Gengxin Li*, Lin Hou, Xiaoyu Liu and Cen Wu

A weighted empirical Bayes risk prediction model using multiple traits

<https://doi.org/10.1515/sagmb-2019-0056>

Received November 14, 2019; accepted July 6, 2020; published online September 4, 2020

Abstract: With rapid advances in high-throughput sequencing technology, millions of single-nucleotide variants (SNVs) can be simultaneously genotyped in a sequencing study. These SNVs residing in functional genomic regions such as exons may play a crucial role in biological process of the body. In particular, non-synonymous SNVs are closely related to the protein sequence and its function, which are important in understanding the biological mechanism of sequence evolution. Although statistically challenging, models incorporating such SNV annotation information can improve the estimation of genetic effects, and multiple responses may further strengthen the signals of these variants on the assessment of disease risk. In this work, we develop a new weighted empirical Bayes method to integrate SNV annotation information in a multi-trait design. The performance of this proposed model is evaluated in simulation as well as a real sequencing data; thus, the proposed method shows improved prediction accuracy compared to other approaches.

Keywords: empirical Bayes; next-generation sequencing; rare variants; risk prediction.

1 Introduction

The advances in the next-generation sequencing (NGS) technology are underway to directly sequence and identify genetic variants that associate with a variety of diseases (Bamshad et al. 2011; Renkema et al. 2014). A few applications of NGS techniques have been well-established for studies including but not limited to Mendelian disorders (Need et al. 2012; Yang et al. 2013), aging-related diseases (Boyd-Kirkup et al. 2013), height (Marouli et al. 2017; Yang et al. 2010), and type 2 diabetes (Fuchsberger et al. 2016). As this new technique (NGS) emerges, it greatly accelerates the discovery of novel variants or new pharmacogenetic markers for precision medicine (Gonzalez-Garay 2014). Previous sequencing studies have revealed that hundreds of novel variants in 82 pharmacogenes (Bush et al. 2016) and tens of millions of novel variants in UK10K sequencing study (The UK10K Consortium 2015) have been successfully identified. Lacking of functional annotations of these novel variants, efficient statistical and computational methods are needed to predict the function of these variants (De Baets et al. 2012; Yourshaw et al. 2015).

To date, tens of thousands of whole-exome sequencing (WES) data and whole-genome sequencing (WGS) data are available at a low cost and low error rate. Scientists can easily access these data for examining the impact of genetic variants on a disease trait, especially for rare variants. In fact, each individual carries over three million genotyped variants (Ball et al. 2012) with respect to the reference genome, where the majority are rare variants with low minor allele frequencies (MAFs). Such a high density of rare variants launches a weak signal, and it is difficult to separate their poor signal from noises in a real sequencing data analysis. Recent studies suggest that the effect of rare variants can be assessed by adopting different prior probabilities (Cirulli and Goldstein 2010). A number of statistical methods have been developed based upon this by collapsing variants in each gene (Li and Leal 2008; Luo et al. 2011; Madsen and

*Corresponding author: Gengxin Li, Department of Mathematics and Statistics, University of Michigan Dearborn, 4901 Evergreen Rd, Dearborn, MI 48128, USA, E-mail: gengxinl@umich.edu

Lin Hou, Center for Statistical Science, Tsinghua University, 30 Shuangqing Rd, Haidian District, Beijing 100084, China

Xiaoyu Liu, Department of Mathematics and Statistics, Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45435, USA

Cen Wu, Department of Statistics, Kansas State University, 1116 Mid-Campus Drive N., Manhattan, KS 66506, USA

Browning 2009; Price et al. 2010; Wu et al. 2011). This strategy can be improved by clarifying the molecular function of genetic variants in biological process (Ng et al. 2009, 2010; Wu et al. 2011). Because of the advancements in methodology, bioinformatic annotation is going to play a crucial role in a real sequencing analysis.

It increasingly recognizes that a few rare variants are likely responsible for the risk assessment of diseases (Golstein et al. 2013). There is evidence that roughly 85% of mutations occur in the protein-coding regions which account for 2% of the human genome, and this small region is very important in understanding the impact of SNV annotation on a disease (Majewski et al. 2011; Ng et al. 2010). When a mutation occurs in a protein-coding region, this change can be classified into a synonymous change (no amino acid change) or a non-synonymous change (an amino acid change). These two changes have quite different impact on a protein sequence, thus annotating the variants may better reveal their biological functions at the sequence level. A few strategies have been proposed to annotate each variant using PolyPhen2, SIFT, ANNOVAR or CADD (Adzhubei et al. 2013; Alexander et al. 2016; Mathur et al. 2018; Kircher et al. 2014; Krupp et al. 2017; Kumar et al. 2009). However, how to efficiently incorporate this information in a risk prediction model is still a challenging question.

Another frequently encountered problem in a sequencing data analysis is “Selection Bias” which may result in the overinflated estimate of the true genetic parameter (Efron 2009). It is caused by the imbalance between a small number of subjects (thousands of individuals) and an extremely large amount of features (millions of genetic variants or tens of thousands of genes). Under this imbalance, traditional least square estimates are unstable, and over-inflated estimates of genetic effects have been observed in the microarray data analysis (Tibshirani et al. 2002; Zhang, 2008). Compared to the microarray data, sequencing data has even larger number of features, and it makes the imbalance worse and results in more serious bias when estimating the genetic effects. To address this limitation, several approaches (e.g., Shrink Centroid (Tibshirani et al. 2002), Ridge Regression (Hoerl and Kennard 1970) and Bias-reduced Estimator (Zhang, 2008)) have been proposed to shrink the estimates to more realistic ones through correcting the bias. These commonly applied shrinkage methods are based on a frequentist perspective. A Bayesian approach may be a more natural choice over the frequentist method since it is immune to the bias through naturally producing smoother estimates. The development of empirical Bayes (EB) (Efron 2009) is motivated by a need to greatly simplify the calculation of traditional Bayesian estimates. Here, EB estimates may efficiently capture the true effects of a large amount of variants in a sequencing data analysis.

Recent studies have revealed that empirical Bayes prediction models have brought impressive achievements in the microarray data analysis (Efron 2009; Singh et al. 2002). Shortly afterward, we extended the empirical Bayes model to a mini exome sequencing data for predicting the risk of cardiovascular disease (Li et al. 2011). These earlier works were developed for a binary trait (disease status), and not applicable to other types of traits. It is known that a simple binary trait may not be adequate to explain the pathogenic mechanism of disease, especially some quantitative traits are more likely to be responsible for the assessment of complex diseases. Later, we developed a joint empirical Bayes model that successfully combines a binary trait with a quantitative trait to more precisely identify the disease-causing mutations in a whole-genome sequencing data (Li et al. 2015). However, this model did not concern the impact of SNV annotation and treated the synonymous and non-synonymous genetic effects equally. A few studies have revealed that an amino acid change may deeply affect the stability of the protein. The model should assign distinct prior probabilities to the synonymous change and non-synonymous change separately. Hence, a new weighted empirical Bayes (WEB_m) method is designed to incorporate the annotation information of variants in a multiple trait risk prediction model for maximizing the functional genetic effects on multi traits. In some aspects, this proposed method combines the annotation information with multiple traits design in one model, and we expect better prediction results.

2 Method

2.1 Weighted gene score

Since rare variants launch poor signals and are hard to detect in a sequencing data analysis, one solution is to aggregate signals by combining multiple variants in a gene or a genomic region. We believe that appropriately joining multiple variants within a gene

may, on average, strengthen the signals of genetic variants, especially for rare variants. Here, each gene consists of many variants, and these variants are classified into two categories, i.e., synonymous SNVs and non-synonymous SNVs, in terms of their functions. Because non-synonymous SNVs may change the protein sequence and are highly associated with the human diseases, we consider to analyze synonymous SNVs and non-synonymous SNVs separately. Thus, we calculate two sub gene scores for each gene. The sub gene made up of synonymous SNVs is called the synonymous gene, and the non-synonymous gene is made up of non-synonymous variants. Here, a weighted method (Madsen and Browning 2009) is applied to these synonymous and non-synonymous SNVs respectively to calculate two sub gene scores,

$$X_{ij}^s = \frac{1}{N_i^s} \sum_{l=1}^{N_i^s} \frac{L_{ijl}^s}{\sqrt{p_{il}(1-p_{il})}}, \quad X_{ij}^o = \frac{1}{N_i^o} \sum_{g=1}^{N_i^o} \frac{L_{ijg}^o}{\sqrt{q_{ig}(1-q_{ig})}} \quad (1)$$

where X_{ij}^s and X_{ij}^o ($i = 1, \dots, N$; $j = 1, \dots, n$) denote the i th synonymous and non-synonymous gene scores for the j th subject, respectively, and the subscripts s and o stand for the synonymous and non-synonymous, respectively. L_{ijl}^s denotes the number of rare alleles of the j th subject at the l th variant ($l = 1, \dots, N_i^s$) within the i th synonymous gene, similarly, L_{ijg}^o is the number of rare allele of the j th subject at the g th variant ($g = 1, \dots, N_i^o$) within the i th non-synonymous gene; Given the i th gene, there are N_i^s synonymous SNVs and N_i^o non-synonymous SNVs; p_{il} and q_{ig} denote the minor allele frequencies of the l th and g th variants in the i th synonymous gene and i th non-synonymous gene respectively. For this scoring scheme, the importance of synonymous and non-synonymous variants is counted by the inverse of its minor allele frequency.

We expect the significance of two sub genes to be different. Here, a logistic regression model is applied to the synonymous and non-synonymous gene-level data separately. The difference between two sub genes is measured by a weight a_i through the following equation,

$$a_i = \frac{-\log(\tau_i^o)}{-\log(\tau_i^o) - \log(\tau_i^s)} \quad (2)$$

where τ_i^s and τ_i^o denote the p values of the i th synonymous and non-synonymous genes when the logistic regression model is fit to the disease status. The above equation is a function of these two p values. Since the numerator is the minus logarithm of p value for the non-synonymous gene, this weight a_i measures the proportion of the gene effect coming from the i th non-synonymous gene. In brief, a larger weight a_i (>0.5) implies that the impact of i th non-synonymous gene on the disease trait is stronger than that of the i th synonymous gene. Hence, the weighted score of the i th gene and the j th subject is calculated by,

$$X_{ij} = (1 - a_i)X_{ij}^s + a_iX_{ij}^o \quad (3)$$

where X_{ij} is a weighted average score between the synonymous gene X_{ij}^s and the non-synonymous gene X_{ij}^o , and $\mathbf{X}_i$ is a vector of the weighted average scores for all individuals, namely the i th weighted gene.

2.2 Estimates of each weighted gene impact on multiple traits

2.2.1 A binary trait: Suppose there are N weighted genes ($\mathbf{X}_1, \dots, \mathbf{X}_N$) and n subjects consisting of n_1 cases and n_0 controls. The best prediction rule 2.4 (Efron 2009) will diagnose individuals to be sick if

$$\sum_{i \in N} y_{i,d} \mathbf{G}_i > \mathbf{b}. \quad (4)$$

where $\mathbf{G}_i = \frac{\mathbf{X}_i - \mu_i \mathbf{1}_n}{\sigma_i}$ is a standardized weighted gene vector where μ_i and σ_i are the true mean and the standard deviation of the i th weighted gene, and $\mathbf{1}_n$ includes n ones. The parameter $y_{i,d} = \frac{n_1 n_0}{n_1 + n_0} (\mu_{i,1} - \mu_{i,0}) / \sigma_i$ measures the mean difference between the cases and controls for the i th weighted gene where the subscript d indicates a binary trait, and $\mu_{i,k}$ ($k = 1$ or 0) denotes the mean score of the i th weighted gene in the case or control group. The difference between two mean scores reflects the genetic contribution of the i th weighted gene to a binary disease trait.

If σ_i is unknown, the parameter $y_{i,d}$ is estimated as follow:

$$T_{i,d} = \frac{\bar{X}_{i,1} - \bar{X}_{i,0}}{\hat{\sigma}_i \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} \sim t_{n-2}(y_{i,d}). \quad (5)$$

Note that $\hat{\sigma}_i = \sqrt{\frac{(n_1-1)s_{i,1}^2 + (n_0-1)s_{i,0}^2}{n_1+n_0-2}}$ is the pooled standard deviation of the i th weighted gene, where $s_{i,1}$ and $s_{i,0}$ are the sample

standard deviations of the cases and controls respectively. $\bar{X}_{i,k} = \frac{\sum_{j=1}^n X_{ij} I(Y_j=k)}{n_k}$ ($k = 1$ or 0) is the sample mean of the i th weighted gene in the case or control group where Y_j ($=1$ or 0) is the j th individual's disease status. $T_{i,d}$ can be converted to a normal variable $Z_{i,d} = \Phi^{-1}(P(T \leq T_{i,d}))$, where Φ^{-1} is an inverse standard normal and $P(T \leq T_{i,d})$ is a cumulative distribution function of the above t statistic with $n-2$ degrees of freedom. Then $Z_{i,d}$ is an estimate of $y_{i,d}$ for a binary trait, and it follows a normal distribution with mean $y_{i,d}$ and the standard deviation close to 1 (Efron 2009, Remark F),

$$Z_{i,d} \sim N(y_{i,d}, \sigma(y_{i,d}))$$

In addition, Equation (5) illustrates that the expected distance between the case and control classes is estimated by $E\left(\frac{\bar{X}_{i,1} - \bar{X}_{i,0}}{\sigma_i}\right) = \frac{y_{i,d}}{w_0}$ where $\frac{1}{w_0} = \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}$. Then the rule (2.1) (Efron 2009) is applied to the weighted genes $(\mathbf{X}_1, \dots, \mathbf{X}_N)$, $\mathbf{G}_i = \frac{\mathbf{X}_i - \mu_i \mathbf{1}_n}{\sigma_i} \sim N\left(\pm \frac{y_{i,d}}{2w_0}, 1\right)$ where + and - indicate the case (disease) and control (healthy) classes, respectively. Threshold $\mathbf{b}$ is the decision boundary $\mathcal{J}$ to distinguish healthy and disease groups. When equal numbers of cases and controls are collected, it is a hyperplane which is orthogonal to the line between $\frac{y_{i,d}}{2w_0}$ and $-\frac{y_{i,d}}{2w_0}$, and the value of $\mathbf{b}$ is equal to 0. If there are unequal numbers of cases and controls, $\mathbf{b}$ is proportional to the logarithm ratio of sample sizes between cases and controls (8.2 Efron 2009).

2.2.2 A quantitative trait: In addition to the disease status (binary trait), a quantitative trait that is related to the disease is often collected in most genetic studies. More importantly, the quantitative trait may offer valuable information for the disease diagnosis. Let $\mathbf{q}$ denote a quantitative trait. For convenience, we standardize $\mathbf{q}$.

A linear regression model is applied to see the impact of $\mathbf{q}$ on each standardized weighted gene $\mathbf{G}_i$,

$$\mathbf{G}_i = \beta_{i,v} \mathbf{q} + \epsilon_i, \quad i = 1, 2, \dots, N. \quad (6)$$

Suppose ϵ_i has mean zero and identity covariance matrix, and is not correlated with $\mathbf{q}$. The coefficient $\beta_{i,v}$ reflects the effect of the quantitative trait ($\mathbf{q}$) on the i th weighted gene ($\mathbf{G}_i$) where the subscript v indicates the quantitative trait. The joint distribution of $\mathbf{G}$ and $\mathbf{q}$ is used to derive the best predictor of $\mathbf{q}$ (Efron 2009, 7.3).

The Bayesian estimate is calculated from the posterior distribution of coefficients $\beta_{i,v}$. According to the definition of $\mathbf{G}_i = (\mathbf{X}_i - \mu_i \mathbf{1}_n)/\sigma_i$ where μ_i and σ_i are the mean and the standard deviation of the i th weighted gene, the best linear prediction model can be expressed as,

$$\mathbf{X}_i = \mu_i \mathbf{1}_n + \sigma_i (\beta_{i,v} \mathbf{q} + \epsilon_i), \quad i = 1, 2, \dots, N.$$

where the least squares method is used to estimate parameters μ_i , σ_i and $\beta_{i,v}$, and a vector $\mathbf{1}_n$ includes n ones. When ϵ_i is normally distributed (see Appendix 1.1), a t -test of hypothesis $\beta_{i,v} = 0$ is inferred from Equation (6), and it follows a noncentral t distribution,

$$T_{i,v} = \frac{\hat{\beta}_{i,v}}{se(\hat{\beta}_{i,v})} = 2w_0 \hat{\beta}_{i,v} \sim t_{n-2}(y_{i,v}), \quad \text{where } w_0 = \sqrt{\frac{\sum_{j=1}^n (q_j - \bar{q})^2}{4}} \left(\bar{q} = \sum_{j=1}^n \frac{q_j}{n} \right)$$

The noncentral t distribution has $n-2$ degrees of freedom with noncentral parameter $y_{i,v} = 2w_0 \beta_{i,v}$, which reflects the impact of the i th weighted gene on the quantitative trait. Then we can obtain a normal scale statistic $Z_{i,v}$ via the transformation equation ($Z_{i,v} = \Phi^{-1}(P(T \leq T_{i,v}))$), and this $Z_{i,v}$ is an estimator of the parameter $y_{i,v}$. It is normally distributed when sample size is large (Efron 2009),

$$Z_{i,v} \sim N(y_{i,v}, \sigma(y_{i,v}))$$

2.3 Weighted EB prediction rule in a multiple trait design

Many medical studies increasingly recognize the importance of some quantitative traits (e.g. blood pressure, weight or blood testing) in diagnosis of complex diseases, especially given the growing understanding that some causal genes may have more immediate effects on a quantitative trait than those on a binary trait. Here, we define a new estimator $Z_{i,c}$ to combine two statistics $Z_{i,d}$ and $Z_{i,v}$ introduced above that correspond to a binary trait and a quantitative trait respectively,

$$Z_{i,c} = \begin{cases} \text{sgn}(Z_{i,d}) \max(|Z_{i,d}|, |Z_{i,v}|) & \text{if } \text{sgn}(Z_{i,d}) \times \text{sgn}(Z_{i,v}) > 0 \\ Z_{i,d} & \text{otherwise} \end{cases} \quad (7)$$

where sgn is a sign function which gives +1 or -1 for each non-zero real number (e.g., $\text{sgn}(5) = 1$ or $\text{sgn}(-5) = -1$). We know that both statistics $Z_{i,d}$ and $Z_{i,v}$ are normally distributed, and the new statistics $Z_{i,c}$ can capture the largest additive genetic effect on both traits if $\text{sgn}(Z_{i,d}) > 0$. In contrast, it can help us aggrandize the protective effects on two traits when $\text{sgn}(Z_{i,d})$ takes a negative value. In case the conflicting genetic effects on multiple traits are detected, the quantitative trait information will be useless at this moment. Directly taking the statistic of the binary trait $Z_{i,d}$ will guarantee the efficiency of the model.

When we investigate the inference of this new statistic, the distribution of $Z_{i,c}$ is similar as that of $\max(Z_{i,d}, Z_{i,v})$ if $Z_{i,d}, Z_{i,v} > 0$ and that of $\min(Z_{i,d}, Z_{i,v})$ if $Z_{i,d}, Z_{i,v} < 0$. In practice, two normal random variables $Z_{i,d}$ and $Z_{i,v}$ might not be independent, then the multivariate normal distribution of a random vector $(Z_{i,d}, Z_{i,v})^T$ is used to derive the exact distribution of this combined statistic $Z_{i,c}$. The exact distribution of $Z_{i,c}$ is not normally distributed (Ker 2001; Nadarajah and Kotz 2008). However, our previous work (Li et al. 2015) has illustrated that $Z_{i,c}$ approximately follows a skewed normal distribution, and the variance of statistics $Z_{i,d}$ and $Z_{i,v}$ close to 1

greatly reduces the extent of the skewness of the normal distribution (Ker 2001; Nadarajah and Kotz 2008). Then empirical Bayes method can shrink this combined statistic ($Z_{i,c}$) to a smoother estimate ($\hat{y}_{i,c}$). In brief, this new statistic maximizes the effects of weighted genes on multiple traits, including genes that are directly acting on a quantitative response instead of the disease trait.

While calculating empirical Bayes estimator, selection bias is inevitable when the number of features (weighted genes) is large. A Bayesian approach is known to offer smoother estimates, and empirical Bayes method (Efron 2009) can greatly simplify the calculation of Bayesian estimates while helping to ensure that the estimates reflect most causal weighted genes' effects. For any pair (Z, γ), the Bayesian estimate $\hat{\gamma}$ is obtained by the first derivative of the cumulant generating function of Z (Efron 2009, Remark E and F),

$$\begin{aligned} Z|y &\sim N(y, \sigma) \\ \hat{K}(z) &= \log(\hat{f}(z)) \\ \hat{\gamma} &= Z + \hat{\sigma}^2 \hat{K}'(z) = \hat{E}(y|z) \end{aligned}$$

Note that $\hat{f}(\cdot)$ estimates the marginal density of Z ; $\hat{K}(\cdot)$ denotes the log of the marginal density function of Z ; $\hat{\sigma}^2$ is the variance estimate of the combined gene score; $\hat{K}'(\cdot)$ is the numerical derivative of $\hat{K}(\cdot)$. $\hat{\gamma}$ is a numerical approximation of the empirical Bayes estimates.

After ranking the estimates of all weighted genes, we select a subset of genes with large effects in the prediction classification,

$$\sum_{i \in I_g} \hat{y}_{i,c} \hat{\mathbf{G}}_i > \mathbf{b}. \quad (8)$$

where genes with the largest weighted EB estimates ($\hat{y}_{i,c}$) on multiple traits are collected in a set I_g .

The covariate variables, such as Age and Smoking status, may play a significant role in predicting disease risk. Obviously, age is a continuous variable, then we can directly adopt z or t test to measure its effect on multiple phenotype traits. In particular, we calculate the smoking proportion among n subjects (Simulation data: $p \approx 0.25$, $np > 10$). This property guarantees that z test is also appropriate to detect the smoking effect on multiple traits. Hence, Sections 2.2.1 and 2.2.2 can be successfully applied to these covariates. According to Equation (7), the combined statistic $Z_{m,c}$ is calculated to maximize the effect of the m th covariate ($m = 1, \dots, M$) on multiple traits. The relative weighted EB estimator ($\hat{y}_{m,c}$) is achieved through shrinking the combined statistic to the smoother estimate. We define a matrix $(\hat{W}_1, \dots, \hat{W}_T)$ including all standardized weighted genes $(\hat{G}_1, \dots, \hat{G}_N)$ and the standardized covariates $(\hat{C}_1, \dots, \hat{C}_M)$ where $T = N+M$.

The final weighted empirical Bayes prediction model is defined by,

$$\sum_{t \in I} \hat{y}_{t,c} \hat{W}_t > \mathbf{b} \quad (9)$$

where I is a set of most important genes and covariates selected from two traits. $\hat{W}_t$ is the t th ($t = 1, \dots, T$) standardized feature.

3 Simulation study

GAW 17 sequencing data (Almasy et al. 2011) is simulated from a complex model, where various deleterious genetic variants from some particular biological pathways (e.g., VEGF) are selected to capture the risk of cardiovascular disease. This data set is made by combining real and simulated data. 1000 Genomes Project (<http://www.1000genomes.org>) provides a real exome sequencing data (genotype data), and this real genotype data is used to simulate the disease trait (D) and the related quantitative traits (e.g., Q_1). In fact, the trait D is simulated by a liability threshold model (its top 30% is declared illness) which is controlled by the latent liability and multiple quantitative traits (Almasy et al. 2011, Equation 1), and nine genes are selected to determine the variation of Q_1 (Almasy et al. 2011, Table 1). Since the simulation model just includes the additive genetic effects, the mean level of a quantitative trait is positively associated with the variants. Moreover, GAW 17 data includes 697 unrelated individuals and 24,487 SNVs where a large number of rare SNVs are widely distributed across the genome. It also generates 200 replicates by changing individuals' smoke status in the simulation model (Almasy et al. 2011), and each replicate includes 697 individuals. In particular, around 87% of SNVs are rare SNVs with minor allele frequencies less than 0.05, and 74% of SNVs are extremely rare with minor allele frequencies less than 0.01. More importantly, the annotation information of SNVs is provided, so the Madsen and Browning method (Madsen and Browning 2009) is applied to the synonymous SNVs and non-synonymous SNVs, separately. It results in 3205 synonymous genes and 3205 non-synonymous genes, and the final weighted gene score is calculated from Equation (3). In this study, our new weighted empirical Bayes

(WEB_m) prediction model is developed to maximize the effects of weighted genes on a disease trait (D) and a quantitative trait (Q_1). The performance of this method is compared with those of the weighted EB (WEB) method based on a single binary trait, empirical Bayes (EB) method, Logistic Regression (LR) and Random Forests (RF: Breiman, 2001).

One key step in developing a risk prediction model is the feature selection, where genes are ranked by their importance to the disease. When we sort genes by their impact, the performances of various methods will be compared by their ranking lists of genes. We expect the genes selected by the simulation model to possess larger effects, and occupy top positions in the gene list. These selected genes are called “true genes” in this study. Table 1 summarizes top 10 features of five classifiers, including WEB_m , WEB, EB, Logistic regression, and Random Forests. It has been demonstrated that one true gene (FLT1) is commonly inferred from four methods (WEB_m , WEB, EB and LR). In addition, the proposed new method (WEB_m) could further identify two more true genes, KDR and PTK2B. As we discussed above, this new weighted empirical Bayes method treats synonymous genes and non-synonymous genes unequally, then more true genes can be detected to better predict the disease trait. Specifically, the effects of non-synonymous FLT1 and KDR are more important than those of synonymous genes. For example, the weight of non-synonymous FLT1 ($a_i = 66\%$) is around two times as large as that of synonymous FLT1 ($1 - a_i = 34\%$). The weighted KDR gene also upweights the effect of non-synonymous change (56%). According to the simulation model, KDR is strongly associated with the quantitative trait Q_1 . Since the new method provides a way to capture causal genes related to Q_1 in addition to the disease status, it is certainly seen that more true genes may be discovered. Besides, Table 1 summarizes the number of SNVs falling into different allele frequency categories, i.e. those with $MAF \geq 0.05$, with MAF between 0.01 and 0.05, and those with $MAF < 0.01$. It is clearly seen that SNVs within the validated genes have a larger proportion of rare variants (FLT1: 91%; KDR: 94%; PTK2B: 61%). Moreover, all methods demonstrate that two covariates (age and smoking status) play more important roles than genes.

When the top feature list is expanded to rank top 50 genes, the comparison of all methods is summarized in Table 2. WEB_m and WEB are capable of detecting one more true gene individually (WEB_m : PIK3C3; WEB: PIK3C2B). Both genes upweight the non-synonymous change (PIK3C3: $a_i = 64\%$; PIK3C2B: $a_i = 72\%$) to calculate the weighted gene score. Specifically, WEB_m and WEB are able to detect the functional genetic impacts on multiple traits and a single trait respectively, then it may be the reason why these two methods can successfully identify more true genes. In contrast, the logistic regression only identifies one true gene, FLT1, among top 50 genes. It is clearly seen that WEB_m combines the annotation information with the multiple trait design to discover the largest number of true genes.

Since GAW 17 data generates 200 replicates through changing individuals' smoke status (Almasy et al. 2011), we can run Cross Validation procedure to calculate the prediction accuracy. These 200 replicates are split into the training and test data, where 100 replicates are randomly selected as the training data to develop the prediction rule, with the remaining 100 replicates as the test data to estimate the prediction accuracy. We repeat this procedure five times to estimate the mean and standard deviation of the prediction accuracy for each method. Table 3 compares areas of under ROC curve (AUC) for different methods, and summarizes their prediction performances. The calculation of AUC is based on two models, where model I considers both genetic and non-genetic features to predict the disease risk, and model II only considers genes. It is clearly seen that WEB_m significantly increases the prediction accuracy in two model settings (0.91(SD:0.020); 0.68(SD:0.016)), over other models (EB: 0.76(0.0102) and 0.6(0.0141); WEB: 0.80(0.015) and 0.64(0.0183)). The improvements can be more readily demonstrated through drawing the ROC curve for each method (Figures 1 and 2). Furthermore, the new model is compared to another classifier (Random Forests). Specifically, WEB_m still performs better than random forest in terms of prediction accuracy (Table 4).

4 Real data analysis

The genetic analysis workshop (GAW) 19 (Blangero et al. 2016; Engelman et al. 2016) offers a whole-exome sequence data for the unrelated Mexican American samples. This study focuses on the identification of novel

Table 1: The prediction rule of five classifiers for a mini-exome sequencing data (WEB_{new} , EB, WEB, RF and LR).

	WEB _m					EB					WEB					RF		LR	
	Predictor	# SNV	WEB _m		# SNV	EB		# SNV	WEB		# SNV	RF		# SNV	MAF	LR			
			Predictor	MAF		Predictor	MAF		Predictor	MAF		Predictor	MAF						
1	SMOKE		AGE			SMOKE			AGE			AGE							
2	AGE		SMOKE			AGE			SMOKE			SMOKE							
3	FLT1	25	<0.01	ATP11A		FLT1			OR13C9			SUSD2			36	<0.01			
		7	0.01–0.05												6	0.01–0.05			
		3	≥0.05												3	≥0.05			
4	KDR	14	<0.01	FLT1		SUSD2			FREM2			ATP11A				<0.01			
		1	0.01–0.05													0.01–0.05			
		1	≥0.05												1	≥0.05			
5	TAS2R48	4	<0.01	SUSD2		ATP11A			TLL1			BUD13				<0.01			
		2	0.01–0.05													0.01–0.05			
		6	≥0.05												1	≥0.05			
6	HLA.B	11	<0.01	BUD13		RIPK3			PPIG			RIPK3			17	<0.01			
		8	0.01–0.05												2	0.01–0.05			
		9	≥0.05												2	≥0.05			
7	OR2T34	5	<0.01	RIPK3		BUD13			BRWD1			FLT1			25	<0.01			
		6	0.01–0.05												7	0.01–0.05			
		5	≥0.05												3	≥0.05			
8	PTK2B	9	<0.01	C10orf107		ADAMTS4			C14orf159			MAP3K12			14	<0.01			
		2	0.01–0.05												3	0.01–0.05			
		7	≥0.05													≥0.05			
9	TAS2R46	3	<0.01	ADAMTS4		WNT16			CARD9			ADAMTS4			33	<0.01			
		5	0.01–0.05												4	0.01–0.05			
		2	≥0.05												3	≥0.05			
10	ZNF91	1	<0.01	MAP3K12		GOLGA1			PCDHGB			C14orf107				<0.01			
		3	0.01–0.05													0.01–0.05			
		5	≥0.05												1	≥0.05			

NOTE: The top 10 important features for five methods (WEB_m , EB, WEB, RF and LR); WEB_m : the Weighted EB model for two traits; WEB: the Weighted EB method for a binary trait; EB: the EB model for a binary trait; RF: Random Forest; LR: Logistic Regression; # SNV indicates the number of SNVs within a specific gene; MAF shows three intervals of Minor Allele Frequency, that is, MAF < 0.01, 0.01 ≤ MAF < 0.05, and MAF ≥ 0.05. The highlighted genes and environmental factors are real causal features which are selected across the proposed models.

genetic effect on a complex disease, hypertension disease. The data includes multiple phenotype traits, the hypertension diagnosis (HTN: yes = 1/no = 0) denoted as a binary trait and the systolic blood pressure (SBP) and diastolic blood pressure (DBP) thought as the quantitative traits. In addition, an important covariate, age, is provided for more accurately predicting the risk of disease. With the aid of the UCSC genome annotation database (<http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/>), we classify around 412,781 SNVs into synonymous SNVs and non-synonymous SNVs. The approach (Madsen and Browning 2009) is applied to two category variants separately, then 6541 synonymous genes and 5727 non-synonymous genes are collected. In this real sequence data, 1851 unrelated individuals are involved in the analysis after removing the missing observations of multiple phenotypes and one covariate. Two prediction approaches, WEB_m and WEB are applied to this gene-level sequence data to evaluate their performances. 10-fold cross-validation procedure is applied to achieve the prediction accuracy. In brief, the proposed model is to fit the training set (1666 subjects), then the remaining 185 subjects is the test set to assess the prediction accuracy. The same procedure repeats 10 times, and the final accuracy is calculated as the average of these 10 prediction accuracies.

Table 5 summarizes top 10 important features selected by three methods (our proposed methods: WEB_m , Random Forest and Logistic Regression). Generally, the covariate, age, is commonly selected by three methods, and its ranking position is number one. It indicates that age is strongly associated with the hypertension disease. More importantly, the gene PBX3 is successfully detected by the new method WEB_m and logistic regression, and this gene may be important to the diagnosis of hypertension disease. When we extend the feature search to top 100 important factors (Table 6), another two genes, ERBB2IP and IGSF9B, are detected by three methods. If we compare the common genes between two methods, more common genes are detected. For example, Supplementary Table 1 lists 6 and 20 common genes among top 100 features detected by WEB_m and Random Forest, WEB_m and Logistic Regression, respectively. The details of these common genes have been summarized by the Supplementary Table 2. If we enlarge the feature search to top 150 and 200 features, a larger number of common genes may be observed (Supplementary Table 1).

Table 7 and Figure 3 summarize the prediction accuracy results. Table 7 compares the prediction accuracy (AUC) between WEB_m and WEB. It is clearly seen that WEB_m can provide a larger prediction accuracy (AUC: 0.78 with SD: 0.029), followed by WEB (AUC: 0.75 with SD: 0.034). The improvement of WEB_m in AUC is most likely due to the fact that WEB_m can maximize the functional effects of genes from multiple traits (HTN and SBP). In the real experiment, WEB_m works well when two category variants have different impacts on the same disease trait or when a quantitative trait is highly related to the binary disease trait. Figure 3 further illustrates the comparison of the ROC curves between WEB_m and WEB under 10-fold cross validation procedure. This figure shows that WEB_m outperforms WEB in terms of prediction accuracy. In summary, we believe our proposed method WEB_m is an efficient prediction tool for the large sequencing data.

Table 2: The ranking position of true genes for five classifiers (WEB_m , EB, WEB, RF and LR).

	Predictor	WEB_m	EB	WEB	RF	LR
		Rank position	Rank position	Rank position	Rank position	Rank position
1	FLT1	1	2	1	–	5
2	KDR	2	–	–	–	–
3	PTK2B	6	–	–	–	–
4	PIK3C3	23	–	–	–	–
5	PIK3C2B	–	–	22	–	–

NOTE: The ranking positions of five true genes FLT1, KDR, PTK2B, PIK3C3 and PIK3C2B detected by five methods (WEB_m , EB, WEB, RF and LR); WEB_m : the Weighted EB method for two traits; WEB: the weighted EB method for a binary trait; EB: the EB for a binary trait; RF: Random Forests; LR: Logistic Regression; “–”: the position of one gene is outside of top 50 important genes.

Table 3: The prediction accuracy of *EB*, *WEB* and *WEB_m* classifiers.

Item	Scenario	Method	EB	WEB	WEB _m
AUC	Gene+Envi	Mean	0.76	0.80	0.91
		SD	0.0102	0.015	0.020
AUC	Gene	Mean	0.60	0.64	0.68
		SD	0.0141	0.0183	0.016

NOTE: AUC is the area under the receiver operating characteristic (ROC) curve when minimizing prediction error; SD is the standard deviation of AUC; *WEB_m*: The Weighted EB method for two traits; *WEB*: The Weighted EB Method for a binary trait; *EB*: EB Method for a binary trait.

Figure 1: ROC curves of *WEB_m* and *WEB* methods for a simulation data (Gene + Covariates). The blue and green lines are ROC curves for *WEB_m* method and *WEB* method, respectively.

Figure 2: ROC curves of *WEB_m* and *WEB* methods for a simulation data (Only Gene). The blue and green lines are ROC curves for *WEB_m* method and *WEB* method, respectively.

Table 4: The prediction accuracy between WEB_m and RF.

Item	Scenario	Method	WEB_m	RF
AUC	Gene+Envi	Mean	0.91	0.66
		SD	0.020	0.019

NOTE: AUC indicates the area under the receiver operating characteristic (ROC) curve when minimizing the prediction error; WEB_m : the Weighted EB Method for two traits; RF: Random Forest.

5 Discussion

With the development of the sequencing technology, scientists can directly identify the disease-causal variants; however, the majority of variants are rare variants with low frequency, which is difficult to be detected in a sequencing data analysis. A few approaches (Li and Leal 2008; Luo et al. 2011; Madsen and Browning 2009; Price et al. 2010; Wu et al. 2011) are applied to strengthen the signals through collapsing all variants in each gene, especially for rare variants. It is increasingly recognized that the bioinformatic annotation of variants plays a crucial role in interpreting the functional impact of a variant. Jointly analyzing multiple phenotype traits may better explore the complicated functions of genetic variants. Hence, incorporating the annotation information in a multi trait model may improve the statistical power to detect the causal genes. Furthermore, it is commonly observed that the estimates of genes are inflated in a sequencing analysis due to the imbalance between a small number of subjects and a large amount of genes. The empirical Bayes (Efron 2009) method provides a natural choice to correct this bias. In this study, we develop a new method to incorporate the SNV annotation in a multiple trait design. Our proposed prediction rule is capable of exploring the functional genetic effects on multiple traits.

While annotating the genetic variants, the variants are classified into two classes, synonymous SNV and non-synonymous SNV. It is well-known that non-synonymous variants are more likely related to the pathogenesis of complex diseases (Sun and Yu 2019). Here, the distinctions of two classes are managed by a weighted gene score (Section 2.1). In reality, two types of variants have strong correlation, but this weighted gene method is based on the condition that two classes are independent. Therefore, developing a method to concern this correlation may be a new goal. Besides, SNV annotation consists of more than two classes. How to incorporate all annotation classes in a risk prediction model is still challenging.

The proposed method can be extended to more complex sequencing data with multiple quantitative traits and one binary trait. Suppose there are V number of quantitative traits ($Q_1, Q_2, \dots, Q_V$) and one binary trait (D) in a sequencing data. When the analysis focuses on the detection of the i th weighted gene effect on multiple traits, three steps are involved here. First, Section 2.2.1 is applied to a disease trait (D), and the estimator ($Z_{i,d}$) is

Table 5: The prediction rule of three classifiers for a real sequencing data (WEB_m , RF and LR).

WEB_m	RF	LR
AGE	AGE	AGE
PBX3	C9	BDP1
NATD1	ASB16	ERN1
YIPF2	GPR98	GID4
XAF1	SSPO	C11orf42
IQCE	PHLDB2	PBX3
ERBB2IP	KIAA1024	APOC4
SEMA3B	ZC3H12A	IGSF9B
APBA3	PLIN3	PPP6C
ABO	TNRC18	ACO69363.1

NOTE: The top 10 important features from the model incorporating genes and environmental factors of three methods (WEB_m , RF and LR); WEB_m : the Weighted EB model for two traits; RF: Random Forest; LR: Logistic Regression; The highlighted genes and environmental factors are common features which are selected by three methods.

Table 6: The ranking position of common genes for three classifiers (WEB_m , RF and LR).

	Predictor	WEB_m	RF	LR
		Rank position	Rank position	Rank position
1	PBX3	2	–	6
2	ERBB2IP	7	92	27
3	IGSF9B	98	34	8

NOTE: The ranking positions of three common genes PBX3, ERBB2IP and IGSF9B detected from three classifiers; WEB_m : the Weighted EB method for two traits; RF: Random Forests; LR: Logistic Regression; “–”: the position of one gene is outside of top 100 important features.

obtained to measure the i th weighted gene effect on a binary trait. Next, we utilize Section 2.2.2 method to multiple quantitative traits separately, and the corresponding gene effects on various quantitative traits are estimated by $Z_{i,v1}, Z_{i,v2}, \dots, Z_{i,vV}$ respectively. Finally, Equation (7) will be changed to measure the largest genetic effect as follows,

$$\begin{aligned}
 Z_{i,c1} &= \begin{cases} \text{sgn}(Z_{i,d})\max(|Z_{i,d}|, |Z_{i,v1}|) & \text{if } \text{sgn}(Z_{i,d}) \times \text{sgn}(Z_{i,v1}) > 0 \\ Z_{i,d} & \text{otherwise} \end{cases} \\
 Z_{i,c2} &= \begin{cases} \text{sgn}(Z_{i,d})\max(|Z_{i,d}|, |Z_{i,v2}|) & \text{if } \text{sgn}(Z_{i,d}) \times \text{sgn}(Z_{i,v2}) > 0 \\ Z_{i,d} & \text{otherwise} \end{cases} \\
 &\vdots \\
 &\vdots \\
 Z_{i,cV} &= \begin{cases} \text{sgn}(Z_{i,d})\max(|Z_{i,d}|, |Z_{i,vV}|) & \text{if } \text{sgn}(Z_{i,d}) \times \text{sgn}(Z_{i,vV}) > 0 \\ Z_{i,d} & \text{otherwise} \end{cases} \\
 Z_{i,c} &= \text{sgn}(Z_{i,d})\max(|Z_{i,c1}|, \dots, |Z_{i,cV}|)
 \end{aligned}$$

It is clearly seen that the final combined statistic $Z_{i,c}$ captures the maximum genetic effect on multiple quantitative traits and one binary trait. However, not all quantitative traits can enrich the gene signal in this study. In fact, only the quantitative traits which are positively correlated with the binary trait can greatly strengthen this genetic signal. In our future study, we plan to develop a better way to incorporate various quantitative traits in a binary trait based risk prediction model for further improving the prediction accuracy.

In this approach, the Z statistic is derived from the transformed T statistic (Equation (5)). But this statistic is inferred from a simple probability model, where we assume a gene vector $\mathbf{G} = (\mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_N)$, with each $\mathbf{G}_i$ approximated by a normal distribution. However, this strategy is based on a balance design where the number of subjects is the same between cases and controls. In real practice, many sequencing data are not balanced, which may largely affect the prediction performance. In the future study, some methods should be proposed to adapt this issue.

It is growingly recognized that the disease pathogenesis is a very complicated process, especially for the interaction between genetic factors and environmental factors. If both genetic and non-genetic predictors are available, a risk prediction may be improved by considering this interaction. Here, two sequencing data are

Table 7: The prediction accuracy of WEB_m and WEB for a real sequencing data.

Item	Scenario	Method	WEB_m	WEB
AUC	Gene+Envi	Mean	0.78	0.75
		SD	0.029	0.034
AUC	Gene	Mean	0.55	0.52
		SD	0.037	0.051

NOTE: AUC is the area under the receiver operating characteristic (ROC) curve when minimizing the prediction error; SD is the standard deviation of AUC; WEB_m : the Weighted EB Method for two traits; WEB : the Weighted EB method for a binary trait.

Figure 3: ROC curves of WEB_m and WEB methods for a real sequencing data (Gene + Covariates). The blue and green lines are ROC curves for WEB_m method and WEB method, respectively.

analyzed, both of them have demographic and environmental covariates, which may provide more information about subjects, in addition to genetic information. Although our proposed approach incorporates both types of information, we have not considered potential interactions between genetic and non-genetic features, which may improve the statistical power to detect causal genetic variants. Further work is needed to include gene–environment and gene–gene interactions in the risk prediction model.

Acknowledgment: Preparation of the Genetic Analysis Workshop 17 Simulated Exome Data Set was supported in part by NIH R01 MH059490 and used sequencing data from the 1000 Genomes Project (www.1000genomes.org). The GAW19 whole genome sequence data were provided by the T2D-GENES Consortium, which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW19 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder study, which are supported by NIH grants P01 HL045222, R01 DK047482, and R01 DK053889. Andrew R. Wood is supported by European Research Council grant SZ-245 50371-GLUCOSEGENES-FP7-IDEAS-ERC. National Natural Science Foundation of China (11601259).

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: The study was supported by NIH R01 MH059490, U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, U01 DK085545, P01 HL045222, R01 DK047482, and R01 DK053889. Andrew R. Wood is supported by European Research Council grant SZ-245 50371-GLUCOSEGENES-FP7-IDEAS-ERC, and National Natural Science Foundation of China (11601259).

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

- Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using polyphen-2. *Curr. Protoc. Hum. Genet.*, <https://doi.org/10.1002/0471142905.hg0720s76>.
- Alexander, J., Potamianou, H., Xing, J., Deng, L., Karagiannidis, I., Tsetsos, F., Drineas, P., Tarnok, Z., Rizzo, R., Wolanczyk, T., et al. (2016). Targeted re-sequencing approach of candidate genes implicates rare potentially functional variants in Tourette syndrome etiology. *Front. Neurosci.* 10: 428.
- Almasy, L., Dyer, T.D., Peralta, J.M., Kent Jr, J.W., Charlesworth, J.C., Curran, J.E., and Blangero, J. (2011). Genetic analysis workshop 17 mini-exome simulation. *BMC Proc.* 5(Suppl. 9): S2.

- Ball, M.P., Thakuria, J.V., Zaranek, A.W., Clegg, T., Rosenbaum, A.M., Wu, X., Angrist, M., Bhak, J., Bobe, J., Callow, M. J., et al. (2012). A public resource facilitating clinical use of genomes. *Proc. Natl. Acad. Sci. U. S. A.* 109: 11920–11927.
- Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* 12: 745–755.
- Blangero, J., Teslovich, T.M., Sim, X., Almeida, M.A., Jun, G., Dyer, T.D., Johnson, M., Peralta, J.M., Manning, A., Wood, A.R., et al. (2016). Omics-squared: human genomic, transcriptomic and phenotypic data for genetic analysis workshop 19. *BMC Proc.* 10(Suppl. 7): 20.
- Boyd-Kirkup, J.D., Green, C.D., Wu, G., Wang, D., and Han, J.D. (2013). Epigenomics and the regulation of aging. *Epigenomics* 5: 205–207.
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45: 5–32.
- Bush, W.S., Crosslin, D.R., Owusu-Obeng, A., Wallace, J., Almoguera, B., Basford, M.A., Bielinski, S.J., Carrell, D.S., Connolly, J.J., Crawford, D., et al. (2016). Genetic variation among 82 pharmacogenes: the PGRNseq data from the eMERGE network. *Clin. Pharmacol. Therap.* 100: 160–169.
- Cirulli, E.T. and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole genome sequencing. *Nat. Rev. Genet.* 11: 415–425.
- De Baets, G., Durme, J.V., Reumers, J., Maurer-Stroh, S., Vanhee, P., Dopazo, J., Schymkowitz, J., and Rousseau, F. (2012). SNPEffect 4.0: online prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res.* 40: D935–D939.
- Efron, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *J. Am. Stat. Assoc.* 104, <https://doi.org/10.1198/jasa.tm08523>.
- Engelman, C.D., Greenwood, C.M.T., Bailey, J.N., Cantor, R.M., Kent, J.W. Jr., Konig, I.R., Bermejo, J.L., Melton, P.E., Santorico, S.A., Schillert, A., et al. (2016). Genetic Analysis Workshop 19: methods and strategies for analyzing human sequence and gene expression data in extended families and unrelated individuals. *BMC Proc.* 10(Suppl. 7): 19.
- Fuchsberger, C., Flannick, J., Teslovich, T.M., Mahajan, A., Agarwala, V., Gaulton, K.J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D.J., et al. (2016). The genetic architecture of type 2 diabetes. *Nature* 536: 41–47.
- Goldstein, D.B., Allen, A., Keebler, J., Margulies, E.H., Petrou, S., Petrovski, S., and Sunyaev, S. (2013). Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.* 14: 460–470.
- Gonzalez-Garay, M.L. (2014). The road from next-generation sequencing to personalized medicine. *J. Personalized Med.* 11: 523–544.
- Hoerl, A.E., and Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12: 55–67.
- Ker, A.P. (2001). On the maximum of bivariate normal random variables. *Extremes* 4: 185–190.
- Kircher, M., Witten, D.M., Jain, P., O’Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46: 310–315.
- Krupp, D.R., Barnard, R.A., Duffourd, Y., Evans, S.A., Mulqueen, R.M., Bernier, R., Rivière, J.B., Fombonne, E., and O’Roak, B.J. (2017). Exonic mosaic mutations contribute risk for autism spectrum disorder. *Am. J. Hum. Genet.* 101: 369–390.
- Kumar, P., Henikoff, S., and Ng, P. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 4: 1073–1081.
- Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* 83: 311–321.
- Li, G.X., Ferguson, J., Zheng, W., Lee, J.S., Zhang, X.H., Li, L., Kang, J., Yan, X.T., and Zhao, H.Y. (2011). Large-scale risk prediction applied to genetic analysis workshop 17 mini-exome sequence data. *BMC Proc.* 5(Suppl. 9): S46.
- Li, G.X., Cui, Y.H., and Zhao, H.Y. (2015). An empirical Bayes risk prediction model for multiple traits using sequencing data. *Stat. Appl. Genet. Mol. Biol.* 1: 551–73.
- Luo, L., Boerwinkle, E., and Xiong, M. (2011). Association studies for next-generation sequencing. *Genome Res.* 21: 1099–108.
- Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 5: e1000384, <https://doi.org/10.1371/journal.pgen.1000384>.
- Majewski, J., Schwartzenruber, J., Lalonde, E., Montpetit, A., and Jabado, N. (2011). What can exome sequencing do for you?. *J. Med. Genet.* 48: 580–589.
- Marouli, E., Graff, M., Lettre, G., Lo, K.S., Wood, A.R., Kjaer, T.R., Fine, R.S., Lu, Y., Schurmann, C., Highland, H.M., et al. (2017). Rare and low-frequency coding variants alter human adult height. *Nature* 542: 186–190.
- Mathur, P., Medicherla, K.M., Chaudhary, S., Patel, M., Bagali, P., and Suravajhala, P. (2018). Whole exome sequencing reveals rare variants linked to congenital pouch colon. *Sci. Rep.* 8: 6646.
- Nadarajah, S. and Kotz, S. (2008). Exact distribution of the Max/Min of two Gaussian random variables. *IEEE Trans. Very Large Scale Integr. Syst.* 16: 2.
- Need, A.C., Shashi, V., Hitomi, Y., Schoch, K., Shianna, K.V., McDonald, M.T., Meisler, M.H., and Goldstein, D.B. (2012). Clinical application of exome sequencing in undiagnosed genetic conditions. *J. Med. Genet.* 49: 353–361.
- Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461: 272–276.

- Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M., Gildersleeve, H., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* 42: 790–793.
- Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* 86: 832–838.
- Renkema, K.Y., Stokman, M.F., Giles, R.H., and Knoers, N.V.A.M. (2014). Next-generation sequencing for research and diagnostics in kidney disease. *Nat. Rev. Nephrol.* 10: 433–444.
- Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D’Amico, A.V., Richie, J.P., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1: 203–209.
- Sun, H. and Yu, G.J. (2019). New insights into the pathogenicity of non-synonymous variants through multi-level analysis. *Sci. Rep.* 9.
- The UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. *Nature* 526: 82–90.
- Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* 99: 6567–6572.
- Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Human Genet.* 89: 82–93.
- Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42: 565–569.
- Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N. Engl. J. Med.* 369: 1502–1511.
- Yourshaw, M., Taylor, S.P., Rao, A.R., Marttin, M.G., and Nelson, S.F. (2015). Rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins. *Briefings Bioinf.* 16: 255–264.
- Zhang, H. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* 9, <https://doi.org/10.1093/biostatistics/kxn001>.

Supplementary Material: Supplementary material to this article can be found online at <https://doi.org/10.1515/sagmb-2019-0056>.