# Longitudinal profiling of low-abundance strains in microbiomes with ChronoStrain

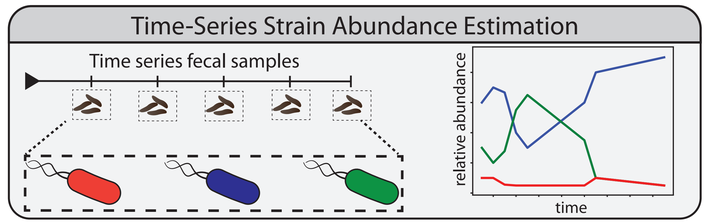

***Younhun Kim*** [1,2,4], Colin Worby [3], Sawal Acharya [2], Lucas R. van Dijk [3], Zachary Gromko [1], Philippe Azimzadeh [5], Karen Dodson [5], Georg K. Gerber [1,2,4], Scott Hultgren [5], Ashlee Earl [3], Bonnie Berger [1,4], Travis E. Gibson [1,2,3,4]

younhun@bwh.harvard.edu
tegibson@bwh.harvard.edu
bab@csail.mit.edu

[1] MIT, [2] Brigham and Women's Hospital, [3] Broad Institute, [4] Harvard Medical School, [5] Washington University in St. Louis


## Abstract

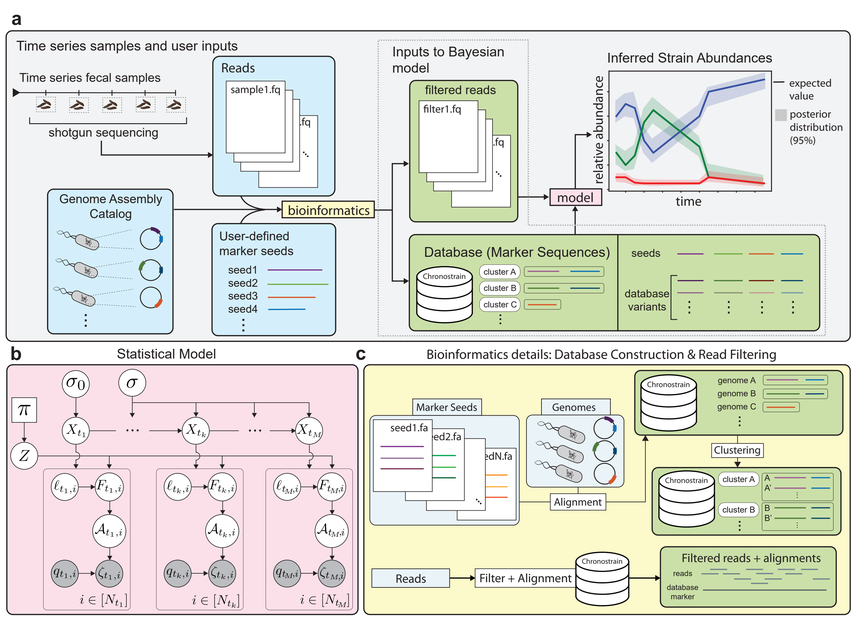

The ability to detect and quantify microbiota over time from shotgun metagenomic data has a plethora of clinical, basic science and public health applications. Given these applications, and the observation that pathogens and other taxa of interest can reside at low relative abundance, there is a critical need for algorithms that accurately profile low-abundance microbial taxa with strain-level resolution. Here we present ChronoStrain: a sequence quality- and time-aware Bayesian model for profiling strains in longitudinal samples. ChronoStrain explicitly models the presence or absence of each strain and produces a probability distribution over abundance trajectories for each strain. Using synthetic and semi-synthetic data, we demonstrate how ChronoStrain outperforms existing methods in abundance estimation and presence/absence prediction. Applying ChronoStrain to two human microbiome datasets demonstrated its improved interpretability for profiling Escherichia coli strain blooms in longitudinal faecal samples from adult women with recurring urinary tract infections, and its improved accuracy for detecting Enterococcus faecalis strains in infant faecal samples. Compared with state-of-the-art methods, ChronoStrain's ability to detect low-abundance taxa is particularly stark.

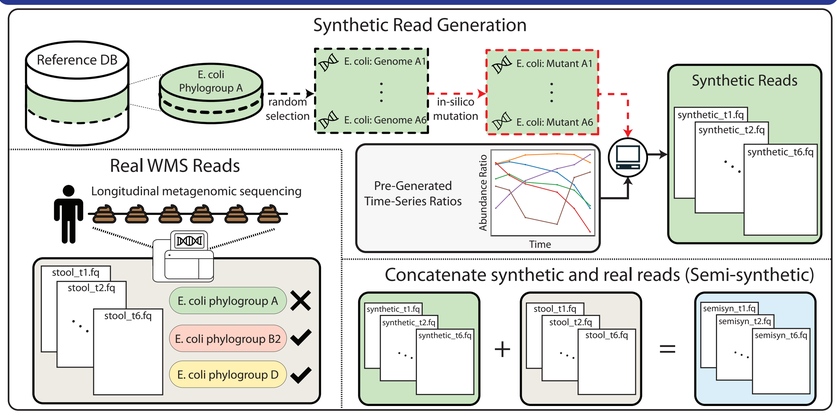

## Method Overview

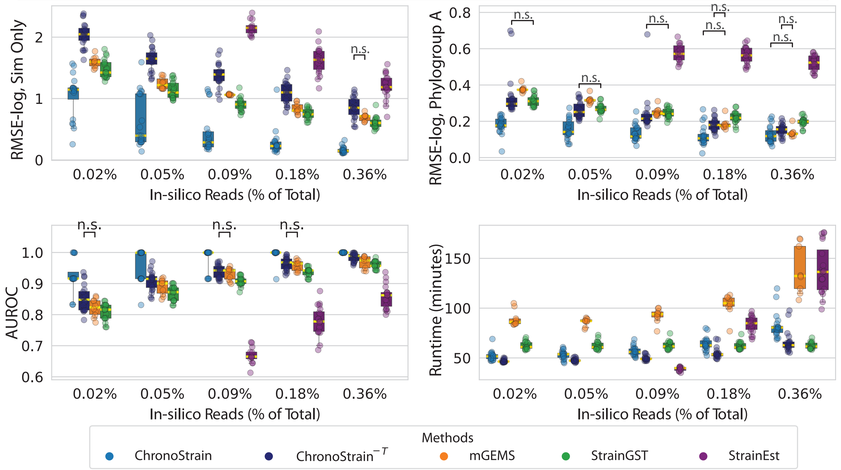
ChronoStrain estimates fully Bayesian time-series strain abundance ratios within a target species, using:

**(1) Arbitrary user-defined genomic markers**
**(e.g. fimbriae, toxins, ABX resistance)**
**(2) Times of sample collection**
**(3) Full read information: nucleotides and quality scores**



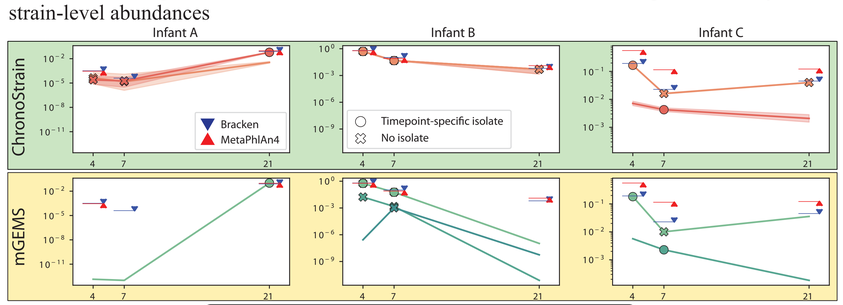

## Benchmarks



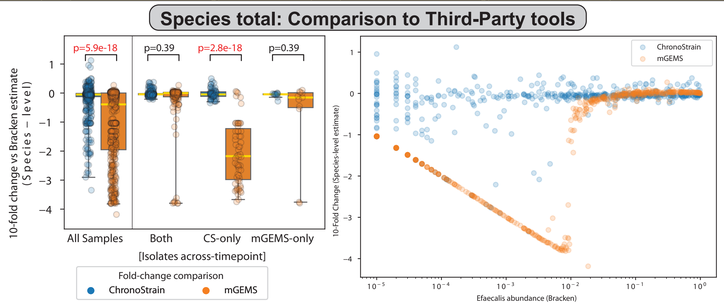




## E. faecalis analysis

We ran analysis on the Baby Biome Study: a cohort of 596 infants, whose stool samples were collected and metagenomically sequenced on days 4, 7, 21 after birth. We analyzed a subset of ~189 infants' samples from which isolates were cultured and assembled. A previous analysis of strain-level abundances

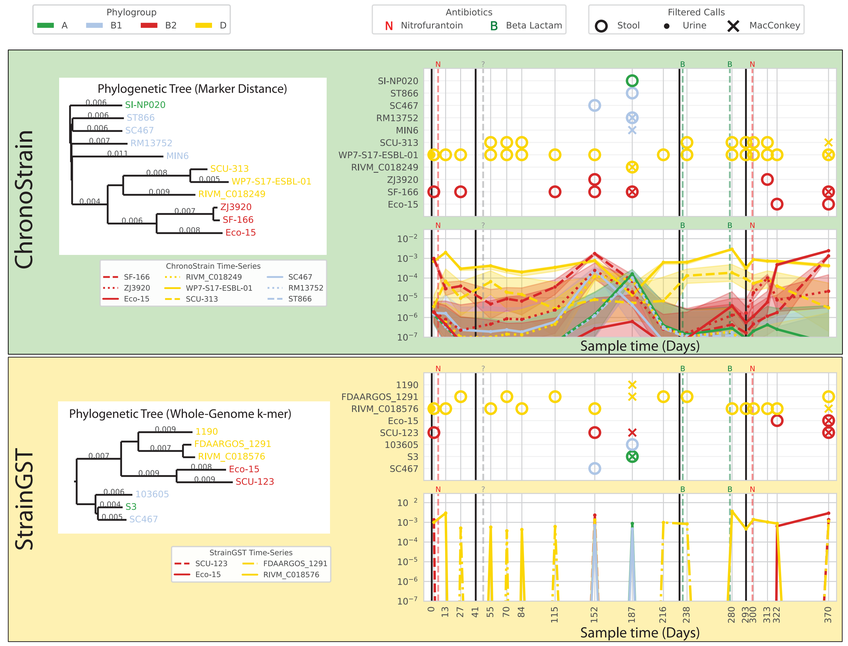



**Species total: Comparison to Third-Party tools**



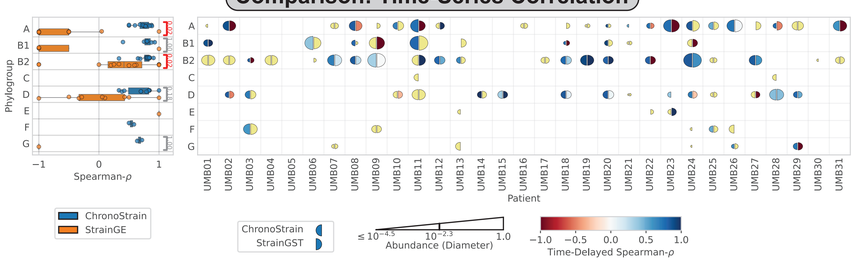

## E. coli analysis

The UMB dataset [5] includes two cohorts: healthy and rUTI-positive. UTI is treated using antibiotics, but recurrence of infection is believed to be due to low-abundance reservoirs of UPEC in the gut. We performed analysis using ChronoStrain, and compared it to a previous analysis result using a different tool (StrainGST), which showed sporadic appearance/disappearance of strains in fecal samples across time. This issue is rectified by analyzing all samples jointly in time-series.




Prior literature reports strain-level stability of microbiomes. ChronoStrain's outputs agree with this: across all of UMB healthy + dysbiotic cohorts, ChronoStrain predicts higher within-phylogroup abundance correlations across time.

**Comparison: Time-Series Correlation**




## Paper & Code

Nature Microbiology Paper:

Software: GitHub