ARTICLE

Genome-wide analysis of common and rare variants via multiple knockoffs at biobank scale, with an application to Alzheimer disease genetics

Zihuai He,[1,2,8,*] Yann Le Guen,[1,3,8] Linxi Liu,[4] Justin Lee,[2] Shiyang Ma,[5] Andrew C. Yang,[1] Xiaoxia Liu,[1] Jarod Rutledge,[6] Patricia Moran Losada,[1] Bowen Song,[7] Michael E. Belloy,[1] Robert R. Butler III,[1] Frank M. Longo,[1] Hua Tang,[6] Elizabeth C. Mormino,[1] Tony Wyss-Coray,[1] Michael D. Greicius,[1] and Iuliana Ionita-Laza[5]

Summary

Knockoff-based methods have become increasingly popular due to their enhanced power for locus discovery and their ability to prioritize putative causal variants in a genome-wide analysis. However, because of the substantial computational cost for generating knockoffs, existing knockoff approaches cannot analyze millions of rare genetic variants in biobank-scale whole-genome sequencing and whole-genome imputed datasets. We propose a scalable knockoff-based method for the analysis of common and rare variants across the genome, *KnockoffScreen-AL*, that is applicable to biobank-scale studies with hundreds of thousands of samples and millions of genetic variants. The application of *KnockoffScreen-AL* to the analysis of Alzheimer disease (AD) in 388,051 WG-imputed samples from the UK Biobank resulted in 31 significant loci, including 14 loci that are missed by conventional association tests on these data. We perform replication studies in an independent meta-analysis of clinically diagnosed AD with 94,437 samples, and additionally leverage single-cell RNA-sequencing data with 143,793 single-nucleus transcriptomes from 17 control subjects and AD-affected individuals, and proteomics data from 735 control subjects and affected indviduals with AD and related disorders to validate the genes at these significant loci. These multi-omics analyses show that 79.1% of the proximal genes at these loci and 76.2% of the genes at loci identified only by *KnockoffScreen-AL* exhibit at least suggestive signal ($p < 0.05$) in the scRNA-seq or proteomics analyses. We highlight a potentially causal gene in AD progression, *EGFR*, that shows significant differences in expression and protein levels between AD-affected individuals and healthy control subjects.

Introduction

Recent advances in whole-genome sequencing (WGS) and genotype imputation technologies provide an exciting opportunity to identify common (minor allele frequency [MAF] $\geq$ 1%) and rare (MAF $<$ 1%) genetic variation in the human genome and to investigate their contribution to complex trait heritability. Large-scale WGS/WG-imputed studies, such as the Trans-Omics for Precision Medicine (TOPMed) study[1] and UK Biobank (UKBB),[2] have collected hundreds of thousands of samples with directly sequenced or imputed whole genomes. However, our ability to analyze and infer causal pathways from these datasets remains limited at this point. The main challenges include the substantial computational cost, the burden of multiple comparisons, and the difficulties with the functional interpretation of the discovered loci. In addition, it is well known that conventional association tests often identify proxy variants that are correlated only with the true causal variants. Identification of causal variants remains challenging, and it usually requires a follow-up statistical fine-mapping analysis.[3]

A considerable proportion of variants identified by genome-wide association studies (GWASs) reside in intergenic regions, and their functional consequences remain unknown. Recent large-scale GWAS analyses have leveraged external multi-omics resources, such as GTEx[4] and ENCODE,[5] for post-GWAS analyses including colocalization, fine-mapping, and functional enrichment analyses, to better interpret their findings and identify putative causal genes and variants. Although important, the success of these analyses has been limited. For example, eQTLs detected in GTEx account only for a minority of GWAS signal,[6] making colocalization analyses less powerful. Fine-mapping methods such as CAVIAR[7] and SuSiE[8] were developed for common variants in GWASs and are not directly applicable to rare variants. Moreover, most public multi-omics resources are not disease specific, making it challenging to connect potential functional consequences of genetic variants to a particular disease or phenotype of interest.

Knockoff-based methods have become increasingly popular due to their enhanced power for locus discovery and their ability to prioritize putative causal variants in a genome-wide analysis,[9–12] in contrast to conventional association

[1]Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA 94305, USA; [2]Quantitative Sciences Unit, Department of Medicine, Stanford University, Stanford, CA 94305, USA; [3]Institut du Cerveau - Paris Brain Institute - ICM, Paris 75013, France; [4]Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA; [5]Department of Biostatistics, Columbia University, New York, NY 10032, USA; [6]Department of Genetics, Stanford University, Stanford, CA 94305, USA; [7]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, USA
[8]These authors contributed equally
*Correspondence: zihuai@stanford.edu
https://doi.org/10.1016/j.ajhg.2021.10.009.

tests. Specifically, whereas conventional fine-mapping methods are applied to individual signal regions in post-GWAS analyses, the knockoff-based approach allows simultaneous genome-wide locus discovery and prioritization of causal variants. The idea of knockoff-based inference is to generate synthetic, noisy copies (knockoffs) of the original genotypes where each sample resembles the original data in terms of linkage disequilibrium (LD) structure but is conditionally independent of the trait of interest, given the original genotypes. The synthetic sequences serve as negative controls for feature selection, which helps enhance power and attenuate the confounding effect of LD. Several knockoff-based methods have been proposed for genetic data. For example, Sesia et al.[9,10] proposed KnockoffZoom based on Hidden Markov Models (HMMs) to generate knockoffs for phased common variants in the UK Biobank (~400,000 samples; 600k variants). They demonstrated that the knockoff-based method exhibits comparable/better performance than state-of-the-art fine-mapping methods such as SuSiE and CAVIAR in terms of the prioritization of causal variants. More recently, He et al.[12] proposed KnockoffScreen based on a sequential conditional independent tuples[13] (SCIT) algorithm for unphased common and rare variants in moderate-scale whole-genome sequencing studies (<50,000 samples; >80,000k variants). It additionally allows for inference based on multiple knockoffs for improved power, stability, and reproducibility. However, these existing approaches are not scalable to the generation of multiple knockoffs for hundreds of thousands of samples and millions of genetic variants.

Here we propose an extension of knockoff-based method, KnockoffScreen-AL, for the analysis of biobank-scale WGS/WG-imputed studies of hundreds of thousands of samples and millions of genetic variants. Since most genome-wide data are unphased, and phasing rare variants is particularly challenging, we propose knockoff-based inference directly using unphased genotype data. To improve the computational efficiency while retaining the advantages over conventional association tests, KnockoffScreen-AL uses a shrinkage algorithmic leveraging[14] (AL) technique to select a subset of "informative" samples to estimate intermediate parameters during the knockoff generation, which makes biobank-scale WGS/WG-imputed studies feasible within the knockoff framework. We also propose low-rank regression and memory-efficient matrix operation to further improve the computational efficiency. We have additionally developed a gene-based inference based on the summary statistics of KnockoffScreen-AL that leverages external transcriptome and epigenome information to help prioritize causal genes nearby.

We applied KnockoffScreen-AL to 388,051 WG-imputed samples from the UK Biobank and identified 31 distinct loci (common variants or rare variant windows) associated with AD, including 14 loci that are missed by conventional association tests on these data. The identified loci correspond to 43 proximal genes. We attempted to replicate these significant findings in an independent meta-analysis of clinically diagnosed AD with 94,437 samples. To gain further insights into the potential causal role of the candidate genes at these loci, we leveraged single-cell RNA-sequencing data with 143,793 single-nucleus transcriptomes from 17 control subjects and AD-affected individuals and proteomics data from 735 control subjects and individuals with AD and related disorders. We observed that 34/43 (79.1%) of the proximal genes exhibit a suggestive effect in either scRNA-seq or proteomics analyses, substantially higher than background genes (46.2%; $p = 1.8 \times 10^{-5}$ by Fisher's exact test). The results demonstrate that KnockoffScreen-AL can identify weaker signals, particularly rare variant loci that are missed by conventional association tests yet with possible functional effects on AD.

Material and methods

We propose a scalable multiple-knockoff based method, KnockoffScreen-AL, to perform whole-genome analysis of unphased genetic data at biobank scale. The idea of knockoff-based inference is to generate a synthetic sequence for each sample while preserving the overall sequence correlation structure. Specifically, for each genetic variant, a knockoff version is created that does not directly affect the trait of interest. By contrasting the original and synthetic data, the knockoff-based method allows the selection of genetic variants/windows related to the phenotype of interest while controlling the false discovery rate (FDR). Like other multiple knockoff-based approaches, the proposed method has several appealing features, including: (1) prioritization of causal variants over associations, (2) ability to distinguish the signal due to rare variants from shadow effects of significant common variants nearby, and (3) improved stability and reproducibility due to multiple knockoffs.[12] We present the technical details in Appendix A.

KnockoffScreen-AL uses state-of-the-art algorithmic leveraging, matrix decomposition techniques, and memory-efficient matrix operations to substantially improve the computational efficiency and memory usage. KnockoffScreen-AL contains four main steps: (1) generate multiple knockoffs per variant; (2) calculate the feature importance score for the original variants and the knockoff variants; (3) calculate the feature statistic by contrasting feature importance scores for the original and their knockoff counterparts; and (4) apply knockoff filter to select significant variants/windows with FDR control. The workflow is shown in Figure 1.

In step 1, KnockoffScreen-AL augments each genetic variant with multiple synthetic variants (knockoffs) by a sequential conditional independent tuples (SCIT) algorithm.[12,14]

Algorithm 1. Sequential conditional independent tuples (multiple knockoffs)

j = 1
while $j \leq p$ do
Sample $\tilde{G}_j^1, \cdots, \tilde{G}_j^M$ independently from
$\mathcal{L}(G_j \mid \boldsymbol{G}_{-\boldsymbol{j}}, \tilde{\boldsymbol{G}}^1_{1:(\boldsymbol{j}-1)}, \cdots, \tilde{\boldsymbol{G}}^{\boldsymbol{M}}_{1:(\boldsymbol{j}-1)})$
$j = j + 1$
End

where G_j denotes the j^{th} variant; $\boldsymbol{G}_{-\boldsymbol{j}}$ denotes all genetic variants except for the j^{th} variant; M is the number of knockoffs per variant;

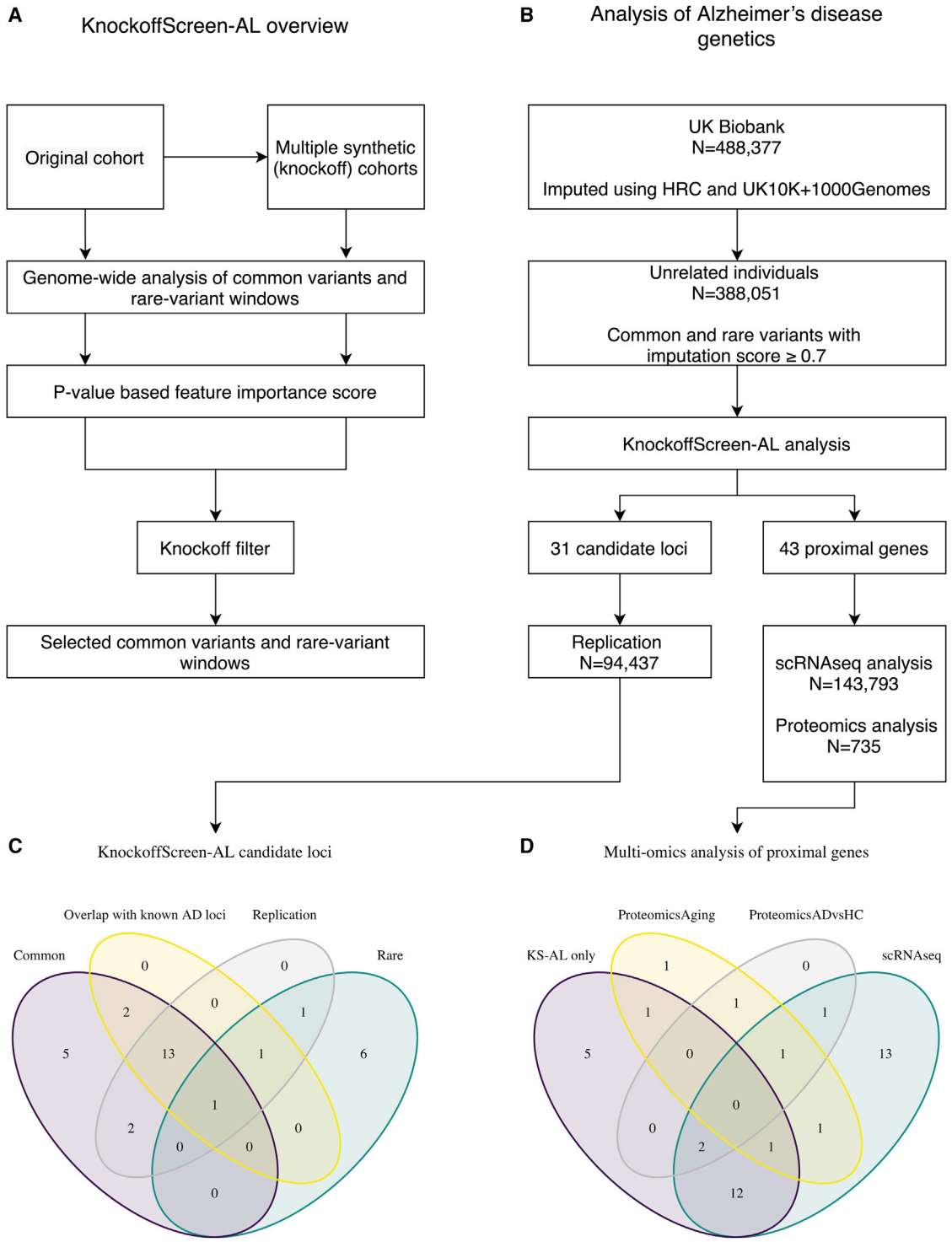

Figure 1. Overview of KnockoffScreen-AL

(A) The KnockoffScreen-AL method.

(B) The application of KnockoffScreen-AL to UK biobank data.

(C) Venn diagrams showing the number of identified loci that overlap with known AD loci or being replicated ($p < 0.05$). Common, common variant loci; rare, rare-variant loci; overlap with known AD loci, overlap with Jansen et al.[15] and Kunkle et al.;[16] replication, replication p value < 0.05 based on summary statistics from Kunkle et al.[16]

(D) Venn diagrams showing the number of implicated genes that are significant ($p < 0.05$) in scRNA-seq or proteomics analysis; KS-AL only: the additional genes identified by KnockoffScreen-AL but missed by conventional association tests; ProteomicsAging: p value $<$ 0.05 in the proteomics analysis of age effect; ProteomicsADvsHC: p value < 0.05 in the proteomics analysis comparing Alzheimer disease-affected individuals to healthy control subjects; scRNA-seq: p value < 0.05 in the scRNA-seq analysis for at least one cell type.

and $\mathcal{L}(G_j \Big| \boldsymbol{G}_{-j}, \tilde{\boldsymbol{G}}^1_{1:(j-1)}, \cdots, \tilde{\boldsymbol{G}}^{\boldsymbol{M}}_{1:(j-1)})$is the conditional distribution of G_j given $\boldsymbol{G}_{-j}$ and $\{\tilde{\boldsymbol{G}}^{\boldsymbol{m}}_{1:(j-1)}\}_{1\le m\le M}$. The SCIT algorithm ensures that all original variants and synthetic variants are simultaneously exchangeable to each other, i.e., the joint distribution of $(G, \tilde{G}^1, \ldots, \tilde{G}^M)$ remains the same if one swaps any subset of variants with their counterparts.

The SCIT algorithm above requires iteratively fitting regression models to estimate the conditional distribution, which can be time consuming when the sample size and/or the number of variants are large, as is the case for biobank-scale WGS/WG-imputed studies. To make such studies feasible, we developed several optimization strategies to make the generation of multiple knockoffs practical for biobank-scale data. First, the naive conditional autoregressive model has been replaced by a low-rank approximation, which will leverage the facts that the covariance matrices used for different iterations largely overlap, and that the knockoffs are highly correlated with the original variants due to the exchangeability property. Second, we have implemented a shrinkage algorithmic leveraging technique, a sampling method to reduce the data size in order to substantially improve the computational efficiency.[14] Third, we have implemented memory-efficient matrix operation using shared memory and memory-mapped files. Details are provided in Appendix A.

In step 2, single-variant tests for common variants (MAF $\geq$ 1%) and window-based tests for rare variants (MAF $<$ 1% and MAC $\geq$ 25) are conducted to scan the genome. Since in the UKBB rare variants are imputed, the imputation quality may be low and thus we chose a relatively conservative threshold for the variants included in the analyses. The window-based tests are applied to every 2 kb window, with half of each window overlapping with adjacent windows of the same size. p values are calculated for the original variants/windows and all the corresponding knockoff counterparts. Feature importance scores are generally defined as

$$T = -\log_{10} p,$$

where a larger T indicates a more significant association. KnockoffScreen-AL is very flexible and can incorporate p values from a variety of tests for rare and common variants. For rare-variant windows, KnockoffScreen-AL performs the aggregated Cauchy association test[17] (ACAT-O) method by default which combines burden, sequence kernel association test[18] (SKAT), and single-variant test for rare variants for enhanced power. For single variants, KnockoffScreen-AL performs score test for quantitative outcomes and the saddle point approximation for binary outcomes which is robust to unbalanced case-control ratio,[19] a common issue for analyses with biobank data. We note that the feature importance score is not restricted to p values. We use this definition here as it can serve as a wrapper method to flexibly use p values from existing or future association tests to construct feature importance scores.

In steps 3 and 4, KnockoffScreen-AL uses the same definition of feature statistic and knockoff filter as in KnockoffScreen to select significant variants/windows with rigorous FDR control. The feature statistic is defined as

$$W = \left(T - \underset{1\le m\le M}{\text{median}} T^m\right) I_{T \geq \max\limits_{1\le m\le M} T^m}, \text{(Equation 1)}$$

and all common variants and rare-variant windows with feature statistic $W > \tau$ are selected, where τ is calculated by the knockoff filter described in Appendix A. Specifically, we select those variants and windows with the original feature importance score having higher value than any of the M knockoffs, and with the gap with the median of knockoff importance score being above some defined threshold. A q value as in the Benjamini-Hochberg procedure[20] can also be computed for each variant/window. Variable selection with $q \le \alpha$ will ensure genome-wide FDR $\le \alpha$.

We show in the next section that KnockoffScreen-AL with algorithmic leveraging and memory-efficient operation exhibits equivalent or slightly higher power than the "exact" SCIT algorithm, but with much lower computational and memory costs, for both quantitative and dichotomous traits.

Results

KnockoffScreen-AL is computationally and memory efficient for biobank-scale data

To evaluate the computational performance of KnockoffScreen-AL, we performed simulations to empirically evaluate the computational time and memory usage for the different methods, varying the sample size and number of variants. We note that the computational cost is reported for the generation of multiple knockoffs for unphased genotype data (as opposed to phased haplotype data). We simulated genetic data using the SKAT package, with varying sample sizes and number of genetic variants. The computing time and memory usage were evaluated on a single CPU (Intel Xeon CPU E5-2640 v3 @ 2.60 GHz). We benchmarked the proposed shrinkage algorithmic leveraging, the proposed method with memory-efficient matrix operation (KnockoffScreen-AL), SCIT with the "exact" linear model (naive SCIT; KnockoffScreen v.1.0), and SCIT with low-rank regression (Low-rank approximation; the revised implementation of KnockoffScreen v.1.1) to generate 5 knockoffs (Figure 2).

The computing time is plotted in Figure 2A. For a study with 100,000 individuals, we observed that the method with shrinkage algorithmic leveraging took 268.48 s to generate 5 knockoffs for 2,000 variants, which is ~9 times faster than the SCIP approach with "exact" linear model (2,396.46 s). The additional memory-efficient matrix operation added negligible computing time to the base method (212.74 versus 267.36 s for 100,000 individuals; 1,428.67 versus 1,450.82 s for 500,000 individuals). An important advantage of our knockoff generation is that it is based on unphased genotype data, unlike existing HMM-based knockoff generators that require phased haplotype data. Indeed, phasing is computationally expensive and accurately phasing rare variants beyond reference panels is challenging. For example, the computing time for phasing 10,000 individuals with 2,000 variants via fastPHASE[21] (number of states 12 as used in KnockoffZoom[10]) is 1,571.2 s. Therefore, the phasing step can take a substantial fraction of the total computing time for knockoff generators that require phased haplotypes, such as the HMM knockoff generator in KnockoffZoom. Furthermore, increasing the number of states to achieve a higher phasing accuracy can lead to substantial increases in computational time (Table S1; e.g., 23,584.8 s when the number of states is 50). This demonstrates the advantage

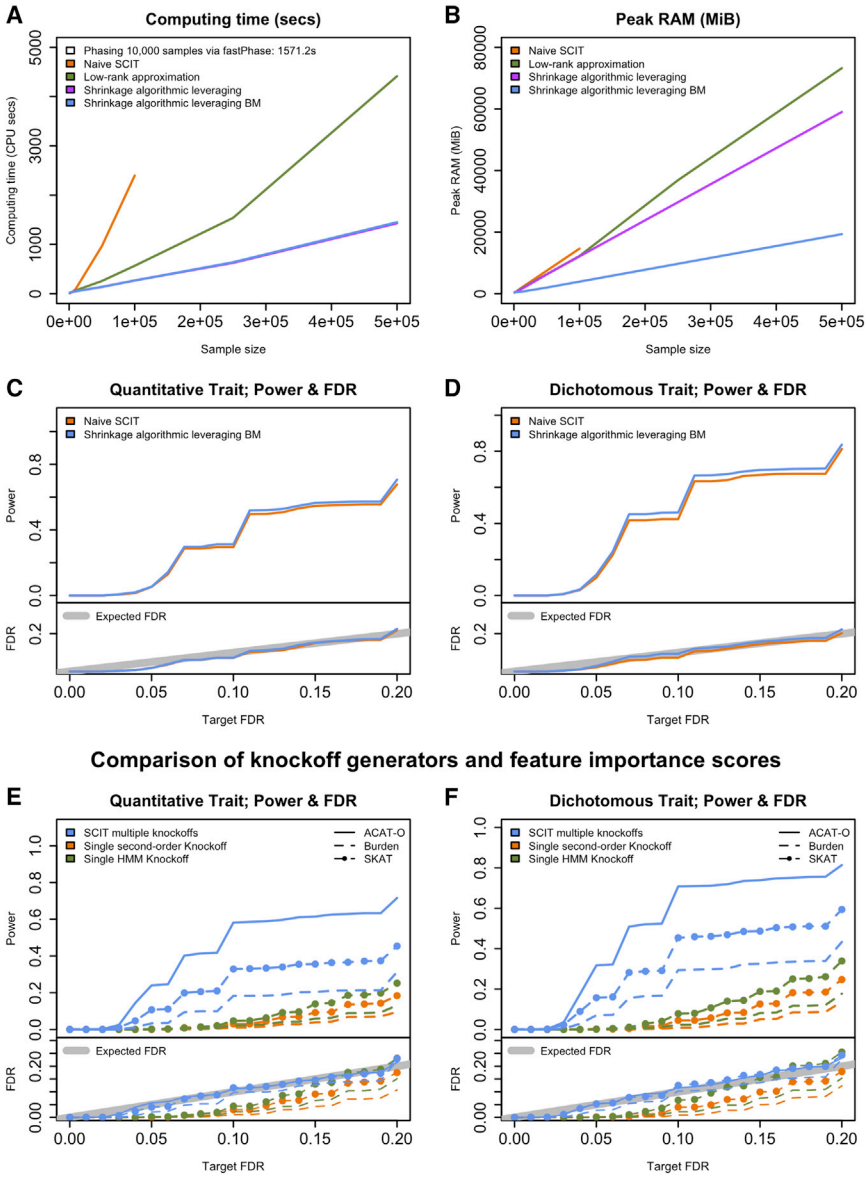

Figure 2. Computing time, peak random-access memory (RAM) use, power, and FDR of different knockoff generators

(A and B) The computing time and RAM were evaluated based on 2,000 variants, varying the sample size from 1,000 to 500,000. Naive SCIT, sequential conditional independent tuples (SCIT) with the "exact" linear model; BM, memory-efficient matrix operation. The shrinkage algorithmic leveraging BM method corresponds to the proposed KnockoffScreen-AL. The computing time for naive SCIT is truncated at sample size 100,000 because it cannot be applied to larger sample size. We also benchmark the computing time for phasing 10,000 samples via fastPhase with number of states $K = 12$.

(C and D) Power/FDR comparison between KnockoffScreen-AL and the naive SCIT.

(E and F) Power/FDR comparison between KnockoffScreen-AL (SCIT multiple knockoffs + ACAT-O) and other existing knockoff generators and feature importance score calculations. The different colors indicate different knockoff generators. The different types of lines indicate different tests to define the importance score.

of the proposed method which directly generates knockoffs for genetic data without phasing. A direct comparison with the HMM method can be found in Table S1.

Memory usage is plotted in Figure 2B. The memory cost for all methods is linear with sample size; the memory-efficient operation reduces the memory cost as the number of variants increases. For a study with ~500,000 individuals like the UK Biobank, KnockoffScreen-AL requires ~20 GB while the base procedure without the memory-efficient matrix operation requires ~60 GB. The results demonstrate that the memory-efficient operation substantially reduces the memory cost, with nearly equivalent computational cost.

We also conducted simulation studies to evaluate the power and FDR of the proposed method, KnockoffScreen-AL, and perform comparisons to the "exact" SCIT method (Figures 2C and 2D). We found that both power and FDR are nearly equivalent to the naive SCIT method, but with substantially reduced computational and memory cost. We additionally present comparisons with other knockoff methods (second-order knockoff generator proposed by Candes et al.[11] and HMM knockoff generator) and other tests for defining feature importance scores (SKAT and burden tests) in Figures 2E and 2F. We show that KnockoffScreen-AL outperforms other methods in terms of power, particularly at low FDR level because of the implementation of multiple knockoffs. More extensive comparisons have been performed previously in He et al.[12]

KnockoffScreen-AL discovers additional loci associated with AD beyond the conventional analyses

We applied KnockoffScreen-AL to the UK Biobank data in order to identify loci associated with AD. The UK Biobank data include 488,377 participants, who were genotyped on single-nucleotide polymorphism (SNP) microarrays and imputed at high resolution based on the reference panels from Haplotype Reference Consortium (HRC) and the UK10K+1000Genomes.[2] After pre-processing steps described in the supplemental material and methods, we focused on 388,051 unrelated participants and 39,091,537 variants with an imputation quality score $\geq$ 0.7. We did not restrict the analysis to British individuals since the KnockoffScreen method was shown to be robust

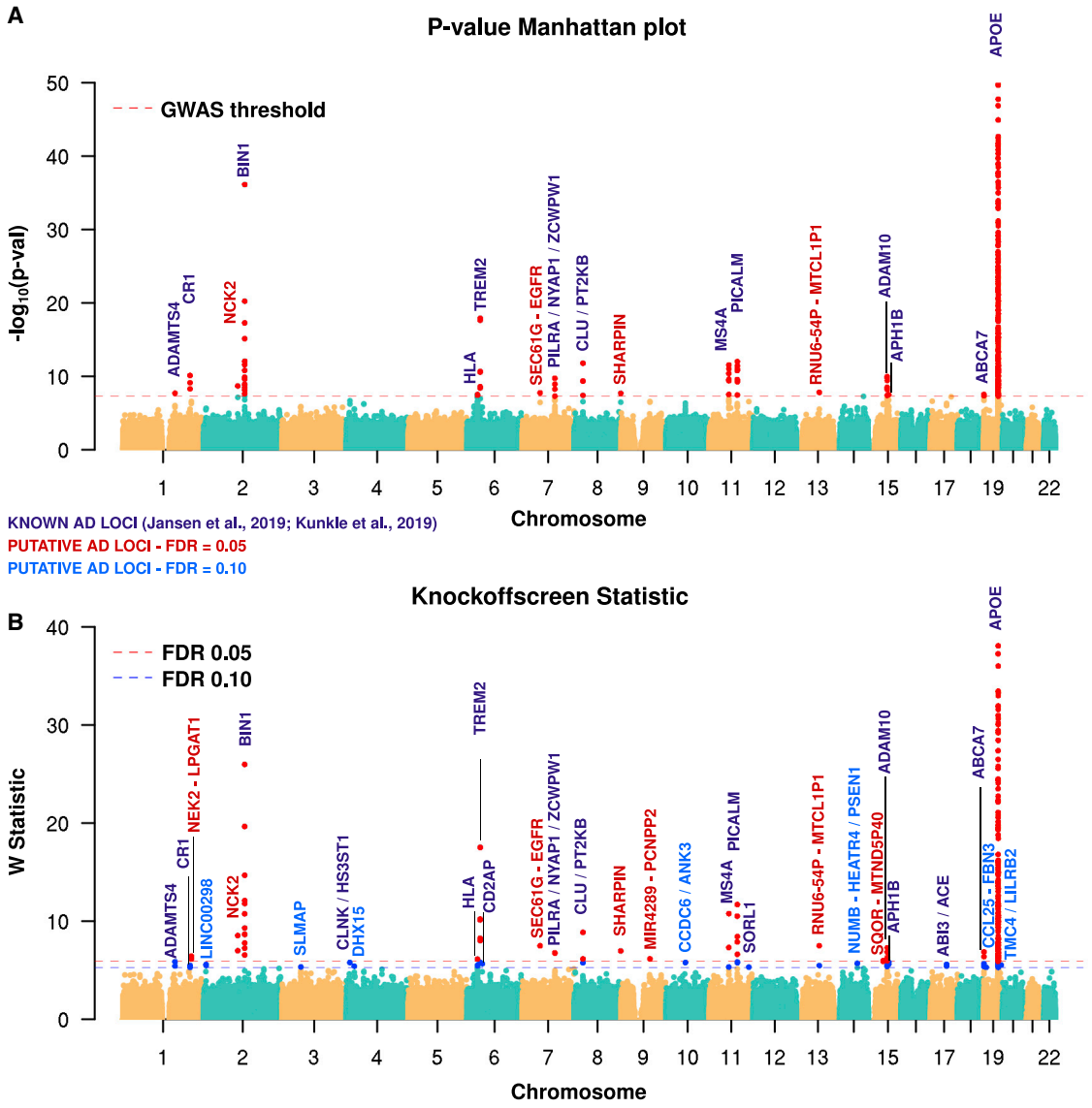

Figure 3. Genome-wide analysis of Alzheimer disease in UK Biobank
(A) The Manhattan plot of p values (truncated at 10^{-50} for clear visualization) from the conventional common-variant and rare-variant association tests with conventional GWAS threshold ($p < 5 \times 10^{-8}$) for FWER control.
(B) The Manhattan plot of KnockoffScreen-AL with target FDR at 0.10. The names of those loci previously reported by GWASs are shown in purple; names of discoveries not included in Jansen et al.[15] and Kunkle et al.[16] are shown in red (FDR = 0.05) and blue (FDR = 0.10).

to population stratification in practice.[12] We used the AD-proxy score defined in Jansen et al.[15] which combines the self-reported parental AD status and the individual AD status. The analyses were adjusted for age at last visit, sex, genotyping array, assessment center, and the first 20 principal components of genetic ancestry as provided by the UK Biobank. We present the workflow in Figure 1. More details on the UK Biobank data and the analyses are available in the supplemental material and methods.

We considered single common variants (MAF $\geq$ 1%) and 2 kb rare-variant windows (MAF $<$ 1% and MAC $\geq$ 25) across the genome.[22] For each variant/window, we computed the p value for the original variant/window and its five knockoffs. Then we defined the feature importance score $T = -\log_{10} p$ and applied the knockoff filter. We compared the results from conventional association tests (i.e., the same combination of single variant and window-based tests as implemented in KnockoffScreen-AL, but with a conventional GWAS threshold of 5×10^{-8} for family-wise-error-rate (FWER) control to the results from KnockoffScreen-AL at an FDR threshold of 0.1 (Figure 3). We note that the comparison is in favor of conventional association tests because the threshold of 5×10^{-8} does not account for the additional rare-variant windows. Inspection of QQ-plots of all tests shows that the type I error rate is well controlled (Figure S1). Tables 1 and 2 summarize the lead variant/window per locus. Known AD loci are named based on their commonly used names in the literature. For other loci, we name them based on the proximal gene(s). The locus and gene annotations can be found in

Table 1. Lead common variant at each locus associated with Alzheimer disease in the UK Biobank at FDR = 0.05 and FDR = 0.10

Chr	Start	RSID	Locus	Gene	p	q	W	MAF	Direction	p. replication
chr1	161186243	rs11585858	*ADAMTS4*	*ADAMTS4*	2.0E−08	0.052	5.869	22.7%	++	2.7E−02
chr1	207512620	rs4562624	*CR1*	*CR1*	7.7E−11	0.076	5.493	17.3%	++	1.8E−15
chr2	127135234	rs6733839	*BIN1*	*(BIN1-NIFKP9)*	7.5E−37	0.003	25.981	39.1%	++	4.0E−28
chr4	11026145	rs4613558	*CLNK/HS3ST1*	*(LINC02498-MIR572)*	2.1E−07	0.055	5.780	26.6%	++	3.9E−05
chr6	32615369	rs4959105	*HLA*	*(HLA-DRB1-HLA-DQA1)*	3.3E−08	0.038	6.135	30.6%	–	1.3E−04
chr6	47627419	rs1385742	*CD2AP*	*(CD2AP-ADGRF2)*	1.6E−06	0.066	5.664	35.4%	++	2.2E−08
chr7	54848587	rs75061358	*SEC61G-EGFR*	*(SEC61G-EGFR)*	1.8E−08	0.020	7.492	7.5%	–	4.7E−02
chr7	100374211	rs1859788	*PILRA/NYAP1*	*PILRA*	1.8E−10	0.024	6.769	31.7%	–	4.2E−05
chr8	27607747	rs35500730	*CLU/PTK2B*	*CLU*	4.4E−10	0.014	8.854	47.2%	–	1.7E−16
chr8	144103704	rs34173062	*SHARPIN*	*SHARPIN*	2.1E−08	0.024	6.965	7.0%	++	3.6E−02
chr10	59882255	rs536782446	*CCDC6/ANK3*	*CCDC6*	1.0E−06	0.055	5.781	44.9%	+	N/A
chr11	60303473	rs55777218	*MS4A*	*MS4A4A*	9.9E−12	0.010	10.762	37.6%	–	4.0E−14
chr11	86157598	rs3851179	*PICALM*	*(RNU6-560P-LINC02695)*	9.8E−13	0.010	11.694	36.8%	–	5.8E−16
chr11	121564744	rs529960410	*SORL1*	*SORL1*	2.8E−07	0.093	5.310	5.1%	–	2.6E−08
chr13	70473040	rs145238220	*RNU6-54P-MTCL1P1*	*(RNU6-54P-MTCL1P1)*	1.6E−08	0.020	7.493	2.0%	+-	9.5E−01
chr14	73461266	rs546214077	*NUMB-HEATR/PSEN1*	*(NUMB-HEATR4)*	1.8E−06	0.064	5.696	40.1%	+	N/A
chr15	58723762	rs6494036	*ADAM10*	*ADAM10*	1.1E−10	0.021	7.280	30.7%	–	1.1E−04
chr15	63343279	rs145859269	*APH1B*	*CA12*	1.3E−06	0.055	5.816	47.9%	–	1.2E−01
chr17	49219935	rs616338	*ABI3/ACE*	*ABI3*	9.6E−07	0.069	5.599	1.1%	+	N/A
chr19	1043639	rs3752231	*ABCA7*	*ABCA7*	4.3E−08	0.024	6.866	25.6%	++	7.4E−08
chr19	8065354	rs7351083	*CCL25-FBN3*	*(CCL25-FBN3)*	3.3E−06	0.096	5.289	38.3%	–	2.6E−01
chr19	44906745	rs769449	*APOE*	*APOE*	<1E−323	0.003	inf	12.3%	++	0.0E+00
chr19	54173120	rs34564463	*TMC4/LILRB2*	*TMC4*	8.7E−07	0.070	5.543	43.9%	++	3.5E−01

For each locus, we present the representative variant/window with the largest W-statistic. The physical positions of each variant/window are given in build hg38. The replication p value is based on summary statistics from Kunkle et al.[16] For variants that cannot be matched to Kunkle et al.,[16] we report the lead variant in LD ($r^2 > 0.4$) if there is any. We assigned each significant variant to its overlapping gene(s) or intergenic region ("gene" column). If it is within a gene, we report the gene's name; if it is intergenic, we report the upstream and downstream genes. The "locus" column presents the locus name corresponds to the one used in these previous GWASs when it is applicable, which often corresponds to the likely causal gene. NA, not applicable.

the supplemental material and methods and Table S2. Additional comparisons to the Benjamini-Hochberg (BH) procedure for FDR control can be found in Figure S2. Although there are many more associations when using the BH procedure, we have shown in He et al.[12] that the BH procedure does not properly control for complex correlations among genetic variants, which can lead to inflated FDR.

The KnockoffScreen-AL analysis identified 23 common-variant loci and 9 rare-variant loci at FDR < 0.1, corresponding to 31 unique loci. APOE locus is the only one that has both common-variant and rare-variant signals (Tables 1 and 2), although conditional analyses adjusting for *APOE*-ε2 and *APOE*-ε4 dosages show that the common-variant and rare-variant signals at the APOE locus are mainly attributed to their LD with the *APOE* alleles ε2 and ε4. Seventeen loci were genome-wide significant ($p < 5 \times 10^{-8}$) in the analysis using our variant/window-based tests. All 17 genome-wide significant loci were also FDR significant in the KnockoffScreen-AL analysis.

Among the 31 loci, 13 were previously reported in Jansen et al. and Kunkle et al.,[15,16] 3 others were reported in a recent large AD GWAS in Bellenguez et al.[23]—including *NCK2, SEC61G-EGFR*, and *SHARPIN* (Table S2)—and 1 locus has not been reported before and is located between *RNU6-54P* and *MTCL1P1*. The knockoff-based analysis based on KnockoffScreen-AL identified 14 additional loci. These included four previously reported[15,16] AD loci (*CLNK/HS3ST1, CD2AP, SORL1, ABI3/ACE*), two loci (*CCDC6/ANK3* and *TMC4/LILRB2*) recently reported in Bellenguez et al.,[23] as well as one locus located in the vicinity (within 500 kb) of a gene associated with early-onset AD

Table 2. Lead rare-variant window at each locus associated with Alzheimer disease in the UK Biobank at FDR = 0.05 and FDR = 0.10

Chr	Start	End	Locus	Gene	p	q	W	p. replication	Lead variant
chr1	211680001	211682001	*NEK2-LPGAT1*	*(NEK2-LPGAT1)*	2.3E−06	0.069	5.58	7.3E−02	rs532925975
chr2	8271001	8273001	*LINC00298*	*LINC00298, LINC00299*	2.0E−09	0.016	8.53	N/A	rs577011164
chr2	105749001	105751001	*NCK2*	*NCK2*	1.3E−06	0.089	5.34	6.1E−05	rs143080277
chr3	57873001	57875001	*SLMAP*	*SLMAP*	3.2E−06	0.084	5.41	4.5E−01	rs546538267
chr4	24552001	24554001	*DHX15*	*DHX15*	1.2E−18	0.003	17.54	7.4E−01	rs181718679
chr6	41161001	41163001	*TREM2*	*TREM2*	3.4E−07	0.038	6.15	4.4E−11	rs75932628
chr9	88852001	88854001	*MIR4289-PCNPP2*	*(MIR4289-PCNPP2)*	6.9E−07	0.044	5.96	7.8E−01	rs577667049
chr15	46118001	46120001	*SQOR-MTND5P40*	*(SQOR-MTND5P40)*	3.1E−39	0.003	33.25	7.7E−01	rs372825762
chr19	45050001	45052001	*APOE*	*CLASRP*	2.3E−06	0.069	5.58	5.6E−02	rs559118614

For each locus, we present the representative variant/window with the largest W-statistic. The physical positions of each variant/window are given in build hg38. The replication p value is based on summary statistics from Kunkle et al.,[16] aggregating all variants within the window via Cauchy's combination test. Lead variant corresponds to the strongest association with the AD proxy among the rare variants (MAF $< 1\%$, MAC ≥ 25) in the window in the UK Biobank. We assigned each significant window to its overlapping gene(s) or intergenic region ("gene" column). If it is within a gene, we report the gene's name; if it is intergenic, we report the upstream and downstream genes. The "locus" column presents the locus name corresponds to the one used in these previous GWASs when it is applicable, which often corresponds to the likely causal gene. NA, not applicable.

(*NUMB-HEATR4/PSEN1*) but for which no common variant associations with late-onset AD have been reported. The remaining seven loci included six loci identified through rare variant window association tests (*NEK2-LPGAT1, LINC00298, SLMAP, DHX15, MIR4289-PCNPP2, SQOR-MTND5P40*) and one locus identified with common variant association: *CCL25-FBN3*.

Replication based on an independent meta-analysis of clinically diagnosed AD

For the 23 common-variant loci and 9 rare-variant loci significant in the KnockoffScreen-AL analysis (31 unique loci in total; APOE locus has both common-variant and rare-variant signals), we assessed their replication using GWAS summary statistics available from Kunkle et al.,[16] an independent meta-analysis of clinically diagnosed AD with 94,437 samples. We evaluated whether the lead variant/window per locus (with largest W feature statistic in the knockoff-based analysis) reached nominal significance ($p < 0.05$) or a more stringent replication threshold $p < \frac{0.05}{23+9} = 0.0016$ based on a Bonferroni correction. For the single common variants, we also evaluated the concordance in the direction of effects. For six common variants which were not present in the replication study, three of them could be linked to a proxy variant ($r^2 > 0.4$) in the replicated study: *CLU* (rs35500730–rs1532276, $r^2 = 0.62$), *SORL1* (rs1385742–rs11218343, $r^2 = 0.89$), and *APH1B* (rs145859269–rs1039289, $r^2 = 0.40$). For rare-variant windows, we calculated the replication p value for the same window but based on summary statistics from Kunkle et al.,[16] aggregating all rare variants (MAF < 0.01) within the window via Cauchy's combination test.

Among the 23 common variants significant in the KnockoffScreen-AL analysis, 20 could be tested for replication and 16 (80%) were found to replicate at $p < 0.05$ with a concordant direction of effect (Table 1) (13/20 reached the more stringent replication threshold $p < 0.0016$). Among the ones which did not reach significance, two loci, *LILRB2* and *APH1B*, were recently reported in Bellenguez et al.[23] Among the three variants that could not be tested, two were previously published (*CCDC6* and *ABI3*).[15,24] For the rare variant window associations, only 2 out of 8 (25%) windows tested reached nominal replication ($p < 0.05$; same 2/8 reached the more stringent replication threshold $p < 0.0016$), using the Cauchy's combination test to aggregate p values of rare variants in each window based on Kunkle et al.[16] summary statistics (Table 2). This is not unexpected given the challenges in replication of rare variant signals. For the Kunkle et al.[16] study, data were imputed based on the 1000 Genomes Project reference panel including a combination of technologies such as low-coverage whole-genome sequencing (mean depth 7.4×), high-coverage whole-exome sequencing (mean depth 65.7×), and microarray genotyping as opposed to high-resolution WGS reference panels.[25] The 1000 Genomes Project (1000G) has a smaller sample size than the imputations panels used by the UK Biobank (HRC and UK10k+1000G).[2] Thus, the imputation of variants with MAF below 1% is of lower quality compared to that based on the UK Biobank data.

Using ToppFun,[26] we have also tested whether the candidate genes in Tables 1 and 2 are enriched in particular Gene Ontology (GO) molecular function, biological process, and cellular component. Interestingly, identified genes were significantly enriched for (1) molecular function (complement component C3a binding, amyloid-beta binding), (2) biological process (amyloid-beta formation, metabolic process as well as amyloid precursor protein catabolic and metabolic process), and (3) cellular component (cell surface, protein complex, clathrin coated vesicle, high-density lipoprotein particle, and endosome)

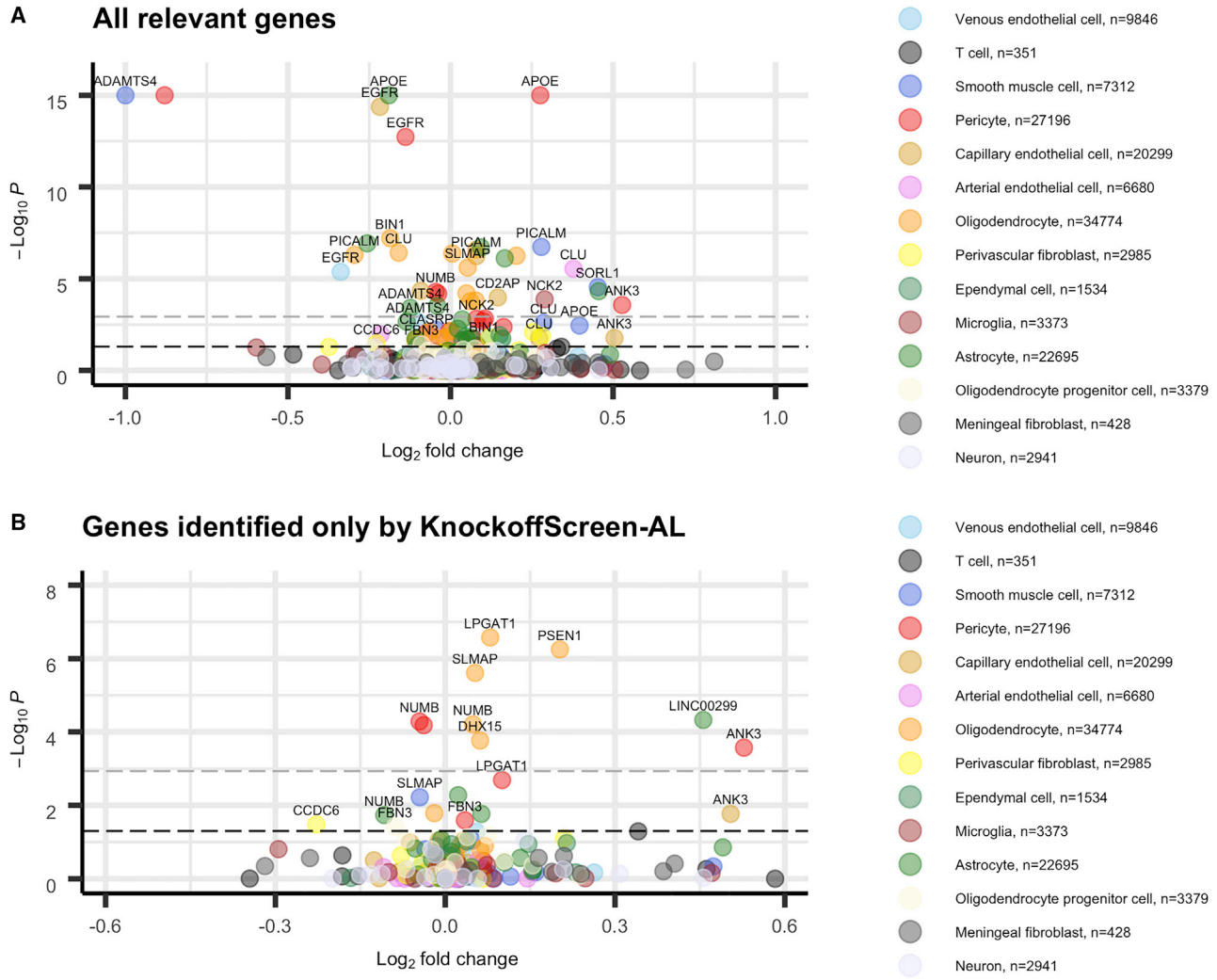

Figure 4. Single-cell RNA-seq data (n = 143,793) analysis of the 43 proximal genes
For each gene, we present the differentially expressed genes (DEG) analysis, comparing Alzheimer disease-affected individuals (AD) with healthy control subjects.
(A) All 43 proximal genes.
(B) The additional genes identified by KnockoffScreen-AL but missed by conventional association tests. Each dot represents a gene. Colors represent different cell types. The black dashed lines present p value cutoff at 0.05; the gray dashed lines present p value cutoff at 0.05/43 (number of candidate genes). For visualization purpose, −log10(p) was capped at 15 and abs(log2(fold change)) was capped at 1.0. Positive log2 fold change corresponds to higher expression level in AD.

(Table S3). These results are consistent with previous reports in Jansen et al.[15]

Single-cell transcriptomics differential expression analyses validate proximal genes

For the genes corresponding to the loci in Tables 1 and 2, we performed differentially expressed gene (DEG) analyses using single-cell RNA sequencing data (scRNA-seq) from 143,793 single-nucleus transcriptomes from 17 hippocampus samples (8 control subjects and 9 AD-affected individuals) and 8 cortex samples (4 control subjects and 4 AD-affected individuals).[27] We observed that 43 out of 59 genes are present in the scRNA-seq dataset, with 21/43 corresponding to the additional loci identified by KnockoffScreen-AL. We performed the DEG analysis stratified by 14 cell types, spanning major brain cell types (e.g., neurons, astrocytes, microglia)—but also including cell types previously missed in prior analyses[28–30] that reside in the vascular, perivascular, and meningeal compartments. These include endothelial cells, pericytes and smooth muscle cells, fibroblasts, and perivascular macrophages and T cells. We included age, batch, and cellular detection rate as covariates. We additionally adjusted for within-sample correlation by including sample dummy variables as covariates. We used this fixed effect model instead of a random effect model because the number of clusters is small relative to the total number of cells. We considered p value threshold 0.05 for suggestive signals and a more stringent Bonferroni correction 0.05/43 = 0.0012 for significant signals. Results are reported in Figure 4. More details on the data and the analyses are available in the supplemental material and methods.

We describe here genes with both large log2 fold change and high statistical significance. The results stratified by brain regions can be found in Figure S3. Overall, we observed that 31/43 (72.1%) genes exhibit suggestive signal ($p < 0.05$) in at least one cell type, a significantly higher proportion compared with the rest of the genes (41.7%; $p = 7.2 \times 10^{-5}$ by Fisher's exact test). Among the 21 genes at loci only identified by KnockoffScreen-AL, 15/21 (71.4%) exhibit suggestive signals ($p < 0.05$), similar to the proportion for the genes identified by the conventional association tests (16/22; 72.7%). *ANK3, SORL1, PSEN1, NUMB, CD2AP, LPGAT1, SLMAP*, and *DHX15* show significant difference in expression ($p < 0.05/43 = 0.0012$) in at least one cell type (Figure S3). Interestingly, *LPGAT1, SLMAP*, and *DHX15* correspond to rare-variant loci that could not be replicated above based on the Kunkle et al.[16] summary statistics.

The scRNA-seq analysis also provides insights into the cell type-specific functional effects of known AD genes, including *APOE, ADAMTS4, BIN1, PICALM, CLU*, and two recently reported genes, *NCK2* and *EGFR*.[23] Notably, *ADAMTS4* is the most significant gene with affected individuals having lower expression in smooth muscle cells (SMCs) and in pericyte cell types in both hippocampus and cortex tissues. *APOE, PICALM, CLU*, and *SORL1* show higher expression in affected individuals in SMCs, as well as arterial cells for *CLU* and pericytes for *APOE*. *APOE* expression in astrocytes is significantly lower in affected individuals which is interesting, given that astrocytes are the main cell types in which *APOE* is expressed in the brain.[31] *NCK2* expression in microglial cells are significantly higher in affected individuals compared with control subjects. *EGFR* expression is significantly lower in affected individuals in pericytes (cortex tissue) and in veinous and capillary cells (hippocampus tissue).

Differential proteomic analyses: Age effect and differential abundance across neurodegenerative diseases and stages

In addition, we performed differential protein level analysis of the proximal genes at the 31 significant loci using Stanford ADRC plasma proteomics data on individuals with neurodegenerative disorders (AD, Parkinson disease, Lewy body dementia, mild cognitive impairment) and healthy control subjects (HC) (see Appendix A). We observed that a relatively small proportion of genes can be linked to proteins. For example, among the 43 genes in the scRNA-seq dataset, 21 can be linked to a protein in this dataset, 8 of which are at the loci identified only by KnockoffScreen-AL. Since the variants at the significant loci may regulate the expression of other nearby genes and thus affect the protein levels, we expand the analysis to include all genes within a ± 200 kb region for each locus. For each protein, we regressed the log2 transformed protein level on neurodegenerative disorders status and age, adjusting for sex, visit, storage days, total protein levels per sample, and 5 principal components calculated by singular value decomposition of the residuals. Since there are multiple visits per sample, we additionally included a random intercept to account for within-subject correlation. Figure 5A shows the results for all proteins available in this dataset that are linked to genes within ± 200 kb of our candidate loci (78 linked genes in total). We performed the differential protein level analysis comparing AD to HC. We also evaluated the effect of aging on protein levels. We considered p value threshold 0.05 for suggestive signals and a more stringent Bonferroni correction $0.05/78 = 0.00064$ for significant signals. More details on the data and the analyses are available in the supplemental material and methods.

For the analysis comparing AD to HC, we observed that CD2AP, TREM2, EGFR, HS3ST1, and PTK2B exhibit suggestive signal ($p < 0.05$), with CD2AP and HS3ST1 corresponding to the loci identified only by KnockoffScreen-AL. CD2AP, EGFR, HS3ST1 and PTK2B also show suggestive signal in the scRNA-seq analysis. Particularly, the lower protein levels for EGFR in AD-affected individuals versus control subjects ($p = 0.0003$) is consistent with the findings from the scRNA-seq data. CD2AP ($p = 0.0079$) corresponds to a locus identified by KnockoffScreen-AL only; it shows significant replication p value in Kunkle et al.[16] (Table 1) and significant difference in expression in the scRNA-seq analysis (Figure 4).

For the analysis of age effect, SEC61G, NUMB, NCK2, TREM2, EGFR, and LILRB2 exhibit suggestive association with aging ($p < 0.05$). NUMB, NCK2, and EGFR also show suggestive signal in the scRNA-seq analysis. Notably, we found that EGFR ($p = 3.4 \times 10^{-5}$ for AD versus HC; $p = 2.6 \times 10^{-29}$ for age effect) and TREM2 ($p = 0.013$ for AD versus HC; $p = 2.0 \times 10^{-21}$ for age effect) exhibit consistent association with AD and aging, with EGFR protein levels significantly decreasing with age across considered neurodegenerative disorders (Figure 5E) and TREM2 protein levels significantly increasing with age (Figure 5F). In addition to the 43 proximal genes, we also observed additional genes associated with AD or age ($p < 0.05$) within the ± 200 kb region of the 31 significant loci, e.g., *USF1* (near *ADAMTS4*), *APOC1* (near *APOE*), *CFD* and *POLR2E* (near *ABCA7*), and *CR2* (near *CR1*).

In summary, the scRNA-seq and proteomic analyses show that 79.1% of all proximal genes exhibit at least suggestive signal ($p < 0.05$) in the scRNA-seq and/or proteomics analyses, a substantially higher fraction than for background genes (79.1% versus 46.2%; $p = 1.8 \times 10^{-5}$ by Fisher's exact test). Similarly, 76.2% of the ones identified only by KnockoffScreen-AL exhibit at least suggestive signal (76.2% versus 46.2%; $p = 0.0074$). These results taken together demonstrate that KnockoffScreen-AL is able to identify weaker signals, particularly rare variant loci, that are missed by conventional association tests and that are supported by evidence from expression and proteomic analyses for a potential functional effect on AD.

Colocalization between *EGFR* and gene expression traits

Given the consistent findings on *EGFR* in our UK Biobank, scRNA-seq, and proteomics analyses, we sought to

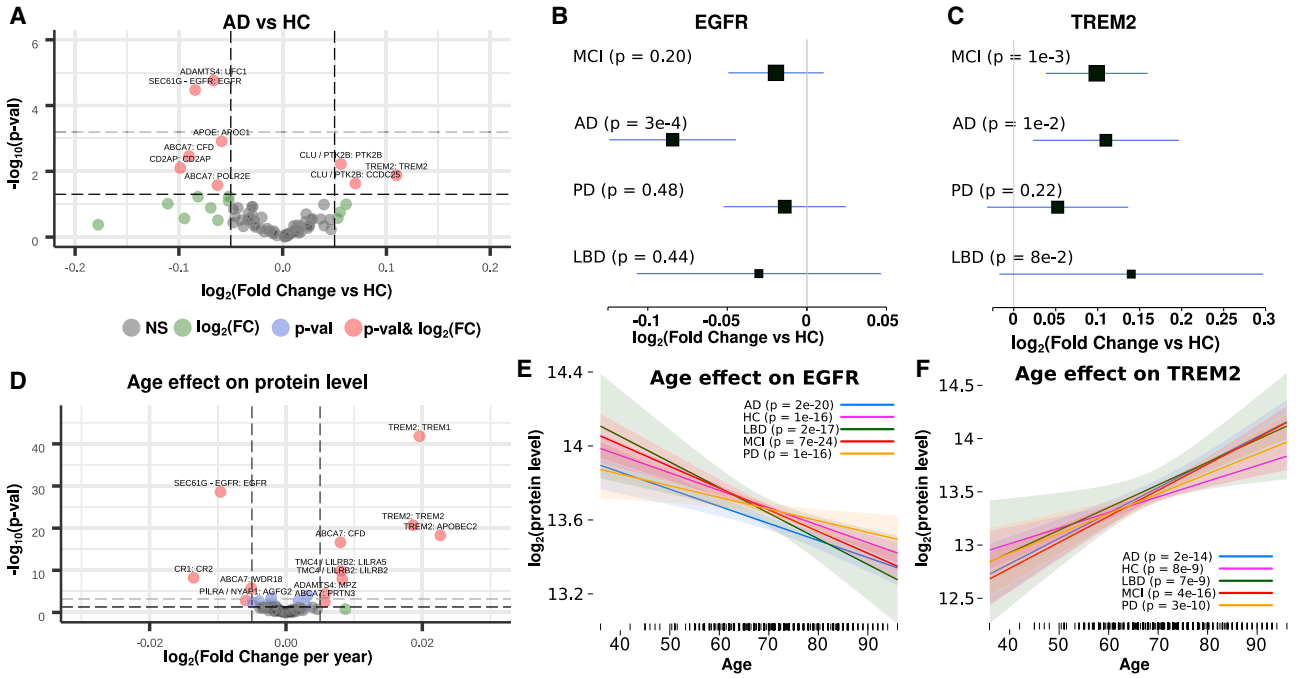

Figure 5. Proteomics data analysis of genes at the 31 significant loci
In addition to the 43 proximal genes, we additionally include genes within ±200 kb at each significant loci that can be matched with proteomics profile.
(A and D) We present the differential abundance analysis comparing Alzheimer disease (AD)-affected individuals with healthy control subjects (HC) (A) and evaluated the age effect (D). Each dot presents a gene. Different colors represent different types of significance. NS, not significant; log2FC: |log2 fold change| ≥ 0.05; p value: p value ≤ 0.05; p value and log2FC: |log2 fold change| ≥ 0.05 and p value ≤ 0.05. The dashed gray lines correspond to the Bonferroni correction p value threshold 0.05/78 = 0.00064.
(B and C) Differential abundance analysis of EGFR/TREM2.
(E and F) Age effect analysis of EGFR/TREM2. MCI, mild cognitive impairment; LBD, Lewy body dementia.

investigate the possible colocalization between brain eQTL p values and our single variant p values at *EGFR* and nearby genes (Figures 6 and S4). We report the posterior probability of colocalization (PP4) and p value computed with R package *coloc*.[32] We found that the GWAS signal at the *SEC61G-EGFR* locus colocalizes with an eQTL associated with decreased *EGFR* expression in the largest brain eQTL meta-analysis available[33] that included 1,433 individuals with brain cortex tissues from ROSMAP, MAYO, MSBB, and CommonMind (PP4 = 1, eQTL $p = 1.0 \times 10^{-27}$) and in multiple GTEx brain tissues, notably cortex (PP4 = 0.99, eQTL $p = 4.6 \times 10^{-9}$), cerebellum (PP4 = 1, eQTL $p = 4.6 \times 10^{-9}$), anterior cingulate cortex (PP4 = 0.72, eQTL $p = 2.5 \times 10^{-5}$), and caudate (PP4 = 0.81, eQTL $p = 3.0 \times 10^{-5}$). Note that the *EGFR* lead SNP in Table 1, rs75061358, was not included in GTEx. Thus, the eQTL p values are reported for the lead variant rs6979446 in high linkage disequilibrium with rs75061358 ($r^2 = 0.87$, in the 1000G EUR[25]). None of the other nearby genes at this locus showed significant colocalization in any of the eQTL datasets considered.

Discussion

We propose a computationally efficient method, KnockoffScreen-AL, for simultaneous genome-wide locus discovery and prioritization of causal variants in biobank-scale WGS/WG-imputed studies. KnockoffScreen-AL can prioritize causal variants over associations due to linkage disequilibrium and integrate multiple knockoffs for improved power, stability, and reproducibility. It allows flexible incorporation of state-of-the-art and future association tests in order to achieve the benefits of knockoff-based inference.

KnockoffScreen-AL builds upon previously developed KnockoffScreen method and uses a shrinkage algorithmic leveraging (AL) technique and several optimization strategies to make the knockoff generation feasible for biobank-scale WGS/WG-imputed studies where both the sample size and the number of variants is exceedingly large. Via simulation studies, we have demonstrated that it leads to equivalent power as the "exact" SCIT algorithm with much lower computational and memory costs. Unlike the HMM knockoff generator in KnockoffZoom that requires first phasing haplotypes and then converting the generated knockoff haplotypes back to genotypes, KnockoffScreen-AL is able to directly generate knockoffs based on unphased genotype data. Therefore, KnockoffScreen-AL can be easily applied to whole-genome studies without the need to phase, a substantial advantage given the high computational cost of phasing and the challenges in accurately phasing rare variants beyond reference panels.

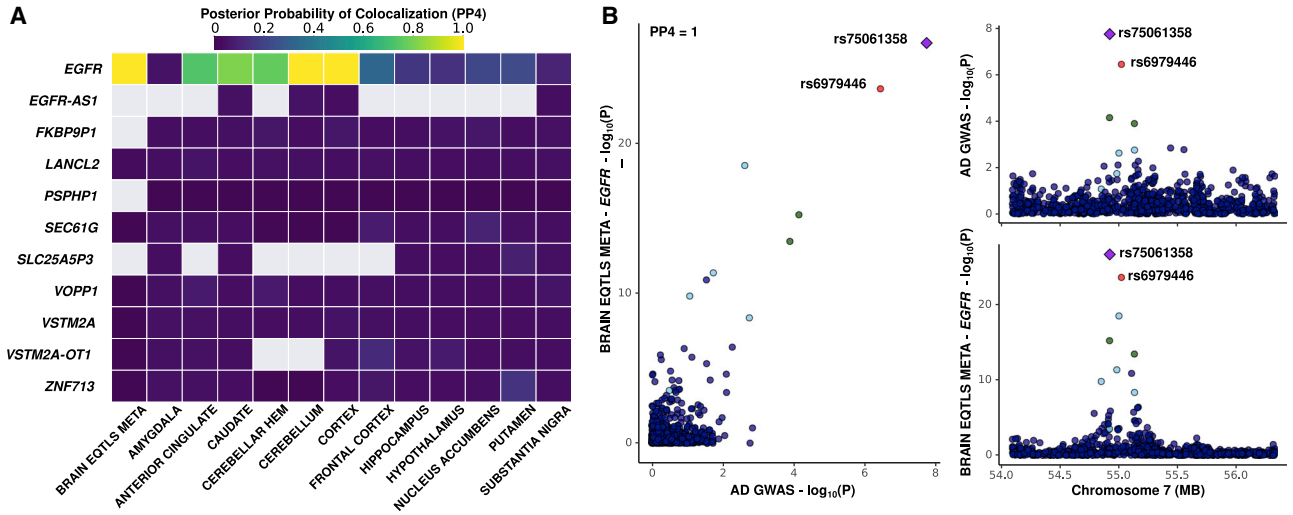

Figure 6. Colocalization analysis of *EGFR*

(A) Colocalization analysis of EGFR and nearby genes with the brain eQTLs meta-analysis and GTEx brain tissue eQTLs.

(B) Colocalization analysis of EGFR with the brain eQTLs meta-analysis. The lead variant rs75061358 and its LD linked variant rs6979446 are highlighted (red and purple, respectively).

Even though the current AD application is based on a smaller sample size compared to recent AD GWAS,[15,23,24,34,35,36] it demonstrates the ability of the proposed Knockoffscreen-AL method to handle large biobank-scale data. Compared to the analysis of the UK Biobank data in Jansen et al.,[15] the data version we used (March 2021 update of the UK Biobank resource) contains more reported AD-affected individuals (ICD10 code) and more reported parental AD (due to 2nd visits for some participants), so our single-variant/window analysis is better powered. Our analyses are also including more individuals based on different ancestry inclusion criteria. We were able to identify additional loci, such as *EGFR, NCK2*, and *SHARPIN* which achieve nominal p values in the replication study. Head-to-head comparisons of the conventional and knockoff-based analyses emphasize the higher statistical power of Knockoffscreen-AL which highlighted additional loci, including known AD loci such as *CLNK/HS3ST1, CD2AP, ABI3/ACE*, and *SORL1*. Therefore, re-analysis of large datasets such as that in de Rojas et al.[34] or other recent AD datasets[23,35,36] has great potential to discover additional loci, and the proposed optimization of computing time and memory usage makes such analyses feasible.

Our independent replication based on Kunkle et al. summary statistics[16] shows that in total 20/31 (64.5%) loci can be replicated ($p < 0.05$). Among them, 80% of the common variant associations are replicated with concordant direction of effect (Table 1). The 11 loci that cannot be replicated include common variants that are not present in the replication cohort (3/11) and rare variants not being considered by the replication study (6/11). These "unreplicated" common variant loci are likely due to the relatively low statistical power to replicate weak associations. Another plausible explanation is phenotypic heterogeneity. The UK Biobank AD-proxy, with self-reported parental AD status, is likely more heterogenous than the clinical diagnosis used in Kunkle et al.[16] We note that these loci are worth investigating in larger datasets. For example, *ABI3/ACE* is a known AD locus although its lead variant did not have a high-LD proxy in our replication study. *CCDC6/ANK3* was reported as genome-wide significant in recent large AD GWAS in Bellenguez et al. and Schwartzentruber et al.[23,24] Our differential expression analysis using scRNA-seq data provided additional support to the *CCDC6/ANK3* locus by showing that AD-affected individuals have significantly higher *ANK3* expression relative to control subjects in multiple cell types in cortex tissue. Additionally, Knockoffscreen-AL enabled analysis of rare variants which were not considered by other GWASs. As such, the identified rare-variant windows (Table 2) are harder to replicate based on the considered Kunkle et al.[16] summary statistics. However, our scRNA-seq analysis shows that several of the corresponding genes exhibit significant differences in expression (*LPGAT1, SLMAP*, and *DHX15*). These associations require additional validations in future whole-exome or whole-genome sequencing studies at scale.

We observed a high proportion of proximal genes identified by KnockoffScreen-AL exhibit suggestive effect in either scRNA-seq or proteomics analyses (Figures 4 and 5). Notably, our DEG analysis of scRNA-seq data highlighted a known main effect of *APOE* in astrocytes[31] and a significant reduced expression of *ADAMTS4* in AD-affected individuals compared with control subjects in pericyte cells within brain tissues. Proteomics analyses highlighted a rare variant locus, *TREM2*. The TREM2 protein was more abundant in AD-affected individuals, although we did not observe significant *TREM2* DEG in scRNA-seq. This effect was also observed in mild-cognitively impaired (MCI) individuals but not when comparing Parkinson disease (PD)-affected individuals to control subjects. Interestingly

TREM2 protein levels were significantly increased with age across neurodegenerative disorders. These findings likely reflect the microglial responses which are more preponderant with age and AD pathology. In addition, although many of the additional loci identified by KnockoffScreen-AL are not replicated in Kunkle et al.,[16] 76.2% of the corresponding genes exhibit suggestive signal ($p < 0.05$) in the scRNA-seq or proteomics analysis. This proportion is similar to that of the genes identified by conventional association tests and the known AD genes. The result demonstrates that KnockoffScreen-AL is able to identify weaker signals, particularly rare variant loci that are missed by conventional association tests yet potentially have a functional effect on AD.

Lastly, we characterized in more detail one genome-wide significant locus located between *SEC61G* and *EGFR*. This association with AD colocalizes with an *EGFR* eQTL in both the largest brain eQTL meta-analysis[33] and GTEx.[4] Our scRNA-seq DEG and proteomics analyses support a role for *EGFR* in AD pathology. Specifically, *EGFR* expression is significantly reduced in AD-affected individuals compared to control subjects in capillary and pericyte cells in brain. Similarly, we noted that EGFR protein levels are significantly reduced in AD-affected individuals compared to control subjects; a similar direction of effect was observed when considering MCI or Lewy-body dementia-affected individuals compared to control subjects. Furthermore, EGFR protein abundance was found to significantly decrease with age. In contrast to our observation, the Accelerating Medicine Partnership in AD (AMP-AD) reported, based on bulk RNA-seq data, that *EGFR* expression was significantly increased in AD-affected individuals compared to control subjects (Figure S5A). Additionally, the AMP-AD database (Figure S5B) also reports significant associations between *EGFR* expression level and both Braak stage and CERAD score. Interestingly, the two associations are in opposite direction: *EGFR* expression is negatively correlated with Braak stage, quantifying tau pathology, and positively correlated with CERAD score, quantifying neuritic plaques aggregates and the likeliness to have pathology corresponding to AD.

EGFR is a known oncogene with existing inhibitory therapeutics designed to curb proliferative potential and induce autophagy.[37,38] In neurodegenerative diseases, upregulation of transcriptomic and proteomic *EGFR* in multiple brain tissues has been associated with increased AD risk.[39,40] Mixed results of treating AD with *EGFR* inhibitors have been observed in animal models, often hinging on their ability to penetrate the blood-brain barrier, resulting in a reduction of reactive astrogliosis and activation of autophagy.[37,41,42] Several single-cell types implicated in maintaining the blood-brain barrier[43,44] show significant AD-associated loss of *EGFR*. Coupled with age-related loss of *EGFR* in plasma being more pronounced in AD-affected individuals versus control subjects, *EGFR* appears to play diverging roles, and its inhibition could negatively impact cell types supporting blood-brain barrier integrity.[45] Conversely, therapeutics that increase levels of EGFR outside of the brain may improve blood-brain barrier integrity but could pose an oncogenic risk. These results further highlight the importance of *EGFR* in AD progression and demonstrate that understanding spatial and cellular context of EGFR signaling will be crucial for targeted therapeutic development.

There are several limitations to the current study. First, the proposed method is developed for unrelated samples. Given the increasing number of studies that include related samples, it would be important to extend the method to handle related samples. A simple modification is to use p values from methods that account for sample relatedness, such as STAAR.[46] However, it is unclear how that affects the FDR and power because the knockoff generation may also need to be modified to account for sample relatedness. Second, the current UKBB analysis only considers genetic variation on the autosomes. The current method would need to be adapted for future analyses of X chromosome. Finally, it would be interesting to consider data-adaptive window sizes and to integrate functional scores to improve the power of UKBB analysis, as described

Appendix A

Overview of the multiple-knockoffs procedure

KnockoffScreen-AL is based on the multiple sequential knockoffs generator proposed in KnockoffScreen (M is the total number of knockoffs), which we describe below.

Algorithm 1. Sequential conditional independent tuples (multiple knockoffs)

j = 1
while $j \le p$ do
Sample $\tilde{G}_j^1, \cdots, \tilde{G}_j^M$ independently from
$\mathcal{L}(G_j \Big| \boldsymbol{G}_{-\boldsymbol{j}}, \tilde{\boldsymbol{G}}^1_{1:(\boldsymbol{j}-1)}, \cdots, \tilde{\boldsymbol{G}}^{\boldsymbol{M}}_{1:(\boldsymbol{j}-1)})$
$j = j + 1$
End

where $\boldsymbol{G}_{-\boldsymbol{j}}$ denotes all genetic variants except for the j^{th} variant; $\mathcal{L}(G_j \Big| \boldsymbol{G}_{-\boldsymbol{j}}, \tilde{\boldsymbol{G}}^1_{1:\boldsymbol{j}-1}, \cdots, \tilde{\boldsymbol{G}}^{\boldsymbol{M}}_{1:\boldsymbol{j}-1})$ is the conditional distribution of G_j given $\boldsymbol{G}_{-\boldsymbol{j}}$ and $\tilde{\boldsymbol{G}}^1_{1:(\boldsymbol{j}-1)}, \cdots, \tilde{\boldsymbol{G}}^{\boldsymbol{M}}_{1:(\boldsymbol{j}-1)}$. We consider the genetic sequence as a Markov chain with memory and approximate $\mathcal{L}(G_j \Big| \boldsymbol{G}_{-\boldsymbol{j}}, \tilde{\boldsymbol{G}}^1_{1:\boldsymbol{j}-1}, \cdots, \tilde{\boldsymbol{G}}^{\boldsymbol{M}}_{1:\boldsymbol{j}-1})$ by $\mathcal{L}(G_j \Big| \boldsymbol{G}_{\boldsymbol{k}\in\boldsymbol{B_j}}, \tilde{\boldsymbol{G}}^1_{1\le\boldsymbol{k}\le\boldsymbol{j}-1,\boldsymbol{k}\in\boldsymbol{B_j}}, \cdots, \tilde{\boldsymbol{G}}^{\boldsymbol{M}}_{1\le\boldsymbol{k}\le\boldsymbol{j}-1,\boldsymbol{k}\in\boldsymbol{B_j}})$, where the index set B_j defines a subset of genetic variants "near" the j^{th} variant, as defined in KnockoffScreen to include "K-nearest" genetic variants within a 100 kb window ($\pm$100 kb from the target variant)[47] using the absolute sample correlation coefficient $|r_{jk}|$ as a similarity measure. Specifically, we include top K variants with $|r_{jk}| > 0.05$ up to $K = n^{1/3}$,

where the choice of K ensures that the coefficient estimations achieve asymptotic normality.[48] To better account for the tightly linked variants, we adopt the same practical strategy in KnockoffScreen and perform a hierarchical clustering such that variants from two different clusters do not have a correlation greater than 0.75. We exclude variants from B_j if they are in the same cluster as the target variant. To avoid over-fitting, we additionally apply an iterative procedure to reduce the number of top variants included in the B_j until the R^2 of the prediction model is less than 0.75.

Knockoff filter to define the threshold τ and Q-value for FDR control. Similar to KnockoffScreen, we define $W = \left(T - \underset{1\le m\le M}{\text{median}} T^m\right) I_{T \ge \max\limits_{1\le m\le M} T^m}$ and

$$\tau = \min\left\{ t > 0 : \frac{\frac{1}{M} + \frac{1}{M}\#\{\kappa \ge 1, \tau \ge t\}}{\#\{\kappa = 0, \tau \ge t\}} \le q \right\}, \quad \text{(Equation A1)}$$

where $T^m = -\log p^m$; I is an indicator function, $I_{T \ge \max\limits_{1\le m\le M} T^m} = 1$ if $T \ge \max\limits_{1\le m\le M} T^m$ and 0 otherwise; $\kappa = \arg\max\limits_{0\le m\le M} T^m$ denote the index of the original (denoted as 0) or knockoff feature that has the largest importance score; $\tau = T^{(0)} - \underset{1\le m\le M}{\text{median}} T^{(m)}$ denote the difference between the largest importance score and the median of the remaining importance scores. In addition, we define the Q-value for a variant/window with statistics $\kappa = 0$ and τ as

$$q = \min_{t \le \tau} \frac{\frac{1}{M} + \frac{1}{M}\#\{\kappa \ge 1, \tau \ge t\}}{\#\{\kappa = 0, \tau \ge t\}}, \quad \text{(Equation A2)}$$

where $\frac{\frac{1}{M}+\frac{1}{M}\#\{\kappa\ge 1,\tau\ge t\}}{\#\{\kappa=0,\tau\ge t\}}$ is an estimate of the proportion of false discoveries if we are to select all windows with feature statistic $\kappa = 0, \tau \ge t$, which is the knockoff estimate of FDR. For variants/windows with $\kappa \ne 0$, we define $q = 1$ and they will never be selected. Selecting variants/windows with $W > \tau$ where τ is calculated at target FDR= α is equivalent to selecting variants/windows with $q \le \alpha$.

Low-rank regression with shared covariance structure

Based on the general SCIT algorithm described above, we iteratively fit linear regressions to estimate

$$\widehat{\boldsymbol{G}}_{\boldsymbol{j}} = \widehat{\alpha} + \sum_{k\ne j, k\in B_j} \widehat{\beta}_k \boldsymbol{G}_{\boldsymbol{k}} + \sum_{1\le m\le M} \sum_{k\le j-1, k\in B_j} \widehat{\gamma}_k^m \tilde{\boldsymbol{G}}_{\boldsymbol{k}}^{\boldsymbol{m}}.$$

We calculate the residual $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{j}} = \boldsymbol{G}_{\boldsymbol{j}} - \widehat{\boldsymbol{G}}_{\boldsymbol{j}}$ and its M permutations $\widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{j}}^{*1}, \ldots, \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{j}}^{*\boldsymbol{M}}$, and then define the knockoff feature for $\boldsymbol{G}_{\boldsymbol{j}}$ to be $\tilde{\boldsymbol{G}}_{\boldsymbol{j}}^{\boldsymbol{m}} = \widehat{\boldsymbol{G}}_{\boldsymbol{j}} + \widehat{\boldsymbol{\varepsilon}}_{\boldsymbol{j}}^{*\boldsymbol{m}}$. This iterative procedure can be time consuming when both the sample size and the number of variants is large. For simplification of notations, we assume that $\boldsymbol{G}_{\boldsymbol{j}}$ and $\tilde{\boldsymbol{G}}_{\boldsymbol{k}}^{\boldsymbol{m}}$ are already centered at 0. The least-squares estimate for regression coefficients is

$$\left(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}, \widehat{\boldsymbol{\gamma}}^{\boldsymbol{T}}\right)^T = \left[cov\left(1, \boldsymbol{G}_{\boldsymbol{B_j}}, \tilde{\boldsymbol{G}}_{\boldsymbol{B_j}}\right)\right]^{-1} \left(1, \boldsymbol{G}_{\boldsymbol{B_j}}, \tilde{\boldsymbol{G}}_{\boldsymbol{B_j}}\right)^T \boldsymbol{G}_{\boldsymbol{j}}$$

where $\boldsymbol{G}_{\boldsymbol{B_j}}$ and $\tilde{\boldsymbol{G}}_{\boldsymbol{B_j}}$ correspond to the original genetic variants $\boldsymbol{G}_{\boldsymbol{k}}$, $k \ne j, k \in B_j$ and all existing knockoffs $\tilde{\boldsymbol{G}}_{\boldsymbol{k}}^{\boldsymbol{m}}$, $k \le j-1, k \in B_j$,

$$cov\left(1, \boldsymbol{G}_{\boldsymbol{B_j}}, \tilde{\boldsymbol{G}}_{\boldsymbol{B_j}}\right) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \boldsymbol{G}_{\boldsymbol{B_j}}^{\boldsymbol{T}} \boldsymbol{G}_{\boldsymbol{B_j}} & \boldsymbol{G}_{\boldsymbol{B_j}}^{\boldsymbol{T}} \tilde{\boldsymbol{G}}_{\boldsymbol{B_j}} \\ 0 & \tilde{\boldsymbol{G}}_{\boldsymbol{B_j}}^{\boldsymbol{T}} \boldsymbol{G}_{\boldsymbol{B_j}} & \tilde{\boldsymbol{G}}_{\boldsymbol{B_j}}^{\boldsymbol{T}} \tilde{\boldsymbol{G}}_{\boldsymbol{B_j}} \end{pmatrix}$$

It is worth noting that $\boldsymbol{G}$ is a sparse matrix and $\boldsymbol{G}_{\boldsymbol{B_j}}^{\boldsymbol{T}} \boldsymbol{G}_{\boldsymbol{B_j}}$ is a submatrix of $\boldsymbol{G}^{\boldsymbol{T}} \boldsymbol{G}$, which only needs to be calculated once prior to the iterations. We can efficiently calculate $cov(1, \boldsymbol{G}_{\boldsymbol{B_j}}, \tilde{\boldsymbol{G}}_{\boldsymbol{B_j}})$ for different j, especially for the early iterations where the number of existing knockoffs is small. To calculate the inverse, we approximate $[cov(1, \boldsymbol{G}_{\boldsymbol{B_j}}, \tilde{\boldsymbol{G}}_{\boldsymbol{B_j}})]^{-1}$ by spectral decomposition

$$\left[cov\left(1, \boldsymbol{G}_{\boldsymbol{B_j}}, \tilde{\boldsymbol{G}}_{\boldsymbol{B_j}}\right)\right]^{-1} \approx V D^{-1} V^T$$

where V contains the leading eigen vectors and D is a diagonal matrix with leading eigen values in decreasing order. We choose V to explain 99.9% variation in $cov(1, \boldsymbol{G}_{\boldsymbol{B_j}}, \tilde{\boldsymbol{G}}_{\boldsymbol{B_j}})$. This low-rank approximation leverages the fact that the original genetic variants and the knockoffs are highly correlated, and it further reduces the computing cost.

Shrinkage algorithmic leveraging

One popular method for dealing with data with ultra-large sample size is sub-sampling. We propose to implement a shrinkage algorithmic leveraging method in SCIT to reduce data size before performing computations. It samples and rescales rows (samples) according to an importance sampling distribution based on the empirical statistical leverage scores.[14] In estimating $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}, \widehat{\boldsymbol{\gamma}}^{\boldsymbol{T}})$, we perform the following steps:

1. Randomly sample r rows (samples) of $(1, \boldsymbol{G}_{\boldsymbol{B_j}}, \tilde{\boldsymbol{G}}_{\boldsymbol{B_j}})$ and the corresponding elements of $\boldsymbol{G}_{\boldsymbol{j}}$, using an importance sampling distribution $\{\pi_i\}_{i=1}^n$.
2. Rescale each sampled individual by $\frac{1}{r\sqrt{\pi_i}}$ to form a weighted linear regression problem.
3. Solve the weighted linear regression to estimate $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}, \widehat{\boldsymbol{\gamma}}^{\boldsymbol{T}})$.

We adopt the shrinkage leveraging estimator[14] to choose

$$\pi_i = 0.5\pi_i^{Lev} + 0.5\pi_i^{Unif}$$

where π_i^{Unif} corresponds to a uniform distribution and π_i^{Lev} corresponds to a distribution defined by empirical statistical leverage scores. Specifically,

$$\pi_i^{Lev} = \frac{h_{ii}}{\sum h_{ii}},$$

where $h_{ii} = \sum_{j=1}^{p} U_{ij}^2$; p is the total number genetic variants and $\boldsymbol{U}$ can be the orthogonal singular vectors of $(1, \boldsymbol{G_{B_j}}, \tilde{\boldsymbol{G}}_{\boldsymbol{B_j}})$. In practice, we calculate the leverage scores based on $(1, \boldsymbol{G})$ to quantify the overall sample importance. The same leverage scores are used across SCIT iterations. To efficiently calculate the leverage scores, we compute the partial singular value decompositions by the augmented implicitly restarted Lanczos bidiagonalization algorithm of Jim Baglama and Lothar Reichel.[14,49] We propose to use the leading $\sqrt{p \log p}$ singular vectors to calculate the leverage scores, and we show that this practical choice controls FDR well in empirical simulation studies. To choose the number of individuals being sampled, we utilize the asymptotic results of Ma et al.[14] that the relative error in estimating $(\hat{\alpha}, \hat{\boldsymbol{\beta}}^{\boldsymbol{T}}, \hat{\boldsymbol{\gamma}}^{\boldsymbol{T}})$ and $\widehat{\boldsymbol{G}}_{\boldsymbol{j}}$ can be bounded if $r = O(K \log K)$, where K is the number of variants in B_j described above. In practice, we combine this with the choice of $K = n^{1/3}$, and choose $r = 10n^{\frac{1}{3}} \log n$. We perform this subsampling procedure once per 200 kb region prior to the iterations. For ultra-rare variants (minor allele counts $\leq$ 25) where a prediction model cannot be properly fitted, we compute direct permutations of the original genotype as knockoffs.

Memory-efficient matrix operation

One challenge for analyzing biobank-scale data is the huge memory cost to extract and store large genotype matrix prior to the computation. This is particularly challenging for the analysis with multiple knockoffs, where each knockoff is a copy of the original genotype matrix. We implemented memory-efficient matrix operation using shared memory and memory-mapped files based on the bigmemory R package.[49,50] During the SCIT procedure to generate knockoffs, the newly generated knockoffs are directly allocated to memory-mapped files, which can be efficiently accessed and manipulated in downstream computations. We also utilized the sparse nature of the genotype matrix to further reduce the memory cost.

Empirical power and FDR simulation

Each replicate consists of 10,000 individuals with genetic data on 1,000 genetic variants from a 200 kb region, simulated using the SKAT package. The SKAT haplotype dataset was generated using a coalescent model (COSI), mimicking the linkage disequilibrium structure of European ancestry samples. The simulations include both rare and common variants. Since the simulations here focus on method comparison for locus discovery to identify relevant clusters of tightly linked variants, we simplify the simulation design by keeping one representative variant from each tightly linked cluster. Specifically, we applied hierarchical clustering such that no two clusters have cross-correlations above a threshold value of 0.75 and then randomly choose one representative variant from each cluster to be included in the simulation study. We set 0.5% variants in the 200 kb region to be causal, all within a 10 kb signal window. Then we generated the quantitative/dichotomous trait as follows:

$$\text{Quantitative trait}: \; Y_i = X_{i1} + \beta_1 g_1 + \ldots + \beta_s g_s + \varepsilon_i^Q,$$

$$\text{Dichotomous trait}: \; g(\mu_i) = \beta_0 + X_{i1} + X_{i2} + \beta_1 g_1 + \ldots + \beta_s g_s,$$

where $X_{i1} \sim N(0,1)$, $\varepsilon_i^Q \sim N(0,3)$, $X_{i2} \sim N(0,1)$ and they are all independent; X_{i1} is the observed covariate that is adjusted in the analysis; X_{i1} is the unobserved covariate; both ε_i^Q and X_{i1} reflect unobserved variation (e.g., unmeasured environment factors that affect the disease risk); $(g_1, \ldots, g_s)$ are selected risk variants; $g(x) = \log\left(\frac{x}{1-x}\right)$ and μ_i is the conditional mean of Y_i; for dichotomous trait, β_0 is chosen such that the prevalence is 10%. We set the effect $\beta_j = \frac{a}{\sqrt{2m_j(1-m_j)}}$, where m_j is the MAF for the j^{th} variant. We define a such that the variance due to the risk variants, $\beta_1^2 var(g_1) + \ldots + \beta_s^2 var(g_s)$, is 0.03. We applied KnockoffScreen-AL to the region as described before, to analyze single common variants and 2 kb rare-variant windows. A window is considered causal if it contains at least one causal variant. For each replicate, the empirical power is defined as the proportion of detected variants/windows among all causal variants/windows; the empirical FDR is defined as the proportion of non-causal variants/windows among all detected variants/windows. We simulated 1,000 replicates and calculated the average empirical power and FDR. We present the comparison between naive SCIT and the proposed modifications in Figure 2. We also present method comparison with other existing knockoff generators (second-order, HMM) and existing tests (SKAT, burden).

Transcriptome and epigenome informed gene-based analyses

The KnockoffScreen-AL framework allows leveraging transcriptome and epigenome information for a gene-based analyses based on the genome-wide summary statistics (p values) from KnockoffScreen-AL.[51] Specifically, for each gene, we extracted p values in our UK Biobank analysis for all variants and windows overlapping with the gene and its predicted enhancer regions. We also extracted our p values for all *cis*-eQTL variants for 49 tissues in a 2 Mb nearby region (± 1 Mb from the transcription start site). The *cis*-eQTLs were identified by a previous analysis of GTEx v.8 data using an elastic-net model as in the Transcriptome Prediction Model Repository (PredictDB).[52] We aggregated all p values using the Cauchy's combination test to compute a combined p value for the gene, $p_{integrative}$. We also compute a p value restricted to those variants/windows overlapping with the gene body itself (i.e., the interval between the transcription start site and the

end of 3′ UTR), p_{gene}. We defined the feature importance score for a given gene as

$$T = -\log_{10} p_{gene} \times I_{p_{gene} > \alpha^*} - \log_{10} p_{integrative} \times I_{p_{gene} \leq \alpha^*},$$

where α^* is a threshold on the gene-based p value that can be used if we are interested only in identifying those genes that have at least some suggestive evidence of association based on the variation in the gene body itself. We set $\alpha^* = 0.0001$ in our applications. We note that this is a relatively conservative threshold as it filters out those genes that are not significant by themselves but could be significant if eQTLs or variants in predicted enhancers were included. To include all genes, we can simply use $\alpha^* = 1$ (see Figure S6 for comparison).

We computed the feature importance score for the original genetic data and its knockoff counterparts, and then applied a knockoff filter to select the genes significant at an FDR threshold of 0.1 (Table S4, Figure S6). The gene-based analysis identified similar associations as the genome-wide screening, and notably emphasized as lead genes *ADAMTS4, CR1, BIN1, TREM2, PILRA, APOE, NCK2,* and *SHARPIN* at their respective loci. It additionally highlighted *SDF2* and *CASP6* which were not part of loci identified in our genome-wide screening and *CASP6* was not identified by the conventional p value-based Bonferroni correction ($p < 2.5 \times 10^{-6}$). We calculated replication p value using MAGMA gene-based analysis using summary statistics of common and rare variants from Kunkle et al.[16] in He et al.[12] and already implemented in the KnockoffScreen-AL software package.

Data and code availability

The manuscript used data from existing studies from the UK Biobank available at https://biobank.ndph.ox.ac.uk/showcase/ and AD GWAS summary statistics available at https://www.niagads.org/datasets/ng00075. The single-cell RNA-seq data for the candidate genes are available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE163577. The proteomics data for the candidate genes are available upon request and will soon be published on its own and made publicly available.

We have implemented KnockoffScreen-AL in a computationally efficient R package that can be applied generally to the analysis of other large biobank dataset or whole-genome sequencing studies. The package can be accessed at https://cran.r-project.org/web/packages/KnockoffScreen/index.html.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.10.009.

Acknowledgments

This research was additionally supported by NIH/NIA awards AG066206 (Z.H.), AG060747 (M.D.G.), AG066515 (Z.H., T.W.-C., E.C.M., M.D.G., F.M.L.), the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie (grant agreement no. 890650, Y.L.G.), and the Alzheimer's Association (AARF-20-683984, granted to M.E.B.). This research has been conducted using the UK Biobank Resource under Application Number 45420.

We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the affected individuals, and their families. The i-Select chips was funded by the French National Foundation on Alzheimer disease and related disorders. EADI (European Alzheimer's Disease Initiative) was supported by the Labex (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Université de Lille 2, and the Lille University Hospital. GERAD/PERADES was supported by the Medical Research Council (grant number 503480), Alzheimer's Research UK (grant number 503176), the Wellcome Trust (grant number 082604/2/07/Z), and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant number 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193, the NIA AG081220 and AGES contract N01–AG–12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC (Alzheimer's Disease Genetic Consortium) was supported by the NIH/NIA grants U01 AG032984, U24 AG021886, and U01 AG016976 and the Alzheimer's Association grant ADGC–10–196728.

Declaration of interests

The authors declare no competing interests.

Received: August 9, 2021
Accepted: October 19, 2021
Published: November 11, 2021

References

1. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299.
2. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
3. Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat. Rev. Genet. *19*, 491–504.
4. Battle, A., Brown, C.D., Engelhardt, B.E., Montgomery, S.B.; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—

RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis &Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; and eQTL manuscript working group (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

5. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
6. Umans, B.D., Battle, A., and Gilad, Y. (2021). Where Are the Disease-Associated eQTLs? Trends Genet. *37*, 109–124.
7. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497–508.
8. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J. R. Stat. Soc. Series B Stat. Methodol. *82*, 1273–1300.
9. Sesia, M., Sabatti, C., and Candès, E.J. (2019). Gene hunting with hidden Markov model knockoffs. Biometrika *106*, 1–18.
10. Sesia, M., Katsevich, E., Bates, S., Candès, E., and Sabatti, C. (2020). Multi-resolution localization of causal variants across the genome. Nat. Commun. *11*, 1093.
11. Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. J. R. Stat. Soc. B *80*, 551–577.
12. He, Z., Liu, L., Wang, C., Le Guen, Y., Lee, J., Gogarten, S., Lu, F., Montgomery, S., Tang, H., Silverman, E.K., et al. (2021). Identification of putative causal loci in whole-genome sequencing data via knockoff statistics. Nat. Commun. *12*, 3152.
13. Gimenez, J.R., and Zou, J. (2019). Improving the Stability of the Knockoff Procedure: Multiple Simultaneous Knockoffs and Entropy Maximization. In The 22nd International Conference on Artificial Intelligence and Statistics (PMLR), pp. 2184–2192.
14. Ma, P., Mahoney, M.W., and Yu, B. (2015). A Statistical Perspective on Algorithmic Leveraging. J. Mach. Learn. Res. *16*, 861–911.
15. Jansen, I.E., Savage, J.E., Watanabe, K., Bryois, J., Williams, D.M., Steinberg, S., Sealock, J., Karlsson, I.K., Hägg, S., Athanasiu, L., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat. Genet. *51*, 404–413.
16. Kunkle, B.W., Grenier-Boley, B., Sims, R., Bis, J.C., Damotte, V., Naj, A.C., Boland, A., Vronskaya, M., van der Lee, S.J., Amlie-Wolf, A., et al.; Alzheimer Disease Genetics Consortium (ADGC); European Alzheimer's Disease Initiative (EADI); Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium (CHARGE); and Genetic and Environmental Risk in AD/Defining Genetic, Polygenic and Environmental Risk for Alzheimer's Disease Consortium (GERAD/PERADES) (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat. Genet. *51*, 414–430.
17. Liu, Y., Chen, S., Li, Z., Morrison, A.C., Boerwinkle, E., and Lin, X. (2019). ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am. J. Hum. Genet. *104*, 410–421.
18. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
19. Dey, R., Schmidt, E.M., Abecasis, G.R., and Lee, S. (2017). A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. Am. J. Hum. Genet. *101*, 37–49.
20. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. B *57*, 289–300.
21. Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. *78*, 629–644.
22. Morrison, A.C., Huang, Z., Yu, B., Metcalf, G., Liu, X., Ballantyne, C., Coresh, J., Yu, F., Muzny, D., Feofanova, E., et al. (2017). Practical Approaches for Whole-Genome Sequence Analysis of Heart- and Blood-Related Traits. Am. J. Hum. Genet. *100*, 205–215.
23. Bellenguez, C., Küçükali, F., Jansen, I., Andrade, V., Moreno-Grau, S., Amin, N., Naj, A.C., Grenier-Boley, B., Campos-Martin, R., Holmans, P.A., et al. (2020). New insights on the genetic etiology of Alzheimer's and related dementia. MedRxiv, 2020.10.01.20200659.
24. Schwartzentruber, J., Cooper, S., Liu, J.Z., Barrio-Hernandez, I., Bello, E., Kumasaka, N., Young, A.M.H., Franklin, R.J.M., Johnson, T., Estrada, K., et al. (2021). Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. Nat. Genet. *53*, 392–402.
25. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
26. Chen, J., Bardes, E.E., Aronow, B.J., and Jegga, A.G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. *37*, W305-11.
27. Yang, A.C., Vest, R.T., Kern, F., Lee, D.P., Maat, C.A., Losada, P.M., Chen, M.B., Agam, M., Schaum, N., Khoury, N., et al. (2021). A human brain vascular atlas reveals diverse cell mediators of Alzheimer's disease risk. bioRxiv. https://doi.org/10.1101/2021.04.26.441262.
28. Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J.Z., Menon, M., He, L., Abdurrob, F., Jiang, X., et al. (2019). Single-cell transcriptomic analysis of Alzheimer's disease. Nature *570*, 332–337.
29. Grubman, A., Chew, G., Ouyang, J.F., Sun, G., Choo, X.Y., McLean, C., Simmons, R.K., Buckberry, S., Vargas-Landin, D.B., Poppe, D., et al. (2019). A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nat. Neurosci. *22*, 2087–2097.
30. Zhou, Y., Song, W.M., Andhey, P.S., Swain, A., Levy, T., Miller, K.R., Poliani, P.L., Cominelli, M., Grover, S., Gilfillan, S., et al. (2020). Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease. Nat. Med. *26*, 131–142.

31. Harris, F.M., Tesseur, I., Brecht, W.J., Xu, Q., Mullendorff, K., Chang, S., Wyss-Coray, T., Mahley, R.W., and Huang, Y. (2004). Astroglial regulation of apolipoprotein E expression in neuronal cells. Implications for Alzheimer's disease. J. Biol. Chem. *279*, 3862–3868.
32. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.
33. Sieberts, S.K., Perumal, T.M., Carrasquillo, M.M., Allen, M., Reddy, J.S., Hoffman, G.E., Dang, K.K., Calley, J., Ebert, P.J., Eddy, J., et al.; CommonMind Consortium (CMC); and The AMP-AD Consortium (2020). Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. Sci. Data *7*, 340.
34. de Rojas, I., Moreno-Grau, S., Tesi, N., Grenier-Boley, B., Andrade, V., Jansen, I.E., Pedersen, N.L., Stringa, N., Zettergren, A., Hernández, I., et al.; EADB contributors; GR@ACE study group; DEGESCO consortium; IGAP (ADGC, CHARGE, EADI, GERAD); and PGC-ALZ consortia (2021). Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. Nat. Commun. *12*, 3417.
35. Naj, A.C., Leonenko, G., Jian, X., Grenier-Boley, B., Dalmasso, M.C., Bellenguez, C., Sha, J., Zhao, Y., Lee, S.J., van der, Sims, R., et al. (2021). Genome-Wide Meta-Analysis of Late-Onset Alzheimer's Disease Using Rare Variant Imputation in 65,602 Subjects Identifies Novel Rare Variant Locus NCK2: The International Genomics of Alzheimer's Project (IGAP). medRxiv. https://doi.org/10.1101/2021.03.14.21253553.
36. Wightman, D.P., Jansen, I.E., Savage, J.E., Shadrin, A.A., Bahrami, S., Rongve, A., Børte, S., Winsvold, B.S., Drange, O.K., Martinsen, A.E., et al. (2020). Largest GWAS (N=1,126, 563) of Alzheimer's Disease Implicates Microglia and Immune Cells. MedRxiv. https://doi.org/10.1101/2020.11.20.20235275.
37. Tavassoly, O., Sato, T., and Tavassoly, I. (2020). Inhibition of Brain Epidermal Growth Factor Receptor Activation: A Novel Target in Neurodegenerative Diseases and Brain Injuries. Mol. Pharmacol. *98*, 13–22.
38. Chong, C.R., and Jänne, P.A. (2013). The quest to overcome resistance to EGFR-targeted therapies in cancer. Nat. Med. *19*, 1389–1400.
39. Wan, Y.-W., Al-Ouran, R., Mangleburg, C.G., Perumal, T.M., Lee, T.V., Allison, K., Swarup, V., Funk, C.C., Gaiteri, C., Allen, M., et al.; Accelerating Medicines Partnership-Alzheimer's Disease Consortium (2020). Meta-Analysis of the Alzheimer's Disease Human Brain Transcriptome and Functional Dissection in Mouse Models. Cell Rep. *32*, 107908.
40. Johnson, E.C.B., Dammer, E.B., Duong, D.M., Ping, L., Zhou, M., Yin, L., Higginbotham, L.A., Guajardo, A., White, B., Troncoso, J.C., et al. (2020). Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. Nat. Med. *26*, 769–780.
41. Chen, J., He, W., Hu, X., Shen, Y., Cao, J., Wei, Z., Luan, Y., He, L., Jiang, F., and Tao, Y. (2017). A role for ErbB signaling in the induction of reactive astrogliosis. Cell Discov. *3*, 17044.
42. Tan, X., Thapa, N., Sun, Y., and Anderson, R.A. (2015). A kinase-independent role for EGF receptor in autophagy initiation. Cell *160*, 145–160.
43. Ross, J.M., Kim, C., Allen, D., Crouch, E.E., Narsinh, K., Cooke, D.L., Abla, A.A., Nowakowski, T.J., and Winkler, E.A. (2020). The Expanding Cell Diversity of the Brain Vasculature. Front. Physiol. *11*, 600767.
44. Blanchard, J.W., Bula, M., Davila-Velderrain, J., Akay, L.A., Zhu, L., Frank, A., Victor, M.B., Bonner, J.M., Mathys, H., Lin, Y.-T., et al. (2020). Reconstruction of the human blood-brain barrier in vitro reveals a pathogenic mechanism of APOE4 in pericytes. Nat. Med. *26*, 952–963.
45. Iivanainen, E., Lauttia, S., Zhang, N., Tvorogov, D., Kulmala, J., Grenman, R., Salven, P., and Elenius, K. (2009). The EGFR inhibitor gefitinib suppresses recruitment of pericytes and bone marrow-derived perivascular cells into tumor vessels. Microvasc. Res. *78*, 278–285.
46. Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K., Aslibekyan, S., et al.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium; and TOPMed Lipids Working Group (2020). 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.
47. Anderson, E.C., and Novembre, J. (2003). Finding haplotype block boundaries by using the minimum-description-length principle. Am. J. Hum. Genet. *73*, 336–354.
48. Wang, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. Ann. Stat. *39*, 389–417.
49. Baglama, J., and Reichel, L. (2005). Augmented Implicitly Restarted Lanczos Bidiagonalization Methods. SIAM J. Sci. Comput. *27*, 19–42.
50. Kane, M., Emerson, J.W., and Weston, S. (2013). Scalable Strategies for Computing with Massive Data. J. Stat. Softw. *55*, 1–19.
51. Ma, S., Dalgleish, J.L., Lee, J., Wang, C., Liu, L., Gill, R., Buxbaum, J.D., Chung, W., Aschard, H., Silverman, E.K., et al. (2021). Powerful gene-based testing by integrating long-range chromatin interactions and knockoff genotypes. medRxiv. https://doi.org/10.1101/2021.07.14.21260405.
52. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., Im, H.K.; and GTEx Consortium (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.