








ARTICLE

<https://doi.org/10.1038/s41467-019-11874-7>

OPEN

# Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits

Wen Zhang <sup>1,2,11</sup>, Georgios Voloudakis <sup>1,11</sup>, Veera M. Rajagopal <sup>1,2,3,4,5</sup>, Ben Readhead<sup>2,6</sup>, Joel T. Dudley<sup>2</sup>, Eric E. Schadt <sup>2</sup>, Johan L.M. Björkegren <sup>2,7,8,9</sup>, Yungil Kim<sup>1,2</sup>, John F. Fullard<sup>1,2</sup>, Gabriel E. Hoffman <sup>2</sup> & Panos Roussos <sup>1,2,10</sup>

Transcriptome-wide association studies integrate gene expression data with common risk variation to identify gene-trait associations. By incorporating epigenome data to estimate the functional importance of genetic variation on gene expression, we generate a small but significant improvement in the accuracy of transcriptome prediction and increase the power to detect significant expression-trait associations. Joint analysis of 14 large-scale transcriptome datasets and 58 traits identify 13,724 significant expression-trait associations that converge on biological processes and relevant phenotypes in human and mouse phenotype databases. We perform drug repurposing analysis and identify compounds that mimic, or reverse, trait-specific changes. We identify genes that exhibit agonistic pleiotropy for genetically correlated traits that converge on shared biological pathways and elucidate distinct processes in disease etiopathogenesis. Overall, this comprehensive analysis provides insight into the specificity and convergence of gene expression on susceptibility to complex traits.

<sup>1</sup> Department of Psychiatry, Pamela Sklar Division of Psychiatric Genomics and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. <sup>2</sup> Department of Genetics & Genomic Sciences and Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York 10029 NY, USA. <sup>3</sup> Department of Biomedicine, Aarhus University, 8000 Aarhus, Denmark. <sup>4</sup> The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Aarhus University, 8000 Aarhus, Denmark. <sup>5</sup> Centre for Integrative Sequencing (iSEQ), Aarhus University, 8000 Aarhus, Denmark. <sup>6</sup> The Biodesign Institute, Arizona State University, Tempe, AZ 85281, USA. <sup>7</sup> Clinical Gene Networks AB, 114 44 Stockholm, Sweden. <sup>8</sup> Vascular Biology Unit, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, 171 77 Stockholm, Sweden. <sup>9</sup> Department of Physiology, Institute of Biomedicine and Translation Medicine, University of Tartu, 50411 Tartu, Estonia. <sup>10</sup> Mental Illness Research Education and Clinical Center (MIRECC), James J. Peters VA Medical Center, Bronx 10468 NY, USA. <sup>11</sup> These authors contributed equally: Wen Zhang, Georgios Voloudakis. Correspondence and requests for materials should be addressed to P.R. (email: [panagiotis.roussos@mssm.edu](mailto:panagiotis.roussos@mssm.edu))

Despite the recent success of genome-wide association studies (GWASs) in furthering our understanding of the genetic basis of disease, the mechanisms through which many of the identified risk variants act remain largely unknown<sup>1</sup>. Disease-associated risk variants are highly enriched in *cis* regulatory elements (CREs), including promoters and enhancers<sup>2,3</sup> and increasing evidence suggests that they affect the regulation of gene expression<sup>2–6</sup>. Multiple computational methods have been developed to perform transcriptome-wide association studies (TWASs) linking risk variants with differential gene expression<sup>7–11</sup>. For instance, using the summary data-based Mendelian randomization method, Zhu and colleagues conducted a TWAS for complex traits by integrating eQTL and GWAS summary data<sup>9</sup>. However, the field is increasingly favoring transcriptomic imputation methods as the basis of TWAS applications, as they enable feature-centered modeling of the combined effect of multiple *cis*-SNPs (SNPs in proximity to the transcription start site) on transcription. The two most widely used methods are PrediXcan and FUSION. Gusev et al. developed the latter method and were the first to apply it to GWAS summary statistics to explore genetic mechanisms for complex traits<sup>11</sup>. On the other hand, PrediXcan<sup>12</sup> is the first, and most widely used, transcriptomic imputation method for individual genotypes that was adapted for use with GWAS summary statistics and, so far, it outperforms similar methods<sup>13</sup>. Briefly, PrediXcan uses elastic net (ENet) regression models, trained in a reference transcriptome, to impute gene expression. The models use a set of *cis*-SNPs as linear predictors of gene expression. The imputed expressions are then correlated with the phenotype of interest to identify gene-trait associations (GTAs). The generated trait-associated imputed transcriptomes can also be leveraged for diverse downstream applications such as the identification of candidate compounds, for which we have reference transcriptomic data, that are predicted to reverse trait-specific, genetically driven, gene expression changes<sup>14</sup>. These downstream applications depend on the prediction accuracy of the genetically regulated expression (GREX) and, thus, any improvements in the transcriptomic imputation performance would translate to higher confidence in the GREX-based drug repositioning predictions.

Here, we present EpiXcan, a method that increases prediction accuracy in transcriptome imputation by integrating epigenetic data to model the prior probability that a SNP affects transcription. EpiXcan specifically leverages annotations derived from the Roadmap Epigenomics Mapping Consortium (REMC) that integrates multiple epigenetic assays, including DNA methylation, histone modification and chromatin accessibility<sup>15</sup>. The rationale of our approach is that SNPs within CREs are more likely to be functionally relevant<sup>16</sup>. We utilize 14 large-scale transcriptome datasets of genotyped individuals to train prediction models and integrate with 58 complex traits and diseases to define significant GTAs. GTAs exhibit significant enrichment for relevant biological pathways and known genes linked to trait-related phenotypes in humans and mice. Imputed transcriptomic changes are used to identify known compounds that can normalize genetically driven expression perturbations. Chemogenomic enrichment analyses are performed and an agnostic approach is proposed to validate drug predictions. Pairwise trait analysis identifies genes that exhibit agonistic pleiotropy for genetically correlated traits that converge on shared biological pathways. Finally, bi-directional regression analysis identifies putative causal relationships among traits. Overall, our analysis provides insight into the specificity and convergence of gene expression mediating the genetic risk architecture underlying susceptibility to complex traits and diseases.

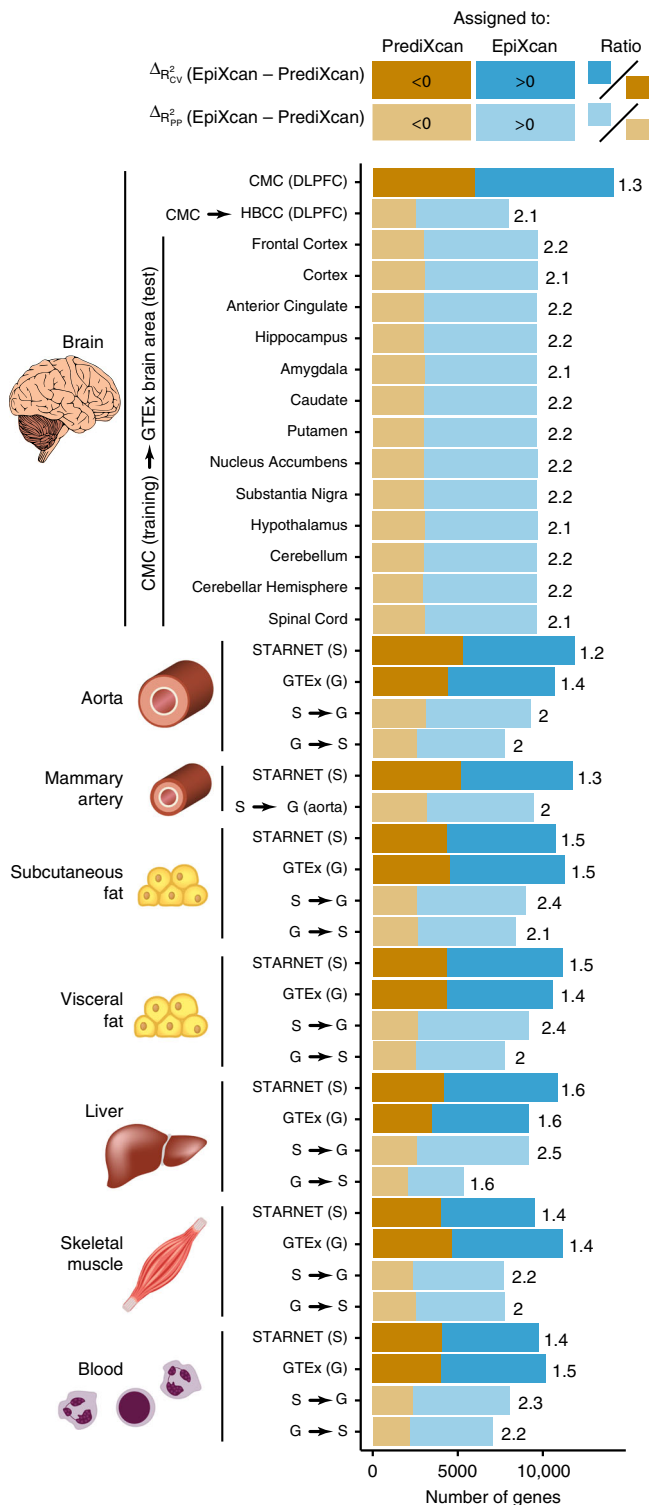
## Results

**EpiXcan outperforms PrediXcan.** Since TWAS is limited to genes that can be accurately predicted from genotype data, increasing prediction accuracy can increase the scope and power of analyses. Here, we integrate biologically relevant data in a single framework to improve performance of gene expression prediction. The overall schematic of EpiXcan is shown in Supplementary Fig. 1. Briefly, EpiXcan leverages epigenetic annotation to inform transcriptomic imputation by employing a three-step process (see Methods and Supplementary Methods): (1) estimate SNP priors that reflect the likelihood of a SNP having a regulatory role in gene expression based on a Bayesian hierarchical model<sup>17</sup> that integrates epigenomic annotation<sup>15</sup> and eQTL summary statistics for *cis*-SNPs (SNPs located  $\pm 1$  Mb from the transcription start site of the gene); (2) rescale the SNP priors to penalty factors by employing an adaptive mapping approach; and (3) use the genotypes and penalty factors in weighted elastic net to perform gene expression prediction.

Using simulated data, we apply EpiXcan and PrediXcan to train prediction models and estimate the adjusted cross-validation R-squared ( $R^2_{CV}$ ), which is the correlation between the predicted and observed expression levels during the nested cross validation. Although the actual  $R^2_{CV}$  achieved by both methods is generally low, in all simulated scenarios, EpiXcan improves the average  $R^2_{CV}$  compared to PrediXcan models (all  $p$  values  $\leq 7 \times 10^{-10}$  based on one-sample sign test; Supplementary Fig. 2). We then train prediction models by applying EpiXcan and PrediXcan in 14 RNAseq datasets, derived from dorsolateral prefrontal cortex (DLPFC) from the CommonMind Consortium (CMC)<sup>18</sup>, seven tissues from Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET)<sup>19</sup> and six tissues from GTEx<sup>20</sup> (Supplementary Table 1). We compare the performance of EpiXcan with PrediXcan by considering the delta value (EpiXcan minus PrediXcan) of two metrics: (1) cross-validation  $R^2$  ( $R^2_{CV}$ ) within each tissue and (2) predictive performance  $R^2$  ( $R^2_{PP}$ ), estimated based on Pearson's correlation between predicted and observed expression in an independent dataset of a relevant tissue. Positive delta values indicate that EpiXcan has higher prediction performance compared to PrediXcan.

Across all datasets, EpiXcan improves the average  $R^2_{CV}$  compared to PrediXcan (all  $p$  values  $\leq 9 \times 10^{-16}$  based on one-sample sign test; Fig. 1; Supplementary Figs. 3, 4; Supplementary Data 1). We predict 4.6% more genes (pairwise Wilcoxon test  $p$  value =  $6.10 \times 10^{-5}$ ) with  $R^2_{CV} > 0.01$  using EpiXcan (average number of genes across tissues is 10,181) compared to PrediXcan (average number of genes across tissues is 9760). To obtain the second metric,  $R^2_{PP}$ , we train prediction models in the training dataset, which are then used to predict expressions in the test dataset. Across all datasets, EpiXcan improves the average  $R^2_{PP}$  compared to PrediXcan (all  $p$  values  $< 9 \times 10^{-16}$  based on one-sample sign test; Fig. 1; Supplementary Figs. 5–7; Supplementary Data 2). Importantly, the ratios of genes predicted more effectively by EpiXcan are higher in the independent dataset evaluation ( $R^2_{PP}$ ) than in the cross-validation (unpaired  $t$ -test,  $p$  value =  $3.3 \times 10^{-17}$ ) (Fig. 1), suggesting that the adaptive rescaling of the penalty factors during model training does not result in significant overfitting that could affect the external validity of the models. Overall, compared to PrediXcan, EpiXcan has improved predictive performance and identifies more genes that can be used for TWAS.

In addition, we compare EpiXcan to recently developed predictive methods such as the Bayesian sparse linear mixed model (BSLMM)<sup>21</sup> and the Dirichlet process regression (DPR) method<sup>22</sup> (Methods; Supplementary Methods). EpiXcan



outperforms BSLMM and DPR in transcriptomic imputation, both in cross-validation and in independent datasets (all  $p$  values  $< 7 \times 10^{-16}$ ) while having on average, depending on the method, from 0.63x to ~240x the computing speed (Supplementary Fig. 8).

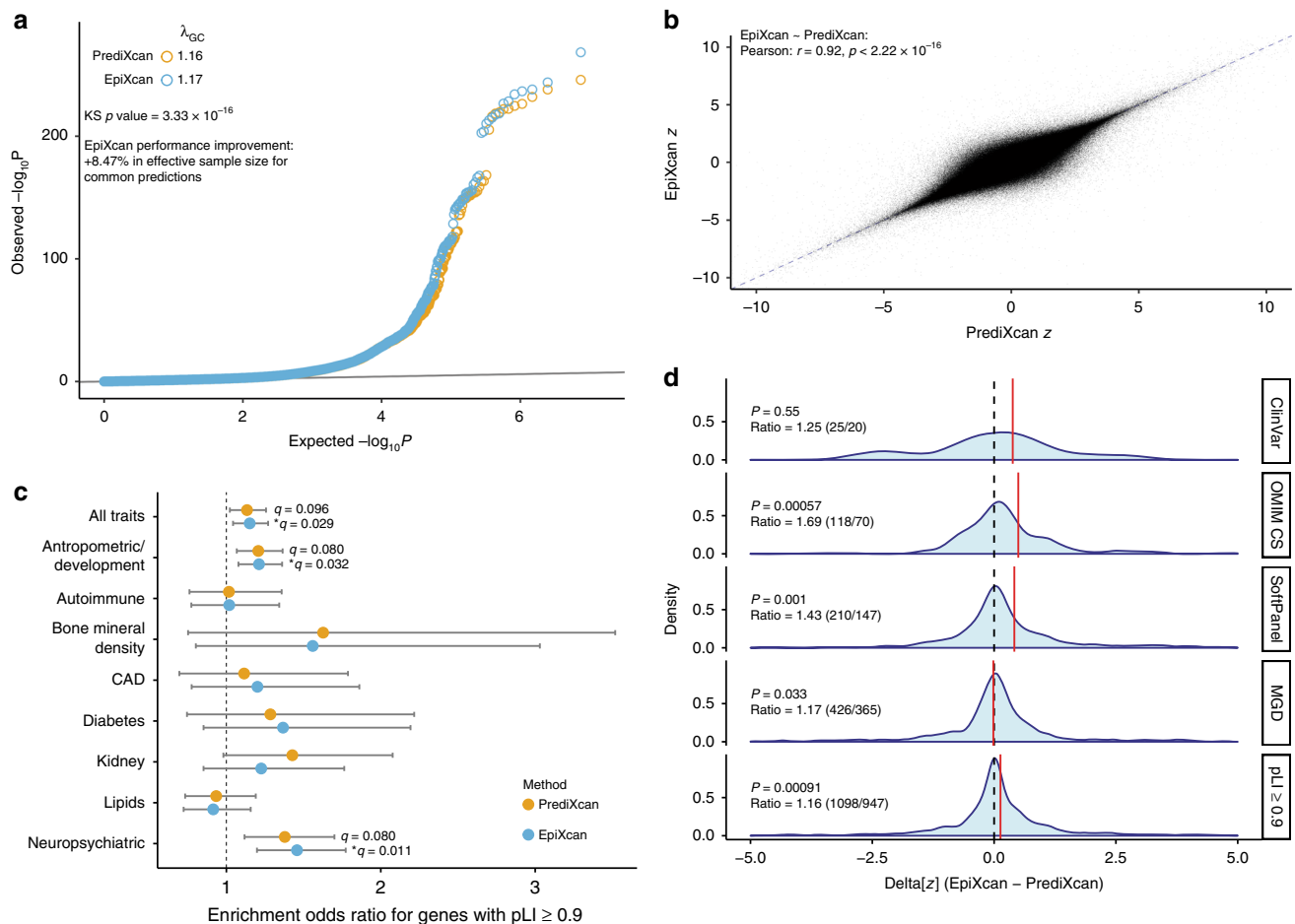
**EpiXcan informs better gene-trait associations.** We apply EpiXcan and PrediXcan prediction models from 14 tissues (Supplementary Table 1) in 58 complex traits (Supplementary Data 3) and examine their performance based on four criteria: the

**Fig. 1** Comparison of prediction performance between EpiXcan and PrediXcan. EpiXcan and PrediXcan models are trained across multiple tissues that include: brain, aorta, mammary artery, subcutaneous fat, visceral fat, liver, skeletal muscle, and blood by leveraging 14 datasets from CMC, STARNET and GTEx. The difference in training performance between EpiXcan and PrediXcan models is compared using the adjusted cross validation  $R^2$  ( $R^2_{CV}$ ) metric. The 14 models are further assessed by estimating the predictive performance ( $R^2_{PP}$ ) in independent datasets; the training dataset is shown before the arrow and the test dataset after the arrow (G = GTEx and S = STARNET). For a given dataset, we compare the  $R^2_{CV}$  and  $R^2_{PP}$  by estimating the delta value of EpiXcan minus PrediXcan for each gene. Positive and negative delta values indicate genes with higher predictive performance in EpiXcan and PrediXcan, respectively. These genes are assigned as “EpiXcan” and “PrediXcan” and counts are shown as barplots. The number on the right indicates the ratio of “EpiXcan” assigned gene counts divided by “PrediXcan” counts. Across all datasets, the ratios are higher than 1 indicating that EpiXcan outperforms PrediXcan.  $p$  value from one-sample sign test indicates that the shift of the delta  $R^2_{CV}$  and  $R^2_{PP}$  values is greater than zero (All  $p$  values  $< 9 \times 10^{-16}$ )

number of GTAs that are: (1) significant after multiple testing correction, (2) positioned outside the GWAS loci (3) unique (i.e., genes identified only by one method), and (4) enriched for clinically relevant genes.

EpiXcan has more power to detect GTAs than PrediXcan (Kolmogorov-Smirnov  $p$  value is  $3.3 \times 10^{-16}$ , Fig. 2a) and achieves an 8.47% average increase in  $\chi^2$  statistic for GTAs where  $\chi^2 \geq 1$  by both methods ( $n = 1,077,801$ , Mann-Whitney  $U$  test  $p$  value  $< 2.2 \times 10^{-16}$ ). Since statistical power is linearly related to the  $\chi^2$  statistic, this corresponds to EpiXcan producing an 8.47% increase in effective sample size. Consequently, we observe a 9.6% increase ( $n = 1202$ ) in the significant GTAs at 0.01 false discovery rate (FDR)<sup>23</sup> using EpiXcan ( $n = 13,724$ ) compared to PrediXcan ( $n = 12,522$ ). One advantage of PrediXcan/EpiXcan methods is that they identify genes within loci that did not reach genome-wide significance ( $p$  value  $< 5 \times 10^{-8}$ ) in GWASs. We detect an 18.3% increase (one sample sign test  $p$  value =  $3.6 \times 10^{-7}$ ) in the GTAs using EpiXcan (mean of 25.4) compared to PrediXcan (mean of 21.5) (Supplementary Fig. 9). The largest difference is observed for height (EpiXcan = 168, PrediXcan = 134), followed by schizophrenia (EpiXcan = 119, PrediXcan = 104) (Supplementary Fig. 10a). The overwhelming majority of the most significant SNPs for these genes identified by both methods have GWAS  $p$  values within the interval ( $5 \times 10^{-8}$ ,  $10^{-3}$ ) and are more likely to be within the interval ( $5 \times 10^{-8}$ ,  $10^{-5}$ ) when adjusted for the  $p$  value distribution of LD-independent genomic regions for each GWAS (Supplementary Fig. 10b). They thus represent ‘borderline’ GWAS results that one might expect to be identified as larger studies become available. Similarly, EpiXcan detects 9.95% more GTAs (one-sample sign test  $p$  value = 0.015) that are not identified by MAGMA gene analysis<sup>24</sup> compared to PrediXcan (Supplementary Fig. 9).

For any given tissue and trait, we find high correlation of GTA  $z$  scores between EpiXcan and PrediXcan (Pearson’s correlation  $r = 0.92$ ) (Fig. 2b), although unique associations are observed for each method. We identify 79.9% ( $n = 327$ ) more unique genes in EpiXcan ( $n = 788$ ) than PrediXcan ( $n = 461$ ) (Supplementary Fig. 11), due to either a lack of a prediction model for a specific gene and/or tissue or insufficient statistical power using PrediXcan models. For example, using the waist-adjusted BMI trait and prediction models from STARNET subcutaneous adipose tissue, overall, we observe high correlation between EpiXcan and PrediXcan genes (Pearson’s  $r = 0.83$ ) (Supplementary Fig. 12). Interestingly, EpiXcan identifies 7 genes (*PPP2R5A*,



**Fig. 2** Comparison of gene-trait associations between EpiXcan and PrediXcan. **a** EpiXcan and PrediXcan pairwise Wilcoxon test  $p$  value distributions for all gene-trait associations. Quantile-quantile (QQ) plot of the  $p$  values for all gene-trait associations show a significant, albeit modest, shift to the left. The genomic inflation factor ( $\lambda$ ) is slightly higher for EpiXcan than PrediXcan (1.17 and 1.16). The two distributions are significantly different (Kolmogorov-Smirnov test  $p$  value is  $3.3 \times 10^{-16}$ ) and EpiXcan achieves an 8.47% improvement in effective sample size for common predictions based on  $\chi^2$  test percentage improvement. **b** EpiXcan and PrediXcan have a high correlation of gene-trait association  $z$  scores. Scatter plot of EpiXcan and PrediXcan  $Z$  values, Pearson  $r = 0.92$  and Spearman  $\rho = 0.91$ ,  $p$  value  $< 2.22 \times 10^{-16}$  for both. Only  $z$  values between  $-10$  and  $10$  are plotted. The dotted blue line corresponds to  $y = x$ . **c** Gene set enrichment analysis (GSEA) for extremely loss-of-function intolerant ( $pLI \geq 0.9$ ) genes. Odds ratio with 95% CI are plotted for combined gene-trait associations from all traits and trait categories for enrichment in genes with  $pLI \geq 0.9$  (\* for  $q$  value  $< 0.05$ ). For all  $pLI$  decile bins enrichment refer to Supplementary Data 4. **d** EpiXcan has more power than PrediXcan to detect expression changes of trait-specific, clinically significant genes. These density plots depict the distribution of the  $\Delta[z]$  (EpiXcan - PrediXcan) values for all gene-trait associations that are significant from either EpiXcan or PrediXcan.  $P$  value is from one sample sign test. Ratio is the number of  $\Delta[z]$  measurements in favor of EpiXcan to that of PrediXcan. The red lines correspond to the mean of each distribution

*ALAS1*, *HOXC8*, *PIEZO1*, *SCD*, *PARP3*, and *EYA1*) that are not detected by PrediXcan, even if we test across all tissue-specific models. *SCD* (stearoyl-CoA desaturase) is of particular interest, as it encodes an enzyme that catalyzes a rate-limiting step in the synthesis of unsaturated fatty acids (mainly oleate and palmitoleate); knocking out the *SCD* ortholog in the mouse results in reduced body adiposity and resistance to diet-induced weight gain<sup>25</sup>. Accordingly, EpiXcan predicts that upregulated *SCD* gene expression is associated with increased waist-adjusted BMI.

To more broadly compare the unique GTAs identified by EpiXcan or PrediXcan, we wanted to see whether they exhibit similar colocalization properties. Several methods (e.g., HEIDI post-SMR<sup>9</sup>, COLOC<sup>7</sup>, eCAVIAR<sup>26</sup>) make use of local LD patterns in an attempt to distinguish: (1) pleiotropy-driven (causal variants affecting both phenotype and gene expression) and causality-driven (causal variants affecting phenotype via gene expression) GTAs from (2) linkage-driven GTAs (one causal variant affecting gene expression and a second causal variant

affecting the phenotype in LD) which can lead to misinterpretation of TWAS-derived GTAs. No method can provide perfect separation of pleiotropy and linkage but, as shown for S-PrediXcan<sup>27</sup>, HEIDI analysis is moderately to highly concordant with COLOC's classification, and PrediXcan performs favorably when compared to other methods. Thus, we utilize our HEIDI post-SMR analysis<sup>5</sup> to identify the proportion of genes with good colocalization properties uniquely identified by either study (Methods) and find no difference between EpiXcan and PrediXcan ( $\chi^2$  test  $p$  value = 0.14, Supplementary Note 1).

**EpiXcan uncovers more clinically relevant genes.** We perform a series of gene set enrichment analyses (GSEA) to determine how well EpiXcan can uncover clinically relevant genes and molecular pathways compared to PrediXcan. For this, we employ five categories of datasets: (1) ExAC gene  $pLI$  (probability of loss-of-function intolerance) dataset<sup>28</sup>, (2) ClinVar dataset—pathogenic

or likely pathogenic genes in the ClinVar database<sup>29</sup>, (3) OMIM CS dataset—genes in OMIM with phenotypes in the clinical synopsis (CS) section<sup>30</sup>, (4) SoftPanel dataset—custom gene panels for our traits created with SoftPanel<sup>31</sup> based on ICD-10 classification and keyword queries (underlying knowledge base is OMIM but gene panel creation is more integrative), and (5) MGD dataset—mouse orthologs of human genes associated with mouse strain-specific phenotypes<sup>32</sup>. GTAs from both PrediXcan and EpiXcan exhibit enrichment for genes that are associated with the traits in the above datasets (Supplementary Fig. 13).

Transcripts identified by EpiXcan ( $q$  value = 0.029), but not by PrediXcan ( $q$  value = 0.096), are enriched for genes that are extremely loss-of-function intolerant ( $pLI \geq 0.9$ ) (Fig. 2c). More specifically, we find significant enrichment of  $pLI$  genes with neuropsychiatric ( $q$  value = 0.012, known association<sup>33,34</sup>) and anthropometric/development ( $q$  value = 0.032) related traits (Supplementary Data 4). Unlike  $pLI$ , for all other gene sets (ClinVar, OMIM CS, SoftPanel, MGD), we define and test for enrichment only for that specific trait. For example, for autism, we generate a gene list from the significant autism-specific GTAs from all tissues for each method. We then perform GSEA for genes in the ClinVar database that are reported to be associated with autism. In so doing, we find that, overall, EpiXcan has more power than PrediXcan to identify clinically relevant genes (Fig. 2d), including those that are more likely to belong to more than one dataset ( $pLI$ , ClinVar, OMIM CS, SoftPanel, MGD) (Supplementary Fig. 14).

In conclusion, TWAS across 58 traits shows that, compared to PrediXcan, EpiXcan has more power to detect significant genes, including unique associations, which are indispensable for life and clinically significant. In the following section, we further explore the EpiXcan-derived GTAs, in terms of: (1) per-tissue contribution of significant genes, (2) gene-set enrichment analysis, (3) computational drug repurposing analysis, and (4) genes shared within, and across, different disease categories.

**Tissues differentially contribute GTAs.** In this study, we employ 3 different training cohorts to generate 14 predictive models for 8 tissue homogenate types and use the predictive models to impute tissue-specific transcriptomes across 58 GWASs. By pooling together imputed transcriptomes for each tissue from all traits, we first determine the robustness of our method by examining the  $z$  score correlation for similar tissues within and across cohorts. As expected, predictions are highly correlated when EpiXcan models are trained in (1) different cohorts (GTEx and STARNET) predicting the same tissue (Spearman's  $\rho$ : 0.89–0.93) and (2) the same cohort predicting similar tissues (Spearman's  $\rho$ : 0.89 when comparing aorta with mammary artery, and 0.92–0.95 when comparing visceral with subcutaneous adipose tissues) (Fig. 3a). In contrast, unrelated tissues, such as blood and brain, exhibit only moderate correlation (Spearman's  $\rho$  0.38–0.42).

Tissue-specificity of GTAs can be used to prioritize biologically relevant tissues for each disease. In contrast to a null model of no trait-associated tissue specificity, significant EpiXcan GTAs are statistically enriched for particular tissues (Pearson's  $\chi^2$  test  $p$  value =  $2.7 \times 10^{-8}$ , Fig. 3b). For example, we find a higher number of contributions than expected from brain tissue in schizophrenia and from blood in inflammatory bowel diseases, which is concordant with previous SMR analysis<sup>5</sup>. This occurs despite the observation that largely similar numbers of GTAs are obtained irrespective of tissue source (Supplementary Fig. 15).

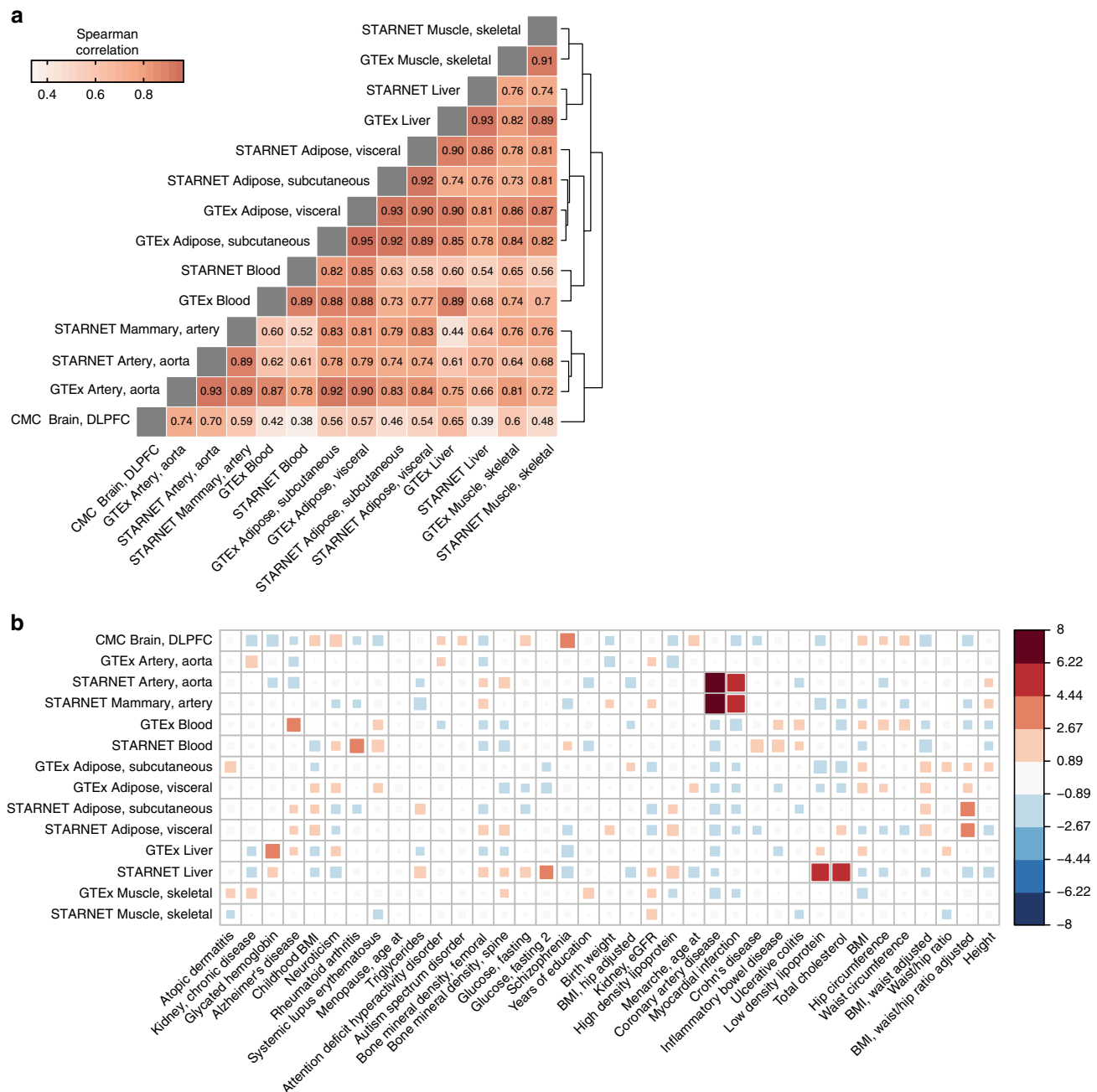
For 48 of the traits, more than 50% of the associated genes are found in only one tissue (Supplementary Fig. 16) and a large proportion ( $32.98 \pm 17.36\%$ ; mean  $\pm$  SD) of these unique GTAs come from the highest contributing tissue type (Supplementary

Fig. 17). A few examples of top tissue type contributors for unique GTAs are as follows; schizophrenia: brain tissue (30.34%, CMC), myocardial infarction and coronary artery disease: arterial tissue (33.33% and 31.88% respectively, STARNET aorta and mammary artery and GTEx aorta), systemic lupus erythematosus: blood (38.89%, STARNET and GTEx blood), most lipid traits: liver (24.06–26.43%, STARNET and GTEx liver). Besides tissue relevance, cohort size and tissue dissimilarity explain 52% of the variation in the number of unique GTAs contributed by different tissues (Supplementary Fig. 18, multiple linear regression model,  $p$  value = 0.007). This indicates that additional GTAs will be uncovered with increased sample size of gene expression datasets in disease-relevant tissues.

**Biological relevance of gene-trait associations.** High confidence GTAs (observed  $p$  value vs. expected  $p$  value) for a given trait get progressively enriched for genes that are more directly implicated in the pathogenesis of diseases with trait-relevant phenotypes. In databases (ClinVar and OMIM) where mostly large effect mutations are cataloged, we observe this progressive increase in enrichment (as indicated by  $\lambda$ , Supplementary Fig. 19), starting with genes that are associated with trait-relevant clinical signs, even if those clinical signs are not the primary symptoms of the disorder (OMIM CS:  $\lambda = 1.29$ ,  $p$  value =  $6.17 \times 10^{-14}$ ). Next, we see enrichment for genes that are driving similar disorders based on ICD10 classification grouping, or phenotype descriptive terms (SoftPanel:  $\lambda = 1.36$ ,  $p$  value <  $2.22 \times 10^{-16}$ ). Finally, we observe the highest enrichment for those genes that are directly driving trait-relevant phenotypes (ClinVar:  $\lambda = 1.83$ ,  $p$  value =  $7.07 \times 10^{-14}$ ). We also observe enrichment ( $\lambda = 1.21$ ,  $p$  value =  $1.69 \times 10^{-13}$ ) for mouse orthologs that produce mouse phenotypes in the same phenotypic category as the relevant human trait.

We perform gene-set enrichment analysis for traits with more than 10 significant GTAs (43 out of 58 traits) to determine if the associated genes can be mapped to biological processes (Supplementary Data 5). After FDR adjustment, 74 highly enriched pathways are obtained with  $p$  values <  $1.70 \times 10^{-5}$  (corresponds to  $q < 0.05$ ). Significantly associated genes are enriched for biological processes relevant to trait pathophysiology. For instance, the enriched pathways for elevated total cholesterol and triglycerides are involved in sterol and lipid homeostasis, as well as lipoprotein digestion, mobilization, and transport. Similarly, for atopic dermatitis the significantly enriched pathway modulates the rate or extent of water loss from an organism via the skin. In addition, genes associated with mineral density of the femoral bone demonstrate a high enrichment for a pathway that positively regulates cartilage development.

**GRex-based computational drug repurposing.** Computational drug repurposing (CDR) offers a systematic approach for relating disease and drug-induced states towards the goal of identifying indications for existing therapeutics<sup>35</sup>. We perform a computational screen against a library of 1309 drug-induced transcriptional profiles<sup>36</sup> to identify small molecules capable of perturbing the expression of our identified trait-associated genes (Fig. 4a). For each trait/compound pair, we calculate a signed connectivity score<sup>36</sup>, which summarizes the transcriptional relationship between each trait and drug signature, thus identifying drugs that might be predicted to “normalize” the gene-trait signature, as well as those expected to induce a “disease-like” state (Fig. 4b–d, Supplementary Data 6). Figure 4e provides example compounds predicted to regulate the expression of genes associated with the “Hip circumference adjusted BMI” trait. This list includes drugs under investigation for treatment of obesity, including ursolic



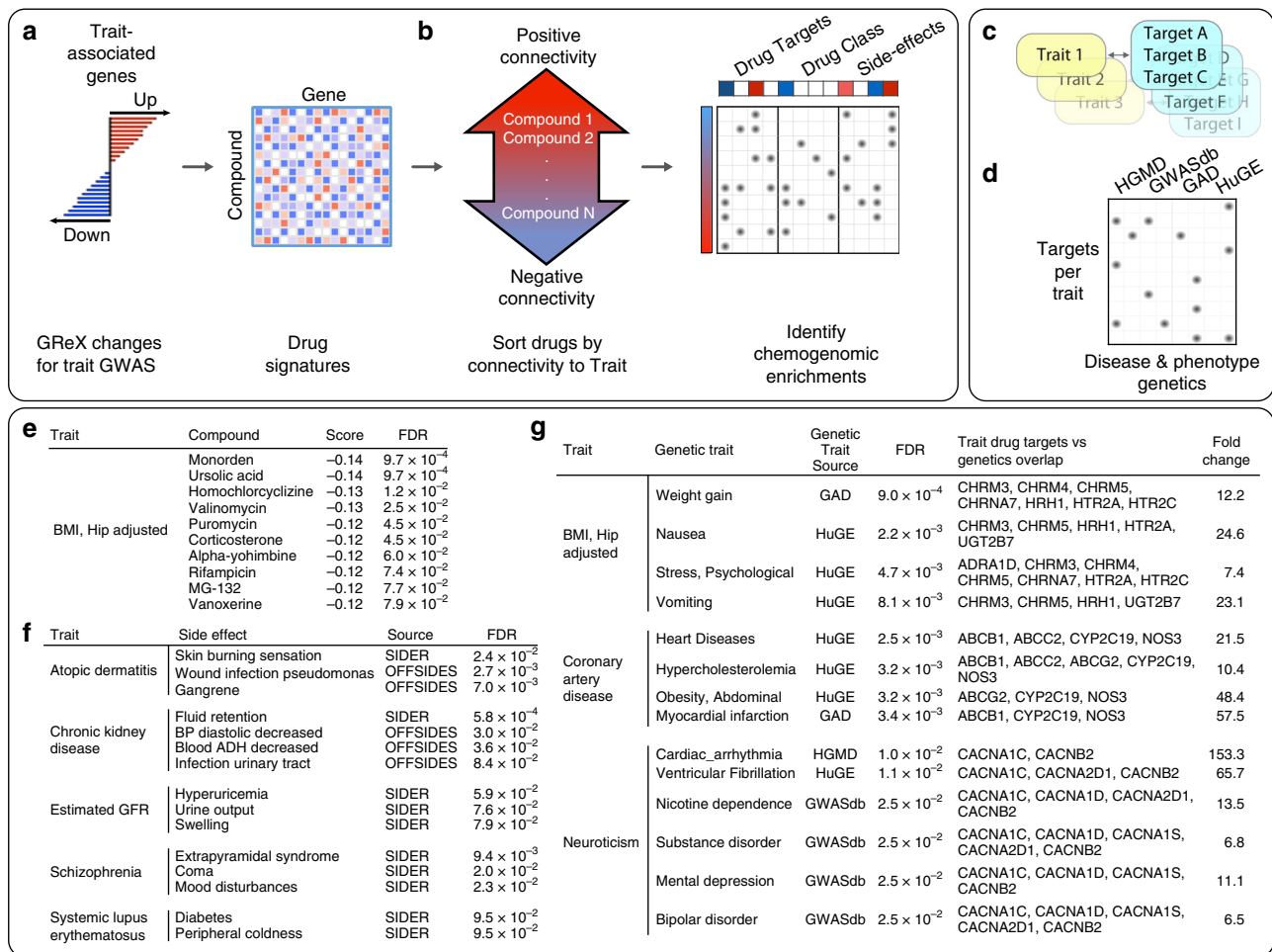
**Fig. 3** Contribution of GWAS and tissues to gene-trait associations. **a** Correlation of genetically regulated expression imputed for different tissues (pooled GTAs for all traits). Correlation is calculated for significant imputed expression changes with the Spearman method. Dendrogram on the right edge is shown from Ward hierarchical clustering. **b** Enrichment of tissue-specificity of significant EpiXcan GTAs compared to a null model, where each tissue contributes equally (Pearson's  $\chi^2$  test  $p$  value =  $2.7 \times 10^{-8}$ ). Statistically, the enrichment is the Pearson standardized residual for each tissue-trait pair from the  $\chi^2$  test. Box size and color indicate enrichment (red) or depletion (blue) for each tissue-trait pair. Only traits with expected frequency of more than 1 significant gene-trait association for each tissue model are evaluated as per Pearson's  $\chi^2$  test requirements. Tissues and traits are ordered based on Ward hierarchical clustering. Right-hand side panel indicates tissue-specificity enrichment score

acid, which is reported to increase skeletal muscle and brown fat while reducing diet-induced obesity<sup>37</sup>.

To explore the higher-level biological context of trait/compound associations, we perform a chemogenomic enrichment analysis to determine whether drugs that regulate particular sets of trait-associated genes might share pharmacological features, such as drug targets, drug classes, side-effects and drug indications (Fig. 4b, Supplementary Data 6). We find multiple significant (FDR < 0.1) chemogenomic trends, including enrichment with phenotypically related side-effects (Fig. 4f), supporting

the potential for these compounds to perturb trait-related molecular networks.

We hypothesized that, in general, trait-associated drug targets would connect to risk-associated genes for phenotypically related diseases<sup>38</sup>. To evaluate this, we identify referenced<sup>39</sup> and predicted<sup>40</sup> drug targets that are enriched (FDR < 0.1) among compounds that modulate the signature of each trait. We identify  $\geq 1$  drug target enrichment, for 53 of the traits considered, and  $\geq 3$  drug targets for 40 traits (Supplementary Data 6). We then perform a further gene set analysis on the targets associated with



**Fig. 4** Leveraging gene-trait associations for computational drug repurposing. **a** Trait-associated genes are used to sort a library of drug induced gene expression signatures according to their connectivity with the trait. GReX: genetically regulated expression. **b** A secondary enrichment analysis on this drug list identifies pharmacological features that are over-represented at the extreme ends of the sorted list, thus presenting a chemogenomic view of the trait. **c** Drug targets linked with each trait ( $FDR < 0.1$ ) are then **(d)** compared with risk loci genes for a range of diseases or phenotypes ( $FDR < 0.1$ ). **e** Top 10 compounds predicted to normalize the expression of “Hip adjusted BMI” associated genes. **f** Subset of side-effect enrichments for phenotypically related traits. **g** Subset of traits with associated drug targets that are enriched for risk associated genes sets with phenotypically related traits

each trait, focusing on disease risk genetic resources that might implicate phenotypes that could then be related to the traits considered within this study. We identify several significant overlaps ( $FDR < 0.1$ ) between trait-associated targets and phenotypically related disease risk gene sets (Fig. 4g, Supplementary Data 6). For example, drug targets enriched among compounds that perturb genes associated with “Hip circumference adjusted BMI” are enriched for risk genes for weight gain, nausea, and psychological stress, and drug targets enriched among compounds that perturb “Coronary Artery Disease” associated genes are enriched for risk genes for heart disease, hypercholesterolemia, abdominal obesity, and myocardial infarction.

Towards an objective assessment of the CDR pipeline performance, we compare the CDR predictions with known physician-curated indications for our traits (Supplementary Note 1) that fall into four groups of increasingly perceived efficacy: (1) non-indication: a drug that neither therapeutically changes the underlying or downstream biology nor treats a significant symptom of the disease, (2) symptomatic: a drug that treats a significant symptom of the disease, (3) FDA-approved for the trait, and (4) disease modifying: a drug that therapeutically changes the underlying or downstream biology of the disease. Compounds that are predicted to normalize the gene-trait

signature demonstrate progressive enrichment for higher indication levels, whereas compounds that are expected to induce a “disease-like” state show a progressive depletion (Supplementary Fig. 20a). When only considering known disease modifying and non-indication compounds for our traits, compounds that are predicted to normalize the gene-trait signature are more likely to be disease modifying (odds ratio 11.37, Barnard’s unconditional test  $p$  value = 0.006, considering only our CDR predictions with  $p$  value  $< 0.3$ , Supplementary Fig. 20b). In addition, the chemogenomic enrichment for drug indications is also able to identify several cases where compounds predicted to normalize the gene-trait signature are enriched for compounds indicated to treat the trait’s comorbidities, e.g. (1) compounds that would reverse the “current versus former smoking” trait are enriched for compounds indicated for congestive heart failure and increased triglycerides and (2) compounds that would reverse childhood obesity are enriched for compounds indicated for coronary artery disease<sup>41</sup>, respectively (considering only chemogenomic enrichments with  $FDR < 0.25$ , Supplementary Data 6).

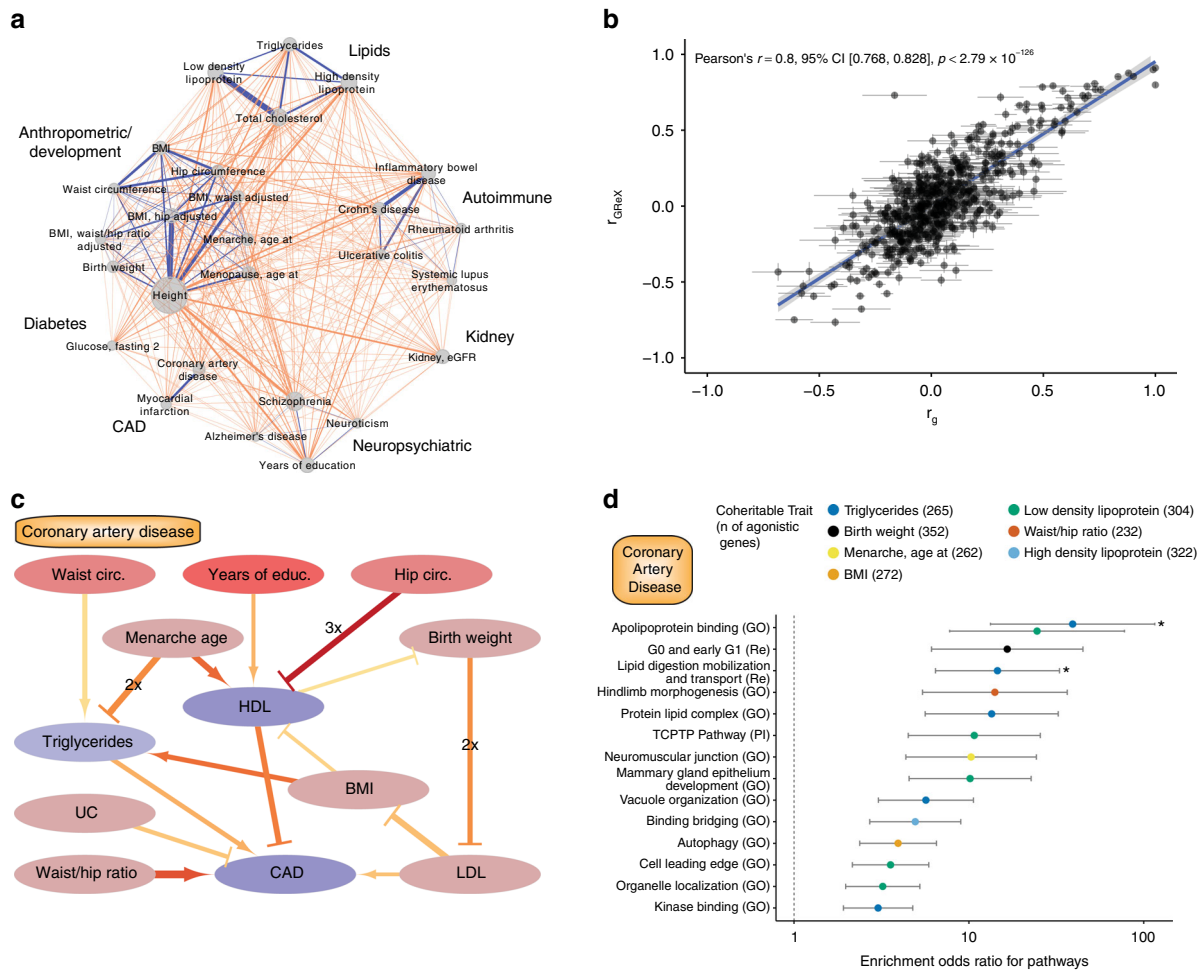
Taken together, these combined analyses illustrate the potential for the approach described in this study to inform drug discovery and drug development efforts. The identification of side effect, drug target and drug indication enrichments linked to known or

plausible trait biology supports the veracity of the repurposing predictions deriving from the accurate prediction of known indications, and, more broadly, the power of integrative genomics approaches to identify molecular networks that underpin disease.

**Trait-trait correlations and gene-trait associations.** To further understand trait relatedness, we construct a network based on pairwise trait comparison of genetically regulated expression (including traits with more than 50 significant associations). By using a broad categorization of traits (Supplementary Data 3), we identify 245 pairs of shared gene associations across trait categories and 66 pairs within trait categories (Fig. 5a, Supplementary Data 7). Higher numbers of genes are shared between traits that belong to the same trait category than those that do not; the

highest number of genes is shared between low density lipoprotein and total cholesterol in the lipids category. Previous studies have shown significant genetic correlation among common traits<sup>42,43</sup>. Pairwise trait GReX correlation shows a positive association with genetic co-heritability<sup>42,43</sup> (Pearson's  $r = 0.8$ ,  $p$  value  $< 2.79 \times 10^{-126}$ ) (Fig. 5b), extending the genetic similarity among traits to specific genes.

We then apply bi-directional regression analyses<sup>44</sup> on the GReX of different traits across all tissues to infer causal relationships among pairs of traits with significant genetic and GReX correlation (Fig. 5c for CAD and Supplementary Fig. 21 for all the traits in our study). We find evidence that CAD is a complex trait whose predicted gene expression changes can be partly, but directly, explained by predicted expression changes found in individuals with elevated triglycerides, elevated LDL,



**Fig. 5** Trait-trait correlations and gene-trait associations. **a** Network indicating shared genes within/across trait categories. Only traits that have more than 50 associated genes are showcased. Edge width denotes number of shared genes for each trait pair. The node size indicates number of gene-trait associations for a given trait. Blue edges denote within-category trait associations and orange edges denote across-category trait associations. The analysis is based on significantly associated genes with  $FDR \leq 0.5\%$ . **b** Scatter plot of genetic correlation ( $r_g$ ) and genetically regulated gene expression ( $r_{GReX}$ ) for each pairwise trait combination. Standard error is shown with gray lines,  $r_g$  and  $r_{GReX}$  are highly correlated (Pearson's  $r = 0.8$ ,  $p$  value  $< 2.79 \times 10^{-126}$ ). **c** Causal trait network of CAD. CAD and up to two traits upstream are plotted in this network graph to demonstrate causal (arrows) and protective (bar-headed lines) relationships as estimated by bi-directional regression analysis. The trait nodes are colored based on the parent causal trait network of all the traits of the study (Supplementary Fig. 21); nodes that have more children than parent nodes are a darker shade of red and blue, respectively. In edges, width denotes absolute beta, redder color denotes lower  $p$  value, and the  $2\times$  or  $3\times$  labels denote that the relationship is identified in 2 or 3 tissues, respectively. The analysis is based on genes with  $FDR \leq 1\%$ , and only the relationships with  $p$  value  $\leq 0.05$  are shown. **d** Graph depicting the odds ratio of pathway enrichment for CAD agonistic genes shared with traits involved in the causal network. Briefly, for causal traits, a list of genes (with unadjusted  $p$  value  $\leq 0.05$ ) that are predicted to change to the same direction (or the opposite direction for protective traits) is used for GSEA for common pathways. In this graph only the top 15 (based on  $q$  value) results are shown and are ranked based on odds ratio; an asterisk (\*) indicates results that have  $q$  value  $\leq 0.05$ . Error bars represent 95% CI for each enrichment



and increased waist/hip ratio. On the other hand, predicted expression changes in individuals with increased HDL, or those suffering from ulcerative colitis (UC), are expected to normalize expression changes in individuals with CAD. By expanding the causal network to include more upstream traits, we can see that another 6 traits (waist and hip circumference, years of education, age at menarche, birth weight, and BMI), which are correlated, or anti-correlated, with CAD may cause, or protect, from the predicted expression changes through effects on intermediate traits (Fig. 5c). For example, waist circumference acts via a causal relationship with triglycerides; other traits follow multiple pathways such as age at menarche, which opposes predicted transcriptomic changes of the increased triglycerides group while promoting imputed transcriptomic changes for individuals with high HDL. We then leverage these causal networks to dissect the pathogenesis of CAD by identifying the molecular pathways shared among all the involved trait pairs. For each trait that can cause or protect from CAD, we identify the agonistic genes—genes whose predicted expression is changing towards the same or opposite direction for causal (e.g., triglycerides) and protective (e.g., HDL) traits, respectively. Gene set enrichment analysis of agonistic genes for biological pathways point towards biologically relevant processes for CAD (Fig. 5d). For example, a subset of CAD genes ( $n = 256$  out of 2806 genes with  $p$  value  $\leq 0.05$ ) is shared with triglycerides and affects biological processes related to apolipoprotein binding and lipid digestion, mobilization, and transport.

Taken together, the pairwise GREX trait correlations illustrate the potential to identify genes that are shared among genetically correlated traits. Agonistic versus antagonistic pleiotropy among two traits can be differentiated by leveraging the directionality of gene expression association in each trait. For traits, such as CAD, this analysis can be applied to dissect the complex phenotype, to identify genes and pathways that are shared with another trait, and potentially identify and develop therapeutic strategies to reverse those perturbations.

## Discussion

The maps of gene expression and regulatory annotations, generated by projects such as REMC<sup>15</sup>, CommonMind<sup>18</sup>, GTEx<sup>20</sup>, and STARNET<sup>19</sup> hold the potential to further our understanding of non-coding risk genetic variation. Here we describe EpiXcan which, compared to PrediXcan, integrates biologically relevant data in a single framework to improve predictive performance of transcriptome imputation. EpiXcan is also better powered to identify clinically significant results such as enrichment for loss-of-function intolerant genes in neuropsychiatric traits<sup>33,34</sup> and can detect more robust gene expression changes in genes associated with severe forms of the trait. Despite improvements in transcriptomic imputation predictive performance, it is important to note that all current imputation methods overall explain a small proportion of gene expression variation. We apply EpiXcan prediction models from 14 tissues in 58 common and complex traits and examine properties of those associations.

First, gene associations are predominantly identified in pathophysiologically relevant tissues and most associations are only identified in one tissue. Considering that the average correlation between genetically regulated gene expression of unrelated tissues such as blood and brain across 58 traits is 0.38–0.42 (Spearman's  $\rho$ ), we highlight the need for trait-relevant tissue datasets for such studies to be more effective.

Second, among genes associated with the traits in this study, we observe significant enrichment for biological pathways involved in trait pathophysiology. Moreover, gene-trait associations are significantly enriched for: (1) pathogenic (or likely pathogenic)

genes for the given trait (clinVar), (2) genes associated with trait-relevant phenotypes (SoftPanel), (3) genes that have been associated with clinical signs relevant to the trait (OMIM CS), and (4) orthologous mouse genes with phenotypes that belong to the same phenotypic category as the given trait. This suggests that common variants partly act via smaller effect size perturbations in genes that lead to more severe forms of the phenotype when subject to larger effect size disruptions, as recently similarly suggested<sup>27</sup>.

Third, by leveraging trait-specific transcriptomic changes, we identify compounds that can reverse trait-specific changes, pointing to potential drug repurposing candidates. We assess the performance of our pipeline by comparing the predictions with known drug indications and find drugs that are predicted to normalize trait-specific changes are more likely than expected to be disease modifying for the trait. Towards further validation of our approach, chemogenomic enrichment analysis reveals trait-specific, phenotypically related, side effects, drug indications and drug target enrichment for risk-associated genes of phenotypically related traits. One recent study<sup>14</sup> applied a similar approach, which was somehow limited in scope (brain tissue –10 regions—transcriptomic imputation with S-PrediXcan for psychiatric traits). It is hard to directly compare the results of the two studies since our approaches differ on many levels: we use EpiXcan, train models on more diverse tissues, employ a different drug repurposing pipeline that includes a set of chemogenomic enrichment analyses and use a more agnostic approach to validate our predictions. Despite the above limitations, both approaches share a lot of similarities, including their use of the same compound signature reference panel source and similar principles for ranking the compound predictions. Among common traits between the two studies, we do not find a particularly high concordance among our predictions (OR range: 0.91–1.31) but we do find that our predictions (1) are more likely to agree for schizophrenia (OR = 1.31,  $p$  value = 0.026) and (2) have higher concordance the higher the brain tissue enrichment score for the trait ( $p$  value  $< 2.2 \times 10^{-16}$ ) compared to other tissues (Supplementary Note 1, Supplementary Fig. 22). For schizophrenia—where our results are most concordant—their studies identify no candidate compounds after adjustment for multiple testing. In contrast, EpiXcan identifies one statistically significant result (phenformin, Supplementary Data 6) that is a very potent anti-diabetic agent (no longer FDA-approved due to safety concerns) which is not surprising given that glucose homeostasis is altered from illness onset in schizophrenia<sup>45</sup>. Within the top 10 results for schizophrenia, we also identify a potent antipsychotic (prochlorperazine), a voltage-gated sodium channel<sup>46</sup> inhibitor (pramocaine) and guanfacine, which was trialed for cognitive impairment in schizophrenia and found to be worthy of further investigation<sup>47</sup>. Although our approach performs remarkably well given the modest percentage of gene expression variation that we are able to explain (despite performance improvements from EpiXcan), there are several limitations that are hampering its translational potential. Towards further improving common variant derived GREX-based CDR pipelines, generating cell-type specific predictive models with spatial and temporal annotation, as well as expanding and improving the repertoire of compound signatures in more relevant cells and in vivo models, holds much promise in establishing a powerful genetically driven drug discovery and repurposing pipeline.

Finally, we use bi-directional regression analysis<sup>44</sup> to construct putative causal trait networks. Causal trait networks built on top of EpiXcan are sufficiently powered to provide valuable insight into the development of complex traits such as CAD. For example, we find that high BMI can influence CAD by two distinct pathways; (1) by positively influencing triglycerides (TG)

which would positively influence CAD, and, conversely, (2) by negatively influencing HDL which would negatively influence CAD. The independent effect of BMI on TG and HDL has been shown in a population with a broad spectrum of BMI values<sup>48</sup> which—as in our study—found no effect of BMI on LDL levels. Downstream, there is genetic evidence to suggest a causal influence of TG on CAD<sup>49</sup>. In addition, a negative correlation of HDL with CAD has been established in observational epidemiology, although a link between genetic loci causal for high levels of HDL and protective for CAD is, at present, elusive<sup>50</sup>. The construction of these causal trait networks allows us to identify genes that exhibit agonistic pleiotropy participating in shared pathways. Such information could potentially be used to develop distinct therapeutic strategies based on individual comorbidities.

Overall, the described method utilizes epigenomic information to further improve prediction of transcriptomes and it provides a framework for TWASs, improved interrogation of trait-associated biological pathway involvement, and a platform for drug repurposing and treatment development.

To facilitate interpretation, we provide the EpiXcan pipeline, trained models and resulting data tables as an online resource.

## Methods

**Genotype and expression data.** Genotype datasets (CMC, GTEx and STARNET) are uniformly processed for quality control (QC) steps before imputation. We restrict our analysis to samples with European ancestry (Supplementary Methods). Genotypes are imputed using the University of Michigan server<sup>51</sup> with the Haplotype Reference Consortium (HRC) reference panel<sup>52</sup>. RNAseq gene level counts are adjusted for known and hidden confounds, followed by quantile normalization. For CMC gene expression, we use the gene level counts generated from DLPPC RNAseq data<sup>18</sup> (<http://commonmind.org/>). For GTEx<sup>53</sup>, we use publicly available, quality-controlled, gene expression datasets from the GTEx consortium (<http://www.gtexportal.org/>). RNAseq data for STARNET were obtained in the form of residualized gene counts from a previously published study<sup>19</sup> [[https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs001203.v1.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001203.v1.p1)]. Additional information for CMC, STARNET and GTEx tissues (for both predictors and observed datasets) including sample sizes is shown in Supplementary Table 1. To compare the prediction accuracy of the CMC-trained predictors, we utilize expression data from the HBCC ( $n = 280$  samples<sup>18</sup>), as well as 13 brain areas from GTEx<sup>53</sup> (Supplementary Table 1). The GTEx data are publicly available de-identified data, whereas ethical approvals of STARNET and CMC data are detailed in the original papers.

**SNP priors and rescaling to WENet penalty factors.** To leverage epigenomic information, we incorporate rescaled SNP priors as penalty factors into a weighted elastic net model. First, we compute eQTLs using MatrixEQTL<sup>54</sup>. Then, epigenomic annotations from REMC<sup>15</sup> are integrated to obtain SNP priors using qtBHM<sup>17</sup> (top panel in Supplementary Fig. 1; Supplementary Methods; Supplementary Data 8). Lastly, the SNP priors are rescaled to penalty factors used in WENet by a data-driven rescaling equation. The optimal rescaling equation is approximated by the best performing quadratic Bézier function, providing both the curve of the rescaling function and the minimum value of the penalty factors. Briefly, to determine the best performing rescaling equation, we simulate genotypes ( $n = 500$  samples) using HAPGEN2<sup>55</sup> and haplotypes from the 1000 Genomes Project<sup>56</sup>. For each gene under consideration, we utilize a shifting window policy to generate quadratic Bézier rescaling equations. In each separate window, we define a minimal penalty factor (Supplementary Fig. 23) and, within that window, evaluate possible intermediate Bézier curve control point locations to test for a wide range of curves for our rescaling equation (Supplementary Fig. 24). The equation that exhibits the highest improvement of  $R^2_{CV}$  when compared to not assigning penalty factors to the SNPs (as in PrediXcan) is selected. The process to evaluate and select the optimal rescaling equation is described in greater detail in Supplementary Methods.

**Simulation analysis and predictive performance comparisons.** Five hundred samples are simulated to verify the model performance. For specific gene, suppose  $X$  is the matrix containing genotypes of all *cis*-SNPs included in the gene. For the  $i$ -th SNP, we choose an effect estimate  $\beta_i$ , so we have vector of estimated effects  $\beta$  for all the SNPs of the gene. Gene expression values are simulated by

$$y = X \times \beta + \text{level} * \epsilon \quad (1)$$

Here ‘ $\times$ ’ denotes matrix-vector product and  $\epsilon$  is normally distributed noise with given standard deviation ( $SD = 0.3$ ). We select ten levels (Level from 0.1 to 1) of noise to simulate expression values for given genes. The CMC eQTL beta values are

used as the effects in the simulation. We use 1000 genes with the highest significance from CMC eQTL studies to perform the simulations. For each gene, we simulate 50 times and take the mean value to evaluate the closeness between simulated and real-world gene expressions.

**Comparison with BSLMM and DPR methods.** We perform a comparison, more limited in scope than in Fig. 1, of EpiXcan with PrediXcan, BSLMM and DPR. We use the CMC dataset for training and cross-validation (CV) and HBCC as an independent test dataset to calculate the gene expression imputation  $R^2$ , as well as the per gene computation duration required by each method. For estimating the  $R^2_{CV}$ , we utilize four folds of the CMC samples for training and then the remaining one fold to test the prediction performance. Similar approaches for predictive performance comparisons were employed in previous studies<sup>22</sup>. To estimate the  $R^2_{pp}$  in independent datasets, we use all CMC samples for training and the HBCC dataset to test the predictive performance. DPR has the option to use two different fitting algorithms: (1) the mean-field variational Bayesian (VB) approximation and (2) the Monte Carlo Markov Chain (MCMC). We perform the above tests and measure the predictive performance and imputation speed for EpiXcan, BSLMM, DPR (VB), DPR (MCMC), as well as PrediXcan. Details about different package implementation parameters are given in Supplementary Methods.

**Large scale gene-trait association analysis.** We train predictors of gene expression by applying EpiXcan and PrediXcan to genotype and RNAseq datasets across 14 tissues (Supplementary Table 1). For each tissue, we keep genes with  $\text{pred.perf } q$  value of the correlation between cross-validated prediction and observed expression ( $\text{pred.perf}^2 \leq 0.01$ ). We identify gene-trait associations by jointly analyzing summary statistics from 58 complex traits (Supplementary Data 3) and gene expression predictors using S-PrediXcan<sup>27</sup>. SNPs in the broad major histocompatibility complex (MHC) region (chromosome 6: 25–35 Mb) are removed.  $p$  values are adjusted using the Benjamini-Hochberg method of controlling the false discovery rate at  $\leq 0.01$ . The gene-trait associations that remain after this filtering are considered significant. The analysis of the GWAS data pertains to de-identified summary-level data and requires no ethical approval.

Uniquely identified genes by EpiXcan (or PrediXcan) are genes that are identified in significant gene-trait associations with one method but not the other. For gene-trait associations found in multiple tissues, we categorize genes as upregulated (or downregulated) in the trait if there are more tissues in which the effects are towards the indicated direction. If there are equivalent numbers of tissues in which the gene is positively and negatively correlated with a given trait, we categorize the gene regulation as ambiguous. Transcriptomic imputation yields approximately the same number of genes predicted to be upregulated or downregulated ( $z$  scores) across each trait (Supplementary Fig. 25). To construct the shared gene network in Fig. 5a: (1) we filter genes so that those with  $\text{pred.perf } q$  values  $\leq 0.5\%$  and FDR-adjusted  $p$  values  $\leq 0.5\%$  are retained, (2) specifically for shared genes across traits of the same category, we only include genes with high effects (e.g.  $|z \text{ score}| \geq \text{mean}_i |z \text{ score}|$ ,  $i$  is number of genes) to limit network density.

For identification of novel genes outside of GWAS loci, we define index SNPs based on LD clumped regions using Plink software (v1.9)<sup>57</sup>. The following settings are used: (a) significance threshold for index SNPs is  $5 \times 10^{-8}$ , (b) significance threshold for clumped SNPs is  $5 \times 10^{-8}$ , (c) clumping window size is 250 Kb and (d) LD threshold for clumping is 0.1. The coordinates of the GWAS loci are defined as 1 Mb on either side of the index SNP in each clump. The genomic coordinates of the significant genes are then extracted from GENCODE (build GRCh37, release 19) and overlapped with the coordinates of GWAS loci. Properties of those genes that lie outside the overlaps are explored. To identify the background distribution of  $p$  values of LD clumped regions, we used Plink, as above, but with no significance thresholds and a clumping window size of 500 kb. In addition, MAGMA gene analysis<sup>24</sup> is performed for 55 GWAS phenotypes. Genes significantly associated with the phenotypes are identified after adjusting for multiple testing correction using Benjamini-Hochberg method. Significant genes identified using EpiXcan and PrediXcan are compared with the significant genes identified using MAGMA (FDR < 0.01) to indicate how many genes are inferred by EpiXcan or PrediXcan but not by MAGMA. The difference in the number of genes that lie outside the overlaps (when compared to GWAS or MAGMA) identified between the two methods is calculated by subtracting the number of genes identified by PrediXcan from the number of genes identified by EpiXcan. The statistical significance is tested with the null hypothesis such that the mean difference is zero using one sample sign test ( $H_0: \bar{X} = 0$ ).

To indicate enrichment or depletion of the trait in a given tissue we use the Pearson standardized residuals as tissue-specificity enrichment score (Standardized residual $_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-p_{i+})(1-p_{+j})}}$ , where  $i$  is row,  $j$  is column,  $n_{ij}$  are observed values,  $\hat{\mu}_{ij}$  are expected values,  $p_{i+}$  is the observed ratio of total row count for  $i$  divided by all observations and  $p_{+j}$  is the observed ratio of total column count for  $j$  divided by all observations as described in Agresti<sup>58</sup>. To see whether there is a deviation from the null hypothesis of statistical independence (e.g. in Fig. 3b, the tissue-trait combination does not affect the number of significant GTAs), we perform Pearson’s  $\chi^2$  test of independence. This method is applied for Fig. 3b and Supplementary Figs. 20a, 22b.

We identify significant GTAs from EpiXcan and PrediXcan as described above (predictive performance  $q$  value  $\leq 0.01$  and  $FDR \leq 0.01$ ) that are also identified in our SMR study ( $p_{\text{SMR}} \leq 0.05$ )<sup>5</sup>. We then classify them into GTAs with either good co-localization properties ( $p_{\text{HET}} \geq 0.05$ ) or not ( $p_{\text{HET}} < 0.05$ , rejecting the null hypothesis that there is a single causal variant affecting both gene expression and trait variation, Supplementary Note 1).

**Gene set enrichment analyses and phenotypic datasets.** To investigate whether the genes associated with a given trait exhibit enrichment for biological pathways, we use gene sets from MSigDB 5.1<sup>59</sup> and filter out non-protein coding genes, as well as genes that do not have eQTL. For the enrichment analysis we only consider traits with >10 genes identified in significant gene-trait associations; this condition is met for 43 traits in our study. Statistical significance is evaluated with one-sided Fisher's exact test and the adjusted  $p$  values are obtained by the Benjamini-Hochberg method. Similarly, for Fig. 2c, we perform gene set enrichment analysis for all decile bins of pLI from ExAC<sup>28</sup> (all results can be found in Supplementary Data 4). The phenotypic datasets: ClinVar, OMIM CS, SoftPanel, and MGD are prepared as described in Supplementary Methods and contain genes that are associated with one or multiple traits. The approximation of known gene-phenotype associations from these datasets allows us to (1) compare the power of EpiXcan vs. PrediXcan in identifying known gene-trait associations (as in Fig. 2d) and (2) evaluate the extent to which common risk variants confer trait risk by affecting gene expression levels of genes associated with monogenic forms of the trait or genes associated with similar-to-the-trait phenotypes in humans and mice.

**Computational drug repurposing.** Compound profiles are sourced from Connectivity map, and are based on gene expression microarray data collected from 6100 individual experiments<sup>36</sup>, each comparing compound-treated with vehicle-treated cell line based gene expression profiles. We download the ranked  $\log_2$  fold change matrix available for the 6100 individual experiments, and merge them into a single representative signature for the 1309 unique small molecule compounds (Supplementary Data 9) according to the prototype-ranked list method<sup>60</sup>.

We iterate over each trait considered in this study, retaining trait/gene associations with an  $FDR < 0.1$ , and converting HGNC gene symbols to NCBI entrez gene identifiers. If a gene is linked with a trait via an association that was detected in multiple tissues, the associations are summarized as the mean  $z$  score. There are 58 traits with a minimum of 5 positively, and negatively, associated genes and each of these query signatures (QS) are used for the subsequent drug repurposing.

For each trait, and each unique compound, we calculate a connectivity score (CS) using an approach described in Lamb et al.<sup>36</sup> The calculation of the CS proceeds as follows: a running sum enrichment score (ES) is calculated for the negative ( $ES_{\text{Neg}}$ ) and positive ( $ES_{\text{Pos}}$ ) components of the QS, separately, reflecting the distribution of the QS component within the ranked gene list of the compound under consideration. ES can assume a value between  $-1$  and  $+1$ , where a negative ES indicates that genes within a QS component are relatively downregulated by a compound, and a positive ES indicates that genes within a QS component are relatively upregulated. The two ES are then combined into a single CS:  $CS = \frac{ES_{\text{Pos}} - ES_{\text{Neg}}}{2}$ . The resulting CS thus assumes a range of  $[-1, +1]$  and aims to summarize the overall transcriptional relationship between a compound and a QS. We estimate statistical significance of a given CS by generating an empirical CS distribution for a given QS against 1000 permutations of compound signatures. Permuted compound signatures are generated by randomizing the ranked  $\log_2$  of gene expression fold change for a given compound, and used to derive two-tailed  $p$  values, which are adjusted by the Benjamini-Hochberg method of controlling the false discovery rate.

For each trait, connectivity scores are then used to sort the list of 1309 compounds and used as the basis for a chemogenomic enrichment analysis. For each compound in the drug signature library, we collect diverse chemogenomic annotations, such as drug target information, side effect, therapeutic class associations, and drug indications. Side-effect associations are downloaded from Offsides<sup>61</sup> and SIDER<sup>61</sup> and connected to compounds in Connectivity map via Stitch identifiers. Drug target associations include targets referenced in DrugBank<sup>39</sup>, and also an augmented set of associations, based on predictions generated using the Similarity Ensemble Approach<sup>40</sup>. Drug disease indications were derived from the Clue Drug Repurposing Hub (<https://clue.io/repurposing>) and used to annotate compounds within the scope of our analysis with available trait indications. This resulted in 139 distinct clinical indications with at least three associated compounds. For each of these features, we calculate a signed running sum enrichment score, which reflects whether that feature is over-represented at the extreme ends of the drug list that has been ordered according to trait. Statistical significance of enrichment scores is based on comparison to a large distribution of permuted null scores, generated by calculating scores from randomized chemogenomic sets that contain an equivalent number of compounds to the true set being evaluated.  $p$  values are adjusted using the Benjamini-Hochberg method of controlling the false discovery rate.

We compile disease and trait risk associations from multiple sources, including HGMD<sup>62</sup>, ClinVar<sup>29</sup>, dbGAP<sup>63</sup>, Genetic Associations Disease<sup>64</sup>, GWAS catalog<sup>65</sup>, GWASdb<sup>66</sup>, Human Phenotype Ontology<sup>67</sup>, HuGE<sup>68</sup>, and OMIM<sup>69</sup>. Many of these are accessed through Harmonizome<sup>70</sup>. We use a Fisher's exact test to compare each set of trait-associated drug targets (that contain at least 3 targets), with each disease

risk gene set. The analysis is performed against a background of 2802 genes, representing the unique set of human drug targets in the combined set of referenced and predicted targets associated with the 1309 compounds. Two-sided  $p$  values are adjusted using the Benjamini-Hochberg method of controlling the FDR.

**Trait co-heritability analysis.** To calculate the genetically regulated gene expression correlation ( $r_{\text{GREX}}$ ), as shown in Fig. 3a, we keep the significant imputed gene expression change ( $z$  score) values with  $q$  value  $\leq 0.01$  and perform pairwise tissue Spearman correlation analysis of the complete cases of  $z$  scores. To cluster the tissues together for plotting, we use hierarchical agglomerative clustering analysis with Ward's method.

For genetically regulated gene expression correlation ( $r_{\text{GREX}}$ ), pairwise genetic correlation ( $r_g$ ), as shown in Fig. 5b, among traits analyzed by GWAS is taken from previously published reports<sup>42,43</sup>. For trait comparisons that appear in both studies we use the more recent study<sup>43</sup>. We consider the genetic correlation between traits significant if  $q$  value  $\leq 0.05$ . To calculate  $r_{\text{GREX}}$ , we keep the imputed gene expression values with unadjusted  $p$  value  $\leq 0.05$  and perform pairwise trait Spearman's correlation analysis with Holm's adjustment for multiple comparisons. To estimate the correlation of  $r_g$  and  $r_{\text{GREX}}$  for the trait pairs in our study we perform Pearson's correlation analysis with Holm's adjustment for multiple comparisons.

We identify all the significantly correlated trait-pairs ( $r_g$  and  $r_{\text{GREX}}$ ,  $q$  value  $\leq 0.05$  as above) and perform bi-directional regression analyses<sup>44</sup> to identify causal relationships among the traits of our study (Supplementary Fig. 21). Then, taking as an example the coronary artery disease (CAD), we graph all the putative causal and protective relationships up to 2 nodes upstream in Fig. 5c (when the causal relationship is bi-directional between 2 traits, the relationship with the higher degrees of freedom is kept) and perform pathway enrichment analysis of shared agonistic genes for this causal network in Fig. 5d. For each causal or protective trait in the network, we generate a list of genes whose expression changes are predicted towards the same direction (or the opposite direction for protective traits) in CAD. These lists of shared agonistic genes are used for GSEA for common pathways. In Fig. 5d, only the top 15 (based on  $q$  value) results are shown and are ranked based on odds ratio.

**URLs.** For CMC, see <http://commonmind.org/>; for Synapse for CMC data, see <https://www.synapse.org/cmc>; for GTEX portal, see <http://www.gtexportal.org/>; for MSigDB, see <http://software.broadinstitute.org/gsea/msigdb>; for EpiXcan website and repository, see <http://icahn.mssm.edu/EpiXcan>; for EpiXcan source code, see <https://bitbucket.org/roussoslab/epixcan>; for qTLBHM package, see <https://github.com/rajanil/qTLBHM>; for RHOGE package, see <https://github.com/bogdanlab/RHOGE>; for PrediXcan pipeline, see <https://github.com/hakyim/PrediXcan>; for PredictDB resource, see [https://github.com/hakyimlab/PredictDB\\_Pipeline\\_GTEEx\\_v7](https://github.com/hakyimlab/PredictDB_Pipeline_GTEEx_v7); for Clue Drug Repurposing Hub, see <https://clue.io/repurposing>; for DPR, see <https://github.com/biostatp2eng/DPR>; for BSLMM, see <https://github.com/genetics-statistics/GEMMA/releases>.

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

## Data availability

The data sets analyzed during the current study are available for download from the links provided in the URLs section; of note is that some are controlled-access data. The data sets generated by the analyses of this study are provided as Supplementary Data files. Intermediate data sets derived from online aggregator databases will be made available from the corresponding author upon reasonable request.

Received: 14 June 2019 Accepted: 5 August 2019

Published online: 23 August 2019

## References

1. Visscher, P. M. et al. 10 Years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
2. Farh, K. K. H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
3. Roussos, P. et al. A role for noncoding variation in schizophrenia. *Cell Rep.* **9**, 1417–1429 (2014).
4. Fullard, J. F. et al. An atlas of chromatin accessibility in the adult human brain. *Genome Res.* **28**, 1243–1252 (2018).
5. Hauberg, M. E. et al. Large-scale identification of common trait and disease variants affecting gene expression. *Am. J. Hum. Genet.* **100**, 885–894 (2017).
6. Hauberg, M. E. et al. Differential activity of transcribed enhancers in the prefrontal cortex of 537 cases with schizophrenia and controls. *Mol. Psychiatry* **8**, 1 (2018).

7. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
8. Giambartolomei, C. et al. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538–2545 (2018).
9. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
10. Nica, A. C. et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).
11. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
12. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
13. Fryett, J. J., Inshaw, J., Morris, A. P. & Cordell, H. J. Comparison of methods for transcriptome imputation through application to two common complex diseases. *Eur. J. Hum. Genet.* **26**, 1658–1667 (2018).
14. So, H. C. et al. Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. *Nat. Neurosci.* **20**, 1342–1349 (2017).
15. Roadmap Epigenomics, Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).
16. Gaffney, D. J. et al. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13**, 1–15 (2012).
17. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
18. Fromer, M. et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
19. Franzén, O. et al. Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science* **353**, 827–830 (2016).
20. GTE Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
21. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11**, 407–409 (2014).
22. Zeng, P. & Zhou, X. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nat. Commun.* **8**, 456 (2017).
23. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B* **64**, 479–498 (2002).
24. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
25. Ntambi, J. M. et al. Loss of stearoyl-CoA desaturase-1 function protects mice against adiposity. *Proc. Natl Acad. Sci.* **99**, 11482–11486 (2002).
26. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
27. Barbeira, A. N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1–20 (2018).
28. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
29. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
30. Online Mendelian Inheritance in Man, OMIM®. *McKusick-Nathans Institute of Genetic Medicine* (Johns Hopkins University, Baltimore, 2018).
31. Wang, L. et al. SoftPanel: a website for grouping diseases and related disorders for generation of customized panels. *BMC Bioinforma.* **17**, 153 (2016).
32. Blake, J. A. et al. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res.* **45**, D723–D729 (2017).
33. Shohat, S., Ben-David, E. & Shifman, S. Varying intolerance of gene pathways to mutational classes explain genetic convergence across neuropsychiatric disorders. *Cell Rep.* **18**, 2217–2227 (2017).
34. Pardiñas, A. F. et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
35. Hodos, R. A., Kidd, B. A., Shameer, K., Readhead, B. P. & Dudley, J. T. In silico methods for drug repurposing and pharmacology. *Wiley Interdiscip. Rev.* **8**, 186–210 (2016).
36. Lamb, J. et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
37. Kunkel, S. D. et al. Ursolic acid increases skeletal muscle and brown fat and decreases diet-induced obesity, glucose intolerance and fatty liver disease. *PLoS ONE* **7**, e39332 (2012).
38. Ruderfer, D. M. et al. Polygenic overlap between schizophrenia risk and antipsychotic response: a genomic medicine approach. *Lancet Psychiatry* **3**, 350–357 (2016).
39. Law, V. et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42**, D1091–D1097 (2014).
40. Keiser, M. J. et al. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197–206 (2007).
41. Nadeau, K. J., Maahs, D. M., Daniels, S. R. & Eckel, R. H. Childhood obesity and cardiovascular disease: Links and prevention strategies. *Nat. Rev. Cardiol.* **8**, 513–525 (2011).
42. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
43. Anttila, V. et al. Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).
44. Mancuso, N. et al. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).
45. Pillinger, T. et al. Impaired glucose homeostasis in first-episode schizophrenia: a systematic review and meta-analysis. *JAMA Psychiatry* **74**, 261–269 (2017).
46. Rees, E. et al. Association between schizophrenia and both loss of function and missense mutations in paralog conserved sites of voltage-gated sodium channels. *bioRxiv* <https://doi.org/10.1101/246850> (2018).
47. Friedman, J. I. et al. Guanfacine treatment of cognitive impairment in schizophrenia. *Neuropsychopharmacology* **25**, 402–409 (2001).
48. Shamai, L. et al. Association of body mass index and lipid profiles: evaluation of a broad spectrum of body mass index patients including the morbidly obese. *Obes. Surg.* **21**, 42–47 (2011).
49. Do, R. et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat. Genet.* **45**, 1345–1353 (2013).
50. Voight, B. F. et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* **380**, 572–580 (2012).
51. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
52. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
53. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
54. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
55. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305 (2011).
56. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
57. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
58. Agresti, A. *An Introduction to Categorical Data Analysis: Second Edition*. (Wiley-Interscience, 2007).
59. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
60. Iorio, F. et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl Acad. Sci.* **107**, 14621–14626 (2010).
61. Bondareff, W., Mountjoy, C. Q., Roth, M. & Hauser, D. L. Neurofibrillary degeneration and neuronal loss in alzheimer's disease. *Neurobiol. Aging* **10**, 709–715 (1989).
62. Stenson, P. D. et al. The human gene mutation database: 2008 update. *Genome Med.* **1**, 13 (2009).
63. Mailman, M. D. et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007).
64. Becker, K. G., Barnes, K. C., Bright, T. J. & Wang, S. A. The genetic association database. *Nat. Genet.* **36**, 431–432 (2004).
65. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
66. Li, M. J. et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **40**, D1047–D1054 (2012).
67. Köhler, S. et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**, D966–D974 (2014).
68. Yu, W., Gwinn, M., Clyne, M., Yesupriya, A. & Khoury, M. J. A navigator for human genome epidemiology. *Nat. Genet.* **40**, 124–125 (2008).
69. Amberger, J. S. & Hamosh, A. Searching online mendelian inheritance in man (OMIM): a knowledge base of human genes and genetic phenotypes. *Curr. Protoc. Bioinforma.* **58**, 1.2.1–1.2.12 (2017).
70. Rouillard, A. D. et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**, baw100 (2016).

## Acknowledgements

The authors would like to thank: Scott Dickinson, Alvaro Barbeira, and Hae Kyung Im for providing valuable feedback on their PrediXcan and S-PrediXcan pipelines; David L. Morris and Timothy J. Vyse for providing access to GWAS summary results in systemic lupus erythematosus; Ping Zeng and Xiang Zhou for providing suggestions on proper

implementations of their BSLMM and DPR pipelines; Hon-Cheong So for providing the top-100 candidate lists of their drug repositioning analysis for psychiatric traits; and the authors Ethan Bahl, Tanner Koomar, and Jacob J Michaelson of the freely available (under the GNU GPLv3 license) cerebroViz R package used to generate an image that served as the basis of the brain artwork in Fig. 1a. This work was supported by the National Institutes of Health (NIH) grants R01AG050986 (P.R.), R01MH109897 (P.R. and E.E.S.) and R01MH109677 (P.R.), Leon Levy Foundation (G.V.; Leon Levy Fellowship in Neuroscience) and the Veterans Affairs Merit grants BX002395 and BX004189 (P.R.). Further, this work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

### Author contributions

P.R., W.Z., and G.V. conceived and designed the study. W.Z. implemented the EpiXcan pipeline and built the website database. W.Z., G.V., V.R., and B.R. did the downstream analysis. Additional feedback and discussion was provided by J.T.D., E.E.S., J.L.M.B., Y. K., J.F.F., and G.E.H. W.Z., G.V., V.R., and P.R. wrote the paper with input from all the authors.

### Additional information

**Supplementary Information** accompanies this paper at <https://doi.org/10.1038/s41467-019-11874-7>.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Peer review information:** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

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



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

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2019