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)[22–27] 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[22,24] (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 reflect sets that are broadly trans-regulated by the selected SNPs and via which the effects of selected SNPs on the trait appear to be mediated. 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

120 *Comparison with Standard trans-eQTL Analysis*. We first compare ARCHIE with standard trans analysis using
121 a series of simulation studies. ARCHIE addresses a composite null hypothesis that there is no association between
122 a group of variants and genes, while standard trans-eQTL mapping tests a series of individual null hypotheses of
123 no association across variant-gene pairs. Thus, these two methods are not directly comparable in terms of power
124 unless the type-I error rates are calibrated with respect to a common benchmark. We first evaluate the type-I error
125 rate for each method under a global null hypothesis of no association between a group of SNPs and a gene network
126 (**Figure 2A**). Type-I error rate for ARCHIE is defined as the proportion of simulation iterations where the p-value
127 for at least one ARCHIE component is less than the chosen α-level. Correspondingly, for standard trans-eQTL
128 analysis, type-I error is calculated as the proportion of simulation iterations where there is at least one p-value
129 across all possible variant-gene pairs is less than the chosen α-level. We found that ARCHIE maintains a highly
130 conservative type-I error rate (**Table 1**) at all levels while standard trans-eQTL mapping maintains the overall
131 familywise error rate of standard trans-eQTL mapping at 0.05 for the approximate Bonferroni correction level
132 ($0.05/(40x9) \approx 1 \times 10^{-04}$). Subsequently, we found that at approximately level 9×10^{-04}, ARCHIE achieves similar
133 type-I error rate as the FWER of trans-eQTL mapping at level 1×10^{-06}. We use these levels for subsequent
134 empirical power estimations.
135
136 We next compared the power of ARCHIE to identify gene-sets associated to trait-related SNPs in comparison to
137 standard trans-eQTL mapping, for several different causal models (**Figure 2B-C**; See Methods for details on the
138 Simulation model). Briefly, each causal network consists of 40 SNPs associated to a trait, with successive 5 SNPs
139 having regulatory effect on a local (cis) gene. In total 8 cis (local) genes (in red) mediate the effect of these SNPs
140 on a network of 9 distal genes. The average cis heritability of the 8 cis-genes were maintained approximately at
141 20%-22% while the average trans-heritability for the 9 distal genes were maintained at about 10%-14%. Each
142 causal scenario was simulated at two sample sizes (N = 1,000 and 30,000) reflecting the approximate sample sizes
143 of GTEx v8 and eQTLGen studies respectively.

144 For sparse causal networks (**Figure 2B**), we found that the power of ARCHIE to detect genes in each layer of
145 network was comparable or somewhat better than trans-eQTL mapping under two sample sizes (**Figure 2D**).
146 However, for the dense causal network (**Figure 2C**), we find that ARCHIE has a major power advantage over
147 trans-eQTL mapping throughout (**Figure 2E**). This is potentially due to the increased weaker effects in the dense
148 causal network (**Figure 2C**) since the average trans-heritability is maintained at the same value as that of the
149 sparse causal network. This demonstrates that ARCHIE can effectively aggregate weaker associations compared
150 to trans-eQTL mapping. Additionally, we estimated power under the presence of a master regulator, where we
151 found that as before ARCHIE had comparable or higher power that trans-eQTL of identifying the downstream
152 genes (**Supplementary Section B2** and **Supplementary Figure 2**). We further compared the sensitivity and
153 specificity of ARCHIE with that of standard trans-eQTL and found that although the sensitivity of ARCHIE was
154 only slightly higher than, the specificity was substantially higher compared to trans-eQTL mapping
155 (**Supplementary Section B3** and **Supplementary Figure 3**).
156
157 It is to be noted that the potential space of alternative null hypothesis for relationship between a group of SNPs
158 and gene networks is extremely large, and a comprehensive evaluation is beyond the scope of this article.
159 However, the above small scale simulations provide us with an intuitive insight that ARCHIE has robust power
160 for detection of trans-association under a variety of plausible models for trans-associations.
161
162 *Assessing trait-specificity*. We note that a primary objective of our analysis was to detect disease specific trans-
163 association pattern in the background of broad trans-associations that are expected to be seen in the genome. Since
164 GWAS variants associated to traits are enriched in trans-eQTLs, we investigated whether ARCHIE captured any
165 trait specific trans-association beyond the general expected pattern of trans-associations for GWAS variants which
166 are not related to the trait in consideration. Thus, we considered this "*competitive null hypothesis*" to be more
167 pertinent for our main analysis and we conducted additional resampling studies to investigate performance of
168 ARCHIE for detecting trait-specific patterns. We performed resampling experiments using trans-eQTL summary
169 statistics reported by eQTLGen consortium[19] across four different traits (See Sample description for details). For

170 a given trait, we used summary statistics for 100 variants across 5,000 genes of which a certain proportion (δ) of
171 the variants were related to the given trait and rest were variants randomly sampled from different traits reflecting
172 the general background of trans-associations expected for GWAS variants (See **Supplementary Section B4** for
173 details). The results from applying ARCHIE, with varying δ (**Supplementary Figure 4**), show that the
174 probability of at least one ARCHIE component to be significant increases with the increase in the proportion of
175 trait-specific variants (δ), consistently across the four different traits. In particular, at $\delta = 0$, the results would
176 correspond to the type-I error of ARCHIE under the competitive null hypothesis. We found that at a level of $1 \times$
177 10^{-04}, the competitive type-I error of ARCHIE is conservative for all the traits under consideration indicating that
178 ARCHIE produces reduced false positives (Supplementary Table 1). Since competitive null hypothesis tests for
179 trait-specific trans-association patterns, the subsequent results presented here are corresponding to the competitive
180 null hypothesis unless otherwise mentioned.
181 The above numerical experiments taken together, demonstrate that ARCHIE can effectively identify weaker trait-
182 specific trans-regulation effects.
183

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

185 We applied ARCHIE on large-scale *trans*-association summary statistics for SNPs associated with 29 different
186 traits reported by the eQTLGen consortium[19] to identify trans-regulated gene-sets associated with the respective
187 traits. The eQTLGen consortium provides summary statistics (Z-values, p-values) from standard trans-eQTL
188 mapping for more than 10,000 variants across numerous loci in the genome, curated from external databases, that
189 have been identified to be associated to traits and diseases in large scale GWAS. For each of the 29 traits, we
190 extracted the *trans*-association summary statistics for the variants associated with the trait and only the genes that
191 were either more 5Mb away from each of the variants or on a different chromosome. Thus, by design, all the
192 genes in the analysis for a trait were distal to all the variants under consideration. We then applied ARCHIE to
193 this *trans*-eQTL summary statistics and selected the trait-specific target genes via the significant gene components
194 (See **Methods** and **Figure 1B** for analysis details). On an average, across these traits, we detect 2 (max = 7 for

195 "Height") significant sets of variant and gene components (ARCHIE components) capturing phenotype-specific
196 *trans*-association patterns (**Supplementary Figure 1**). Of the target genes selected by ARCHIE in the significant
197 gene-components for each trait, approximately only 49.3% genes displayed a strong association in standard
198 analysis (*trans*-eQTL p-value $< 1 \times 10^{-06}$ reported in eQTLGen) with any variant associated to that traits. The
199 remaining 50.7% genes (termed "novel genes") harbors only weaker ($0.05 >$ p-value $> 1 \times 10^{-06}$) associations and
200 hence cannot be detected by standard *trans*-eQTL mapping alone; these genes display a similar pattern of *trans*-
201 association with corresponding selected trait-related variants and are detectable only via the significant ARCHIE
202 components. In fact, a large majority of the novel genes (89%) harbor *multiple* weaker trans-associations with the
203 variants selected in the variant component. We made the list of target genes and variants selected by ARCHIE for
204 each phenotype publicly available through an openly accessible database (See URL). Here, we focus on results
205 for three different phenotypes, reflecting three different classes of diseases, their corresponding *trans*-association
206 patterns, the selected target gene-sets and the novel genes detected by ARCHIE.
207
208 *Schizophrenia.* Schizophrenia is a neuropsychiatric disorder that affects perception and cognition. The eQTLGen
209 consortium reports complete (non-missing) *trans*-association statistics for 218 SNPs, curated from multiple large-
210 scale GWAS, associated with Schizophrenia (SCZ) across 7,756 genes. Of these, 7,047 genes were expressed in
211 whole blood of Genotype-Tissue Expression (GTEx)[28] v8 individuals. Testing against the competitive null
212 hypothesis (See previous section), we identified one significant ARCHIE component capturing *trans*-association
213 patterns significantly related to SCZ (**Figure 3A-B**) consisting of 27 variants across several different genomic
214 loci and 75 genes. Of the selected genes, only 16 (21.4%) had evidence of at least one strong association (p-value
215 $< 1 \times 10^{-06}$) and possibly multiple weaker ($0.05 >$ p-value $> 1 \times 10^{-06}$) association as reported by eQTLGen. The
216 remaining 59 genes (78.6%) only had weaker *trans*-associations with SCZ-related variants (**Table 2**,
217 **Supplementary Table 2**). These novel genes were not identified using traditional *trans*-eQTL mapping and were
218 not reported as significant findings by the eQTLGen consortium. We further investigated the robustness of the
219 results using coexpression estimates from gene expressions levels reported in Whole Blood by the Depression

220 Genes and network (DGN) study[29]. The results show that out of 75 genes and 59 novel genes identified by
221 ARCHIE, 62 and 45 were also selected in the replication analysis using data from DGN, demonstrating the
222 potential robustness of the results by ARCHIE.
223
224 Next, we tested whether the selected variants and genes were enriched in trans-heritability for SCZ. Using an
225 expression imputation approach (See **Methods** for details), we estimated the approximate heritability of SCZ
226 explained by the trans-associations between the selected genes and variants, using individual level data from UK
227 Biobank. We compared the estimate to (1) the expected distribution of heritability explained by the selected SNPs
228 and randomly chosen genes (excluding the selected genes) for SCZ and (2) the expected distribution of heritability
229 explained by the selected SNPs and genes for a randomly chosen trait. The results showed that the selected SNPs
230 and genes are significantly enriched in trait heritability (p-value < 0.001) than expected by chance (**Figure 3D**).
231
232 Several of the 59 identified novel genes have previously been reported to be associated with neurological
233 functions. For example, chemokine receptor 4 (*CXCR4*), a gene that underlies interneuron migration and several
234 neurodegenerative diseases[30], was identified by aggregating weaker associations from 20 SCZ-related SNPs in
235 the variant component, but does not have any significant *trans*-associations. Similarly, caveolin-1 (*CAV1*), which
236 is a known regulator of a SCZ risk gene (*DISC1*)[31], aggregates 13 weaker association to SCZ-related variants in
237 the variant component. Notably, the target genes identified by ARCHIE include genes such as *HSPA5* and *AP5S1*,
238 which not only harbor multiple *trans*-associations from SCZ-related variants but have also been reported to have
239 *cis*-variants associated with psychiatric disorders[32 33]. We investigated whether in general the genes selected by
240 ARCHIE had have evidence of association with SCZ through *cis* variants. Aggregating results from several large-
241 scale cis-eQTL studies across tissues[9,34], we found that 12 of the 59 of the (enrichment p-value = 2.8×10^{-05}) novel
242 genes have nominally significant (p-value < 1×10^{-04}) evidence of cis-regulatory SNPs to be associated with SCZ
243 or other different neuropsychiatric diseases. Further, eQTLGen consortium reports the association of gene
244 expressions with several publicly available polygenic risk scores (PRS) of SCZ at different p-value thresholds,

curated from external databases. We found that out of the 59 novel genes identified, 38 (64.4%) had a nominal association with at least one reported PRS for SCZ.

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[35], 36 canonical pathways[36–39] and 4 hallmark pathways[40] 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 3C** 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 3C**). For example, among the enriched gene ontology (GO) terms, GO-0034976: response to endoplasmic reticulum stress[41] (FDR adjusted p-value=0.013), GO-055065 metal ion homeostasis[42] (adjusted p-value=0.029), GO-0006915: apoptotic process[43] (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[44], hypoxia[45], mTORC1 signaling[46] and unfolded protein response[47], all of which have suggestive evidence of being associated to SCZ. Using numerous TF databases[48,49], 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[50,51]. While connection between immune pathways and SCZ is less obvious, we believe this is also a noteworthy finding given recent studies have increasingly pointed out complex interactions between the immune system, inflammation, and the brain[52]. However, relative importance of different pathways is likely to have been biased in our analysis because blood is more likely to contain stronger signature of immune mechanisms than processes that are directly related to brain.

.

Protein-protein interaction (PPI) enrichment analysis using STRING (v11.0)[53] showed a significant enrichment (p-value = 1.1×10^{-03}) indicating that the corresponding proteins may physically interact. Although the analysis is performed using trans-associations in whole blood, we found several of the selected target genes were

271 differentially expressed in tissues directly related with SCZ. For example, 3 novel genes (*PADI2, KCNJ10,*
272 *MLC1*), were highly upregulated in several brain tissues (**Supplementary Figure 5**), in comparison to their
273 expression in rest of the tissues in GTEx v8. To systematically evaluate this, we performed a differential
274 expression enrichment analysis to investigate whether the target genes were differentially expressed in any of the
275 54 tissues in GTEx v8 dataset. Since SCZ is a complex disease, we expect the selected genes to be differentially
276 expressed or regulated in numerous potentially distinct tissues. For each tissue, we curated lists of differentially
277 expressed genes across the genome. We defined a gene to be differentially expressed in a tissue if the
278 corresponding gene expression level in that tissue was significantly different from that across the rest of the tissues
279 (See **Supplementary Section C** for details). Using such pre-computed lists of differentially expressed genes for
280 each tissue, we found that the target genes selected by ARCHIE were enriched within the set of differentially
281 expressed genes in 12 different tissues including 7 brain tissues in GTEx v8 (**Supplementary Figure 6**). Majority
282 of these tissues had been previously reported to be involved in the pathophysiology of SCZ[54–56].
283
284 *Ulcerative Colitis.* Ulcerative colitis (UC) is a form of inflammatory bowel disease, affecting the innermost lining
285 of colon and rectum, causing inflammation and sores in the digestive tract and can lead to several colon-related
286 symptoms and complications including colon cancer[57–59]. The eQTLGen consortium reports complete (non-
287 missing) *trans*-association summary statistics for 163 SNPs associated with Ulcerative Colitis, curated from
288 multiple large-scale GWAS, across 12,010 genes. Of these, 10,307 genes were expressed in Whole Blood from
289 GTEx v8 individuals. Using ARCHIE, we detected two significant variant-gene components comprising of 74
290 SNPs and 148 genes in total (**Figure 4A** and **Supplementary Figure 7**; **Supplementary Table 2**) that reflect
291 *trans*-association patterns specific to UC. Of the selected genes, 68 genes (45.9%) were novel, meaning they did
292 not have any strong *trans*-association (**Table 2, Supplementary Table 2**) with the variants related to UC. Further,
293 similar to SCZ, we found the associations of the SNPs with target genes was strongly enriched (p-value < 0.001)
294 for heritability of UC than expected by chance alone and also the selected trans-associations explained heritability
295 of UC more than expected for a random trait or disease (**Figure 4D**).
296

297 Several of the novel target genes detected have been previously linked to intestinal inflammations and diseases.
298 For example, glycoprotein A33 (*GPA33*) is known to impact intestinal permeability[60] and is an established colon
299 cancer antigen[61]. Recent research using mouse-models have reported a connection between the regulation of
300 *GPA33* and the development of colitis and other colon related inflammatory syndromes[62]. We also identify
301 spermine oxidase (*SMOX*) through its weaker association with 9 UC-related variants. *SMOX* is significantly
302 upregulated in individuals with inflammatory bowel diseases[63] and has been implicated in gastric and colon
303 inflammations as well as carcinogenesis[64].
304
305 Using a series of follow-up analyses, we identify several pathways to be enriched (FDR adjusted p-value < 0.05)
306 for the selected target genes (**Supplementary Table 5-6**), majority of them being immune related (59.6%).
307 Among others, the hallmark interleukin-2-STAT5 signaling pathway (FDR adjusted p-value = 1.6×10^{-08}) has
308 previously been reported to be associated to development of UC via suppression of immune response [65]. Various
309 GO pathways related to endocytosis, lymphocyte activation, T-cell activation are found to be overrepresented in
310 the selected target genes as well (**Figure 4B** and **Table 4**). Further enrichment analysis using broad TF databases,
311 we found the selected target genes across both gene-components are enriched (adjusted p-value < 0.01) for targets
312 of 18 different TFs, majority of which have been previously reported to be involved in mucosal inflammation,
313 inflammation of the intestine and epithelial cells and in immune-related responses (**Supplementary Table 7**).
314
315 Protein-protein interaction (PPI) enrichment analysis shows that the resultant proteins interact more often than
316 random (p-value= 8.8×10^{-03} and 1.3×10^{-03} respectively for two significant ARCHIE components) (**Figure 4C**).
317 Additionally, the selected genes were found to be enriched for genes significantly differentially expressed in
318 several relevant tissues like colon-sigmoid and small-intestine ileum among others (**Supplementary Figure 8**).
319
320 We further investigated if any known mechanism can explain how the selected genes are associated with the
321 selected variants, including mechanisms reflecting *cis* mediation[15]. In one example from our analysis, we observe
322 that, among the 41 variants selected by variant-component 1, one UC-related variant rs3774959 is a *cis*-eQTL of

323 *NFKB1* (p-value = 6.2×10^{-41} in eQTLGen and 6.3×10^{-05} in GTEx in whole blood). The Nuclear factor κB (NF-
324 κB) family of transcription factors (TF) including *NFKB1*, has been extensively reported to be involved in
325 immune[66] and inflammatory responses[67]. In particular, mutations in the promoter region of Nuclear factor κB1
326 (*NFKB1*) have been strongly implicated to be associated to UC[68], although the downstream target genes of *NFKB1*
327 that are associated with UC, are largely unknown. Among 106 target genes selected in the first gene component,
328 there are 6 genes (*CD74, CD83, IL1B, Il2RA, PTPN6, FOXP3*) that are reported targets for *NFKB1* (adjusted
329 enrichment p-value = 7.5×10^{-03}) in TRRUST v2.0[48]. Thus, it can be conceptualized that the selected UC-related
330 variant may regulate the expression levels of the 6 selected targets of *NFKB1* via *cis*-regulation of *NFKB1*
331 expression levels, influencing UC-status downstream.
332
333 *Prostate Cancer*. Prostate cancer (PC) is one of the most common types of cancers in middle aged and older men,
334 having a high public health burden with more than 3 million new cases in USA per year. The eQTLGen
335 consortium reports complete (non-missing) *trans*-association summary statistics for 122 SNPs associated with
336 prostate cancer, curated from multiple large-scale GWAS, across 12,951 genes. Of these, 11,385 genes were
337 expressed in Whole Blood from GTEx v8 individuals. Using ARCHIE, we detected two significant variant-gene
338 components comprising 33 SNPs, spanning 14 different loci across the genome, and 53 genes in total (Figure 5A;
339 **Supplementary Table 2**) that reflect *trans*-association patterns specific to PC, of which 44 genes (83.1%) were
340 novel (**Table 2**, **Supplementary Table 2**). Additionally, similar to SCZ and UC, we found evidence of enrichment
341 of trans-heritability of PC that can be mediated by the target genes and also that the trans-heritability mediated by
342 the selected SNPs and genes was significantly more than that for a randomly chosen trait (**Figure 5D**), but the
343 level of significance achieved was relatively weaker (p-value = 0.001 and 0.007; See **Methods** for details).
344
345 Among the novel genes, we identified several key genes that are generally implicated in different types of cancers.
346 For example, *TP53* aggregates weaker *trans*-associations with 9 PC-related variants in ARCHIE component 1
347 (Figure 5B). The *TP53* gene encodes tumor protein p53 which acts as a key tumor suppressor and regulates cell

348 division in general. *TP53* is implicated in a large spectrum of cancer phenotypes and has been considered to be
349 one of the most important cancer genes studied[69]. Further, genes associated with the second gene component
350 included *SMAD3* (**Figure 5C**) which is also a well-known tumor suppressor gene that plays a key role in
351 transforming growth factor β (TGF-β) mediated immune suppression and also in regulating transcriptional
352 responses suitable for metastasis[70–72]. *TP53* and *SMAD3* belong to two different ARCHIE components meaning
353 that they might pertain to two relatively distinct biological processes that are independently affected by different
354 sets of PC related variants. Additionally, the second gene-component included *EEA1* which is reported to have
355 significantly altered expression levels in prostate cancer patients[73].
356
357 Using enrichment analyses (**Supplementary Table 8-9**), we found several pathways, including broadly
358 ubiquitous pathways to be significantly overrepresented in the selected genes for both the gene-components like
359 regulation of intracellular transport (adjusted p-value=0.017) and mRNA 3'-UTR binding (adjusted p-value =
360 0.008). Notably, we found the selected genes to be enriched for targets of several transcription factors many of
361 which have been associated with different types and subtypes of cancer (**Supplementary Table 10**). For example,
362 we found a TF target enrichment for *SPAG9* (adjusted p-value = 0.016) which has been identified to be associated
363 to breast cancer, ovarian cancer, colorectal cancer and others [74]. We also found enrichment for targets of *SSRP1*
364 (adjusted p-value = 0.016), which is differentially regulated in a wide spectrum of malignant tumors [75] along with
365 enrichment for targets of *MYC* and *TP73* which are well established cancer related genes (**Supplementary Table**
366 **10**). However, we did not find any evidence for significance enrichment of PPI among identified genes.
367
368 The downstream analysis suggests that a majority of the pathways (78.1%) enriched for the selected genes are
369 immune related as observed in the previous examples as well. This might have been driven by the fact that
370 eQTLGen reports summary *trans*-associations in whole blood. In general, whole blood might not be the ideal
371 candidate tissue to identify *trans*-associations pertaining to PC. It is conceivable that relevant tissue-specific
372 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. ARCHIE involves estimating LD and coexpression from publicly available reference datasets. However, we found that the results can be reasonably robust to the choice of reference data and estimation errors arising due to sampling variations (**Supplementary Figure 9**). In general, robust performance can be achieved under an optimal ratio of the sample sizes of the reference transcriptomic study and the study from which summary statistics of standard trans-eQTL associations are being analyzed. 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 potentially mediate the effects of multiple 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

397 broader trait-specific trans-regulation patterns that were detected by underlying gene and variant components.
398 The set of selected target genes in the gene component is one of the key outputs of ARCHIE. Using a series of
399 follow-up analyses for three different types of traits, we showed that the selected genes are often overrepresented
400 in known disease-relevant pathways, enriched in protein-protein interaction networks, shows co-regulations
401 across tissues and contains targets for known transcription factors implicated to the disease (SCZ and UC) and
402 key tumor suppressor genes (PC). Further, using a *trans*-expression imputation approach, we demonstrated that
403 the selected genes can significantly mediate heritability associated trait related variants. All of these analyses
404 point out that the *trans*-association patterns we detect are likely to have trait specific biological basis.

405

406 There are several limitations of the proposed method and current analysis. First, in the current version ARCHIE,
407 we begin with a set of genetic variants associated with a trait, but we do not incorporate the underlying association
408 directions and effect sizes in the analysis. This approach allowed us to independently investigate identified target
409 genes through testing for consistency of directions of association of the SNPs with the trait and those with the
410 expressions of the target through the *trans*-heritability analysis. However, it is likely that incorporation of the
411 GWAS effect sizes (value and direction) of trait association for the SNPs in the sCCA itself will lead to improved
412 power for detection of the trait specific target genes. One of the immediate ways to incorporating that might be
413 to weight the variants with weights proportional to squared (or absolute value) of the GWAS effect size. However,
414 incorporating the direction of the GWAS effect with the sCCA framework remains an interesting problem and
415 merits further research. Additionally, incorporation of information on cis-genes and known functional annotation
416 of genetic variants can improve the power of the analysis as well.

417

418 Although the estimation of the ARCHIE components is computationally efficient, in the current implementation,
419 the resampling-based testing method is computationally intensive. In the future, further research is merited to
420 develop analytical approximation techniques to reduce the computational burden of ARCHIE. Since variable

421 selection is a major goal in ARCHIE, we have introduced regularization via the L1 penalty due to its proven
422 theoretical selection consistency[76]. In fact, in presence of correlation between SNPs as well gene expressions,
423 elastic net regularization approach can be effective in accurate selection. In the future, usefulness of alternative
424 types of penalty functions in the selecting the genes and variants merits further research. Further, due to the lack
425 of existing methods to identify trait specific trans associations, we have compared the performance against the
426 standard trans-eQTL mapping. Although the eQTLGen data analysis shows that ARCHIE can identify a broader
427 range of gene-sets trans-regulated by GWAS variants as compared to standard trans-eQTL mapping, the goals of
428 ARCHIE and the standard analysis are different and therefore not directly comparable – the goal of the standard
429 analysis is to identify association in individual variant-gene pair irrespective of trait specificity. In contrast,
430 ARCHIE can identify trait-specific trans-regulated target genes harboring multiple weaker associations with trait-
431 related variants. As newer methods are developed to detect trait-relevant trans-regulated genes, more
432 comprehensive analyses comparing alternative approaches will allow us to understand their benefits and
433 limitations.

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

442 In this article we have analyzed summary statistics reported by eQTLGen in the whole blood. This is primarily
443 because of the substantial effective sample size of eQTLGen. While the approach can be applied to eQTL results
444 from other tissues, the underlying sample sizes may be too limited to yield sufficient power. Although blood
445 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 [77]. 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. Individual level data for the DGN cohort is available by application through the NIMH Center for Collaborative Genomic Studies on Mental Disorders. Instructions for requesting access to data can be found at https://www.nimhgenetics.org/access_data_biomaterial.php, and inquiries should reference the "Depression Genes and Networks study (D. Levinson, PI)".

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.

469 **URL**

470 eQTLGen (*trans*-eQTL summary statistics): https://www.eqtlgen.org/*trans*-eqtls.html

471 GTEx: https://www.gtexportal.org/home/

472 1000 Genomes: https://www.internationalgenome.org/data/

473 UKBiobank: https://www.ukbiobank.ac.uk/

474 FUMA: https://fuma.ctglab.nl/

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

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

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

478
479

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.* **7**, e1002003 (2011).
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* **10**, 515–534 (2009).
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. Rosa, M. J. *et al.* Estimating multivariate similarity between neuroimaging datasets with sparse canonical

531 correlation analysis: An application to perfusion imaging. *Front. Neurosci.* (2015).
532 doi:10.3389/fnins.2015.00366
533 26. Feng, H. *et al.* Leveraging expression from multiple tissues using sparse canonical correlation analysis
534 and aggregate tests improves the power of transcriptome-wide association studies. *PLOS Genet.* **17**,
535 e1008973 (2021).
536 27. Rodosthenous, T., Shahrezaei, V. & Evangelou, M. Integrating multi-OMICS data through sparse
537 canonical correlation analysis for the prediction of complex traits: a comparison study. *Bioinformatics*
538 **36**, 4616–4625 (2020).
539 28. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*
540 (2019). doi:10.1101/787903
541 29. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of
542 922 individuals. *Genome Res.* (2014). doi:10.1101/gr.155192.113
543 30. Volk, D. W., Chitrapu, A., Edelson, J. R. & Lewis, D. A. Chemokine receptors and cortical interneuron
544 dysfunction in schizophrenia. *Schizophr. Res.* (2015). doi:10.1016/j.schres.2014.10.031
545 31. Kassan, A. *et al.* Caveolin-1 regulation of disrupted-in-schizophrenia-1 as a potential therapeutic target
546 for schizophrenia. *J. Neurophysiol.* (2017). doi:10.1152/jn.00481.2016
547 32. Kakiuchi, C. *et al.* Functional polymorphisms of HSPA5: Possible association with bipolar disorder.
548 *Biochem. Biophys. Res. Commun.* (2005). doi:10.1016/j.bbrc.2005.08.248
549 33. Martin, J. *et al.* A brief report: de novo copy number variants in children with attention deficit
550 hyperactivity disorder. *Transl. Psychiatry* (2019). doi:10.1101/2019.12.12.19014555
551 34. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes
552 Associated with 30 Complex Traits. *Am. J. Hum. Genet.* (2017). doi:10.1016/j.ajhg.2017.01.031
553 35. Hill, D. P., Smith, B., McAndrews-Hill, M. S. & Blake, J. A. Gene Ontology annotations: What they
554 mean and where they come from. in *BMC Bioinformatics* (2008). doi:10.1186/1471-2105-9-S5-S2
555 36. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* (2000).
556 doi:10.1093/nar/28.1.27
557 37. Croft, D. *et al.* Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Res.*
558 (2011). doi:10.1093/nar/gkq1018
559 38. Schaefer, C. F. *et al.* PID: The pathway interaction database. *Nucleic Acids Res.* (2009).
560 doi:10.1093/nar/gkn653
561 39. Adriaens, M. E. *et al.* The public road to high-quality curated biological pathways. *Drug Discovery*
562 *Today* (2008). doi:10.1016/j.drudis.2008.06.013
563 40. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*
564 **1(6)**, 417–425 (2015).
565 41. Patel, S., Sharma, D., Kalia, K. & Tiwari, V. Crosstalk between endoplasmic reticulum stress and
566 oxidative stress in schizophrenia: The dawn of new therapeutic approaches. *Neuroscience and*
567 *Biobehavioral Reviews* (2017). doi:10.1016/j.neubiorev.2017.08.025
568 42. Landek-Salgado, M. A., Faust, T. E. & Sawa, A. Molecular substrates of schizophrenia: Homeostatic
569 signaling to connectivity. *Molecular Psychiatry* (2016). doi:10.1038/mp.2015.141
570 43. Chen, X. *et al.* Apoptotic engulfment pathway and schizophrenia. *PLoS One* (2009).
571 doi:10.1371/journal.pone.0006875
572 44. Liu, M.-L. *et al.* Severe disturbance of glucose metabolism in peripheral blood mononuclear cells of
573 schizophrenia patients: a targeted metabolomic study. *J. Transl. Med.* **13**, 226 (2015).
574 45. Cannon, T. D., Yolken, R., Buka, S. & Torrey, E. F. Decreased Neurotrophic Response to Birth Hypoxia
575 in the Etiology of Schizophrenia. *Biol. Psychiatry* **64**, 797–802 (2008).
576 46. Ryskalin, L., Limanaqi, F., Frati, A., Busceti, C. & Fornai, F. mTOR-Related Brain Dysfunctions in
577 Neuropsychiatric Disorders. *Int. J. Mol. Sci.* **19**, 2226 (2018).
578 47. Kim, P., Scott, M. R. & Meador-Woodruff, J. H. Dysregulation of the unfolded protein response (UPR)
579 in the dorsolateral prefrontal cortex in elderly patients with schizophrenia. *Mol. Psychiatry* (2019).
580 doi:10.1038/s41380-019-0537-7
581 48. Han, H. *et al.* TRRUST v2: An expanded reference database of human and mouse transcriptional

582 regulatory interactions. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1013
583 49. Zheng, G. *et al.* ITFP: An integrated platform of mammalian transcription factors. *Bioinformatics* (2008).
584 doi:10.1093/bioinformatics/btn439
585 50. Aberg, K. A. *et al.* Methylome-Wide Association Study of Schizophrenia. *JAMA Psychiatry* **71**, 255
586 (2014).
587 51. Martínez, G. *et al.* Regulation of Memory Formation by the Transcription Factor XBP1. *Cell Rep.* **14**,
588 1382–1394 (2016).
589 52. Khandaker, G. M. *et al.* Inflammation and immunity in schizophrenia: implications for pathophysiology
590 and treatment. *The Lancet Psychiatry* **2**, 258–270 (2015).
591 53. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage,
592 supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* (2019).
593 doi:10.1093/nar/gky1131
594 54. Rasetti, R. *et al.* Evidence That Altered Amygdala Activity in Schizophrenia Is Related to Clinical State
595 and Not Genetic Risk. *Am. J. Psychiatry* **166**, 216–225 (2009).
596 55. Tognin, S. *et al.* Enlarged hypothalamic volumes in schizophrenia. *Psychiatry Res. Neuroimaging* **204**,
597 75–81 (2012).
598 56. Small, S. A., Schobel, S. A., Buxton, R. B., Witter, M. P. & Barnes, C. A. A pathophysiological
599 framework of hippocampal dysfunction in ageing and disease. *Nat. Rev. Neurosci.* **12**, 585–601 (2011).
600 57. Bassotti, G. *et al.* Gastrointestinal motility disorders in inflammatory bowel diseases. *World J.*
601 *Gastroenterol.* (2014). doi:10.3748/wjg.v20.i1.37
602 58. Farrokhyar, F., Marshall, J. K., Easterbrook, B. & Irvine, E. J. Functional gastrointestinal disorders and
603 mood disorders in patients with inactive inflammatory bowel disease: Prevalence and impact on health.
604 *Inflamm. Bowel Dis.* (2006). doi:10.1097/01.MIB.0000195391.49762.89
605 59. Lakatos, P. L. & Lakatos, L. Risk for colorectal cancer in ulcerative colitis: Changes, causes and
606 management strategies. *World Journal of Gastroenterology* (2008). doi:10.3748/wjg.14.3937
607 60. Van Der Post, S. *et al.* Structural weakening of the colonic mucus barrier is an early event in ulcerative
608 colitis pathogenesis. *Gut* (2019). doi:10.1136/gutjnl-2018-317571
609 61. Rageul, J. *et al.* KLF4-dependent, PPARγ-induced expression of GPA33 in colon cancer cell lines. *Int. J.*
610 *Cancer* (2009). doi:10.1002/ijc.24683
611 62. Williams, B. B. *et al.* Glycoprotein A33 deficiency: A new mouse model of impaired intestinal epithelial
612 barrier function and inflammatory disease. *DMM Dis. Model. Mech.* (2015). doi:10.1242/dmm.019935
613 63. Gobert, A. P. *et al.* Distinct immunomodulatory effects of spermine oxidase in colitis induced by
614 epithelial injury or infection. *Front. Immunol.* (2018). doi:10.3389/fimmu.2018.01242
615 64. Hu, T. *et al.* Spermine oxidase is upregulated and promotes tumor growth in hepatocellular carcinoma.
616 *Hepatol. Res.* (2018). doi:10.1111/hepr.13206
617 65. Sadlack, B. *et al.* Ulcerative colitis-like disease in mice with a disrupted interleukin-2 gene. *Cell* **75**, 253–
618 261 (1993).
619 66. Hayden, M. S. & Ghosh, S. NF-κB in immunobiology. *Cell Research* (2011). doi:10.1038/cr.2011.13
620 67. Liu, T., Zhang, L., Joo, D. & Sun, S. C. NF-κB signaling in inflammation. *Signal Transduction and*
621 *Targeted Therapy* (2017). doi:10.1038/sigtrans.2017.23
622 68. Borm, M. E. A., Van Bodegraven, A. A., Mulder, C. J. J., Kraal, G. & Bouma, G. A NFKB1 promoter
623 polymorphism is involved in susceptibility to ulcerative colitis. *Int. J. Immunogenet.* (2005).
624 doi:10.1111/j.1744-313X.2005.00546.x
625 69. Wang, X. & Sun, Q. TP53 mutations, expression and interaction networks in human cancers. *Oncotarget*
626 (2017). doi:10.18632/oncotarget.13483
627 70. Millet, C. & Zhang, Y. E. Roles of Smad3 in TGF-β signaling during carcinogenesis. *Critical Reviews in*
628 *Eukaryotic Gene Expression* (2007). doi:10.1615/CritRevEukarGeneExpr.v17.i4.30
629 71. Tang, P. M.-K. *et al.* Smad3 promotes cancer progression by inhibiting E4BP4-mediated NK cell
630 development. *Nat. Commun.* **8**, 14677 (2017).
631 72. Lu, S., Lee, J., Revelo, M., Wang, X. & Dong, Z. Smad3 is overexpressed in advanced human prostate
632 cancer and necessary for progressive growth of prostate cancer cells in nude mice. *Clin. Cancer Res.*

633 (2007). doi:10.1158/1078-0432.CCR-07-1078
634 73. Johnson, I. R. D. *et al.* Endosomal gene expression: a new indicator for prostate cancer patient
635 prognosis? *Oncotarget* **6**, 37919–37929 (2015).
636 74. Kanojia, D., Garg, M., Gupta, S., Gupta, A. & Suri, A. Sperm-Associated Antigen 9 Is a Novel
637 Biomarker for Colorectal Cancer and Is Involved in Tumor Growth and Tumorigenicity. *Am. J. Pathol.*
638 **178**, 1009–1020 (2011).
639 75. Garcia, H. *et al.* Facilitates Chromatin Transcription Complex Is an "Accelerator" of Tumor
640 Transformation and Potential Marker and Target of Aggressive Cancers. *Cell Rep.* **4**, 159–173 (2013).
641 76. Zhao, P. & Yu, B. On model selection consistency of Lasso. *J. Mach. Learn. Res.* (2006).
642 77. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet.* 1–16
643 (2020). doi:10.1016/j.tig.2020.08.009
644 78. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–
645 209 (2018).
646 79. Schäfer, J. & Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and
647 implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* (2005). doi:10.2202/1544-6115.1175
648 80. Saha, A. & Battle, A. False positives in trans-eQTL and co-expression analyses arising from RNA-
649 sequencing alignment errors. *F1000Research* (2018). doi:10.12688/f1000research.17145.1
650

651 # Methods

652 ## Sample Description

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

660 **GTEx.** The Genotype-Tissue Expression (GTEx) project[28] aims to study tissue-specific gene expression and
661 regulation. We used individual level data from GTEx (v8) whole blood to construct the co-expression matrix
662 (Σ_{EE}) and further downstream validation of the gene-sets selected using ARCHIE. In our analysis, we used the
663 latest version (v8) of GTEx having gene-expression and genotype data with samples from 54 different tissues. In
664 particular, 755 individuals had expression data on 20,315 genes for whole blood. Of these, we used 670
665 individuals with genotype data present.

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

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

675

676 # Methods

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

678 Our proposed method, Aggregative *trans* association to detect phenotype specific gene-sets (ARCHIE), can select
679 target genes *trans*-associated with trait-related variants using summary statistics in a sparse canonical correlation
680 framework. To introduce the details of ARCHIE, first we briefly describe sparse canonical correlation analysis
681 (sCCA). For *n* individuals, let $\boldsymbol{G^{n\times p}}$ be the normalized genotype matrix for *p* variants and $\boldsymbol{E^{n\times g}}$ be the normalized
682 gene-expression matrix for *g* genes all of which are distant (*trans*) to the variants. sCCA seeks to estimate sparse
683 linear combinations of variants ($\boldsymbol{u^{p\times 1}}$) and genes ($\boldsymbol{v^{g\times 1}}$) such that the correlation between $\boldsymbol{Gu}$ and $\boldsymbol{Ev}$ is
684 maximized i.e.,

685 $$(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{argmax}\, \tilde{\boldsymbol{v}}^{\boldsymbol{T}}\boldsymbol{E}^{\boldsymbol{T}}\boldsymbol{G}\tilde{\boldsymbol{u}}$$

686 $$\text{with } \boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{E}^{\boldsymbol{T}}\boldsymbol{Ev} \le \boldsymbol{1} \text{ and } \boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{G}^{\boldsymbol{T}}\boldsymbol{Gu} \le \boldsymbol{1} \text{ and } ||\tilde{\boldsymbol{u}}||_{\boldsymbol{1}} \le \boldsymbol{c_u}\ ;||\tilde{\boldsymbol{v}}||_{\boldsymbol{1}} \le \boldsymbol{c_v}$$

687 where $||\boldsymbol{x}||_{\boldsymbol{h}}$ is the L_h norm of a vector *x*; $\boldsymbol{c_u}$ (or $\boldsymbol{c_v}$) is the sparsity parameter on the variant (or gene) component
688 for the lasso-type L_1 penalty. The subsequent pairs of sCCA components are obtained by similarly maximizing
689 the correlation between $\boldsymbol{Gu}$ and $\boldsymbol{Ev}$ and under the constraint of being uncorrelated or orthogonal to the previous
690 components. In practice, the optimization problem can be reformulated in terms of covariance matrices as:

691 $$(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{argmax}\, \tilde{\boldsymbol{v}}^{\boldsymbol{T}}\boldsymbol{W}\tilde{\boldsymbol{u}}$$

692 $$\text{with } ||\tilde{\boldsymbol{u}}||_{\boldsymbol{1}} \le \boldsymbol{c_u}\ ;||\tilde{\boldsymbol{v}}||_{\boldsymbol{1}} \le \boldsymbol{c_v} \text{ and } ||\tilde{\boldsymbol{u}}||_{\boldsymbol{2}} = \boldsymbol{1}, ||\tilde{\boldsymbol{v}}||_{\boldsymbol{2}} = \boldsymbol{1}$$

693 where $\boldsymbol{W} = \boldsymbol{\Sigma}_{\boldsymbol{EE}}^{-\boldsymbol{1/2}}\boldsymbol{\Sigma}_{\boldsymbol{GE}}\boldsymbol{\Sigma}_{\boldsymbol{GG}}^{-\boldsymbol{1/2}}$ and $\boldsymbol{v_k} = \boldsymbol{Wu}$; $\boldsymbol{\Sigma}_{\boldsymbol{GG}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{EE}}$ are the column-correlations of $\boldsymbol{G}$ and $\boldsymbol{E}$ respectively
694 and $\boldsymbol{\Sigma}_{\boldsymbol{GE}}$ is the cross-covariance matrix between the variants and the gene-expressions. However, the matrices can
695 be estimated from publicly available summary statistics and reference data. We approximate $\boldsymbol{\Sigma}_{\boldsymbol{GG}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{EE}}$ by the
696 empirical LD-matrix and a penalized co-expression matrix (See Supplementary Section A3) using external
697 reference samples and $\boldsymbol{\Sigma}_{\boldsymbol{GE}}$ can be obtained from the summary statistics of the regression for trans-eQTL mapping
698 across all pairs of variants and gene-expressions (See Supplementary Section A3).

To apply ARCHIE, thus, 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, estimated from reference panels with individual level genotype (or dosage) and gene expression data respectively, as follows:

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

where Σ_{GG} and Σ_{EE} are estimates of LD-matrix and co-expression matrix (see **Supplementary Section A**), and Σ_{GE} is the cross-correlation matrix obtained 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

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

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

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

723 components (jointly termed ARCHIE components) can be interpreted as the sparse latent factors that explain the
724 majority of the aggregated association between all the trait-related variants and all the genes. The corresponding
725 sparse canonical correlation (cc-value) between each pair of variant and gene components, defined as $q^2 =$
726 $\frac{(v^T W u)^2}{\sqrt{(u^T W^T W u)(v^T W W^T v)}}$ would be a measure of the cumulative association between the selected sets of variants and
727 genes by aggregating multiple (possibly weaker) associations (**Figure 1A** shows an illustration using *P* variants
728 and *G* genes). Multiple such components (u, v) can be extracted to reflect approximately orthogonal latent factors
729 of the aggregative correlation, corresponding to possibly distinct mechanisms of trans-regulation (See
730 **Supplementary Section A**).
731 At suitable levels of sparsity (See **Supplementary Section A2**), ARCHIE components produce a much smaller
732 number of selected target genes which harbor multiple moderate to weak *trans*-association from a selected set of
733 trait-associated variants, thus reflecting a trait-specific pattern of *trans*-association. A detailed algorithm for the
734 estimation of the ARCHIE components is provided in the **Supplementary Section A1**. We can interpret the
735 ARCHIE output as the subset of genes (selected by the gene component) having possibly multiple weaker trans-
736 associations to the subset of variants (selected by the variant component).
737

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

740 To test which ARCHIE components significantly capture the phenotype-specific *trans*-association pattern we
741 evaluated the results from the original analysis against a *competitive* null hypothesis. Since trait-related variants
742 are expected to be enriched for *trans*-eQTLs in general, we test whether the cc-values obtained in the original
743 analysis are higher than that obtained using the *trans*-summary statistics between a random set of GWAS-
744 identified variants and genes of similar size, that do not reflect any trait-specific pattern. For this, we first construct
745 a *null matrix* by taking a random sample of *p* variants from the pool of all variants available and extracting the
746 corresponding *trans*-summary statistics for another set of randomly chosen *g* genes. Since eQTLGen reports the
747 *trans*-summary statistics across about 10,000 variants associated with different traits, we can construct the *null*

748 *matrix* using the *trans*-summary statistics from these variants that are associated with different traits and not with
749 the trait of interest. This matrix of *trans*-associations, by design, should not reflect phenotype-specific patterns.
750 For example, in the analysis for Schizophrenia (SCZ) using summary statistics across 218 variants and 7,047
751 genes, we construct the null matrix using 1 variant selected at random from 218 randomly chosen traits and
752 extracting their corresponding *trans*-summary statistics across 7,047 randomly chosen genes.
753
754 Then we use ARCHIE with the same sparsity levels as the original analysis, to extract the gene and variant
755 components and calculate corresponding cc-values. We repeat this step multiple (M) times to generate a
756 *competitive* null distribution of cc-values. We then evaluate the observed cc-values from the original analysis
757 against the corresponding *competitive* null distributions to calculate the p-value. In particular, the p-value of the
758 k^{th} ARCHIE component is given as

759 $$p_k = \frac{\sum_{i=1}^{M} I(q_k^2 > q_{k;null(i)}^2)}{M}$$

760 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
761 of the kth cc-value. We declare that the top L components significantly capture phenotype-specific *trans*-
762 association patterns if

763 $$L = min\{k: p_k > \alpha; k = 1,2, \dots, \min(p, g)\} - 1$$

764 The random set of p variants should be carefully chosen so that none of the variants associated to the phenotype
765 in consideration or any phenotype sharing substantial genetic correlation, are included. Further the set should be
766 such that it does not include a large fraction of the variants from the same phenotype (different from the original
767 phenotype), which may bias the *competitive* null distribution towards the *trans*-association cc-values for that
768 phenotype.
769

770 **Simulation Model**

771 To demonstrate that ARCHIE can identify downstream trans-associations, we simulate individual level gene-
772 expression data for N individuals (N=1,000 or 30,000). First, we randomly sampled N unrelated individuals with

genotypes at 50 independent and randomly selected SNPs from the UK Biobank with minor allele frequencies ranging from 10%-40%. The SNPs were then arranged in sets of 5 each. Each set of 5 SNPs were then used to simulate the cis-gene expression of 8 genes in total marked in red (Figure 2A-C). For type-I error simulations (Fig 2A) the expressions for genes 1-4 were simulated from a univariate standard gaussian distribution and the subsequent downstream genes 5-9 were simulated using the causal regulatory model. The gene-gene regulatory causal effects on genes 5-9 were chosen such that the approximately 20-30% of the variation in gene expression was explained by causal upstream genes.

For power simulations (Figure 2B, C), the direct cis-regulatory effects of the 5 SNPs on the corresponding cis-gene expression were chosen from a gaussian distribution such that the average cis-heritability explained by the SNPs were maintained at approximately 20-22%. Then we simulated gene-expression for genes 1 through 9 using a causal regulatory model as shown in Figure 2B or C, using the simulated cis-gene expressions. The regulatory effects between the gene expressions were chosen such that the total variance explained for the expression of Genes 1-9 by indirect effect of the SNPs (trans-heritability) were approximately 10-14%. For the global null scenario, the downstream genes 1-9 were simulated independent from the cis genes (Figure 2A) using the causal regulatory model. We further simulated a separate sample of 700 ($\mathbf{N}_{\boldsymbol{Ref}}$) from which the gene expressions were used to estimate the coexpression matrix ($\boldsymbol{\Sigma}_{\boldsymbol{EE}}$). The $\boldsymbol{\Sigma}_{\boldsymbol{GG}}$ matrix was taken to be a diagonal matrix with the diagonal elements being the estimated sample variance of the SNPs. Using this simulated data, we applied ARCHIE and compared the results to that obtained from standard trans-eQTL mapping. To test against the global null hypothesis, that there was no trans-association between the SNPs and any gene in the gene network, we simulated null $\boldsymbol{\Sigma}_{\boldsymbol{GE;null}}$ as $\mathbf{vec}(\boldsymbol{\Sigma}_{\boldsymbol{GE;null}}) \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{EE}} \otimes \boldsymbol{\Sigma}_{\boldsymbol{GG}})$ where vec(.) represents the stacked row vectorization of a matrix and $\otimes$ represents the Kronecker product. We simulated applied ARCHIE to the $\boldsymbol{\Sigma}_{\boldsymbol{GE;null}}$ multiple times to obtain a distribution of the sparse canonical correlation scores (q^2). Any ARCHIE component with p-value less than the chosen level was declared significant and the corresponding selected genes were noted (See Supplementary Section B for more details).

798 Next, through resampling experiments we assess whether ARCHIE can potentially identify trait specific trans-
799 associations. Using data from eQTLGen consortium, we construct a matrix of trans-eQTL summary statistics (Z-
800 values) across for p variants and g genes. Out of the p variants, we set δ proportion of them to be related to a
801 particular trait. For high values of δ we expect that the trans-summary statistics matrix would reflect trans-
802 association patterns pertaining to the trait and hence should be captured by the ARCHIE components. We then
803 applied ARCHIE on this matrix and evaluate the significance of the components. We replicate this experiment
804 multiple times in a given setting, to estimate the empirical probability of at least one ARCHIE component to be
805 significant (See **Supplementary Section B** for details) and reported the empirical probability of at least one
806 ARCHIE component to be significant across varying values of δ. We repeated this resampling experiment with
807 four different traits (**Supplementary Figure 4**).

808

809 **Analysis of eQTLGen data**

810 To identify phenotype-specific *trans*-associations, we applied ARCHIE on the *trans*-association summary
811 statistics for 10,317 trait-related variants across 19,942 genes reported by the eQTLGen consortium[19] (See Sample
812 Description for details on the study). In line with the consortium, we defined any gene to be trans to a variant if
813 the variant was located at least 5Mb from the transcription start site of the gene or on another chromosome. The
814 data contains multiple variants associated with the same trait analyzed for *trans*-eQTL mapping. Our analysis was
815 restricted to phenotypes that had at least 100 associated variants tested for *trans*-mapping in the consortium,
816 producing 29 phenotypes. **Figure 1B** shows a graphical representation of the major steps of our workflow. Briefly,
817 for each phenotype, we extracted the summary *trans*-eQTL association statistics (Z-score, p-value) and removed
818 all genes that were in within 5Mb of any of the trait related variants. In the preprocessing step, we filtered for any
819 missing data and retained the genes that were also expressed in GTEx (v8)[28] whole blood. This produced a list of
820 approximately 129 (min: 112; max: 533) variants and 10,219 (min: 3426; max: 13910) genes on an average per
821 phenotype. ARCHIE requires two additional matrices representing the correlation among the variants themselves
822 (a linkage-disequilibrium matrix) and among the gene-expression levels (a co-expression matrix), which can be

823 estimated using reference data. We constructed the LD-matrix for the variants from individual-level genotype
824 information using 5,000 randomly selected, unrelated European samples in UK Biobank[78]. For the correlation
825 between gene-expressions, we used a penalized co-expression matrix[79] of the corresponding genes constructed
826 from the covariate-adjusted quantile normalized gene-expression levels for individuals in GTEx v8 data.
827 Subsequently, for the given trait, we extracted the selected variants and genes using the significant components
828 and were evaluated for presence of false-positives due to cross-mapping.
829
830 **Cross-Mappability:** Alignment errors due to similarity in sequenced reads can lead a substantial rise in false
831 positives for detecting *trans*-eQTL associations [80]. With the selected ARCHIE components, we extracted the
832 nearby genes expressed in GTEx v8 whole blood for the selected variants (TSS within ±500 kb of the variant)
833 and evaluated the cross-mapping scores for these genes with the selected target genes. Across the 3 traits analyzed
834 in this article, we found that all such gene-pairs were mostly non cross-mappable (SCZ: 99.98%, UC: 99.17%,
835 PC: 99.93%), indicating that the *trans*-association patterns were less likely to be affected by false positive arising
836 from alignment errors.
837
838
839 **Follow-up Analysis.**
840 **Quantifying and Testing for Enrichment for Trait Heritability Explained by Identified Target Genes.** In
841 the following, we propose a method for quantifying trait heritability explained by the GWAS variants that would
842 be mediated by the identified target-genes and develop a corresponding test for enrichment through comparison
843 of such estimates of mediated heritability associated with that from random genes. For or a particular trait of
844 interest, we start with the Z-scores for regression-based *trans*-eQTL mapping for a set of underlying p variants
845 and g genes. We will assume that, using ARCHIE, we have identified G target genes that capture trait-specific
846 *trans*-association patterns. To perform the test as proposed above, we require individual level phenotype and
847 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

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

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**.

Tables and Figures

Figure 1. Overview of ARCHIE and comparison to standard trans-eQTL mapping. (A) An illustrative example for ARCHIE. Association statistics (-$\log_{10}$ p-value) from trans-eQTL mapping for *P* variants and *G* genes are shown in the left panel heatmap. Using LD and co-expression estimated from reference datasets, ARCHIE aggregates multiple weaker trans-eQTL associations to select a subset of variants and genes which capture trait specific trans-association patterns (right panel heatmap). (B) Conceptual pipeline for analysis of summary statistics of trans-eQTL mapping provided by the eQTLGen consortium by integrating external estimates from GTEx (v8) and UK Biobank. See Results and Methods for more details.

Figure 2. Simulation model and results. (A) causal network model for global null scenario with no association between the SNPs 1-40 and the Genes 1-9 (marked in blue). See Results and Methods: Simulation Model for more details. (B) Sparse causal network: causal regulatory network of downstream Genes 1-9 (in blue) regulated by SNPs 1-40 mediated via 8 local (cis) genes (marked in red). (C) Dense causal network: causal regulatory network of downstream Genes 1-9 (in blue) regulated by SNPs 1-40 mediated via 8 local (cis) genes (marked in red). For both (B) and (C) the heritability of cis gene expressions explained was maintained at 20-22% and the heritability of downstream genes (1-9) explained was maintained at 10-14%. (D) Empirical power to identify different gene sets within sparse causal regulatory network (2B) using ARCHIE compared to standard trans-eQTL mapping. (E) Empirical power to identify different gene sets within dense causal regulatory network (2C) using ARCHIE compared to standard trans-eQTL mapping.

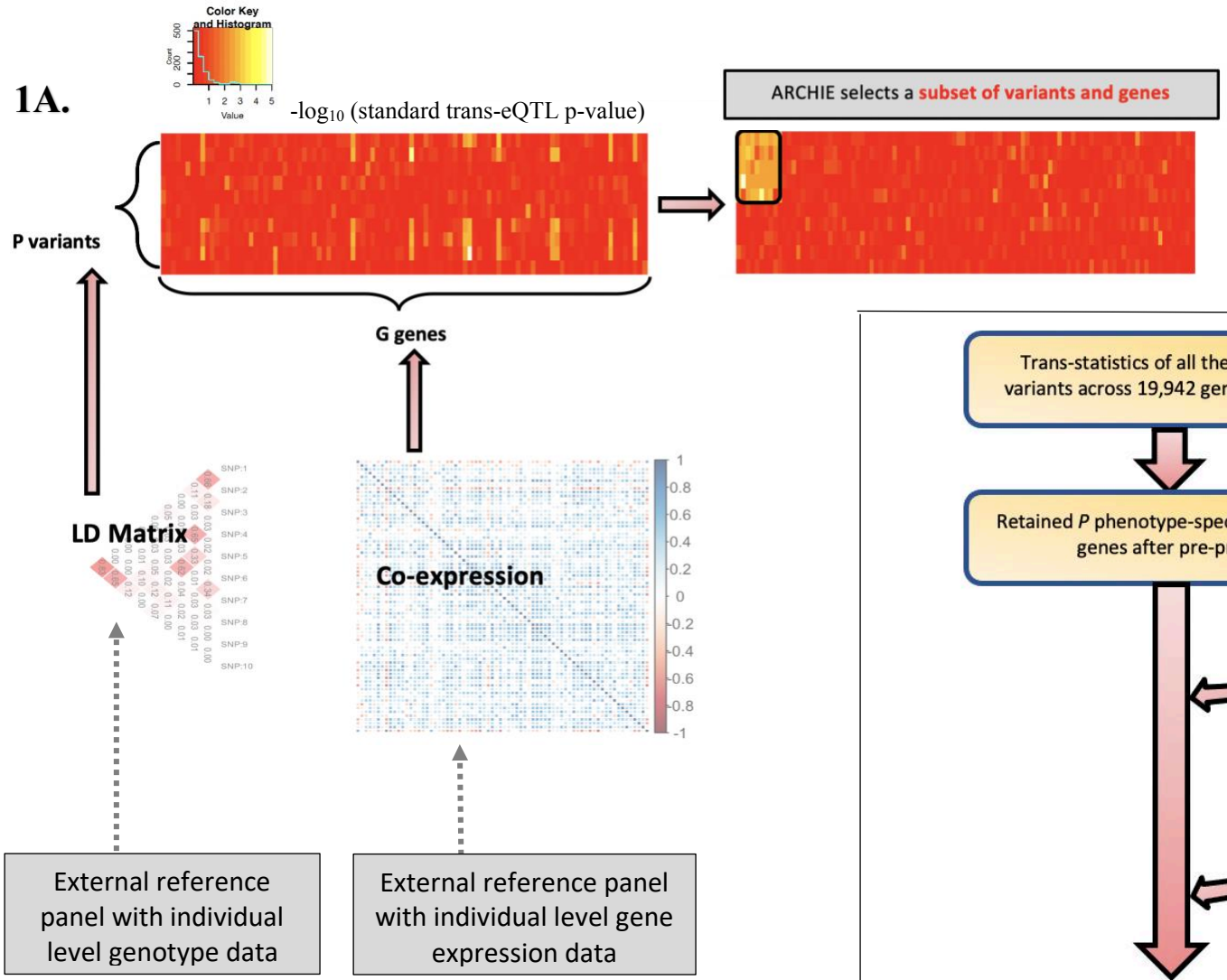
Figure 3. Trans-association pattern and properties of selected target gene-set for SCZ. (A) Top 10 sparse canonical correlation values (cc-values; red) and the corresponding *competitive* null distributions (black box plot) for ARCHIE analysis of SCZ. Top 1 ARCHIE component is significant. The competitive null distribution denotes the general pattern of sparse canonical correlation (and trans-associations) expected in analysis of variants that are not specific to SCZ. The significant ARCHIE

components are marked by a blue asterisk. (B) Reported $-\log_{10}$(p-values) for trans-eQTL association between variants and genes selected in ARCHIE component 1. Any association p-value $< 10^{-08}$ is collapsed to 10^{-08} for the ease of viewing. (C) Pathway enrichment results for some selected top pathways for the target genes selected in gene-component 1. For full pathway enrichment results for the target genes selected see **Supplementary Tables 3-4**. The dashed vertical line corresponds to a suggested FDR threshold of 0.001. (D) trans-heritability enrichment analysis for target genes selected for SCZ. The violin plots represent the distribution of pseudo-r^2 for the selected SNPs and a random gene-set of the same size on SCZ (left) and the distribution of pseudo-r^2 of the selected genes and SNPs on a random trait in UK Biobank (right). The red dashed line represents the observed pseudo-r^2 for the selected SNPs and target genes for SCZ (See **Results** and **Methods** for details).

Figure 4. Analysis of selected target gene-set for UC. (A) Top 10 cc-values (red) and the corresponding *competitive* null distributions (black box plot) for ARCHIE analysis of UC. Top 2 ARCHIE components are significant. The competitive null distribution denotes the general pattern of sparse canonical correlation (and trans-associations) expected in analysis of variants that are not specific to UC. The significant ARCHIE components are marked by a blue asterisk. (B) Pathway enrichment results for some selected top pathways for the selected target genes. The dashed vertical line corresponds to a suggested FDR threshold of 0.001. For full pathway enrichment results for the target genes selected for both the components see **Supplementary Tables 5**-7. (C) PPI enrichment for the two significant ARCHIE components for UC. (D) trans-heritability enrichment analysis for target genes selected for UC. The violin plots represent the distribution of pseudo-r^2 for the selected SNPs and a random gene-set of the same size on UC (left) and the distribution of pseudo-r^2 of the selected genes and SNPs on a random trait in UK Biobank (right). The red dashed line represents the observed pseudo-r^2 for the selected SNPs and target genes for UC (See **Results** and **Methods** for details).

Figure 5. Analysis of selected target gene-set for PC. (A) Top 10 cc-values (red) and the corresponding competitive null distributions (black box plot) for ARCHIE analysis of PC. Top 2 ARCHIE components are significant. The competitive null distribution denotes the general pattern of sparse canonical correlation (and trans-associations) expected in analysis of variants that are not specific to PC. The significant ARCHIE components are marked by a blue asterisk. (B) Reported $-\log_{10}$ p-value for trans-eQTL association between variants and genes selected in (B) ARCHIE component 1 and (C) component 2. *TP53* and *SMAD3* are highlighted. Any association p-value $< 10^{-08}$ is collapsed to 10^{-08} for the ease of viewing. (D) trans-heritability enrichment analysis for target genes selected for PC. The violin plots represent the distribution of pseudo-r^2 for the selected SNPs and a random gene-set of the same size on PC (left) and the distribution of pseudo-r^2 of the selected genes and SNPs on a random trait in UK Biobank (right). The red dashed line represents the observed pseudo-r^2 for the selected SNPs and target genes for PC (See **Results** and **Methods** for details).

Figure 1


1A.
Color Key and Histogram
Count
Value
-log10 (standard trans-eQTL p-value)
ARCHIE selects a subset of variants and genes
P variants
G genes
LD Matrix
Co-expression
External reference panel with individual level genotype data
External reference panel with individual level gene expression data

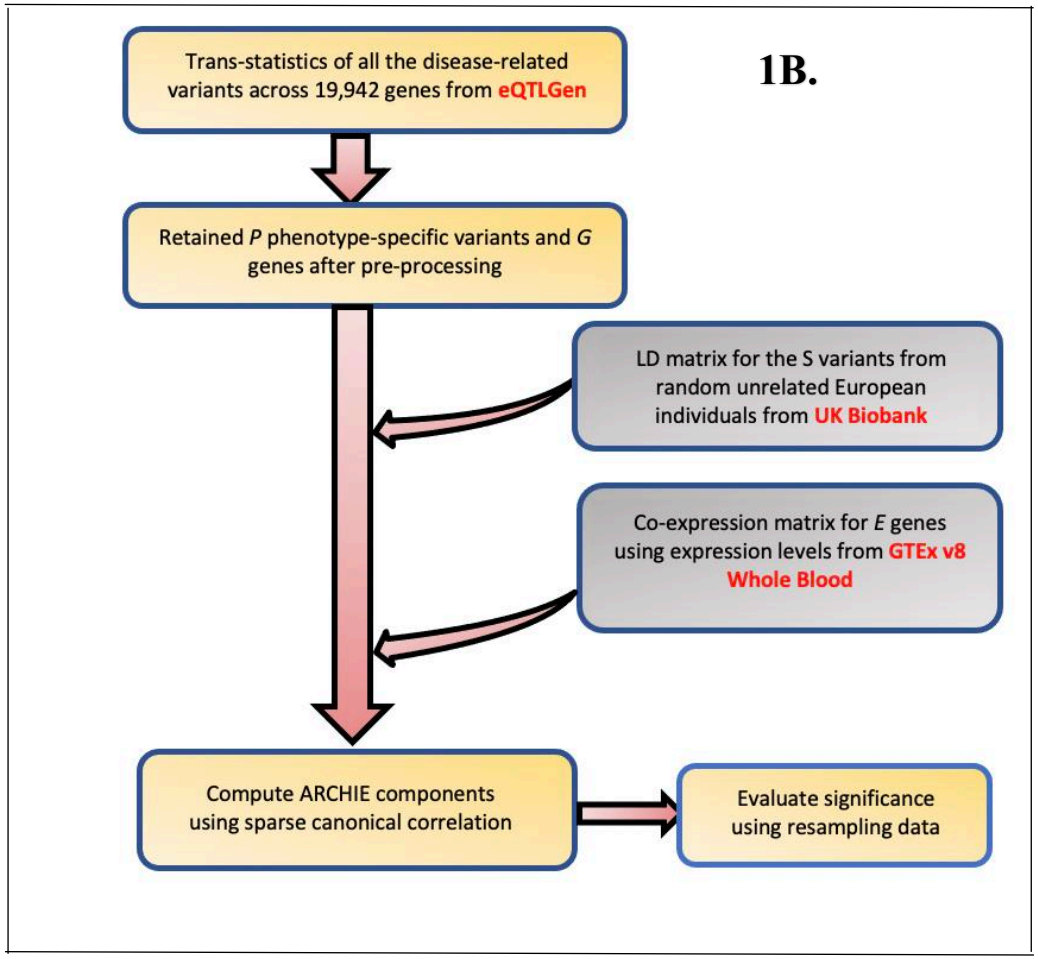

1B.
Trans-statistics of all the disease-related variants across 19,942 genes from eQTLGen
Retained P phenotype-specific variants and G genes after pre-processing
LD matrix for the S variants from random unrelated European individuals from UK Biobank
Co-expression matrix for E genes using expression levels from GTEx v8 Whole Blood
Compute ARCHIE components using sparse canonical correlation
Evaluate significance using resampling data

Figure 2

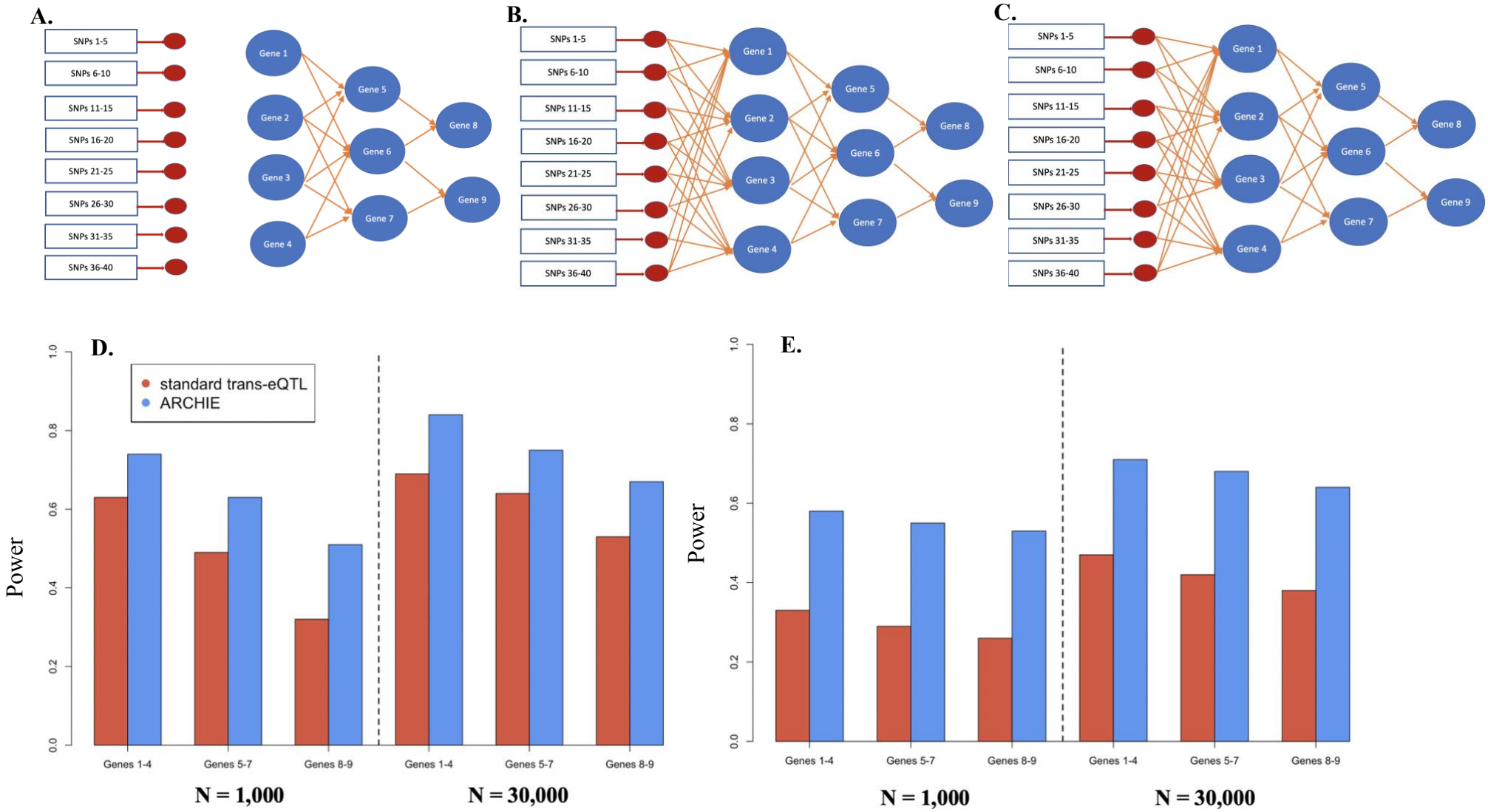
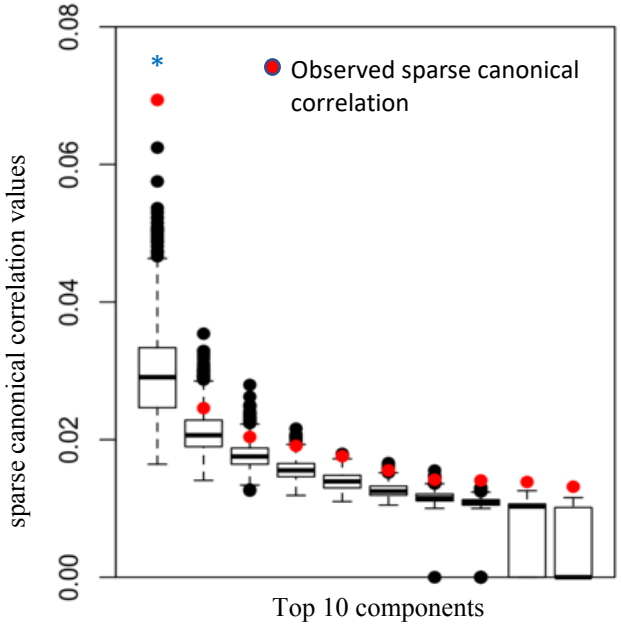
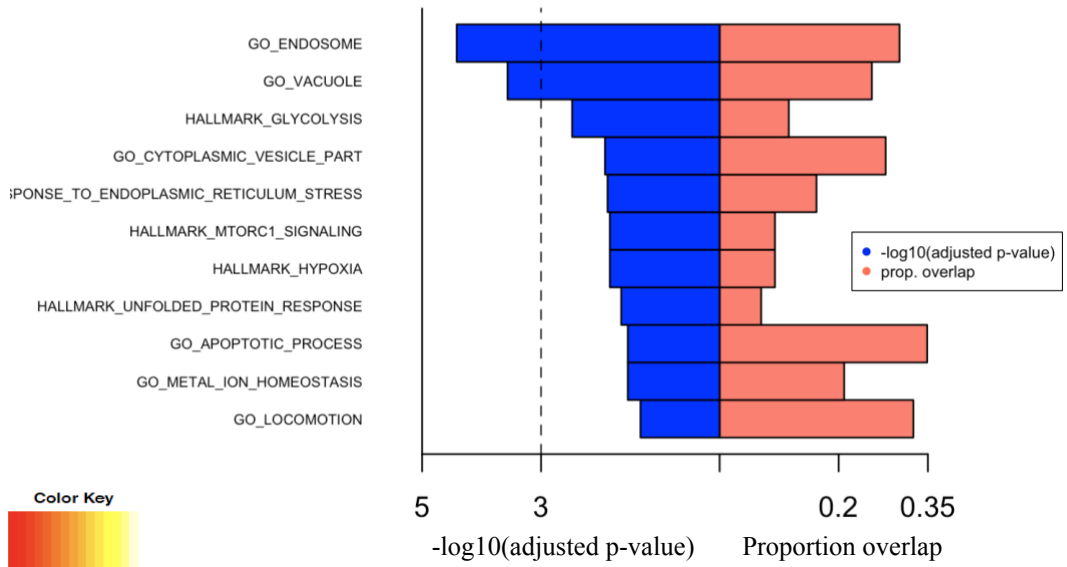

A.
B.
C.
SNPs 1-5
SNPs 6-10
SNPs 11-15
SNPs 16-20
SNPs 21-25
SNPs 26-30
SNPs 31-35
SNPs 36-40
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
Gene 7
Gene 8
Gene 9
D.
E.
standard trans-eQTL
ARCHIE
Power
0.0
0.2
0.4
0.6
0.8
1.0
Genes 1-4
Genes 5-7
Genes 8-9
N = 1,000
N = 30,000

Figure 3

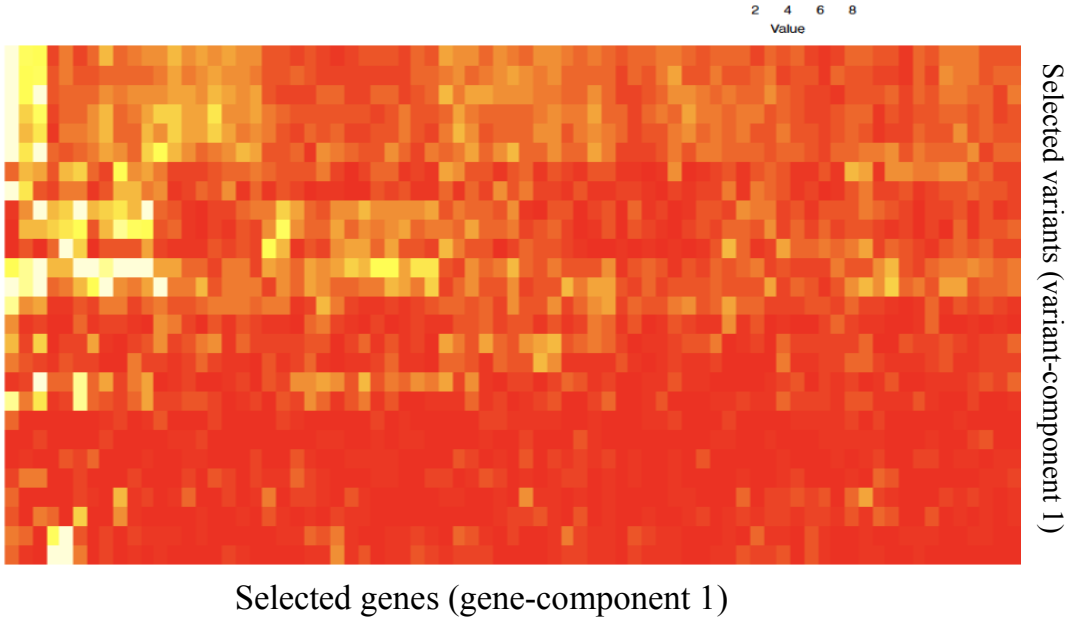
3A.



3C. Pathway enrichment analysis (SCZ)



3B.



3D.

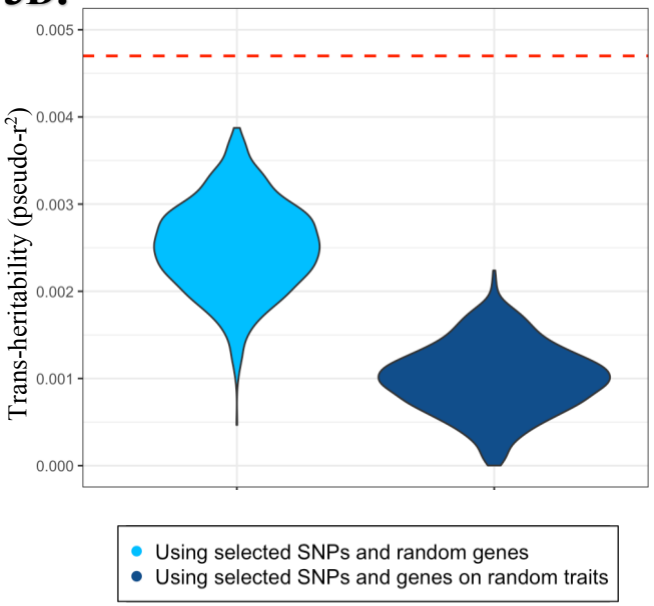


Figure 4

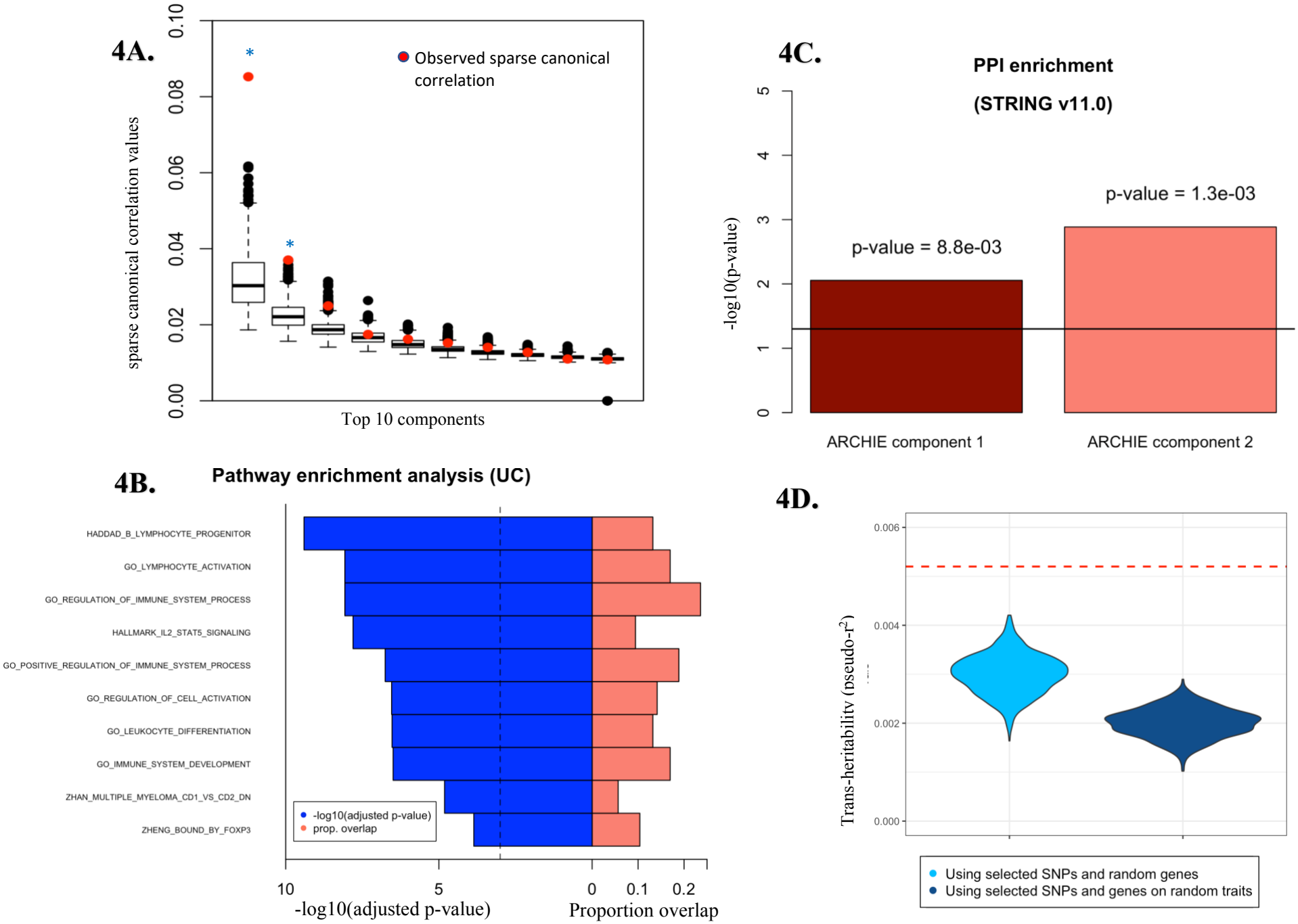

4A.
Observed sparse canonical correlation
sparse canonical correlation values
0.00
0.02
0.04
0.06
0.08
0.10
Top 10 components
4C.
PPI enrichment
(STRING v11.0)
-log10(p-value)
0
1
2
3
4
5
p-value = 8.8e-03
p-value = 1.3e-03
ARCHIE component 1
ARCHIE ccomponent 2
4B.
Pathway enrichment analysis (UC)
HADDAD_B_LYMPHOCYTE_PROGENITOR
GO_LYMPHOCYTE_ACTIVATION
GO_REGULATION_OF_IMMUNE_SYSTEM_PROCESS
HALLMARK_IL2_STAT5_SIGNALING
GO_POSITIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS
GO_REGULATION_OF_CELL_ACTIVATION
GO_LEUKOCYTE_DIFFERENTIATION
GO_IMMUNE_SYSTEM_DEVELOPMENT
ZHAN_MULTIPLE_MYELOMA_CD1_VS_CD2_DN
ZHENG_BOUND_BY_FOXP3
-log10(adjusted p-value)
prop. overlap
10
5
0
0.1
0.2
-log10(adjusted p-value)
Proportion overlap
4D.
Trans-heritability (pseudo-r²)
0.000
0.002
0.004
0.006
Using selected SNPs and random genes
Using selected SNPs and genes on random traits

Figure 5

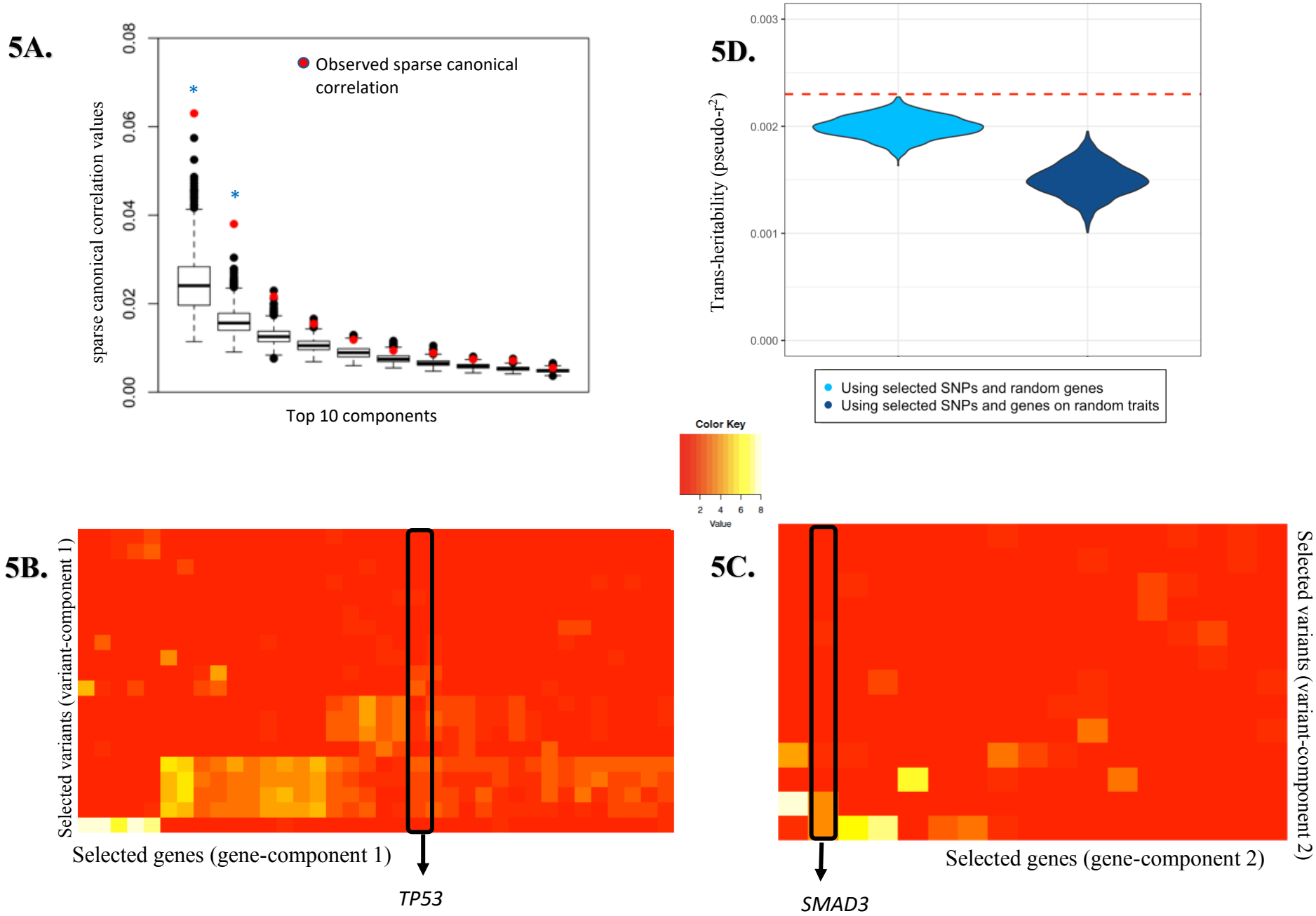

5A.
Observed sparse canonical correlation
sparse canonical correlation values
0.00
0.02
0.04
0.06
0.08
Top 10 components
5D.
Trans-heritability (pseudo-r²)
0.000
0.001
0.002
0.003
Using selected SNPs and random genes
Using selected SNPs and genes on random traits
Color Key
2
4
6
8
Value
5B.
Selected variants (variant-component 1)
Selected genes (gene-component 1)
TP53
5C.
Selected variants (variant-component 2)
Selected genes (gene-component 2)
SMAD3

Table 1: Type-I error of ARCHIE under the global null hypothesis of no association between the SNPs and gene network as shown in Figure 2A. See Methods and Supplementary Section B for details on the simulation model.

Hypothesis		Level			
	Method	1×10^{-03}	1×10^{-04}	1×10^{-05}	1×10^{-06}
At least one variant-gene association significant	Standard trans-eQTL	0.15	0.044	0.004	2.1×10^{-04}
At least one ARCHIE component significant	ARCHIE	1.8×10^{-04}	9.1×10^{-06}	3.9×10^{-07}	2.5×10^{-07}

Table 2: The number of variants, genes and novel genes selected by ARCHIE for each the significant components for the analysis of SCZ, UC and PC. For a full list of selected variants and genes see **Supplementary Table 2**.

	Schizophrenia (SCZ)	**Ulcerative Colitis (UC)**		**Prostate Cancer (PC)**	
	Component 1	**Component 1**	**Component 2**	**Component 1**	**Component 2**
Variants	27	41	33	20	13
Genes	75	106	42	36	17
Novel genes (proportion %)	59 (78.7%)	56 (52.8%)	16 (38.1%)	31 (86.1%)	13 (76.5%)

Table 3: Pathway enrichment results for the target genes selected for SCZ. Several selected top pathways containing novel genes across different categories are shown. See **Supplementary Tables 3 and 4** for results on all significant pathways and transcription factor target gene-sets for the selected genes.

Category	Pathway	Adjusted p-value	Genes in pathway	Genes overlap	Novel genes overlap
GO	Endosome (GO: 0005768)	3.8×10^{-05}	885	13	11
	Vacuole (GO: 0005773)	2.7×10^{-04}	760	11	9
	Response to endoplasmic reticulum stress (GO: 0034976)	1.3×10^{-02}	272	7	6
Hallmark	Glycolysis	3.3×10^{-03}	200	5	4
	Hypoxia	1.4×10^{-02}	200	4	3
	MTORC1 signaling	1.4×10^{-02}	200	4	3
Curated Gene sets	Wierenga_STAT5A_TARGETS_DN	5.9×10^{-03}	189	6	5
	Zhou_INFLAMMATORY_RESPONSE_LIVE_DN	6.5×10^{-03}	372	7	3
Oncogenic Signatures	TBK1. DN.48HRS_DN	3.1×10^{-02}	50	3	3
	ATF2_S_UP. V1_DN	4.4×10^{-02}	185	4	4
Immunologic Signatures	GSE6269_E_COLI_VS_STREP_PNEUMO_INF_PBMC_DN	3.9×10^{-07}	166	9	8
	GSE6269_E_COLI_VS_STAPH_AUREUS_INF_PBMC_DN	2.1×10^{-04}	174	7	5

Table 4: Pathway enrichment results for the target genes selected for UC. Several selected top pathways containing novel genes across different categories are shown here. See **Supplementary Tables 5 -7** for results on all the significant pathways for the selected in ARCHIE components 1 and 2 respectively.

Component	Category	Pathway	Adjusted p-value	Genes in pathway	Genes overlap	Novel genes overlap
1	GO	Regulation of Immune system (GO: 0002683)	8.5×10^{-09}	1606	25	11
		Lymphocyte activation (GO: 0046649)	8.5×10^{-09}	709	18	8
		Regulation of T-cell activation (GO: 0050863)	2.9×10^{-06}	311	11	6
	Hallmark	IL2 STAT5 signaling	3.2×10^{-10}	200	10	5
		G2M checkpoint	2.6×10^{-03}	200	4	4
	Curated	PILON_KLF1_TARGETS_DN	2.5×10^{-05}	854	11	4
		KEGG: Endocytosis	3.2×10^{-02}	181	5	3
		Reactome: Adaptive immune system	2.1×10^{-02}	807	10	3
	Immunologic Signatures	GSE7460_WT_VS_FOXP3_HET_ACT_WITH_TGFB_TCONV_UP	6.9×10^{-08}	200	10	8
		GSE4984_UNTREATED_VS_GALECTIN1_TREATED_DC_DN	4.5×10^{-16}	191	16	7
		GSE22886_NAIVE_CD4_TCELL_VS_MONOCYTE_UP	3.2×10^{-07}	196	11	6
2	GO	Defense Response (GO: 0006952)	2.0×10^{-04}	1684	11	4
		Inflammatory Response (GO: 0022610)	1.2×10^{-03}	714	7	2
	Curated	JAATINEN_HEMATOPOIETIC_STEM_CELL_DN	4.7×10^{-04}	233	6	3
	Immunologic Signatures	GSE22886_NAIVE_BCELL_VS_MONOCYTE_UP	2.1×10^{-12}	195	11	4
		GSE10325_CD4_TCELL_VS_BCELL_DN	5.8×10^{-08}	193	8	3