



GhostKnockoff inference empowers identification of putative causal variants in genome-wide association studies

Received: 26 November 2021

Accepted: 9 November 2022

Published online: 23 November 2022

 Check for updates

Zihuai He^{1,2}  , Linxi Liu³, Michael E. Belloy¹, Yann Le Guen^{1,4}, Aaron Sossin⁵, Xiaoxia Liu¹, Xinran Qi¹, Shiyang Ma⁶, Prashna K. Gyawali¹, Tony Wyss-Coray¹, Hua Tang⁷, Chiara Sabatti⁵, Emmanuel Candès^{8,9}, Michael D. Greicius¹ & Iuliana Ionita-Laza⁶

Recent advances in genome sequencing and imputation technologies provide an exciting opportunity to comprehensively study the contribution of genetic variants to complex phenotypes. However, our ability to translate genetic discoveries into mechanistic insights remains limited at this point. In this paper, we propose an efficient knockoff-based method, GhostKnockoff, for genome-wide association studies (GWAS) that leads to improved power and ability to prioritize putative causal variants relative to conventional GWAS approaches. The method requires only Z-scores from conventional GWAS and hence can be easily applied to enhance existing and future studies. The method can also be applied to meta-analysis of multiple GWAS allowing for arbitrary sample overlap. We demonstrate its performance using empirical simulations and two applications: (1) a meta-analysis for Alzheimer's disease comprising nine overlapping large-scale GWAS, whole-exome and whole-genome sequencing studies and (2) analysis of 1403 binary phenotypes from the UK Biobank data in 408,961 samples of European ancestry. Our results demonstrate that GhostKnockoff can identify putatively functional variants with weaker statistical effects that are missed by conventional association tests.

Recent advances in genome sequencing technologies and improvement in genotype imputation accuracy enable large-scale genetic studies with hundreds of thousands of samples and tens of millions of variants. The ultimate goal of such studies is to provide a credible set of putative causal variants that could lead to novel targets for the development of genomic-driven medicine. However, our ability to identify causal genetic variants and to translate genetic discoveries into mechanistic insights and drug targets remains limited at this

point¹. Conventional genome-wide association studies (GWAS) are based on marginal association models that regress an outcome of interest on a single genetic variant at a time, using Bonferroni correction for the number of independent tests to control the family-wise error rate (FWER). Although this approach based on marginal test statistics has successfully discovered many disease-associated variants, its statistical power can be suboptimal, and it often identifies proxy variants that are only correlated with the true causal variants².

¹Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA 94305, USA. ²Quantitative Sciences Unit, Department of Medicine, Stanford University, Stanford, CA 94305, USA. ³Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA. ⁴Institut du Cerveau - Paris Brain Institute - ICM, Paris 75013, France. ⁵Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA. ⁶Department of Biostatistics, Columbia University, New York, NY 10032, USA. ⁷Department of Genetics, Stanford University, Stanford, CA 94305, USA. ⁸Department of Statistics, Stanford University, Stanford, CA 94305, USA. ⁹Department of Mathematics, Stanford University, Stanford, CA 94305, USA. ✉e-mail: zihuai@stanford.edu

Multiple lines of genetic research suggest that small effect risk loci that currently lie below the genome-wide significance threshold even in large GWAS can be informative to understand complex phenotypes. First, the widely used polygenic models for complex phenotypes are based on the idea that heritability can be explained by a large number of loci, each with small or infinitesimal effects^{3,4}. Second, because of the polygenic nature of complex traits, inclusion of loci of small effects improves the predictive power of polygenic risk scores (PRS) for many traits⁵. Third, although the effect sizes of such loci are small in populations, they can still provide important biological insights. In particular, their effect on molecular phenotypes can be large and they can lead to effective drug targets (e.g. statins)⁶. However, small effect loci are difficult to distinguish from noisy loci, especially with suboptimal marginal association tests commonly used in GWAS.

Knockoff inference is a recently proposed statistical framework for variable selection in high-dimensional settings^{7,8}. Unlike marginal association testing in GWAS, the knockoff-based inference performs genome-wide conditional tests that account for linkage disequilibrium (LD) thereby reducing false positive findings due to LD confounding. It provides rigorous control of the false discovery rate (FDR), i.e. the expected proportion of discoveries which correspond to truly null hypotheses. The idea of the knockoff-based inference is to generate synthetic, noisy copies (knockoffs) of the original genetic variants that resemble the true variants in terms of preserving correlations but are conditionally independent of the disease phenotype given the original variants. The knockoffs serve as negative controls for the conditional tests to select significant genetic risk loci and to attenuate the confounding effect of LD. Unlike the conventional Benjamini-Hochberg procedure that does not account for LD, the knockoff framework appropriately accounts for arbitrary correlations among the conditional tests while guaranteeing control of the FDR⁹.

Several knockoff-based methods have already been proposed for genetic studies including Candès et al.⁷, Sesia et al.¹⁰, Sesia et al.¹¹, He et al.¹² and Sesia et al.¹³. These showed that controlling FDR can be more powerful to identify causal variants with weaker effect sizes relative to conventional GWAS, under the assumption of a polygenic model. In particular, they demonstrated that the variants identified by the knockoff inference are more likely to be the causal ones. Despite these appealing features, individual level data needed for the knockoff generation are often not available in large meta-analyses GWAS; instead, summary statistics that do not contain individual identifiable information are usually available. Furthermore, the high computational and memory cost needed to generate individual data knockoffs represents a major bottleneck to achieve its full potential. Finally, unlike for the traditional GWAS, there is currently a lack of standardized, efficient pipelines to facilitate the application of knockoff-based inference to genetic studies.

In this paper, we propose a novel method, *GhostKnockoff*, to allow efficient knockoff-based inference using freely available GWAS summary statistics for enhanced locus discovery and genome-wide prioritization of causal variants. Methodologically, we show that for the conventional score test in genetic association studies, one can directly generate the knockoff feature importance score per variant without the need to generate individual-level knockoffs for hundreds of thousands of samples. The method requires only summary statistics (i.e., Z-scores) from conventional GWAS while retaining the useful features of knockoff-based inference. The method additionally allows meta-analysis of studies with arbitrary sample overlap. We demonstrate its performance in empirical simulations and two applications: (1) meta-analysis study of Alzheimer's disease (AD) aggregating five genome-wide association studies, three whole-exome sequencing studies and one whole-genome sequencing study; (2) analysis of 1403 individual binary phenotypes from UK Biobank data on 408,961 samples with European ancestry. These analyses demonstrate the appealing properties of the proposed method for robust discovery of

additional loci and ability to localize putative causal variants at each locus. Additional discoveries made by the proposed method are further supported by functional enrichment analyses and single-cell transcriptomic analyses. The method is computationally efficient and required only 5.45 h on one central processing units (CPU) to analyze genome-wide summary statistics from the nine AD studies.

Results

Summary statistics-based multiple knockoff inference

We assume a study population of n independent individuals and p genetic variants. Let $\mathbf{X}_i = (X_{i1}, \dots, X_{iq})$ be a vector of covariates, $\mathbf{G}_i = (G_{i1}, \dots, G_{ip})$ be a vector of genotypes for the i th individual, and Y_i be the phenotype with conditional mean μ_i given \mathbf{X}_i and \mathbf{G}_i . A commonly used statistical model for modeling genetic association is the generalized linear model:

$$g(\mu_i) = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X}_i + \boldsymbol{\beta}^T \mathbf{G}_i, \quad (1)$$

where $g(\mu) = \mu$ for a continuous trait, and $g(\mu) = \text{logit}(\mu)$ for a binary trait. Without loss of generality, we assume that both phenotype and genotype are centered and standardized to have mean 0 and variance 1. If there are covariates involved, \mathbf{Y} can be centered at the conditional mean given the covariates. Conventional GWAS performs a marginal association test for each variant and controls for FWER. It tests against hypothesis $H_0 : Y_i \perp G_{ij}$ for $j=1, \dots, p$ via a score test. The per-sample score statistic can be written as $\mathbf{G}_i^T Y_i$. The Z-scores aggregating all samples can be written as

$$\mathbf{Z}_{\text{score}} = \frac{1}{\sqrt{n}} \mathbf{G}^T \mathbf{Y} \quad (2)$$

where \mathbf{G} is a $n \times p$ genotype matrix; $\mathbf{Y} = (Y_1, \dots, Y_n)^T$. Typical knockoff-based inference contains four main steps: (1) generate one or multiple knockoffs per variant and per sample; (2) calculate the feature importance score for both original and knockoff variants, e.g. square of a Z-score, $\mathbf{Z}_{\text{score}}^2$ and the knockoff counterpart $\tilde{\mathbf{Z}}_{\text{score}}^2$; (3) calculate the feature statistic by contrasting the feature importance scores for the original and knockoff variants; (4) implement the knockoff filter procedure to select significant variants with FDR control⁷. Although the Z-scores above are derived from a generalized linear model, it is worth noting that knockoff inference holds without the explicit model assumption. Unlike conventional score test or permutation analysis, knockoff-based inference performs conditional tests that account for linkage disequilibrium and controls FDR. It tests against hypothesis $H_0 : Y_i \perp G_{ij} | \mathbf{G}_{i,-j}$ for $j=1, \dots, p$, where $\mathbf{G}_{i,-j}$ includes all genetic variants except the j th variant.

The knockoff generation in step (1) can be computationally intensive. The main contribution of this paper is to show that for this particular form of feature importance score (e.g., the conventional score test as in genetic association studies), one can directly generate the knockoff feature importance score per variant without the need to generate individual-level knockoffs for hundreds of thousands of samples. Our method takes simple Z-scores as input and retains many useful features of knockoff-based inference, except for the flexibility to incorporate more sophisticated (non-linear) machine learning algorithms.

For a multiple-knockoff-based inference where each genetic variant is paired with M knockoffs, we show that the knockoff counterpart for $\mathbf{Z}_{\text{score}}$ can be directly generated by

$$\tilde{\mathbf{Z}}_{\text{score}} = \mathbf{P} \mathbf{Z}_{\text{score}} + \mathbf{E}, \text{ with } \mathbf{E} \sim \mathbf{N}(\mathbf{0}, \mathbf{V}), \quad (3)$$

where $\tilde{\mathbf{Z}}_{\text{score}} = (\tilde{\mathbf{Z}}_{\text{score}}^m)_{1mM}$ is a pM dimensional vector and each $\tilde{\mathbf{Z}}_{\text{score}}^m$ is a p dimensional vector of Z-scores corresponding to the m th group of knockoffs; \mathbf{P} and \mathbf{V} are $pM \times p$ and $pM \times pM$ matrices respectively obtained by solving a convex optimization problem (see "Methods")

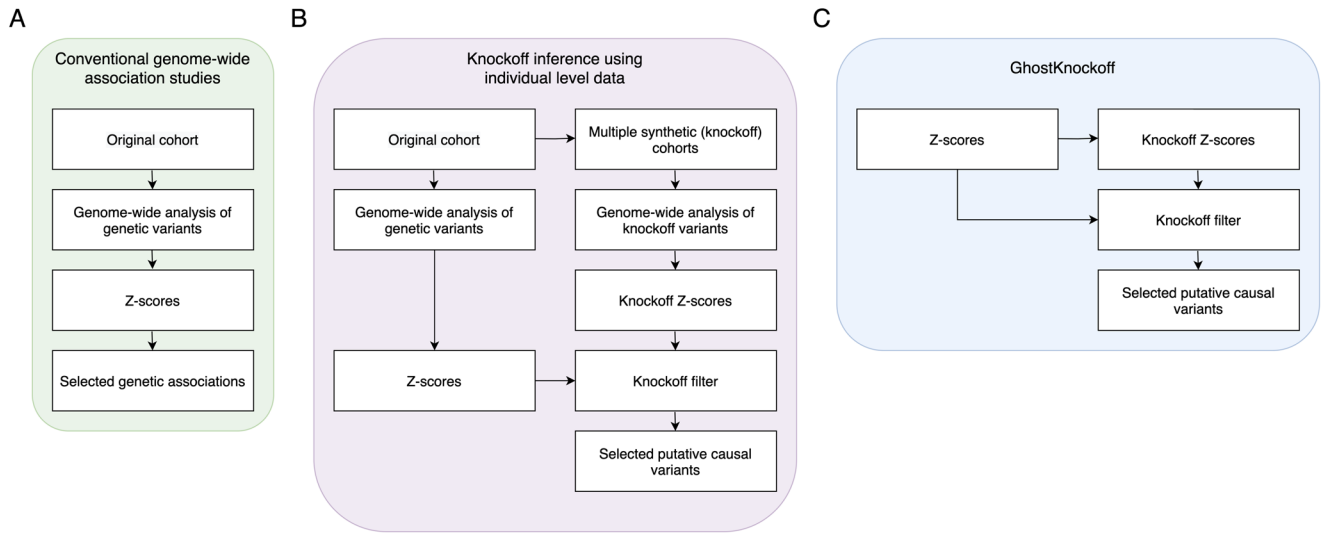


Fig. 1 | Overview of GhostKnockoff. We present the workflow of GhostKnockoff compared to conventional GWAS and knockoff inference based on same marginal test statistics using individual level data. **A** Conventional GWAS. **B** Knockoff inference using individual level data. We present the approach based on marginal test statistics. **C** The proposed GhostKnockoff using Z-scores from conventional GWAS as input.

section). Note that the matrices \mathbf{P} and \mathbf{V} are derived from the LD structure of the variants, which can be estimated by means of an external reference panel when individual-level data are not available. Intuitively, \mathbf{P} can be viewed as a “projection matrix” that maps each Z-score from a marginal test to a Z-score that quantifies indirect effects through other variants due to linkage disequilibrium. Therefore, the contrast between \mathbf{Z}_{score} and $\tilde{\mathbf{Z}}_{score}$ can prioritize causal variants that have a direct effect on the outcome of interest. This way, the Z-scores for knockoff variants can be efficiently obtained by “projecting” and sampling from a multivariate normal distribution. We show that the knockoff Z-scores generated by this approach are equivalent in distribution to those calculated based on individual-level knockoffs, thus enjoying all the desirable properties thereof. We present the details of knockoff filter in “Methods” and the workflow in Fig. 1.

GhostKnockoff was derived for a particular form of Z-score described above. In practice, we may obtain p -values from different statistical models (e.g. linear model, logistic model, mixed model etc., with different covariate adjustments) and from different tests (e.g. Wald’s test, likelihood ratio test, score test etc.). In such cases, we can apply GhostKnockoff to Z-scores obtained by an inverse normal transformation of p -values multiplied by the direction of effect, i.e., $\text{sign} \times \Phi^{-1}(pvalue/2)$. In simulations, we have observed that the FDR control remained robust as long as the p -values were computed using tests with valid type I error rate. For example, for studies with related samples, the p -values can be computed using a mixed model that accounts for sample relatedness, such as GMMAT¹⁴, SAIGE¹⁵ and fastGWA-GLMM¹⁶. We further discuss the robustness of GhostKnockoff in the Discussion section.

Meta-analysis of possibly overlapping studies

Suppose Z-scores from K independent studies with sample sizes n_1, \dots, n_K are available. We denote them as $\mathbf{Z}_{1,score}, \dots, \mathbf{Z}_{K,score}$; $N = n_1 + \dots + n_K$ is the total number of samples including possible duplicates. In general, the meta-analysis Z-score can be written as a weighted sum of individual study Z-scores

$$\mathbf{Z}_{score} = \sum_k w_k \mathbf{Z}_{k,score} \tag{4}$$

where w_k is the weight assigned to the k th study^{17,18}. The choice $w_k = \sqrt{n_k/N}$ corresponds to a conventional meta-analysis Z-score

weighted by sample size. Studies with overlapping samples are common in meta-analyses of genetic data, therefore we consider a weighting scheme that accounts for possible sample overlap and that maximizes the power of the meta-analysis. Intuitively, the optimal weights will up-weight those studies with higher independent contribution and down-weight the studies that largely overlap with others. In the “Methods” section, we show that the optimal weights w_k are given by solving

$$\text{minimize } \sum_{1 \leq i,j \leq K} w_i w_j \text{cor}.S_{ij}, \text{ subject to } \sum_k w_k \sqrt{n_k} = 1, w_k \geq 0, \tag{5}$$

where $\text{cor}.S_{ij}$ quantifies the study correlations due to sample overlap. We propose a method based on the knockoff framework to estimate $\text{cor}.S_{ij}$ (see “Methods” section). Note that, for case-control studies, n_k can be replaced by $4 / (\frac{1}{n_{k,case}} + \frac{1}{n_{k,control}})$ to better account for case-control imbalance¹⁷.

Our proposal for meta-analysis of possibly overlapping studies via knockoffs is to mimic a pooled mega-analysis, by revising the knockoff generation to account for the possibility of duplicated samples. We first consider a scenario where all studies in the meta-analysis are homogeneous and have a shared LD structure. In the meta-analysis setting we compute the knockoff Z-scores as

$$\tilde{\mathbf{Z}}_{score} = \sum_k w_k * (\mathbf{P}\mathbf{Z}_{k,score} + \gamma \mathbf{E}_k), \mathbf{E}_k \sim \mathbf{N}(\mathbf{0}, \mathbf{V}) \text{ independently for all } k \tag{6}$$

where

$$\gamma = \sqrt{1 + \frac{N}{\hat{N}_{eff}} - \frac{\hat{N}_{eff}}{N}}, \frac{\hat{N}_{eff}}{N} = \frac{\sum_k w_k^2}{\sum_{1 \leq i,j \leq K} w_i w_j \text{cor}.S_{ij}} \tag{7}$$

$\text{cor}.S_{ij}$ quantifies the correlation between studies i and j ; $\{w_k\}_{1 \leq k \leq K}$ are the solution of the above quadratic optimization problem. Note that $\gamma \geq 1$ can be thought of as a “dependency factor” that accounts for sample overlap; meta-analysis of independent studies corresponds to $\gamma = 1$, when $\text{cor}.S_{ij} = 0$, $i \neq j$ and subsequently $\hat{N}_{eff} = N$. We present the details in the “Methods” section.

Next, we consider a more general scenario where there are L groups of studies with different LD structures across groups (e.g. each

group could be of a different ancestry). We assume that each group l includes K_l (homogenous) possibly overlapping studies with sample sizes n_{lk} and Z-scores $\mathbf{Z}_{lk, score}$; $n_l = \sum n_{lk}$. In the “Methods” section, we show that the overall Z-score and its knockoff counterpart can be computed as

$$\mathbf{Z}_{score} = \frac{1}{\sqrt{N}} \sum_{1 \leq l \leq L} \sqrt{n_l} * \mathbf{Z}_{l, score}; \tilde{\mathbf{Z}}_{score} = \frac{1}{\sqrt{N}} \sum_{1 \leq l \leq L} \sqrt{n_l} * \tilde{\mathbf{Z}}_{l, score} \quad (8)$$

where for the l th group $\mathbf{Z}_{l, score} = \sum w_{lk} \mathbf{Z}_{lk, score}$; $\tilde{\mathbf{Z}}_{l, score} = \sum w_{lk} * (\mathbf{P}_l \mathbf{Z}_{lk, score} + \gamma_l \mathbf{E}_{lk})$ with $\mathbf{E}_{lk} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_l)$ independently for all l and k ; \mathbf{P}_l and \mathbf{V}_l can be obtained by solving the same convex optimization problem in “Methods” section, using the LD structure of the corresponding group. Intuitively, we perform the knockoff generation for each group separately, and then aggregate the group Z-scores to compute an overall Z-score and its knockoff.

Computational efficiency

GhostKnockoff is computationally efficient. Empirically, it only required 11.5 hours on average with one CPU to analyze a phenotype from the UK Biobank, and 5.45 hours to meta-analyze the nine AD genetic studies. This is significantly faster than the existing knockoff methods that require individual level data, which can take several days as reported in Sesia et al.¹¹ and He et al.¹². Note that both the HMM method and the SCIT method for individual level knockoff generation has a model complexity $O(np)$. By comparison, directly generating knockoff Z-scores as in GhostKnockoff has a model complexity $O(p)$. A primary gain of computational efficiency is from using a reference panel to pre-compute required matrices (\mathbf{P} and \mathbf{V} as described above) for generating knockoff Z-scores; \mathbf{P} and \mathbf{V} are pre-computed using the correlation (LD) structure estimated from a reference panel. Moreover, the random Gaussian term \mathbf{E} can be pre-sampled given \mathbf{V} . Therefore, the generation of knockoff Z-scores for a new study only involves few steps of simple matrix multiplications. For scenarios where we have Z-scores from multiple phenotypes or from multiple studies that share the same LD structure, the same pre-calculated matrices can be simultaneously applied. For new studies where the Z-scores are not readily available, there are many new advances in computing variant-level Z-scores and p -values efficiently for biobank scale data. The proposed method can leverage other analytical tools that efficiently compute variant-level Z-scores.

Power and FDR simulations

We performed simulations to empirically evaluate the performance of GhostKnockoff, which include: (1) comparing knockoff inference based on summary statistics vs. knockoff inference based on individual level data; (2) evaluating the proposed method that accounts for sample overlap; (3) comparing multiple-knockoff inference vs. single knockoff inference; (4) comparing knockoff inference vs. conventional marginal association tests in terms of the prioritization of causal variants. Note that the power comparisons between knockoff FDR control and usual FWER control in a genome-wide setting have been extensively studied by Sesia et al. (2020) and He et al. (2021), and therefore we did not focus on these existing comparisons in this paper. Instead, our simulation study focuses on method comparison in a local region. We simulated genetic data directly using whole-genome sequencing (WGS) data from the Alzheimer’s Disease Sequencing Project (ADSP). The ADSP WGS data (NG00067.v5) are jointly called by the ADSP consortium following the SNP/Indel Variant Calling Pipeline and data management tool (VCPA)¹⁹. We restricted the sampling to individuals with >80% European ancestry (estimated by SNPWeights v2.1 using reference populations from the 1000 Genomes Consortium^{20,21}). For each replicate, we randomly drew individuals for two overlapping studies with 2500 individuals per study and 2000 genetic variants

randomly selected from a 1Mb region near the *APOE* region (chr19:44,909,011-45,912,650; hg38). We then restricted the simulation studies to variants with minor allele counts > 25 and to variants that are not tightly linked. We considered three levels of sample overlap: 0% (independent), 25% (moderate) and 50% (high), with 0%, 25% and 50% samples in each study being present in the other study. Details on the simulation studies can be found in the “Methods” section.

First, we compare GhostKnockoff (GhostKnockoff-S and GhostKnockoff-M, where S or M represents single and multiple knockoff (five knockoff copies per variant), respectively) with the second-order knockoff generator proposed by Candès et al. that requires individual-level data (IndividualData Knockoff-S). We also extend the second-order knockoff generator to the multiple knockoff setting and include it in the comparison (IndividualData Knockoff-M). The methods that require individual level data were applied to an oracle pooled dataset that only contains unique samples. For a fair comparison, all methods are based on the same feature importance scores, i.e. the squared Z-scores from a marginal score test for association. We present the results in Fig. 2 (A–F). As shown, all methods have valid FDR control. In terms of power, the proposed methods based on summary statistics (GhostKnockoff-M/S) have consistent power as methods that require individual level data (IndividualData Knockoff-M/S). The results also hold when there is 25%/50% sample overlap. When there is sample overlap, the power of GhostKnockoff-M/S becomes slightly lower than IndividualData Knockoff-M/S. A likely explanation is that the study correlation in this simulation setting is over-estimated and, subsequently, the proposed method becomes slightly conservative. In real data applications, we will use genome-wide Z-scores to estimate the sample correlation, which should be more accurate. This simulation study also confirms the higher power of multiple knockoff inference (GhostKnockoff-M) relative to single knockoff inference (GhostKnockoff-S) at low target FDR (e.g. 0.05/0.10) and in the scenario with small number of causal variants (10 in our setting) as shown by Gimenez and Zou (2018)²². This is because the detection threshold of the knockoff filter (the necessary number of independent signals $\approx \frac{1}{M \times \text{target FDR}}$, where M is the number of knockoff copies per variant) is lower for multiple knockoffs compared to single knockoff. The power of multiple knockoff inference will be eventually comparable to single knockoff inference at high target FDR or in a genome-wide analyses of polygenic traits as shown in He et al.¹².

Second, we compare GhostKnockoff-M (target FDR = 0.1) with conventional marginal tests used in GWAS, adjusted by Bonferroni correction for FWER control (IndividualData marginal test-Bonferroni; target FWER = 0.05) or Benjamini-Hochberg procedure for FDR control (IndividualData marginal test- Benjamini-Hochberg; target FDR = 0.1). We show results for the setting with 25% sample overlap. Since the FDR control is more liberal than the FWER control, we focus on the prioritization of causal variants in this simulation study. For each replicate, we calculate the proportion of identified variants being causal (1-FDR) for each method. We observed that the causal proportion for GhostKnockoff-M is substantially higher than for conventional marginal tests, because GhostKnockoff-M performs a conditional test, and therefore can properly account for LD (Fig. 2G, H). Furthermore, the causal proportion using the Benjamini-Hochberg procedure is lower than that using the Bonferroni correction. Since the Benjamini-Hochberg procedure assumes positive dependence among tests, it may fail to control FDR under more complex LD structure, which might result in increased false positive rates. Therefore the conventional marginal testing with Benjamini-Hochberg adjustment does not provide a valid approach for GWAS.

Finally, we compare multiple knockoff inference (GhostKnockoff-M) with single knockoff inference (GhostKnockoff-S) in terms of the randomness due to sampling knockoff copies, referred to as stability of knockoff inference. We show results for the setting with 25% sample

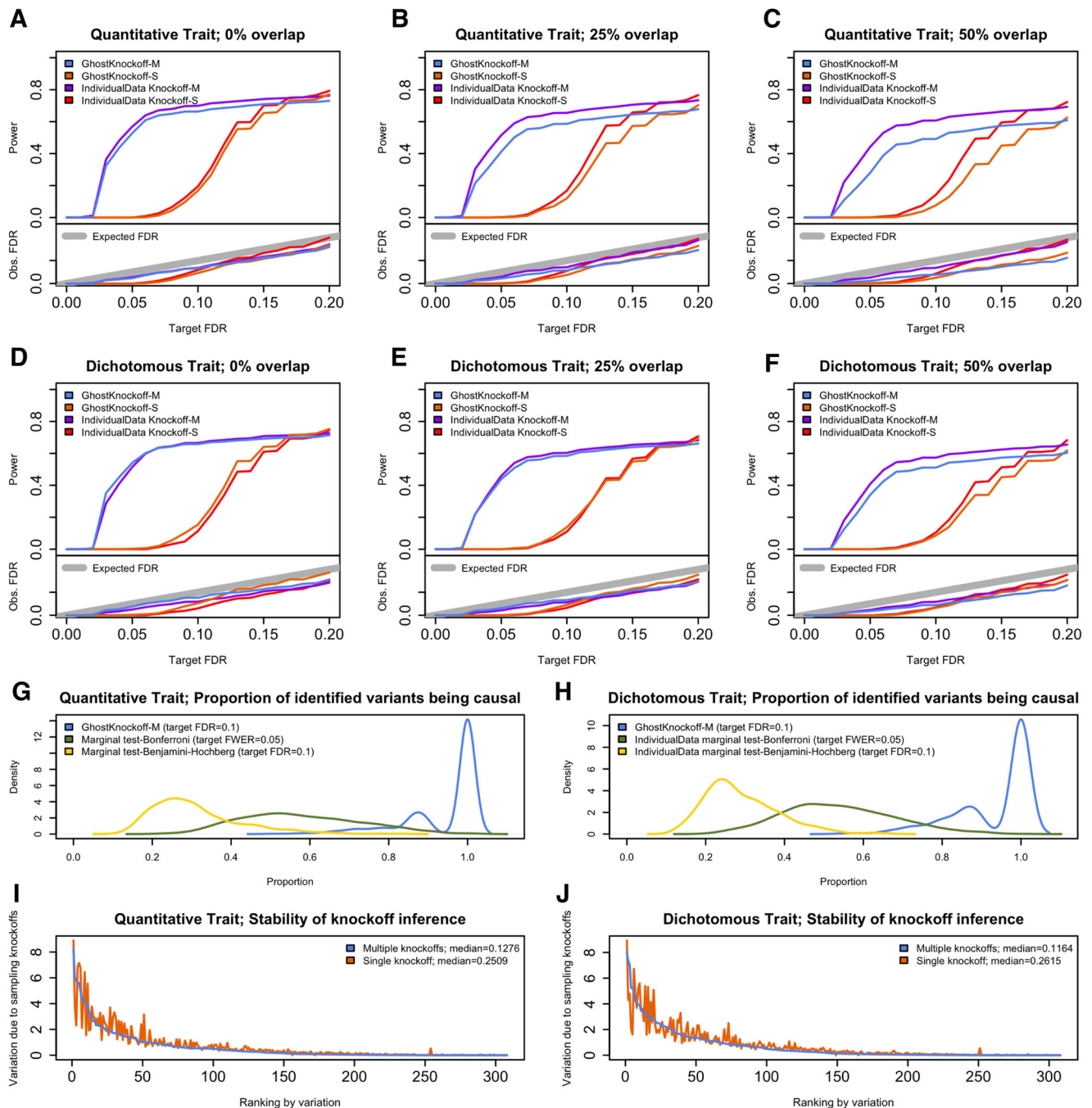


Fig. 2 | Empirical simulation studies for power, FDR and stability. Two cohorts are randomly sampled from the same population. **A–F.** Power and FDR based on 1000 replicates for different types of traits (quantitative and dichotomous) and different levels of sample overlap (0%/25%/50%), with different target FDR varying from 0 to 0.2. GhostKnockoff-M/S: the proposed multiple/single knockoff method based on the meta-analysis of Z-scores calculated separately from each individual

cohort. IndividualData Knockoff-M/S: knockoff inference based on individual level data. **G, H.** Prioritization of causal variants. **I, J.** Stability of knockoff inference, with 25% overlap and 20% unobserved variants per study. The stability is quantified as the standard deviation of feature statistics across 1000 replicates due to randomly sampling knockoffs for a given dataset.

overlap. We fixed the genotype data and phenotype data, and repeatedly performed knockoff inference. For each replicate, we computed the W-statistic for each variant. Then we calculated the standard deviation of the W-statistics per variant to quantify the variation due to sampling knockoffs. The multiple knockoff procedure shows less randomness due to sampling knockoffs compared to single knockoff inference (Fig. 2I–J).

We have also performed additional comparisons with other existing knockoff generators that require individual level data, including the knockoff generator for Hidden Markov Models

(HMMs) proposed by Sesia et al. (2019) with number of states $S = 50$, and the sequential knockoff generator proposed by He et al. (2021). We observed similar results as in comparison with the second-order knockoff generator above (Supplementary Fig. 1). In Supplementary Fig. 2, we present an empirical evaluation of the robustness to study-specific rare variants, where 10%/20% rare variants (minor allele frequency < 0.01) are randomly set to be unobserved in each cohort. We observed that the method remains valid, but requires slight modifications as described in the “Methods” section.

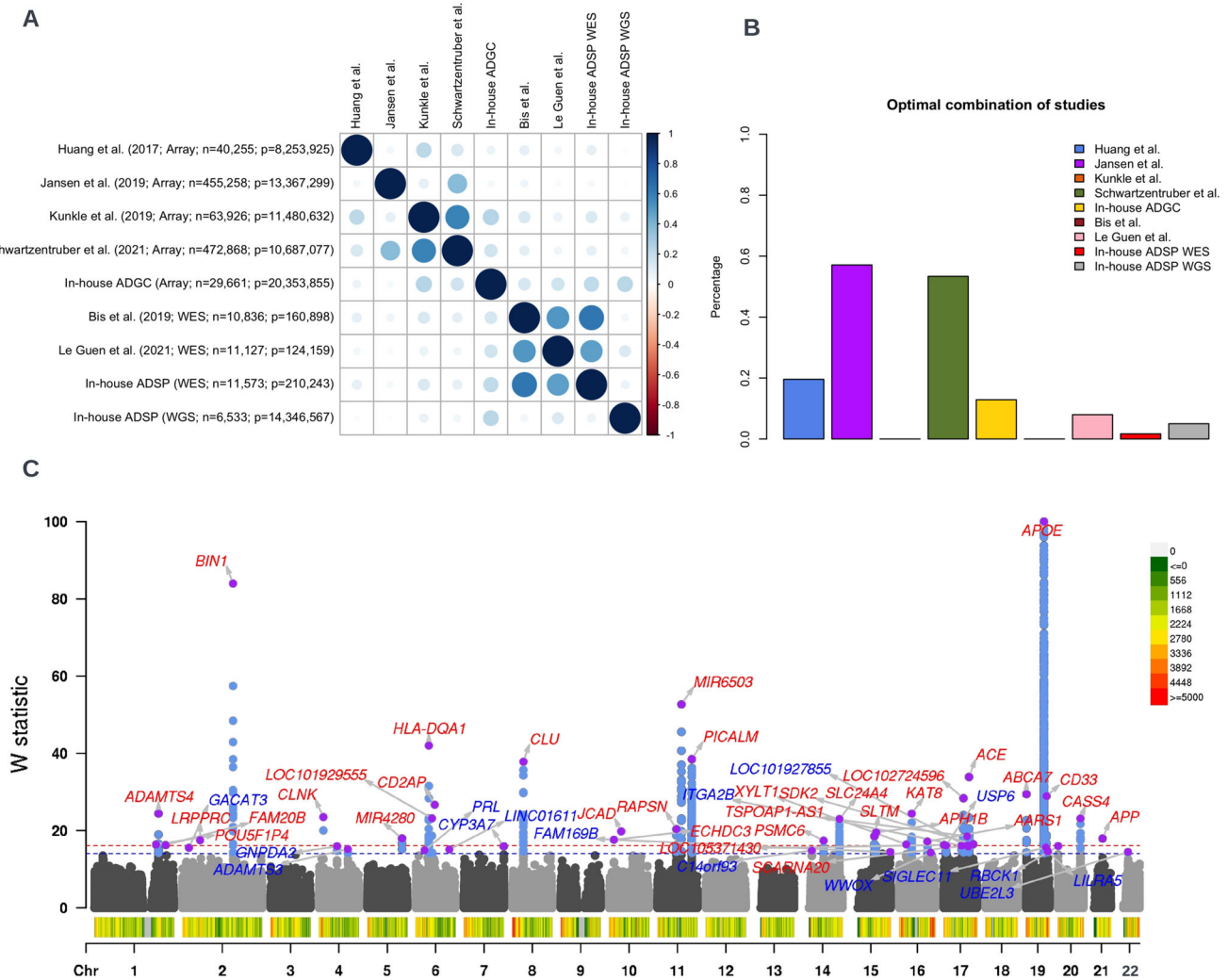


Fig. 3 | Meta-analysis of Alzheimer's disease studies. **A** Study correlations estimated using the proposed method. For each study, we present sequencing technology, sample size and number of variants. **B** Optimal combination of studies estimated using the proposed method. Each bar presents the weight per study in percentage, i.e. weight per study divided by the summation of all weights. **C** We

present the Manhattan plot of W statistics (truncated at 100 for clear visualization) from GhostKnockoff with target FDR at 0.05 (red) and 0.10 (blue). The results are based on the optimal weights combining the nine studies. Variant density is shown at the bottom of Manhattan plot (number of variants per 1Mb).

Meta-analysis of Alzheimer's disease genetics

We applied GhostKnockoff to aggregate summary statistics from nine overlapping large-scale array-based genome-wide association studies, and whole-exome/-genome sequencing studies. Specifically, the studies include (1) The genome-wide survival association study performed on 14,406 AD case samples and 25,849 control samples by Huang et al.²³; (2) The genome-wide meta-analysis of clinically diagnosed AD and AD-by-proxy (71,880 cases, 383,378 controls) by Jansen et al.²⁴; (3) The genome-wide meta-analysis of clinically diagnosed AD (21,982 cases, 41,944 controls) by Kunkle et al.²⁵; (4) The genome-wide meta-analysis by Schwartzentruber et al.²⁶, aggregating Kunkle et al. 2019 and UK Biobank based on a proxy AD phenotype; (5) In-house genome-wide associations study of 15,209 cases and 14,452 controls aggregating 27 cohorts across 39 SNP array datasets, imputed using the TOPMed reference panels²⁷; (6, 7) Two whole-exome sequencing analyses of data from The Alzheimer's Disease Sequencing Project (ADSP) by Bis et al.²⁸ (5740 cases, 5096 controls), and Le Guen et al.²⁹ (6008 cases, 5119 controls); (8) In-house whole-exome sequencing analysis of ADSP (6155 cases, 5418 controls); (9) In-house whole-genome sequencing analysis of the 2021 ADSP release³⁰ (3584 cases, 2949 controls). All studies focused on individuals with European ancestry. We used LD matrices estimated using the high coverage whole-

genome sequencing data of the expanded 1000 Genomes Project³¹. Due to the relatively small sample size of the 1000 Genomes Project (503 individuals of European ancestry) used to estimate the LD matrices, we restrict the analyses to common and low-frequency variants with minor allele frequency >1%.

We present the estimated study correlations $cor_{.S_{ij}}$ in Fig. 3A, and the estimated optimal weights in Fig. 3B. The correlation results are consistent with our knowledge of overlap and other factors, such as differences in phenotype definition, analysis strategies (e.g. statistical model), and quality control, that can affect the correlations between these studies (see more details in Supplementary Materials). Similarly, the weighting scheme up-weighted studies that are large in size and carry independent information, and down-weighted studies that largely overlap with others. In the Supplementary Materials, we discuss the consistency between the estimated study correlations and similarities in the design of these studies.

We present the results of the meta-analyses of the nine studies in Fig. 3C. We define two loci as independent if they are at least 1Mb away from each other. We adopt the most proximal gene's name as the locus name, recognizing that it is not necessarily the most likely causal gene. Our analysis identified in total 34 loci significant at FDR 0.05 and 50 loci significant at FDR 0.1. Supplementary Table 2 summarizes the 34

loci at FDR 0.05. The results show that most of the associations exhibit suggestive signals in individual studies, and most identified loci, including existing and new ones, have the same direction of effects across all studies, except very few loci where one dataset has an opposite direction of effect, although not significant, relative to other studies. Several new genes that were not reported in the latest AD meta-analyses including Jansen et al. (2019), Kunkle et al. (2019) and Schwartzenuber et al. (2021) are worth mentioning. For example, *LRPPRC* (leucine-rich pentatricopeptide repeat motif containing protein) and *APP* (Amyloid beta precursor protein) have support for their possible involvement in AD from multiple studies³². Furthermore, Hosp et al. (2015) identified *LRPPRC* as a preferential interactor of *APP* carrying the so-called Swedish mutation (*APP*^{sw}), which causes early-onset AD³³. Among the new genes, *TREML1*, *CYP3A7*, *SIGLEC11*, *IL34*, *RBCK1*, *C16orf92*, *WWOX* are within 1MB of novel loci reported in recent studies by Wightman et al. (2021)³⁴ and Bellenguez et al.³⁵

To validate the results from the knockoff inference, we adopted an alternative strategy when an independent replication study is not available. Specifically, we applied the method to a subset of samples and show that the identified variants are also replicated when we increase the sample size. We considered GhostKnockoff analysis of Kunkle et al. (2019), Schwartzenuber et al. (2021), and all nine studies. Note that data from Kunkle et al. (2019) is a subset of Schwartzenuber et al. (2021); Schwartzenuber et al. (2021) is a subset of the meta-analysis. We considered the replication of genetic variants. A genetic variant is replicated if the same variant is also identified in the larger study with the same direction of effect and a smaller *p*-value. We present the results for FDR = 0.05 and FDR = 0.10 in Supplementary Table 1 and Supplementary Fig. 3.

At FDR=0.05, we observed that 338 out of 385 (87.8%) variants by GhostKnockoff analysis of Kunkle et al. (2019) are also replicated in the analysis of Schwartzenuber et al. (2021); 447 out of 634 (70.5%) variants identified by GhostKnockoff analysis of Schwartzenuber et al. (2021) are also replicated in the proposed meta-analysis of all nine studies. At FDR=0.10, 370 out of 448 (82.6%) variants identified by GhostKnockoff analysis of Kunkle et al. (2019) are also replicated in the analysis of Schwartzenuber et al. (2021); 510 out of 724 (70.4%) variants identified by GhostKnockoff analysis of Schwartzenuber et al. (2021) are also replicated in the proposed meta-analysis of all nine studies. Overall, we conclude that the proposed GhostKnockoff (conditional test + FDR control), though systematically different from the conventional GWAS, is a valid approach to make reproducible genetic discoveries.

Finally, we present the results based on sample size weighted combination as opposed to the proposed optimal weights in Supplementary Fig. 4. We observed that GhostKnockoff with standard weights identified 29 loci at FDR 0.05 and 46 loci at FDR 0.10, while GhostKnockoff with proposed weighting scheme identified 34 loci at FDR 0.05 and 50 loci at FDR 0.10.

Single-cell transcriptomics differential expression analyses validate proximal AD genes

For the proximal genes (nearest genes to the lead variant) corresponding to the 50 loci identified at FDR 0.10 in Fig. 3, we performed differentially expressed gene (DEG) analyses using single-cell RNA sequencing data (scRNAseq) from 143,793 single-nucleus transcriptomes from 17 hippocampus (8 controls and 9 AD cases) and 8 cortex samples (4 controls and 4 AD cases)³⁶. We performed the DEG analysis stratified by 14 cell types, spanning major brain cell types (e.g., neurons, astrocytes, microglia) and cell types that reside in the vascular, perivascular, and meningeal compartments, including endothelial cells, pericytes and smooth muscle cells, fibroblasts, perivascular macrophages and T cells. We adjusted for age, batch, cellular detection rate, and 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. Among the 50 proximal genes, 38 had expression measurements in the scRNAseq dataset. We considered *p*-value threshold 0.05 for suggestive signals and a more stringent Bonferroni correction $0.05/38 = 0.0013$ for significant signals (more details on the analyses are available in the “Methods” section).

We show the scRNAseq results ($-\log_{10}(p\text{-value})$ vs. \log_2 fold change) in Fig. 4A. Overall, we observed a consistently higher proportion of differentially expressed genes for the proximal genes compared to all other 23496 genes that are observed in the scRNAseq data (background genes), across the 14 cell types (Fig. 4B). Specifically, we found that 25/38 (65.79%) genes exhibit suggestive signal ($p < 0.05$) in at least one cell type, a significantly higher proportion compared with the background genes (41.73%; $p = 4.4 \times 10^{-3}$ by Fisher's exact test; Fig. 4C). Among the genes identified by GhostKnockoff at FDR = 0.05, 69.23% exhibit suggestive signals ($p < 0.05$), similar to the proportion for the proximal genes identified at FDR = 0.10. We also observed that the DEG signals are more pronounced for the genes identified by GhostKnockoff compared to background genes (Fig. 4D; *p*-values are generally smaller), showing a strong enrichment of DEG signals for the proximal genes that reside in the loci identified by GhostKnockoff.

Phenome-wide analysis of UK Biobank data

We applied GhostKnockoff separately to each of 1403 binary phenotypes from the UK Biobank data with 408,961 white British participants (European ancestry). In this analysis, GhostKnockoff reduces to a knockoff inference based on summary statistics from a single study. We collected existing Z-scores calculated by Zhou et al. (2018) using SAIGE, a method that controls for case-control imbalance and sample relatedness⁴⁵. GhostKnockoff was applied to each phenotype separately to select associated genetic variants at FDR 0.1. Similar to the AD genetics analysis, we restrict the analyses to common and low-frequency variants with minor allele frequency $>1\%$.

We aim to illustrate two properties of GhostKnockoff inference that are systematically different from a conventional GWAS analysis. First, GhostKnockoff controls for FDR while conventional GWAS controls FWER; Second, GhostKnockoff performs a conditional test while conventional GWAS performs a marginal association test. We expect that the more liberal FDR control will lead to more associated loci, but the conditional test will reduce false positive findings due to LD confounding at each locus. For each phenotype, we count the number of independent associated loci, i.e. loci more than 1Mb away from each other. Within each locus, we count the number of genetic variants passing the FDR 0.1 threshold. We compare GhostKnockoff results and results from conventional GWAS using SAIGE (with *p*-value threshold 5×10^{-8}) in terms of the number of identified independent loci (FDR vs. FWER) and the number of identified variants per locus (conditional test vs. marginal test) (Fig. 5).

GhostKnockoff identifies generally more loci per disease phenotype (4.18-fold more discoveries on average, Figs. 5A, C) relative to GWAS. This is not surprising given that we identify significant variants at a more liberal threshold of FDR 0.1. More interesting, GhostKnockoff identifies less genetic variants within each locus (52% less variants on average after accounting for LD, Fig. 5B, C), even though FDR control is more liberal than the FWER control. We additionally present the results stratified by phecode category (phecodes are grouped into different categories as in SAIGE) in Fig. 5D, E¹⁵. Again, we observed that the proposed method consistently exhibits more loci and less genetic variants within each locus across disease categories.

To evaluate the functional effect of the identified variants, we performed functional enrichment analysis using 19 functional scores included in regBase³⁷, including: CADD³⁸, DANN³⁹, FATHMM-MKL⁴⁰, FunSeq2⁴¹, Eigen⁴², Eigen_PC⁴², GenoCanyon⁴³, FIRE⁴⁴, ReMM⁴⁵, LINSIGHT⁴⁶, fitCons⁴⁷, FATHMM-XF⁴⁸, CScape⁴⁹, CDTS⁵⁰, DVAR⁵¹, FitCons2⁵², ncER⁵³, Orion⁵⁴ and PAFA⁵⁵. All scores are on the Phred

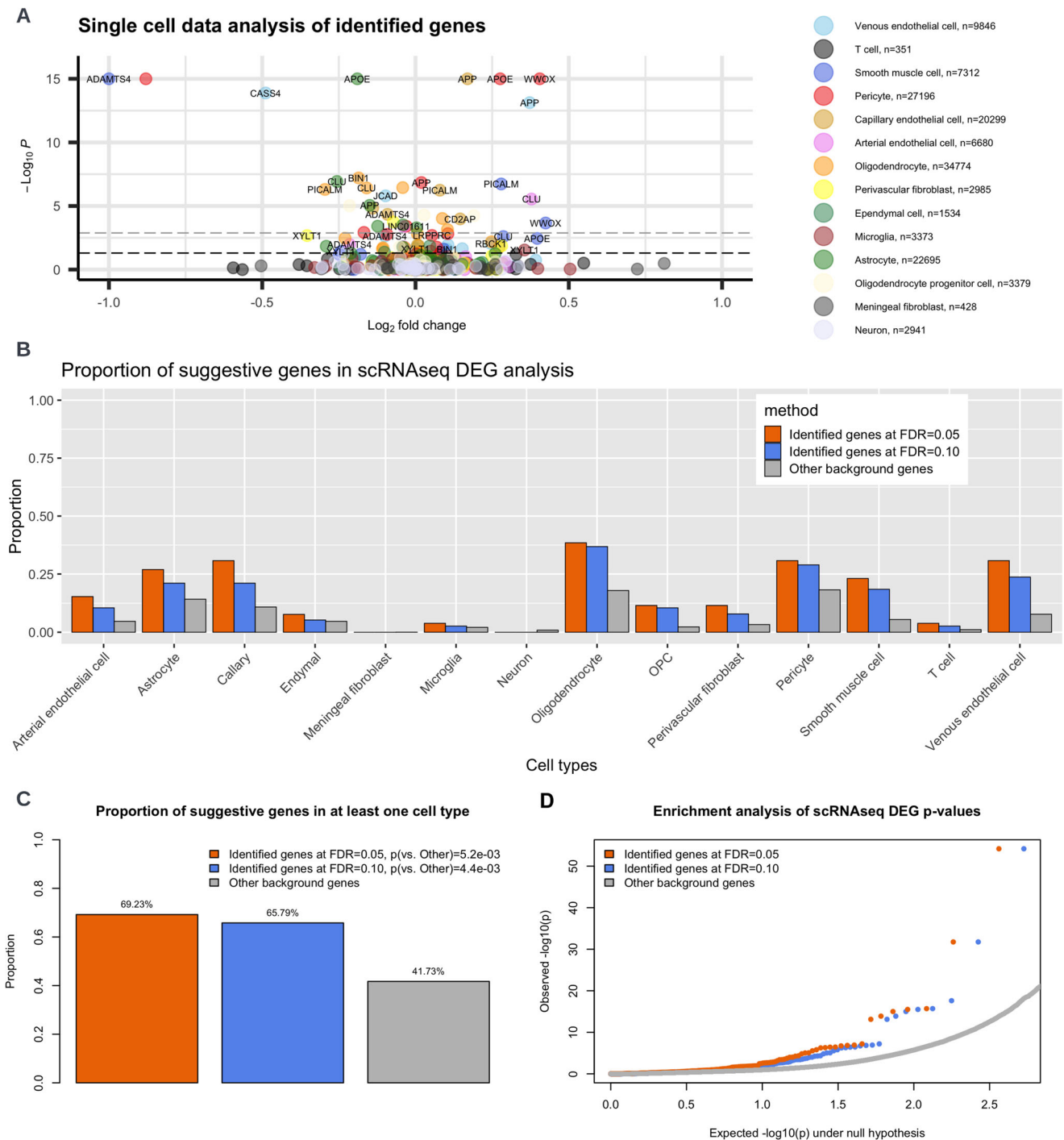


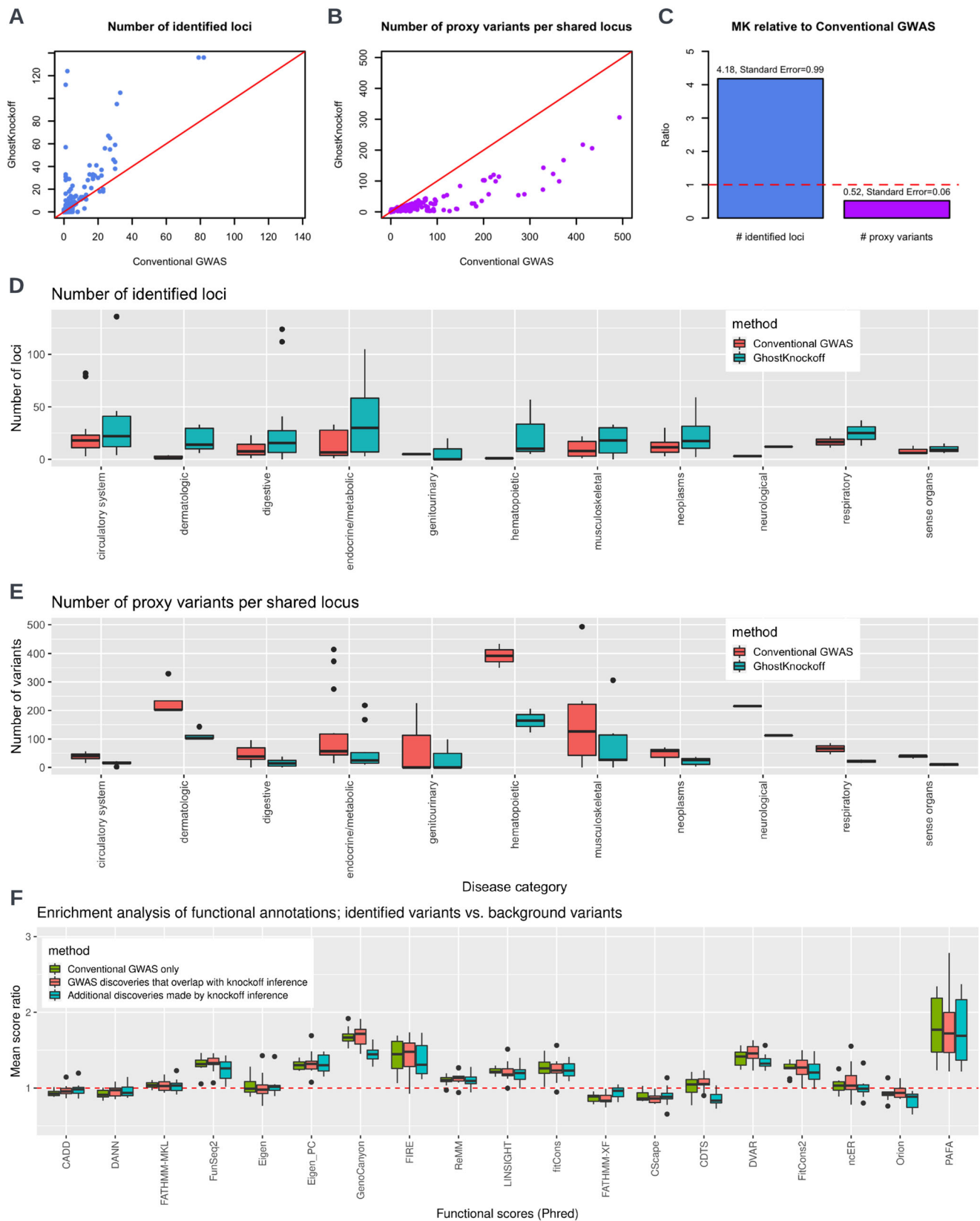
Fig. 4 | Single-cell RNAseq data ($n = 143793$) analysis of the identified proximal AD genes. **A** Differentially expressed genes (DEG) analysis using MAST implemented in Seurat, comparing Alzheimer’s disease cases (AD) with healthy controls. Each dot represents a gene. Colors represent different cell types. OPC: Oligodendrocyte progenitor cell. The black dashed line corresponds to p -value cutoff 0.05; the gray dashed line corresponds to p -value cutoff 0.05/38 (number of candidate genes) which accounts for multiple comparisons. For visualization purposes,

$-\log_{10}(p)$ values are capped at 15 and $\text{abs}(\log_2(\text{fold change}))$ values are capped at 1.0. Positive \log_2 fold change corresponds to higher expression level in AD. **B** Proportion of suggestive genes stratified by cell types. **C** Proportion of suggestive genes in at least one cell type. P -values are calculated with two-sided Fisher’s exact test. **D** Enrichment analysis of DEG nominal p -values relative to background genes. P -values are calculated by MAST implemented in Seurat.

scale. We partitioned the identified variants into three sets: 1. Variants identified by conventional GWAS only; 2. GWAS discoveries that overlap with knockoff inference; 3. Additional discoveries made by knockoff inference. Each identified variant was MAF matched with 10 randomly selected background variants on the same chromosome. For each major disease category, we calculated the ratio between the average functional score of variants in a set and the average functional

score of background variants. A ratio higher than one indicates enriched functional effects of the identified variants. We present the results in Fig. 5F.

We observed that the additional discoveries made by knockoff inference are significantly enriched in higher functional scores of FunSeq2, Eigen_PC, GenoCanyon, FIRE, ReMM, LINSIGHT, fitCons, DVAR, fitCons2, and PAFA (p -values with one-sample t-test are shown



in Supplementary Fig. 5). Note that most of these functional annotation scores were proposed to predict regulatory effects of non-coding variants. This is expected given that most variants identified by GWAS and knockoff inference are in non-coding regions. Interestingly, PAFA, which prioritizes non-coding variants associated with complex diseases, shows the highest enrichment. These results illustrate that the

additional discoveries made by knockoff inference, though weaker in terms of effect sizes, have putative regulatory effects on traits.

Discussion

We have proposed GhostKnockoff to perform knockoff-based inference without generating any individual-level knockoff variants.

Fig. 5 | Phenome-wide Analysis of 1403 binary phenotypes from UK biobank data with 408,961 white British participants with European ancestry. **A, B** Comparison between conventional GWAS and GhostKnockoff. **C** Summary of (A) and (B). For each phenotype, we calculated the ratio between the total number of identified loci/ the average number of proxy variants per shared locus by GhostKnockoff and by conventional GWAS (capped at 500 for better visualization). Panel (C) presents the average ratio (as in (A) and (B)) across 1403 phenotypes. The standard error is calculated as $\frac{\text{standard deviation of the ratio}}{\sqrt{\text{total number of phenotypes}-1}}$. **D** Distribution of the number of identified loci. We present boxplot (median and 25%/75% quantiles) for each

disease category. **E** For loci identified by both conventional GWAS and the proposed method, we present median and 25%/75% quantiles of the number of identified variants per locus. For visualization purposes, we present the results for disease phenotypes with ≥ 5 loci identified by either conventional GWAS or the proposed multiple knockoff inference for panels (D, E). **F** Functional score of variants identified by GhostKnockoff compared to that of genome-wide background variants. Each data point in the boxplot corresponds to the average score of one disease category. The boxplot presents median and 25%/75% quantiles.

GhostKnockoff can be applied to commonly available summary statistics from conventional GWAS to improve the power to identify additional, potentially weaker, associations and to prioritize the causal variants at each locus. Additionally, GhostKnockoff can be applied to meta-analysis of possibly overlapping studies. In applications to phenome-wide analyses of UK Biobank data and a meta-analysis of Alzheimer’s disease studies, we identified loci that were missed by conventional marginal association tests with improved precision. The additional discoveries are supported by functional enrichment analyses and single-cell transcriptomic analyses. These results demonstrate the improved performance of GhostKnockoff in distinguishing small effect loci that are potentially functional from noisy background genome.

As with fine-mapping studies, using in-sample LD information is best⁵⁶. Oftentimes with large meta-analyses in-sample LD is unavailable, then the LD structure can be estimated from an external reference panel. Similar to other summary statistics-based methods, the current method assumes a matched reference panel to ensure the equivalence between knockoff inference based on individual level data and GhostKnockoff. Mismatch between the LD in the target cohort and the reference panel can increase the FDR. In our empirical studies and real data analyses, we found that the LD structure estimated from the appropriate population in the 1000 Genome data is a reasonably good approximation for common and low-frequency variants (MAF $\geq 1\%$). The effect of LD misspecification is local, and therefore may affect more the ability to prioritize the causal variants at each locus, and less so the genome-wide locus discovery. For lower frequency variants, larger reference panels, such as the TOPMed, gnomAD and Pan UKBB, have become increasingly available. Recent studies have shown that imputation quality of rare variants can be significantly improved by using these reference panels. We expect that the performance of GhostKnockoff on lower frequency variants can be improved similarly by leveraging these reference panels to better estimate the LD structure^{57–59}.

Another limitation of the current method is the practical aspect on how to deal with highly correlated variants. Although the knockoff method helps to prioritize causal variants over associations due to LD, it is difficult or impossible to distinguish causal genetic variants from highly correlated variants, e.g. the variants in regions of high LD such as the major histocompatibility complex (MHC). The presence of tightly linked variants can diminish the power to identify the causal ones. The current implementation applied a hierarchical clustering of genetic variants prior to the analysis and then randomly selected one variant in each cluster. Although this strategy ensures that each variant has a representative variant included in the analysis, the statistical power can be suboptimal when the underlying causal variant is not selected. Alternatively, the group knockoff filter which groups variants and thus requires exchangeability at the group rather than variant level can be used^{60,61}. It would be of interest to incorporate group knockoffs into GhostKnockoff for improved power.

We focus on sample overlap because it is one of the main sources of study correlation, but the method can be more general to quantify study correlation due to other factors such as data generation, genotype coverage, imputation, and phenotype definition. In fact, the proposed estimation of study correlation is valid if the correlation of

genome-wide Z-scores is correctly inferred. Our proposed method for estimating the correlation is data driven. If other factors increase/decrease the correlation, we think that the data-driven estimation will remain valid. However, we do require that *i*-values from different datasets are valid, and correctly reflecting the same disease-genetic association. If the *p*-values in the original study are deflated or inflated, the meta-analysis results can be biased subsequently. In addition, the dependency factor for knockoff generation was derived based on sample overlapping. In practice, we found that the analysis based on this dependency factor reasonably reflects other factors, but the theoretical guarantee will require future investigations.

We note that GhostKnockoff was derived for a particular form of Z-score, where both features and outcomes are standardized with mean zero and standard deviation one, and samples within each study are assumed independent. In practice available Z-scores can be based on different statistical models (e.g. linear model, logistic model, mixed model etc., with different covariate adjustments) and different tests (e.g. Wald’s test, likelihood ratio test, score test etc.). Using empirical simulation studies, we observed that GhostKnockoff is robust to such variations (Fig. 2). Intuitively, the Z-scores from different analytical procedures share very similar joint distribution, with a similar marginal distribution and correlations mainly determined by LD. The proposed generation of knockoff Z-scores derived based on a particular form of Z-score $\frac{1}{\sqrt{n}}\mathbf{G}^T\mathbf{Y}$ provides a reasonable approximation and therefore the empirical FDR is under control. However, the theoretical justification of the robustness of GhostKnockoff remains unclear and it will be important to study in the future.

Methods

Knockoff-based inference using summary statistics

Our meta-analysis method is based on the second-order knockoff generator, which was initially developed for Gaussian distribution and then shown to remain empirically robust to certain deviations if the estimated second-order moments are sufficiently close to those of the underlying distribution.

The proposed knockoff-based inference using summary statistics attempts to mimic the inference based on individual-level data. For single-knockoff, we show in Supplementary Materials that the two methods are equivalent if we directly generate

$$\tilde{\mathbf{Z}}_{score} = \mathbf{P}\mathbf{Z}_{score} + \mathbf{E}, \text{ with } \mathbf{E} \sim \mathbf{N}(\mathbf{0}, \mathbf{V}) \tag{9}$$

$$\mathbf{P} = (\mathbf{I} - \mathbf{D}\boldsymbol{\Sigma}^{-1}), \mathbf{V} = 2\mathbf{D} - \mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D} \tag{10}$$

where \mathbf{I} is a $p \times p$ identity matrix; $\boldsymbol{\Sigma}$ is the correlation matrix of \mathbf{G}_i that characterizes the linkage disequilibrium; $\mathbf{D} = \text{diag}(s_1, \dots, s_p)$ is a diagonal matrix obtained by solving the following convex optimization problem:

$$\text{minimize } \sum_{j=1}^p |1 - s_j|, \text{ subject to } \begin{cases} 2\boldsymbol{\Sigma} - \mathbf{D} \succcurlyeq \mathbf{0}, \\ s_j \geq 0, 1 \leq j \leq p \end{cases} \tag{11}$$

This way, we can directly simulate the knockoff Z-scores from a multivariate normal distribution. For a given dataset, the knockoff

Z-scores will follow the same distribution as those calculated based on generating individual-level knockoffs. Therefore, the summary statistics-based knockoff generation retains the properties of knockoff-based inference.

When p is ultra-high-dimensional as in genetic studies with millions of variants, it is not feasible to operate with genome-wide correlation matrix Σ . In practice, we divided the genome into blocks that can be loaded into memory and performed the intermediate calculation of the knockoff statistics. Then the knockoff statistics from all blocks are aggregated for a genome-wide feature selection. This practical solution does not model inter-block correlation. Consequently, the current method cannot attenuate the confounding effect of long-range LD.

Extension to multiple knockoffs

Here we extend the single-knockoff approach to the case with multiple knockoffs where the original feature and multiple knockoffs are simultaneously exchangeable^{12,62}. We show in Supplementary Materials that the multiple-knockoffs counterpart for a Z-score can be directly generated by

$$\tilde{\mathbf{Z}}_{score} = \mathbf{P}\mathbf{Z}_{score} + \mathbf{E}, \text{ with } \mathbf{E} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}) \quad (12)$$

$$\mathbf{P} = \begin{pmatrix} \mathbf{I} - \mathbf{D}\Sigma^{-1} \\ \dots \\ \mathbf{I} - \mathbf{D}\Sigma^{-1} \end{pmatrix}, \mathbf{V} = \begin{pmatrix} \mathbf{C} & \mathbf{C} - \mathbf{D} & \dots & \mathbf{C} - \mathbf{D} \\ \mathbf{C} - \mathbf{D} & \mathbf{C} & \dots & \mathbf{C} - \mathbf{D} \\ \dots & \dots & \dots & \dots \\ \mathbf{C} - \mathbf{D} & \mathbf{C} - \mathbf{D} & \dots & \mathbf{C} \end{pmatrix} \quad (13)$$

where $\tilde{\mathbf{Z}}_{score}$ is a pM -dimensional vector; \mathbf{I} is a $p \times p$ identity matrix; Σ is the correlation matrix of \mathbf{G}_i that characterizes the linkage disequilibrium; $\mathbf{C} = 2\mathbf{D} - \mathbf{D}\Sigma^{-1}\mathbf{D}$; $\mathbf{D} = \text{diag}(s_1, \dots, s_p)$ is a diagonal matrix given by solving the following convex optimization problem:

$$\text{minimize } \sum_j |1 - s_j|, \text{ subject to } \begin{cases} \frac{M+1}{M}\Sigma - \mathbf{D} \succcurlyeq \mathbf{0}, \\ s_j \geq 0, 1 \leq j \leq p. \end{cases} \quad (14)$$

Knockoff filter to define the threshold τ and Q-value for FDR control

After the knockoff Z-scores are generated, we calculate the feature importance score as the element-wise square of Z-scores,

$$\mathbf{T} = \mathbf{Z}_{score}^2, \mathbf{T}^m = (\tilde{\mathbf{Z}}_{score}^m)^2 \quad (15)$$

and the knockoff statistics

$$\kappa_j = \arg \max_{0 \leq m \leq M} T_j^m, \tau_j = T_j^{(0)} - \text{median}_{1 \leq m \leq M} T_j^{(m)} \quad (16)$$

where m indicates the m th knockoff. For the j th variant, κ_j denote the index of the original (denoted as 0) or the knockoff feature that has the largest importance score; τ_j denotes the difference between the largest importance score and the median of the remaining importance scores; $T_j^{(m)}$ is corresponding to the order statistics with $T_j^{(0)} \geq \dots \geq T_j^{(m)}$. κ and τ obey the a property similar to the “flip-sign” property in the single knockoff scenario^{12,62}. In the multiple knockoff scenario, κ_j plays a role as the sign, and τ_j quantifies the magnitude that is invariant to swapping. Subsequently, we define a W -statistic to quantify the magnitude of effect on the outcome as

$$\mathbf{W} = \left(\mathbf{T} - \text{median}_{1 \leq m \leq M} \mathbf{T}^m \right) \mathbf{1}_{\mathbf{T} \geq \max_{1 \leq m \leq M} \mathbf{T}^m} \quad (17)$$

Variants with $W > \tau$ are selected, where τ is the threshold calculated by the knockoff filter. We note that this W -statistic is different from the knockoff statistics in the original model-X knockoff paper⁷. Instead, it is a function of the exact knockoff statistics that obey the “flip sign” property in the context of multiple knockoff inference. We use it as a convenient and intuitive representation of the magnitude of association. We present the exact knockoff statistics that obey the “flip sign” property and the corresponding knockoff filter in the “Methods” section.

We define the threshold for the knockoff filter as

$$\tau = \min \left\{ t > 0 : \frac{\frac{1}{M} + \frac{1}{M} \#\{\kappa_j \geq 1, \tau_j \geq t\}}{\#\{\kappa_j = 0, \tau_j \geq t\}} \leq q \right\} \quad (18)$$

In addition, we define the Q-value for a variant with statistics $\kappa = 0$ and τ as

$$q = \min_{t \leq \tau} \frac{\frac{1}{M} + \frac{1}{M} \#\{\kappa_j \geq 1, \tau_j \geq t\}}{\#\{\kappa_j = 0, \tau_j \geq t\}} \quad (19)$$

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

Meta-analysis of independent studies via knockoffs

Suppose Z-scores from K independent studies with sample sizes n_1, \dots, n_K are available, denoted as $\mathbf{Z}_{1,score}, \dots, \mathbf{Z}_{K,score}$. We define the meta-analysis Z-score as

$$\mathbf{Z}_{score} = \frac{1}{\sqrt{N}} \sum_k \sqrt{n_k} * \mathbf{Z}_{k,score} = \frac{1}{\sqrt{N}} \sum_k \mathbf{S}_k \quad (20)$$

where $N = \sum_k n_k$ is the total number of samples; \mathbf{S}_k is the score test statistic for the k -th study. Note that the meta-analysis Z-score is also a summation of sample score statistics, where the correlation structures across different studies are assumed to be the same. Following the same derivation for a single study, we can generate the knockoff feature importance by

$$\begin{aligned} \tilde{\mathbf{Z}}_{score} &= \frac{1}{\sqrt{N}} \sum_k \sqrt{n_k} * (\mathbf{P}\mathbf{Z}_{k,score} + \mathbf{E}_k) \\ &= \frac{1}{\sqrt{N}} \mathbf{P} \sum_k \sqrt{n_k} * \mathbf{Z}_{k,score} + \frac{1}{\sqrt{N}} \sum_k \sqrt{n_k} * \mathbf{E}_k, \end{aligned} \quad (21)$$

where $\mathbf{E}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ independently for all k .

Given independence between studies, $\frac{1}{\sqrt{N}} \sum_k \sqrt{n_k} * \mathbf{E}_k$ still follows the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{V})$. Therefore, the knockoff feature importance statistic at meta-analysis level can be equivalently generated as

$$\tilde{\mathbf{Z}}_{score} = \mathbf{P}\mathbf{Z}_{score} + \mathbf{E}, \text{ where } \mathbf{E} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}) \quad (22)$$

The same knockoff filter procedure as before can be applied in this setting.

Meta-analysis of possibly overlapping studies via knockoffs

The principle for meta-analysis of overlapping studies is to mimic a pooled mega-analysis where the knockoff generation should be revised to account for the presence of possibly duplicated samples. One sufficient condition for a valid knockoff inference is that for a

sample that is present in more than one study its knockoff version for different studies should be identical instead of being independently generated for each study.

Let N_{effect} be the effective number of samples, i.e. the total number of unique samples; N be the total number of records (including duplicates); $d_1, \dots, d_{N_{effect}}$ be the number of occurrences for each unique sample, $N = d_1 + \dots + d_{N_{effect}}$. The feature importance score is then defined as

$$\mathbf{Z}_{score} = \frac{1}{\sqrt{N}} \sum_k \mathbf{S}_k = \frac{1}{\sqrt{N}} \sum_{1 \leq i \leq N_{effect}} d_i * \mathbf{G}_i^T Y_i \quad (23)$$

Since $\tilde{\mathbf{G}}_i | \mathbf{G}_i \sim \mathbf{G}_i \mathbf{P}^T + \mathbf{e}_i^T$ with $\mathbf{e}_i \sim \mathbf{N}(\mathbf{0}, \mathbf{V})$,

$$\tilde{\mathbf{Z}}_{score} = \frac{1}{\sqrt{N}} \sum_k \tilde{\mathbf{S}}_k = \frac{1}{\sqrt{N}} \sum_{1 \leq i \leq N_{effect}} d_i * (\mathbf{P} \mathbf{G}_i^T Y_i + \mathbf{e}_i Y_i) \quad (24)$$

As \mathbf{e}_i 's are independent for all i , in distribution it is equivalent to generate

$$\begin{aligned} \tilde{\mathbf{Z}}_{score} &= \frac{1}{\sqrt{N}} \mathbf{P} \sum_k \mathbf{S}_k + \sqrt{\frac{\sum_{1 \leq i \leq N_{effect}} d_i^2 Y_i^2}{N}} \mathbf{E} \approx \frac{1}{\sqrt{N}} \mathbf{P} \sum_k \mathbf{S}_k \\ &+ \sqrt{\frac{\sum_{1 \leq i \leq N_{effect}} d_i^2}{N}} \mathbf{E} = \mathbf{P} \mathbf{Z}_{score} + \gamma \mathbf{E} \end{aligned} \quad (25)$$

where $\mathbf{E} \sim \mathbf{N}(\mathbf{0}, \mathbf{V})$. The approximation “ \approx ” is because $\frac{\sum_{1 \leq i \leq N_{effect}} d_i^2 Y_i^2}{N}$ is an approximation of $E(d_i^2 Y_i^2) = E(d_i^2) E(Y_i^2) = E(d_i^2)$ since Y_i has mean 0 and variance 1, and d_i and Y_i are independent. Then $E(d_i^2)$ can be approxi-

mated by $\frac{\sum_{1 \leq i \leq N_{effect}} d_i^2}{N_{effect}}$; and we define $\gamma = \sqrt{\frac{\sum_{1 \leq i \leq N_{effect}} d_i^2}{N}}$, which can be thought of as a “dependency” factor that accounts for sample overlapping. When $d_1 = \dots = d_{N_{effect}} = 1$, we have $N_{effect} = N$ and $\gamma = 1$, i.e. the scenario with independent studies. It is worth noting that our derivation is based on the following two assumptions: first, different studies are combined by using weights proportional to $\sqrt{n_k/N}$; second, each data point can be observed in different studies with the same probability.

Calculation of the dependency factor γ

The number of duplicates per sample is typically unknown. We propose an approximation of γ as

$$\gamma = \sqrt{\frac{\sum_{1 \leq i \leq N_{effect}} d_i^2}{N}} = \sqrt{\frac{N_{effect}}{N} * \frac{\sum_{1 \leq i \leq N_{effect}} d_i^2}{N_{effect}}} = \sqrt{\frac{N_{effect}}{N} * \bar{d}_i^2} \quad (26)$$

Under the assumption that the $N - N_{effect}$ duplicates are randomly distributed, d_i follows a distribution $1 + B(N - N_{effect}, \frac{1}{N_{effect}})$, where $B(\cdot)$ denotes a binomial distribution with $N - N_{effect}$ trials and success probability $\frac{1}{N_{effect}}$. Thus

$$\bar{d}_i^2 \approx E(d_i^2) = var(d_i) + E(d_i)^2 = \frac{(N - N_{effect})(N_{effect} - 1)}{N_{effect}^2} + \left(\frac{N}{N_{effect}}\right)^2 \quad (27)$$

Since N is sufficiently large,

$$\gamma = \sqrt{\frac{N}{N_{effect}} + \left(\frac{N}{N_{effect}} - 1\right) \left(\frac{N_{effect}}{N} - \frac{1}{N}\right)} \approx \sqrt{1 + \frac{N}{N_{effect}} - \frac{N_{effect}}{N}} \quad (28)$$

Study correlations and effective sample size

We propose a technique based on the proposed knockoff framework to identify study correlations due to sample overlap. The method requires GWAS summary statistics only and it naturally accounts for LD. Specifically, we first calculate the study correlation matrix

$$\mathbf{cor.S} = \mathbf{COR}(\mathbf{Z}_{1,score} - \mathbf{PZ}_{1,score}, \dots, \mathbf{Z}_{K,score} - \mathbf{PZ}_{K,score}) \quad (29)$$

where $\mathbf{Z}_{k,score} - \mathbf{PZ}_{k,score} = \mathbf{D}\Sigma^{-1}\mathbf{Z}_{k,score}$, which is the expected difference between the original Z-score and the knockoff Z-score. It quantifies the putative causal effect adjusting for nearby correlated variants. Under the null hypothesis that genetic variants are independent of the outcome of interest, the correlation between two independent studies is expected to be 0. Thus non-zero off-diagonal elements of $\mathbf{cor.S}$ quantify the sample overlap. In practice, we use variants with $|Z\text{-score}| \leq 1.96$ to calculate $\mathbf{cor.S}$ to remove the correlation due to polygenic effects.

We then estimate N_{effect} in a similar way as when estimating effective sample sizes in association studies with sample relatedness⁶³. We calculate effective sample size as

$$\hat{N}_{effect} = N * \frac{N}{\sum_{1 \leq i, j \leq K} \sqrt{n_i n_j} \mathbf{cor.S}_{ij}} \quad (30)$$

where n_i is the sample size of the i th study. For example, $\hat{N}_{effect} = N$ if all studies are independent; $\hat{N}_{effect} = N/K$ if all studies are identical. $\frac{N}{\sum_{1 \leq i, j \leq K} \sqrt{n_i n_j} \mathbf{cor.S}_{ij}}$ is the ratio of the variance of $\sum_k \mathbf{S}_k$ ignoring sample overlap over that accounting for sample overlap. Thus, we propose an approximation of γ as

$$\gamma \approx \sqrt{1 + \frac{N}{\hat{N}_{effect}} - \frac{\hat{N}_{effect}}{N}}, \quad \frac{\hat{N}_{effect}}{N} = \frac{N}{\sum_{1 \leq i, j \leq K} \sqrt{n_i n_j} \mathbf{cor.S}_{ij}} \quad (31)$$

Connection with existing meta-analysis methods that allow overlapping samples

A common approach in meta-analysis is to sum Z-scores and weight them properly based on sample sizes, i.e.^{16,17}

$$\mathbf{Z}_{score} = \frac{1}{\sqrt{N}} \sum_k \sqrt{n_k} * \mathbf{Z}_{k,score} \quad (32)$$

When the Z-scores are independent, \mathbf{Z}_{score} follows $N(0,1)$ under the null hypothesis. When there are overlapping samples, the variance is no longer 1. Instead,

$$var(\mathbf{Z}_{score}) = \frac{1}{N} \sum_{1 \leq i, j \leq K} \sqrt{n_i n_j} cov(\mathbf{Z}_{i,score}, \mathbf{Z}_{j,score}) \quad (33)$$

$var(\mathbf{Z}_{score}) > 1$ when there are overlapping studies and can be used to compute an effective sample size. Specifically, the \mathbf{Z}_{score} should be reduced to

$$\mathbf{Z}'_{score} = \sqrt{\frac{N_{eff}}{N}} * \frac{1}{\sqrt{N}} \sum_k \sqrt{n_k} * \mathbf{Z}_{k,score} \quad (34)$$

where

$$N_{eff} = \frac{N}{var(\mathbf{Z}_{score})} = N * \frac{N}{\sum_{1 \leq i, j \leq K} \sqrt{n_i n_j} cov(\mathbf{Z}_{i,score}, \mathbf{Z}_{j,score})} \quad (35)$$

$cov(Z_{i,score}, Z_{j,score})$ quantifies the sample overlap between studies i and j , and can be estimated using genome-wide Z-scores. Note that individual Z-scores are normalized, therefore $cov(Z_{i,score}, Z_{j,score}) = cor(Z_{i,score}, Z_{j,score})$. As discussed above, in the proposed approach, we estimate it using the knockoff method that accounts for LD.

Meta-analysis of overlapping studies with heterogeneous LD structure

Suppose there are L groups (e.g. different ancestries) with different LD structure $\Sigma_1, \dots, \Sigma_L$, and each group includes K_l overlapping studies with sample size n_{lk} and Z-scores $\mathbf{Z}_{lk,score}$; $n_l = \sum n_{lk}$. We assume that each group contains studies with the same LD structure. The overall Z-score is computed as

$$\mathbf{Z}_{score} = \frac{1}{\sqrt{N}} \sum_{l,k} \sqrt{n_{lk}} * \mathbf{Z}_{lk,score} = \frac{1}{\sqrt{N}} \sum_l \sqrt{n_l} * \mathbf{Z}_{l,score} \tag{36}$$

where the Z-score for each group is $\mathbf{Z}_{l,score} = \frac{1}{\sqrt{n_l}} \sum \sqrt{n_{lk}} * \mathbf{Z}_{lk,score}$. Let $I_{effect,l}$ be the index of unique samples corresponding to the l th group, and N_l be its size. The knockoff Z-score can be obtained as

$$\tilde{\mathbf{Z}}_{score} = \frac{1}{\sqrt{N}} \sum_{1 \leq i \leq N_{effect}} d_i * \tilde{\mathbf{G}}_i^T \mathbf{Y}_i = \frac{1}{\sqrt{N}} \sum_{1 \leq l \leq L} \sum_{i \in I_{effect,l}} d_i * (\mathbf{P}_l \mathbf{G}_i^T \mathbf{Y}_i + \mathbf{e}_{il} \mathbf{Y}_i), \tag{37}$$

where $\mathbf{e}_{il} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_l)$ independently for all i and l . Note that $\sum_{i \in I_{effect,l}} d_i \mathbf{Y}_i \mathbf{e}_{il}$

still follows a normal distribution $\sqrt{\sum_{i \in I_{effect,l}} d_i^2 \mathbf{Y}_i^2} \mathcal{N}(\mathbf{0}, \mathbf{V}_l)$, and in distribution

$$\begin{aligned} \tilde{\mathbf{Z}}_{score} &= \frac{1}{\sqrt{N}} \sum_{1 \leq l \leq L} \sqrt{N_l} * \mathbf{P}_l \mathbf{Z}_{l,score} + \frac{1}{\sqrt{N}} \sum_{1 \leq l \leq L} \sqrt{N_l} * \sqrt{\frac{\sum_{i \in I_{effect,l}} d_i^2 \mathbf{Y}_i^2}{N_l}} \mathbf{E}_l \\ &= \frac{1}{\sqrt{N}} \sum_{1 \leq l \leq L} \sqrt{N_l} * (\mathbf{P}_l \mathbf{Z}_{l,score} + \gamma_l \mathbf{E}_l) =: \frac{1}{\sqrt{N}} \sum_{1 \leq l \leq L} \sqrt{N_l} * \tilde{\mathbf{Z}}_{l,score} \end{aligned} \tag{38}$$

where $\mathbf{E}_l \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_l)$ independently for all l . Thus, the knockoff Z-score can be generated by

$$\tilde{\mathbf{Z}}_{score} = \frac{1}{\sqrt{N}} \sum_{1 \leq l \leq L} \sqrt{N_l} * \tilde{\mathbf{Z}}_{l,score} \tag{39}$$

which means that it can be generated by the weighted summation of the knockoff Z-scores generated for each individual group.

Meta-analysis with optimal weights

When meta-analyzing studies with overlapping samples, we would like to down-weight studies that are largely overlapping with others. In general, the proposed meta-analysis approach can be written as

$$\mathbf{Z}_{score} = \sum_k w_k \mathbf{Z}_{k,score} = \sum_k w_k \sqrt{n_k} * \frac{1}{\sqrt{n_k}} \mathbf{Z}_{k,score}. \tag{40}$$

Assuming that the effect sizes per variant from different studies are the same, $\frac{1}{\sqrt{n_k}} \mathbf{Z}_{k,score}$ is entry-wise in the same order of $\boldsymbol{\mu}$ where $\boldsymbol{\mu}$ is constant that quantifies the marginal association under the alternative hypothesis. We aim to maximize

$$\frac{\mathbf{Z}_{score}}{sd(\mathbf{Z}_{score})} = \frac{\sum_k w_k \mathbf{Z}_{k,score}}{\sqrt{\sum_{1 \leq i,j \leq K} w_i w_j cor.S_{ij}}} \sim \frac{\sum_k w_k \sqrt{n_k}}{\sqrt{\sum_{1 \leq i,j \leq K} w_i w_j cor.S_{ij}}} * \boldsymbol{\mu} \tag{41}$$

which is equivalent to

$$\text{minimize } \sum_{1 \leq i,j \leq K} w_i w_j cor.S_{ij}, \text{ subject to } \sum_k w_k \sqrt{n_k} = 1, w_k \geq 0. \tag{42}$$

We note that this is similar to the optimal weights proposed by Lin and Sullivan (2009), except for the additional constraint $w_k \geq 0^{17}$. It is a convex optimization problem with a unique solution which can be efficiently solved by standard software such as the CVXR package in R. With the proposed weights, we revise the calculation above of the dependency factor as

$$\gamma = \sqrt{1 + \frac{N}{\hat{N}_{eff}} - \frac{\hat{N}_{eff}}{N}} = \frac{\sum_k w_k^2}{\sum_{1 \leq i,j \leq K} w_i w_j cor.S_{ij}} \tag{43}$$

It is worth noting that the form of γ was derived for weights $\sqrt{\frac{n_k}{N}}$. Theoretically, the exact form of γ should be further revised to reflect the change in weights, which can be complicated to compute in practice. Here we use this form as an approximation, and we found that it empirically controls FDR very well in the presence of sample overlap.

Practical strategy for study-specific rare variants

Another possible and often overlooked complication in meta-analyses of genetic studies is the unequal coverage of variants across studies due to different genotyping platforms and/or different imputation panels for individual studies, which results in the presence of “study-specific” rare variants (MAF < 0.01) that are measured in only some of the studies. One suboptimal solution is to only include variants measured across all studies. Alternatively, to maximize the power of genetic discovery, all variants may be included in the meta-analyses. In the presence of study-specific rare variants, we propose to calculate the overall Z-score and its knockoff counterpart as

$$\mathbf{Z}_{score} = \sum_k w_k * \mathbf{C}_k \mathbf{Z}_{k,score}, \tilde{\mathbf{Z}}_{score} \sim \sum_k w_k * \mathbf{C}_k (\mathbf{PZ}_{k,score} + \gamma \mathbf{E}_k) \tag{44}$$

where \mathbf{C}_k is a diagonal matrix with $c_{kj} = 1$ if the j th variant is observed in study k and $c_{kj} = 0$ otherwise; $\mathbf{E}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ independently for all k . Intuitively, we generate modified knockoff Z-scores $\mathbf{PZ}_{k,score} + \gamma \mathbf{E}_k$ for each study and combine them as one meta-analysis knockoff Z-score. When a variant is not measured in a study, we propose coding both its Z-score and knockoff Z-score for that study as 0. This way the study does not contribute to the meta-analysis Z-score/knockoff Z-score. It is worth noting that knockoff inference is scale-free because the feature selection is based on a contrast between \mathbf{Z}_{score} and $\tilde{\mathbf{Z}}_{score}$. Therefore, it does not require rescaling \mathbf{Z}_{score} and $\tilde{\mathbf{Z}}_{score}$ to account for the reduced variation due to study-specific variants.

Practical strategy for tightly linked variants

Although the knockoff method helps to prioritize causal variants over associations due to LD, it is difficult or impossible to distinguish causal genetic variants from highly correlated variants. The presence of tightly linked variants can diminish the power to identify the causal ones. We applied a hierarchical clustering of genetic variants prior to the analysis, where variants in the same cluster have a pair-wise correlation ≥ 0.75 . Then we restricted the analysis to one randomly selected representative variant in each cluster. This ensures that each genetic variant in the genome has a highly correlated representative to be included in the analysis and the analysis is unbiased. On average (based on the nine AD studies), we observed that 37.9% variants can be matched with the 1000 Genome reference panel for the proposed GhostKnockoff analysis after this pruning procedure. That is, on average, each variant represents 2.64 variants in the same cluster.

Empirical power and FDR simulations

For each replicate, we first generated two overlapping studies (2500 individuals per study) with genetic data on 2000 common and rare genetic variants randomly selected from a 1Mb region near the *APOE* region (chr19:44909011-45912650; hg38) in the ADSP study. We then restricted the simulations to variants with minor allele counts >25 to ensure stable calculation of summary statistics (e.g. *p*-values). Since the simulations here focus on method comparison 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. To simulate multiple causal variants, we randomly set 10 variants in the 1Mb region to be causal, with a positive effect on the quantitative/dichotomous trait as follows:

$$\text{Quantitative trait : } Y_i = X_{i1} + \beta_1 G_{i,1} + \dots + \beta_{10} G_{i,10} + \varepsilon_i^Q \quad (45)$$

$$\text{Dichotomous trait : } g(\mu_i) = \beta_0 + X_{i1} + X_{i2} + \beta_1 G_{i,1} + \dots + \beta_{10} G_{i,10} \quad (46)$$

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; ε_i^Q and X_{i2} reflect variation due to unobserved covariates; ($G_{i,1}, \dots, G_{i,10}$) 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, $\beta_1^2 \text{var}(g_1) + \dots + \beta_{10}^2 \text{var}(g_{10})$, is 1. We applied the proposed methods to the region as described before to analyze single variants. For each replicate, the empirical power is defined as the proportion of detected variants among all causal variants; the empirical FDR is defined as the proportion of non-causal variants among all detected variants. We simulated 1000 replicates and calculated the average empirical power and FDR.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The manuscript used summary statistics from existing studies from the UK Biobank available at <https://pheweb.org/UKB-SAIGE/>. The summary statistics from each GWAS for Alzheimer's disease can be found at https://ctg.cncr.nl/software/summary_statistics, <https://www.niagads.org/datasets>, and <https://www.ebi.ac.uk/gwas/>. Specifically, (1) The genome-wide survival association study performed by Huang et al. 2017 (NIAGADS ID: NG00058); (2) The genome-wide meta-analysis by Jansen et al. 201944 (available through: https://ctg.cncr.nl/software/summary_statistics); (3) The genome-wide meta-analysis by Kunkle et al. 2019 (NIAGADS ID: NG00075); (4) The genome-wide meta-analysis by Schwartzenuber et al. 2021 (GWAS catalog ID: GCST90012877); (5) In-house genome-wide associations study imputed using the TOPMed reference panels (see Supplementary Table 3); (6,7) Two whole-exome sequencing analyses of data from ADSP by Bis et al. 2020 (NIAGADS ID: NG00065), and Le Guen et al. 2021 (NIAGADS ID: NG000112); (8) In-house whole-exome sequencing analysis of ADSP (NIAGADS ID: NG00067.v5); (9) In-house whole-genome sequencing analysis of ADSP (NIAGADS ID: NG00067.v5). The single-cell RNASeq data for the candidate genes are available in the GEO database under accession code [GSE163577](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE163577). The results of our analysis of UK Biobank and AD genetics can be downloaded at: [zihuaihelab.github.io](https://github.com/zihuaihelab).

Code availability

We have implemented GhostKnockoff in a computationally efficient R package that can be accessed at <https://cran.r-project.org/web/packages/GhostKnockoff/>.

References

- Sierksma, A., Escott-Price, V. & De Strooper, B. Translating genetic risk of Alzheimer's disease into mechanistic insight and drug targets. *Science* **370**, 61–66 (2020).
- Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491 (2018).
- Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
- Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **2010** 427 **42**, 565–569 (2010).
- Sims, R., Hill, M. & Williams, J. The multiplex model of the genetics of Alzheimer's disease. *Nat. Neurosci.* **2020** 233 **23**, 311–322 (2020).
- Visscher, P. M. et al. 10 years of gwas discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
- Candès, E., Fan, Y., Janson, L. & Lv, J. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)* **80**, 551–577 (2018).
- Barber, R. F. & Candès, E. J. Controlling the false discovery rate via knockoffs. *Ann. Statistics* **43**, 2055–2085 (2015).
- Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
- Sesia, M., Sabatti, C. & Candès, E. J. Gene hunting with hidden Markov model knockoffs. *Biometrika* **106**, 1–18 (2019).
- Sesia, M., Katsevich, E., Bates, S., Candès, E. & Sabatti, C. Multi-resolution localization of causal variants across the genome. *Nat. Commun.* **11**, 1–10 (2020).
- He, Z. et al. Identification of putative causal loci in whole-genome sequencing data via knockoff statistics. *Nat. Commun.* **2021** 121 **12**, 1–18 (2021).
- Sesia, M., Bates, S., Candès, E., Marchini, J. & Sabatti, C. False discovery rate control in genome-wide association studies with population structure. *Proc. Natl. Acad. Sci. USA* **118**, e2105841118 (2021).
- Chen, H. et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.* **98**, 653–666 (2016).
- Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
- Jiang, L., Zheng, Z., Fang, H. & Yang, J. A generalized linear mixed model association tool for biobank-scale data. *Nat. Genet.* **53**, 1616–1621 (2021).
- Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
- Lin, D. Y. & Sullivan, P. F. Meta-analysis of genome-wide association studies with overlapping subjects. *Am. J. Hum. Genet.* **85**, 862 (2009).
- Leung, Y. Y. et al. VCPA: genomic variant calling pipeline and data management tool for Alzheimer's Disease Sequencing Project. *Bioinformatics* **35**, 1768–1770 (2019).
- Chen, C. Y. et al. Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29**, 1399–1406 (2013).
- Auton, A. et al. A global reference for human genetic variation. *Nat* **2015** 5267571 **526**, 68–74 (2015).
- Gimenez, J. R. & Zou, J. Improving the stability of the knockoff procedure: multiple simultaneous knockoffs and entropy

- maximization. In: *AISTATS 2019—22nd Int. Conf. Artif. Intell. Stat.* (eds Chaudhuri, K. & Sugiyama, M.) Vol. 89, 2184–2192 (2018).
23. Huang, K. L. et al. A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease. *Nat. Neurosci.* **20**, 1052–1061 (2017).
 24. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
 25. Kunkle, B. W. et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A β , tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
 26. Schwartzenuber, J. et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat. Genet.* **53**, 392–402 (2021).
 27. Belloy, M. E. et al. Challenges at the APOE locus: A robust quality control approach for accurate APOE genotyping. *medRxiv* <https://doi.org/10.1101/2021.10.19.21265022> (2021).
 28. Bis, J. C. et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. *Mol. Psychiatry* **2018** 258 **25**, 1859–1875 (2018).
 29. Le Guen, Y. et al. A novel age-informed approach for genetic association analysis in Alzheimer's disease. *Alzheimer's Res. Ther.* **13**, 1–14 (2021).
 30. Belloy, M. E. et al. A fast and robust strategy to remove variant level artifacts in Alzheimer's Disease Sequencing Project data. *medRxiv* <https://doi.org/10.1101/2021.10.28.21265577> (2021).
 31. Byrska-Bishop, M. et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* <https://doi.org/10.1101/2021.02.06.430068> (2021)
 32. O'Brien, R. J. & Wong, P. C. Amyloid precursor protein processing and Alzheimer's disease. *Annu. Rev. Neurosci.* **34**, 185 (2011).
 33. Hosp, F. et al. Quantitative interaction proteomics of neurodegenerative disease proteins. *Cell Rep.* **11**, 1134–1146 (2015).
 34. Wightman, D. P. et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).
 35. Bellenguez, C., Küçükali, F., Jansen, I., MedRxiv, V. A.- & 2020, undefined. New insights on the genetic etiology of Alzheimer's and related dementia. *medrxiv.org*.
 36. Yang, A. C. et al. A human brain vascular atlas reveals diverse cell mediators of Alzheimer's disease risk. *bioRxiv* <https://doi.org/10.1101/2021.04.26.441262> (2021).
 37. Zhang, S. et al. regBase: whole genome base-wise aggregation and functional prediction for human non-coding regulatory variants. *Nucleic Acids Res.* **47**, e134–e134 (2019).
 38. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
 39. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).
 40. Shihab, H. A. et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536–1543 (2015).
 41. Fu, Y. et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).
 42. IONITA-LAZA, I., MCCALLUM, K., XU, B. & BUXBAUM, J. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214 (2016).
 43. Lu, Q. et al. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci. Rep.* **5**, 1–13 (2015).
 44. Ioannidis, N. M. et al. FIRE: functional inference of genetic variants that regulate gene expression. *Bioinformatics* **33**, 3895 (2017).
 45. Smedley, D. et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. *Am. J. Hum. Genet.* **99**, 595 (2016).
 46. Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
 47. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. Probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276 (2015).
 48. Rogers, M. F. et al. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**, 511–513 (2018).
 49. Rogers, M. F., Shihab, H. A., Gaunt, T. R. & Campbell, C. CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Sci. Rep.* **7**, 1–10 (2017).
 50. Di Iulio, J. et al. The human noncoding genome defined by genetic diversity. *Nat. Genet.* **50**, 333–337 (2018).
 51. Yang, H. et al. De novo pattern discovery enables robust assessment of functional consequences of non-coding variants. *Bioinformatics* **35**, 1453 (2019).
 52. Gulko, B. & Siepel, A. An evolutionary framework for measuring epigenomic information and estimating cell-type specific fitness consequences. *Nat. Genet.* **51**, 335 (2019).
 53. Wells, A. et al. Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat. Commun.* **10**, 5241 (2019).
 54. Gussow, A. B. et al. Orion: detecting regions of the human non-coding genome that are intolerant to variation using population genetics. *PLoS ONE* **12**, e0181604 (2017).
 55. Zhou, L. & Zhao, F. Prioritization and functional assessment of noncoding variants associated with complex diseases. *Genome Med.* **10**, 53 (2018).
 56. Benner, C. et al. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* **101**, 539 (2017).
 57. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
 58. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
 59. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779. (2015).
 60. Dai, R. & Barber, R. The knockoff filter for FDR control in group-sparse and multitask regression. In: *Proc. 33rd International Conference on Machine Learning*, (eds Balcan, M. F. & Weinberger, K. Q.) Vol. 48, 1851–1859 (PMLR, 2016).
 61. Katsevich, E. & Sabatti, C. Multilayer knockoff filter: controlled variable selection at multiplex resolutions. *Ann. Appl. Stat.* **13**, 1 (2019).
 62. Gimenez, J. R., Ghorbani, A. & Zou, J. Knockoffs for the mass: new feature importance statistics with false discovery guarantees. In: *Proc. 22nd International Conference on Artificial Intelligence and Statistics* 2125–2133 (2019).
 63. Yang, Y. et al. Effective sample size: quick estimation of the effect of related samples in genetic case-control association analyses. *Comput. Biol. Chem.* **35**, 40 (2011).

Acknowledgements

This research was additionally supported by NIH/NIA award AG066206 (Z.H.), AG060747 (M.D.G.), AG066515 (Z.H., T.W.-C., M.D.G.). We gratefully acknowledge the studies which provided summary statistics.

Author contributions

Z.H. developed the concepts for the manuscript and proposed the method. Z.H. L.L., H.T. M.G., and I.I.-L designed the analyses and

applications and discussed results. L.L., C.S., E.C. significantly improved the mathematical rigor of the method. Z.H., M.E.B., Y.L.G., A.S., X.L., X.Q., P.G., and S.M. conducted the analyses. M.E.B. and M.D.G. helped with design and interpretation of AD genetics. T.W.C. helped with design and interpretation of the scRNAseq analyses. Z.H. prepared the manuscript and all authors contributed to editing the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-34932-z>.

Correspondence and requests for materials should be addressed to Zihuai He.

Peer review information *Nature Communications* thanks Doug Speed and Jennifer Yokoyama for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022