Supplementary Information for:

Chapter 3 | Genetic control of the transcriptional response to active tuberculosis disease and treatment

This PDF file includes:

Supplementary Figure S3.14

Supplementary Figure S3.15

Supplementary Figure S3.16

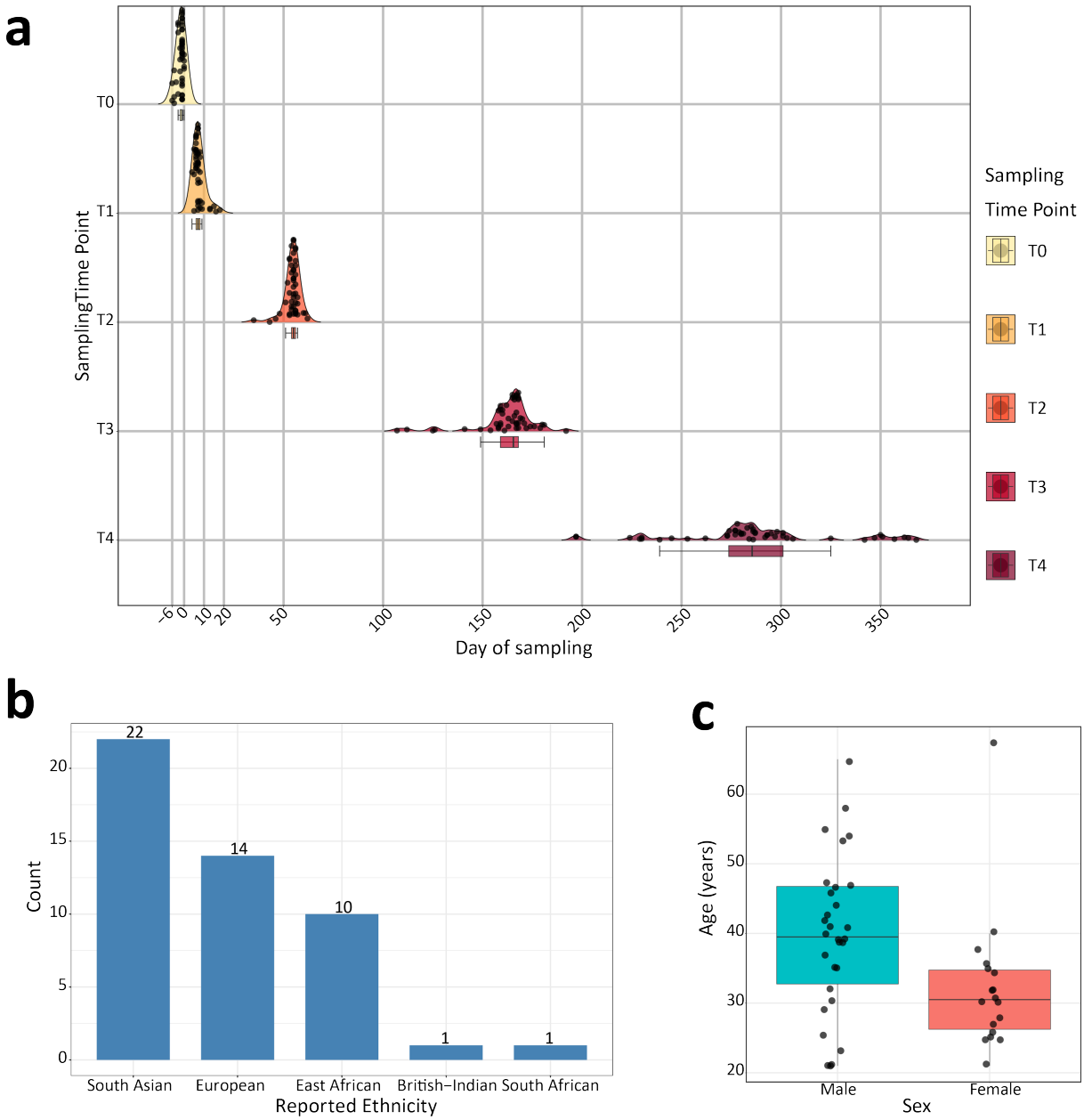
Supplementary Figure S3.17

Supplementary Notes

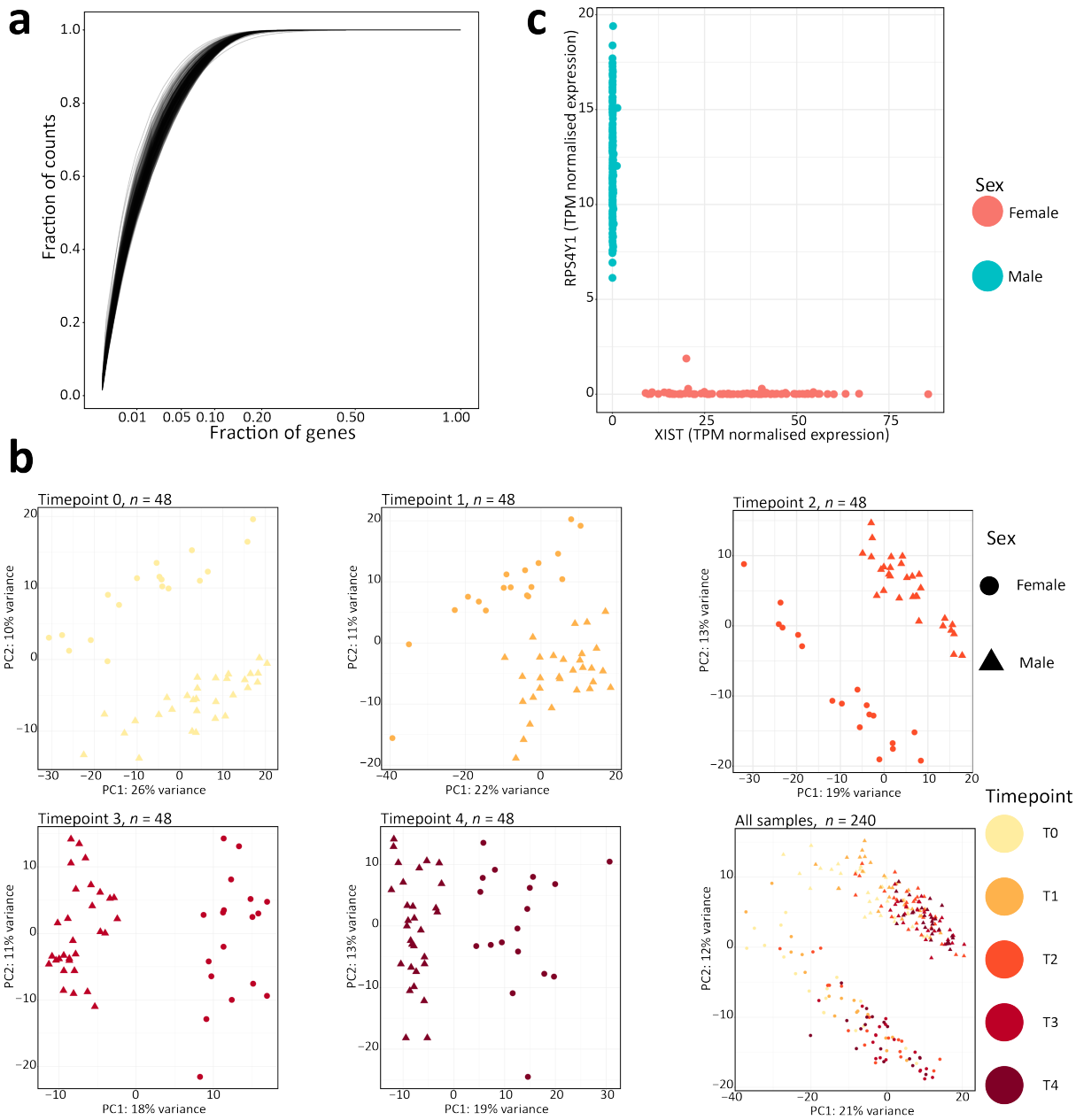
Supplementary Note 3.1: Benchmarking deconvolution algorithms

Prior to deconvolving the PB RNA-seq data, we performed a benchmarking assessment of four deconvolution algorithms: Non-negative least squares (NNLS), MuSic, Bisque, and CIBERSORT with various parameter modifications for a total of 12 different analyses (See Methods). We then assessed deconvolution performance using the average Pearson correlation (ρ) and average root mean square error (RMSE) estimates across all 15 cell types. To do this, we first generated pseudo-bulk RNA-seq data for $n = 100$ individuals derived from the scRNA-seq data with known cell type proportions using the SimBu R/Bioconductor package (**Supplementary Table S3.28**, **Supplementary Figure S3.14**). **Supplementary Figure S3.15** details the results of the benchmarking assessment for MuSic, NNLS, and Bisque with, and without specifying marker genes from the signature matrix. **Supplementary Figure S3.16** illustrates the deconvolution results for the pseudo-bulk RNA-seq data using CIBERSORT with raw or TPM-normalised scRNA-seq data and raw, CPM-, or TPM-normalised bulk data. The least well performing algorithm was CIBERSORT when specifying the raw signature matrix and the TPM normalised pseudo-bulk RNA-seq data ($\rho = 0.382$; RMSE = 0.085). The best performing algorithm was CIBERSORT when specifying the raw signature matrix and the CPM-normalised pseudo-bulk RNA-seq data ($\rho = 0.998$, RMSE = 0.003). We observed that CPM-normalising the pseudo-bulk RNA-seq data slightly improved the deconvolution performance but overall, had minimal impact on the performance metrics (**Supplementary Figure S3.16**; **Supplementary Table S3.29**). MuSiC performed well with, and without specifying marker genes ($\rho_{markers(+)} = 0.991$, $\rho_{markers(-)} = 0.961$; $RMSE_{markers(+)} = 0.009$, $RMSE_{markers(-)} = 0.017$); however, Bisque ($\rho_{markers(+)} = 0.76$, $\rho_{markers(-)} = 0.66$; $RMSE_{markers(+)} = 0.062$, $RMSE_{markers(-)} = 0.072$) and NNLS ($\rho_{markers(+)} = 0.897$, $\rho_{markers(-)} = 0.419$; $RMSE_{markers(+)} = 0.032$, $RMSE_{markers(-)} = 0.082$) were consistently poor (**Supplementary Figure S3.15)**. For CIBERSORT with the CPM-normalised pseudo-bulk RNA-seq data, separating out the results from the deconvolution by cell type highlighted that CIBERSORT was efficient in inferring all the cell types present in the pseudo-bulk RNA-seq data ($\rho = 0.989–1.000$; RMSE = 0.001–0.007 across all cell types, respectively) (**Supplementary Figure S3.17**). The complete results from the deconvolution benchmarking assessment are provided in **Supplementary Table S3.29.**

Supplementary Figures

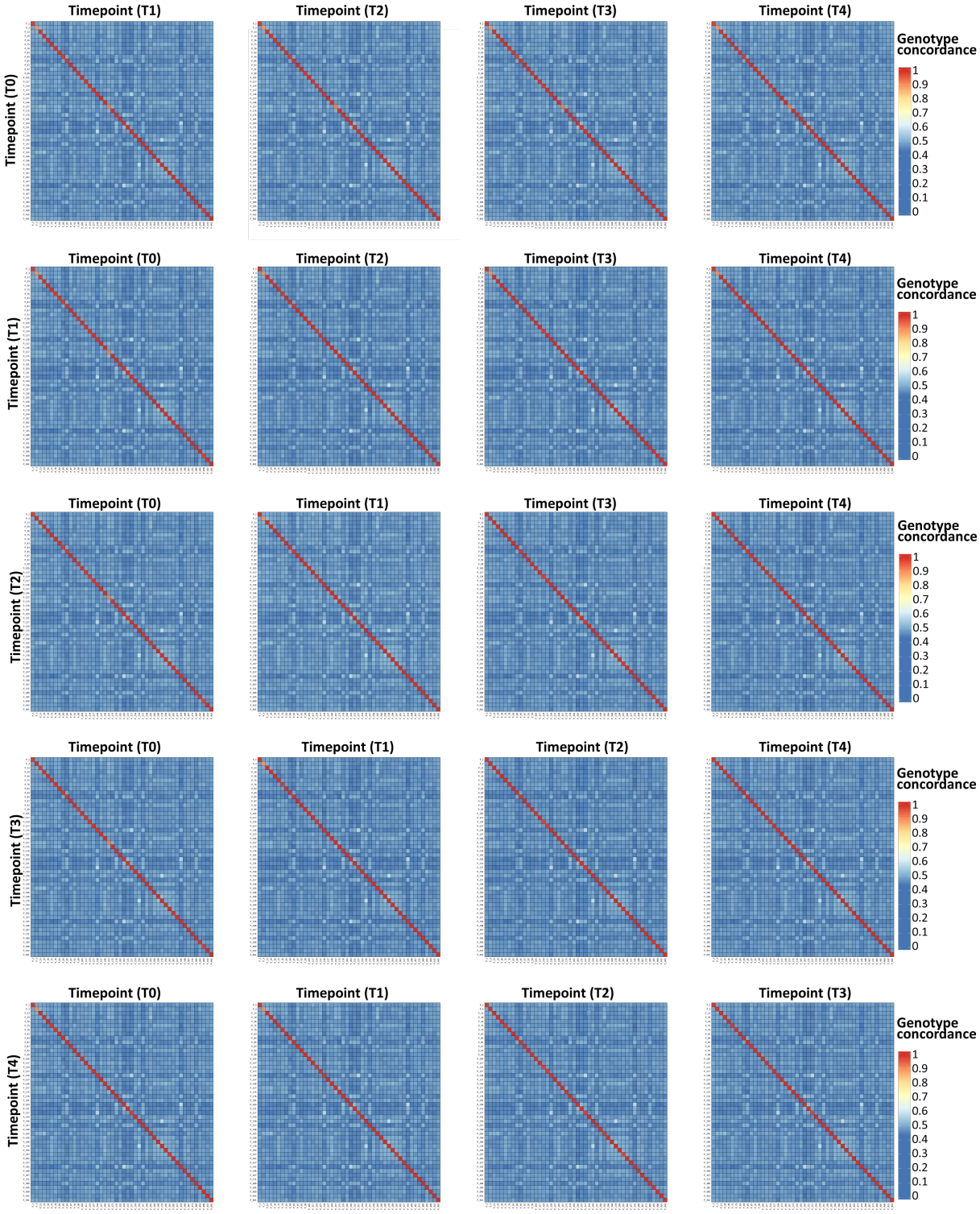


Supplementary Figure S3.1: Sample metadata information.
(**a**) Semi-raincloud plot showing the distribution of day of sampling for all *n* = 48 individuals across all five timepoints analysed (T0–T4). The line in the box plots indicates the median. Whiskers are minima (Q1 − 1.5 × IQR) and maxima (Q3 + 1.5 × IQR), where IQR is the interquartile range (Q3–Q1). (**b**) Bar plot showing the reported ethnicity of all *n* = 48 participants. (**c**) Box plot showing the age distribution of participants plotted by sex. The median of the distributions is indicated by a solid horizontal line within the box plot. Whiskers are minima (Q1 − 1.5 × IQR) and maxima (Q3 + 1.5 × IQR) where IQR is the interquartile range (Q3–Q1).

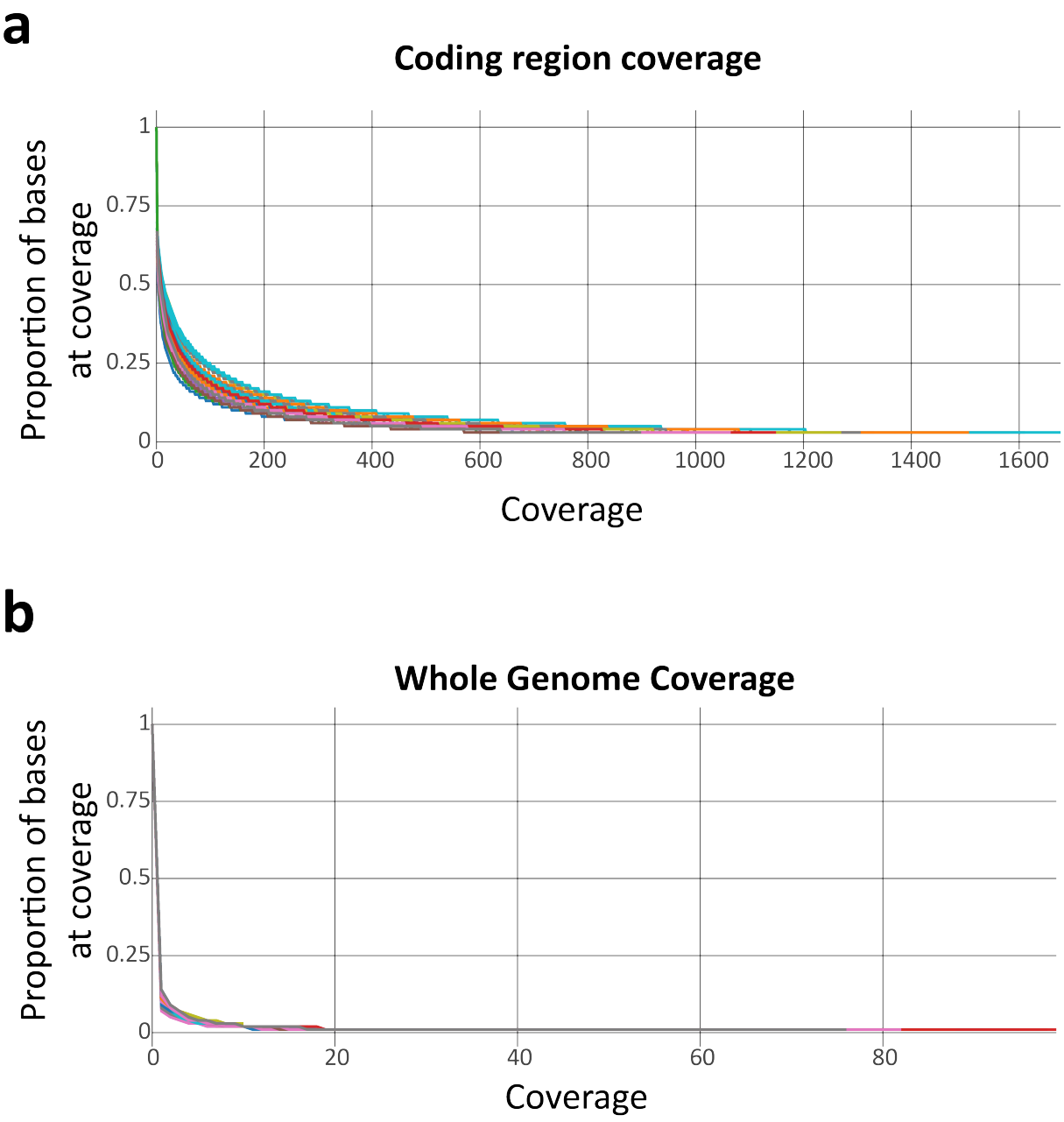


Supplementary Figure S3.2: RNA-seq quality control.

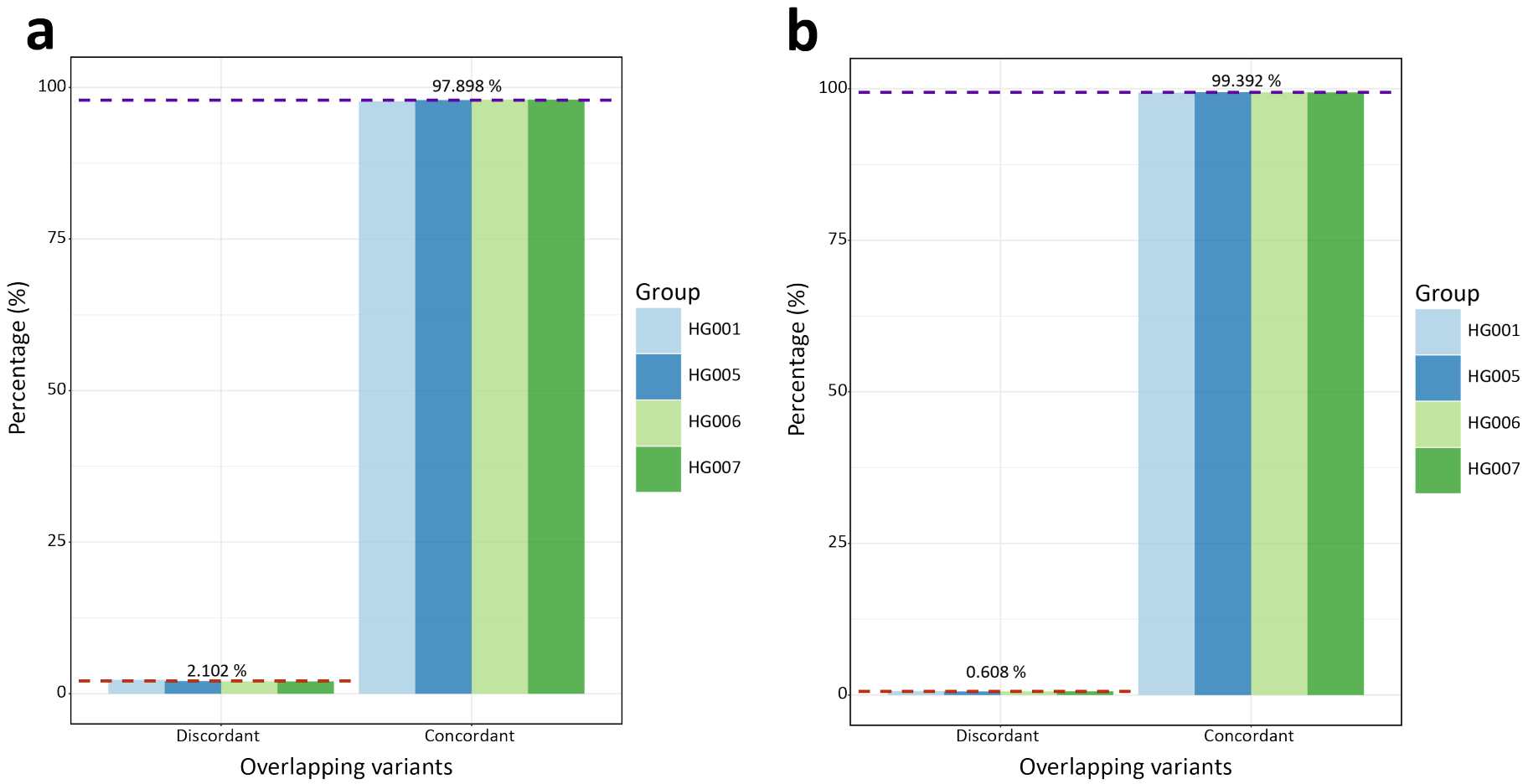
(**a**) Line plot showing the fraction of counts captured by the proportion of genes after expression quantification in each library (*n* = 240). (**b**) Principal component analysis (PCA) of the top 1500 most variable genes as determined from the variance stabilised transformed RNA-seq data in DESeq2 (Love *et al.* 2014) for each timepoint and all RNA-seq samples (*n* = 240) together with principal components 1 and 2 (PC1 and PC2) shown in each comparison. Samples are coloured based on their designated cohort. (**C**) Scatter plot showing the relationship between a female-specific gene (*XIST*) and a Y-linked gene (*RPS4Y1*) with individuals coloured based on their reported sex.



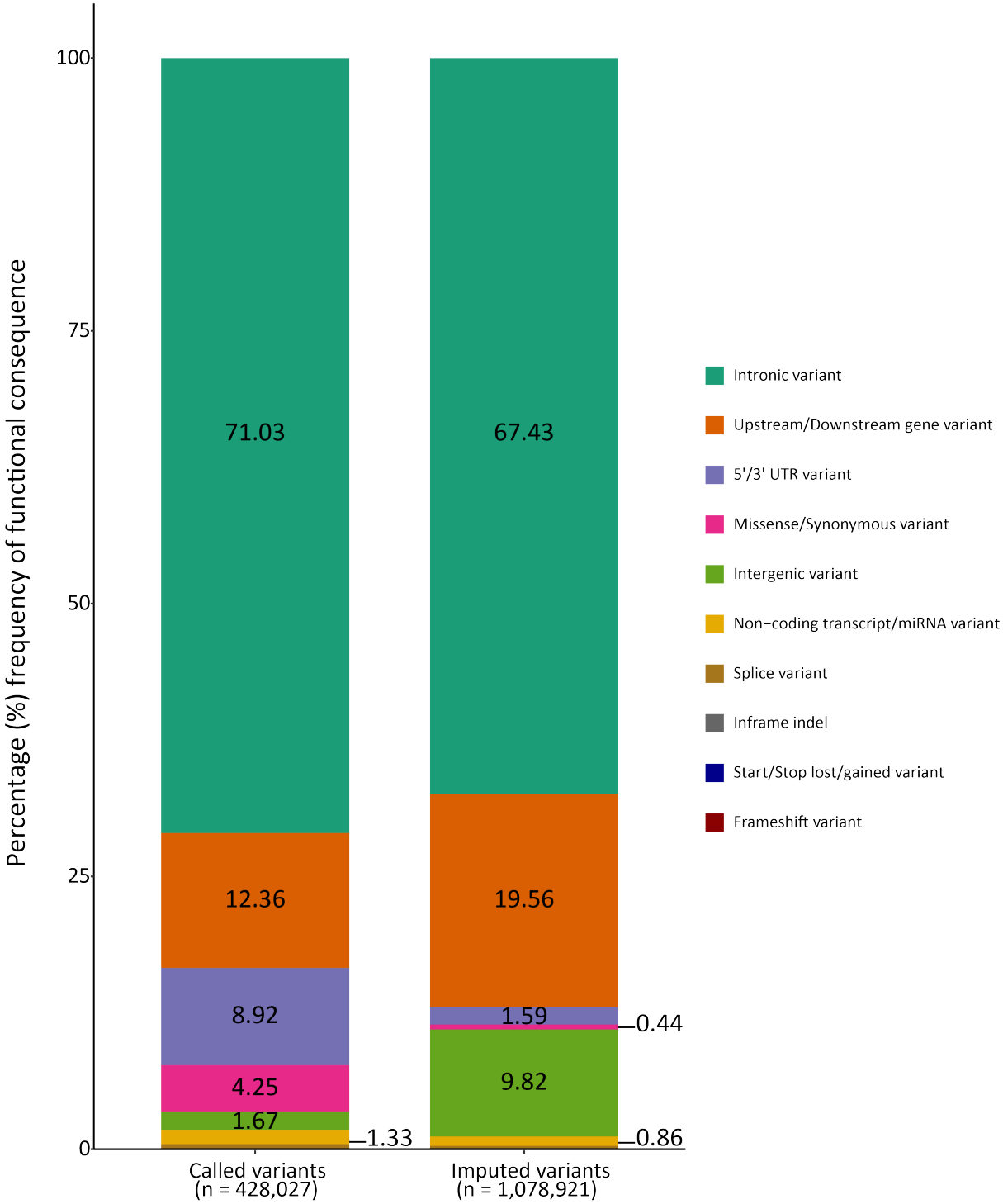
Supplementary Figure S3.3: Sample mismatch assessment.
Genotype concordance using raw variants that were called from all individuals (*n* = 48) in timepoint 0 (T0), timepoint 1 (T1), timepoint 2 (T2), timepoint 3 (T3), and timepoint 4 (T4) versus all individuals in the other cohorts, respectively. A high genotype concordance indicates a sample match.



Supplementary Figure S3.4: Variant call coverage assessment.
(**a**) Plot showing the relationship between the proportion of bases covered in coding regions and at what coverage using the merged BAM files for all *n* = 48 individuals. (**b**) Same as (**a**) but displaying coverage across the entire genome.

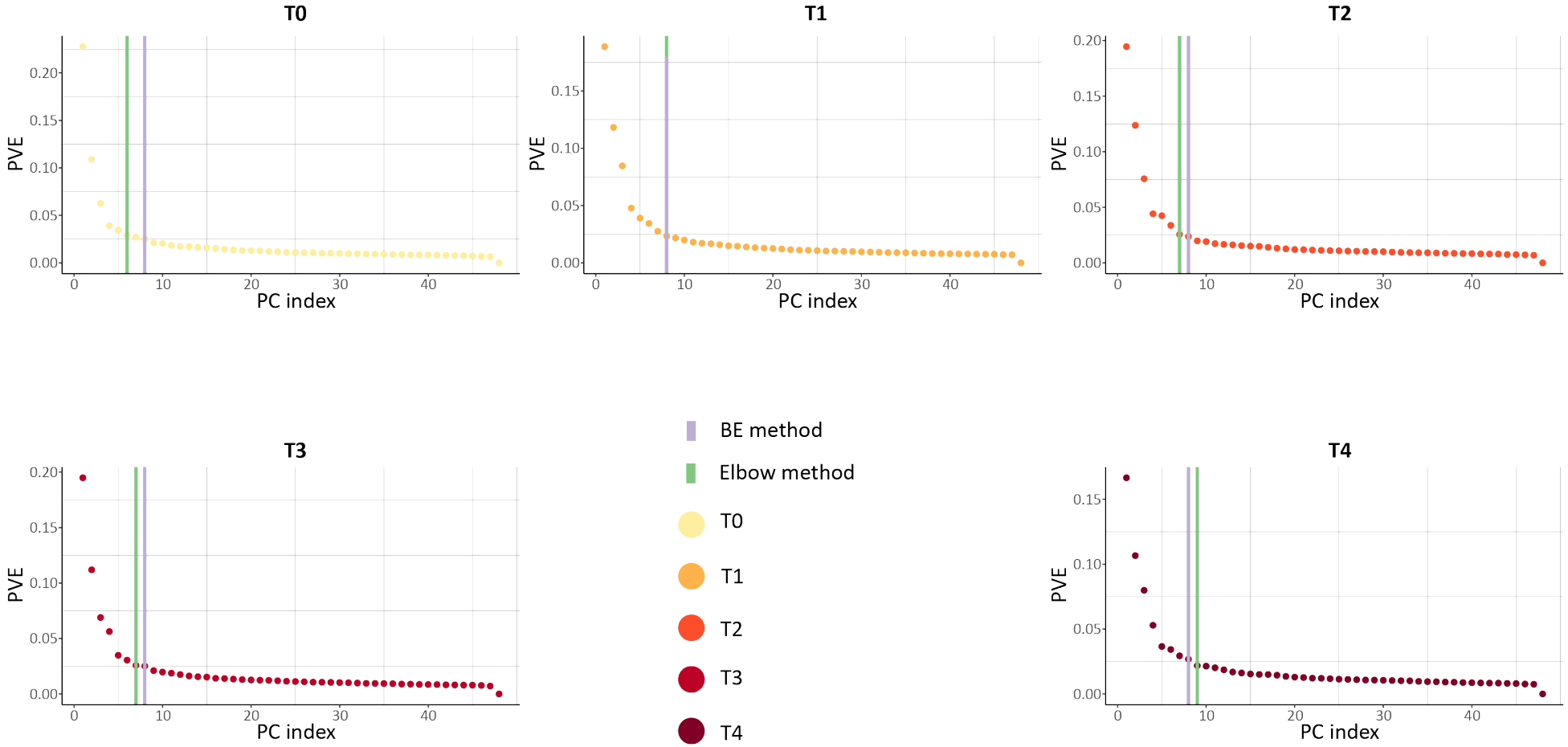


Supplementary Figure S3.5: Genome in a bottle (GIAB) comparison.
(**a**) Bar plot showing the percentage of concordant (same reference and alternative allele pair) alleles and discordant alleles at matched variant sites present in the set of *n* = 48 individuals here and in benchmarked regions of *n* = 4 GIAB data sets (HG001, HG005, HG006, and HG007) (Wagner *et al.* 2022). Dashed red and purple lines indicate the mean percentage across all four data sets for discordance and concordance respectively. (**b**) Same as (**a**) but when only considering sites when only considering biallelic sites.

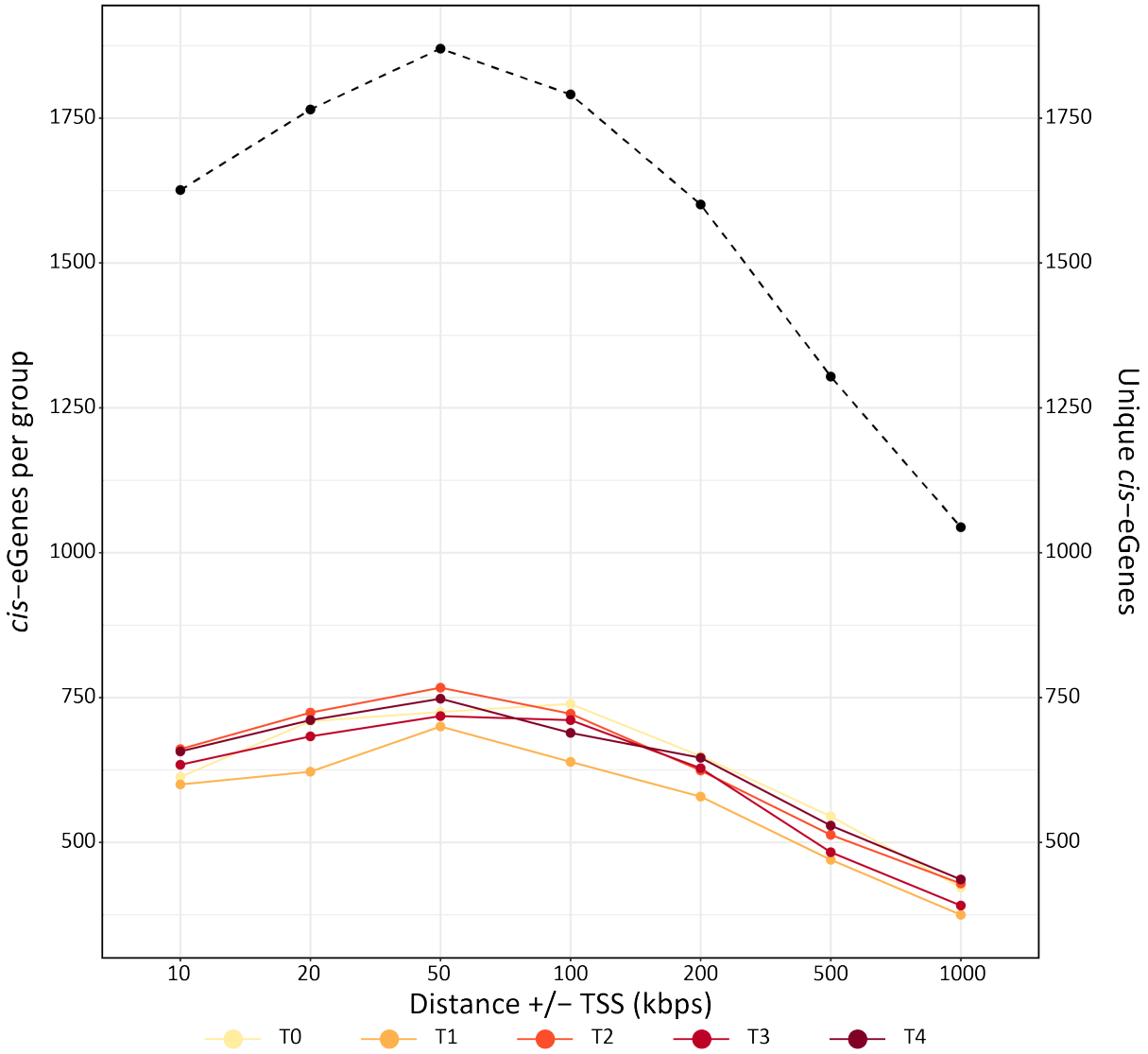


Supplementary Figure S3.6: Functional consequences of variants directly called from RNA-seq data or imputed using the multi-ancestry WGS panel.

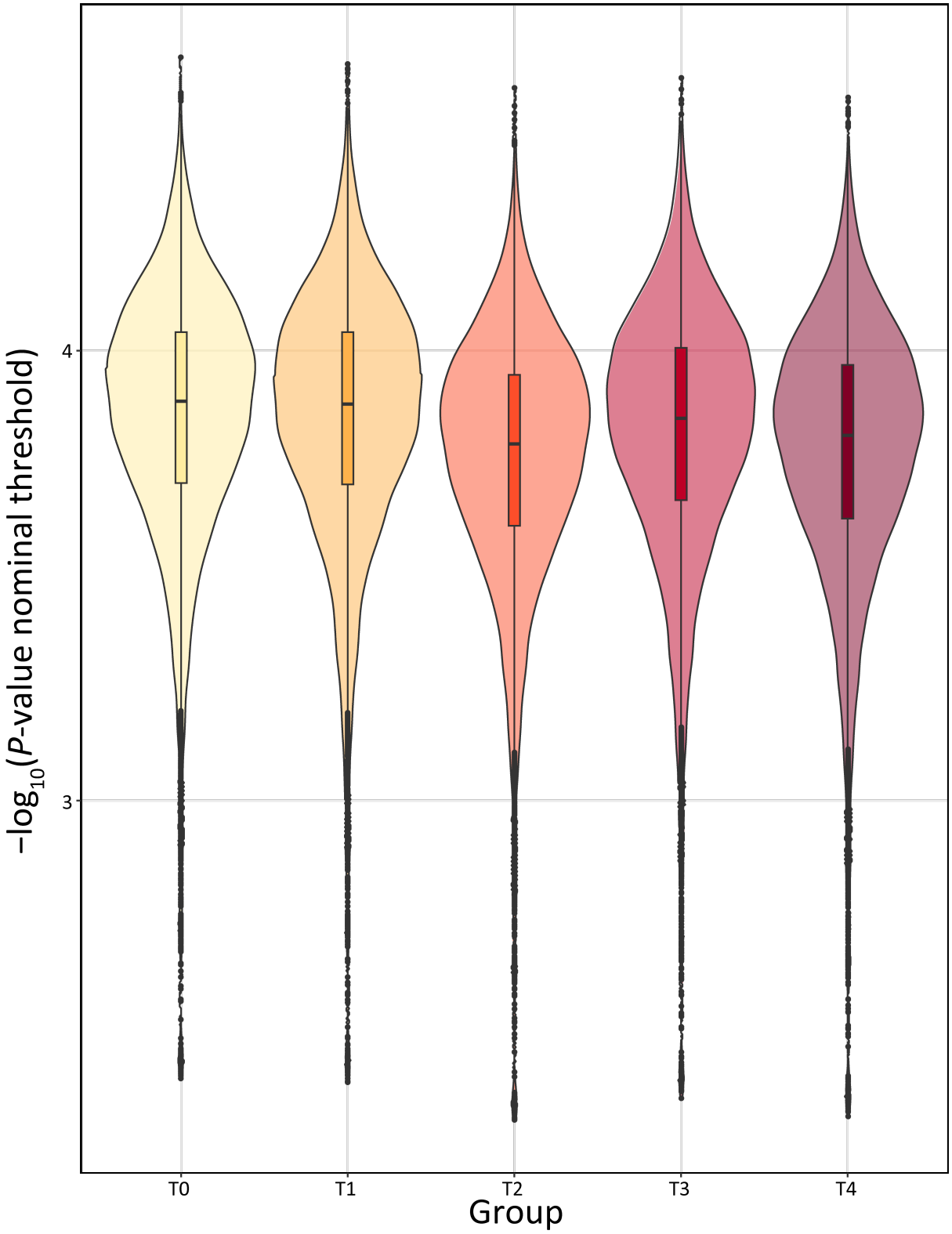
The stacked bar chart illustrates the percentage occurrence of 10 predicted functional impact categories for variants either directly called from the RNA-seq data (*n* = 428,027) or imputed (*n* = 1,078,921) determined using Ensemble's variant effect predictor (VEP). Values of categories are shown if percentage occurrence was > 1% in either the called or imputed cohort, respectively.



Supplementary Figure S3.7: Transcriptomic PCs inferred for eQTL analysis.
Scree plots from PCA4QTL (Zhou *et al.* 2022) for each of the five timepoints (T0–T4) showing the proportion of variation explained (PVE) by each principal component (PC) index. The green line indicates the number of PCs selected using the elbow method and the purple line indicates the number of PCs inferred using the Buja and Eyuboglu (BE) method (Buja & Eyuboglu 1992).

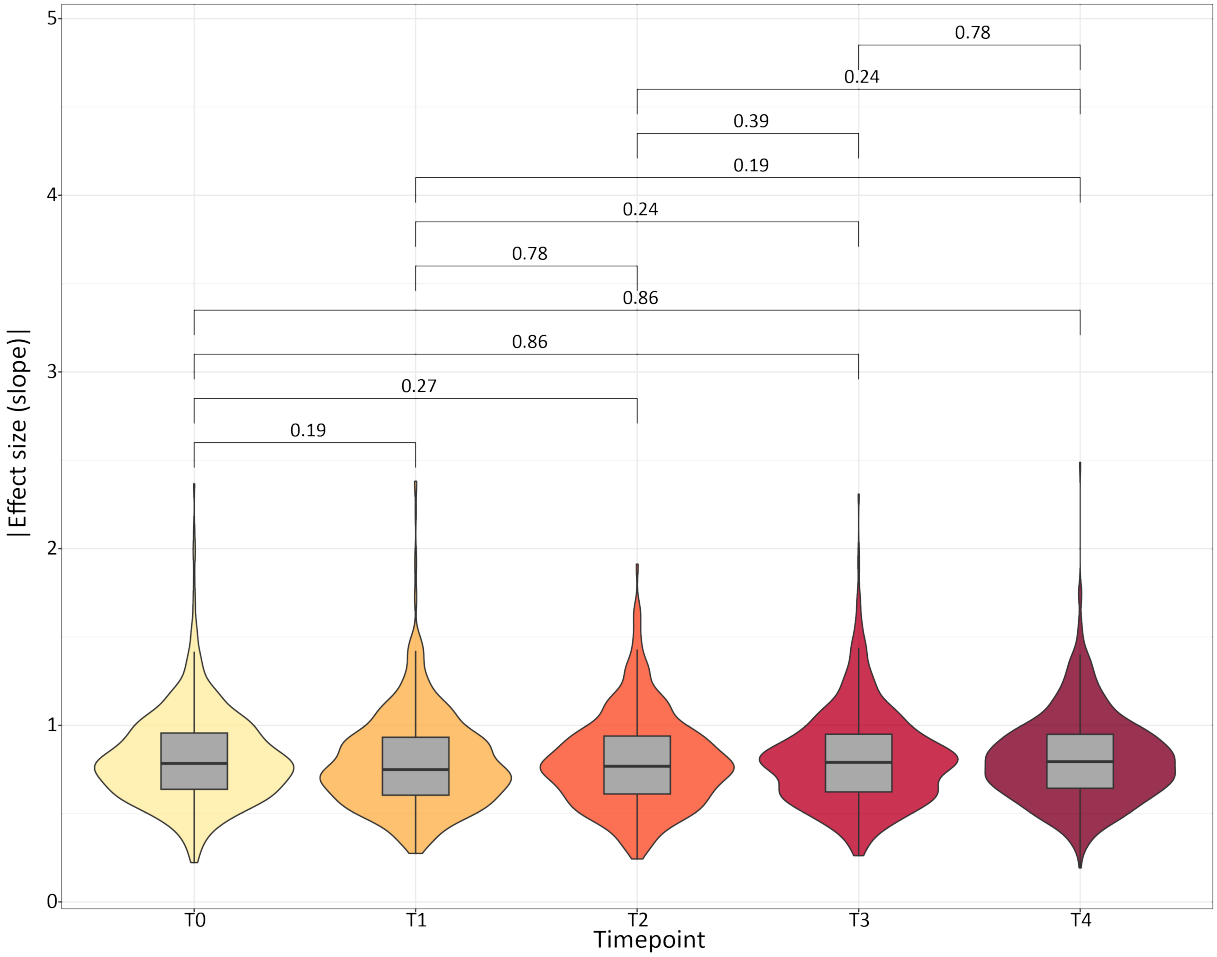


Supplementary Figure S3.8: Selection of optimum distance for cis-eQTL mapping.
The number of *cis*-eGenes displayed on the first *y*-axis inferred by TensorQTL (Taylor-Weiner *et al.* 2019) for each cohort (T0–T4) using a *cis*-window size varying form 10 kb up to 1 Mb (1000 kb). The dashed line indicates the number of unique *cis*-eGenes identified across all the five groups and is illustrated by the second *y*-axis.



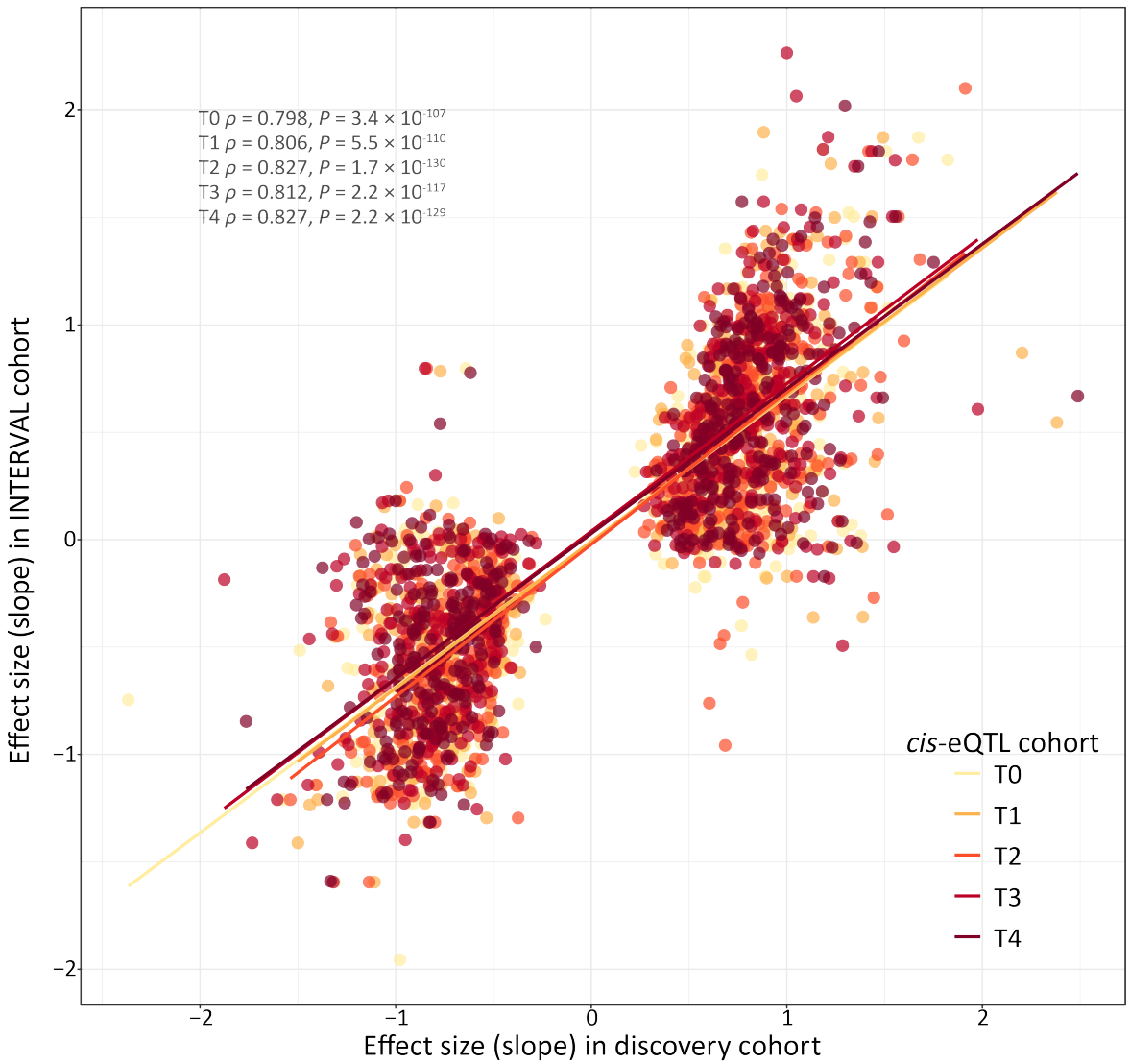
Supplementary Figure S3.9: Nominal P-value threshold for determining TensorQTL cis-eVariants

Violin plots showing the distribution of nominal *P*-value threshold per *cis*-eGene for all groups (T0–T4), respectively.



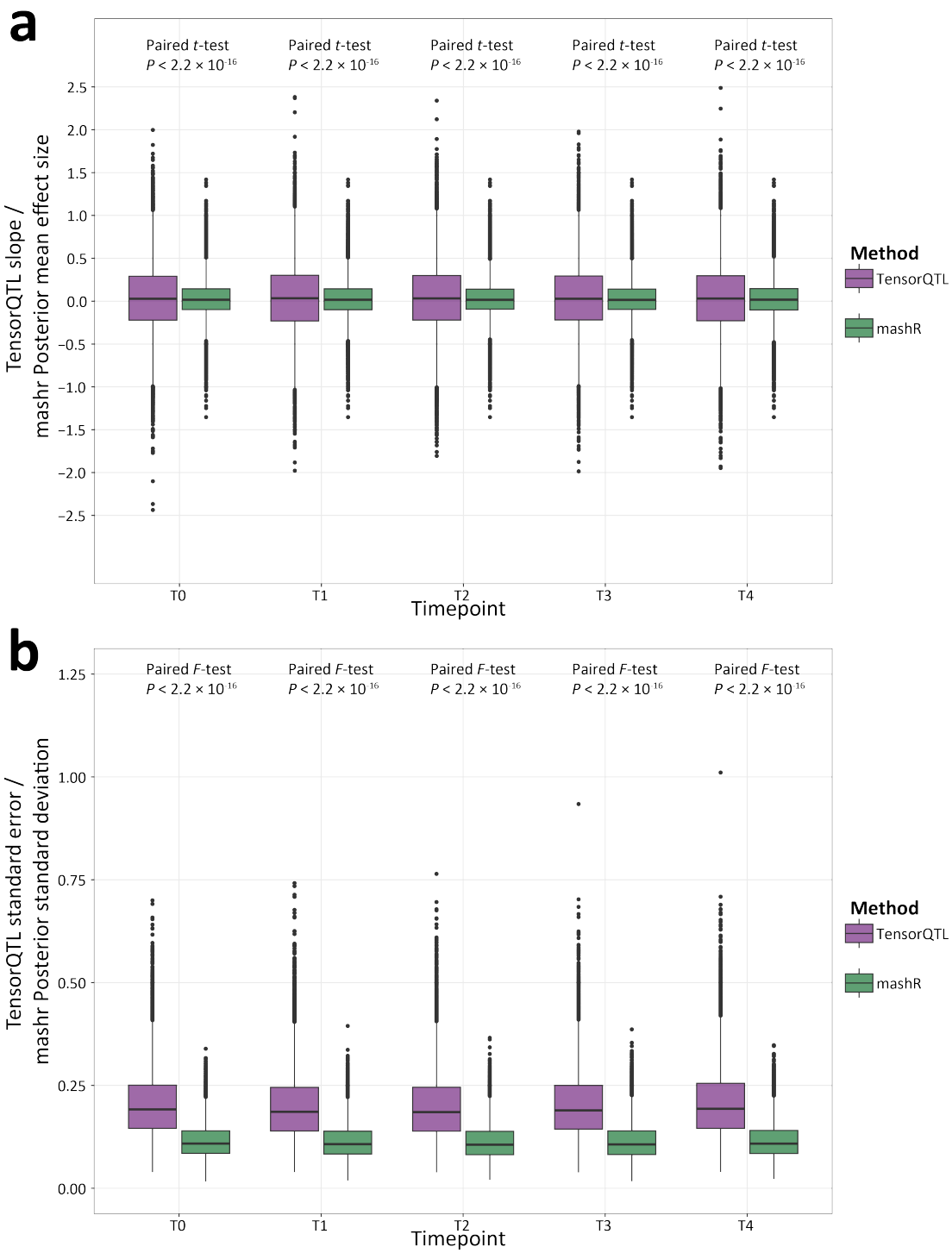
Supplementary Figure S3.10: Distribution of absolute effect size (slope) values of top cis-eQTLs (Padj. < 0.1) identified in each timepoint.

Joint box- and violin-plots showing the distribution of the absolute effect size (slope) for top significant *cis*-eQTLs ($P_{adj.} < 0.1$) identified in each timepoint (T0, T1, T2, T3, T4), respectively. Values indicate adjusted p-values from pairwise comparisons of each distribution using a Wilcoxon rank-sum test. The box plots cover the interquartile range with the median line denoted at the centre, and the whiskers extend to the most extreme data point that is no more than 1.5 × IQR from the edge of the box.



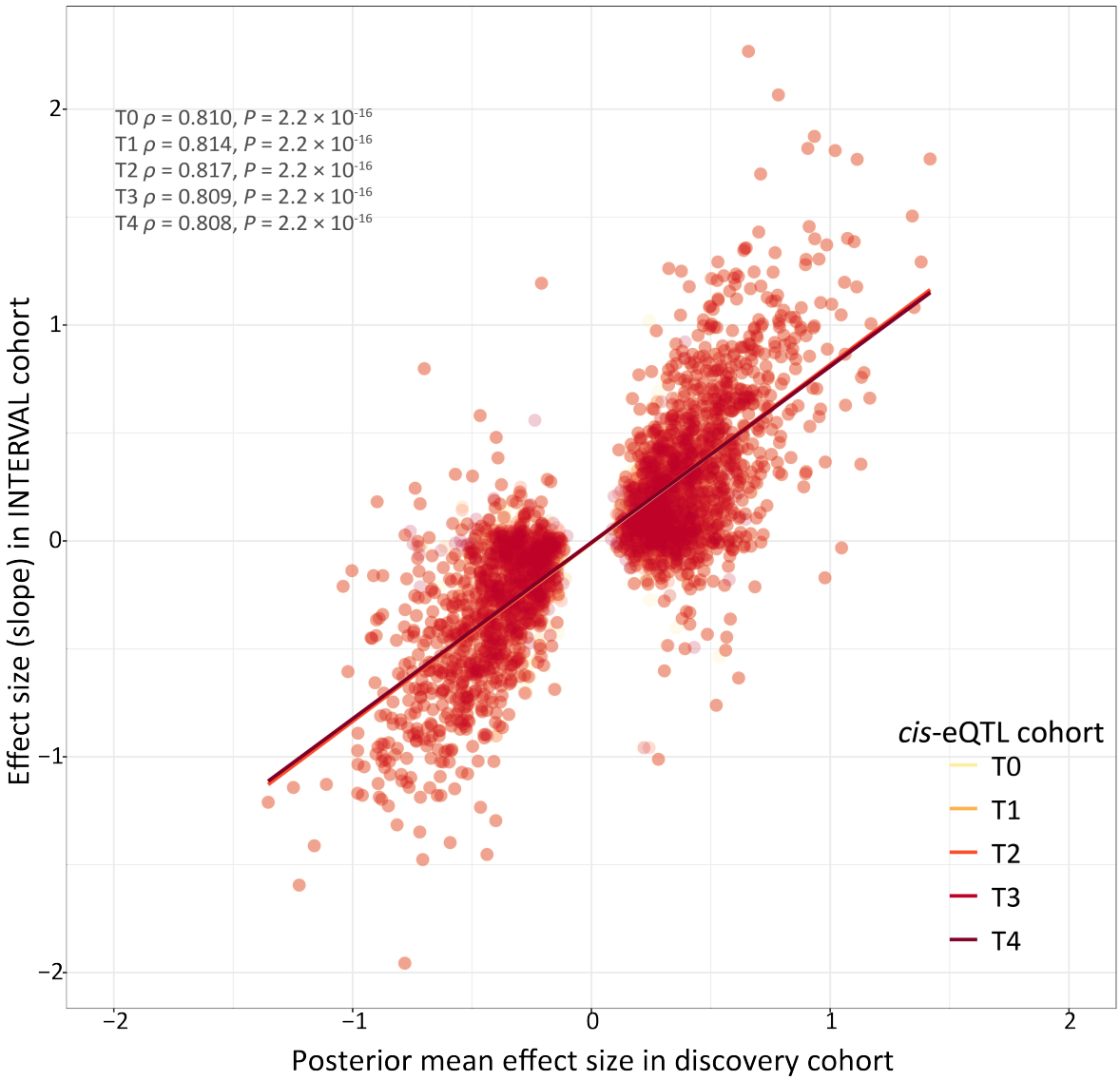
Supplementary Figure S3.11: Effect size comparison of cis-eQTLs and matched variant-gene pairs identified in the INTERVAL cohort.

Scatterplot illustrating the effect sizes of significant *cis*-eQTLs ($P_{adj.}$ < 0.1) identified in this study (*x*-axis) and matched variant-gene pairs identified in the INTERVAL cohort (*y*-axis). Spearman correlation values are also reported in addition to the corresponding *P*-value representing the significance level of each respective correlation. The coloured lines indicate lines of best fit within each timepoint, respectively.



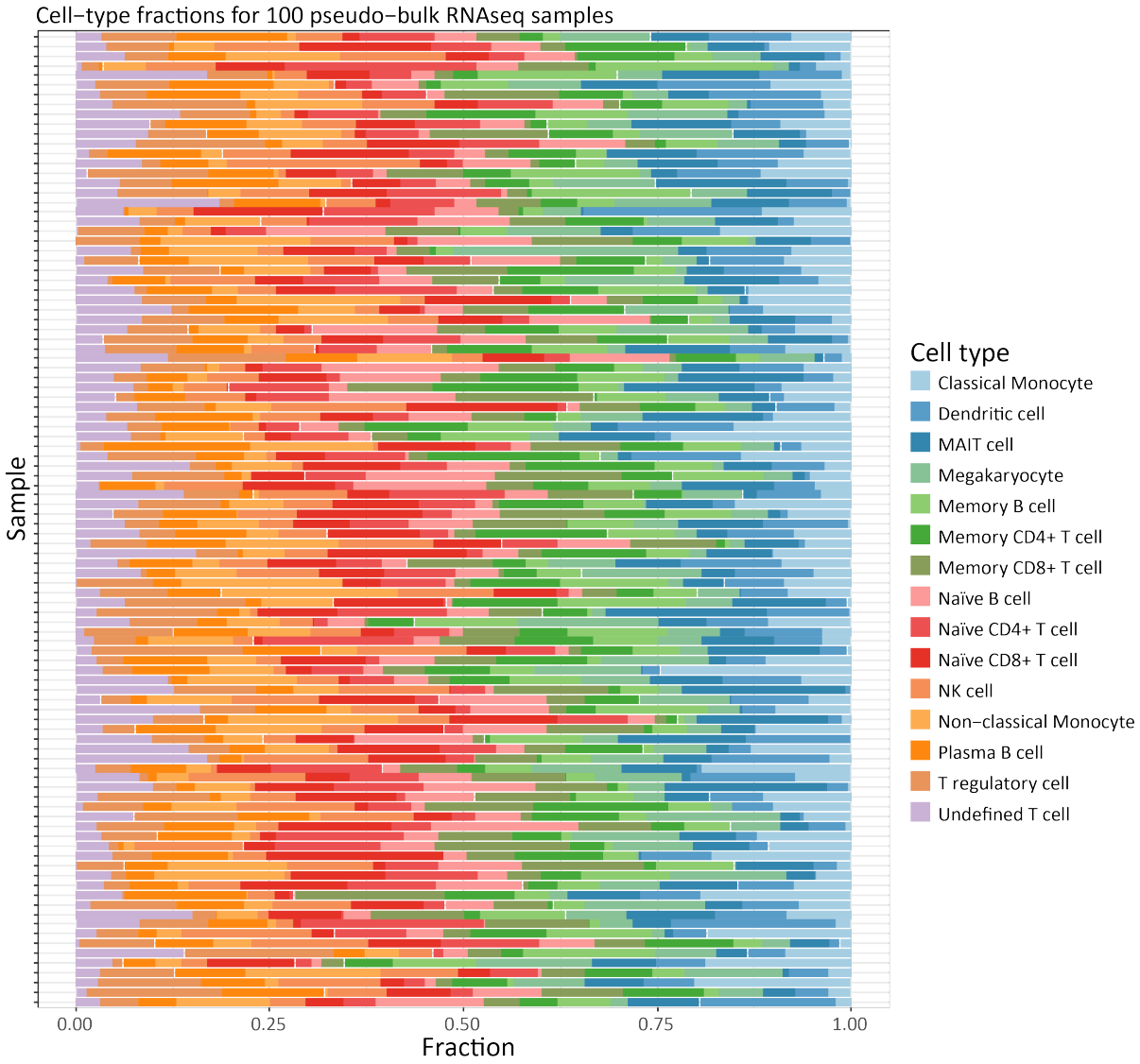
Supplementary Figure S3.12: Re-estimation of *cis*-eQTL effect sizes and standard error estimates using mashr.

(**a**) Box plots showing the distribution of effect size for the most significant variant tested across all five groups and coloured based on the original TensorQTL (Taylor-Weiner *et al.* 2019) effect size (slope) and marshr (Urbut *et al.* 2019) posterior mean effect size. *P*-values are inferred from *t*-tests comparing the mean effect size and the mean posterior mean effect size estimate in each timepoint, respectively. (**b**) Same as (**a**) but illustrating the standard error estimates for the most significant variant tested across all five groups and the mashr posterior standard deviation. *P*-values are inferred from *F*-tests comparing the standard error estimates and posterior standard deviation estimates in each group separately.

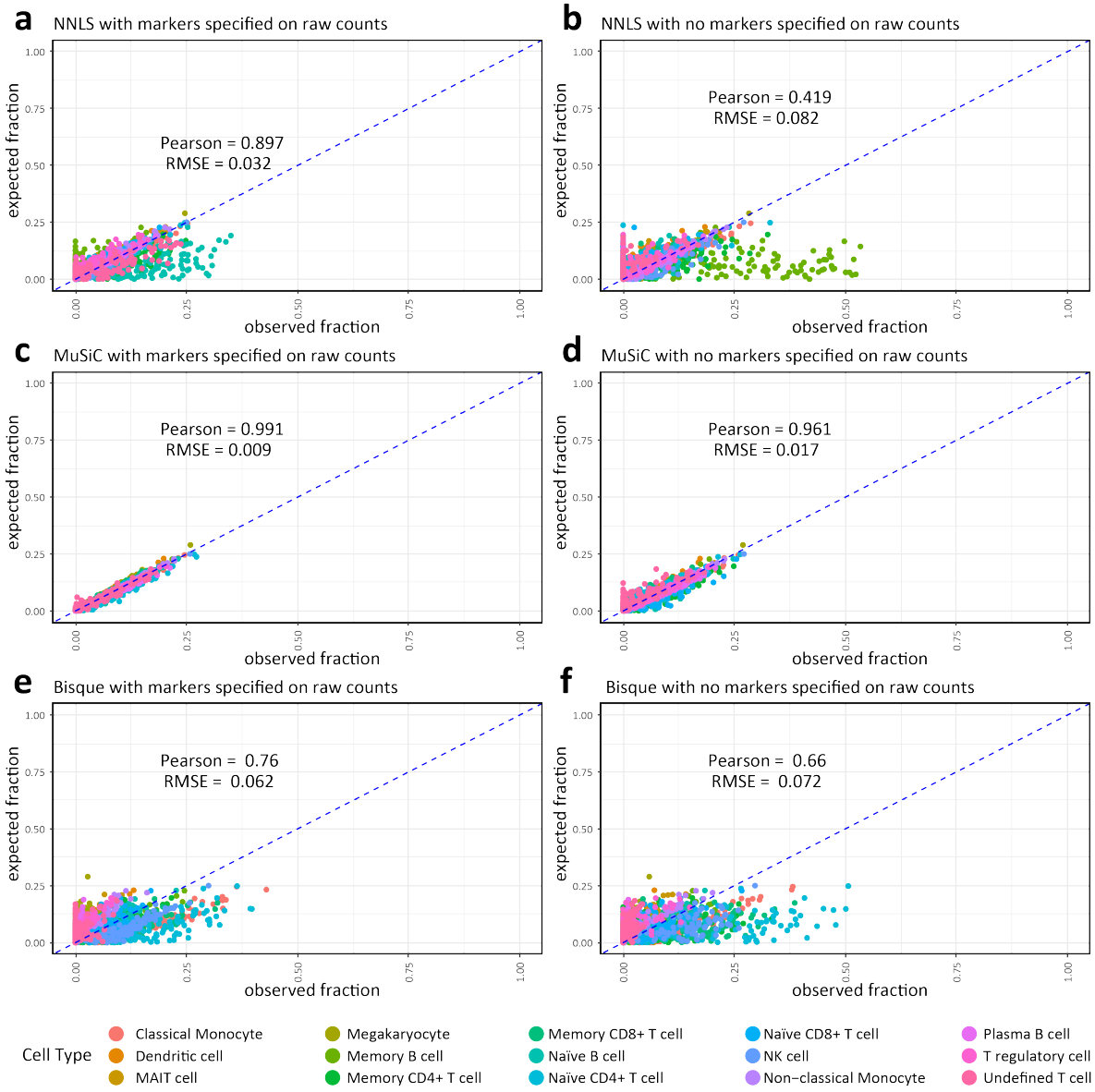


Supplementary Figure S3.13: Effect size comparison of *cis*-eQTLs after mashr analysis to matched variant-gene pairs identified in the INTERVAL cohort.

Scatterplot illustrating the posterior mean effect sizes determined from mashr of significant *cis*-eQTLs (LFSR < 0.05) identified in this study (*x*-axis) and matched variant-gene pairs identified in the INTERVAL cohort (*y*-axis). Spearman correlation values are also reported in addition to the corresponding *P*-value representing the significance level of each respective correlation. The coloured lines indicate lines of best fit within each timepoint, respectively.

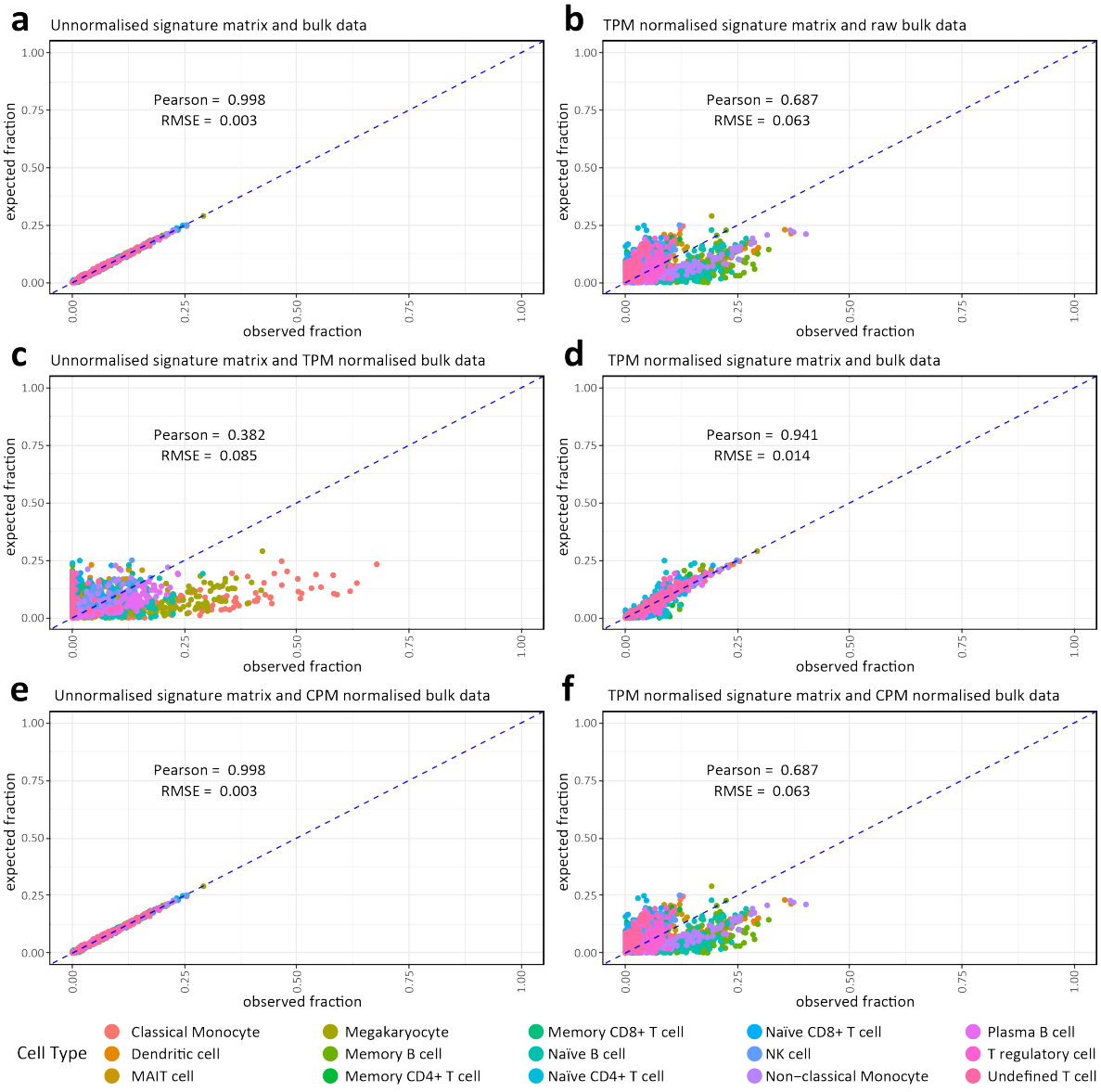


Supplementary Figure S3.14: Simulation of known cell type proportions using the Simbu R/Bioconductor R package based on pseudo-bulk RNA-seq data.
Bar plots of 100 pseudo-bulk RNA-seq samples on the *y*-axis with the fraction of each cell type designated and separated based on colour illustrated on the *x*-axis. Cell type proportions were generated using the Simbu R/Bioconductor R package (Dietrich *et al.* 2022) based on the scRNA-seq data used in this study.



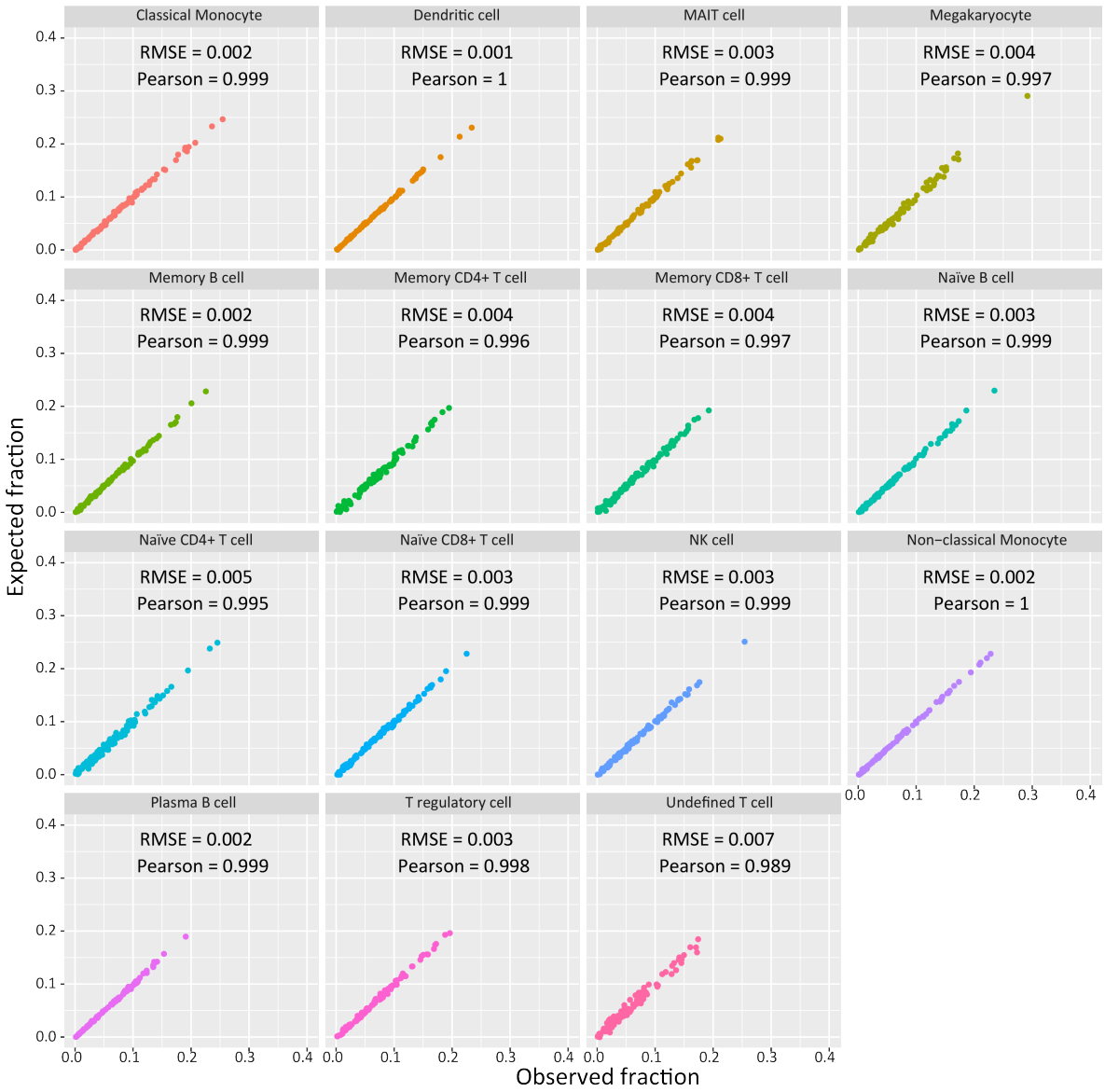
Supplementary Figure S3.15: Deconvolution performance of NNLS, MuSiC and Bisque with and without marker genes

Scatter plot showing the expected fraction of cell types on the *y*-axis versus the observed fraction of cell types following deconvolution on the *x*-axis using (**a**) non-negative least squares (NNLS) with marker genes specified, and (**b**) without marker genes specified. (**c**) MuSiC with marker genes specified and (**d**) without marker genes specified. (**e**) Bisque with marker genes specified, and (**f**) without marker genes specified. Pearson, Pearson correlation for all 15 cell types; RMSE = average root mean square error for all 15 cell types.



Supplementary Figure S3.16: Deconvolution performance of CIBERSORT with various normalisation strategies applied to the genes in the signature matrix and pseudo-bulk RNA-seq data

Scatter plot showing the expected fraction of cell types on the *y*-axis versus the observed fraction of cell types following deconvolution on the *x*-axis using (**a**) CIBERSORT with raw unnormalised signature matrix counts and pseudo-bulk data. (**b**) CIBERSORT with transcripts per million (TPM) normalised signature matrix and unnormalised pseudo-bulk data matrix counts and pseudo-bulk data. (**c**) CIBERSORT with raw unnormalized signature matrix and TPM-normalised pseudo-bulk data. (**d**) CIBERSORT with TPM-normalised signature matrix and TPM-normalised pseudo-bulk data. (**e**) CIBERSORT with raw unnormalized signature matrix and counts per million (CPM) normalised pseudo-bulk data. (**f**) CIBERSORT with TPM-normalised signature matrix and CPM-normalised pseudo-bulk data. Pearson, Pearson correlation for all 15 cell types; RMSE, average root mean square error for all 15 cell types.



Supplementary Figure S3.17: Deconvolution performance of each cell type using CIBERSORT with raw unnormalized signature matrix and CPM-normalised pseudo-bulk RNA-seq data
Deconvolution performance of CIBERSORT with raw unnormalized signature matrix and counts per million (CPM) normalised pseudo-bulk data. Pearson, Pearson correlation for all 15 cell types; RMSE, root mean square error.

Supplemental References

Buja A. & Eyuboglu N. (1992) Remarks on Parallel Analysis. *Multivariate Behav Res* 27, 509-40.

Dietrich A., Sturm G., Merotto L., Marini F., Finotello F. & List M. (2022) SimBu: bias-aware simulation of bulk RNA-seq data with variable cell-type composition. *Bioinformatics* 38, ii141-ii7.

Love M.I., Huber W. & Anders S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550.

Taylor-Weiner A., Aguet F., Haradhvala N.J., Gosai S., Anand S., Kim J., Ardlie K., Van Allen E.M. & Getz G. (2019) Scaling computational genomics to millions of individuals with GPUs. *Genome Biol* 20, 228.

Urbut S.M., Wang G., Carbonetto P. & Stephens M. (2019) Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet* 51, 187-95.

Wagner J., Olson N.D., Harris L., Khan Z., Farek J., Mahmoud M., Stankovic A., Kovacevic V., Yoo B., Miller N., Rosenfeld J.A., Ni B., Zarate S., Kirsche M., Aganezov S., Schatz M.C., Narzisi G., Byrska-Bishop M., Clarke W., Evani U.S., Markello C., Shafin K., Zhou X., Sidow A., Bansal V., Ebert P., Marschall T., Lansdorp P., Hanlon V., Mattsson C.A., Barrio A.M., Fiddes I.T., Xiao C., Fungtammasan A., Chin C.S., Wenger A.M., Rowell W.J., Sedlazeck F.J., Carroll A., Salit M. & Zook J.M. (2022) Benchmarking challenging small variants with linked and long reads. *Cell Genom* 2, 100128.

Zhou H.J., Li L., Li Y., Li W. & Li J.J. (2022) PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol* 23, 210.