Explaining the genetic causality for complex diseases *via* deep association kernel learning

Feng Bao[1, *], Yue Deng[2, *, #], Mulong Du[3, 4], Zhiquan Ren[1], Sen Wan[1], Junyi Xin[5], Feng Chen[4], David C. Christiani[3, 6], Meilin Wang[5, #] and Qionghai Dai[1, #]

[1] Institute for Brain and Cognitive Sciences and Department of Automation, Tsinghua University, Beijing 100084, China.

[2] Institute of Artificial Intelligence, Beihang University, Beijing 100191, China.

[3] Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA.

[4] Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China.

[5] Department of Environmental Genomics, School of Public Health, Nanjing Medical University, Nanjing 211166, China.

[6] Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA.

*These authors contributed equally to this work.

# Corresponding authors: Yue Deng (ydeng@buaa.edu.cn), Meilin Wang (mwang@njmu.edu.cn) and Qionghai Dai (qhdai@tsinghua.edu.cn)

ABSTRACT

The genetic effect explains the causality from genetic mutation 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 that existing approaches fail. When applied to real-world GWAS data, our approach discovered potential casual pathways that could be explained by alternative biological studies.

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 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 statistic scores by combing P-values from single-locus tests [8-10]; SKAT [11] and its variants [12, 13] perform association test using kernel regression. However, these existing approaches rely on some pre-assumed genetic models to conduct hand-crafted 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 highly needed.

We introduce the Deep Association Kernel learning (DAK) framework to conduct pathway-level GWAS (**Fig. 1**). Our DAK framework incorporates convolutional layers to encode raw SNPs as latent genetic representation. Then, kernel regression layers are connected with these encoded genetic representations to predict the disease status. More importantly, this kernel regression layer allows performing 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)[14], enrichment methods (GATES, HYST and aSPU)[9, 15, 16] 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 including gastric cancer, colorectal cancer, lung cancer and psychiatric disorder.

Results

Deep association kernel learning

We introduced deep association kernel (DAK) learning to achieve the detection of complex associations and enhance the interpretability of GWAS (**Fig. 1** and **Methods**). 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. Then, pathway-level features are extracted by convolutional layers (**Supplementary Fig. 1**), followed by a kernel regression layer to derive the statistical significance (**Supplementary Fig. 2**). To allow learning from labels at 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 (**Supplementary Figs. 3** and **4**).

Type-I error

In each simulation experiment, we simulated dataset under null (no causal pathway) or alternative (disease was caused by different genetic associations) hypothesis (**Fig. 2a** and **Methods**). 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) (**Methods**) in 100 replicates.

We first report the Type-I error. If no causal loci existed in all pathways (null hypothesis), all methods showed low error rate level (**Supplementary Fig. 5**). Changing the sample size had little effects on the results. The training curve showed DAK converged within several iterations (**Supplementary Fig. 6**).

Single effect

We then considered that the disease was caused by a single common variant. To illustrate different functional pathway of genes to the disease, we assumed the allele of the causal locus contributed to the disease in five different genetic models: 1) additive model, minor homozygous genotype had two-fold effect than 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 (**Fig. 2a**).

On the most widely-used additive disease mode, we found that all methods showed reasonable accuracy to identify the pathway with disease locus (**Fig. 2b** and **Supplementary Fig. 7**). However, when the fundamental genetic model changes, the power of all comparing methods dropped dramatically while DAK maintained reliable performances with best power across all conditions. Specifically, for the challenging recessive genetic model, accuracies of all comparing 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 (**Supplementary Fig. 8**). We further noted that when the sample size was increased to 5,000, the power of all methods were increased and DAK was still the best (**Fig. 2b** and **Supplementary Fig. 7**).

The discovery of rare variants (minor allele frequency < 0.5%) 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 (**Fig. 2c** and **Supplementary Fig. 9**). When the effect size was decreased, other comparing approaches failed but DKA can still maintain very reliable performances (**Supplementary Fig. 9**). We demonstrated DAK could discover the causal rare variant at power around 0.8 on datasets even only with 3,000 samples (**Fig. 2d** and **Supplementary Fig. 9**), which was a challenging task for other methods.

Joint effect

Most diseases are results 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 3 causal common variants and generated phenotype under 5 genetic models (**Methods**). Performances of all methods were much lower compared with results under single variant. However, DAK still dramatically outperformed other methods and achieved the most stable performance among all experiments (**Fig. 3a** and **Supplementary Fig. 10**). The performances of all methods was enhanced when the effect size was increased. The advantages of DAK were more obvious when the causal positions were rare variants. (**Figs. 3b, c** and **Supplementary Fig. 11**)

Applications to real datasets

We performed DAK on four disease datasets: gastric cancer, colorectal cancer, lung cancer and schizophrenia (**Supplementary Table 1**). After the quality control steps, we divided all SNPs into pathway groups by their genetic coordinates (**Methods**). DAK was optimized on one-hot coded pathways and score test was conducted on each pathway using learned neural network parameters to get the statistical P-value.

For the gastric cancer (GC) dataset, three KEGG pathways exhibited genome-wide significance after Bonferroni correction (α = 0.05/186 = 2.68E-4). Two of them (*Terpenoid backbone biosynthesis* and *oxidative phosphorylation*) showed strong associations (**Fig. 4a** and **Supplementary Table 2**). In a previous study, *terpenoid backbone biosynthesis* was identified as a strong relation to Hepatocellular carcinoma (HCC) using miRNA and mRNA high-throughput sequencing[17]. *Oxidative phosphorylation* is closely related to the biological process in mitochondria and it plays an essential role in the development of tumors [18]. Existing studies have shown its association to endometrial carcinoma, leukemias, lymphomas, etc [19]. Recent work also indicated it could be an important target to treat cancer using relevant inhibitor [20]. For the *focal adhesion* pathway, it is important for cell proliferation, cell survival and cell migration. In cancer, activities of focal adhesion were altered during tumor formation and developing[21]. It is also a widely known target for cancer therapy development[22]. For the other three pathways showing borderline significances, *alpha linolenic acid metabolism* was discovered to downregulated human and mouse colon cancers[23]; Function of *ubiquitin mediated proteolysis* on cancers was also widely known[24].

For the colorectal cancer (CRC) dataset, DAK identified two KEGG pathways showing genome-wide significance (**Fig. 4b** and **Supplementary Table 3**). 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 at early time[25]. The other significant pathway *glyoxylate and dicarboxylate metabolism* was recently identified to be related to the metabolic switch in colorectal cancer cells [26]. Other three pathways, *one carbon pool by folate*, *oocyte meiosis* and *amino sugar and nucleotide sugar metabolism* were also discovered as high risky pathways to CRC. The mechanism between one-carbon metabolism and CRC has been studied[27] and several key mutations in this pathway has been related to CRC[28]. Oocyte meiosis was identified to be associated to colonic diseases in previous study based on expression data[29] and amino sugar and nucleotide sugar metabolism may contribute to the lipid metabolism abnormality in CRC[30].

For the lung cancer (LC) dataset, DAK reported two significant pathways: *lysine degradation* and *proteasome* (**Fig. 4c** and **Supplementary Table 4**). In LC treatment, proteasome inhibitor has been used to non-small cell and small cell LC[31-33] while lysine modification was discovered to impact a wide range of cancer types[34]. Other three pathways also had relatively small P-values. Colorectal cancer pathway indicates that LC may share causal genes with certain types of CRC. Lysosome was reported to support the development LC[35]. For primary immunodeficiency pathway, it is known to lead to infections and cancers[36].

For the schizophrenia (SP) dataset, we did not identify pathways reaching genome-wide significance after statistical correction (**Fig. 4d** and **Supplementary Table 5**). Interestingly, one pathway, *dilated cardiomyopathy* (DCM), showed borderline significance with SP. This pathway is related to the 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 of SP, patients had a significantly increased possibility to get DCM[37]. In other detailed case reports, the usage of clozapine as the treatment to SP finally lead to DCM [38-40]. This implies that SP and DCM may share biological pathways and the treatment may target at the process that is important to both.

Taken together, DAK efficiently discovered pathways that were known to be associated with diseases and also revealed new potential causal pathways.

Methods

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

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

$$f_i^{(p)} = cov\left(g_i^{(p)}\middle|\Theta_c\right) =$$
$$[\max\left[f_{c_1}\left(g_i^{(p)}|\theta_{c_1}\right)\right], \max\left[f_{c_2}\left(g_i^{(p)}|\theta_{c_2}\right)\right], \dots, \max\left[f_{c_M}\left(g_i^{(p)}|\theta_{c_M}\right)\right]]^T \in \mathbb{R}^M,$$

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

By applying the output of the convolutional layers through a h_∞ layer[41], we obtained the kernel representation of the p-th pathway for individual i,

$$h_\infty\left(f_i^{(p)}\right) = [k\left(f_i^{(p)}, f_1^{(p)}\right), \dots k\left(f_i^{(p)}, f_j^{(p)}\right) \dots k\left(f_i^{(p)}, f_N^{(p)}\right)]\ \forall\, j \neq\ i$$

where $k(\cdot,\cdot)$ is a kernel function[12] and N is the number of samples.

We then define a pathway-level kernel regression function:

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

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

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

$$L_i = \max\,[l_i^{(1)} \dots l_i^{(P)}]$$

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 pathways is associated with the disease. The training loss is defined as:

$$C = \sum_{i=1}^{N} \left|\left|y_i - L_i\right|\right|^2$$

This loss function is optimized by TensorFlow in batches.

After well training, we performed score test to quantify the statistical significance of each pathway using the same approach in SKAT[12]. For each pathway p, the statistic score was derived from the kernel similarity matrix $\mathcal{K}^{(p)} = \left[h_\infty\left(f_1^{(p)}\right), \dots h_\infty\left(f_i^{(p)}\right) \dots h_\infty\left(f_N^{(p)}\right)\right]^T$ via:

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

where $L = [l_1^{(p)}, \dots, l_N^{(p)}]$ (*resp*. $Y = [y_1, \dots, y_N]$) 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 CEU population from 1000 Genomes Project[42]. Based on this reference, we simulated full genome data of 10,000 samples using HapGen 2 software[43]. 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 <1e-5. After that, all data were converted into raw files.

Simulation of phenotype

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:

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

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 mentioned before 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 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 odd ratio of 1.22.

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

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

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 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were downloaded from the database of The Molecular Signatures Database (MSigDB) in the items of “C2: curated gene sets”[44]. The whole-genome SNPs were firstly mapped to genes based on their positions (RefSeq hg19[45]). Then genes grouped in the same pathway were further assembled together.

Real dataset collections

The genotyping data of GWASs of gastric cancer and schizophrenia were deposited in the database of Genotypes and Phenotypes (dbGaP; phs000361 and phs000021, separately). The genotyping data of GWASs of colorectal cancer and lung cancer were derived from previous studies[46, 47].

All GWAS datasets were firstly imputed using SHAPEIT and IMPUTE2 based on the 1000 Genomes Project (Phase I, version 3, 1092 individuals. Then the imputed SNPs were cleaned with the criteria of (i) MAF < 0.01; (ii) call rate < 95%; (iii) Hardy–Weinberg equilibrium $P < 1.0 \times 10^{-6}$; (V) info score < 0.3. The population structure was estimated by a PCA using EIGENSOFT 5.0.1, and the principle components were extracted as covariates, corresponding with age, sex and variables if appropriate for modeling adjustment.

Evaluation

Performances of all methods were quantified under two metrics: type I error rate and empirical power, corresponding to experiments conducted under assumptions that no disease existed or causal pathway existed. On simulated datasets, all comparing 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. Then, the type I error rate / empirical power was defined as the proportion of experiments detecting significant pathways among 100 repeats.

Comparison methods

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

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

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

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

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

DAK: The detail structure of DAK was illustrated in **Supplementary Fig. 1**. We also employed linear kernel to be comparable with SKAT. The model was constructed in TensorFlow framework and was run on machine with Nvidia Titan X GPU. We set the training epoch to 100.

Software availability

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

Acknowledgement:

We thank Prof. Xihong Lin from Harvard T.H. Chan School of Public Health for the valuable discussion.

Conflict of Interests:

Authors declare no conflict of interests.

References:

1. Visscher, P.M. et al. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* **101**, 5-22 (2017).
2. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nature reviews genetics* **6**, 95 (2005).
3. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics* **11**, 843 (2010).
4. Chang, C.C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
5. Peng, G. et al. Gene and pathway-based second-wave analysis of genome-wide association studies. *European Journal of Human Genetics* **18**, 111 (2010).
6. Jin, L. et al. Pathway-based analysis tools for complex diseases: a review. *Genomics, proteomics & bioinformatics* **12**, 210-220 (2014).
7. White, M.J. et al. Strategies for Pathway Analysis Using GWAS and WGS Data. *Current protocols in human genetics* **100**, e79 (2019).
8. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545-15550 (2005).
9. Li, M.X., Gui, H.S., Kwan, J.S. & Sham, P.C. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am J Hum Genet* **88**, 283-293 (2011).
10. Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic acids research* **45**, W130-W137 (2017).
11. Wu, M.C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
12. Lee, S. et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics* **91**, 224-237 (2012).
13. Lin, X. et al. Test for rare variants by environment interactions in sequencing association studies. *Biometrics* **72**, 156-164 (2016).
14. Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* **83**, 311-321 (2008).
15. Li, M.X., Kwan, J.S. & Sham, P.C. 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 (2012).
16. Pan, W., Kwak, I.-Y. & Wei, P. A powerful pathway-based adaptive test for genetic association with common or rare variants. *The American Journal of Human Genetics* **97**, 86-98 (2015).
17. Ding, M. et al. Integrated analysis of miRNA, gene, and pathway regulatory networks in hepatic cancer stem cells. *Journal of translational medicine* **13**, 259 (2015).
18. Maiuri, Maria C. & Kroemer, G. Essential Role for Oxidative Phosphorylation in Cancer Progression. *Cell Metabolism* **21**, 11-12 (2015).

19. Ashton, T.M., McKenna, W.G., Kunz-Schughart, L.A. & Higgins, G.S. Oxidative Phosphorylation as an Emerging Target in Cancer Therapy. *Clin Cancer Res* **24**, 2482-2490 (2018).
20. Molina, J.R. et al. An inhibitor of oxidative phosphorylation exploits cancer vulnerability. *Nature medicine* **24**, 1036 (2018).
21. Eke, I. & Cordes, N. Focal adhesion signaling and therapy resistance in cancer. *Semin Cancer Biol* **31**, 65-75 (2015).
22. McLean, G.W. et al. The role of focal-adhesion kinase in cancer - a new therapeutic opportunity. *Nat Rev Cancer* **5**, 505-515 (2005).
23. Chamberland, J.P. & Moon, H.-S. Down-regulation of malignant potential by alpha linolenic acid in human and mouse colon cancer cells. *Familial Cancer* **14**, 25-30 (2014).
24. Salghetti, S.E., Kim, S.Y. & Tansey, W.P. Destruction of Myc by ubiquitin-mediated proteolysis: cancer-associated and transforming mutations stabilize Myc. *The EMBO journal* **18**, 717-726 (1999).
25. Weiden, P.L., Bean, M.A. & Schultz, P. Perioperative blood transfusion does not increase the risk of colorectal cancer recurrence. *Cancer* **60**, 870-874 (1987).
26. Charitou, T. et al. Transcriptional and metabolic rewiring of colorectal cancer cells expressing the oncogenic KRAS G13D mutation. *British journal of cancer*, 1 (2019).
27. Hanley, M.P. & Rosenberg, D.W. One-carbon metabolism and colorectal cancer: Potential mechanisms of chemoprevention. *Current pharmacology reports* **1**, 197-205 (2015).
28. Myte, R. et al. One-carbon metabolism biomarkers and genetic variants in relation to colorectal cancer risk by KRAS and BRAF mutation status. *PloS one* **13**, e0196233 (2018).
29. Wu, D., Li, Q., Song, G. & Lu, J. Identification of disrupted pathways in ulcerative colitis-related colorectal carcinoma by systematic tracking the dysregulated modules. *Journal of BU ON.: official journal of the Balkan Union of Oncology* **21**, 366-374 (2016).
30. Han, S. et al. Intestinal microorganisms involved in colorectal cancer complicated with dyslipidosis. *Cancer biology & therapy* **20**, 81-89 (2019).
31. Scagliotti, G. Proteasome inhibitors in lung cancer. *Critical reviews in oncology/hematology* **58**, 177-189 (2006).
32. Escobar, M., Velez, M., Belalcazar, A., Santos, E.S. & Raez, L.E. The role of proteasome inhibition in nonsmall cell lung cancer. *BioMed Research International* **2011** (2011).
33. Sooman, L. et al. Synergistic effects of combining proteasome inhibitors with chemotherapeutic drugs in lung cancer cells. *BMC research notes* **10**, 544 (2017).
34. Chen, L. et al. Pan-cancer analysis reveals the functional importance of protein lysine modification in cancer development. *Frontiers in genetics* **9**, 254 (2018).
35. Patra, K.C., Weerasekara, V.K. & Bardeesy, N. AMPK-Mediated Lysosome Biogenesis in Lung Cancer Growth. *Cell metabolism* **29**, 238-240 (2019).

36. Salavoura, K., Kolialexi, A., Tsangaris, G. & Mavrou, A. Development of cancer in patients with primary immunodeficiencies. *Anticancer research* **28**, 1263-1269 (2008).
37. Volkov, V. & Volkov, V. Dilated cardiomyopathy in patients with schizophrenia. *Terapevticheskii arkhiv* **85**, 43-46 (2013).
38. Longhi, S. & Heres, S. Clozapine-induced, dilated cardiomyopathy: a case report. *BMC research notes* **10**, 338 (2017).
39. Tanner, M. & Culling, W. Clozapine associated dilated cardiomyopathy. *Postgraduate medical journal* **79**, 412-413 (2003).
40. Bobb, V.T., Jarskog, L.F. & Coffey, B.J. Adolescent with treatment-refractory schizophrenia and clozapine-induced cardiomyopathy managed with high-dose olanzapine. *Journal of child and adolescent psychopharmacology* **20**, 539-543 (2010).
41. Wilson, A.G., Hu, Z., Salakhutdinov, R. & Xing, E.P. in Artificial Intelligence and Statistics 370-378 (2016).
42. Siva, N. (Nature Publishing Group, 2008).
43. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304-2305 (2011).
44. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell systems* **1**, 417-425 (2015).
45. Rosenbloom, K.R. et al. The UCSC genome browser database: 2015 update. *Nucleic acids research* **43**, D670-D681 (2014).
46. Xin, J. et al. Combinations of single nucleotide polymorphisms identified in genome - wide association studies determine risk for colorectal cancer. *International journal of cancer* (2019).
47. Wang, Z. et al. Multi-omics analysis reveals a HIF network and hub gene EPAS1 associated with lung adenocarcinoma. *EBioMedicine* **32**, 93-101 (2018).

Figures

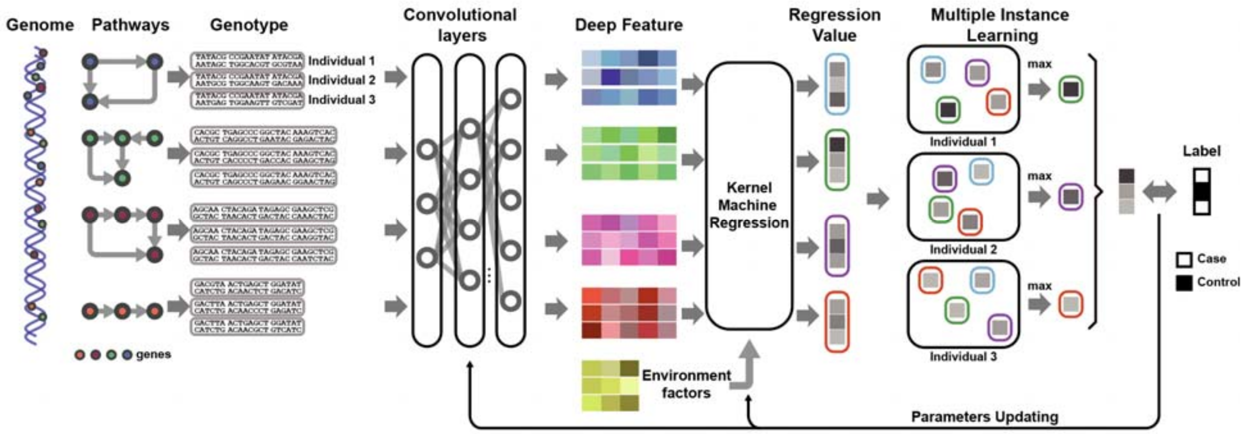


Figure 1. The framework of DAK. SNPs are grouped into pathway-level gene set and coded into one-hot format. Convolutional layers are employed to encode causal loci into deep features. Kernel machine regression is incorporated to enable statistical tests of association via SKAT framework. Multiple instance learning selects the most suspicious pathway at individual level. Parameters of the whole framework are optimized in an end-to-end manner through backpropagation.

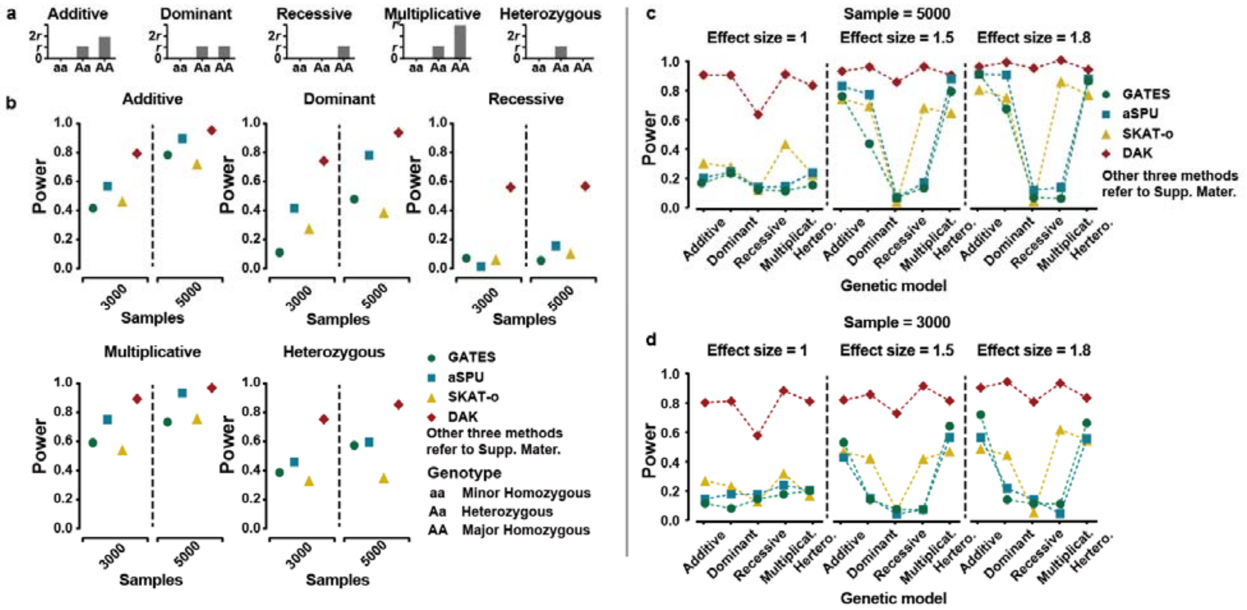


Figure 2. Disease risk levels for different genotypes in five genetic models. (**a**) Performances to discover the disease pathway resulted from single common variant. Effect size was set to 0.2 and simulated phenotypes were generated under five effect models. Under each sample size (3,000, 5,000), seven methods (four showed here and three in **Supplementary Figures**) were used to discover the disease pathway. Power was calculated from 100 replicates after Bonferroni correction. (**b**) Performances to discover the disease pathway resulted from single rare variant. Effect size was set to 1, 1.5 and 1.8, respectively to simulate phenotypes. 5,000 samples were considered. (**c**) Performances of DAK to discover the disease pathway resulted from single rare variant under small sample size (3,000).

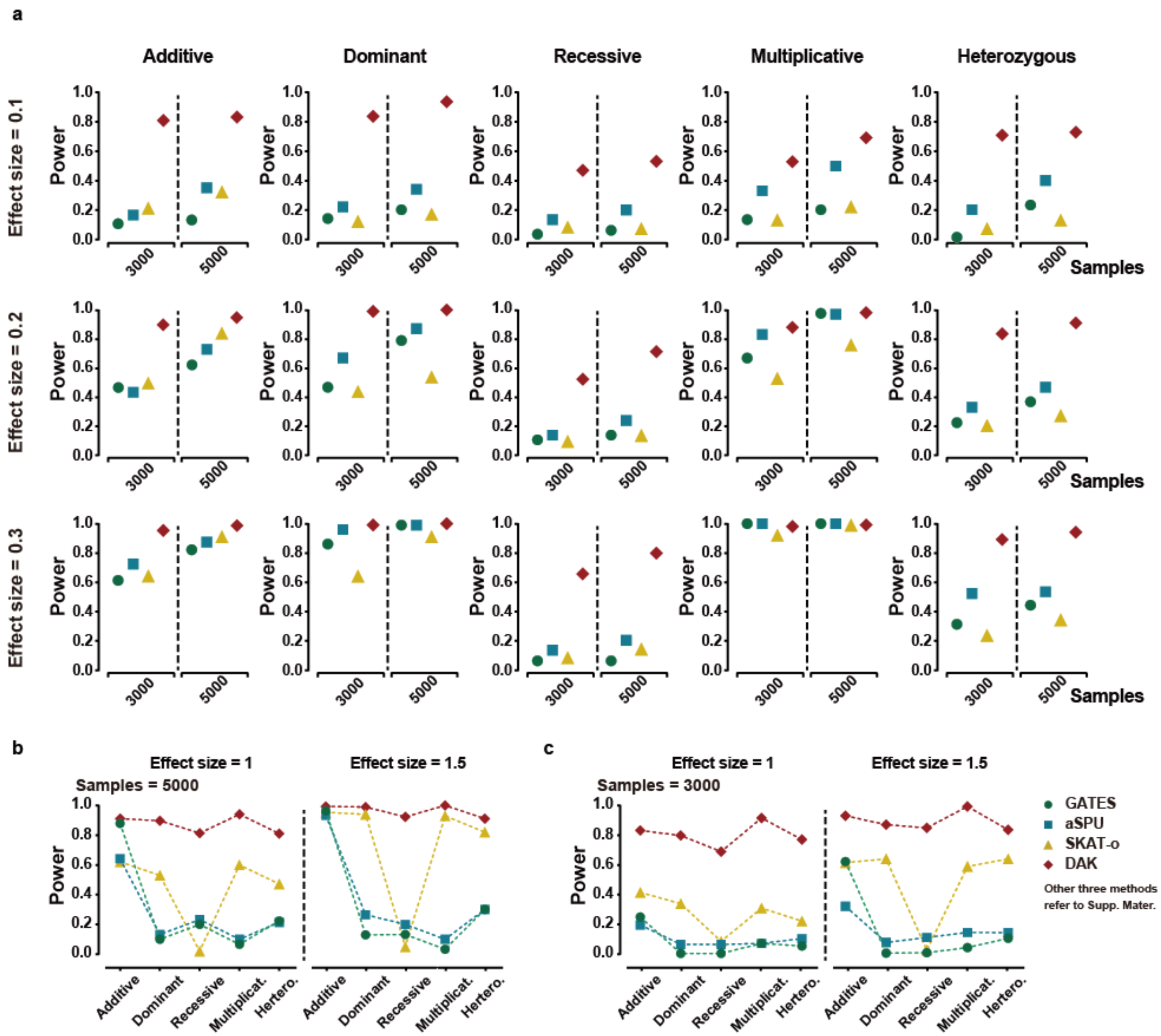
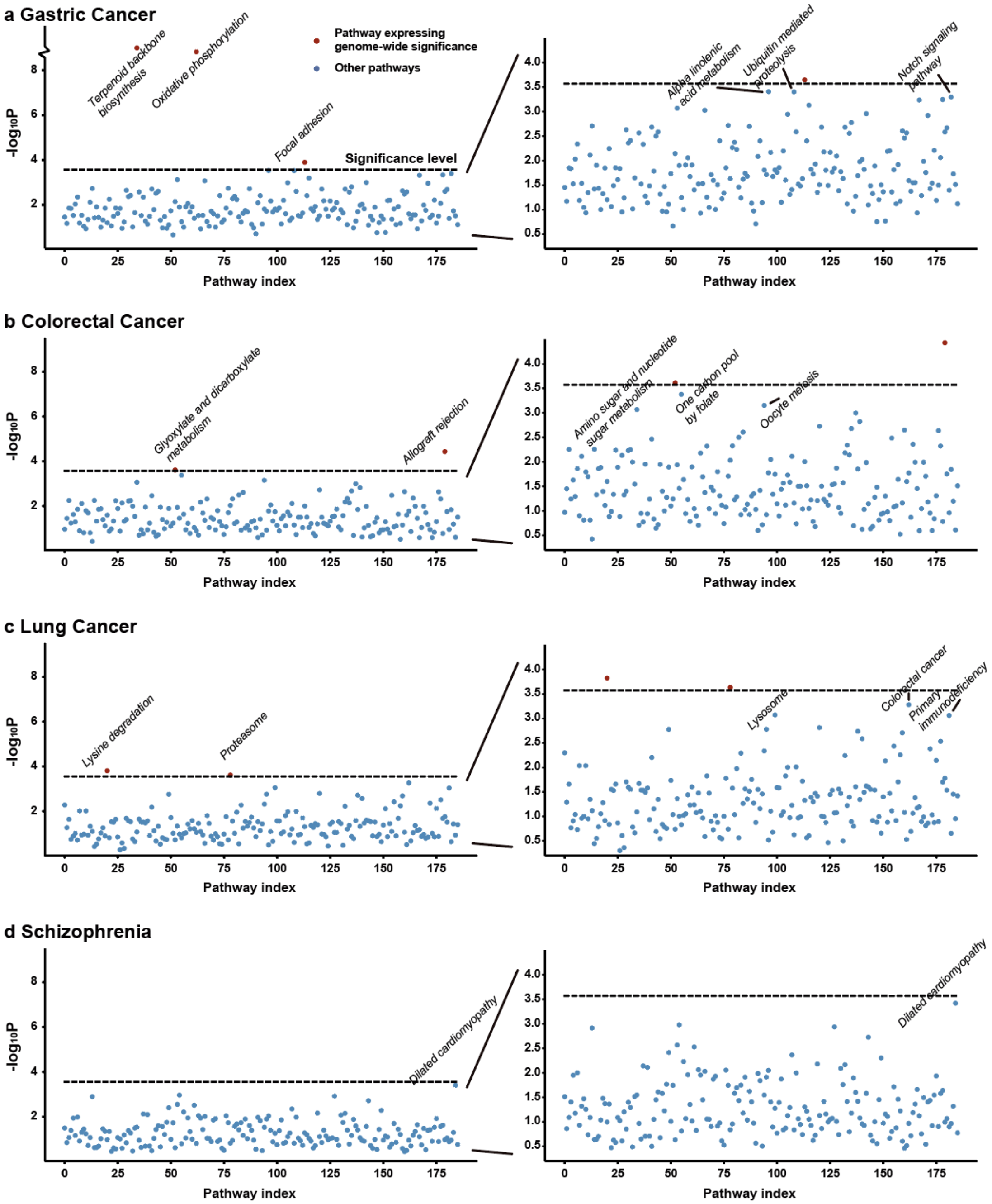


Figure 3. (**a**) Performances to discover the disease pathway resulted from three common variant. Effect size was set to 0.1, 0.2, 0.3 and simulated phenotypes were generated under five effect models. Under each sample size (3,000, 5,000), seven methods (four illustrated here) were used to discover the disease pathway. The power was calculated from 100 repeats after Bonferroni correction. (**b**) Performances to discover the disease pathway resulted from three rare variant. Effect size was set to 1, 1.5 and only 5,000 samples were considered. (**c**) Performances of DAK to discover the disease pathway resulted from three rare variant when only a small sample size (3,000) is available.



a Gastric Cancer

b Colorectal Cancer

c Lung Cancer

d Schizophrenia

Figure 4. Scatter plots of P-values of KEGG pathways by DAK on four real datasets. Pathways showing genome-wide significances after Bonferroni correction (= 0.05/186 = 2.68E-4) were marked in red.