

# Genome-wide large-scale multi-trait analysis characterizes global patterns of pleiotropy and unique trait-specific variants

Received: 23 April 2023

Accepted: 29 July 2024

Published online: 14 August 2024

 Check for updatesGuanghao Qi<sup>1</sup>, Surya B. Chhetri<sup>2,3</sup>, Debashree Ray<sup>4,5</sup>, Diptavo Dutta<sup>6</sup>, Alexis Battle<sup>2,7,8</sup>, Samsiddhi Bhattacharjee<sup>9</sup>✉ & Nilanjan Chatterjee<sup>10,11</sup>✉

Genome-wide association studies (GWAS) have found widespread evidence of pleiotropy, but characterization of global patterns of pleiotropy remain highly incomplete due to insufficient power of current approaches. We develop fastASSET, a method that allows efficient detection of variant-level pleiotropic association across many traits. We analyze GWAS summary statistics of 116 complex traits of diverse types collected from the GRASP repository and large GWAS Consortia. We identify 2293 independent loci and find that the lead variants in nearly all these loci (~99%) to be associated with  $\geq 2$  traits (median = 6). We observe that degree of pleiotropy estimated from our study predicts that observed in the UK Biobank for a much larger number of traits ( $K = 4114$ ) (correlation = 0.43,  $p$ -value  $< 2.2 \times 10^{-16}$ ). Follow-up analyses of 21 trait-specific variants indicate their link to the expression in trait-related tissues for a small number of genes involved in relevant biological processes. Our findings provide deeper insight into the nature of pleiotropy and leads to identification of highly trait-specific susceptibility variants.

Genome-wide association studies (GWAS) have identified thousands of susceptibility loci across individual complex traits and diseases<sup>1</sup>. Studies have also pointed to the evidence of widespread pleiotropy<sup>2–5</sup>, i.e., genetic variants within individual loci are often associated with multiple traits. The discovery of pleiotropy has transformed the analysis and interpretation of GWAS data. It has, for example, led to the development of more powerful statistical methods for association testing<sup>6–11</sup> and polygenic prediction<sup>10</sup>, robust methods for causal inference accounting for pleiotropic associations<sup>12–15</sup>, as well as new study designs to investigate multiple or even hundreds of traits simultaneously<sup>16–19</sup>. The presence of pleiotropy also poses unique

opportunities and challenges for developing or/and repurposing drugs while minimizing their “off-target” antagonistic effects<sup>20</sup>.

There has been a long quest in genetics to characterize the degree and nature of pleiotropy for complex traits<sup>21</sup>. Early population genetic models postulated universal pleiotropy where genetic variants at any given locus have the potential to affect all traits<sup>21</sup>. More recent experimental studies, however, have suggested that while pleiotropy is highly prevalent, it is likely to be modular in nature, i.e. any given gene is likely to affect a relatively small number of traits<sup>21</sup>. Data from recent GWAS for human traits have also indicated pleiotropy is pervasive<sup>2–5</sup>, but quantifying the true extent of pleiotropy has been challenging.

<sup>1</sup>Department of Biostatistics, School of Public Health, University of Washington, Seattle, WA, USA. <sup>2</sup>Department of Biomedical Engineering, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA. <sup>3</sup>Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA. <sup>4</sup>Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. <sup>5</sup>Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. <sup>6</sup>Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA.

<sup>7</sup>Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA. <sup>8</sup>Department of Genetic Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD, USA. <sup>9</sup>Biotechnology Research and Innovation Council-National Institute of Biomedical Genomics (BRIC-NIBMG), Kalyani, India. <sup>10</sup>Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA. ✉e-mail: [sb1@nibmg.ac.in](mailto:sb1@nibmg.ac.in); [nchatte2@jhu.edu](mailto:nchatte2@jhu.edu)

A number of recent studies<sup>4,22</sup> have quantified the degree of pleiotropy associated with a variant based on the number of associated traits that reach genome-wide significance ( $p < 5 \times 10^{-8}$ ) in individual trait analysis, but such analysis inevitably leads to serious underestimation of the extent of pleiotropy due to lack of power for GWAS of individual traits for the detection of smaller effect-sizes. Another recent study adopted use of a more liberal threshold ( $z$ -statistic  $> 2$  or  $< -2$ )<sup>23</sup> for large scale detection of pleiotropy, but such an analysis can introduce a large number of false positives.

In addition, while previous studies have mostly focused on detecting highly pleiotropic loci and variants<sup>17,22,24</sup>, we believe that given the evidence of highly abundant pleiotropy, an interesting line of investigation would be to detect highly trait-specific loci and variants and explore their unique biological characteristics, if any. Identification of trait specific genetic association may facilitate identification of “core genes” under the omnigenic model for complex traits<sup>25,26</sup>, distinguish the genetic architecture of related traits and potentially help develop drug targets with fewer side effects. Detection of trait-specific associations, however, requires highly powerful methods for detecting pleiotropy as undetected weaker associations would lead to an increase in findings for trait-specific variants. The most comprehensive analysis of pleiotropy based on current GWAS<sup>4</sup> has reported that almost 70% of identified individual SNPs to be associated with a single trait-domain, but these are likely to be highly overestimated due to the lack of power of the underlying analytic method.

In this paper, we develop fastASSET, an extension of the ASSET association analysis based on subSETs (ASSET), which allows detection of any association between a variant and an underlying subset of traits that contribute to the association signal<sup>7</sup>. A major advantage of ASSET, compared to other multi-trait association tests<sup>6,8,9</sup>, is that it not only allows powerful detection of SNPs that show any association across a group of traits, but also readily maps significant SNPs to sets of associated traits. The subset selection feature, which has shown to have robust sensitivity for the detection of true traits under association even when power of individual studies vary<sup>7</sup>, makes the method ideally suited for the investigation of the extent of pleiotropy in current GWAS. While the method has been successfully applied to a number of multi-trait GWAS analyses<sup>27–29</sup> involving a limited number of traits, it is not feasible to implement for the analysis of very large number of traits because of computational burdens associated with all subset search. Here, we develop fastASSET that allows association testing for individual SNPs across many traits by first incorporating a pre-screening step and then performing ASSET-testing based on subSETs (ASSET) on selected traits with suitable adjustment for the pre-screening for  $p$ -value evaluation. Pre-screening excludes traits with large  $p$ -values from subset search and reduces the computational burden. The subset selection feature gives fastASSET unique advantage to map significant variants to sets of associated traits. The vast majority of multi-trait methods such as metaCCA<sup>30</sup>, MOSTest<sup>31</sup>, and JASS<sup>32</sup>, MultiPhen<sup>8</sup>, metaMANOVA, metaUSAT<sup>6</sup>, and HIPO<sup>9</sup> focuses on identifying variants associated with at least one of the traits without specifying which traits are associated (Table 1). Such discoveries are harder to interpret when many traits are analyzed simultaneously, considering that genetic associations have been reported across most part of the genome. MTAG<sup>10</sup> uses multi-trait summary statistics to estimate SNP effects on each trait but does not account for multiple testing across traits which is critical for the analysis of many traits simultaneously. fastASSET returns both multiple-testing adjusted  $p$ -values for association testing and a subset of selected traits, enhancing the interpretation of pleiotropic associations (Table 1).

We use fastASSET to analyze 116 traits collected from large GWAS Consortia and the Genome-Wide Repository of Associations Between SNPs and Phenotypes (GRASP) hosted by the National Institutes of Health (NIH). We identify 2293 independent loci that are

associated with at least one trait and show that lead variants at nearly all of these loci are associated with two or more traits. We show that the degree of pleiotropy we estimate for the underlying variants based on the 116 traits predicts the level of their pleiotropy associated with a much larger number of traits ( $> 4000$ ) in the UK Biobank Study. We conduct a series of follow-up analyses to examine whether the degree of pleiotropy of genetic variants may be related to functional mechanisms, including cis-regulatory effects, chromatin states, transcription factor (TF) binding, and enhancer-gene connections. We further provide detail characterization of 21 highly unique trait-specific variants, i.e., those were associated with only one trait in the fastASSET analysis. In addition, we apply fastASSET to study the patterns of pleiotropy in East Asian population and compare with the patterns in European data. Finally, we discuss the limitations of the study and the different types of pleiotropy. We caution that the pleiotropy discussed in the study are based on associations and not necessarily causal patterns.

## Results

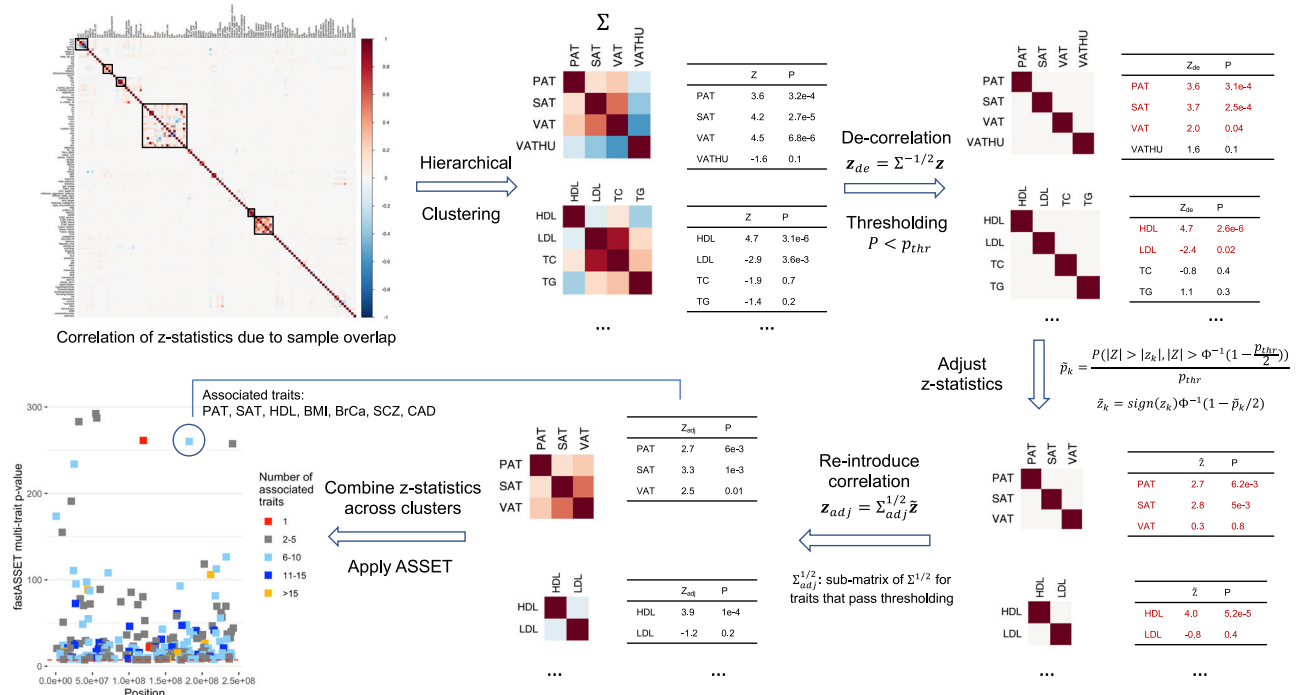
### Overview of datasets and methods

We collect 338 summary-level datasets from the NIH GRASP repository and supplement it with 20 summary-level datasets from large GWAS Consortia. See Supplementary Fig. 1 for data preprocessing pipeline and Methods for details. After filtering out duplicated and highly correlated traits, removing studies with small sample sizes and data quality issues, we retain 116 well-powered studies consisting of primarily participants of European ancestry (Supplementary Data 1 and 2). The studies cover a wide range of complex traits and diseases in 16 domains (Supplementary Data 1). Genetic correlation analysis using linkage disequilibrium (LD) score regression<sup>3</sup> reveals widespread genome-wide pleiotropy (Supplementary Fig. 2). Here, we use LD scores for European ancestry downloaded from the LDSC GitHub repository (see “Methods” for details). We restrict further analysis to 7,462,466 single-nucleotide polymorphisms (SNP) for which summary-statistics were available for at least 50 out of the 116 traits and have minor allele frequency (MAF)  $> 0.01$ .

We develop fastASSET, an extension of the ASSET method<sup>7</sup>, to conduct multi-trait association testing across a large number of traits (see Fig. 1, Methods and Supplementary Notes for details). ASSET was originally designed to conduct single-SNP association tests by performing meta-analysis across all subsets of traits and then evaluating the significance of the maximum of meta-analysis  $z$ -scores over all subsets<sup>7</sup>. In this method, for SNPs that reach a desired significance threshold for association, the set of traits for which the underlying meta-analysis  $z$ -statistics is maximized defines the set of underlying associated traits. For the analysis of a large number of traits, the original ASSET, which searches through all subsets, is computationally intractable. The fastASSET method reduces the computational burden by only searching among the traits which show suggestive evidence of associations. The method first de-correlates the  $z$ -statistics for a given SNP associated with different traits using estimates of phenotypic correlations available from the LD score regression (see Fig. 1 and Methods). Next, it selects the set of traits that shows suggestive level of association (e.g.  $p < 0.05$ ) with the given SNP based on the de-correlated  $z$ -statistics. The de-correlated  $z$ -statistics are then adjusted for the pre-selection step as  $\tilde{z}_k = \text{sign}(z_k) \Phi^{-1}(1 - \frac{p_k}{2})$  where  $\tilde{p}_k = P(|Z_k| > |z_k| | |Z_k| > \Phi^{-1}(0.975))$  and  $\Phi$  is the quantile function of standard normal distribution. The adjusted  $z$ -statistics are then further transformed back to the original scale of the traits, also using estimates of phenotypic correlations available from the LD score regression, and then these  $Z$ -statistics are incorporated as input into the original ASSET method. For each SNP, fastASSET outputs a  $p$ -value for the association with any trait under consideration (“global association”) and a set of associated traits.

**Table 1 | Summary of existing methods for multi-trait association testing using GWAS summary statistics**

Method	Description	Output	Reference (PMID)
metaCCA	Perform canonical correlation analysis (CCA) using correlation matrices estimated from summary statistics and reference panel	<i>P</i> -value for association with a set of traits (global association)	27153689
MOSTest	Mahalanobis norm of the vector of z-statistics ( $z^T \Sigma^{-1} z$ ) with correlation matrix $\Sigma$ estimated from randomly permuted genotypes		32665545
JASS	Omnibus test statistic based on $z^T \Sigma^{-1} z$ and sumZ statistic $\frac{(w^T z)^2}{w^T \Sigma w}$		32002517
MultiPhen	Regression with the genotype as dependent variable and phenotype for multiple traits as independent variable		22567092
metaMANOVA	Test association using multivariate analysis of variance statistic, highly similar to MOSTest and JASS Omnibus test. Correlation matrix $\Sigma$ is estimated using SNPs with no association with the traits.		29226385
metaUSAT	Optimal combination of metaMANOVA and sum of squared score (SSU) statistics		29226385
HIPO	Search for the linear combination of multi-trait summary statistics that maximizes average non-centrality parameter across SNPs		30289880
MTAG	Use multi-trait summary statistics to obtain single-trait effect size estimates by incorporating a prior distribution on the effect size	Estimate of individual-trait GWAS effect size and associated test statistic	29292387
ASSET/fastASSET	Search for optimal subset of traits that maximizes meta-analysis z-statistic	<i>P</i> -value for association with a set of traits (global association) and <b>a subset of selected traits</b>	22560090



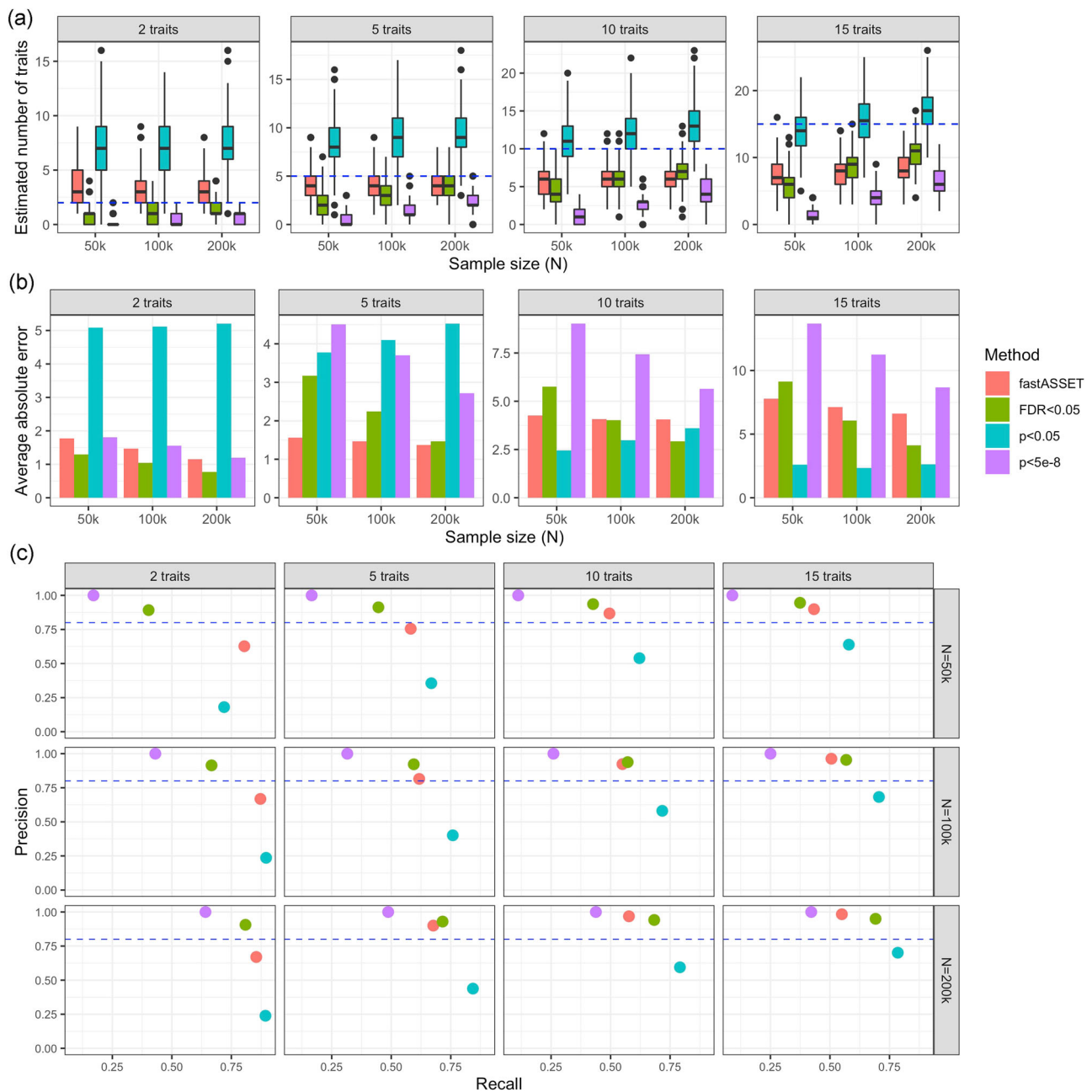
**Fig. 1 | Pipeline of fastASSET.** fastASSET is a statistical method for large-scale multi-trait analysis across many traits. As an extension of ASSET, fastASSET searches for the subset of traits that leads to maximum meta-analysis z-statistic, among those that pass a liberal pre-screening threshold  $P < p_{thr}$  (e.g.,  $p_{thr} = 0.05$ ). It adjusts the z-statistics for pre-screening to avoid “double dipping” while accounting for sample overlap across studies. The procedure is as follows. 1) Apply hierarchical clustering to the LD score regression intercept matrix (top-left heatmap),

which represents correlation of z-statistics induced by sample overlap. 2) De-correlate z-statistics within each cluster using Cholesky decomposition. Select the traits with  $P < p_{thr}$  corresponding to de-correlated z-statistics (red rows in the tables). 3) Adjust the z-statistics for pre-screening using conditional p-values. 4) Re-introduce the correlation by multiplying by the Cholesky decomposition. 5) Combine z-statistics across clusters and feed to ASSET as input.

## Simulation studies

We simulate GWAS summary statistics for 116 traits (see Supplementary Notes for simulation settings). We compare fastASSET to three ad-hoc approaches: selecting traits whose individual-trait associations reach 1) genome-wide significance  $p < 5 \times 10^{-8}$ , 2)  $p < 0.05$ , or 3)  $FDR < 0.05$  across traits. Averaged across scenarios, fastASSET outperforms  $p < 5 \times 10^{-8}$  and  $p < 0.05$  in estimating the degree of

pleiotropy (Fig. 2a, b). The genome-wide threshold  $p < 5 \times 10^{-8}$  is highly conservative and consistently underestimates the degree of pleiotropy. The liberal threshold  $p < 0.05$  introduces many false positives and tends to overestimate pleiotropy by a large margin, especially when the true degree of pleiotropy is low (2 or 5 traits). The other method,  $FDR < 0.05$ , have comparable performance with fastASSET when  $N = 100k$ . fastASSET tends to be more accurate than  $FDR < 0.05$



**Fig. 2 | Performance of fastASSET for estimation of pleiotropy and trait selection in simulation studies. a** Estimated degree of pleiotropy (number of traits) vs. true degree of pleiotropy (dashed lines). Boxplots show the median (centerline) and first and third quartiles (lower and upper hinges) of the distribution ( $n = 300$  replications). The top of upper whisker represents the largest value no more than  $1.5 \times$  interquartile range (IQR) from the hinge; the bottom of lower whisker represents the smallest value no more than  $1.5 \times$  IQR of the hinge. **b** Average absolute error for estimating the degree of pleiotropy, defined as the mean(|estimate-truth|). The individual-trait-based methods (false discovery rate (FDR)  $< 0.05$ ,  $p < 0.05$ ,  $p < 5 \times 10^{-8}$ ) are based on two-sided z-test. **c** Precision-recall for identifying associated traits. For **a–c**, GWAS regression coefficients for one SNP  $j$  and 116 traits ( $\beta_j$ , vector of length 116) are simulated from model  $\hat{\beta}_j = \beta_j + e_j$ , where  $e_j$  is the error

term generated from multivariate normal distribution that reflects realistic correlation across traits. True effect  $\beta_j$  is simulated by first randomly selecting a set of  $K = 2, 5, 10$  or 15 traits (columns) which have true associations with the SNP, followed by generating the effect size from  $\beta_{fix} + N(\text{mean} = 0, \text{variance} = 0.01^2)$ . Around half of the  $K$  traits (round( $K/2$ )) have  $\beta_{fix} = -0.01$  and the rest have  $\beta_{fix} = 0.01$ , which reflects bidirectional fixed effects. Effect size heterogeneity is reflected by  $N(\text{mean} = 0, \text{variance} = 0.01^2)$ . We vary the sample size  $N$ , assumed to be the same across all traits, to 50k, 100k and 200k (rows). This simulation procedure is repeated 300 times for each setting. See Supplementary Notes for details. Calculation of precision and recall is restricted to SNPs that reach genome-wide significance for global association: fastASSET  $p$ -value  $< 5 \times 10^{-8}$  for fastASSET, metaUSAT  $p$ -value  $< 5 \times 10^{-8}$  for the other three methods.

under lower sample size ( $N = 50k$ ) but less accurate when  $N = 200k$  (Figs. 2a, b). This is likely due to the increased power of individual-trait-based approach under large sample size, hence there is less benefit in using a multi-trait method. The performance for identifying the subset of associated traits tracks closely with that for estimating pleiotropy (Fig. 2c). When  $N = 50k$  or  $100k$  and  $K = 2$  or 5, fastASSET maintains

high precision and substantially improves recall compared to more conservative methods  $p < 5 \times 10^{-8}$  and  $FDR < 0.05$ . fastASSET performs similarly to  $FDR < 0.05$  under larger  $N$  and  $K$  and appears less powerful (lower recall) when  $N = 200k$  and  $K = 10$  or 15. The Genome-wide threshold  $p < 5 \times 10^{-8}$  has low recall across scenarios, while the liberal threshold  $p < 0.05$  has low precision. fastASSET is conservative in



detecting trait-specific variants (associated with 1 trait), with high precision and very low recall (Supplementary Data 3). This property appears more desirable when the sample size is low. For example, when  $N=20k$ , the precision for fastASSET is 47.8% and that for  $FDR < 0.05$  is 13%. Though the recall for fastASSET is much lower (3.7% vs 15% for  $FDR < 0.05$ ), the high precision allows us to restrict follow-up analyses to high-confidence trait-specific SNPs. When the sample size is larger, the precision of  $FDR < 0.05$  catches up with that of fastASSET with a higher recall.

In addition, fastASSET has well controlled type I error for detecting global association. metaUSAT (v1.17) and metaMANOVA<sup>6</sup>, two methods that only conduct global association testing but not subset selection, have moderately inflated type I error (Supplementary Data 4). fastASSET has comparable power to metaUSAT and metaMANOVA (Supplementary Fig. 3), with slight power gain at  $N=200k$  and slight loss when  $N=50k$ . We observe that power reaches near saturation for several scenarios. This indicates that power difference is less distinguishable across methods for pleiotropic analysis across many traits. Hence the subset selection feature provided by fastASSET is an important advantage over other multi-trait methods. We also compare the running time of fastASSET and the original ASSET (Supplementary Data 5). The running time of ASSET grows rapidly with the number of traits analyzed. For 25 traits, ASSET uses ~46 min (2733.89 s) to analyze 100 SNPs. However, fastASSET can complete analysis of the same data within a few seconds, even if the number of traits reaches 100. Compared to two other multi-trait methods, fastASSET is faster than metaUSAT but slower than metaMANOVA (Supplementary Data 5).

We conduct additional simulations to evaluate the performance of fastASSET to estimate pleiotropy in more general scenarios that allow heterogeneous sample size and genetic correlation across traits (Supplementary Fig. 4). In this scenario, genetic effects are simulated as the sum of a homogeneous component (same effect size across associated traits) and a heterogeneous component (varying effect size across traits). See the legend of Supplementary Fig. 4 for details. The patterns are largely similar to the previous simulation setting: fastASSET is more accurate than  $FDR < 0.05$  in estimating degree of pleiotropy when  $N=50k$ , but could be less accurate in some scenarios when  $N=200k$ . When there are only heterogeneous effects (homogeneous effect = 0), fastASSET has a clear advantage over  $FDR < 0.05$  for estimating pleiotropy, but can have low precision for identifying associated traits in some settings (when only two traits have true associations). When the homogeneous effect is strong (Supplementary Fig. 4c), fastASSET appears to have comparable performance with  $FDR < 0.05$  when averaged across all scenarios. However, it has better precision-recall tradeoff compared to the scenario with no homogeneous effect.

### Quantifying and validating levels of pleiotropy

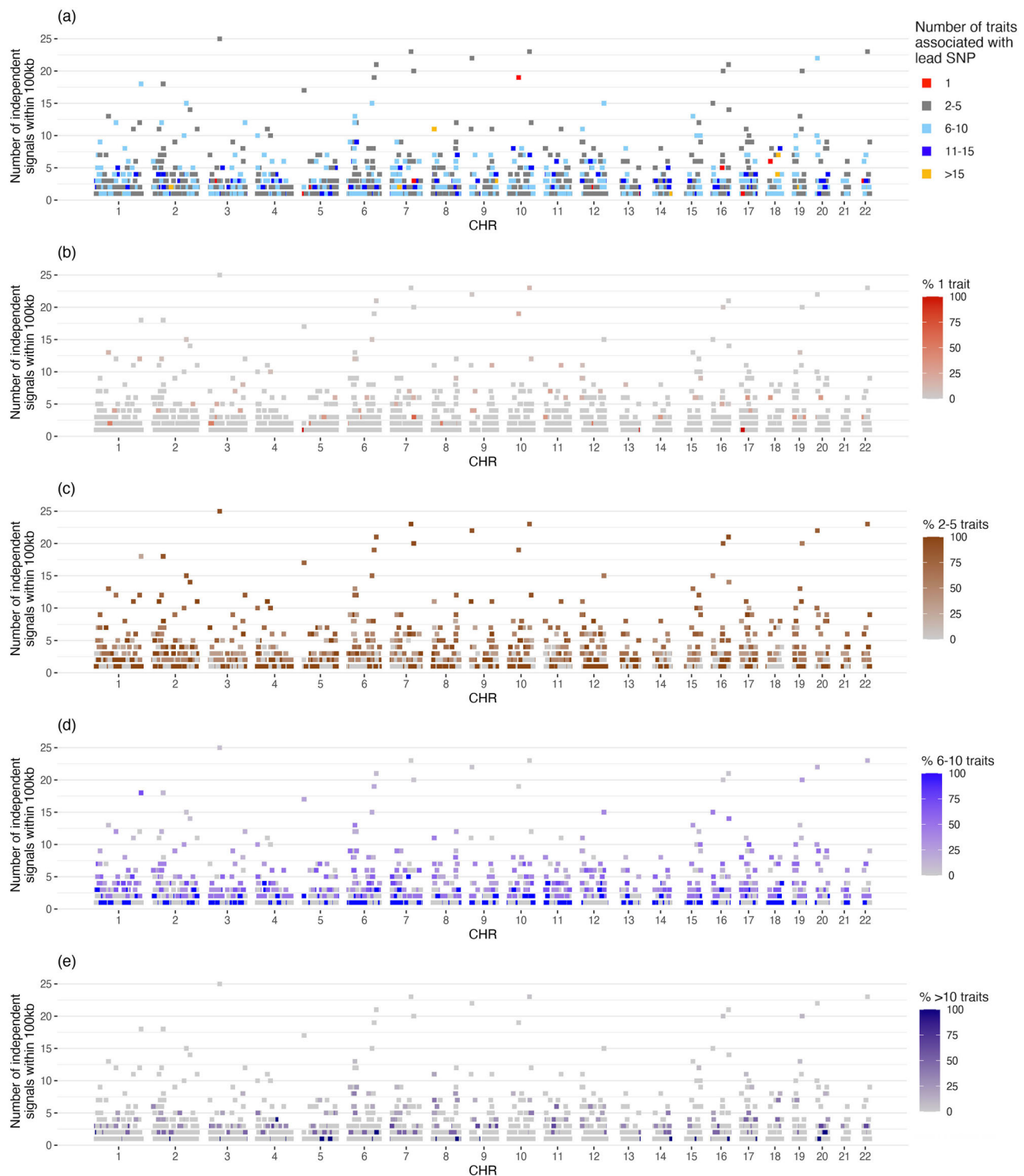
We find widespread genetic associations and varying degree of pleiotropy across the genome (Fig. 3). Identified associations represent 2293 loci (fastASSET  $p$ -value  $< 5 \times 10^{-8}$ ,  $r^2 < 0.1$  and  $> 500$  kb apart, Supplementary Data 6). We further investigate the signal density within each locus by counting the number of independent genome-wide significant SNPs (fastASSET  $p$ -value  $< 5 \times 10^{-8}$ ,  $r^2 < 0.1$ ) within 100 kb of the lead SNP (see “Methods” for details). We observe substantial variation in signal density (Fig. 3). Multiple independent signals were present at nearly half (48% out of 2293) of the loci; 11% of the loci harbored at least 5 independent signals and some loci can harbor up to 25 signals (Fig. 3). For the loci with multiple associated SNPs, levels of pleiotropy can vary within the locus (Fig. 3). For example, SNP rs3760047 (chr16:281299) is associated with 4 traits, but 8 out of the 15 significant SNPs within 100 kb are associated with 2–5 traits and the remaining 7 SNPs are associated with 6–10 traits (Supplementary Data 6). In the following, we use the lead SNPs detected by fastASSET

for each locus to study the level of pleiotropy across the genome and its relationship with different types of variant annotations.

The vast majority of the lead SNPs are associated with 2–10 traits with a median of 6 (Fig. 4a). At the two ends of the spectrum, we found 21 SNPs to be associated with only one trait, representing highly trait-specific genetic mechanisms and 58 SNPs to be highly pleiotropic defined as those which are associated with 16 or more traits. Next, we investigate whether the degree of pleiotropy we estimate for the lead variants based on the 116 traits also predicts degree of pleiotropy that the same variants will manifest across a much wider spectrum of traits. To test this hypothesis, we collect the summary statistics for 4114 traits from the Neale lab UK Biobank (UKB) GWAS<sup>33,34</sup>, and quantify the levels of pleiotropy for each of the 2293 lead SNPs by the number of associated traits at per-SNP false discovery rate ( $FDR < 0.05$ ). See “Methods” for details. We observe that degree of pleiotropy estimated from our study predicts that observed in the UK Biobank (spearman correlation = 0.43,  $p$ -value  $< 2.2 \times 10^{-16}$ , Fig. 4b). The relationship remains highly significant even after adjusting for LD score (partial correlation = 0.42,  $p$ -value  $< 2.2 \times 10^{-16}$ ). The analysis indicates that pleiotropic characteristics of the detected SNPs is not specific to the selected traits in our discovery analysis and likely represent a much broader property related to their roles in gene regulation.

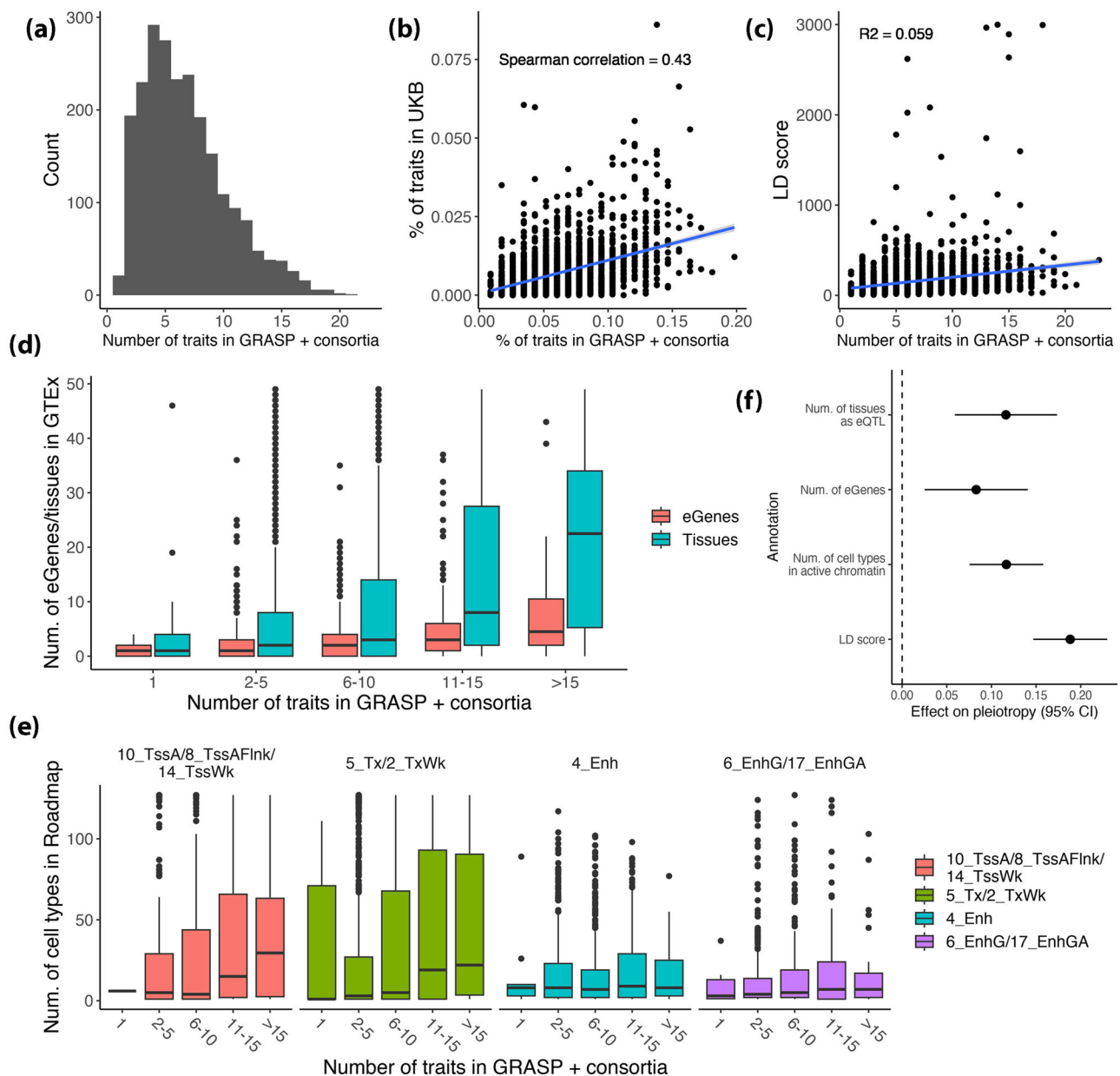
### Relationship with functional and LD annotations

We investigated how the level of pleiotropy is correlated with different types of variant annotations. First, we found that the numbers of associated traits detected by fastASSET to be positively correlated with LD score values across the 2293 lead SNPs ( $R^2 = 0.059$ ,  $p$ -value =  $4.95 \times 10^{-32}$ , Fig. 4c). The pattern is expected since lead SNPs which tag more SNPs due to LD will appear to be associated with a larger number of traits when there are distinct causal variants for distinct traits within the loci. We also found that SNPs associated with a larger number of traits tend to be significant expression quantitative trait loci (eQTL) for a larger number of tissues and for a larger number of eGenes (both  $R^2 = 0.069$ ,  $p$ -value  $< 2.2 \times 10^{-16}$ ). Trait-specific SNPs are significant eQTL for a median of 1 tissue and 1 gene, while highly pleiotropic SNPs ( $> 15$  traits) are significant eQTL for a median of 22.5 tissues and 4.5 genes (Fig. 4d). While such pattern has been reported earlier<sup>4</sup>, we observe a much sharper dose-response trend and higher level of statistical significance (e.g. compared to Fig. 1e in ref. 4), arising likely due to the higher accuracy of the fastASSET analysis for the detection of degree of pleiotropy. In our dataset, eQTLs explain less variation of pleiotropy if it is quantified by the number of traits reaching  $p < 5 \times 10^{-8}$  (the approach adopted by Watanabe et al.<sup>4</sup>), with  $R^2 = 0.031$  for number of tissues and  $R^2 = 0.052$  for number of eGenes. We also found more pleiotropic SNPs tend to be in regions of active chromatin state in a larger number of tissue or cell types, reflected by results from the IDEAS method<sup>35,36</sup> (Fig. 4e). This trend is especially pronounced for the promoter (10\_TssA, 8\_TssAFlnk and 14\_TssWk) and transcription (5\_Tx and 2\_TxWk) related chromatin states, but weaker for enhancer-related states (4\_Enh, 6\_EnhG and 17\_EnhGA). We observe similar relationship using chromatin states learned by ChromHMM (Supplementary Fig. 5). In a multivariate regression analysis that accounts for LD and all functional annotations simultaneously, we found the relationship between degree of pleiotropy and three types of functional characteristics of the SNPs each remain highly significant ( $p$ -value  $< 0.005$ ), with eQTL tissue specificity showing the largest effect size in per standard deviation unit (Fig. 4f). Further, analysis based on the JASPAR and HOCOMOCO databases<sup>37,38</sup> revealed that more pleiotropic SNPs ( $> 10$  traits) are more likely to overlap with transcription factor binding sites (TFBS) ( $p$ -value = 0.022, Supplementary Fig. 6). The relationship remains significant ( $p$ -value = 0.021) after adjusting for the annotations in Fig. 4 (Supplementary Fig. 6). Finally, analysis based on activity-by-contact (ABC) model revealed that among the SNPs that overlap with enhancers, level of pleiotropy



**Fig. 3 | Brisbane plot showing signal density and levels of pleiotropy of 2,293 genome-wide significant loci identified by fastASSET analysis across 116 traits.** We define significant SNPs as those that have fastASSET  $p$ -value  $< 5 \times 10^{-8}$  (two-sided) and clump them to obtain 10,628 independent SNPs ( $r^2 < 0.1$ ), which we then group into 2293 loci whose lead SNPs (strongest association within a locus) are at least 500 kb apart. Each dot represents the lead SNP of a locus, and the y axis is the number of independent significant SNPs within 100 kb from the lead SNP. In **a** dots are colored by the number of traits associated with the lead SNP as reported by fastASSET. In **b–e** dots are colored by the proportion of significant SNPs in a locus

(within 100 kb from the lead SNP) that fall into each category of pleiotropy. **b** Color represents proportion of significant SNPs that are trait-specific (associated with 1 trait). **c** Color represents proportion of significant SNPs that are associated with 2–5 traits. **d** Color represents proportion of significant SNPs that are associated with 6–10 traits. **e** Color represents proportion of significant SNPs that are associated with >10 traits. Nearly half (1094) of the loci harbor at least two independent genome-wide significant SNPs within 100 kb of the lead SNP (including the lead SNP itself); 248 (11%) loci harbor at least 5 significant SNPs within 100 kb of the lead SNP. See Supplementary Data 6 for details.



**Fig. 4 | Number of associated traits detected by fastASSET and its relationship with different variant annotations.** Results are shown for only the lead SNPs of 2,293 independent loci identified by fastASSET. Sub-figures show (a) Distribution of number of associated traits; (b) Relationship with the number of associated traits in the UK Biobank (per-SNP FDR < 0.05, each point represents a lead SNP); (c) Relationship with linkage disequilibrium (LD score). For (b) and (c) shared area along the blue line is the 95% confidence band. (d) Relationship with number of tissues and genes for which the SNP is a significant eQTL in GTEx v8; (e) Relationship with the number of cell types in which the SNP is in each active chromatin state; (f) Effect of each annotation on pleiotropy (number of associated traits) conditional on other

annotations, and 95% confidence intervals. Chromatin states are learned by IDEAS method (Zhang et al., 2017) using data from the Roadmap Epigenomics Consortium. Coefficients in (f) are estimated using multiple linear regression model of the form  $pleiotropy \sim (LD\ score) + (num.of\ tissues\ as\ eQTL) + (num.of\ eGenes) + (num.of\ cell\ types\ in\ active\ chromatin)$  where all dependent and independent variables are standardized to have unit variance. Boxplots show the median (centerline) and first and third quartiles (lower and upper hinges) of the distribution; the top of the upper whisker represents the largest value no more than 1.5 \* inter-quartile range (IQR) from the hinge; the bottom of lower whisker represents the smallest value no more than 1.5 \* IQR of the hinge.

increases with the number of active tissues (correlation = 0.09,  $p$ -value = 0.069, Supplementary Fig. 6). However, this relationship disappears after adjusting for other annotations ( $p$ -value = 0.65, Supplementary Fig. 6).

### Trait-specific Variants

The fastASSET analysis identified 21 independent trait-specific lead SNPs, defined as those for which associated subset included only one trait (Table 2). We observe that the association  $p$ -value between a trait-specific SNP and the primary trait are at least  $10^8$  times lower than the

next most strongly associated trait (Fig. 5 and Supplementary Fig. 7). Highly pleiotropic SNPs, however, often have comparable level of associations with a large number of traits (Fig. 5 and Supplementary Fig. 7). We also validate the nature of trait specificity of these SNPs in the external UK Biobank study (Table 2). Notably, even though UK Biobank covered a much larger of traits, the trait-specific SNPs we identified for Alzheimer's disease, breast cancer, prostate cancer, Crohn's disease (CD) are only associated with traits related to these diseases, except rs4631223 detected for primary trait CD was also found to be associated with Urea. Some of the disease-specific SNPs we

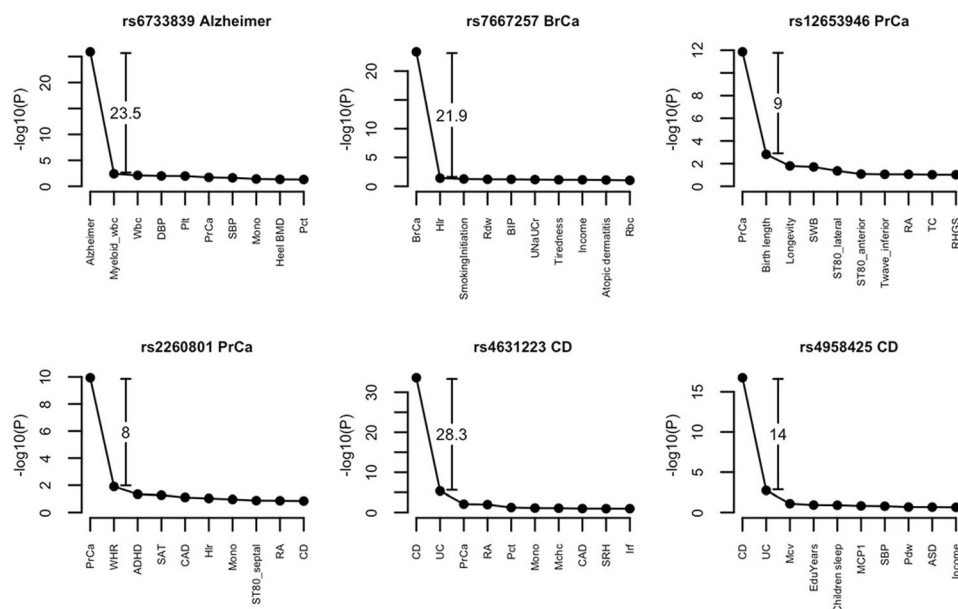
Table 2 | Trait-specific lead SNPs identified by fastASSET and validation analysis in UK Biobank

CHR	BP	SNP	fastASSET P-value	Trait	Number of associated traits in UKB	Associated traits in UKB (similar traits are grouped and the number of traits in group is shown in the parentheses)
2	127892810	rs6733839	1.16E-22	Alzheimer's disease	2	Illnesses of father: Alzheimer's disease/dementia, Illnesses of mother: Alzheimer's disease/dementia
4	175828082	rs7667257	2.73E-22	Breast cancer	3	Malignant neoplasm of breast (3)
5	1895829	rs12653946	5.05E-10	Prostate cancer	5	Malignant neoplasm of prostate (3), malignant neoplasm of male genital organs (2)
10	122759891	rs2260801	4.80E-09	Prostate cancer	0	None
5	40438471	rs4631223	2.47E-31	Crohn's disease	2	Crohn's disease, Urea
5	150253398	rs4958425	1.36E-15	Crohn's disease	0	None
2	119507607	rs115242848	5.63E-262	Heel BMD	16	Osteoporosis, Sitting height, Fractured/broken bones in last 5 years, FVC, Heel BMD and ultrasound (12)
7	121389859	rs149472274	4.26E-19	Heel BMD	13	Heel BMD and ultrasound (12), Heel bone mineral density (6), Calcium
8	70182266	rs283326	2.45E-23	Heel BMD	13	Blepharitis/eyelid infection, Heel BMD and ultrasound (12)
10	54419939	rs12250150	6.17E-278	Heel BMD	14	FEV1, Fractured/broken bones in last 5 years, Heel BMD and ultrasound (12)
11	35096310	rs2098878	8.99E-52	Heel BMD	14	Heel BMD and ultrasound (12), Ever had hysterectomy, Leiomyoma of uterus
16	51903981	rs72805220	4.57E-53	Heel BMD	28	Weight (2), Body fat measurements (8), Impedance of leg (2), High light scatter reticulocyte (2), Heel BMD and ultrasound (12), Waist/Hip circumference (2)
17	10031642	rs9914258	1.31E-09	Heel BMD	26	Glaucoma and related measurements (8), xalatan 0.005% eye drops, 6 mm weak meridian (right), Sitting/ Standing height (2), FVC, Heel BMD and ultrasound (8), Other diseases of anus and rectum, #Arthrosis, Albumin, Calcium, Phosphate
18	13722308	rs1941749	1.63E-126	Heel BMD	18	Fractured bones (2), Heel BMD and ultrasound (12), Fracture of forearm, Alkaline phosphatase, C-reactive protein, Triglycerides
12	50767037	rs61926181	1.37E-19	DBP	12	Blood pressure and hypertension (5), Number of treatments/medications taken, Blood pressure medication (2), Mean platelet (thrombocyte) volume, Malignant neoplasm of tonsil (2), Congenital malformations/ deformations and chromosomal abnormalities
3	38019428	rs9826470	7.92E-35	Height	19	Height measurements (3), FVC (2), FEV1, Whole body fat-free/water mass (2), Arm/Trunk mass measurements (6), Diseases of the skin and subcutaneous tissue (2), Hand grip strength (right), Basal metabolic rate, Lymphocyte count
8	121293376	rs10107534	2.03E-17	Height	12	Height measurements (3), Nervous feelings, Answered sexual history questions, Arm mass measurements (4), Viral meningitis, Parkinson's disease, Glaucoma suspect
13	106636056	rs2478292	1.03E-13	Intelligence	1	Fluid intelligence score
6	9327556	rs9357047	1.89E-33	Male baldness	4	Hair/balding patterns (4)
15	23803985	rs62000036	2.20E-22	Age at menarche	2	Impedance of arm (right), Age when periods started (menarche)
22	17579495	rs41409548	9.65E-23	Monocyte count	2	Monocyte count/percentage (2)

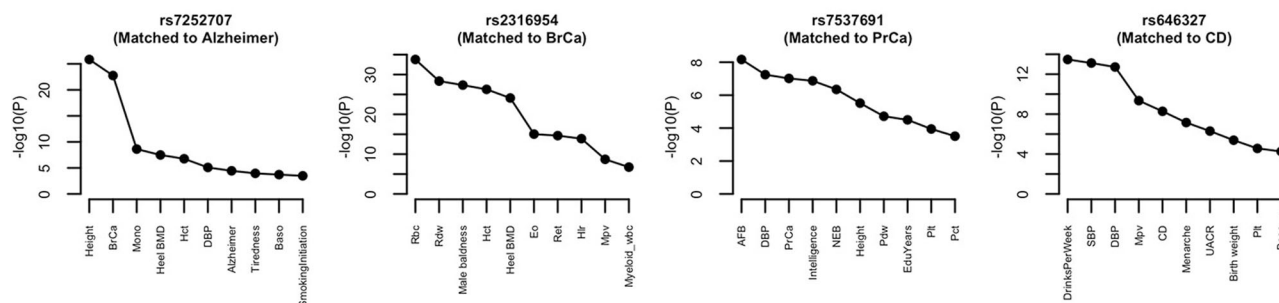
BMD: bone mineral density, DBP: diastolic blood pressure. Associated traits in UKB are identified by per-variant FDR < 0.05. P-values are derived from two-sided fastASSET tests.



## (a) Trait-specific



## (b) Highly pleiotropic (&gt;15 traits)



**Fig. 5 | Top 10 associated traits for trait-specific SNPs for binary diseases and matched highly pleiotropic SNPs. a** Trait-specific SNPs for binary disease traits. **b** Matched highly pleiotropic SNPs. Traits are ordered by descending order of  $-\log_{10}(p\text{-value})$  in each panel. See Supplementary Data 1 for sample size of individual-trait GWASs.  $P$ -values are from two-sided z-tests for individual traits.

Each disease-specific SNP is matched to the highly pleiotropic SNP (>15 traits) that has the smallest association  $p$ -value with the disease (matched trait shown in the parentheses). See Supplementary Fig. 7 for trait-specific SNPs for quantitative traits and matched highly pleiotropic SNPs. See “Methods” for details of the matching procedure.

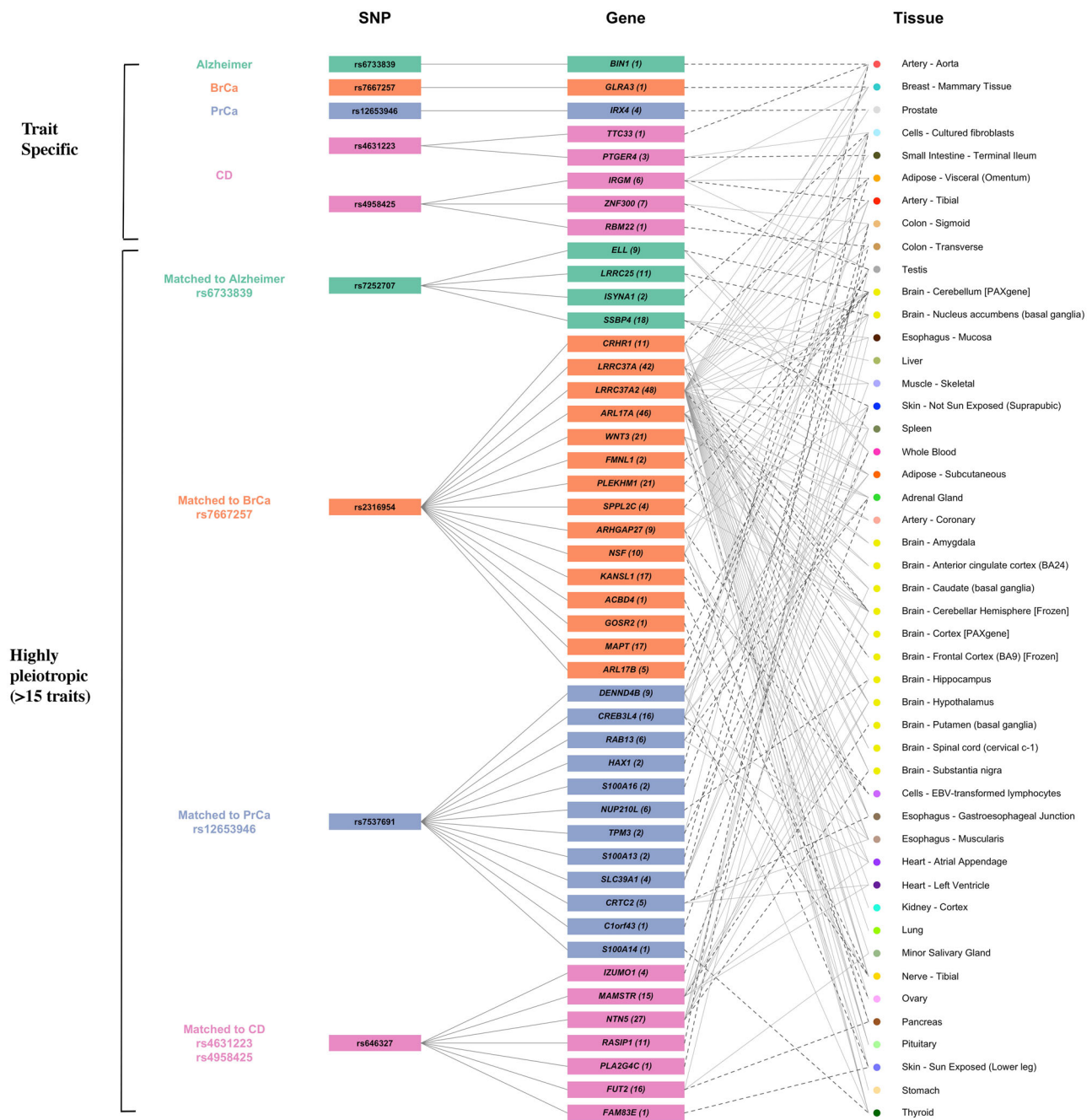
detected were not significant ( $FDR \geq 0.05$ ) in the UK Biobank (Table 2) likely due to its lower number of cases for rare diseases compared to large case-control studies that contributed to our discovery analysis. For quantitative traits, we observe replication of trait-specificity in the UK biobank for most of the trait-specific SNPs for intelligence, male baldness, age at menarche and monocyte count (Table 2). The trait-specific SNPs we detected for heel bone mineral density (BMD), height and diastolic blood pressure (DBP), however, were associated with a larger number of traits in the UKB, but the majority of the additional traits were related to the primary trait.

### Cis-regulatory effects for trait-specific SNPs

To gain further insight into the mechanisms driving trait-specific SNPs, we search for their *cis*-eGenes and corresponding tissues in GTEx<sup>39</sup>, as well as the biological functions of the eGenes in the GeneCards database<sup>40</sup>. Although eQTL studies may be biased towards specific types of variants<sup>41</sup>, they still explain a significant fraction of trait heritability<sup>42</sup>, which could increase further as more eQTL data are collected on new cell types or contexts. Among the 21 trait-specific SNPs, 11 SNPs are eQTLs for at least one gene-tissue pair ( $q$ -value <

0.05) in GTEx v8 (Fig. 6 and Supplementary Fig. 8). Five of them are associated with the expression of one single gene. rs6733839 (Alzheimer’s disease) is a significant eQTL for *BINI* gene in aorta (Fig. 6). Though the aorta does not appear to be the relevant tissue for Alzheimer’s disease, the SNP is also moderately associated the expression of the *BINI* in brain cerebellum (Supplementary Fig. 9). rs7667257 (breast cancer) is an eQTL for only *GLRA3* in breast mammary tissue, which is exactly the relevant issue for breast cancer. rs12653946 (prostate cancer) is associated with the expression of *IRX4* in 4 tissues but has the largest effect size in prostate—the disease’s tissue of occurrence (Fig. 6). rs9914258 (heel BMD) is associated with the expression of *GAS72* with the strongest effect in cultured fibroblasts (Supplementary Fig. 8). rs41409548 (monocyte count) is associated with the expression of *IL17RA* in brain cortex (Supplementary Fig. 8) and a number of other brain tissues (Supplementary Fig. 9).

The other trait-specific SNPs, though not exactly tissue or gene specific, are generally associated with a small number of genes in a small number of tissues compared to highly pleiotropic “matched” SNPs that are also associated with the same traits. rs4631223 (Crohn’s disease) is associated with the expression of *TTC33* and *PTGER4*, and



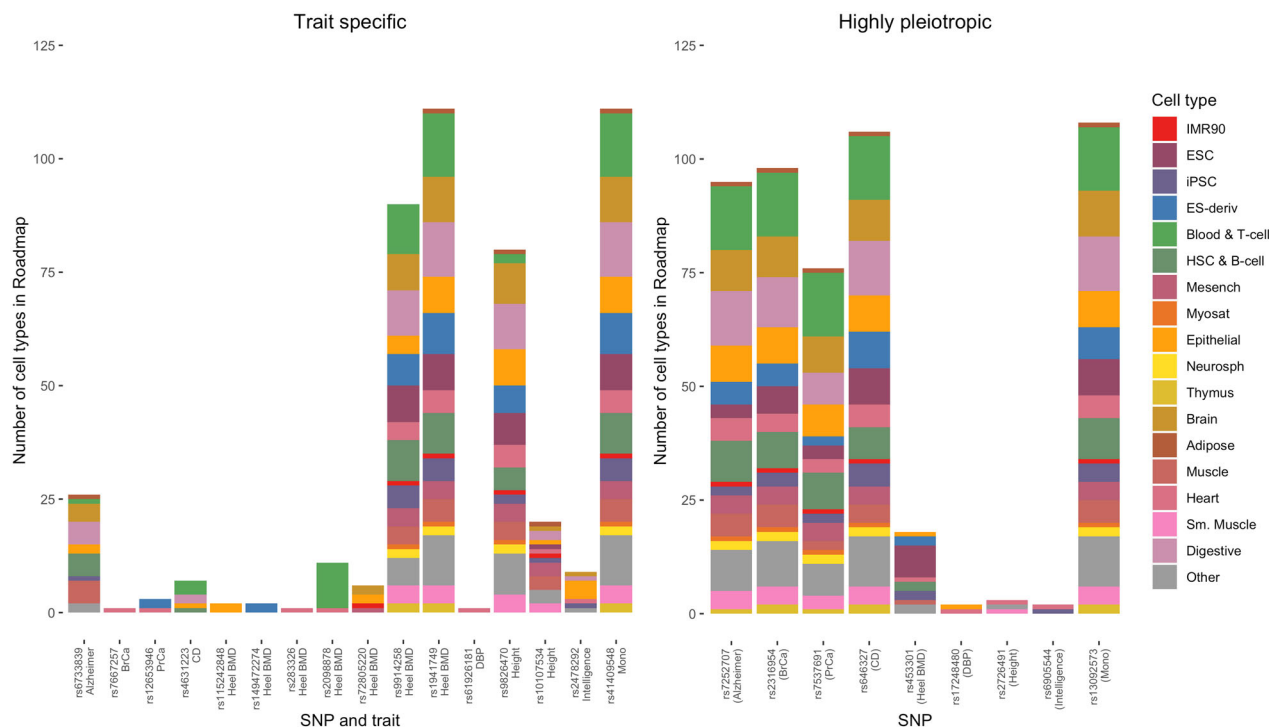
**Fig. 6 | Cis-regulatory effects of trait-specific SNPs for binary diseases and matched highly pleiotropic SNPs.** For each SNP, we show its protein-coding eGenes ( $q$ -value  $< 0.05$ ) and the corresponding top tissues in GTEx v8. The top tissues for each variant-gene pair are defined as those with eQTL effect size is  $> 0.7$  (the largest effect size among significant tissues for this variant-gene pair); the tissue harboring the largest effect is highlighted by darker dashed lines. In addition, we annotate each gene name by the total number of associated tissues

( $q$ -value  $< 0.05$ ) regardless of effect size. Each trait-specific SNP is matched to the highly pleiotropic SNP ( $> 15$  traits) that have the strongest association with the trait (lowest  $p$ -value). See “Methods” for details of the matching procedure. Pseudo-genes and non-coding RNAs are excluded. See Supplementary Fig. 8 for trait-specific SNPs for quantitative traits and highly pleiotropic SNPs. BrCa: breast cancer; PrCa: prostate cancer; CD: Crohn’s disease.

the largest effect size for *PTGER4* is in small intestine (Fig. 6), the organ where Crohn’s disease usually occurs. Another trait-specific SNP for Crohn’s disease, rs4958425, is associated with the expression of *IRGM*, *ZNF300* and *RBM22*. The association for *RBM22* occurs in colon transverse and a secondary association for *ZNF300* occurs in colon sigmoid (Fig. 6), both of which are in the digestive system. *IRGM* is moderately associated with rs4958425 in small intestine (Supplementary Fig. 9). Previous studies have supported the connections between the trait mechanisms and the functions of some of the genes above,

including the role of *BIN1* in Alzheimer’s disease, *PTGER4* and *IRGM* in Crohn’s disease and *IL17RA* in monocyte count<sup>43–46</sup> (Supplementary Data 7).

We further used COLOC<sup>47</sup> to conduct colocalization analysis for the loci indexed by 11 trait-specific SNPs that are also significant eQTLs in GTEx. We find evidence of colocalization ( $PP4 > 0.8$ ) between genetic effects on the trait and on gene expression for 8 trait-specific SNPs (Supplementary Data 8), indicating shared causal SNPs. Many of the genes and tissues we highlighted as potential mechanisms for trait-



**Fig. 7 | Cell types in which the trait-specific and matched highly pleiotropic SNPs have active chromatin state.** Each trait-specific SNP is matched to the highly pleiotropic SNP (>15 traits) that have the strongest association with the trait (lowest *p*-value). See Methods for details of the matching procedure. Chromatin states are

learned by applying the IDEAS method to data from the Roadmap Epigenomics Project. BrCa: breast cancer; PrCa: prostate cancer; CD: Crohn's disease; BMD: bone mineral density; DBP: diastolic blood pressure; mono: monocyte count.

specific SNPs (Fig. 6 and Supplementary Fig. 8) are also supported by colocalization, e.g. *IRX4* and prostate, *PTGER4* and small intestine, *RBM22* and transverse colon, etc. (Supplementary Data 8). These results provide further evidence that trait-specificity may be driven by gene- and tissue-specific cis-regulatory effects.

### Chromatin state of trait-specific SNPs

We explore another potential mechanism of trait-specificity using chromatin states learned by the IDEAS method<sup>35</sup> using data from Roadmap Epigenomics Consortium<sup>48</sup>. Though the number of SNPs is small, we find suggestive evidence that trait specific SNPs tend to be in active chromatin state only in specific tissue or cell types (Fig. 7). On the other hand, highly pleiotropic SNPs tend to have active chromatin state in a wide range of tissue and cell types (Fig. 7). For a number of trait-specific SNPs, chromatin state results strongly link the SNP to trait/disease related tissues. For example, rs4631223 (Crohn's disease) is in active chromatin state in blood and T-cells, B-cells and digestive tissues, which are related to the immune mechanism of Crohn's disease and the organ where it occurs. rs6733839 (Alzheimer's disease) is active in brain and muscle tissues (Fig. 7), both of which were shown to be involved in the disease<sup>49–51</sup>, though it is also active in immune cells and digestive tissues, whose role in the disease is less clear. rs61926181 (DBP) is active in heart tissues which is also closely related to DBP.

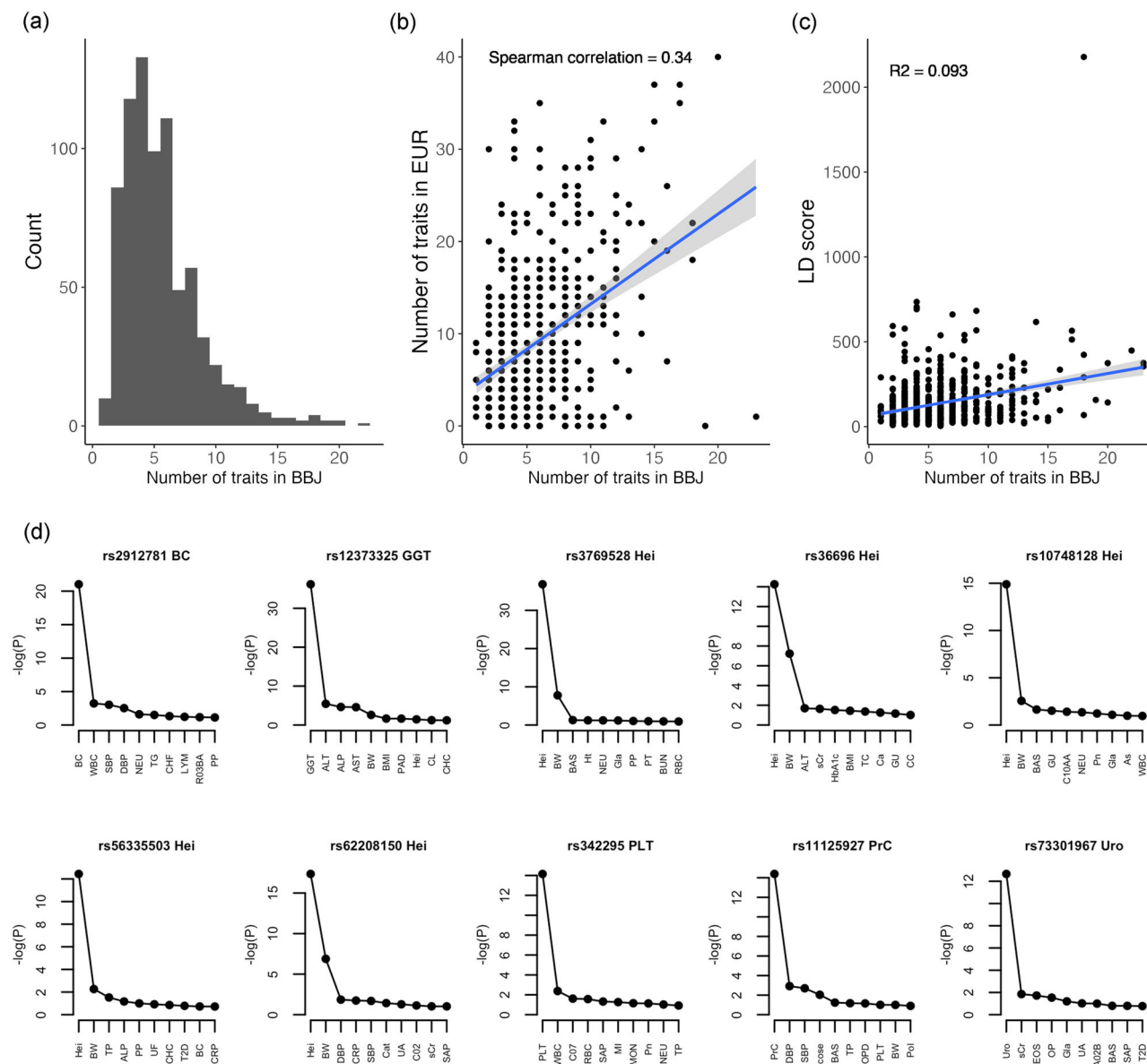
### Pleiotropy in East Asian population

To study the patterns of pleiotropy in a different population, we collect GWAS summary statistics for 220 traits in Biobank Japan (BBJ)<sup>22</sup>. After quality control (see “Methods” for details), 75 traits remain for the final statistical analysis (Supplementary Data 9). Using fastASSET, we identify 778 loci with significant global association ( $p < 5 \times 10^{-8}$ ,  $r^2 < 0.1$ , lead SNPs are >500 kb apart). See Supplementary Data 10 for the list of lead SNPs. The number of associated traits ranges from 1 to 20, with mode at 4 traits (Fig. 8a). For each variant identified in BBJ, we compute

the number of associated traits in European population (SNP-specific FDR < 0.05). The level of pleiotropy in BBJ and that in European sample has a correlation of 0.34 ( $p < 2.2 \times 10^{-16}$ ) among approximately the same set of traits, showing evidence of replication and heterogeneity across populations. Consistent with the results for European ancestry (Fig. 4c), the level of pleiotropy is correlated with LD score with  $R^2 = 0.093$  (Fig. 8c). We identify 10 trait-specific lead SNPs in the BBJ (Fig. 8d). These SNPs are associated with breast cancer, prostate cancer, height, G-glutamyl transpeptidase (GGT), platelet count (PLT), and urolithiasis. In the GTEx data, these trait-specific SNPs are eQTLs for a small number of genes and tissues, though they appear less tissue- and gene-specific than the trait-specific SNPs identified in the European data (Supplementary Data 11). This could be due to the lack of eQTL data for East Asian ancestry, and GTEx (primarily European) cannot accurately reflect ancestry-matched eQTL effects possibly due to differences in allele frequencies and LD with causal variants.

## Discussion

In this paper, we present fastASSET for large-scale multi-trait genome-wide association analysis. fastASSET possesses two features simultaneously: 1) a *p*-value for global association and 2) a subset of traits associated with each variant. The power of fastASSET for global association is comparable to existing multi-trait methods (Supplementary Fig. 3). The subset selection feature, which is not possessed by the vast majority of multi-trait methods, is a unique advantage that enables us to study patterns of pleiotropy. In particular, for joint analysis of a wide range of traits, global association testing without specifying the set of associated traits is difficult to interpret. We observe that fastASSET estimates degree of pleiotropy and identifies associated traits more accurately than ad-hoc methods based on individual-trait *p*-values (Fig. 2). Compared to these ad-hoc methods for trait selection, fastASSET offers a formal way to screen out variants without any phenotypic effects. Although it is possible to extend fastASSET for multi-



**Fig. 8 | Results of analysis of Biobank Japan data.** **a** Distribution of level of pleiotropy. **b** Level of pleiotropy in Biobank Japan (BBJ) and European population. **c** Pleiotropy in BBJ vs. LD score for East Asian population. Shaded area along the blue line is the 95% confidence band. **d** Ten trait-specific SNPs identified in BBJ and

their top 10 associated traits. BC: breast cancer; GGT: G-glutamyl transpeptidase; PLT: platelet count; PrC: prostate cancer; Uro: urolithiasis.  $P$ -values are from two-sided z-tests for individual traits.

ancestry multi-trait analysis, it requires further development of methods and hence is left as a future direction. More specifically, genetic effect sizes are expected to be less heterogeneous for the same trait across ancestries than across different traits. Such varying levels of heterogeneity is not reflected by the current fastASSET.

A related type of methods try to deconvolve the genetic components that act together<sup>52,53</sup>. These methods typically focus on pleiotropic components that have effects on a widespread range of traits or at least a module of related traits. They tend to ignore trait-specific variants since those variants are unlikely to have major contributions to the components. fastASSET can be used to study both pleiotropic and trait-specific variants. In addition, deconvolution methods<sup>52,53</sup> do not directly quantify the level of pleiotropy for each variant, which would require further development of downstream methods.

We apply fastASSET to summary statistics of 116 traits to quantify degree of pleiotropy at the level of individual SNPs across the genome and their relationship with various annotation characteristics.

Specifically, we detect a much higher degree of pleiotropy at individual variant level than earlier studies and show that patterns of pleiotropy may be driven by multiple functional mechanisms. In a first-of-its-kind effort, we identify 21 highly trait-specific variants and conduct extensive follow-up studies to show that they can often be linked to the functions of specific genes in trait-related tissues.

While our analysis confirms the ubiquitous nature of pleiotropy reported in prior studies<sup>4,22</sup>, we are able to provide several insights due to both the use of a rapid powerful method for cross-trait association analysis and detailed follow-up investigation of relationship between the level of pleiotropy and functional characteristics across the SNPs. First, we observed that pleiotropy is ubiquitous, not only at a locus level but also at the level of individual variants. A prior large-scale study of pleiotropy across 558 traits based on UK Biobank had noted earlier that while >90% of loci are pleiotropic, about 60% of the individual SNPs show pleiotropic association with more than one trait and only 32% show association across multiple domains of traits<sup>4</sup>. Our analysis,



though based on a smaller number of traits, reveals that lead variants across almost all the identified loci are associated with more than one trait. The nature of pleiotropy, however, is “modular”<sup>21</sup>, i.e. a given variant has detectable effect only on a small fraction of the traits studied.

Second, our study confirms observations reported in previous studies<sup>4,39</sup> that variants affecting expression in multiple genes and tissues tend to be more pleiotropic, but we are able to detect the trend at a stronger resolution due to more precise characterization of the nature of pleiotropy. We further show that the level of pleiotropy is associated with the tissue-specificity of active chromatin states, and the relationship appears to be stronger for promoter and transcription related states than enhancers. Among SNPs located in enhancer regions, we found pleiotropy can be driven by interaction with multiple genes in multiple cell types. Finally, we also demonstrate enrichment of TF binding sites among pleiotropic SNPs.

Finally, a unique contribution of this paper is the identification of the highly unique trait-specific variants. We show that the trait-specific SNPs detected by fastASSET have dominant associations with the primary trait that are far stronger than any secondary associations (Fig. 5 and Supplementary Fig. 7) and they show similar pattern of trait specificity even when examined against a much larger number of traits in the UK Biobank study (Table 2). Despite their small number, the trait-specific SNPs can provide unique insight to biological mechanisms. We find that they often have clearly interpretable regulatory effects (Figs. 6 and 7). Some of them regulate the expression of a single gene in a single tissue. Others are active in multiple tissues, but the strongest regulatory effects occur in trait-related tissues. For some other SNPs the mechanisms are less obvious, but still clearly contrast with highly pleiotropic SNPs which can affect many genes in many tissues.

Our study has several limitations. In our discovery analysis, we restricted the analysis to a diverse but relatively small set of ( $K = 116$ ) traits that have large associated GWAS ( $N$  ranging from 7000 to 1.2 million;  $> 50\%$  studies have  $N > 100k$ ). As we have excluded vast number of other traits for which GWAS data are also available, we have not been able to provide a more complete genotype-phenotype maps. Nevertheless, we observe that the level of pleiotropy we observe for the detected SNPs with respect to a smaller number of traits correlates well with the level of pleiotropy observed in the UK Biobank study with respect to a much larger number of traits. It is particularly notable that the trait-specific SNPs we detect based on only 116 traits largely show same or similar trait specific effect when validated against more than 4000 traits in the UK Biobank. The correlation could be even higher if pleiotropy were quantified by fastASSET in UK Biobank. However, pleiotropic analysis of more than 4000 traits will be a massive task as it requires careful filtering of traits based on correlations. This analysis is beyond the scope of the current paper. A related limitation is to use the number of traits as the metric for pleiotropy. While also adopted by several other studies<sup>4,22</sup>, this metric depends on the set of traits being studied. We strive to eliminate the impact by selecting a broad range of traits and eliminating highly correlated traits. However, this issue could still have an impact and future studies are warranted on the optimal selection of traits or alternative metric of pleiotropy.

Although fastASSET has robust performance for estimating the level of pleiotropy when considering all the scenarios, it can be less accurate in some scenarios. For example, the estimated level of pleiotropy can be miscalibrated (Fig. 2). This is likely due to the meta-analysis procedure that underpins fastASSET, which is optimal when the SNP have similar effect size on all traits. In the original ASSET paper<sup>7</sup>, we acknowledged that the subset selection feature can be conservative. It tends to select the stronger effects and leave out some traits with weaker effects. However, selecting only the stronger associations allows us to focus on the main pathways underlying the genetic effects of the SNP, and ignore less important peripheral effects. Therefore, the results from fastASSET needs to be interpreted

carefully. SNPs could have effects on traits other than those selected by fastASSET. However, the association with main traits is much stronger, which is enough to justify that regulating these traits is the primary function of the SNP. For example, rs61926181 is selected as a trait-specific SNP for DBP, though it is also associated with SBP based on raw  $p$ -values. Nevertheless, the association with DBP is substantially stronger despite the two traits being highly correlated. In contrast, rs6035355 is classified to be associated with both DBP and SBP (Supplementary Data 6) since the significance level is nearly identical ( $p = 1.58 \times 10^{-8}$  for DBP,  $p = 1.77 \times 10^{-8}$  for SBP).

Another limitation of our study is that we have primarily identified pleiotropic associations based on single SNP association analysis without further exploring the underlying causal relationship among traits and SNPs. For example, pleiotropy between one SNP and two traits can be classified into multiple types<sup>54</sup>: 1) vertical pleiotropy due to causal relationship between two traits; 2) horizontal pleiotropy that arises from the SNP having independent effects on both traits 3) linkage disequilibrium between two distinct causal variants for different traits; 4) disease misclassification. Categories 3 and 4 are referred to as spurious pleiotropy in some literature<sup>54</sup>. The pleiotropy discussed in this paper can be viewed as pleiotropy at the locus level instead of the variant level. Identifying patterns of pleiotropy at the level of causal variants across many traits accounting for LD and heterogeneity in effect size is a challenging task. A more detailed characterization of pleiotropy at the level of each individual variant will ultimately require joint fine-mapping analysis across many traits by each locus, but methods for large-scale multi-trait fine-mapping are currently not available.

In summary, we develop a method and carry out a large-scale pleiotropic analysis across GWAS of a diverse set of traits. Our study provides insights into functional characteristics of the genome that contribute to pleiotropy and leads to the identification of unique trait-specific genetic variants which have not been previously explored. In the future, large-scale cross-trait fine-mapping studies are needed to pinpoint causal variants and the underlying nature of pleiotropy.

## Methods

### Data and initial filtering

We collect 338 full summary level datasets published between 2007 and 2019 from the NIH Genome-Wide Repository of Associations Between SNPs and Phenotypes (GRASP). GRASP includes a wide range of phenotypes including anthropometric traits, biomarkers, blood cell levels, adipose volume, early growth traits, social science Indices, cardiometabolic diseases, psychiatric and neurological diseases, autoimmune diseases, etc. We further collect 20 summary-level datasets from GWAS Consortia for the traits included in GRASP but when the Consortia offer a study with larger sample sizes. See Supplementary Data 1 and 2 for a full list of datasets. Thus restricting the analysis to the GRASP traits allowed us to carry out the pleiotropic analysis of across a diverse set of traits with large available GWAS, but avoid including many overlapping traits such as those in UK Biobank.

We apply several filtering steps sequentially (see Supplementary Fig. 1 for a flow chart): 1) remove datasets in which the genetics variants do not have genome-wide coverage (genotyped by exome array, metabochip and immunochip); 2) remove datasets that only report  $p$ -values without direction of effect; 3) remove studies with small sample size: for continuous traits, we remove studies with sample size  $< 5,000$ ; for binary traits, we remove studies with  $< 5000$  cases or  $< 5000$  controls; 4) remove studies where  $> 25\%$  individuals are of non-European ancestry; 5) remove duplicated traits: among the studies for the same trait, we keep the study with the largest sample size and discard the rest of them; 6) remove traits that are deterministic functions of other traits in our study, or other traits adjusted of covariates. After the filtering, 150 traits entered downstream analysis (see Supplementary Data 2 for all traits removed from analysis).

## LD score regression and further filtering

We applied LD score regression<sup>3,55</sup> to estimate the heritability and genetic correlation across 150 traits. LD scores based on 1000 Genomes European data and the list of HapMap3 SNPs were downloaded from the LDSC GitHub repository (<https://github.com/bulik/ldsc>). Only SNPs in HapMap3 were used to calculate the LD score and perform the regression.

To ensure the traits have a substantial genetic component, we remove 21 traits that do not have a heritability estimate that is significantly different from 0 ( $z$  statistic  $> 1.96$ ). To further reduce the genetic overlaps of the traits, we remove 13 traits that have high genetic correlation ( $r_g$ ) with others. In short, if two traits have genetic correlation  $|r_g| > 0.95$ , we remove the trait less enriched of genetic associations, quantified by the product of the sample size of the study ( $N$ ) and the heritability of the trait ( $h^2$ ). The algorithm is as follows:

1. Sort the traits in decreasing order of  $Nh^2$ .
2. Start from the trait with the highest  $Nh^2$ , remove the traits that have  $|r_g| > 0.95$  with this trait.
3. Proceed to the trait with the next highest  $Nh^2$ , repeat until no pairs of traits have  $|r_g| > 0.95$ .

After the above filtering steps, we retain a final list of 116 traits for statistical analysis (Supplementary Data 1). We further restricted to the variants that are available for  $\geq 50$  traits, and in 1000 Genomes Phase 3 European sample with minor allele frequency (MAF)  $> 0.01$ . This leads to a total of 7,462,466 variants. Since our goal is not to identify causal variants, the discrepancy of reference panels across studies should not have a major impact.

## Statistical Analysis Using fastASSET

The first step to study the patterns of pleiotropy is to quantify the level of pleiotropy across the genome. Previous studies often have quantified the pleiotropy of a SNP by counting the number of traits that reach  $p < 5 \times 10^{-8}$  in individual trait analysis. This approach, however, is likely to miss many weaker associations. Here we described fastASSET, an accelerated version of the ASSET method which 1) detects SNPs associated with any trait in our collection ("global association") and 2) reports a subset of traits associated with each SNP.

The original ASSET is developed based on meta-analysis across different phenotypes and searches for the subset of traits with maximum meta-analysis  $z$ -statistic<sup>7</sup>. However, since we are analyzing a large number of traits, the original ASSET which searches through all subsets is computationally intractable. To reduce the computational burden, we pre-select the traits that show suggestive evidence of association with the given SNP using a liberal  $p$ -value threshold, denoted by  $p_{thr}$  (e.g. 0.1 or 0.05). We only conduct subset search among the traits with  $p$ -value  $< p_{thr}$ . Denote the corresponding threshold for  $z$ -statistic by  $z_{thr} = \Phi^{-1}(1 - \frac{p_{thr}}{2})$ .

Before the subset search, the  $z$ -statistics  $z_k$  ( $k = 1, \dots, K$  traits) need to be adjusted for pre-selection to prevent type I error inflation. If the SNP is not associated with any of the given traits and none of the studies have overlapping samples,  $z_k$ 's are independent across studies and can be adjusted independently. The adjusted  $p$ -value is

$$\tilde{p}_k = P(|Z| > |z_k| \mid |Z| > z_{thr}) = \frac{P(|Z| > |z_k|, |Z| > z_{thr})}{P(|Z| > z_{thr})} = \frac{P(|Z| > |z_k|, |Z| > z_{thr})}{p_{thr}} \quad (1)$$

which is equal to  $\frac{P(|Z| > |z_k|)}{p_{thr}}$  if  $|z_k| > z_{thr}$ , and equal to 1 if  $|z_k| \leq z_{thr}$ . Hence the adjusted  $z$ -statistic is  $\tilde{z}_k = \text{sign}(z_k) \Phi^{-1}(1 - \frac{p_k}{2})$ .

However, the data used for our study have complex sample overlaps, hence  $\tilde{z}_k$ 's of overlapping studies are no longer independent. To apply the above adjustment, we first de-correlate the  $z$ -statistics. The correlation of  $z$ -statistics due to sample overlap can be estimated

using bivariate LD score regression<sup>3</sup>:

$$E[z_{ij}z_{2j} \mid l_j] = \frac{\sqrt{N_1N_2}\rho_g}{M}l_j + \rho_{12}^{(z)} \quad (2)$$

where  $\rho_{12}^{(z)}$  is the correlation of  $z_{ij}$  and  $z_{2j}$  under the null hypothesis and  $l_j$  is the LD score. Since sample overlap usually occurs between traits in the same study or consortium, the correlation matrix of  $z$ -statistics  $\rho^{(z)} = \{\rho_{kl}^{(z)}\}_{k,l=1,\dots,K}$  roughly follows a block diagonal structure. We first partition the traits into blocks using hierarchical clustering based on distance matrix  $1 - |\rho^{(z)}|$ . We cut the hierarchical clustering tree at 0.8 such that  $|\rho_{kl}^{(z)}| < 0.2$  if studies  $k$  and  $l$  are in different clusters. We de-correlate the  $z$ -statistics within each cluster and ignore between-cluster correlation.

Within each cluster, we first order the studies by effective sample size from smallest to largest. For continuous traits, the effective sample size is the total sample size; for binary traits, the effective sample size is defined as  $\frac{N_{case}N_{control}}{N_{case} + N_{control}}$ . Denote by  $\rho^{(z,t)}$  the LD score regression intercept matrix of the traits in cluster  $t$ . Denote by vector  $\mathbf{Z}^{(t)}$  the  $z$ -statistics of traits in cluster  $t$ , and by  $\mathbf{N}^{(t)}$  the effective sample size. If trait  $k$  is continuous,  $N_k^{(t)}$  is the total sample size; if trait  $k$  is case-control,  $N_k^{(t)} = \frac{N_{case}N_{control}}{N_{case} + N_{control}}$ . Define  $\mathbf{S}^{(t)} = \frac{1}{\sqrt{\mathbf{N}^{(t)}}}$  as the standard error when genotype and traits are standardized to have unit variance. The de-correlation and adjustment algorithm proceeds as follows:

1. Apply Cholesky decomposition to  $\rho^{(z,t)} = U^T U$ .
2. De-correlated the  $z$ -statistics by  $\tilde{\mathbf{Z}}^{(t)} = (U^T)^{-1}\mathbf{Z}^{(t)}$ . Select the traits with  $|\tilde{z}_k^{(t)}| > z_{thr}$  and adjust the  $z$ -statistics independently for each trait using the conditional  $p$ -value approach described by Eq. (1). Denote the adjusted  $z$ -statistics by  $\tilde{\mathbf{Z}}_{adj}^{(t)}$ .
3. Convert the  $z$ -statistics back to the original scale by  $\mathbf{Z}_{adj}^{(t)} = U^T \tilde{\mathbf{Z}}_{adj}^{(t)}$ . Leave the standard errors  $\mathbf{S}^{(t)}$  unchanged. Retain the elements of  $\mathbf{Z}_{adj}^{(t)}$  and  $\mathbf{S}^{(t)}$  corresponding to the selected traits and denote them as  $\mathbf{Z}_{scr}^{(t)}$  and  $\mathbf{S}_{scr}^{(t)}$ .

Finally, we combine the adjusted  $z$ -statistics and standard errors ( $\mathbf{Z}_{scr}^{(t)}$  and  $\mathbf{S}_{scr}^{(t)}$ ,  $t = 1, \dots, T$ ) for the selected traits from all clusters. The combined summary statistics are used as input for ASSET analysis (see Supplementary Notes for a brief description of ASSET)<sup>7</sup>.

## Analysis of data for 116 traits

Since different studies are relatively independent, LD score regression<sup>3</sup> analysis did not reflect substantial sample overlap across different studies. The clusters, based on sample overlap and phenotypic correlation are generally small and restricted to traits within the same study (Supplementary Fig. 2). For a small proportion of the SNPs, a large number of traits pass the pre-selection threshold (single-trait  $p$ -value  $< 0.05$ ) which makes the subsequent subset search computationally intractable. We removed a total of 16,686 SNPs that were associated with more than 16 traits in one direction (either positive or negative). Under the global null hypothesis of no association, the probability of observing such pattern is small ( $1.13 \times 10^{-8}$ ) and thus we consider these SNPs to be pleiotropic though they are not further analyzed by ASSET. In fact, the majority of these SNPs can be tagged by SNPs that are analyzed by ASSET through LD ( $r^2 > 0.5$ ) and hence removing them does not lead to significant loss of information, except 309 SNPs from 74 independent regions ( $r^2 < 0.1$ ,  $> 500$  kb apart). We choose the SNP with the largest number of traits that pass  $p < 0.05$  threshold as the index SNP for each locus. As a sensitivity analysis, we allocate additional computational resources to run fastASSET analysis

for the 74 lead SNPs. These SNPs are enriched in pleiotropic SNPs, but some of the SNPs are associated with only a few traits (Supplementary Fig. 10). Due to their small number, they should not have a major impact on the characterization of pleiotropy.

Among the SNPs we analyzed by ASSET, we further removed 2620 SNPs for which one of the one-sided ASSET  $p$ -values is smaller than the  $p$ -values from standard meta-analysis of the selected traits, indicating convergence issues in the underlying  $p$ -value approximation method.

For each of the remaining 7,443,160 SNPs, fastASSET reports a  $p$ -value for global association and a set of traits associated with the SNP. For each SNP, we use the number of its associated traits reported by fastASSET as the metric of pleiotropy. We consider SNPs with fastASSET  $p$ -value  $p < 5 \times 10^{-8}$  as genome-wide significant. LD clumping leads to 10,628 independently associated SNPs with  $r^2 < 0.1$ . We then group these SNPs into 2293 independent loci of which the lead SNP (lowest  $p$ -value within the locus) are at least 500 kb apart. As a metric of signal density within each locus, we count the number of independently associated SNPs within 100 kb of the lead SNP.

We are especially interested in trait-specific SNPs (associated with only 1 trait as identified by fastASSET) and secondarily highly pleiotropic variants with associated with  $>15$  traits.

### Validation in UK Biobank

We download the UK Biobank (UKB) GWAS summary statistics from Neale lab<sup>34</sup>. Among the 2293 index SNPs of the independent loci identified by fastASSET, 6 SNPs are not included in the UK Biobank summary statistics. For each of those SNPs, we identify a proxy SNP in the UKB summary statistics as the one with the highest  $r^2$  within 100 kb of the lead SNP. If  $r^2 < 0.8$  between the lead SNP and the proxy SNP, we exclude the locus from the validation study in UKB. We successfully found proxies for 5 index SNPs and one other (rs12203592) does not have a proxy SNP with  $r^2 \geq 0.8$ .

Among the 11,934 summary level datasets published by Neale lab (round 2 GWAS, August 1st, 2018), we removed the GWAS results for age, sex and 22 datasets that do not have a phenotype code. For continuous traits, we use the summary dataset for the inverse rank normalized trait (variable type continuous\_irnt) and discard the dataset for the raw trait (variable type continuous\_raw). We only keep the datasets for joint analyzes across both sexes and discard the sex-stratified results. In addition, there are two versions of GWASs for a subset of 166 traits. We keep the most recent version (v2) and remove the first version. After filtering, we retain 4114 summary level datasets for our validation study. For each of the 2292 index SNPs, we select the associated traits among the remaining 4114 UKB traits by a per-SNP FDR threshold of 0.05. Note that we are not able to conduct an independent validation study due to substantial sample overlap between our primary discovery data and UKB. Instead, we use UKB data to test the generalizability of the estimated degree of pleiotropy from 116 traits to a larger number of traits.

### Relationship between pleiotropy and LD

To study the relationship between pleiotropy and LD, we estimate LD scores from 1000 Genomes Phase 3 EUR sample using SNPs with MAF  $> 0.01$  and 1 centiMorgan (cM) window. The calculation is performed using ldsc software<sup>55</sup>. Note that here we choose to estimate LD scores using all the SNPs in our analysis instead of using those downloaded from LDSC, which include only HapMap3 SNPs.

### eQTL status lookup and colocalization

To explore potential cis-regulatory mechanisms that may drive pleiotropy, we look up the lead SNPs of the 2293 loci identified by fastASSET to explore the relationship between the pleiotropy and eQTL tissue/gene specificity. We accessed eQTL summary statistics for 49 tissues in GTEx v8<sup>39</sup>. We lift over the base pair coordinates from hg19 to hg38 to match the genome build of GTEx v8. For each

of the lead variant, we count the number of tissues in which it is a significant eQTL for at least one gene with  $q$ -value  $< 0.05$ , and the total number of unique genes for which it is a significant eQTL regardless of tissues.

For each trait-specific locus, we conduct colocalization analysis between the GWAS signal for the corresponding trait and eQTL signals in GTEx v8. We restrict to protein-coding genes and gene-tissue pairs for which the lead SNP of the locus is a significant eQTL at  $q$ -value  $< 0.05$ . We use the SNPs within 50 kb from the lead SNP as input to COLOC<sup>47</sup>. We consider the GWAS and eQTL signals to be colocalized if the posterior probability of colocalization (PP4)  $> 0.8$ .

### Chromatin state

To explore potential relationship between pleiotropy and chromatin states, we queried their chromatin state using 15-state chromHMM model<sup>56</sup> in the HaploReg v4.1 database<sup>37</sup>. The chromatin state was learned using a core set of 5 histone marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3) for 111 epigenomes from the Roadmap Epigenomics Project<sup>48</sup>. We defined chromatin states with state number  $\leq 7$  as an active and open chromatin state, including active transcription start site (1\_TssA), flanking active TSS (2\_TssAFlnk), transcription at gene 5' and 3' (3\_TxFlnk), strong transcription (4\_Tx), weak transcription (5\_TxWk), genic enhancers (6\_EnhG) and enhancers (7\_Enh). For each variant, we count the number of datasets in Roadmap whether it falls in active chromatin state as well as the number of broad categories provided by Roadmap ([https://egg2.wustl.edu/roadmap/web\\_portal/meta.html](https://egg2.wustl.edu/roadmap/web_portal/meta.html)). We also study and evaluate the association of pleiotropy with each active chromatin state separately. For example, we select the lead SNPs identified by fastASSET that are 1\_TssA in at least one Roadmap dataset and investigate the relationship between the number of traits selected by fastASSET and the number of tissue/cell type groups in which this variant is in 1\_TssA. We perform same analysis for other active chromatin states to learn the association. The tissue and cell type groups are defined by the Roadmap Epigenomics Consortium as provided in the link: (<https://docs.google.com/spreadsheets/d/1yikGx4MsO9Ei36b64yOy9Vb6oPC5IBGfBYEt-N6gOM/edit#gid=15>).

In addition to chromHMM based model, we conduct similar analysis using the chromatin states learned by the IDEAS, an integrative and discriminative epigenome annotation system employing 2D genome segmentation method<sup>35,36</sup> to jointly characterize the chromatin states across many different cell types. Since the IDEAS classification of chromatin states is marginally different from ChromHMM, we consider 10\_TssA (Active Transcription Start Site), 8\_TssAFlnk (Flanking Active TSS), 14\_TssWk (Weak TSS), 5\_Tx (Strong Transcription), 2\_TxWk (Weak Transcription), 4\_Enh (Enhancers), 6\_EnhG (Genic Enhancers), 17\_EnhGA (Active Genic Enhancers) as an active and open chromatin state. IDEAS algorithm here integrates epigenomes by preserving the position-dependency and cell type specific epigenetic events at fine scales<sup>36</sup>.

The IDEAS bigBed files were downloaded from [http://bx.psu.edu/~yuzhang/Roadmap\\_ideas/](http://bx.psu.edu/~yuzhang/Roadmap_ideas/).

The bigBed files were converted to bed format using UCSC program bigBedToBed program fetched from the directory of binary utilities in <http://hgdownload.cse.ucsc.edu/admin/exe/>.

The bed files thus processed were used for further downstream analysis, including the association of pleiotropy and epigenetic states. Blacklisted genomic region was filtered out where applicable as provided by ENCODE.

### Transcription factor binding sites

To understand potential relationship between pleiotropy transcription factor (TF) binding, we downloaded and referenced the JASPAR 2018<sup>37</sup> and HOCOMOCO V11<sup>38</sup>, homosapiens comprehensive model collection of motif database. Intersection of variants with TF binding sites was



performed by Bedtools v2.29.2<sup>58</sup> to observe the association between SNPs and transcription factor binding sites.

To associate level of pleiotropy with TF binding profiles, we divide the 2,293 lead SNPs into two pleiotropy bins: 1) associated with 1–10 traits 2) associated with >10 traits. We calculate the proportion of SNPs in each bin that overlaps with transcription binding sites (TFBS). To explore whether this relationship is independent of the effects of other annotations, we fit the following logistic regression: (associated with >10 traits)  $\sim$  (overlapping with TFBS) + (LD score) + (number of tissues for which the variant is an eQTL) + (number of eGenes) + (number of cell types for which the variant is in active chromatin state) and check the significance of the first predictor.

### Enhancer-gene connection

The activity-by-contact (ABC) model combines chromatin states with 3-dimensional contacts to map enhancers to their target genes<sup>59,60</sup>. We download the enhancer-gene map for 131 human cell types and tissues constructed by the ABC model from the Engreitz lab website (<https://www.engreitzlab.org/resources/>). This dataset includes all enhancer-gene connections with ABC scores  $\geq 0.015$ .

Among the lead SNPs that overlap with enhancers, we compute the correlation between the level of pleiotropy and the number of cell types for which the overlapping enhancer affects at least one target gene by the ABC model. We further compute the partial correlation adjusting for other annotations: LD score, number of tissues for which the variant is an eQTL, number of eGenes, number of cell types for which the variant is in active chromatin state.

### Matching trait-specific SNPs to highly pleiotropic SNPs

In the functional follow-up studies of 21 trait-specific SNPs, we match each of them to one SNP that is also associated with the given trait but show high-degree of pleiotropy (associated with >15 traits). For each trait-specific SNP (e.g. for Crohn's disease), we examine the individual-trait p-value between the trait (e.g. Crohn's disease) and all the highly pleiotropic SNPs and choose the one that shows the strongest association (lowest p-values) with the given trait (e.g. Crohn's disease). In this procedure, multiple trait-specific SNPs for one trait could be matched to the same highly pleiotropic SNP. This matching procedure applies to the examination of top 10 associated traits for lead SNPs (Fig. 5), eQTL analysis (Fig. 6) and chromatin state analysis (Fig. 7).

### Analysis of Biobank Japan data

We download the summary statistics for 220 phenotypes in Biobank Japan (BBJ)<sup>22</sup>. We run LD score regression<sup>3</sup> to estimate the genetic correlation (slope) and phenotypic correlation (intercept) of these traits. LD scores of East Asian population are obtained from GitHub repository *ldsc* (<https://github.com/bulik/ldsc>). We apply similar quality control pipeline as in the analysis of European data: 1) remove continuous traits with total sample size <5,000 or case-control disease traits with <2,000 cases or <2,000 controls; 2) remove traits with heritability z-score <1.96; 3) remove highly correlated traits with genetic correlation > 0.95 or < -0.95; 4) remove SNPs with MAF < 0.01 in BBJ. After filtering, we retain 75 traits and 4,965,789 SNPs for fastASSET analysis. Genome-wide significant SNPs are defined as those with fastASSET global association p-value <  $5 \times 10^{-8}$ . We conduct LD clumping based on fastASSET p-values for global association using genotype data of 1000 Genomes Phase 3 East Asian population (N = 504) as the reference panel. We require the lead SNPs of different loci to be nearly independent ( $r^2 < 0.1$ ) and at least 500 kb apart. To compare patterns of pleiotropy between East Asian and European ancestry, we obtain the summary statistics for European ancestry GWAS used in Sakaue et al.<sup>22</sup> (UK Biobank and FinnGen). We successfully match 72 (out of 75) traits from BBJ in our analysis to traits in UK Biobank and FinnGen.

### Inclusion & ethics statement

This study uses publicly available data. Ethics approvals were obtained by the original studies that generated the data.

### Reporting summary

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

### Data availability

All data used in this paper are publicly available. Summary statistics from the GRASP repository are available for download at <https://grasp.nhlbi.nih.gov/Overview.aspx>. Other summary data are available through links provided by the original publication (see Supplementary Data 1 for references). Biobank Japan GWAS summary statistics are available at the National Bioscience Database Center (NBDC) Human Database with the accession code hum0197 (<https://humandbs.dbcls.jp/en/hum0197-v18>). Functional genomic data can be accessed via links below. GTEx Consortium: <https://www.gtexportal.org/home/>. Roadmap Epigenomics Project: [https://egg2.wustl.edu/roadmap/web\\_portal/index.html](https://egg2.wustl.edu/roadmap/web_portal/index.html). Neale lab UK Biobank GWAS: <http://www.nealelab.is/uk-biobank>. JASPAR database for transcription factor binding profiles: <http://jaspar.genereg.net/>. ABC model: <https://www.engreitzlab.org/resources/>. Biobank Japan GWAS and European ancestry validation GWAS summary statistics: <https://humandbs.dbcls.jp/en/hum0197-v18>. Source data are provided with this paper.

### Code availability

fastASSET (v0.1.0)<sup>61</sup> is available as an R package via GitHub repository <https://github.com/gqi/fastASSET>. It is also available as a function in the ASSET R package (<https://github.com/sbstatgen/ASSET>, function *fast\_asset*). LD score regression was performed using LDSC (v1.0.1): <https://github.com/bulik/ldsc>. Data collection and analysis were conducted using R v4.0.2 and 4.3.2.

### References

- Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–D1012 (2019).
- Visscher, P. M. & Yang, J. A plethora of pleiotropy across complex traits. *Nat. Genet.* **48**, 707–708 (2016).
- Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
- Watanabe, K. et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
- Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
- Ray, D. & Boehnke, M. Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genet. Epidemiol.* **42**, 134–145 (2018).
- Bhattacharjee, S. et al. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.* **90**, 821–835 (2012).
- O'Reilly, P. F. et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLOS ONE* **7**, e34861 (2012).
- Qi, G. & Chatterjee, N. Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits. *PLOS Genet* **14**, e1007549 (2018).
- Turley, P. et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
- Ray, D. & Chatterjee, N. A powerful method for pleiotropic analysis under composite null hypothesis identifies novel shared loci between Type 2 Diabetes and Prostate Cancer. *PLOS Genet* **16**, e1009218 (2020).



12. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?\*. *Int. J. Epidemiol.* **32**, 1–22 (2003).
13. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–R98 (2014).
14. Zheng, J. et al. Recent developments in mendelian randomization studies. *Curr. Epidemiol. Rep.* **4**, 330–345 (2017).
15. Qi, G. & Chatterjee, N. Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nat. Commun.* **10**, 1941 (2019).
16. Verma, A. et al. PheWAS and beyond: the landscape of associations with medical diagnoses and clinical measures across 38,662 individuals from geisinger. *Am. J. Hum. Genet.* **102**, 592–608 (2018).
17. Diogo, D. et al. Phenome-wide association studies across large population cohorts support drug target validation. *Nat. Commun.* **9**, 4285 (2018).
18. Verma, A. et al. Human-disease phenotype map derived from PheWAS across 38,682 individuals. *Am. J. Hum. Genet.* **104**, 55–64 (2019).
19. Shen, X. et al. A phenome-wide association and Mendelian randomisation study of polygenic risk for depression in UK biobank. *Nat. Commun.* **11**, 2301 (2020).
20. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).
21. Stearns, F. W. One hundred years of pleiotropy: a retrospective. *Genetics* **186**, 767–773 (2010).
22. Sakaue, S. et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).
23. Jordan, D. M., Verbanck, M. & Do, R. HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. *Genome Biol.* **20**, 222 (2019).
24. Chen, C.-Y. et al. Analysis across Taiwan Biobank, Biobank Japan, and UK Biobank identifies hundreds of novel loci for 36 quantitative traits. *Cell Genom.* **3**, 100436 (2023).
25. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
26. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022–1034.e6 (2019).
27. Li, Y. R. et al. Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nat. Med.* **21**, 1018–1027 (2015).
28. Kar, S. P. et al. Genome-wide meta-analyses of breast, ovarian, and prostate cancer association studies identify multiple new susceptibility loci shared by at least two cancer types. *Cancer Discov.* **6**, 1052–1067 (2016).
29. Hung, R. J. et al. Cross cancer genomic investigation of inflammation pathway for five common cancers: lung, ovary, prostate, breast, and colorectal cancer. *JNCI J. Natl. Cancer Inst.* **107**, djv246 (2015).
30. Cichonska, A. et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinforma. Oxf. Engl.* **32**, 1981–1989 (2016).
31. van der Meer, D. et al. Understanding the genetic determinants of the brain with MOSTest. *Nat. Commun.* **11**, 3512 (2020).
32. Julienne, H. et al. JASS: command line and web interface for the joint analysis of GWAS results. *NAR Genomics Bioinforma.* **2**, lqaa003 (2020).
33. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
34. Neale, B. M. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank. <http://www.nealelab.is/uk-biobank> (2018).
35. Zhang, Y., An, L., Yue, F. & Hardison, R. C. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res* **44**, 6721–6731 (2016).
36. Zhang, Y. & Hardison, R. C. Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. *Nucleic Acids Res* **45**, 9823–9836 (2017).
37. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res* **46**, D260–D266 (2018).
38. Kulakovskiy, I. V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* **46**, D252–D259 (2018).
39. Consortium, T. Gte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
40. Stelzer, G. et al. The genecards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinforma.* **54**, 1.30.1–1.30.33 (2016).
41. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet.* **55**, 1866–1875 (2023).
42. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).
43. Mandelkow, E.-M. & Mandelkow, E. Tau in Alzheimer's disease. *Trends Cell Biol.* **8**, 425–427 (1998).
44. Butcher Matthew, J., Gjurich Breanne, N., Phillips, T. & Galkina Elena, V. The IL-17A/IL-17RA axis plays a proatherogenic role via the regulation of aortic myeloid cell recruitment. *Circ. Res.* **110**, 675–687 (2012).
45. Kabashima, K. et al. Prostaglandin E2-EP4 signaling initiates skin immune responses by promoting migration and maturation of Langerhans cells. *Nat. Med.* **9**, 744–749 (2003).
46. Singh, S. B. et al. Human IRGM regulates autophagy and cell-autonomous immunity functions through mitochondria. *Nat. Cell Biol.* **12**, 1154–1165 (2010).
47. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLOS Genet* **10**, e1004383 (2014).
48. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
49. Cummings, J. L. & Cole, G. Alzheimer Disease. *JAMA* **287**, 2335–2338 (2002).
50. K., I., F., L., C. -X., G. & I., G.-I. Tau in Alzheimer disease and related tauopathies. *Curr. Alzheimer Res.* **7**, 656–664 (2010).
51. Boyle, P. A., Buchman, A. S., Wilson, R. S., Leurgans, S. E. & Bennett, D. A. Association of muscle strength with the risk of alzheimer disease and the rate of cognitive decline in community-dwelling older persons. *Arch. Neurol.* **66**, 1339–1344 (2009).
52. Tanigawa, Y. et al. Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. *Nat. Commun.* **10**, 4064 (2019).
53. Ballard, J. L. & O'Connor, L. J. Shared components of heritability across genetically correlated traits. *Am. J. Hum. Genet.* **109**, 989–1006 (2022).
54. van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H. & Wray, N. R. Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet.* **20**, 567–581 (2019).
55. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
56. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).
57. Ward, L. D. & Kellis, M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res* **44**, D877–D881 (2016).
58. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841 (2010).

59. Fulco, C. P. et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
60. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
61. Qi, G. et al. Genome-wide large-scale multi-trait analysis characterizes global patterns of pleiotropy and unique trait-specific variants. Zenodo <https://doi.org/10.5281/zenodo.12210515> (2024).

## Acknowledgements

Research of Guanghao Qi, Diptavo Dutta and Nilanjan Chatterjee was supported by an R01 grant from the National Human Genome Research Institute 1 R01 HG010480-01 (NC). Debashree Ray was supported by grant R03DE029254 (DR) from the NIH.

## Author contributions

G.Q., S.B., and N.C. conceived the idea. G.Q. and S.B. implemented the statistical method. G.Q. performed the simulation studies and data analyzes. S.B.C. performed the analyzes on chromatin accessibility and transcription factor binding sites. D.R., D.D., and A.B. provided feedback on the methods and data analyzes. S.B. and N.C. supervised the project. G.Q. drafted the paper. All authors edited the paper.

## Competing interests

AB is a co-founder of CellCIPHER, Inc, is a shareholder in Alphabet, Inc, and has consulted for Third Rock Ventures, LLC. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-51075-5>.

**Correspondence** and requests for materials should be addressed to Samsiddhi Bhattacharjee or Nilanjan Chatterjee.

**Peer review information** *Nature Communications* thanks Raymond Walters, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is 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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024