Bacterial genotype deconvolution in shotgun metagenomic reads using fuzzy genotypes

Byron J. Smith[1], Katherine S. Pollard[1,2,3]

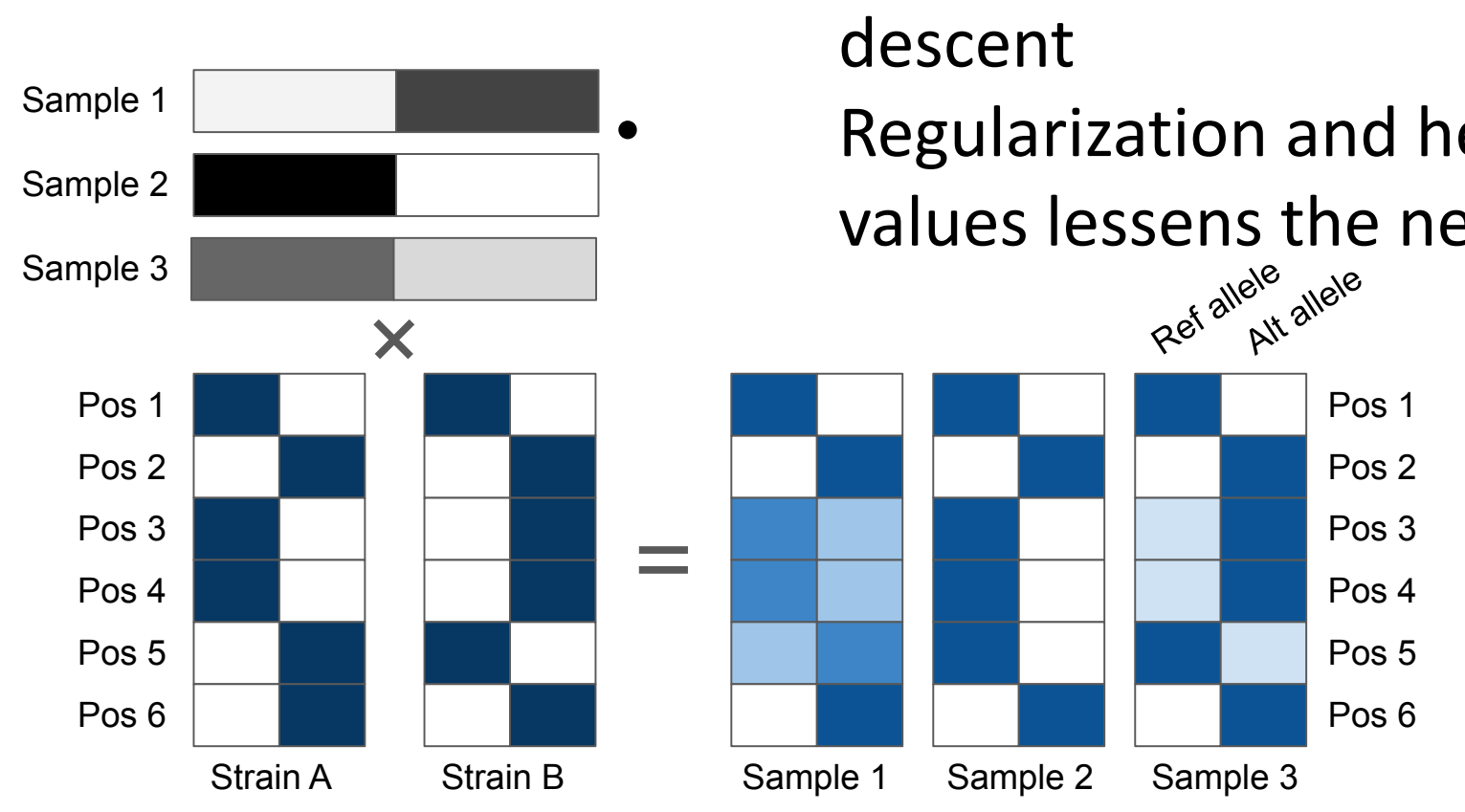
[1]Gladstone Institute of Data Science and Biotechnology, [2]UCSF Epidemiology & Biostatistics, [3]Chan Zuckerberg Biohub

Probabilistic Modeling
in Genomics - CSHL 2021

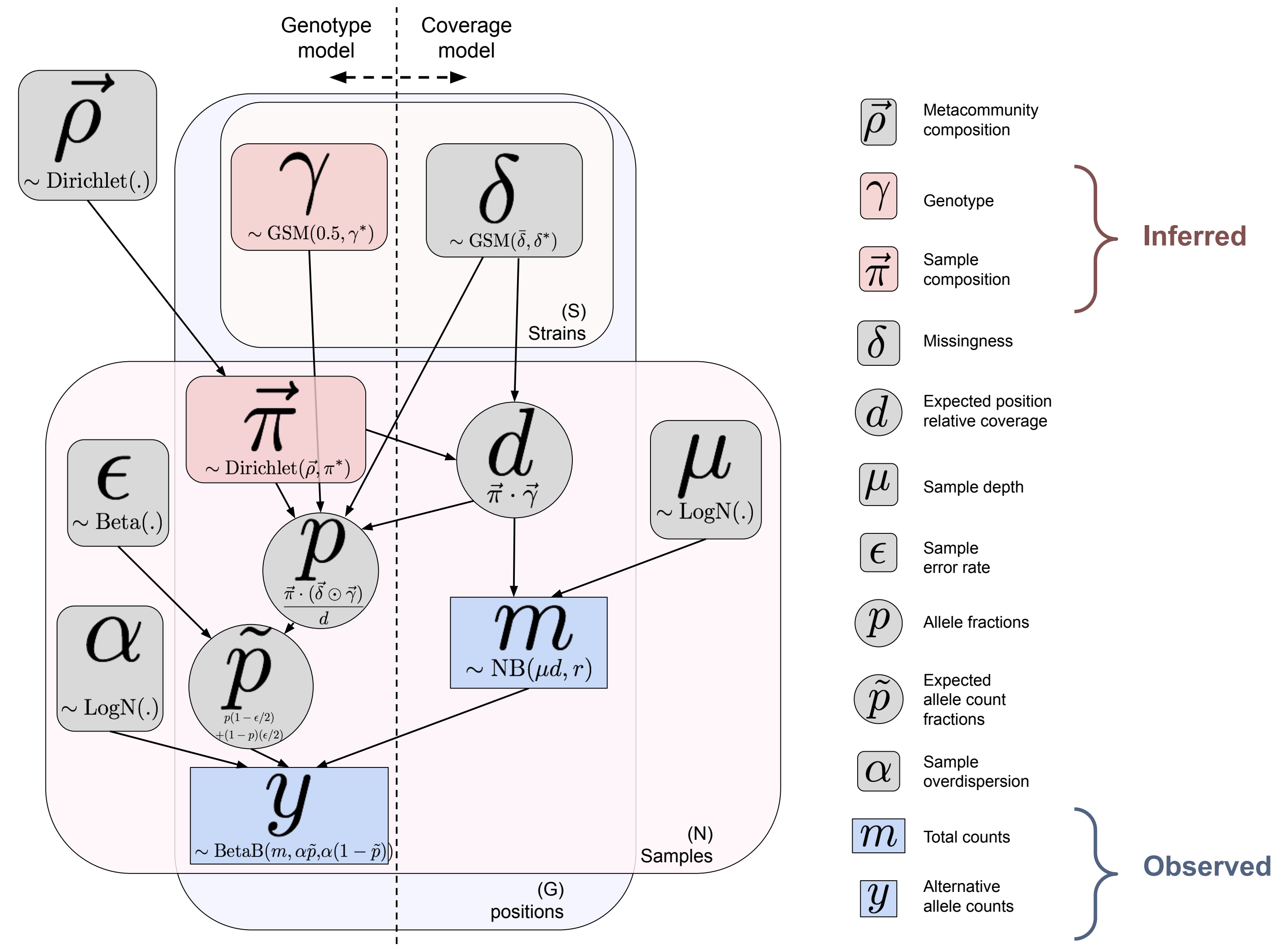
https://byronjsmith.com/probgen2021_poster.pdf

TODO: Very brief summary

- Summarize the summary: taxonomic estimation usually ignores strain diversity, or approximates it using SNP/gene-content similarity as a proxy.
- Factorization methods are a more principled approach:
  - Enable analysis of mixtures of genotypes
  - Easily accommodates missing data
  - Better ways to asses confidence
- Available tools for strain factorization are slow and require fitting multiple models to choose best parameterization
- Here I describe a new model-based approach which harnesses a fully differentiable (fuzzy) genotype model
  - This allows models to be fit very quickly using gradient descent
  - Regularization and heuristic algorithms for selecting initial values lessens the need for multiple fits to be compared.



TODO: The Model/Method


Genotype model
Coverage model
$\vec{\rho}$ ~ Dirichlet(.)
γ ~ GSM(0.5, γ^*)
δ ~ GSM($\bar{\delta}$, δ^*)
(S) Strains
$\vec{\pi}$ ~ Dirichlet($\vec{\rho}$, π^*)
d $\vec{\pi} \cdot \vec{\gamma}$
ϵ ~ Beta(.)
μ ~ LogN(.)
p $\frac{\vec{\pi} \cdot (\vec{\delta} \odot \vec{\gamma})}{d}$
α ~ LogN(.)
$\tilde{p}$ $p(1-\epsilon/2) + (1-p)(\epsilon/2)$
m ~ NB(μd, r)
y ~ BetaB(m, $\alpha\tilde{p}$, $\alpha(1-\tilde{p})$)
(N) Samples
(G) positions
$\vec{\rho}$ Metacommunity composition
γ Genotype
$\vec{\pi}$ Sample composition
Inferred
δ Missingness
d Expected position relative coverage
μ Sample depth
ϵ Sample error rate
p Allele fractions
$\tilde{p}$ Expected allele count fractions
α Sample overdispersion
m Total counts
y Alternative allele counts
Observed

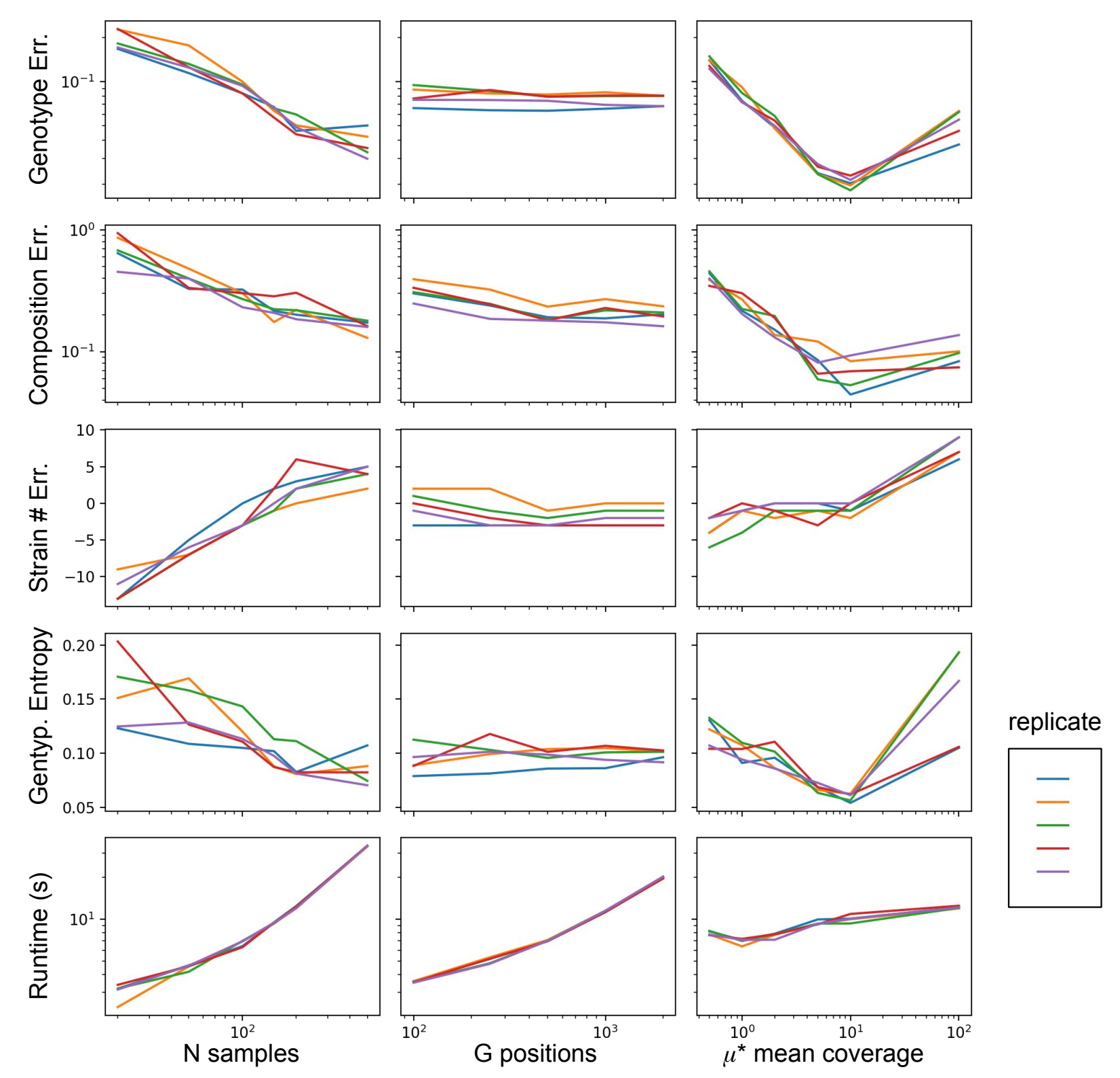
TODO: Major results

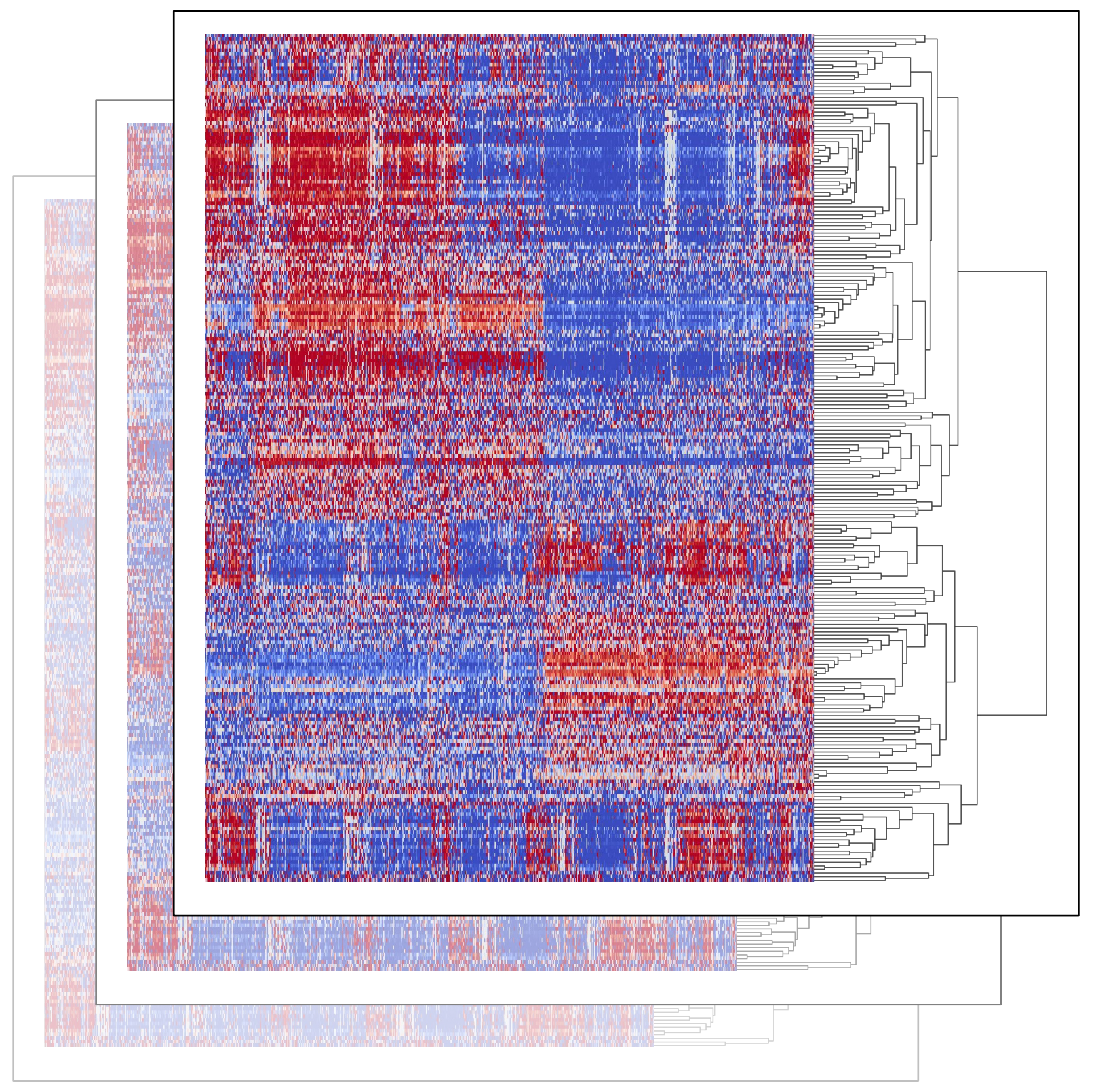
Model is **fast and accurate** in simulations with realistically challenging data: low-coverage, noisy, heavily admixed, and high strain diversity.

Key performance measures are (1) genotype error (mean abundance weighted squared deviation from ground truth with “fuzziness” adjustment), (2) compositional error (RMSE of all, pairwise sample Bray-Curtis dissimilarities normalized to the expected value), and (3) total runtime to parameter convergence.

Increasing the number of samples, the number of genome positions, or the mean sample coverage all generally improve model fit with approximately **linear runtime**.

Fit to metagenotype data produced by e.g. GT-PRO, StrainPhlAn, MIDAS, inStrain, etc., our model produces plausible and interpretable results.





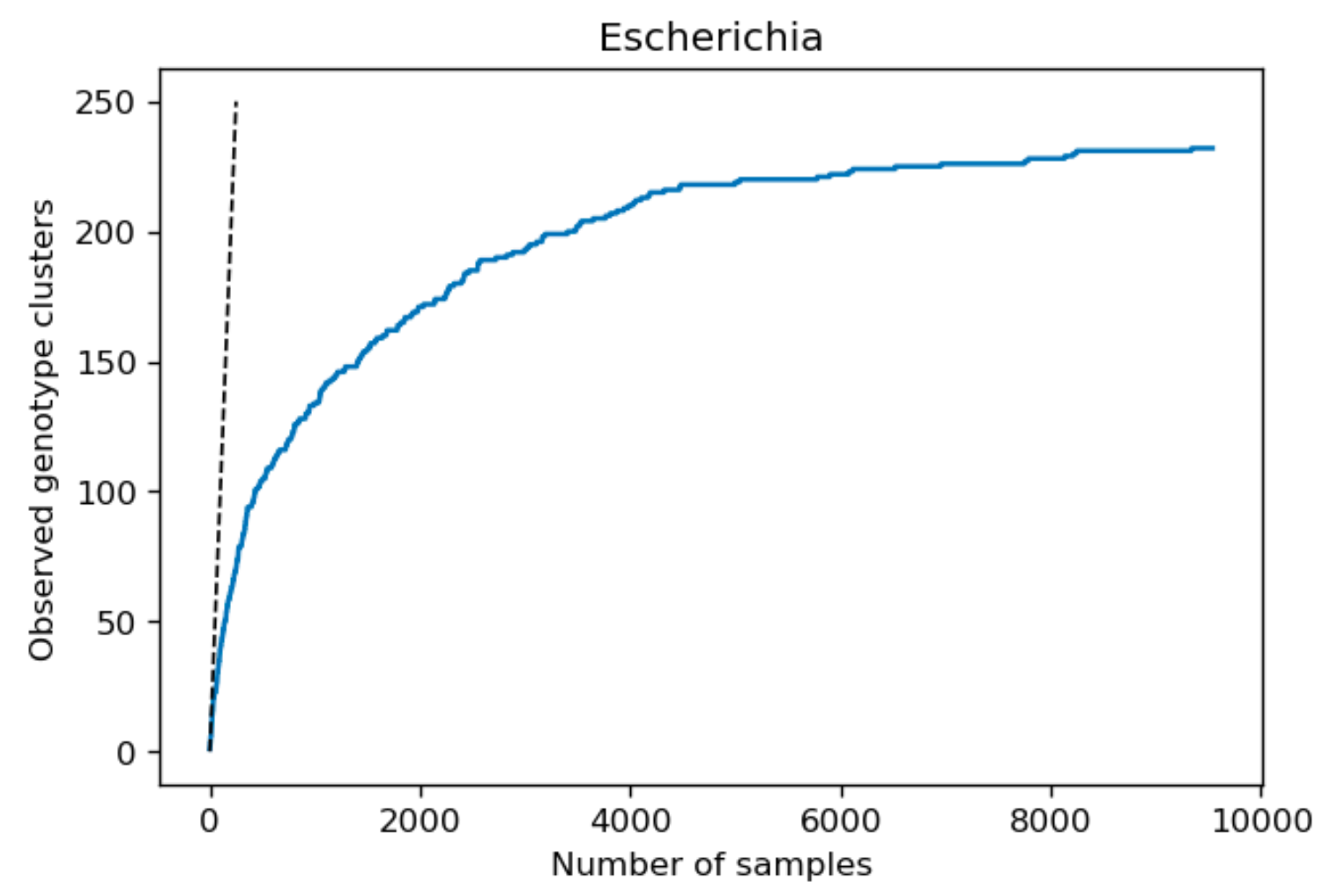
For instance, we fit this model to data previously processed by GT-PRO for the genus *Escherichia* from 9540 human microbiome metagenomes across a subsample of 1000 SNP positions (maximum major allele frequency of 90%). Fitting the model took 185 seconds on a GPU.

Inferred genotypes can be further filtered using a variety of metrics (such as estimated abundance, genotype entropy, source sample error, etc.), and further consolidated into clusters of highly similar genotypes.

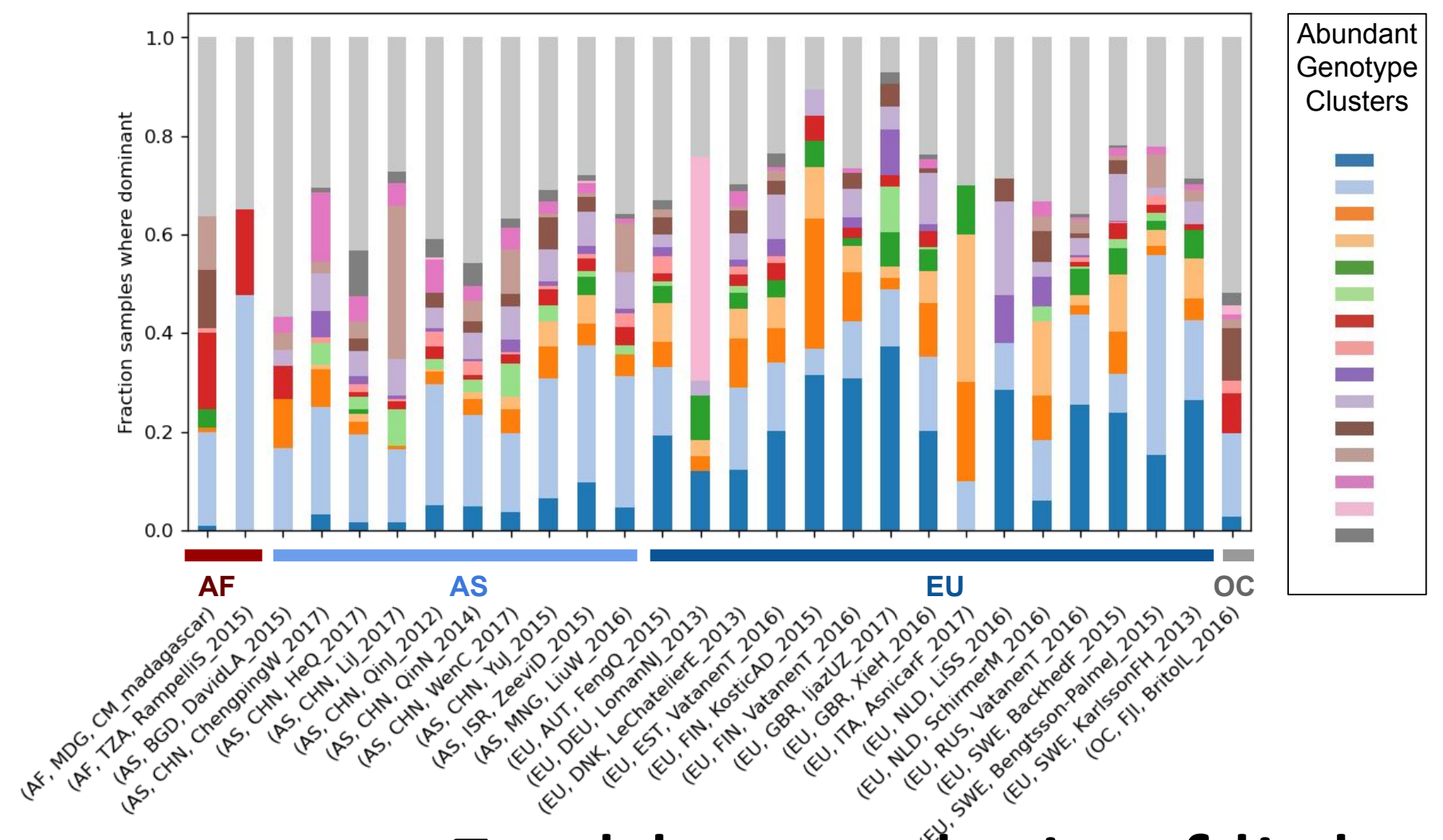
In our analysis of Escherichia we identified 232 distinct genotype clusters.

Analyses of inferred genotypes may yield insights into the population structure and evolution of strains without the need for cultured representatives, and can be carried out in parallel for any species for which metagenotype data can be generated.

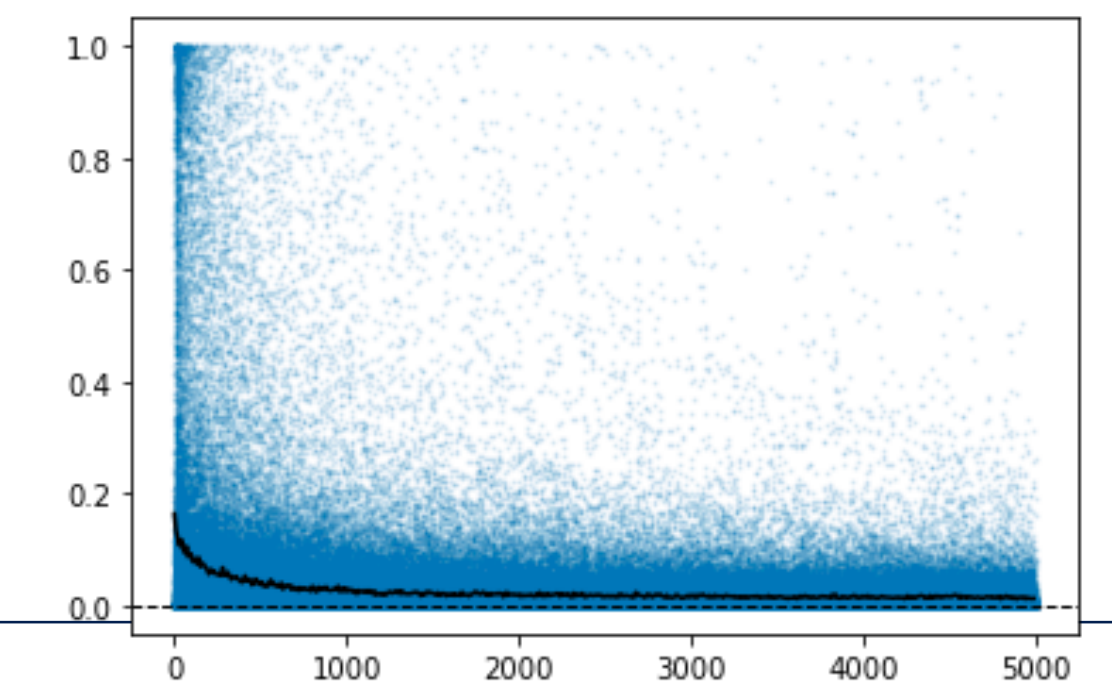
Based on rarefaction curves and a Chao1 richness estimator, (and subject to strong assumptions) this analysis suggests that we may have identified a majority of the *Escherichia* genotype diversity in the surveyed host population.



Across all of the studies combined in these data, some *Escherichia* groups are more likely to be dominant within specific continents and countries, although most high incidence strains are dominant in at least one sample in he majority of included studies.



Enables analysis of linkage disequilibrium / recombination / etc.

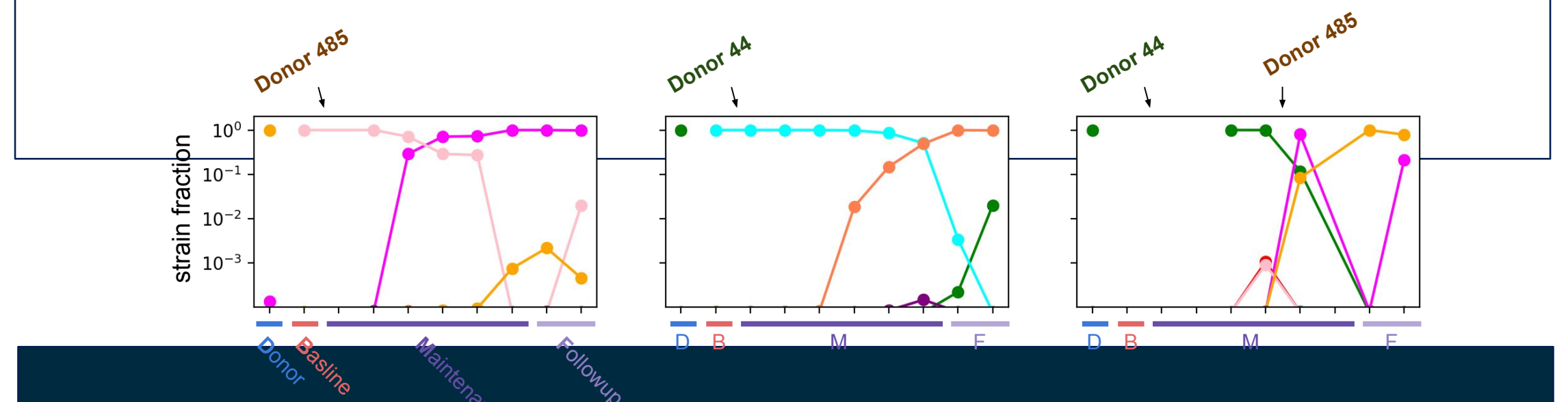


TODO: Impact

Bullets with bold text for most important impacts:

- Accurate, sensitive, and far more resolution than other taxonomic methods.
- Careful modeling of biological noise and tunable regularization enable interpretable results.
- Scales easily to very large data, especially using GPUs
- This enables study of diversity and evolution without culturing.

Strain tracking works well:



Footnotes and Citations

TODO: GT-PRO/StrainPhlan/MIDAS StrainFinder, UHGG, Pyro

Acknowledgments & Contact

This work was supported by an NIH T32 training grant TODO.

TODO: Funding from Biohub, Gladstone

TODO: n ByronJSmith

Updated Poster: https://byronjsmith.com/probgen2021_poster.pdf