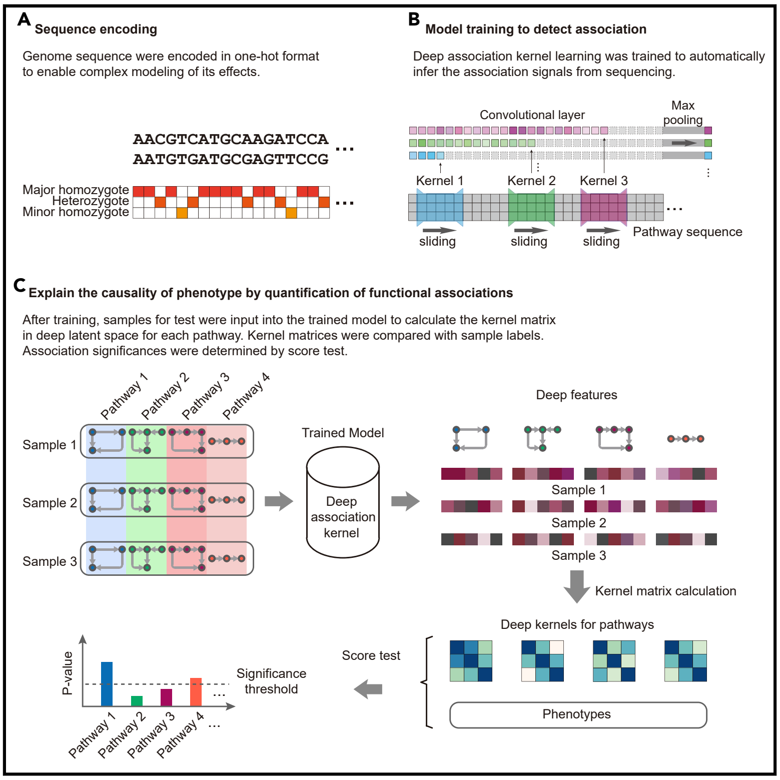
Patterns

Article

Explaining the Genetic Causality for Complex Phenotype via Deep Association Kernel Learning

Graphical Abstract



Genome sequence were encoded in one-hot format to enable complex modeling of its effects.

Deep association kernel learning was trained to automatically infer the association signals from sequencing.

After training, samples for test were input into the trained model to calculate the kernel matrix in deep latent space for each pathway. Kernel matrices were compared with sample labels. Association significances were determined by score test.

Highlights

- Utilize deep learning to infer complicated causal signals from genome
- Validate the model on different types of causal variants
- Explain the rationale of the model by interpretable analysis of the framework
- Apply the model to four real datasets with various diseases

Authors

Feng Bao, Yue Deng, Mulong Du, ..., David C. Christiani, Meilin Wang, Qionghai Dai

Correspondence

ydeng@buaa.edu.cn (Y.D.),
mwang@njmu.edu.cn (M.W.),
qhdai@tsinghua.edu.cn (Q.D.)

In Brief

Genetic mutations are key factors for complex diseases. Comprehensively understanding the genetic contribution will improve the mechanism study and treatment of diseases. However, genetic causalities are complex and mutation specific. To extensively dissect the unknown genetic causality, we propose deep association kernel learning (DAK) that utilizes the power of deep learning to automatically infer complex, non-linear, various causal loci from gene sequence at pathway level. On four real datasets covering cancers and mental disease, we demonstrate that DAK can discover unseen yet meaningful suspicious pathways.

Bao et al., 2020, Patterns *1*, 100057
August 14, 2020 © 2020 The Authors.
https://doi.org/10.1016/j.patter.2020.100057

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Explaining the Genetic Causality for Complex Phenotype via Deep Association Kernel Learning

Feng Bao,[1,2,10] Yue Deng,[3,4,10,*] Mulong Du,[5,6] Zhiquan Ren,[1] Sen Wan,[1] Kenny Ye Liang,[1] Shaohua Liu,[3] Bo Wang,[3] Junyi Xin,[7,8] Feng Chen,[6] David C. Christiani,[5,9] Meilin Wang,[7,8,*] and Qionghai Dai[1,2,11,*]

[1]Department of Automation, Tsinghua University, Beijing 100084, China
[2]Institute for Brain and Cognitive Sciences, Tsinghua University, Beijing 100084, China
[3]School of Astronautics, Beihang University, Beijing 100191, China
[4]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China
[5]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
[6]Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China
[7]Department of Environmental Genomics, Jiangsu Key Laboratory of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing 211166, China
[8]Department of Genetic Toxicology, The Key Laboratory of Modern Toxicology of Ministry of Education, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China
[9]Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA
[10]These authors contributed equally
[11]Lead Contact
*Correspondence: ydeng@buaa.edu.cn (Y.D.), mwang@njmu.edu.cn (M.W.), qhdai@tsinghua.edu.cn (Q.D.)
https://doi.org/10.1016/j.patter.2020.100057

THE BIGGER PICTURE Genetic mutations cause complex diseases in many different ways. Comprehensively identifying the genetic causality can lead to valuable insights into the development and treatment of diseases. However, existing genome-wide association study (GWAS) approaches are always built under linear assumption and simple disease models, restricting their generalization in discovering the complicated causality. DAK (deep association kernel learning) is a GWAS method that is constructed in a deep-learning framework and can simultaneously identify multiple types of genetic causalities without any modifications to the model. For biological contributions, the proposed approach enables the understanding of non-linear, complex genetic causalities and improves functional studies of the disease; for computational contributions, our method unifies kernel learning and association analysis in a joint explainable deep-learning framework.

1 2 3 4 5 **Proof-of-Concept:** Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

The genetic effect explains the causality from genetic mutations to the development of complex diseases. Existing genome-wide association study (GWAS) approaches are always built under a linear assumption, restricting their generalization in dissecting complicated causality such as the recessive genetic effect. Therefore, a sophisticated and general GWAS model that can work with different types of genetic effects is highly desired. Here, we introduce a deep association kernel learning (DAK) model to enable automatic causal genotype encoding for GWAS at pathway level. DAK can detect both common and rare variants with complicated genetic effects where existing approaches fail. When applied to four real-world GWAS datasets including cancers and schizophrenia, our DAK discovered potential casual pathways, including the association between dilated cardiomyopathy pathway and schizophrenia.

INTRODUCTION

The genome-wide association study (GWAS) is extensively used for uncovering potential causal loci from complex biological phenotypes.[1–3] The classical GWAS models assume that single locus contributes to the disease independently and the risk increases linearly with the number of minor alleles. These linear models are only powerful in discovering variants with strong

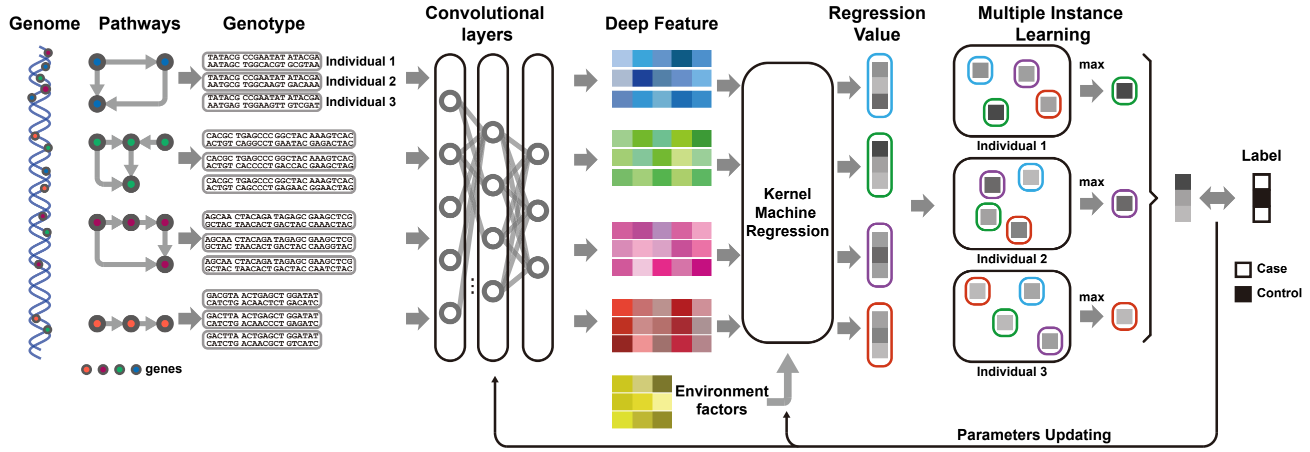


Figure 1. The Framework of DAK

SNPs are grouped into pathway-level gene set and coded into one-hot format. Convolutional layers are employed to encode causal loci into deep features. Kernel machine regression is incorporated to enable statistical tests of association via SKAT framework. Multiple-instance learning selects the most suspicious pathway at individual level. Parameters of the whole framework are optimized in an end-to-end manner through back-propagation. For ease of illustration, three individuals and four pathways are shown in the figure ($N = 3$, $P = 4$). Genotype of each SNP was further encoded into one-hot format before feeding into DAK model (Experimental Procedures).

and direct associations.[4] As an improvement, pathway-based methods were proposed by taking groups of biologically meaningful genes into consideration.[5–7] For instance, gene-set enrichment methods derive pathway-level statistical scores by combing p values from single-locus tests,[8–10] SKAT (sequence kernel association test),[11] and its variants[12,13] perform association tests using kernel regression. However, these existing approaches rely on some pre-assumed genetic models to conduct handcrafted genotype encoding. Unfortunately, in practice, the genetic effect of complex disease is unknown and can hardly be appropriately modeled in advance. Therefore, a genetic-model-free GWAS approach that can reasonably model the inherent relation between genotype and phenotype is urgently needed.

We introduce a deep-learning framework, deep association kernel learning (DAK), to conduct pathway-level GWAS (Figure 1). While the successes of deep learning for genomic studies has been witnessed in variant calling,[14] mutation effects prediction,[15] and binding motif identifications,[16] it has not been established for solving general GWAS problems. Our DAK framework incorporates convolutional layers to encode raw SNPs as latent genetic representation. Kernel regression layers are then connected with these encoded genetic representations to predict the disease status. More importantly, this kernel regression layer allows one to perform statistical significance tests on the learned genetic representations to uncover the disease-associated pathways. Both the convolutional and kernel regression layers are trained jointly using multiple-instance loss in an end-to-end manner. Therefore, DAK relies on no pre-assumed genetic model and can learn all model parameters in a pure data-driven manner.

We compared DAK with seven representative gene/pathway-based methods: classical statistic method (Burden test),[17] enrichment methods (GATES, HYST, and aSPU)[9,18,19] and kernel methods (SKAT and SKAT-o).[11,12] DAK is the only approach that consistently performs well under a wide range of genetic models including additive, multiplicative, dominant, recessive, and heterozygous effects. We further applied our method to four disease datasets, namely gastric cancer (GC), colorectal cancer (CRC), lung cancer (LC), and psychiatric disorder.

RESULTS

Deep Association Kernel Learning

We introduced DAK to achieve the detection of complex associations and enhance the interpretability of GWAS (Figure 1 and Experimental Procedures). Here, alleles are coded in the one-hot representations to enable flexible modeling of genotype effects for each locus. Variants in the same biological pathway are grouped together and the combinational effects of multiple SNPs within a pathway are considered at the same time. Next, pathway-level features are extracted by convolutional layers (Figure S1), followed by a kernel regression layer to derive the statistical significance (Figure S2). To allow learning from labels at the individual level, the whole framework is trained with a multiple-instance loss in an end-to-end manner. Finally, the variance tests used in SKAT are performed on the learned kernel matrix to derive statistical p values (Figures S3 and S4).

Type I Errors on Non-causal Pathways on Simulated Datasets

In each simulation experiment, we simulated datasets under null (no causal pathway) or alternative (disease was caused by different genetic associations) hypothesis (Figure 2A and Experimental Procedures). All seven methods were tested on simulated datasets. Performances of different approaches were evaluated using type I error rates (corresponding to null hypothesis) and empirical powers (corresponding to alternative hypothesis) (Experimental Procedures) in 100 replicates.

We first report the type I error. If no causal loci existed in all pathways (null hypothesis), all methods showed a low error-rate level (Figure S5). Changing the sample size had little

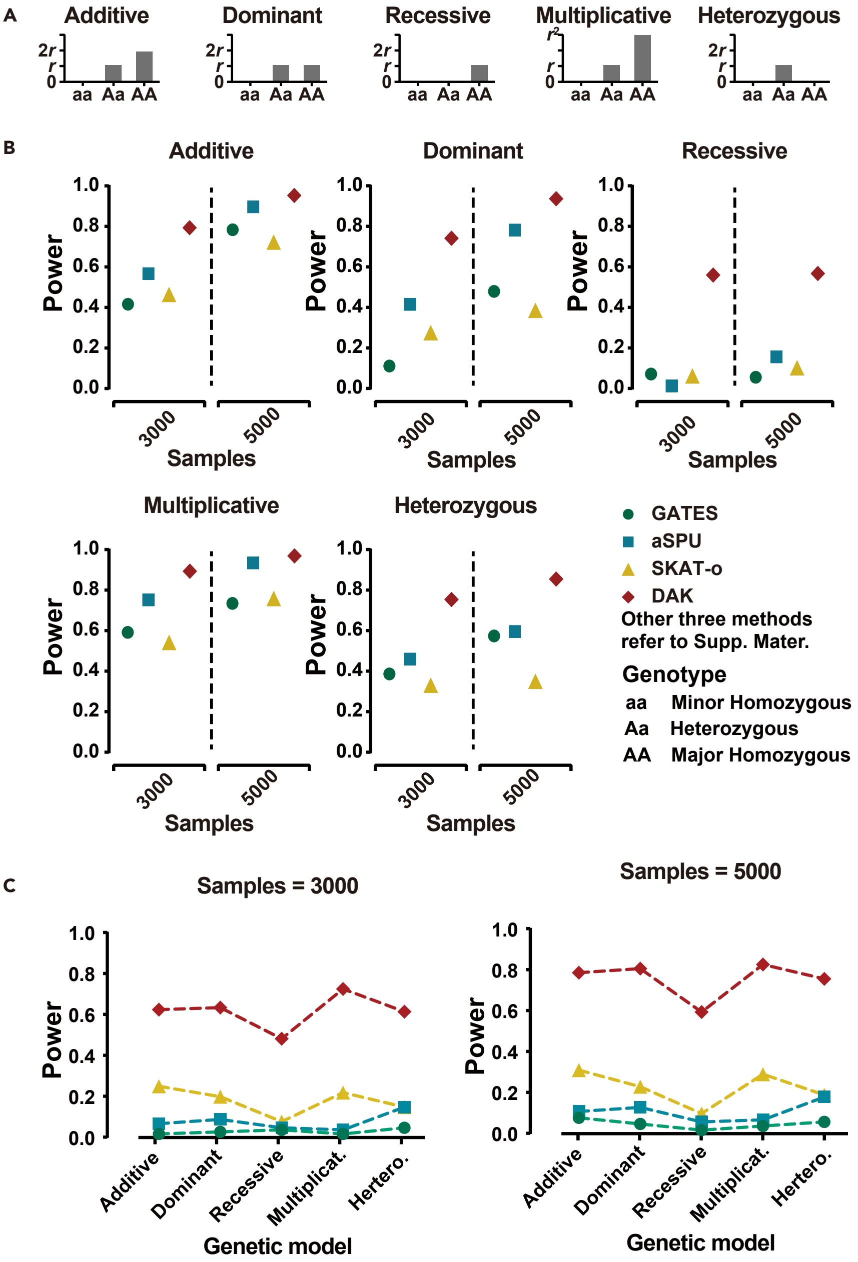


Figure 2. Performance Evaluations on Associations with Single Variant

(A) Disease risk levels for different genotypes in five genetic models.

(B) Performances to discover the disease pathway resulting from single common variant. Effect size was set to 0.2 and simulated phenotypes were generated under five effect models. Under each sample size (3,000, 5,000), seven methods (four showed here and three in Figure S13) were used to discover the disease pathway. Power was calculated from 100 replicates after Bonferroni correction.

(C) Performances to discover the disease pathway resulting from single rare variant. Effect size was set to 0.8 to simulate phenotypes; 3,000 and 5,000 samples were considered.

effect on the results. The training curve showed that DAK converged within several iterations (Figure S6).

Powers on Pathways with Single Effects on Simulated Datasets

We then considered that the disease was caused by a single common variant. To illustrate different functional pathways of genes to the disease, we assumed that the allele of the causal locus contributed to the disease in five different genetic models: (1) additive model, minor homozygous genotype had 2-fold effect over the heterozygous type; (2) dominant mode, two genotypes showed the same effect size; (3) multiplicative model, minor alleles increased the disease risk exponentially; (4) recessive model, only minor homozygous genotypes had effects; and (5) heterozygous model, only heterozygous alleles had effects (Figure 2A).

On the most widely used additive disease mode, we found that all methods showed reasonable accuracy in identifying the pathway with disease locus (Figures 2B and S7). However, when the fundamental genetic model changed, the power of all comparison methods dropped dramatically while DAK maintained a reliable performance with best power across all conditions. Specifically, for the challenging recessive genetic model, accuracies of all comparison methods greatly decreased and were far below the performances of DAK. The performance of DAK was further improved when increasing the effect size while other methods were still of low accuracy (Figure S8). We further noted that when the sample size was increased to 5,000, powers of all methods were increased and DAK maintained the best performance (Figures 2B and S7). With further increase in sample size (to 100,000), DAK is capable of detecting associations as weak as 0.01 (Figure S9). We also calibrated the performance of DAK on imbalanced datasets (Figure S10) and in datasets with known strong/weak linkage disequilibrium (LD) structures and LD scores (Figures S11 and S12).

The discovery of rare variants (minor allele frequency <1%) is a challenging task in GWAS due to the low gene frequency. We simulated a rare dataset of 5,000 samples where the disease was caused by single rare variant under five genotype models. Again, DAK obtained much higher performances than others on recessive and multiplicative genetic models (Figures 2C [bottom] and S13). We demonstrated that DAK could discover the causal rare variant at power around 0.8 on datasets even with only 3,000 samples (Figure 2C, top), which was a challenging task for other methods.

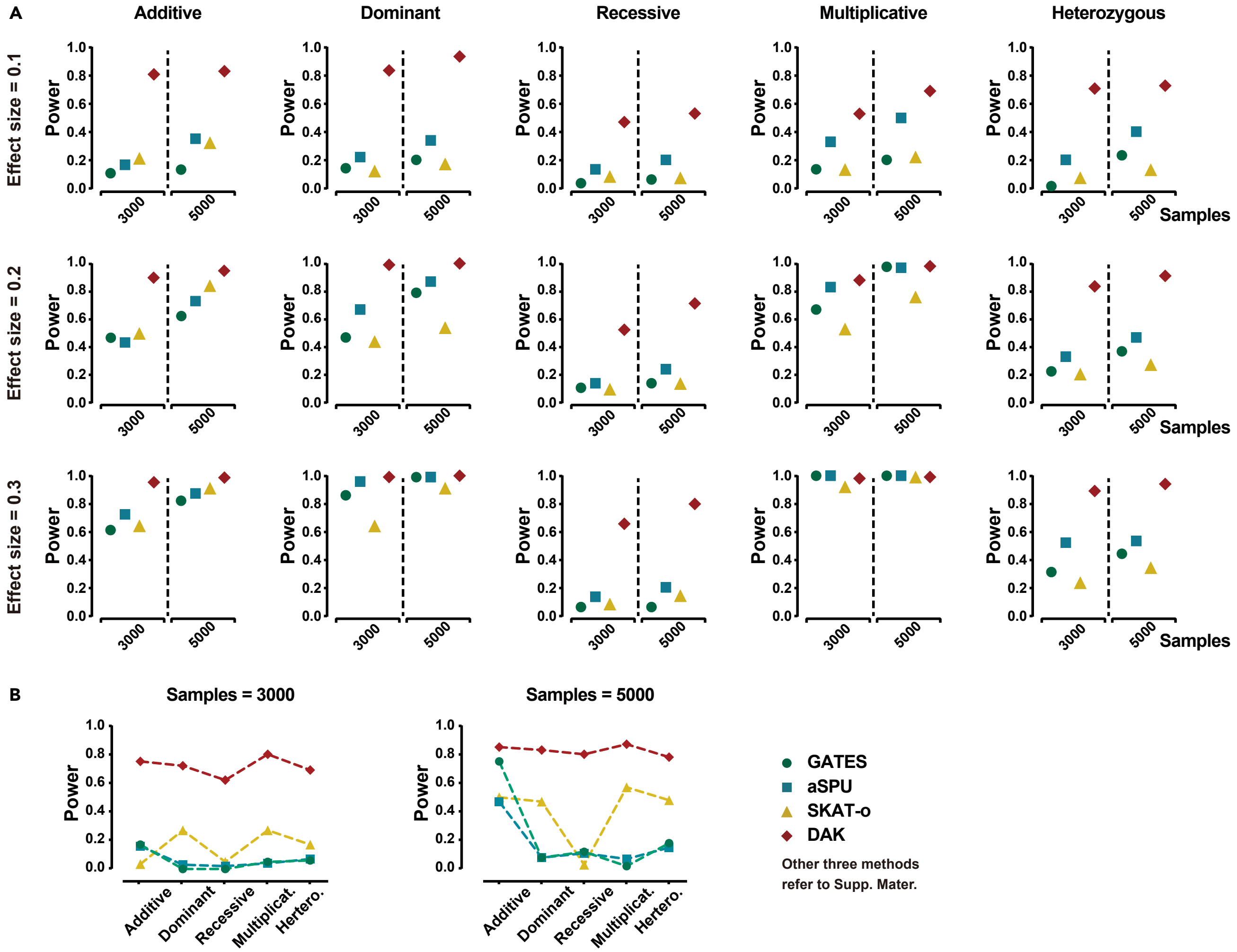


Figure 3. Performance Evaluations on Associations with Multiple Variants

(A) Performances to discover the disease pathway resulting from three common variants. Effect size was set to 0.1, 0.2, and 0.3 and simulated phenotypes were generated under five effect models. Under each sample size (3,000, 5,000), seven methods (four illustrated here) were used to discover the disease pathway. The power was calculated from 100 repeats after Bonferroni correction.

(B) Performances to discover the disease pathway resulting from three rare variants. Effect size was set to 0.8; 3,000 and 5,000 samples were considered.

We further analyzed the performance of DAK on causal variants with different minor allele frequency (MAF) ranges. DAK maintained high-power performances even with a small effect size (0.2) when MAF was >0.005 (Figure S14A). In simulations focusing on human leukocyte antigen regions, DAK also maintained similar accuracy with both common and rare variants (Figure S15). Lengths of pathways also showed little effect on the power of DAK (Figure S16). We also considered experiments with complex phenotype by hundreds of SNPs with small effect sizes (0.005). DAK showed greatly advantageous results compared with competitors (Figure S17).

Powers on Pathways with Joint Effects on Simulated Datasets

Most diseases are the result of the joint effect of multiple genes. However, it can be more challenging to identify the combined and mixed effect signals from multiple causal variants. Here, we simulated joint effects by randomly assigning three causal common variants and generated phenotype under five genetic models (Experimental Procedures). Performances of all methods were much lower compared with results under the single variant. However, DAK still dramatically outperformed other methods and achieved the most stable performance among all experiments (Figures 3A and S18). The performances of all methods were enhanced when the effect size was increased. The advantages of DAK were more obvious when the causal positions were rare variants (Figures 3B and S19).

To analyze the effect from LD structures, we further quantified the power of DAK on two simulated datasets with known strong or weak LD patterns. DAK also showed promising performances in discovering associations by multiple variants with small effect size (Figure S11). Further analysis of DAK on multiple causal variants with various MAF ranges was also performed (Figure S14B).

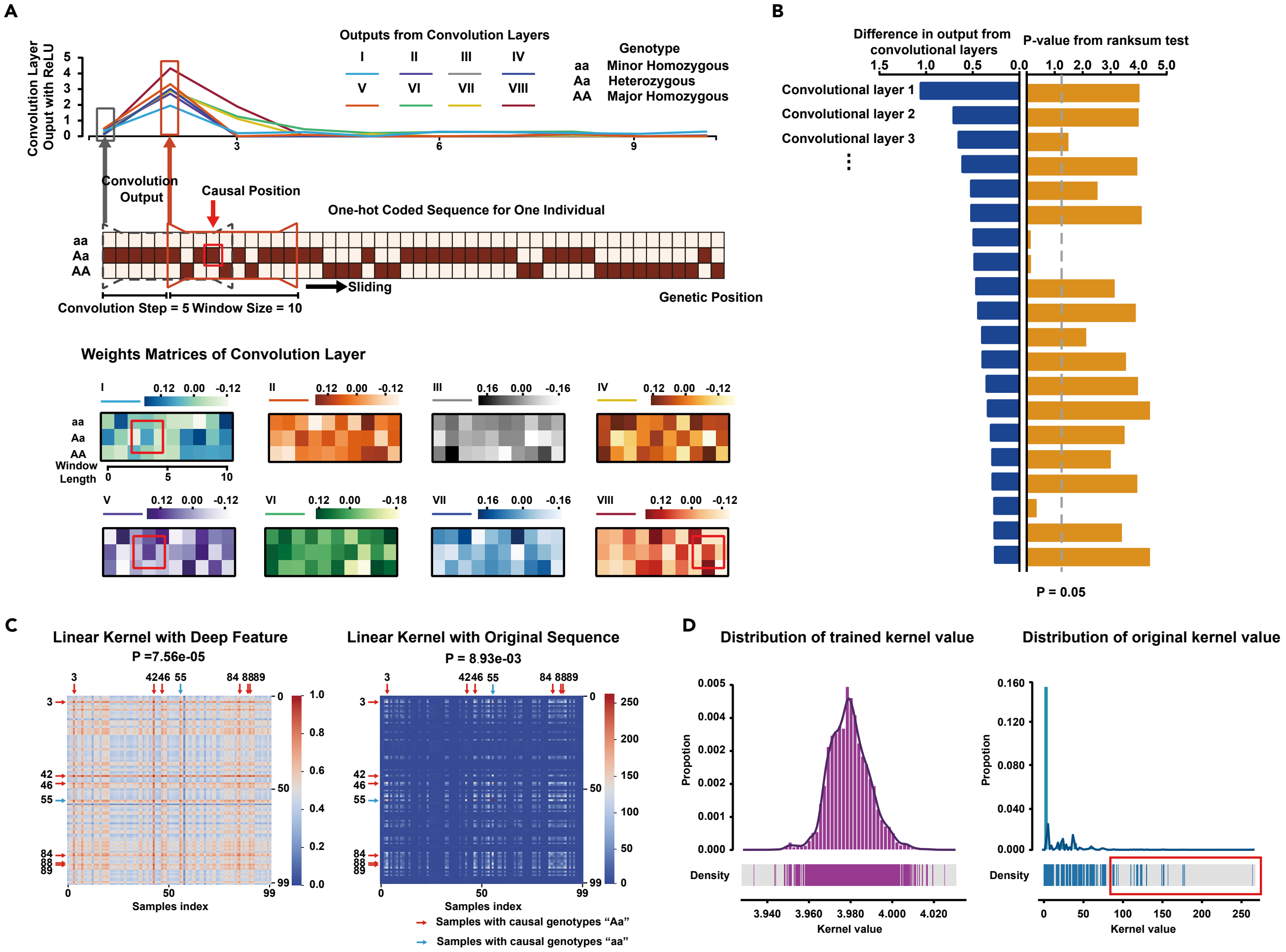


Figure 4. Explanable Analysis of DAK on Identifying Association Signals

DAK improves the detection ability of causal pathways by increasing the difference of convolution outputs between causal regions and non-causal regions (A and B) and enlarging similarities between samples carrying causal alleles (C and D).

(A) Locus indicated by the red arrow was selected as the causal position in the pathway (top). The learned weights of convolution layers (bottom) exhibit large responses in the causal position.

(B) Rank-sum tests on convolution outputs show significant differences between causal and non-causal regions.

(C) Sequence kernel association test (SKAT) on deep features obtains smaller p values than on original sequence (7.56×10^{-5} versus 8.93×10^{-3}). Samples with disease alleles ("Aa"/"aa," indicated by red/blue arrows) show higher similarity in deep features.

(D) Deep kernel matrix shows a near-Gaussian distribution; while original SKAT kernel shows a long-tail distribution with several extremely large outliers (in red box).

Explaining the Rationale of DAK with Simulated Pathways

To explain the rationale of how DAK improves the detection of association, we visualized and analyzed the functions of different deep layers. We simulated pathway sequences and phenotypes with a randomly assigned causal position (indicated by the red arrow in Figure 4A) using an additive genomic model.

We firstly showed that convolution layers could efficiently identify the causal regions. With learned weight matrices (Figure 4A, bottom), convolution layers exhibited larger responses in the region of the causal locus (Figures 4A [top curve] and S20). To statistically quantify changes between causal and non-causal regions, we employed rank-sum statistical tests[20] to calculate the rank difference of outputs from convolutional layers. p values indicated that most kernels had significantly different outputs between two regions (Figure 4B).

We next showed that deep kernel matrices could better define the sample similarity than original kernel matrices. Samples with disease alleles showed stronger similarities in deep-feature kernel than in original-sequence kernel (Figure 4C). When comparing sample similarities with and without disease genotypes ("AA" versus "Aa/aa"), the differences are minor in the original kernel matrix and but are obviously reflected in the deep kernel matrix (Figure S21). All entries in deep and original kernel matrices exhibited a near-Gaussian and long-tail distribution, respectively (Figures 4D and S22).

In the subsequent significance test on kernels, the large values in long-tail distribution can reduce the power and lead to weak association results.

The multiple-instance learning layer in DAK selected the pathway with maximal signal into the loss function. To evaluate whether DAK can prioritize pathways with true associations, we output indices of selected pathways and compared them with the true index of associated pathways based on the experiments in Figure 2B. From the precision score, we observed that DAK could accurately identify the pathway with true association from all candidates for most genetic models (Figure S23).

Applications to Real Datasets

We performed DAK on four disease datasets: GC, CRC, LC, and schizophrenia (SP) (Table S1). After the quality control steps, we divided all SNPs into pathway groups by their genetic coordinates (Experimental Procedures). DAK was optimized on one-hot coded pathways, and the score test was conducted on each pathway using learned neural network parameters to obtain the statistical p value.

For the GC dataset, three Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways exhibited genome-wide significance after Bonferroni correction ($\alpha = 0.05/186 = 2.68 \times 10^{-4}$). Two of them (*terpenoid backbone biosynthesis* and *oxidative phosphorylation*) showed strong associations (Figure 5A and Table S2). In a previous study, *terpenoid backbone biosynthesis* was identified as having a strong relation with hepatocellular carcinoma using microRNA and mRNA high-throughput sequencing.[21] *Oxidative phosphorylation* is closely related to the biological process in mitochondria and plays an essential role in the development of tumors.[22] Existing studies have shown its association with endometrial carcinoma, leukemias, and lymphomas.[23] Recent work also indicated that it could be an important target to treat cancer using a relevant inhibitor.[24] The *focal adhesion* pathway is important for cell proliferation, cell survival, and cell migration. In cancer, activities of focal adhesion are altered during tumor formation and development.[25] It is also a widely known target for cancer therapy development.[26] For the other three pathways showing borderline significance, *alpha linolenic acid metabolism* was discovered to downregulate human and mouse colon cancers;[27] the function of *ubiquitin mediated proteolysis* on cancers is also widely known.[28]

For the CRC dataset, DAK identified two KEGG pathways showing genome-wide significance (Figure 5B and Table S3). The most significant pathway, *allograft rejection*, is well known as an immune action pathway. The relation between allograft rejection, blood transfusion, and colorectal cancer recurrence was reported as early as 1987.[29] The other significant pathway, *glyoxylate and dicarboxylate metabolism*, was recently identified to be related to the metabolic switch in colorectal cancer cells.[30] Another three pathways, *one carbon pool by folate*, *oocyte meiosis*, and *amino sugar and nucleotide sugar metabolism*, were also discovered as high-risk pathways to CRC. The mechanism between one-carbon metabolism and CRC has been studied,[31] and several key mutations in this pathway have been related to CRC.[32] *Oocyte meiosis* was identified to be associated with colonic diseases in a previous study based on expression data,[33] and *amino sugar and nucleotide sugar metabolism* may contribute to the lipid metabolism abnormality in CRC.[34] For this dataset, we also ran DAK with and without the adjustment of population structures. DAK maintained stable performances in both conditions (Table S4 and Figure S24).

For the LC dataset, DAK reported two significant pathways: *lysine degradation* and *proteasome* (Figure 5C and Table S5). In LC treatment, proteasome inhibitor has been used to treat non-small cell LC and small cell LC[35–37] while lysine modification was discovered to affect a wide range of cancer types.[38] The other three pathways also had relatively small p values. The CRC pathway indicates that LC may share causal genes with certain types of CRC. Lysosome was reported to support the development LC.[39] The primary immunodeficiency pathway is known to lead to infections and cancers.[40] To evaluate the stability of associated pathways, we further performed analysis on another independent LC dataset with 14,803 cases and 12,262 controls. In the new dataset, we successfully replicated significantly associated pathways identified from the previous LC dataset (Table S5). We also discovered two interesting pathways in the new dataset showing strong associations with LC: *drug metabolism cytochrome P450* ($p = 0.00229$) and *nicotinate and nicotinamide metabolism* ($P = 0.00103$) (Table S6). These two pathways were closely related to the metabolism of chemicals in smoking, which is widely known as a major risk factor for LC.

For the SP dataset, we did not identify pathways reaching genome-wide significance after statistical correction (Figure 5D and Table S7). Interestingly, one pathway, *dilated cardiomyopathy* (DCM), showed borderline significance with SP. This pathway is related to heart muscle disease and can lead to heart failure. There is no existing study indicating its biological connection to SP. However, one clinical investigation has shown that after neuroleptics to treat SP, patients had a significantly increased possibility of developing DCM.[41] In other detailed case reports, the use of clozapine as treatment for SP finally led to DCM.[42–44] This implies that SP and DCM may share biological pathways and that the treatment may target the process that is important to both.

We also performed analysis on these real datasets with permuted labels to assess null distributions (Figure S25). Taken together, DAK efficiently discovered pathways that were known to be associated with diseases and also revealed potential associated pathways.

DISCUSSION

The identification of genetic causality can lead to valuable insights into the development of complex diseases. In this work, we employed DAK to discover disease-associated pathways by deep kernel learning. We demonstrated that DAK had promising and stable accuracies in discovering different types of causal variants, including common/rare loci, single/joint causal effects, various gene-disease models, and strong/weak effect levels, meanwhile controlling well the overfitting problem. DAK is computationally efficient (Figure S26) and is able to work with large-scale datasets due to the batch-training mechanism. To our knowledge, this is the first work that takes all of these important disease conditions into consideration. We also demonstrated the usability of DAK on four real datasets including cancers and mental disease.

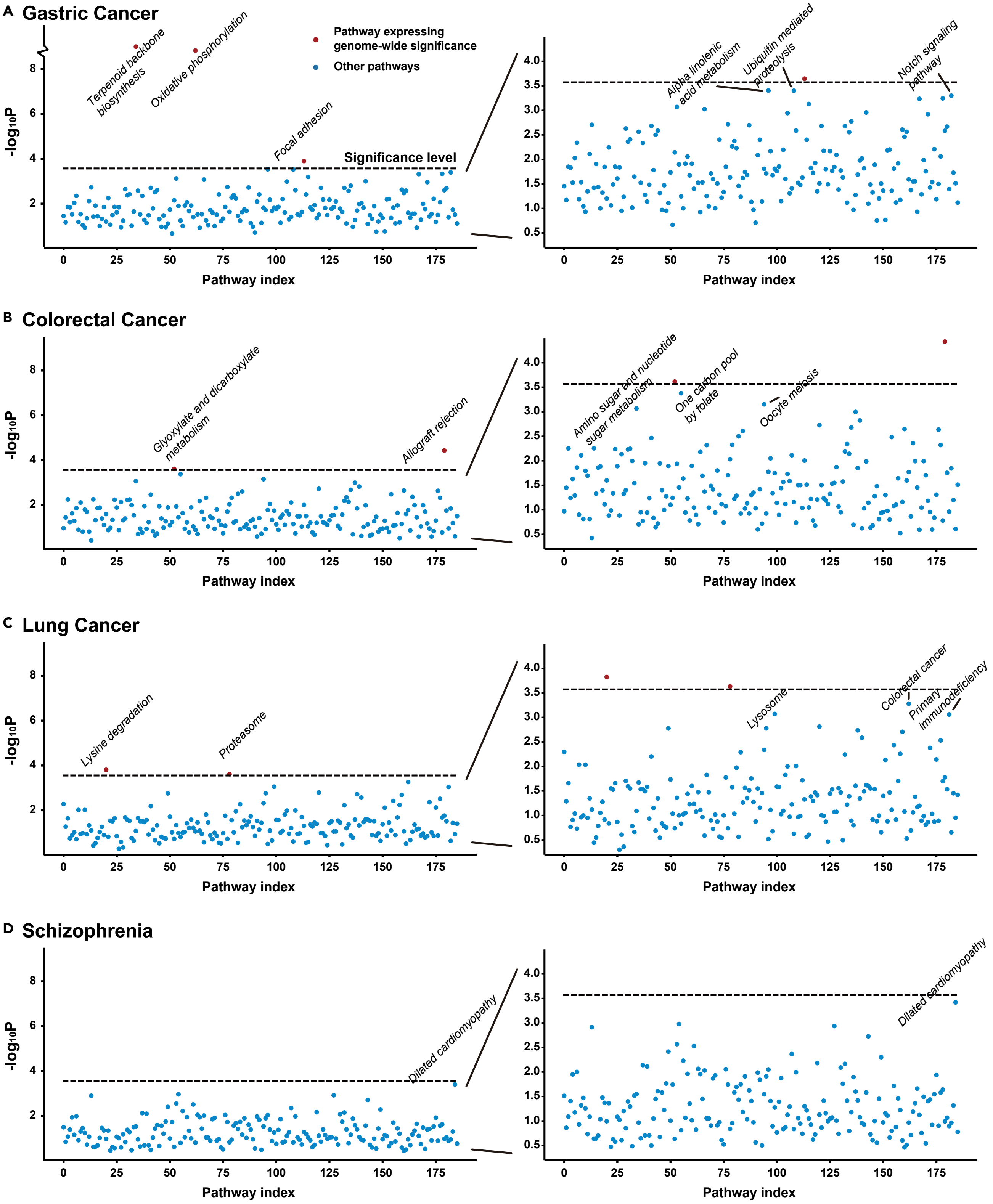


Figure 5. Scatterplots of p Values of KEGG Pathways by DAK on Four Real Datasets

Datasets from (A) gastric cancer, (B) colorectal cancer, (C) lung cancer, and (D) schizophrenia. Pathways showing genome-wide significances after Bonferroni correction ($\alpha = 0.05/186 = 2.68 \times 10^{-4}$) are marked in red.

Beyond current analyses, it is potentially interesting to explore DAK's performances from other directions in the future, given the availability of proper datasets. Large-scale datasets can be more informative in association analyses and can cover more complex population structures. In this work, we have not fully considered complex genetic variations such as similar biological functions from multiple SNPs and single genetic variation with multiple functional consequences. It would be meaningful to incorporate such complexity with the development of new simulation tools. DAK also shows potential to be used for other genomic research problems including disease risk predictions and gene-level GWASs. For real experiments, we discussed results from existing studies to gather evidence to support our discoveries. However, we note here that these can only be viewed as "partial evidence" and cannot yet be regarded as ground truth for evaluations. Future analyses of datasets with clinical evidence would an ideal way to evaluate the performance of DAK on real data.

Taken together, DAK offers an advanced and interpretable tool for GWASs at pathway level.

EXPERIMENTAL PROCEDURES

Resource Availability

Lead Contact

Qionghai Dai, PhD; qhdai@tsinghua.edu.cn.

Materials Availability

This study did not include new materials.

Data and Code Availability

The genotyping data of GWASs of GC and SP were deposited in dbGaP: phs000361 and phs000021, separately. The genotyping data of GWAS of colorectal cancer and LC were derived from previous studies.[45,46]

DAK is available from Github: https://github.com/fbaothu/DAK.

Other tools used in this work can be downloaded from:

Plink: http://zzz.bwh.harvard.edu/plink/; HAPGEN 2: https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html; The 1000 Genomes Project: http://www.1000genomes.org/; UCSC Genome Browser: https://genome.ucsc.edu/; SKAT and SKAT-o: https://www.hsph.harvard.edu/skat/; GATES, HYST, and aSPU: https://cran.r-project.org/web/packages/aSPU/index.html.

DAK Architecture

For the *i*th individual from a total number of *N* samples, y_i denotes the phenotype (such as disease or control); $x_i \in \mathbb{R}^K$ is an adjusted vector composed of *K* environmental related factors (e.g., gender, stratification, and bias). The genotype of each SNP belongs to one of three types: major homozygous, heterozygous, and minor homozygous genotypes. Therefore, it is natural to represent the genotype of each SNP by a one-hot vector with the non-zero entry indicating its particular genotype.

We group all $I^{(p)}$ SNPs on the *p*th pathway of individual *i* together and obtain the corresponding pathway-level genotype matrix $g_i^{(p)} \in \mathbb{R}^{I^{(p)} \times 3}$. After pathway assembling, we obtain a total number of *P* pathways for all samples.

We transform each $g_i^{(p)}$ through convolutional layers $conv(\bullet|\Theta_c)$ with *M* convolutional operators:

$$f_i^{(p)} = cov(g_i^{(p)}|\Theta_c) = \left[\max\left[f_{c_1}\left(g_i^{(p)}|\theta_{c_1}\right)\right],\right.$$

$$\left.\max\left[f_{c_2}\left(g_i^{(p)}|\theta_{c_2}\right)\right], \ldots, \max\left[f_{c_M}\left(g_i^{(p)}|\theta_{c_M}\right)\right]\right]^T \in \mathbb{R}^M,$$

where $f_{c_j}(\bullet|\theta_{c_j})$ represents the *j*th convolutional operator with parameter θ_{c_j} and $\max[\bullet]$ is the max-pooling operator. $\Theta_c = \{\theta_{c_1}, \ldots \theta_{c_M}\}$ denotes all learnable parameters of the convolutional layer.

By applying the output of the convolutional layers through a h_∞ layer,[47] we obtained the kernel representation of the *p*th pathway for individual *i*,

$$h_\infty\left(f_i^{(p)}\right) = \left[k\left(f_i^{(p)}, f_1^{(p)}\right), \ldots k\left(f_i^{(p)}, f_j^{(p)}\right) \ldots k\left(f_i^{(p)}, f_N^{(p)}\right)\right],$$

where $k(\bullet, \bullet)$ is a kernel function[12] (Supplemental Experimental Procedures) and *N* is the number of samples. Because the kernel function is applied to deep features $f_i^{(p)}$ instead of raw sequences, we note here that weighed kernel functions by MAF are not applicable.

We then define a pathway-level kernel regression function:

$$l_i^{(p)} = \mathcal{L}\left(x_i, h_\infty\left(f_i^{(p)}\right)|\omega\right) = \alpha x_i + \beta h_\infty\left(f_i^{(p)}\right),$$

where $\omega = \{\alpha, \beta\}$ contains learnable regression coefficients for environment factor and genotype features, respectively. For individual *i*, we can obtain $[l_i^{(1)} \ldots l_i^{(P)}]$ from a total number of *P* pathways.

We noticed that the labels (disease versus non-disease) are only provided at the individual level while not at each single pathway level. We hence consider multiple-instance learning loss[48] and define the individual level label for sample *i* as

$$L_i = \max\left[l_i^{(1)} \ldots l_i^{(P)}\right].$$

Multiple-instance learning selects the pathway with the maximal response from all pathways into the next layer. This multiple-instance learning loss is naturally explained in the context of GWAS: a sample is treated as a patient if at least one of his or her pathways is associated with the disease. The training loss is defined as

$$C = \frac{1}{N}\sum_{i=1}^{N} cost(y_i, \sigma(L_i)),$$

where $\sigma(\bullet)$ is the sigmoid function that converts regression outcomes into probabilities and $cost(\bullet)$ is the cost function that calculates losses between true labels and predicted labels. Here we used cross entropy. This loss function is optimized by TensorFlow in batches.

After well training, the kernel machine regression is used to model the relation between phenotype and kernel matrix. Kernel method has been validated as a powerful approach to quantify the statistical significance of each pathway and is widely used in a number of GWAS methods[12,19] (Supplemental Experimental Procedures). For each pathway *p*, the statistical score was derived from the kernel similarity matrix $\mathcal{K}^{(p)} = [h_\infty(f_1^{(p)}), \ldots h_\infty(f_i^{(p)}) \ldots h_\infty(f_N^{(p)})]^T$ via

$$Q_p = (L - Y)^T \mathcal{K}^{(p)} (L - Y),$$

where $L = [l_1^{(p)}, \ldots, l_N^{(p)}]$ (resp. $Y = [y_1, \ldots, y_N]$) is the predicted (resp. ground truth) disease statues for the pathway *p* across *N* samples. As introduced in SKAT, the Q_p was compared with the mixture of χ^2 distributions to obtain p value.

Simulation of Genotype and Data Preprocessing

We downloaded haplotypes of the CEU population from the 1000 Genomes Project.[49] Based on this reference, we simulated full genome data of 10,000 samples using HapGen 2 software.[50] On simulated dataset, we performed the following data quality control steps using Plink:[4] removing individuals with missingness >0.05; removing SNPs with missing rate >0.05 or Hardy-Weinberg equilibrium <1 × 10^{-5}. Thereafter, all data were converted into raw files.

Simulation of Phenotypes

Phenotypes for samples were simulated based on statistical hypothesis. Under null hypothesis that no causal pathway existed, case/control (represented in 1/0) labels were assigned randomly. Under alternative hypotheses, phenotypes were generated using linear models:

$$\log\left(\frac{r_k}{1 - r_k}\right) = \alpha + \beta^T x_k + \gamma c_k + \epsilon,$$

where r_k is the probability for sample *k* being a disease; $x_k \in \mathbb{R}^K$ is the vector of environmental factors as already mentioned and $\beta \in \mathbb{R}^K$ is the corresponding effect weights; $c_k \in \mathbb{R}$ is the genotype of pre-selected causal SNP and is coded according to the genetic model assumption: for example, $c_k = 0, 1, 0$ for the genotype "AA," "Aa," "aa," respectively. For a multiplicative genetic model where the disease increased exponentially, we first determine the risk r_k for samples with "Aa" allele and then exponentially increase the risk for "aa" samples. γ is the effect size of genotype. We followed the same setting in SKAT,[13] with a 0.2 effect size equivalent to odds ratio of 1.22. We note here that in type I error analysis, different genetic models will have no effect to the simulated phenotype because the γ was set to zero. Therefore, we did not evaluate the error-rate performance with different genetic models.

For simulation of disease caused by joint effects, we extend the linear model to

$$\log\left(\frac{r_k}{1-r_k}\right) = \alpha + \beta^T x_k + \sum_{j=1}^{N_c} \gamma^{(j)} c_k^{(j)} + \epsilon,$$

where N_c is the number of causal SNPs. After simulating phenotypes, we randomly selected 50% cases and 50% controls for analyses.

Pathway Set Assembling

A total of 186 KEGG pathways were downloaded from the Molecular Signatures Database (MSigDB) in the items of "C2: curated gene sets."[51] The whole-genome SNPs were firstly mapped to genes based on their positions (RefSeq hg19),[52] then genes within the same pathway were further assembled together. Finally, pathway-level SNP sets were used as input for analysis. If variants had multiple gene mappings, we assigned them to different gene sets. We also tested the performance of DAK on pathway sets with random gene orders and on regulatory regions (Figures S27 and S28).

Real Dataset Collections

All GWAS datasets are described in Table S1. In brief, the raw genotypes were firstly imputed using SHAPEIT and IMPUTE2 based on the 1000 Genomes Project (Phase I, version 3, 1,092 individuals). The imputed SNPs were then cleaned with the criteria of (1) MAF <0.01, (2) call rate <95%, (3) Hardy-Weinberg equilibrium $p < 1.0 \times 10^{-6}$, (4) info score <0.3. The population structure was estimated by a principal components analysis using EIGENSOFT 5.0.1, and the principal components were extracted as covariates, corresponding with age, sex, and variables if appropriate for modeling adjustment. Performances with different MAF filtering depths were also provided (Figure S29). The study protocol was performed in accordance with the Institutional Review Board of Nanjing Medical University and Massachusetts General Hospital, the Human Subjects Committee of the Harvard School of Public Health, and the research use statements in the database of Genotypes and Phenotypes (dbGaP).

Evaluation

Performances of all methods were quantified under two metrics, type I error rate and empirical power, corresponding to experiments conducted under the assumption that no disease existed or no causal pathway existed. On simulated datasets, all comparison methods were used to derive pathway-level p values. Under each experimental setting, the association analysis was repeated 100 times on different datasets that were randomly sampled from simulated data. The type I error rate/empirical power was then defined as the proportion of experiments detecting significant pathways among 100 repeats.

Comparison Methods

HYST combines extended Simes' test and scaled χ^2 test from single SNP association results.

Burden test uses MAF as weights and additively combines all SNPs.

GATES takes extended Simes' test to aggregate single SNP test results.

SKAT employs kernels to model the similarity between individuals and directly calculates the association significance between sample kernels and sample phenotypes. Here we used the default kernel setting ("linear.-weighted") and default parameters.

aSPU is a method for adaptive testing of association analysis. It employs the sum of powered score tests to combine single SNPs.

SKAT-o combines SKAT and Burden test and selects the best results from them. We also used the default settings for SKAT.

The detailed structure of DAK is illustrated in Figure S1. We also employed linear kernel to be comparable with SKAT and provided performance evaluations of DAK using other alternative kernels (Figure S30). The model was constructed in TensorFlow framework and was run on a machine with Nvidia Titan X GPU. We set the training epoch to 100 and optimized parameters using ADAM optimizer. Performances with changing structure parameters were also provided (Figure S31).

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.patter.2020.100057.

ACKNOWLEDGMENTS

We thank Prof. Xihong Lin from Harvard T.H. Chan School of Public Health for valuable discussion. This study was supported by the National Natural Science Foundation of China (no. 61327902 to Q.D. and no. 61971020 to Y.D.), the Project of BMSTC (no. Z181100003118014 to Q.D.), Tsinghua University Initiative Scientific Research Program (Q.D.), Science and Technology on Space Intelligent Control Laboratory (no. HTKJ2019KL502006 to Y.D.), US National Institutes of Health (U01CA209414 to D.C.C.), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (Public Health and Preventive Medicine grant to M.W.).

AUTHOR CONTRIBUTIONS

F.B., Y.D., and Q.D. developed the algorithms. F.B., Y.D., M.D., Z.R., S.W., S.L., B.W., K.Y.L., and Q.D. conducted experimental analysis on both simulated and biological datasets. K.Y.L. packed the algorithm into the software package. M.D., J.X., F.C., D.C.C., and M.W. collected real datasets and performed data processing. The manuscript was written by F.B., Y.D., M.D., M.W., and Q.D. All authors read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 26, 2020
Revised: May 25, 2020
Accepted: June 1, 2020
Published: July 1, 2020

REFERENCES

1. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. *101*, 5–22.
2. Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. *6*, 95.
3. Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. Nat. Rev. Genet. *11*, 843.
4. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.
5. Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J.D., Jin, L., and Amos, C.I. (2010). Gene and pathway-based second-wave analysis of genome-wide association studies. Eur. J. Hum. Genet. *18*, 111.
6. Jin, L., Zuo, X.Y., Su, W.Y., Zhao, X.L., Yuan, M.Q., Han, L.Z., Zhao, X., Chen, Y.D., and Rao, S.Q. (2014). Pathway-based analysis tools for complex diseases: a review. Genomics Proteomics Bioinformatics *12*, 210–220.
7. White, M.J., Yaspan, B.L., Veatch, O.J., Goddard, P., Risse-Adams, O.S., and Contreras, M.G. (2019). Strategies for pathway analysis using GWAS and WGS data. Curr. Protoc. Hum. Genet. *100*, e79.

8. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., and Lander, E.S. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U S A *102*, 15545–15550.

9. Li, M.X., Gui, H.S., Kwan, J.S., and Sham, P.C. (2011). GATES: a rapid and powerful gene-based association test using extended Simes procedure. Am. J. Hum. Genet. *88*, 283–293.

10. Wang, J., Vasaikar, S., Shi, Z., Greer, M., and Zhang, B. (2017). WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res. *45*, W130–W137.

11. 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. Hum. Genet. *89*, 82–93.

12. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Team, E.L., Christiani, D.C., Wurfel, M.M., and Lin, X.; NHLBI GO Exome Sequencing Project (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. *91*, 224–237.

13. Lin, X., Lee, S., Wu, M.C., Wang, C., Chen, H., Li, Z., and Lin, X. (2016). Test for rare variants by environment interactions in sequencing association studies. Biometrics *72*, 156–164.

14. Ainscough, B.J., Barnell, E.K., Ronning, P., Campbell, K.M., Wagner, A.H., Fehniger, T.A., Dunn, G.P., Uppaluri, R., Govindan, R., Rohan, T.E., et al. (2018). A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. Nat. Genet. *50*, 1735–1743.

15. Sundaram, L., Gao, H., Padigepati, S.R., McRae, J.F., Li, Y., Kosmicki, J.A., Fritzilas, N., Hakenberg, J., Dutta, A., Shon, J., and Xu, J. (2018). Predicting the clinical impact of human mutation with deep neural networks. Nat. Genet. *50*, 1161–1170.

16. Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

17. 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.

18. Li, M.X., Kwan, J.S., and Sham, P.C. (2012). HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. Am. J. Hum. Genet. *91*, 478–488.

19. Pan, W., Kwak, I.-Y., and Wei, P. (2015). A powerful pathway-based adaptive test for genetic association with common or rare variants. Am. J. Hum. Genet. *97*, 86–98.

20. Steel, R.G. (1959). A multiple comparison rank sum test: treatments versus control. Biometrics, 560–572.

21. Ding, M., Li, J., Yu, Y., Liu, H., Yan, Z., Wang, J., and Qian, Q. (2015). Integrated analysis of miRNA, gene, and pathway regulatory networks in hepatic cancer stem cells. J. Transl. Med. *13*, 259.

22. Maiuri, M.C., and Kroemer, G. (2015). Essential role for oxidative phosphorylation in cancer progression. Cell Metab. *21*, 11–12.

23. Ashton, T.M., McKenna, W.G., Kunz-Schughart, L.A., and Higgins, G.S. (2018). Oxidative phosphorylation as an emerging target in cancer therapy. Clin. Cancer Res. *24*, 2482–2490.

24. Molina, J.R., Sun, Y., Protopopova, M., Gera, S., Bandi, M., Bristow, C., McAfoos, T., Morlacchi, P., Ackroyd, J., and Agip, A.N. (2018). An inhibitor of oxidative phosphorylation exploits cancer vulnerability. Nat. Med. *24*, 1036.

25. Eke, I., and Cordes, N. (2015). Focal adhesion signaling and therapy resistance in cancer. Semin. Cancer Biol. *31*, 65–75.

26. McLean, G.W., Carragher, N.O., Avizienyte, E., Evans, J., Brunton, V.G., and Frame, M.C. (2005). The role of focal-adhesion kinase in cancer—a new therapeutic opportunity. Nat. Rev. Cancer *5*, 505–515.

27. Chamberland, J.P., and Moon, H.-S. (2014). Down-regulation of malignant potential by alpha linolenic acid in human and mouse colon cancer cells. Fam. Cancer *14*, 25–30.

28. Salghetti, S.E., Kim, S.Y., and Tansey, W.P. (1999). Destruction of Myc by ubiquitin-mediated proteolysis: cancer-associated and transforming mutations stabilize Myc. EMBO J. *18*, 717–726.

29. Weiden, P.L., Bean, M.A., and Schultz, P. (1987). Perioperative blood transfusion does not increase the risk of colorectal cancer recurrence. Cancer *60*, 870–874.

30. Charitou, T., Srihari, S., Lynn, M.A., Jarboui, M.A., Fasterius, E., Moldovan, M., Shirasawa, S., Tsunoda, T., Ueffing, M., Xie, J., et al. (2019). Transcriptional and metabolic rewiring of colorectal cancer cells expressing the oncogenic KRAS G13D mutation. Br. J. Cancer *121*, 37–50.

31. Hanley, M.P., and Rosenberg, D.W. (2015). One-carbon metabolism and colorectal cancer: potential mechanisms of chemoprevention. Curr. Pharmacol. Rep. *1*, 197–205.

32. Myte, R., Gylling, B., Häggström, J., Schneede, J., Löfgren-Burström, A., Huyghe, J.R., Hallmans, G., Meyer, K., Johansson, I., Ueland, P.M., et al. (2018). One-carbon metabolism biomarkers and genetic variants in relation to colorectal cancer risk by KRAS and BRAF mutation status. PLoS One *13*, e0196233.

33. Wu, D., Li, Q., Song, G., and Lu, J. (2016). Identification of disrupted pathways in ulcerative colitis-related colorectal carcinoma by systematic tracking the dysregulated modules. J. BUON *21*, 366–374.

34. Han, S., Pan, Y., Yang, X., Da, M., Wei, Q., Gao, Y., Qi, Q., and Ru, L. (2019). Intestinal microorganisms involved in colorectal cancer complicated with dyslipidosis. Cancer Biol. Ther. *20*, 81–89.

35. Scagliotti, G. (2006). Proteasome inhibitors in lung cancer. Crit. Rev. Oncol. Hematol. *58*, 177–189.

36. Escobar, M., Velez, M., Belalcazar, A., Santos, E.S., and Raez, L.E. (2011). The role of proteasome inhibition in nonsmall cell lung cancer. Biomed. Res. Int. https://doi.org/10.1155/2011/806506.

37. Sooman, L., Gullbo, J., Bergqvist, M., Bergström, S., Lennartsson, J., and Ekman, S. (2017). Synergistic effects of combining proteasome inhibitors with chemotherapeutic drugs in lung cancer cells. BMC Res. Notes *10*, 544.

38. Chen, L., Miao, Y., Liu, M., Zeng, Y., Gao, Z., Peng, D., Hu, B., Li, X., Zheng, Y., Xue, Y., and Zuo, Z. (2018). Pan-cancer analysis reveals the functional importance of protein lysine modification in cancer development. Front. Genet. *9*, 254.

39. Patra, K.C., Weerasekara, V.K., and Bardeesy, N. (2019). AMPK-mediated lysosome biogenesis in lung cancer growth. Cell Metab. *29*, 238–240.

40. Salavoura, K., Kolialexi, A., Tsangaris, G., and Mavrou, A. (2008). Development of cancer in patients with primary immunodeficiencies. Anticancer Res. *28*, 1263–1269.

41. Volkov, V., and Volkov, V. (2013). Dilated cardiomyopathy in patients with schizophrenia. Ter. Arkh. *85*, 43–46.

42. Longhi, S., and Heres, S. (2017). Clozapine-induced, dilated cardiomyopathy: a case report. BMC Res. Notes *10*, 338.

43. Tanner, M., and Culling, W. (2003). Clozapine associated dilated cardiomyopathy. Postgrad. Med. J. *79*, 412–413.

44. Bobb, V.T., Jarskog, L.F., and Coffey, B.J. (2010). Adolescent with treatment-refractory schizophrenia and clozapine-induced cardiomyopathy managed with high-dose olanzapine. J. Child Adolesc. Psychopharmacol. *20*, 539–543.

45. Xin, J., Du, M., Gu, D., Ge, Y., Li, S., Chu, H., Meng, Y., Shen, H., Zhang, Z., and Wang, M. (2019). Combinations of single nucleotide polymorphisms identified in genome-wide association studies determine risk for colorectal cancer. Int. J. Cancer *145*, 2661–2669.

46. Wang, Z., Wei, Y., Zhang, R., Su, L., Gogarten, S.M., Liu, G., Brennan, P., Field, J.K., McKay, J.D., Lissowska, J., and Swiatkowska, B. (2018). Multi-omics analysis reveals a HIF network and hub gene EPAS1 associated with lung adenocarcinoma. EBioMedicine *32*, 93–101.

47. Wilson, A.G., Hu, Z., Salakhutdinov, R., and Xing, E.P. (2016). Deep kernel learning. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (PMLR 51), pp. 370–378.

48. Maron, O., and Lozano-Pérez, T. (1998). A framework for multiple-instance learning. In Advances in Neural Information Processing Systems 10 (NIPS 1997), M.I. Jordan, M.J. Kearns, and S.A. Solla, eds. (MIT Press), pp. 570–576.

49. Siva, N. (2008). 1000 Genomes Project. Nat. Biotechnol. *26*, https://doi.org/10.1038/nbt0308-256b.

50. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304–2305.

51. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. Cell Syst. *1*, 417–425.

52. Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. (2014). The UCSC genome browser database: 2015 update. Nucleic Acids Res. *43*, D670–D681.