Bacterial genotype deconvolution in shotgun metagenomic reads using fuzzy genotypes

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

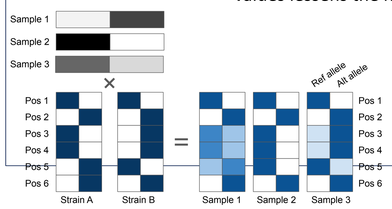
[1]Gladstone Institute of Data Science and Biotechnology, [2]UCSF Epidemiology & Biostatistics, [3]Chan Zuckerberg Biohub

Probabilistic Modeling in Genomics - CSHL 2021

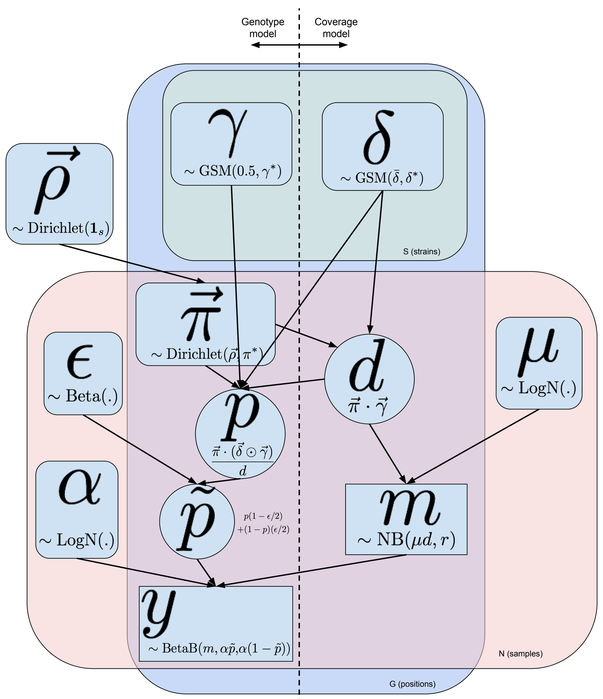
https://byronjsmith.com/probgen2021_poster.pdf

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.

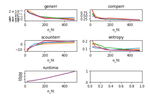


TODO: The Model/Method

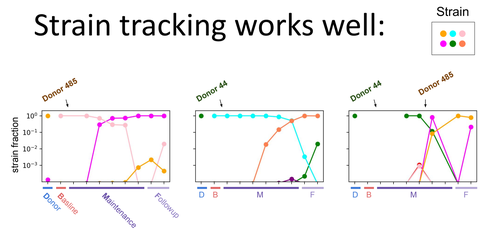


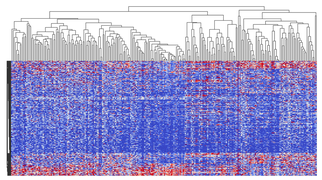
TODO: Major results

Model fits very quickly and accurately even on low quality data



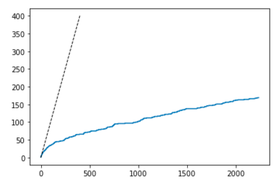
Strain tracking works well:



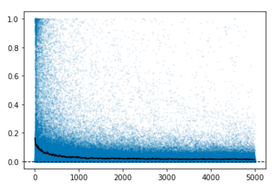


Can fit very large data and produce meaningful results

Huge strain diversity: Chao2 ~ 700 for e.g. F. prausnitzii



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.

Footnotes and Citations

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

Acknowledgments & Contact

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

GLADSTONE INSTITUTES
SCIENCE OVERCOMING DISEASE

ByronJSmith

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

CC BY