ChronoStrain: Sequence quality and time aware strain tracking with shotgun metagenomic data

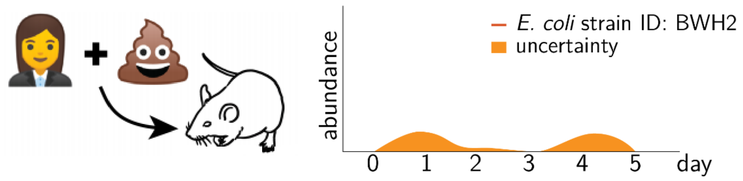
Younhun Kim[1], Sawal Acharya[2], Daniel Alfonsetti[1], Georg K. Gerber[2,3], Bonnie Berger[1], Travis E. Gibson[2,3]

younhun@mit.edu

[1]MIT, [2]Massachusetts Host Microbiome Center, Brigham and Women's Hospital, [3]Harvard Medical School

Introduction

We present a sequence quality and time aware model for tracking microbial strains in shotgun metagenomic data. The motivating application of this model is the tracking of low abundance pathogens in longitudinal human and murine studies. We use a maximum a posteriori (MAP) inference algorithm and illustrate its efficacy on synthetic data. We explicitly include quality scores (beyond simply trimming or removing low quality reads before mapping) and time of sample collection, together, for learning the abundance of microbial strains from shotgun data. Our results show that using time-correlations and accounting for quality scores results in a more efficient use of samples.



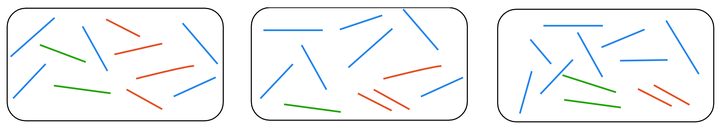
Problem Statement

Given: Time-indexed collection of metagenomic shotgun reads from an individual.

Output: Time-indexed collection of relative abundance vectors (of strains).

Assumptions:

- Reference database of strain-identifying markers & their copy numbers.
- Pre-filtering of reads that align to reference markers.

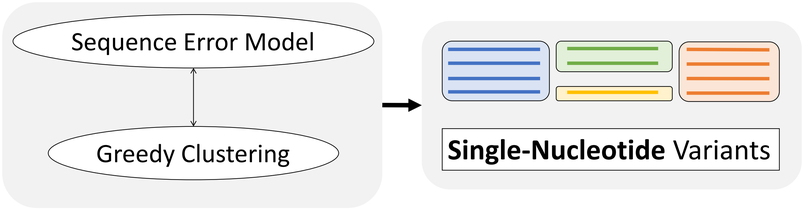


A: 40% B: 40% C: 20%

A: 20% B: 70% C: 10%

A: 15% B: 70% C: 15%

Single Timepoint Example: DADA2 (16S Amplicon)



Idea: Sequence error model allows for SNVs in strain-level markers to be incorporated into inference. Exclude commonly-used "Reference Alignment"-based preliminary bucketing by guesses of strain origin.

The Model

How to enable fine-grained inference?

As a motivational point, we are concerned with the loss of information when using quality-naive alignment or mapping tools to categorize reads.

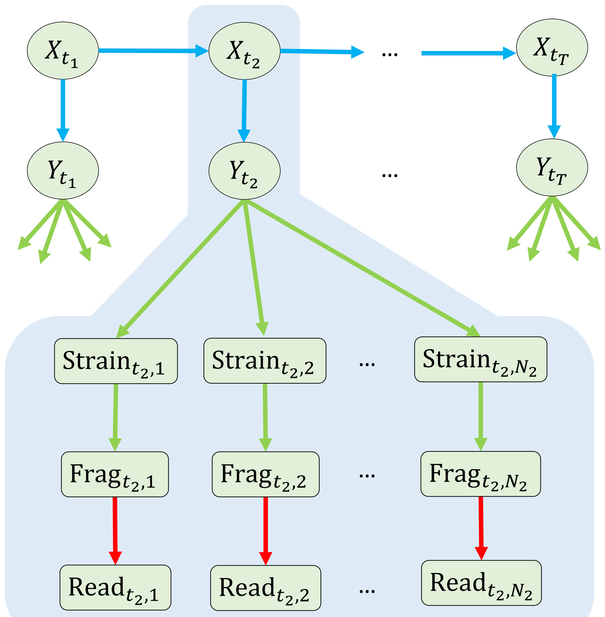
When is joint inference across time helpful?

A common issue when using single time-point methods independently across samples is *sample-deficiency of low-abundance strains.*

Bayesian Model

Relative abundance: $Y_1, Y_2, \ldots, Y_T$

Latent representation: $X_1, X_2, \ldots, X_T$



Latent Dynamics

$$X_{t_1} \sim \mathcal{N}(\mathbf{0}, \tau_0 I)$$
$$X_{t_j} | X_{t_{j-1}} \sim \mathcal{N}(\mathbf{0}, \tau(t_j - t_{j-1}) I)$$
$$Y_t = \text{softmax}(X_t)$$

Shotgun Sampling

$$\text{Strain}_j \sim \text{Categorical}(Y_j)$$
$$\text{Frag}_j \sim \text{Uniform}(\text{Frag}(\text{Strain}_j))$$

Error Model

$$\text{Read} = (\text{Seq}, \text{Quality})$$
$$\mathbb{P}(Seq = s | \text{Frag} = f, \text{Qual} = q) \propto \prod_{\ell=1}^{L} e(f_\ell \to s_\ell | q_\ell)$$

Inference Method

Goal: Maximum a Posteriori:

$$\hat{X}_{MAP} = \underset{X}{\text{argmax}}\, \mathbb{P}(X \mid \text{Reads})$$

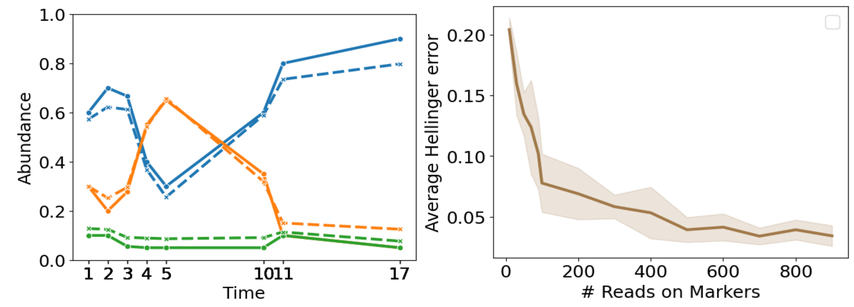
Algorithm: Expectation-Maximization*

$$\hat{X}^{j+1} \leftarrow \underset{X}{\text{argmax}}\, \mathbb{E}_F[\log p(X, Frag = F, Reads) \mid \hat{X}^j, Reads]$$

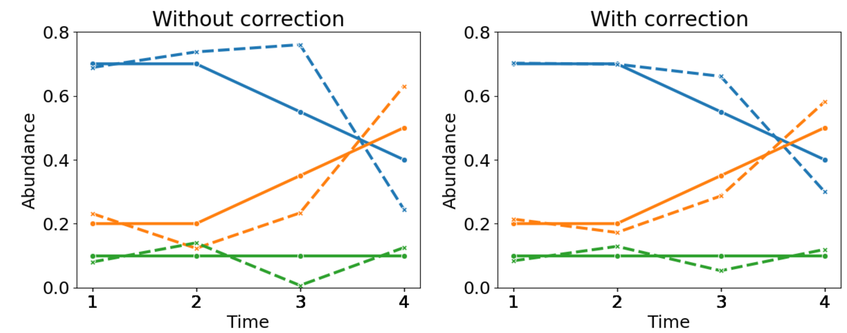
**Caveat: nonlinearity of softmax function makes explicit argmax impossible. We use a gradient-ascent update scheme instead.*

Test on Synthetic Data

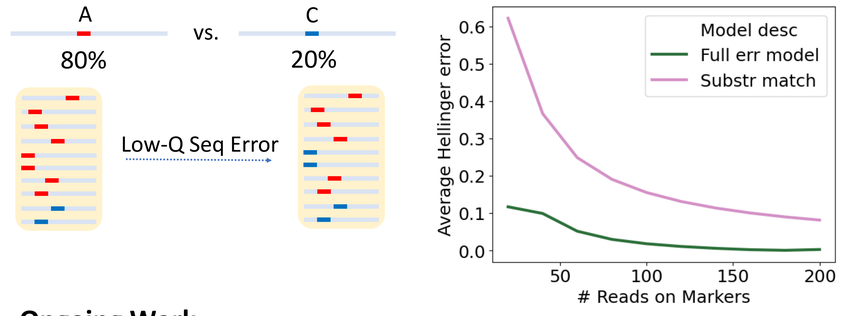
Tests for correctness & sample efficiency $(\text{Hellinger}(P, Q) = \|\sqrt{P} - \sqrt{Q}\|_2)$



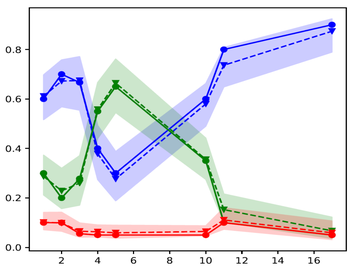
Time correlation effect at low samples (200 marker reads / time pt.)



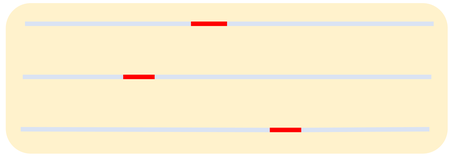
Error model effect on single-nucleotide errors at low samples



Ongoing Work



Full posterior approximation



de novo learning of marker variants

- Development as a bioinformatics tool, with marker-based filtering
- Learning error model on-the-fly
- Longitudinal study in mice