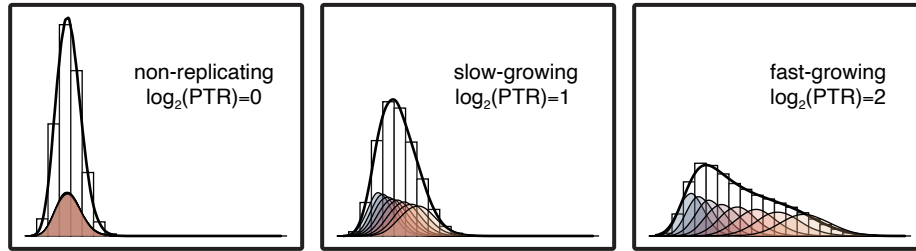


1 **Pilea: profiling bacterial growth dynamics from**
2 **metagenomes with sketching**

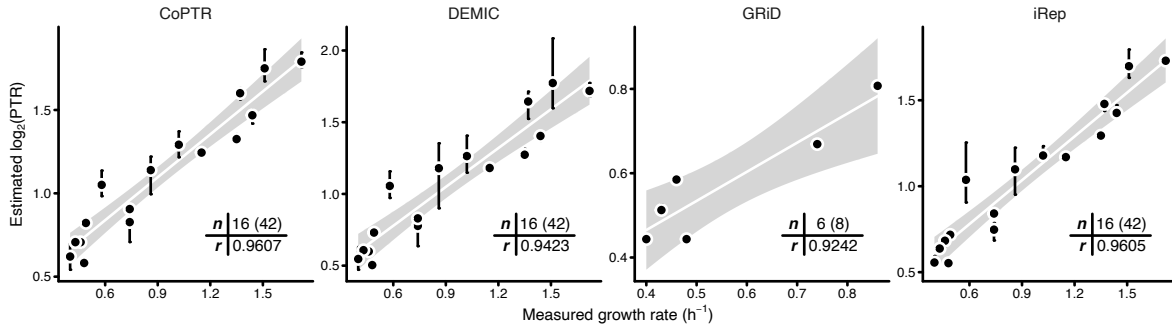
3 *Supplementary information*

4 **List of Figures**

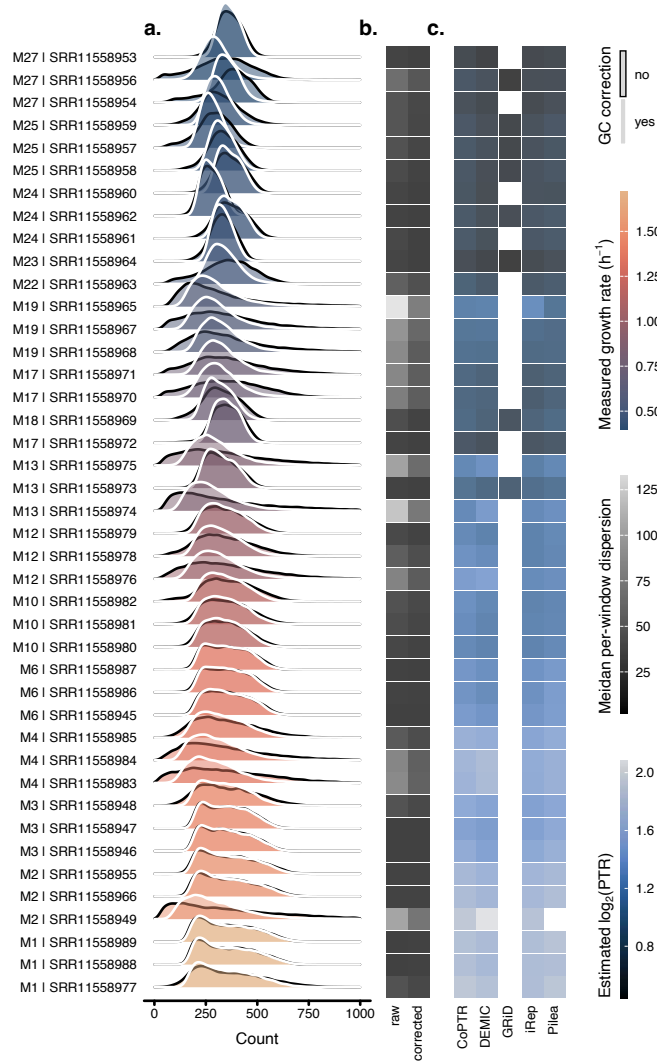
5	S1	Graphical illustration of Pilea’s Poisson assumption	2
6	S2	Performance of existing PTR inference tools on the <i>Escherichia coli</i> dataset .	3
7	S3	Details of PTR estimates for the <i>Escherichia coli</i> dataset	4
8	S4	Relationships between Pilea’s PTR estimates and measured growth rates for	
9		four bacterial isolates	5
10	S5	ANI estimates for <i>Escherichia</i> assemblies	6
11	S6	Performance details of tools aggregated by number of strains or coverage on	
12		the simulated dataset	7
13	S7	Associations between Pilea’s PTR estimates and observed growth rates for	
14		MAGs from the marine dataset	8
15	S8	Associations between other tools’ PTR estimates and observed growth rates	
16		for MAGs from the marine dataset	9
17	S9	Pairwise comparison of PTR estimates between Pilea and other tools for the	
18		marine dataset	10
19	S10	Times for processing the marine dataset using bacterial MAGs versus GTDB	
20		R226 representatives	11
21	S11	Comparison of coverage profiles between Illumina and Oxford Nanopore Tech-	
22		nologies (ONT) sequencing	12



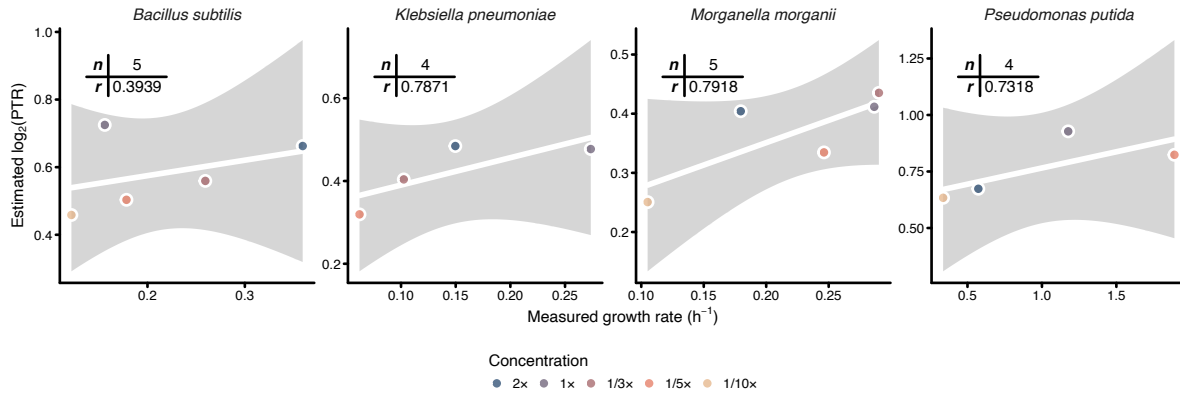
Supplementary Fig. S1: Graphical illustration of Pileup’s Poisson assumption. Pileup assumes k -mer counts within a sufficiently small genomic window can be approximated by a Poisson distribution, even for fast-growing species with multifork replication ($\text{PTR} > 2$). As a consequence, the observed counts of all sketched k -mers can be described using a series of Poisson distributions, with each corresponding to a window and characterized by a distinct rate parameter (λ). In practice, instead of fitting a simple Poisson distribution, Pileup fits a finite mixture of ZTP distributions to each window and uses the rate parameter of the highest-weight component to represent the window’s expected coverage, thereby increasing its robustness to noise caused by e.g., inter- and intraspecific repeats.



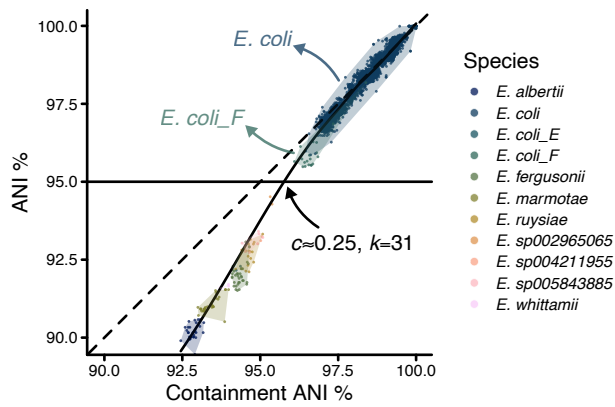
Supplementary Fig. S2: Performance of existing PTR inference tools on the *Escherichia coli* dataset. Tables show the number of medium/sample estimates (n) returned by each tool and Pearson correlation coefficients (r). White lines represent linear regression fits. DEMIC used [GCF_028618655.1](#) (contig-level) as its reference, while the other tools used [GCF_000005845.2](#) (complete-genome). GRiD produced no estimates for samples with growth rates $> 1 \text{ h}^{-1}$.



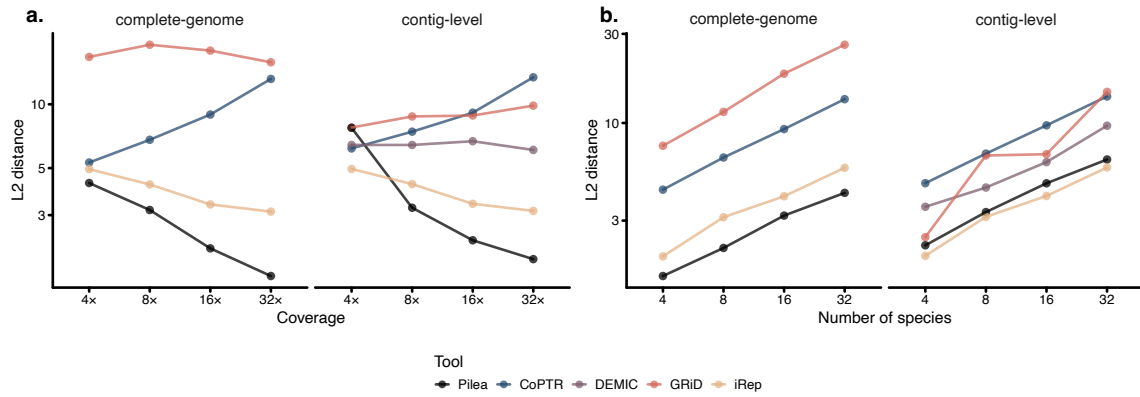
Supplementary Fig. S3: Details of PTR estimates for the *Escherichia coli* dataset. **a** Count distributions of Pileup's sketched k -mers before (black outline) and after (white outline) GC-bias correction. Colors reflect measured growth rates. **b** Median per-window dispersions (variance-to-mean ratios) before and after GC-bias correction. **c** PTR estimates from different tools. Blank cells indicate no output. All tools were run with default parameters. Pileup returned no estimates for one sample that exhibited excessive per-window dispersion, even after GC-bias correction.



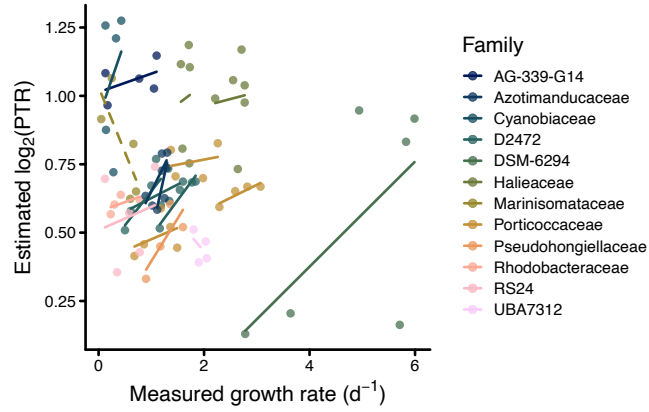
Supplementary Fig. S4: Relationships between Pilea's PTR estimates and measured growth rates for four bacterial isolates. Colors indicate LB media nutrient concentrations (2 $\times$, 1 $\times$, 1/3 $\times$, 1/5 $\times$, and 1/10 $\times$) used for bacterial cultivation. Tables list sample sizes (n) and Pearson correlation coefficients (r). Samples with negative measured growth rates were excluded. All four isolates exhibited moderately linear relationships between $\log_2(\text{PTR})$ estimates and measured growth rates.



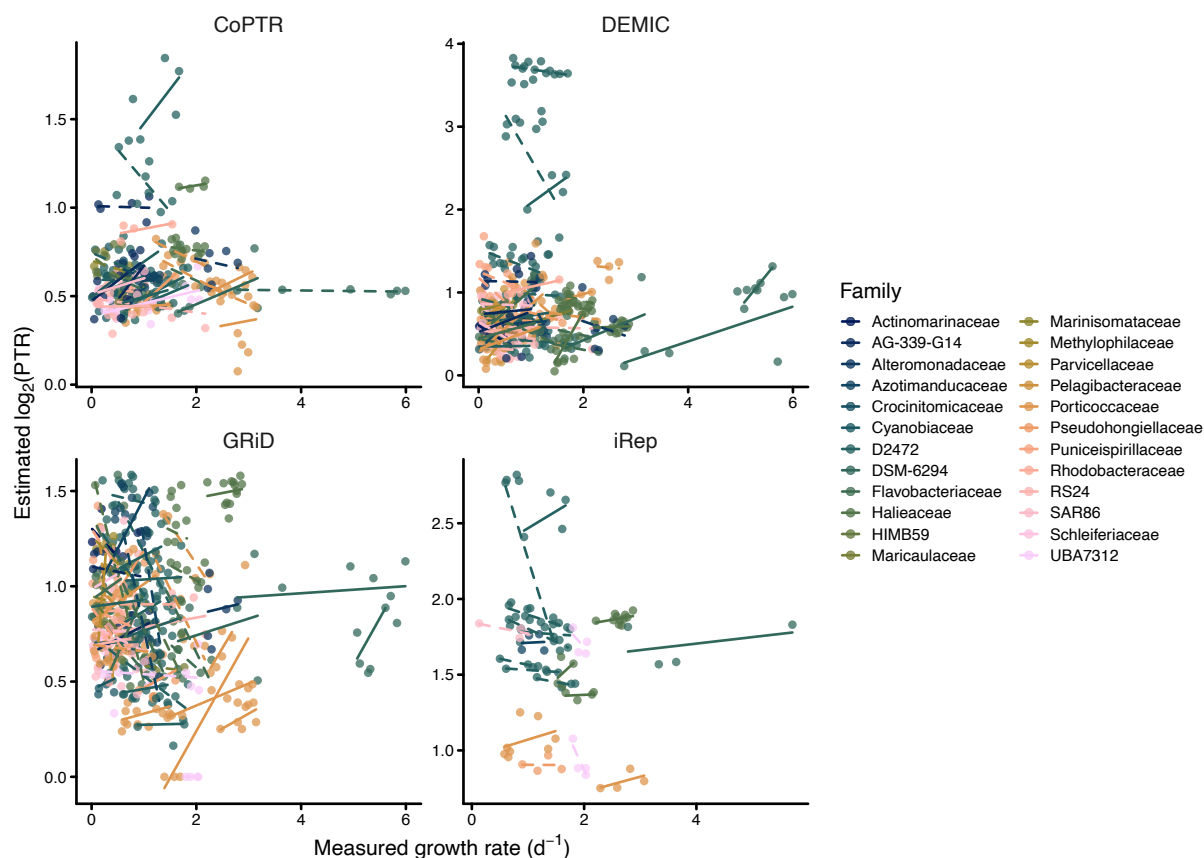
Supplementary Fig. S5: ANI estimates for *Escherichia* assemblies. ANI between [GCF_000005845.2](#) (reference) and GTDB 45,529 *Escherichia* assemblies (query) estimated by skani or based on containment. Containment ANI tended to overestimate at low identity ranges.



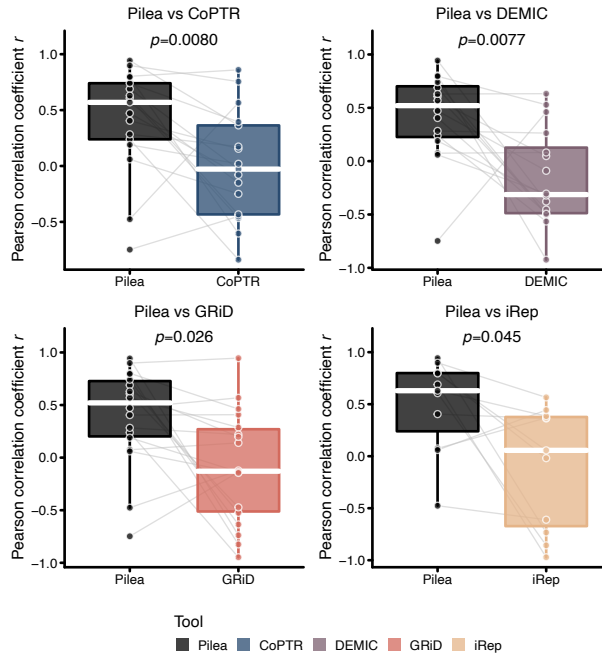
Supplementary Fig. S6: Performance details of tools aggregated by number of strains or coverage on the simulated dataset. a,b L2 distances at different coverages (a) or with different numbers of species (b) using complete-genome or contig-level assemblies as references. Colors indicate tools. Pilea and iRep showed improved performance at high coverages, while the other tools (CoPTR, DEMIC, and GRiD) did not.



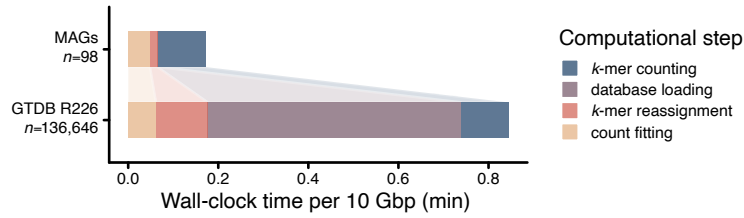
Supplementary Fig. S7: Associations between Pilea's PTR estimates and observed growth rates for MAGs from the marine dataset. MAGs are colored according to their taxonomic family predicted by GTDB-Tk. Solid and dashed lines indicate positive ($n = 16$) and negative ($n = 2$) linear associations, respectively.



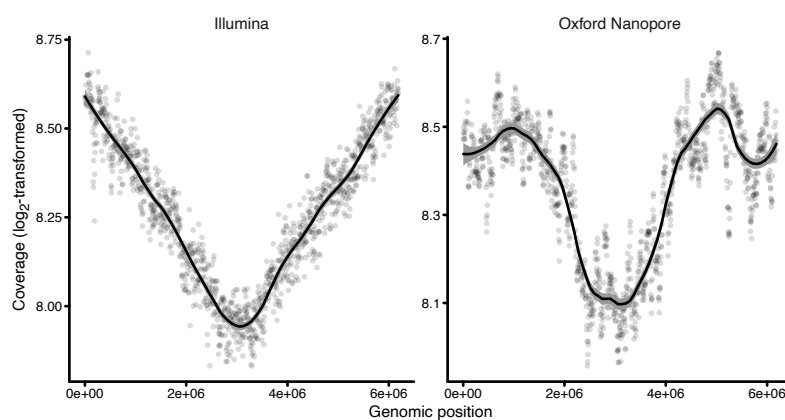
Supplementary Fig. S8: Associations between other tools' PTR estimates and observed growth rates for MAGs from the marine dataset. MAGs are colored according to their taxonomic family predicted by GTDB-Tk. Solid and dashed lines indicate positive and negative linear associations, respectively.



Supplementary Fig. S9: Pairwise comparison of PTR estimates between Pilea and other tools for the marine dataset. Statistical significance was assessed using p -values obtained from two-sided Wilcoxon signed-rank (paired) tests. Although Pilea returned the fewest PTR estimates ($n = 18$), its estimates showed significantly higher correlations with observed growth rates than those of all other tools.



Supplementary Fig. S10: Times for processing the marine dataset using bacterial MAGs versus GTDB R226 representatives. Times for each computational step were extracted from Pilea’s log files. Substantial increases in run-time were observed for database loading (from 0.002 to 0.563 min) and *k*-mer reassignment (from 0.017 to 0.115 min) when using GTDB R226 representatives, primarily due to the larger number of sketched *k*-mers.



Supplementary Fig. S11: Comparison of coverage profiles between Illumina and Oxford Nanopore Technologies (ONT) sequencing. Raw Illumina ([SRX18086120](#)) and ONT ([SRX18086121](#)) sequencing reads for a pure culture spike-in species *Pseudomonas allopuntida* ([SAMN31524854](#)) were mapped to the assembled genome using minimap2, and mean per-base coverage was calculated in 5,000 bp non-overlapping windows. While the Illumina coverage profile displayed the expected V-shaped coverage pattern (origin-to-terminus coverage gradient), the ONT profile exhibited systematic fluctuations that persisted even after smoothing.