. These prominent tests require that the number of markers in a testing set is much smaller than the sample size. However, in population deep sequencing studies, one encounters quite often high dimensional data sets (HDS), where the number of marker loci is larger than the sample size and the number of functional variants is very small. The aforementioned tests are incapable of identifying the functional variants on such sparse HDS. With high-dimensional sparse functional marker data sets, the aforesaid tests are incapable to identify them. Some sparse regression methods were developed to localize individual functional markers from high-dimensional marker sets, jointly modeling pedigree structure and population structure. They include Lasso (Rakitsch, et al., 2013), Ridge regression (Endelman, 2011), Elastic-net (Zou and Hastie, 2005) and the USR that we proposed recently (Cao, et al., 2014). However, these methods yield biased solutions and are ineffective to prevent false discoveries of random markers and high-dimensional marker sets irrelevant to functional variants.

In this article, we first present a unified test (uFineMap) for accurately localizing causal loci. The uFineMap is a marker wise test under a scaled sparse linear mixed regression, which jointly models marker wise effect, relatedness and population stratification. It applies scaled L_p ($0 < p < 1$) norm regularization to generate a *de-biased* solution. Next, we present an additional significant test (uHDSet) for identifying high-dimensional sparse associations in deep sequencing genomic data of related individuals. The uHDset integrates the marker wise statistics of the uFineMap to identify susceptible high-dimensional marker sets. In the uHDSet, the dependence among markers is modeled to appropriately control set-based Type I error rates. Under extensive simulations, the uFineMap outperformed the GEMMA (Zhou and Stephens, 2012) and a Scaled Lasso based method (Javanmard and Montanari, 2014). The uHDSet yields higher statistical power than famSKAT and GEMMA. Applications to Framingham Heart Study also show that our method yields novel interesting candidate genes and pathways for follow-up studies, showing its advantages over the two compared prominent alternative methods. Finally, caveats of the proposed methods and perspective future efforts are discussed.

2. METHODS

Our method focuses on constructing statistical tests for high-dimensional genetic data with cryptic relatedness. We propose two significance tests: uFineMap test (single marker/variant test) and uHDSet test (unified high-dimensional set test or whole regional test). Similar to (Bühlmann, 2013; Javanmard and Montanari, 2014), we develop uFineMap significance test for single variants based on the scaled sparse regression (Sun and Zhang, 2012), which is a generalization of ordinary sparse regression. Furthermore, we build new statistics for the uHDSet test based on a combination of marker wise statistics. The uHDSet test facilitates us to identify susceptible genes or genetic regions instead of single variants.

2.1 Unified scaled L_p norm regularized regression

At first, we need to define some basic notations. Let *n* denote the number of subjects; *m* denotes the number of independent variables (SNPs); and *L* represents the number of covariates. Suppose we have dependent variable $\mathbf{Y}=(y_1,y_2,...,y_n)^T$, which stands for phenotype for each subject. $\mathbf{X}=(x_1,x_2,...,x_n)$ is a *n*x*m* matrix where the row $x_i=(x_{i1},x_{i2},..,x_{im})^T$ represents genotype data for the *i*th subject. Typically, genotypes are coded as 0, 1 or 2 which denote the number of copies the minor allele. $\mathbf{\Phi}_{n\times n}=(\varphi_{i,j})$ is the kinship matrix or IBD (identity-by-descent) matrix. The kinship coefficient $\varphi_{i,j}$ measures the relatedness between individual *i* and *j*. $\mathbf{W}=(w_1,w_2,...,w_n)$ is an *n*x*L* matrix, where $w_i=(w_{i1},w_{i2},...,w_{iL})^T$ represents the covariates, e.g., age, sex, height, and weight.

We assume that the phenotypes, genotypes and covariates are associated with the following linear mixed model:

$$\mathbf{Y}=\mathbf{W\alpha}+\mathbf{X\beta}+\boldsymbol{\varepsilon} \quad (1)$$

where $\boldsymbol{\varepsilon}\sim N(\mathbf{0},\mathbf{\Sigma}=\sigma_\Phi^2\mathbf{\Phi}+\sigma_e^2\mathbf{I}_n)$, $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,...,\alpha_L)^T$ and $\boldsymbol{\beta}=(\beta_1,\beta_2,...,\beta_m)^T$ are the corresponding regression coefficients. Both Emma and Gemma methods can evaluate the variance component ratio $\sigma_\Phi^2/\sigma_\varepsilon^2$ of covariant matrix $\mathbf{\Sigma}$. In this paper, we use Gemma method to evaluate the $\sigma_\Phi^2/\sigma_\varepsilon^2$ ratio.

In model (1), the regression coefficients $\boldsymbol{\beta}=(\beta_1,\beta_2,...,\beta_m)^T$ represent the effect of variants which is the most important variable we are interested in. However, the high-dimensionally of genetic data will easily lead to over-fitting problem under regular regression model. To overcome this issue, a general form of the unified sparse regression model with L_p ($0<p<1$) norm regularization was proposed by USR paper with the following minimization problem(Cao, et al., 2014):

$$(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}})=\arg\min_{\boldsymbol{\beta}\in R^m,\boldsymbol{\alpha}\in R^L}(\mathbf{Y}-\mathbf{W\alpha}-\mathbf{X\beta})^T\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{W\alpha}-\mathbf{X\beta})+\lambda\|\boldsymbol{\beta}\|_p^p \quad (2)$$

where the L_p ($0<p<1$) norm regularization is defined by

$$\lambda\|\boldsymbol{\beta}\|_p^p=\lambda\sum_{i=1}^m|\beta_i|^p,\ 0<p<1$$

As is well known (Cao, et al., 2014; Chen, et al., 2010; XU, et al., 2012) that L_p ($0<p<1$) norm regularization results in a sparser solution than L1 norm regularization, which was widely popularized by the Lasso (least absolute shrinkage and selection operator)(Tibshirani, et al., 1996). In particular, previous simulation results in (Cao, et al., 2014) suggest the use of the $L_{0.3}$ norm regularization, in order to achieve a proper sparsity level of the solution with great computational efficiency. To keep the method flexible, we also offer users different choices for the L_p ($0<p<1$) norm in our R code.

In addition to the selection of the L_p norm, the regularization (tuning) parameter λ largely affects the solution of Equation (2) as well. In general, the choice of λ is regarded as a difficult problem. Popular methods for this purpose include the minimization of either the Bayesian information criterion (BIC) or the Akaike information criterion (AIC) as a function of λ, cross-validation, and stability selection(Meinshausen and Bühlmann, 2010) to select λ. However, none of these methods can be applied to control the Type I error, especially for a region based significance test.

By adopting the idea of scaled sparse linear regression (Sun and Zhang), which jointly estimates the regression coefficients and the noise level of the data, we avoid the regularization parameter selection problem. The estimated noise level is used for bias correction. The obtained de-biased estimator is applied to perform marker wise significance tests for each variant.

The scaled L_p norm based sparse regression model is given by

$$(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\sigma})=\arg\min_{\boldsymbol{\beta}\in R^m,\boldsymbol{\alpha}\in R^L,\sigma>0}\{\frac{(\mathbf{Y}-\mathbf{W\alpha}-\mathbf{X\beta})^T\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{W\alpha}-\mathbf{X\beta})}{2n\sigma}+\frac{\sigma}{2}+\lambda\|\boldsymbol{\beta}\|_p^p\} \quad (3)$$

In the unified scaled sparse regression the tuning parameter λ is updated iteratively, which requires an initial value λ_0. However, the sensitivity of the results to the selection of λ_0 is low. Moreover, de-biased estimators can be constructed to balance out the bias in the estimated noise level $\hat{\sigma}$ and the bias caused by the L_p norm regularization, which are both proportional to the initial λ_0. The asymptotic distribution of the de-biased estimators can then be derived without major difficulties.

To solve the optimization problem (3), we combine the algorithm for unified L_p norm based sparse regression with that for the general scaled sparse regression (Sun and Zhang, 2012) and propose the following algorithm.

The Algorithm for Unified Scaled Sparse Regression (3)

Step 1: Data centralization: $\sum_{i=1}^{n} x_{ij} = 0$, for j=1,2,…m

Step 2: Initialize $\lambda^{(0)} = \lambda_0 = 2\sqrt{\frac{\log(m)}{n}}$, $\sigma^{(0)} = \sqrt{\frac{\mathbf{Y}^T\mathbf{\Sigma}^{-1}\mathbf{Y}}{n}}$, $\hat{\boldsymbol{\alpha}}^{(0)} = \mathbf{0}$ and $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, set iterative index r=0, $\varepsilon = 0.0001$; Initialize, $\beta_j^{(0)} = 0$, for j=1,2,…m

Step 3: Update $\hat{\sigma}, \lambda, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}$ coordinately

$$\hat{\sigma}^{(r+1)} = \sqrt{\frac{1}{n}(\mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\alpha}}^{(r)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(r)})^T\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{W}\hat{\boldsymbol{\alpha}}^{(r)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(r)})}$$

$$\lambda^{(r+1)} = \sigma^{(r+1)}\lambda_0$$

Update the regression coefficients by USR method (Cao, et al., 2014)

$$(\hat{\boldsymbol{\beta}}^{(r+1)}, \hat{\boldsymbol{\alpha}}^{(r+1)}) = \arg\min_{\boldsymbol{\beta}\in R^m, \boldsymbol{\alpha}\in R^L} \{\frac{1}{2n\hat{\sigma}^{(r+1)}}(\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha}^{(r)} - \mathbf{X}\boldsymbol{\beta}^{(r)})^T\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{W}\boldsymbol{\alpha}^{(r)} - \mathbf{X}\boldsymbol{\beta}^{(r)}) + \lambda^{(r)} \| \boldsymbol{\beta}^{(r)} \|_p^p\}$$

Step 4: If $\| \hat{\boldsymbol{\beta}}^{(r+1)} - \hat{\boldsymbol{\beta}}^{(r)} \|_2 < \varepsilon$ stop; otherwise return to Step 3

2.2 The bias correction of unified scaled L_p norm regularized sparse regression

Lasso, Ridge regression, and many other popular regression methods utilize a regularization term, in order to obtain a stable solution on an HDS. The L_1 norm regularization term used in Lasso typically shrinks many regression coefficients to zero. This, however, introduces a bias making the non-zero regression coefficients smaller in magnitude.

Adopting the idea of unbiased estimation (Javanmard and Montanari, 2014), we develop a unbiased estimator to recover the unbiased regression coefficients, and to assess the corresponding asymptotic Gaussian distribution. A detailed algorithm is presented below.

The Algorithm for Unbiased Estimator

Step 1: Set $\gamma = \frac{\hat{\lambda}}{\hat{\sigma}}$, where $\hat{\lambda}$ and $\hat{\sigma}$ are the estimated parameters of the unified scaled sparse regression (3)

Step 2: Set $\mathbf{Z} = (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})/n$

Step 3: For i=1,2,…,m, solve $\mathbf{u}_i$ by the following constraint convex program:

$$\begin{aligned} &\text{minimize} && \mathbf{u}_i^T\mathbf{Z}\mathbf{u}_i \\ &\text{subject to} && \| \mathbf{Z}\mathbf{u}_i - \mathbf{e}_i \|_\infty \le \gamma \end{aligned}$$

Because the calculation of each $\mathbf{u}_i$ is independent. To increase the computation speed, we parallelize the calculation.

Step 4: Set $\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_m)^T$ (4)

If any of the above problems is not feasible, then set $\mathbf{M} = \mathbf{I}_{m\times m}$

Step5: Define the unbiased estimator by $\hat{\boldsymbol{\beta}}^u = \hat{\boldsymbol{\beta}} + \frac{1}{n}\mathbf{M}\mathbf{X}^T\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ (5)

where $\hat{\boldsymbol{\beta}}$ is the solution of formula (3).

2.3 Hypothesis tests and confidence intervals

To clarify the problem, we assume $\mathbf{Y}$ is the covariates adjusted phenotype. After ignore the covariates, the true model becomes:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{\Sigma} = \sigma_\Phi^2\mathbf{\Phi} + \sigma_e^2\mathbf{I}_n) \quad (6)$$

where $\boldsymbol{\beta}_0$ is the ground truth regression coefficients and stands for true signal.

We define the sparse level of $\boldsymbol{\beta}_0$ as $S_0 = \{i \in \{1,2,...,m\} \mid \beta_{0,i} \neq 0\}$. In this paper, we apply a weak assumption for the sparse model, which is $s_0 = | S_0 | = o(\sqrt{n/\log(m)})$. Without any further notice, we always assume that this assumption holds. Although the sparse ground truth is preferred, our method is also robust for the non-sparse setting, according to the simulation result in Fig.5.8S and 5.9S in the appendix.

2.3.1 uFineMap test

For each predictor i, we need to develop a significance test to determine whether the corresponding regression coefficient β_i is significant or not. For a specific $i \in \{1,2,...,m\}$, we define the null hypothesis H_0: $\beta_i = 0$ versus the alternative hypothesis H_1: $\beta_i \neq 0$

Supposing the model (6) stands and considering the unbiased estimator (5), we prove that the following asymptotic distribution holds

$$n(\hat{\boldsymbol{\beta}}^u - \boldsymbol{\beta}^0) \xrightarrow{d} N(0, \sigma^2\mathbf{M}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{M}^T), \quad (7)$$

where $\mathbf{M}$ is defined by formula (4). The detailed proof is given in Theorem 1 in **Appendix**.

With this theorem, we can directly derive the significance test for each marker, e.g., uFineMap test. The p-value for each variable can be calculated by the following:

$$P(i) = 2(1 - \Phi(\frac{n|\hat{\beta}_i^u|}{\hat{\sigma}\sqrt{[\mathbf{M}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{M}^T]_{i,i}}})), i = 1,2,...,m \quad (8)$$

where Φ is the cumulative distribution function of a standard normal distribution.

2.3.2 uHDSet test

The next major question is how to control the family-wise error rate (FWER) to claim the whole significant genetic region. Besides Bonferroni-Holm correction or some existing multiple testing correction methods to control the FWER or false discovery rate (Benjamini and Hochberg, 1995; Benjamini and Hochberg, 2000). We are commitment to developing a powerful and efficient multiple testing adjustment, taking dependence into consideration, which would be more powerful than uncorrelated adjustment.

For uHDSet test, the null hypothesis is H_0: $\beta_1 = \beta_2 = ... = \beta_m = 0$, and the alternative hypothesis is H_1: $\exists \beta_i \neq 0, i \in \{1,2,...,m\}$.

Inspired by the idea of van de Geer et al. (van de Geer, et al., 2013), we develop a new statistic for uHDSet significance test by utilizing the previous uFineMap statistics: $S = \max_{i\in\{1,2,...,m\}} \frac{n|\hat{\beta}_i^u|}{\hat{\sigma}\sqrt{[\mathbf{M}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{M}^T]_{i,i}}}$.

For an arbitrary $z \in R$, the following equation holds

$$P(S \le z \mid \mathbf{X}) - P(\max_{i\in\{1,2,...,m\}} \frac{|W_i|}{\hat{\sigma}\sqrt{[\mathbf{M}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{M}^T]_{i,i}}} \le z \mid \mathbf{X}) \to 0$$

where $W \sim N(\mathbf{0}, \hat{\sigma}^2\mathbf{M}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{M}^T)$. The proof is presented in Theorem 2 in **Appendix**.

Under null hypothesis H_0: $\beta_1 = \beta_2 = ... = \beta_m = 0$, statistic S is asymptotically equivalent to the maximum of a series of dependent $\chi^2(1)$ variables, whose distribution relies on the design matrix $\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}$. For any

fixed matrix $\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}$, we simulate its distribution and use its quantile to estimate the p-value of the uHDSet statistic *S*.

3. RESULTS

To validate our proposed tests, we run simulations under different types of pedigree structures to demonstrate their performances comprehensively, in terms of both Type I error rate control and statistical power.

3.1 Nuclear family simulation

To explicitly evaluate the tests for common family structure while control the heritability level and sample size. We use the following linear model to generate simulation data with nuclear family structure (each family consists of two children and their parents).

$$\mathbf{Y} = b\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \quad (9)$$

where *b* is the effect size for causal marker; $\boldsymbol{\Sigma} = 1/3\boldsymbol{\Phi} + 2/3\mathbf{I}$.

We randomly assign 30% of variables to be rare variants (minor allele frequency (MAF) < 1%), 20% of variables to be low frequency variants (1% < MAF< 5%) and the rest variables to be common variants (5% < MAF< 50%).

3.1.1 Data generation

The basic procedure of performing nuclear family simulation is as follows:

Step1: Given MAF for each variable, set the ground truth $\boldsymbol{\beta}_0$ with 10 causal variants (5 of them are rare variants); set the correlation matrix $\mathbf{K}_{ij} = \rho^{|i-j|}$, where $i, j \in \{1,2,...,m\}$ and the coefficient ρ determines the correlation for each pair of variables. We set ρ=0.6 throughout the simulation.

Step2: Sampling $\mathbf{E}^1_{n\times m} \sim N(\mathbf{0}, \mathbf{I}\otimes\mathbf{K})$ and $\mathbf{E}^2_{n\times m} \sim N(\mathbf{0}, \mathbf{I}\otimes\mathbf{K})$

Step3: For each subject *i* and variable *j*, update the genotype matrix by: $X_{ij} = I(E^1_{ij} > \Phi^{-1}(maf(j))) + I(E^2_{ij} > \Phi^{-1}(maf(j)))$.

Step4: Generate the vector of trait values of *n* subjects according to model (9) for a given *b*. The selection of *b* is discussed at *3.1.3*

3.1.2 Type I error rate evaluation

To validate if the proposed significant tests can control the Type I error rate, we generated genotype data by the procedure in Section *3.1.1*, setting *n*=500 and *m*=1000. The trait value is generated by $\mathbf{Y} = \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. We replicated this simulation 1000 times and recorded the corresponding p-values to draw quantile-quantile (Q-Q) plots. Under null hypothesis, the quantile of the p-value should follow the uniform distribution *U*(0,1).

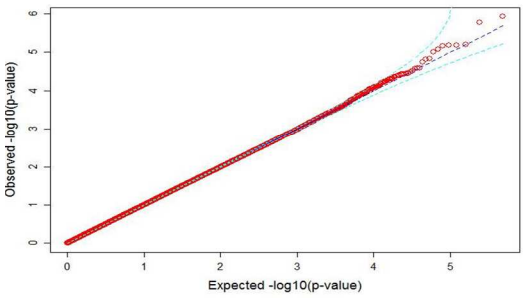

Fig 1. The Q-Q plot for uFineMap test. The x axis is negative log10 of expected p-values, and the y axis represents negative log10 of observed p-values.

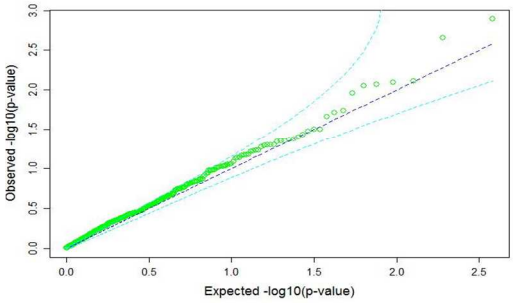

Fig 2. The Q-Q plot for uHDSet test. The x axis is negative log10 of expected p-values, while the y axis represents negative log10 of observed p-values

Fig 1 illustrates most points are aligned near the diagonal line, which is expected. The two dashed curves represent 95% concentration band (CB). With all the points concentrated within the 95% CB, we concluded that the observed p-values follow the uniform distribution over interval (0,1). The Q-Q plot assures that the Type I error rate of uFineMap test is appropriately controlled.

Fig 2 shows that the distribution of uHDSet test's p-values agrees with the uniform distribution, indicating the validity of the adjustment of multiple testing. Therefore, we can draw a conclusion that both of our uFineMap test and uHDSet test can control the Type I error rate appropriately.

3.1.3 Statistical power analysis

The design matrix is simulated by the same procedure as in Section *3.1.1*. As typical, we set the nominal significance level at 0.05 and generated the trait values with respect to various values of heritability *H*. We define the heritability *H* to be the ratio of variance between true signal and the total variance of trait value, which can be explicitly written as:

$$H = \frac{b^2 Var(\mathbf{X}\boldsymbol{\beta}_0)}{Var(\mathbf{Y})} = \frac{b^2 Var(\mathbf{X}\boldsymbol{\beta}_0)}{b^2 Var(\mathbf{X}\boldsymbol{\beta}_0) + Var(\boldsymbol{\varepsilon})}$$

Then we have $b = \sqrt{\frac{HVar(\boldsymbol{\varepsilon})}{(1-H)Var(\mathbf{X}\boldsymbol{\beta}_0)}}$

Let the ground truth signal to be $\beta_0(i) = \begin{cases} 1 & i \in \{1,3,5,7,9,11\} \\ 0 & otherwise \end{cases}$,

i.e., the true marker set to be recovered. We set two of the causal variants to be rare variants and the other four as common variants. We increased the heritability *H* from 0 to 1 and calculated its power at each point. For the sake of saving computational time, we only repeated the procedure 2000 times for each given *H*.

The statistical power for the uFineMap test is defined as $Power = \sum_{t=1}^{T} s_0^{-1} \sum_{i \in S_0} I[P(i) < 0.05] / T$, where T is the simulation replicates; $P(i)$ is the p-value from uFineMap test of ith marker and $I()$ is the indicator function.

For uHDSet test, $P(t)$ represents the p-value calculated at t-th simulation. We define the empirical statistical power to be

$$Power = \sum_{t=1}^{T} I[P(t) < 0.05] / T$$

To evaluate our method, we compare the uFineMap test with other high-dimensional inference methods (e.g., Scaled Lasso (Javanmard and Montanari, 2014), single marker Chi-square test and Gemma (Zhou and Stephens, 2012)). For the uHDSet test comparison, we additionally consider a popular regional based association test, famSKAT (Wu, et al. 2011). The results are shown in Fig 3 and Fig 4 respectively.

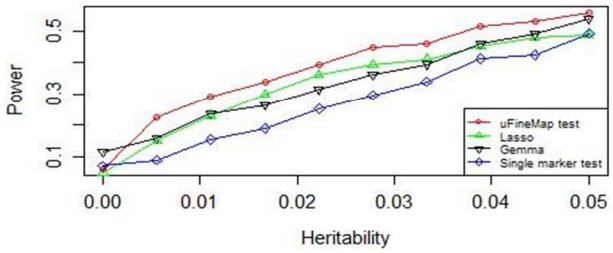

Fig 3. Power vs. heritability for marker wise tests. The legend "uFineMap test" stands for our proposed method; "Lasso" is the Scaled Lasso test; "Gemma" refers to Gemma method and "Single marker test" represents Chi-square single marker test.

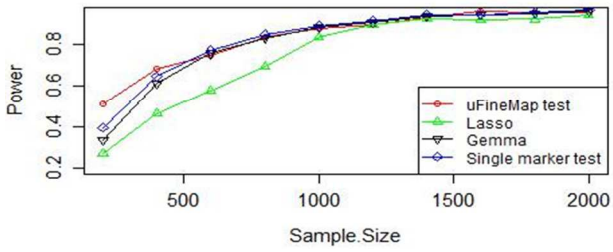

Fig 4. Power vs. sample size for uFineMap tests.

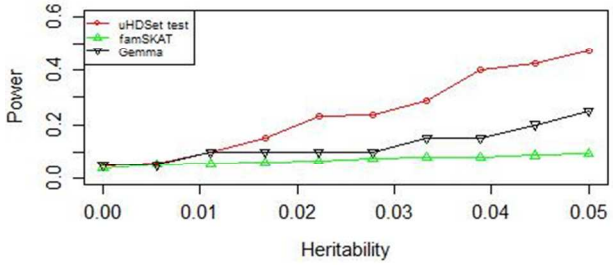

Fig 5. Power vs. heritability for regional tests. The legend "uHDSet" stands for our proposed method.

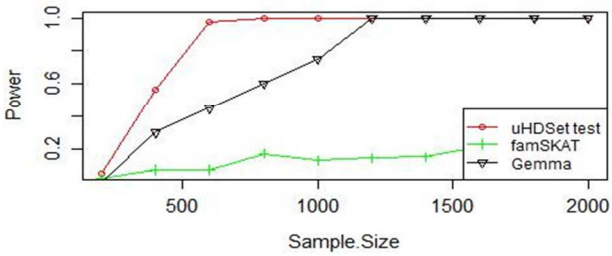

Fig 6. Power vs. sample size for uHDSet tests.

In Fig 3, the uFineMap test performs uniformly better than Scaled Lasso test, Gemma and the single marker test. It indicates that the uFineMap test has a noticeable power gain to identify both common and rare causal variants.

Fig 4 evaluates different methods' performance with respect to sample size changes. It illustrates that our uFineMap test overall outperforms other two methods especially when the sample size are small. Meanwhile, all the competing methods show a similar pattern for a large sample problem.

Similar to Fig 3 and Fig 4, Fig 5 and Fig 6 indicate that the statistical power of all regional tests will increase with the growth of sample size and heritability, which is consistent with our expectation. In addition, at the lower sample size area, our uHDSet test performs much better than famSKAT and Gemma. With the increase of the sample size, the powers of the three methods converge to the same value.

In conclusion, our proposed tests have higher power than competing existing methods regardless of heritability. Meanwhile, it performs almost equally well for large sample size data.

3.2 Complex family simulation

To further compare different methods fairly, instead of using our own or over-simplified simulation data, we used the software SeqSIMLA2. SeqSIMLA2 can simulate sequence data in families under quantitative disease models.

Using SeqSIMLA2, we generate quantitative traits for 8 large families with 67 individuals (the family tree for each family is shown in **Appendix** Figure 5.1S) with 1000 SNPs in total.

3.2.1 Type I error rate evaluation

To verify the validity of our proposed tests, we need to evaluate if the Type I error is well controlled under the null hypothesis. Figure 5.1S and 5.2S (in Appendix) show the Q-Q plots for uFineMap test and uHDSet test respectively. The results indicate that the Type I error rate is appropriately controlled in complex family structure.

3.2.2 Power comparison

We randomly assign 50 causal variants (25 common, 25 rare) to generate the continuous phenotype. Additionally, we proposed two simulation setting for markers effects. We assign all causal

markers to be positively related to the trait value for the same causal direction setting. For the different causal direction setting, half of the causal markers are randomly given a negative relationship with the trait value.

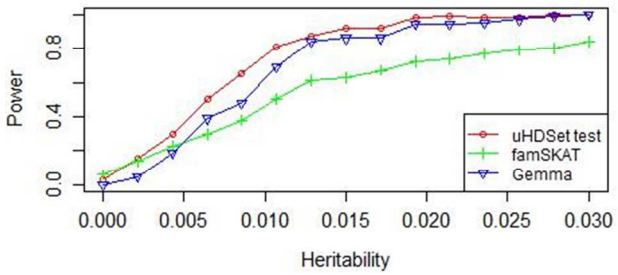

Fig 7. Power comparison with same causal direction.

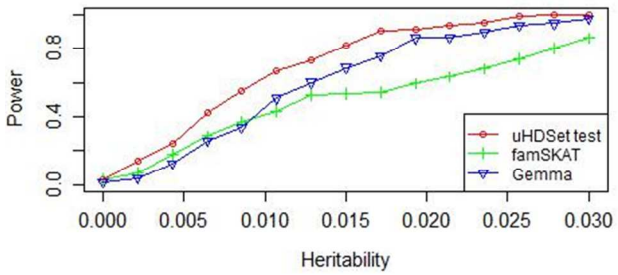

Fig 8. Power comparison with different causal direction.

Fig 7 and Fig 8 present the comparison of three competing methods under same direction and different direction settings respectively. The similar patterns also occurred at a marker wise tests comparison. To make the presentation concise, we only show the result of regional tests, and the result of marker wise tests can be found in the **Appendix** (Figure 5.3S and 5.4S). We can draw the conclusion that all three methods are robust with respect to causal variants direction. But our uHDSet test is almost uniformly more powerful than Gemma and famSKAT for SeqSIMLA simulation data.

3.3 Analysis of sequence data from Framingham Heart Study

To demonstrate the effectiveness of our methods for real genetic variants detection, we applied them to the analysis of sequence data of Framingham Heart Study. This dataset contains both GWAS and next generation sequencing (NGS) data from 4229 subjects with HipBMD data. We used the FISH (Zhang, et al., 2014) method for genotype imputation and selected HipBMD as the phenotype data. After quality control, we obtained 3322 individuals with 6,500,475 SNPs in total. We apply two kinds of data analysis strategies: whole genome analysis and pathway based analysis.

3.3.1 Whole genome analysis

We separate each chromosome into several genetic windows and then apply our uFineMap and uHDSet tests in each window. We set the window size to be 100kb base pairs. After the separation, the whole genome is separated by a total number of 16514 sets of markers. The phenotype is adjusted by the covariates and the top 10 principle components of the genotype before the application of the proposed method. Following the same process as in the simulation studies, we obtain the results and draw the Manhattan plots for 22 chromosomes, as shown in Fig 9 and Fig 10 respectively. Additional results of Manhattan plots for the whole genome (i.e., from chromosome 1 to 22) with higher resolution are presented in **Appendix**.

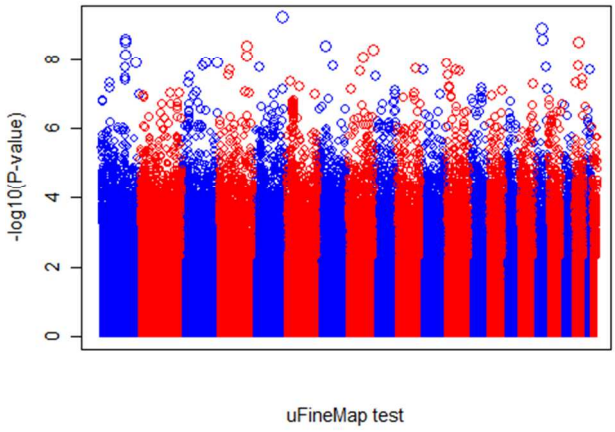

Fig 9. The Manhattan plot for uFineMap test of 22 chromosomes. Each point represents p-value of its corresponding SNP.

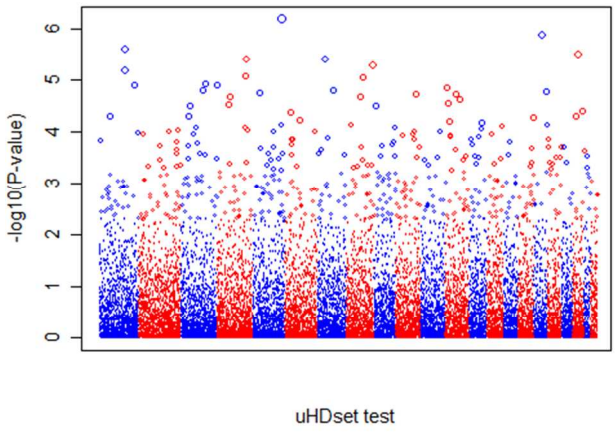

Fig 10. The Manhattan plot for uHDSet test of 22 chromosomes. Each point represents p-value of a 100kb window SNPs region.

By combining the overlapped region of Fig 9 and Fig 10, the uHDSet test report 68 regions of highest susceptibility that exceed

a p-value threshold of 0.001. The reported p-value is based on the whole regional test. According to GeneCards websites, there are 11 genes (Table 1) within the selected regions that are associated with BMD or osteoporosis disease, which further confirms our findings. However, these 11 genes are missed by the use of famSKAT and Gemma method. The reported p-value of Gemma is generated by the minimal p-value after Bonferroni correction for the SNPs within the region.

Table 1. The selected susceptibility genes by uHDSet test

Gene	Chromosome	uHDSet p-value	famSKAT p-value	Gemma p-value
DNM3	1	2.47E-06	0.071107033	0.963871
APOB	2	7.43E-05	0.018075521	0.044156
ERC1	12	0.000154572	0.075876014	0.54554
SRD5A1	5	0.000267385	0.227392554	1
NR3C2	4	0.000317415	0.884812719	0.287339
PLCG1	20	0.000487724	0.022591921	1
INSIG2	2	0.00067805	0.73450689	0.29285
CYP24A1	20	0.000719511	0.132626874	1
ITGA1	5	0.000794757	0.143515502	1
BMPR2	2	0.000901023	0.762703102	0.729078
WNT4	1	0.000940191	0.602006435	0.718623

For the marker wise test, the uFineMap test report 82 susceptible SNPs that exceed a p-value threshold of 10^{-5}. Table 2 shows the 6 reported SNPs that are associated with BMD or osteoporosis disease according to GeneCards websites.

Table 2. The selected susceptibility SNPs by uFineMap test

SNPs	Gene	Chromosome	uFineMap	Gemma
rs11571334	ALOX12	17	4.47E-07	4.68E-05
rs3136452	F2	11	5.39E-07	8.37E-05
rs1264891	OVGP1	1	2.36E-06	5.53E-05
rs10513003	ITGA1	5	4.38E-06	2.99E-05
rs1491717	GC	4	5.17E-06	7.43E-05
rs235766	BMP2	20	5.67E-06	2.99E-05

3.3.2 Pathway analysis

To further illustrate the benefit of the uHDSet test, we collect 880 pathways from KEGG, REACTOME and BIOCARTA pathway analysis databases. We first extract genes belonging to each pathway, then select the corresponding SNPs. The selected SNPs of a specific pathway are combined to form the design matrix for association tests. We list 6 most significant pathways that pass p-value cut-off 10^{-3} in Table 3 for which the prominent famSKAT methods fails to detect. The two P38/MAPK pathways were previously found to play a critical role by other publications (Kim, et al., 2013; Lee, et al., 2008). Endogenous Sterols pathway is also related with BMD reported by another study (Warriner and Saag, 2013). Chemokines pathway is important regulator in development, homeostasis and pathophysiological processes associated with osteoporosis (Lazennec and Richmond, 2010).

Table 3. The selected functional pathways by uHDSet test only

Pathway name	uHDSet p-value	famSKT p-value
REACTOME_FACILITATIVE_NA_INDEPENDENT_GLUCOSE_TRANSPORTERS	5.00E-05	0.05809
REACTOME_ACTIVATED_TAK1_MEDIATES_P38_MAPK_ACTIVATION	7.00E-05	0.05635
REACTOME_P38MAPK_EVENTS	8.00E-05	0.09401
REACTOME_ENDOGENOUS_STEROLS	0.00016	0.00110
REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	3.00E-04	0.07827
KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_GLOBO_SERIES	0.00065	0.13751

Each p-value in Table 3 is generated based on a whole pathway-based region. It can be seen that, our uHDSet method is more powerful than famSKAT in identifying significant pathways which contain a relatively large number of genetic markers.

4. CONCLUSION

Some promising association tests with the adjustment of family structure have been established on the LDSs (low dimensional sets). However, these methods would suffer power loss in high dimensional data. To overcome the limitations of these tests, we propose the uFineMap and uHDSet test for assessing the significance of the HDSs with cryptic relatedness, which are based on novel scaled linear mixed sparse regression. The proposed tests are designed to address the challenge of variants detection under complex pedigree structures, which implement an explicit way to appropriately control the Type I error rate at both single marker level and SNPs set level.

The promising results of testing on both simulated and real data indicate that the uFineMap and uHDSet test yield a considerably higher statistical power gain in comparison to other competing methods, especially for high dimensional data with cryptic relatedness. The uFineMap test can pinpoint single susceptible variants with higher resolutions, even for rare functional variants. In addition, our methods also maintain substantial power for detecting susceptibility variants in low dimensional data or large samples. Last but not least, our methods can identify both rare and common variants efficiently.

One limitation of the proposed method is that we assume a linear mixed relationship between phenotype and genotype, which might not be true in the real world. Therefore, nonlinear regression models with adjustment of relatedness and population stratification may be more suitable. In addition, the overall

computational complexity is $O(n^2m^3)$, which is much longer than simply solving the sparse linear mixed model or other efficient methods designed for LDSs, particularly for extremely large data. To compensate for this issue, parallel computing is implemented to reduce the total computational time for large scale genetic data analysis.

ACKNOWLEDGEMENT:

Our work is partially supported by NIH R01 GM109068 and R01 MH104680. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute NHLBI in collaboration with Boston University (Contract No. N01-HC-25195).

5. REFERENCES

Amos, C.I. (1994) Robust variance-components approach for assessing genetic linkage in pedigrees, *American journal of human genetics*, **54**, 535.

Bühlmann, P. (2013) Statistical significance in high-dimensional linear models, *Bernoulli*, **19**, 1212-1242.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Benjamini, Y. and Hochberg, Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics, *Journal of Educational and Behavioral Statistics*, **25**, 60-83.

Cao, S.*, et al.* (2014) A Unified Sparse Representation for Sequence Variant Identification for Complex Traits, *Genetic epidemiology*, **38**, 671-679.

Chen, H., Meigs, J.B. and Dupuis, J. (2013) Sequence kernel association test for quantitative traits in family samples, *Genetic epidemiology*, **37**, 196-204.

Chen, X., Xu, F. and Ye, Y. (2010) Lower Bound Theory of Nonzero Entries in Solutions of \ell_2-\ell_p Minimization, *SIAM Journal on Scientific Computing*, **32**, 2832-2852.

Devlin, B. and Roeder, K. (1999) Genomic control for association studies, *Biometrics*, **55**, 997-1004.

Endelman, J.B. (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP, *The Plant Genome*, **4**, 250-255.

Javanmard, A. and Montanari, A. (2014) Confidence intervals and hypothesis testing for high-dimensional regression, *The Journal of Machine Learning Research*, **15**, 2869-2909.

Kim, H.K., Kim, M.-G. and Leem, K.-H. (2013) Osteogenic activity of collagen peptide via ERK/MAPK pathway mediated boosting of collagen synthesis and its therapeutic efficacy in osteoporotic bone by back-scattered electron imaging and microarchitecture analysis, *Molecules*, **18**, 15474-15489.

Lazennec, G. and Richmond, A. (2010) Chemokines and chemokine receptors: new insights into cancer-related inflammation, *Trends in molecular medicine*, **16**, 133-144.

Lee, H.W.*, et al.* (2008) Berberine promotes osteoblast differentiation by Runx2 activation with p38 MAPK, *Journal of Bone and Mineral Research*, **23**, 1227-1237.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 417-473.

Price, A.L.*, et al.* (2010) New approaches to population stratification in genome-wide association studies, *Nature Reviews Genetics*, **11**, 459-463.

Rakitsch, B.*, et al.* (2013) A Lasso multi-marker mixed model for association mapping with population structure correction, *Bioinformatics*, **29**, 206-214.

Sun, T. and Zhang, C.-H. (2012) Scaled sparse linear regression, *Biometrika*, **99**, 879-898.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

van de Geer, S., Bühlmann, P. and Ritov, Y.a. (2013) On asymptotically optimal confidence regions and tests for high-dimensional models, *arXiv preprint arXiv:1303.0518*.

Voight, B.F. and Pritchard, J.K. (2005) Confounding from cryptic relatedness in case-control association studies, *PLoS Genet*, **1**, e32.

Warriner, A.H. and Saag, K.G. (2013) Glucocorticoid-related bone changes from endogenous or exogenous glucocorticoids, *Current Opinion in Endocrinology, Diabetes and Obesity*, **20**, 510-516.

Wu, M.C.*, et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test, *The American Journal of Human Genetics*, **89**, 82-93.

XU, Z.-B.*, et al.* (2012) Representative of L1/2 Regularization among Lq (0< q $\leqslant$ 1) Regularizations: an Experimental Study Based on Phase Diagram, *Acta Automatica Sinica*, **38**, 1225-1228.

Zhang, L.*, et al.* (2014) FISH: fast and accurate diploid genotype imputation via segmental hidden Markov model, *Bioinformatics*, btu143.

Zhou, X. and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies, *Nature genetics*, **44**, 821-824.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301-320.