

1 **Tissue specificity-aware TWAS (TSA-TWAS) framework identifies novel associations with**  
2 **metabolic, immunologic, and virologic traits in HIV-positive adults**

3 Binglan Li<sup>1</sup>, Yogasudha Veturi<sup>2</sup>, Anurag Verma<sup>2</sup>, Yuki Bradford<sup>2</sup>, Eric S. Daar<sup>3</sup>, Roy M. Gulick<sup>4</sup>,  
4 Sharon A. Riddler<sup>5</sup>, Gregory K. Robbins<sup>6</sup>, Jeffrey L. Lennox<sup>7</sup>, David W. Haas<sup>8,9</sup>, Marylyn D.  
5 Ritchie<sup>1,2,10\*</sup>

6 <sup>1</sup> Genomics and Computational Biology Program, University of Pennsylvania, Philadelphia,  
7 Pennsylvania, United States of America

8 <sup>2</sup> Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania, United States  
9 of America

10 <sup>3</sup> Lundquist Institute at Harbor-UCLA Medical Center, Torrance, California, United States of  
11 America

12 <sup>4</sup> Weill Cornell Medicine, New York, New York, New York, United States of America

13 <sup>5</sup> University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

14 <sup>6</sup> Harvard Medical School, Boston, Massachusetts, United States of America

15 <sup>7</sup> Emory University School of Medicine, Atlanta, Georgia, United States of America

16 <sup>8</sup> Departments of Medicine, Pharmacology, Pathology, Microbiology & Immunology, Vanderbilt  
17 University School of Medicine, Nashville, Tennessee, United States of America

18 <sup>9</sup> Department of Internal Medicine, Meharry Medical College, Nashville, Tennessee, United  
19 States of America

20 <sup>10</sup> Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, Pennsylvania,  
21 United States of America

22 \* Corresponding author

23 E-mail: [marylyn@pennmedicine.upenn.edu](mailto:marylyn@pennmedicine.upenn.edu)

## 24 **Abstract**

25 As a type of relatively new methodology, the transcriptome-wide association study (TWAS) has  
26 gained interest due to capacity for gene-level association testing. However, the development of  
27 TWAS has outpaced statistical evaluation of TWAS gene prioritization performance. Current  
28 TWAS methods vary in underlying biological assumptions about tissue specificity of  
29 transcriptional regulatory mechanisms. In a previous study from our group, this may have  
30 affected whether TWAS methods better identified associations in single tissues versus multiple  
31 tissues. We therefore designed simulation analyses to examine how the interplay between  
32 particular TWAS methods and tissue specificity of gene expression affects power and type I  
33 error rates for gene prioritization. We found that cross-tissue identification of expression  
34 quantitative trait loci (eQTLs) improved TWAS power. Single-tissue TWAS (i.e., PrediXcan) had  
35 robust power to identify genes expressed in single tissues, but, had high false positive rates for  
36 genes that are expressed in multiple tissues. Cross-tissue TWAS (i.e., UTMOST) had overall  
37 equal or greater power and controlled type I error rates for genes expressed in multiple tissues.  
38 Based on these simulation results, we applied a tissue specificity-aware TWAS (TSA-TWAS)  
39 analytic framework to look for gene-based associations with pre-treatment laboratory values  
40 from AIDS Clinical Trial Group (ACTG) studies. We replicated several proof-of-concept  
41 transcriptionally regulated gene-trait associations, including *UGT1A1* (encoding bilirubin uridine  
42 diphosphate glucuronosyl transferase enzyme) and total bilirubin levels ( $p = 3.59 \times 10^{-12}$ ), and  
43 *CETP* (cholesterol ester transfer protein) with high-density lipoprotein cholesterol ( $p = 4.49 \times$   
44  $10^{-12}$ ). We also identified several novel genes associated with metabolic and virologic traits, as  
45 well as pleiotropic genes that linked plasma viral load, absolute basophil count, and/or  
46 triglyceride levels. By highlighting the advantages of different TWAS methods, our simulation  
47 study promotes a tissue specificity-aware TWAS analytic framework that revealed novel aspects  
48 of HIV-related traits.

49 publicly available.

50

## 51 **Introduction**

52 Translating fundamental genetics research discoveries into clinical research and clinical practice  
53 is a challenge for biomedical studies of complex human traits [1,2]. Greater than 90% of  
54 complex trait-associated single-nucleotide polymorphisms (SNPs) identified via genome-wide  
55 association studies (GWAS) are located in noncoding regions of the human genome [3,4]. The  
56 difficulty in making connections between noncoding variants and downstream affected genes  
57 can hinder the translatability of GWAS discoveries to clinical research. The emerging  
58 transcriptome-wide association studies (TWAS) are a type of recently developed bioinformatics  
59 methodology that provide a means to address the challenge of GWAS translatability. TWAS  
60 mitigates the translational issue by integrating GWAS data with expression quantitative trait loci  
61 (eQTLs) information to perform gene-level association analyses. TWAS hypothesizes that SNPs  
62 act as eQTLs to collectively moderate the transcriptional activities of genes and thus influence  
63 complex traits of interest [5,6]. Accordingly, TWAS methods in general comprise two steps. The  
64 first step in TWAS is to impute the genetically regulated gene expression (GR<sub>EX</sub>) for research  
65 samples in a tissue-specific manner. The second step is to conduct association analyses  
66 between GR<sub>EX</sub> and the trait of interest to evaluate the gene-trait relationship for statistical  
67 significance [7-9]. Genome-wide eQTLs data are now available for various primary human  
68 tissues (e.g., liver, brain and heart) thanks to large-scale eQTL consortia including the  
69 Genotype-Tissue Expression (GT<sub>EX</sub>) project [10] and the eQTLGEN consortium [11]. The  
70 considerable centralized eQTL data have been fostering the development and application of  
71 TWAS.

72

73 While TWAS is an innovative and potentially powerful computational approach, several factors  
74 can influence TWAS. The choice of eQTL datasets matters for the performance of TWAS [12].  
75 Most available eQTLs to date are identified in a tissue-by-tissue manner [5,10]. This approach,  
76 however, does not leverage the potential for shared transcriptional regulatory mechanisms  
77 across tissues, and can be limited by sample sizes of single tissues. One way to overcome this  
78 limitation is to take into consideration all available tissues, so as to increase sample sizes and  
79 improve the quality of eQTL datasets. We referred this type of eQTL detection method as the  
80 integrative tissue-based eQTL detection method [13-15]. Without a simulation study, however, it  
81 was unclear *how the choice of eQTL detection methods will impact TWAS*.

82

83 Another prominent question in TWAS studies is the choice of the association approaches.  
84 TWAS started with single-tissue association approaches, such as PrediXcan [5] and FUSION  
85 [6]. The most recent TWAS methods, such as UTMOST [15] and MulTiXcan [16], perform cross-  
86 tissue association analyses. Such TWAS methods evaluate whether a gene is significantly  
87 associated with a trait by integrating association data across tissues and adjusting for the  
88 statistical correlation structure among tissues. However, genes may vary substantially with  
89 regard to how they are regulated across tissues. When a gene is specifically expressed in a  
90 single or few tissues versus expressed in multiple tissues, *how will tissue specificity of gene*  
91 *expression affect TWAS power and type I error rates?*

92

93 Another appealing feature of TWAS is its capacity for tissue-specific association analyses  
94 thanks to the availability of tissue-specific eQTLs in a variety of primary human tissues.  
95 However, several recent studies revealed shared regulatory mechanisms across multiple  
96 human tissues [17] and showed that *cis*-eQTLs are less tissue-specific than other regulatory  
97 elements [10,11]. This suggests that TWAS can possibly identify genes in tissues that share  
98 biology with the causal tissue(s), but in fact are not the causal tissues for the trait of interest

99 [18]. While TWAS is likely to identify false positive tissues, to date, the false positive rates of  
100 tissues in TWAS is unknown.

101  
102 The above TWAS challenges can be summarized in two questions — *How does tissue*  
103 *specificity affect TWAS performance? How would this impact the choices of TWAS methods?*  
104 Available simulation strategies can be limited in answering these questions. Some have not  
105 taken into consideration the gene expression correlation structure across tissues [19,20]. Some  
106 assume a monogenic structure of transcriptional regulation [13-15,21], rather than the polygenic  
107 structure suggested by recent studies [10,22,23]. To address these issues, we applied a novel  
108 strategy to simulate eQTLs and gene expression of a wide range of tissue specificity (see  
109 **Methods**). We then applied different TWAS methods on the simulated datasets to assess  
110 power, type I error rates, and false positive rates of tissues. We found that the tissue specificity  
111 affected TWAS performance, with no single type of TWAS method being best for every type of  
112 genetic background of transcriptional regulation.

113  
114 The simulation results motivated the development and implementation of an enhanced, tissue  
115 specificity-aware TWAS (TSA-TWAS) analytic framework. We tested the performance of TSA-  
116 TWAS analytic framework using AIDS Clinical Trials Group (ACTG) data (described in  
117 **Methods**). We showed that the TSA-TWAS was able to both replicate proof-of-concept gene-  
118 trait associations and identify novel trait-related genes. The simulation scheme highlighted the  
119 effects of tissue specificity on TWAS performance, and that TSA-TWAS could help better  
120 understand regulatory mechanisms that underlie complex human traits.

## 121 Results

### 122 Simulation design

123 We designed a novel simulation framework to investigate how the tissue specificities of eQTLs  
 124 and gene expression affected TWAS power and type I error rates, and the choices of TWAS  
 125 methods (Fig 1). We tested two representative eQTL detection methods, elastic net  
 126 (implemented in PrediXcan [5]) and group LASSO (implemented in UTMOST [15]); and two  
 127 gene-trait association approaches, Principal Component Regression (PC Regression;  
 128 implemented in MultiXcan [16]) and Generalized Berk-Jones test (GBJ test; implemented in  
 129 UTMOST [15]) (Table 1).

130 **Table 1. TWAS methods tested in this simulation study**

eQTL detection methods		Gene-trait association approaches		Equivalent developed TWAS methods	PMID
Type	Name	Type	Name		
Single tissue-based	Elastic net	Single-tissue association	Linear or logistic regression	PrediXcan	31086352
Integrative tissue-based	Group LASSO	Single-tissue association	Linear or logistic regression	Single-tissue UTMOST	30804563
Single tissue-based	Elastic net	Cross-tissue association	Principal component regression	MultiXcan	30668570
Integrative tissue-based	Group LASSO	Cross-tissue association	Generalized Berk-Jones test	Cross-tissue UTMOST	30804563

131  
 132 Tissue-specific eQTLs were defined as those that were only functioning in one single tissue.  
 133 Multi-tissue eQTLs were defined as those that had regulatory effect across all gene-expressing  
 134 tissues (see **Methods**). We generated genes that had different genetic makeup of tissue-  
 135 specific and multi-tissue eQTLs in a gene to evaluate the influence of tissue specificity of eQTLs  
 136 on TWAS performance.

137

138 Tissue specificity of gene expression was determined by the number of gene-expressing tissues  
139 and the similarity of gene expression levels across tissues. (see **Methods**). Tissue-specific  
140 genes were those specifically expressed in only one or two tissues. Ubiquitously expressed  
141 genes were those expressed in all ten simulated tissues with high gene expression similarity  
142 (expression similarity = 60%, 80%). Differentially expressed and similarly expressed genes were  
143 those having distinctive gene expression levels (gene expression similarity = 0, 20% 40%) or  
144 highly correlated gene expression levels across tissues (gene expression similarity = 60%,  
145 80%), respectively, regardless of the number of gene-expressing tissues. To evaluate the  
146 impact of tissue specificity of gene expression on TWAS performance, we generated genes that  
147 were expressed in varied numbers of tissues and had diverse gene expression similarities  
148 across tissues.

149

150 In addition, we designed different strength of gene-trait associations defined by  $R^2_{expression-trait}$   
151 (the proportion of phenotypic variation explained by gene expression levels), but the reported  
152 results by default were the cases under  $R^2_{expression-trait} = 1\%$ . Only continuous traits were  
153 evaluated in this simulation study, in accordance with ACTG baseline laboratory values in the  
154 real-world application dataset.

155

## 156 **Power of different TWAS methods**

157 We did not observe any obvious effect of tissue-specificity of eQTLs on TWAS power, except for  
158 ubiquitously expressed genes. TWAS, specifically group LASSO (implemented in  
159 UTMOST[15]), had greater power to prioritize ubiquitously expressed genes that were mostly  
160 regulated by multi-tissue eQTLs than those that were not (Fig 2, bottom row).

161

162 We then asked how eQTL detection methods affected TWAS gene-prioritization power, and  
163 whether one eQTL detection method was preferred over another. We found that the integrative  
164 tissue-based eQTL detection method had, on average, approximately 2% greater power than  
165 the single-tissue method. Take differentially expressed genes for instance, eQTLs identified via  
166 the Group LASSO led to 53.8% gene prioritization power of TWAS and eQTLs identified via the  
167 Elastic Net led to 50.7% power (Wilcoxon Signed-rank Test  $p = 5.85 \times 10^{-4}$ ; S4 Fig, top right  
168 corner). More pairwise comparison results among all TWAS methods can be found in S1 Table.  
169 Overall, TWAS gained slightly more power when using eQTLs identified in an integrative tissue  
170 context.

171  
172 Gene-trait association approaches affected TWAS power more so than did choice of eQTL  
173 detection method. For tissue-specific genes or differentially expressed genes, SLR consistently  
174 had equal or greater power (average 70%) than the cross-tissue association approaches (PC  
175 regression and GBJ test; Fig 2, top left triangle). For ubiquitously expressed genes or similarly  
176 expressed genes, GBJ test had equal or greater power than the single-tissue association  
177 approach (SLR; Fig 2, bottom right triangle). Especially for ubiquitously expressed genes, GBJ  
178 test had statistically significant greater power (62%) compared to SLR (51%) (Fig 2, bottom right  
179 corner, Wilcoxon Signed-rank Test  $p = 9.4 \times 10^{-5}$ ).

180  
181 The group LASSO-GBJ test (implemented in UTMOST) had a greater power to prioritize  
182 similarly or ubiquitously expressed genes. For genes that were expressed in five tissues, power  
183 of the group LASSO-GBJ test increased from 62.2% for differentially expressed genes (Fig 2,  
184 top left corner) to 66.6% for similarly expressed gene (Fig 2, bottom right corner). For genes that  
185 were expressed in all ten tissues, power of the group LASSO-GBJ test increased from 51.2%  
186 for differentially expressed genes (Fig 2, top left corner) to 61.9% for ubiquitously expressed  
187 gene (Fig 2, bottom right corner). Moreover, the group LASSO-GBJ test showed equal or



188 statistically significant greater power than other TWAS methods in 65 of the 76 simulated  
189 scenarios (~84%). Black brackets in Fig 2 showed cases where Group LASSO-GBJ had higher  
190 power than other three methods; red brackets showed cases where Group LASSO-GBJ had  
191 lower power than other three methods. Comprehensive statistical test results of power  
192 differences are available in S4 Fig and S1 Table. However, GBJ test cannot handle the case  
193 where the gene was only expressed in one single tissue.

194

195 Overall, the group LASSO-GBJ test had equal or greater power in prioritizing genes that were  
196 expressed in multiple tissues. Single-tissue association approaches (e.g. SLR) had greater  
197 power and robust performance in prioritizing tissue-specific genes.

198

199 The strength of gene-trait associations affected TWAS gene prioritization power. The stronger  
200 the gene-trait associations, the greater the power for TWAS gene prioritization (Fig 2, S5-7  
201 Figs).

202

### 203 **Type I error rates of various TWAS methods**

204 All TWAS methods had well-controlled type I error rates ( $\leq 5\%$ ; Fig 3, S2 Table). Significance  
205 thresholds in this simulation were corrected using the Bonferroni approach to control for family-  
206 wise error rate. All single-tissue association approaches (Elastic Net-SLR and Group LASSO-  
207 SLR) had less type I error rates than the cross-tissue associations approaches (Wilcoxon  
208 Signed-rank Test  $p < 0.01$ , S8 Fig). Both GBJ test and PC regression had average type I error  
209 rates of approximately 5%. The GBJ test showed statistically significant lower type I error rates  
210 than PC regression for ubiquitously expressed genes (Wilcoxon Signed-rank  $p < 0.05$ , S8 Fig,  
211 S2 Table).

212

## 213 **False positives of statistically significant tissues**

214 If not corrected for the number of tested tissues, single-tissue TWAS would have greater power  
215 (S9 Fig), but also a higher false positive rate for tissues (S10 Fig). False positive rates of tissues  
216 were at least 10% for genes that were expressed in more than one tissue. In effect, while the  
217 genes might be related to a trait of interest, 10% of statistically significant results pointed to  
218 wrong tissues. The false positive rate of tissues proportionally increased with the number of  
219 gene-expressing tissues. The highest false positive rates were seen in the case of ubiquitously  
220 expressed genes (S10 Fig, bottom right corner), which on average, had an 84% false positive  
221 rate based on 20 random replications. This suggested that any single-tissue TWAS may have  
222 10-84% false positive rate tissues associations if not adjusted for the number of tested tissues.

223  
224 Adjusting for the number of tested tissues reduced the false positive rates somewhat, but  
225 number-wise, the false positive rate may remain quite high. False positive rates of tissues were  
226 relatively controlled at approximately 5% for tissue-specific genes (Fig 4, top left corner). False  
227 positive rates still increased with the number of tissues in which a gene was expressed (Fig 4).  
228 Genes expressed in ten tissues had at least on average a 24% false positive rate. False positive  
229 rates were as high as 77% for ubiquitously expressed genes (Fig 4, bottom right corner).

230

## 231 **Validation and support of simulation design**

232 To evaluate whether our simulation findings would translate from in silico parameter designs to  
233 real world scenarios, we designed a Monte Carlo simulation process to estimate the trait  
234 heritability behind various genetic scenarios (S11 Fig). The results suggested that  
235  $R^2_{expression-trait}$  increased with trait heritability (S12 Fig). Heritability of traits with  
236  $R^2_{expression-trait} = 1\%$  were estimated to be on average 1% (standard error (s.e.) = 0.059%)  
237 which were derived from multiple, repeated random sampling. In contrast, the minor allele

238 frequencies (MAF) of eQTLs had almost no effect on trait heritability. This suggested that trait  
239 heritability positively influenced the strength of gene-trait associations in TWAS. In other words,  
240 if a trait was moderated by genetic factors through differential gene expression, the greater a  
241 trait's heritability is, the stronger the associations were in TWAS.

242

### 243 **Designing the TSA-TWAS analytic framework**

244 Our simulation suggested an influence of tissue specificity on TWAS performance. Thus, we  
245 designed a TSA-TWAS analytic framework to balance trade-offs among power, type I error  
246 rates, and false positive rates of tissues and to take into consideration the distribution of GReX  
247 (S13 Fig). The idea was illustrated in Fig 5. When trait-related tissue(s) are known, we  
248 recommend single-tissue TWAS in the known related tissues only. Additionally, we recommend  
249 using eQTLs identified by integrative tissue-based eQTL detection methods (for example, group  
250 LASSO), which showed slightly greater power. In contrast, if trait-related tissue(s) are uncertain,  
251 it may be better to stratify genes based on the number of tissues in which the genes are  
252 predicted to be expressed. For genes predicted to be expressed in just one tissue, single-tissue  
253 TWAS will have greater power and can provide information on trait-related tissues. For genes  
254 that are expressed multiple tissues, cross-tissue TWAS will provide overall equal or greater  
255 power, as well as controlled type I error rates.

256

### 257 **TSA-TWAS replicated known associations**

258 We applied TSA-TWAS to 37 baseline laboratory values from a combined dataset of five clinical  
259 trials from AIDS Clinical Trials Group (ACTG) with available genotype data (N = 4,360; Table 2).  
260 We first imputed the GReX to distinguish genes whose GReX were only expressed in one tissue  
261 versus multiple tissues. Genes expressed in just one tissue comprised 2,812 (23%) of 12,038  
262 genes on which data were available. The remaining 9,226 (77%) genes had GReX in multiple

263 tissues. Genes expressed in one, and in more than one tissue were tested for associations with  
 264 baseline laboratory values using single-tissue, and by cross-tissue gene-trait association  
 265 approaches, respectively (see **Methods**). TSA-TWAS found in total 83 statistically significant  
 266 gene-trait associations, comprising 45 distinct genes and 10 traits (Fig 7).

267 **Table 2. Summary statistics of the ACTG genotyping phase I-IV baseline laboratories.**

Trait	Sample Size	Mean	Std. Dev.	Min	Max	Transformation	Unit	Description
Albumin	1216	4.05	0.44	1.80	5.30		g/dL	week 0 albumin (Alb, g/dL)
Bicarbonate	3971	26.01	2.94	12.00	35.00		mmol/L	week 0 bicarbonate (Bicarb, mmol/L)
Calcium	1336	9.17	0.44	7.40	10.80		mg/dL	week 0 calcium (Ca, mg/dL)
Chloride	4048	103.27	2.94	88.00	117.00		mmol/L	week 0 chloride (Cl, mmol/L)
Cholesterol	4286	159.27	36.80	5.90	414.00		mg/dL	week 0 cholesterol (Chol, mg/dL)
Creatinine	4100	0.91	0.20	0.05	2.80		mg/dL	week 0 creatinine (Creat, mg/dL)
HDL-c	2376	37.31	12.78	3.90	148.00		mg/dL	week 0 HDL-c (HDL-c, mg/dL)
Hemoglobin	4293	13.49	1.77	6.00	20.20		g/dL	week 0 hemoglobin (Hgb, g/dL)
Absolute basophil count	2526	1.44	0.32	0.00	3.39	Log <sub>10</sub>	cells/mm <sup>3</sup>	log <sub>10</sub> transformed week 0 absolute basophil count (Baso, cells/mm <sup>3</sup> )
Absolute eosinophil count	3932	2.06	0.40	0.18	3.55	Log <sub>10</sub>	cells/mm <sup>3</sup>	log <sub>10</sub> transformed week 0 absolute eosinophil count (Eos, cells/mm <sup>3</sup> )
Alkaline phosphatase	4226	1.88	0.15	0.70	2.72	Log <sub>10</sub>	U/L	log <sub>10</sub> transformed week 0 alkaline phosphatase (AlkP, U/L)
ALT	4233	1.48	0.27	0.04	2.81	Log <sub>10</sub>	U/L	log <sub>10</sub> transformed week 0 ALT (ALT, U/L)
Absolute lymphocyte count	4149	3.11	0.24	0.92	4.03	Log <sub>10</sub>	cells/mm <sup>3</sup>	log <sub>10</sub> transformed week 0 absolute lymphocyte count (Lymph, cells/mm <sup>3</sup> )
Absolute monocyte count	4116	2.58	0.21	0.66	3.69	Log <sub>10</sub>	cells/mm <sup>3</sup>	log <sub>10</sub> transformed week 0 absolute monocyte count (Mono, cells/mm <sup>3</sup> )
Amylase	1026	1.85	0.20	1.11	2.89	Log <sub>10</sub>	U/L	log <sub>10</sub> transformed week 0 amylase (Amyl, U/L)
Absolute neutrophil count	4277	3.32	0.21	2.28	4.67	Log <sub>10</sub>	cells/mm <sup>3</sup>	log <sub>10</sub> transformed week 0 absolute neutrophil count (ANC, cells/mm <sup>3</sup> )
AST	4235	1.49	0.21	0.48	2.81	Log <sub>10</sub>	U/L	log <sub>10</sub> transformed week 0 AST (AST, U/L)
BUN	4221	1.08	0.15	-0.22	2.17	Log <sub>10</sub>	mg/dL	log <sub>10</sub> transformed week 0 BUN (BUN, mg/dL)
CK	1360	1.97	0.38	-0.05	3.79	Log <sub>10</sub>	U/L	log <sub>10</sub> transformed week 0 CK (CK, U/L)
Fasting glucose	3233	1.93	0.08	1.52	2.64	Log <sub>10</sub>	mg/dL	log <sub>10</sub> transformed week 0 fasting glucose (Gluc fasting, mg/dL)
Glucose (Log <sub>10</sub> )	3031	1.93	0.08	1.70	2.77	Log <sub>10</sub>	mg/dL	log <sub>10</sub> transformed week 0 glucose (Gluc, mg/dL)
LDL-c	3539	1.95	0.16	0.00	2.57	Log <sub>10</sub>	mg/dL	log <sub>10</sub> transformed week 0 LDL-c (LDL-c, mg/dL)
Lipoprotein	1118	1.58	0.32	0.30	2.85	Log <sub>10</sub>		log <sub>10</sub> transformed week 0 lipoprotein
Platelet count	4263	2.30	0.15	1.15	3.34	Log <sub>10</sub>	x10E9/L	log <sub>10</sub> transformed week 0 platelet count (Plat, x10E9/L)
Total bilirubin	4202	-0.31	0.21	-1.00	0.49	Log <sub>10</sub>	mg/dL	log <sub>10</sub> transformed week 0 total bilirubin (TBili, mg/dL)
Triglyceride	4318	2.07	0.25	1.08	3.45	Log <sub>10</sub>	mg/dL	log <sub>10</sub> transformed week 0 triglyceride (Trig, mg/dL)
White blood cell count	4279	0.62	0.16	-0.05	1.49	Log <sub>10</sub>	x10E3 cells/cu	log <sub>10</sub> transformed week 0 white blood cell count (WBC, x10E3)

							mm	cells/cu mm)
Hematocrit	4274	39.83	5.10	1.00	62.10		percent	week 0 hematocrit (Hct, percent)
Phosphate	3261	3.44	0.61	0.80	7.70		mg/dL	week 0 phosphate (Phos, mg/dL)
Potassium	4062	4.15	0.39	2.00	8.00		mmol/L	week 0 potassium (K, mmol/L)
Sodium	4067	138.88	2.80	123.00	151.00		mmol/L	week 0 sodium (Na, mmol/L)
CD4 count	4358	14.78	6.46	0.00	36.55	Square root	cells/mm <sup>3</sup>	square root of absolute CD4 count at week 0
Viral load	4358	4.75	0.72	2.02	7.11	Log <sub>10</sub>	copies/dL	week 0 viral load RNA
Fasting cholesterol	4136	158.42	36.24	6.10	414		mg/dL	week 0 fasting cholesterol
Fasting HDL-c	4126	1.56	0.15	0.60	2.20	Log <sub>10</sub>	mg/dL	log <sub>10</sub> transformed week 0 fasting HDL-c
Fasting LDL-c	4042	1.95	0.15	0.85	2.57	Log <sub>10</sub>	mg/dL	log <sub>10</sub> transformed week 0 fasting LDL-c
Fasting triglyceride	3888	2.05	0.24	1.08	2.45	Log <sub>10</sub>	mg/dL	log <sub>10</sub> transformed week 0 fasting triglycerides

268

269 TSA-TWAS replicated several previously reported risk genes for certain baseline lab values  
 270 (Table 3). The lowest p-values for association were observed between total plasma bilirubin  
 271 levels and several genes on chromosome 2, nearby or overlapping *UGT1A1*. These included  
 272 *MROH2A* ( $p = 1.39 \times 10^{-12}$ ), which has been previously reported by GWAS of various  
 273 populations [24-27], *UGT1A6* ( $p = 2.78 \times 10^{-15}$ ), *UGT1A7* ( $p = 4.51 \times 10^{-12}$ ) and *UGT1A1* ( $p =$   
 274  $3.59 \times 10^{-12}$ ) [24,25,27,28]. We replicated the well-known association between *CETP* and high-  
 275 density lipid-cholesterol levels (HDL-c;  $p = 4.49 \times 10^{-12}$ ) [29]. Association was also found  
 276 between *GPLD1* and plasma alkaline phosphatase levels ( $p = 1.08 \times 10^{-11}$ ) [30]. *GPLD1*  
 277 encodes a glycosylphosphatidylinositol-degrading enzyme that releases attached proteins from  
 278 the plasma membrane and engages in regulation of alkaline phosphate activities. Other  
 279 replicated discoveries included association between *ALDH5A1* and plasma alkaline  
 280 phosphatase levels ( $p = 1.79 \times 10^{-11}$ ) [31], *C6orf48* and absolute basophil count ( $p = 1.69 \times$   
 281  $10^{-12}$ ) [32], *KCNJ15* and plasma triglyceride levels ( $p = 3.18 \times 10^{-13}$ ) [33].

282 **Table 3. Replicated associations related to HIV baseline laboratory values identified by**  
 283 **TSA-TWAS.**

Trait	Gene	Chromosome	TSS	P
Alkaline phosphatase	<i>GPLD1</i>	6	24428177	1.08E-11

	<i>ALDH5A1</i>	6	24494852	1.79E-11
Fasting HDL	<i>CETP</i>	16	56961850	4.49E-12
HDL	<i>CETP</i>	16	56961850	4.49E-12
Total bilirubin	<i>UGT1A6</i>	2	233692866	2.78E-15
	<i>MROH2A</i>	2	233775679	1.39E-12
	<i>UGT1A1</i>	2	233760248	3.59E-12
	<i>UGT1A7</i>	2	233681938	4.51E-12
Triglyceride	<i>KCNJ15</i>	21	38256698	3.18E-13
Viral load	<i>C4B</i>	6	32014762	4.11E-15
	<i>GABBR1</i>	6	29602228	1.14E-12
	<i>ABCB4</i>	7	87401697	1.07E-11
	<i>HLA-B</i>	6	31269491	1.15E-11
	<i>C6orf48</i>	6	31834608	2.32E-11
	<i>A4GALT</i>	22	42692121	8.39E-11

284

285 We have additionally replicated several genes' association with plasma viral loads in HIV-  
 286 positive adults, including *A4GALT* ( $p = 8.39 \times 10^{-11}$ ) [34], *ABCB4* ( $p = 1.07 \times 10^{-11}$ ) [35], *C4B*  
 287 ( $p = 4.11 \times 10^{-15}$ ) [36], *GABBR1* ( $p = 1.14 \times 10^{-12}$ ) [37], and *HLA-B* ( $p = 1.15 \times 10^{-11}$ ) [38].

288

289

290

## 291 **Novel genes prioritized by the TSA-TWAS**

292 In addition to the above replications, TSA-TWAS identified novel associations with plasma viral  
 293 load (Table 4). For instance, *PRDX5* ( $p = 7.01 \times 10^{-14}$ , which encodes a member of the  
 294 peroxiredoxin family of antioxidant enzymes) was associated with plasma viral load with great  
 295 significance. Several novel genes were first time reported to be associated with certain baseline  
 296 laboratory values, which were otherwise associated with other traits by previous studies. For  
 297 instance, *ATF6B* is a protein-coding gene that encodes a transcription factor in the unfolded  
 298 protein response (UPR) pathway during ER stress and it has been associated with HIV-

299 associated neurocognitive disorders in previous research. In our study, ATF6B associates with  
 300 plasma viral load ( $p = 2.83 \times 10^{-9}$ ).

301 **Table 4. Novel associations related to HIV baseline laboratory values identified by TSA-**  
 302 **TWAS.**

Trait	Gene	Chromosome	TSS	P
Absolute basophil count	<i>KCTD7</i>	7	66628767	3.08E-14
	<i>CNBD2</i>	20	35955360	3.83E-13
	<i>CD2AP</i>	6	47477789	7.27E-13
	<i>RP11-385F7.1</i>	6	47477243	1.32E-12
	<i>C6orf48</i>	6	31834608	1.69E-12
	<i>PARM1</i>	4	74933095	1.84E-11
	<i>USP19</i>	3	49108046	1.51E-10
	<i>GPATCH4</i>	1	156594487	2.24E-10
	<i>GPR22</i>	7	107470018	1.81E-09
	<i>HIST1H1E</i>	6	26156354	2.19E-09
	<i>RPS28</i>	19	8321500	2.87E-09
	<i>KCNJ15</i>	21	38256698	4.72E-09
	<i>TTI2</i>	8	33473423	6.35E-09
	<i>CDK5RAP3</i>	17	47967810	1.05E-08
	<i>F2RL1</i>	5	76818933	2.99E-08
<i>C4B</i>	6	32014762	8.92E-08	
Absolute neutrophil count	<i>PMVK</i>	1	154924734	3.63E-08
Alkaline phosphatase	<i>PCDHB3</i>	5	141100756	7.44E-09
	<i>KCNJ15</i>	21	38256698	2.77E-08
Fasting HDL	<i>NLRC5</i>	16	56989485	1.70E-09
Sodium	<i>CNBD2</i>	20	35955360	7.71E-08
Triglyceride	<i>PCDHB3</i>	5	141100756	5.78E-14
	<i>GPATCH4</i>	1	156594487	2.12E-12
	<i>CNBD2</i>	20	35955360	7.21E-12
	<i>C6orf48</i>	6	31834608	9.13E-12
	<i>PARM1</i>	4	74933095	1.88E-11
	<i>TTI2</i>	8	33473423	2.69E-11
	<i>USP19</i>	3	49108046	9.69E-11

	<i>HIST1H1E</i>	6	26156354	1.20E-10
	<i>CD2AP</i>	6	47477789	1.04E-09
	<i>RP11-385F7.1</i>	6	47477243	1.17E-09
	<i>C4B</i>	6	32014762	1.23E-09
	<i>KCTD7</i>	7	66628767	1.34E-08
	<i>RPS28</i>	19	8321500	1.40E-08
	<i>C11orf74</i>	11	36594493	1.94E-08
	<i>ATAT1</i>	6	30626842	5.32E-08
Viral load	<i>PPP1R18</i>	6	30676389	6.27E-14
	<i>PRDX5</i>	11	64318088	7.01E-14
	<i>F2RL1</i>	5	76818933	1.81E-12
	<i>CDK5RAP3</i>	17	47967810	1.95E-12
	<i>RPS28</i>	19	8321500	3.50E-12
	<i>USP19</i>	3	49108046	3.60E-12
	<i>KCTD7</i>	7	66628767	3.84E-12
	<i>TTI2</i>	8	33473423	4.27E-12
	<i>TSTD1</i>	1	161037631	4.57E-12
	<i>UBFD1</i>	16	23557732	5.27E-12
	<i>RP11-385F7.1</i>	6	47477243	1.05E-11
	<i>KCNJ15</i>	21	38256698	1.08E-11
	<i>CD2AP</i>	6	47477789	1.54E-11
	<i>CNBD2</i>	20	35955360	1.70E-11
	<i>PARM1</i>	4	74933095	1.86E-11
	<i>ATAT1</i>	6	30626842	2.20E-11
	<i>HIST1H1E</i>	6	26156354	8.44E-11
	<i>MTRF1L</i>	6	152987362	1.14E-10
	<i>MLF1</i>	3	158571163	1.23E-10
	<i>PCDHB3</i>	5	141100756	2.42E-09
	<i>ATF6B</i>	6	32115335	2.83E-09
	<i>GPR22</i>	7	107470018	3.75E-09
	<i>RBM17</i>	10	6088987	5.39E-09
	<i>PLA2G7</i>	6	46704320	6.34E-09
	<i>GPATCH4</i>	1	156594487	1.81E-08
	<i>NDUFS4</i>	5	53560633	2.07E-08
	<i>C11orf74</i>	11	36594493	2.14E-08



	<i>CSNK2B</i>	6	31665391	2.76E-08
	<i>GPR18</i>	13	99254714	4.10E-08
	<i>FEZ2</i>	2	36531805	4.54E-08

303

## 304 **Pleiotropic genes associated with baseline laboratory values**

305 We also found several pleiotropic genes which were statistically significantly associated with  
306 plasma viral load, triglyceride levels, and/or absolute basophil count (Fig 7). These included  
307 *ABCB4, ATAT1, C11orf74, C4B, C6orf48, CD2AP, CDK5RAP3, CNBD2, F2RL1, GPATCH4,*  
308 *GPR22, KCNJ15, KCTD7, PARM1, PCDHB3, RPS28, TTI2, USP19.* Some of them were  
309 located on chromosome 6, surrounding the major histocompatibility complex (MHC) region,  
310 while the rest scattered across the human genome. Meanwhile, we did not observe correlations  
311 among plasma viral load, triglyceride levels, or absolute basophil count. The strongest  
312 correlation was observed between plasma viral load and triglyceride levels ( $r^2 = 0.24$ ),  
313 suggesting only weak correlation, and correlations for the other pairs of laboratory values were  
314 approximately 0. Overall, there were potential pleiotropic genes for plasma viral load,  
315 triglyceride levels, and/or absolute basophil count in HIV-positive adults.

## 316 **Discussion**

### 317 **Novel design of the simulation framework**

318 In this report, we described a novel simulation framework for TWAS, and evaluated TWAS gene  
319 prioritization performance for genes with various degrees of tissue specificity. Our simulation  
320 results validated conclusions from several previous eQTL or TWAS studies [13-15,21], and also  
321 generated new findings that warrant attention in future TWAS. First, TWAS methods tested in  
322 this study all had well-controlled type I error rates ( $\leq 5\%$ ) for genes with any degrees of tissue-  
323 specificity. Second, single-tissue TWAS tended to have higher false positive rates of tissues.  
324 The phenomenon became more obvious when genes had more correlated expression levels

325 across tissues. For tissue-specific genes, false positive rates of tissues could be controlled ( $\leq$   
326 5%) by adopting a more stringent multiple testing correction approach. However, for  
327 ubiquitously expressed genes, false positive rates of tissues remained significant (~77%) even  
328 after a stringent multiple testing adjustment. Third, TWAS gene prioritization power was  
329 improved by eQTLs that were identified by jointly analyzing transcriptomic data across tissues.  
330 Fourth, for tissue-specific genes, single-tissue and cross-tissue gene-level association  
331 approaches had similar power. For ubiquitously expressed and similarly expressed genes,  
332 cross-tissue association approaches had greater power.

333

334 We further tested our simulation designs for how they would translate to real-world data by  
335 evaluating trait heritability in our simulated datasets. We found no apparent effect of MAF  
336 distribution on trait heritability under TWAS models. Instead, trait heritability increased with  
337  $R^2_{expression-trait}$ . When  $R^2_{expression-trait} = 1\%$ , trait heritability was approximately 1% (s.e. =  
338 0.059%). The estimated trait heritability was within a reasonable range and supported our  
339 simulation design.

340

### 341 **Associations in the clinical trials dataset**

342 TSA-TWAS successfully replicated proof-of-concept gene-trait associations, including  
343 associations between *CETP* and HDL-c, and between *GPLD1* and plasma alkaline phosphatase  
344 levels. For total plasma bilirubin levels, our TSA-TWAS framework prioritized *UGT1A1* and  
345 genes near *UGT1A1*. *UGT1A1* encodes the hepatic protein that glucuronidates bilirubin [28],  
346 and has been known to affect bilirubin levels [24,25,27]. Other genes have been associated with  
347 total bilirubin levels in numerous studies [24-27]. These genes span 1Mbp at the 2q37.1 locus  
348 and are within the same topologically associating domain (TAD), which suggests that a

349 regulatory mechanism may affect expression of the entire *KCNJ13-UGT1A-MROH2A* gene  
350 region.

351  
352 TSA-TWAS has also identified several pleiotropic genes that linked plasma viral load, absolute  
353 basophil count, and/or triglyceride levels, which were otherwise independent from each other.  
354 Plasma viral load is a strong predictor of clinical outcome and is highly variable among people  
355 living with HIV. Individuals vary in their ability in suppressing viral loads, in the absence of  
356 antiretroviral treatments. Moreover, people living with HIV experience dyslipidemia to different  
357 degrees. Grunfeld *et al.* [39] found that AIDS patients experienced different lipid changes from  
358 HIV-infected patients without AIDS. The discovery of pleiotropic genes suggests the complexity  
359 of HIV pathogenesis and provides a future direction for research on the complex inter-individual  
360 variability among people living with HIV.

361

## 362 **Limitations & future directions**

363 Our simulations revealed high false positive rates of tissues for single-tissue TWAS. The high  
364 false positive rates seen with single-tissue TWAS may be due to limited sample sizes for eQTL  
365 discovery. GTEx analysis has shown that discovery of tissue-specific eQTLs is contingent on  
366 the sample sizes of tissues [10]. Unfortunately, many tissues still have limited sample sizes for  
367 the identification of tissue-specific eQTLs. Consequently, single-tissue TWAS may not have  
368 ample power to prioritize potential trait-related tissues. Adopting stricter multiple testing  
369 adjustment strategies for single-tissue TWAS is one practical approach to help reduce false  
370 positive rates in prioritized tissues, but this will sacrifice power.

371

372 The evaluation of TWAS power and type I error rates estimated from this simulation study might  
373 be limited due to the small sample sizes ( $N = 2,000$  for association analyses). We selected this

374 sample size for simulation in order to make it comparable to the average sample size of the  
375 ACTG phase I-IV combined clinical traits interrogated in this study. TWAS gene prioritization  
376 power can be improve with greater sample, but also under influence of many other factors as  
377 shown in Veturi *et al.* [21] and this study. Thus, TWAS performance can differ from dataset to  
378 dataset when using different TWAS methods. It was difficult to take every factor into  
379 consideration in this work. We dedicated this study to explore tissue specificity's impact on  
380 TWAS performance, and, for future TWAS studies, suggest customized simulation to better  
381 understand TWAS performance on specific datasets and diseases of interest.

382

## 383 **Conclusions**

384 Gene-level association studies offer the opportunity to better understand the genetic  
385 architecture of complex human traits by leveraging regulatory information from both noncoding  
386 and coding regions of the genome. This may expedite translation of basic research discoveries  
387 to clinical applications. We provide a comprehensive simulation algorithm to fully investigate  
388 TWAS performance for diverse biological scenarios. Based on our simulation, we promote a  
389 TSA-TWAS analytic framework. TSA-TWAS framework on ACTG clinical trials data ascribed  
390 statistical significance to proof-of-concept gene-trait associations, and also found several novel  
391 associations and pleiotropic genes, suggesting the complexity of HIV-related traits that latest  
392 bioinformatics methods can reveal.

393

394 Additional work is needed to fully understand the tissue and genetic architecture underlying  
395 complex traits. The simulation algorithm and schema developed for this study is versatile  
396 enough to answer other questions regarding causal genes and tissues for complex traits.  
397 Overall, our work provides and tests a novel, flexible simulation framework and an TSA-TWAS  
398 analytic framework for future complex trait studies.

399

## 400 **Materials and Methods**

### 401 **TWAS simulation design**

402 The simulation study systematically evaluated how the tissue-specificity of eQTLs and gene  
403 expression levels influences TWAS gene prioritization performance. We assumed additive  
404 genetic effects of eQTLs on gene expression levels, and of gene expression levels on traits.  
405 The TWAS simulation scripts are available in R programming language at GitHub  
406 ([https://github.com/BinglanLi/multi tissue twas sim](https://github.com/BinglanLi/multi_tissue_twas_sim)).

407

408 **Genotype.** We started by simulating genotypes for one gene in 1,500 individuals, which  
409 include eQTL and non-eQTL SNPs. Genotypes are denoted as  $X_{N \times M}$  throughout this paper,  
410 where  $N$  denotes the total number of individuals and  $M$  denotes the total number of SNPs in a  
411 gene that include tissue-specific eQTLs, multi-tissue eQTLs and non-eQTL SNPs. These  
412 individuals were later stratified into an eQTL discovery dataset ( $N_{eQTL} = 500$ ) and a TWAS  
413 testing dataset ( $N_{TWAS} = 1000$ ), sample sizes comparable to those of current GTEx and ACTG  
414 datasets used in this analysis, respectively. Genotypes were simulated as biallelic SNPs and  
415 then converted into allele dosages as is done in most eQTL detection methods. MAF assigned  
416 to SNPs ranged from 1% to 50% and were randomly drawn from a uniform distribution,  
417  $U(0.01, 0.5)$ . Parameter settings of eQTLs in this simulation were drawn from observations in  
418 different eQTL databases (S1-3 Figs).

419

420 **Gene expression level.** We simulated one gene's standardized expression levels at a time  
421 such that it was expressed in a fixed number of tissues. Let  $P$  denote the number of tissues  
422 where the gene is expressed,  $P = 1, 2, 5, \text{ or } 10$ . If a gene is only expressed in a single tissue ( $P$

423 = 1), then, only single-tissue eQTLs were simulated for this given gene and no multi-tissue  
424 eQTLs were present.

425

426 A previous study showed that the number of eQTLs in a gene does not have as pronounced an  
427 effect on the TWAS power in comparison to other parameters [21]. Hence, we assumed that a  
428 given gene was regulated by the same total number of eQTLs in each of the  $P$  tissues, which is  
429 denoted by  $M_{eQTLs}$  ( $M_{eQTLs} = 30$ ). eQTLs can be tissue-specific or have effect across multiple  
430 tissues. Here, we defined tissue-specific eQTLs as those that had effects in one and only one  
431 tissue. Multi-tissue eQTLs were defined as those who had effects in all  $P$  tissues in which the  
432 given gene is simulated to be expressed. We allowed multi-tissue eQTLs to have different effect  
433 sizes in different tissues. Assuming that a gene was expressed in  $P$  tissue(s) (say  $P = 5$ ), then,  
434 this gene is regulated by both, tissue-specific eQTLs and multi-tissue eQTLs, in any of the  $P$   
435 tissues. Let  $M_{ts-eQTLs}$  denote the number of tissue-specific eQTLs, and  $M_{mt-eQTLs}$  the number  
436 of multi-tissue eQTLs. A simulated gene had the same  $M_{ts-eQTLs}$  across  $P$  tissues, and the  
437 same  $M_{mt-eQTLs}$  across  $P$  tissues, such that  $M_{ts-eQTLs}$  and  $M_{mt-eQTLs}$  added up to  $M_{eQTLs}$  in  
438 each of the  $P$  tissues. Five different numbers of  $M_{mt-eQTLs}$  (0, 6, 12, 18, 24, corresponding  
439  $M_{ts-eQTLs} = 30, 24, 18, 12, 6$ ) were evaluated, except when a gene was simulated to be  
440 expressed only in one gene, in which case  $M_{mt-eQTLs}$  always equaled 0.

441

442 Each gene was simulated under an additive genetic model per tissue. Let  $E_{N \times P}$  denote the  
443 simulated gene expression levels for one gene, of  $N$  individuals, and across  $P$  tissues. For the  
444 given simulated gene, let  $E_{np}$  represent the simulated expression level of the  $n$ th individual in  
445 the  $p$ th tissue, which is an aggregate of tissue-specific eQTLs, multi-tissue eQTLs and non-  
446 eQTL effects in individual  $n$  for tissue  $p$ . The multivariate normal random effects model to  
447 simulate one gene's expression levels is then expressed as follows:

$$E = X_{ts-eQTLs}\beta_{ts-eQTLs} + X_{mt-eQTLs}\beta_{mt-eQTLs} + \varepsilon_1$$

448 where  $E$  is the  $N \times P$  matrix of standardized gene expression levels for a gene in  $N$  individuals  
 449 across  $P$  tissues.  $X_{ts-eQTLs}$  is the  $N \times M_{ts-eQTLs}$  matrix of standardized tissue-specific eQTL  
 450 genotypes. Similarly,  $X_{mt-eQTLs}$  is the  $N \times M_{mt-eQTLs}$  matrix of standardized multi-tissue eQTL  
 451 genotypes.  $\beta_{ts-eQTLs}$  is a  $M_{ts-eQTLs} \times P$  matrix of tissue-specific eQTL effects.  $\beta_{ts-eQTLs,ip}$   
 452 represents the  $i$ th tissue-specific eQTL in the  $p$ th tissue, which could be a different eQTL across  
 453  $P$  tissues. Each value in the  $\beta_{ts-eQTLs}$  is independent of the others.  $\beta_{mt-eQTLs}$  is a  $M_{mt-eQTLs} \times P$   
 454 matrix of multi-tissue eQTL effects wherein  $\beta_{mt-eQTLs,jp}$  represents the  $j$ th multi-tissue eQTL in  
 455 the  $p$ th tissue. In contrast to tissue-specific eQTLs,  $\beta_{mt-eQTLs,j}$  denotes the same  $j$ th multi-  
 456 tissue eQTL in all  $P$  tissues, and is allowed to have similar or dissimilar effect sizes across  $P$   
 457 tissues (explained later in this section).  $vec(\beta_{ts-eQTLs}) \sim N(\mathbf{0}_{M_{ts-eQTLs} \times P}, \Sigma_{P \times P}^{ts-eQTLs} \otimes I_{M_{ts-eQTLs}})$

458 where  $\Sigma_{P \times P}^{ts-eQTLs} = \begin{cases} h_{SNP-expression}^2 \times \frac{M_{ts-eQTLs}}{M_{eQTLs}}, & p = p' \\ 0, & p \neq p' \end{cases}$ . The constant,  $h_{SNP-expression}^2 \times \frac{M_{ts-eQTLs}}{M_{eQTLs}}$ ,

459 represents the proportion of variation in gene expression that can be explained by tissue-  
 460 specific eQTLs.  $vec(\beta_{mt-eQTLs}) \sim N(\mathbf{0}_{M_{mt-eQTLs} \times P}, \Sigma_{P \times P}^{mt-eQTLs} \otimes I_{M_{mt-eQTLs}})$

461 where  $\Sigma_{P \times P}^{mt-eQTLs} = \begin{cases} h_{SNP-expression}^2 \times \frac{M_{mt-eQTLs}}{M_{eQTLs}}, & p = p' \\ cor(tissue_p, tissue_{p'}) \times h_{SNP-expression}^2 \times \frac{M_{mt-eQTLs}}{M_{eQTLs}}, & p \neq p' \end{cases}$ . The constant,

462  $h_{SNP-expression}^2 \times \frac{M_{mt-eQTLs}}{M_{eQTLs}}$ , represents the proportion of gene expression variation that can be

463 explained by multi-tissue eQTLs.  $cor(tissue_p, tissue_{p'})$  represents the extent of similarity  
 464 between  $\beta_{mt-eQTLs,p}$  and  $\beta_{mt-eQTLs,p'}$ , i.e. the Pearson Correlation Coefficient between multi-  
 465 tissue eQTL effect sizes in the  $p$ th and  $p'$ th tissues, respectively. The simulation algorithm  
 466 allows multi-tissue eQTLs to have five different levels of  $cor(tissue_i, tissue_j)$  (0, 0.2, 0.4, 0.6,

467 and 0.8).  $\varepsilon_1$  is the  $N \times P$  matrix of residual errors that represent non-eQTL effects on a gene's  
468 expression level and  $vec(\varepsilon_1) \sim N(\mathbf{0}_{N \times P}, \Sigma_{P \times P}^e \otimes I_n)$

469 where  $\Sigma_{P \times P}^e = \begin{cases} 1 - h_{SNP-expression}^2, & p = p' \\ cor(tissue_p, tissue_{p'}) \times 1 - h_{SNP-expression}^2, & p \neq p' \end{cases}$ . The constant,  $1 -$

470  $h_{SNP-expression}^2$ , represents the proportion of gene expression variation that can be explained by  
471 factors other than eQTLs that can also regulate a gene's final transcription isoforms and levels.  
472 We designed the error term to have such a covariance structure that the final aggregate  
473 expression levels of the given gene in  $p$ th tissue ( $E_p$ ) was correlated with that in the  $p'$ th tissue  
474 ( $E_{p'}$ ) due to multi-tissue eQTLs as well as other biological factors. These other biological factors  
475 (such as alternative splicing events, post-transcriptional modifications and regulation of mRNA  
476 degradation) can either be shared or different across tissues. We adopted a simple assumption  
477 that the more similar a gene's expression levels are across tissues, the more likely multi-tissue  
478 eQTLs (and non-eQTL biological factors) will share effect sizes across tissues. Thus, correlation  
479 of gene expression across tissues (for example, correlation between  $E_p$  and  $E_{p'}$ ) is expected to  
480 be similar to, if not the same as, the correlation of multi-tissue eQTL effect sizes (for example,  
481 correlation between  $\beta_{mt-eQTLs,p}$  and  $\beta_{mt-eQTLs,p'}$ ) as well as the correlation between non-eQTL  
482 biological factors. All three random effect terms, i.e.  $\beta_{ts-eQTLs}$ ,  $\beta_{mt-eQTLs}$ , and  $\varepsilon_1$  were simulated  
483 using the *rmvnorm* function from the R package, *mvtnorm*. We evaluated the extent of bias  
484 between assumed combination of simulation parameters and those estimated from the empirical  
485 distribution of simulated  $E_{N \times P}$ , which met the expectation (S16 Fig).

486

487 In the special case where a gene was simulated to be expressed only in a single tissue, the  
488 model was equivalent to a univariate normal distribution with mean 0 and variance equal to the  
489 expression heritability of that gene.

490



491 Tissue specificity of genes was characterized by the number of tissues in which genes are  
492 expressed as well as the similarity of gene expression levels across tissues. Tissue specificity of  
493 eQTLs was characterized by the proportion of multi-tissue eQTLs in a gene, the number of  
494 tissues where multi-tissue eQTLs were effective, and the similarity of eQTL effect sizes across  
495 tissues.

496

497 **Phenotype.** We assumed one and only one causal tissue for a phenotype and simulated  
498 phenotype datasets for the TWAS testing dataset ( $N = 1,000$ ). This design was adopted from  
499 the simulation work of Dr. Yiming Hu *et al.* in the paper that described UTMOST [15]. Let  $E_{eQTLs}$   
500 denote the standardized genetically regulated expression component in the causal tissue. The  
501 model to simulate traits from gene expression levels can be expressed as  $Y = E_{eQTLs}b_1 + \varepsilon_2$ ,  
502 where  $Y$  is a  $1000 \times 1$  vector of standardized responses for the 1,000 individuals in the TWAS  
503 testing dataset,  $b_1$  is the  $M_{eQTLs} \times 1$  vector of gene expression effect drawn from a normal  
504 distribution with mean zero and variance  $R^2_{expression-trait}$ , and  $\varepsilon_2$  is the vector of normally-  
505 distributed errors with mean zero and variance  $1 - R^2_{expression-trait} \cdot R^2_{expression-trait}$  was  
506 assigned values in 0.001%, 0.05%, 0.5% and 1%, to represent different strengths of gene  
507 expression level-trait relations. To evaluate type I error rates,  $R^2_{expression-trait} = 0\%$   
508 corresponded to the null model where gene and trait were unrelated.

509

510 **eQTL detection.** We adopted two types of eQTL detection methods, 1) elastic net  
511 (implemented in PrediXcan [5]) and 2) group LASSO (implemented in UTMOST [15]). For ease  
512 of parallel computation, these two algorithms were adapted and integrated into the TWAS  
513 simulation tool scripts. eQTLs detected in a single tissue context (elastic net) and those  
514 detected in an integrative tissue context (group LASSO) were then used to impute GReX, and  
515 for gene-trait association analyses.

516

517 **Imputation of GReX.** Expression level of a gene can be imputed using a linear model as

518  $E = X\beta$ , where  $E$  is the  $N \times 1$  vector of imputed gene expression levels of the gene,  $X$  is the

519  $N \times M$  matrix of genotypes, and  $\beta$  is the  $M \times 1$  vector of eQTLs' estimated regulatory effects on

520 the gene, and can be obtained by either elastic net or group LASSO.

521

522 **Association analysis.** Single-tissue gene-trait associations were then estimated using SLR

523 model, i.e., *lm* function in R. Cross-tissue gene-trait association analyses were also conducted

524 in R but using PC Regression (implemented in MulTiXcan [16]) and GBJ test (implemented in

525 UTMOST [15]).

526

527 **Measures of TWAS performance.** Each combination of simulation parameters was

528 repeated 100 times independently to assess power and type I error rates at  $\alpha = 0.05$ . Estimation

529 of TWAS power was calculated as the percentage that a simulated causal gene was

530 successfully identified as statistically significant in the causal tissue in the hundred simulations.

531 Estimation of TWAS type I error rates were calculated as the percentage that a gene was falsely

532 identified as statistically significant when there was no gene-trait signal simulated in the hundred

533 simulations. We assumed that a gene is related to a trait in a single tissue, which is often the

534 case for non-pleiotropic genes. In the simulation, we knew the causal tissues for the simulated

535 traits. We calculated the false positive rates of tissues by counting the proportion of statistically

536 significant results that were in non-causal tissues.

537

538 The entire process was repeated 20 times for each combination of simulation parameters to

539 avoid sampling variability and to determine distributions of power, type I error rates, and false

540 positive rates of tissues. We further evaluated the statistical significance of the differences in

541 power and type I error rate between every pairs of TWAS methods using Wilcoxon Signed-rank  
542 test.

543

#### 544 **Evaluation of simulated genetic scenarios**

545 Trait heritability assessment validated and supported our design of simulation parameters. We  
546 designed a Monte Carlo simulation approach to randomly generate eQTL-gene-trait relations  
547 using the aforementioned simulation tool. Each replication simulated one genotypic dataset and  
548 one subsequent GReX profile for a gene. We simulated 30 non-eQTL and 30 eQTL SNPs for  
549 5,000 individuals in which MAF followed a uniform distribution of 1-50% and eQTLs explained  
550 30% of gene expression variation. The GReX profile was then used to generate 50 different  
551 traits using different random seeds. Thus, each simulated genotypic dataset had 50 estimated  
552 trait heritability values available; we took the average of these as the point estimate of the trait  
553  $h^2$  for each genotypic dataset. GCTA [40] was not appropriate for our simulation as it assumes  
554 genome-wide genotypic data. Instead, we used the R package, *regress*, to estimate trait  
555 heritability in the simulated datasets. The entire process was repeated 30 times to generate a  
556 distribution of estimated trait heritability for a given combination of simulation parameters.

557

558 To determine the influence of MAF on trait heritability, we designed different ranges of MAF  
559 distributions. MAF of SNPs followed a uniform distribution of 1-50% as in the primary TWAS  
560 performance evaluation, and also 1-20% and 1-5%. We also simulated traits where  
561  $R^2_{expression-trait} = 0\%$  (negative control), 2%, or 5% (positive controls) to support the estimation  
562 of trait heritability when  $R^2_{expression-trait} = 0.001\%$ , 0.05%, 0.5%, and 1%.

563

#### 564 **AIDS Clinical Trials Group studies**

565 The ACTG is the world's largest HIV clinical trials network. It has conducted major clinical trials

566 and translational biomedical research that have improved treatments and standards of care for  
567 people living with HIV in the United States and worldwide. In this study, we used data from four  
568 separate genotyping phases of specimens from ACTG studies in a combined dataset that  
569 comprises HIV treatment-naïve participants at least 18 years of age enrolled in randomized  
570 treatment trials [41-47]. Participants enrolled into ACTG protocols A5095, A5142, ACTG 384,  
571 A5202 or A5257. Informed consent for genetic research was obtained under ACTG protocol  
572 A5128. Clinical trial designs and outcomes, and results of a genome-wide pleiotropic study for  
573 baseline laboratory values have been described elsewhere[24,25].

574

### 575 **Genotypic data and quality control**

576 A total of 4,411 individuals were genotyped in four phases. Phase I (samples from study A5095)  
577 was genotyped using Illumina 650Y array; Phase II (studies ACTG384 and A5142) and III (study  
578 A5202) were genotyped using Illumina 1M duo array; Phase IV (study 5257) was genotyped  
579 using Illumina HumanCoreExome BeadChip. Preparation of genotypic data included pre-  
580 imputation quality control (QC), imputation, and post-imputation QC. Pre- and post-imputation  
581 QC followed the same guidelines [48] and used PLINK1.90 [49] and R programming language.  
582 Imputation was performed on the combined ACTG phase I-IV genotype dataset after pre-  
583 imputation QC, which used IMPUTE2 [50] with 1000 Genomes Phase 1 v3 [51] as the  
584 reference panel. Combined ACTG phase I-IV imputed data comprised 27,438,241 variants. The  
585 following procedures/parameters were used in the post-imputation QC by PLINK1.90: sample  
586 inclusion in the ACTG genotyping phase I-IV phenotype collection, biallelic SNP check,  
587 imputation score ( $> 0.7$ ), concordance of genetic and self-reported sex, genotype call rate ( $>$   
588 99%), sample call rate ( $> 98\%$ ), MAF ( $> 5\%$ ), and relatedness check ( $\hat{\pi} > 0.25$ ; one individual  
589 was dropped from each related pair). Subsequent principal component analysis (EIGENSOFT  
590 [52]) projected remaining individuals onto the 1000 Genomes Project Phase 3 sample space to

591 examine for population stratification. Based on percent of variance explained, the first three  
592 principal components estimated by SmartPCA in EIGENSOFT were used as covariates to  
593 adjust for population structure in the subsequent analyses. The final QC'ed ACTG phase I-IV  
594 combined imputed data comprised 2,185,490 genotyped and imputed biallelic SNPs for 4,360  
595 individuals.

596

### 597 **Phenotypic data and QC**

598 Data for 37 baseline (i.e., pre-treatment) laboratory measures were available from 5,185 HIV  
599 treatment-naive individuals in the ACTG genotyping phase I-IV datasets. We assembled these  
600 laboratory traits using a MySQL database and applied QC using R. We retained only individuals  
601 with available genotype data, and traits that were normally distributed and met the criterion of  
602 phenotype missing rate < 80%. Frequency distributions of traits were inspected using  
603 *hist\_plot.R* that facilitates manual inspection of continuous traits by providing fast, high-  
604 throughput visualization along with necessary summary statistics of each visualized traits[53].  
605 *hist\_plot.R* is part of the CLARITE [53], which is available at <https://github.com/HallLab/clarite>.  
606 We also cross-referenced the retained traits to other published work that analyzed the same  
607 traits using these clinical trials datasets [24,25]. Non-fasting serum lipid measures were retained  
608 based on data from several studies [54-56]. The final combined dataset for ACTG genotyping  
609 phases I-IV comprised 37 baseline laboratory traits (Table 2).

610

### 611 **Description of a general TSA-TWAS analytic framework**

612 The TSA-TWAS analytic framework has the following general steps.

- 613 1) Impute the GReX for the gene based on the input eQTL database(s) and the genotypic  
614 dataset.

- 615        2) Determine whether the gene is predicted to be expressed in only one tissue or in  
616            multiple tissues.
- 617        3) If the gene is predicted to be expressed in only one tissue, perform single-tissue TWAS  
618            using simple linear or logistic regression depending on the trait.
- 619        4) If the gene is predicted to be expressed in multiple tissues, perform cross-tissue TWAS  
620            using the GBJ test.
- 621        5) Repeat step 2-4 for the next gene.
- 622        6) (Optional) If there is more than one trait, repeat step 1-5 for the next trait.

623

### 624 **Imputation of GReX for genes**

625 We used GTEx v8 MASHR-based eQTLs models [57] to impute gene expression levels in a  
626 tissue-specific manner. MASHR-based eQTLs models selected variants that have biological  
627 evidence of a potential causal role in gene expression, and estimated these variants' effect  
628 sizes on gene expression levels in 49 tissues, using GTEx v8 as the reference dataset  
629 (available at <http://predictdb.org/>). The GTEx v8 MASHR-based eQTLs models were  
630 downloaded from their website on October 31, 2019. The QC'ed ACTG phase I-IV combined  
631 imputed data was used to impute the individual-level GReX in 49 human tissues.

632

### 633 **Statistical analysis for Gene-level associations**

634 We tested for single-tissue gene-trait associations by performing association tests on imputed  
635 GReX and ACTG baseline lab traits using PLATO [58,59] in 49 tissues, separately. All baseline  
636 laboratory traits were continuous and thus were modeled by linear regression with covariates.  
637 Covariates included age, sex, and the first three principal components calculated by  
638 EIGENSOFT to adjust for sampling biases and underlying population structure. For cross-tissue  
639 association analyses, we adapted the UTMOST script in R programming language and

640 performed the GBJ test for the individual-level ACTG data. The lowest p-value that can be  
641 generated by GBJ test in R is approximately  $1 \times 10^{-15}$ . No obvious inflation was observed in the  
642 TSA-TWAS framework. ACTG phenome-wide TWAS results were visualized using PhenoGram  
643 [60], a web-based, versatile data visualization tool to create chromosomal ideograms with  
644 customized annotations, available at <http://visualization.ritchielab.org/phenograms/plot>.  
645 Supplementary manhattan plot was created by *hudson*, a R package available at  
646 <https://github.com/anastasia-lucas/hudson>.

647

### 648 **Statistical correction**

649 Two strategies to correct for multiple testing were implemented in the ACTG analysis, method-  
650 wise and family-wise Bonferroni significance thresholds. The method-wise approach ascribes  
651 significance to statistical tests by controlling for the number of tests conducted in one type of  
652 method. For single-tissue gene-trait associations, the method-wise Bonferroni significance  
653 threshold was corrected for the number of genes ( $n = 483$ ) and traits ( $n = 37$ ), which resulted in  
654  $\alpha = \frac{0.05}{2,812 \times 37} \approx 4.8 \times 10^{-7}$ . For cross-tissue gene-trait associations, the method-wise Bonferroni  
655 significance threshold corrected for the number of genes and traits, which gave  $\alpha = \frac{0.05}{9,226 \times 37} \approx$   
656  $1.46 \times 10^{-7}$ . The family-wise approach assigns significance to tests by accounting for all tests  
657 performed in this study to control for FWER. Hence, single-tissue and cross-tissue association  
658 tests shared the same family-wise Bonferroni significance threshold,  $\frac{0.05}{12,038 \times 37} \approx 1.12 \times 10^{-7}$ .  
659 The significance threshold for interpreting results, by default, referred to the family-wise  
660 threshold. All results reported are exact p-values and thus, can be easily compared to either  
661 multiple testing threshold.

662

## 663 **Acknowledgements**

664 The authors are grateful to the many persons living with HIV who volunteered for ACTG  
665 protocols A5095, A5142, ACTG 384, A5202, and A5257. In addition, they acknowledge the  
666 contributions of study teams and site staff for these protocols. We thank Paul J. McLaren, PhD  
667 (Public Health Agency of Canada, Winnipeg, Canada) for prior involvement and collaborations  
668 that used these genome-wide genotype data. Study drugs were provided by DuPont  
669 Pharmaceutical Company, Bristol-Myers Squibb, Inc., Agouron Pharmaceuticals, Inc.,  
670 GlaxoWellcome, Inc., Merck and Co., Inc. Boehringer-Ingelheim Pharmaceuticals, Inc., Gilead  
671 Sciences, Inc., GlaxoSmithKline, Inc., Abbott Laboratories, Inc., Tibotec Therapeutics. The  
672 clinical trials were ACTG 384 (ClinicalTrials.gov: NCT00000919), A5095 (NCT00013520),  
673 A5142 (NCT00050895), A5202 (NCT00118898), and A5257 (NCT00811954). We thank Dr.  
674 Yiming Hu and Zhaolong Yu from Yale University and Alvaro Barbeira and Dr. Hae Kyung Im  
675 from the University of Chicago for technical support. We thank Dr. Shefali S. Verma and  
676 Anastasia Lucas from Dr. Marylyn Ritchie's lab at the University of Pennsylvania for suggestions  
677 and assistance in data visualization.

678

679 This project was supported by the National Institute of Allergy and Infectious Diseases (NIAID  
680 award number U01AI068636), the National Institute of Mental Health, and the National Institute  
681 of Dental and Craniofacial Research.

682

683 Grant support included TR000124 (to E.S.D.); AI077505, TR000445, AI069439 (to D.W.H.); and  
684 the National Institute of Allergy and Infectious Disease (NIAID award AI077505 and AI116794  
685 (to M.D.R.).

686



687 Clinical research sites that participated in ACTG protocols ACTG 384, A5095, A5142, A5202 or  
688 A5257, and collected DNA under protocol A5128 were supported by the following grants from  
689 the National Institutes of Health (NIH): A1069412, A1069423, A1069424, A1069503, AI025859,  
690 AI025868, AI027658, AI027661, AI027666, AI027675, AI032782, AI034853, AI038858,  
691 AI045008, AI046370, AI046376, AI050409, AI050410, AI050410, AI058740, AI060354,  
692 AI068636, AI069412, AI069415, AI069418, AI069419, AI069423, AI069424, AI069428,  
693 AI069432, AI069432, AI069434, AI069439, AI069447, AI069450, AI069452, AI069465,  
694 AI069467, AI069470, AI069471, AI069472, AI069474, AI069477, AI069481, AI069484,  
695 AI069494, AI069495, AI069496, AI069501, AI069501, AI069502, AI069503, AI069511,  
696 AI069513, AI069532, AI069534, AI069556, AI072626, AI073961, RR000046, RR000425,  
697 RR023561, RR024156, RR024160, RR024996, RR025008, RR025747, RR025777, RR025780,  
698 TR000004, TR000058, TR000124, TR000170, TR000439, TR000445, TR000457, TR001079,  
699 TR001082, TR001111, and TR024160.

700

## 701 Reference

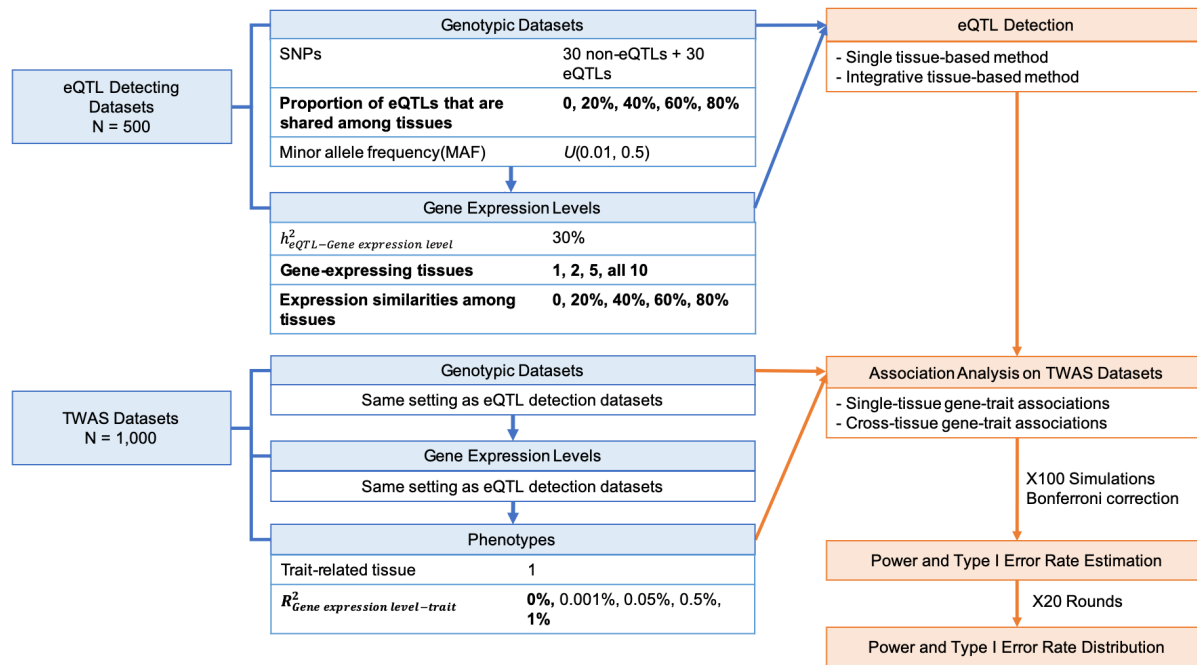
- 702 1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of  
703 GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human*  
704 *Genetics*. ElsevierCompany; 2017 Jul 6;101(1):5–22.
- 705 2. Lappalainen T. Functional genomics bridges the gap between quantitative genetics and  
706 molecular biology. *Genome Research*. Cold Spring Harbor Lab; 2015 Oct;25(10):1427–31.
- 707 3. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI  
708 Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids*  
709 *Research*. 2017 Jan 4;45(D1):D896–D901.
- 710 4. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic  
711 localization of common disease-associated variation in regulatory DNA. *Science*. American  
712 Association for the Advancement of Science; 2012 Sep 7;337(6099):1190–5.
- 713 5. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al.  
714 A gene-based association method for mapping traits using reference transcriptome data. *Nat*  
715 *Genet*. Nature Publishing Group; 2015 Aug 10;47(9):1091–8.
- 716 6. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches  
717 for large-scale transcriptome-wide association studies. *Nat Genet*. Nature Publishing Group;  
718 2016 Mar;48(3):245–52.
- 719 7. Thériault S, Gaudreault N, Lamontagne M, Rosa M, Boulanger M-C, Messika-Zeitoun D,  
720 et al. A transcriptome-wide association study identifies PALMD as a susceptibility gene for  
721 calcific aortic valve stenosis. *Nature Communications*. Nature Publishing Group; 2018 Mar  
722 7;9(1):988.
- 723 8. Wu L, Shi W, Long J, Guo X, Michailidou K, Beesley J, et al. A transcriptome-wide  
724 association study of 229,000 women identifies new candidate susceptibility genes for breast  
725 cancer. *Nat Genet*. Nature Publishing Group; 2018 Jun 18;50(7):968–78.
- 726 9. Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasaniuc B. Integrating Gene  
727 Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex  
728 Traits. *American journal of human genetics*. Elsevier; 2017 Mar 2;100(3):473–87.
- 729 10. Battle A, Brown CD, Engelhardt BE, Montgomery SB. Genetic effects on gene expression  
730 across human tissues. *Nature*. Nature Publishing Group; 2017 Oct 12;550(7675):204–13.
- 731 11. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Unraveling the  
732 polygenic architecture of complex traits using blood eQTL meta-analysis. 2018.
- 733 12. Li B, Veturi Y, Bradford Y, Verma SS, Verma A, Lucas AM, et al. Influence of tissue  
734 context on gene prioritization for predicted transcriptome-wide association studies. *Pac Symp*  
735 *Biocomput*. 2019;24:296–307.
- 736 13. Flutre T, Wen X, Pritchard J, Stephens M. A statistical framework for joint eQTL analysis  
737 in multiple tissues. Gibson G, editor. *PLoS Genet*. Public Library of Science; 2013  
738 May;9(5):e1003486.

- 739 14. Sul JH, Han B, Ye C, Choi T, Eskin E. Effectively identifying eQTLs from multiple tissues  
740 by combining mixed model and meta-analytic approaches. Schork NJ, editor. PLoS Genet.  
741 Public Library of Science; 2013 Jun;9(6):e1003491.
- 742 15. Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, et al. A statistical framework for cross-  
743 tissue transcriptome-wide association analysis. Nat Genet. Nature Publishing Group; 2019  
744 Mar;51(3):568–76.
- 745 16. Barbeira AN, Pividori M, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted  
746 transcriptome from multiple tissues improves association detection. Plagnol V, editor. PLoS  
747 Genet. 2019 Jan;15(1):e1007889.
- 748 17. Liu X, Finucane HK, Gusev A, Bhatia G, Gazal S, O'Connor L, et al. Functional  
749 Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues.  
750 American journal of human genetics. 2017 Apr 6;100(4):605–16.
- 751 18. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al.  
752 Opportunities and challenges for transcriptome-wide association studies. Nat Genet. Nature  
753 Publishing Group; 2019 Apr;51(4):592–9.
- 754 19. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al.  
755 Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics  
756 and antibody-based proteomics. Mol Cell Proteomics. American Society for Biochemistry and  
757 Molecular Biology; 2014 Feb;13(2):397–406.
- 758 20. Kryuchkova-Mostacci N, Robinson-Rechavi M. Tissue-Specificity of Gene Expression  
759 Diverges Slowly between Orthologs, and Rapidly between Paralogs. Ouzounis CA, editor. PLoS  
760 Comput Biol. Public Library of Science; 2016 Dec;12(12):e1005274.
- 761 21. Veturi Y, Ritchie MD. How powerful are summary-based methods for identifying  
762 expression-trait associations under different genetic architectures? Pac Symp Biocomput.  
763 2018;23:228–39.
- 764 22. Wheeler HE, Shah KP, Brenner J, Garcia T, Aquino-Michaels K, GTEx Consortium, et al.  
765 Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human  
766 Tissues. Montgomery SB, editor. PLoS Genet. 2016 Nov 11;12(11):e1006423–3.
- 767 23. Ongen H, Brown AA, Delaneau O, Panousis NI, Nica AC, GTEx Consortium, et al.  
768 Estimating the causal tissues for complex traits and diseases. Nat Genet. 2017  
769 Dec;49(12):1676–83.
- 770 24. Moore CB, Verma A, Pendergrass S, Verma SS, Johnson DH, Daar ES, et al. Phenome-  
771 wide Association Study Relating Pretreatment Laboratory Parameters With Human Genetic  
772 Variants in AIDS Clinical Trials Group Protocols. Open Forum Infect Dis. 2015 Jan;2(1):ofu113.
- 773 25. Verma A, Bradford Y, Verma SS, Pendergrass SA, Daar ES, Venuto C, et al.  
774 Multiphenotype association study of patients randomized to initiate antiretroviral regimens in  
775 AIDS Clinical Trials Group protocol A5202. Pharmacogenetics and Genomics. 2017  
776 Mar;27(3):101–11.

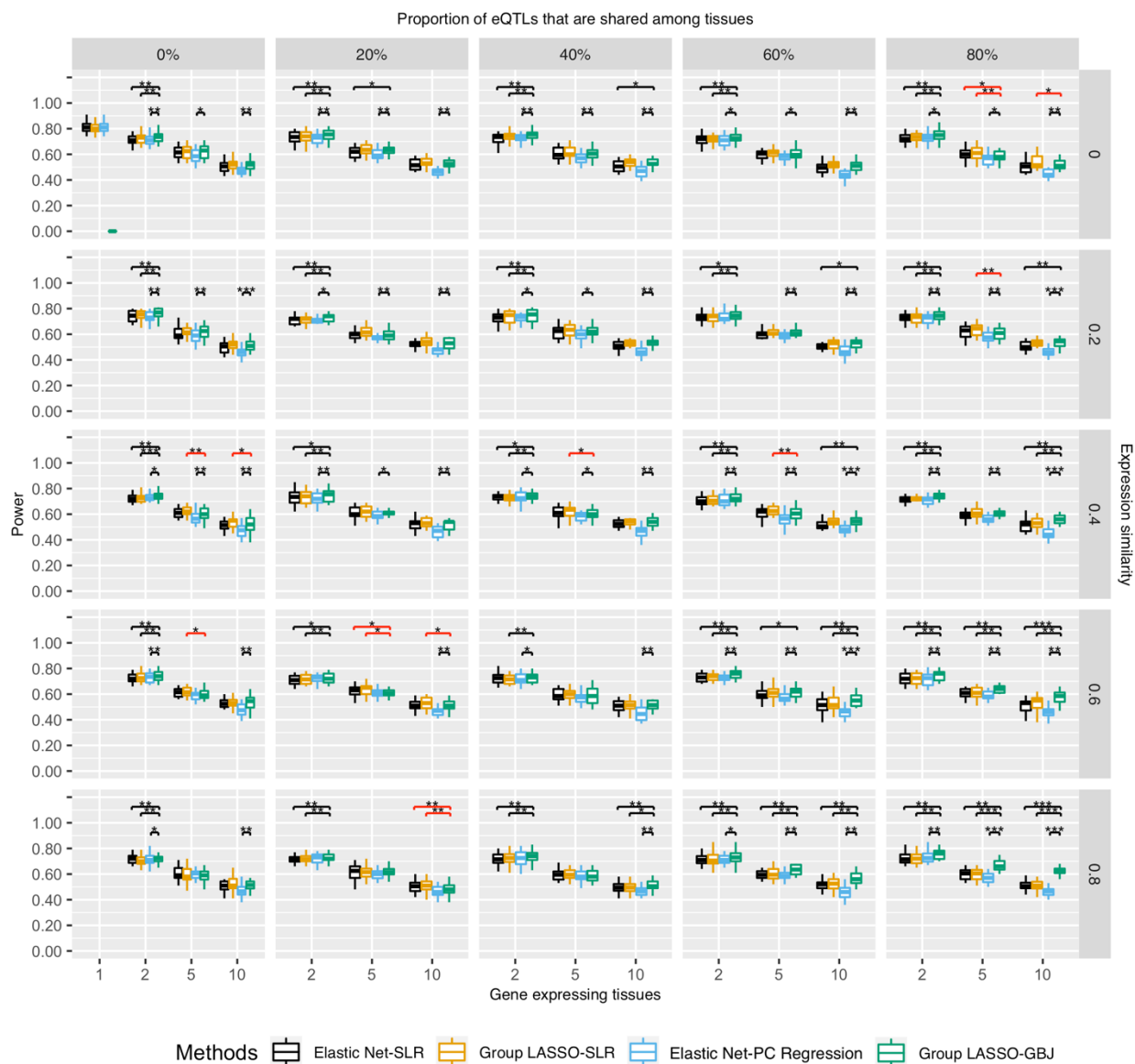
- 777 26. Coltell O, Asensio EM, Sorlí JV, Barragán R, Fernández-Carrión R, Portolés O, et al.  
778 Genome-Wide Association Study (GWAS) on Bilirubin Concentrations in Subjects with  
779 Metabolic Syndrome: Sex-Specific GWAS Analysis and Gene-Diet Interactions in a  
780 Mediterranean Population. *Nutrients*. Multidisciplinary Digital Publishing Institute; 2019 Jan  
781 4;11(1):90.
- 782 27. Dai X, Wu C, He Y, Gui L, Zhou L, Guo H, et al. A genome-wide association study for  
783 serum bilirubin levels and gene-environment interaction in a Chinese population. *Genet  
784 Epidemiol*. 2013 Apr;37(3):293–300.
- 785 28. Tukey RH, Strassburg CP. Human UDP-glucuronosyltransferases: metabolism,  
786 expression, and disease. *Annu Rev Pharmacol Toxicol*. 2000;40:581–616.
- 787 29. Barter PJ, H Bryan Brewer J, Chapman MJ, Hennekens CH, Rader DJ, Tall AR.  
788 Cholesteryl Ester Transfer Protein. *Arterioscler Thromb Vasc Biol*. Lippincott Williams & Wilkins;  
789 2003 Feb 1;23(2):160–7.
- 790 30. Chambers JC, Zhang W, Sehmi J, Li X, Wass MN, Van der Harst P, et al. Genome-wide  
791 association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat  
792 Genet*. 2011 Oct 16;43(11):1131–8.
- 793 31. Kanai M, Akiyama M, Takahashi A, Matoba N, Momozawa Y, Ikeda M, et al. Genetic  
794 analysis of quantitative traits in the Japanese population links cell types to complex human  
795 diseases. *Nat Genet*. Nature Publishing Group; 2018 Mar;50(3):390–400.
- 796 32. Le Clerc S, Coulonges C, Delaneau O, van Manen D, Herbeck JT, Limou S, et al.  
797 Screening low-frequency SNPs from genome-wide association study reveals a new risk allele  
798 for progression to AIDS. *J Acquir Immune Defic Syndr*. 2011 Mar 1;56(3):279–84.
- 799 33. Rhee EP, Ho JE, Chen M-H, Shen D, Cheng S, Larson MG, et al. A genome-wide  
800 association study of the human metabolome in a community-based cohort. *Cell Metab*. 2013 Jul  
801 2;18(1):130–43.
- 802 34. Lingwood CA, Branch DR. The role of glycosphingolipids in HIV/AIDS. *Discov Med*.  
803 *Discov Med*; 2011 Apr;11(59):303–13.
- 804 35. van Til NP, Heutinck KM, van der Rijt R, Paulusma CC, van Wijland M, Markusic DM, et  
805 al. Alteration of viral lipid composition by expression of the phospholipid floppase ABCB4  
806 reduces HIV vector infectivity. *Retrovirology*. BioMed Central; 2008 Feb 1;5(1):14–9.
- 807 36. Wu B, Ouyang Z, Lyon CJ, Zhang W, Clift T, Bone CR, et al. Plasma Levels of  
808 Complement Factor I and C4b Peptides Are Associated with HIV Suppression. *ACS Infect Dis*.  
809 2017 Dec 8;3(12):880–5.
- 810 37. Dunn SJ, Khan IH, Chan UA, Scearce RL, Melara CL, Paul AM, et al. Identification of cell  
811 surface targets for HIV-1 therapeutics using genetic screens. *Virology*. 2004 Apr 10;321(2):260–  
812 73.
- 813 38. Migueles SA, Sabbaghian MS, Shupert WL, Bettinotti MP, Marincola FM, Martino L, et al.  
814 HLA B\*5701 is highly associated with restriction of virus replication in a subgroup of HIV-  
815 infected long term nonprogressors. *PNAS*. 2000 Mar 14;97(6):2709–14.

- 816 39. Grunfeld C, Pang M, Doerrler W, Shigenaga JK, Jensen P, Feingold KR. Lipids,  
817 lipoproteins, triglyceride clearance, and cytokines in human immunodeficiency virus infection  
818 and the acquired immunodeficiency syndrome. *J Clin Endocrinol Metab.* 1992 May;74(5):1045–  
819 52.
- 820 40. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait  
821 analysis. *American journal of human genetics.* 2011 Jan 7;88(1):76–82.
- 822 41. Robbins GK, De Gruttola V, Shafer RW, Smeaton LM, Snyder SW, Pettinelli C, et al.  
823 Comparison of sequential three-drug regimens as initial therapy for HIV-1 infection. *N Engl J*  
824 *Med.* Massachusetts Medical Society; 2003 Dec 11;349(24):2293–303.
- 825 42. Gulick RM, Ribaldo HJ, Shikuma CM, Lustgarten S, Squires KE, Meyer WA, et al. Triple-  
826 nucleoside regimens versus efavirenz-containing regimens for the initial treatment of HIV-1  
827 infection. *N Engl J Med.* Massachusetts Medical Society; 2004 Apr 29;350(18):1850–61.
- 828 43. Gulick RM, Ribaldo HJ, Shikuma CM, Lalama C, Schackman BR, Meyer WA, et al.  
829 Three- vs four-drug antiretroviral regimens for the initial treatment of HIV-1 infection: a  
830 randomized controlled trial. *JAMA.* 2006 Aug 16;296(7):769–81.
- 831 44. Riddler SA, Haubrich R, DiRienzo AG, Peeples L, Powderly WG, Klingman KL, et al.  
832 Class-sparing regimens for initial treatment of HIV-1 infection. *N Engl J Med.* Massachusetts  
833 Medical Society; 2008 May 15;358(20):2095–106.
- 834 45. Sax PE, Tierney C, Collier AC, Fischl MA, Mollan K, Peeples L, et al. Abacavir-  
835 lamivudine versus tenofovir-emtricitabine for initial HIV-1 therapy. *N Engl J Med.* Massachusetts  
836 Medical Society; 2009 Dec 3;361(23):2230–40.
- 837 46. Daar ES, Tierney C, Fischl MA, Sax PE, Mollan K, Budhathoki C, et al. Atazanavir Plus  
838 Ritonavir or Efavirenz as Part of a 3-Drug Regimen for Initial Treatment of HIV-1: A Randomized  
839 Trial. *Ann Intern Med.* American College of Physicians; 2011 Apr 5;154(7):445–56.
- 840 47. Lennox JL, Landovitz RJ, Ribaldo HJ, Ofotokun I, Na LH, Godfrey C, et al. A Phase III  
841 Comparative Study of the Efficacy and Tolerability of Three Non-Nucleoside Reverse  
842 Transcriptase Inhibitor-Sparing Antiretroviral Regimens for Treatment-Naïve HIV-1-Infected  
843 Volunteers: A Randomized, Controlled Trial. *Ann Intern Med.* NIH Public Access; 2014 Oct  
844 7;161(7):461–71.
- 845 48. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al.  
846 Quality control procedures for genome-wide association studies. Haines JL, Korf BR, Morton  
847 CC, Seidman CE, Seidman JG, Smith DR, editors. *Curr Protoc Hum Genet.* 2011 Jan;Chapter  
848 1(1):Unit1.19–1.19.18.
- 849 49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a  
850 tool set for whole-genome association and population-based linkage analyses. *The American*  
851 *Journal of Human Genetics.* 2007 Sep;81(3):559–75.
- 852 50. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method  
853 for the Next Generation of Genome-Wide Association Studies. Schork NJ, editor. *PLoS Genet.*  
854 *Public Library of Science;* 2009 Jun 19;5(6):e1000529.

- 855 51. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD,  
856 Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature  
857 Publishing Group. 2010 Oct 28;467(7319):1061–73.
- 858 52. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal  
859 components analysis corrects for stratification in genome-wide association studies. Nat Genet.  
860 Nature Publishing Group; 2006 Aug;38(8):904–9.
- 861 53. Lucas AM, Palmiero NE, McGuigan J, Passero K, Zhou J, Orie D, et al. CLARITE  
862 Facilitates the Quality Control and Analysis Process for EWAS of Metabolic-Related Traits.  
863 Front Genet. Frontiers; 2019 Dec 18;10:1164.
- 864 54. Langsted A, Nordestgaard BG. Nonfasting versus fasting lipid profile for cardiovascular  
865 risk prediction. Pathology. 2019 Feb;51(2):131–41.
- 866 55. Nordestgaard BG. A Test in Context: Lipid Profile, Fasting Versus Nonfasting. J Am Coll  
867 Cardiol. 2017 Sep 26;70(13):1637–46.
- 868 56. Mora S, Chang CL, Moorthy MV, Sever PS. Association of Nonfasting vs Fasting Lipid  
869 Levels With Risk of Major Coronary Events in the Anglo-Scandinavian Cardiac Outcomes Trial-  
870 Lipid Lowering Arm. JAMA Intern Med. American Medical Association; 2019 May  
871 28;179(7):898–905.
- 872 57. Barbeira AN, Bonazzola R, Gamazon ER, Liang Y, Park Y, Kim-Hellmuth S, et al.  
873 Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. bioRxiv. Cold Spring  
874 Harbor Laboratory; 2020 May 23;42(D1):814350.
- 875 58. Hall MA, Wallace J, Lucas A, Kim D, Basile AO, Verma SS, et al. PLATO software  
876 provides analytic framework for investigating complexity beyond genome-wide association  
877 studies. Nature Communications. Nature Publishing Group; 2017 Oct 27;8(1):1167.
- 878 59. Grady BJ, Torstenson E, Dudek SM, Giles J, Sexton D, Ritchie MD. Finding unique filter  
879 sets in PLATO: a precursor to efficient interaction analysis in GWAS data. Pac Symp Biocomput.  
880 2010;;315–26.
- 881 60. Wolfe D, Dudek S, Ritchie MD, Pendergrass SA. Visualizing genomic information across  
882 chromosomes with PhenoGram. BioData Min. BioMed Central; 2013 Oct 16;6(1):18–12.
- 883
- 884
- 885



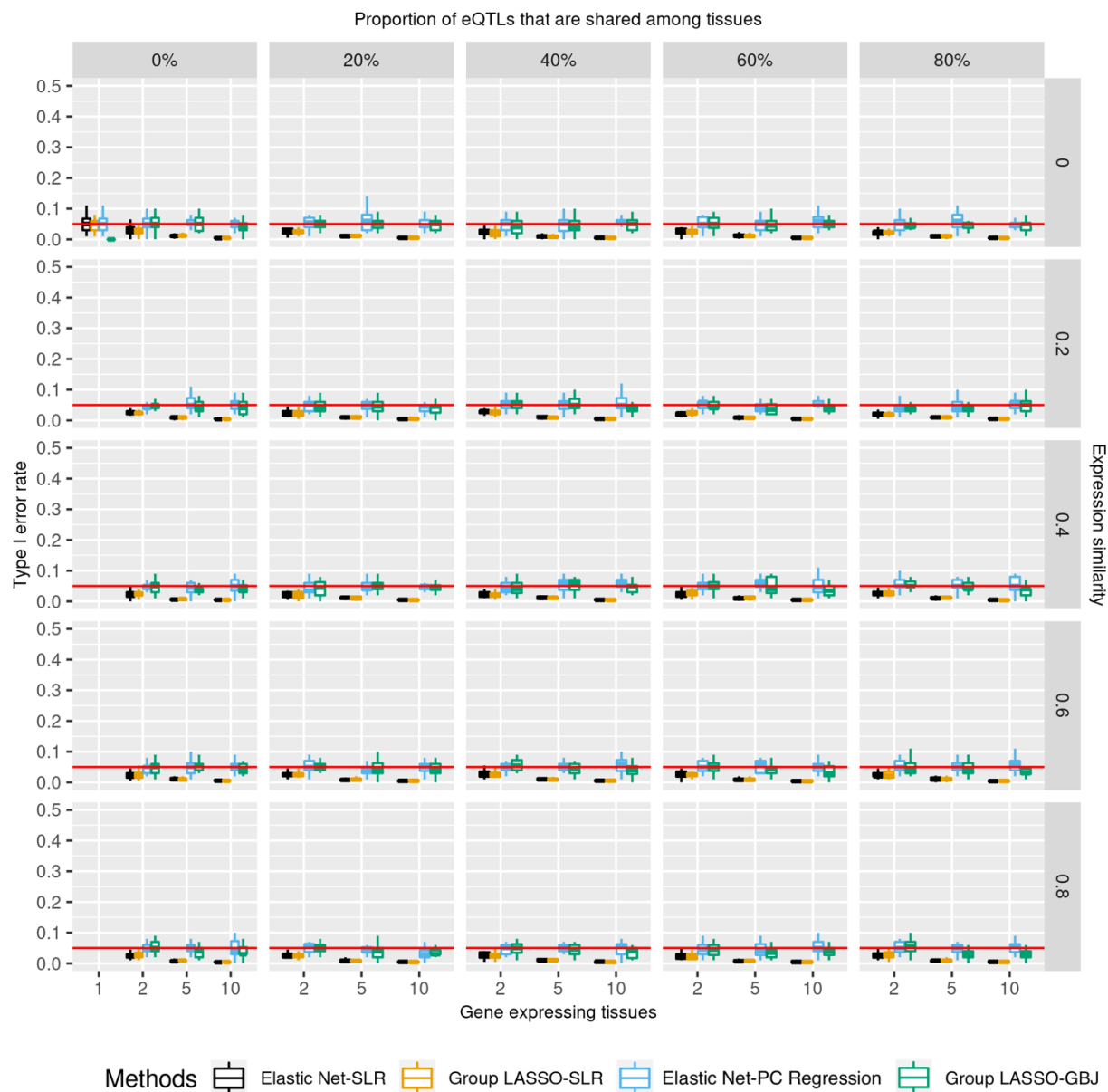
**Fig 1. Cross-tissue TWAS simulation scheme.** With the simulation parameters, we were able to generate SNP-gene-trait relations of varied tissue specificity backgrounds. In each replication, simulated datasets were divided into an eQTL detection dataset and a TWAS dataset. The former was used to identify eQTLs using different eQTL detection methods and the sample size was equivalent to that of GTEx. The detected eQTLs were then passed, separately, to the TWAS dataset to assist gene-level association tests. The TWAS dataset sample size was equivalent of that of the ACTG clinical trial dataset. Two types of gene-level association approaches estimated and ascribed p-values to the simulated gene-trait relations. In each replication, we simulated 100 different SNP-gene-trait pairs for one single point estimation of TWAS gene prioritization performance. All association p-values had been adjusted for the number of genes and tissues in each replication. 20 independent replications were conducted to obtain the distribution of TWAS performance statistics.



**Fig 2. Power of different TWAS methods in prioritizing genes of varied tissue specificity properties.** Power was the proportion of successfully identified gene-trait associations in the causal tissue out of all simulations. X-axis is the number of gene-expressing tissues. Each column stands for the proportion of eQTLs that are shared among tissues for a gene. Each row is the similarity of gene expression profiles across tissues which is estimated by correlation. Moving from the top left to the bottom right is a gradient spectrum from tissue-specific genes to broadly expressed genes. The colors represent different TWAS methods and y-axis is the

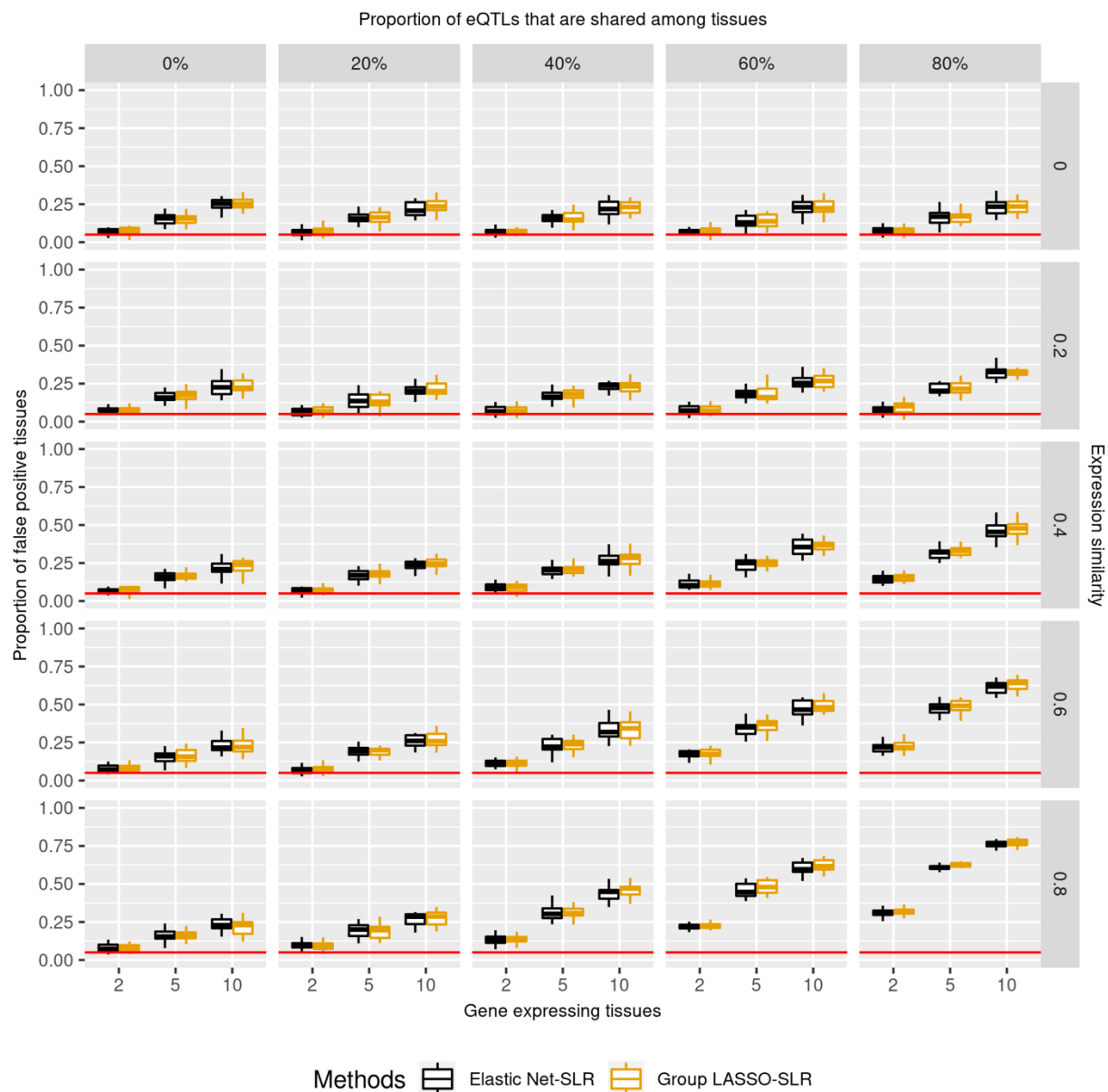


power. For tissue-specific genes at the top left, single-tissue TWAS (Elastic Net-SLR) and cross-tissue TWAS (Group LASSO-GBJ) had similar power. For broadly expressed genes at the bottom right, cross-tissue TWAS (Group LASSO-GBJ) had greater power. Brackets showed pairwise comparison of power between the Group LASSO-GBJ and other TWAS methods using Wilcoxon Signed-rank Test. Black brackets were cases where Group LASSO-GBJ had higher power than other three methods; red brackets were cases where Group LASSO-GBJ had lower power than other three methods. \*p-value < 0.05, \*\*p-value < 0.01, \*\*\*p-value < 0.0001.



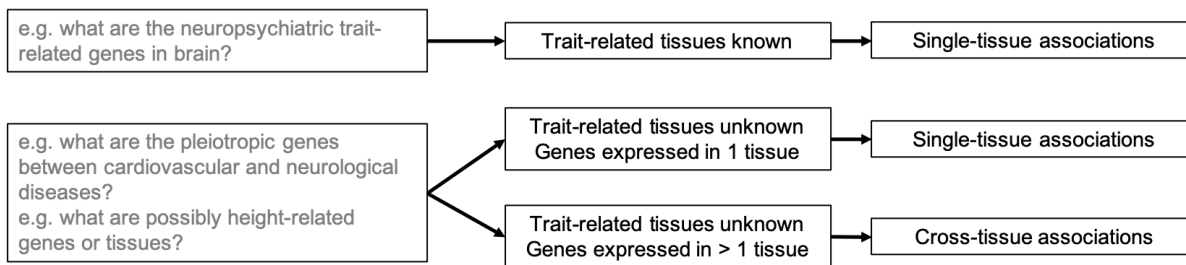
**Fig 3. Type I error rates of different TWAS methods in prioritizing genes of diverse tissue specificity properties.** Type I error rate was the probability that TWAS wrongly identified a gene-trait association as significant while there was not any signal simulated in the dataset. Association p-values were controlled for the number of genes and tested tissues. X-axis is the number of gene-expressing tissues. Each column stands for the proportion of eQTLs that are shared among tissues for a gene. Each row is the similarity of gene expression profiles across tissues which is estimated by correlation. Moving from the top left to the bottom right is a

gradient spectrum from tissue-specific genes to broadly expressed genes. The colors represent different TWAS methods and y-axis is the type I error rate. All TWAS methods had controlled type I error rates ( $\leq 5\%$ ).

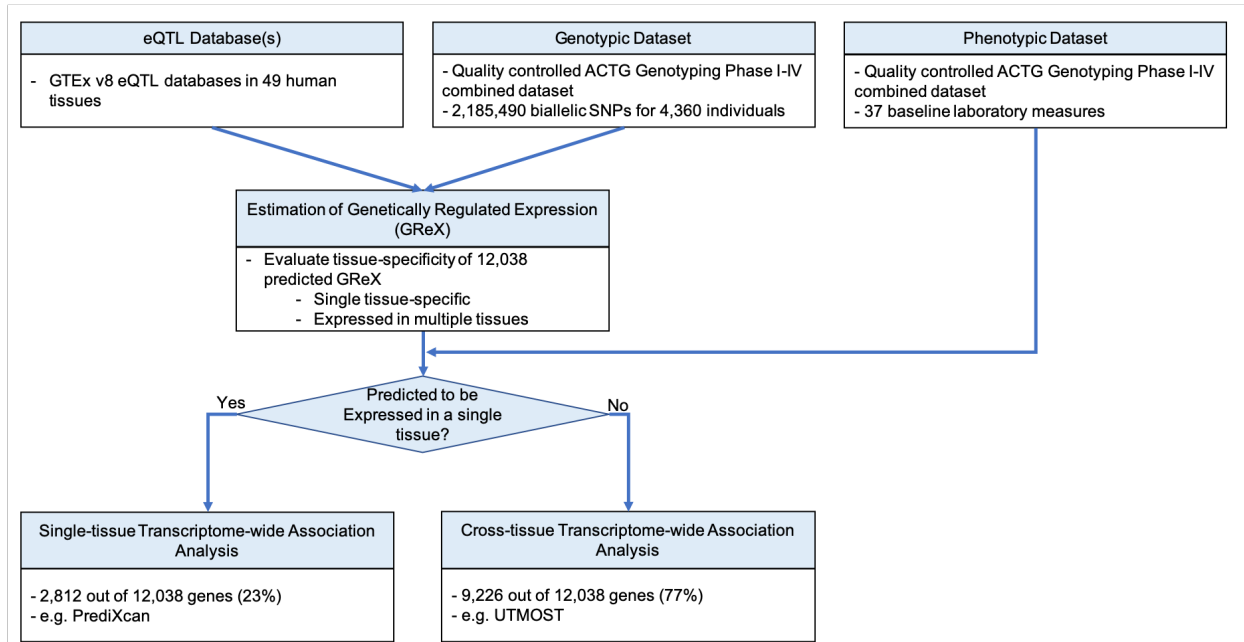


**Fig 4. False positive rates of tissues among statistically significant results.** False positive rates were the proportion of significant associations found in trait-irrelevant tissues amongst all significant results. Association p-values were controlled for the number of genes and tested tissues. X-axis is the number of gene-expressing tissues. Each column stands for the proportion of eQTLs that are shared among tissues for a gene. Each row is the similarity of gene expression profiles across tissues which is estimated by correlation. Moving from the top left to the bottom right is a gradient spectrum from tissue-specific genes to broadly expressed genes.

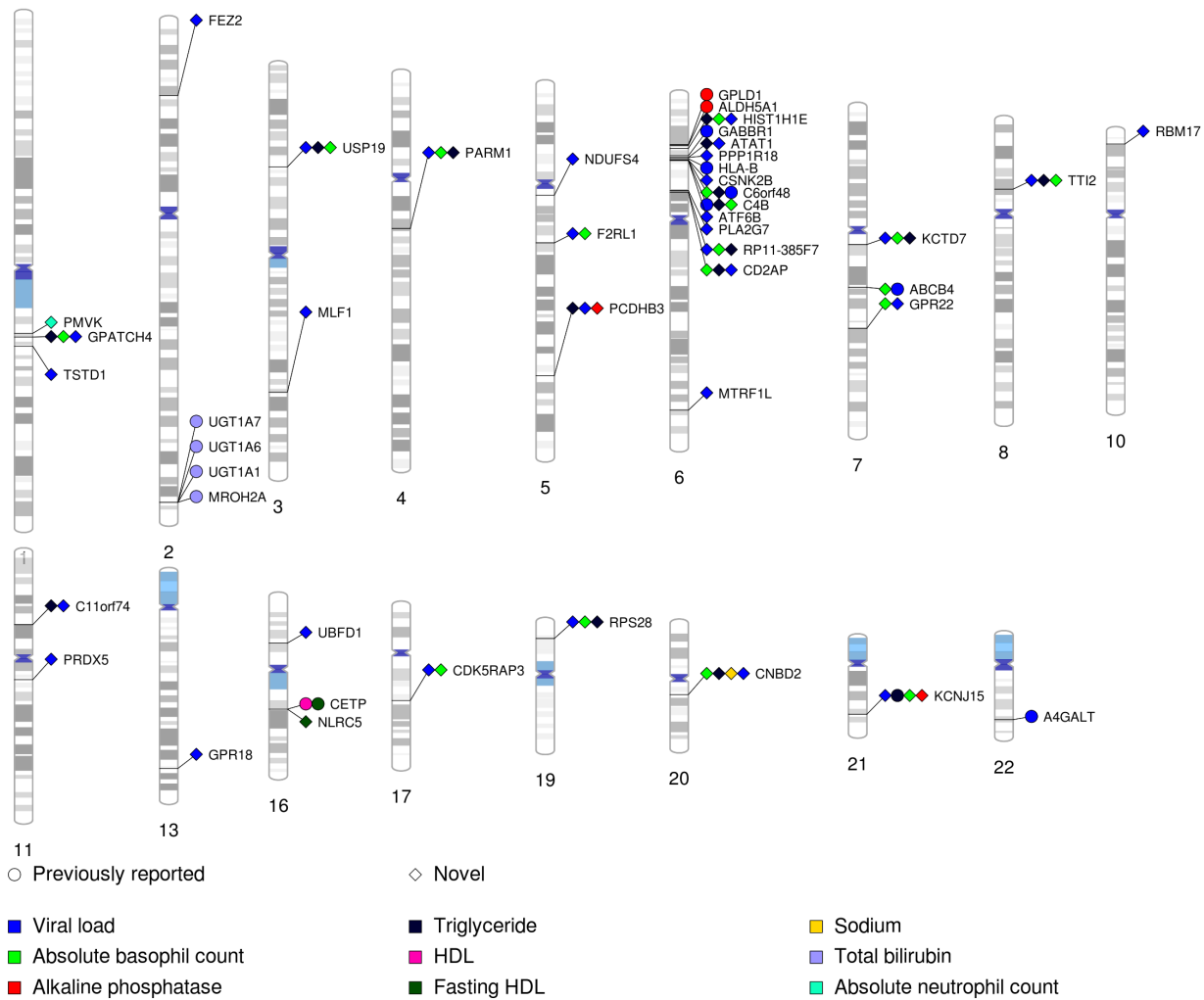
Colors represent different TWAS methods and y-axis is the false positive rate of tissues among statistically significant results. Single-tissue TWAS wrongly identified 5% and 77% trait-irrelevant tissues for tissue-specific and broadly expressed genes, respectively.



**Fig 5. A proposed TSA-TWAS analytic framework that leverages TWAS performance on genes of different tissue specificity properties.** The framework proposed based on our simulations is as follows: If trait-related tissue(s) are known for a trait or disease of interest, run single-tissue TWAS, for example, PrediXcan. If trait-related tissue(s) are unknown, run cross-tissue TWAS (UTMOST) on the genes that are expressed in more than one tissue and run single-tissue TWAS (PrediXcan) on the genes that are expressed in one single tissue.



**Fig 6. The TSA-TWAS analytic framework for the ACTG combined genotyping phase I-IV baseline laboratory traits.** Approximately 2.2 million SNPs, 4,360 individuals, and 37 baseline laboratory traits survived the QC. UTMOST eQTL models were used to impute GReX of a total of 12,038 genes in 49 tissues. 2,812 genes (23%) had GReX in one single tissue, and 9,226 genes (77%) had GReX in more than one tissue.



**Fig 7. PhenoGram of statistically significant gene-trait associations identified by the TSA-TWAS analytic framework.** We plotted the associations with  $p$ -value  $< 1.12 \times 10^{-7}$ . Each association is arranged according to the SNP location on each chromosome and the points are color-coded by baseline laboratory values. Diamonds represented previously reported or replicated associations, and circle represented novel associations identified in this study.