Powerful eQTL mapping through low coverage RNA sequencing

Tommer Schwarz [1,2], Toni Boltz [3], Kangcheng Hou [1], Merel Bot [4], Chenda Duan [5], Loes Olde Loohuis [7,8], Marco P. Boks [9], René S. Kahn [9,10], Roel A. Ophoff [1,3,11], Bogdan Pasaniuc [1,2,3,6]

1. Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, CA, USA
2. Department of Computational Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
3. Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
4. Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, California, USA
5. Department of Computer Science, University of California, Los Angeles, Los Angeles, CA USA
6. Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
7. Department of Psychiatry, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
8. Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
9. Department of Psychiatry, Brain Center University Medical Center Utrecht, University Utrecht, Utrecht, The Netherlands
10. Department of Psychiatry, Icahn School of Medicine, Mount Sinai, NY, the United States of America
11. Department of Psychiatry, Erasmus University Medical Center, Rotterdam, The Netherlands

Corresponding authors: Tommer Schwarz (tomschwarz@g.ucla.edu), Bogdan Pasaniuc (pasaniuc@ucla.edu)

ABSTRACT

Mapping genetic variants that regulate gene expression (eQTLs) in large-scale RNA sequencing (RNA-seq) studies is often employed to understand functional consequences of regulatory variants. However, the high cost of RNA-Seq limits sample size, sequencing depth, and therefore, discovery power. In this work, we demonstrate that, given a fixed budget, eQTL discovery power can be increased by lowering the sequencing depth per sample and increasing the number of individuals sequenced in the assay. We perform RNA-Seq of whole blood tissue across 1490 individuals at low-coverage (5.9 million reads/sample) and show that the effective power is higher than an RNA-Seq study of 570 individuals at high-coverage (13.9 million reads/sample). Next, we leverage synthetic datasets derived from real RNA-Seq data to explore the interplay of coverage and number individuals in eQTL studies, and show that a 10-fold reduction in coverage leads to only a 2.5-fold reduction in statistical power. Our study suggests that lowering coverage while increasing the number of individuals is an effective approach to increase discovery power in RNA-Seq studies.

KEYWORDS

RNA-Seq, Gene expression, Association testing

BACKGROUND

The vast majority of risk loci identified in genome-wide association studies (GWAS) are difficult to interpret as they lie in noncoding regions of the genome. Variants that regulate gene expression abundance, as measured through expression quantitative trait locus (eQTL) studies, provide insightful information about the functional interpretation of GWAS signals [1-2] by integrating eQTL associations with GWAS to identify target genes that are driving the GWAS signal at a locus [3-6]. RNA sequencing (RNA-Seq) is the state-of-the-art assay for measuring gene expression in bulk

73 tissue and is therefore the assay of choice for eQTL mapping [7-8] to identify population-level
74 associations between genetic variants and gene expression. However, the high cost of RNA-Seq
75 often limits the sample size and therefore reduces the discovery power of existing eQTL studies
76 based on RNA-Seq [2,6,9].

77 Traditional RNA-Seq study design prioritizes sequencing depth per individual (targeted levels of
78 coverage in the range of 30-50 million reads) over the number of individuals (samples) included
79 in the study [10-12]. However, given that high levels of coverage per individual limits the sample size
80 of the study, this results in a loss of statistical power in eQTL mapping. Previous studies have
81 established that the low-coverage whole genome sequencing of a larger number of individuals
82 attains increased power of association over higher-coverage studies of smaller sample sizes in
83 GWAS [13-17]. This raises the hypothesis that, similarly as for whole genome sequencing and
84 GWAS, lower coverage RNA-seq with a considerable increase in the number of individuals
85 sequenced could increase power of discovery in eQTL studies [18-21]. Currently, there is no
86 systematic approach for determining the optimal sample size (in terms of number of sequenced
87 individuals) and coverage to maximize eQTL discovery power.

88 In this work, we perform RNA-Seq in 1490 individuals at a lower coverage (average mapped read
89 depth of 5.9 million reads/sample) and find that eQTL discovery power is comparable to a similar
90 sized budget experiment with fewer individuals and higher coverage. Compared to high-coverage
91 RNA-Seq, we find a high degree of consistency in both the gene expression as well as eQTL
92 effects. We assess the interplay of coverage per sample and accuracy of expression estimates
93 using synthetic RNA-Seq datasets generated by the down-sampling of high-coverage data. Our
94 analyses show that a sequencing experiment conducted at a target coverage of 10 million
95 reads/sample has an average correlation per-gene of 0.40, when compared to an experiment
96 conducted with a target coverage of 50 million reads/sample. We provide evidence to show that
97 under a fixed budget, sequencing at lower coverage levels (< 10 million reads/sample) and

98 increased sample size can boost the effective sample size per unit of cost compared to standard
99 approaches to eQTL study design.

100 **RESULTS**

101 ***Low-coverage RNA-sequencing is accurate for eQTL mapping***

102 To validate the utility of low-coverage RNA-sequencing, we sequenced whole blood tissue from
103 N = 1490 unrelated individuals (**Methods**) at 5.9 million reads per individual across the coding
104 genome (**Supplementary Figure 1A** and **Supplementary Figure 1B**). We target a sequencing
105 coverage of 9.5 million reads per sample, yielding M=5.9 million reads mapped to RefSeq genes
106 on average (sd across samples of 1.96 million, **Supplementary Figure 2**). We refer to this dataset
107 as the lower-coverage RNA-Seq, or the M=5.9 million reads/sample dataset. We contrast this
108 data with a similar-sized budget RNA-Seq data at 2.4-fold higher-coverage data (M = 13.9 reads)
109 across N = 570 individuals (**Supplementary Figure 1C** and **Supplementary Figure 1D**) [22], which
110 we refer to as the higher-coverage whole blood RNA-Seq, or the M = 13.9 million reads/sample
111 dataset (Table 1).

112

Referred to as:	Coverage (million reads per sample)	Tissue	Number of samples	Library prep method
Lower-Coverage or M=5.9M reads/sample (Whole Blood)	5.9	Whole blood	1490	TruSeq Stranded plus rRNA and GlobinZero
Higher-Coverage or M=13.9M reads/sample (Whole Blood) [19]	13.9	Whole blood	570	Meta-analysis of (1) TruSeq Stranded plus rRNA and GlobinZero and (2) TruSeq Stranded polyA selected
High-	50.3	Fibroblast	155	TruSeq Stranded polyA selected

coverage (Fibroblast)				
GTEX[12]	82	Whole blood	670	TruSeq Non-stranded polyA selected
eQTLGen[13]	N/A	Whole blood	31684	Meta-analysis consisting of RNA-Seq and microarray

Table 1: RNA-Seq datasets discussed in this paper. The coverage refers to the average number of reads that successfully map to the transcriptome, except for GTEX, which refers to the median number of total reads per sample (average mapped not available). Further description of sample overlaps among cohorts in **Supplementary Note**.

First, we assess the number of genes quantified in the two datasets. We observe 40459 genes with at least one mapped read on average across samples in the whole blood high-coverage dataset, and 27308 genes with at least one mapped read on average across samples in the whole blood low-coverage dataset. Notably, when restricting to protein coding genes with at least one mapped read in high-coverage and low-coverage dataset, we find a much stronger overlap between the data sets with 18329 versus 15605 genes, respectively. This is likely due to the very sparse abundance of the non-protein coding genes leading to low-coverage being most likely to miss them. Indeed, we observe similar effects across the high vs low coverage datasets when assessing the genes with sufficient expression to be included in eQTL analysis (TPM > 0.1 in 20% of individuals, see **Methods**): 26566 genes (15496 protein coding genes) in high coverage data versus 19039 (13339 protein coding genes) in low coverage data. Most importantly we observe a high correlation in the abundance levels across the two data sets thus demonstrating that high and low coverage recover similar abundance levels (R^2= 0.91, **Figure 1A**).

Next, we investigate the power of low-coverage RNA-Seq for eQTL mapping. We conducted cis-eQTL mapping with a 1 Mb window using FastQTL, restricting to the 1490 unrelated individuals in the low-coverage data (**Methods**), to identify 7587 genes (eGenes) with a significant

135 association at FDR correction level of 5%. As expected, eQTL distribution is concentrated at
136 transcription start sites (TSS), with 73% of eGenes TSS within 250kb of the associated SNP
137 (eSNP). Repeating this approach using high-coverage whole blood data in 570 individuals, we
138 only find 5971 genes with a significant association at FDR correction level of 5%. 4969 of the
139 7587 eGenes found using the low-coverage data are also significant in the high-coverage data.
140 Of these, 2151 of the eGenes are protein coding eGenes that share the same associated eSNP,
141 and we see an extremely high level of concordance between effect sizes for these eGenes across
142 the two datasets (R^2 = 0.93, **Figure 1B**). This further indicates that low-coverage RNA-Seq is
143 robust in capturing eQTL effect sizes. 1002 genes were detected in the high-coverage eQTL
144 analysis and not in the low-coverage analysis, with 573 (of the 1002) not passing expression
145 levels (TPM >0.1 in 20% individuals) to be included in the low-coverage eQTL analysis; only 234
146 of the 573 were protein coding genes. Similar concordance is observed at the level of p-values
147 for the associations in both datasets (**Figure 1C**). Comparing the p-values for eGenes detected
148 in both eQTL analyses, the corresponding regression line has a slope of 0.39, consistent with
149 lower-coverage dataset having superior statistical power to detect associations over lower-
150 coverage dataset, and consistent with overall number of significant eQTL discoveries. We report
151 the results from using typed SNPs in these eQTL analyses (**Methods**), but observe similar
152 patterns when using the full set of imputed SNPs.

153 To further validate the performance of eQTL analysis using lower coverage RNA-Seq (coverage
154 5.9M, n = 1490), we compared the resulting eQTLs to the ones found by GTEx consortium in
155 whole blood [12] (**Supplementary Figure 3**). Restricting to the 12247 protein coding genes with
156 sufficient expression to be included in both studies (> 0.1 TPM in 20% of samples) we find that
157 4529 out of the 5957 protein coding genes (76%) with a significant association using the lower-
158 coverage data also had a significant association in GTEx, correcting at an FDR level of 5%.
159 Further restricting to eGenes with the same leading SNP in both of these datasets (140 genes)

160 (**Figure 1D**), we observe a correlation (R^2) of 0.85 between their effect sizes. Looking into the
161 1428 protein coding genes with a significant association in eQTL analysis using the lower-
162 coverage RNA-Seq but not in GTEx using an FDR cutoff of 5%, we observe that 372 have
163 significant association in GTEx using an FDR cutoff of 10%. To further ensure that these eGenes
164 are not false positives, we compare the set of 1428 genes with eQTL analysis conducted by the
165 eQTLGen Consortium [23] and find that all but 190 of these genes contain a significant association
166 in eQTLGen. This suggests that the additional associations found using lower-coverage data that
167 are not found in GTEx are not false positives, but fall just below the significance threshold in the
168 GTEx analysis.

169 Finally, we explore the impact of RNA-Seq at lower coverages for cell type expression
170 estimation. We use Cibersort [24] to compare cell-type proportion estimates between the lower-
171 coverage data and higher-coverage data (**Methods**). We find that the median estimated cell
172 type proportions are conserved across both datasets, suggesting that deconvolution of cell type
173 specific signal from gene expression profiles of whole blood samples is not impacted when
174 coverage is reduced by half (**Supplementary Figure 7**).

175 ***Impact of RNA-seq coverage on eQTL power***

176 Having demonstrated the accuracy of low-coverage RNA-Seq in eQTL mapping in real data, we
177 next focused on exploring the interplay of number of individuals and coverage for optimizing
178 power for discovery. As simulating RNA-Seq data is challenging [25-26], we down-sample reads from
179 higher-coverage RNA-Seq data to create synthetic datasets at various coverages (**Methods**). We
180 observe that with just a fraction of the reads, it is still possible to estimate gene expression (**Figure**
181 **2A**). For example, using just 10% of the data (5.0 million reads/sample) retains a per gene R^2 of
182 0.40, on average. The results from our analyses using these synthetic lower-coverage RNA-Seq
183 datasets suggest that under simplified settings of no per-sample library preparation cost, the
184 statistical power can be increased up to fourfold by spending more resources on increasing

185 sample size and fewer resources on increasing coverage. In practice, increasing the number of
186 samples in an RNA-Seq study leads to increased library preparation costs, making the increase
187 in obtainable statistical association power less obvious.

188 It has been established that statistical power in association studies is a function of sample size,
189 phenotype measurement accuracy, and genotype measurement accuracy [13,19,29]. This means that
190 the power of a study with sample size N and estimated gene expression is approximately the
191 same as the power of a study with sample size R^2*N, using the true gene expression
192 measurements. In this scenario, R^2 is the correlation between the true expression and the
193 expression estimates. We therefore report the squared correlation (R^2) between synthetic
194 datasets at various coverages and the full data at an average of 50 million reads/sample (which
195 is assumed to be the true gene expression). While these results show the mean R^2 for all genes
196 obtained under one synthetic dataset (one draw) per coverage level, we find that the synthetic
197 datasets are consistent across multiple draws at the same coverage level (**Supplementary**
198 **Figure 4A**) and each show similar correlations with the ground truth gene expression
199 (**Supplementary Figure 4B**).

200 Next, we quantified how well lower-coverage RNA-Seq can be used to detect eGenes [30]. We
201 explore the number of genes with significant associations after FDR correction at 5% under
202 various levels of simulated coverage (**Figure 2B**). Using synthetic data, as the number of reads
203 per sample decreases, we find that many eGenes are still detectable. For example, at 10 million
204 reads per sample, just 20% of the full coverage, 60% of the eGenes are still detected. In the
205 context of eQTL studies, synthetic RNA-Seq supports the idea that sequencing at lower
206 coverages over a higher number of individuals is a promising approach to boosting statistical
207 power.

208 Finally, we explore the estimation accuracy in the synthetic data as a function of relative gene
209 expression abundance, since less abundant genes may not be captured altogether at lower

210 sequencing coverages. We stratify genes into five groups based on their relative expression in
211 the full dataset (M=50.3 million reads/sample) and report the R^2 for genes in each of these groups
212 in synthetic data (**Figure 2C**). We observe that in the synthetic RNA-Seq dataset at 10 million
213 reads/sample, we capture expression of highly expressed genes better than lower expressed
214 genes. Specifically, for genes in the lowest through the highest quintiles of relative gene
215 abundance, we find the average correlation (R^2) to the ground truth of expression to be 0.36, 0.44,
216 0.61, 0.73, 0.86, respectively. We observe the same effect for synthetic datasets at coverages of
217 1 million reads/sample and 25 million reads/sample (**Supplementary Figure 5A** and
218 **Supplementary Figure 5B**). These results suggest that the ability to achieve similar power in
219 eQTL analysis studies will differ per gene, and is a function of relative expression. We further
220 investigate the properties of genes that are accurately quantified at lower coverage levels of
221 sequencing and find that that protein coding genes are more accurately quantified at lower
222 coverage levels (**Supplementary Figure 6A**). Conversely, the number of transcripts per gene,
223 gene length, and GC content do not appear to be factors that broadly influence the gene
224 quantification accuracy when sequencing coverage is reduced (**Supplementary Figure 6B**,
225 **Supplementary Figure 6C,** and **Supplementary Figure 6D**).

226 ***Optimal association power for eQTLs is attained at lower coverage with a larger number***
227 ***of samples***

228 In the context of reducing experimental costs, we explored the trade-off between the number of
229 samples sequenced and the average coverage. We evaluated the expected effective sample size
230 obtained with lower coverage per sample and compared this with a conventional approach of 50
231 million reads/sample. We down-sample reads (as described in Section 1 and **Methods**) from a
232 high-coverage RNA-Seq experiment derived from Fibroblast tissue in order to create lower-
233 coverage RNA-Seq synthetic data. This is done to match actual low coverage sequencing as

234 closely as possible. To evaluate the relationship between cost, coverage, and sample size, we
235 use the following equation to model the budget: $B = n * a + \frac{n*b*c}{d}$

236 Where *B* is the cost/budget (in US dollars), *a* is the library preparation cost per sample, *b* is the
237 target coverage of each sample (in millions of reads), *c* is the cost per lane (which contains *d*
238 million reads), and *d* is the number of reads per sequencing lane (in millions). We compute the
239 effective sample size of an eQTL study as a function of average coverage, which determines the
240 number of samples sequenced under a fixed budget (**Figure 3A**). As an example, at a fixed
241 budget of $300,000, the highest effective sample size is achieved by sequencing 2045 individuals
242 using 10 million reads per sample, which leads to a corresponding effective sample size of 1107.
243 An experiment achieving the sample effective sample size, using 50 million reads per sample,
244 would cost $426,564 (N = 1107, R^2 = 1.0). Therefore, by lowering the coverage of each sample
245 and increasing sample size, we achieve the same effective sample size at just 70.3% of the cost.

246 In practice, it is common to observe a considerable discrepancy between the target number of
247 reads in an experiment and the number of reads that successfully map to genes. This can be
248 attributed to different library prep techniques, quality of samples, or tissue type. To show how
249 mapping rate can influence the effective sample size of an experiment, we model effective sample
250 size with varying levels of mapping rates (**Methods**). As expected, we observe that as the
251 mapping rate increases, there is a corresponding increase in effective sample size (**Figure 3C**).
252 We provide a webtool as a practical approach for selecting cost-effective designs for maximizing
253 eQTL power: https://tomschwarz.shinyapps.io/RNASeqCoverageCalculator/.

254 With a budget of ~$300k and an expected mapping rate of 0.60 (chosen based on mapping rate
255 of similar experiments using TruSeq Stranded plus rRNA and GlobinZero in whole blood tissue),
256 we see the maximum effective sample size would be achieved at a target coverage of 7 million
257 reads per sample, including 2227 individuals in the study. We estimate that achieving the same

258 effective sample size using data with 50 million reads per sample would cost ~$500k (N = 1328),
259 or 1.78x the cost of sequencing 2227 individuals at a coverage of 7 million reads/sample. To
260 explore other cost scenarios we provide a webtool we created a webtool where one can enter
261 budget, costs, and other details about the experiment, in order to see how to achieve optimal
262 effective sample size (https://tomschwarz.shinyapps.io/RNASeqCoverageCalculator/).

263 **DISCUSSION**

264 In this work, we generate RNA-Seq data at a lower coverage than typically used in eQTL studies
265 (5.9M reads/sample) and demonstrate how this approach boosts effective sample size per unit
266 cost in an association study. To further validate this approach, we use synthetic RNA-Seq data to
267 show that the optimal level of coverage in an RNA-Seq project for the purpose of identifying eQTL
268 associations is lower than is commonly practiced [10-12]. Based on our findings, we recommend
269 increasing sample size while lowering sequencing depth per sample in order to achieve optimal
270 statistical power in association studies.

271 We conclude with some notes, caveats, and future directions. First, synthetic RNA-Seq via
272 downsampling reads is potentially limited in several ways. These synthetic datasets of lower
273 coverage RNA-Seq are created by uniformly sampling from real RNA-Seq data with an average
274 of 50 million reads mapped per sample. However, in practice, it is possible that sequencing biases
275 are not captured by uniform sampling due to the different experimental setup compared to the
276 dataset from which we sample [27]. Additionally, these synthetic datasets are based on data
277 obtained from fibroblast tissue with different transcriptomic profiles from whole blood, potentially
278 influencing the sequencing depth required to detect associations with gene expression. Finally,
279 this is optimized for eQTL discovery. Other mechanisms that are detected using RNA-Seq, such
280 as RNA splicing, have different mechanisms and will likely have different optimal coverages for
281 detection. The fact that we identify different sets of eGenes depending on which gene expression
282 measurements we consider (GTEX vs eQTLGen vs lower-coverage RNA-Seq), shows that we

283 need to increase cohort sizes in order to fully understand the connection between genetics and
284 gene expression in blood. Furthermore, the results in **Figure 3A** (figure showing effective sample
285 size at various coverages) indicate that even including 1490 individuals under this fixed budget is
286 not enough to achieve the optimal effective sample size. Current approaches are not sufficient to
287 understand the full landscape of eQTLs in whole blood tissue, even while only considering a single
288 genetic ancestry group. We compare the eGenes identified by GTEx, eQTLGen, and the lower-
289 coverage RNA-Seq (**Supplementary Figure 8**) and find that no single study is sufficient in
290 capturing all of the associations in whole blood. As observed in GWAS, much larger sample sizes
291 including far more ancestral diversity in these samples will enable discovery of novel associations
292 in transcriptomics. Including non-European populations and considering the temporal aspect of
293 gene expression will help us gain a better understanding of the blood transcriptome landscape in
294 the entire population.

295 **CONCLUSIONS**

296 In summary, we show that reducing coverage and increasing the number of samples in an eQTL
297 study is a valid approach for increasing effective sample size of the association study. We use
298 both real and synthetic RNA-Seq data to confirm the benefit of increased sample sizes in eQTL
299 studies. This approach can be applied to any dataset for which genotypes are available and will
300 help scientists optimize resources when measuring gene expression for the purpose of integration
301 with genetics. We provide an online tool to assist with improved design of eQTL studies at
302 https://tomschwarz.shinyapps.io/RNASeqCoverageCalculator/.

303

304 **METHODS**

305 **Cohort Description**

306 The samples included are from a study with individuals ascertained for bipolar disorder (BP). The
307 cohort consists of 916 individuals with BP, 358 controls, and 216 relatives of the individuals with
308 BP.

309 **Connection between effect size and** R^2

310 If *g* is the genotype at the SNP that we are testing for associations, and is the effect size of that
311 SNP when regressing on the true gene expression, y, and is the effect size of that SNP when
312 regressing on the estimated gene expression, $\tilde{y}$. The relationship between *y* and $\hat{y}$ is as follows
313 that $r^2 = corr(y, \hat{y})$. It follows that the estimates of effect size for a SNP on the true gene
314 expression, $\hat{\beta}$, are related to the estimate of effect size for a SNP on the estimated gene
315 expression, $\hat{\tilde{\beta}}$ as $\hat{\tilde{\beta}} = cov(g, \tilde{y}) = cov(g, ry + \varepsilon) = cov(g, ry) + cov(g, \varepsilon) = r\hat{\beta}$ where ε is a
316 random variable with mean 0 and variance 1. The association test statistics at low-coverage is
317 $x_{ground} = Ncor^2(g, y)$ thus implying that the association statistic at low coverage
318 is $x_{low-coverage} = Ncor^2(g, \tilde{y}) = N\hat{\tilde{\beta}}^2 = N(r\hat{\beta})^2 = r^2 * Ncor^2(g, y) = r^2 x_{ground}$

319 **Budget model**

320 We modeled the cost of a large-scale bulk RNA-Seq experiment based on parameters from two
321 different library prep techniques: (1) TruSeq Stranded plus rRNA and GlobinZero and (2) TruSeq
322 Stranded polyA selected, both from the UCLA Neuroscience Genomics core. Cost, or *B,* is a
323 function of the following: *a,* the library preparation cost per sample, *b,* which is the target coverage
324 of each sample (in millions of reads). *c,* the cost per lane (which contains *d* million reads), and *d*
325 is the number of reads per sequencing lane (in millions). Altogether, we model the budget as $B =$
326 $n * a + \frac{n*b*c}{d}$

327 **Genotyping pipeline**

328 Genotypes for the low-coverage whole blood samples were obtained from the following platforms:
329 OmniExpressExome (N = 810), PSYCH (N = 523), and COEX (N = 163). Given that the SNP-

330 genotype data for both the fibroblast and whole blood samples came from numerous studies using
331 various genotyping platforms (including GSA, Illumina550, OmniExpress Exome, COEX, and
332 PsychChip) the number of overlapping SNPs across all platforms was < 80k, prompting us to
333 perform imputation separately for each genotyping platform. Genotypes were first filtered for
334 Hardy-Weinberg equilibrium p value < 1.0e-6 for controls and p value < 1.0e-10 for cases, with
335 minor allele frequency (MAF) > 0.01, leaving 148613 typed SNPs.
336 Genotypes were imputed using the 1000 Genomes Project phase 3 reference panel [33] by
337 chromosome using RICOPILI v.1 [34] separately per genotyping platform, then subsequently
338 merged. Imputation quality was assessed by filtering variants where genotype probability > 0.8
339 and INFO score > 0.1, resulting in 2289732 autosomal SNPs. We restricted to only autosomal
340 due to sex chromosome dosage, as commonly done [12].

341 **Synthetic low coverage RNA-Seq**

342 We use high-coverage RNA-Seq (average of 50 million reads/sample, TruSeq Stranded plus
343 rRNA and GlobinZero library prep) from a set of 152 cell lines derived from human fibroblast cells.
344 We assume this to be the ground truth of gene expression. We used seqtk
345 (https://github.com/lh3/seqtk) to randomly sample reads at various coverages, uniformly. We
346 performed five iterations of downsampling at each level of coverage in order to account for
347 potential variability in the sampling and sequencing errors.

348 **RNA-Seq processing pipeline**

349 We used FASTQC to visually inspect the read quality from the lower-coverage whole blood RNA-
350 Seq (5.9M reads/sample) and the higher-coverage fibroblast RNA-Seq (13.9M reads/sample).
351 We then used kallisto to pseudoalign reads to the GRCh37 transcriptome and quantify estimates
352 for transcript expression. We aggregated transcript counts using scripts from the GTEX
353 consortium (https://github.com/broadinstitute/gtex-pipeline) [12].

354 **cis-eQTL mapping**

355 Excluding related individuals (pi_hat > 0.2) from the analysis, we perform cis-eQTL analysis
356 mapping using FastQTL [30], using a defined window of 1 Mb both up and downstream of every
357 gene's TSS, for sufficiently expressed genes. We run the eQTL analysis in permutation pass
358 mode (1000 permutations, and perform multiple testing corrections using the q-value FDR
359 procedure, correcting at 5% unless otherwise specified. We then restrict our associations to the
360 top (or leading) SNP per eGene.

361 **Cell type proportion estimation**

362 We estimate the proportion of cell types of both the lower-coverage and higher-coverage bulk
363 whole blood RNA-seq datasets using CIBERSORTx [35] with batch correction applied and LM22
364 signature matrix as the reference gene expression profile. The LM22 signature matrix uses 547
365 genes to distinguish between 22 human hematopoietic cell phenotypes.

366 R^2 **adjustment**

367 To account for the variability in mapping rate across different library prep techniques [37] and
368 different tissue types, we look at the mean R^2 at the expected coverage, which is calculated as
369 *expected coverage = target coverage * estimated mapping rate.* Using mean R^2 values from
370 comparing lower-coverage synthetic RNA-Seq to higher-coverage RNA-Seq real data, we fit a
371 log curve to estimate the adjusted mean R^2 (R^2_{adj}) at the expected coverage.

372 **Effective Sample Size**

373 Under a fixed-budget setting, we calculate effective sample size (N_{eff}) for a given coverage using
374 the adjusted mean R^2 (R^2_{adj}) and the number of samples included at a given coverage level (N)

375 $N_{eff} = R^2_{adj} * N$

376 **DECLARATIONS**

377 **Ethics approval and consent to participate**

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees of human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Consent for publication

Not applicable.

Availability of data and materials

Gene expression data will be made available upon publication

Competing interests

All authors declare they have no competing interest

Funding

T.S. was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under Award Number T32NS048004. T.B. was supported by the NIH (grant number 5T32HG002536-19). This research was supported by the National Institute of Mental Health of the National Institutes of Health under Award number 5R01MH115676-04. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

T.S., B.P., and R.O. initialized the study. B.P. and R.O. directed and supervised the project. R.O., R.K., and M.P.B. collected samples. M.B. prepared samples for sequencing. Bioinformatics analysis was conducted by T.S., T.B., K.H., C.D., and L.O.L. . The first draft of the manuscript was drafted by T.S. and all authors contributed to editing, revisions, and approval.

Acknowledgements

We thank the study subjects for their willingness to provide specimens and clinical data. We thank Yi Ding, Kathryn Burch, Ruthie Johnson, Arjun Bhattacharya, and Malika Freund for meaningful discussion in helping make this work possible.

SUPPLEMENTAL INFORMATION

Supplementary Figure 1: *Distribution of ancestry among samples*. **(S1A)** MDS plot of 2000 samples in 5.9M read/sample cohort. **(S1B)** Distribution of ancestry among sample in 5.9M read/sample cohort. **(S1C)** MDS plot of 759 samples in 13.9M read/sample cohort: Genotype PC1 and PC2 are projected onto PCs from 1000 Genomes Project. **(S1D)** Distribution of ancestry among sample in 13.9M read/sample cohort.

Supplementary Figure 2: *Number of pseudoaligned reads per sample.* **(S2A)** Number of pseudoaligned reads per sample in low-coverage RNA-Seq. **(S2B)** Number of pseudoaligned reads per sample in high-coverage RNA-Seq.

Supplementary Figure 3: *Real data p-value comparison scatterplot with GTEX*

Supplementary Figure 4: *Variability in correlations in synthetic data.* **(S4A)** Scatterplot of log TPM of different draws in synthetic data. **(S4B)** Distribution of correlations observed between synthetic lower-coverage RNA-Seq and ground truth.

Supplementary Figure 5: *Using synthetic data, how well do we capture expression as a function of average expression in a given gene*. **(S5A)** Correlation as a function of relative expression, at 25 million reads/sample. (**S5B)** Correlation as a function of relative expression, at 1 million reads/sample.

Supplementary Figure 6: *Using synthetic data, how well do we capture expression in different gene categories*. (**S6A)** Using synthetic data, how well do we capture expression as a function of whether a gene is protein coding or not. **(S6B)** Using synthetic data, how well do we capture expression as a function of number of isoforms in a given gene. **(S6C)** Using synthetic RNA-Seq, how well do we capture expression as a function of gene length in a given gene Gene expression estimation accuracy simulated at 10 million reads/sample as a function of relative gene length. **(S6D)** Using synthetic RNA-Seq, how well do we capture expression as a function of GC content.

Supplementary Figure 7: *Estimation of cell-type proportions*.

Supplementary Figure 8: *Overlap of significant eGenes using RNA-Seq from three different datasets*

REFERENCES

449 1. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in
450 a Caucasian population. *Nature* vol. 464 773–777 (2010).
451 2. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression
452 variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
453 3. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to
454 Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).
455 4. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association
456 studies. *Nat. Genet.* **48**, 245–252 (2016).
457 5. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using
458 reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
459 6. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait
460 loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
461 7. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics.
462 *Nat. Rev. Genet.* **10**, 57–63 (2009).
463 8. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **7**,
464 862–872 (2006).
465 9. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to
466 Function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
467 10. Hoffman, G. E. *et al.* CommonMind Consortium provides transcriptomic and epigenomic
468 data for Schizophrenia and Bipolar Disorder. *Sci Data* **6**, 180 (2019).
469 11. Franzén, O. *et al.* Cardiometabolic risk loci share downstream cis- and trans-gene
470 regulation across tissues and diseases. *Science* **353**, 827–830 (2016).
471 12. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across
472 human tissues. *Science* **369**, 1318–1330 (2020).
473 13. Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power
474 for genome-wide association studies. *Nat. Genet.* **44**, 631–635 (2012).
475 14. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and
476 coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–132 (2014).
477 15. CONVERGE consortium. Sparse whole-genome sequencing identifies two loci for major
478 depressive disorder. *Nature* **523**, 588–591 (2015).
479 16. Homburger, J. R. *et al.* Low coverage whole genome sequencing enables accurate
480 assessment of common variants and calculation of genome-wide polygenic scores. *Genome*
481 *Med.* **11**, 74 (2019).

482 17. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and
483 imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 412
484 (2021).
485 18. Baccarella, A., Williams, C. R., Parrish, J. Z. & Kim, C. C. Empirical assessment of the
486 impact of sample number and read depth on RNA-Seq analysis workflow performance. *BMC*
487 *Bioinformatics* **19**, 423 (2018).
488 19. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data.
489 *Am. J. Hum. Genet.* **69**, 1–14 (2001).
490 20. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential
491 expression in RNA-seq: a matter of depth. *Genome Res.* **21**, 2213–2223 (2011).
492 21. Robinson, D. G. & Storey, J. D. subSeq: determining appropriate sequencing depth
493 through efficient read subsampling. *Bioinformatics* **30**, 3424–3426 (2014).
494 22. Krebs, C. E. *et al.* Whole blood transcriptome analysis in bipolar disorder reveals strong
495 lithium effect. *Psychol. Med.* **50**, 2575–2586 (2020).
496 23. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Unraveling the
497 polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv*. 447367
498 (2018)
499 24. Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling
500 Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol. Biol.* **1711**, 243–259 (2018).
501 25. Williams, A. G., Thomas, S., Wyman, S. K. & Holloway, A. K. RNA-seq Data: Challenges
502 in and Recommendations for Experimental Design and Analysis. *Curr. Protoc. Hum. Genet.* **83**,
503 11.13.1–20 (2014).
504 26. Frazee, A. C., Jaffe, A. E., Langmead, B. & Leek, J. T. Polyester: simulating RNA-seq
505 datasets with differential transcript expression. *Bioinformatics* **31**, 2778–2784 (2015).
506 27. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
507 quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
508 28. Liu, Y. *et al.* Evaluating the impact of sequencing depth on transcriptome profiling in
509 human adipose. *PLoS One* **8**, e66883 (2013).
510 29. Mandric, I. *et al.* Optimized design of single-cell RNA sequencing experiments for cell-
511 type-specific eQTL analysis. *Nat. Commun.* **11**, 5504 (2020).
512 30. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient
513 QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).

514 31. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an
515 assessment of technical reproducibility and comparison with gene expression arrays. *Genome*
516 *Res.* **18**, 1509–1517 (2008).
517 32. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-
518 Based Linkage Analyses. *The American Journal of Human Genetics* vol. 81 559–575 (2007).
519 33. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. A global reference for
520 human genetic variation. *Nature* vol. 526 68–74 (2015).
521 34. Lam, M. *et al*. RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics* **36**,
522 930-933 (2019).

523 35. Newman, A.M. *et al*. Determining cell type abundance and expression from bulk tissues
524 with digital cytometry. *Nature Biotechnology* **37**, 773-782 (2019).

525 36. Hoffman, G. E. et al. CommonMind Consortium provides transcriptomic and epigenomic data
526 for Schizophrenia and Bipolar Disorder. Sci. Data 6, 180 (2019).

527 37. Aguet, F. et al. Genetic effects on gene expression across human tissues. Nature 550, 204–
528 213 (2017)

529
530 **DISPLAY ITEMS**
531
532
533 Table 2: Sequencing cost scenarios (Figure 3):
534

	Cost per lane	Cost per sample
Scenario 1	$1790	$87
Scenario 2	$1790	$30
Scenario 3	$1790	$150
Scenario 4	$1000	$150

535
536 The cost parameters corresponding to the effective sample size scenarios in Figure 3. Cost per
537 sample reflects the cost of library prep to include an additional sample. Cost per lane reflects the
538 cost per sequencing lane, which allows for 300 million reads.
539

Figure 1: *Concordance of eQTL discovery when using lower-coverage RNA-Seq vs higher-coverage RNA-Seq* **(1A):** Restricting to the 20735 genes with sufficient expression levels to be included in eQTL analysis in both the 5.9M read/sample and 13.9M read/sample dataset, comparison of the median expression (log TPM) across samples, of every gene. $R^2 = 0.91$. **(1B):** In real data, scatterplot of effect sizes of most significant eQTL hits for the 2151 protein coding genes with the same eQTL hit in both eQTL analyses performed (low-coverage and high-coverage). On the x-axis, we show the effect sizes for these genes using low-coverage RNA-Seq, on the y-axis we show the effect sizes for these genes using high-coverage RNA-Seq. **(1C):** Real data p-value comparison scatterplot: In real data, scatterplot of -log p-values of most significant eQTL hit for 13950 genes included in both eQTL analyses performed (low-coverage and high-coverage). On the x-axis, we show the -log p-values for these genes using low-coverage RNA-Seq, on the y-axis we show the -log p-values for these genes using high-coverage RNA-Seq. The dotted line shows $y = x$, while the solid line shows the line of best fit for the 3985 protein-coding eGenes with a significant eQTL hit in both datasets. **(1D):** In real data, scatterplot of effect sizes of the most significant eQTL hit for the 140 eGenes with the same leading SNP identified in both eQTL analyses performed (lower-coverage RNA-Seq with 5.9M reads/sample and GTEX). On the x-axis, we show the effect size for these eGenes from eQTL analysis conducted using the 1490 individuals of EUR ancestry and imputed genotypes, and on the y-axis we show the effect sizes for these eGenes from eQTL analysis published by the GTEX Consortium.

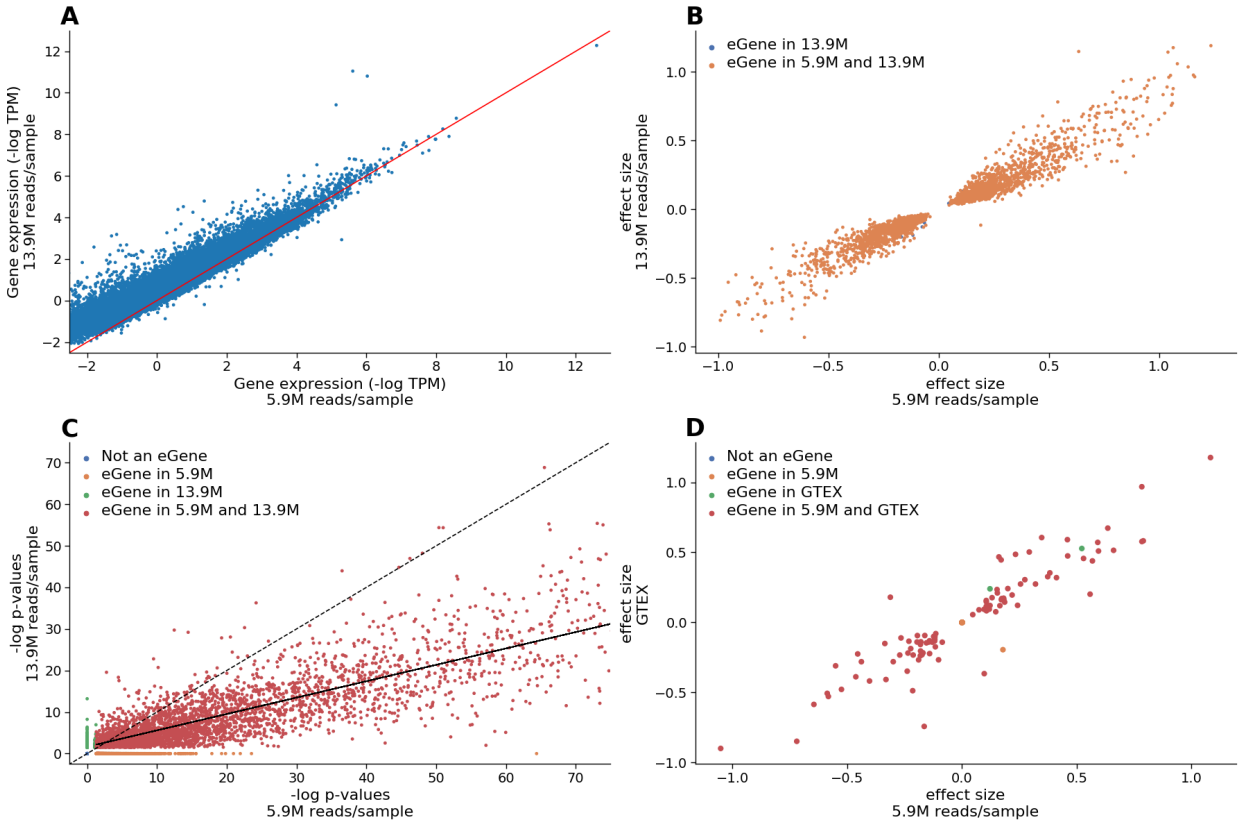


Figure 2: *Synthetic lower-coverage RNA-Seq captures expression signal* **(2A):** On the x-axis, we show the level of simulated coverage, and on the y-axis we show the mean Pearson correlation of every gene. We calculate this value by finding the R^2 values for the TPM values of each of

45,910 genes across 155 samples between the high coverage data (average of 50 million reads per sample) and the simulated data, and reporting the mean R^2 value per gene. **(2B):** For a fixed number of individuals, absolute number and percentage of eGenes captured at 5% FDR, for synthetic RNA-Seq at varying levels of coverage. **(2C):** Gene expression accuracy as a function of relative gene expression observed in actual RNA-Seq data with 50 million reads/sample. 23,540 genes (with average expression < 0.1 TPM) are divided into five ascending quintiles of expression based on their average expression in 155 samples.

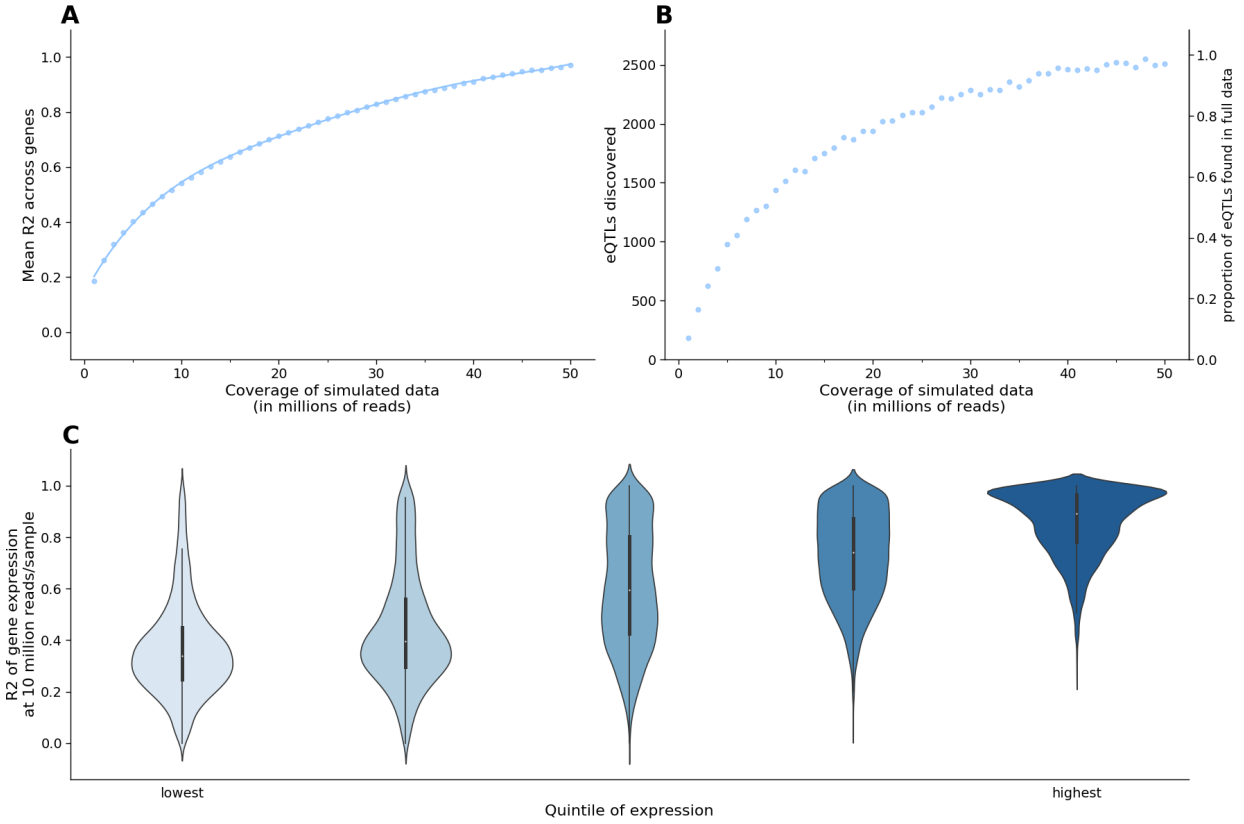
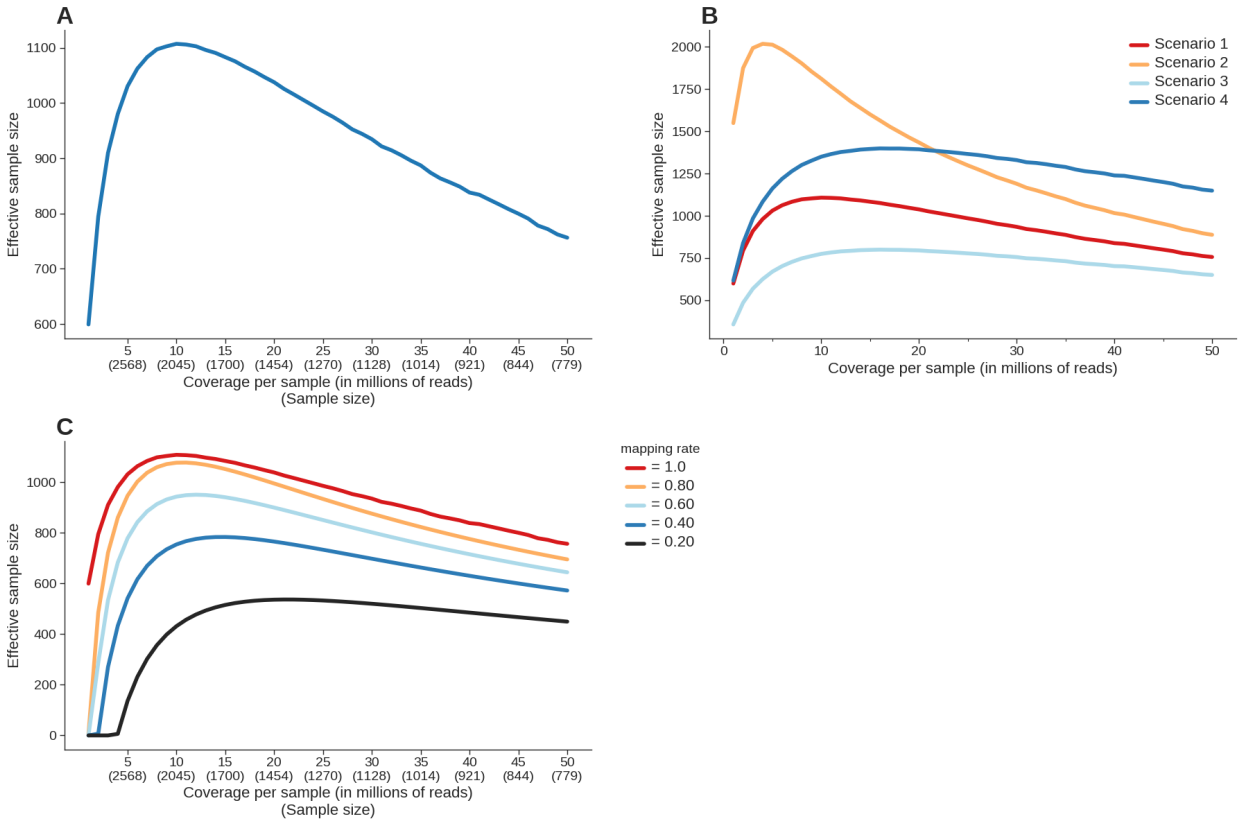


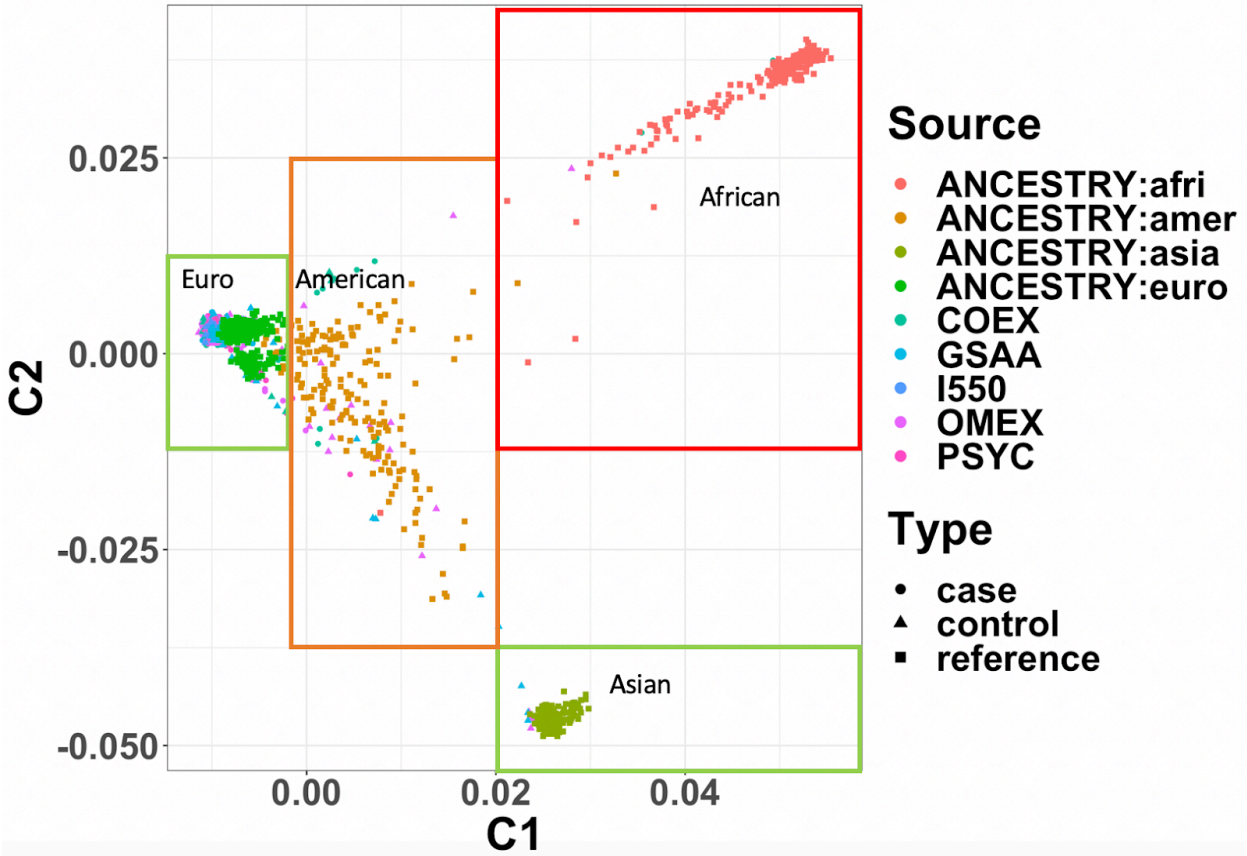
Figure 3: *Effective sample size under various budget parameters*. **(3A):** Effective sample size in RNA-Seq under a fixed budget ($300,000) as a function of the number of samples and the resulting coverage. Cost assumptions: $87 per library prep per sample, $1790 per lane of sequencing (300 million reads). **(3B):** Effective sample size in RNA-Seq under a fixed budget ($300,000) as a function of the number of samples and the resulting coverage. Cost assumptions vary and are reflected in Table 2. **(3C):** Effective sample size under a fixed budget ($300,000) as a function of the number of samples and the results coverage. A global mapping rate parameter is used to simulate actual experimental conditions (Methods).



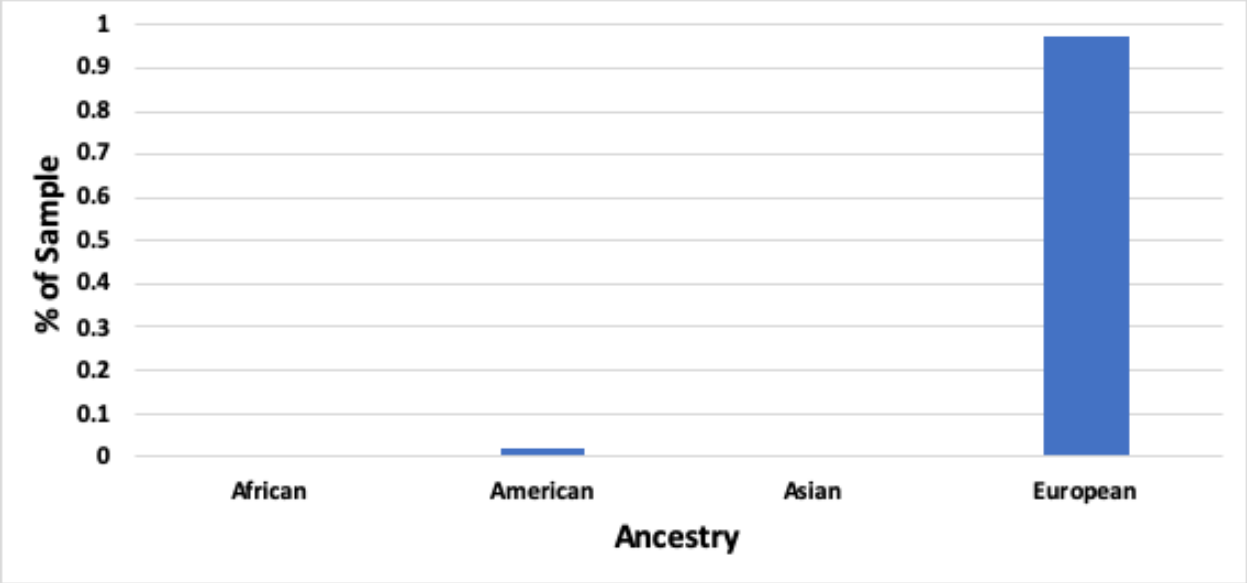
SUPPLEMENTARY FIGURES

Supplementary Figure 1: *Distribution of ancestry among samples*. **(S1A)** Genotype PC1 and PC2 are projected onto PCs from 1000 Genomes Project. Points labeled with "ANCESTRY" are from 1000 Genomes Project, remaining points designate the specific genotyping platform used in our cohort. Boxes are drawn around the centers to show where samples from the n = 2000 / 5.9M reads/sample cohort lie. **(S1B)** A barplot showing the distribution of ancestry observed in the n = 2000 / 5.9M reads/sample cohort, according to the MDS plot. Note that only the 1963 samples that pass genotype QC thresholds are included here. Exact numbers of samples per ancestry group are: African - 4, American - 34, Asian - 9, European – 1916. **(S1C)**: Genotype PC1 and PC2 are projected onto PCs from 1000 Genomes Project. Points labeled with "ANCESTRY" are from 1000 Genomes Project, remaining points designate the specific genotyping platform used in our cohort. Boxes are drawn around the centers to show where samples from the n = 759 / 13.9M reads/sample cohort lie. **(S1D)** A barplot showing the distribution of ancestry observed in the n = 759 / 13.9M reads/sample cohort, according to the MDS plot. Exact numbers of samples per ancestry group are: African - 4, American - 19, Asian - 1, European – 735.

607 **S1A**:

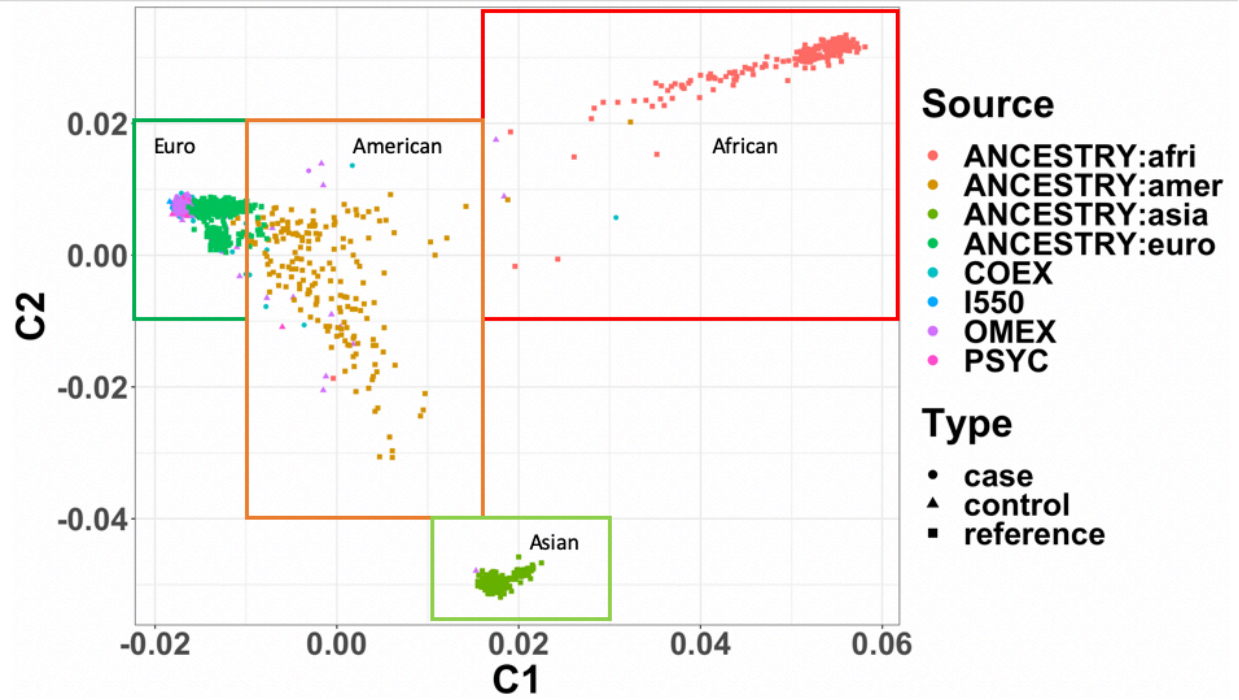


608
609
610 **S1B**:

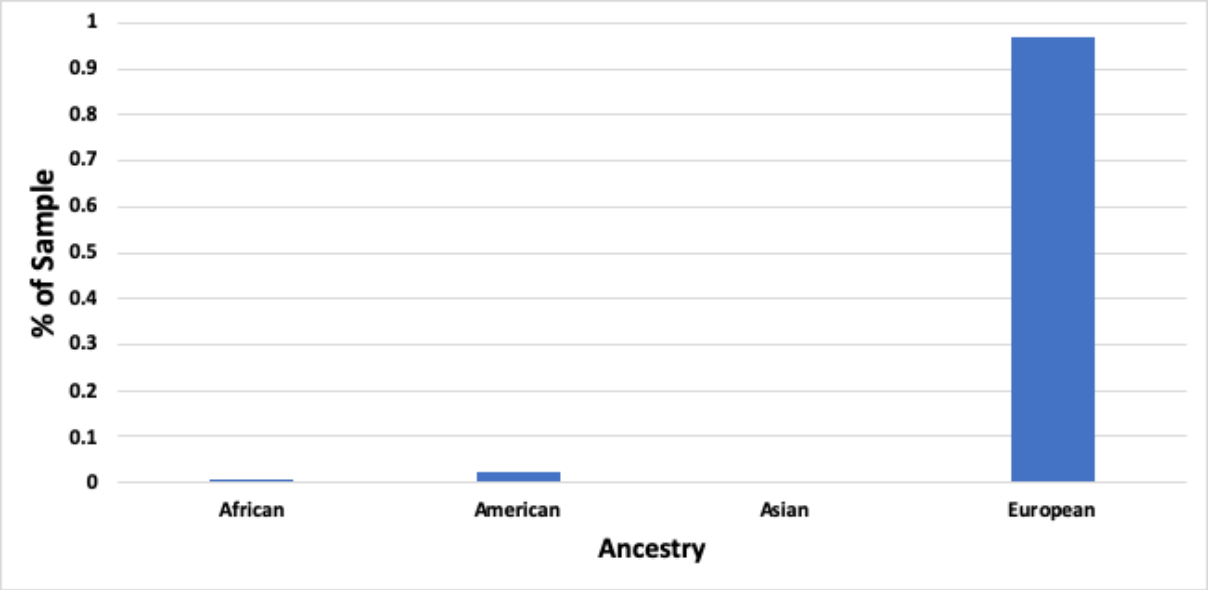


611
612

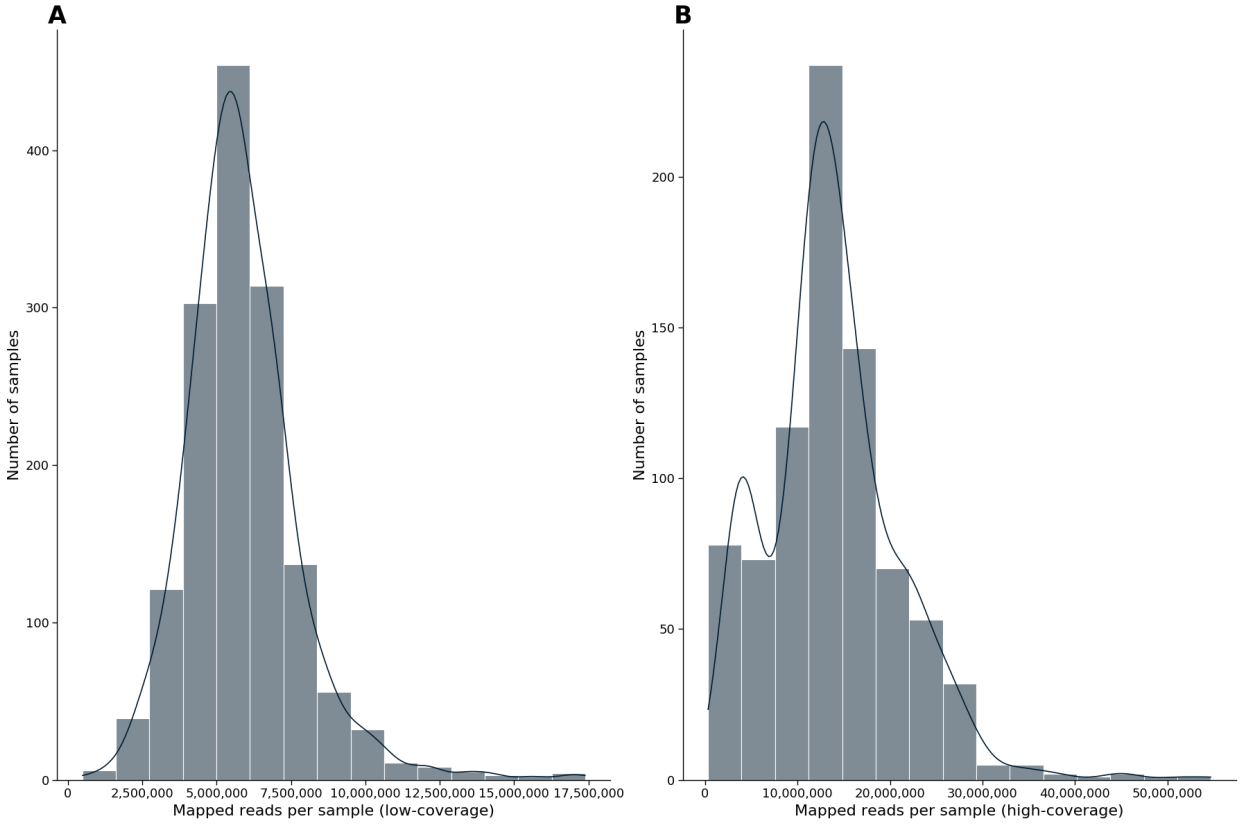
613 **S1C**:



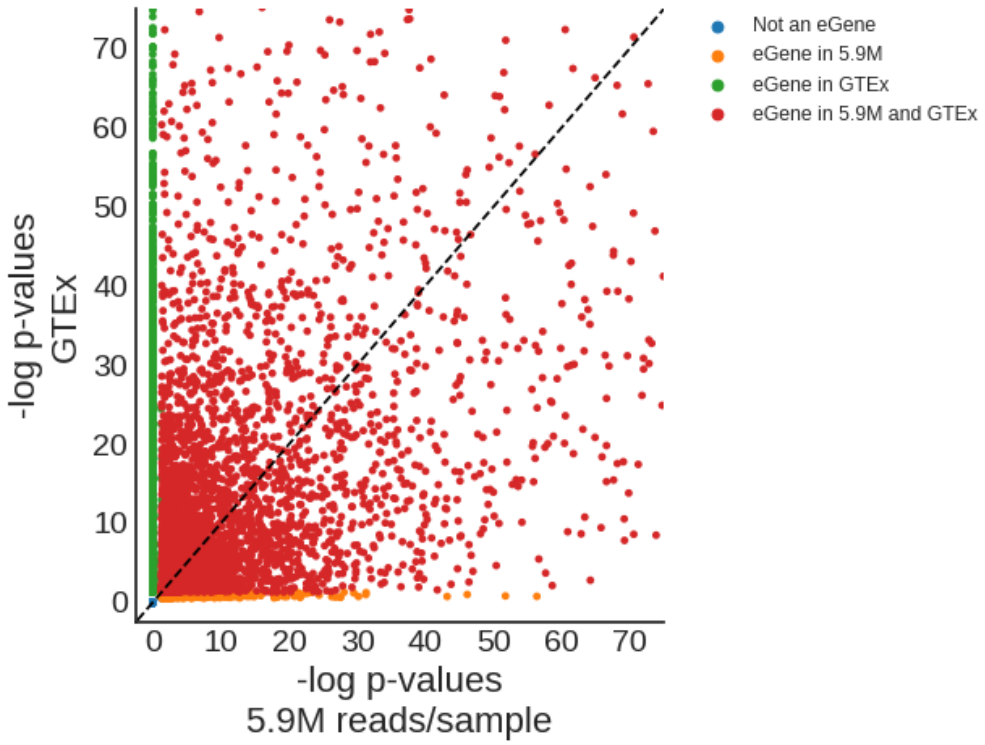
614
615
616 **S1D**:



617
618
619 **Supplementary Figure 2:** *Number of pseudoaligned reads per sample for low-coverage and*
620 *high-coverage experiments*. **(S2A)** In real data, a histogram showing the number of reads mapped
621 to genes (or in kallisto terms: number of reads for which transcriptome successfully mapped), per
622 sample. **(S2B)** In real data, a histogram showing the number of reads mapped to genes (or in
623 kallisto terms: number of reads for which transcriptome successfully mapped), per sample.
624

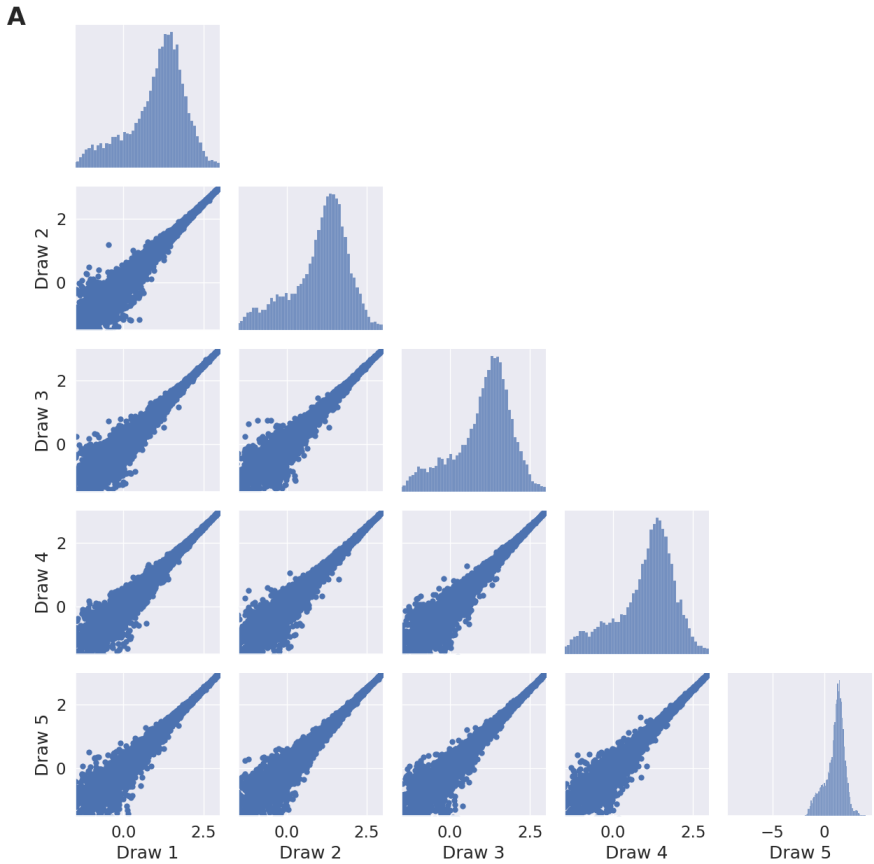


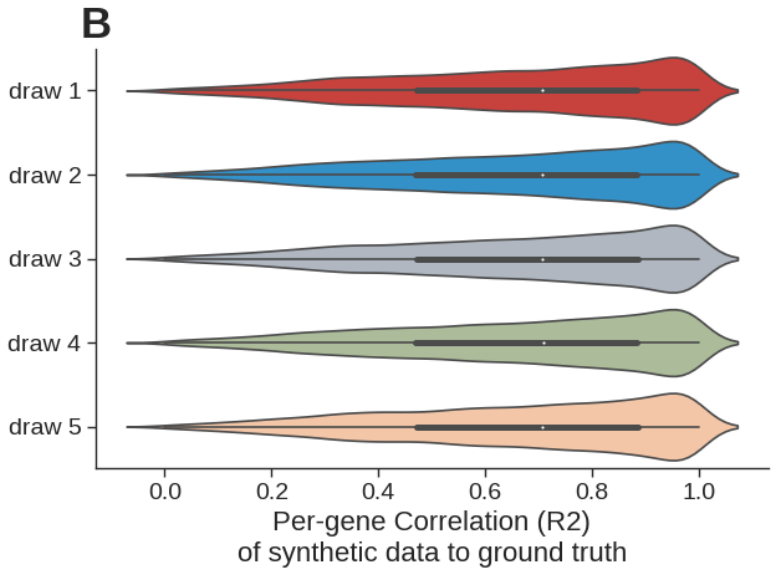
Supplementary Figure 3: *Real data p-value comparison scatterplot with GTEX*. **(S3A)** Using the 12,496 protein-coding genes included both in GTEX and the low-coverage datasets, on the x-axis, we show the -log p-values for leading SNP eQTL associations in the low-coverage dataset. On the y-axis, we show the -log p-values for leading SNP eQTL associations in the GTEX dataset.



634

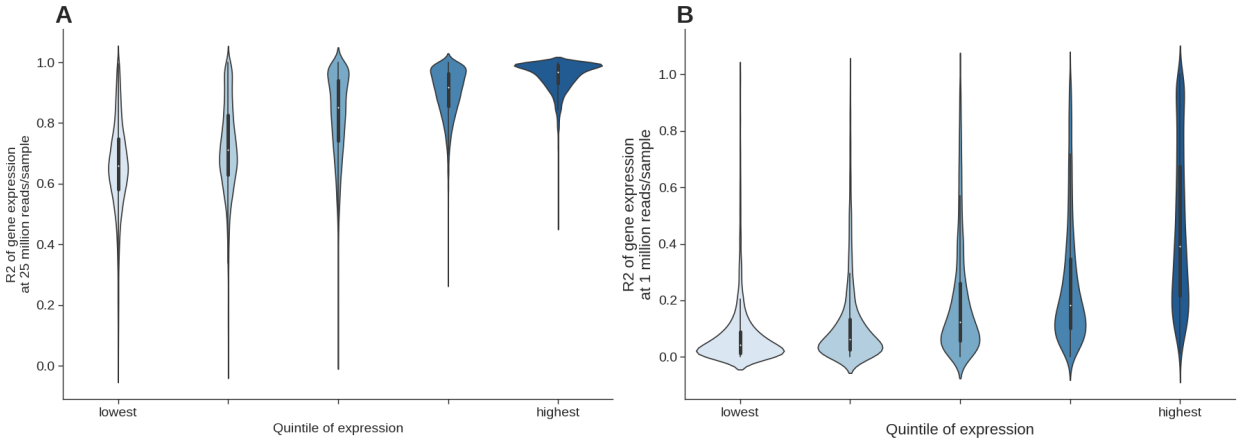
635 **Supplementary Figure 4:** V*ariability in correlations in synthetic data*. **(S4A)** For synthetic data
636 corresponding to one sample, a comparison of estimated log TPM values between five different
637 uniform sampling draws at 10 million reads/sample, for 14,948 protein-coding genes. **(S4B)** For
638 14,948 protein-coding genes estimated across five different uniform sampling draws at 10 million
639 reads/sample, we compare the distribution of correlation (R^2) between the estimated expression
640 of the samples and the ground truth gene expression.
641



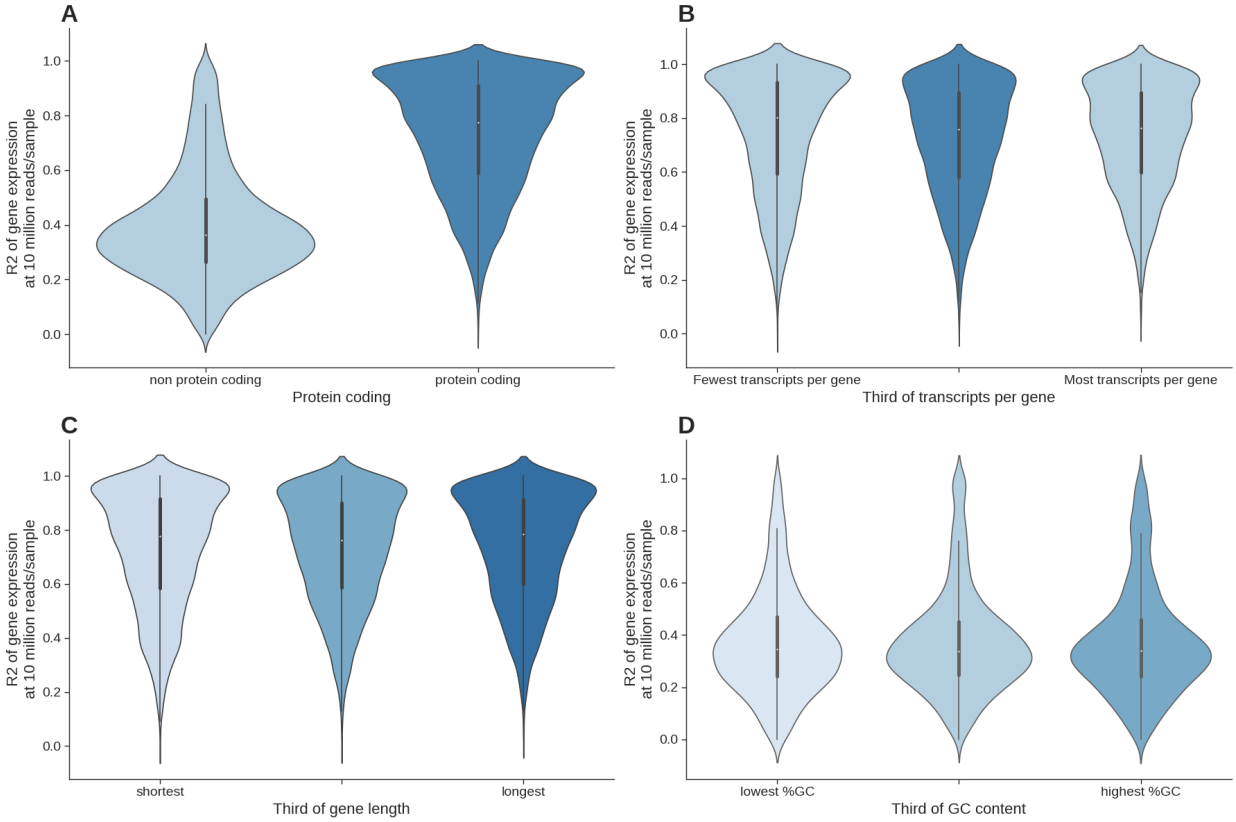


Supplementary Figure 5: *Using synthetic data, how well do we capture expression as a function of average expression in a given gene*. **(S5A)** Gene expression accuracy using data simulated with 1 million reads/sample, as a function of relative gene expression observed in actual RNA-Seq data with 50 million reads/sample. 23,043 genes (with average expression < 0.1 TPM) are divided into five ascending quintiles of expression based on their average expression in 155

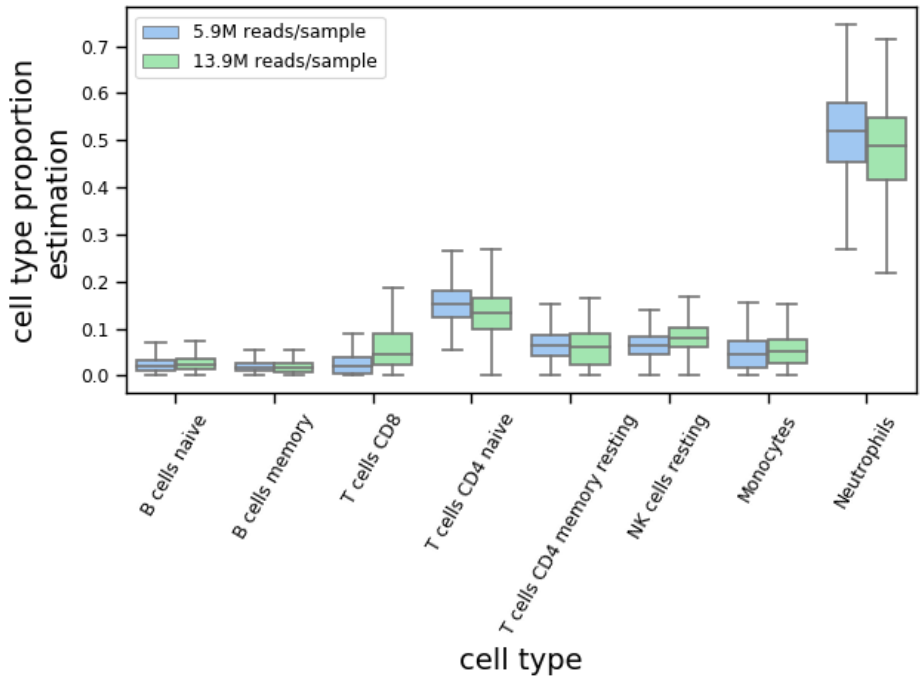
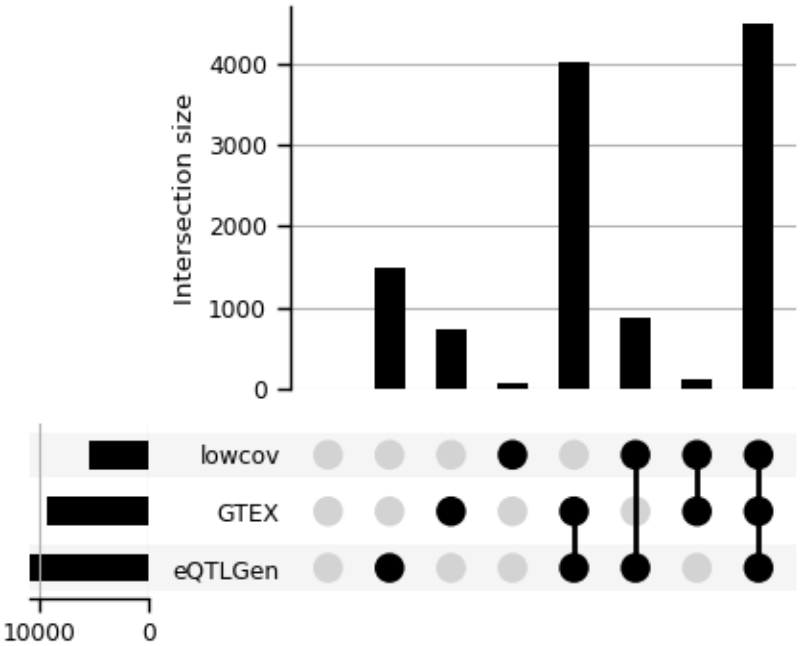
samples. **(S5B)** Gene expression accuracy using data simulated with 1 million reads/sample, as a function of relative gene expression observed in actual RNA-Seq data with 50 million reads/sample. 23,043 genes (with average expression < 0.1 TPM) are divided into five ascending quintiles of expression based on their average expression in 155 samples.



Supplementary Figure 6: *Using synthetic data, how well do we capture expression in different gene categories*. **(S6A)** Gene expression estimation accuracy simulated at 10 million reads/sample as a function of whether a gene codes for a protein. 24,093 genes (with average expression < 0.1 TPM) are divided into two groups. **(S6B)** Gene expression estimation accuracy simulated at 10 million reads/sample as a function of how many transcripts each gene has. 23,540 genes (with average expression < 0.1 TPM) are divided into three ascending groups based on the number of transcripts contained in each gene. **(S6C)** Gene expression estimation accuracy simulated at 10 million reads/sample as a function of relative gene length. 14,484 genes (with average expression < 0.1 TPM, protein coding) are divided into three groups based on the length of each gene. **(S6D)** Gene expression accuracy as a function of relative GC content. 5,771 genes (with average expression < 0.1 TPM and GC content reported) are divided into three groups based on the length of each gene.



Supplementary Figure 7: *Estimation of cell-type proportions.* **(S7)** In real data, a comparison of estimated cell type proportions from Cibersortx between lower-coverage (5.9M reads/sample) and higher-coverage (13.9M reads/sample) RNA-Seq data for the eight most common cell types in whole blood tissue.



Supplementary Figure 8: *Overlap of significant eGenes using RNA-Seq from three different datasets*. **(S8)** Comparing number of genes with significant associations between three datasets: (1) Lower-coverage RNA-Seq (5.9M reads/sample on average, across 1,496 individuals), (2) GTEX (83M reads/sample on average, across 670 samples), (3) eQTLGen (31,684 individuals, mix of RNA-Seq and MicroArray assays used).