nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-018-0182-0

Quantifying and comparing bacterial growth dynamics in multiple metagenomic samples

Yuan Gao and Hongzhe Li*

The accurate quantification of microbial growth dynamics for species without complete genome sequences is biologically important, but computationally challenging in metagenomics. Here we present dynamic estimator of microbial communities (DEMIC; https://sourceforge.net/projects/demic/), a multi-sample algorithm based on contigs and coverage values, to infer the relative distances of contigs from the replication origin and to accurately compare bacterial growth rates between samples. We demonstrate robust performances of DEMIC for various sample sizes and assembly qualities using multiple synthetic and real datasets.

The growth dynamics of microbial populations is an important feature that reflects their physiological status and drives variation in their composition. Available approaches for estimating the growth dynamics of bacteria make use of phenotypic markers, sequencing tag or fluorescence dilution and involve additional experimental steps[1–4]. Such methods are often limited by low stability, population complexity or aerobic environment. Recently, peak-to-trough ratio (PTR) was reported as a promising index for species with complete genome sequences[5]. PTR measures growth dynamics of a bacterial population by calculating the difference in sequencing coverage that results from bidirectional replication from a fixed replication origin in the genome (Supplementary Fig. 1).

For species with only genome assemblies, the accurate locations of the assemblies on the original genome are unknown, making it infeasible to calculate the PTR of the coverages[5,6]. In addition, contigs that are assembled from metagenomic sequencing data are usually fragmented owing to intraspecific variations, interspecific and/or intraspecific repeated sequences as well as limited sequencing depths[7,8]. Moreover, binning algorithms sometimes fail to cluster all of the contigs from the same species into one group, or erroneously include a fraction of contigs from other species[9–11]. These noisy features complicate the estimation of growth dynamics for genome assemblies.

Here, we present dynamic estimator of microbial communities (DEMIC), which takes advantage of highly fragmented contigs that are assembled in multiple metagenomic samples, such as those from different time points or different host subjects, to accurately compare growth dynamics of a given species that is observed in multiple samples. In DEMIC, for a given contig cluster, relative distances from the replication origin that contribute most to the variability of read coverages of different contigs are inferred by dimension reduction of the contig coverage matrix (Fig. 1, Methods). This is combined with GC bias correction, contig and sample filtering to achieve the final estimates of the growth dynamics of different samples. The method can be applied to a wide range of bacterial communities with closely related species and is robust to sample sizes, contig contamination and completeness of contig clusters.

To evaluate the performance of DEMIC, we used multiple sequencing datasets from four bacterial species grown in different medium, including 36, 36, 50 and 19 datasets of *Lactobacillus gasseri*, *Enterococcus faecalis*, *Citrobacter rodentium* and *Escherichia coli*, respectively (Supplementary Fig. 2, Methods). When applied to contig clusters of three species (*L. gasseri*, *E. faecalis* and *C. rodentium* with completeness and contamination shown in Supplementary Table 1) that were generated from the synthetic datasets by coassembly and binning[12,13], DEMIC was able to estimate the growth rates in all 122 species–experiment combinations (Supplementary Fig. 3).

PTRC[5], the method to calculate the PTR, relies on the availability of complete reference genomes and has been demonstrated experimentally to be accurate at estimating the growth dynamics for the datasets analyzed above. PTRs from PTRC were therefore chosen as the gold standard for our evaluations. As shown in Fig. 2a,b and Supplementary Fig. 4, estimates from DEMIC and PTRC were highly correlated for all 122 growth rates of all three species. By contrast, iRep[6], the algorithm based on the draft genomes, had relatively low and unstable correlations with PTRC. For example, *E. faecalis* had a moderate growth rate in sample 24 based on the estimates from PTRC and DEMIC, but it was classified by iRep as fast growing (Fig. 2a). For *C. rodentium*, although contig contamination accounted for about 15% of the contig clusters[14] (Methods) (Supplementary Table 1), estimates from DEMIC still showed a correlation of 0.97 with PTRs (Fig. 2b).

For growth dynamic estimation, one of the keys steps is the inference of the relative distance of a contig to the replication origin. In DEMIC, this step is based on principal component analysis (PCA) of contig coverages in multiple samples (Methods). For all three species, the inferred relative distances based on multiple samples were more accurate than direct sorting of contigs based on their coverages in a single sample (Supplementary Table 2). For example, the inferred relative distances of *C. rodentium* contigs achieved a high correlation of 0.964 with the true distances, whereas direct sorting of the contigs by coverages only had a mean correlation of 0.756 in all 50 samples.

We next evaluated how assembly contamination, completeness and sample size affect the performance of DEMIC. First, to assess the effect of assembly contamination, we randomly added different fractions of assembled sequences from *E. coli* into the contig clusters of *L. gasseri* and *E. faecalis*, and used the mixed assemblies to compare the performance. Remarkably, all estimates from DEMIC still showed a high correlation of 0.98 even when as many as 30% of contigs were from contamination (Fig. 2c), suggesting high effectiveness and robustness of the contig-filtering steps used by DEMIC. Second, we observed that increasing the fraction of contigs led to an improved accuracy when estimating

Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
*e-mail: hongzhe@pennmedicine.upenn.edu

growth dynamics (Fig. 2d). In 93.3% of test cases with 40–50% completeness of the contigs, estimates from DEMIC showed a high correlation ($r > 0.9$) with the PTR values, and such a high correlation was observed in all 120 tests for which completeness was 60% or more. Finally, we observed more stable performances of DEMIC with an increase in sample size (Supplementary Fig. 5). DEMIC generated estimates that were highly consistent with PTRs in 93.3% of test cases with only three samples ($r > 0.9$), and in 99.3% of test cases with six or more samples. By contrast, iRep showed clearly decreased performances with increased assembly contamination and little improvement with increased assembly completeness (Fig. 2c,d).

To further assess the accuracy of DEMIC at estimating the growth dynamics from more complex and diverse bacterial communities, especially those composed of closely related species, we simulated a dataset of 45 species with randomly assigned PTRs and average coverages in 50 samples. These 45 species were from 15 genera of five phyla including Actinobacteria, Bacterioidetes, Firmicutes, Proteobacteria and Spirochaetes (Supplementary Figs. 6–8) and each genus included three species with an average nucleotide identity (ANI) ranging from 66.6% to 91.2% (Methods). The coassembly and binning pipeline generated a total of 41 contig clusters with different completeness (48.4–100%) and contamination (0–81.6%) and each was dominated by one simulated species. DEMIC successfully estimated almost all growth rates of these 41 species (1,220 out of 1,222; Fig. 3a) without estimating any spurious rate for species that were absent from a sample. Moreover, the mean of correlations between DEMIC estimates and the true PTRs ($r = 0.992$) achieved a similar level to those from PTRC ($r = 0.995$) based on complete genomes, and greatly outperformed iRep ($r = 0.888$) (Fig. 3b, Supplementary Fig. 9).

Phylogenetically related species affect assembly and binning qualities because of their similar genome sequences; this not only resulted in failure to bin four species in the above simulated datasets, but also led to the mixture of their contigs into clusters dominated by other species. We evaluated the effects of these related species on the performance of DEMIC by comparing results in different ANI groups. As shown in Fig. 3c,d, no significant change was observed between any two of the three ANI groups of species ($P > 0.1$ for all comparisons). By contrast, outputs by iRep was markedly affected by increased ANI ($P < 0.001$ and $P < 0.05$ for comparisons of two neighboring groups). For example, because the species *Paenibacillus polymyxa* and *Paenibacillus terrae* shared a high ANI (87.4%), the binning algorithm generated a mixed contig cluster, which had *P. polymyxa* as the dominant species (53.1%), but that also contained contigs from *P. terrae*. With such a high proportion of contamination, iRep failed to generate any accurate outputs, resulting in estimates that were inconsistent with either of the two species ($r < 0.3$, Supplementary Fig. 10a,b). However, by iteratively filtering contigs according to the distribution of their first principal component (PC1) of the stepwise PCA (Fig. 1d, Methods), DEMIC successfully improved the contig cluster (99.7% from *P. polymyxa*, Supplementary Fig. 10c) and thus accurately estimated the growth dynamics of *P. polymyxa* ($r = 0.994$, Supplementary Fig. 10d).

To compare PTRs, DEMIC and iRep using real metagenomic data, we analyzed the sequencing datasets from seawater samples of eight Red Sea stations[15] and fecal samples of 26 healthy subjects[16]. PTRs could be calculated using PTRC for 7 and 34 bacterial species with complete reference genomes for the two datasets, respectively (Supplementary Fig. 11a). By constrast, DEMIC could effectively estimate growth dynamics for 34 and 110 species with contig clusters, compared to estimations for 8 and 57 species using iRep, respectively, indicating that DEMIC can quantify growth dynamics of a larger set of bacteria. DEMIC outperformed the other two methods by 133% to 437% (Supplementary Fig. 11b) for the number of estimated growth rates. As DEMIC and iRep have the same input requirements, we also compared the computational resources that were needed. When using eight threads, DEMIC completed its estimation for these datasets (90 and 77 billion bases) in about 2 h using 10 Gb random-access memory (RAM), about one-fifth of the time and one-third of the RAM needed by iRep (Supplementary

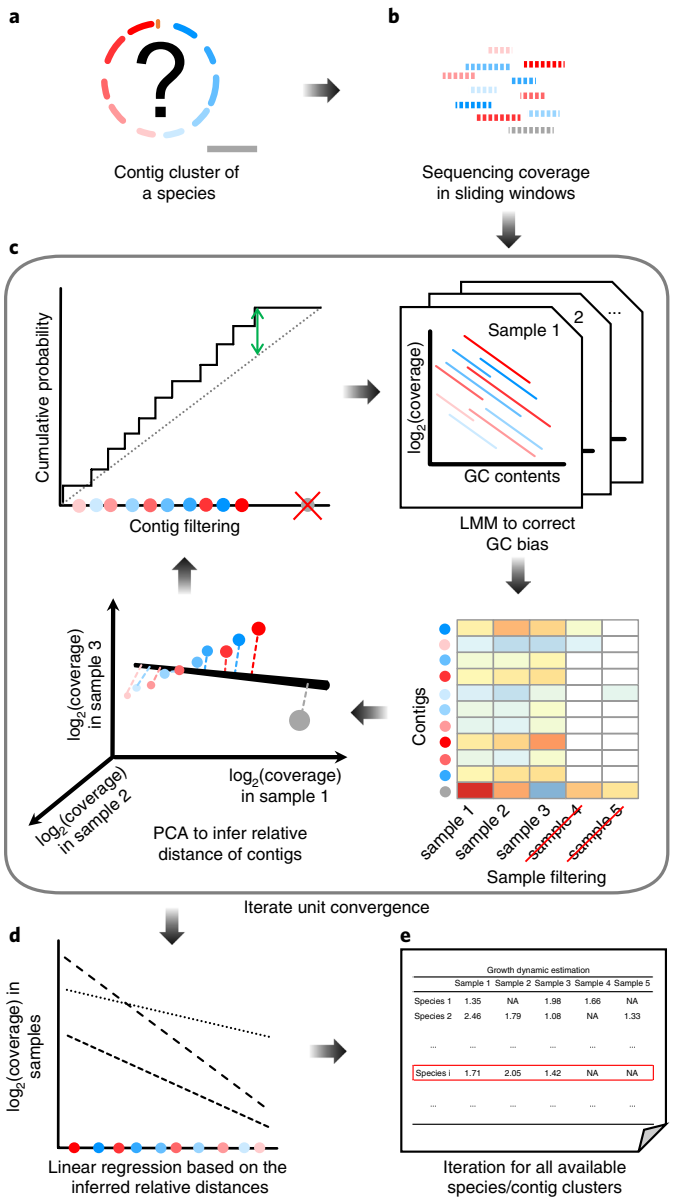

Fig. 1 | Computational pipeline of DEMIC. a, In a contig cluster identified by the binning algorithm, the genomic locations and potential contamination of contigs represented by different colors are unknown. **b**, The coverage (sequencing depth) of contigs in a cluster was first calculated for all sliding windows. **c**, The inference is an iterative process that includes GC bias correction using linear mixed-effects models (LMM), identification of informative samples, relative distance inference using PCA and filtering of contaminated contigs. Colored dots represent different contigs in the cluster. **d**, After convergence of both sample and contig sets, growth rates are estimated for the informative samples. Dashed lines represent linear regressions of log-transformed coverages of contigs in different samples and their inferred relative distances to the replication origin. **e**, The same pipeline is applied to each of the contig clusters identified by the binning algorithm. NA represents samples that are not informative for estimating growth rate of a cluster.

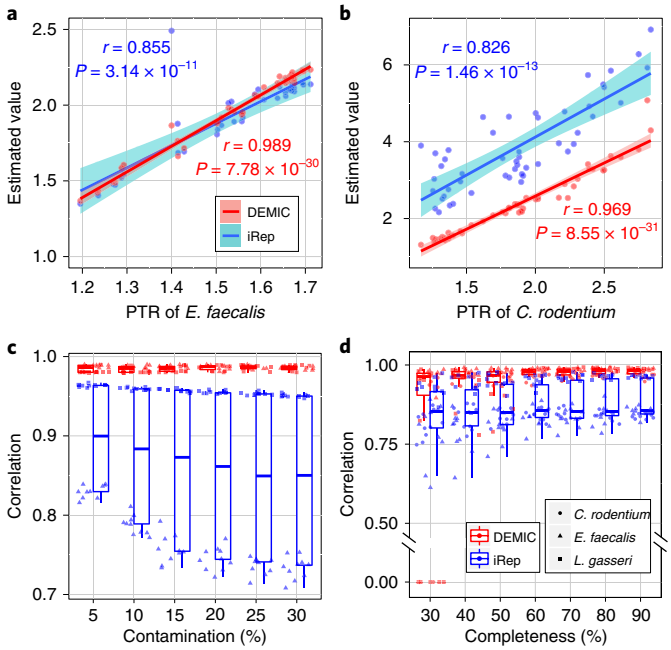

Fig. 2 | Performance evaluation of DEMIC based on sequencing datasets of three species. a,**b**, Correlations of estimates from DEMIC and iRep with PTR values (Pearson's *r* value) in 36 datasets of *E. faecalis* (**a**) and 50 datasets of *C. rodentium* (**b**). The shaded areas indicate 99% confidence intervals. **c**,**d**, Evaluation of the effects of contig contamination (**c**) and completeness (**d**) of the contig cluster on the performances of DEMIC and iRep. Evaluations of the sample size and contig cluster completeness were based on *L. gasseri*, *E. faecalis* and *C. rodentium* ($n=10$ for each) and evaluation of contig contaminations was based on *L. gasseri* and *E. faecalis* ($n=10$ for each). Correlations of all evaluations were plotted, and the box plots indicate the median (center line), first and third quartiles (box edges) and 1.5 times the interquartile range (whiskers).

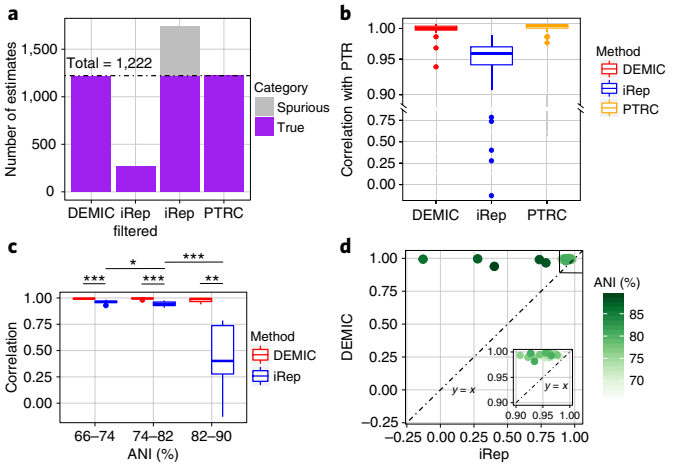

Fig. 3 | Performance evaluation of DEMIC based on simulated data of 45 closely related species from five phyla. a, Number of estimates of simulated PTRs. **b**, Correlation between DEMIC estimates for the 41 contig clusters and PTRs (Pearson's *r* value). **c**, Estimation accuracy of iRep and DEMIC for different ANI groups of species ($n=19$, 17 and 5 for percentage ANI group 66–74, 74–82 and 82–90, respectively). Two-sided Mann–Whitney U-tests, $^{*}P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$. **b**,**c**, The box plots indicate the median (center line), first and third quartiles (box edges) and 1.5 times the interquartile range (whiskers). **d**, Correlations (Pearson's *r* value) between DEMIC and iREP estimates and PTRs for all 41 species. The inset shows species that had a correlation with the corresponding PTR that was greater than 0.9 for both methods.

Fig. 11c,d). When using different binning algorithms, we observed similar results (Supplementary Fig. 12).

Depth-dependent gradients of physicochemical properties explain the most variation in microbial compositions in Red Sea[15]. Using the estimates from DEMIC, we observed a strong association between bacterial growth dynamics and sea depth (Supplementary Fig. 13a). For example, DEMIC estimated growth rates for a contig cluster, with about 60% completeness and an average identity of 92% to *Marinobacter adhaerens* in 22 seawater samples from 7 stations. At a depth of 500 m, the estimated growth rates were between 1.06 and 1.15 for all stations, significantly lower than those for 10-m and 100-m estimates, which ranged from 1.37 to 1.92 ($P<0.005$; Supplementary Fig. 13b,c).

When applied to metagenomic datasets of fecal samples of 26 healthy and 86 children with Crohn's disease[16], DEMIC estimated the growth dynamics of 278 species with contig clusters that had a wide range of completeness and contamination, of which more than 20% were estimated in 50 samples or more (Supplementary Fig. 14a,b). The high sensitivity of DEMIC made it possible to compare growth dynamics among different groups. For example, we found six (and one) species (Supplementary Table 3) that had significantly higher (and lower) growth rates in healthy subjects compared to patients with Crohn's disease. Notably, after treatment with an anti-TNF antibody or with an enteral diet for 1–8 weeks, the corresponding growth dynamics of three out of the seven species indicated above from subjects with Crohn's disease showed a significant shift toward the healthy subjects (Supplementary Fig. 14c; $P<0.05$ after false-discovery rate (FDR) correction).

Shotgun metagenomic sequencing data offer new insights into bacterial growth dynamics in microbiome studies. We present DEMIC, which effectively utilizes the data from multiple samples of each species in order to infer relative distances of contigs to the replication origin. Closely related organisms are one of the main factors that affect the completeness and contamination of a metagenomics pipeline, including assembly and binning. DEMIC uses a stepwise filtering strategy to iteratively update contig clusters, which provides an effective way of removing the high proportion of contig contamination (Supplementary Fig. 10). Owing to a substantially lower fraction of the genomes that is recovered by assembly and binning methods for different strains of the same species[17,18], DEMIC, like other available PTR estimation methods, is currently not able to provide estimates of growth dynamics at the level of a strain. As continuous efforts are being made to improve assembly, binning and other related methods[19,20], we expect that DEMIC may eventually be extended to the level of strains.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41592-018-0182-0.

Received: 20 March 2018; Accepted: 18 August 2018;
Published online: 12 November 2018

References

1. Myhrvold, C., Kotula, J. W., Hicks, W. M., Conway, N. J. & Silver, P. A. *Nat. Commun.* **6**, 10039 (2015).
2. Helaine, S. et al. *Proc. Natl Acad. Sci. USA* **107**, 3746–3751 (2010).
3. Claudi, B. et al. *Cell* **158**, 722–733 (2014).
4. Abel, S. et al. *Nat. Methods* **12**, 223–226 (2015).
5. Korem, T. et al. *Science* **349**, 1101–1106 (2015).
6. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. *Nat. Biotechnol.* **34**, 1256–1263 (2016).

7. Breitwieser, F. P., Lu, J. & Salzberg, S. L. *Brief. Bioinform.* https://doi.org/10.1093/bib/bbx120 (2017).
8. Alneberg, J. et al. *Nat. Methods* **11**, 1144–1146 (2014).
9. Albertsen, M. et al. *Nat. Biotechnol.* **31**, 533–538 (2013).
10. Rearick, D. et al. *Nucleic Acids Res.* **39**, 2357–2366 (2011).
11. Wu, Y. W., Tang, Y. H., Tringe, S. G., Simmons, B. A. & Singer, S. W. *Microbiome* **2**, 26 (2014).
12. Wu, Y. W., Simmons, B. A. & Singer, S. W. *Bioinformatics* **32**, 605–607 (2016).
13. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. *Bioinformatics* **31**, 1674–1676 (2015).
14. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. *Genome Res.* **25**, 1043–1055 (2015).
15. Thompson, L. R. et al. *ISME J.* **11**, 138–151 (2017).
16. Lewis, J. D. et al. *Cell Host Microbe* **18**, 489–500 (2015).
17. Sangwan, N., Xia, F. & Gilbert, J. A. *Microbiome* **4**, 8 (2016).
18. Sczyrba, A. et al. *Nat. Methods* **14**, 1063–1071 (2017).
19. Luo, C. et al. *Nat. Biotechnol.* **33**, 1045–1052 (2015).
20. Beaulaurier, J. et al. *Nat. Biotechnol.* **36**, 61–69 (2018).

Acknowledgements

This research was supported by grant R01GM123056 (H.L.) from the National Institutes of Health.

Author contributions

H.L. and Y.G. conceived and designed the project. Y.G. implemented the method. Both authors analyzed the data, and wrote and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0182-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to H.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

DEMIC implementation. This algorithm was implemented in Perl and R, and has been extensively tested on Linux and Mac OS X. No dependency is needed to run DEMIC except two non-default packages, lme4[21] and FactoMineR[22], in R. Multithreading is available to process both multiple metagenomic samples and multiple contig clusters in large datasets.

Calculation of contig coverage for sliding windows. DEMIC is designed to process sorted alignments of metagenomic shotgun sequencing reads against assembled contigs in SAM format (Fig. 1a). To estimate growth dynamics for a species, sequencing coverage values are first calculated from the read alignments of each sample, for all sliding windows of the same size within contigs (Fig. 1b). Thresholds for mapping length (≥50 bp by default), mapping quality (≥5 by default) and mismatch ratio (≤0.03 by default) are used during the following process to filter out spurious or ambiguous alignments.

Specifically, reads that are aligned to the jth contig with length $l_j \geq l' + p + 2l_r$ are used for coverage calculation using sliding windows, where p is the sliding window step size (100 bp by default), l' is the window size (5,000 bp by default, an integer multiple of p) and l_r is the read length that is excluded from each side of the contig. The total steps within a window is $q = l'/p$. For the ith sample, the average coverage of the kth window Y_{ijk} is calculated as

$$Y_{ijk} = \frac{T_{ijk-1} + T'_{ijkq} - T'_{ij(k-1)1}}{l'}$$

where T_{ijk-1} represents the total base coverage for the previous $(k-1)$th window, $T'_{ij(k-1)1}$ represents the total base coverage for the first p bases of the previous $(k-1)$th window, and T'_{ijkq} represents the total read coverage of the last p bases of the current kth window. Using this calculation, the average coverage of all sliding windows in a contig except the first one can be efficiently calculated while the sorted alignments of a sample are being scanned, avoiding repetitively counting the aligned reads for the bases that are in the previous sliding windows. As another filter step to remove chimeric contigs, only contigs with a coverage larger than 0 in all sliding windows were kept for a sample, and these coverage values were log-transformed for the subsequent analyses.

A linear mixed-effects model for the correction of sequencing bias. GC content has been reported to result in bias in next-generation sequencing platforms such as Illumina[23]. To detect and eliminate such biases, GC content for each window was first calculated during the scanning of contig sequences, using the same pattern as described above for the coverage calculation. For a given contig cluster, a linear mixed-effects model was then fitted to the coverage values calculated above with GC contents as the fixed effect and a sample- and contig-specific random intercept. Specifically,

$$\log_2 Y_{ijk} = a_0 + (X_{jk} - \overline{X})a + Z_{ij} + e_{ijk}$$

where Y_{ijk} is the average sequencing coverage of the kth window of the jth contig of the ith sample, a_0 is the intercept, X_{jk} is the GC content of the jth window of the kth contig, $\overline{X}$ is the average GC content of all the contigs, a is the regression coefficient, Z_{ij} is the sample- and contig-specific random intercept and e_{ijk} is the random error. This model was fitted for each contig cluster to estimate the intercept, the fixed effects a of the GC content and the random effects Z_{ij} for contig j and sample i using the best linear unbiased predictor. The resulting best linear unbiased predictor of Z_{ij}, denoted by $\hat{z}_{ij}$, corrects the average coverage of contig according to the difference in GC content of the average GC content of all contigs and therefore eliminates sequencing bias. We define $Y'_{ij} = \hat{a}_0 + \hat{z}_{ij}$ as the GC-adjusted log-transformed coverage of sample i and contig j and Y' as the final log-transformed coverage matrix, where $\hat{a}_0$ is the estimate of a_0.

Estimation of growth dynamics based on multiple samples. For the accurate inference of the relative distances between contigs and the replication origin, samples with low coverage of the given species were excluded from the following steps. Specifically, because the majority of contigs in each cluster are expected to be from the same species, an informative sample should have coverage for more than half of the contigs. Samples with an average of coverages lower than 0 for all contigs were also excluded from this step because of their relatively large variation. If two or more informative samples achieved the above thresholds, a preliminary filtering of the contigs was then used to remove contigs with no coverage in any of the informative samples.

To infer the relative distance of contigs from the replication origin, a dimension reduction method was applied to the log-transformed coverage matrix (Y') of the informative samples and contigs. Suppose that the log-transformed coverage matrix has a dimension of $N_s \times N_c$, where N_c and N_s represent the number of contigs and informative samples, respectively. A PCA was performed to reduce the dimension to $1 \times N_c$ so that PC1 accounts for the largest contribution to the variability of coverages among the N_c contigs across all N_s samples. This variability across different contigs is expected to result from different relative distances of the contigs to the replication origin. PC1 values of the N_c contigs, denoted as a vector $\mathbf{U}$, are therefore expected to be highly correlated with the contig locations relative to the replication origin. We then sorted the N_c contigs and determined their relative distances on the basis of their PC1 values. These sorted values were used to estimate the PTR in the next step.

The contig group needed further filtering to make sure that the PCA was not affected by the contigs from other species. Specifically, the assembled contigs were expected to be evenly distributed along a bacterial genome, and such a uniform distribution would be distorted if a few contigs from the other species were mixed into the group. Therefore, the distribution of PC1 of all contigs $\mathbf{U}$ was examined against the putative uniform distribution, unif(min(U), max(U)), by a Kolmogorov–Smirnov test. If a difference was found between the current distribution and the uniform distribution at a significance level of $P = 0.05$, the two contigs with maximum and minimum PC1 values were compared with respect to their distance from the adjacent contig, and the one with a larger distance was regarded as the contamination and removed in this step (Fig. 1c).

All of the remaining contigs were then used to fit an ordinary linear regression model for each sample (Fig. 1d). Specifically, for the ith sample, we fitted the following linear regression

$$Y'_{ij} = b_{0i} + b_i U_j,\ j = 1, 2, \ldots, N_c$$

where b_{0i} and b_i were the intercept and slope parameters. Let $\hat{b}_{0i}$ and $\hat{b}_i$ be the least-squares estimates of the coefficients. From these models, the growth dynamics of the species in these samples were calculated as the ratio of the exponential of model-fitted coverages of the two contigs with the maximum $(U_{(N_c)})$ and minimum $(U_{(1)})$ values of the PC1. We called this quantity the estimated PTR. Specifically, for the ith sample, its estimated PTR was defined as

$$\frac{\exp(\hat{b}_{0i} + \hat{b}_i U_{(N_c)})}{\exp(\hat{b}_{0i} + \hat{b}_i U_{(1)})},\ i = 1, 2, \ldots, N_s$$

Iteration and random strategies. For the implementation of DEMIC, several iteration and random strategies were adopted to ensure robustness of the pipeline before the final estimation of PTR. First, the four steps in the previous sections were repeated until convergence (Fig. 1c), including GC bias correction based on linear mixed-effects models, identification of informative samples, relative distance inference based on PCA and filtering of contigs. Both sets of contigs and samples were required to be the same between the current and the last iteration to achieve convergence of the four steps, which is designed to avoid a potential influence of less informative samples or contig contamination on a linear mixed-effects model and PCA. Second, to eliminate potential local optimum of the iteration steps, one can test the consistency between two different subsets of the contigs. In brief, after calculation of coverages for contigs within the sliding windows, two subsets were randomly selected so that each of the sets contained the same fraction (80% by default) of the total contigs and their union represented the total contigs. Each subset was independently used for relative distance inference using the four steps described above, and their consistency with each other was tested by correlation of linear regression slopes (b) in all remaining samples. Third, these linear regression slopes were used to estimate growth dynamics only when the correlation was above the designated threshold (0.98 by default), otherwise another two subsets were randomly selected and the above steps were iterated.

Datasets. Different types of datasets were downloaded or generated to evaluate the performance of DEMIC. We first used a synthetic dataset composed of 141 real sequencing datasets generated in a previous study[5]. The sequencing datasets were downloaded from the European Nucleotide Archive (accession number PRJEB9718) with the corresponding metadata, and each of them was from *L. gasseri* (ERR969426–ERR969461), *E. faecalis* (ERR969335–ERR969370), *C. rodentium* (ERR930224–ERR930295, ERR969253–ERR969278) and *E. coli* (ERR969315–ERR969334), which were grown separately in vitro. The synthetic dataset contained 50 simulated samples, and each sample was set to randomly contain 2–4 of the above sequencing datasets from different species in order to mimic the composition of microbiota (Supplementary Fig. 2). The synthetic dataset contained 6.1 billion base pairs in total, and each species present in a sample had a sequencing depth ranging from 0.17- to 96-fold.

A simulated sequencing dataset was next generated to test the effects of phylogenetically related species on the performance. A list of species with RefSeq ID, taxonomy and replication origin recorded in a previous study[24] was downloaded. A total of 15 genera in the list with at least three species in each were randomly selected. Reference genome sequences of randomly selected sets of three species in each genus were downloaded from NCBI to generate sequencing reads. According to the replication origin and genome size, for a given randomly assigned PTR (<3), we first generated read coverages along the genome based on an exponential distribution. A function of accumulative distribution of read coverages along the genome was then calculated. Sequencing reads were next generated one by one using the above accumulative distribution function and a random number to determine the location of each read on the genome, until the total read number

achieved a randomly assigned average coverage (between 0.5- and 10-fold) for the species in a sample. Sequencing errors, including substitutions, insertions and deletions, were simulated in a position- and nucleotide-specific pattern according to a recent study on metagenomic sequencing error profiles of Illumina[25]. The generated dataset contained 45 species from 15 genera of 5 different phyla (Supplementary Fig. 6, generated by iTOL[26]) and the ANI between species within each genus ranged from 66.6% to 91.2% according to Integrated Microbial Genomes and Microbiomes[27]. The probability of one species existing in each of the 50 simulated samples was set to 0.6, and a total of 1,336 average coverages and the corresponding PTRs were randomly and independently assigned (Supplementary Figs. 7 and 8). The final simulated sequencing dataset was about 20 billion bases.

The PLEASE dataset[16] included sequencing data from the fecal samples of 26 healthy children and 86 children with Crohn's disease. Samples from healthy children were sequenced for one time point, whereas samples from the patients with Crohn's disease were sequenced at four time points including baseline, one week, four weeks and eight weeks after treatment with anti-TNF antibodies or enteral diet treatment. The reads were downloaded from the NCBI Sequence Read Archive (SRP057027) with the corresponding metadata. We used the subset of 26 healthy subjects (77 billion bases) to compare the ability of DEMIC, PTRC and iRep to estimate bacterial growth dynamics and used the whole datasets (859 billion bases) to compare estimates of growth dynamics of the same species in different samples using DEMIC.

The Red Sea dataset[15] included 45 metagenomic samples of seawater sampled from different depths at eight stations in the Red Sea. The reads were downloaded from the NCBI Sequence Read Archive (SRP061183) with the corresponding metadata. We used the whole datasets (90 billion bases) to compare the ability of DEMIC, PTRC and iRep to estimate bacterial growth dynamics and to compare estimates of growth dynamics of the same species in different samples using DEMIC.

Coassembly, binning and mapping. For both synthetic and real datasets, coassembly was performed to facilitate binning as well as analysis of DEMIC and iRep. MEGAHIT[13] version 1.1.1 was used as the assembler because of its advantages for generating longer total assembly length[18] and because of its controllable memory usage, which is convenient for large metagenomic datasets. The default settings of MEGAHIT were used for all of the datasets.

After coassembly, contigs were clustered into groups using binning algorithms. MaxBin[12] version 2.2.4 was used to cluster contigs in the synthetic datasets, simulated datasets, all RedSea and 26 healthy PLEASE datasets because of its outstanding performances in medium- and low-complexity datasets[18]. MetaBAT[28] version 2.12.1 was used for binning of the RedSea and PLEASE datasets because of its overall performances and high speed when processing high-complexity datasets. CheckM[14] was used to assess the contig completeness and contamination of the contig clusters using the default settings.

For all of the datasets above, Bowtie 2[29] version 2.3.2 was used to align reads back to assembled contigs. The output alignment results were then sorted by samtools[30] version 0.1.19 and used as input for both DEMIC and iRep.

Evaluation based on the synthetic datasets and random tests. After coassembly and binning of the constructed contigs, contigs from three species were successfully clustered, including *L. gasseri*, *E. faecalis* and *C. rodentium*. Neither MaxBin nor MetaBAT generated a contig cluster corresponding to *E. coli*, because of its relatively low-sequencing depth compared to *C. rodentium* in the same family. The following evaluations were therefore based on contig clusters of *L. gasseri*, *E. faecalis* and *C. rodentium*. Bacterial growth rates in the synthetic datasets were first estimated by PTRC, DEMIC and iRep using the respective default settings. For a total of 122 growth rates of the three species (36, 36 and 50, respectively), correlations between PTRC and DEMIC as well as between PTRC and iRep were calculated using Pearson's *r* correlations.

To generalize our evaluation to diverse metagenomic datasets, three different types of random tests were performed to test the effects of sample counts, fraction of contig contamination and completeness of contig clusters on the performance. Specifically, groups of 3, 6, 10, 15, 20 and 25 samples, groups with 5%, 10%, 15%, 20%, 25% and 30% of contig contamination and groups of 30%, 40%, 50%, 60%, 70%, 80% and 90% completeness of contig clusters were considered. For each random test, DEMIC was applied to the selected subset of samples or contig clusters that were randomly generated according to a given fraction, so that these random tests in the same group were independent of each other.

After coassembly and binning for the simulated dataset of 45 species in 50 samples, contigs from 41 species were successfully clustered. Four species failed to be binned into separate clusters as dominant species, including *Caldicellulosiruptor lactoaceticus*, *P. terrae*, *Xanthomonas axonopodis* and *Xanthomonas oryzae*. The subsequent evaluations were therefore all based on contig clusters of the 41 clusters and the corresponding 1,222 PTRs (Supplementary Fig. 9). A window size of 3,000 and a mismatch threshold of 0.02 were used in DEMIC with all other settings as default. PTRC were provided with complete reference genomes, and the default settings were used for both PTRC and iRep.

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

Code availability. Source codes are freely available at https://sourceforge.net/projects/demic/ under the GNU General Public License.

Data availability

The accession numbers and weblinks for all real datasets are provided in the Methods. Simulated data are available upon request from the corresponding author.

References

21. Bates, D., Mächler, M., Bolker, B. M. & Walker, S. C. *J. Stat. Softw.* **67**, 1–48 (2015).
22. Lê, S., Josse, J. & Husson, F. *J. Stat. Softw.* **25**, 1–18 (2008).
23. Ross, M. G. et al. *Genome. Biol.* **14**, R51 (2013).
24. Gao, F., Luo, H. & Zhang, C. T. *Nucleic Acids Res.* **41**, D90–D93 (2013).
25. Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N. & Quince, C. *BMC Bioinformatics* **17**, 125 (2016).
26. Letunic, I. & Bork, P. *Nucleic Acids Res.* **44**, W242–W245 (2016).
27. Markowitz, V. M. et al. *Nucleic Acids Res.* **40**, D115–D122 (2012).
28. Kang, D. D., Froula, J., Egan, R. & Wang, Z. *PeerJ* **3**, e1165 (2015).
29. Langmead, B. & Salzberg, S. L. *Nat. Methods* **9**, 357–359 (2012).
30. Li, H. et al. *Bioinformatics* **25**, 2078–2079 (2009).

nature research

Corresponding author(s): Hongzhe Li

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	The paper uses both public data sets and simulated data sets. Links to access the public data are provided. DEMIC_simulator was used for generating the simulated data sets, and made available in a repository (see code availability).
Data analysis	The following software were used for data analysis: Bowtie 2 v2.3.2, samtools v0.1.19, MEGAHIT v1.1.1, MaxBin v2.2.4, MetaBAT v2.12.1, CheckM v1.0.7, PTRC v1.1 and iRep v1.10. Source code of DEMIC is available in a repository (see code availability).

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

The accession numbers and weblinks for all real data sets are provided in Methods. Simulated data are available on request from the corresponding author.

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

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

Sample size	This is a methodological paper reporting a computational method for estimating bacterial growth rates using shotgun metagenomic sequencing data. We used a synthetic dataset generated from 36, 36 and 50 sequencing data sets of Lactobacillus gasseri, Enterococcus faecalis and Citrobacter rodentium to evaluate the performances of our method (all available data sets of L. gasseri, E. faecalis generated previously were used). In the random tests, we also showed our method has stable performance when six or more samples were available. In the simulation tests, we generated 50 samples for totally 45 species from 15 genera of five phyla. For the real public data sets, all samples were used.
Data exclusions	No data were excluded.
Replication	The methods was evaluated for different groups of contig contamination, completeness and sample counts (See Randomization) by 10 times in each group. Finally, DEMIC was able to accurately estimate growth dynamics when 40% of contigs or more and when six samples or more were provided in these evaluations. The accuracy of DEMIC was not affected by contamination in the evaluations.
Randomization	To generalize our evaluation to diverse metagenomic data sets, three different types of random tests were performed to evaluate the effects of sample counts, fraction of contig contaminations and completeness of contig clusters on the performance. Specifically, groups of 3, 6, 10, 15, 20, 25 samples, groups with 5%, 10%, 15%, 20%, 25% and 30% of contig contaminations and groups of 30%, 40%, 50%, 60%, 70%, 80% and 90% completeness of contig clusters were considered. In the simulation tests, we randomly selected 15 genera with three species in each and generated a total of 50 samples for the 45 species. Average coverages and PTRs were randomly assigned for each species in each sample independently.
Blinding	Not relevant because this is a methodological study.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Research animals
☒	☐ Human research participants

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging