SNP Set Analysis for Detecting Disease Association Using Exon Sequence Data

Ru Wang[1,2,*], Jie Peng[1], Pei Wang[2]

[1] Department of Statistics, University of California, Davis, CA, 95616
[2]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109

[*] Corresponding author

Email Addresses:
RW: ruwang@ucdavis.edu
JP: jie@wald.ucdavis.edu
PW: pwang@fhcrc.org

Abstract

Rare variants are believed to play important roles in disease etiology. Recent advances in high throughput sequencing technology enable one to systematically characterize the genetic effects of both common and rare variants. In this paper, we introduce several approaches which simultaneously test the effects of common and rare variants within a SNP set based on logistic regression models and logistic kernel machine models. Gene-environment interactions and SNP-SNP interactions are also considered in some of these models. We illustrate the performance of these methods using the unrelated individual data from Genetic Analysis Workshop 17. Three true disease genes, *FLT1*, *PIK3C3*, and *KDR*, have been consistently selected by the proposed methods. In addition, compared to logistic regression models, the logistic kernel machine models are more powerful, presumably because the latter reduce effective number of parameters through regularization. Our results also suggest that, a screening step is effective in decreasing the number of false positive findings which is often a big concern for association studies.

Background

High-throughput sequencing technologies have been evolving extraordinarily fast in the past few years. They have been recently applied to genome-wide association studies (GWAS) to study the effects of both common and rare variants. The different natures of these two types of variants call for distinct methods. For common variants, association tests based on individual SNPs are still widely used. However, such approaches suffer from multiple comparison problems and do not take into account possible interactions among variants. To overcome these limitations, SNP set based analysis have been developed for testing the joint effect (either linear or nonlinear) of variants within a SNP set. For instance, in [1], a kernel machine based method is proposed for association studies, which is flexible in modeling various interactions and nonlinear effects. In [2], similarity scores of genotypes between pairs of individuals are first derived using a kernel and then these scores are used as the response variable in an ANOVA model to establish association between genotypes and phenotypes. Such methods tend to be more powerful and flexible than individual SNP analysis. While many GWAS studies in the past focus on common variants, it is now widely believed that, for complex diseases, rare variants are more likely to be functional than common variants [3]. Since rare variants usually have very low marginal effects, multiple rare variants within a SNP set (e.g., a gene or a pathway) are thus often combined into a single variable to be used in tests for association. For example, [4] propose a method by collapsing multiple rare variants to a single indicator recording whether the genome contains any rare variant for the SNP set under consideration or not; [5] propose a weighted sum score where the weight for each variant indicator (0-absent, 1-present) is proportional to the inverse of its estimated standard deviation in the population. An overview of the rare variants collapsing methods is provided by [6].

In order to effectively detect the association signals, it could be beneficial to jointly model the common and rare variants, as well as account for correlations among both variants. For this purpose, in this paper, we introduce several methods to jointly model the common and rare variants within a SNP set. Note, throughout this paper, SNPs with minor allele frequency less than 1% are treated as rare variants and all other SNPs are treated as common variants. We start with logistic regression models including gene-environment interaction terms, and derive score statistics for testing the presence of any marginal or interaction effects. We then consider logistic kernel machine models which can incorporate both interactions among SNPs and gene-environment interactions. This model is an extension of the method proposed in [1,7]. We also introduce a summary score for combining common variants based on the idea of principal fitted components [8], which is then used to reduce dimensionality of the logistic regression model. We then use the 200 independently simulated data sets for unrelated individuals from GAW17 [9] to illustrate these methods, where a SNP set is defined as the observed SNPs (common and rare) within a gene. We also employ a two-stage procedure consisting of a screening stage and a testing stage when analyzing the GAW17 data. The results suggest that the kernel machine methods enjoy better power than the score tests, and the screening stage helps to reduce the number of false positive findings.

Methods

Logistic regression models and score tests

For the *i*th individual ($i = 1,\cdots,n$), let response y_i be 0 if unaffected, and 1 if affected. Let X_i be a $q\times 1$ covariates vector (including an intercept term), z_i be a $p \times 1$ vector of SNP genotypes (or summary scores) for a given gene (SNP set) under testing, and s_i be the environment covariate which is also included in X_i. We consider the logistic regression model with gene-environment interactions,

$$logit(p_i) = X_i^T\beta + a^T z_i + s_i \cdot b^T z_i, i = 1,\cdots,n, \qquad (1)$$

where $p_i = \Pr(y_i = 1|X_i, z_i)$. The goal is to test the null hypothesis $H_0\colon a = b = 0$, and we consider the corresponding score statistic. For a detailed derivation and expression of the score statistic, see [10]

Logistic kernel machine models

Following [7, 1], we now extend (1) to a semi-parametric logistic regression model

$$logit(p_i) = X_i^T\beta + h(z_i) + s_i \cdot g(z_i), i = 1,\cdots,n, \qquad (2)$$

where $h(\cdot)$ and $g(\cdot)$ belong to *reproducing kernel Hilbert spaces* H_K and $H_{\widetilde{K}}$ generated by kernels $K(\cdot,\cdot)$ and $\widetilde{K}(\cdot,\cdot)$, respectively. Considering penalized likelihood, $h(\cdot)$ and $g(\cdot)$ can be estimated by

$$(\hat{h},\hat{g}) = argmax_{h\in H_K, g\in H_{\widetilde{K}}}\left\{\sum_{i=1}^{n}\left(y_i\log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)\right) - \frac{1}{\lambda}\|h\|^2_{H_K} - \frac{1}{\tilde{\lambda}}\|g\|^2_{H_{\widetilde{K}}}\right\} \qquad (3)$$

Following [7], the above solutions have the same form as the Penalized Quasi-Likelihood estimators from the logistic mixed model:

$$logit(p_i) = X_i^T\beta + h_i + s_i \cdot g_i,\ i = 1,\cdots,n, \qquad (4)$$

where $h_i \sim_{i.i.d.} N_n(0, \frac{1}{\lambda}K)$, $g_i \sim_{i.i.d.} N_n(0, \frac{1}{\tilde{\lambda}}\widetilde{K})$, and $K := (K(z_i, z_j))$, $\widetilde{K} :=$

$(\widetilde{K}(z_i, z_j))$, and h_i's and g_i's are independent. Denote $\tau = 1/\lambda$, and $\tilde{\tau} = 1/\tilde{\lambda}$. Now, testing the null hypothesis of no genetic effects $H_0: h(\cdot) = g(\cdot) = 0$ in (2) can be reformulated as testing the absence of the variance components $H_0: \tau = \tilde{\tau} = 0$ in model (4). As in [7, 1], we consider the (two-dimensional) test statistic $Q^* = \binom{Q^*_\tau}{Q^*_{\tilde{\tau}}}$ based on the score statistic of $(\tau, \tilde{\tau})$. The two components of Q^* can be approximated by scaled chi-square distributions $\kappa^*_\tau \chi^2_{(\upsilon^*_\tau)}$, $\kappa^*_{\tilde{\tau}} \chi^2_{(\upsilon^*_{\tilde{\tau}})}$, respectively, through matching the means and variances [7]. Finally, we construct a combined test statistic $Q^*_{max} = \max\left(\frac{Q^*_\tau}{\kappa^*_\tau}, \frac{Q^*_{\tilde{\tau}}}{\kappa^*_{\tilde{\tau}}}\right)$. The corresponding p-value is then

$$\text{p-value} = 1 - F_{\chi^2}(Q^*_{max}, \upsilon^*_\tau) \cdot F_{\chi^2}(Q^*_{max}, \upsilon^*_{\tilde{\tau}}),$$

where $F_{\chi^2}(\cdot, \upsilon)$ is the cumulative distribution function of a chi-square distribution with υ degrees of freedom. For detailed derivations and expressions of Q^*, κ^*_τ, $\kappa^*_{\tilde{\tau}}$, υ^*_τ, and $\upsilon^*_{\tilde{\tau}}$, see [10]. Note, when both K and $\widetilde{K}$ are linear kernels, i.e., $K(z_i, z_j) = \widetilde{K}(z_i, z_j) = z_i^T z_j$, models (1) and (2) have the same form. However, they are treated differently and consequently the corresponding test statistics are different.

Summary score for common variants

For a gene with *p* common variants, we introduce a summary score:

$$comm.score_i \coloneqq \sum_{k=1}^{p} I_{ik} \cdot logit(\hat{p}_k^A)/logit(\hat{p}_k^U)\,, i = 1, \cdots, n, \qquad (5)$$

where I_{ik}= the number of times the *k*th variant being observed in the *i*th individual;

$$\hat{p}_k^A = \frac{m_k^A + 1}{2n^A + 2}, \hat{p}_k^U = \frac{m_k^U + 1}{2n^U + 2},$$

with m_k^A, m_k^U being the number of times the *kth* variant being observed among affected and unaffected individuals respectively; and n^A, n^U being the total numbers of affected and unaffected individuals, respectively. This summary score is derived based on the idea of principal fitted components for dimension reduction [8].

Two-stage procedure

We propose a two-stage procedure to analyze the GAW17 data. In the screening stage, genes that do not show any statistical significance are filtered out. The main purpose of this stage is to achieve dimension reduction and at the same time to retain genes that are more likely to be associated with the disease. In the testing stage, we apply various methods described above to test the subset of genes that have passed the screening criteria.

Stage I: Screening In this stage, both genetic effects and gene-environment interaction effects are investigated, while common and rare variants are handled differently. Common variants are tested in the three sub-populations (i.e., European, Asian and African) separately, while rare variants are studied based on the whole population. For each gene, the genotypes of the common variants (coded as 0, 1, 2 denoting the number of minor alleles) are treated as a vector and a Hotelling's T^2 test is used to test whether there is a mean difference between the affected and unaffected individuals [4]. For rare variants, weighted-sum scores [5] are derived for the synonymous and non-synonymous groups, respectively, denoted by *WS.syn* and *WS.nonsyn*. Then a two-dimension Hotelling's T^2 test is performed based on *WS.syn* and *WS.nonsyn*. To test gene-environment interactions, we consider the null hypothesis, $Corr(G,E|Y=0)=Corr(G,E|Y=1)$. We take the difference between Fisher's z-transformations of sample correlations for the affected and unaffected groups as the test statistic: $T=\log\left(\frac{1+\widehat{Corr}(G,E|Y=1)}{1-\widehat{Corr}(G,E|Y=1)}\right)-\log\left(\frac{1+\widehat{Corr}(G,E|Y=0)}{1-\widehat{Corr}(G,E|Y=0)}\right)$. Again, instead of testing each variant individually, we use combined scores for both common variants (5) and rare variants (the weighted sum score) and test gene-environment interactions for each SNP set as a whole. In addition, for rare variants, we only consider the non-synonymous variants.

In all the above tests, the p-values are determined through permuting disease status (while keeping the total numbers of affected and unaffected individuals unchanged). Finally, genes are deemed to pass the screening and become candidates for the testing stage if they have (unadjusted) p-values smaller than a pre-specified threshold (e.g. 0.1) for at least one of the above tests.

<u>*Stage II: Testing*</u> In this stage, two kinds of models are considered — logistic regression models (1) and logistic kernel machine models (4). For all models, the covariates vector consists of age, sex, two principal component scores to account for population structures (see Results section for more details), as well as an environmental factor – the smoke status. For rare variants, we further introduce a combined weighted-sum score:

WS.combined = WS.syn + 2WS.nonsyn,

where non-synonymous variants receive more weights.

For logistic regression models, two different scenarios are considered for the common variants, one using the original genotypes (referred to as *logistic regression*) and the other using the common score (5) with the weights calculated based on the corresponding screening data set (referred to as *logistic common.score*). In addition, *WS.combined* is used for both scenarios. Finally, score statistics are calculated and the p-values are determined by theoretical χ^2 distributions.

For logistic kernel machine models (4), the original genotypes are used for common variants. We consider two different schemes for the kernels. One uses linear kernels for both K and $\widetilde{K}$, and the other uses quadratic kernel for K which models interactions among variants and linear kernel for $\widetilde{K}$. One would expect that quadratic kernel is more powerful if there are SNP-SNP interactions, while linear kernel may be more powerful if such interactions are absent. For the quadratic kernel case, *WS.combined* is used and the method is referred to as *quad rare.WS.combined*. While for the linear kernel case, two scenarios are considered for combining rare variants, one using *WS.combined* (referred to *linear rare.WS.combined*), another using

WS.nonsyn (referred to as *linear rare.WS.nonsynonymous*). Moreover, for the kernel machine methods, the weighted sum scores for rare variants and the genotypes of the common variants are both standardized (to have mean zero and standard deviation one) before model fitting.

In total, we consider five different methods in the testing stage, which are summarized in Table 1.

Results

GAW17 Data description

The GAW17 data we analyzed in this paper have 200 replicates, each consisting of data for 697 unrelated individuals. The genotypes, age and sex of these individuals are from real studies and are kept fixed across the 200 replicates. One environmental risk factor – the smoke status and a binary disease status were simulated for each replicate [9]. Moreover, in all these replicates, the total numbers of affected and unaffected individuals are fixed to be 209 and 488, respectively, which reflects the population prevalence of this disease.

The 697 individuals were from seven different sources: Denver-Chinese, Han-Chinese, Japanese, Luhya, Yoruba, CEPH, and Tuscan. Through principal component analysis on about 1000 common variants (distance≥50,000 bp) with minor allele frequency (MAF) larger than 10%, the first two principal components clearly divide the sample into three distinct clusters, corresponding to African (Luhya and Yoruba), Asian (Chinese and Japanese) and Caucasian (CEPH and Tuscan).

The genotype data consist of 24487 SNPs from 3205 genes on 22 autosomal chromosomes. MAF for 74% of SNPs is less than 1%. In our analysis, these are treated as rare variants, while all other SNPs are treated as common variants. Moreover, 2208 genes contain at least one common variant and the maximum number of common variants within a gene is 52. 2476 genes contain at least one rare variant and the maximum number is 179. 162 rare variants are removed from the subsequent analysis since they only appear in one individual. Genes with rare variant event occurring in < 1% individuals are removed and 2534 genes are left for subsequent analysis. In the end, genotypes are coded as 0, 1 and 2, indicating the number of minor alleles at each locus.

Findings

We randomly divide the 200 simulated replicates into 100 pairs. For each pair, one data set is used for screening and the other is used for testing. Across the 100 screening date sets, if 0.1 threshold is used, the mean number of genes passing screening is 1307 and eight genes (*RUNX2*, *MUC3A*, *TMEM67*, *NIBP*, *AKAP2*, *GOLGA1*, *USP5*, and *FLT1*) are selected at least 95 times. If the 0.05 threshold is used, the mean number of genes passing screening is 824 and one gene (*FLT1*) is selected 95 times. For each pair of screening and test data sets,

genes that pass the screening step are tested using the five methods described in the previous section. P-values are adjusted by the Holm's procedure [11] which is an improvement of the Bonferroni's procedure and controls the family wise error rate (FWER). A gene is then said to be selected by a method, if its corresponding adjusted p-value is less than 0.1. Throughout the 100 pairs of screening and test data sets, if threshold 0.1 is used in the screening step, then four genes (*FLT1, PIK3C3, KDR, PRR4*) are selected for more than 10 times by at least one of the five testing methods. In contrast, if no screening is employed (i.e., all 2534 genes are passed to the testing stage), nine genes are selected for more than 10 times by at least one of the five testing methods. The selection frequencies of these genes are illustrated in Figure 1. As can be seen there, Gene *FLT1* is selected over 40 times by *linear rare.WS.combined*, and over 50 times by *linear rare.WS.nonsynonymous*. Moreover, Gene *PIK3C3* and Gene *KDR* are selected for about 20 times by *linear rare.WS.combined* and *quad rare.WS.combined*, respectively. Note, quadratic kernel model is capable of capturing some of the SNP-SNP interaction effects, while the linear kernel model does not. Thus, quadratic kernel working better for Gene *KDR* may imply that there are potential SNP-SNP interaction effects in this gene, which may result from the complicated disease model and/or correlation structure among the SNPs. Compared with the kernel machine methods, the two logistic regression methods give less consistent results in terms of gene selection across the replicates. Furthermore, summarizing information of common variants by *common.score* seems to improve the power of the logistic regression model slightly.

Gene *FLT1* is on chromosome 13, and it contains 35 SNPs, among which 25 are rare variants. Applying the logistic regression model with gene-environment interaction (1) on the first replicate indicates that the (common) variant *C13S523* associates with disease status highly significantly (nominal p-value= 0:000817). This variant is non-synonymous with MAF being 6.7%. The weighted sum score of the rare variants in *FLT1* also shows evidence of association (nominal p-value= 0.0033). Gene *KDR* is on chromosome 4 with 14 rare variants and 2 common variants. Gene *PIK3C3* has 7 variants, 6 rare and 1 nonsynonymous common variant. It also seems that this common variant is the reason that Gene *PIK3C3* is picked by *linear rare.WS.combined* about 20 times across the 100 replicates.

The above results were obtained without the knowledge of the underlying disease model. Afterwards, we examine the GAW17 simulation model [9]. It turns out that, *FLT1, PIK3C3, KDR* are true disease susceptible genes. However, other genes reported in Figure 1 are not directly related to disease status. By comparing the top and bottom panels in Figure 1, it appears that the procedure with screening step is effective in eliminating such genes. A closer look of the results reveals that, these genes are mainly filtered by the screening step. For instance, Gene *TAS2R48* was detected as a significant gene among 18 (out of 100) data pairs by the *linear.rare.WS.combined* method when no screening is applied. However for 15 out of 18 pairs, *TAS2R48* would not pass the screening step if 0.1 threshold was used.

Conclusions

In this paper, we consider SNP set analysis for detecting disease susceptible variants using exon sequence data. In large scale association studies, there is often a necessity to combine information across variants to improve detection power. This is especially the case for rare variants. In this paper, we adopt the weighted sum score by [5] to summarize information across rare variants within each SNP set. In addition, we propose a summary score based on principal fitted components [8] to combine information across common variants. Moreover, large number of variants also poses challenges such as multiple comparisons, modeling various interactions, etc. To address this issue, we extend the logistic kernel machine methods in [1, 7] to include gene-environment interactions. Compared to logistic regression models, the logistic kernel machine models are more powerful, which estimate the degrees of freedom in a data adaptive way by accounting for correlations among the SNPs. Thus they reduce the effective number of parameters and consequently enjoy improvements in power. Kernel machine models also have greater degrees of flexibility in modeling interactions and nonlinearity. We also apply a two-step procedure consisting of a screening stage and a testing stage to the GAW17 data. The results suggest that, the screening stage is effective in decreasing the number of false positive findings which is often a big concern for association studies.

Acknowledgements

This work is supported by grant R01GM082802 from the National General Institute of Medicine. The Genetic Analysis Workshop is supported by NIH grant R01 GM031575.

References

1. M Wu, P Kraft, M Epstein, D Taylor, S Chanock, D Hunter, X Lin: **Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies**. *The American Journal of Human Genetics* 2010, 86:929-942.
2. I Mukhopadhyay, E Feingold, D Weeks, A Thalamuthu: **Association Tests Using Kernel-Based Measures of Multi-Locus Genotype Similarity Between Individuals.** *Genetic Epidemiology* 2010, 34: 213-221.
3. W Bodmer, C Bonilla: **Common and Rare Variants in Multifactorial Susceptibility to Common Diseases.** *Natural Genetics* 2008,40: 695-701.
4. B Li, S Leal: **Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data.** *The American Journal of Human Genetic* 2008, 83: 311-321.
5. B Madsen, S Browning: **A Group-wise Association Test for Rare Mutation Using a Weighted Sum Statistics.** *PLoS Genetics* 2009, 5(2): e1000384. Doi:10.1371/journal.pgen.1000384
6. C Dering, E Pugh, A Ziegler: **Statistical analysis of rare sequence variants: An overview of collapsing methods.** *Genetic Epidemiology* 2011, *GAW SUPPL.*
7. D Liu, D Ghosh, X Lin: **Estimation and Testing for the Effect of a Genetic**

Pathway on a Disease Outcome using Logistic Kernel Machine Regression via Logistic Mixed Models. *BMC Bioinformatics* 2008, 9: 292. doi: 10.1186 / 1471-2105-9-292

8. R Cook, L Forzani: **Principal Fitted Components for Dimension Reduction in Regression.** *Statistical Science* 2008, 23: 485-501.
9. J Blangero et al.: **Genetic Analysis Workshop 17 mini-exome simulation.** *BMC Proceedings* 2011, this volume.
10. R Wang, J Peng, P Wang: **A note on logistic regression and logistic kernel machine models.** arXiv: submit/0207473.
11. S Holm: **A simple sequentially rejective multiple test procedure.** *Scand. J. Statist* 1979, 6: 65-70.

Figures

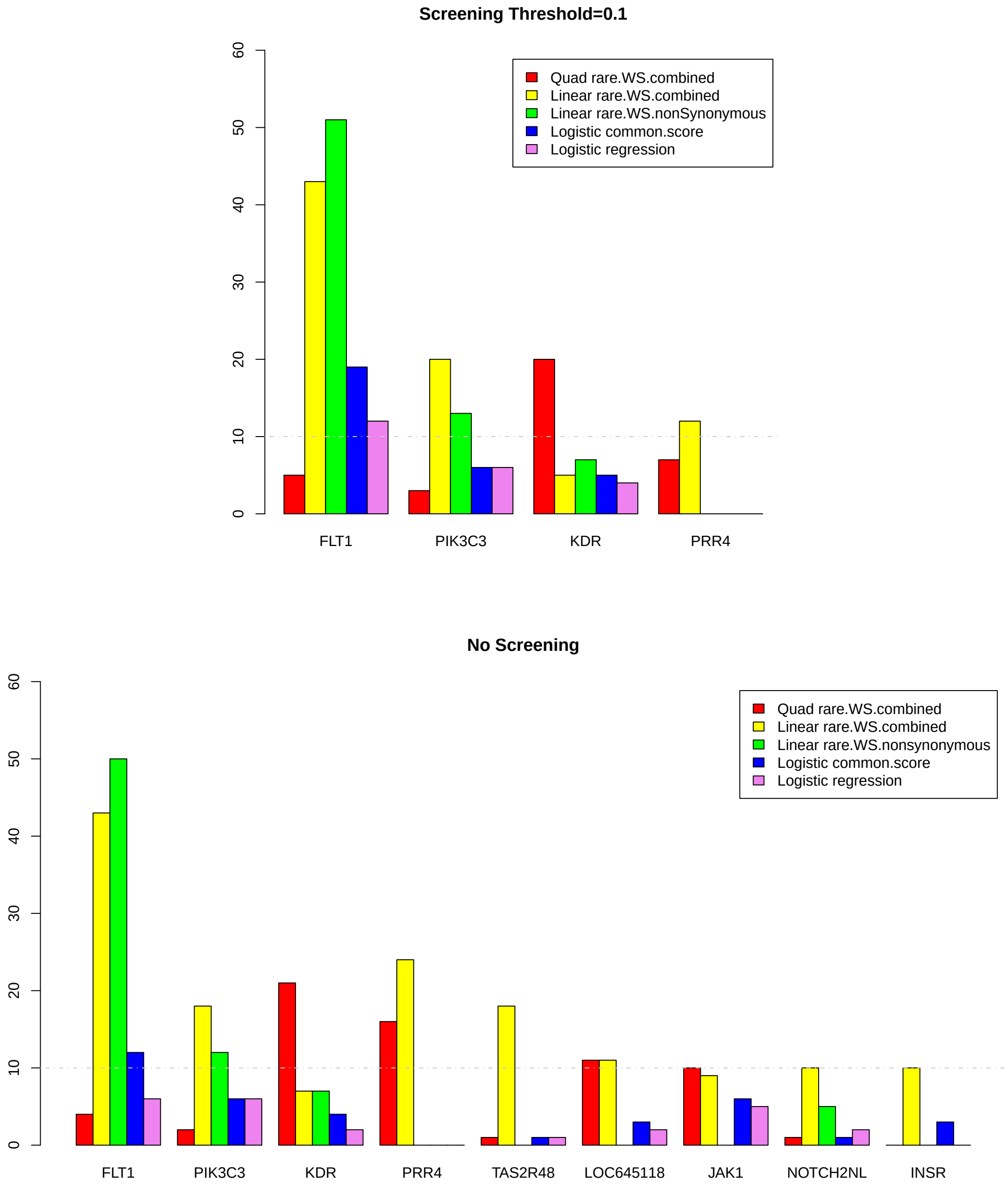
Figure 1 - Frequently selected genes and their selection frequencies.

For each gene, the height of the bar represents the number of times it's being selected across the 100 screening-testing pairs. Top panel: 0.1 threshold in screening stage; Bottom panel: No screening.

Tables

Table 1 - Methods in the testing stage.

Method	Model	Kernel	Common variants	Rare variants
logistic regression	logistic regression	NA	genotypes	WS.combined
logistic common.score	logistic regression	NA	common.score	WS.combined
linear rare.WS.combined	kernel machine	linear	genotypes	WS.combined
linear rare.WS.nonsynonymous	kernel machine	linear	genotypes	WS.nonsyn
quad rare.WS.combined	kernel machine	quadratic	genotypes	WS.combined


Screening Threshold=0.1

No Screening

Figure 1: **Frequently selected genes and their selection frequencies**. For each gene, the height of the bar represents the number of times its being selected across the 100 screening-testing pairs. Top panel: 0.1 threshold in screening stage; bottom panel: No screening.