nature microbiology

Article

https://doi.org/10.1038/s41564-025-01983-z

Longitudinal profiling of low-abundance strains in microbiomes with ChronoStrain

Received: 18 January 2023

Accepted: 13 March 2025

Published online: 6 May 2025

Check for updates

Younhun Kim[1,2,3,4], Colin J. Worby[3], Sawal Acharya[2], Lucas R. van Dijk[3,5], Daniel Alfonsetti[6], Zackary Gromko[6], Philippe N. Azimzadeh[7], Karen W. Dodson[7], Georg K. Gerber[2,4,8], Scott J. Hultgren[7], Ashlee M. Earl[3], Bonnie Berger[1,3,6,8] ✉ & Travis E. Gibson[2,3,4,6] ✉

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.

The human microbiome is involved in many aspects of human health and disease and exhibits a great level of diversity within and across host environments[1]. One of the most basic forms of analysis performed on any sample in a microbiome study is determining what bacteria are present and at what abundance. Although some applications call for coarser-grained taxa identification at the operational taxonomic unit (OTU) or species level[2,3], newer studies increasingly focus on more fine-grained resolution at the strain, or even single nucleotide variant (SNV) level[4–7]. Since these studies try to draw conclusions about strain fitness, stability and/or competition, they rely on accurate quantifications of strains in time series.

The process of converting bulk shotgun sequencing reads to taxa abundances ('metagenomic profiling') usually involves some aspect of mapping or aligning reads to reference sequences[8–13]. An alternative is to perform metagenomic assembly[14], although for low-abundance taxa including most gastrointestinal pathogens of interest, this is unlikely to generate scaffolds of sufficient quality to produce reliable strain-level insights. Unfortunately, state-of-the-art methods quantifying strain-level abundances have a multitude of shortcomings when used to track low-abundance taxa, and these shortcomings become more evident when used to study longitudinal samples.

[1]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA. [3]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Harvard Medical School, Boston, MA, USA. [5]Delft Bioinformatics Lab, Delft University of Technology, Delft, the Netherlands. [6]Computer Science and AI Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [7]Department of Molecular Microbiology and Center for Women's Infectious Disease Research, Washington University School of Medicine, St Louis, MO, USA. [8]Harvard-MIT Health Sciences and Technology, Cambridge, MA, USA. ✉e-mail: bab@mit.edu; tegibson@bwh.harvard.edu

https://doi.org/10.1038/s41564-025-01983-z

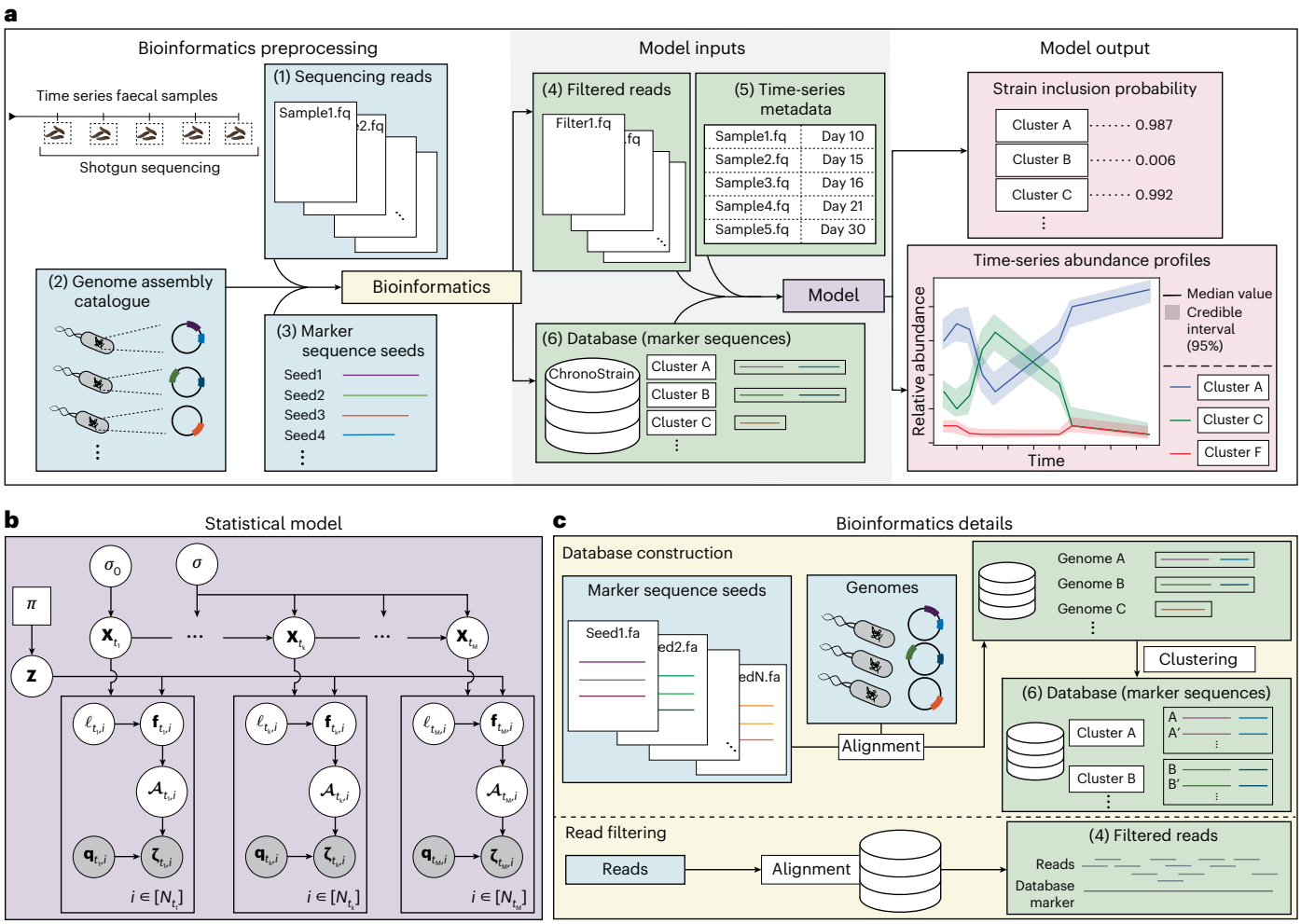


Fig. 1 | Overview of ChronoStrain. a, A high-level schematic of ChronoStrain's analysis pipeline showing the inputs to the bioinformatics preprocessing step (blue), the model inputs (green) and the model outputs (pink). **b**, A graphical representation of the probabilistic model (Methods 'Latent abundance model', 'Sequencing fragment model' and 'Sequencing noise model') used by ChronoStrain. White circles are latent random variables, grey circles are observations, squares are hyperparameters (not all model parameters shown). **c**, A detailed schematic of the bioinformatics preprocessing step illustrating how the marker sequence seeds and the reference genomes are used to construct the strain (cluster) database along with an additional illustration of the initial read filtering process.

Several methods report a statistic that can be directly interpreted as a strain's predicted abundance[11,15–18]. Others report 'pile-up' statistics for SNPs across reference genomes or gene-specific loci[12,15], which require further algorithms to produce strain abundances[9,19]. However, no existing method simultaneously leverages the temporal information in a longitudinal study design while also leveraging the per-base uncertainty in the reads while profiling taxa (quality scores are often only used for preprocessing low-quality reads[20]). Utilizing base-call uncertainty can help overcome ambiguity when mapping or aligning reads.

To address these gaps, we developed ChronoStrain—an uncertainty-aware, timeseries strain abundance estimation algorithm. Our Bayesian algorithm fits all the above specifications. Using raw reads with associated quality score information, sample metadata (host and time of collection) and marker sequence seeds (to construct a strain database), ChronoStrain learns a presence/absence probability and a probabilistic abundance trajectory estimate for each strain being profiled. In this work, we define a 'strain' as a cluster of marker sequences (subsequences from reference genomes) where the threshold used to define these clusters is an arbitrary user-specified variable.

We demonstrate the superior performance of our algorithm on synthetic benchmarking data[21], semi-synthetic data and data from two human studies. The first human study is the rUTI microbiome project (UMB), a year-long longitudinal study of women with a history of recurrent urinary tract infections (rUTI) with a matched healthy cohort[7]. With the UMB study we focus on the increased utility and interpretability one has with ChronoStrain compared with other state-of-the-art methods when tracking strains over time. The second set of samples we apply ChronoStrain to is the Baby Biome Study (BBS)[22], which collected and sequenced between 1 and 6 faecal samples per subject (mean of 2.5) over the first few months of a child's life. With the BBS data, we demonstrate an improved lower limit of detection with ChronoStrain for *E. faecalis* strains, using paired sample isolates to aid in validating our method.

Results

Overview of ChronoStrain

The ChronoStrain pipeline is outlined in Fig. 1. Three components are processed in an initial bioinformatics step (Fig. 1a, blue shaded boxes):

(1) Raw FASTQ files from the experiment
(2) A database of genome assemblies
(3) A database of marker sequence 'seeds'

During initial bioinformatics processing, two tasks are performed. First, components (2) and (3) are used to generate a custom database of marker sequences for each strain that will be profiled (Methods—'ChronoStrain database'). Then, the raw reads from (1) are filtered against this database, resulting in a set of filtered reads (Methods—'Read filtering'). The inputs for the ChronoStrain Bayesian model are then (Fig. 1a, green shaded boxes):

(4) Filtered read files with quality scores (FASTQ format)
(5) A metadata file containing sample timepoint information
(6) A custom database of marker sequences for each strain being profiled

Two of the most useful outputs from the model are a probability for the presence/absence of each strain in the samples as well as a (probabilistic) timeseries abundance profile for each strain in the database (Fig. 1a, pink shaded boxes). We now discuss core components of the pipeline and model in more detail.

Our definition of strains in this work is an operational one: a strain is simply a collection of marker sequences. Instead of manually specifying markers for each genome, the user specifies marker sequence 'seeds'. The seeds need not be genes per se; they can be arbitrary nucleotide sequences. Candidate marker seeds found in this work include MetaPhlAn core marker genes[13], sequence typing genes[23], fimbrial genes and other known virulence factors. Each seed is aligned to the reference database genomes, and each sufficiently similar match is identified as a marker sequence for the corresponding genome in the database.

As a final step, the user gets to decide whether the reference sequences should be clustered and what the threshold for clustering should be, thus picking the granularity for distinguishing distinct strains[24]. In this work, we use different thresholds for strain clustering, ranging from 99.8% sequence similarity to ~100% (every unique marker sequence combination is a different strain). We reiterate that 'strain cluster' and 'strain' are used interchangeably.

Our Bayesian model, for a single time series, is shown in Fig. 1b. Strain abundances are modelled using a stochastic process $\mathbf{X}_{t_k}$ (which is indexed by each strain s as $X_{t_k,s}$) across timepoints $t_k \in \{t_1, \ldots, t_M\}$, together with model inclusion variables $\mathbf{Z}$ (indexed as Z_s). Then, at each timepoint t_k, the ith read is modelled as a nucleotide sequence $\boldsymbol{\zeta}_{t_k,i}$ with its corresponding quality score vector $\mathbf{q}_{t_k,i}$. The sequence $\boldsymbol{\zeta}_{t_k,i}$ is modelled through the variables $\mathbf{f}_{t_k,i}$ (the source nucleotide sequence fragment of the read), $\ell_{t_k,i}$ (the random length for a sliding window along the markers that determines which fragment is measured) and $\mathcal{A}_{t_k,i}$ (the fragment-to-read substitution/indel error profile). The output of our Bayesian inference is not one abundance estimate; it is actually a full probability distribution. These distributions in turn can then be directly interrogated to assess model uncertainty (Supplementary Text B.5). A complete description of the model can be found in Methods—'Latent abundance model', 'Sequencing fragment model' and 'Sequencing noise model'.

ChronoStrain outperforms other methods in benchmarking

We benchmarked ChronoStrain on both synthetic and semi-synthetic data. The synthetic benchmark is based on the CAMI2 'strain-madness' challenge[21]. Our semi-synthetic benchmark combines real reads from a participant in the UMB study[15] with synthetic in silico reads. We first present the semi-synthetic benchmark before closing with the CAMI2-based benchmark.

Our semi-synthetic data generation process is outlined in Fig. 2a with further details provided in Methods 'Semi-synthetic data generation'. Real reads are taken from the first six longitudinal stool samples from UMB participant 18 (UMB18) where only phylogroup B2 and D *Escherichia coli* strains had been detected. The synthetic reads are generated from six phylogroup A strains that are synthetically mutated to be distinct from genomes in the reference database. Then, using a predefined temporal (ground truth) abundance profile, synthetic reads are generated from the six mutant strains and then combined with the real reads. These combined read sets are realistic, while having a well-defined notion of ground truth abundance ratios for evaluation.

For comparison, we included StrainGST[15], StrainEst[25] and the mGEMS pipeline[16]. For a discussion on methods that did not make it into the benchmark, refer to Supplementary Text A. For all semi-synthetic benchmarking, we ran ChronoStrain in two different modes: timeseries-aware and timeseries-agnostic (ChronoStrain^{-T}) where the latter refers to running ChronoStrain on each sample independently.

ChronoStrain significantly outperforms all other methods for all simulated read depths in terms of root mean squared error of log-abundances (RMSE-log) and area under receiver-operator curve (AUROC), except for one scenario, all while maintaining a comparable runtime to the other methods (Fig. 2). As expected, ChronoStrain^{-T} performs worse than ChronoStrain, but is still significantly better or visually on par with the other comparator methods with respect to AUROC (Fig. 2d) and the phylogroup A RMSE-log (Fig. 2c). Even though ChronoStrain^{-T} does not encode sample timepoint information, it still explicitly models presence/absence for each strain with an indicator variable Z_s, which can help control for false positives.

When RMSE-log is only computed over the six target strains (Fig. 2b), the methods are not penalized for false positives and ChronoStrain^{-T} performs significantly worse than mGEMS and StrainGST. The increased performance for the full timeseries run of ChronoStrain comes from its more accurate estimates for low-abundance strains, visible when we bin the RMSE-log contributions according to the synthetic strain's sample abundance (Extended Data Fig. 1). We also performed a formal sensitivity analysis, varying the models' hyperparameters and the ground-truth genomes' mutation rates which can be found in Supplementary Text C.2.

We based our fully synthetic benchmark on the Critical Assessment of Metagenome Interpretation II (CAMI2 (ref. 21)) 'strain-madness' challenge. The original strain-madness challenge generated reads for 100 different abundance profiles from a set of 408 genomes. For our analysis, we focused on species with multiple conspecific strains that had valid multi-locus sequence typing (MLST) schemes (Methods 'CAMI2 strain-madness benchmark'). This inclusion criterion resulted in strain-level profiling of five species: *S. pneumoniae* (174 genomes), *E. coli* (97 genomes), *K. pneumoniae* (47 genomes), *S. aureus* (21 genomes) and *E. faecium* (21 genomes).

Under RMSE-log, ChronoStrain^{-T} significantly outperformed all other methods across all species (Extended Data Fig. 2). For the L1-norm error, which was employed in the original CAMI2 challenge[21], no method had superior performance simultaneously across all five species; a similar pattern for the L1 metric appears in the semisynthetic results (Extended Data Fig. 3). The apparent performance discrepancy between L1 and RMSE-log arises because L1 largely ignores error contributions from the numerous low-abundance genomes. For those genomes, ChronoStrain^{-T} consistently outperformed other methods across all profiled species. Given that real microbial communities' constituent abundances span many orders of magnitude, we recommend not solely relying on L1 when evaluating abundance estimation on complex microbial communities.

Improved interpretability of UMB longitudinal samples

The UMB project monitored 31 women in two cohorts, 'rUTI' (multiple UTIs in past year) and 'healthy' (no recent history of UTI), over the course of a full year[7]. Each participant provided a stool sample once a month, with outgrowth cultures grown from rectal and urine samples taken at the first month for all participants. For those participants who were diagnosed with a UTI, additional urine samples and outgrowth cultures were taken on the days of diagnoses when possible.

https://doi.org/10.1038/s41564-025-01983-z

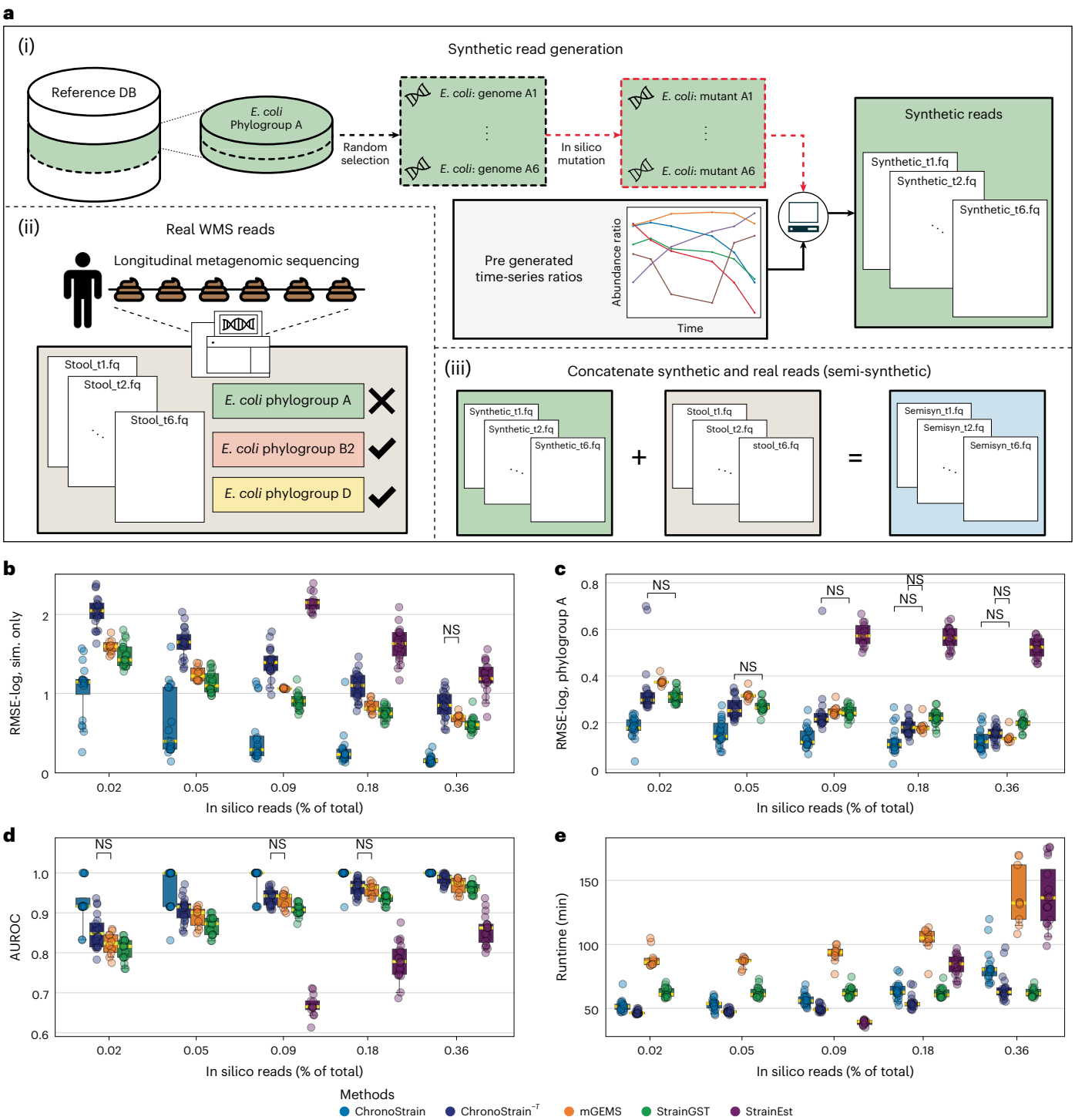


Fig. 2 | ChronoStrain outperforms other state-of-the-art methods with semi-synthetic data. a, Outline of the semi-synthetic data generation process: (i) synthetic reads are generated from 6 mutant phylogroup A *E. coli* genomes; (ii) real reads are then taken from the first six longitudinal samples from participant UMB18, which only had *E. coli* phylogroup B2 and D identified; (iii) finally, real and synthetic reads are concatenated. **b**,**c**, RMSE for model log predictions averaged over the six synthetic strains (**b**) and evaluated over all phylogroup A strains in the database (**c**) to account for each method's estimate including false positives. **d**, AUROC for synthetic strain detection normalized over phylogroup A. **e**, Algorithm runtimes. Each read depth has $n = 20$ replicates. All comparisons to ChronoStrain are statistically significant at level 0.05 after paired two-sided Wilcoxon tests with Benjamini–Hochberg (BH) correction, unless noted with an NS (*P* values in Supplementary Table 1). Medians are coloured yellow, boxes are 25% and 75% quantiles, and whiskers are 2.5% and 97.5% quantiles.

Beyond this, metadata about participants' self-reported dates of last-known antibiotic administration and the dates of infection are available. In addition to the original samples, we have added a new data modality. For a subset of samples from the rUTI cohort for which blooms were identified by the original StrainGST analysis[7], cultures from stool samples plated on MacConkey agar (favouring Gram-negative bacteria including *E. coli*) were sequenced.

We applied ChronoStrain and StrainGST to all 31 time series in the UMB study (Supplementary Figs. G1–G31) with model outputs for UMB participant 18 shown in Fig. 3. Note that the ChronoStrain strain

https://doi.org/10.1038/s41564-025-01983-z

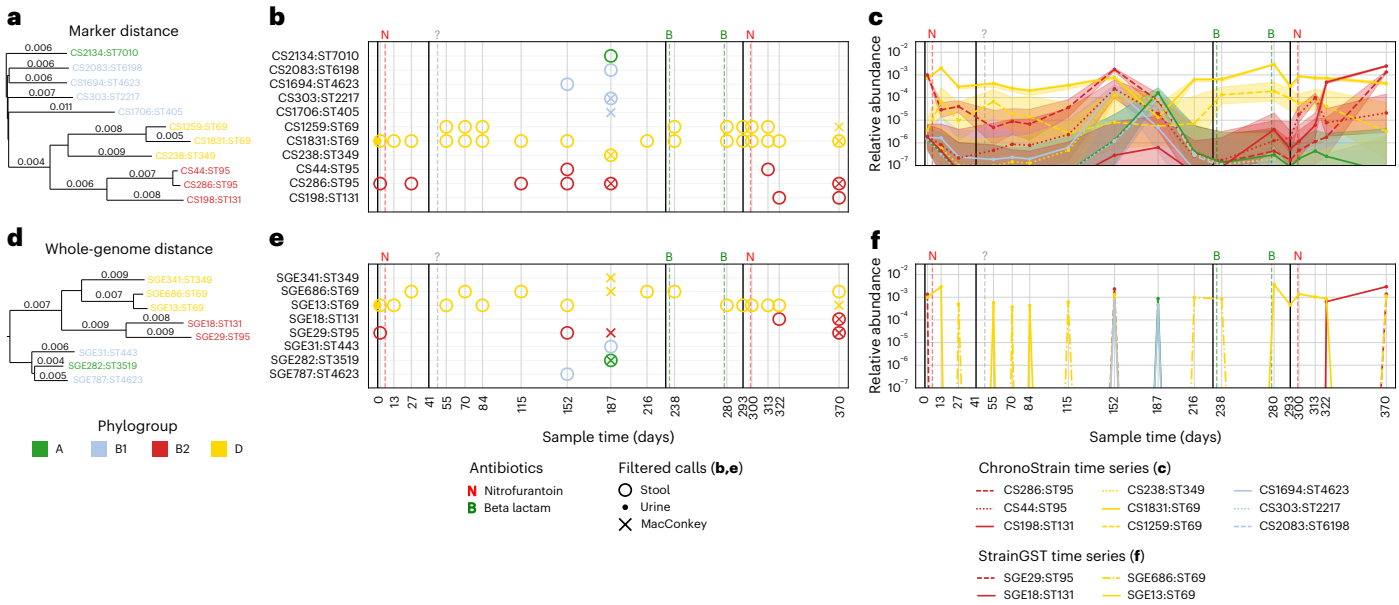


Fig. 3 | Visualization of ChronoStrain's and StrainGST's outputs for UMB18, an rUTI-positive participant. a,**d**, Phylogenetic subtrees of strains computed using two different metrics: marker-specific *k*-mer proportion distance[41] (**a**) and whole-genome *k*-mer Mash distance (**d**). Clusters are labelled with the prefix 'CS' or 'SGE' to denote respective clustering methods. MLST labels (Achtman *E. coli* scheme[27]) are attached as indicated by 'ST' prefix. **b**,**e**, Scatterplot of strain detections across time series for CS (**b**) and SGE (**e**) clustering methods. Different markers indicate sample modality (stool, MacConkey culture from stool, urine). Solid vertical lines indicate dates of UTI diagnoses. Dashed vertical lines marked at the top with a letter (for example, 'N' for nitrofurantoin) indicate self-reported last-known dates of antibiotic administration. **c**,**f**, Plots of estimated 'overall' relative abundances in stool for CS (**c**) and SGE (**f**). Shaded regions with ChronoStrain are centred 95% credible intervals using $n = 5{,}000$ posterior samples; centres (solid lines) are medians.

clusters have a 'CS' prefix, whereas the StrainGST strain clusters have an 'SGE' prefix. For both methods, we have annotated the clusters with their respective MLST labels[26,27] using an 'ST' prefix as well. See Methods 'UMB *E. coli* analysis' for full details on the UMB analysis pipeline.

The output of ChronoStrain (Fig. 3b,c) suggests that the initial infection most probably came from a Phylogroup D/ST69 strain (the filled yellow circle next to CS1831 on day 0). After multiple rounds of antibiotics, CS1831 is no longer detectable in the urine but still persists in the gastrointestinal tract (GIT). Indeed, in stool, two ST69 clusters are called across multiple timepoints (CS1831, solid yellow line; and CS1259, dashed yellow line), where the dominant cluster is the same as the one called in the urine sample and is also the most abundant strain in 13 of 17 timepoints.

Another prominent strain cluster is the phylogroup B2/ST95 cluster CS286 (dashed red line), which shows differing responses to the antibiotics; this cluster is recapitulated in both of the enriched MacConkey-culture samples (Xs in Fig. 3b; culture-specific abundance estimates in Extended Data Fig. 4). The initial dose of nitrofurantoin and the unknown antibiotic reported by the participant before day 55 fail to clear this strain from the GIT; it is present with an abundance above or near 10^{-5}. Around the time of the third and fourth round of antibiotics, which were beta-lactam inhibitors, all the phylogroup B2 strains' abundances drop well below 10^{-6}. However, the fifth round of antibiotics is a redosage of nitrofurantoin near day 300, for which ChronoStrain identifies the rapid growth of both the old CS286 but also a different phylogroup B2 strain CS198 (solid red line) which is a newly dominant strain that was previously undetected. These results suggest that beta-lactam more effectively cleared the B2 taxa in the GIT than nitrofurantoin. This is consistent with previous literature[28,29] suggesting that nitrofurantoin has higher host bioavailability and thus accumulates less in the GIT in comparison to beta-lactam.

Interpreting the output of StrainGST (Fig. 3e,f), one sees that adjacent timepoints call two phylogroup D, ST69 clusters in a mutually exclusive manner. This time series/temporal inconsistency (alternating presence/absence over time) is something that is absent in the joint analysis done by ChronoStrain (Supplementary Text F). Furthermore, the faecal sample analysis suggests that an ST95 strain is present up to but not including day 187, yet the MacConkey-culture sample on that same timepoint suggests otherwise. In ChronoStrain's analysis, the corresponding dominant ST95 cluster was still detected on that particular date, which suggests that our method's joint analysis had the correct detection call. The lack of a coherent cross-sample consensus makes it difficult to evaluate the sensitivity of different strains to antibiotics or to determine the presence of new strains from a bloom. Furthermore, the lack of a credibility (or confidence) interval hampers the interpretability of StrainGST.

In the analysis of UMB18 just presented, we defined distinct strain clusters as those with less than 0.998 nucleotide identity over the database marker sequences. This choice in threshold was so as to coincide with StrainGST's level of nucleotide identity used to define clusters in their original work[7]. To demonstrate the utility and interpretability of our method, we performed the same analysis as above but with a much more fine-grained clustering (Fig. 4). This was thresholded at $1\text{–}10^{-10}$ nucleotide identity, effectively capturing single-nucleotide difference over our markers.

The overall story from before is the same: there is a dominant Clade D strain, a Clade B1 (or C) strain blooms in the middle of the time series, and a previously undetected B2 strain becomes dominant at the end. With this fine-resolution view, however, we do call more strain clusters. However, one can directly see that several of the strain's credible intervals are entirely overlapping, for example, dashed yellow trajectories for ST69 clusters, or dashed blue trajectories for phylogroup B1 clusters. This suggests that the model is having trouble differentiating those strains and probably should be clustered together as in the coarser threshold.

Limit-of-detection improvement in infant samples

The Baby Biome Study collected and sequenced longitudinal faecal samples from 774 full term babies during the neonatal and infancy period, with additional paired samples from a subset of mothers[22].

https://doi.org/10.1038/s41564-025-01983-z

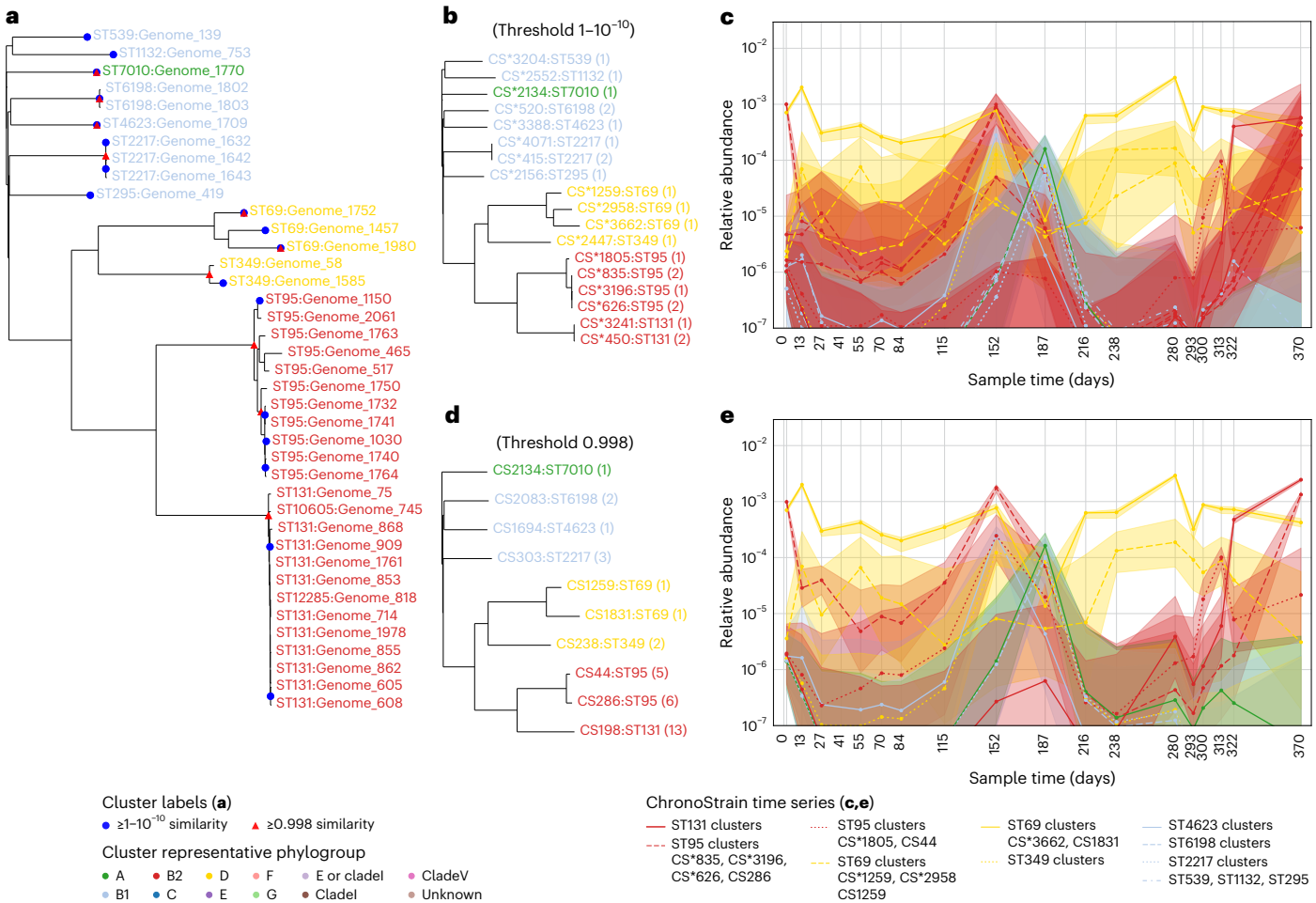


Fig. 4 | ChronoStrain's credibility intervals are a direct visualization of uncertainty. Analysis for UMB18 was run a second time with a much fine-grained database, clustered at 1–10^{-10} similarity threshold. The fine-grained clustering is indicated with the prefix 'CS*'. **a**, A phylogenetic tree showing the clusters called for UMB18 across both granularities. Each leaf node is a genome; internal nodes indicate least common ancestors for each cluster. **b**,**d**, The threshold-specific subtree of clusters called by ChronoStrain in the time series, showing both the fine-grained clustering (**b**) and coarse-grained clustering (**d**). The size of each cluster is between parentheses: for example, 'CS*835:ST95 (2)' denotes that the cluster CS*835 has two genome constituents. **c**,**e**, The threshold-specific timeseries relative abundances for the fine-grained threshold (**c**) and coarse-grained threshold (**e**). Near-perfect overlap of multiple trajectories' shaded regions (centred 95% credible intervals using $n = 5{,}000$ posterior samples) across time suggests that the clustering needs to be coarsened, such as the ST95 dashed red median trajectories in **c**, which merge together in **e**. Centres (solid lines) are medians. Non-perfect overlaps of solid and dashed yellow trajectories suggest the presence of two distinct ST69 strains, one being an order of magnitude more abundant than the other.

Each infant in this study had between 1 and 6 faecal samples collected, with a majority of the neonatal samples taken on days 4, 7 and 21. From the faecal samples of 189 infants, 805 isolates were obtained. Of these, 349 isolates were *E. faecalis* (321 from infants, 28 from mothers). We applied the mGEMS pipeline as well as ChronoStrain to the subset of 189 infants' timeseries faecal samples with a database that incorporated the isolate genomes.

After ensuring that both methods' databases were on equal footing (Methods 'BBS *E. faecalis* analysis'), we performed inference on the infant faecal metagenomic samples. For mGEMS, we used the same hyperparameters as described in ref. 30. To compare the methods at roughly equal sensitivities, we tuned ChronoStrain's post-inference threshold so that the two methods reported the same number of infant *E. faecalis* isolates (Fig. 5).

As intended, the number of strain calls per sample (Fig. 5c) as well as the number of strain clusters corresponding to a cultured isolate from the same timepoint (Fig. 5e) were similar for both methods. However, we did notice a stark difference in the abundance estimates. We illustrate this with example trajectories from three infants A01077, B00053 and B02273 in Fig. 5a,b (complete set in Supplementary Figs. H1–H21). To provide an independent comparison, we used Kraken2+Bracken[31,32] and MetaPhlAn4 (ref. 33) to estimate *E. faecalis* species abundances (triangles in Fig. 5a,b). mGEMS often produces underestimates relative to Bracken across the BBS infant dataset (Fig. 5g, 'All samples'), with the largest discrepancy between Bracken and mGEMS occurring for those samples where ChronoStrain makes a strain call to a paired sample isolate but mGEMS does not ('CS-only'). To better understand this discrepancy, we plotted *E. faecalis* abundance fold-changes relative to Bracken (and MetaPhlAn) for each sample (Extended Data Fig. 5). At ~0.01 relative abundance and below, mGEMS outputs zero (or near-zero) *E. faecalis* abundances, unlike Bracken, MetaPhlAn and ChronoStrain.

Finally, we tested the robustness of the methods when the reference database no longer contains genomes identical to the strains we are trying to track. For this experiment, we mutated 117 of the BBS isolate genomes, chosen from those already called by mGEMS (Supplemental Text E.1), and then performed inference with both methods using the same hyperparameters and thresholds as before. Overall, mGEMS calls decreased from 117/117 to only 45/117 strains, but ChronoStrain's results were largely unchanged from 108/117 to 109/117

https://doi.org/10.1038/s41564-025-01983-z

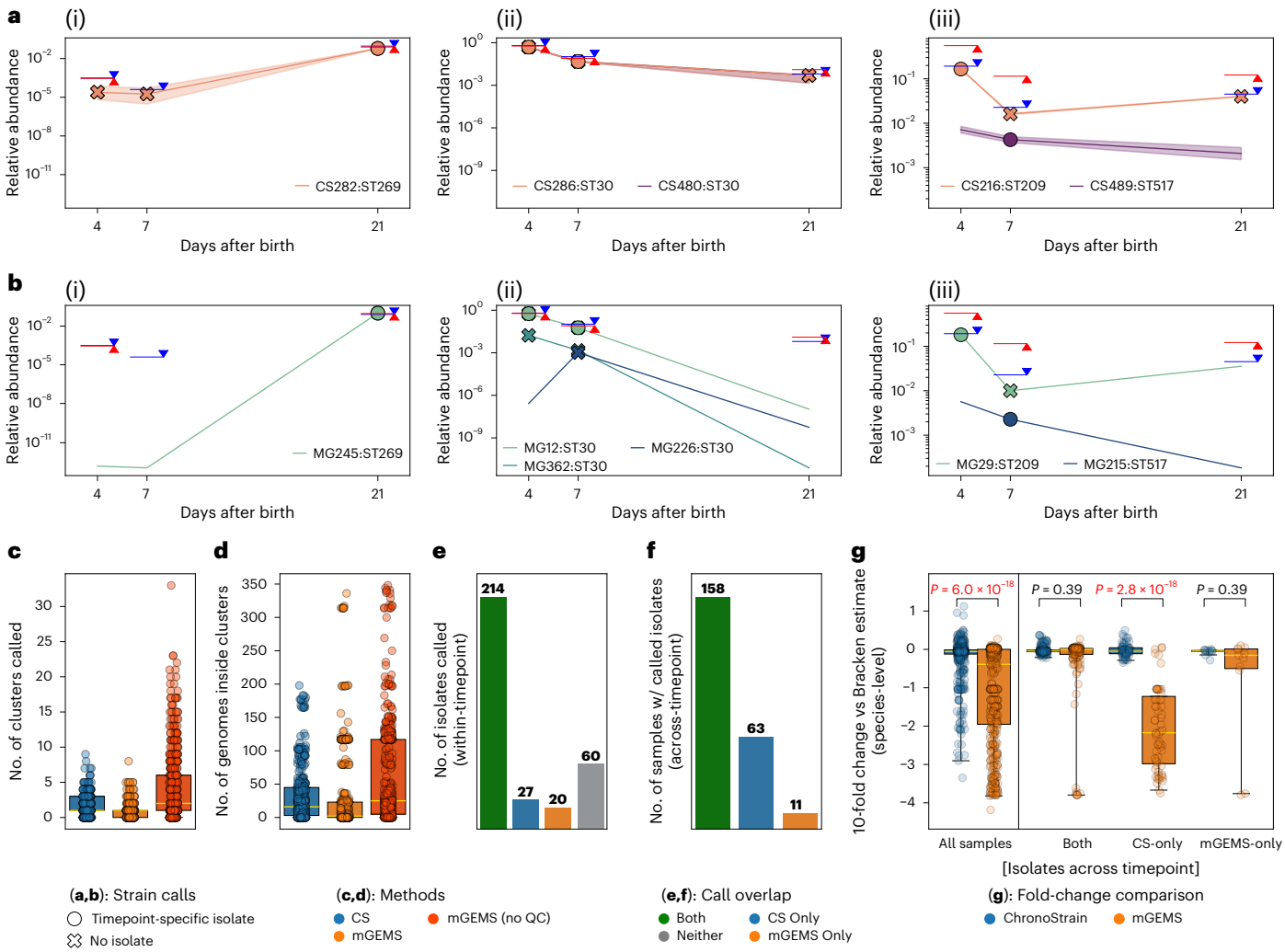


Fig. 5 | ChronoStrain calls isolates from infants across time with more accurate abundance estimation than mGEMS. a, Example ChronoStrain *E. faecalis* strain relative abundance estimates for infants A01077 (i), B00053 (ii) and B02273 (iii). Shaded intervals are the 95% credibility intervals using $n = 5{,}000$ posterior samples. **b**(i–iii), mGEMS estimates for the same infants. For each cluster, we drew its trajectory only if it passed the method's respective filter in at least one sample. Each sample-specific strain call is marked, depending on whether the cluster contains an isolate culture from that donor sample (filled circle/cross). Blue triangles + horizontal lines are Bracken *E. faecalis* species abundance estimates; red triangles + lines are MetaPhlAn4 species estimates. **c**,**d**, Number of clusters passing the filter (**c**) and total genomes within those clusters (**d**) on $n = 486$ samples. **e**, Number of strain calls with an isolate from that same sample (321 total). **f**, Number of samples with a called isolate by either method or both, where the isolate was sourced from a 'different' sample from the same infant, labelled as 'across-timepoint' predictions. **g**, For each sample categorized in **f** (for example, $n = 158$ for 'Both'; $n = 486$ under 'All'), we checked how far the species predictions are from Bracken. Paired two-sided Wilcoxon test *P* values with BH correction are displayed. In **c**, **d** and **g**, medians are coloured yellow, boxes are 25% and 75% quantiles, and whiskers are 2.5% and 97.5% quantiles.

(Fig. 6 and Extended Data Figs. 6 and 7). In Supplementary Text E.2 we discuss these results in more detail.

We emphasize that we intentionally chose isolates with a paired sample that already had an mGEMS call, regardless of what ChronoStrain had identified using the unmutated database, which is why we reported 108/117 calls for the original ChronoStrain run instead of the full 117. The drop in the number of strain calls by mGEMS is by design of the method's demix_check diagnostic (which tries to measure the novelty of genomes within the sample as compared with those in the database): the scores became worse (increased) with mutated isolates in the database (Fig. 6b). One can increase the number of correct strain calls by allowing larger demix_check scores, but this comes at the cost of specificity. With the demix_check score threshold increased from 2 to 4, mGEMS correctly calls 93/117 of the strains, but the median number of calls per sample becomes six times larger than that of ChronoStrain (Extended Data Fig. 6c).

These experiments demonstrate that ChronoStrain has more accurate abundance estimates for lower-abundance strains and that it is more robust to database discrepancies when making strain calls, without losing specificity. mGEMS is sensitive to having the strain of interest being isolated and sequenced for database inclusion, and the pipeline does not reliably estimate abundances below 10^{-2}. These differences can affect how the strain dynamics are interpreted in statistically significant ways. For instance, when looking at strain turnover within *E. faecalis* (pair of adjacent timepoints where the most abundant strain is different), ChronoStrain estimates that twice as many infants had at least one turnover occurrence in the first month compared with mGEMS (40/189 versus 19/189, χ^2 test $P = 0.0046$, Supplementary Table 5).

Discussion

There are three major differences between ChronoStrain and the other strain profiling methods we have highlighted in this work.

https://doi.org/10.1038/s41564-025-01983-z

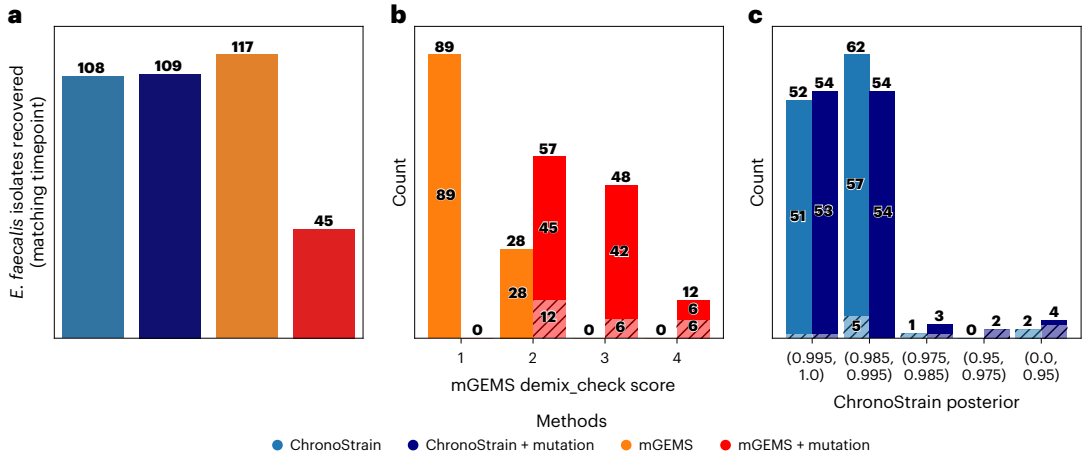


Fig. 6 | ChronoStrain is more robust than mGEMS to mismatches between database genomes and sample reads. We performed inference on the BBS data where 117 of the BBS isolates were mutated (genome mutation rate 0.002) before including them in the database (Supplementary Text E.1). **a**, The raw number of isolate clusters called by each method. Note that mGEMS calls more isolates due to the experimental design: the 117 isolates were initially chosen using mGEMS predictions, even if ChronoStrain did not call them. **b**, The demix_check score distribution for all 117 isolate clusters; '1' is best, '4' is worst. Each bar is divided into two sections: the solid upper region indicates strains with an abundance ratio ≥0.01, and the diagonal-lined lower region indicates strains with abundance ratio <0.01. The mutated genomes caused a precipitous increase in the demix_check scores. **c**, ChronoStrain's posterior probabilities; solid upper region indicates strain calls with an abundance ratio ≥0.065, and diagonal-lined lower region indicates strain calls with abundance ratio <0.065.

First, ChronoStrain explicitly models strain presence/absence through a dedicated parameter, enabling a direct interpretation of the model's confidence in strain detection. Second, ChronoStrain generates probability distributions for strain profiles over time. ChronoStrain's timeseries credible intervals not only aid in temporal interpretation but also help evaluate the appropriateness of the strain cluster resolution. Third, our method is marker based and thus we use a much smaller portion of the reference genomes as part of our pipeline. This approach allows for more intensive computational analysis of marker-aligned reads while maintaining comparable overall processing times to whole-genome *k*-mer-based methods. Ultimately, ChronoStrain was more interpretable than other methods and had superior performance particularly when profiling lower-abundance strains.

ChronoStrain has two primary limitations. The first is our requirement of providing marker seeds, which, in turn, determine the marker sequences. Users should carefully select markers on the basis of their specific applications. However, MetaPhlAn[33] core genes and MLST[27] genes provide a strong starting point, as demonstrated in the synthetic CAMI2 'strain-madness' benchmark, and our semi-synthetic results show viability of non-core genes (for example, variants output by PanPhlAn[34]). The second limitation is ChronoStrain's reliance on reference genomes.

In the future, we plan to address some of these limitations by incorporating de novo gene assembly directly into the model. We also plan to address the use of long reads (for example, Oxford Nanopore, PacBio Hifi) as these technologies become more commonplace in metagenomic studies. Even with state-of-the-art accuracy (for example, 60% of reads being Q30 or better, as recently attained by ref. 35), this still leaves 40% of reads with 20 errors or more (assuming a 20 kb read length). In this work, we annotated *E. coli* using phylogroup and ST numbers. Other types of annotations, such as ST131 sublineages[36], would require inclusion of extra marker genes; we leave the study of these other potential strain groupings as future work.

We have introduced a reference-based strain profiling tool called ChronoStrain. It performs joint Bayesian inference across longitudinal samples and exhibits statistically significant performance improvements over current state of the art. The model outputs can be directly interrogated to assess the model's confidence in strain calls or if the model is having trouble discriminating between different strains in the database. We believe these results will have direct impacts on biological insights particularly when profiling lower-abundance taxa.

Methods

Human participants

The UTI Microbiome Project[7] was conducted with the approval and under the supervision of the Institutional Review Board of Washington University School of Medicine in St Louis, MO. The rUTI study arm consisted of women from the St Louis area with at least three UTIs in the past 12 months. Women with no history of UTI, or at most one UTI ever, were recruited into the control arm via the Department of Urological Surgery at Barnes-Jewish Hospital in St Louis. A total of 16 control and 15 rUTI women aged between 18 and 45 were recruited to the study. All participants provided informed consent.

None of the authors in this work were involved in the original Baby Biome Study[22]. That study was approved by the NHS London - City and East Research Ethics Committee. Participants were recruited through the University Hospital system: Barking, Havering and Redbridge University Hospitals NHS Trust (BHR), the University Hospitals Leicester NHS Trust (LEI) and the University College London Hospitals NHS Foundation Trust (UCLH). All mothers provided informed consent for themselves and their children to participate in the study. Enrolment consisted of 774 babies total; 178 of these babies were paired with 175 mothers who provided maternal faecal samples.

Overview of ChronoStrain

To specify our Bayesian model, one provides a database $\mathcal{S}$ of strains and their marker sequences, a list of timepoints $\mathcal{T}$ and each timepoint t's collection of N_t corresponding reads. For each timepoint $t \in \mathcal{T}$ and $i \in 1, \ldots, N_t$, each observed read $r_{t,i} = (\boldsymbol{\zeta}_{t,i}, \mathbf{q}_{t,i})$ is specified as a (nucleotide sequence, phred quality score vector) pair.

The full Bayesian model, outlined below, describes the joint distribution of $(\mathbf{X}, \mathbf{Z}, R)$: $\mathbf{X} = (\mathbf{X}_t)_{t\in\mathcal{T}}$ is a latent representation of the unobserved abundance profile at time t, $\mathbf{Z}$ is a vector that decides which strains are included in the population, and $R = (R_t)_{t\in\mathcal{T}}$ each is the subcollection of reads (out of N_t original ones) that align to our database. We perform variational inference to approximate the posterior distribution $p(\mathbf{X}, \mathbf{Z} \mid R)$. We note that parts of the model implemented in this paper are specifically tailored for short Illumina reads;

the next section points out exactly where this assumption is encoded. This leaves room for a model variant that accommodates long reads for future work.

Latent abundance model

First, we model the unobserved abundances using latent representations $\mathbf{X}_{t_1}, \ldots, \mathbf{X}_{t_{|\mathcal{T}|}} \in \mathbb{R}^S$ and a single vector $\mathbf{Z} \in \{0,1\}^S$. The $\mathbf{X}_t$'s are discretized observations of a Weiner process:

$$\begin{aligned} \mathbf{X}_{t_1} &\sim \mathcal{N}\left(\mathbf{0}, \sigma_0^2 I\right) \\ \mathbf{X}_{t_k} | \mathbf{X}_{t_{k-1}} &\sim \mathcal{N}\left(\mathbf{X}_{t_{k-1}}, \sigma^2 (t_k - t_{k-1}) I\right). \end{aligned} \tag{1}$$

The scalar variance parameters σ_0^2, σ^2 are each assigned independent instances of Jeffrey's improper prior[37] for the variance of a Gaussian with known mean:

$$p(\sigma_0) \propto \frac{1}{\sigma_0} \quad \text{and} \quad p(\sigma) \propto \frac{1}{\sigma} \tag{2}$$

This prior is chosen to fulfil the role of a 'non-informative' prior and for its transformational invariance property. For users of our method, this invariance means that the choice of measurement units of time—whether it is minutes, hours or days—does not matter. During inference, the variables σ, σ_0 are integrated out of the posterior.

The elements of vector $\mathbf{Z} = (Z_s)_{s \in \mathcal{S}}$ are independent BERNOULLI(π); each Z_s indicates whether the strain cluster is present ($Z_s = 1$) or absent ($Z_s = 0$). To transform the real-valued vectors $\mathbf{X}_t$ into relative abundances, we take $Y_{t,s} = \frac{Z_s e^{X_{t,s}}}{\sum_{\hat{s}} Z_{\hat{s}} e^{X_{t,\hat{s}}}}$, so that $\mathbf{Y}_t \in \Delta^{S-1}$, the S-component probability simplex. This transformation is often called 'masked softmax' in machine learning; the softmax transformation has similarly been used in continuous-time dynamic topic models[38,39].

Sequencing fragment model

Conditioned on $\mathbf{Y}_t$, we model the read $r_{t,i}$ (timepoint t, index i) independently from all other reads as described below. We model the reads in two steps: the random position of the reads' source fragments and then the sequencing noise. In our model, a 'fragment' is a substring of a marker sequence, representing the nucleotides which later get measured into reads. Each read $r = (\boldsymbol{\zeta}, \mathbf{q})$ without a mate pair is modelled as being a noisy measurement of a single randomly chosen nucleotide sequence fragment $\mathbf{f}$; the mate–pair model is discussed in Supplementary Text B.1.3. The primary assumption is that each read's source fragment overlaps with 'some' marker in the database, necessitating a filtering step as described in 'Read filtering'.

First, we introduce a few definitions. For a nucleotide sequence $\mathbf{x}$, let $|\mathbf{x}|$ denote its length. Allowing each marker sequence of strain genome s to be padded with $\beta|\boldsymbol{\zeta}|$ of 'empty' nucleotides on both ends, let $\mathcal{W}_s^{(\ell)}$ be the collection of all length-ℓ sliding windows of markers of s. We say that $\mathbf{w} \in \mathcal{W}_s^{(\ell)}$ 'induces' fragment $\mathbf{f}$ if $\mathbf{f}$ is the string obtained from $\mathbf{w}$ by removing all padded bases; in particular, $\mathbf{f}$ is always at most as long as $\mathbf{w}$ ($|\mathbf{f}| < |\mathbf{w}|$). For instance, $\beta = 0.5$ guarantees that we only consider fragments $\mathbf{f}$ induced by ≥50% of the (short) read. Let $\Sigma_s^{(\ell)}$ denote the set of all f induced by each $\mathbf{w} \in \mathcal{W}_s^{(\ell)}$. Finally, let $n_{\mathbf{f},s}^{(\ell)} = |\{\mathbf{w} \in \mathcal{W}_s^{(\ell)} : \mathbf{w} \text{ induces } \mathbf{f}\}|$ be the number of times $\mathbf{f}$ is induced, and let $n_s^{(\ell)} = \sum_{\mathbf{f} \in \Sigma_s^\ell} n_{\mathbf{f},s}^{(\ell)}$ be its sum across all $\mathbf{f}$.

Using the above definitions, we describe the fragment model. For each read $r_{t,i}$, let $\ell_{t,i}$ be NEGATIVE_BINOMIAL($R_{\text{NB}} > 0$, $P_{\text{NB}} \in [0, 1]$)-distributed. We model $\mathbf{f}_{t,i}$ as being sampled proportional to the frequency at which it is represented in the population at time t. More precisely (dropping the subscripts t, i to make it easier to read):

$$p(\mathbf{f} | \mathbf{Y}_t, \ell) \propto \sum_{s \in \mathcal{S}} Y_{t,s} n_{\mathbf{f},s}^{(\ell)} \tag{3}$$

This proportionality represents a normalization across all fragments $\mathbf{f}$, and the normalization denominator $\sum_f \sum_{\hat{s}} Y_{t,\hat{s}} n_{\mathbf{f},\hat{s}}^{(\ell)} = \sum_{\hat{s} \in \mathcal{S}} Y_{t,\hat{s}} n_{\hat{s}}^{(\ell)}$ is a function of $\mathbf{Y}_t$, the quantity we are trying to estimate. Algorithmically, a certain approximation of this (Supplementary Text B.1.2) results in an efficient correction for strains whose markers are overrepresented in the database.

We remark that this is precisely the part of our method specially tailored for short reads. When operating on long reads (typically ~10–25 kb or longer), our approximation fails to hold and thus our algorithm must be adapted to a different strategy. Furthermore, long reads can span multiple markers, thus requiring an adjustment in the fragment count definition.

Sequencing noise model

Conditioned on $\mathbf{f}_{t,i}$, we describe the per-base sequencing error model for read $r_{t,i} = (\boldsymbol{\zeta}_{t,i}, \mathbf{q}_{t,i})$ using the mathematical language of sequence alignments. For any alignment $\mathcal{A}_{t,i}$, meaning an arbitrary alignment of $\boldsymbol{\zeta}_{t,i}$ to $\mathbf{f}_{t,i}$ represented by a $2 \times K$ array (for some K) of three symbols: 'Match', 'Mismatch' and 'Gap' (representing either an insertion or deletion), we model

$$\mathcal{A}_{t,i} | \mathbf{f}_{t,i}; \mathbf{q}_{t,i} \sim \text{PHRED\_WITH\_INDELS}(\mathbf{f}_{t,i}, \mathbf{q}_{t,i}). \tag{4}$$

We drop the subscripts for exposition. The PHRED_WITH_INDELS($\mathbf{f}, \mathbf{q}$) distribution is supported over feasible alignments in the theoretical search space of the Needleman–Wunsch dynamic programming algorithm[40]. Its likelihood function is given by a formula assuming fixed indel error rates ϵ_{ins} and ϵ_{del} and the standard phred score model:

$$\begin{aligned} p(\mathcal{A} | \mathbf{f}; \mathbf{q}) = (\epsilon_{\text{del}})^{(\#\text{ deletions})} (\epsilon_{\text{ins}})^{(\#\text{ insertions})} \\ \prod_{j \in \text{Matches}(\mathcal{A})} (1 - 10^{-q_j/10}) \prod_{j \in \text{ Mismatches}(\mathcal{A})} (10^{-q_j/10}) \end{aligned} \tag{5}$$

for any feasible alignment $\mathcal{A}$. The parameters $\epsilon_{\text{ins}}, \epsilon_{\text{del}}$ are specific to the sequencing machine and may depend on whether the reads are forward or reverse in the mate pair.

Finally, conditional on $\mathcal{A}$ and treating quality scores $\mathbf{q}$ as a fixed observation, each mismatched/inserted base of $\boldsymbol{\zeta}$ is sampled uniformly at random; the likelihood of the read's nucleotides $\boldsymbol{\zeta}$ is the product

$$p(\boldsymbol{\zeta}_{t,i} | \mathbf{f}_{t,i}, \mathcal{A}_{t,i}) = \left(\frac{1}{4}\right)^{\#\text{ insertions}} \left(\frac{1}{3}\right)^{\#\text{ mismatches}} \prod_{j \in \text{Matches}} \mathbb{1}\{\boldsymbol{\zeta} \text{ and } \mathbf{f} \text{ matchat } j\}. \tag{6}$$

In its entirety, this model has hyperparameters $\pi, R_{\text{NB}}, P_{\text{NB}}, \beta, \epsilon_{\text{ins}}, \epsilon_{\text{del}}$. Our choices are explained in Supplementary Text B.2.

ChronoStrain database

In ChronoStrain's model, a strain $s \in \mathcal{S}$ is specified by a (multi-)set of markers $\mathcal{M}_s$, where a 'marker' $m \in \mathcal{M}_s$ is simply a nucleotide sequence specific, but not necessarily unique, to that strain. Such a sequence could be, for example, a variant of a known gene encoding a particular target function of interest. The fact that we are using sets implies that we pay no attention to markers' ordering on the chromosome, but we do care about genes' copy numbers, potential homologies and their exact nucleotide sequence.

To construct our database, we require a FASTA file containing a reference sequence (or a multifasta file containing multiple known variants) for each gene; these sequences are called 'seeds'. To allow usage of non-core genes as seeds, the pipeline addresses potential complications: genes may have homologues within other species, vary in copy number, are possibly mis-annotated, or possibly missed by in silico PCR primer searches (for example, *E. coli* O-antigen gene cluster).

The construction pipeline begins by downloading all available genome assemblies (excluding metagenome-assembled genomes of

potential low quality) from the same family as our target species (for example, Enterobacteriaceae when analysing *E. coli*). For each seed, we run a local BLAST query; the hits are thresholded by per cent alignment identity (by default, 75%) and minimum length 150 (a typical read length), with --max_target_seqs = 10 × #(database genomes) to report a generous number of hits per marker seed query.

Next, to address redundancy in the marker seeds, we merge BLAST hits that overlap or are contiguous. For instance, if positions (35,000–42,000) and (41,000–50,000) are BLAST hits for genome g on the same contig/chromosome, then we merge them into a single hit spanning positions (35,000–50,000) forming a single marker on g.

Lastly, to address redundancy in the collection of strain genomes, we clustered them. We used the tool dashing2 (ref. 41) on the multifasta file of markers for each genome, which computes approximations of multiplicity-aware k-mer (multi)-set distances. Using the pairwise distance matrix output by this tool, we run scikit-learn's implementation of agglomerative clustering with 'complete' linkage; this is parametrized by distance threshold. Each cluster $\mathcal{C}$'s representative strain $s_{\text{rep}}(\mathcal{C})$ was chosen as the strain whose distances most closely resemble those of the whole cluster:

$$s_{\text{rep}}(\mathcal{C}) = \underset{s\in\mathcal{C}}{\operatorname{argmin}} \sum_{\mathcal{C}'} |d(\mathcal{C},\mathcal{C}') - d(\{s\},\mathcal{C}')| \tag{7}$$

where $d(\mathcal{C},\mathcal{C}') = \frac{1}{|\mathcal{C}||\mathcal{C}'|}\sum_{x\in\mathcal{C},y\in\mathcal{C}'} \texttt{dashing2}(x,y)$ is the average distance between two clusters. The above sum is over all clusters, including $\mathcal{C}$ itself.

Read filtering

We use bwa-mem2 to quickly align reads to the marker database. The match/mismatch penalties are assigned the $\log_2$-odds ratio of errors from the PHREDWITHINDELS model (assuming a pessimistic quality score of 20) in relation to a uniformly random sequence of nucleotides:

$$\begin{aligned}\text{Match bonus} &= 2 \approx \log_2\left(\tfrac{1-10^{-2}}{1/4}\right)\\ \text{Mismatch penalty} &= -5 \approx \log_2\left(\tfrac{(3/4)\times 10^{-2}}{1/4}\right)\end{aligned} \tag{8}$$

Assuming that indels are randomly distributed across each read according to indel error rates $\epsilon_{\text{ins}}, \epsilon_{\text{del}}$ (Supplementary Text B.2), we set the gap open penalty to zero and the extend penalties to $-\log_2(\epsilon_{\text{ins}}), -\log_2(\epsilon_{\text{del}})$. On the basis of these alignments, we only keep reads that aligned to some marker sequence, where the alignment maps the read with at least 97.5% nucleotide identity. The % identity here is computed after re-attaching soft-clipped bases as mismatches to the local alignments output by bwa-mem2.

Target posterior approximation

On the basis of the Bayesian generative model, we aim to estimate the posterior distribution $p(\mathbf{X},\mathbf{Z}|R)$. We employ ADVI[42], which uses stochastic optimization on Monte Carlo estimates of the evidence lower bound (ELBO) objective. Using standard VI notation q to denote a generic approximate distribution: we use the factorized family $q(\mathbf{X})q(\mathbf{Z})$ of densities for our variational fit. $q(\mathbf{X})$ is by default given a full covariance matrix, but for longer time series (for example, UMB stool samples), we apply a further mean-field factorization $q(\mathbf{X}) = q(\mathbf{X}_{t_1})\cdots q(\mathbf{X}_{t_{|\mathcal{T}|}})$.

The main difficulty for inference is in making the data likelihood function $p(R|\mathbf{X},\mathbf{Z})$ efficiently computable. We employ a heuristic sparsification, mathematically derived in Supplementary Text B.1. The objective function is implemented and optimized using the JAX library[43] and the Adam gradient descent algorithm. The posterior approximation of $\mathbf{X}$ is initialized to have mean zero (corresponding to a uniform abundance profile for all timepoints) and covariance equal to the identity matrix. The posterior approximation of $\mathbf{Z}$ is a Gumbel-Softmax relaxation[44], where the temperature τ is initialized to 10.0 and is annealed by a factor of 0.95 every epoch, down to a minimum of 10^{-4}. In this work, we calculated all statistics using n = 5,000 samples from this estimated posterior.

Detection classifier

For all real-data analyses, we applied the following method to interpret the approximated posterior distribution $p(\mathbf{X},\mathbf{Z}|R)$. First, we computed the collection of strain clusters $\overline{s}$ where each $s\in\overline{s}$ satisfies $p(Z_s|R) > \overline{\pi}$. We chose $\overline{\pi} = 0.95$, equivalent to a Bayes Factor threshold of 10^5 when the prior is $\pi = 10^{-3}$. Then, we sampled from the conditional posterior $p(\mathbf{X}|\mathbf{Z} = \mathbb{1}_{\overline{s}})$, meaning that we conditioned on only $\overline{s}$ appearing in the model; the partial mean-field factorization of the variational solution makes this sampling trivial. For each timepoint t_k in the time series, a strain s is marked as 'detected' (as in Figs. 3b and 5) if the resulting database-normalized relative abundance estimate median($Y_{t,s}$) exceeds a cut-off ρ = 5%.

Analysis details

Here we describe the methodology used to analyse each dataset (UMB, BBS and semi-synthetic), including marker seeds for the two databases used in this work and the settings that were used for each method. The precise database construction workflow for ChronoStrain is implemented as a Jupyter notebook for each dataset, available to view in our codebase. Analysis on real data for all methods (including the background reads from UMB18 for semi-synthetic) were run on trimmed and decontaminated data (see 'Sequencing and real-data processing').

UMB *E. coli* analysis

For *E. coli* strain abundance estimation found in this work, our database seeds were:

(1) Genes from all *E. coli* MLST schemes on PubMLST[26]
(2) Genes used by the ClermonTyping tool[45] for phylotyping
(3) The O-antigen gene cluster, flanked by the JumpSTART and gnd primers[46]
(4) H-antigen encoding (flagellar) genes annotated with names fliC, flk*, fll*, flm*
(5) Annotated fimbrial genes fim*, and
(6) Annotated Shigatoxin genes stx*

E. coli currently has two ST schemes on PubMLST; we simply included all of the genes from both. Our catalogue of reference genomes consisted of 5,405 whole-chromosome assemblies from the Enterobacteriaceae family, of which 2,063 were *E. coli*. After the BLAST and redundancy and overlap correction, our markers made up ~1.5% of the genome when averaged across *E. coli* entries. For both UMB and semi-synthetic analyses, we chose a 99.8% weighted marker k-mer frequency similarity (dashing2 with ProbMinHash sketching) as a cut-off for the agglomerative clustering heuristic. After this process, we ended up with a database of 2,325 Enterobacteriaceae strain cluster representatives and their marker sequences; 842 are *E. coli*.

The systematic BLAST search and overlap correction steps were critical. For steps 2, 4, 5 and 6, we relied on genbank annotations; we found that several genes suffered from mis-annotations and/or unconventional naming schemes (for example, stx1A versus stxA1), and thus the overlap correction helped correct for redundancies. Furthermore, the primers for the O-antigen gene cluster are known to have mutations in different subclades of *E. coli*[47], so the systematic BLAST search helped identify genes missed by the in silico primer matching step (we used EMBOSS primersearch[48] which helps, but does not guarantee, finding all hits).

We ran ChronoStrain using this database with default inference settings, with prior π = 0.001, and interpreted using posterior threshold $\overline{\pi}$ = 0.95. Per-timepoint calls (Fig. 3b,d) were made using an abundance cut-off ('Detection classifier'). We ran StrainGST with default settings

(*k*-mer length 23, 5 iterations and score threshold of 0.02) and using a database of *Escherichia* and *Shigella* genomes. We did not run the next tool in the StrainGE pipeline (StrainGR) which characterizes novel SNVs from the reads, since our goal was only to compare abundances.

BBS *E. faecalis* analysis

For *E. faecalis* strain abundance estimation, our database seeds were:

(1) Genes from all *E. faecalis* MLST schemes on PubMLST[26]
(2) PCR primer-specified pathogenicity/virulence-marking/polymorphic genes from ref. 49
(3) A subset of 39 genes from the infants' *E. faecalis* isolates

We first performed database construction and clustering at $1–10^{-8}$ similarity using just database seed sets 1 and 2 listed above (resulting in ~450 *E. faecalis* clusters out of ~660 total). Using these genes, many of the BBS isolates across 'different' infants were co-clustered even at this extreme of a threshold.

To finely separate these isolates, we constructed seed set #3 using the following heuristic. We annotated these infant isolate genomes using pgap[50], and for each gene name g, we used MAFFT[51] to perform multiple alignment. Let $\mathcal{C}[i]$ denote the set of isolates from infant i contained within $\mathcal{C}$ (a cluster produced using just seed sets 1 and 2), and for any isolate x, let $g[x]$ denote the aligned sequence of the gene g in x (if x has multiple copies of g, we picked the last one in the GFF annotation; if x has zero copies of g, we took a sequence of gap characters). We formed a distance metric $d_g^{(\mathcal{C})}(i,j)$:

$$d_g^{(\mathcal{C})}(i,j) = \min_{x\in\mathcal{C}[i],y\in\mathcal{C}[j]} \texttt{Hamming}(g[x],g[y]) \quad (9)$$

which quantifies how well g distinguishes the isolates of infants i and j in $\mathcal{C}$. Finally, we picked the top $k = 3$ genes $g_1, \ldots, g_3$ maximizing the number of non-zero entries of the distance matrix $d_{g_i}^{(\mathcal{C})}$; repeating this for each $\mathcal{C}$ gave us 39 genes for marker group 3. We made no attempt to 'optimize' k in this work; $k = 3$ gave a reasonable database size that fit within the memory constraints of our machines.

Our reference genomes consisted of 1,087 complete chromosomal assemblies from the Enterococcaceae family (excluding *E. faecalis*) to account for sequence similarity-induced confounders, plus the 2,026 + 350 *E. faecalis* isolates from a separate European study and BBS as in the original mGEMS analysis[30]. Averaged across *E. faecalis* entries, our markers made up ~1.6% of the genome. We used a 99.8% marker similarity cut-off, which resulted in 533 Enterococcaceae strain cluster representatives, of which 387 are *E. faecalis*. Of these, 83 contained at least one BBS isolate. ChronoStrain was run on this database with prior $\pi = 0.001$. Results were interpreted using posterior threshold $\overline{\pi} = 0.95$; per-timepoint calls (Fig. 5a–f) were made using an abundance cut-off ('Detection classifier').

To compare to the previous study, we reran the mGEMS pipeline (Themisto+mSWEEP+mGEMS bin extraction) in a hierarchical style as in ref. 30. This means that we first ran the pipeline to bin reads by species (using a published index of ~640,000 genomes[52] compatible with Themisto v.3.2.1), then an analysis on the *E. faecalis* bin, and finally demix_check to check the quality of the strain bins. Finally, we only kept strain bins with abundance greater than 0.01 and removed bins with poor confidence scores (3 or 4). For mGEMS, we compiled the index for *E. faecalis* quantification using all 2,376 *E. faecalis* isolates. We ran PopPUNK using the threshold method with threshold 0.00036; this was tuned manually, so that PopPUNK produced exactly 83 clusters containing at least one infant isolate, roughly matching ChronoStrain's granularity (in the original *E. faecalis* index from ref. 30, clustered using PopPUNK's dbscan implementation, BBS isolates were concentrated within only 37 out of 168 clusters, meaning that the original clustering was quite coarse). Additional species-level quantification was done twice, once using Kraken2+Bracken and once using MetaPhlAn4 v.4.0.6.

Semi-synthetic data generation

Before generating reads, we first created databases for *E. coli* abundance estimation. To do so, we used the same set of reference genomes from the UMB analysis and reused the *E. coli* marker set from the UMB analysis for ChronoStrain. Each method performs clustering. We tuned the clustering parameters (Supplementary Text C.1) so that the clustering methods resulted in the same number of clusters ±1 relative to the 99.8%-threshold clustering output by ChronoStrain. This was done to ensure that all methods were making predictions at the same clustering granularity.

The synthetic portion of our dataset was made up of 10 'genome' replicates (sampling 6 genomes and random mutations) times 2 'sequencing' replicates (read simulation was rerun with a different seed) across 5 possible simulated read depths, totalling 100 distinct overall replicates. For each 'genome' replicate seed, we selected 6 random phylogroup A *E. coli* genomes from the collection of reference genomes used to build the UMB analysis database. To ensure that each cluster (regardless of clustering method) was fairly represented across the replicates, we assigned to each genome a probability weight proportional to the reciprocal of the root mean-square of the respective databases' cluster sizes. Then, we picked 6 random genomes one at a time without replacement using these weights. After a genome was chosen, we removed all other genomes that shared a cluster with it, for all clustering schemes, so that no cluster was represented twice. Finally, to each chosen reference genome, we introduced independent random mutations by flipping a coin of bias p(Heads) = 0.002 for each base. If Heads, we chose one of the three remaining nucleotides at random to substitute.

For each choice of six mutated phylogroup A genomes, we simulated reads using ART[53] and its built-in HiSeq (length 150) error profiles. The reads were sampled according to the counts given by a MULTINOMIAL($N, \mathbf{y}_t$) distribution, where N is the chosen simulation read count (N = 2,500, 5,000, 10,000, 20,000, 40,000) and each $\mathbf{y}_t$ is the vector of ground-truth abundance ratios summing to 1 for the 6 synthetic genomes for timepoint t. The timeseries ratios $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_6)$ were hand picked beforehand and fixed for all replicates; the choice of the genomes above decides which genome gets assigned to which fixed trajectory out of the 6. These trajectories were chosen to span a wide range from 0.1 to 10^{-3}, where some strains fluctuated between different orders of magnitude, so as to provide a challenging dataset where strains were hard to detect in some timepoints and easier in others.

The simulated reads spanned six timepoints and were merged with the first six timepoints from UMB18's stool sample sequences. This choice was made on the basis of a preliminary analysis using the benchmarked methods, which suggested that phylogroup A was either absent or undetectable in these samples. Note that the lowest simulated read count (N = 2,500) approximately amounted to ~10^{-7} overall relative abundance of phylogroup A after accounting for the ~10 million real reads.

Semi-synthetic inference and analysis

We ran StrainGST to report up to 20 strains instead of the default of 5, so as to infer strains beyond what it would have returned on the background samples alone. For StrainEst, we raised its sensitivity slightly (-p 5 -a 3), incurring some runtime cost but returning non-trivial outputs; default settings caused the programme to recall no ground-truth clusters. mGEMS was run hierarchically as described in ref. 30: the first analysis used the same ~640,000 genome index, and then the *E. coli* bin analysis was run using the database described above. We did not factor demix_check scores into the analysis because all scores were 4 (the worst possible score) for all *E. coli* bins. ChronoStrain was run using default hyperparameters. After inference, ChronoStrain was interpreted using $\overline{\pi} = 0.95$.

For measuring the error in abundance estimation, we computed the RMSE-log after renormalizing across a subset of clusters (the six

source clusters, or all the clusters labelled as phylogroup A). For an estimate $\hat{\mathbf{y}}$, the RMSE-log metric is given by the formula

$$\text{RMSE-log}(\mathbf{y},\hat{\mathbf{y}}) = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\log(y_{t,i}+\epsilon)-\log(\hat{y}_{t,i}+\epsilon)\right)^2} \quad (10)$$

where $\epsilon = 10^{-4}$, an order of magnitude smaller than the smallest (relative) simulated ratio in y. To evaluate the classification metric (AUROC), we turned each evaluated method into a classifier by applying a threshold on their abundance outputs to determine 'detection'.

Before evaluating any error for mGEMS, we thresholded the raw abundance estimate[54] to mitigate the harmful effect of noise on RMSE-log. To do this, we first restricted the abundances of the raw mSWEEP output to the target subset (either phylogroup A or the six ground-truth clusters) and renormalized. Then, we zeroed out all entries below a limit-of-detection (LOD) threshold ϵ_{LOD} and renormalized once more to produce $\hat{\mathbf{y}}$; we chose $\epsilon_{\text{LOD}} = 0.001$ so that the lowest ground-truth ratio of ~0.0025 was allowed to appear. We show the sensitivity of the errors to the choice of ϵ_{LOD} in Supplementary Fig. C1b.

CAMI2 strain-madness benchmark

When evaluating strain-level profiling on the CAMI2 strain-madness dataset, we constructed each method's database (mGEMS, StrainGE and ChronoStrain) using only the gold-standard genomes; see Supplementary Text D for a discussion on why this choice was made. We did not cluster these genomes for any method, so that each method would have to deal with the full extent of genome-level granularity.

We set out to evaluate methods on those taxa that had more than one conspecific gold-standard strain. We also wanted to standardize ChronoStrain's database construction across different taxa, and so we further narrowed it down to taxa with published MLST schema as of June 2024. As mentioned in the main text, these taxa were *S. pneumoniae*, *E. coli*, *K. pneumoniae*, *S. aureus* and *E. faecium*.

For mGEMS and StrainGE, we constructed a single pan-species database using all of the gold-standard genomes. For ChronoStrain, we designed five separate databases, one for each species. Per species, we constructed a set of marker seeds from two sources: (1) MetaPhlan4 markers in all species-level genome bins (SGBs) labelled with that species name and (2) reference gene sequences for each gene in that species' MLST schema. For instance, to construct a database for *K. pneumoniae* classification, we used *K. pneumoniae* MLST genes and MetaPhlAn4 SGB marker genes as seeds. All gold-standard strains are present in every sample, thus: for ChronoStrain we did not apply a $\overline{\pi}$ threshold, for StrainGST we lowered the score threshold to 0.0, and for mGEMS no abundance thresholding was done.

Computational resources

For benchmarking, all four methods were run on stock Alienware Aurora R15s (Intel 12900KF with 128 GB of RAM). ChronoStrain's inference step, in particular, was run on a single RTX 3090; other benchmarked methods were not designed with GPU hardware in mind. GPU memory size is the primary limiting factor that requires us to cluster the database. If one includes many more markers, more database clusters and/or more samples, a GPU with more memory is required. All benchmark analyses were able to fit on the RTX 3090 (typical CPU memory footprint was less than ~10 GB CPU RAM during inference).

Sequencing and real-data processing

MacConkey-cultured samples were sequenced in the same manner as outlined for the original UMB dataset[7]. Starting with the raw reads, we used the demultiplexed, whole-genome portion of the UMB dataset for all participants. Just before analysis, all reads from UMB were preprocessed using the KneadData pipeline (v.0.11.0, https://huttenhower.sph.harvard.edu/kneaddata/), which invokes Trimmomatic v.0.39 (ref. 20) to trim adapters and low-quality bases at the ends (phred ≤ 10), and Bowtie2 (ref. 55) to discard reads that align to the human genome. The BBS metagenomic reads available online did not require trimming and decontamination. CAMI strain-madness reads were also quality trimmed using Trimmomatic, but using a more conservative setting (phred ≤ 5) to better retain paired-end information. All processing details can be found in our codebase.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All UMB-related sequencing data, including the new MacConkey-culture sequencing experiments, are available under BioProject ID PRJNA400628. Publicly available BBS sequencing reads were downloaded from the European Nucleotide Archive under accession PRJEB32631, and isolates under accession PRJEB22252. The ~640k genome Themisto index was downloaded from Zenodo (https://doi.org/10.5281/zenodo.7736981 (ref. 56). Databases and raw outputs for all real-data analyses were uploaded in Zenodo (https://doi.org/10.5281/zenodo.10932689 (ref. 57) and https://doi.org/10.5281/zenodo.10932761 (ref. 58), along with semi-synthetic benchmark inputs (https://doi.org/10.5281/zenodo.14593703 (ref. 59).

Code availability

The latest version of the ChronoStrain software is available on GitHub at https://github.com/gibsonlab/chronostrain. This paper used ChronoStrain v0.6.0, available on GitHub and archived on Zenodo at https://doi.org/10.5281/zenodo.15116549 (ref. 60). This paper's analyses also used Themisto v.3.2.1, mSWEEP v.2.0.0-3-gbd20c93, StrainGE v.1.3.8, StrainEst v.1.2.4, Kraken2 v.2.1.3 (database k2_standard_16gb_20240112), Bracken v.2.9, and MetaPhlAn v.4.0.6 (database Jun23_CHOCOPhlAnSGB_202307). Biological reads were preprocessed via KneadData v.0.11.0; in silico reads were generated using ART v.2016-Jun-06.

References

1. Lloyd-Price, J., Abu-Ali, G. & Huttenhower, C. The healthy human microbiome. *Genome Med.* **8**, 51 (2016).
2. Schloss, P. D. & Westcott, S. L. Assessing and improving methods used in operational taxonomic unit-based approaches for 16s rRNA gene sequence analysis. *Appl. Environ. Microbiol.* **77**, 3219–3226 (2011).
3. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
4. Yan, Y., Nguyen, L. H., Franzosa, E. A. & Huttenhower, C. Strain-level epidemiology of microbial communities and the human microbiome. *Genome Med.* **12**, 71 (2020).
5. Worley, J. et al. Genomic determination of relative risks for *Clostridioides difficile* infection from asymptomatic carriage in intensive care unit patients. *Clin. Infect. Dis.* **73**, e1727–e1736 (2020).
6. Yassour, M. et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci. Tansl. Med.* **8**, 343ra81 (2016).
7. Worby, C. J. et al. Longitudinal multi-omics analyses link gut microbiome dysbiosis with recurrent urinary tract infections in women. *Nat. Microbiol.* **7**, 630–639 (2022).
8. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
9. Luo, C. et al. ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* **33**, 1045–1052 (2015).

10. Schaeffer, L., Pimentel, H., Bray, N., Melsted, P. & Pachter, L. Pseudoalignment for metagenomic read assignment. *Bioinformatics* **33**, 2082–2088 (2017).
11. Reppell, M. & Novembre, J. Using pseudoalignment and base quality to accurately quantify microbial community composition. *PLoS Comput. Biol.* **14**, e1006096 (2018).
12. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
13. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
14. Quince, C. et al. Desman: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* **18**, 181 (2017).
15. van Dijk, L. R. et al. Strainge: a toolkit to track and characterize low-abundance strains in complex microbial communities. *Genome Biol.* **23**, 74 (2022).
16. Mäklin, T. et al. Bacterial genomic epidemiology with mixed samples. *Microb. Genom.* **7**, 000691 (2021).
17. Ahn, T.-H., Chai, J. & Pan, C. Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics* **31**, 170–177 (2015).
18. Sankar, A. et al. Bayesian identification of bacterial strains from sequencing data. *Microb. Genom.* **2**, e000075 (2016).
19. Smith, B. J., Li, X., Abate, A., Shi, Z. J. & Pollard, K. S. Scalable microbial strain inference in metagenomic data using StrainFacts. *Front. Bioinform.* https://doi.org/10.3389/fbinf.2022.867386 (2022).
20. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
21. Meyer, F. et al. Critical assessment of metagenome interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
22. Shao, Y. et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
23. Enright, M. C. & Spratt, B. G. Multilocus sequence typing. *Trends Microbiol.* **7**, 482–487 (1999).
24. Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting strains in microbiomes. *Nat. Rev. Microbiol.* **18**, 491–506 (2020).
25. Albanese, D. & Donati, C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nat. Commun.* **8**, 2260 (2017).
26. Jolley, K. A., Bray, J. E. & Maiden, M. C. Open-access bacterial population genomics: Bigsdb software, the pubmlst. org website and their applications. *Wellcome Open Res.* **3**, 124 (2018).
27. Achtman, M. & Pluschke, G. Clonal analysis of descent and virulence among selected *Escherichia coli*. *Annu. Rev. Microbiol.* **40**, 185–210 (1986).
28. Squadrito, F. J. & del Portal, D. *Nitrofurantoin* (StatPearls Publishing, 2025).
29. Khanna, N. R. & Gerriets, V. *Beta-Lactamase Inhibitors* (StatPearls Publishing, 2025).
30. Mäklin, T. et al. Strong pathogen competition in neonatal gut colonisation. *Nat. Commun.* **13**, 7417 (2022).
31. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
32. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
33. Blanco-Miguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
34. Scholz, M. et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* **13**, 435–438 (2016).
35. Baid, G. et al. Deepconsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat. Biotechnol.* **41**, 232–238 (2023).
36. Petty, N. K. et al. Global dissemination of a multidrug resistant *Escherichia coli* clone. *Proc. Natl Acad. Sci. USA* **111**, 5694–5699 (2014).
37. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A* **186**, 453–461 (1946).
38. Blei, D. M. & Lafferty, J. D. Dynamic topic models. In *Proc. 23rd International Conference on Machine Learning* 113–120 (ICML, 2006).
39. Wang, C., Blei, D. & Heckerman, D. Continuous time dynamic topic models. In *Proc. 24th Conference on Uncertainty in Artificial Intelligence* 579–586 (AUAI Press, 2008).
40. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
41. Baker, D. N. & Langmead, B. Genomic sketching with multiplicities and locality-sensitive hashing using dashing 2. *Genome Res.* **33**, 1218–1227 (2023).
42. Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. & Blei, D. M. Automatic differentiation variational inference. *J. Mach. Learn. Res.* **18**, 1–45 (2017).
43. Bradbury, J. et al. JAX: composable transformations of Python+NumPy programs. *GitHub* http://github.com/google/jax (2018).
44. Jang, E., Gu, S. & Poole, B. Categorical reparameterization with Gumbel-Softmax. In *Proc. International Conference on Learning Representation* (ICLR, 2017); https://openreview.net/forum?id=rkE3y85ee
45. Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E. & Clermont, O. Clermontyping: an easy-to-use and accurate in silico method for *Escherichia* genus strain phylotyping. *Microb. Genom.* **4**, e000192 (2018).
46. Liu, Y. et al. *Escherichia coli* o-antigen gene clusters of serogroups o62, o68, o131, o140, o142, and o163: DNA sequences and similarity between o62 and o68, and PCR-based serogrouping. *Biosensors* **5**, 51–68 (2015).
47. Fratamico, P. M., Briggs, C. E., Needle, D., Chen, C.-Y. & DebRoy, C. Sequence of the *Escherichia coli* o121 o-antigen gene cluster and detection of enterohemorrhagic *E. coli* o121 by PCR amplification of the *wzx* and *wzy* genes. *J. Clin. Microbiol.* **41**, 3379–3383 (2003).
48. Curwen, V. EMBOSS primersearch. *EMBOSS* https://emboss.sourceforge.net/apps/cvs/emboss/apps/primersearch.html (2000).
49. McBride, S. M., Fischetti, V. A., LeBlanc, D. J., Moellering Jr, R. C. & Gilmore, M. S. Genetic diversity among *Enterococcus faecalis*. *PLoS ONE* **2**, e582 (2007).
50. Tatusova, T. et al. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624 (2016).
51. Katoh, K., Misawa, K., Kuma, K.-i & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
52. Blackwell, G. A. et al. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLoS Biol.* **19**, e3001421 (2021).
53. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
54. Mäklin, T. et al. High-resolution sweep metagenomics using fast probabilistic inference. *Wellcome Open Res.* **5**, 14 (2020).

55. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
56. Alanko, J. N. Species-colored Themisto v3 index with 640k bacterial genomes. *Zenodo* https://doi.org/10.5281/zenodo.7736981 (2023).
57. Kim, Y. & Gibson, T. E. Chronostrain manuscript: BBS analysis database and output. *Zenodo* https://doi.org/10.5281/zenodo.10932689 (2025).
58. Kim, Y. & Gibson, T. E. Chronostrain manuscript: UMB analysis database and output. *Zenodo* https://doi.org/10.5281/zenodo.10932761 (2025).
59. Kim, Y. Semi-synthetic *E. coli* abundance estimation benchmark. *Zenodo* https://doi.org/10.5281/zenodo.14593703 (2025).
60. Kim, Y. & Gibson, T. E. ChronoStrain. *Zenodo* https://doi.org/10.5281/zenodo.15116549 (2025).

Acknowledgements

The work was performed with partial funding from NIH R35GM143056 (T.E.G.), NIH R21AI154075 (T.E.G.), NIH R35GM141861 (B.B.) NIH R01GM130777 (G.K.G.), NIH U19AI110818 (A.M.E.) and NIH R01DK121822 (S.J.H., A.M.E.). We thank T. Mäklin and J. Corander for feedback and assistance regarding mGEMS, Themisto and mSWEEP; and T. Mäklin and A. K. Pöntinen for providing *E. faecalis* isolate genomes from the Baby Biome Study.

Author contributions

Y.K. and T.E.G. conceived the statistical model in consultation with B.B. Y.K., S.A., D.A. and Z.G. implemented the software. Y.K. and T.E.G. performed synthetic and real-data analyses. C.J.W., L.R.v.D., P.N.A., K.W.D., G.K.G., S.J.H., A.M.E. and B.B. provided major consultation regarding benchmarking, interpretation of analysis on UMB cohorts, and the paper. K.W.D., S.J.H. and A.M.E. acquired the UMB data, including the new sequencing of MacConkey cultures. Y.K. and T.E.G. prepared the paper. G.K.G. and B.B. provided major structural suggestions for the draft and final versions. Review and approval of the final manuscript was provided by all authors.

Competing interests

The authors declare no competing interests.

Additional information

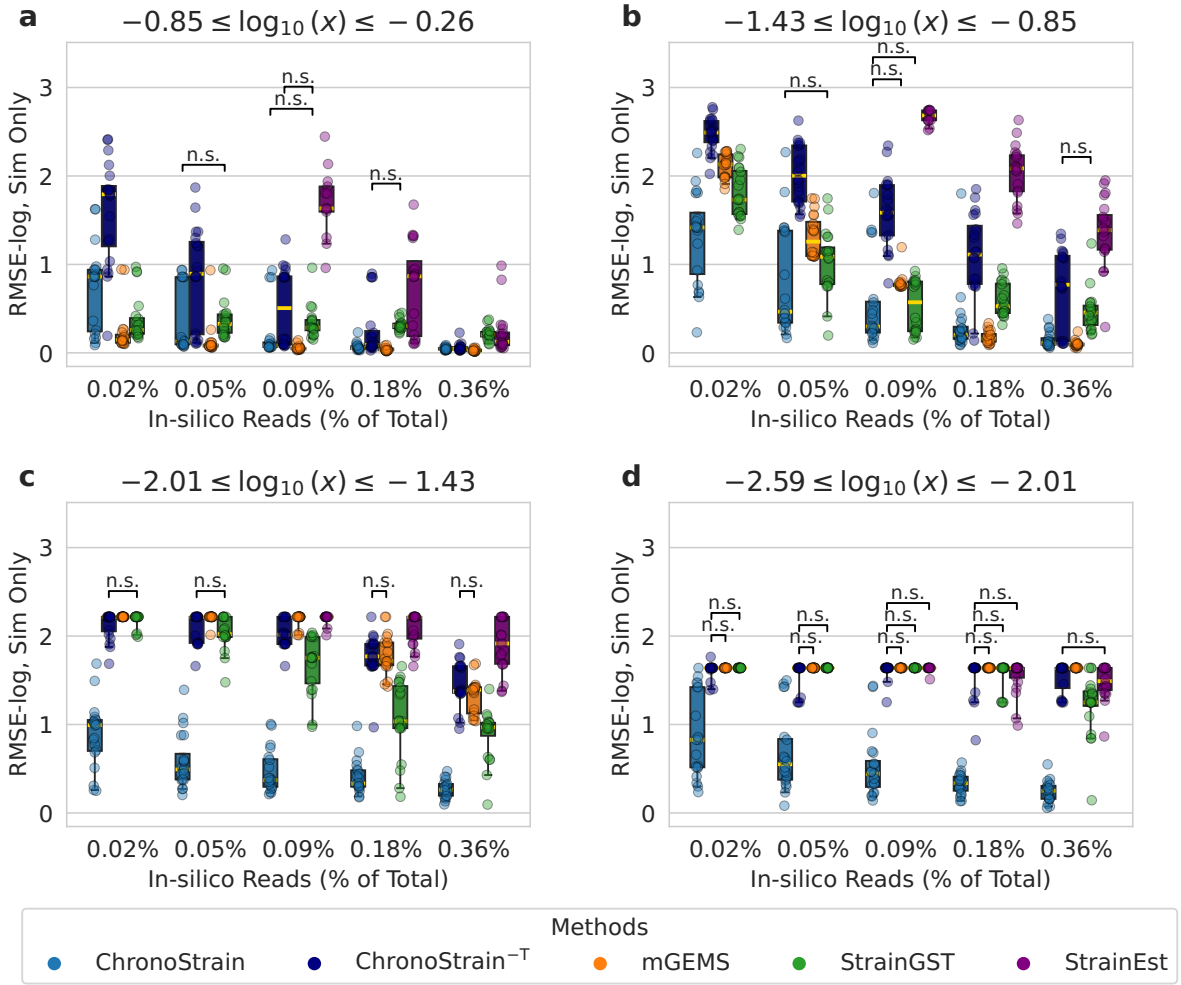
Extended data is available for this paper at https://doi.org/10.1038/s41564-025-01983-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-025-01983-z.

Correspondence and requests for materials should be addressed to Bonnie Berger or Travis E. Gibson.

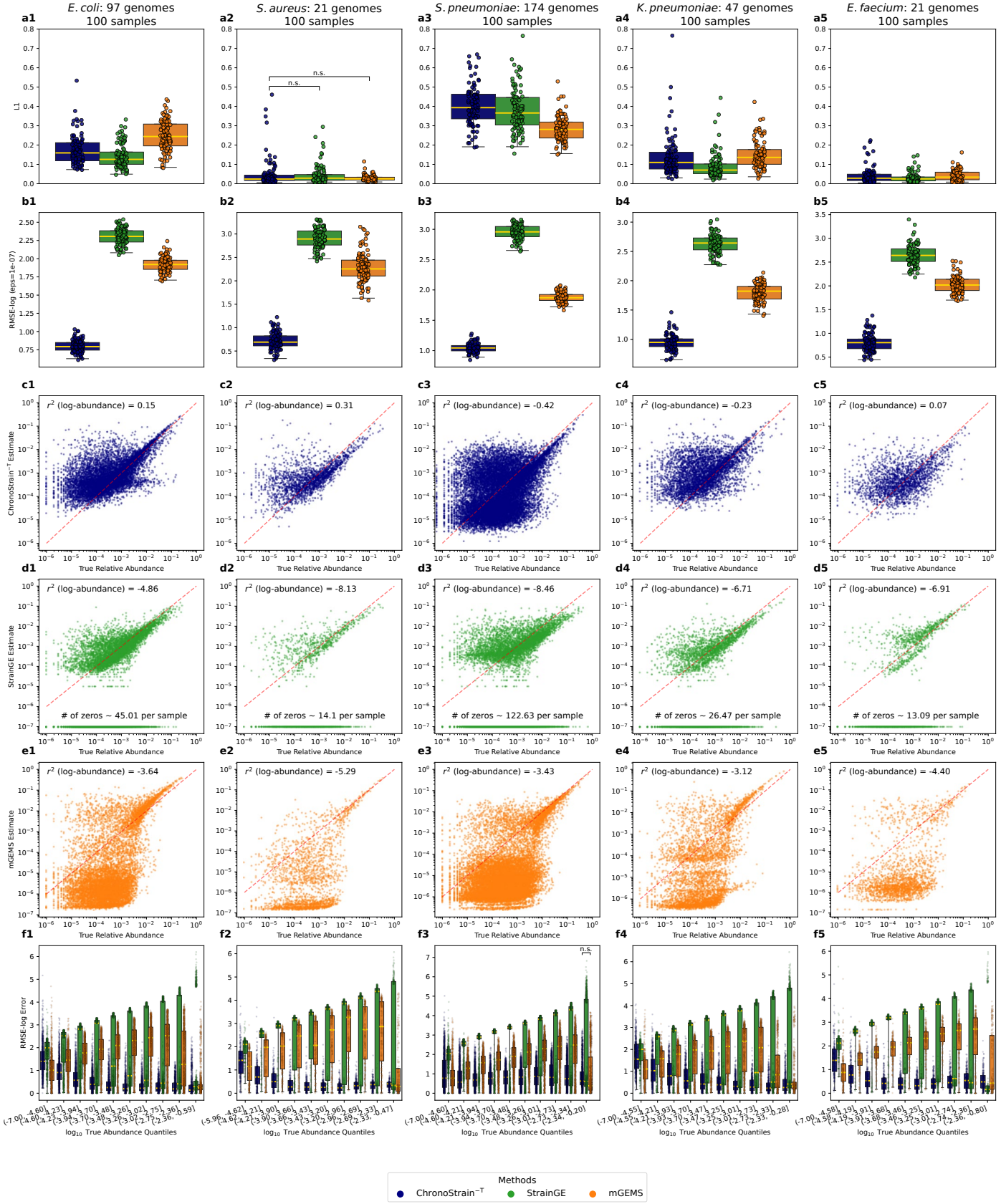
Peer review information *Nature Microbiology* thanks the anonymous reviewers for their contribution to the peer review of this work.

https://doi.org/10.1038/s41564-025-01983-z



Extended Data Fig. 1 | (Semisynthetic benchmark) Abundance-specific RMSE-log error performance. To break down the contribution of the estimates to the RMSE error by abundances, we evaluated the RMSE-log error after binning (timepoint, synthetic strain) pairs by ground-truth abundance ratios. **(a-d)** Bins are ordered from "most abundant" strain in timepoint to "least abundant", where $\log_{10}(x)$ indicates the log-ratio. The RMSE-log was evaluated after adding 10^{-4} to predictions to handle zeroes. All comparisons to Chronostrain are statistically significant at level 0.05, after two-sided, paired Wilcoxon tests with Benjamini-Hochberg (BH) correction, unless noted with an n.s. (*p*-values in Supplemental Table 2) Medians are colored yellow, boxes are 25% and 75% quantiles, whiskers are 2.5% and 97.5% quantiles.

https://doi.org/10.1038/s41564-025-01983-z

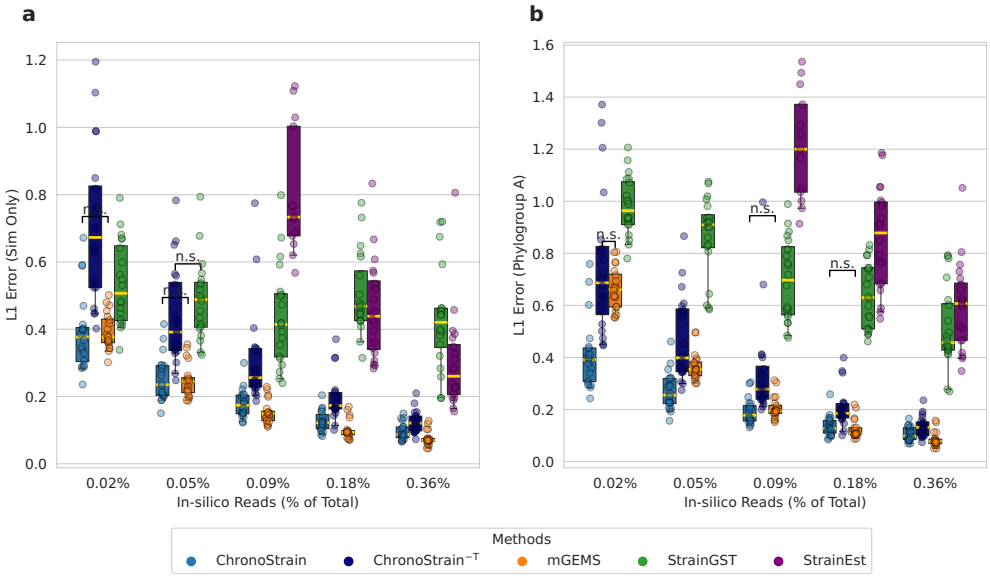


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41564-025-01983-z

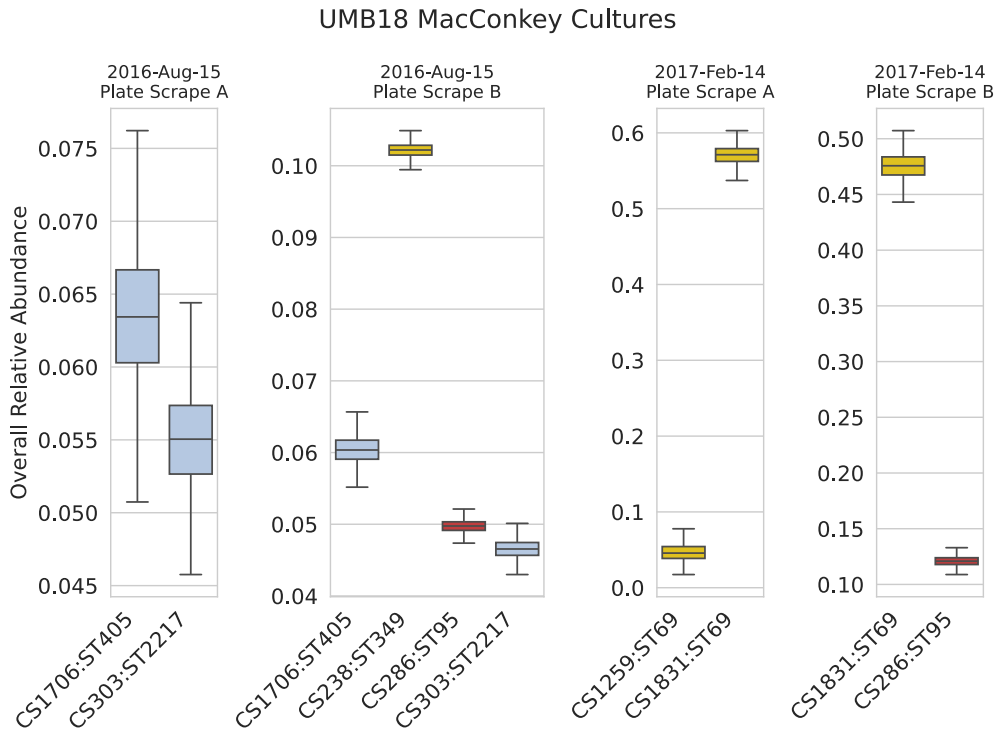
Extended Data Fig. 2 | (CAMI2) Evaluation of methods on CAMI2's strain-madness dataset. We evaluated both the L1 error (**a1**-**a5**) and the RMSE-log error (**b1**-**b5**) on five taxa. ChronoStrain^{-T} (marker database of MetaPhlAn + MLST markers) is typically middle-of-the-pack in L1 but performs the best in RMSE-log. The L1 error tends to be dominated by high-abundance predictions, and hides each method's errors for lower abundance ratios. (**c**-**e**) are plots of ground-truth abundance versus prediction. StrainGSTcontains many spurious zero-abundance predictions, and mGEMS predictions have a stark estimation drop-off when abundance is around 10^{-3} ~ 10^{-2}, depending on the taxa. (**f**) Quantile-binned errors show the stratification across different abundance levels. ChronoStrain performs consistently the best across the vast majority of bins, excluding the lowest abundance bin (where it is difficult for all methods) and the highest (where whole-genome methods are expected to perform better than ChronoStrain). All comparisons to Chronostrain are statistically significant at level 0.05, after two-sided, paired Wilcoxon tests with Benjamini-Hochberg (BH) correction, unless noted with an n.s. (*p*-values in Supplemental Tables 6, 7) In (**a**, **b**, **f**), medians are colored yellow, boxes are 25% and 75% quantiles, whiskers are 2.5% and 97.5% quantiles.

https://doi.org/10.1038/s41564-025-01983-z



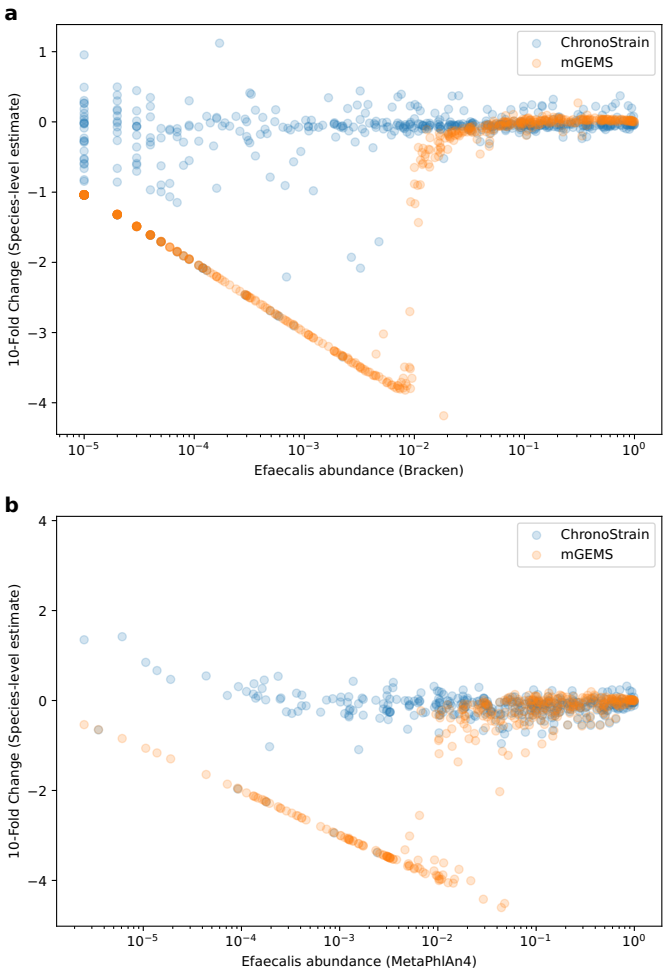
Extended Data Fig. 3 | (Semisynthetic benchmark) The L1 errors of the methods. In addition to the RMSE-log shown in Fig. 2 of the main text, we also evaluated the L1 distance to the ground truth. Note that the L1 metric has traditionally been used for benchmarking taxonomic profiling, but the lack of log-scaling means that this metric heavily favors getting the largest abundances correct while ignoring low-abundance taxa. **(a)** The L1 error evaluated after re-normalizing on the six ground-truth clusters. **(b)** The L1 error evaluated after re-normalizing on all phylogroup A clusters. All comparisons to Chronostrain are statistically significant at level 0.05, after two-sided, paired Wilcoxon tests with Benjamini-Hochberg (BH) correction, unless noted with an n.s. (*p*-values in Supplemental Table 1) Medians are colored yellow, boxes are 25% and 75% quantiles, whiskers are 2.5% and 97.5% quantiles.

https://doi.org/10.1038/s41564-025-01983-z



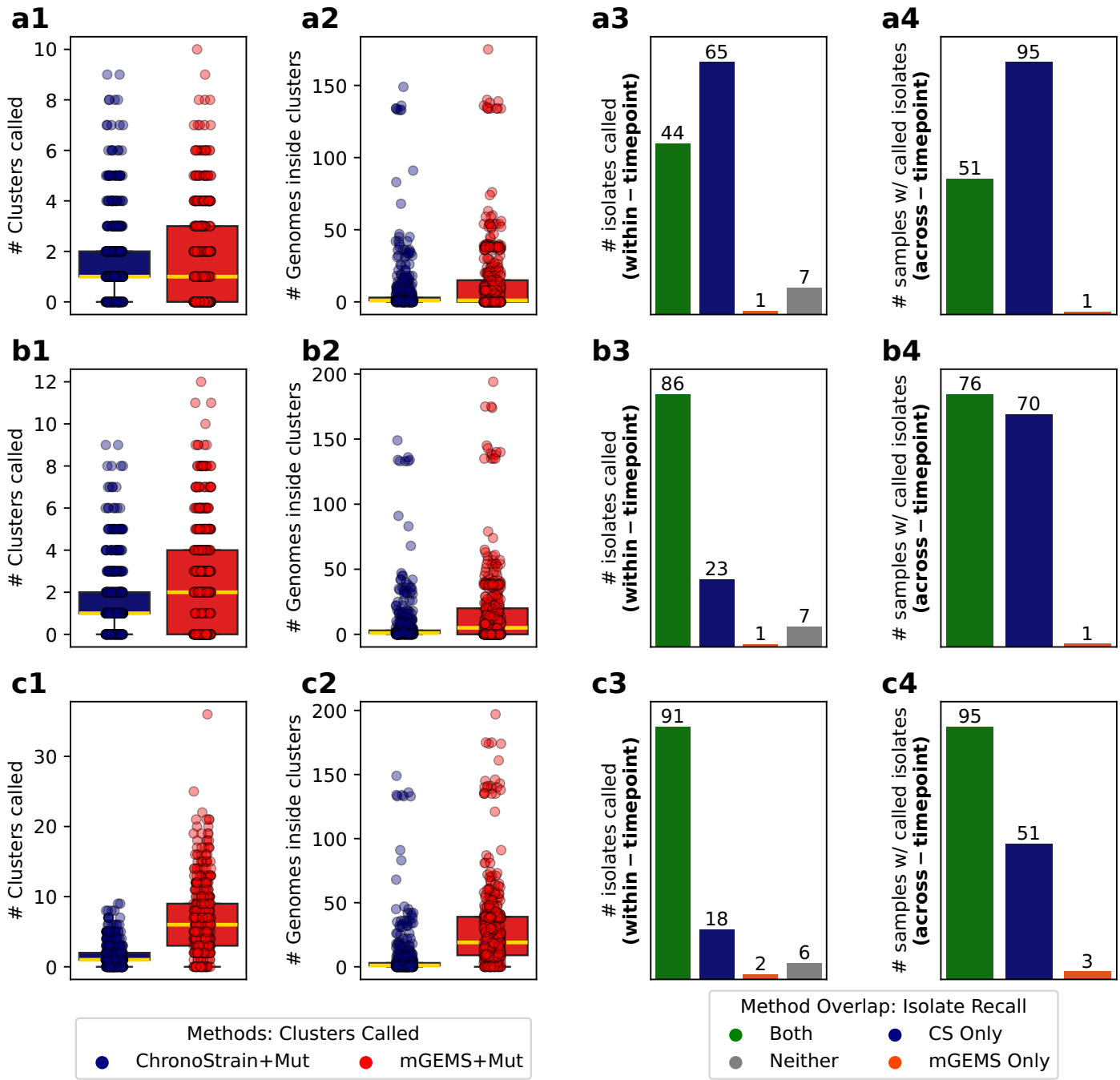
Extended Data Fig. 4 | (UMB) ChronoStrain's raw estimates for the *E. coli* abundances of MacConkey-cultures grown from UMB18 stool samples. The *y*-axis quantifies the strain clusters marked with X shown on Main Fig. 3b. Colors are chosen using the same phylogroup-based color palette. In the boxplots, the boxes denote the 25%, 50% and 75% quantiles from the posterior samples and the whiskers denote the 2.5% and 97.5% quantiles.

https://doi.org/10.1038/s41564-025-01983-z



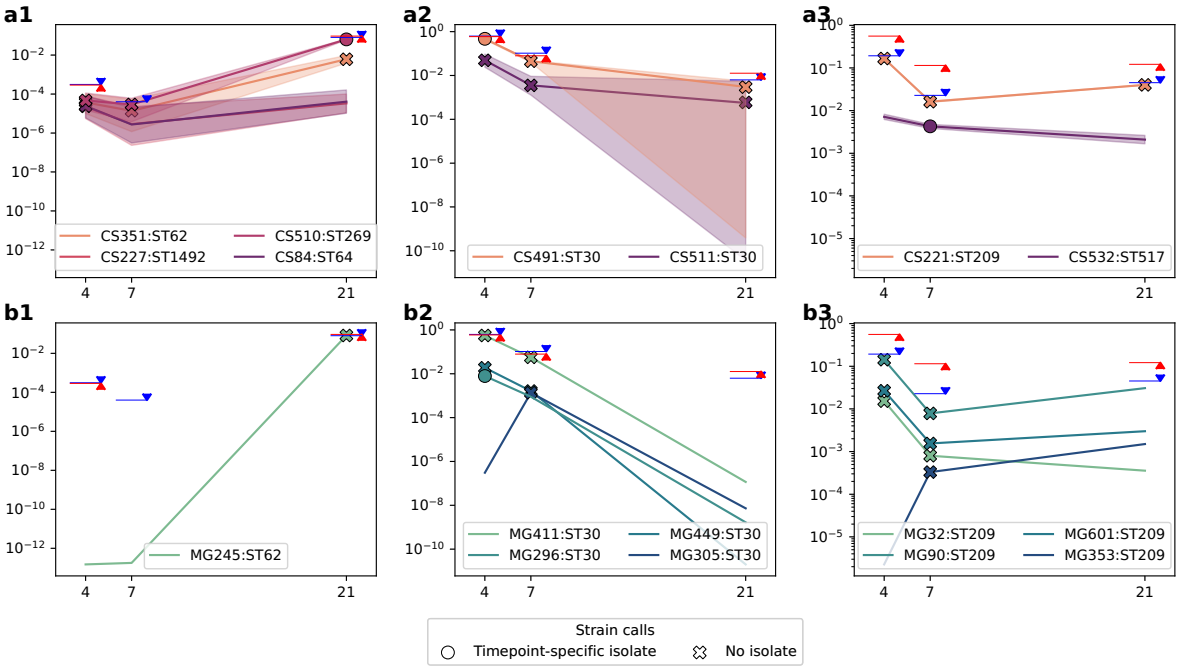
Extended Data Fig. 5 | (BBS) A comparison of the fold-change versus third-party methods for ChronoStrain and mGEMS' *E. faecalis* species estimates. We ran two methods for species-level comparison: **(a)** Kraken2 + Bracken and **(b)** MetaPhlAn4. Each dot is a single infant sample. In both plots, the fold change is computed as the difference between $\log_{10}$ of the predictions plus $\epsilon = 10^{-6}$ to avoid NaN's. The dots forming a straight, diagonal line on the bottom of the scatterplots (the line $\log_{10}(x) + y = -6$) are an artifact of this ϵ, where ChronoStrain (resp. mGEMS) – but not Bracken (resp. MetaPhlAn4) – estimates zero/near-zero abundance for *E. faecalis* on the same metagenomic sequencing input. For mGEMS, a sharp drop-off just above 10^{-2} is visible in both **(a)** and **(b)**, suggesting that the method has a detection threshold at that location. In contrast, ChronoStrain generally agrees with both third-party methods all the way down to 10^{-5}, with higher spread as *E. faecalis* becomes rarer in the sample.

https://doi.org/10.1038/s41564-025-01983-z



Extended Data Fig. 6 | (BBS) Inference result comparison using the mutated database of 117 infant isolates, in the style of Figure 5. These are the results after running the mutated-database experiment on the BabyBiome dataset ($n = 486$ samples), as discussed in Supplemental Information, Section E. The first row (**a**) only retains mGEMS predictions with `demix_check` quality scores 2 or better. The second row (**b**) retains 3 or better, third row (**c**) is 4 or better. ChronoStrain thresholds are held fixed in all three rows (ChronoStrain: $\overline{\pi} = 0.95$, ratio ≥ 0.065). One may loosen the `demix_check` threshold in order to obtain comparable numbers of isolate calls (**a3**,**b3**,**c3**) at the cost of calling more clusters (**a1**,**b1**,**c1**), whereas ChronoStrain remained largely the same from the unmodified run from Fig. 5 (Fig. 6). In the boxplots, medians are colored yellow, boxes are 25% and 75% quantiles, whiskers are 2.5% and 97.5% quantiles.

https://doi.org/10.1038/s41564-025-01983-z



Extended Data Fig. 7 | (BBS) Abundance trajectories of clusters using mutated databases. Using the mutated analysis results, we plotted the inferred trajectories for the same three examples shown in Fig. 5a,b, drawn in the same style. Just like before, the trajectory for a cluster is rendered only if it passes the filter for the respective method in at least one timepoint. For each trajectory's timepoint, if it passed the filter, we place a marker. It is either an O (has an isolate cultured at that timepoint) or X (no isolate for that timepoint). After mutation, mGEMS' filter no longer calls the corresponding isolates in infants `A01077`, `B02273` (**b1, b3**) whereas ChronoStrain remains unchanged from the original inference (**a1, a3**). For `B00053` (**a2, b2**), the isolate is called correctly at the first timepoint for both methods but in mGEMS is no longer the dominant strain, whereas both methods in the original analysis agreed that it was the dominant strain in the sample (Fig. 5, panels a2, b2).

nature portfolio

Corresponding author(s): Travis Gibson, Bonnie Berger

Last updated by author(s): 2025-Feb-28

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Reference sequences were downloaded from NCBI using the datasets API, and reads were downloaded using sra-tools.
Data analysis	Analysis and Benchmarking used Chronostrain v0.6.0 (https://github.com/gibsonlab/chronostrain), mGEMS (https://github.com/PROBIC/mGEMS), Themisto (https://github.com/algbio/themisto), StrainGE (https://github.com/broadinstitute/StrainGE) and StrainEst v1.2.4 (https://github.com/compmetagen/strainest). Biological reads were pre-processed via KneadData v0.11.0; in-silico reads were generated using ART v2016-Jun-06.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All UMB-related sequencing data, including the new MacConkey-culture sequencing, are available under BioProject ID PRJNA400628. Publically available BBS

sequencing reads were downloaded from the European Nucleotide Archive under accession PRJEB32631, and isolates under accession PRJEB22252. The 640k genome Themisto index was downloaded from Zenodo (7736981). Databases and raw outputs for all real-data analyses are available on Zenodo (10932689, 10932761). The semisynthetic benchmark inputs are available on Zenodo (14593703).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	All UMB study participants were identified as women in the original study.
Reporting on race, ethnicity, or other socially relevant groupings	No socially relevant categorization variables were used in the UMB study.
Population characteristics	Women from the St. Louis, MO area reporting three or more UTIs in the past 12 months were recruited into the rUTI studyarm, while women with no history of UTI (at most one UTI ever) were recruited into the control arm via the Department of Urological Surgery at Barnes-Jewish Hospital in St. Louis, MO. We excluded participants who: i) had inflammatory bowel disease (IBD) or urological developmental defects (e.g., ureteral reflux, kidney agenesis, etc.), ii) were pregnant, iii) take antibiotics as prophylaxis for rUTI, and iv) were younger than 18 years or older than 45 at the time of enrollment.
Recruitment	rUTI participants were recruited based on clinical history via the Department of Urological Surgery, along with age-matched control participants with no history of rUTI. Flyers were posted around Wash U Medical School, Wash U in St. Louis campus, and the Barnes-Jewish Hospital. Participants were remunerated midway through the study, and at the end of the study upon completion, via gift cards. Self selection biases may therefore exist; in particular we did not collect socio-economic data on participants. However, given age matching and the similarity in self-reported dietary habits between cohorts, we do not anticipate any such bias to have a significant impact on the composition of the gut microbiome.
Ethics oversight	The UMB study was conducted with the approval and under the supervision of the Institutional Review Board of Washington University School of Medicine in St. Louis, MO (IRB No. 201401068)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size calculation for longitudinal microbiome analyses is not straightforward. Each high-risk patient represents their own controlled experiment. No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous longitudinal microbiome studies which were able to detect significant effects, e.g. Dethlefsen & Relman PNAS 2011. 108 Suppl 1: p.4554-61; Turnbaugh et al., Nature, 2006. 444(7122): p. 1027-31.
Data exclusions	No data were excluded from the UMB study.
Replication	The UMB study was an observational cohort study and no replication was performed, although we have described the recruitment process and sampling strategy sufficiently such that the study may be replicated.
Randomization	The UMB study was an observational study with no intervention and cohorts based on pre-determined criteria; as such, no randomization was required. Control participants were age-matched to rUTI participants, and few dietary differences existed between the cohorts based on survey responses.
Blinding	The UMB study was an observational study with no intervention and cohorts based on pre-determined criteria; as such, blinding was not relevant.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.