

Powerful gene-based testing by integrating long-range chromatin interactions and knockoff genotypes

Shiyang Ma, James Dalgleish, Justin Lee, Chen Wang, Linxi Liu, Richard Gill, Joseph Buxbaum, Wendy Chung, Hugues Aschard, Edwin Silverman, et al.

▶ **To cite this version:**

Shiyang Ma, James Dalgleish, Justin Lee, Chen Wang, Linxi Liu, et al.. Powerful gene-based testing by integrating long-range chromatin interactions and knockoff genotypes. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118 (47), pp.e2105191118. 10.1073/pnas.2105191118 . pasteur-03526216

HAL Id: pasteur-03526216

https://pasteur.hal.science/pasteur-03526216

Submitted on 14 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Powerful gene-based testing by integrating long-range chromatin interactions and knockoff genotypes

Shiyang Ma[a], James Dalgleish[a], Justin Lee[b], Chen Wang[a], Linxi Liu[c], Richard Gill[d,e], Joseph D. Buxbaum[f,g,h], Wendy K. Chung[i,j], Hugues Aschard[k], Edwin K. Silverman[l,m], Michael H. Cho[l,m], Zihuai He[b,n], and Iuliana Ionita-Laza[a,1]

[a]Department of Biostatistics, Columbia University, New York, NY 10032; [b]Quantitative Sciences Unit, Department of Medicine, Stanford University, Stanford, CA 94305; [c]Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260; [d]Department of Human Genetics, Genentech, South San Francisco, CA 94080; [e]Department of Epidemiology, Columbia University, New York, NY 10032; [f]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [g]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [h]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [i]Department of Pediatrics, Columbia University, New York, NY 10032; [j]Department of Medicine, Columbia University, New York, NY 10032; [k]Department of Computational Biology, Institut Pasteur, 75015 Paris, France; [l]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115; [m]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115; and [n]Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA 94305

Edited by Michael P. Epstein, Emory University School of Medicine, and accepted by the Editorial Board October 7, 2021 (received for review March 17, 2021)

Gene-based tests are valuable techniques for identifying genetic factors in complex traits. Here, we propose a gene-based testing framework that incorporates data on long-range chromatin interactions, several recent technical advances for region-based tests, and leverages the knockoff framework for synthetic genotype generation for improved gene discovery. Through simulations and applications to genome-wide association studies (GWAS) and whole-genome sequencing data for multiple diseases and traits, we show that the proposed test increases the power over state-of-the-art gene-based tests in the literature, identifies genes that replicate in larger studies, and can provide a more narrow focus on the possible causal genes at a locus by reducing the confounding effect of linkage disequilibrium. Furthermore, our results show that incorporating genetic variation in distal regulatory elements tends to improve power over conventional tests. Results for UK Biobank and BioBank Japan traits are also available in a publicly accessible database that allows researchers to query gene-based results in an easy fashion.

gene-based association tests | long-range chromatin interactions | knockoff statistics | GWAS and whole-genome sequencing | fine-mapping

Gene-based association tests are commonly used to identify genetic factors in complex traits. Relative to individual variant or window-based tests, they have appealing features, including improved functional interpretation and potentially higher power due to lower penalty for multiple testing. Due to the recent advances in massively parallel sequencing technologies, a large number of gene-based tests have been proposed in the literature to test for association with genetic variation identified in sequencing studies (1–6). One important limitation of the current gene-based tests is that they often fail to incorporate the epigenetic context in noncoding regions. Moreover, how to best analyze the noncoding part of the genome to increase power remains unclear. Recently, several sliding window approaches have been proposed to scan the genome with flexible window sizes and appropriate adjustments for multiple testing, while accounting for correlations among test statistics (7, 8). However, these approaches are essentially scanning the genome in a one-dimensional (1D) fashion and fail to take into account the three-dimensional (3D) structure of the genome (9). Furthermore, because they scan the genome agnostically, the burden of multiple testing is high, which may lead to low power to identify true associations. These 1D approaches also suffer from interpretability issues similar to genome-wide association studies (GWAS) and therefore require follow-up analyses to be performed in order to identify the target genes. Several existing tests, such as multimarker analysis of genomic annotation (MAGMA), high-throughput chromosome conformation capture (Hi-C)-coupled MAGMA (H-MAGMA), and an omnibus test in the variant-set test for association using annotation information framework (STAAR-O) (10–12), attempt to link variants to their cognate genes based on physical proximity or chromatin-interaction data. We will compare our proposed tests to these existing approaches both conceptually and empirically, and we will show that our tests are more flexible and powerful than these existing tests. Furthermore, when individual-level data are available, the proposed tests can produce a more narrow list of associated genes at a locus by reducing the confounding effect of linkage disequilibrium (LD), a unique aspect of our gene-based test.

A related and popular gene-based strategy is the transcriptome-wide association studies (TWASs) (13, 14) that use GWAS data for a specific trait combined with genetic-variation gene-expression repositories, such as GTEx (15), to perform gene-based association tests. However, TWASs are limited to expression quantitative trait loci (eQTLs) being present in the reference datasets, and the majority of genetic associations

Significance

Gene-based tests are important tools for elucidating the genetic basis of complex traits. Despite substantial recent efforts in this direction, the existing tests are still limited, owing to low power and detection of false-positive signals due to the confounding effects of linkage disequilibrium. In this paper, we describe a gene-based test that attempts to address these limitations by incorporating data on long-range chromatin interactions, several recent technical advances for region-based testing, and the knockoff framework for synthetic genotype generation. Through extensive simulations and applications to multiple diseases and traits, we show that the proposed test increases the power over state-of-the-art gene-based tests and provides a narrower focus on the possible causal genes involved at a locus.

Author contributions: S.M. and I.I.-L. designed research; S.M. and I.I.-L. performed research; J.D., L.L., R.G., J.D.B., E.K.S., M.H.C., and Z.H. contributed new reagents/analytic tools; S.M., J.D., J.L., C.W., J.D.B., W.K.C., H.A., E.K.S., M.H.C., Z.H., and I.I.-L. analyzed data; and S.M. and I.I.-L. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. M.P.E. is a guest editor invited by the Editorial Board.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: ii2135@columbia.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2105191118/-/DCSupplemental.

Published November 19, 2021.

Downloaded at Institut Pasteur - CeRIS on January 14, 2022

cannot be clearly assigned to existing eQTLs (16, 17). Therefore, they may have reduced power to identify the relevant genes for the trait of interest.

Regulatory elements, including enhancers and promoters, play an important role in controlling when, where, and to what degree genes will be expressed. Most of the disease-associated variants in GWAS lie in noncoding regions of the genome, and it is believed that a majority of causal noncoding variants reside in enhancers (18). However, identifying enhancers and linking them to the genes they regulate is challenging. A number of methods have emerged in recent years to identify promoter–enhancer interactions. These techniques range from chromatin conformation capture (3C), which is limited to the detection of a single interaction, to circular chromosome conformation capture (which can detect all loci that interact with a single locus), to many-to-many mapping technologies possible using targeted enrichment. Hi-C maps the complete DNA interactome and elucidates the spatial organization of the human genome (19–21). Hi-C provides direct physical evidence of interactions that may mediate gene-regulatory relationships and can aid in identifying putative regulatory elements for a gene of interest. However, due to the prohibitive sequencing costs of the Hi-C experimental technique, it is challenging to obtain high-resolution (e.g., 1 Kb) Hi-C data in a large number of cell types and tissues at multiple developmental times.

We propose here comprehensive gene-based association tests for common and rare genetic variation in both coding and noncoding regions, putative regulatory elements, and which incorporate several recent advances for region-based tests, including 1) scanning the genic and regulatory regions with varied window sizes; 2) the aggregated Cauchy association test (ACAT) to combine *P* values from single-variant, burden, and dispersion (sequence kernel association test [SKAT]) tests; 3) incorporation of multiple functional annotations; and 4) the saddlepoint approximation for unbalanced case-control data (22–25). To further improve the power and the ability to prioritize putative causal genes at significant loci when individual-level data are available, we leverage a recent development in statistics, namely, the knockoff framework for knockoff genotype generation (26) that helps control the false discovery rate (FDR) under arbitrary correlation structure and attenuates the confounding effect of LD. One can think of the knockoff genotypes as synthetic, noisy copies of the original genotypes, which resemble the original data in terms of LD structure, but are conditionally independent of the trait of interest, given the original genotypes. Although conventional methods, such as the Benjamini–Hochberg (BH) procedure, are also designed to control the FDR (27), they cannot guarantee FDR control at the target level with arbitrarily correlated *P* values. Furthermore, unlike the knockoff framework implemented here, the conventional methods do not naturally account for correlations due to LD. The proposed gene-based test is related to a recently proposed window-based test, KnockoffScreen (8). Specifically, we employ the knockoff generation algorithm for genotype data that we have introduced in KnockoffScreen (8) and develop knockoff-based inference for gene-based tests. We demonstrate below that the proposed test has important advantages compared with the window-based test KnockoffScreen in terms of controlling the FDR at gene level. While KnockoffScreen can identify significant windows with valid FDR control at window level, functional interpretation of significant windows is still needed, which means that post hoc analyses need to be done to link those windows to relevant genes. However, as we show in simulations, this procedure can lead to highly inflated FDR at gene level.

We evaluate the performance relative to existing methods using simulations and applications to multiple studies, including GWAS studies for neuropsychiatric and neurodegenerative diseases, whole-genome sequencing studies for Alzheimer's disease (AD) from the Alzheimer's Disease Sequencing Project (ADSP), and for lung function from the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Program. We also provide results of applications to UK Biobank and BioBank Japan binary and continuous traits.

Results

Overview of the Proposed Gene-Based Association Tests. We provide here a brief overview of the proposed gene-based tests that aim to comprehensively evaluate the effects of common and rare, coding, and proximal and distal regulatory variation on a trait of interest. A workflow depicting the overall gene-based testing approach proposed here is shown in Fig. 1. Briefly, we build our final test, GeneScan3DKnock, progressively, starting with a test focused on scanning the gene body region (i.e., the interval between the transcription start site [TSS] and the end of the 3′ untranslated region [UTR]) with varied window sizes. We refer to this test as GeneScan1D. We extend this test by incorporating genetic variants residing in putative regulatory elements, such as promoters and enhancers. In particular, we use chromatin immunoprecipitation sequencing (ChIP-seq) peak data extracted from the ChIP-Atlas database to identify promoter regions and data from the GeneHancer database to link enhancers to their target genes (28). We also use the activity-by-contact (ABC) model to predict functional enhancer–gene connections for five cell types and tissues (29). This is the GeneScan3D test. Finally, when individual-level data are available, we implement the knockoff framework for a more powerful gene-discovery and fine-mapping approach and refer to this test as GeneScan3DKnock.

We take advantage of recent advances in region-based tests for sequencing data (4, 22, 24) to perform computationally efficient and comprehensive tests with genetic variation in a gene (including variants located in proximal and distal regulatory elements), while scanning the gene with a range of window sizes for improved power. The framework allows for the incorporation of a variety of functional genomics annotations as weights for individual variants included in the tests. Furthermore, an aspect

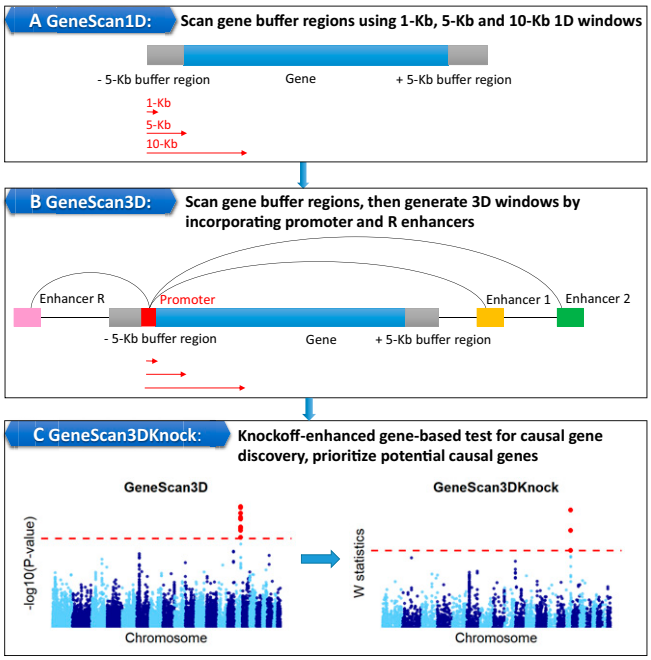


Fig. 1. Workflow of the proposed gene-based tests. (*A*) GeneScan1D, a 1D scan of the gene and buffer region. (*B*) GeneScan3D, a 3D scan of the gene and regulatory elements linked to it. (*C*) GeneScan3DKnock, the knockoff-enhanced test, implementing a knockoff-based version of GeneScan3D.

Downloaded at Institut Pasteur - CeRIS on January 14, 2022

of our testing framework is the derivation of knockoff statistics based on the generation of knockoff (synthetic) genetic data that resemble the original genotypes in terms of correlation structure, but are conditionally independent of the outcome variable given the true genotypes (8, 26, 30). The knockoff genotypes are essentially noisy copies of the original genotypes and serve as negative controls for the original genotype data; they help to select important genes while controlling the FDR. GeneScan3DKnock computes for each gene a knockoff statistic W that measures the importance of each gene (similar to a P value) and then uses the knockoff filter to detect genes that are significant at a specified FDR target level (26). We also compute a q value for each gene. A q value is similar to a P value, except that it measures significance in terms of FDR rather than family-wise error rate (FWER) and already incorporates correction for multiple testing. The knockoff version of the gene-based test has unique features relative to the standard gene-based tests, including improved power and the ability to prioritize causal genes over associations due to LD. The details on these specific tests can be found in *Materials and Methods*.

We compare with the nearest competitor gene-based tests in the literature, namely, MAGMA/H-MAGMA (10, 11), TWAS/FUSION (14), and STAAR-O (12). We also show comparisons with the recently proposed window-based test KnockoffScreen (8).

Simulation Studies.

Power and type I error-rate evaluation for a single gene. We conducted simulation studies in order to 1) examine the type I error rates of the proposed tests, GeneScan1D and GeneScan3D, under different significance levels; and (2) evaluate the potential power gain by incorporating the regulatory elements. For power comparisons, we considered the nearest competitor gene-based tests, MAGMA/H-MAGMA and STAAR-O.

For type I error-rate simulations, we used real Genetic Epidemiology of COPD (COPDGene) TOPMed whole-genome sequencing data, with $n = 5{,}593$ for a continuous trait and $n = 4{,}450$ for a binary trait. We randomly selected 10 genes (average gene length 25 Kb), and for each selected gene, we randomly selected $R = 2$ GeneHancer and ABC enhancers (average enhancer length 1.35 Kb). For each selected gene and the corresponding enhancers, we used the real genotype data, while the phenotype data are simulated as below:

- For a continuous trait: $Y_i = Z_i + \epsilon_i$,
- For a binary trait: $\text{logit}(P(Y_i = 1|\ Z_i)) = \alpha_0 + Z_i$,

where $Z_i \sim N(0, 1)$ is a covariate and $\epsilon_i \sim N(0, 1)$ is the standard normal error; Z_i and ϵ_i are independent. For the binary trait, an equal number of cases and controls were simulated. For GeneScan1D and GeneScan3D, we used two window sizes, 1 Kb and 5 Kb, to scan the gene region. All variants and common variants only were considered in the type I error-rate simulation studies.

To evaluate power and compare with existing tests such as MAGMA/H-MAGMA and STAAR-O, we used the same whole-genome sequencing data. We randomly selected 10 genes (average length 25 Kb), and, for each selected gene, we randomly selected $R = 10$ GeneHancer and ABC enhancers (average enhancer length 1.87 Kb). Power was computed for each gene separately, and the average over the 10 genes was reported. We made use of the real genotypes for the selected genes plus and minus a 5-Kb buffer region and for the corresponding enhancers. For each gene, the phenotype data were generated as follows:

- For a continuous trait: $Y_i = \beta_1 G_{i1} + \cdots \beta_s G_{is} + Z_i + \epsilon_i$,
- For a binary trait: $\text{logit}(P(Y_i = 1|\ \ Z_i, \mathbf{G_i})) = \alpha_0 + Z_i + \beta_1 G_{i1} + \cdots \beta_s G_{is}$,

where G_{ij} denotes the genotypes of randomly selected causal variants and β_j 's are the corresponding effect sizes. For the binary trait, an equal number of cases and controls were simulated. We set 2% of the variants in the gene and buffer region to be causal, all within a 2-Kb signal window. For each enhancer, we set 2% (uniformly distributed) variants to be causal. The effect size of the causal variant j was assumed to be $\beta_j = c|\log_{10}\text{MAF}_j|$, where MAF is the minor allele frequency. We assumed $c = 0.25$ for the continuous trait and $c = 0.6$ (e.g., OR $= 6.05$, when MAF $= 0.001$) for the binary trait.

For GeneScan1D and GeneScan3D, we used three window sizes for scanning: 1, 5, and 10 Kb. We applied MAGMA on the gene plus and minus the 5-Kb buffer region. For GeneScan3D and H-MAGMA, we incorporated $R = \{2,\ 5,\ 10\}$ enhancers. We also conducted STAAR-O gene-centric analyses on 1) the entire gene body and 2) the same $R = \{2,\ 5,\ 10\}$ enhancers, and then we combined the STAAR-O P values for these elements using the Cauchy combination method. As detailed in *SI Appendix*, to allow for fair comparisons, for STAAR-O we used the same weighting and MAF/minor allele count (MAC) thresholds, as used for the proposed tests. For the sake of completeness, we also ran the default setting of STAAR-O gene-centric analyses focused on rare variants. Finally, we adjusted for 10 principal components of ancestry.

Type I error rate. We conducted 10^7 replications to examine the empirical type I error rate under both continuous and binary traits (*SI Appendix*, Table S1). For continuous traits, the type I error rates were well controlled in all analyses under moderate significance levels 10^{-3}, 10^{-4}, and 10^{-5}. Even for a stringent significance level of 2.5×10^{-6}, the type I error rates fell within the 95% CI: $(1.52 \times 10^{-6}, 3.48 \times 10^{-6})$. For binary phenotypes, the type I error rates were slightly conservative at different levels.

Power. We evaluated the empirical power at significance level 2.5×10^{-6} using 10,000 replications (Fig. 2*A* and *SI Appendix*, Fig. S1*A*). As shown, GeneScan3D and STAAR-O have important power advantages relative to H-MAGMA, likely due to their better tolerance of noisy variation, as also demonstrated below. GeneScan3D also exhibited higher power than STAAR-O, likely due to the sliding window scanning implemented in GeneScan3D. The 3D tests overall tended to be more powerful than the 1D tests, with the relative benefits diminishing as the number of signal enhancers decreased. STAAR-O with the default settings (focused on rare variants only) had lower-power performance (*SI Appendix*, Fig. S2*A*), as expected, given that our simulations included common causal variants, in addition to rare causal variants.

Robustness to noisy enhancers. When performing the 3D analyses, it is likely that some of the putative regulatory elements do not contain any signal variants. We conducted additional power simulation studies to evaluate the performance when only $R = \{2,\ 5\}$ enhancers of a total of 10 enhancers for a gene contained any signal variants. We compared with the power of the oracle approach, i.e., when only the signal-containing enhancers were included in the analyses (Fig. 2*A* and *SI Appendix*, Figs. S1*A* and S2*A*). GeneScan3D and STAAR-O exhibited negligible power loss, suggesting that they are robust to inclusion of noisy enhancers, unlike H-MAGMA, which is less robust in such realistic settings. This empirical observation is consistent with the theoretical expectation: While GeneScan3D/STAAR-O combined signal from individual enhancers using the Cauchy P value combination method and, hence, are expected to maintain strong power in the presence of noisy enhancers, H-MAGMA is based on a principal component regression approach and, hence, combines genetic variants across multiple enhancers, rendering it less robust in the presence of noisy enhancers.

Downloaded at Institut Pasteur - CeRIS on January 14, 2022

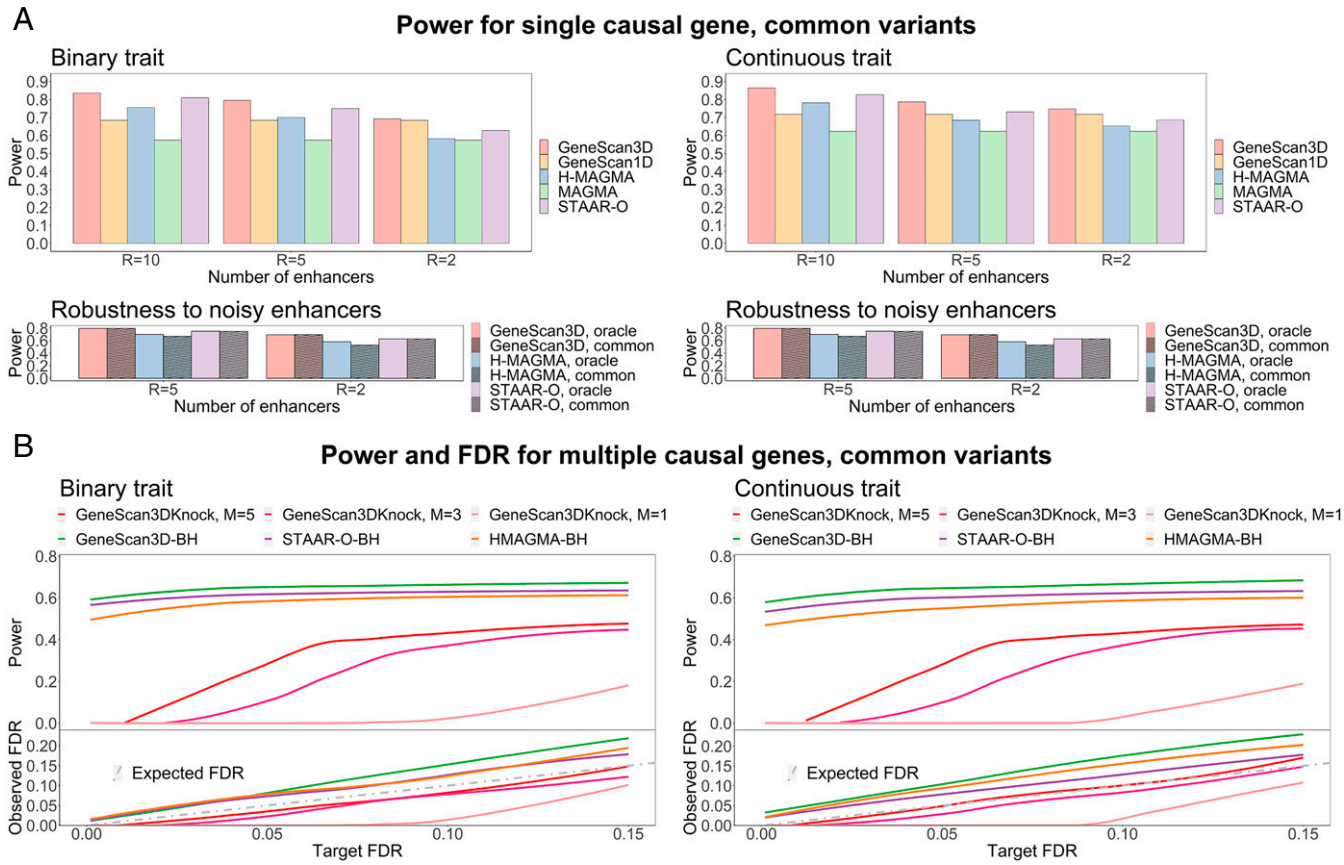


Fig. 2. Power and FDR of the proposed gene-based tests and binary and continuous traits with tests including only common variants. (*A*) Power and robustness to noisy enhancers. *A*, *Upper* shows power for the GeneScan3D, GeneScan1D, H-MAGMA, MAGMA, and STAAR-O tests. The number of enhancers (*R*) ranges from 2 to 10. *A*, *Lower* shows power for the GeneScan3D, H-MAGMA, and STAAR-O tests, assuming causal variants in $R = \{2,\ 5\}$ causal enhancers. Power is compared between using only the $R = \{2,\ 5\}$ causal enhancers (the oracle approach) vs. using all 10 enhancers (including noisy enhancers). (*B*) Power and FDR for GeneScan3DKnock using different numbers of knockoffs and the BH procedure for GeneScan3D, STAAR-O, and H-MAGMA.

Power and FDR evaluation with multiple causal genes. One aspect of the proposed knockoff-based test is that it allows for selecting significant genes by controlling the FDR in the presence of complex correlations due to LD. For the knockoff-based test, GeneScan3DKnock, we evaluated the empirical FDR and power, assuming multiple causal and noise genes. We randomly selected 10 causal genes and 250 noise genes (gene length 10 to 100 Kb, average length 39 Kb) as follows. Among the noise genes, some were selected to be physically close to the causal genes, i.e., within ± 2-Mb region, and others were randomly selected across the genome. For each gene, we only included the corresponding GeneHancer and ABC enhancers that fell within a 150-Kb region (± 75 Kb from the gene midpoint). This restriction to a 150-Kb region was done for computational reasons and only for these power simulations. On average, there were 10 enhancers for each gene, with an average length of 1.25 Kb. We generated multiple knockoff genotypes for 250-Kb regions spanning each gene (± 50 Kb on either side of the 150-Kb region), as detailed in *Materials and Methods*. Note that to avoid enhancer sharing across genes and too-strong LD among causal and noise genes (which leads to false discoveries for all the statistical tests considered here), we selected the genes such that the corresponding 150-Kb regions were disjoint.

For each replicate, we randomly selected a 10-Kb causal window in each causal gene ±5-Kb buffer region and set 3.5% variants in the window to be causal. We also set 3.5% variants in all enhancers to be causal. We generated the continuous/binary traits using the selected causal variants as follows:

- For a continuous trait: $Y_i = Z_i + \beta_1 G_{i1} + \cdots + \beta_s G_{is} + \epsilon_i$,
- For a binary trait: $\text{logit}(P(Y_i = 1|\ \ Z_i, \mathbf{G_i})) = \alpha_0 + Z_i + \beta_1 G_{i1} + \cdots + \beta_s G_{is}$.

As above, $Z_i \sim N(0,1)$ is a covariate, and $\epsilon_i \sim N(0,1)$ is the standard normal error; Z_i and ϵ_i are independent; α_0 is chosen such that the prevalence is 10%. Again, we set the effect size $\beta_j = c|\log_{10}\text{MAF}_j|$ for the j-th causal variant, with $c = 0.2$ for the continuous trait and $c = 0.6$ for the binary trait.

The empirical power and FDR for GeneScan3DKnock were averaged over 100 replicates. We present results for single knockoffs, as well as multiple knockoffs ($M = 3$ and 5). We calculated the original and knockoff P values from the GeneScan3D test (for all variants and common variants), adjusting for 10 principal components of ancestry. We computed q values for 10 causal and 250 noise genes in order to identify significant genes using the GeneScan3DKnock test at different target FDR levels, up to 0.15. The empirical power was defined as the proportion of causal genes being identified; the empirical FDR was defined as the proportion of detected genes that are noise. We show that GeneScan3DKnock can control the FDR at the target level and that using multiple knockoffs can improve power substantially, especially at lower levels of target FDR, where the single-knockoff approach has very low power, as expected (Fig. 2*B* and *SI Appendix*, Fig. S1*B*).

For comparisons, we evaluated the empirical power and FDR for competitor methods, including STAAR-O and H-MAGMA, using the standard BH procedure for FDR control.

Downloaded at Institut Pasteur - CeRIS on January 14, 2022

A gene is significant if the corresponding q value is $\leq$ the target FDR level. The results show that the conventional BH procedure may not control the FDR at the target level, and, therefore, the proposed knockoff-based approach provides a valid alternative when FDR control is desirable, such as for polygenic traits with multiple underlying causal genes (Fig. 2*B* and *SI Appendix*, Figs. S1*B* and S2*B*).

Although our main comparisons are with gene-based tests, we performed additional comparisons with the recently proposed window-based method KnockoffScreen (8) in order to illustrate the need for a gene-based knockoff filter when our interest is controlling the FDR at the gene level. We applied KnockoffScreen by scanning each 150-Kb region using several window sizes (1, 5, and 10 Kb), and we computed the empirical power and FDR of KnockoffScreen at both gene level and window level (*SI Appendix*). Although KnockoffScreen can control the window-level FDR, as shown in *SI Appendix*, Fig. S3, the empirical gene-level FDR can be quite high, suggesting that the proposed framework designed to control the FDR at the gene level is more appropriate for gene discovery. Essentially, as a window-based test, KnockoffScreen leads to a larger number of rejections, i.e., higher power, but also higher FDR at gene level (*SI Appendix*, Fig. S3).

Applications to Individual-Level Data from Whole-Genome Sequencing Studies.

AD. We present results from an application to whole-genome sequencing data from the ADSP. The data include 3,085 whole genomes from the ADSP Discovery Extension Study and 809 whole genomes from the Alzheimer's Disease Neuroimaging Initiative, for a total of 3,894 whole genomes (more details are available in *SI Appendix*). We adjusted for age, age^2, gender, ethnic group, sequencing center, and the leading 10 principal components of ancestry. Seven tissue/cell-type specific GenoNet functional scores (31) related to brain were incorporated, including E071 (brain hippocampus [HIP] middle), E074 (brain substantia nigra), E073 (brain dorsolateral prefrontal cortex [DLPFC]), E068 (brain anterior caudate), E067 (brain angular gyrus), E069 (brain cingulate gyrus), and E072 (brain inferior temporal lobe).

We show results for several tests, including GeneScan1D and GeneScan3D tests; the proposed knockoff-based approach, GeneScan3DKnock, based on five random knockoffs; as well as existing tests, including MAGMA/H-MAGMA, STAAR-O, and TWAS. We do not include the results from KnockoffScreen since we have shown in the simulations that at the gene level, it can have inflated FDR. For all the tests except for the knockoff-based GeneScan3DKnock test, we identified significant genes using the Bonferroni method for FWER control since, as shown in the simulations, the conventional BH procedure does not control the FDR at the target level. For GeneScan3DKnock, we used the implemented knockoff filter procedure to identify significant genes at an FDR threshold of 10%. We present results for common variants only (those with MAF $> 1/\sqrt{2n}$, where n is the sample size) and all variants (rare variant-only analyses are not well powered at these sample sizes).

Overall, all tests considered identify the well-known signal at the apolipoprotein E (APOE) locus (Fig. 3 and *SI Appendix*, Table S2 and Fig. S4). The GeneScan3D, H-MAGMA, and STAAR-O tests detected additional significant genes on chromosome 19 (chr19), mostly due to signals residing in the promoters and/or enhancers overlapping genes at the *APOE* locus (Fig. 4). These results suggest that the APOE region is a central nucleating point for loops that regulate expression of potentially AD-associated genes. Therefore, it is possible that the strong signal observed at the *APOE* locus can be linked to genes that are farther away (Fig. 4 and *SI Appendix*, Table S3).

GeneScan3DKnock has improved power and reduces false-positive associations relative to existing tests. False-positive signals can arise due to possible coregulation of multiple genes by the same "causal " enhancers, or simply due to LD among causal and noncausal variants in genes or associated regulatory elements. The knockoff-based test GeneScan3DKnock cannot help eliminate false positives due to coregulation, but can attenuate the effect of LD-induced confounding. We computed the knockoff statistic W and the q value for each gene. A scatterplot of genome-wide W knockoff statistics vs. $-\log10(P$ values) based on the GeneScan3D test illustrates how almost half of the significant genes at the *APOE* locus based on GeneScan3D (using the conventional FWER control) are no longer significant in the GeneScan3DKnock test, despite the less stringent FDR control (Fig. 3 and *SI Appendix*, Fig. S4 and Tables S2 and S4). Similarly, GeneScan3DKnock identified a lower number of significant genes relative to STAAR-O with stringent FWER control. These include a large number of genes linked to GH19F044889 and several overlapping ABC enhancers, which contain variants in high LD with variants in the *APOE* gene (Fig. 4). Indeed, we obtained a narrower list of significant genes on chr19 related to the *APOE* locus, which includes the main genes from the 1D tests, i.e., *APOE, TOMM40, APOC1*, and *NECTIN2*, but other interesting genes as well, including *BCAM*, *RELB*, and *QPCTL*. For example, rare variants in *BCAM* and *RELB* have recently been identified to be associated with AD and neuroimaging biomarkers of AD after adjusting for APOE genotypes (32, 33). *QPCT*, an important paralog of *QPCTL*, has been shown to be involved in AD pathogenesis and cognitive decline by glutaminyl cyclase-catalyzed pGlu-Aβ formation (34). This ability to remove a substantial proportion of false-positive signals due to LD is a unique and appealing feature of the proposed GeneScan3DKnock test.

Interestingly, GeneScan3DKnock detected several associations outside chr19 that were missed by the competitor gene-based tests. These include *NECTIN1*, *ZNF843*, *ZNF646*, and *PPP1R17* (for common variants) and *HIPK3* (for all variants), which were previously found to be involved in AD-related pathophysiology. Nectin-1 is a member of the immunoglobulin superfamily and a Ca(2+)-independent adherens junction protein involved in synapse formation (35). The important role of nectin in synaptic development and maintenance can explain how genetic variation in *NECTIN1* can perturb synaptic activity and play a role in AD. *ZNF646* lies within the *KAT8* locus, recently identified in two large-scale GWAS studies focused on clinically diagnosed AD and AD-by-proxy individuals (36, 37). Furthermore, *ZNF646* was prioritized at the *KAT8* locus based on high posterior probability for the colocalization between AD GWAS single-nucleotide polymorphisms (SNPs) at the *KAT8* locus and eQTLs from both brain (DLPFC) and microglia (38). Similarly, *PPP1R17* was found to be significantly underexpressed in the brains of 14-mo-old Sgo1$-/+$ mice (a murine AD model of chromosome instability with chromosomal and centrosomal cohesinopathy) compared with age-matched wild-type animals (39). The protein encoded by this gene is found primarily in Purkinje cell bodies and projections in the cerebellum and subsets of neurons in the hypothalamus. An SNP located in the promoter of *PPP1R17* was previously found to be associated with hypercholesterolemia (40). Finally, homeodomain-interacting protein kinase 3 (HIPK3) belongs to a group of HIPKs, including HIPK2, which is down-regulated by elevated amount of Amyloid β (Aβ), a hallmark of AD (41). Protein HIPK3 levels were also found to be significantly different between individuals with mild cognitive impairment that converted to AD vs. the nonconverters (42).

Replication of significant genes using summary statistics from a large meta-analysis AD study. To provide more objective evidence

Downloaded at Institut Pasteur - CeRIS on January 14, 2022

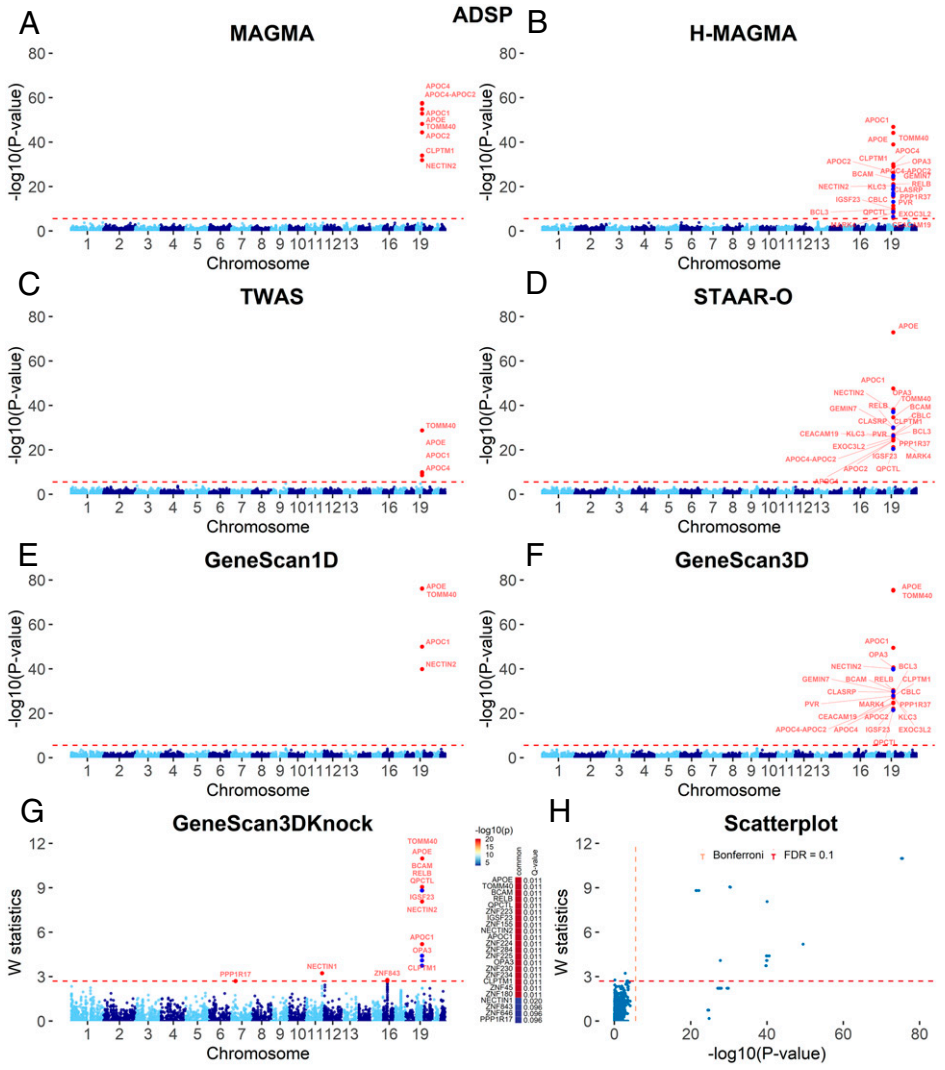


Fig. 3. Applications to ADSP whole-genome sequencing data, common variants only. (*A–F*) Manhattan plots of MAGMA, H-MAGMA, TWAS, STAAR-O, GeneScan1D, and GeneScan3D results, respectively. (*G*) Manhattan plot of GeneScan3DKnock results. Genes within the zinc-finger-containing (*ZNF*) gene cluster on chr19 are unlabeled and shown in blue in H-MAGMA, GeneScan3D, and GeneScan3DKnock analyses for clear visualization. G, *Right* shows a heatmap with *P* values of the GeneScan3D test (truncated at 10^{-20}) for all genes passing the FDR = 0.1 threshold and the corresponding *q* values that already incorporate correction for multiple testing. The genes are shown in descending order of the knockoff statistics. (*H*) Scatterplot of *W* knockoff statistics (GeneScan3DKnock) vs. $-\log_{10}$ (*P* value) (GeneScan3D) for common variants. Each dot represents a gene. The dashed lines show the significance threshold defined by Bonferroni correction (for *P* values) and the data-adaptive threshold for FDR control (for *W* statistic).

of replication, we leveraged a large meta-analysis of clinical AD and AD-by-proxy studies [71,880 AD or proxy cases and 383,378 controls (36)] and performed gene-based tests (GeneScan1D and GeneScan3D) using the available GWAS summary statistics for 58 significant genes identified for AD across all the different tests, including the BH-adjusted tests. We constructed 3D windows using the same procedure as before (*Materials and Methods*). For each 3D window, we then applied the ACAT procedure (22) to combine *P* values for single variants within the 3D window. Since we did not have access to individual-level data, we did not conduct Burden and SKAT tests for these replication studies. Results are shown in *SI Appendix*, Table S5. Note that most of the genes identified by GeneScan3DKnock at FDR 10%, including genes at the APOE locus, *ZNF646* and *ZNF843*, had a replication *P* value based on GeneScan3D $< 0.05/58 = 8.62 \times 10^{-4}$, while genes identified by conventional BH controlling procedures (*SI Appendix*, Figs. S5 and S6) failed to replicate for the most part, concordant with simulation studies showing that the BH procedure can result in inflated empirical FDR values, and, therefore, it is not a rigorous procedure to identify significant genes at a desired FDR level.

Lung function (FEV1). The COPDGene study includes chronic obstructive pulmonary disease (COPD) cases, controls, and additional smokers with varied lung function. In addition to COPD case/control status, lung-function measurements are also available, including forced expiratory volume in 1 s (FEV1), forced vital capacity (FVC), and their ratio (FEV1/FVC). We analyzed whole-genome sequencing data from the TOPMed freeze5b dataset, which includes a subset of 5,593 Non-Hispanic White individuals for continuous traits and 4,450 individuals for COPD case/control binary trait. We present results from the application to FEV1, adjusting for sequencing center, 10 principal components of ancestry, age, age^2, gender, height, height^2, smoking pack-years, and current smoking. We incorporated five tissue/cell-type specific GenoNet functional scores (31) related to lung, namely: E017 (IMR90), E088 (fetal lung), E096 (lung), E114 (A549), and E128 (normal human lung fibroblast).

As with the AD example, GeneScan3DKnock identified significant genes that were missed by the other tests. Specifically, we identified a cluster of significant genes on chromosome 12 that included *FRS2*, *CCT2*, *RAB3IP*, *LRRC10*, *and BEST3* and that was missed by all the other gene-based tests considered (Fig. 5). Notably, an intronic SNP (rs10444582) in *FRS2* was identified to be significantly associated with FEV1 in the UK Biobank and SpiroMeta ($P = 1.2 \times 10^{-10}$, $n = 396,723$) (43). This locus was not included in the final list of loci released by Shrine et al. (43), as the *P* value in the replication cohort (SpiroMeta) was only 3.5×10^{-3}, above the predetermined significance threshold. As COPDGene was not part of the Shrine et al. study (43), our findings on chromosome 12 provide additional, independent evidence for this signal. Another significant and potentially in-

Downloaded at Institut Pasteur - CeRIS on January 14, 2022

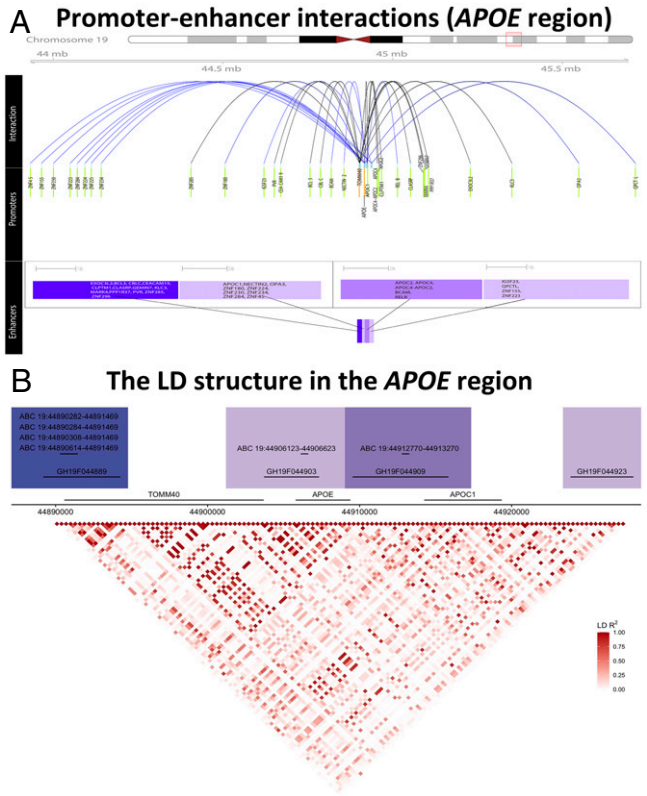


Fig. 4. Visualization of promoter–enhancer interactions of significant genes at the *APOE* locus. (*A*) The promoter–enhancer links are shown for the significant genes in the GeneScan3D analyses for common variants, where arcs in blue point to those genes identified by the knockoff procedure only. Genes with a signal enhancer are shown in green; those with no signal enhancer are in orange (*A*, *Middle*). The *APOE* locus is the location of a tight cluster of several enhancers (shown in purple in *A*, *Bottom*), with arcs connecting the enhancers to many different gene promoters. (*B*) The LD structure in the *APOE* region. The locations of the enhancers in the region are also shown.

teresting gene is *RAB7A*. A common SNP (rs9847178) residing in the promoter-flanking region of *RAB7A* has been found to be genome-wide significant in a recent large GWAS study on smoking (44). A nearby SNP (rs7650872) in the same promoter-flanking region has been found to be genome-wide significant with eosinophil counts in the UK Biobank (45). High eosinophil counts predict decline in FEV1 (46). Interestingly, a recent study has shown that loss of *RAB7A* confers resistance to SARS–CoV-2 by reducing ACE2 levels (47), concordant with reports in the literature of loci associated with susceptibility and/or response to infection that have been previously associated with lung-function phenotypes (48).

Replication of significant genes using summary statistics from the UK Biobank data. We performed similar replication studies for 40 significant genes identified for FEV1 across all the different tests, including the BH-adjusted tests, using 383,471 European individuals with FEV1 measurements in the UK Biobank (*SI Appendix*, Table S6). The covariates adjusted for in the analyses include 10 principal components of ancestry, age, age^2, gender, age· gender, and age^2· gender. Note that the number of available covariates in the UK Biobank is limited, and some important covariates for FEV1, such as height and smoking, are not adjusted for in these analyses. Despite this caveat, most of the genes identified as significantly associated with FEV1 in our COPDGene study replicated in the UK Biobank study (replication P value based on GeneScan3D $< 0.05/40 = 1.25 \times 10^{-3}$).

Applications to GWAS Summary Statistics.

GWAS for neuropsychiatric and neurodegenerative diseases. We applied the different gene-based tests to summary statistics from nine GWAS studies of brain disorders, including five neuropsychiatric traits: attention-deficit/hyperactivity disorder (ADHD) (49), autism spectrum disorder (ASD) (50), bipolar disorder (51), schizophrenia (52), and major depressive disorder (53); and four neurodegenerative traits: AD (36), Parkinson's disease (54), amyotrophic lateral sclerosis (ALS) (55), and multiple sclerosis (MS) (56). We do not include STAAR-O here since the current implementation is not applicable to summary statistics.

Since these applications focus on brain disorders, we leveraged two existing Hi-C human brain datasets for the DLPFC in adult brain (57) and for the germinal zone (GZ) and cortical and subcortical plate (CP) in fetal brain (58) and used the Fit-Hi-C method to identify statistically significant promoter–enhancer interactions from these data (59) (*SI Appendix*). Similarly, we applied MAGMA/H-MAGMA (11) to the same datasets using the same significant Hi-C interactions. We also applied TWAS/FUSION (14) based on 13 brain regions from GTEx version 7 (amygdala [AMY], anterior cingulate cortex, caudate, cerebellar hemisphere, cerebellum, cortex, frontal cortex, HIP, hypothalamus, nucleus accumbens, putamen, spinal cord, and substantia nigra). The Cauchy P value combination method was used to combine TWAS P values from different brain regions.

We used a liberal significance threshold (10^{-3}) to select genes from these analyses (because some of the GWAS studies, e.g., AD, ADHD, and MS, are underpowered) and investigated their expression patterns using spatiotemporal and single-cell transcriptomics data, as described below. The number of significant genes at this threshold and the overlap across the different tests are shown in *SI Appendix*, Table S7 and Fig. S11. Compared with 1D analyses (GeneScan1D and MAGMA), GeneScan3D and H-MAGMA detected a much larger number of disease-associated genes, as expected, given that they incorporate signals from distal regulatory elements. GeneScan3D and H-MAGMA also detected a substantially higher number of significant genes relative to TWAS, a possible reflection of the limitation of eQTL-based approaches to discover significant associations that cannot be explained by eQTLs in the reference datasets.

Developmental and single-cell expression profiles. We used spatiotemporal transcriptomic data from embryonic and adult brains measured at 15 time periods, ranging from four postconceptional weeks to age ≥ 60 y (60). The gene-expression data are available for six brain regions: neocortex (NCX), mediodorsal nucleus of the thalamus, cerebellar cortex, HIP, AMY, and striatum. We focused here on the cortical expression profiles (NCX area), with 410 samples for the prenatal stages (periods 1 to 7) and 526 samples for the postnatal stages (periods 8 to 15). We centered the developmental expression matrix to the mean expression level for each sample.

We computed the average expression values across the significant genes for each brain sample and then compared the values for prenatal and postnatal brain samples. For psychiatric diseases, the genes detected by GeneScan3D tended to have significantly higher expression in prenatal relative to postnatal periods, as expected, and the trajectories highlight developmental windows in early or midgestation periods (Fig. 6*A*). For neurodegenerative diseases, the pattern was reversed, with higher expression in the postnatal periods (except for MS), concordant with the expectation that genes for neurodegenerative disorders have increased expression with aging. The results for H-MAGMA suggest similar patterns, but with reversed patterns for ASD and ALS and less significant differences for AD (*SI Appendix*, Fig. S12*A*). Results for GeneScan1D, MAGMA, and TWAS showed similar patterns (*SI Appendix*, Figs. S13–S15), although there were some discrep-

Downloaded at Institut Pasteur - CeRIS on January 14, 2022

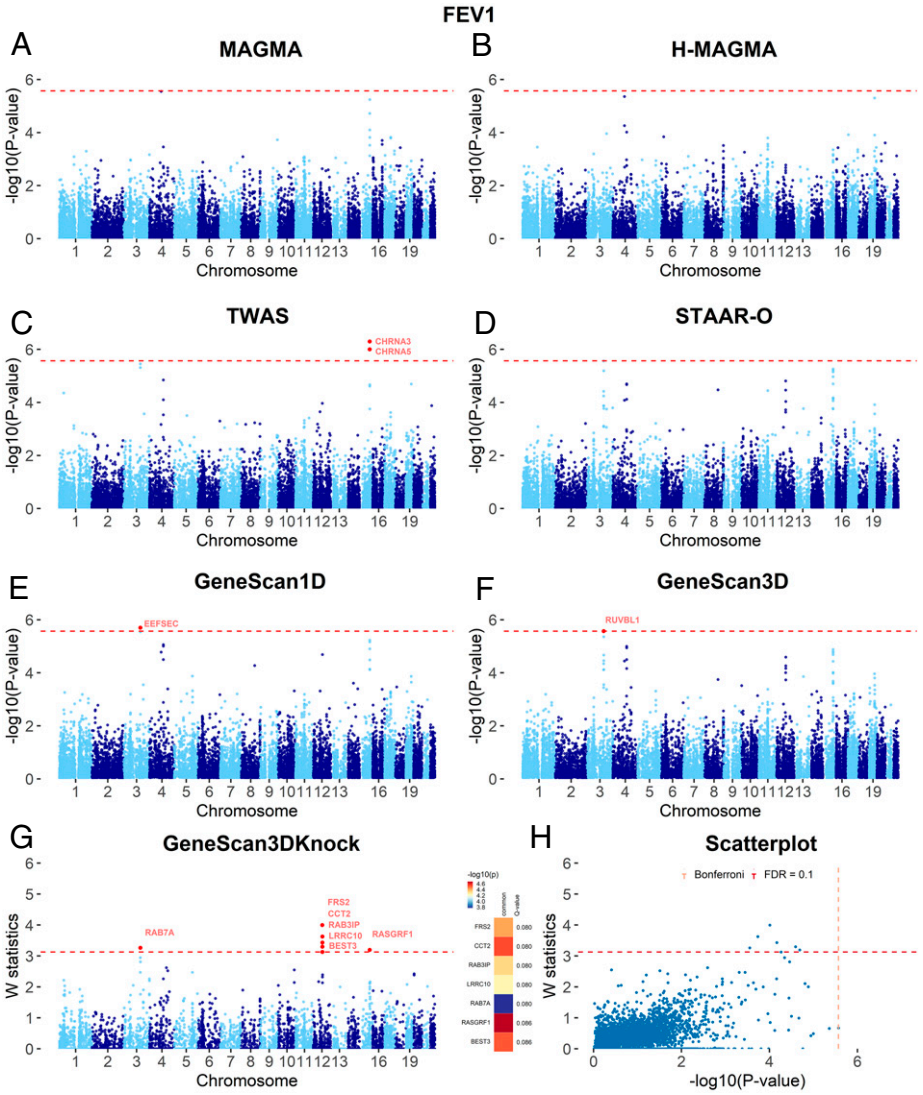


Fig. 5. Applications to COPDGene whole-genome sequencing data (trait FEV1) for common variants only. (*A–F*) Manhattan plots of MAGMA, H-MAGMA, TWAS, STAAR-O, GeneScan1D, and GeneScan3D results, respectively. (*G*) Manhattan plot of GeneScan3DKnock results. *G*, *Right* shows a heatmap with *P* values of GeneScan3D test for all genes passing the FDR = 0.1 threshold and the corresponding *q* values that already incorporate correction for multiple testing. The genes are shown in descending order of the knockoff statistics. (*H*) Scatterplot of *W* knockoff statistics (GeneScan3DKnock) vs. $-\log_{10}$ (*P* value) (GeneScan3D) for common variants. Each dot represents a gene. The dashed lines show the significance threshold defined by Bonferroni correction (for *P* values) and the data-adaptive threshold for FDR control (for *W* statistic).

ancies, including the higher postnatal expression vs. prenatal expression for the ASD-significant genes and significantly higher prenatal expression for ALS-significant genes (MAGMA).

Additionally, we also leveraged existing single-cell expression profiles (57) on 285 single cells from six adult brain-cell types, including neurons (131 cells), astrocytes (62 cells), microglia (16 cells), endothelial (20 cells), oligodendrocytes (38 cells), and oligodendrocyte progenitor cells (18 cells). For each single cell, we centered the expression data to the mean level of genes and then computed the average across the significant genes for a given disease. For each specific cell type, we averaged across the multiple cells in this cell type. We computed standardized expression levels (i.e., subtracted the mean and divided by the SD) for the six adult cell types. Genes identified by GeneScan3D for psychiatric disorders tended to show higher expression levels primarily in neurons and, to some extent, in astrocytes compared to other cell types, whereas genes for neurodegenerative disorders tended to show higher expression levels primarily in microglia (Fig. 6*B*). In particular, genes significant for ADHD showed the highest expression in astrocytes, consistent with recent evidence suggesting a key role of astrocytes in the regulation of attention-deficit disorder and hyperactivity (61).

Results for the other tests showed similar overall patterns as the GeneScan3D (*SI Appendix*, Figs. S12–S15), although with some differences, including less pronounced evidence for the role of astrocytes in ADHD (except for TWAS). These results serve as a proof of concept for the proposed 3D test, showing that genes identified by GeneScan3D and other existing tests exhibit expression patterns consistent with existing literature, i.e., an important role for neurons for neuropsychiatric diseases and microglia for neurodegenerative diseases (62).

Browser for results on UK/Japan Biobank data.

UK Biobank. We applied GeneScan3D to 1,403 UK Biobank binary phecodes and 827 continuous phenotypes using summary statistics on 28 million imputed variants. We have created a browser that displays phenome-wide results for a given gene and genome-wide gene-based results for a given trait and provides summary tables for significant genes. These gene-based results for the UK Biobank traits complement existing databases for single-variant tests (63) and rare variant-focused tests, such as SAIGE-GENE, a scalable generalized mixed-model region-based association test (6).

BioBank Japan. For non-European populations, we applied GeneScan3D to BioBank Japan binary phenotypes using available case-control GWAS summary statistics on 8,712,794 autosomal variants and 207,198 × chromosome variants with 212,453 Japanese individuals across 42 diseases (64). Results can be queried using the aforementioned browser.

Downloaded at Institut Pasteur - CeRIS on January 14, 2022

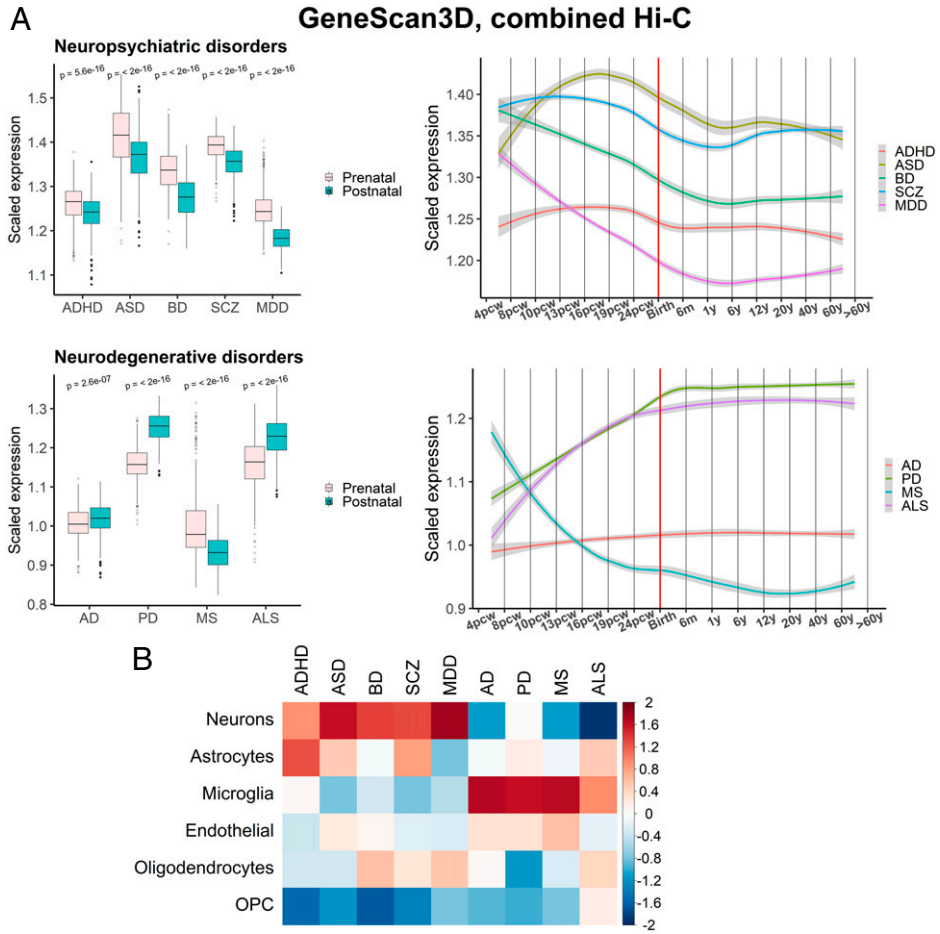


Fig. 6. (*A*) Human brain developmental expression of GeneScan3D significant genes for each brain disorder, combined Hi-C for adult brain and fetal brain GZ and CP layers. *P* values of Wilcoxon rank sum tests are shown in the boxplots to compare independent prenatal and postnatal samples. (*B*) Cell-type expression profiles of GeneScan3D significant genes. BD, bipolar disorder; MDD, major depressive disorder; PD, Parkinson's disease; SCZ, schizophrenia.

Discussion

We propose gene-based tests that integrate genetic variation residing in putative regulatory regions and implement the knockoff framework for increased power and improved causal gene prioritization. This framework provides a rich toolkit for the analysis of GWAS and whole-genome sequencing data with applications to gene discovery and fine mapping. Based on empirical studies, we show that the proposed gene-based tests are more powerful and help attenuate the confounding effect of LD relative to state-of-the-art gene-based tests. They also have distinct advantages compared with the recently proposed window-based test, KnockoffScreen, in terms of functional interpretation and appropriate FDR control at the gene level. Indeed, our simulation results suggest that the knockoff filter procedure needs to be performed at the gene, rather than window, level if our interest is in identifying genes and controlling FDR at the gene level.

Our gene-based tests can be seen as complementary to the TWAS approach. Like TWAS, they attempt to incorporate the effect of distal regulatory elements into the test. TWAS, however, is limited to common eQTLs detectable in reference datasets, which appear to account for a minority of GWAS signal (16, 17). In contrast, our approach has the ability to assess the effects of coding, noncoding, rare, and common variants, including those with no detectable effects on gene expression, and can scan the gene with varied window sizes. Furthermore, the knockoff framework can attenuate the confounding effect of LD and is able to produce a narrower list of possible causal genes, likely removing some of the false gene discoveries.

In this paper, we have focused on using existing external data on gene–enhancer links, and we recognize the limitations of these databases, both in terms of the accuracy of these links and the number of cell types with available data. Single-cell Hi-C is an emerging technology that could help overcome issues of tissue heterogeneity and expand these maps across many more cell types (65).

Like all gene-based tests that incorporate genetic variation in distal regulatory elements, our tests are also susceptible to false-positive associations due to, for example, causal variants residing in putative enhancers that may show significant interactions with promoters of multiple genes based on Hi-C data. Identifying the actual causal gene(s) requires follow-up experimental studies, such as CRISPR gene-perturbation experiments (66).

In summary, we propose comprehensive gene-based tests for common and rare variation, both coding and regulatory variation, that are more powerful than competitor gene-based tests in the literature. The GeneScan3DKnock approach is implemented in a computationally efficient R package.

Materials and Methods

GeneScan3DKnock: Proposed Gene-Based Association Test Incorporating Regulatory Elements and Knockoff Statistics. We describe here the details of the proposed gene-based test that aims to comprehensively evaluate the effects of rare and common, coding, and proximal and distal regulatory variation on a trait of interest. Details of existing gene-based association tests and additional tests for comparison, including GeneScan1D, MAGMA/H-MAGMA, and STAAR-O, as well as KnockoffScreen, are in *SI Appendix*.

Joint testing of rare and common variation within a window. For a fixed window Φ, we incorporated several recent advances for association tests for sequencing studies to compute the corresponding *P* value p_Φ, as follows. For each window, we conducted:

a. Burden and SKAT tests for common and low-frequency variants (MAF$\geq 1/\sqrt{2n}$) with Beta(MAF$_j$; 1, 25) weights. These tests aimed to detect the combined effect of common and low-frequency variants.

Downloaded at Institut Pasteur - CeRIS on January 14, 2022

b. Burden and SKAT tests for rare variants (MAF$< 1/\sqrt{2n}$ and MAC $\geq t$) with Beta(MAF$_j$; 1, 25) weights. These tests aimed to detect the combined effect of rare variants.
c. Burden and SKAT tests for rare variants, weighted by cell-type-specific functional annotations. These tests aimed to utilize functional annotations for improved power (24, 67).
d. Burden tests for aggregation of ultrarare variants (MAC$< t$). These tests aimed to aggregate effects from extremely rare variants (e.g., singletons, doubletons, etc.).
e. Single-variant score tests for common, low-frequency, and rare variants (MAC$\geq t$) in the window.

We then applied the ACAT (22) to combine P values from tests in a–e to compute P values of each 1D window Φ for all variants, including common and rare variants. Note that for the current analyses, we used MAC threshold 10.

GeneScan3D: Integrating proximal and distal regulatory elements for a gene. For a given gene G, we considered the gene body (i.e., the interval between the TSS and the end of 3′ UTR) $\pm$ 5-Kb buffer regions and integrated a single ChIP-seq promoter and R putative enhancers into the analyses (Fig. 1). A set of overlapping 1D windows $\Phi_m, m = 1, \ldots, M$ with window sizes 1 Kb, 5 Kb, and 10 Kb were generated to scan the gene and buffer regions together (each 1D window was overlapping with half of its adjacent windows for a given window size). Then, we constructed 3D windows for the gene by adding a ChIP-seq promoter and R putative enhancers to each 1D window $\Phi_m, m = 1, \ldots, M$ as follows (details on how to identify the regulatory elements for each gene are in *SI Appendix*):

$$\begin{aligned}
&\Phi^{3D}_{m,0} = \Phi_m + \text{ChIP-seq promoter}\\
&\Phi^{3D}_{m,1} = \Phi_m + \text{ChIP-seq promoter} + \text{Enhancer1}\\
\text{3D windows:}\quad &\ldots\ldots\\
&\ldots\ldots\\
&\Phi^{3D}_{m,R} = \Phi_m + \text{ChIP-seq promoter} + \text{Enhancer}R.
\end{aligned}$$

For each such 3D window, we computed a P value $p_{\Phi^{3D}_{m,r}}$, with $1 \leq m \leq M$, $0 \leq r \leq R$ using the proposed combined test for a window. Finally, we computed a gene-level P value p_G by combining the $(1 + R) \times M$ P values using the Cauchy's combination method (22), as follows:

$$Q = \frac{\sum_{r=0}^{R}\sum_{m=1}^{M}\tan[(0.5 - p_{\Phi^{3D}_{m,r}})\pi]}{(1+R)\times M}.$$

The P value of the Cauchy statistic is $p_G = 1/2 - \arctan(Q)/\pi$.

Note that the GeneScan3D analysis can be easily adapted for GWAS summary statistics by applying the ACAT procedure to combine P values for single variants within the 3D window.

GeneScan3DKnock: Knockoff-enhanced gene-based test for causal gene discovery. An advantage of the proposed GeneScan3D test is that it allows the discovery of multiple possible causal genes by incorporating information from proximal and distal regulatory elements. However, it is likely that some of those genes are false positives, owing to confounding due to LD and/or coregulation. Extensive LD at a locus of interest can confound the results and lead to many genes being significant. For example, if the LD region overlaps several enhancers, all genes regulated by such enhancers may show a significant signal. The knockoff framework (26), a recent advance in statistics, can be leveraged to reduce the effect of LD in such cases and can help prioritize a narrow list of potential causal genes. Furthermore, the knockoff-based test is of independent interest, as by design, it controls the FDR at a target level under arbitrary correlation structure and can have higher power to identify additional significant genes that are missed by the conventional gene-based test, as we show empirically in the applications. The knockoff-based test has two steps: the knockoff generation and the filtering of the results using the knockoff filter.

Model-X Knockoff Generation. The idea of the knockoff-based procedure is to generate artificial or knockoff genotypes $\tilde{G}$ such that for any subset K of variants the distribution of $(G, \tilde{G})$ is invariant when swapping G_K and $\tilde{G}_K$, i.e., $(G, \tilde{G})_{\text{swap(K)}} \overset{d}{=} (G, \tilde{G})$. Additionally the knockoff genotypes have the property that $\tilde{G} \perp Y|G$. Note that the well-known permutation procedure that permutes the samples does not guarantee these exchangeability properties between the original and knockoff genotypes. To generate valid knockoff genotypes, we can use a sequential model for knockoff generation that leverages the local patterns of LD, as previously proposed based on the Hidden Markov Models (HMMs) (30, 68), or an auto-regressive model (8), in such a way that the knockoff genotypes are exchangeable with the original (true) genotypes G, but are independent of the phenotype conditional on the original genotypes. The knockoff genotypes serve as negative controls and are designed to mimic the correlation or LD structure found within the original genotypes. Specifically, we sequentially sampled for each variant j the corresponding knockoff genotype $\mathcal{L}(G_j|G_{-j}, \tilde{G}_{1\ldots(j-1)})$, independent of the observed value of G_j. Because of the HMM's significant computational complexity with unphased genetic data, in order to generate knockoff genotypes, we relied on a recently introduced, computationally efficient, auto-regressive model that follows from the assumption that genotypes can be approximately modeled by a multivariate normal distribution:

$$G_j = \alpha + \sum_{k\neq j}\beta_k G_k + \sum_{k\leq j-1}\gamma_k \tilde{G}_k + \epsilon_j,$$

where ϵ_j is a random error term. (Note that we can leverage the approximate block structure for LD in the genome to only include variants in a neighborhood of the current variant j.) We estimated (α, β, γ) by minimizing the mean squared loss. We calculated the residual $\hat{\epsilon}_j = G_j - \hat{G}_j$ and its permutation $\hat{\epsilon}_j{}^*$, and then we defined the knockoff feature as $\tilde{G}_j = \hat{G}_j + \hat{\epsilon}_j{}^*$. More details on this knockoff-generation procedure and its theoretical and empirical properties can be found in ref. 8.

Knockoff Filter. Once the knockoff genotypes $\tilde{G}$ were generated, the knockoff filter was used to select significant genes. Specifically, we performed a gene-based test, as described above (GeneScan3D), in both the original cohort and the knockoff one. Let p_G and $p_{\tilde{G}}$ be the resulting P values. We defined a feature statistic by contrasting the observed P value for each gene to its counterpart based on the knockoff data. More precisely, the feature statistic for a gene G is defined as $W_G = T_G - T_{\tilde{G}}$, where $T_G = -\log_{10}(p_G)$ and $T_{\tilde{G}} = -\log_{10}(p_{\tilde{G}})$ are the importance scores for gene G in the original and knockoff cohort, respectively. This feature statistic has the flip-sign property, meaning that swapping the genetic variants in gene G with their knockoff counterparts changes the sign of W_G. A data-adaptive threshold τ for W_G can be determined by the knockoff filter (26) so that the FDR is controlled at the nominal level q, as follows:

$$\tau = \min\left\{t > 0 : \frac{1 + \#\{G : W_G \leq -t\}}{\#\{G : W_G \geq t\}} \leq q\right\}.$$

We selected all genes with $W_G \geq \tau$ since genes with large feature statistics are more likely to be causal (nonnull) genes. This follows from the exchangeability property between the original and the knockoff genotypes, which ensures that the importance scores (T_G and $T_{\tilde{G}}$) for the null genes are exchangeable, and therefore the feature statistic W_G is symmetric around zero for the null genes, but tends to be larger than zero for nonnull genes.

We additionally computed the corresponding q value for a gene, q_G. The q value already incorporates correction for multiple testing and is defined as the minimum FDR that can be attained when all tests showing evidence against the null hypothesis at least as strong as the current one are declared as significant. In particular, we define the q value for a gene G with feature statistic $W_G > 0$ as

$$q_G = \min_{t\leq W_G}\frac{1 + \#\{G' : W_{G'} \leq -t\}}{\#\{G' : W_{G'} \geq t\}},$$

where $\frac{1+\#\{G':W_{G'}\leq -t\}}{\#\{G':W_{G'}\geq t\}}$ is an estimate of the proportion of false discoveries if we were to select all genes with feature statistic $> t$ (with $t > 0$). For genes with feature statistic $W_G \leq 0$, we set $q_G = 1$.

Multiple Knockoffs. To improve the power and stability of the knockoff procedure, we implemented a multiple-knockoff procedure (8, 69), where the inference is based on generating multiple, independent knockoff datasets. Gimenez and Zou (69) proposed an extension of the sequential model for knockoff generation to multiple knockoffs and showed the validity of the multiple-knockoff-generation procedure in controlling the FDR. We implemented this procedure here to generate multiple independent knockoff datasets. Briefly, we sequentially sample for each variant j: $\tilde{G}_j^1, \ldots, \tilde{G}_j^M$ from $\mathcal{L}(G_j|\ G_{-j}, \tilde{G}^1_{1\ldots j-1}, \ldots, \tilde{G}^M_{1\ldots j-1})$, where M is the number of knockoffs. With multiple knockoffs, the feature statistic for a gene G is defined as

$$W_G = (T_G - \text{median}T^m_{\tilde{G}})I_{T_G \geq \max_{1\leq m\leq M} T^m_{\tilde{G}}},$$

where $T^m_{\tilde{G}}$ is the gene-importance score for gene G in the m-th knockoff replicate, and I is an indicator function. We define

Downloaded at Institut Pasteur - CeRIS on January 14, 2022

$$\tau = \min\left\{t > 0 : \frac{\frac{1}{M} + \frac{1}{M}\#\{G : \kappa_G \geq 1, \tau_G \geq t\}}{\#\{G : \kappa_G = 0, \tau_G \geq t\}} \leq q\right\},$$

where $\kappa_G = \text{argmax}_{0 \leq m \leq M} T_G^m$ (note that $T_G^0 = T_G$) and $\tau_G = T_G - \text{median} T_G^m$. We selected genes with $W_G \geq \tau$, i.e., those genes that have importance scores greater than any of those corresponding to the M knockoffs ($\kappa_G = 0$) and for which the difference from the median importance score is above some threshold ($\tau_G \geq \tau$). A q value for a gene G can be computed for the multiple-knockoff scenario, similar to the single-knockoff case.

The multiple-knockoff procedure helps improve power because at a target FDR of q, the single-knockoff approach needs to make a minimum of $1/q$ discoveries, while a multiple-knockoff approach with M knockoffs decreases this detection threshold to $1/Mq$. Therefore, in situations where the signal is sparse and the target FDR level q is low, a single-knockoff procedure will have very low power. In such cases, the multiple-knockoff procedure will tend to improve power. Furthermore, the multiple-knockoff procedure also helps with improving the stability of the selected genes, given that each knockoff generation is random, and, therefore, the results from a single knockoff can be unstable.

Data Availability. We used data from existing studies from COPDGene (TopMED; Database of Genotypes and Phenotypes [dbGaP] accession no. phs000951.v4.p4) and the ADSP (dbGaP accession no. phs000572.v8.p4) and summary-level GWAS results on neuropsychiatric and neurodegenerative traits are available from refs. 36 and 49–56. Details of web-based resources are in *SI Appendix*. All study data are included in the article and/or supporting information. We have implemented GeneScan3DKnock in a computationally efficient R package, available on GitHub (https://github.com/Iuliana-Ionita-Laza/GeneScan3DKnock), that can be applied generally to the analysis of other whole-genome sequencing or GWAS studies. Details of software implementation are in *SI Appendix*.

ACKNOWLEDGMENTS. This research was supported by NIH/National Institute of Mental Health Awards MH106910 and MH095797 (to I.I.-L.); and NIH/National Institute on Aging Award AG066206 (to Z.H.). We gratefully acknowledge the studies and participants who provided biological samples and data for the ADSP and TOPMed projects. The full study-specific acknowledgments are detailed in *SI Appendix*.

1. B. Li, S. M. Leal, Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
2. M. C. Wu *et al.*, Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
3. I. Ionita-Laza, R. Ottman, Study designs for identification of rare disease variants in complex diseases: The utility of family-based designs. *Genetics* **189**, 1061–1068 (2011).
4. S. Lee *et al.*, 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 (2012).
5. I. Ionita-Laza, S. Lee, V. Makarov, J. D. Buxbaum, X. Lin, Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* **92**, 841–853 (2013).
6. W. Zhou *et al.*, Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat. Genet.* **52**, 634–639 (2020).
7. Z. He, B. Xu, J. Buxbaum, I. Ionita-Laza, A genome-wide scan statistic framework for whole-genome sequence data analysis. *Nat. Commun.* **10**, 3018 (2019).
8. Z. He *et al.*, Identification of putative causal loci in whole-genome sequencing data via knockoff statistics. *Nat. Commun.* **12**, 3152 (2021).
9. A. Lesne, J. Riposo, P. Roger, A. Cournac, J. Mozziconacci, 3D genome reconstruction from chromosomal contacts. *Nat. Methods* **11**, 1141–1143 (2014).
10. C. A. de Leeuw, J. M. Mooij, T. Heskes, D. Posthuma, MAGMA: Generalized gene-set analysis of GWAS data. *PLOS Comput. Biol.* **11**, e1004219 (2015).
11. N. Y. A. Sey *et al.*, A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat. Neurosci.* **23**, 583–593 (2020).
12. X. Li *et al.*, Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52**, 969–983 (2020).
13. E. R. Gamazon *et al.*, A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
14. A. Gusev *et al.*, Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
15. S. Kim-Hellmuth *et al.*, Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
16. S. Chun *et al.*, Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).
17. B. D. Umans, A. Battle, Y. Gilad, Where are the disease-associated eQTLs? *Trends Genet.* **37**, 109–124 (2021).
18. J. Nasser *et al.*, Genome-wide maps of enhancer regulation connect risk variants to disease genes. *Nature* **593**, 238–243 (2021).
19. A. Mishra, R. D. Hawkins, Three-dimensional genome architecture and emerging technologies: Looping in disease. *Genome Med.* **9**, 87 (2017).
20. J. Hawkins *et al.*, Development of a framework for the co-production and prototyping of public health interventions. *BMC Public Health* **17**, 689 (2017).
21. S. S. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
22. Y. Liu *et al.*, ACAT: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.* **104**, 410–421 (2019).
23. R. Dey, E. M. Schmidt, G. R. Abecasis, S. Lee, A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *Am. J. Hum. Genet.* **101**, 37–49 (2017).
24. Z. He, B. Xu, S. Lee, I. Ionita-Laza, Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in Metabochip data. *Am. J. Hum. Genet.* **101**, 340–352 (2017).
25. M. Halvorsen *et al.*, Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia. *Nat. Commun.* **11**, 1842 (2020).
26. E. Candès, Y. Fan, L. Janson, J. Lv, Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *J. R. Stat. Soc. B* **80**, 551–577 (2018).
27. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
28. S. Fishilevich *et al.*, GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**, bax028 (2017).
29. C. P. Fulco *et al.*, Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
30. M. Sesia, E. Katsevich, S. Bates, E. Candès, C. Sabatti, Multi-resolution localization of causal variants across the genome. *Nat. Commun.* **11**, 1093 (2020).
31. Z. He, L. Liu, K. Wang, I. Ionita-Laza, A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using MPRAs. *Nat. Commun.* **9**, 5199 (2018).
32. E. E. Blue, A. Cheng, S. Chen, C. E. Yu; Alzheimer's Disease Genetics Consortium, Association of uncommon, Noncoding variants in the APOE region with risk of Alzheimer disease in adults of European ancestry. *JAMA Netw. Open* **3**, e2017666 (2020).
33. K. Nho *et al.*, Association analysis of rare variants near the APOE region with CSF and neuroimaging biomarkers of Alzheimer's disease. *BMC Med. Genomics* **10**(suppl. 1), 29 (2017).
34. M. Morawski *et al.*, Glutaminyl cyclase in human cortex: Correlation with (pGlu)-amyloid-β load and cognitive decline in Alzheimer's disease. *J. Alzheimers Dis.* **39**, 385–400 (2014).
35. D. Y. Kim, L. A. Ingano, D. M. Kovacs, Nectin-1alpha, an immunoglobulin-like receptor involved in the formation of synapses, is a substrate for presenilin/gamma-secretase-like cleavage. *J. Biol. Chem.* **277**, 49976–49981 (2002).
36. I. E. Jansen *et al.*, Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
37. R. E. Marioni *et al.*, GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* **8**, 99 (2018).
38. K. de Paiva Lopes *et al.*, Atlas of genetic effects in human microglia transcriptome across brain regions, aging and disease pathologies. *bioRxiv* [Preprint] (2020). https://doi.org/10.1101/2020.10.27.356113 (Accessed 28 October 2020).
39. C. V. Rao, M. Farooqui, Y. Zhang, A. S. Asch, H. Y. Yamada, Spontaneous development of Alzheimer's disease-associated brain pathology in a Shugoshin-1 mouse cohesinopathy model. *Aging Cell* **17**, e12797 (2018).
40. S. Ono *et al.*, A promoter SNP (-1323T>C) in G-substrate gene (GSBS) correlates with hypercholesterolemia. *J. Hum. Genet.* **48**, 447–450 (2003).
41. C. Lanni *et al.*, Homeodomain interacting protein kinase 2: A target for Alzheimer's beta amyloid leading to misfolded p53 and inappropriate cell survival. *PLoS One* **5**, e10171 (2010).
42. A. Maruszak *et al.*, Cellular phenotyping of hippocampal progenitors exposed to patient serum predicts conversion to Alzheimer's disease. *bioRxiv* [Preprint] (2017). https://doi.org/10.1101/175604 (Accessed 13 August 2017).
43. N. Shrine *et al.*, New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.* **51**, 481–493 (2019).
44. M. Liu *et al.*, Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).
45. G. Kichaev *et al.*, Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).
46. W. C. Tan *et al.*, High eosinophil counts predict decline in FEV1: Results from the CanCOLD study. *Eur. Respir. J.* **57**, 2000838 (2020).
47. Z. Daniloski *et al.*, Identification of required host factors for SARS-CoV-2 infection in human cells. *Cell* **184**, 92–105.e16 (2021).
48. L. Wang *et al.*, An atlas connecting shared genetic architecture of human diseases and molecular phenotypes provides insight into COVID-19 susceptibility. *medRxiv* [Preprint] (2020). https://doi.org/10.1101/2020.12.20.20248572 (Accessed 22 December 2020).
49. D. Demontis *et al.*, Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).
50. J. Grove *et al.*, Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
51. E. A. Stahl *et al.*, Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.* **51**, 793–803 (2019).

Downloaded at Institut Pasteur - CeRIS on January 14, 2022

52. A. F. Pardiñas *et al.*, Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
53. D. M. Howard *et al.*, Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).
54. M. A. Nalls *et al.*, Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: A meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).
55. W. van Rheenen *et al.*, Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).
56. T. F. M. Andlauer *et al.*, Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. *Sci. Adv.* **2**, e1501678 (2016).
57. D. Wang *et al.*, Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
58. H. Won *et al.*, Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).
59. F. Ay, T. L. Bailey, W. S. Noble, Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* **24**, 999–1011 (2014).
60. H. J. Kang *et al.*, Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).
61. J. Nagai *et al.*, Hyperactivity with disrupted attention by activation of an astrocyte synaptogenic cue. *Cell* **177**, 1280–1292.e20 (2019).
62. . S. Hickman, S. Izzy, P. Sen, L. Morsett, J. El Khoury, Microglia in neurodegeneration. *Nat. Neurosci.* **21**, 1359–1369 (2018).
63. O. Canela-Xandri, K. Rawlik, A. Tenesa, An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
64. K. Ishigaki *et al.*, Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat. Genet.* **52**, 669–679 (2020).
65. J. Dekker *et al.*, The 4D nucleome project. *Nature* **549**, 219–226 (2017).
66. E. Cano-Gamez, G. Trynka, From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Front. Genet.* **11**, 424 (2020).
67. A. Kundaje *et al.*, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
68. M. Sesia, C. Sabatti, E. J. Candès, Rejoinder: Gene hunting with hidden Markov model knockoffs. *Biometrika* **106**, 35–45 (2019).
69. J. R. Gimenez, J. Zou, Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. *arXiv* [Preprint] (2018). https://arxiv.org/abs/1810.11378 (Accessed 26 October 2018).

Downloaded at Institut Pasteur - CeRIS on January 14, 2022