Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

Aggregative trans-eQTL analysis detects trait-specific target gene sets in whole blood

Diptavo Dutta
Johns Hopkins University https://orcid.org/0000-0002-6634-9040

Yuan He
Johns Hopkins University

Ashis Saha
Johns Hopkins University

Marios Arvanitis
Department of Medicine, Division of Cardiology, Johns Hopkins School of Medicine, Baltimore MD USA

Alexis Battle
Johns Hopkins University https://orcid.org/0000-0002-5287-627X

Nilanjan Chatterjee (nchatte2@jhu.edu)
Johns Hopkins Bloomberg School of Public Health

Article

Keywords: genome-wide association studies (GWAS), genetic traits, trans-eQTL

Posted Date: May 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-523532/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Aggregative *trans*-eQTL analysis detects trait-specific target gene sets in whole blood

Diptavo Dutta[1], Yuan He[2], Ashis Saha[3], Marios Arvanitis[2,4], Alexis Battle[2,3,*] & Nilanjan Chatterjee[1,5,*]

[1] Department of Biostatistics, Johns Hopkins University.

[2] Department of Biomedical Engineering, Johns Hopkins University.

[3] Department of Computer Science, Johns Hopkins University.

[4] Department of Cardiology, Johns Hopkins University.

[5] Department of Oncology, Johns Hopkins University.

[*]Contributed equally

Correspondences to: ajbattle@jhu.edu and nilanjan@jhu.edu

Abstract

Large scale genetic association studies have identified many trait-associated variants and understanding the role of these variants in downstream regulation of gene-expressions can uncover important mediating biological mechanisms. In this study, we propose Aggregative *tRans* assoCiation to detect pHenotype specIfic gEne-sets (ARCHIE), as a method to establish links between sets of known genetic variants associated with a trait and sets of co-regulated gene-expressions through *trans* associations. ARCHIE employs sparse canonical correlation analysis based on summary statistics from *trans*-eQTL mapping and genotype and expression correlation matrices constructed from external data sources. A resampling based procedure is then used to test for significant trait-specific trans-association patterns in the background of highly polygenic regulation of gene-expression. Simulation studies show that compared to standard trans-eQTL analysis, ARCHIE is better suited to identify "core"-like genes through which effects of many other genes may be mediated and which can explain disease specific patterns of genetic associations. By applying ARCHIE to available *trans*-eQTL summary statistics reported by the eQTLGen consortium, we identify 71 gene networks which have significant evidence of *trans*-association with groups of known genetic variants across 29 complex traits. Around half (50.7%) of the selected genes do not have any strong *trans*-associations and could not have been detected by standard trans-eQTL mapping. We provide further evidence for causal basis of the target genes through a series of follow-up analyses. These results show ARCHIE is a powerful tool for identifying sets of genes whose *trans* regulation may be related to specific complex traits. The method has potential for broader applications for identification of networks of various types of molecular traits which mediates complex traits genetic associations.

Introduction

Genome-wide association studies (GWAS) have identified tens of thousands of common variants associated with a variety of complex traits [1] and a majority of these identified trait-related variants are in the non-coding regions of the genome[2–4]. It has been shown that these GWAS identified variants have a substantial overlap with variants that are associated with the expression levels of genes (eQTL) [5–7]. A number of tools [8–11] have been developed to identify potential target genes through which genetic associations may be mediated by investigating the effect of variants on local genes (*cis*-eQTL), typically within 1Mb region around the variant, but underlying causal interpretation remains complicated due to linkage disequilibrium and pleiotropy. A recent study has shown that a modest fraction of trait-heritability can be explained cis-mediated bulk gene-expressions [12], but future studies with more cell-type specific information has the potential to explain further.

Compared to *cis*-eQTL, studies of *trans*-eQTL have received less attention though they have the potential to illuminate downstream genes and pathways that would shed light on disease mechanism. A major challenge has been the limited statistical power for detection of *trans*-eQTL effects due to much weaker effects of SNPs on expressions of distal genes compared to those in cis-regions and a very large burden of multiple testing. However, *trans*-effects, when detected, has been shown to be more likely to have tissue-specific effects[13,14] and are more enriched than cis-eQTLs among disease loci[15]. *Trans*-eQTLs are, in general, known to act on regulatory circuits governing broader groups of genes[16] and thus have the potential to uncover gene networks and pathways consequential to complex traits [17,18]. Limited studies of *trans*-eQTL effect of known GWAS loci have identified complex downstream effects on known consequential genes for diseases[15,19]. In fact, an "omnigenic" model of complex traits has been hypothesized under which a large majority of genetic associations is mediated by cascading *trans*-effects on a few "core genes" [20,21]. Thus, given the increasing scope of eQTL studies, it has become even more important to comprehensively identify trait-specific trans-associations to highlight biological processes and mechanisms underlying phenotypic change. However, to the best of our knowledge, no framework has been developed to detect such trait-specific trans-association patterns and sets of trait-relevant genetically

regulated genes, specifically leveraging summary statistics from transcriptomic studies, which are more readily available than individual-level genotype data.

In this article, we propose a novel summary-statistics based method using sparse canonical correlation analysis (sCCA)[22–24] framework, termed Aggregative *t*R*ans* assoCiation to detect pHenotype specIfic gEne-sets (ARCHIE), which identifies sets of distal genes whose expression levels are trans-associated to (or regulated by) groups of GWAS SNPs associated to a trait. The method requires several summary statistics from standard SNP-gene expression *trans*-eQTL mapping, estimates of linkage disequilibrium (LD) between the variants and co-expression between genes, which can be estimated using publicly available datasets. Together, the selected variants and genes (jointly termed ARCHIE components) reflect significant trait-specific patterns of *trans*-association (**Figure 1A** shows an illustration for the functionality of ARCHIE). Compared to standard *trans*-eQTL mapping, the proposed method improves power for detection of signals by aggregating multiple *trans*-association signals across GWAS loci and genes. Moreover, we propose a resampling-based method to assess the statistical significance of the top components of sCCA for testing enrichment of trait-specific signals in the background of broader genome-wide *trans*-associations. If multiple ARCHIE components are significant, they reflect approximately orthogonal patterns of *trans*-associations for the trait-related variants, with the selected target genes pertaining to distinct downstream mechanisms of *trans*-regulation.

We apply the proposed method to analyze large-scale *trans*-association summary statistics for SNPs associated with 29 traits reported by the eQTLGen consortium[19]. The results show that ARCHIE can identify trait-specific patterns of *trans*-associations and relevant sets of variants and co-regulated target genes. The majority (50.7%) of the detected target genes are novel, meaning they would not have been identified by standard *trans*-eQTL mapping alone. We provide independent evidence supporting our results, using a series of downstream analysis to show that the selected target genes are enriched in known trait-related pathways and define directions of associations for the SNPs that are more enriched for underlying trait heritability than expected by chance. The proposed methods can be further applied in the future to association statistics data on other types of high

throughput molecular traits, such as proteins and metabolites, to understand their mediating role in genetic architecture of complex trait.

Results

Overview of Methods

We assume that we have summary statistics data (Z-values and p-values) available for a set of variants identified through large-scale GWAS of a given trait of interest, from standard trans-eQTL analysis across large number of distal genes. We further assume that we have additional reference datasets to estimate correlation (linkage disequilibrium) among the SNPs and among gene expressions in the underlying population of interest. ARCHIE uses these datasets to employ a sparse canonical correlation analysis[20,22] (sCCA) which produces sparse linear combinations of the trait-related variants (termed variant-component) that is associated with sparse linear combination of genes (termed gene-component) where each non-zero element of the variant (or gene)-component indicates that the respective variant (or gene) is selected. (**Figure 1A** shows a toy example of ARCHIE's functionality). The selected genes are broadly trans-regulated by the selected SNPs and mediate their effect to the trait. Further, we evaluate whether the genes and variants selected in the ARCHIE components reflect significant trait-specific trans-association patterns, through a resampling method by comparing the observed sparse canonical correlation values to that expected from trans-associations of GWAS variants not specific to a trait (for details see **Methods** and **Supplementary Section A**). We show through simulation and resampling studies that by jointly analyzing multiple GWAS variants associated to a trait, ARCHIE can identify broader downstream trans-regulatory mechanisms relevant to the trait compared to standard trans-eQTL mapping which identifies general trans-associations that might not be trait-specific and can arise due to factors like pleiotropy, correlated expressions and others (See **Supplementary Section B**).

Simulation Study Results

We performed simulation and resampling experiments to compare the performance of ARCHIE with standard trans-eQTL mapping.

Identification of downstream genes. The principal aim of ARCHIE is to aggregate multiple, possibly weaker, trans-associations of downstream genes with trait-related variants. We simulated data from a simple underlying causal network of genes with Genes 5-6 having stronger trans-regulation and Genes 7-9 having weaker trans association with the upstream variants (**Figure 1B** and **Supplementary Section B**). Across different sample sizes (N = 1,000 and 30,000), we found that ARCHIE selects the downstream genes (Genes 7-9) that has trans-associations to multiple upstream genetic variants, with high probability. In comparison, standard trans-eQTL mapping has a lower power to detect such cascaded trans-associations on an average (**Figure 1C** and **Supplementary Table 1**). Standard trans-eQTL mapping can only identify stronger trans-associations irrespective of the underlying causal model (Gene 5-6), which might not be relevant to the genetic architecture of the trait overall. Hence, ARCHIE can be viewed as a complementary approach to standard trans-eQTL mapping and detects downstream target genes by aggregating multiple weaker trans-associations.

Assessing trait-specificity. To demonstrate that ARCHIE can identify trait-specific trans-associations as compared to standard trans-eQTL mapping, we performed resampling experiments using trans-eQTL summary statistics reported by eQTLGen consortium[19] across four different traits (See Sample description for details). For a given trait, we used summary statistics for 100 variants across 5,000 genes of which a certain proportion (δ) of the variants were related to the given trait and rest were variants randomly sampled from different traits reflecting the general background of trans-associations expected for GWAS variants (See **Supplementary Section B** for details). The results from applying ARCHIE, with varying δ (**Supplementary Figure 2**), show that the probability of at least one ARCHIE component to be significant increases with the increase in the proportion of trait-specific variants (δ), consistently across the four different traits.

The above numerical experiments taken together, demonstrate that ARCHIE can effectively identify weaker trait-specific trans-regulation effects.

Trait-specific patterns of *trans*-associations in Whole Blood

We applied ARCHIE on large-scale *trans*-association summary statistics for SNPs associated with 29 different traits reported by the eQTLGen consortium[19] to identify trans-regulated gene-sets associated with the respective traits. For each of the 29 traits, we applied ARCHIE to the *trans*-eQTL summary statistics for the set of GWAS loci identified for that trait and tested in eQTLGen, across all distal genes, and selected the trait-specific target genes via the significant gene components (See **Methods** and **Supplementary Figure 1A** for analysis details). On an average, across these traits, we detect 2 (max = 7 for "Height") significant sets of variant and gene components (ARCHIE components) capturing phenotype-specific *trans*-association patterns (**Supplementary Figure 1B**). Of the target genes selected by ARCHIE in the significant gene-components for each trait, approximately only 49.3% genes displayed a strong association in standard analysis (*trans*-eQTL p-value $< 1 \times 10^{-06}$ reported in eQTLGen) with any variant associated to that traits. The remaining 50.7% genes (termed "novel genes") harbors only weaker ($0.05 >$ p-value $> 1 \times 10^{-06}$) associations and hence cannot be detected by standard *trans*-eQTL mapping alone; these genes display a similar pattern of *trans*-association with corresponding selected trait-related variants and are detectable only via the significant ARCHIE components. We made the list of target genes and variants selected by ARCHIE for each phenotype publicly available through an openly accessible database (See URL). Here, we focus on results for three different phenotypes their corresponding *trans*-association patterns, the selected target gene-sets and the novel genes detected by ARCHIE.

<u>*Schizophrenia.*</u> Schizophrenia is a neuropsychiatric disorder that affects perception and cognition. The eQTLGen consortium reports complete (non-missing) *trans*-association statistics for 218 SNPs, curated from multiple large-scale GWAS, associated with Schizophrenia (SCZ) across 7,756 genes. Of these, 7,047 genes were expressed in whole blood of Genotype-Tissue Expression (GTEx)[25] v8 individuals. We identified one significant ARCHIE

component capturing *trans*-association patterns significantly related to SCZ (**Figure 2A-B**) consisting of 27 variants and 75 genes. Of the selected genes, only 16 (21.4%) had evidence of at least one strong association (p-value < 1×10^{-06}) and possibly multiple weaker (0.05 > p-value > 1×10^{-06}) association as reported by eQTLGen. The remaining 59 genes (78.6%) only had weaker *trans*-associations with SCZ-related variants and could not have been identified using traditional *trans*-eQTL mapping (**Table 1**, **Supplementary Table 2**). Using an expression imputation approach (See **Methods** for details), we found the selected trans-associations, in particular, the target genes mediate significant trait heritability (p-value < 0.001) than expected by chance (**Figure 2D**).

Several of the 59 identified novel genes have been previously been reported to be associated with neurological functions. For example, chemokine receptor 4 (*CXCR4*), a gene that underlies interneuron migration and several neurodegenerative diseases[26], was identified by aggregating weaker associations from 20 SCZ-related SNPs in the variant component, but does not have any significant *trans*-associations. Similarly, caveolin-1 (*CAV1*), which is a known regulator of a SCZ risk gene (*DISC1*)[27], aggregates 13 weaker association to SCZ-related variants in the variant component. Notably, the target genes identified by ARCHIE include genes such as *HSPA5* and *AP5S1*, which not only harbor multiple *trans*-associations from SCZ-related variants but have also been reported to have *cis*-variants associated with psychiatric disorders[28 29]. We investigated whether in general the genes selected by ARCHIE had have evidence of association with SCZ through *cis* variants. Aggregating results from several large-scale cis-eQTL studies across tissues[9,30], we found that 12 of the 59 of the (enrichment p-value = 2.8×10^{-05}) novel genes have nominally significant (p-value < 1×10^{-04}) evidence of cis-regulatory SNPs to be associated with SCZ or other different neuropsychiatric diseases.

By performing pathway enrichment analysis of the target genes, we investigated if the selected genes represented known SCZ-related biological mechanisms (See **Methods** for details). Among the significantly enriched pathways, the majority (51.3%) were immune related. In particular, we identified 42 GO pathways[31], 36 canonical pathways[32–35] and 4 hallmark pathways[36] to be strongly enriched (FDR adjusted p-value < 0.05) for the selected

genes with 73 (89.0%) of them containing at least one novel gene (**Figure 2C** and **Supplementary Table 3**). Several pathways, previously reported in connection to SCZ, are identified to be enriched (FDR adjusted p-value < 0.05) for the selected genes (**Figure 2C**). For example, among the enriched gene ontology (GO) terms, GO-0034976: response to endoplasmic reticulum stress[37] (FDR adjusted p-value=0.013), GO-055065 metal ion homeostasis[38] (adjusted p-value=0.029), GO-0006915: apoptotic process[39] (adjusted p-value = 0.029), GO-0043005: neuron projection (adjusted p-value = 0.021) have previously been suggested to be linked to SCZ. Four hallmark gene-sets are also found to be significantly enriched for the selected genes including glycolysis[40], hypoxia[41], mTORC1 signaling[42] and unfolded protein response[43], all of which have suggestive evidence of being associated to SCZ. Using numerous TF databases[44,45], we found that the selected target genes were enriched (adjusted p-value < 0.05) for targets of 10 TFs (**Supplementary Table 4**), several of which have been previously reported to be associated with neuropsychiatric disorders[46,47].

Protein-protein interaction (PPI) enrichment analysis using STRING (v11.0)[48] showed a significant enrichment (p-value = 1.1×10^{-03}) indicating that the corresponding proteins may physically interact. Next, we performed a differential expression enrichment analysis to investigate whether the target genes were differentially expressed in any of the 54 tissues in GTEx v8 dataset. For each tissue, we curated lists of differentially expressed genes across the genome. We defined a gene to be differentially expressed in a tissue if the corresponding gene expression level in that tissue was significantly different from that across the rest of the tissues (See **Supplementary Section C** for details). Using such pre-computed lists of differentially expressed genes for each tissue, we found that the target genes selected by ARCHIE were enriched within the set of differentially expressed genes in 12 different tissues including 4 brain tissues in GTEx v8 (**Supplementary Figure 3**). For example, 3 novel genes (*PADI2, KCNJ10, MLC1*), were highly differentially expressed in several brain tissues (**Supplementary Figure 4**), in comparison to their expression in rest of the tissues.

<u>*Ulcerative Colitis.*</u> Ulcerative colitis (UC) is a form of inflammatory bowel disease, affecting the innermost lining of colon and rectum, causing inflammation and sores in the digestive tract and can lead to several colon-related

symptoms and complications including colon cancer[49–51]. The eQTLGen consortium reports complete (non-missing) *trans*-association summary statistics for 163 SNPs associated with Ulcerative Colitis, curated from multiple large-scale GWAS, across 12,010 genes. Of these, 10,307 genes were expressed in Whole Blood from GTEx v8 individuals. Using ARCHIE, we detected two significant variant-gene components comprising of 74 SNPs and 148 genes in total (**Figure 3A** and **Supplementary Figure 5**; **Supplementary Table 2**) that reflect *trans*-association patterns specific to UC. Of the selected genes, 68 genes (45.9%) were novel, meaning they did not have any strong *trans*-association (**Table 1, Supplementary Table 2**) with the variants related to UC. Further, similar to SCZ, we found the associations of the SNPs with target genes was strongly enriched (p-value < 0.001) for heritability of UC than expected by chance alone (**Figure 3D**).

Several of the novel target genes detected have been previously linked to intestinal inflammations and diseases. For example, glycoprotein A33 (*GPA33*) is known to impact intestinal permeability[52] and is an established colon cancer antigen[53]. Recent research using mouse-models have reported a connection between the regulation of *GPA33* and the development of colitis and other colon related inflammatory syndromes[54]. We also identify spermine oxidase (*SMOX*) through its weaker association with 9 UC-related variants. *SMOX* is significantly upregulated in individuals with inflammatory bowel diseases[55] and has been implicated in gastric and colon inflammations as well as carcinogenesis[56].

Using a series of follow-up analyses, we identify several pathways to be enriched (FDR adjusted p-value < 0.05) for the selected target genes (**Supplementary Table 5-6**), majority of them being immune related (59.6%). Among others, the hallmark interleukin-2-STAT5 signaling pathway (FDR adjusted p-value = 1.6×10^{-08}) has previously been reported to be associated to development of UC via suppression of immune response [57]. Various GO pathways related to endocytosis, lymphocyte activation, T-cell activation are found to be overrepresented in the selected target genes as well (**Figure 3B** and **Table 3**). Further enrichment analysis using broad TF databases, we found the selected target genes across both gene-components are enriched (adjusted p-value < 0.01) for targets

of 18 different TFs, majority of which have been previously reported to be involved in mucosal inflammation, inflammation of the intestine and epithelial cells and in immune-related responses (**Supplementary Table 7**).

Protein-protein interaction (PPI) enrichment analysis shows that the resultant proteins interact more often than random (p-value= 8.8×10^{-03} and 1.3×10^{-03} respectively for two significant ARCHIE components) (Figure 3C). Additionally, the selected genes were found to be enriched for genes significantly differentially expressed in several relevant tissues like colon-sigmoid and small-intestine ileum among others (**Supplementary Figure 6**).

We further investigated if any known mechanism can explain how the selected genes are associated with the selected variants, including mechanisms reflecting *cis* mediation[15]. In one example from our analysis, we observe that, among the 41 variants selected by variant-component 1, one UC-related variant rs3774959 is a *cis*-eQTL of *NFKB1* (p-value = 6.2×10^{-41} in eQTLGen and 6.3×10^{-05} in GTEx in whole blood). The Nuclear factor κB (NF-κB) family of transcription factors (TF) including *NFKB1*, has been extensively reported to be involved in immune[58] and inflammatory responses[59]. In particular, mutations in the promoter region of Nuclear factor κB1 (*NFKB1*) have been strongly implicated to be associated to UC[60], although the downstream target genes of *NFKB1* that are associated with UC, are largely unknown. Among 106 target genes selected in the first gene component, there are 6 genes (*CD74, CD83, IL1B, Il2RA, PTPN6, FOXP3*) that are reported targets for *NFKB1* (adjusted enrichment p-value = 7.5×10^{-03}) in TRRUST v2.0[44]. Thus, it can be conceptualized that the selected UC-related variant may regulate the expression levels of the 6 selected targets of *NFKB1* via *cis*-regulation of *NFKB1* expression levels, influencing UC-status downstream.

Prostate Cancer. Prostate cancer (PC) is one of the most common types of cancers in middle aged and older men, having a high public health burden with more than 3 million new cases in USA per year. The eQTLGen consortium reports complete (non-missing) *trans*-association summary statistics for 122 SNPs associated with prostate cancer, curated from multiple large-scale GWAS, across 12,951 genes. Of these, 11,385 genes were

expressed in Whole Blood from GTEx v8 individuals. Using ARCHIE, we detected two significant variant-gene components comprising 33 SNPs and 53 genes in total (Figure 4A; **Supplementary Table 2**) that reflect *trans*-association patterns specific to PC, of which 44 genes (83.1%) were novel (**Table 1**, **Supplementary Table 2**). Additionally, similar to SCZ and UC, we found evidence of enrichment of trans-heritability of PC that can be mediated by the target genes (Figure 4D), but the level of significance achieved was relatively weaker (p-value = 0.002 and 0.008; See **Methods** for details).

Among the novel genes, we identified several key genes that are generally implicated in different types of cancers. For example, *TP53* aggregates weaker *trans*-associations with 9 PC-related variants in ARCHIE component 1 (Figure 4B). The *TP53* gene encodes tumor protein p53 which acts as a key tumor suppressor and regulates cell division in general. *TP53* is implicated in a large spectrum of cancer phenotypes and has been considered to be one of the most important cancer genes studied[61]. Further, genes associated with the second gene component included *SMAD3* (Figure 4C) which is also a well-known tumor suppressor gene that plays a key role in transforming growth factor β (TGF-β) mediated immune suppression and also in regulating transcriptional responses suitable for metastasis[62–64]. *TP53* and *SMAD3* belong to two different ARCHIE components meaning that they might pertain to two relatively distinct biological processes that are independently affected by different sets of PC related variants. Additionally, the second gene-component included *EEA1* which is reported to have significantly altered expression levels in prostate cancer patients[65].

Using enrichment analyses (**Supplementary Table 8-9**), we found several pathways, including broadly ubiquitous pathways to be significantly overrepresented in the selected genes for both the gene-components like regulation of intracellular transport (adjusted p-value=0.017) and mRNA 3'-UTR binding (adjusted p-value = 0.008). Notably, we found the selected genes to be enriched for targets of several transcription factors many of which have been associated with different types and subtypes of cancer (**Supplementary Table 10**). For example, we found a TF target enrichment for *SPAG9* (adjusted p-value = 0.016) which has been identified to be associated

to breast cancer, ovarian cancer, colorectal cancer and others [66]. We also found enrichment for targets of *SSRP1* (adjusted p-value = 0.016), which is differentially regulated in a wide spectrum of malignant tumors [67] along with enrichment for targets of *MYC* and *TP73* which are well established cancer related genes (**Supplementary Table 10**). However, we did not find any evidence for significance enrichment of PPI among identified genes.

The downstream analysis suggests that a majority of the pathways (78.1%) enriched for the selected genes are immune related as observed in the previous examples as well. This might have been driven by the fact that eQTLGen reports summary *trans*-associations in whole blood. In general, whole blood might not be the ideal candidate tissue to identify *trans*-associations pertaining to PC. It is conceivable that relevant tissue-specific analysis for PC could have illuminated further *trans*-association patterns and identified key tissue-specific target genes. Despite that, we can identify several genes which have been elaborately reported to be key target genes for various cancers as well as some novel trans-associations. This underlines the utility of our aggregative approach and that it can illuminate important target genes pertaining to a trait.

Conclusion

While modern genome-wide association studies have been successful in identifying large number of genetic variants associated with complex traits, the underlying biological mechanisms by which these association arise has remained elusive. Although trans genetic regulations, mediated through cis- or otherwise, has been proposed for detecting important target genes for GWAS variants, identification of trans associations using standard univariate SNP vs gene-expression association analysis is notoriously difficult due to weak effect-sizes and large multiple testing burdens. In this article, we have proposed ARCHIE, a novel method for identifying groups of trait-associated genetic variants whose effects may be mediated through *trans*-associations with groups of coregulated genes. Further, we develop a resampling-based method to test the statistical significance of trait-specific enrichment patterns in the background of expected highly polygenic broad trans association signals. We have shown through simulation studies that compared to standard trans-eQTL analysis, ARCHIE is more

powerful for the detection of "core"-like genes which may mediate the effects of many upstream genes and variants and can explain trait specific genetic associations.

Application of the method to eQTLGen consortium *trans*-eQTL statistics not only identified many novel *trans*-associations for trait-related variants, but also it helped to contextualize the individual associations in terms of broader trait-specific trans-regulation patterns that were detected by underlying gene and variant components. The set of selected target genes in the gene component is one of the key outputs of ARCHIE. Using a series of follow-up analyses for three different types of traits, we showed that the selected genes are often overrepresented in known disease-relevant pathways, enriched in protein-protein interaction networks, shows co-regulations across tissues and contains targets for known transcription factors implicated to the disease (SCZ and UC) and key tumor suppressor genes (PC). Further, using a *trans*-expression imputation approach, we demonstrated that the selected genes can significantly mediate heritability associated trait related variants. All of these analyses point out that the *trans*-association patterns we detect are likely to have trait specific biological basis.

There are several limitations of the proposed method and current analysis. First, in the current version ARCHIE, we begin with a set of genetic variants associated with a trait, but we do not incorporate the underlying association directions and effect sizes in the analysis. This approach allowed us to independently investigate identified target genes through testing for consistency of directions of association of the SNPs with the trait and those with the expressions of the target through the *trans*-heritability analysis. However, it is likely that incorporation of the direction of trait association for the SNPs in the sCCA analysis itself will lead to improved power for detection of the trait specific target genes. Further, incorporation of known functional annotation of genetic variants and other prior information regarding the relationship between genes can improve the power of the analysis as well. Although the estimation of the ARCHIE components is computationally efficient, in the current implementation, the resampling-based testing method is computationally intensive. In the future, further research is merited to

develop analytical approximation techniques to reduce the computational burden of ARCHIE. Further, due to the lack of existing methods to identify trait specific trans associations, we have compared the performance against the standard trans-eQTL mapping. Although the eQTLGen data analysis shows that ARCHIE can identify a broader range of gene-sets trans-regulated by GWAS variants as compared to standard trans-eQTL mapping, the goals of ARCHIE and the standard analysis are different and therefore not perfectly comparable – the goal of the standard analysis is to identify association in individual variant-gene pair irrespective of trait specificity. In contrast, ARCHIE can identify trait-specific trans-regulated target genes harboring multiple weaker associations with trait-related variants. As newer methods are developed to detect trait-relevant trans-regulated genes, more comprehensive analyses comparing alternative approaches will allow us to understand their benefits and limitations.

Currently, not many transcriptomic studies have made the summary statistic from trans-eQTL mapping available. However, as these studies grow in size and with improved methods of data sharing, we expect more studies to make the summary statistics from trans-eQTL mapping available, allowing researchers to investigate a broader range of diseases and traits. Further, ARCHIE can be broadly applicable to understand role of other types of molecular traits, such as proteins and metabolites, in mediating complex trait genetic associations. In the future, as data on molecular biomarkers become increasingly available in large biobanks, tools like ARCHIE will be increasingly needed to understand common pathways through which genes and biomarkers interact to cause specific diseases.

In this article we have analyzed summary statistics reported by eQTLGen in the whole blood. This is primarily because of the substantial effective sample size of eQTLGen. While the approach can be applied to eQTL results from other tissues, the underlying sample sizes may be too limited to yield sufficient power. Although blood might not be the most relevant tissue for a number of traits, our analyses did detect trans association patterns that appear to have a broader biological basis in the disease genetics, from multiple independent lines of evidence. Nevertheless, it is likely that our analysis has missed many *trans*-association patterns that will be present only in specific disease-relevant tissues, cell types or/and dynamic stages [68]. In the future, we will seek applications for

ARCHIE in various types of emerging eQTL databases to provide a more complete map of networks of genetics variants and *trans*-regulated gene expressions and relevant contexts.

In summary, in this article we have developed a novel summary-based method, ARCHIE, to detect trait-specific gene-sets by aggregating *trans*-associations from multiple trait-related variants. ARCHIE is a powerful tool for identifying target gene sets through which the effect of genetic variants on a complex trait may be mediated. In the future, applications of the methods to a variety of existing and new data on association between genetic variants with high-throughput molecular traits can provide insights to biological mechanisms underlying genetic basis of complex traits.

Acknowledgements:

The UK BioBank data was obtained under the UK BioBank resource application 17712. GTEx data was obtained from dbGAP (accession id: phs000424.v8.p2). Drs. Chatterjee, Dutta and Battle were supported by NIH R01-HG010480-01. Dr. Battle was additionally supported by 1R01MH109905 (NIMH). The authors declare no competing interests.

Author Contributions

D.D., N.C. and A.B. conceived the project. D.D. obtained the data and carried out the data analysis. D.D., Y.H., A.S., M.A., N.C. and A.B. interpreted the results. D.D. wrote the manuscript under the supervision of N.C. and A.B. All the authors critically read and approved the manuscript.

URL

eQTLGen (*trans*-eQTL summary statistics): https://www.eqtlgen.org/*trans*-eqtls.html
GTEx: https://www.gtexportal.org/home/
1000 Genomes: https://www.internationalgenome.org/data/
UKBiobank: https://www.ukbiobank.ac.uk/
FUMA: https://fuma.ctglab.nl/

ShinyGO: http://bioinformatics.sdstate.edu/go/

STRING: https://string-db.org/cgi/

GitHub: https://github.com/diptavo/ARCHIE (initial release)

References:

1. 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).
2. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
3. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
4. Eicher, J. D. *et al.* GRASP v2.0: An update on the Genome-Wide Repository of Associations between SNPs and Phenotypes. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gku1202
5. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* (2012). doi:10.1101/gr.136127.111
6. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (80-.).* **337**, 1190–1195 (2012).
7. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genet.* (2010). doi:10.1371/journal.pgen.1000888
8. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47(9)**, 1091–1098 (2015).
9. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* (2016). doi:10.1038/ng.3506
10. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).
11. Barbeira, A. N. *et al.* Integrating predicted transcriptome from multiple tissues improves association detection. *PLOS Genet.* **15**, e1007889 (2019).
12. 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.* (2020). doi:10.1038/s41588-020-0625-2
13. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* (2017). doi:10.1038/nature24277
14. Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: The muTHER study. *PLoS Genet.* (2011). doi:10.1371/journal.pgen.1002003
15. Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* (2013). doi:10.1038/ng.2756
16. Yao, C. *et al.* Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. *Am. J. Hum. Genet.* (2017). doi:10.1016/j.ajhg.2017.02.003
17. Brynedal, B. *et al.* Large-Scale *trans*-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. *Am. J. Hum. Genet.* (2017). doi:10.1016/j.ajhg.2017.02.004
18. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nature Reviews Genetics* (2006). doi:10.1038/nrg1964
19. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* (2018). doi:10.1101/447367
20. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
21. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* (2019). doi:10.1016/j.cell.2019.04.014
22. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* (2009). doi:10.1093/biostatistics/kxp008
23. Hardoon, D. R. & Shawe-Taylor, J. Sparse canonical correlation analysis. *Mach. Learn.* (2011). doi:10.1007/s10994-010-5222-7
24. Witten, D. M. & Tibshirani, R. J. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.* (2009). doi:10.2202/1544-6115.1470

25. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv* (2019). doi:10.1101/787903
26. Volk, D. W., Chitrapu, A., Edelson, J. R. & Lewis, D. A. Chemokine receptors and cortical interneuron dysfunction in schizophrenia. *Schizophr. Res.* (2015). doi:10.1016/j.schres.2014.10.031
27. Kassan, A. *et al.* Caveolin-1 regulation of disrupted-in-schizophrenia-1 as a potential therapeutic target for schizophrenia. *J. Neurophysiol.* (2017). doi:10.1152/jn.00481.2016
28. Kakiuchi, C. *et al.* Functional polymorphisms of HSPA5: Possible association with bipolar disorder. *Biochem. Biophys. Res. Commun.* (2005). doi:10.1016/j.bbrc.2005.08.248
29. Martin, J. *et al.* A brief report: de novo copy number variants in children with attention deficit hyperactivity disorder. *Transl. Psychiatry* (2019). doi:10.1101/2019.12.12.19014555
30. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.* (2017). doi:10.1016/j.ajhg.2017.01.031
31. Hill, D. P., Smith, B., McAndrews-Hill, M. S. & Blake, J. A. Gene Ontology annotations: What they mean and where they come from. in *BMC Bioinformatics* (2008). doi:10.1186/1471-2105-9-S5-S2
32. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* (2000). doi:10.1093/nar/28.1.27
33. Croft, D. *et al.* Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Res.* (2011). doi:10.1093/nar/gkq1018
34. Schaefer, C. F. *et al.* PID: The pathway interaction database. *Nucleic Acids Res.* (2009). doi:10.1093/nar/gkn653
35. Adriaens, M. E. *et al.* The public road to high-quality curated biological pathways. *Drug Discovery Today* (2008). doi:10.1016/j.drudis.2008.06.013
36. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1(6)**, 417–425 (2015).
37. Patel, S., Sharma, D., Kalia, K. & Tiwari, V. Crosstalk between endoplasmic reticulum stress and oxidative stress in schizophrenia: The dawn of new therapeutic approaches. *Neuroscience and Biobehavioral Reviews* (2017). doi:10.1016/j.neubiorev.2017.08.025
38. Landek-Salgado, M. A., Faust, T. E. & Sawa, A. Molecular substrates of schizophrenia: Homeostatic signaling to connectivity. *Molecular Psychiatry* (2016). doi:10.1038/mp.2015.141
39. Chen, X. *et al.* Apoptotic engulfment pathway and schizophrenia. *PLoS One* (2009). doi:10.1371/journal.pone.0006875
40. Liu, M.-L. *et al.* Severe disturbance of glucose metabolism in peripheral blood mononuclear cells of schizophrenia patients: a targeted metabolomic study. *J. Transl. Med.* **13**, 226 (2015).
41. Cannon, T. D., Yolken, R., Buka, S. & Torrey, E. F. Decreased Neurotrophic Response to Birth Hypoxia in the Etiology of Schizophrenia. *Biol. Psychiatry* **64**, 797–802 (2008).
42. Ryskalin, L., Limanaqi, F., Frati, A., Busceti, C. & Fornai, F. mTOR-Related Brain Dysfunctions in Neuropsychiatric Disorders. *Int. J. Mol. Sci.* **19**, 2226 (2018).
43. Kim, P., Scott, M. R. & Meador-Woodruff, J. H. Dysregulation of the unfolded protein response (UPR) in the dorsolateral prefrontal cortex in elderly patients with schizophrenia. *Mol. Psychiatry* (2019). doi:10.1038/s41380-019-0537-7
44. Han, H. *et al.* TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1013
45. Zheng, G. *et al.* ITFP: An integrated platform of mammalian transcription factors. *Bioinformatics* (2008). doi:10.1093/bioinformatics/btn439
46. Aberg, K. A. *et al.* Methylome-Wide Association Study of Schizophrenia. *JAMA Psychiatry* **71**, 255 (2014).
47. Martínez, G. *et al.* Regulation of Memory Formation by the Transcription Factor XBP1. *Cell Rep.* **14**, 1382–1394 (2016).
48. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1131

49. Bassotti, G. *et al.* Gastrointestinal motility disorders in inflammatory bowel diseases. *World J. Gastroenterol.* (2014). doi:10.3748/wjg.v20.i1.37
50. Farrokhyar, F., Marshall, J. K., Easterbrook, B. & Irvine, E. J. Functional gastrointestinal disorders and mood disorders in patients with inactive inflammatory bowel disease: Prevalence and impact on health. *Inflamm. Bowel Dis.* (2006). doi:10.1097/01.MIB.0000195391.49762.89
51. Lakatos, P. L. & Lakatos, L. Risk for colorectal cancer in ulcerative colitis: Changes, causes and management strategies. *World Journal of Gastroenterology* (2008). doi:10.3748/wjg.14.3937
52. Van Der Post, S. *et al.* Structural weakening of the colonic mucus barrier is an early event in ulcerative colitis pathogenesis. *Gut* (2019). doi:10.1136/gutjnl-2018-317571
53. Rageul, J. *et al.* KLF4-dependent, PPARγ-induced expression of GPA33 in colon cancer cell lines. *Int. J. Cancer* (2009). doi:10.1002/ijc.24683
54. Williams, B. B. *et al.* Glycoprotein A33 deficiency: A new mouse model of impaired intestinal epithelial barrier function and inflammatory disease. *DMM Dis. Model. Mech.* (2015). doi:10.1242/dmm.019935
55. Gobert, A. P. *et al.* Distinct immunomodulatory effects of spermine oxidase in colitis induced by epithelial injury or infection. *Front. Immunol.* (2018). doi:10.3389/fimmu.2018.01242
56. Hu, T. *et al.* Spermine oxidase is upregulated and promotes tumor growth in hepatocellular carcinoma. *Hepatol. Res.* (2018). doi:10.1111/hepr.13206
57. Sadlack, B. *et al.* Ulcerative colitis-like disease in mice with a disrupted interleukin-2 gene. *Cell* **75**, 253–261 (1993).
58. Hayden, M. S. & Ghosh, S. NF-κB in immunobiology. *Cell Research* (2011). doi:10.1038/cr.2011.13
59. Liu, T., Zhang, L., Joo, D. & Sun, S. C. NF-κB signaling in inflammation. *Signal Transduction and Targeted Therapy* (2017). doi:10.1038/sigtrans.2017.23
60. Borm, M. E. A., Van Bodegraven, A. A., Mulder, C. J. J., Kraal, G. & Bouma, G. A NFKB1 promoter polymorphism is involved in susceptibility to ulcerative colitis. *Int. J. Immunogenet.* (2005). doi:10.1111/j.1744-313X.2005.00546.x
61. Wang, X. & Sun, Q. TP53 mutations, expression and interaction networks in human cancers. *Oncotarget* (2017). doi:10.18632/oncotarget.13483
62. Millet, C. & Zhang, Y. E. Roles of Smad3 in TGF-β signaling during carcinogenesis. *Critical Reviews in Eukaryotic Gene Expression* (2007). doi:10.1615/CritRevEukarGeneExpr.v17.i4.30
63. Tang, P. M.-K. *et al.* Smad3 promotes cancer progression by inhibiting E4BP4-mediated NK cell development. *Nat. Commun.* **8**, 14677 (2017).
64. Lu, S., Lee, J., Revelo, M., Wang, X. & Dong, Z. Smad3 is overexpressed in advanced human prostate cancer and necessary for progressive growth of prostate cancer cells in nude mice. *Clin. Cancer Res.* (2007). doi:10.1158/1078-0432.CCR-07-1078
65. Johnson, I. R. D. *et al.* Endosomal gene expression: a new indicator for prostate cancer patient prognosis? *Oncotarget* **6**, 37919–37929 (2015).
66. Kanojia, D., Garg, M., Gupta, S., Gupta, A. & Suri, A. Sperm-Associated Antigen 9 Is a Novel Biomarker for Colorectal Cancer and Is Involved in Tumor Growth and Tumorigenicity. *Am. J. Pathol.* **178**, 1009–1020 (2011).
67. Garcia, H. *et al.* Facilitates Chromatin Transcription Complex Is an "Accelerator" of Tumor Transformation and Potential Marker and Target of Aggressive Cancers. *Cell Rep.* **4**, 159–173 (2013).
68. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet.* 1–16 (2020). doi:10.1016/j.tig.2020.08.009
69. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
70. Schäfer, J. & Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* (2005). doi:10.2202/1544-6115.1175
71. Saha, A. & Battle, A. False positives in *trans*-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Research* (2018). doi:10.12688/f1000research.17145.1

Methods

Sample Description

eQTLGen. The eQTLGen consortium[19] is a large-scale multi-study effort to identify to study the downstream effects of trait-related variants via their effects on gene-expression in whole blood. The consortium consists of 37 individual studies with a collective sample size of 31,684 participants. With this sample size, the study has relatively higher power to detect moderate to weaker effects of variants on gene-expression. 10,317 variants related to complex traits, compiled from several GWAS databases, were tested for *trans*-associations with the expression levels of 19,964 genes in whole blood. The authors have made summary statistics (Z-score, p-value) for these *trans*-eQTL mapping analyses freely available to public.

GTEx. The Genotype-Tissue Expression (GTEx) project[25] aims to study tissue-specific gene expression and regulation. We used individual level data from GTEx (v8) whole blood to construct the co-expression matrix (Σ_{EE}) and further downstream validation of the gene-sets selected using ARCHIE. In our analysis, we used the latest version (v8) of GTEx having gene-expression and genotype data with samples from 54 different tissues. In particular, 755 individuals had expression data on 20,315 genes for whole blood. Of these, we used 670 individuals with genotype data present.

UK Biobank. UK Biobank is a large biobank study with above 500,000 participants. Among several data resources available, the genotype data, hospitalization records and health-records data are available. We used individual level genotype data from UK Biobank to construct LD matrix (Σ_{GG}) and for further downstream analysis of the selected target genes.

The phenotype data constructed from hospitalization and health-data records were used in the quantification and testing of enrichment in *trans*-heritability explained by the selected target genes (See Methods). We included the individuals with European ancestry in the analysis. For example, in the analysis of schizophrenia (SCZ), we used a sample of 366,326 participants from UK Biobank to construct the imputed gene-expression levels and evaluate the corresponding regression r^2 as an estimate of *trans*-heritability on SCZ as a binary phenotype.

Methods

Estimating trait-specific pattern of *trans*-associations.

Our proposed method, Aggregative *t*R*ans* assoCiation to detect pHenotype specIfic gEne-sets (ARCHIE), can select target genes *trans*-associated with trait-related variants using summary statistics in a sparse canonical correlation framework. In particular, we aim to identify the downstream target genes that harbor relatively independent associations from variants related to a trait. To apply ARCHIE, we start with the summary statistics from *trans*-eQTL mapping (Z-value, p-value). Given the *trans*-association summary statistics across the variants related with the trait and all the corresponding distant genes (variant > 5Mb away from the transcription start site of the gene), we first adjust for the correlation within the variants and genes through appropriate linkage disequilibrium (LD) and co-expression matrices respectively as follows:

$$W = \Sigma_{EE}^{-1/2}\Sigma_{GE}\Sigma_{GG}^{-1/2}$$

where Σ_{GG} and Σ_{EE} are estimates from LD-matrix and co-expression matrix (see **Supplementary Section A**), and Σ_{GE} is the cross-correlation matrix obtainable using the Z-values from the standard *trans*-eQTL mapping across all pairs of variants and gene-expressions. It is important to adjust for the dependence within the variants and gene expression levels using the LD and co-expression matrices respectively, since we aim to identify trait-relevant gene-sets and variants through independent trans-associations. Furthermore, gene-expression levels in bulk tissues may appear to be correlated due to cell composition effects as well, which needs to be adjusted for by incorporating the estimated co-expression between the genes.

Using W, the correlation-adjusted matrix of *trans*-associations, ARCHIE employs sparse canonical correlation analysis[22,24] (sCCA) which produces a sparse linear combination of the variants (u; termed variant-component) that is strongly correlated with a sparse linear combination of genes (v; termed gene-component) by solving the following optimization problem

$$(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{argmax}\, \tilde{\boldsymbol{v}}^{\boldsymbol{T}} \boldsymbol{W} \tilde{\boldsymbol{u}}$$

with $||\tilde{\boldsymbol{u}}||_1 \leq \boldsymbol{c_u}$;$||\tilde{\boldsymbol{v}}||_1 \leq \boldsymbol{c_v}$ and $||\tilde{\boldsymbol{u}}||_2 = \mathbf{1}$, $||\tilde{\boldsymbol{v}}||_2 = \mathbf{1}$

where $||x||_h$ is the L_h norm of a vector x; c_u (or c_v) is the sparsity parameter on the variant (or gene) component for the lasso-type L_1 penalty. Sparsity aids in interpretation since each non-zero element of a variant or gene component indicates that the respective variant (or gene) is selected in that component. Thus, (u, v), which are the resultant variant and gene components (jointly termed ARCHIE components) can be interpreted as the sparse latent factors that explain the majority of the aggregated association between all the trait-related variants and all the genes. The corresponding sparse canonical correlation (cc-value) between each pair of variant and gene components, defined as $q^2 = \frac{(v^T W u)^2}{\sqrt{(u^T W^T W u)(v^T W W^T v)}}$ would be a measure of the cumulative association between the selected sets of variants and genes by aggregating multiple (possibly weaker) associations (**Figure 1A** shows an illustration using P variants and G genes). Multiple such components (u, v) can be extracted to reflect approximately orthogonal latent factors of the aggregative correlation, corresponding to possibly distinct mechanisms of trans-regulation (See **Supplementary Section A**).

At suitable levels of sparsity (See **Supplementary Section A**), ARCHIE components produce a much smaller number of selected target genes which harbor multiple moderate to weak *trans*-association from a selected set of trait-associated variants, thus reflecting a trait-specific pattern of *trans*-association. A detailed algorithm for the estimation of the ARCHIE components is provided in the **Supplementary Section A**.

Testing Hypothesis of Enrichment of Trait-Specific *trans*-Association using a Competitive Null Hypothesis Framework

To test which ARCHIE components significantly capture the phenotype-specific *trans*-association pattern we evaluated the results from the original analysis against a *competitive* null hypothesis. Since trait-related variants are expected to be enriched for *trans*-eQTLs in general, we test whether the cc-values obtained in the original analysis are higher than that obtained using the *trans*-summary statistics between a random set of GWAS-identified variants and genes of similar size, that do not reflect any trait-specific pattern. For this, we first construct a *null matrix* by taking a random sample of p variants from the pool of all variants available and extracting the corresponding *trans*-summary statistics for another set of randomly chosen g genes. Since eQTLGen reports the

trans-summary statistics across about 10,000 variants associated with different traits, we can construct the *null matrix* using the *trans*-summary statistics from these variants that are associated with different traits and not with the trait of interest. This matrix of *trans*-associations, by design, should not reflect phenotype-specific patterns. For example, in the analysis for Schizophrenia (SCZ) using summary statistics across 218 variants and 7,047 genes, we construct the null matrix using 1 variant selected at random from 218 randomly chosen traits and extracting their corresponding *trans*-summary statistics across 7,047 randomly chosen genes.

Then we use ARCHIE with the same sparsity levels as the original analysis, to extract the gene and variant components and calculate corresponding cc-values. We repeat this step multiple (M) times to generate a *competitive* null distribution of cc-values. We then evaluate the observed cc-values from the original analysis against the corresponding *competitive* null distributions to calculate the p-value. In particular, the p-value of the kth ARCHIE component is given as

$$p_k = \frac{\sum_{i=1}^{M} I(q_k^2 > q_{k;null(i)}^2)}{M}$$

where q_k^2 denotes the kth cc-value in the original analysis and $q_{k;null(.)}^2$ denotes the elements of the null distribution of the kth cc-value. We declare that the top L components significantly capture phenotype-specific *trans*-association patterns if

$$L = min\{k: p_k > \alpha; k = 1,2, \dots, \min(p, g)\} - 1$$

The random set of p variants should be carefully chosen so that none of the variants associated to the phenotype in consideration or any phenotype sharing substantial genetic correlation, are included. Further the set should be such that it does not include a large fraction of the variants from the same phenotype (different from the original phenotype), which may bias the *competitive* null distribution towards the *trans*-association cc-values for that phenotype.

Simulation set-up

To demonstrate that ARCHIE can identify downstream trans-associations, we simulate individual level gene-expression data for N individuals (N=1,000 or 30,000). First, we randomly sampled 50 independent SNPs from the UK Biobank. Then we simulated gene-expression data using the causal regulatory model shown in **Figure 1B** where the red arrows denote cis-regulatory effect of SNPs on gene expressions and blue arrows denote gene-gene regulatory effects. Given the genotypes at the sampled SNPs, the expressions levels for genes 1-4 were simulated from a gaussian error model, with effects sizes from a ***Gaussian* (0.7,0.1)** distribution. Given the simulated expression levels for genes 1-4, the expression levels for genes 5-9 were again simulated using a gaussian error model similarly (See **Supplementary Section B** for details). Using this simulated data, we applied ARCHIE and compared the results to that obtained from standard trans-eQTL mapping (significance level 1×10^{-06}) using the genotype data for SNPs and expression levels of genes 5-9, excluding the cis-genes (genes 1-4).

Next, through resampling experiments we assess whether ARCHIE can potentially identify trait specific trans-associations. Using data from eQTLGen consortium, we construct a matrix of trans-eQTL summary statistics (Z-values) across for p variants and g genes. Out of the p variants, we set δ proportion of them to be related to a particular trait. For high values of δ we expect that the trans-summary statistics matrix would reflect trans-association patterns pertaining to the trait and hence should be captured by the ARCHIE components. We then applied ARCHIE on this matrix and evaluate the significance of the components. We replicate this experiment multiple times in a given setting, to estimate the empirical probability of at least one ARCHIE component to be significant (See **Supplementary Section B** for details) and reported the empirical probability of at least one ARCHIE component to be significant across varying values of δ. We repeated this resampling experiment with four different traits (**Supplementary Figure 2**).

Analysis of eQTLGen data

To identify phenotype-specific *trans*-associations, we applied ARCHIE on the *trans*-association summary statistics for 10,317 trait-related variants across 19,942 genes reported by the eQTLGen consortium[19] (See Sample Description for details on the study). In line with the consortium, we defined any gene to be trans to a variant if the variant was located at least 5Mb from the transcription start site of the gene or on another chromosome. The data contains multiple variants associated with the same trait analyzed for *trans*-eQTL mapping. Our analysis was restricted to phenotypes that had at least 100 associated variants tested for *trans*-mapping in the consortium, producing 29 phenotypes. **Figure 1B** shows a graphical representation of the major steps of our workflow. Briefly, for each phenotype, we extracted the summary *trans*-eQTL association statistics (Z-score, p-value) and removed all genes that were in within 5Mb of any of the trait related variants. In the preprocessing step, we filtered for any missing data and retained the genes that were also expressed in GTEx (v8)[25] whole blood. This produced a list of approximately 129 (min: 112; max: 533) variants and 10,219 (min: 3426; max: 13910) genes on an average per phenotype. ARCHIE requires two additional matrices representing the correlation among the variants themselves (a linkage-disequilibrium matrix) and among the gene-expression levels (a co-expression matrix), which can be estimated using reference data. We constructed the LD-matrix for the variants from individual-level genotype information using 5,000 randomly selected, unrelated European samples in UK Biobank[69]. For the correlation between gene-expressions, we used a penalized co-expression matrix[70] of the corresponding genes constructed from the covariate-adjusted quantile normalized gene-expression levels for individuals in GTEx v8 data. Subsequently, for the given trait, we extracted the selected variants and genes using the significant components and were evaluated for presence of false-positives due to cross-mapping.

Cross-Mappability: Alignment errors due to similarity in sequenced reads can lead a substantial rise in false positives for detecting *trans*-eQTL associations [71]. With the selected ARCHIE components, we extracted the nearby genes expressed in GTEx v8 whole blood for the selected variants (TSS within ±500 kb of the variant) and evaluated the cross-mapping scores for these genes with the selected target genes. Across the 3 traits analyzed in this article, we found that all such gene-pairs were mostly non cross-mappable (SCZ: 99.98%, UC: 99.17%,

PC: 99.93%), indicating that the *trans*-association patterns were less likely to be affected by false positive arising from alignment errors.

Follow-up Analysis.

Quantifying and Testing for Enrichment for Trait Heritability Explained by Identified Target Genes. In the following, we propose a method for quantifying trait heritability explained by the GWAS variants that would be mediated by the identified target-genes and develop a corresponding test for enrichment through comparison of such estimates of mediated heritability associated with that from random genes. For or a particular trait of interest, we start with the Z-scores for regression-based *trans*-eQTL mapping for a set of underlying p variants and g genes. We will assume that, using ARCHIE, we have identified G target genes that capture trait-specific *trans*-association patterns. To perform the test as proposed above, we require individual level phenotype and genotype data independent of the samples used in the original analysis. Given genotypes (or dosages) at the p variant sites for an individual k, for each target gene, we define the *trans*-imputed expression scores (TIES) as the predicted expression value for the jth target gene as

$$TIES(p)_{jk} = \sum_{i=1}^{p} \frac{Z_{ij} x_{ik}}{\sqrt{2m_i(1-m_i)}},$$

where Z_{ij} is the z-score for the effect of the ith trait-related variant on the jth gene, x_{ik} is the genotype or dosage for the kth individual at the ith variant and m_i is the minor allele frequency of the ith variant. We construct the TIES under two different schemes:

1. Using all the trait-related variants with complete *trans*-association statistics reported in eQTLGen
2. Using only the trait-related variants selected in the significant components

To evaluate how strongly the TIES for the G target genes are associated with the phenotype levels, we use the following multiple regression model

$$g[E(y_k)] = \beta_0 + \sum_{j=1}^{G} TIES(p)_{jk}$$

where y_k is the phenotype value (e.g. disease status) for the kth individual; $g[.]$ is a canonical link function and can be set to be the identity function for continuous phenotype or the logistic function for binary (disease status) phenotypes. We record the pseudo-r^2 from this regression model as a measure of association between the TIES and the phenotype value. The pseudo-r^2 would provide an estimate of trans-heritability, meaning it can quantify the variance explained by the trait-related variants that is expected to be mediated via the selected target genes in context of the trans-associations reported. To test whether the observed r^2 is significant in comparison to what is expected at random, we adopt a resampling based approach. We sampled *g* genes (excluding the originally selected target genes) from the genome, constructed the corresponding TIES for the individuals and recorded the r^2 for the regression model. This would a null estimate of trans-heritability of the SNPs expected to be mediated by a set of *g* genes. If the observed pseudo-r^2 is substantially higher than the null estimates, we can infer that the trans-associations selected by ARCHIE explain higher variance compared to that expected through random trans-associations. We performed resampling multiple (1000) times to generate a control (null) distribution of r^2 to reflect the associations expected from random genes. We then calculated the p-value of the observed r^2 using the originally selected *g* genes from this control distribution to evaluate whether the TIES have any significant association with the phenotype.

Approximately, the observed r^2 reflects the proportion of trait-variance explained by the TIES. Thus, a significantly higher r^2 would imply that the selected genes harbor several *trans*-associations and mediate the effects of the trait-related variants more than any random set of genes. As the analysis of association between TIES and trait (for both the selected trans genes and random genes) is performed in an independent dataset, and no information on directions or magnitudes of trait association for the SNPs are used in the original ARCHIE analysis, the test for heritability enrichment provides independent validation of the relevance of selected target-genes in explaining genetic associations for the trait. In our application, we used individual level phenotype and genotype data from UK Biobank participants to estimate association between TIES and traits.

We also performed several other follow up analyses including PPI enrichment, pathway enrichment, differentially expressed genes enrichment. These analyses were carried out using pre-established standard pipelines. For full details on these see **Supplementary Section C**.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesfinal.xlsx
- Supplementaryfinal.docx