Appendix A

Supplementary information

A.1 Data access and code availability

The following new data generated as part of this work is available publicly:

- *S. pyogenes* sequence reads from section 2.6.3 are available on the European Nucleotide Archive under study accession IDs PRJEB2839 (isolates from Fiji) and PRJEB3313 (isolates from Kilifi).
- From the paired blood and CSF isolates in section 4.5 read data, assembled and annotated contigs were deposited in the European Nucleotide Archive (ENA): study accession number ERP004245.
- Sample metadata used from these paired blood and CSF isolates has been deposited in Figshare (DOI: 10.6084/m9.figshare.4329809).

Relevant code for each section can be found on github:

- Testing of tree inference methods, section 2.3.1: https://github.com/johnlees/which_tree
- SEER, section 2.5: https://github.com/johnlees/seer
- Carriage duration analysis and results, chapter 3: https://github.com/johnlees/carriage-duration
- Paired sample analysis, section 4.5: https://github.com/johnlees/paired-samples
- Calculation of Tajima's D, section 4.4.2: https://github.com/johnlees/tajima-D
- Fix to subtest code, section 5.2.2: https://github.com/johnlees/subtest
- Code to perform all-by-all variant association in genome-to-genome analysis, section 5.3.1: https://github.com/johnlees/epistasis-code
- Miscellaneous code and scripts, referred to throughout: https://github.com/johnlees/bioinformatics

A.2 Supplementary figures

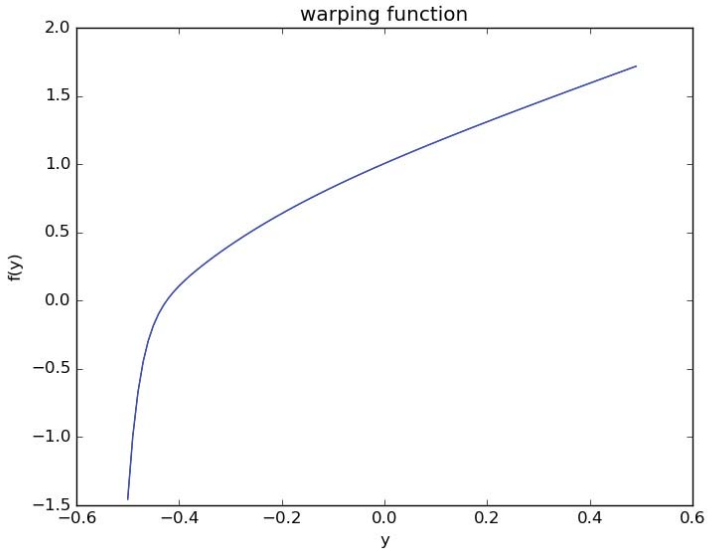

Figure A.1: Monotonic warping function from warped-lmm. x-axis shows the centred and normalised input phenotype; y-axis shows corresponding warped value.

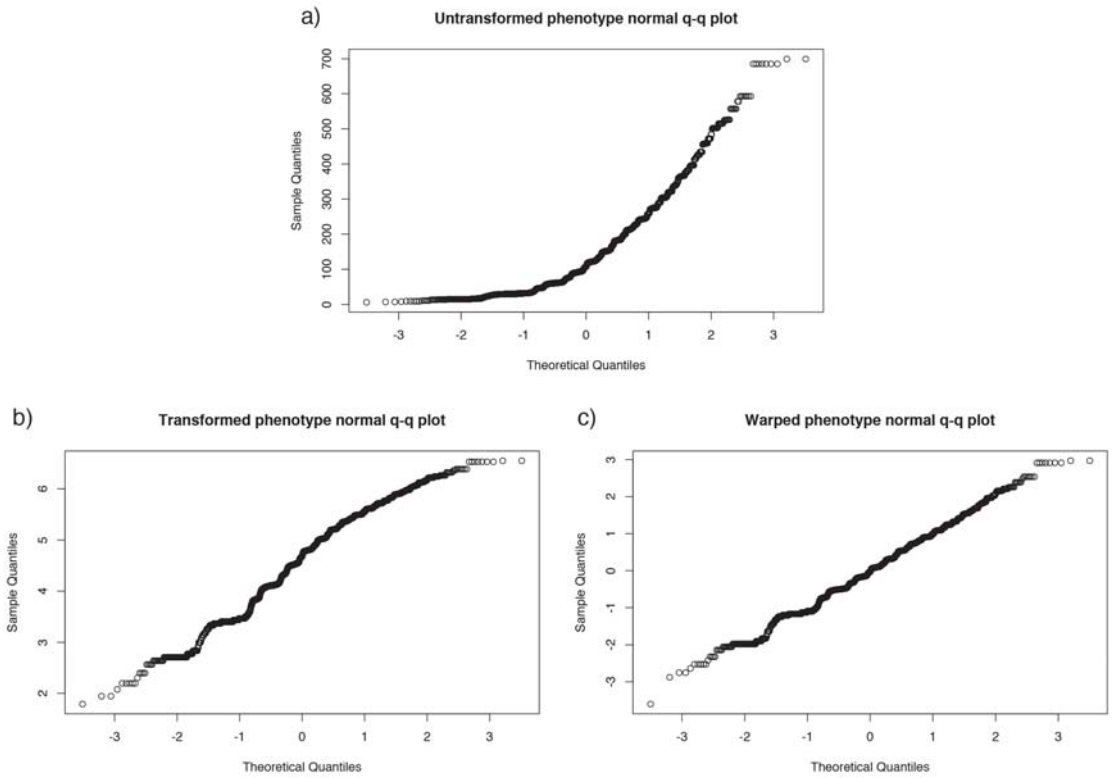

Figure A.2: Normal quantile-quantile plot of carriage length, and effect of monotonic transformation. Panel **a)** the inferred carriage duration, **b)** after the natural logarithm is taken, and **c)** after the warping function is applied.

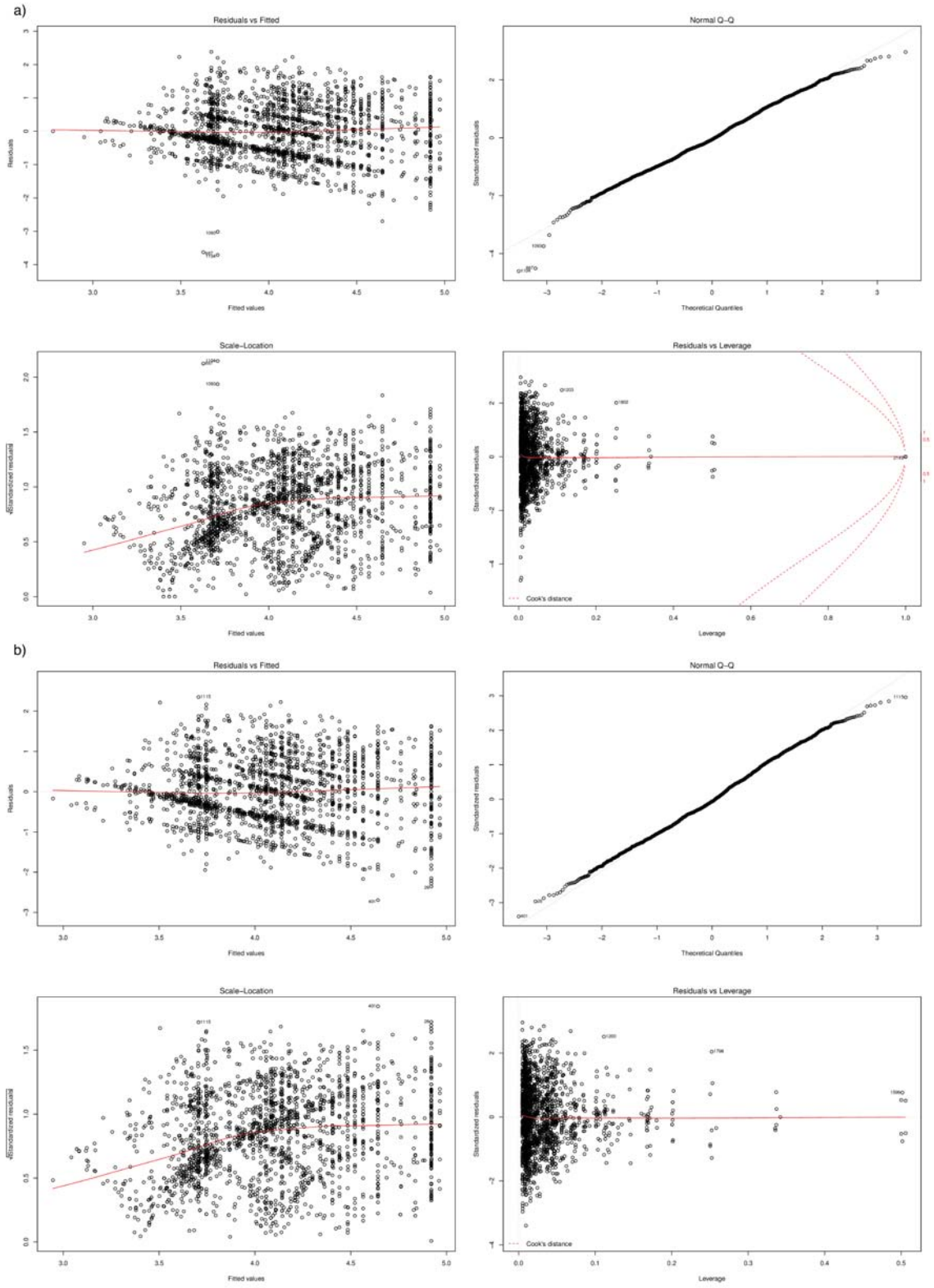

Figure A.3: Regression diagnostics and outlier removal. Panel **a)** shows prior to outlier removal, **b)** after outlier removal as produced by `plot.lm()` in R. Points deviating from normal residuals (top right plot), and at high leverage (bottom right plot) were removed. These observations appeared to be due to swabs not taken at the prescribed monthly intervals.

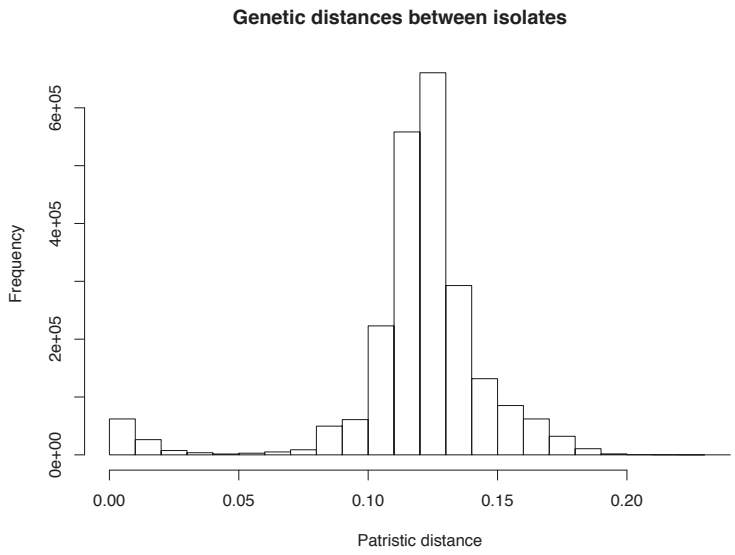

Figure A.4: Histogram of pairwise patristic distances on the inferred phylogeny. A cut-off for heritability estimation was chosen at 0.04, under which a clear second maxima corresponds to closely related isolates on the tree.

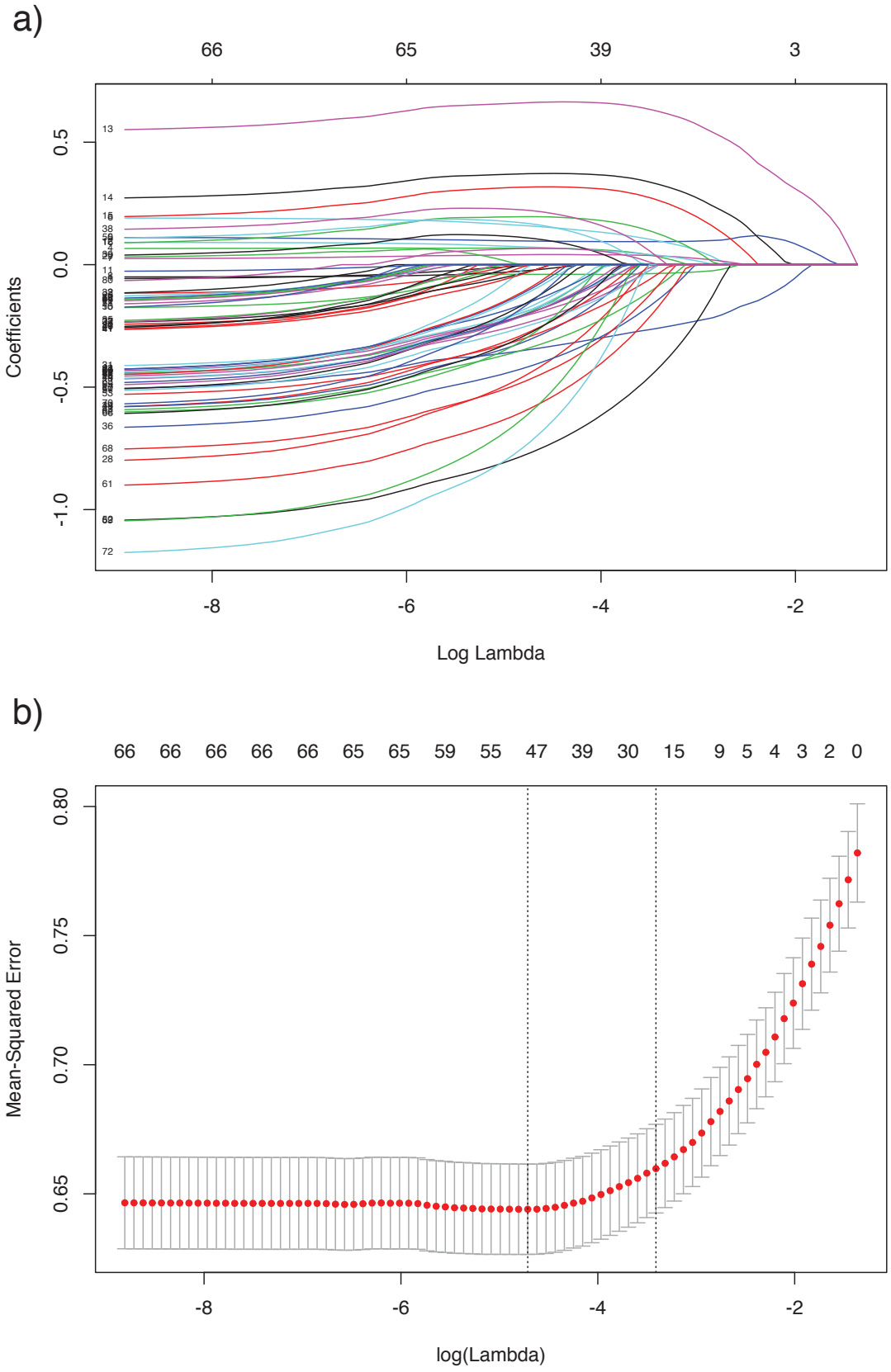

Figure A.5: Lasso regression plots for lineage effects. Panel **a)** shows the value of each predictor on the y-axis for different values of the ℓ_1 penalty λ on the x-axis, which increases from left to right. The labels along the top are the number of predictors remaining in the model for each λ. Panel **b)** shows the results of leave-one-out cross validation on the mean-squared error, along the same x-scale. The λ at minimum error is shown by the left dashed line, and the λ within one standard error is shown by the right dashed line.

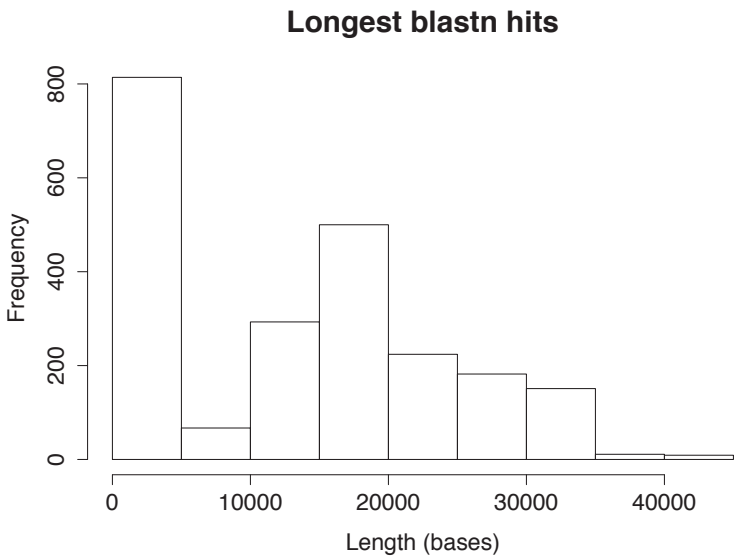

Figure A.6: Identification of phage in assemblies by `blastn` hit length. Histogram of the length of top hits against a database of phage sequence by `blastn`. Isolates with >5000bp hits were defined as having phage present.

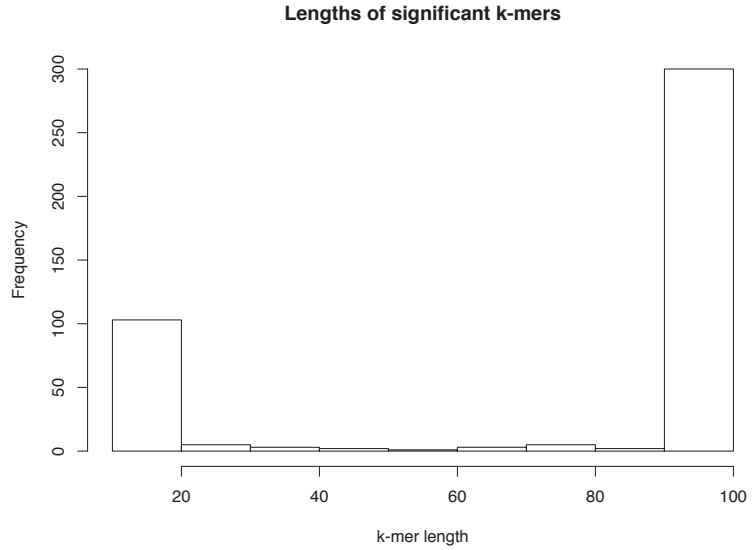

Figure A.7: The lengths of those k-mers reaching significance in the LMM analysis, binned by frequency. Lengths below 20 bases were filtered from downstream analysis, due to having low specificity.

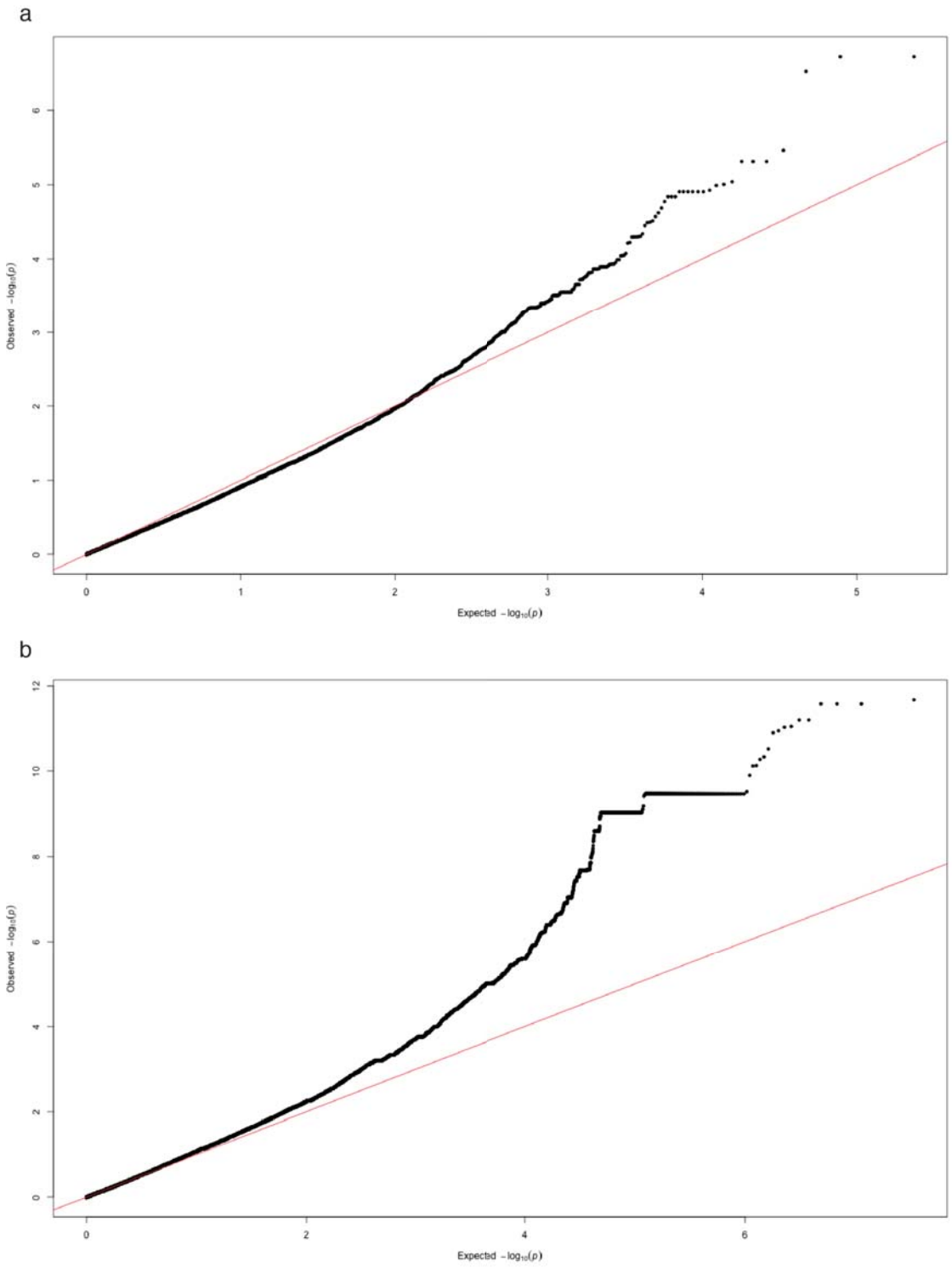

Figure A.8: Quantile-quantile plots of association p-values. For `fast-lmm` results on **a)** SNPs passing quality filters and **b)** k-mers of all lengths passing quality filters.

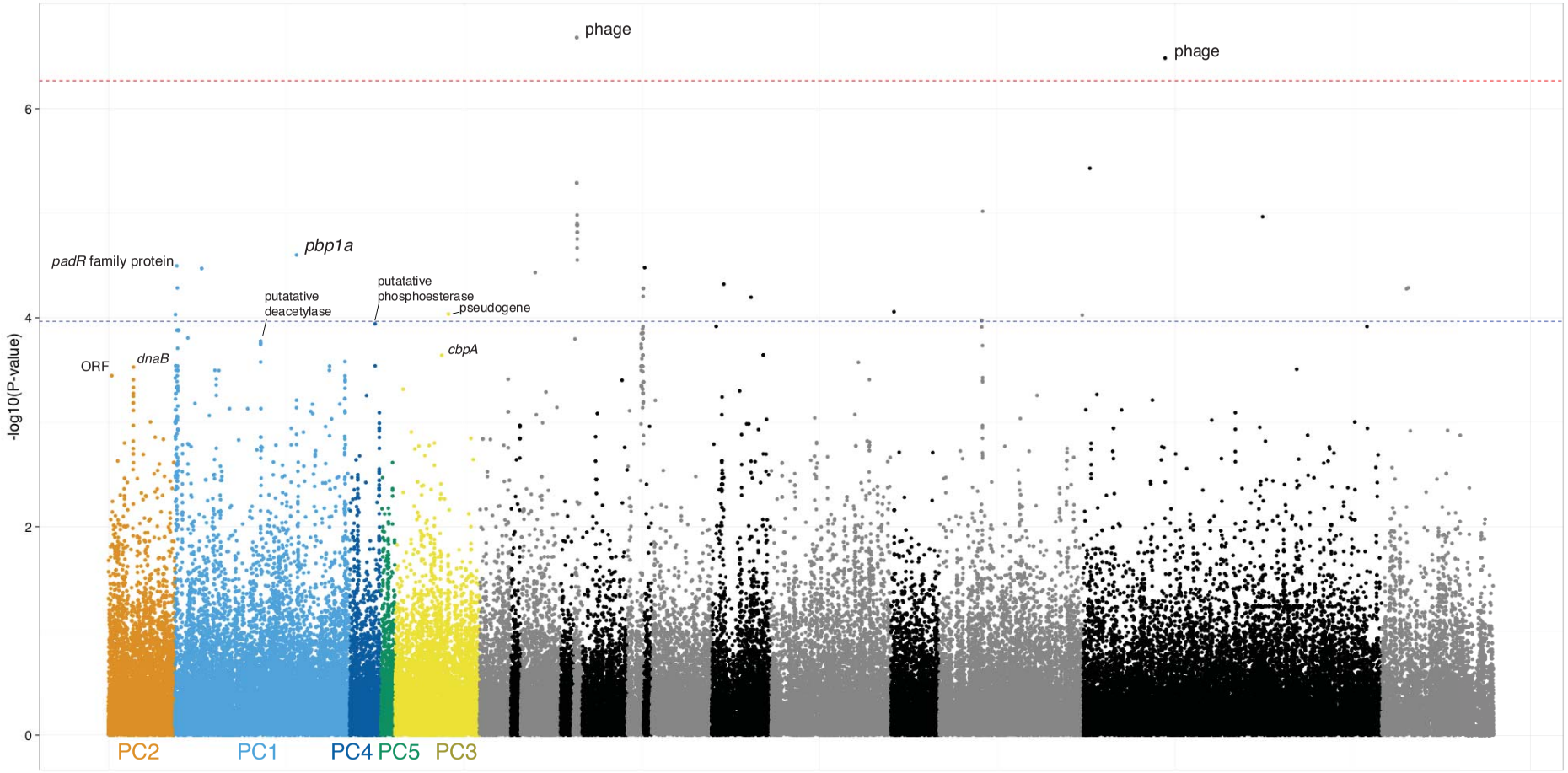

Figure A.9: Possible SNPs associated with lineage and carriage duration. The SNPs and p-values as shown in fig. 3.6, however the x-axis is now ordered by strength of association of lineage (defined by principal component) with carriage duration. The left most lineages are those most associated, those in black/grey were not significantly associated. SNPs are coloured by the lineage they are most associated with.

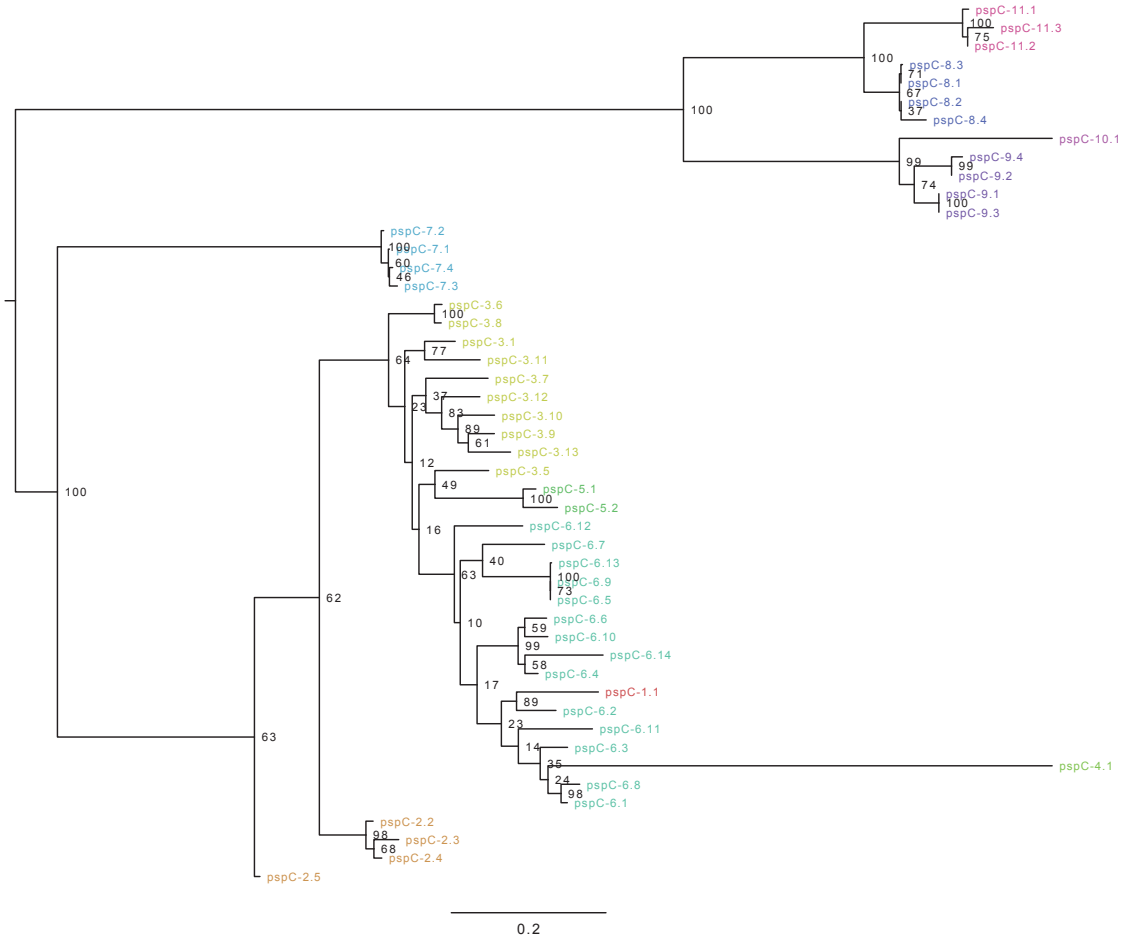

Figure A.10: Maximum likelihood tree of *pspC* protein alignment, with 100 bootstrap replicates (nodes are labelled with bootstrap supports). Tips are coloured by allele group.

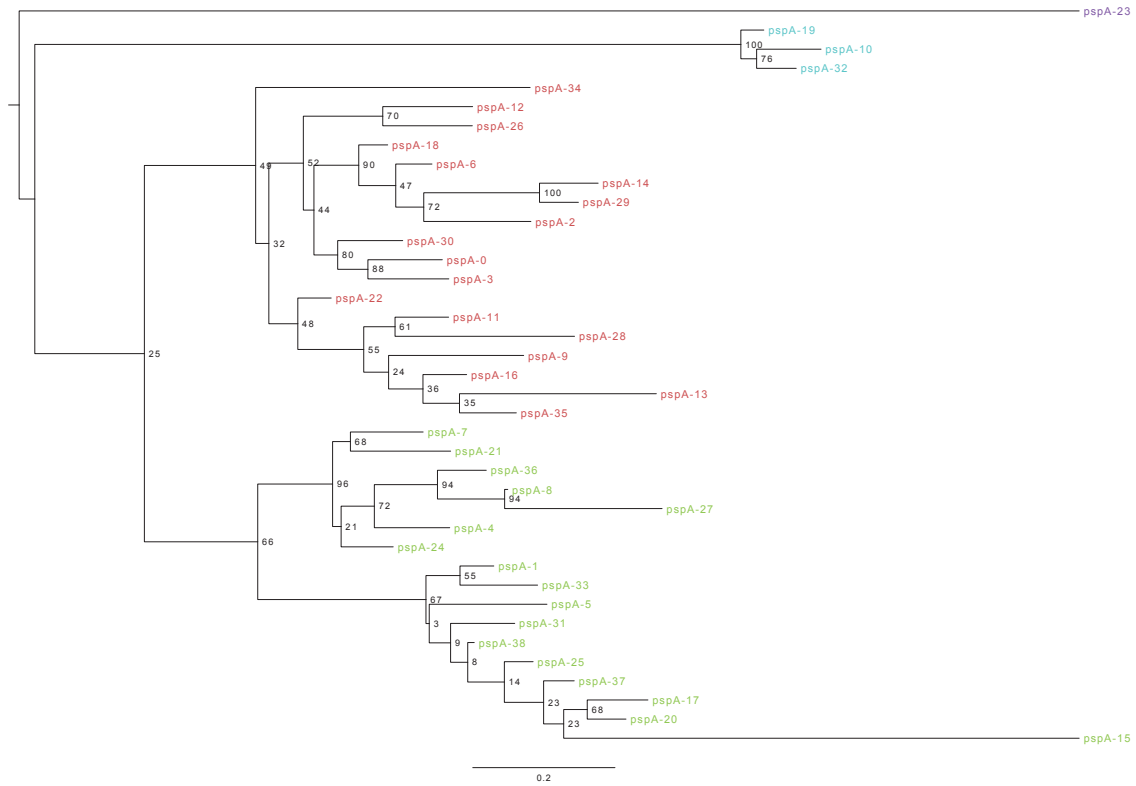

Figure A.11: Maximum likelihood tree of *pspA* protein alignment, with 100 bootstrap replicates (nodes are labelled with bootstrap supports). Tips are coloured by allele group.

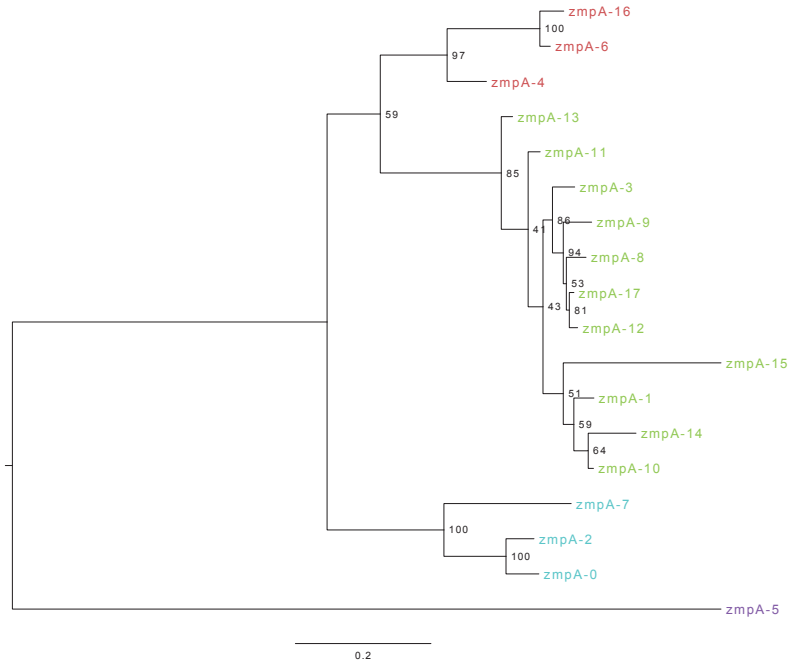

Figure A.12: Maximum likelihood tree of *zmpC* protein alignment, with 100 bootstrap replicates (nodes are labelled with bootstrap supports). Tips are coloured by allele group.

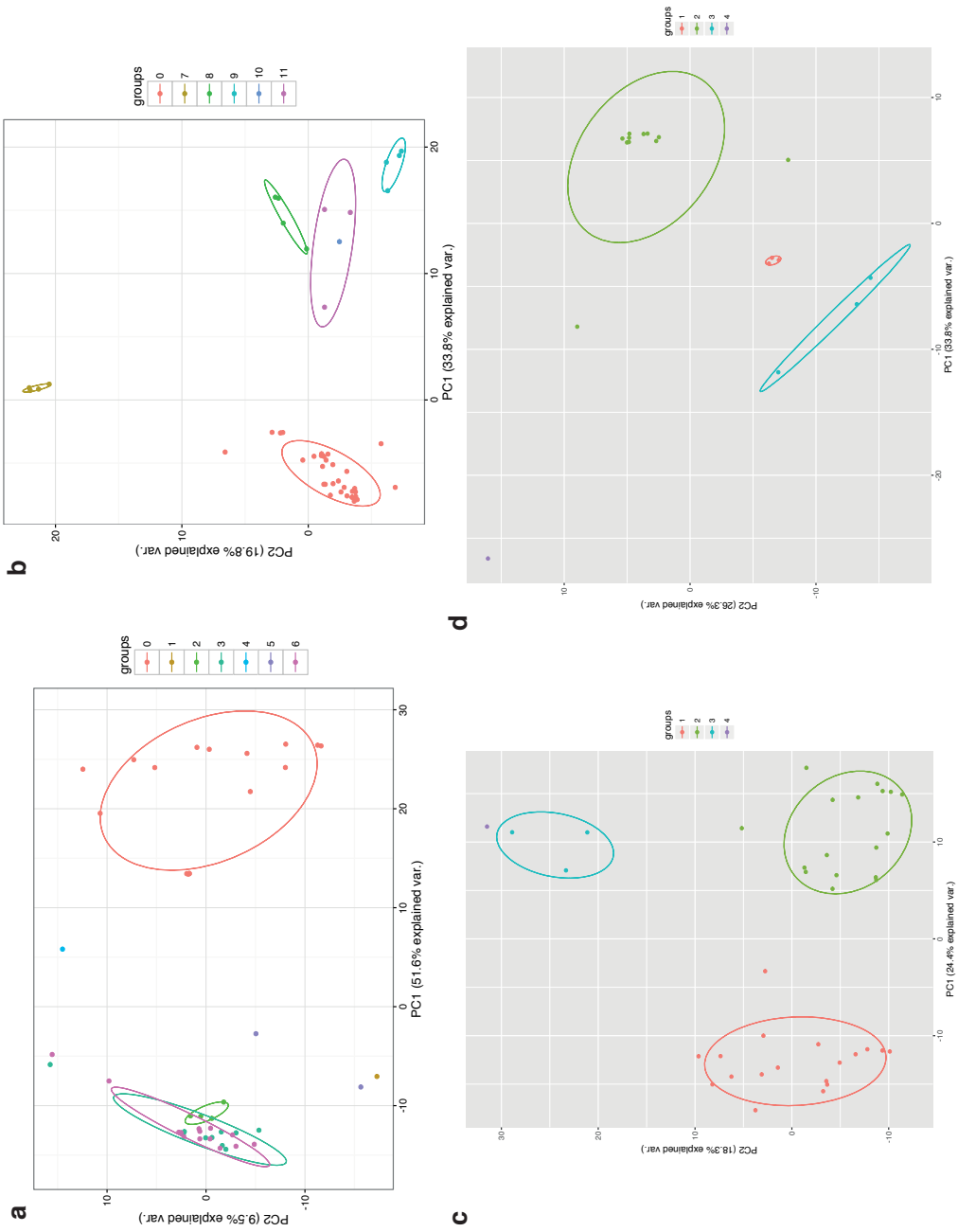

Figure A.13: PCA plots of classifiers used on antigen training data, first two principal components shown in each case. Points are coloured by the allele number, where 0 is a genome without the antigen. Where more than one point is available for a class, an ellipse has been drawn around its centroid. **a)** PspC, alleles 0–6; **b)** PspC, alleles 0,7–11; **c)** PspA, alleles 1–4; **d)** ZmpA, alleles 1–4

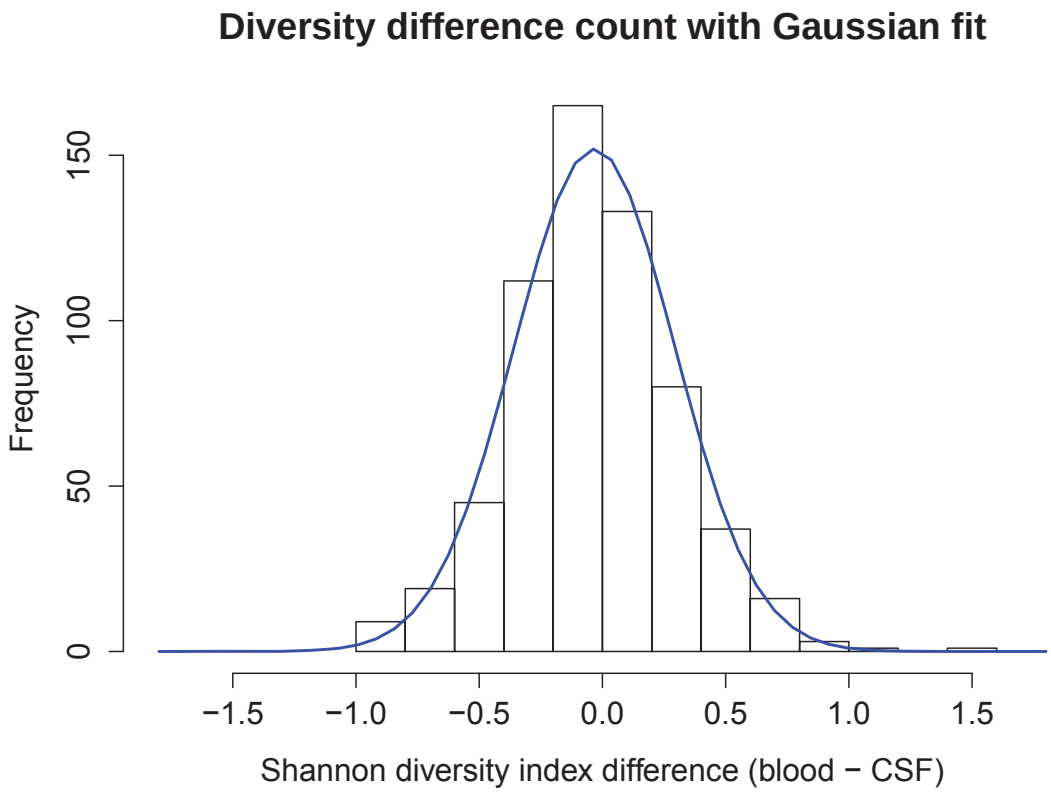

Figure A.14: Distribution of difference in Shannon diversity index between the *ivr* locus model π_{blood} and π_{CSF}. A Gaussian distribution was fitted to the data, which has a mean of roughly zero and little skew. The maximum possible Shannon diversity index (for equal amounts of each allele A-F) is 1.8.

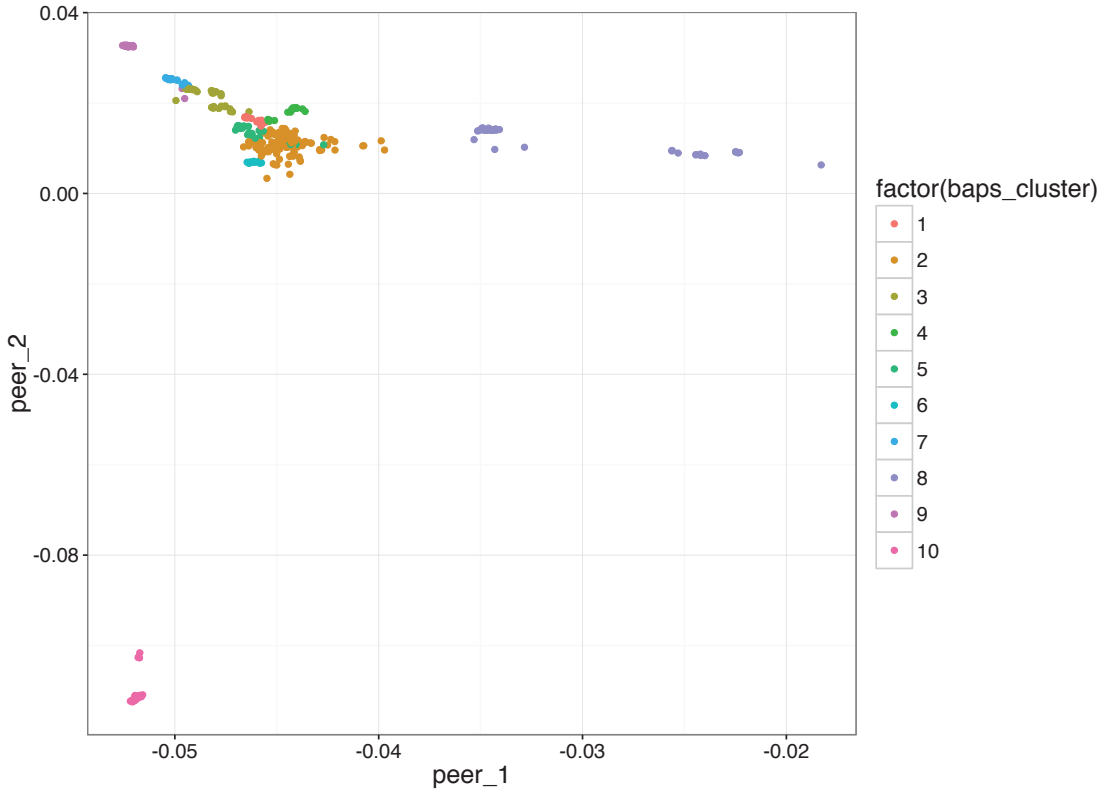

Figure A.15: Plot of the samples in the genome-to-genome analysis. x-axis is the first PEER factor loading, y-axis is the second PEER factor loading. Sample are coloured by the BAPS cluster they were assigned to.