kTWAS: integrating kernel-machine with transcriptome-wide association studies improves statistical power and reveals novel genes

Chen Cao[1], Devin Kwok[2], Shannon Edie[3], Qing Li[1], Bowei Ding[2], Pathum Kossinna[1], Simone Campbell[1,4], Jingjing Wu[2], Matthew Greenberg[2], Quan Long[1, 2,5,6#].

[1]Department of Biochemistry & Molecular Biology, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Canada.
[2]Department of Mathematics & Statistics, University of Calgary, Calgary, Canada.
[3]Department of Biology, Queen's University, Kingston, Ontario, Canada.
[4]Heritage Youth Researcher Summer Program.
[5]Department of Medical Genetics, University of Calgary, Calgary, Canada.
[6]Hotchkiss Brain Institute, O'Brien Institute for Public Health, University of Calgary, Calgary, Canada.
[#]Correspondence should be addressed to quan.long@ucalgary.ca

Abstract

The power of genotype-phenotype association mapping studies increases significantly when the contributions of multiple variants in a focal genetic region are aggregated effectively. Currently, two categories of frequently used methods are used to aggregate variants. Transcriptome-wide association studies (TWAS) represent a category of emerging methods that utilize gene expressions to select genetic variants, before using a pretrained linear combination of selected variants for downstream association mapping. In contrast, kernel methods such as SKAT measure the genetic similarity in a focal region as modelled by various types of kernels to associate genotypic and phenotypic variance, allowing such methods to model nonlinear effects. Thus far, no thorough comparison has been made between these categories, and there are also no methods that integrate these two approaches. In this work we have developed a novel method called kTWAS that leverages TWAS-like feature selection followed by a SKAT-like kernel-based score test, to combine advantages from both approaches. We demonstrate the improved power of kTWAS against TWAS and multiple SKAT-based protocols through extensive simulations, and identify novel disease associated genes in WTCCC genotyping array data and MSSNG (Autism) sequence data. The source code for kTWAS and our simulations are available in our GitHub repository (https://github.com/theLongLab/kTWAS).

Keywords

Transcriptome-wide association studies (TWAS), Kernel methods, Power analysis, Nonlinear genetic effects.

Introduction

Transcriptome-wide association studies (TWAS) have emerged as an important technique for associating genetic variants and phenotypic changes[1-5]. Pioneered by Gamazon et al.[6], TWAS is typically conducted in two steps: First, a model is trained to predict gene expression from genotypes, using a reference dataset which contains both expression and genotype information for each sampled individual. Techniques including ElasticNet[6], Bayesian sparse linear mixed mode (BSLMM)[7-9], deep auto-encoder mode[10] and deep learning regression model[11] are used to fit this genotype-expression model. The pretrained genotype-expression model is then used to predict expression activity from the main dataset for genotype-phenotype association mapping (referred to as GWAS dataset hereafter), which contains genotype and phenotype information, but not expression data, for each case or control in the GWAS cohort. Initiated by Gusev et al.[8], methods using summary statistics in the GWAS dataset (i.e., meta-analysis) to conduct TWAS were also developed[12]. The key insight of TWAS is that transcriptomic data can be used to select for genetic variants critical to gene expression (i.e., eQTLs), which improve the quality of downstream GWAS. TWAS effectively aggregates sensible genetic variants as linear combinations by modelling genetic contributions to expression. This approach is generally effective regardless of the performance of the genotype-expression model. As a result, despite the low predictive power of the genotype-expression models used in TWAS (with an average R^2around 1%), TWAS has led to significant successes in real data[3, 4, 13-17]. Indeed, as demonstrated in our simulations[18], the use of predicted expressions from a genotype-phenotype model in TWAS can result in better power than the use of experimentally measured expressions. This may be because predicted gene expressions capture the genetic contributions more precisely than actual expression levels, which are subject to noise from environmental factors.

The popularity of TWAS has overshadowed another well-established branch of kernel-machine based models for analyzing genetic association, such as the sequence association kernel test[19, 20](**Table 1**). The key insight of kernel methods is that the similarity of a genetic region between different subjects can be used to associate the variance of the genotypes and phenotypes in that region without knowing which specific genetic variants are causal in the focal region. As a result, kernel-machine based methods can model the aggregated effects of multiple genetic variants and capture genetic interactions within a local region while being robust to noise. At first glance, TWAS and kernel-based models appear quite different, as TWAS utilizes expression whereas kernel methods are purely DNA-based. Intuitively, TWAS appears to be more powerful as it integrates more information in the form of expression data. However, in our opinion, TWAS and kernel methods are quite comparable because they are both gene-set analyses which test an aggregated set of genetic variants for associations with phenotype. Essentially, TWAS selects and weights genetic variants for aggregation using a linear model, whereas kernel methods rank the relative importance of pre-selected genetic variants using various kernel machines. From the perspective of machine learning, these two methods cover two complementary aspects of feature engineering: how to select and how to organize features in high-dimensional data analysis. Kernel methods organize genetic variants in a flexible way to cope with unknown genetic architecture, but do not provide a quantitative way to pre-select sensible variants, while TWAS models select sensible variants by utilizing gene expressions but assume that variants must be linearly associated to phenotype.

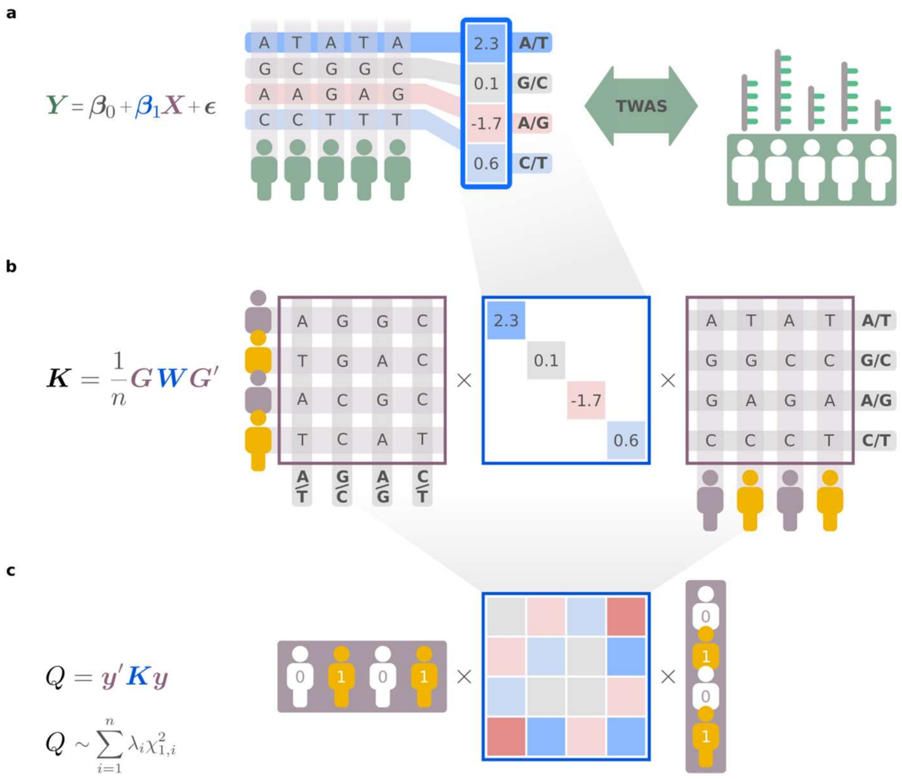

Figure 1. The protocol of kTWAS. (a) Taking advantage of the pretrained genotype-expression model ElasticNet based on PrediXcan. **(b)** the kernel K_W from SKAT, where K_W is based on the weight of each variant $W = (\beta_1, \dots, \beta_m)$ in the pretrained genotype-expression model in **a**. **(c)** Q score test from SKAT using the TWAS-informed kernel K from above, where Q follows a mixture of chi-squared distributions under the null hypothesis which is shown in **c**.

Surprisingly, a thorough comparison has yet to be made between TWAS and kernel methods. In their pioneering work on PrediXcan, Gamazon et al.[6] applied SKAT and PrediXcan to Wellcome Trust Case Control Consortium (WTCCC) data[21], reporting that PrediXcan produced an elevated proportion of significant variants across all P values . However, it is unclear whether PrediXcan's ability to detect more significant variants can translate directly to prediction of real phenotypic outcomes, as the authors did not conduct simulations (under which the ground truth is known) to quantify the power of competing methods. Further developments in TWAS have incorporated multiple tissues[22], better models for predicting expression[9] combating artifacts caused by co-expressed genes[23] and extending TWAS to other middle-omes such as proteins[24, 25] and images[26]. These later developments include comparisons to the seminal TWAS tools[6, 8], but have not been compared directly against kernel methods such as the flagship tool SKAT. Not only has this lack of comparisons unfairly discouraged the use of kernel methods but has also missed a golden opportunity to integrate the advantages of both approaches to better model the genetic basis of complex traits.

Year	SKAT[19]	SKAT[20]	PrediXcan[6]	PrediXcan[8]
2010	1	4	0	0
2011	14	32	0	0
2012	70	55	0	0
2013	161	53	0	0
2014	242	61	0	0
2015	201	73	21	3
2016	272	62	55	28
2017	263	66	119	96
2018	223	65	154	116
2019	239	46	188	188

Table 1. Number of citations for SKAT and PrediXcan of the last ten years according to Google Scholar.

In this work, we propose a novel model called kTWAS (kernel-based transcriptome-wide association study), which integrates TWAS and kernel methods. We expect that kTWAS will take advantage of TWAS-based feature selection, which is directed by expression data, as well as a kernel-based association test, which is robust to the underlying genetic architecture of the focal phenotype. As a result, the power of kTWAS should be equivalent to TWAS, due to its ability to select genetic variants regulating gene expressions; and also as robust as SKAT to noise and interactions between associated genetic variants.

Using simulated data, we have conducted thorough power comparisons between six protocols: PrediXcan, kTWAS, and four different applications of SKAT under various assumed distributions of genetic effects. (A detailed description and justification of the chosen assumptions is presented in Materials & Methods). Although our main focus is on cases for which subject-level genotypes are available, we have also tested six corresponding protocols in which genotypes are unavailable, and as such the association mapping is conducted using summary statistics (i.e., meta-analysis). We simulate phenotypes based on four representative genetic architectures: an additive model, heterogeneity model, and two interaction models. Multiple effect sizes and heritabilities are tested. While each protocol has unique strengths, in most cases our kTWAS method outperforms alternatives with significant margins. As expected, the relative power of each method is similar when the corresponding meta-analysis method is applied. Moreover, we have conducted extensive real data analysis using kTWAS, which identified a larger number of significant genes with literature support than standard TWAS. Evidence from annotations and literature search support the significance of the novel variants discovered by kTWAS.

In the following section Materials & Methods, we present the design of kTWAS and details of the simulations and power analysis. The outcome of simulations and discoveries from real data are presented in Results. Finally, we conclude the main messages and discuss additional literatures and future work.

Materials & Methods

Mathematical details of SKAT, PrediXcan, and kTWAS

The popular sequence kernel association test (SKAT) tool was selected to represent kernel-based methods. SKAT utilizes a score test that aggregates contributions of multiple genetic variants using a kernel machine[20]. In particular, SKAT employs the following test statistic:

$$Q = \mathbf{Y}'\mathbf{K}\,\mathbf{Y}$$

Where ***Y*** is the phenotype, and ***K*** is a kernel calculated from the centralized genotype matrix ***G*** in the focal region, consisting of G_{ij} samples from *i* individuals with *j* variants per individual. A simple example of a linear kernel is given by $\mathbf{K} = \mathbf{G}'(\mathbf{I})\mathbf{G}/\mathrm{n}$ (where n is the number of variants)

While Wu et al.[20], originally subtracted cofactors such as sex and age from the phenotype vector Y before conducting the score test, to simplify comparisons this work will not include any cofactors to evaluate each model. Furthermore, while additional extensions to SKAT have been developed to handle rare variants[19, 27] and the combined effect of rare and common variants[28], this paper will focus only on common variants, and apply the original SKAT method under four different variant selection methods.

As the earliest TWAS method, PrediXcan was selected to represent TWAS in this paper. PrediXcan is composed of two steps: First, a linear model is trained to predict genetically regulated gene expression (called GReX[6]) using a reference panel in which both genotype and expression data are available:

$$Z \sim \sum \beta_i G_i + \varepsilon$$

Where β_i are the regression parameters to be trained; and $G = (G_1, G_2, \dots, G_m)$ is the genotype matrix in the focal region. Various methods can be used to train this predictive model[8, 29]. In particular, PrediXcan uses ElasticNet[30] to conduct this training. We use the pre-trained PrediXcan genotype-expression model which are available for download on the authors' website[31].

Using the above model, GReX expressions are then estimated from the genotype information of the main GWAS dataset, which provides both genotype and phenotype information:

$$\hat{Z} \sim \sum \beta_i G_i$$

The estimated GReX $\hat{Z}$ is then associated to the phenotype:

$$Y \sim \hat{Z} + \varepsilon$$

Various extensions to PrediXcan have since been developed, covering many cases in association mapping[8, 22, 23, 32, 33]. In particular, Gusev et al. pioneered the first tool utilizing summary statistics[8] to conduct TWAS. To ensure theoretical and technical consistency, in this paper we chose S-PrediXcan, the meta-analysis version of PrediXcan[12], to represent meta-analysis tools in our model comparisons.

To test our hypothesis that SKAT and TWAS have different advantages which can be integrated, we developed the novel method kTWAS, or kernel-based transcriptome-wide association study. The protocol of kTWAS is illustrated in **Fig. 1**. Mathematically,

We first take advantage of the pretrained genotype-expression model ElasticNet based on PrediXcan (**Fig. 1a**):

$$Z \sim \sum \beta_i G_i + \varepsilon$$

We then prepare the kernel $\boldsymbol{K_W}$ from SKAT, where $\boldsymbol{K_W}$ is based on the weight of each variant $W = (\beta_1, \ldots, \beta_m)$ in the pretrained genotype-expression model above(**Fig. 1b**):

$$\boldsymbol{K_W} = \boldsymbol{G'WG}$$

Finally, we conduct the Q score test from SKAT using the TWAS-informed kernel K from above, where Q follows a mixture of chi-squared distributions under the null hypothesis:

$$Q = \boldsymbol{Y'K_WY}$$

As outlined in the introduction, kTWAS is expected to enjoy the advantage of both kernel methods and TWAS so that both feature selections and feature organizations are considered.

Protocols compared

We selected a total of six genotype-based protocols for power comparisons, including kTWAS: (1) SKAT-naive applies the default setting of SKAT, which gives equal weight to all genetic variants in a region. In practice, researchers may use this “naive” version of SKAT if they have no prior knowledge selecting which variants in the focal region. (2) SKAT-S-LM pre-selects genetic variants based on their marginal associations to phenotype, which is assessed by associating each individual variant in the region independently to the phenotype. A linear model, such as the model implemented in PLINK[34], is used to pre-select an arbitrary number of variants. As different genes have wildly varying numbers of causal genotypes contributing to the phenotype, we chose the pragmatic approach of matching the number of markers selected by SKAT-S-LM to the number of variants selected by the ElasticNet model in PrediXcan. (3) SKAT-S-LMM is similar to SKAT-S-LM, but uses a linear mixed model (LMM) to perform variant selection, as implemented by EMMAX[35]. Since LM and LMM are both representative models for conducting single-variant GWAS, we chose to test both to cover a wider spectrum of variant selection methods. (4) SKAT-eQTL pre-selects genetic variants based on published eQTLs[36], instead of by screening for marginal effects in the current GWAS data. Since this method uses independent eQTLs, the variants selected are not jointly modeled linearly in the pre-selection, whereas associating markers directly to phenotype in the GWAS data under study may cause overfitting. As such, we expect SKAT-eQTL covers a different perspective to SKAT-S-LM and SKAT-S-LMM, depending on the marginal effects of individual variants. (5) PrediXcan, as discussed previously, is the first and most representative TWAS tool (6) kTWAS is our novel tool integrating TWAS and SKAT.

Additionally, we conducted an equivalent power analysis of the above six models under meta-analysis, based on the protocols MetaSKAT[27] and S-PrediXcan[12]. In all of the above protocols, the default linear kernel is used in SKAT (https://cran.r-project.org/web/packages/SKAT/SKAT.pdf).

Data simulation procedure and power analysis

Genotype data and selected gene region. We used genotype data from (http://hgdownload.cse.ucsc.edu/gbdb/hg38/1000Genomes/) of the 1000 Genomes Project[37], with a sample size is N = 2548. We used the pretrained genotype-expression models for PrediXcan[6, 12], available at http://predictdb.org/, which are trained for all tissue types available in GTEx (v8). As the sample size and data quality varies between different GTEx tissues, the number of genes for which PrediXcan is applicable also varies. Whole-blood tissue is the largest and most frequently used gene set, with 7252 available genes (together with 1Mb flanking genetic regions) on which PrediXcan can be applied. We therefore simulated 7252 datasets based on the whole-blood tissue gene set.

Genetic architecture and parameterizations. For each gene, we simulated phenotypes based on four different genetic architectures. Their definitions and parameterizations are described below.

(1) “Additive” model. This model associates phenotype with the sum of genetic effects. For each gene, we select a genetic region that includes the gene body and 1Mb of flanking sequences. From this region, 4 single nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) higher than 1% were randomly selected, where 2 SNPs are chosen from variants preselected by the genotype-expression model used in TWAS, and the remaining 2 SNPs are selected from known eQTLs excluding those identified by the genotype-expression model. The first category of SNPs (those identified by the genotype-expression model) favor the performance of PrediXcan, while the second category of SNPs (from eQTLs not identified by the genotype-expression model) favor SKAT related models, as kernels better capture the effects of unsampled variants. To simplify simulations, we fix the number of SNPs from each category to 2, and we further rescale the phenotypic variance components contributed by ElasticNet SNPs versus other eQTL SNPs by a simulation parameter which is set at six different scale factors: 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0.

(2) “Heterogeneity” model. In this model, we randomly select two SNPs in the focal region. Subjects carrying an alternate allele at either or both SNPs will have an associated phenotypic change, while subjects carrying both alternate alleles will not have more changes than the ones have only a single alternate allele. As in the additive model, the use of TWAS-favored SNPs (those selected by ElasticNet) versus kernel-favored SNPs (those selected from random eQTLs) is adjustable. We introduce a parameter to set the number of SNPs selected by ElasticNet as 0, 1, and 2. This parameter is analogous to the scale factor applied to variance contributed by ElasticNet SNPs in the additive model.

(3) & (4) “Both” and “Compensatory” interaction models. We randomly select two SNPs in the focal region in the same way as the heterogeneity model. We also include the same parameter which selects for 0, 1, and 2 ElasticNet SNPs. The effects of the SNPs are differently modeled though. For the interaction model called “Both”, a phenotypic change is made when both alternate alleles are present. For the “Compensatory” interaction model, the phenotypic change is made only when there is exactly one alternate allele in the two SNPs. Subjects carrying both alternate alleles will have the same genetic component (contributing to phenotype) as subjects carrying neither alternate allele. This model reflects the phenomena where a given mutation’s effects are compensated by the presence of another mutation, which is frequently observed in many organisms[38-41].

Heritability. Using one of the four models above, we generated a phenotype whose variance component or heritability equals a preselected value h^2. That is, given the variance of the phenotype’s genetic component as σ_g^2, we calculate σ_e^2 so that $\frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2} = h^2$, and then randomly sample the normal distribution $N(0, \sigma_e^2)$ to determine the contribution of non-genetic components such as noise or environmental effects. Finally, the sum of the genetic and non-genetic components is stored as the simulated phenotype for use in association mappings and power calculations.

Power calculations & Adjustment of type-I errors. For each of the above models and their associated parameters, we simulated 7252 datasets (each containing 2548 subjects) for which the causal variants are randomly selected from the variants a focal region (centralized by a gene and flanked by 1Mb genetic regions). We then test each protocol’s ability to successfully identify the causal gene in each dataset, where we define success as when the association test produces a Bonferroni-corrected[42] p-value lower than a predetermined critical point. We aimed to fix the type-I error across all protocols to α = 0.05. However, due to various reasons including the uneven distribution of genetic variants among the 7252 genes and inherent biases between the protocols (e.g., overfitting caused by SKAT-S-LM and SKAT-S-LMM models), we discovered the type-I errors varied widely when different protocols were tested with a fixed critical value of 0.05. To equalize the type-I error across all protocols, we conducted simulations using random phenotypes with no genetic component whatsoever to empirically determine the null distribution for each protocol. We then analyzed data from all 7252 genetic regions with each protocol and determined the critical point where the smallest (most significant) 5% of all p-values are located. This ensures that all protocols have a fair comparison at a type-I error of 0.05.

The statistical power of each protocol is given by the number of successes divided by the total number of datasets (7252). For the six protocols that utilize genotype data, we conduct association mapping using the simulated genotypes. For the six protocols utilizing summary statistics, we first calculate summary statistics from the simulated genotypes according to the S-PrediXcan and MetaSKAT instruction manuals, before running the protocols.

Real data analysis

We compared the performance of kTWAS and PrediXcan on the WTCCC and MSSNG datasets. WTCCC contains genotype data for 7 complex diseases, with 2,000 cases per disease and approximately 3,000 shared controls of primarily European ancestry. : The diseases surveyed by WTCCC are bipolar disease (BD), coronary artery disease (CAD), Crohn’s disease (CD), rheumatoid arthritis (RA), type 1 diabetes (T1D), type 2 diabetes (T2D), and hypertension (HT). Genotype data was collected from individuals using Affymetrix GeneChip 500K arrays. MSSNG is the largest available whole genome sequencing dataset for Autism Spectrum Disorder (ASD), containing 7065 sequences from ASD patients and controls[43].

Results

Simulations

Type-I errors. The type-I errors estimated by sampling the simulated distribution of the null hypothesis are generally close to the targeted rate $\alpha = 0.05$, except in the case of SKAT-S-LM and SKAT-S-LMM (**Table 2**). This is consistent with the intuition that the pre-selection process of SKAT-S-LM and SKAT-S-LMM aggregates random false effects which inflates type I errors for these protocols. We apply a more stringent cutoff (determined from the simulations described above), to ensure fairness in the power comparisons below.

Model	Type I Error	Model	Type I Error
SKAT-naive	5.50E-02	MetaSKAT-naive	5.36E-02
SKAT-eQTL	5.22E-02	MetaSKAT-eQTL	5.42E-02
kTWAS-S-LM	1.12E-14	MetaSKAT -S-LM	1.04E-14
kTWAS-S-LMM	1.27E-14	MetaSKAT -S-LMM	1.22E-14
PrediXcan	5.04E-02	S-PrediXcan	5.79E-02
kTWAS	5.29E-02	MetakTWAS	4.97E-02

Table 2. Type I error for different models used in our comparison.

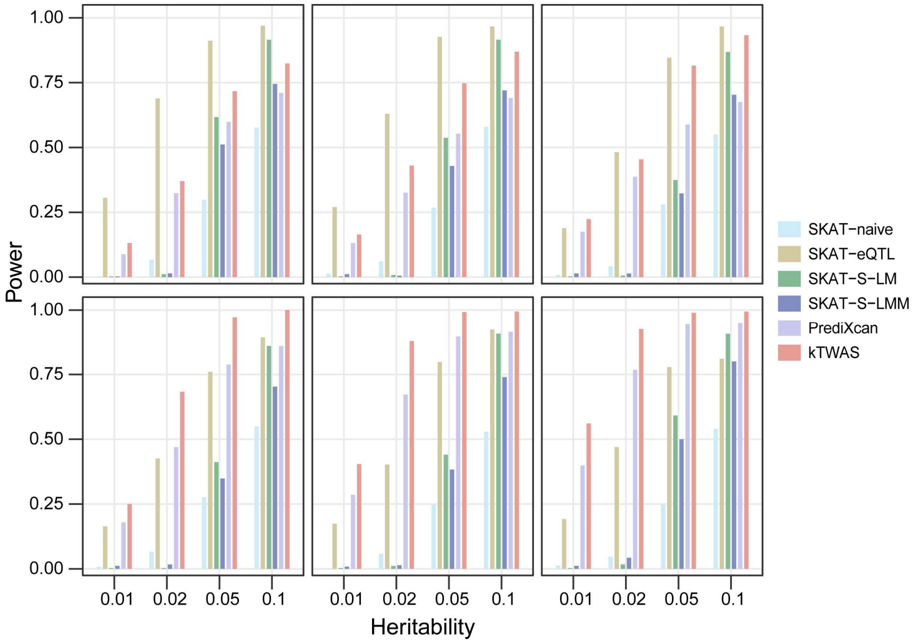

Figure 2. Statistical power (y-axis) of protocols on “Additive” model simulated genotypes, at varying levels of trait heritability (x-axis) and different proportions of ElasticNet-selected SNPs. The compared protocols are SKAT-naive, SKAT-eQTL, SKAT-S-LM, SKAT-S-LMM, PrediXcan, and kTWAS. The proportions of ElasticNet-selected SNPs are 0.0, 0.2, 0.4 for the top row of plots (left to right), and 0.6, 0.8, 1.0 for the bottom row (left to right).

Additive architecture. **Fig. 2** and **Fig. 3** plot the power of the genotype and summary statistic-based protocols under the additive model. kTWAS clearly outperforms PrediXcan at all levels of contribution from ElasticNet-selected SNPs, showing that kernel methods integrated with TWAS-based feature selection can always outperform the linear model utilized by TWAS, even when the underlying genetic architecture is linear.

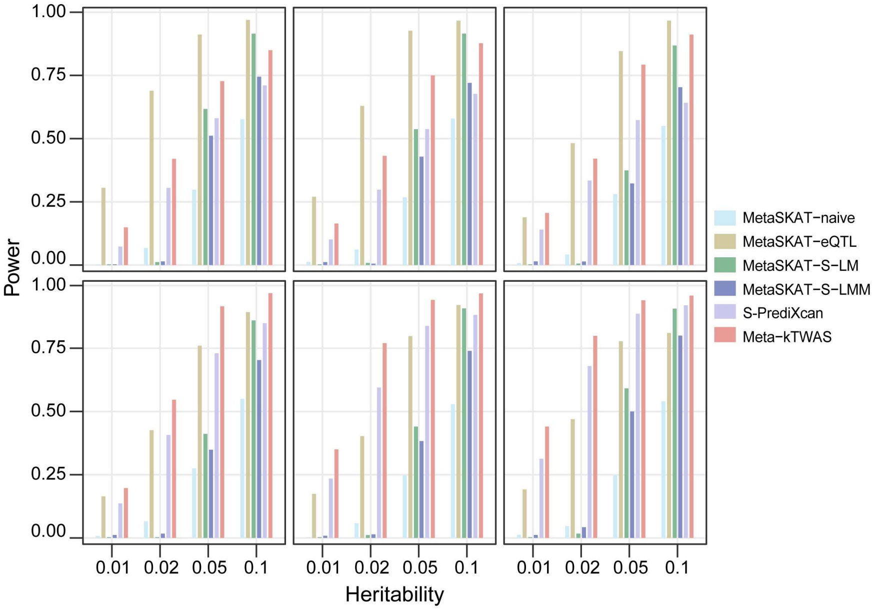

Figure 3. Statistical power (y-axis) of protocols on "Additive" model simulated meta-analysis summary statistics, at varying levels of trait heritability (x-axis) and different proportions of ElasticNet-selected SNPs. The compared protocols are MetaSKAT-naive, MetaSKAT-eQTL, MetaSKAT-S-LM, MetaSKAT-S-LMM, S-PrediXcan, and Meta-kTWAS. The proportions of ElasticNet-selected SNPs are 0.0, 0.2, 0.4 for the top row of plots (left to right), and 0.6, 0.8, 1.0 for the bottom row (left to right).

The comparison between the four SKAT methods show that SKAT-eQTL is the best performing kernel method when the proportion of ElasticNet selected SNPs is low. This is expected, since the non-ElasticNet SNPs favor kernel methods. When the proportion of ElasticNet-selected SNPs is high, which is expected to favor the PrediXcan model, SKAT-eQTL has worse performance than both kTWAS and PrediXcan. SKAT-S-LM and SKAT-S-LMM, which select SNPs based on marginal effects, are generally less powerful than SKAT-eQTL and kTWAS, indicating that their pre-selection process may be overfitting and therefore reducing power (after adjusting for equivalent type I error rates). When regional heritability is low, the power of SKAT-S-LM and SKAT-S-LMM are both extremely low, likely because noise caused by random artifacts. Overall, SKAT-naive, which conducts no pre-selections, has the lowest power when heritability is greater than 0.05, but outperforms SKAT-S-LM and SKAT-S-LMM when heritability is less than 0.05. In particular, at very high heritability where h^2 = 0.1, SKAT-S-LM and its meta-analysis equivalent have very high power approaching that of SKAT-eQTL and kTWAS. This may be because SNPs have strong marginal associations when overall genetic effects are high, leading to reduced noise in the linear pre-selection process employed by SKAT-S-LM and SKAT-S-LMM. We do not have a clear interpretation on why SKAT-S-LM consistently outperforms SKAT-S-LMM.

The meta-analysis protocols utilizing summary statistics have similar trends compared to the genotype-based protocols, although the power of the meta-analysis protocols is slightly lower than the genotype-based protocols.

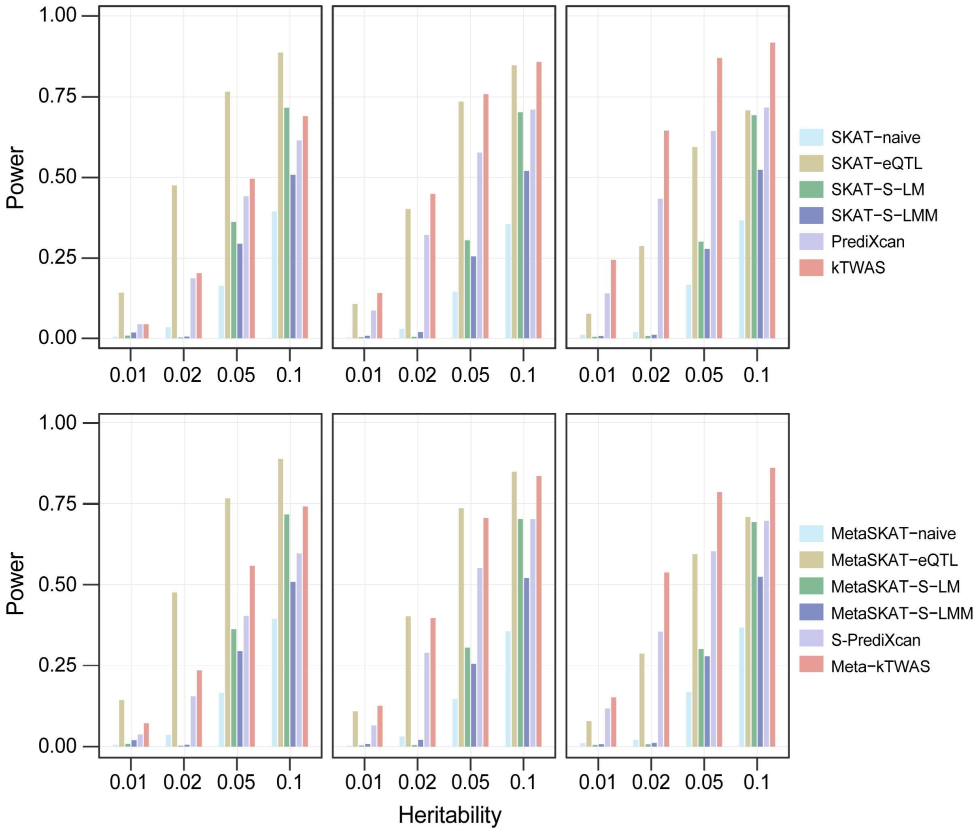

Figure 4. Statistical power (y-axis) of protocols on “Heterogeneity” model simulated genotypes (top row) and summary statistics (bottom row), at varying levels of trait heritability (x-axis) and different numbers of GTEx whole-blood ElasticNet predictors. The number of GTEx whole-blood Elastic Net predictors for both rows is 0, 1, 2, and the corresponding number of eQTL-selected SNPs is 2, 1, 0 (left to right).

Non-linear architectures. ***Figs. 4, 5, and 6*** *plot t*he power of genotype and summary statistic-based protocols under the Heterogeneity, Both Interactions, and Compensatory Interactions models. Although the genetic architectures are fundamentally different, there are several trends which are consistent across all architectures: 1) kTWAS always outperforms PrediXcan; 2) SKAT-eQTL outperforms kTWAS when both causal SNPs are non-ElasticNet selected SNPs; 3) SKAT-S-LM has high power only when heritability is high. Notably, kTWAS and SKAT-eQTL outperform PrediXcan with larger margins under the non-linear architectures than in the additive model. This is consistent with our expectation that kernel methods are more robust and flexible when genetic interactions are present, even when using a linear kernel.

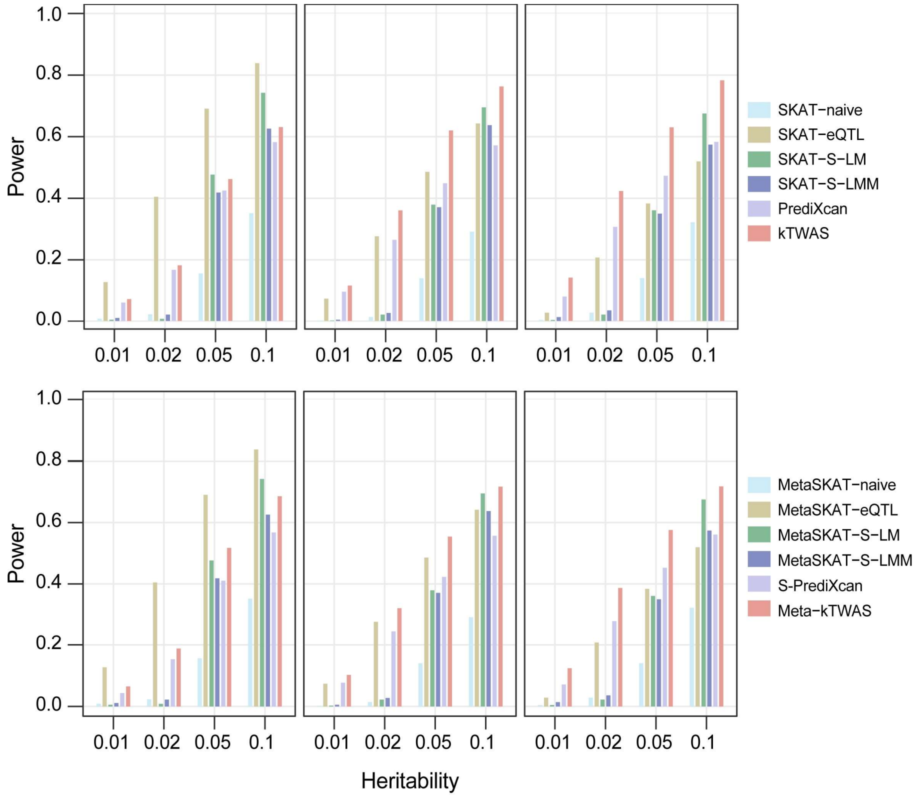

Figure 5. Statistical power (y-axis) of protocols on “Both inheritance” model simulated genotypes (top row) and summary statistics (bottom row), at varying levels of trait heritability (x-axis) and different numbers of GTEx whole-blood ElasticNet predictors. The number of GTEx whole-blood Elastic Net predictors for both rows is 0, 1, 2, and the corresponding number of eQTL-selected SNPs is 2, 1, 0 (left to right).

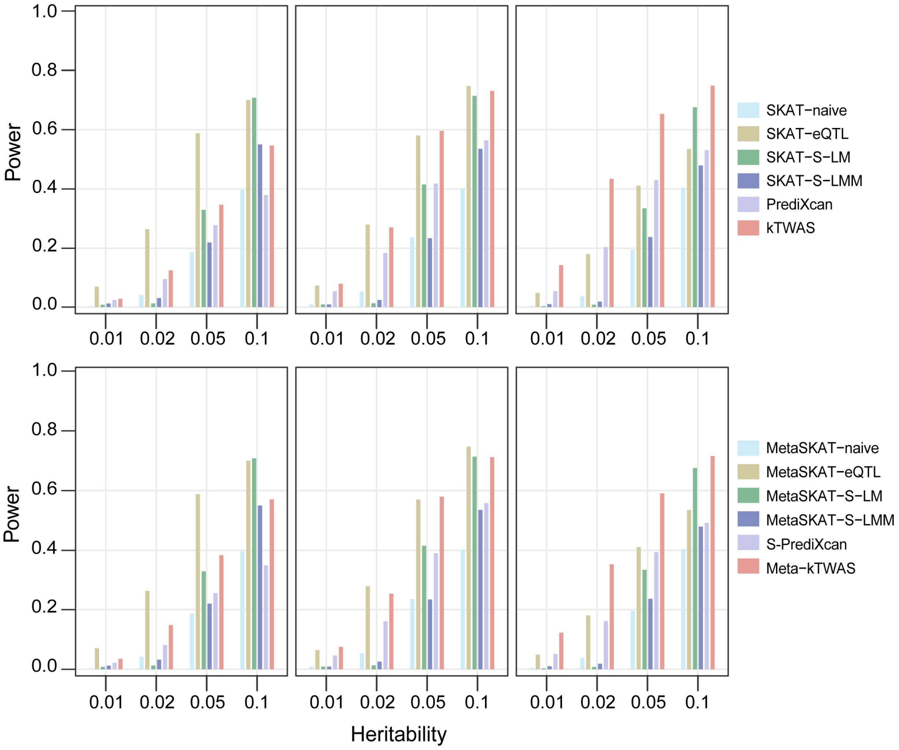

Figure 6. Statistical power (y-axis) of protocols on “Compensatory inheritance” model simulated genotypes (top row) and summary statistics (bottom row), at varying levels of trait heritability (x-axis) and different numbers of GTEx whole-blood ElasticNet predictors. The number of GTEx whole-blood Elastic Net predictors for both rows is 0, 1, 2, and the corresponding number of eQTL-selected SNPs is 2, 1, 0 (left to right).

Applying kTWAS to real data

ASD whole genome data provided by MSSNG. **Fig. 7a** shows the Manhattan plot for the output of kTWAS and PrediXcan. Based on a Bonferroni-corrected p-value < 0.05, we observed 6 peaks corresponding to RP11-575H3.1 (nominal P=1.73×10^{-6}), NDUFV1 (P=2.06×10^{-6}), PPP1R32 (P=2.59×10^{-6}), NBPF15 (P=3.11×10^{-6}), NBPF9 (P=5.82×10^{-6}), and SRGAP2B (P=6.44×10^{-6}). **Fig. 7b** shows the corresponding Manhattan plot for PrediXcan. Two genes (RP11-575H3.1 and NBPF15) identified by kTWAS were also discovered by PrediXcan, but at lower significance levels (nominal P-values of 2.74×10^{-6} and 7.17×10^{-6}, respectively). The remaining four genes are not identified as significant with PrediXcan (nominal P-values 0.66 for NDUFV1, 4.67×10^{-4} for PPP1R32, 1.31×10^{-3} for NBPF9, and 0.23 for SRGAP2B).

Three of the four genes detected by kTWAS alone have literature support as candidate genes for association with ASD. The inhibition of SRGAP2 function by its human-specific paralogs has contributed to the evolution of the human neocortex and plays an important role during human brain development[44, 45]. NBPF9 is a member of the neuroblastoma breakpoint family (NBPF) which consists of dozens of recently duplicated genes primarily located in segmental duplications on human chromosome 1. Members of this gene family are characterized by tandemly repeated copies of DUF1220 protein domains. Gene copy number variations in the human chromosomal region 1q21.1, where most DUF1220 domains are located, have been implicated in a number of developmental and neurogenetic diseases such as autism[46]. In particular, rare variants located in NBPF9 are reported to be associated with ASD[47]. Additionally, evidence shows that NDUFV1 is a 'developmental/neuropsychiatric' susceptibility gene when a rare duplication CNV occurs at 11p13.3[47]. The only gene not supported by literature is PPP1R32, which may be a novel gene for ASD research. The genes jointly identified by kTWAS and PrediXcan are both supported by literature[48, 49].

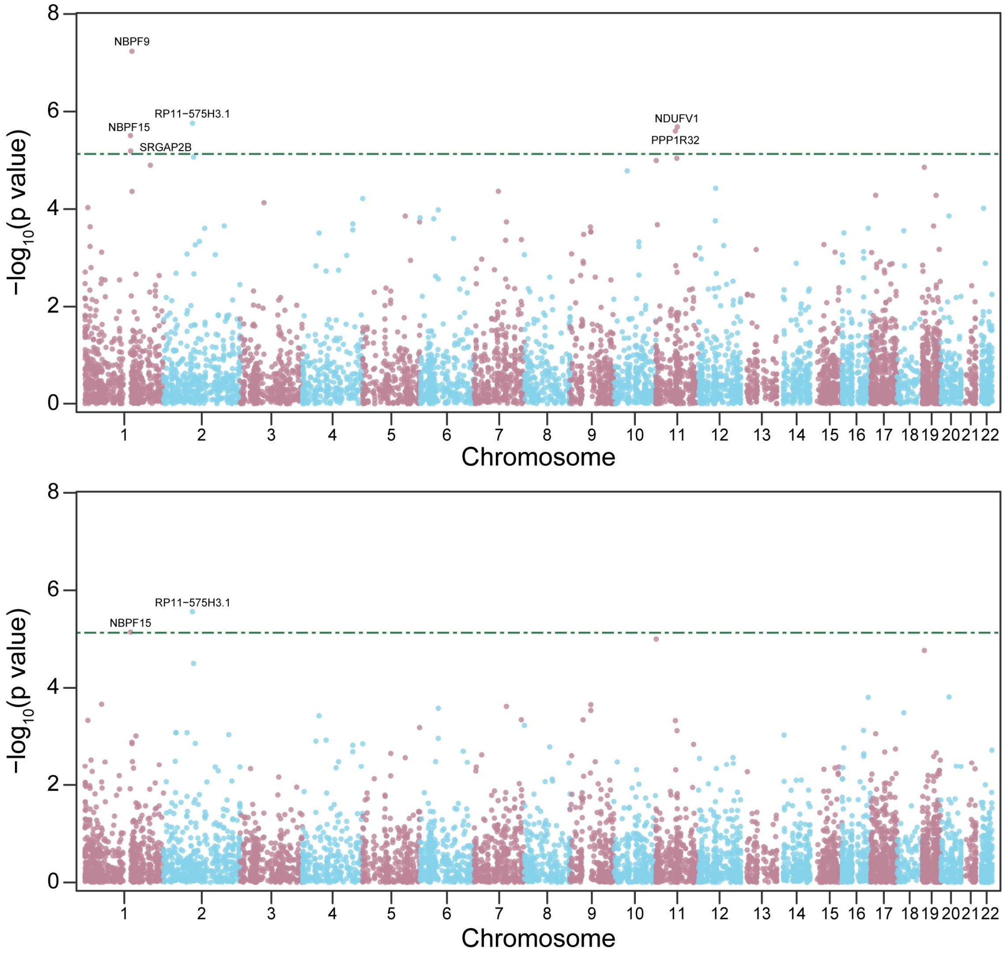

Figure 7. GWAS Manhattan plots of negative log P-values (y-axis) for autism-associated SNPs at genome coordinates (x-axis) of MSSNG consortium whole genome data. Plots show associations generated by kTWAS (**a**, top) and PrediXcan (**b**, bottom), with gene expressions predicted using the GTEx whole-blood ElasticNet model. The Bonferroni-corrected significance threshold (green dash line) is $0.05/6794=7.37\times10^{-6}$.

WTCCC genotyping data. We applied kTWAS to type 1 diabetes (T1D) data, identifying 52 genes significantly associated with the risk of T1D (Bonferroni-corrected p-value < 0.05). In contrast, PrediXcan identified 32 genes, of which 31 are also detected by kTWAS (**Table 3**). Among the 21 genes identified by kTWAS alone, 19 are within the MHC region, which has been shown to influence susceptibility to complex, autoimmune, and infectious diseases, including

T1D in particular[50]. Most of these genes have been reported to have association with T1D[51-63].

The PMIDs of these supporting literatures were annotated in the **Table 3**. The remaining two genes are BTN3A2 and ALDH2. BTN3A2 is reported to play important roles in regulating the immune response, and is a potential novel susceptibility gene for T1D[51] . ALDH2 is known to offer myocardial protection against stress conditions such as diabetes mellitus[64], although the underlying mechanism is unclear.

Gene	**Chr**	**Start**	**End**	**PrediXcan p-value**	**kTWAS p-value**	**PMID**
C1orf216	chr1	35713875	35719472	**3.22E-06**	**8.41E-07**	
HIST1H3E	chr6	26224199	26227473	**8.10E-08**	**8.40E-08**	
BTN3A2	chr6	26365159	26378320	0.014022	**1.35E-08**	19295542
HIST1H2B	chr6	27815044	27815424	**5.08E-08**	**2.73E-07**	
ZSCAN9	chr6	28224886	28233482	**6.58E-11**	**1.86E-06**	
ZFP57	chr6	29672392	29681110	2.06E-05	**3.44E-07**	27075368
PPP1R11	chr6	30066709	30070265	0.001064	**3.93E-14**	25422764
TRIM10	chr6	30151945	30160934	**1.83E-28**	**4.71E-28**	
TRIM15	chr6	30163206	30172696	**2.34E-10**	**2.45E-16**	
PPP1R18	chr6	30676389	30687895	**2.37E-07**	**2.44E-07**	
NRM	chr6	30688047	30691420	**1.82E-24**	**3.09E-24**	
FLOT1	chr6	30727709	30742733	**3.48E-17**	**4.07E-18**	
IER3	chr6	30743199	30744554	**1.20E-21**	**7.77E-20**	
LINC00243	chr6	30798654	30830659	0.000443	**3.32E-14**	
DDR1	chr6	30876421	30900156	7.35E-05	**1.84E-11**	
CCHCR1	chr6	31142439	31158238	**1.36E-06**	0.000535	
HLA-B	chr6	31269491	31356442	**1.82E-24**	**3.09E-24**	
MICB	chr6	31494881	31511124	**5.76E-27**	**3.38E-29**	
ATP6V1G2	chr6	31544462	31546848	**7.35E-64**	**6.92E-48**	
NFKBIL1	chr6	31546870	31558829	**8.26E-13**	**2.08E-14**	
NCR3	chr6	31588910	31592985	**4.18E-40**	**1.00E-22**	
AIF1	chr6	31615184	31617021	**9.87E-13**	**4.62E-11**	
LY6G5B	chr6	31670167	31673776	**3.54E-06**	**2.40E-13**	
LY6G5C	chr6	31676684	31684040	**1.29E-13**	**1.48E-13**	
ABHD16A	chr6	31686949	31703444	**3.40E-16**	**8.55E-23**	
DDAH2	chr6	31727038	31730580	**9.68E-58**	**3.23E-64**	
CLIC1	chr6	31730618	31739763	**1.47E-30**	**1.31E-34**	
VWA7	chr6	31765590	31777294	0.032341	**1.86E-07**	31932636
C6orf48	chr6	31834608	31839766	1.46E-05	**1.26E-11**	20221424
C2	chr6	31897785	31945649	0.027457	**8.99E-08**	1684365

SKIV2L	chr6	31959111	31969755	**4.28E-27**	**1.72E-24**	
DXO	chr6	31969810	31972292	0.009281	**3.17E-39**	
C4A	chr6	31982024	32002681	**3.00E-159**	**2.80E-131**	
C4B	chr6	32014762	32035418	**1.81E-69**	**4.23E-77**	
CYP21A2	chr6	32038265	32041670	0.44389	**1.17E-30**	
AGER	chr6	32180968	32184324	**1.99E-08**	**2.08E-08**	
NOTCH4	chr6	32194843	32224067	0.000112	**4.38E-07**	22414874
HLA-DRB5	chr6	32517343	32530287	**5.29E-81**	**1.03E-121**	
HLA-DRB1	chr6	32578769	32589848	4.31E-05	**2.64E-54**	
HLA-DQB1	chr6	32659467	32668383	**5.09E-58**	**1.70E-61**	
HLA-DQA2	chr6	32741342	32747215	0.004502	**2.64E-35**	19143816
HLA-DQB2	chr6	32756098	32763534	0.975179	**1.29E-15**	15256073
HLA-DOB	chr6	32812763	32817048	**1.19E-14**	**2.17E-20**	
TAP2	chr6	32821833	32838780	**4.83E-130**	**7.04E-228**	
PSMB8	chr6	32840717	32844047	0.025	**3.37E-06**	20221424
TAP1	chr6	32845209	32853978	0.454086	**1.04E-06**	8248212
HLA-DOA	chr6	33004178	33009612	0.007979	**8.12E-10**	19458622
RPS18	chr6	33272048	33276510	0.007684	**8.13E-10**	19609442
RPS26	chr12	56041853	56044675	**1.66E-11**	**1.44E-11**	
CNPY2	chr12	56309842	56316222	**2.25E-10**	**3.09E-10**	
ALDH2	chr12	111766887	111817529	0.001816	**1.63E-07**	27882330

Table 3. PrediXcan and kTWAS results for Bonferroni-corrected significant gene associations with type 1 diabetes in WTCCC consortium. To account for multiple testing, we used a significance threshold of 6.89×10^{-6} (0.05/7252) for all diseases. The significant genes are in bold. Chromosome and gene start positions are based on GENCODE version 26.

The other diseases in WTCCC have limited numbers of significant genes, except for rheumatoid arthritis (RA). kTWAS identified 24 genes associated with RA, while PrediXcan detected 19 significant genes, of which 18 are also detected by kTWAS (**Table 4**). All six genes detected by kTWAS alone (VARS2, NCR3, NOTCH4, TAP2, HLA-DQB2, LY6G5B) are in the MHC region and have substantial literature support. In particular, a nonsynonymous change in the VARS2 locus (rs4678) is strongly associated with RA[65]. One SNP in NCR3 could regulate the expression of two genes in RA cases and increased NCR3 expression is significantly associated with reduced RA susceptibility[66]. NOTCH4 is also reported to be RA-susceptible by different researchers[67, 68]. Yu et al. provided genetic evidence that TAP2 gene codon 565 polymorphism could play a role in RA[69]. A study on Italian patients found a mutation in HLA-DQA2 (rs9275595) could contribute to RA pathogenesis. Although there is no direct evidence to show LY6G5B is associated with RA, strong associations have been found between RA and a 126-kb region in the MHC class III region between BAT2 and CLIC1 which contains the five Ly-6 members including LY6G5B[70], indicating that LY6G5B might be a novel RA risk gene.

Disease	Gene	Chr	Start	End	PrediXcan p-value	kTWAS p-value
BD	CTD-2589	chr11	46238382	46239267	**5.43E-07**	0.196195
BD	SLC48A1	chr12	47753916	47782721	**7.24E-07**	**1.12E-06**
BD	RP11-382	chr15	83112738	83208018	**5.82E-06**	9.46E-06
BD	ERVK3-1	chr19	58305319	58315663	**1.85E-06**	**6.78E-08**
CAD	C12orf43	chr12	121000486	121016502	0.000590514	**5.23E-07**
CAD	RP11-347I19.8	chr12	121797511	121797872	**1.09E-06**	0.410928711
CD	APEH	chr3	49674002	49683946	**2.08E-06**	**2.13E-06**
RA	NT5DC2	chr3	52524496	52535054	**4.53E-08**	0.000142
RA	TRIM7	chr5	181193924	181205293	**6.50E-06**	**6.65E-06**
RA	TRIM26	chr6	30184455	30213427	**3.15E-11**	**3.46E-11**
RA	FLOT1	chr6	30727709	30742733	**1.34E-06**	**3.61E-07**
RA	IER3	chr6	30743199	30744554	**1.48E-09**	**1.91E-07**
RA	VARS2	chr6	30914205	30926459	0.004841	**5.19E-07**
RA	ATP6V1G2	chr6	31544462	31546848	**1.57E-06**	**3.85E-07**
RA	NCR3	chr6	31588910	31592985	0.000505	**1.28E-14**
RA	PRRC2A	chr6	31620720	31637771	**4.75E-18**	**6.26E-18**
RA	BAG6	chr6	31639028	31652705	**4.03E-12**	**2.24E-09**
RA	LY6G5B	chr6	31670167	31673776	0.723351	**6.37E-09**
RA	DDAH2	chr6	31727038	31730580	**4.47E-07**	**3.92E-07**
RA	MSH5	chr6	31739948	31762798	**1.78E-11**	**5.92E-18**
RA	C6orf48	chr6	31834608	31839766	**7.55E-22**	**1.44E-23**
RA	SKIV2L	chr6	31959111	31969755	**4.72E-21**	**2.10E-12**
RA	STK19	chr6	31971166	31981451	**1.03E-17**	**1.34E-17**
RA	CYP21A2	chr6	32038265	32041670	**2.89E-07**	**5.72E-08**
RA	NOTCH4	chr6	32194843	32224067	0.003209	**1.43E-10**
RA	HLA-DRB5	chr6	32517343	32530287	**8.82E-09**	**4.66E-17**
RA	HLA-DRB1	chr6	32578769	32589848	**3.29E-33**	**1.21E-14**
RA	HLA-DQA1	chr6	32628179	32643652	**2.03E-10**	**1.76E-10**
RA	HLA-DQA2	chr6	32741342	32747215	**4.11E-07**	**3.94E-15**
RA	HLA-DQB2	chr6	32756098	32763534	0.229988	**1.24E-10**
RA	TAP2	chr6	32821833	32838780	0.189783	**1.86E-07**
RA	C12orf43	chr12	121000486	121016502	**2.24E-06**	**1.82E-15**
T2D	C1orf216	chr1	35713875	35719472	**1.37E-07**	**2.20E-08**
T2D	CTD-2589M5.5	chr11	46238382	46239267	**4.68E-06**	0.102425
T2D	KCNMB4	chr12	70366276	70434292	**2.22E-06**	**2.27E-06**

Table 4. PrediXcan and kTWAS results for Bonferroni-corrected significant gene associations with five diseases in WTCCC consortium. To account for multiple testing, we used a significance threshold of

6.89×10^{-6} (0.05/7252) for all diseases. bipolar disease (BD), coronary artery disease (CAD), Crohn's disease (CD), rheumatoid arthritis (RA), type 2 diabetes (T2D). The significant genes are in bold. Chromosome and gene start positions are based on GENCODE version 26.

Taken together, the analysis of MSSNG sequence data and T1D and RA genotyping array data from WTCCC illustrates that kTWAS is able to identify a larger number of significant and sensible genes versus PrediXcan. These results confirm that the inclusion of kernel methods in TWAS increases statistical power in real data, compared to standard TWAS which only uses linear combinations of selected SNPs and is therefore not robust to non-linear effects.

Conclusion

In this work, we have thoroughly highlighted the essential advantages and differences between TWAS and kernel methods in terms of their ability to select and organize genetic features. From this perspective we designed a novel protocol kTWAS to integrate advantages of both methods, leading to a tool that takes advantage of expression data and is robust to non-linear effects. We demonstrate that kTWAS results in increased power by conducting extensive simulations and real data analyses. This work will help researchers understand the conditions under which TWAS and kernel methods perform best, and how to integrate both to capture non-linear effects. This work also reveals that the application of a linear kernel is more powerful than simple linear regression for detecting nonlinear effects.

Other researchers have also investigated the link between SKAT and TWAS. Xu et al. have designed a power testing framework with both TWAS and SKAT as special cases of their power test[29]. Although , their framework does not directly compare the power of the two protocols, or suggest a means for integrating them.

As shown in **Figs. 2-6**, it is evident that SKAT-eQTL also has high power, despite not taking advantage of the TWAS-style feature pre-selection which is achieved by ElasticNet or the multiple-regression based methods found in SKAT-S-LM and SKAT-S-LMM. In essence, SKAT-eQTL only selects for genetic variants with good marginal effects without considering linear combinations of such variants. The effectiveness of SKAT-eQTL invites more thorough testing via theoretical and real data analyses, and this will be an immediate area of study in our future work.

Reference:

1. Hormozdiari F, Gazal S, van de Geijn B et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits, Nat Genet 2018;50:1041-1047.
2. Zeng B, Lloyd-Jones LR, Montgomery GW et al. Comprehensive Multiple eQTL Detection and Its Application to GWAS Interpretation, Genetics 2019;212:905-918.

3. Gusev A, Mancuso N, Won H et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights, Nat Genet 2018;50:538-548.
4. Mancuso N, Gayther S, Gusev A et al. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions, Nat Commun 2018;9:4079.
5. Huckins LM, Dobbyn A, Ruderfer DM et al. Gene expression imputation across multiple brain regions provides insights into schizophrenia risk, Nat Genet 2019;51:659-674.
6. Gamazon ER, Wheeler HE, Shah KP et al. A gene-based association method for mapping traits using reference transcriptome data, Nat Genet 2015;47:1091-1098.
7. Zeng P, Zhou X, Huang S. Prediction of gene expression with cis-SNPs using mixed models and regularization methods, BMC Genomics 2017;18:368.
8. Gusev A, Ko A, Shi H et al. Integrative approaches for large-scale transcriptome-wide association studies, Nat Genet 2016;48:245-252.
9. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models, PLoS Genet 2013;9:e1003264.
10. Xie R, Wen J, Quitadamo A et al. A deep auto-encoder model for gene expression prediction, BMC Genomics 2017;18:845.
11. Xie R, Quitadamo A, Cheng J et al. A predictive model of gene expression using a deep learning framework. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2016, p. 676-681. IEEE.
12. Barbeira AN, Dickinson SP, Bonazzola R et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics, Nat Commun 2018;9:1825.
13. Theriault S, Gaudreault N, Lamontagne M et al. A transcriptome-wide association study identifies PALMD as a susceptibility gene for calcific aortic valve stenosis, Nat Commun 2018;9:988.
14. Gong L, Zhang D, Lei Y et al. Transcriptome-wide association study identifies multiple genes and pathways associated with pancreatic cancer, Cancer Med 2018;7:5727-5732.
15. Ratnapriya R, Sosina OA, Starostik MR et al. Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration, Nat Genet 2019;51:606-610.
16. Atkins I, Kinnersley B, Ostrom QT et al. Transcriptome-Wide Association Study Identifies New Candidate Susceptibility Genes for Glioma, Cancer Res 2019;79:2065-2071.
17. Zhang W, Voloudakis G, Rajagopal VM et al. Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits, Nat Commun 2019;10:3834.
18. Ding B. Conditions under which transcriptome-wide association studies will be more powerful. University of Calgary, 2020.
19. Wu MC, Lee S, Cai T et al. Rare-variant association testing for sequencing data with the sequence kernel association test, Am J Hum Genet 2011;89:82-93.
20. Wu MC, Kraft P, Epstein MP et al. Powerful SNP-set analysis for case-control genome-wide association studies, Am J Hum Genet 2010;86:929-942.
21. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, Nature 2007;447:661-678.
22. Hu Y, Li M, Lu Q et al. A statistical framework for cross-tissue transcriptome-wide association analysis, Nat Genet 2019;51:568-576.
23. Wainberg M, Sinnott-Armstrong N, Mancuso N et al. Opportunities and challenges for transcriptome-wide association studies, Nat Genet 2019;51:592-599.
24. Brandes N, Linial N, Linial M. PWAS: Proteome-Wide Association Study. Cham, 2020, p. 237-239. Springer International Publishing.

25. Okada H, Ebhardt HA, Vonesch SC et al. Proteome-wide association studies identify biochemical modules associated with a wing-size phenotype in Drosophila melanogaster, Nat Commun 2016;7:12649.
26. Xu Z, Wu C, Pan W et al. Imaging-wide association study: Integrating imaging endophenotypes in GWAS, Neuroimage 2017;159:159-169.
27. Lee S, Teslovich TM, Boehnke M et al. General framework for meta-analysis of rare variants in sequencing association studies, Am J Hum Genet 2013;93:42-53.
28. Ionita-Laza I, Lee S, Makarov V et al. Sequence kernel association tests for the combined effect of rare and common variants, Am J Hum Genet 2013;92:841-853.
29. Xu Z, Wu C, Wei P et al. A Powerful Framework for Integrating eQTL and GWAS Summary Data, Genetics 2017;207:893-902.
30. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. Springer series in statistics New York, 2001.
31. PredictDB Data Repository, URL http://predictdb.org/ 2019.
32. Mancuso N, Shi H, Goddard P et al. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits, Am J Hum Genet 2017;100:473-487.
33. Zhu Z, Zhang F, Hu H et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets, Nat Genet 2016;48:481-487.
34. Purcell S, Neale B, Todd-Brown K et al. PLINK: a tool set for whole-genome association and population-based linkage analyses, Am J Hum Genet 2007;81:559-575.
35. Kang HM, Sul JH, Service SK et al. Variance component model to account for sample structure in genome-wide association studies, Nat Genet 2010;42:348-354.
36. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans, Science 2015;348:648-660.
37. Genomes Project C, Auton A, Brooks LD et al. A global reference for human genetic variation, Nature 2015;526:68-74.
38. Long Q, Rabanal FA, Meng D et al. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden, Nat Genet 2013;45:884-890.
39. Brown KM, Costanzo MS, Xu W et al. Compensatory mutations restore fitness during the evolution of dihydrofolate reductase, Mol Biol Evol 2010;27:2682-2690.
40. Kulathinal RJ, Bettencourt BR, Hartl DL. Compensated deleterious mutations in insect genomes, Science 2004;306:1553-1554.
41. Tomala K, Zrebiec P, Hartl DL. Limits to Compensatory Mutations: Insights from Temperature-Sensitive Alleles, Mol Biol Evol 2019;36:1874-1883.
42. Weisstein EW. Bonferroni correction, https://mathworld. wolfram. com/ 2004.
43. RK CY, Merico D, Bookman M et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder, Nat Neurosci 2017;20:602-611.
44. Alqallaf AK, Alkoot FM, Mash'el S A. Discovering the Genetics of Autism. Recent Advances in Autism Spectrum Disorders-Volume I. IntechOpen, 2013.
45. Dennis MY, Nuttle X, Sudmant PH et al. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication, Cell 2012;149:912-922.
46. O'Bleness MS, Dickens CM, Dumas LJ et al. Evolutionary history and genome organization of DUF1220 protein domains, G3 (Bethesda) 2012;2:977-986.
47. Woodbury-Smith M, Paterson AD, Thiruvahindrapduram B et al. Using extended pedigrees to identify novel autism spectrum disorder (ASD) candidate genes, Human genetics 2015;134:191-201.
48. Parikshak NN, Swarup V, Belgard TG et al. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism, Nature 2016;540:423-427.
49. Wu H, Zhai LT, Guo XX et al. The N-terminal of NBPF15 causes multiple types of aggregates and mediates phase transition, Biochem J 2020;477:445-458.

50. Matzaraki V, Kumar V, Wijmenga C et al. The MHC locus and genetic susceptibility to autoimmune and infectious diseases, Genome Biol 2017;18:76.
51. Viken MK, Blomhoff A, Olsson M et al. Reproducible association with type 1 diabetes in the extended class I region of the major histocompatibility complex, Genes Immun 2009;10:323-333.
52. Bak M, Boonen SE, Dahl C et al. Genome-wide DNA methylation analysis of transient neonatal diabetes type 1 patients with mutations in ZFP57, BMC Med Genet 2016;17:29.
53. Qiu YH, Deng FY, Li MJ et al. Identification of novel risk genes associated with type 1 diabetes mellitus using a genome-wide gene-based association analysis, J Diabetes Investig 2014;5:649-656.
54. Hebbar P, Abu-Farha M, Alkayal F et al. Genome-wide association study identifies novel risk variants from RPS6KA1, CADPS, VARS, and DHX58 for fasting plasma glucose in Arab population, Sci Rep 2020;10:152.
55. Brorsson C, Tue Hansen N, Bergholdt R et al. The type 1 diabetes - HLA susceptibility interactome--identification of HLA genotype-specific disease genes for type 1 diabetes, PLoS One 2010;5:e9576.
56. Simon S, Awdeh Z, Campbell RD et al. A restriction fragment of the C2 gene is a unique marker for C2 deficiency and the uncommon C2 allele C2* B (a marker for type 1 diabetes), The Journal of clinical investigation 1991;88:2142-2145.
57. Bonegio R, Susztak K. Notch signaling in diabetic nephropathy, Exp Cell Res 2012;318:986-992.
58. Brorsson C, Hansen NT, Lage K et al. Identification of T1D susceptibility genes within the MHC region by combining protein interaction networks and SNP genotyping data, Diabetes, Obesity and Metabolism 2009;11:60-66.
59. Guja C, Guja L, Nutland S et al. Type 1 diabetes genetic susceptibility encoded by HLA DQB1 genes in Romania, J Cell Mol Med 2004;8:249-256.
60. Jackson DG, Capra JD. TAP1 alleles in insulin-dependent diabetes mellitus: a newly defined centromeric boundary of disease susceptibility, Proc Natl Acad Sci U S A 1993;90:11079-11083.
61. Santin I, Castellanos-Rubio A, Aransay AM et al. Exploring the diabetogenicity of the HLA-B18-DR3 CEH: independent association with T1D genetic risk close to HLA-DOA, Genes Immun 2009;10:596-600.
62. Bergholdt R, Brorsson C, Lage K et al. Expression profiling of human genetic and protein interaction networks in type 1 diabetes, PLoS One 2009;4:e6250.
63. Pan G, Deshpande M, Thandavarayan RA et al. ALDH2 Inhibition Potentiates High Glucose Stress-Induced Injury in Cultured Cardiomyocytes, J Diabetes Res 2016;2016:1390861.
64. Guo Y, Yu W, Sun D et al. A novel protective mechanism for mitochondrial aldehyde dehydrogenase (ALDH2) in type i diabetes-induced cardiac dysfunction: role of AMPK-regulated autophagy, Biochim Biophys Acta 2015;1852:319-331.
65. Vignal C, Bansal AT, Balding DJ et al. Genetic association of the major histocompatibility complex with rheumatoid arthritis implicates two non-DRB1 loci, Arthritis Rheum 2009;60:53-62.
66. Liu G, Hu Y, Jin S et al. Cis-eQTLs regulate reduced LST1 gene and NCR3 gene expression and contribute to increased autoimmune disease risk, Proc Natl Acad Sci U S A 2016;113:E6321-E6322.
67. AlFadhli S, Nanda A. Genetic evidence for the involvement of NOTCH4 in rheumatoid arthritis and alopecia areata, Immunol Lett 2013;150:130-133.
68. Mitsunaga S, Hosomichi K, Okudaira Y et al. Exome sequencing identifies novel rheumatoid arthritis-susceptible variants in the BTNL2, J Hum Genet 2013;58:210-215.
69. Yu MC, Huang CM, Wu MC et al. Association of TAP2 gene polymorphisms in Chinese patients with rheumatoid arthritis, Clin Rheumatol 2004;23:35-39.

70. Mallya M, Campbell RD, Aguado B. Characterization of the five novel Ly-6 superfamily members encoded in the MHC, and detection of cells expressing their potential ligands, Protein Sci 2006;15:2244-2256.