Probabilistic Modeling in Genomics

CSHL 2021

Bacterial genotype deconvolution in shotgun metagenomic reads using fuzzy alleles

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

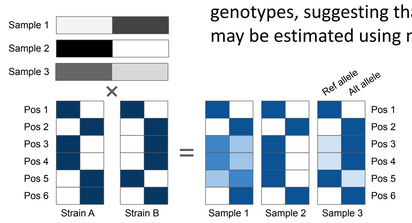
The most recent version of this poster is available at:

https://byronjsmith.com/probgen2021_poster.pdf

This is version: 2021-04-12a

Introduction

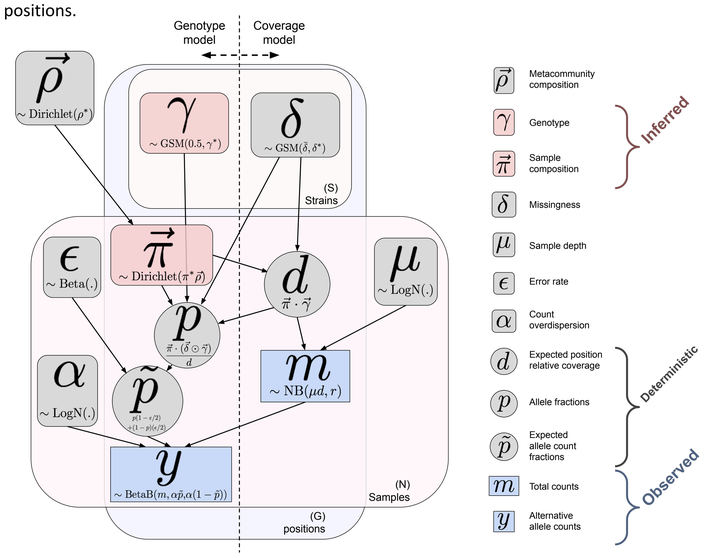
- **Intraspecific variation** is widespread and functionally important in microbial systems including the **human microbiome**.
- Standard methods to study communities often do not measure this level of taxonomic detail. Nonetheless, metagenomic sequence data in principle encodes rich, strain-level information.
- Here we refer to the observed counts of various alleles at polymorphic positions as a "**metagenotype**". Both read mapping and exact K-mer based approaches have been developed to process metagenomic reads into metagenotypes for species of interest[1,2,3,4].
- The expected proportions of allele counts across sites can be approximated as a linear combination of unmeasured strain proportions and their respective genotypes, suggesting that these latent parameters of interest may be estimated using methods related to **matrix factorization**



- However, existing tools[5,6] to do so have not yet been widely adopted, due in part to computational limitations and biological deviations from this simple model.

Metagenotype model

We constructed a detailed, generative model for metagenotype data across **bi-allelic sites**, which extends the simple multiplicative model described above. Importantly, genotypes in our model need not be purely the reference or alternative allele, but instead are **"fuzzy"** varying between 0.0 (entirely reference) and 1.0 (entirely alternative). Unlike similar, previously published approaches[5,6], this makes parameters in our model **fully differentiable**, and allows us to harness gradient-based optimization methods for maximum a posteriori (MAP) estimation. Of note, our approach models allele counts as over-dispersed compared to a binomial process, and also allows for the possibility that some strains are missing genome positions.

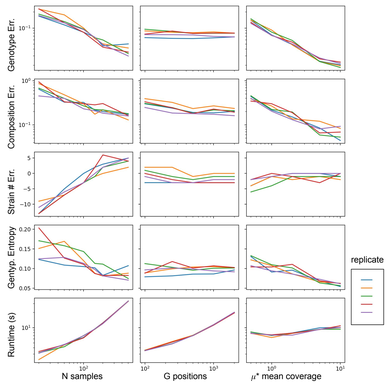


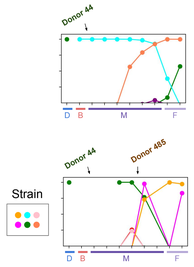
Major results

Inference of genotypes and their relative abundances using this model are **fast and accurate** in simulations, even with challenging data: low-coverage, noisy, heavily admixed, and high strain diversity, reflecting a frequent reality of metagenomic libraries.

We measured performance based on three metrics: (1) genotype error (mean abundance weighted squared deviation from ground truth with adjustment for "fuzziness"), (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 (N), the number of genome positions (G), or the mean sample coverage (μ^*) all generally improve model accuracy with approximately **linear increases in runtime**.





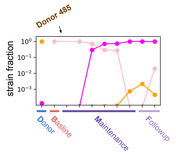
Our model produces plausible and interpretable results when fit to metagenotype data produced by **GT-PRO**[1].

In a study of fecal microbiome transplant (FMT) as a treatment for ulcerative colitis, we found that genotypes were quickly and stably transferred from donors to patients.

We went on to fit our model to a large corpus of publicly available human microbiome metagenomes previously processed by GT-PRO. For the genus *Escherichia*, **9540 samples** had sufficient coverage to be included, and we subsampled **1000 SNP positions** (major allele frequency of <90%). Fitting the model took **185 seconds on a GPU**.

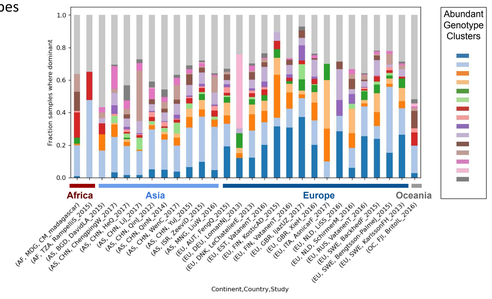
Inferred genotypes can be filtered using a variety of metrics (such as estimated abundance, genotype entropy, source sample error, etc.), and further consolidated into clusters of highly similar genotypes.

In our analysis of *Escherichia* we identified **232 distinct genotype clusters**. Based on rarefaction curves and the Chao1 richness estimator, our analysis suggests that we may have cataloged a majority of the *Escherichia* genotype diversity in the surveyed host population.

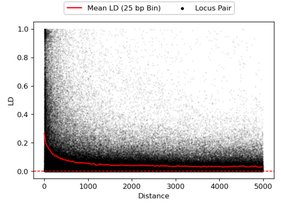


Analysis of inferred genotypes may yield insights into the **population structure and evolution of strains without the need for cultured representatives**, and can be carried out for any species where metagenotype data can be collected.

These inferences can also inform analysis of **microbial biogeography**. Across all of the samples combined here[1], certain *Escherichia* groups were more likely to be dominant within specific studies, and have visibly different dominance rates across four continents. Nonetheless, all high incidence strains are dominant in at least one sample in the majority of included studies.

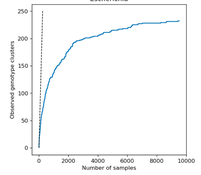


To demonstrate the potential value of genotype inferences for evolutionary studies, we re-fit metagenotype data for 13394 SNP positions with major allele frequencies of <90%. Conditioning on parameters estimated in the previous run from subsampled data, we fit genotypes by iterating over all positions in blocks of 1000 positions. This process took an additional 360 seconds.



Linkage disequilibrium was estimated for all locus pairs across the 232 inferred genotype clusters. Mean LD over all pairs was 0.022. LD was stronger for locus pairs closer together in the reference genome for the species. For loci within 100 bases of each other, mean LD was 0.217, and between 100 and 200 bases it was 0.143, indicating that this trend was not an artifact of nearby loci being found on the same read.

This rapid decay of statistical linkage with linear distance may reflect extensive genome recombination between populations of *Escherichia*, a phenomenon that has been previously described using isolate genomes[7,8].

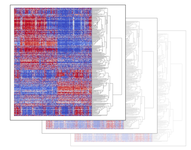


Challenges & Lessons Learned

- The **Pyro probabilistic programming framework**[9] makes model development and MAP estimation easy and fast. Scalability on GPUs is an added bonus.

Open problems

- **Strain number misspecification** negatively impacts inference. Attempts to induce sparsity with the Dirichlet prior on ρ seems to be ineffective.
- It is not clear how to best incorporate parameter uncertainty. Full **Bayesian inference** is challenging given the multi-modal posterior, and attempts to apply variational methods have had numerical problems.
- Some nuisance parameters (e.g. α) collapse to **biologically implausible values** even with strong priors. It is not yet apparent how this affects estimates.



Footnotes

1. Shi ZJ, Dimitrov B, Zhao C, Nayfach S, Pollard KS. 2020. Ultra-rapid metagenotyping of the human gut microbiome. *bioRxiv* 2020.06.12.149336.
2. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. 2017. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research* 27:626–638.
3. Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Research* 26:1612–1625.
4. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. 2021. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nature Biotechnology* 1–10.
5. Smillie CS, Sauk J, Gevers D, Friedman J, Sung J, Youngster I, Hohmann EL, Staley C, Khoruts A, Sadowsky MJ, Allegretti JR, Smith MB, Xavier RJ, Alm EJ, 2018. Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. *Cell Host and Microbe* 23:229-240.e5.
6. Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, Eren AM. 2017. DESMAN: A new tool for de novo extraction of strains from metagenomes. *Genome Biology* 18:1–22.
7. Arevalo P, VanInsberghe D, Elsherif M, Gore J, Polz MF. 2019. A reverse ecology approach based on a biological definition of microbial populations. *Cell* 178:820-834.e14.
8. Frazão N, Sousa A, Lässig M, Gordo I. 2019. Horizontal gene transfer overrides mutation in Escherichia coli colonizing the mammalian gut. *PNAS* 116:17906–17915.
9. Bingham E, Chen JP, Jankowiak M, Obermeyer F, Pradhan N, Karaletsos T, Singh R, Szerlip P, Horsfall P, Goodman ND. 2019. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research* 20:1–6.

Acknowledgments & Contact

This work was supported by an NIH T32 training grant 5T32DK007007.

Special thanks to Zhou Jason Shi for the contribution of and extensive discussion about GT-PRO metagenotype data.

@ByronJSmith