bioRxiv preprint doi: https://doi.org/10.1101/2023.10.21.562237; this version posted March 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

SUPPLEMENTARY MATERIALS

Genotype prediction of 336,463 samples from public expression data

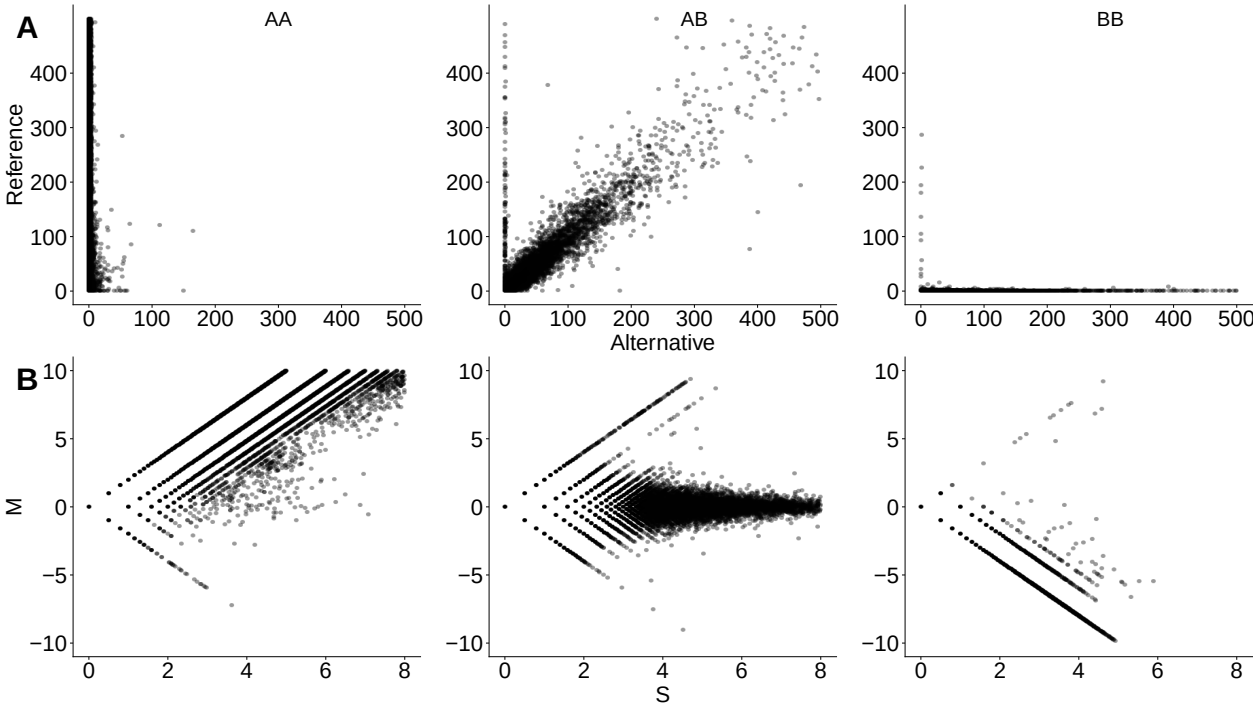
Afrooz Razi, Christopher C. Lo, Siruo Wang, Jeffrey T. Leek*, Kasper D. Hansen*

* Correspondence to jtleek@fredhutch.org (LAG), khansen@jhsph.edu (KDH)

Contents

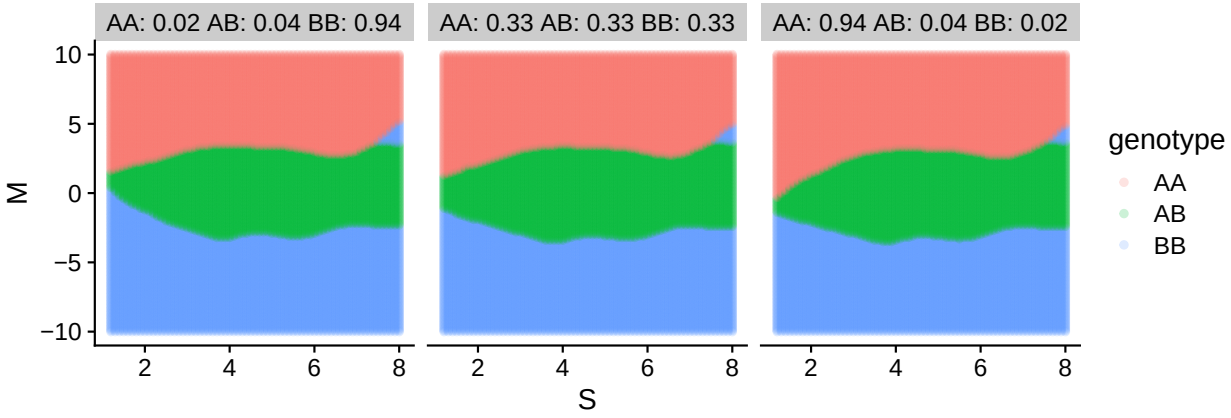
bioRxiv preprint doi: https://doi.org/10.1101/2023.10.21.562237; this version posted March 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1 Supplemental Figures

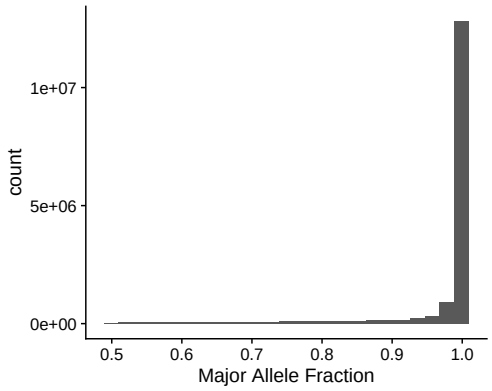


Supplementary Figure S1. Transformation of SNP coverage values. We depict 200 random GTEx samples in the training set. **(A)** Raw reference and alternative read counts where each point represents a SNP of a sample. Each plot is faceted based on genotype (A being reference and B being alternative). **(B)** The same samples and SNPs are plotted again using S vs. M transformation.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.21.562237; this version posted March 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

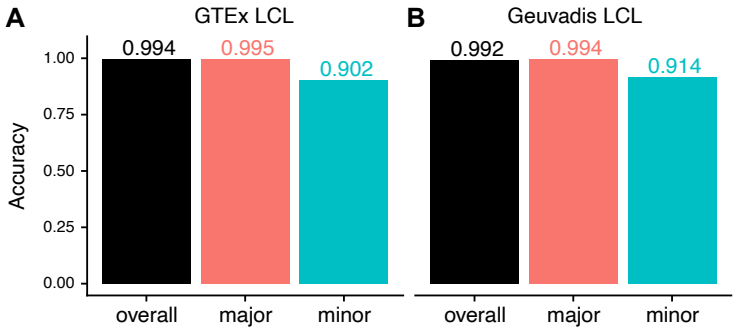


Supplementary Figure S2. Sensitivity analysis for the choice of prior distribution. Each facet depicts the decision boundaries of the three genotypes based on a prior genotyping distribution. The prior genotyping distribution are specified in each facet's title. The reference allele is "A" and the alternative allele is "B".

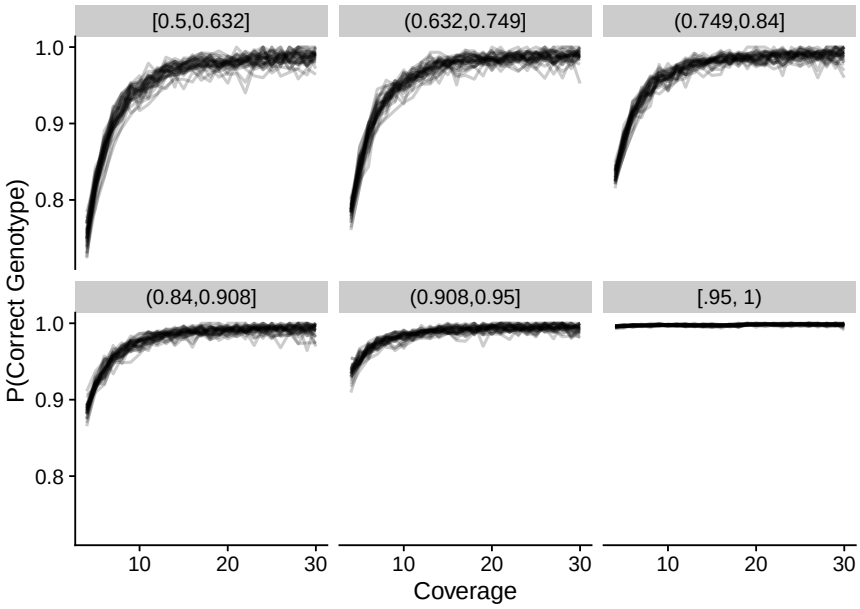


Supplementary Figure S3. Distribution of major allele frequency of GTEx test set. The major allele frequency for a SNP is the number of most prevalent allele for the SNP across the 200 GTEx test set individuals divided by the number of total alleles for the SNP across the individuals.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.21.562237; this version posted March 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

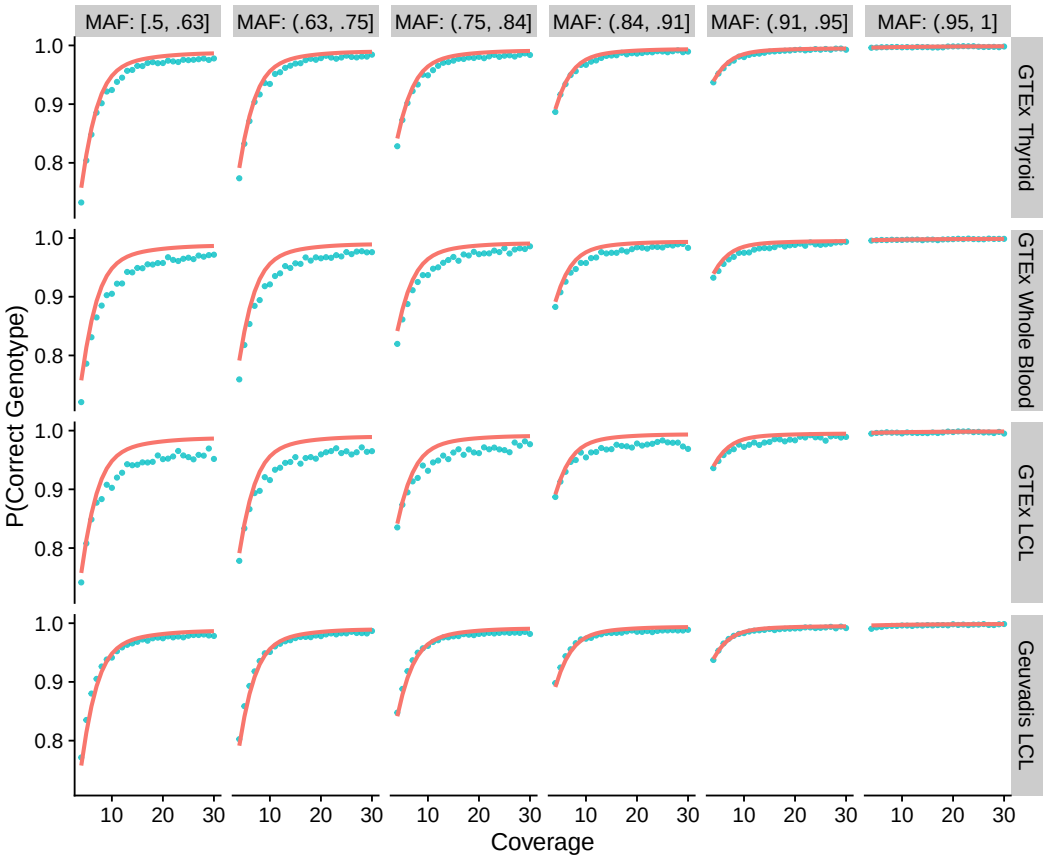


Supplementary Figure S4. Model performance for lymphoblastoid cell lines. (A) Model performance for 35 GTEx LCL samples in the GTEx test set. **(B)** Model performance for 462 Geuvadis samples as an out-of-study set.



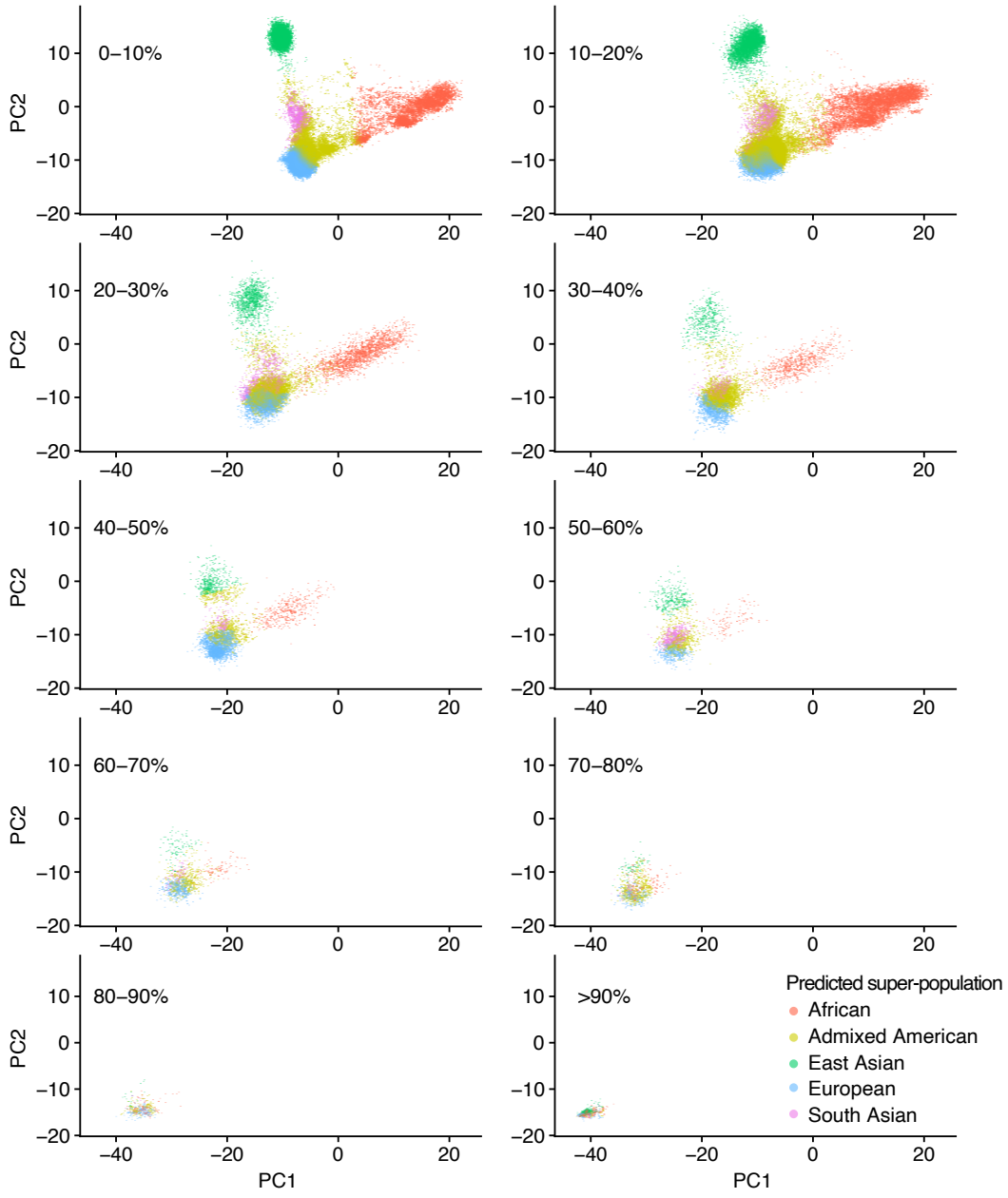
Supplementary Figure S5. Genotyping accuracy as a function of sequencing coverage and major allele frequency. Genotyping accuracy at the SNP level is grouped by discrete coverage values and major allele frequency bin to get a continuous averaged accuracy value. Each facet denotes a major allele frequency bin, and each line represents a tissue type from GTEx training set.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.21.562237; this version posted March 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



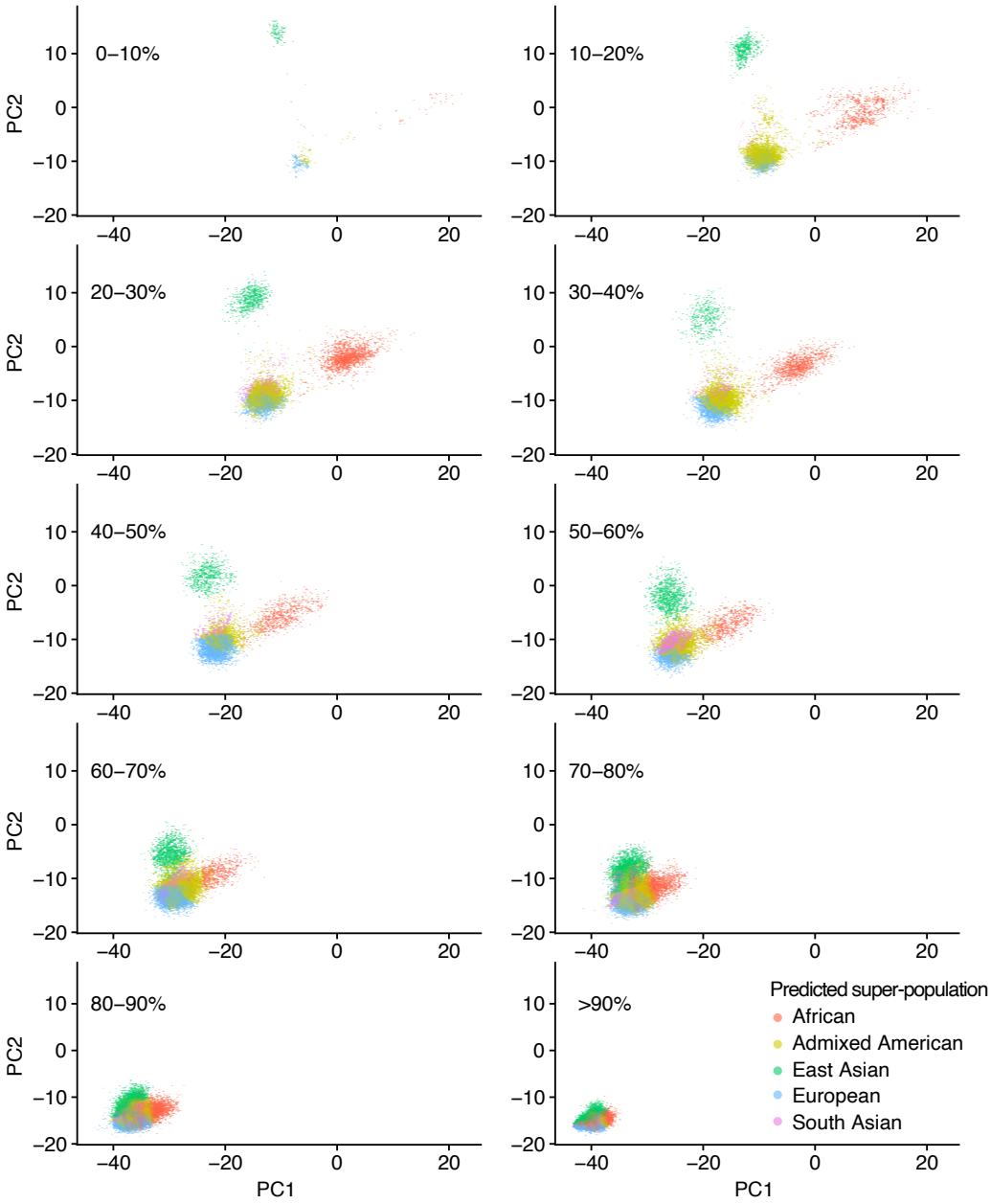
Supplementary Figure S6. Genotyping accuracy as a function of sequencing coverage and major allele frequency for several tissues. Genotyping accuracy at the SNP level is grouped by discrete coverage values and major allele frequency bin to get a continuous averaged accuracy value (blue-green points). The model prediction is shown in red lines. Each facet along the x-axis denotes a major allele frequency bin. Each facet along the y-axis denotes a different tissue type. We contrast a well fitted tissue type (GTEx Thyroid) against less well fitted tissue types (GTEx Whole Blood, GTEx LCL). We also contrast model predictions of LCL tissues between two studies (GTEx LCL vs. Geuvadis LCL).

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.21.562237; this version posted March 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



Supplementary Figure S7. SRA Bulk RNA-seq samples ancestry prediction. Bulk RNA-seq samples from SRA are separated based on their percent missing genotype in 10% increments. Each color corresponds to the predicted super-population based on 1000 Genome reference.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.21.562237; this version posted March 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



Supplementary Figure S8. SRA single-cell RNA-seq samples ancestry prediction. Single-cell samples from SRA are separated based on their percent missing genotype in 10% increments. Each color corresponds to the predicted super-population based on 1000 Genome reference.