Probabilistic model based on circular statistics for quantifying coverage depth dynamics originating from DNA replication

Shinya Suzuki, Takuji Yamada

Department of Life Science and Technology, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro, Tokyo 152-8550, Japan

Corresponding Author:
Takuji Yamada
2-12-1 Ookayama, Meguro, Tokyo 152-8550, Japan
Email address: takuji@bio.titech.ac.jp

Abstract

Background. With the development of DNA sequencing technology, static omics profiling in microbial communities, such as taxonomic and functional gene composition determination, has become possible. Additionally, the recently proposed *in situ* growth rate estimation method allows the applicable range of current comparative metagenomics to be extended to dynamic profiling. However, with this method, the applicable target range is presently limited. Furthermore, the characteristics of coverage depth during replication have not been sufficiently investigated.

Results. We developed a probabilistic model that mimics coverage depth dynamics. This statistical model explains the bias that occurs in the coverage depth due to DNA replication and errors that arise from coverage depth observation. Although our method requires a complete genome sequence, it involves a stable to low coverage depth ($>0.01\times$). We also evaluated the estimation using real whole-genome sequence datasets and reproduced the growth dynamics observed in previous studies. By utilizing a circular distribution in the model, our method facilitates the quantification of unmeasured coverage depth features, including peakedness, skewness, and degree of density, around the replication origin. When we applied the model to time-series culture samples, the skewness parameter, which indicates the asymmetry, was stable over time; however, the peakedness and degree of density parameters, which indicate the concentration level at the replication origin, changed dynamically. Furthermore, we demonstrated the activity measurement of multiple replication origins in a single chromosome.

Conclusions. We devised a novel framework for quantifying coverage depth dynamics. Our study is expected to serve as a basis for replication activity estimation from a broader perspective using the statistical model.

Introduction

The development of high-throughput DNA sequencers has enabled massive and exhaustive microbiome analyses. By mapping fragmented reads onto databases, the taxonomic and functional gene composition of a sample can be measured. Several researchers have utilized this procedure to investigate samples from various environments, such as those of human and animal bodies as well as those of other types of environmental samples (Hildebrand et al., 2013; Kato et al., 2015; Zhu et al., 2015; Higashi et al., 2018). One possible means to progressing beyond the profiling of static information involves investigating microbial dynamics. Although time-series microbiome profiling via quantitative polymerase chain reaction (PCR) or cell-sorting may allow dynamics measurement, such methods do not easily provide a comprehensive understanding of growth dynamics from the single sample involved therein (Tourlousse et al., 2017; Vandeputte et al., 2017). Meanwhile, the peak-to-trough ratio (PTR) of the coverage of whole genome sequencing (WGS) reads mapped to a reference genome sequence provides an estimate of growth; notably, this approach uses WGS reads from just a single sample (Korem et al., 2015).

79   This method is based on the considerable increase in DNA around the replication origin via
80   bidirectional DNA replication (Cooper & Helmstetter, 1968; Bremer & Churchward, 1977). As
81   quantitative pipelines continuously undergo re-evaluation and extension, a draft quality genome
82   sequence may be applied. Few methods have been proposed to quantify the growth of bacteria
83   from genomic data. iRep uses a mechanism in which the slope of the sorted coverage on contig
84   sequences was correlated with the growth rate (Brown et al., 2016). GRiD enables more robust
85   estimation by sorting the coverage depths of multiple contigs (Emiola & Oh, 2018). DEMIC
86   performed accurate estimation by using the coverage depths of multiple samples and estimating
87   the appropriate position via principal component analysis (Gao & Li, 2018). Some studies using
88   such pipelines have revealed associations between growth estimates and factors such as disease,
89   24-hour oscillations, and diet (Olm et al., 2017; Forsyth et al., 2018). Thus, such an approach
90   quantifying the growth of bacteria from coverage depth is expected to facilitate the investigation
91   of new fields of microbial research. However, some questions associated with coverage depth
92   modeling remain unresolved.
93   The first challenge regarding growth estimation from coverage depth is related to the
94   application scope of the method. Previous studies have enabled growth estimation for a broad
95   range of samples, but the range of applicability remains limited. Taking coverage depth as an
96   example, even the most robust method currently requires 0.05× average coverage with a
97   complete sequence or 0.2× with a *de novo*-assembled sequence. A novel growth rate estimation
98   method that is less sensitive to decreases in coverage depth could be utilized in broader
99   applications. Second, the previously proposed pipelines are not applicable to microbes with
100 multiple replication origins as these pipelines use a model based on a single peak and trough.
101 This approach narrows the range of measurement targets as some taxa such as archaea have two
102 or more replication origins in a single replicon (Lundgren et al., 2004; Robinson et al., 2004;
103 Andersson et al., 2010). It has also been suggested that some bacteria have multiple replication
104 origins (Gao, 2015; Ohbayashi et al., 2016). In addition to growth estimation based on coverage
105 depth, it is also difficult to predict replication origins from sequence features such as GC-skew in
106 some microbes (Gao & Zhang, 2008; Sernova & Gelfand, 2008; Vieira-Silva & Rocha, 2010).
107 To overcome this challenge, a previous report proposed a method for predicting the positions of
108 multiple replication origins based on the amount of chromosomal DNA (Xu et al., 2012).
109 However, no method with a statistical background has been introduced. Third, the characteristics
110 of coverage depth distributions themselves have not been sufficiently investigated. Some
111 previous studies have reported non-linear DNA quantity trends (Hawkins et al., 2013; Pelve et
112 al., 2013; Wu et al., 2014; Akiyama et al., 2016), such as those including a significant increase
113 around the replication origin. Based on these studies, it has been suggested that replication
114 affects not only the ratio of maximum to minimum depth, but also changes the degree of density
115 around the replication origin. Modeling this phenomenon could be useful for both molecular
116 biologists and microbiologists, enabling them to quantify the extra dynamics of replication.
117 Furthermore, it is unclear whether the coverage depth trend is skewed toward the 5' direction, is
118 skewed toward the 3' direction, or is symmetric. As some previous studies have suggested that

119   the asymmetry of replisome progression is associated with the phenotype (Rodriguez-Lopez et
120   al., 2002), it would be valuable to develop a method of symmetric level detection.
121   Here, we propose a method of modeling coverage depth dynamics using probabilistic
122   statistics. Focusing on data generation when mapping fragmented reads to a circular genome
123   sequence, we combined multinomial and directional distributions to mimic the read sampling
124   process and bias of the DNA quantity. When applied to a dataset from a culture experiment, our
125   method provided a stable and robust estimation of even a small number of reads and mutated
126   reference sequences. To observe the degree of correlation between the growth estimates and
127   experimental growth rates, we applied our method to WGS reads, which were obtained from a
128   previous time-series culture experiment (Korem et al., 2015); this led to the observation of a high
129   degree of correlation between the growth estimates and experimental growth rates. *In vivo* data
130   sets were used to confirm the reproducibility of the growth dynamics in previous studies. Using
131   the previous *in vitro* and *in vivo* samples, we ensured that our method is sufficiently robust to
132   coverage and noise. Furthermore, by extending these models to enable them to form tapered and
133   skewed coverage depth shapes, we designed a method of measuring coverage depth bias. Using a
134   mixture of directional distributions allows growth estimation to be applied to sequences with
135   multiple replication origins. We also demonstrate such estimations in relation to genome
136   sequences of *Sulfolobus solfataricus* and *Haloferax volcanii* (McCarthy et al., 2015).
137
138   ## Materials & Methods
139   **Circular distributions and statistics**
140   The distributions and statistics used in this study are shown in Table 1. The location parameter
141   has the highest probability and corresponds to the replication origin in the chromosome in this
142   model. We changed the character of the concentration parameter by changing the distribution, as
143   it can be aligned by $\rho_{c.} = \frac{\tanh(\kappa)}{2}$ or $\rho_{w.C.} = \tanh\left(\frac{\kappa}{2}\right)$ (Jones & Pewsey, 2005; Pewsey, Neuhäuser &
144   Ruxton, 2013). In addition to major distributions, we introduced a linear cardioid distribution and
145   exponential linear cardioid distribution to evaluate the coverage depth trend. These functions are
146   symmetric around the location parameter, and the integral around a unit circle is 1; i.e.,
147     $\int_{-\pi}^{\pi} P(\theta|\mu,\rho)d\theta = \int_{0}^{2\pi} P(\theta|\mu,\rho)d\theta = 1$. For each distribution, the probabilistic PTR can be
148   analytically defined as the ratio between the maximum and minimum value of the probability
149   density function (see the *Statistical model to estimate replication rate* section for details).
150   Some of the models (von Mises, cardioid, wrapped Cauchy, and Jones-Pewsey) were
151   symmetrically or asymmetrically extended with or without inverse transformation, as previously
152   described (Abe, Pewsey & Shimizu, 2013; Pewsey, Neuhäuser & Ruxton, 2013; Abe, Pewsey &
153   Fujisawa, 2013). To make the shape near the mode of a distribution variable, we used Batschelet
154   or inverse-Batschelet transformation. Batschelet transformation (symmetric extension; SE)
155   transforms the angular variable into $g_\lambda(\theta) = (\theta - \mu) + \lambda\sin(\theta - \mu)$, where λ is the peakedness
156   parameter. Using this transformation, the normalization constant was calculated using the
157   composite Simpson's law as the integral around a unit circle cannot be maintained as 1. Inverse-

158   Batschelet transformation (inverse symmetric extension; InvSE) transforms the angular variable
159   into $g_\lambda(\theta) = \frac{1-\lambda}{1+\lambda}\theta + \frac{2\lambda}{1+\lambda}t_{1,\lambda}^{-1}(\theta)$, where $t_{1,\lambda}(\theta) = \theta - \frac{1}{2}(1+\lambda)\sin(\theta - \mu)$. This transformation does
160   not change the normalization constant. To make the distribution asymmetric with mode-
161   invariance, we used the mode-invariance asymmetric transformation extension (MIAE) or
162   inverse-transformed mode-invariance asymmetric transformation extension (InvMIAE). These
163   transformations satisfy the requirement that asymmetricity be analyzed in replication. As
164   replication begins at the origin irrespective of the rapidity of bacterial growth, the highest
165   coverage depth position is preserved regardless of the asymmetry level. The skewness parameter
166   must not affect the pPTR when the pPTR and skewness are measured independently. In these
167   transformations, the symmetricity around the mode changes via an additional skewness
168   parameter, while the location parameter, concentration parameter, and pPTR remain unchanged
169   even if the skewness parameter changes. MIAE transforms the angular variable into $g_\nu(\theta) = \theta - \nu$
170   $\sin^2(\theta - \mu)$, where ν is the skewness parameter. This transformation also requires a normalization
171   constant. InvMIAE transforms the angular variable into $g_\nu(\theta) = s_\nu^{-1}(\theta)$, where $s_\nu(\theta) = \theta + \nu\sin^2$
172   $(\theta - \mu)$. This transformation does not change the normalization constant, and the position of the
173   mode is preserved. To compute the inverse transformation, we used several root-finding
174   algorithms (see *Parameter estimation* section). The significance of fitness improvement via
175   additional parameters was evaluated using the likelihood ratio test with a chi-squared distribution
176   based on the theorem of Wilks in addition to the Akaike Information Criterion (AIC) and the
177   Bayes Information Criterion (BIC). To compute the likelihood ratios, the original distribution
178   was compared with the extended distributions. The Jones-Pewsey distribution was compared
179   with the von Mises distribution.
180
181
182   **Statistical model to estimate replication rate**
183   A statistical model that simulates the coverage depth dynamics along the genome position was
184   constructed. Let $d_{i,s}$ be the coverage depth of the ith position obtained when mapping the WGS
185   reads of sample s onto the genome sequence. Here i represents the binned position of coverage.
186   If the coverage depth is not compressed or binned, i matches a nucleotide position. We fit a
187   generalized linear model (GLM) for each $d_{i,s}$ as follows.
188   The starting point of our model is the conversion of $d_{i,s}$ into the frequency of observation of the
189   ith position. Here, the total number of observations T_s for sample s is calculated as the sum of $d_{i,s}$
190   over sequence length, I, i.e., $T_s = \sum_{i=1}^{I} d_{i,s}$. We did not directly fit the coverage depth to a model
191   because it could fail to fit with a low coverage depth dataset. Instead, we modeled a bias to
192   observe the base positions by mapping reads with a probability distribution P and parameter set
193   ω, which defines the potential of observation probability, $i_{t,s} \sim P(\omega_s)$, where t is a unique identifier
194   of nucleotides for all reads mapped to the genome. Focusing on the genome structure of bacteria
195   and archaea, the observation probability is supposed to be circular. For compatibility with the
196   structure, the position i is converted into an angle θ, following $\theta = \frac{i}{I}2\pi$. Here, the coverage depths

197   d_i are stacking counts of the observed angle θ. Circular statistics, instead of an ordinal real-value
198   approach, are required to quantify the bias based θ. In circular statistics, the first possible means
199   of analyzing an angle dataset involves expressing the bias as a simple statistic without any
200   model. For example, the mean resultant length (MRL) represents how data are concentrated
201   around the sample mean direction. The second approach, which was mainly used in this study,
202   involves the modeling of a phenomenon via probability distributions that generate positions. We
203   introduced the following four distribution types from the circular distribution family: von Mises,
204   wrapped Cauchy, cardioid, and Jones-Pewsey distributions. These distributions are widely used
205   in circular statistics and are versatile in terms of implementation and inference (Jones & Pewsey,
206   2005; Pewsey, Neuhäuser & Ruxton, 2013). Additionally, these distributions are useful for
207   representing known coverage depth characteristics. For example, some researchers have
208   described non-linear coverage depth trends over genome sequences in both bacteria and archaea
209   (Chen et al., 2005; Watanabe et al., 2012; Hawkins et al., 2013; Pelve et al., 2013; Rudolph et al.,
210   2013; Wendel, Courcelle & Courcelle, 2014; Wu et al., 2014; Maduike et al., 2014; Yang et al.,
211   2015; Akiyama et al., 2016; Ohbayashi et al., 2016; Forsyth et al., 2018; Retkute et al., 2018). It
212   was expected that this trend could be quantified by extension for the distributions proposed in the
213   previous studies. For these circular distributions with likelihood $L_{dist.}$, the overall log-likelihood
214    $\log L$ of the model can be calculated as follows:

$$\log L(\omega|\theta,d,I,S) = \Sigma_{s=1}^{S}\Sigma_{i=1}^{I}d_{i,s}\log L_{dist.}(\omega|\theta_i). \qquad (1)$$

216   As θ is considered to be continuous rather than discrete for the purpose of these distributions, we
217   confirmed that the parameters could be estimated appropriately (Supplementary Text S1).
218   Equation (1) coincides with a term that changes with the probability parameter in the log-
219   likelihood equation of the multinomial distribution shown in equation (2) although the sum of the
220   probability over all nucleotides is not 1 because the distribution is not discrete:

$$\log \text{Multinomial}(d|T,p) = \Sigma_{s=1}^{S}\Sigma_{t=1}^{T_s}\log t - \Sigma_{s=1}^{S}\Sigma_{i=1}^{I}\Sigma_{j=1}^{d_{s,i}}\log j + \Sigma_{s=1}^{S}\Sigma_{i=1}^{I}d_{s,i}\log p_i. \quad (2)$$

222   Following the model with the likelihood represented by equation (1), the location
223   parameter corresponds to the position of the replication origin as long as the concerned
224   chromosome does not have multiple replication origins. Contrastingly, the concentration
225   parameter is associated with growth as it determines the shape of the distribution. Therefore, we
226   allowed the location parameter to be shared among all of the samples and the concentration
227   parameter for each sample to be independent. For the concentration parameters, we set the half-
228   Student's t-distribution as a prior distribution (Gelman, 2006). We set 2.5 as the shape
229   parameter; 0 as the location parameter; and 0.2 (von Mises, Jones-Pewsey), 0.1 (cardioid), 0.17
230   (wrapped Cauchy), 0.105 (linear cardioid), and 0.1103(exponential linear cardioid) as the scale
231   parameters. These were selected such that the value of the cumulative probability density
232   function became nearly 0.8 when the PTR was 2.0. This characteristic suggests that most of the
233   PTRs are distributed between 1.0 and 2.0 in an environment. The distribution of the coverage
234   depth PTR in a previous study rationalizes this suggestion (Korem et al., 2015). For the degree of

235   density of the Jones-Pewsey distribution, the peakedness of the symmetric extended distribution,
236   and the skewness parameter of the asymmetry extended distribution, we set the Gaussian
237   distribution with a location parameter of 0 and a scale parameter of 1.0 as the prior distribution to
238   avoid overfitting. From the model, we introduced an estimate that expresses the degree of
239   growth. It is known that many microbes in prokaryotes replicate their chromosomal DNA on
240   both sides from the origin such that the apparent amount of DNA increases near the origin. This
241   behavior introduces a latent bias that makes DNA near the origin more likely to be observed
242   during replication. This bias is simply expressed by the concentration parameter in a circular
243   distribution. However, for consistency with the previous study (Korem et al., 2015), we defined a
244   probabilistic PTR (pPTR), which is the ratio of the maximum of probability density function to
245   the minimum, i.e., $PTR_{\text{probability}} = \frac{p_{\max}(\theta)}{p_{\min}(\theta)}$, as a growth dynamics index. This score represents the
246   latent bias of the probability for the position at which a nucleotide is observed around the
247   replication origin. Unlike the original PTR, which is directly estimated from the coverage depth,
248   pPTR is obtained by modeling the bias based on a circular distribution and the probability
249   framework of interest. Following the model, in which the coverage depth is a result of discrete
250   sampling expressed as in equation (2), the coverage depth at a given position can be modeled by
251   the binomial distribution by formula (3):

$$d_{i,s} \sim \text{Binomial}\left(T_s, P\left(\theta = \tfrac{i}{I}2\pi \,|\, \omega_s\right)\right). \tag{3}$$

253   This equation hints at the benefit of using probability rather than coverage depth directly
254   (Supplementary Text S2; Supplementary Fig. 1).

255
256

257   **Extending the model to multiple origins of replication**
258   We constructed a statistical model for multiple origins of the replication-based mixture model
259   and circular distributions. Let α_m be the ratio of the mixture to the mth replication origin and M
260   be the number of the replication origin, then the probability of obtaining the angle θ from the
261   sample s is formulated to $P(\theta|\omega_s,\alpha) = \sum_{m=1}^{M} \alpha_m P_{circle}(\theta|\omega_{s,m})$ with the circular distribution P_{circle}.
262   Note that the sum of the ratio is 1, i.e., $\sum_{m=1}^{M} \alpha_m = 1$. Based on the model, the overall log-
263   likelihood $\log L$ can be calculated as follows:

$$\log L(\omega,a|\theta,d,I,S) \propto \sum_{s=1}^{S}\sum_{i=1}^{I} d_{i,s}\log(\sum_{m=1}^{M}\alpha_m L_{\text{dist.}}(\omega_{m,s}|\theta = \tfrac{i}{I}2\pi, s = s)). \tag{4}$$

265   The equation is compatible with a genome sequence with a single replication origin as it takes
266   the same form as equation (1) when we set $\alpha_1 = 1$ and $M = 1$. We set a Dirichlet distribution as
267   the prior distribution of α as $\alpha \sim \text{Dirichlet}(A)$ and employed $50/M$ as A, as previous studies have
268   implied that each replication origin shows similar activity (Robinson et al., 2004; Andersson et
269   al., 2010; Hawkins et al., 2013). Then, the ratio is likely to assume a similar value, which defines
270   the equality of the mixture. As the activity index for multiple replication origins, we defined a

271   weighted pPTR (wPTR) and mean-weighted pPTR (mwPTR). The wPTR of the mth replication
272   origin is computed via a weighted concentration parameter using a mixture ratio, where mwPTR
273   is the average of these. For example, the wPTR of the von Mises-based model is given by exp(2
274    $\alpha_m \kappa_m$), and mwPTR is given by $\frac{1}{M}\sum_{m=1}^{M} wPTR_m$. These scores are based on the model that
275   replication stops if the replisome comes across another replisome, as mainly reported in
276   prokaryotes (Leman & Noguchi, 2013; Wendel, Courcelle & Courcelle, 2014); however, this
277   model has not been sufficiently investigated in archaea or eukaryotes. This model assumes that
278   the effect of multiple origins at each location can be expressed as their sum. Following the mixed
279   effects of multiple origins, the coverage depth is probabilistically sampled. This assumption
280   results in the probability of regions that are not related to the origin approaching low values.
281   Hence, the probability distribution of each origin becomes very steep, with increases in the
282   concentration parameter and unweighted PTR. By weighting the parameters, we approximated
283   the degree of activity in the case in which only the single origin worked in the chromosome. We
284   used the average of the wPTR as the representative score of the chromosome because a previous
285   study reported that the growth rate decreased when a replication origin on the chromosome was
286   knocked out, whereas the deviation of the DNA amount between the origins and terminus did not
287   change significantly (Wu et al., 2014). By using the average, the effect of activation on multiple
288   origins could be determined.
289
290

Parameter estimation
292   For all data, we estimated the model parameters using an implemented software package. This
293   package fits the parameters to the data by maximizing the joint posterior (optimizing mode) or
294   generating samples from the posterior distribution of the parameters (sampling mode). Briefly,
295   after maximizing the log-likelihood of the model for the data via each method, we adopted the
296   value that yielded the maximum log-likelihood via the optimizing mode or the expected *a*
297   *posteriori* (EAP) of the parameter posterior distribution via the sampling mode. Unless otherwise
298   noted, we used the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm
299   for the optimizing mode and the No-U-Turn Sampler (NUTS) algorithm (Hoffman & Gelman,
300   2014), which is the quasi-Markov chain Monte Carlo (MCMC) method, for the sampling mode.
301   For sampling, we set the number of chains to 1 and the number of iterations to 500; the first 300
302   iterations were considered a warm-up and discarded. From the samples, we calculated the EAP
303   as a representative model parameter. The convergence of the sampling was checked using Rhat
304   statistics. If the statistics were equal to or less than 1.1, the result was accepted (Gelman et al.,
305   2013). We used composite Simpson's algorithm with 20 subintervals to calculate the integrals in
306   the models. We computed the inverse transformation of the functions using Newton's method
307   when the skewness ν or peakedness parameter λ was less than or greater than 0.8. These
308   thresholds were evaluated by performing manual simulations such that the transformations did
309   not oscillate. We checked the convergence of the function as to whether the error was less than
310   the machine precision of a float-type variable defined using Stan. The Illinois method was used

311   in other cases (Dowell & Jarratt, 1971). Then, we checked the convergence of the function, as to
312   whether the error was less than 1.0×10^{-13}. In both methods, we defined a maximum iteration
313   count that terminated the calculation at 30 iterations for Newton's method and 100 iterations for
314   the Illinois method. We set the mathematics transformation to reasonably estimate the
315   parameters. If a location parameter is estimated directly from 0 to 2π, and the true location
316   parameter is located near the edge of the range, the parameter estimation is likely to fail as it
317   does not detect cyclicity. We re-parameterized the location parameter as $\mu = \mathrm{arctan2}\left(\overrightarrow{\theta_\mu}\right)$ for the
318   two-dimensional unit vector $\overrightarrow{\theta_\mu}$ to overcome this estimation difficulty (Pewsey, Neuhäuser &
319   Ruxton, 2013). The unit vector can be estimated directly as each element has continuity from -1
320   to 1. To calculate information criteria, 1+number of the sample×2 was used as the number of
321   parameters for the Jones-Pewsey distribution-based model, and 1+number of the samples was
322   used for the others.

323
324

325   **Coverage depth calculation**
326   The coverage depth was calculated by aligning the WGS reads to the template genome sequence.
327   We downloaded WGS via the SRA Toolkit. After converting the WGS reads into the FASTQ
328   format, we aligned them to the genome sequence using Bowtie2 with a "--very-sensitive"
329   parameter set (Langmead & Salzberg, 2012). We sorted the resulting SAM files and calculated
330   the depth using SAMtools (Li et al., 2009). Next, a moving median filter with a 100 nt window
331   size and a 100 nt stride length was applied. The moving median filter runs through the coverage
332   depths, replacing each coverage with the median of neighboring locations. We applied the
333   moving median filter for the following two reasons: to reduce the noise and outliers and to
334   reduce the data size, which affects the computational time of the model fitting. If the coverage
335   depth seemed to have noise regions that increased the coverage due to highly conserved regions
336   such as ribosomal genes, an additional filtering for outliers was performed; specifically, the top
337   1% of the coverage depth was removed and replaced with blank coverage. This decision and the
338   threshold, which were determined based on a previous study (Brown et al., 2016), were
339   independently evaluated using a frequency histogram of the coverage depth containing a noise
340   region, which increases the coverage in multiple datasets (Supplementary Text S3). Certain
341   regions that remained blank following filtering were filled with 0. As the WGS reads of *H.*
342   *volcanii* were separated into multiple FASTQ files, we concatenated them into a single file based
343   on growth conditions prior to alignment. To evaluate the error in the sequence edge parts, we
344   copied 263 nt in the head portion of the genome sequence to the tail portion. We also constructed
345   a graph genome sequence, circularized it, and mapped WGS reads to the graph genome sequence
346   using the variation graph toolkit (vg) with the default parameter set (Garrison et al., 2018). We
347   omitted the GC-content correction procedure to simplify the pipeline as the method using the
348   full-length genome sequence seemed to be robust (Brown et al., 2016) against the local coverage
349   depth bias attributed to the GC-content (Ross et al., 2013). If a user requires greater accuracy, we

350 recommend using the correction method or PCR-free library preparation (Benjamini & Speed,
351 2012; Brown et al., 2016; Gao & Li, 2018).
352
353
354 **Growth rate evaluation with experimental growth rate**
355 The accuracy of the growth estimates was evaluated via comparisons with experimental growth
356 rates. Unless otherwise noted, the experimental growth rates were calculated from the colony
357 formation unit (CFU)/ml, optical density (OD), or relative abundance using $gr_i =$
358 $\frac{\log_2(abun_{i+1}) - \log_2(abun_{i-1})}{t_{i+1} - t_{i-1}}$ following the approach taken in a previous study (Korem et al., 2015);
359 e.g., the experimental growth for Supplementary Fig. 3c uses this equation. For relative
360 abundance, we used the mOTUs2 pipeline with the default parameter set (Milanese et al., 2019).
361 When comparing methods, we additionally calculated the experimental growth rate (shown in
362 Supplementary Fig. 3d) in a differential manner to obtain the dynamics in a short time span; i.e.,
363 $gr_i = \frac{\log_2(abun_{i+1}) - \log_2(abun_i)}{t_{i+1} - t_i}$. For comparison, we performed growth estimation using tools
364 developed in previous studies (Korem et al., 2015; Brown et al., 2016; Emiola & Oh, 2018; Gao
365 & Li, 2018). We added a time delay for the correlation coefficient between the experimental and
366 growth estimates following the approach of the previous study (Korem et al., 2015). The time
367 delay, which provided three or more combinations of the growth estimates and experimental
368 growth rates and yielded the highest correlation coefficient, was accepted.
369
370
371 **Effect of normalization on the model**
372 We evaluated the effect of normalization on the parameter estimation by checking the difference
373 between an estimated distribution and the true one. For the evaluation datasets, we generated
374 continuous angles from the von Mises distribution. These angles were binned into discrete angles
375 following the defined discrete length. We selected the location parameter from $-\pi$, $-\frac{\pi}{2}$, 0, and $\frac{\pi}{2}$
376 ; the concentration parameter from 0.1, 0.4, and 0.7; the discrete length to be segmented from 5,
377 10, 30, 120, 600, and 1,000; and the average coverage depth to be observed from 0.5×, 1.0×,
378 2.0×, 4.0×, 8.0×, and 16.0×. Next, we fitted the unnormalized and normalized von Mises
379 distribution-based models to the simulated data. Finally, we evaluated the error of the estimated
380 distribution using the Kullback-Leibler divergence (Kullback & Leibler, 1951). When evaluating
381 the parameter of interest, the other parameters were fixed to 0 for the location parameter, 0.7 for
382 the concentration parameter, 30 for the discrete length, and 16 for the average coverage depth.
383 For the distribution with the true parameter, we set 10,000 as the discrete length. Calculations
384 were performed 10 times with different seeds for each parameter set. This part of the procedure
385 was performed using Scipy (Virtanen et al., 2019). We introduced a constraint c_ω to normalize
386 the continuous circular distribution, i.e., $P_{\text{discrete}}(\theta_n|\omega) = \frac{1}{c_\omega}P_{\text{continuous}}(\theta|\omega)$ for discrete circular
387 data $\theta_n(n = 1,2,...,N)$ and the parameter set ω. For the von Mises and wrapped Cauchy

388 distributions, we calculated the sum of the likelihood directly as $\log(c_\omega) = \log\left(\sum_{n=1}^{N} P(\theta_n|\omega)\right)$

389 because we could not have a closed-form equation. For the cardioid distribution, it was

390 formulated as $c_\omega = \sum_{n=1}^{N} \frac{1}{12\pi}(1 + 2\rho\cos(\theta_n - \mu)) = \frac{N}{2\pi}$ owing to $\sum_{n=1}^{N}\cos(\theta) = 0$. For the linear

391 cardioid distribution, it was formulated as $c_\omega = \sum_{n=1}^{N} \frac{1}{12\pi}\left(1 + 2\rho\left(||\theta - \mu| - \pi| - \frac{\pi}{2}\right)\right) = \frac{N+2}{2\pi}$. The

392 rearrangement of the formulas was performed using Sympy (Meurer et al., 2016).

393

394

395 **Simulation of skewness after coverage depth sorting and investigation of the causes**

396 The cause of the skewed shapes appearing at both ends of the coverage depth after sorting

397 required investigation. The probable phage and duplicate gene regions are considered to generate

398 outliers and to form the skewed shape. Thus, the probable regions on the chromosomal sequence

399 were annotated. To identify probable phage regions in the genome, we used PHAST (Zhou et al.,

400 2011). For the duplicate gene regions, first, we used Prokka to predict the coding sequences

401 (Seemann, 2014). Next, we mapped these predicted sequences to the genome sequence of

402 *Lactobacillus gasseri* using Bowtie2 and annotated regions as overlapping if two or more hits

403 were obtained. The parameter set of Bowtie2 was "-a --very-sensitive." Thereafter, we extracted

404 the partial coverage depth from 1.0–1.4 Mnt regions in which annotated features did not exist. To

405 obtain the odds ratio, 36 metagenomic sequences obtained from the previous study (Korem et al.,

406 2015) were mapped, and the coverage depth was calculated via the method described above.

407 Next, the coverage depth for each sample was sorted, and the number of annotated bases in 5%

408 of the upper, lower, and total sequence was counted. We modeled the number n to follow a

409 binomial distribution with the total number of bases N of the target sequence and the appearance

410 probability p as parameters: $n \sim \text{Binomial}(N,p)$. The odds ratio was calculated as $odds = \frac{p}{1-p}$ from

411 the appearance probability p. To estimate these parameters, we employed an MCMC algorithm

412 with NUTS using PyStan. Four chains were utilized, and 20,000 iterations were performed,

413 where the first 1,000 of these were discarded as warm-ups. The posterior distribution of the EAP

414 was used as the representative.

415

416

417 **Evaluation of robustness in terms of coverage depth using culture dataset**

418 To investigate the robustness of our proposed method, we compared the growth estimates

419 calculated using a sufficient amount of reads and those using rarefied reads. To evaluate

420 coverage depth dynamics from a single origin, we first used *L. gasseri* WGS samples with an

421 average of more than 20× coverage (n = 20). After confirming that many variations occur at

422 lower than 5.0× coverage, we selected *Escherichia coli*, *Enterococcus faecalis*, or *L. gasseri*

423 WGSs reads with more than 5× coverage from the dataset by Korem and colleagues. To evaluate

424 multiple origins of replication data, the WGS of *S. solfataricus* was used. We randomly sampled

425 reads from the FASTQ files using seqtk (Li, 2013) such that the average coverage depth would

426   be 0.001×, 0.005×, 0.01×, 0.05×, 0.1×, 0.5×, 1×, or 5×. In the first evaluation of a replication
427   dataset from a single origin, we additionally sampled 10× and 20× coverage. In the evaluation of
428   multiple origins, we additionally sampled 100× coverage. The pPTR, wPTR, and mwPTR were
429   calculated using the rarefied reads and compared with those obtained from 20× and the full
430   coverage depth. For the *S. solfataricus* dataset, we specified the number of components as three
431   and performed estimation via the optimizing mode using 30 different seeds. We selected
432   representative results with the highest likelihood and compared pPTR and mwPTR with those
433   corresponding to no modifications. As DEMIC cannot work with a single genome sequence even
434   if it is complete, we used a genome sequence obtained by co-assembling all of the reads using
435   MEGAHIT with the default parameter set (Li et al., 2015). We utilized the default parameter set
436   for PTRC, DEMIC, bPTR, iRep, and GRiD. Finally, we calculated the error rates as

$$\left| \frac{Estimate_{\text{modified}} - Estimate_{\text{reference}}}{Estimate_{\text{reference}}} \right| \tag{5}$$

438   The results from original WGS reads were used for reference estimates. To validate the error
439   rate, we defined 15% as a threshold, as was done in a previous study (Brown et al., 2016).
440
441

Evaluation of robustness in terms of coverage depth using metagenomics reads

443   To evaluate the robustness using metagenomic datasets, we employed the inflammatory bowel
444   disease (IBD) dataset from a previous study (Franzosa et al., 2018). We searched for
445   combinations of species and WGSs with an average cover depth of 20× or more. In order for the
446   first screening to satisfy the scope of the method, we used Kraken2 (Wood, Lu & Langmead,
447   2019) and Bracken (Lu et al., 2017) with the default parameters; the objective of this was to
448   count the number of reads to be assigned to the genome sequences. As a collection of complete
449   chromosomal sequences, we constructed a database with species-level resolution (see *Complete*
450   *genome sequence database* section). Based on the taxonomic profile, the combinations with
451   more than 0.1× coverage depth were selected as candidates (n = 16,413 combinations from 220
452   WGS samples) in the first screening. Next, we aligned the WGS reads of the datasets to the
453   database using Bowtie2 and calculated the coverage depth using SAMtools. After applying the
454   moving median filter and outlier elimination, we calculated the average coverage depth. If the
455   combination had more than 20.0× average coverage and passed the first screening, we concluded
456   that the combination was eligible to serve as an evaluation target (n = 676). For the targets, we
457   extracted paired-end aligned reads using SAMtools with the "view -f 2" option for PTRC. This
458   procedure was required because the GEM-mapper, which is used in PTRC, did not allow
459   singletons to be aligned. After that, we rarefied the reads using SAMtools with the "view -s"
460   option and converted the alignment result into a FASTQ file for the input. For our method, we
461   counted the coverage depth from the alignment result using SAMtools. After applying each
462   method, we calculated the error rate following equation (5). We used PTRC as a benchmark
463   method, as it provided the most stable estimation with low coverage in the culture WGS dataset.
464

465

Evaluation of robustness in terms of mutation rate

The robustness of our method in terms of the mutation rate with a single replication origin was assessed by evaluating how much of the value was maintained when the reference genome was used for mutation estimation. The reference genome sequence was mutated in three ways. The first involved nucleotide-level mutation at random positions in the genome sequence. Based on the length of the genome sequence, every 5% portion, ranging from 5 to 30% of the nucleotides, was randomly selected and mutated to an ambiguous nucleotide N. The second way involved block-level mutation at random positions in the genome sequence. We used msbar in EMBOSS to create block-level mutated sequences (Rice, Longden & Bleasby, 2000). Based on the length of the genome sequence, every 5% portion, ranging from 5 to 50% of the nucleotides, was randomly mutated. We used 5,000 nt as the block size. The third way involved mutation at the block level at a specific region in the genome sequence. We randomly selected the position to be mutated. The size of the region was determined based on the sequence length. After that, every 5% portion, ranging from 5 to 30% of the nucleotides, was mutated to an ambiguous nucleotide N. As the positional relationship of the contig sequence was unknown, the evaluation of DEMIC in this regard was not performed. The first and third mutations were performed using in-house scripts with Biopython. After constructing pseudo-mutated genome sequences, we estimated and compared the growth following the above procedure. We assessed the robustness of the estimation with multiple replication origins, mutating the *S. solfataricus* genome sequence using the first and second methods. The estimations and evaluations for multiple origins of replication were performed in the same manner as the robustness evaluation with low coverage depth.

Evaluation of robustness in terms of peak noise

To investigate the influence of peak noise, we generated an artificial dataset. We used the short-read sequence of *E. faecalis*, *L. gasseri*, and *S. solfataricus* published previously (Korem et al., 2015; Payne et al., 2018). Firstly, a 100 bp region was randomly selected from a reference genome sequence using the seqkit sample command. The sequence was copied every 10 times from 10 to 100 times and added to the FASTQ files. We set 93 as the quality score. Using the mixed file, we calculated the coverage depth and estimated the growth in optimization mode following the procedure described above. For filtering, the top 1% of the depth was removed. Following this step, we computed the error rate using equation (5) and compared the results with the noiseless results. The evaluation for multiple replication origins was performed in the same manner as the evaluation for low coverage depth. Secondly, we randomly selected 1,000 bp regions from the reference genome sequence every 10 regions from 10 to 100 regions. This length was determined based on the average gene length of prokaryotes. These sequences were mixed with the WGS such that the coverage depth amounted to 20×, 40×, 60×, 80×, and 100× in each region. Finally, we evaluated the error of the estimates following the same procedure.

504

505

Evaluation of robustness in terms of sample size

506 **Evaluation of robustness in terms of sample size**
507 To assess the effect of the sample size on the estimates, partial sample sets were generated from
508 the full sample set, and the results were compared. We used the short-read sequences of *E. coli,*
509 *E. faecalis*, and *L. gasseri* published previously (Korem et al., 2015). The partial set was
510 configured to include 1, 4, 8, 12, 16, and 20 samples. Each set was distributed in a manner that
511 avoided duplication of the same sample. Except for the sample set with only one sample, the sets
512 were constructed to contain each sample at least 10 times. For each sample set, preprocessing
513 and inference were conducted according to the above procedure, and the error rate was
514 calculated in comparison with the results obtained when all of the samples were used
515 simultaneously.

516

517

Skewness in Watson and Crick strands

518 **Skewness in Watson and Crick strands**
519 To count the coverage depth in Watson strands and Crick strands, we used the SAMtools view
520 command with the "-f" option set to 0 for the Watson strand and 16 for the Crick strand. This
521 was done after mapping reads to the template genome sequence. The procedures that followed
522 were the same for both strands. Finally, the highest log-likelihood in 30 independent trials was
523 used as a representative estimate.

524

525

Growth estimation of species with multiple replication origins

526 **Growth estimation of species with multiple replication origins**
527 We used the von Mises distribution for mixing because it has an intermediate degree of density
528 around the mode and an open range of concentration parameters; i.e., $\kappa > 0$. Both genomic and
529 short-read sequences were obtained according to the procedures described previously
530 (Ausiannikava et al., 2018; Payne et al., 2018). The coverage depth was calculated according to
531 the above-mentioned procedure. As the deletion was confirmed in the genome sequence of *S.*
532 *solfataricus* by the Integrative Genomics Viewer (Robinson et al., 2011), we deleted regions
533 from 1,443,200 nt to 1,485,069 nt on CP011055, from 1,443,192 nt to 1,485,075 nt on
534 CP011056, and from 1,443,197 nt to 1,485,072 nt on CP011057. We used Dfast to annotate cdc*6*
535 in the genome sequences (Tanizawa, Fujisawa & Nakamura, 2018). The number of components
536 *M* was determined by using AIC and Widely Applicable Information Criterion (WAIC) as the
537 mixture model was a singular model, and there was a possibility that a decision based only on
538 AIC could produce incorrect results. The MCMC algorithm was applied to the constructed
539 model, and WAIC was calculated from the log-likelihood (Watanabe, 2010; Gelman et al.,
540 2013). The calculation was performed using an in-house Stan script and CmdStan with the
541 threading option. The sampling was performed 1,500 times on a single chain, where the first
542 1,000 samplings were excluded as warm-ups. We fitted the model to the data assuming a number
543 from 1 to 4 for the distribution. To calculate the AIC, we used the EAP of the posterior
544 distribution to represent the log-likelihood.

545

546

Growth rate estimation for infected *Citrobacter rodentium*

WGS reads of mice fecal samples from the original growth dynamics analysis study were used
(Korem et al., 2015). These sequences were aligned to a complete genome sequence database
(see *Complete genome sequence database* section) to validate the applicability of using multiple
reference genome sequences using Bowtie2 with the "--very-sensitive" option. After that, we
counted the coverage depth by SAMtools. After extracting the coverage of the *C. rodentium*
chromosome sequence, we cleaned and compressed the coverage using the moving median filter
after removing the top 1%. Finally, we fitted the von Mises model to the coverage. The pPTRs
were compared via Welch's *t*-test.

556

557

Complete genome sequence database

The Genome Taxonomy Database version 89.0 was used to control the fineness of the taxonomy
on the species level (Parks et al., 2018). For each species, when the representative species had a
complete genome, it was used. When it did not, the sequence with the highest CheckM
completeness score (Parks et al., 2015) and largest genome size was used. Species without
complete genome sequences or those with multiple chromosome sequences were excluded.
Mobile genetic elements were excluded from the database by checking the sequence label using
seqkit; we filtered out the sequences labeled "plasmid," "Plasmid," "phage," "chromid,"
"pMLa," and "Linear."

567

568

Growth estimate evaluation on metagenomic dataset

The growth dynamics were estimated at species-level resolution. We filtered low-quality reads in
WGS via Trimmomatic and then removed human-derived reads by aligning them to the GRCh38
reference human genomic sequence using Bowtie2. We used "SLIDINGWINDOW:4:15 MIN
LEN:36" as a parameter in Trimmomatic. After quality control, we aligned qualified
metagenomic reads with the complete genome sequence database using Bowtie2 with the "--
very-fast" option. After extracting the alignment results of the target reference sequence, we
counted the coverage depth. As the metagenomic sequences were not clean compared to the
culture datasets, we performed additional filtering as described in the coverage depth calculation
method. After preprocessing was completed, we fitted the model to the coverage depth of each
sequence via the optimizing mode. For the estimations, we selected samples with greater than or
equal to 0.0001% relative abundance of the taxon and greater than 0.01× average coverage
depth. We used Kraken2 and Bracken with the complete genome sequence database to estimate
the relative abundance of the species. After filtering out the ultra-low coverage depth samples,
we excluded the samples that might not achieve random sampling from the chromosomal DNA
sequence. This is because the estimates of these samples would have an error. To detect the

585 invalid samples, we focused on the difference of actual zero coverage fraction f and a theoretical
586 score $\hat{f}$ based on the Lander-Waterman theory (Supplementary Text S4). This theoretical score
587 can be obtained as $\hat{f} \approx exp(-a)$, where a denotes the average coverage depth. For samples with
588 average coverage less than 5.0×, we excluded samples with log-scale fractions greater than 0.56
589 times the theoretical score. Assuming there to be uneliminated noise coverage depth, samples
590 with estimated PTRs greater than or equal to 3.0 were excluded. Welch's t-test for independent
591 groups was used to examine the differences between the growth estimates, and Hedges' g was
592 used to evaluate the effect size for the two groups.
593
594

595 **Software**
596 We implemented the statistical model using Stan (Carpenter et al., 2017). Wrapped by Python
597 scripts, this model is available for use in the command-line environment. This package also
598 contains a moving median filter, a visualizer, a statistics profiler based on directional statistics,
599 an information criterion calculator with estimated results, an asymmetric test calculator using
600 Pewsey's method, and other utilities required to analyze the coverage depth over replicon. Other
601 software versions are summarized in Supplementary Table 1. Our package for growth estimation
602 is available from https://github.com/TaskeHAMANO/SPHERE. This software was implemented
603 using Python3 (≥ 3.6) and Stan. The wrapper software used in this study for PTRC, DEMIC, and
604 GRiD is available from https://github.com/TaskeHAMANO/PTRC-in-cwl,
605 https://github.com/TaskeHAMANO/DEMIC-in-cwl, and
606 https://github.com/TaskeHAMANO/GRiD-in-cwl, respectively. This software is distributed
607 under the BSD-3-Clause license. The wrapped software of msbar in EMBOSS is available from
608 https://github.com/TaskeHAMANO/msbar-in-cwl. This software is distributed under the GPL-
609 3.0 license. These wrapper scripts were implemented using the Common Workflow Language
610 (CWL) v1.1. These scripts have been tested on Linux and macOS.
611
612

613 **Availability of data and material**
614 The WGS data of time-series-cultured *E. coli*, *E. faecalis*, *L. gasseri*, *S. solfataricus*, and *H.*
615 *volcanii* are available from BioProject (PRJEB9718, PRJNA250819, PRJNA250820,
616 PRJNA250827, PRJNA346830, PRJNA250832, PRJNA250833, and PRJNA422812). The
617 genome sequences of *E. coli* NMC3722, *E. faecalis* ATCC 29212, *L. gasseri* ATCC33323, *S.*
618 *solfataricus* SULA, SARC-B, SARC-C, USLG, SARC-H, SARC-I, and *H. volcanii* DS2 are
619 available from GenBank and RefSeq (CP011495, CP008816, NC_008530, CP011057,
620 CP011055, CP011056, CP033235, CP033236, CP033237, and NC_013967). The genome
621 sequence of *H. volcanii* H26 was modified from DS2 as previously described (Hawkins et al.,
622 2013). The genome and metagenome sequences used in the cohort studies analysis are listed in
623 Supplementary Table 2. The final chromosome sequences we used to construct the genome
624 sequence database are listed in Supplementary Table 3.

625
626
Results

628 **Creating an artificial coverage depth**
629 We constructed a statistical model for coverage depth dynamics based on circular distributions.
630 To validate our model visually, we generated an artificial coverage depth using the above-
631 mentioned circular distributions (Supplementary Text S5). The generated coverage depth
632 reproduced high variance and concentration at the replication origin, expressed as the location
633 parameter of the circular distribution. Interestingly, when sorted, this artificial coverage depth
634 showed a distorted trend in both the upper and lower orders regardless of the circular distribution
635 type (Fig. 1a). The same shape was visualized previously (Brown et al., 2016), wherein it was
636 stated that this shape was formed by a specific sequence feature, such as a phage. However, our
637 model generated artificial depths from smooth probability trends and did not include any
638 artificial noise. To investigate in detail the cause of the distorted regions seen at both ends,
639 potential prophage sequences and duplicate genes in the genome sequence of *Lactobacillus*
640 *gasseri* were analyzed. Among them, only the intact prophage region was abundant at the lower
641 end (Supplementary Table 4). Moreover, a similar distorted structure was reproduced on a partial
642 genome sequence that did not contain suspicious regions (Fig. 1b). We evaluated the best model
643 for this distribution and used it to determine the threshold for outlier removal (Supplementary
644 Table 5; Supplementary Text S3).

645
646

647 **Performance evaluation with experimental growth rates**
648 Using a statistical model based on a circular distribution, we first evaluated the model's accuracy
649 by estimating the correlation between the computational and experimental growth rates, as had
650 been done previously (Korem et al., 2015). We estimated the coverage depth by mapping the
651 WGS reads to the genome sequence and counting the coverage depth (Fig. 2a). After reducing
652 the variance, outliers, and data size with a moving median filter, the proposed model was fitted
653 to the cleaned coverage depths (Fig. 2b). To evaluate the accuracy of the method, we used the
654 WGSs of the three species (*E. coli*, *E. faecalis*, and *L. gasseri*) previously obtained from culture
655 experiments under aerobic and anaerobic conditions (Korem et al., 2015). These data were
656 accompanied by CFU/ml or OD in time series for evaluation. As substantiated in the previous
657 study, we observed a high correlation coefficient between the growth estimates and experimental
658 growth rates (Fig. 2c; Supplementary Fig. 2ab). Regardless of the culture state, *E. coli* and *E.*
659 *faecalis* exhibited high degrees of correlation, without requiring time delay adjustments ($r \geq 0.5$).
660 In contrast, *L. gasseri* required a time delay adjustment of 90–120 minutes. Our growth estimates
661 yielded correlation coefficients equivalent to those obtained using the previous methods, with
662 experimental growth rates of both 60 minutes (Supplementary Fig. 2c) and 30 minutes
663 (Supplementary Fig. 2d). Our estimates were correlated with the temporal growth based on the
664 relative abundance ($r = 0.76 \pm 0.04$, n = 4, each with 10 timepoints; Supplementary Fig. 2e)

665   previously obtained (Korem et al., 2015) even when the samples originated from mixed cultures
666   with multiple intestinal species.
667   Secondly, we tuned the parameters of interest. The window and stride size of the
668   moving median filter were optimized to 100 bp by comparing the growth estimates with the
669   experimental growth rates (Supplementary Figs. 3 and 4; Supplementary Texts S5 and S6). Our
670   pipeline, which used a sequence aligner that did not take circular structures into account,
671   confirmed that the decrease in coverage at both edges, termed the edge effect in a previous study
672   (Brown et al., 2016), exerted only a small effect on the estimation (Supplementary Fig. 5;
673   Supplementary Text S7). Our method performed well regarding memory usage and computation
674   time with the exception of the Jones-Pewsey distribution-based model (Supplementary Fig. 6).
675   Finally, we evaluated the applicability of the method using artificial datasets. When applied
676   to modified datasets from culture experiments, most methods, including those used in previous
677   studies, have performed adequately with low coverage depth WGSs until at least 0.5×, whereas
678   our growth estimates remained stable even at coverage depths of 0.01× (Fig. 3a and
679   Supplementary Fig. 7a). According to the bPTR in the coverage depth mode, the error
680   percentage was not correlated with the number of reads. Although GRiD seemed to maintain a
681   low error rate from 0.005× to 0.01× coverage depth, the number of samples with growth
682   estimated as 1.0 increased (0.005×: 10/20, 0.01×: 17/20). As DEMIC requires multiple contigs
683   for estimation and thus is not applicable to a single template genome sequence even when the
684   sequence is complete, we performed the evaluation using assembled genome sequences.
685   Moreover, this approach cannot be applied to low coverage less than or equal to 1.0× coverage.
686   Given the culture dataset results, we also performed the evaluation on metagenomic reads and
687   confirmed that the error rate was less than 15% on average until 0.01× coverage (Fig. 3b). Our
688   method was also stable for mutations at both the 5,000 bp block and single nucleotide levels
689   (Supplementary Figs. 7b and c). Because the approach uses the entire sequence structure, the
690   results obtained from a sequence mutated on a single specific region deviated from those of full-
691   length sequences (Supplementary Fig. 7d). To evaluate the effect in human intestine WGS reads,
692   we quantified the deletion size on the chromosome sequences in a metagenomic dataset
693   (Supplementary Fig. 8; Supplementary Text S8). Our estimates were as stable as those generated
694   using the previous methods when a single peak noise was contaminated (Supplementary Fig. 7e).
695   With more noise, although the error rate of our estimates remained less than 15% on average,
696   some samples showed substantial error as the amount of artificial noise increased
697   (Supplementary Fig. 7f). To address this, we investigated the relationship between the error rate
698   and the zero coverage fraction (Lander & Waterman, 1988; Roach, 1995) and determined the
699   threshold to exclude invalid samples with an error rate of more than 15%. As a result, we
700   detected the noise-contaminating coverage samples with a recall score of 0.81 (Supplementary
701   Texts S4 and S9; Supplementary Figs. 9 and 10; Supplementary Table 6). Finally, the number of
702   samples was related to the variation, but the effect was not substantial compared with those of
703   the other factors (Supplementary Fig. 7g).
704

705

Performance evaluation using *in vivo* dataset

706
707   To evaluate the accuracy, we compared the growth estimates with the known growth dynamics
708   using previous datasets. For the *in vivo* sample setting, we tuned the window size of the moving
709   median filter based on the coefficient of variance and concluded that 100 nt was the best
710   (Supplementary Fig. 11). First, we checked the reproducibility of the growth estimates using *C.*
711   *rodentium*-infected mice fecal samples. As was also reported previously (Korem et al., 2015), *tir*
712   mutant *C. rodentium* had a higher pPTR than the wild-type (WT) strain (Supplementary Fig. 12;
713   *p*-value by Welch's *t*-test between WT and mutant on days 6–9: 8.72×10^{-5}, $n_{WT} = 12$, $n_{\Delta tir} = 12$).
714   Second, we evaluated the growth dynamics in the fecal microbiome in IBD patients (Franzosa et
715   al., 2018). When we compared the estimates between Crohn's disease subjects and healthy
716   volunteers, we reproduced the significant high growth estimates of *Eggerthella lenta* in the
717   patients (*p*-value: 1.26×10^{-7}, Hedges' $g = -1.21$, $n_{healthy} = 42$, $n_{Crohn} = 49$). Although this
718   difference was limited to the remission and active patients in the small sample size dataset used
719   in the previous study (Korem et al., 2015), we observed this difference in the large cohort dataset
720   even by PTRC (*p*-value: 1.31×10^{-2}, Hedge's $g = -0.57$, $n_{healthy} = 35$, $n_{Crohn} = 57$). Finally, we
721   confirmed the growth dynamics of *Bifidobacterium breve* and *Bifidobacterium adolescentis* in
722   the neonates and their mothers fecal microbiome cohort (Bäckhed et al., 2015). It is well known
723   that *B. breve* is abundant in infant guts, whereas *B. adolescentis* is abundant in adult guts.
724   Moreover, a previous experimental study demonstrated that *B. breve* grows well in a medium
725   containing formula based on soy, milk, or casein hydrolysate. These biological signals were also
726   reproduced in the estimates obtained using our model (Supplementary Fig. 13; Supplementary
727   Table 7).

728

729

Shape, peakedness, and skewness of coverage depth

730
731   As an additional application of our model, we investigated the shape of the coverage depth by
732   comparing the kinds of circular distributions(Supplementary Tables 8 and 9). In a comparison of
733   the fitness of multiple models, the Jones-Pewsey distribution model exhibited the highest fitness
734   among the vanilla models (those without argument transformation) on average. The shape
735   parameter of the Jones-Pewsey distribution in the datasets of Korem et al, 2015. changes
736   considerably with time (Supplementary Fig. 14a). For example, in the *E. faecalis* dataset, the
737   distribution was dense around the replication origin in the first phase; however, it gradually
738   dispersed over time. In contrast, the trend was reversed in the anaerobically cultured *L. gasseri*.
739   To evaluate the coverage depth concentration phenomenon further, we implemented the
740   InvSE von Mises distribution model. Along with the Jones-Pewsey distribution model, the
741   peakedness parameter of the InvSE von Mises distribution changed considerably with time
742   (Supplementary Fig. 14b). Comparing the model with the vanilla von Mises-based model, the
743   InvSE von Mises-based model exhibits a lower AIC and BIC (Supplementary Table 8). As in the
744   Jones-Pewsey model case, the peakedness was initially high. However, in *E. faecalis*, it became

745   lower later on (Fig. 4). These parameters in the Jones-Pewsey and InvSE von Mises distributions
746   showed high correlation coefficients, but their trends were not identical (min r = -0.793, p =
747   2.93×10^{-7}; Supplementary Fig. 15).
748   Next, we evaluated the symmetricity of replication. Although several methods that
749   extend circular probability densities toward asymmetricity have been described (Batschelet,
750   1981; Pewsey, 2002; Abe, Pewsey & Fujisawa, 2013), a few requirements must be satisfied to
751   adapt to replication dynamics. Therefore, the InvMIAE von Mises distribution-based model was
752   used in this study. First, we evaluated the robustness of the asymmetric extended method
753   (Supplementary Text S10; Supplementary Fig. 16). As a result, we concluded that this extension
754   was not suitable for use with the short-read sequences that were largely mutated from the
755   template genome sequence. We therefore determined the applicability of the dataset by
756   estimating the mutation rate from the frequency of the zero-coverage depth using a zero-inflated
757   model (Supplementary Text S11). The results indicated that the *E. faecalis* dataset did not satisfy
758   the criteria (Supplementary Fig. 17a). Finally, we fitted the model to the actual coverage depth.
759   The skewness parameter had a low variance with time and was nearly 0 except for the *E. faecalis*
760   data (Supplementary Fig. 14b). Although *E. faecalis* showed high skewness, the InvMIAE model
761   fitted 0 or outlier coverage depths rather than the skewness of the whole sequence
762   (Supplementary Fig. 17b). Furthermore, we measured skewness using only Watson and Crick
763   strands. The skewness parameters showed strong correlations, and no specific skewness in any
764   specific strand was found (Supplementary Fig. 18). These results support the hypothesis that
765   coverage depth is symmetrical, contrary to our expectations.
766
767

768   **Extension of the model to multiple replication origins**
769   To demonstrate the extendibility of our method, we modeled the coverage depth behavior of
770   multiple replication origins, using a mixture of circular distributions. To validate our model, we
771   applied it to WGS data of *Sulfolobus solfataricus* and *Haloferax volcanii,* which contain three
772   replication origins (Ausiannikava et al., 2018; Payne et al., 2018). Based on the AIC and WAIC,
773   we determined the number of components in both datasets to be three because this number
774   yielded the best score on average (Supplementary Table 10). Of the seven datasets, five matched
775   the true number of active replication origins. The estimated location parameters of *S. solfataricus*
776   were distributed close to *cdc6,* which is a marker gene for the replication origin (the average
777   error rate of the location parameter with respect to the marker gene is $6.55 \pm 4.48\%$, n = 6)
778   (Supplementary Figs. 19a–f) (Lundgren et al., 2004; Robinson et al., 2004). In contrast, although
779   there is a distinct peak around 2 Mbp, *cdc6* is not evident in the *H. volcanii* genome sequence
780   (Figs. 5a and b). We compared the weighted PTRs between the exponential growth and
781   stationary phases. All of the origins in the exponential growth phase increased the estimates
782   (exponential growth phase: 3.59, 3.18, and 2.66; stationary phase: 1.64, 2.26, and 1.43). We also
783   checked the difference in the wPTR among multiple origins. Notably, the middle of the
784   replication origin position nearly coincided with the position at which the genomes were split by

785   the ratio of wPTR (Figs. 5c and d; Supplementary Figs. 19g–l). As was done when the model
786   was applied for a single replication origin, the robustness of the estimates was evaluated using
787   the artificially modified dataset, which was an *S. solfataricus* dataset in this case. As a general
788   trend, the individually weighted PTRs were more sensitive to the modifications than the mean
789   weighted PTRs. When the number of reads was limited to 0.1× coverage depth on average, the
790   error of the estimates was less than 15% at the median (Supplementary Fig. 19m). Although it
791   was more susceptible to noise than the model for a single origin of the replication origin, this
792   estimate was robust so long as the mutation rate was less than 7% at the point level or 4% at the
793   block level (Supplementary Figs. 19n and o). Our method avoided the effect of a single noise
794   region which increases coverage in the conserved region (Supplementary Fig. 19p).
795
796

797   **Discussion**
798   Here, we introduced a generative statistical model of coverage depth based on circular statistics
799   and evaluated the estimated growth dynamics, replication trend, and differences in wPTR among
800   multiple origins. In directional statistics, the simplest approach to expressing angular bias may be
801   the use of the MRL. Although the MRL of the coverage depth was correlated with the
802   experimental growth rate in the culture datasets, it is not as robust as estimates obtained via
803   statistical models. This statistic can be easily calculated even with poor computational resources,
804   but is not suitable for metagenomic datasets. Our proposed method was as accurate as the
805   previous methods when compared with the experimental growth rates, and furthermore, it was
806   robust against random mutations in the reference sequence and decreases in the coverage depth.
807   Conversely, it was sensitive to a decrease in coverage depth due to mutations concentrated in a
808   specific direction as well as to an increase in coverage depth due to conserved regions. In future
809   research, it is expected that the rapid increase or decrease in coverage depth will be modeled to
810   more accurately estimate the dynamics of the coverage depth. The simplest approach is not to
811   use coverage depth in regions that are expected to be ineligible, as has been done in previous
812   studies. However, this filtering approach alone does not provide a reasonable estimation for the
813   proposed model as it also uses the absence of observation for parameter estimation. If a valid
814   region [*a,b*] can be assumed, ineligible regions could be excluded by normalizing the likelihood
815   function to satisfy $\int_a^b p(\theta)d\theta = 1$. In applying the proposed method to the coverage depth obtained
816   from the metagenomic sequence, the average coverage depth and random sampling properties
817   must be examined, as was done here. Although we did not utilize it in this study, one of the
818   advantages of a GLM is its ability to incorporate covariate effects into the model. If one wants to
819   evaluate the relationship between the covariates **x** and pPTR, it is suitable to use a link function
820   for the concentration parameter. For example, when the von Mises distribution model is used, let
821   β be the coefficient of the covariates; then $\kappa = \exp(\beta_0 + \beta_1 x_1 + \cdots)$ satisfies the requirement, i.e.,
822    $\kappa > 0$. For a wrapped Cauchy distribution, the inverse logit function is appropriate to satisfy
823    $0 < \rho < 1$.

824   We demonstrated the generation of artificial coverage depths using our statistical model.
825   It was confirmed that the shape after sorting in ascending order was similar to the experimentally
826   obtained replication profiles, which were presented previously (Brown et al., 2016). This shape
827   did not depend on the type of circular distribution. These results demonstrate that the distorted
828   shape could be generated not only by prophage sequences, strain variations, and highly
829   conserved regions but also by the randomness of observation (DNA sequencing). When we
830   evaluated this shape using actual data, we observed its appearance even in partial sequences that
831   did not include these regions. This finding suggested that the shape is attributable not only to
832   specific regions but also to the variance in observations, which is modeled by a multinomial
833   distribution in our model. Filtering these parts out undoubtedly reduces the noise in the coverage
834   depth.
835   In the evaluation using the *in vivo* dataset, we successfully confirmed consistency with
836   previous studies. The species in the *Bifidobacterium* genus showed growth diversity when our
837   method was evaluated using fecal WGS from infants and their mothers. Previous studies have
838   revealed that *Bifidobacterium adolescentis* is abundant in adults and *Bifidobacterium breve* is
839   abundant in infants (Turroni et al., 2012; Ruiz-Moyano et al., 2013; Kato et al., 2017); this trend
840   was also reflected by the growth estimates. Although this finding was not reproduced by PTRC,
841   previous studies have indicated that *B. breve* grows faster than other *Bifidobacterium* species in
842   formula milk (Dubey & Mistry, 1996) and human breast milk (Turroni et al., 2011). Since not all
843   of the infants in the dataset had been weaned at the time of the study (Bäckhed et al., 2015), it is
844   suggested that our method appropriately interpreted the dynamics.
845   When we compared the non-extended directional distributions for the replication trends,
846   the Jones-Pewsey distribution exhibited the best fitness. This result implies that the additional
847   parameter could contribute to the coverage depth dynamics that had been overlooked. The
848   additional shape parameter implied that more reads were concentrated around the replication
849   origin in the early stage of the exponential growth phase, except for *L. gasseri* in an anaerobic
850   culture. We additionally applied the InvSE model to evaluate this phenomenon based on another
851   quantification; this model reproduced the trend obtained using the Jones-Pewsey distribution
852   model. We provide two possible explanations for the above phenomenon. The first is the effect
853   of multiple replication forks. As the cell division phase is shorter than the genome replication
854   phase in bacteria, the genome begins replication before finishing the current replication origin
855   (Cooper & Helmstetter, 1968; Bremer & Churchward, 1977; Yoshikawa & Wake, 1993;
856   Wallden et al., 2016), allowing multiple rounds of replication to occur around the replication
857   origin while rapid replication is occurring. Emiola and Oh also discussed the effect of multiple
858   fork replication on the coverage depth (Emiola & Oh, 2018). The second hypothesis is that, as
859   the entire chromosome is not affected at the start of DNA replication, some DNA appears only
860   around the replication origin. However, this concept does not explain the generation time of
861   bacteria. Under laboratory conditions, DNA replication of *E. coli* is reportedly completed within
862   approximately 30 minutes (Helmstetter & Cooper, 1968). If the second hypothesis was valid, the
863   additional coverage depth concentration around the origin should be finished within 30 minutes

864   from the beginning of the culture. However, the degree of density remained low for an hour in
865   our study. The trend observed in the anaerobic cultured *L. gasseri* was the opposite of what was
866   seen in the others; however, it is worth noting that *L. gasseri* required 90–120 minutes of
867   adjustment to have a sufficient correlation between the experimental growth rate and estimated
868   growth dynamics. This suggests that both the activation of DNA replication and cell division are
869   required to decrease the degree of density. Accordingly, we inferred that the degree of density
870   and peakedness may indicate the activity of multiple replication forks. In contrast, the skewness
871   parameters in *E. coli* and *L. gasseri* did not change dynamically during the experimental
872   duration. Additionally, we confirmed the presence of a strong correlation between the skewness
873   of the Watson and Crick strands, implying that the amount of DNA remains symmetric between
874   the Watson and Crick strands as well as between the leading and lagging strands.
875   In addition to the application to microbes with a single replication origin, we extended
876   the model's application to microbes with multiple origins in a single chromosome. One
877   interesting finding was the difference in wPTR among multiple origins. From the relationship
878   between the intermediate position of the origins and the split position of the chromosome
879   sequence based on wPTR, the efficiency, in terms of the activity of the origins, was
880   quantitatively confirmed. If only a single replication origin was active in a chromosome,
881   considerable time could be required for whole-genome replication, which would be a
882   disadvantage for survival. By properly activating the origins at a distance, replication may be
883   efficiently completed. However, our results indicated that not all replication origins exhibit
884   similar activity. There are various characteristics that cause the activity to differ, such as
885   (a)synchronous initiation (Lundgren et al., 2004), replication fork speed (Elshenawy et al., 2015),
886   and so on. Therefore, the mechanism underlying the differences observed for each replication
887   origin must be clarified, and the characteristics of neighboring genes must be investigated.
888   The current study was affected by certain limitations. First, the proposed method requires
889   circular genome sequences for accurate estimation. As several methods involving contig or
890   scaffold-level sequences have already been proposed for estimating the *quasi*-growth of bacteria
891   (Brown et al., 2016; Emiola & Oh, 2018; Gao & Li, 2018), it is recommended that these methods
892   be properly used depending on the accuracy requirements. It is difficult to detect trends in the
893   amount of DNA other than the coverage depth bias or to estimate the bias in chromosomes with
894   multiple replication origins using these methods. We consider our method to be appropriate for
895   data analyses related to detailed replication profiles.
896   Second, the taxonomic resolution is limited to the species level in our method, on account
897   of the first limitation. When the growth estimates of a reference strain were calculated using
898   metagenome samples containing different but closely related strains, their growth dynamics were
899   found to be different, indicating that the pPTR distributions may be mixed. This difficulty
900   regarding the taxonomic resolution has yet to be solved via growth rate estimation, which may
901   give rise to major challenges in environments such as soil, wherein many closely related species
902   are contained because of empty niches and/or microstructures (Dumbrell et al., 2010; Thompson
903   et al., 2017). However, this challenge may be less serious in environments devoid of close

904   relatives on account of the filling of niche space and/or strong selective pressure. The human
905   intestine likely corresponds to the latter case (Jeraldo et al., 2012; Li & Ma, 2016; Thompson et
906   al., 2017) but may shift to the former case in situations in which the population is being
907   reconstructed because of an environmental change (Langenheder & Szekely, 2011). This
908   resolution problem may be solved by constructing pan-genome sequences from metagenomic
909   reads and allocating coverage depth appropriately. Third, the evaluation scope of the extended
910   model is limited. Although we evaluated and eliminated the possibility of overfitting in our
911   dataset, we cannot deny the possibility that the dynamics of the peakedness and stability of the
912   skewness around the origin are specific to the three strains we used. External validations are
913   expected to confirm the variability of the peakedness or stability of the skewness over the growth
914   phase. Finally, in our method as well as all currently proposed methods for estimating bacterial
915   growth, the estimate itself is only a proxy of the growth rate. Theoretically, (p)PTR for a taxon t

916   in a sample s is represented by $(p)PTR_{s,t} = 2^{C_t/\tau_{s,t}}$, where C_t is the replication period and $\tau_{s,t}$ is
917   the doubling time (Cooper & Helmstetter, 1968; Bremer & Churchward, 1977; Korem et al.,
918   2015). Our interest is in the doubling time, but the estimate is also influenced by the replication
919   period. This period may vary from species to species depending on the genome size and other
920   factors. Therefore, it is not appropriate to compare estimates between species. It is necessary to
921   analyze the effects of the replication period C and to propose a method that yields a doubling
922   time that is comparable between species (Gibson et al., 2018).
923

Conclusions

925   We developed a probabilistic model based on circular statistics to model the coverage depth
926   behavior in DNA replication using WGS reads. This method was demonstrated to be robust for a
927   small number of reads ($\geq 0.01\times$). The probabilistic PTR from our model demonstrated a
928   significant correlation with the experimental growth rates in the culture dataset. In addition to
929   facilitating quantification of the ratio differences, this method enables detailed measurement of
930   DNA quantity changes by using circular distributions in the model. Moreover, by combining
931   multiple distributions, it became possible to estimate the growth of organisms with multiple
932   replication origins, such as archaea. Therefore, this method further extends the applicability of
933   growth estimation from fragmented reads. We expect that the growth estimation method
934   presented herein will help elucidate factors that have not yet been observed in studies of
935   microbiome formation.
936

Acknowledgements

938   We thank Ken Kurokawa, Hiroshi Mori, and Koichi Higashi for discussions regarding the
939   research design during the early stages of this work. We thank Yuya Nakamura for discussions
940   regarding the statistical modeling experiment and design of the linear cardioid model. We thank
941   Darzi Youssef for his review of the *in vivo* dataset section. We thank Anton Björk for his
942   essential advice related to the skewable asymmetric extended model. We thank Thorsten Allers

943 for providing the *H. volcanii* genome sequence. We are grateful to the reviewers for their
944 numerous helpful comments, which helped us to improve this report.
945

References

947 Abe T, Pewsey A, Fujisawa N. 2013. Asymmetric distribution family on circumference with mode invariance. In:
948   *JFSAA*. Osaka, 1.
949 Abe T, Pewsey A, Shimizu K. 2013. Extending circular distributions through transformation of argument. *Annals of*
950   *the Institute of Statistical Mathematics* 65:833–858. DOI: 10.1007/s10463-012-0394-5.
951 Akiyama MT, Oshima T, Chumsakul O, Ishikawa S, Maki H. 2016. Replication fork progression is paused in two
952   large chromosomal zones flanking the DNA replication origin in Escherichia coli. *Genes to Cells* 21:907–914.
953   DOI: 10.1111/gtc.12388.
954 Andersson AF, Pelve EA, Lindeberg S, Lundgren M, Nilsson P, Bernander R. 2010. Replication-biased genome
955   organisation in the crenarchaeon Sulfolobus. *BMC Genomics* 11:454. DOI: 10.1186/1471-2164-11-454.
956 Ausiannikava D, Mitchell L, Marriott H, Smith V, Hawkins M, Makarova KS, Koonin E V, Nieduszynski CA,
957   Allers T. 2018. Evolution of genome architecture in archaea: spontaneous generation of a new chromosome in
958   Haloferax volcanii. *Mol. Biol. Evol.* 35:1855–1868. DOI: 10.1093/molbev/msy075.
959 Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, Li Y, Xia Y, Xie H, Zhong H, Khan MT,
960   Zhang J, Li J, Xiao L, Al-Aama J, Zhang D, Lee YS, Kotowska D, Colding C, Tremaroli V, Yin Y, Bergman
961   S, Xu X, Madsen L, Kristiansen K, Dahlgren J, Jun W. 2015. Dynamics and stabilization of the human gut
962   microbiome during the first year of life. *Cell Host and Microbe* 17:690–703. DOI:
963   10.1016/j.chom.2015.04.004.
964 Batschelet E. 1981. *Circular statistics in biology*. Academic Press Inc.
965 Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing.
966   *Nucleic Acids Research* 40:1–14. DOI: 10.1093/nar/gks001.
967 Bremer H, Churchward G. 1977. An examination of the Cooper-Helmstetter theory of DNA replication in bacteria
968   and its underlying assumptions. *Journal of Theoretical Biology* 69:645–654. DOI: 10.1016/0022-
969   5193(77)90373-3.
970 Brown CT, Olm MR, Thomas BC, Banfield JF. 2016. Measurement of bacterial replication rates in microbial
971   communities. *Nature Biotechnology* 34:1256–1263. DOI: 10.1038/nbt.3704.
972 Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A.
973   2017. Stan: A probabilistic programming language. *J. Stat. Softw.* 76. DOI: 10.18637/jss.v076.i01.
974 Chen L, Brügger K, Skovgaard M, Redder P, She Q, Torarinsson E, Greve B, Awayez M, Zibat A, Klenk H-P,
975   Garrett RA, Bru K, Redder P, Awayez M. 2005. The genome of Sulfolobus acidocaldarius, a model organism
976   of the Crenarchaeota. *J. Bacteriol.* 187:4992–4999. DOI: 10.1128/JB.187.14.4992-4999.2005.
977 Cooper S, Helmstetter CE. 1968. Chromosome replication and the division of Escherichia coli B/r. *Journal of*
978   *Molecular Biology* 31:519–540. DOI: 10.1016/0022-2836(68)90425-7.
979 Dowell M, Jarratt P. 1971. A modified regula falsi method for computing the root of an equation. *BIT Numerical*
980   *Mathematics* 11:168–174. DOI: 10.1007/bf01934364.
981 Dubey UK, Mistry VV. 1996. Growth characteristics of bifidobacteria in infant formulas. *Journal of Dairy Science*
982   79:1146–1155. DOI: 10.3168/jds.s0022-0302(96)76468-8.
983 Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH. 2010. Relative roles of niche and neutral processes in
984   structuring a soil microbial community. *ISME Journal* 4:337–345. DOI: 10.1038/ismej.2009.122.
985 Elshenawy MM, Jergic S, Xu Z-QQ, Sobhy MA, Takahashi M, Oakley AJ, Dixon NE, Hamdan SM. 2015.
986   Replisome speed determines the efficiency of the Tus−Ter replication termination barrier. *Nature* 525:394–
987   398. DOI: 10.1038/nature14866.
988 Emiola A, Oh J. 2018. High throughput in situ metagenomic measurement of bacterial replication at ultra-low
989   sequencing coverage. *Nature Communications* 9:4956. DOI: 10.1038/s41467-018-07240-8.
990 Forsyth VS, Armbruster CE, Smith SN, Pirani A, Springman AC, Walters MS, Nielubowicz GR, Himpsl SD,
991   Snitkin ES, Mobley HLTT. 2018. Rapid growth of uropathogenic Escherichia coli during human urinary tract
992   infection. *mBio* 9:e00186-18. DOI: 10.1128/MBIO.00186-18.
993 Franzosa EA, Sirota-madi A, Avila-pacheco J, Fornelos N, Haiser HJ, Reinker S, Vatanen T, Hall AB, Mallick H,
994   Mciver LJ, Sauk JS, Wilson RG, Stevens BW, Scott JM, Pierce K, Deik AA, Bullock K, Imhann F, Porter JA,
995   Zhernakova A, Fu J, Weersma RK, Wijmenga C, Clish CB, Vlamakis H, Huttenhower C, Xavier RJ. 2018.

996   Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology* 4:293–
997   305. DOI: 10.1038/s41564-018-0306-4.
998   Gao F. 2015. Bacteria may have multiple replication origins. *Front. Microbiol.* 6:1–4. DOI:
999   10.3389/fmicb.2015.00324.
1000   Gao Y, Li H. 2018. Quantifying and comparing bacterial growth dynamics in multiple metagenomic samples.
1001   *Nature Methods*. DOI: 10.1038/s41592-018-0182-0.
1002   Gao F, Zhang CT. 2008. Ori-Finder: A web-based system for finding oriCs in unannotated bacterial genomes. *BMC*
1003   *Bioinformatics* 9:1–6. DOI: 10.1186/1471-2105-9-79.
1004   Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten
1005   B, Durbin R. 2018. Variation graph toolkit improves read mapping by representing genetic variation in the
1006   reference. *Nat. Biotechnol.* 36:875–879. DOI: 10.1038/nbt.4227.
1007   Gelman A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1:515–533.
1008   DOI: 10.1214/06-BA117A.
1009   Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D. 2013. *Bayesian data analysis*. CRC Press.
1010   Gibson B, Wilson DJ, Feil E, Eyre-Walker A. 2018. The distribution of bacterial doubling times in the wild.
1011   *Proceedings of the Royal Society B: Biological Sciences* 285. DOI: 10.1098/rspb.2018.0789.
1012   Hawkins M, Malla S, Blythe MJ, Nieduszynski CA, Allers T. 2013. Accelerated growth in the absence of DNA
1013   replication origins. *Nature* 503:544–547. DOI: 10.1038/nature12650.
1014   Helmstetter CE, Cooper S. 1968. DNA synthesis during the division cycle of rapidly growing Escherichia coli B/r.
1015   *Journal of Molecular Biology* 31:507–518. DOI: 10.1016/0022-2836(68)90424-5.
1016   Higashi K, Suzuki S, Kurosawa S, Mori H, Kurokawa K. 2018. Latent environment allocation of microbial
1017   community data. *PLoS Comput. Biol.* 14:e1006143. DOI: 10.1371/journal.pcbi.1006143.
1018   Hildebrand F, Nguyen TLA, Brinkman B, Yunta RG, Cauwe B, Vandenabeele P, Liston A, Raes J. 2013.
1019   Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota
1020   variation in common laboratory mice. *Genome Biology* 14:R4. DOI: 10.1186/gb-2013-14-1-r4.
1021   Hoffman MD, Gelman A. 2014. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte
1022   Carlo. *Journal of Machine Learning Research* 15:1593–1623. DOI: abs/1111.4246.
1023   Jeraldo P, Sipos M, Chia N, Brulc JM, Dhillon AS, Konkel ME, Larson CL, Nelson KE, Qu A, Schook LB, Yang F,
1024   White BA, Goldenfeld N. 2012. Quantification of the relative roles of niche and neutral processes in
1025   structuring gastrointestinal microbiomes. *Proceedings of the National Academy of Sciences* 109:9692–9698.
1026   DOI: 10.1073/pnas.1206721109.
1027   Jones MC, Pewsey A. 2005. A family of symmetric distributions on the circle. *J. Am. Stat. Assoc.* 100:1422–1428.
1028   DOI: 10.1198/016214505000000286.
1029   Kato H, Mori H, Maruyama F, Toyoda A, Oshima K, Endo R, Fuchu G, Miyakoshi M, Dozono A, Ohtsubo Y,
1030   Nagata Y, Hattori M, Fujiyama A, Kurokawa K, Tsuda M. 2015. Time-series metagenomic analysis reveals
1031   robustness of soil microbiome against chemical disturbance. *DNA Research* 22:413–424. DOI:
1032   10.1093/dnares/dsv023.
1033   Kato K, Odamaki T, Mitsuyama E, Sugahara H, Xiao J zhong, Osawa R. 2017. Age-related changes in the
1034   composition of gut bifidobacterium species. *Current Microbiology* 74:987–995. DOI: 10.1007/s00284-017-
1035   1272-4.
1036   Korem T, Zeevi D, Suez J, Weinberger A, Avnit-Sagi T, Pompan-Lotan M, Matot E, Jona G, Harmelin A, Cohen N,
1037   Sirota-Madi A, Thaiss CA, Pevsner-Fischer M, Sorek R, Xavier RJ, Elinav E, Segal E. 2015. Growth
1038   dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science*
1039   349:1101–1106. DOI: 10.1126/science.aac4812.
1040   Kullback S, Leibler RA. 1951. On information and sufficiency. *Ann. Math. Stat.* 22:79–86. DOI:
1041   10.1214/aoms/1177729694.
1042   Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis.
1043   *Genomics* 2:231–239. DOI: 10.1016/0888-7543(88)90007-9.
1044   Langenheder S, Szekely AJ. 2011. Species sorting and neutral processes are both important during the initial
1045   assembly of bacterial communities. *Isme Journal* 5:1086–1094. DOI: 10.1038/Ismej.2010.207.
1046   Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357–359. DOI:
1047   10.1038/nmeth.1923.
1048   Leman AR, Noguchi E. 2013. The replication fork: Understanding the eukaryotic replication machinery and the
1049   challenges to genome duplication. *Genes* 4:1–32. DOI: 10.3390/genes4010001.
1050   Li H. 2013.lh3/seqtk. *Available at https://github.com/lh3/seqtk* (accessed December 10, 2018).

1051 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence
1052   alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079. DOI: 10.1093/bioinformatics/btp352.
1053 Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: An ultra-fast single-node solution for large and
1054   complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31:1674–1676. DOI:
1055   10.1093/bioinformatics/btv033.
1056 Li L, Ma ZS. 2016. Testing the neutral theory of biodiversity with human microbiome datasets. *Scientific Reports*
1057   6:1–10. DOI: 10.1038/srep31448.
1058 Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: Estimating species abundance in metagenomics data.
1059   *PeerJ* 2017:1–17. DOI: 10.7717/peerj-cs.104.
1060 Lundgren M, Andersson A, Chen L, Nilsson P, Bernander R. 2004. Three replication origins in Sulfolobus species:
1061   synchronous initiation of chromosome replication and asynchronous termination. *Proceedings of the National*
1062   *Academy of Sciences of the United States of America* 101:7046–51. DOI: 10.1073/pnas.0400656101.
1063 Maduike NZ, Tehranchi AK, Wang JD, Kreuzer KN. 2014. Replication of the Escherichia coli chromosome in
1064   RNase HI-deficient cells: Multiple initiation regions and fork dynamics. *Molecular Microbiology* 91:39–56.
1065   DOI: 10.1111/mmi.12440.
1066 McCarthy S, Gradnigo J, Johnson T, Payne S, Lipzen A, Martin J, Schackwitz W, Moriyama E, Blum P, Julien G,
1067   Tyler J, Payne S, Anna L, Joel M, Wendy S, Etsuko M, Paul B. 2015. Complete genome sequence of
1068   sulfolobus solfataricus strain 98/2 and evolved derivatives. *Genome Announc.* 3:61779. DOI:
1069   10.1128/genomeA.00549-15.
1070 Meurer A, Smith CP, Paprocki M, Čertík O, Rocklin M, Kumar Am, Ivanov S, Moore JK, Singh S, Rathnayake T,
1071   Vig S, Granger BE, Muller RP, Bonazzi F, Gupta H, Vats S, Johansson F, Pedregosa F, Curry MJ, Saboo A,
1072   Fernando I, Kulal S, Cimrman R, Scopatz A, Kirpichev SB, Rocklin M, Kumar Am, Ivanov S, Moore JK,
1073   Singh S, Rathnayake T, Vig S, Granger BE, Muller RP, Bonazzi F, Gupta H, Vats S, Johansson F, Pedregosa
1074   F, Curry MJ, Terrel AR, Roučka Š, Saboo A, Fernando I, Kulal S, Cimrman R, Scopatz A. 2016. SymPy:
1075   Symbolic computing in Python. *PeerJ Computer Science* 3:e103. DOI: 10.7287/peerj.preprints.2083v2.
1076 Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh HJ, Cuenca M, Hingamp P, Alves R, Costea PI, Coelho
1077   LP, Schmidt TSB, Almeida A, Mitchell AL, Finn RD, Huerta-Cepas J, Bork P, Zeller G, Sunagawa S. 2019.
1078   Microbial abundance, activity and population genomic profiling with mOTUs2. *Nature Communications* 10.
1079   DOI: 10.1038/s41467-019-08844-4.
1080 Ohbayashi R, Watanabe S, Ehira S, Kanesaki Y, Chibazakura T, Yoshikawa H. 2016. Diversification of DnaA
1081   dependency for DNA replication in cyanobacterial evolution. *ISME Journal* 10:1113–1121. DOI:
1082   10.1038/ismej.2015.194.
1083 Olm MR, Brown CT, Brooks B, Firek B, Baker R, Burstein D, Soenjoyo K, Thomas BC, Morowitz M, Banfield JF.
1084   2017. Identical bacterial populations colonize premature infant gut, skin, & oral microbiomes & exhibit
1085   different in situ growth rates. *Genome Research* 27:601–612. DOI: 10.1101/gr.213256.116.
1086 Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A standardized
1087   bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*
1088   36:996–1004. DOI: 10.1038/nbt.4229.
1089 Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: Assessing the quality of
1090   microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* 25:1043–1055.
1091   DOI: 10.1101/gr.186072.114.
1092 Payne S, McCarthy S, Johnson T, North E, Blum P, Pattwell SS, Bath KG, Casey BJ, Ninan I, Lee FS, Mcewen BS,
1093   Pattwellab SS, Bathacd KG, Caseya BJ, Ninan I, Leea FS. 2018. Nonmutational mechanism of inheritance in
1094   the archaeon Sulfolobus solfataricus. *Proceedings of the National Academy of Sciences* 115:12271–12276.
1095   DOI: 10.1073/PNAS.
1096 Pelve EA, Martens-Habbena W, Stahl DA, Bernander R. 2013. Mapping of active replication origins in vivo in
1097   thaum- and euryarchaeal replicons. *Molecular Microbiology* 90:538–550. DOI: 10.1111/mmi.12382.
1098 Pewsey A. 2002. Testing circular symmetry. *The Canadian Journal of Statistics* 30:591–600. DOI:
1099   10.2307/3316098.
1100 Pewsey A, Neuhäuser M, Ruxton GD. 2013. *Circular statistics in R*. Oxford University Press.
1101 Retkute R, Hawkins M, Rudolph CJ, Nieduszynski CA. 2018. Modeling of DNA replication in rapidly growing
1102   bacteria with one and two replication origins. *bioRxiv*:1–22. DOI: 10.1101/354654.
1103 Rice P, Longden L, Bleasby A. 2000. EMBOSS: The European molecular biology open software suite. *Trends in*
1104   *Genetics* 16:276–277. DOI: 10.1016/S0168-9525(00)02024-2.
1105 Roach JC. 1995. Random subcloning. *Genome Research* 5:464–473. DOI: 10.1101/gr.5.5.464.

1106   Robinson NP, Dionne I, Lundgren M, Marsh VL, Bernander R, Bell SD. 2004. Identification of two origins of
1107   replication in the single chromosome of the archaeon Sulfolobus solfataricus. *Cell* 116:25–38. DOI:
1108   10.1016/S0092-8674(03)01034-1.
1109   Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative
1110   genomics viewer. *Nature Biotechnology* 29:24–26. DOI: 10.1038/nbt.1754.
1111   Rodriguez-Lopez AM, Jackson DA, Iborra F, Cox LS, Rodríguez-López AM, Jackson DA, Iborra F, Cox LS. 2002.
1112   Asymmetry of DNA replication fork progression in Werner's syndrome. *Aging cell* 1:30–39. DOI:
1113   10.1046/j.1474-9728.2002.00002.x.
1114   Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. 2013. Characterizing
1115   and measuring bias in sequence data. *Genome Biology* 14. DOI: 10.1186/gb-2013-14-5-r51.
1116   Rudolph CJ, Upton AL, Stockum A, Nieduszynski CA, Lloyd RG. 2013. Avoiding chromosome pathology when
1117   replication forks collide. *Nature* 500:608–611. DOI: 10.1038/nature12312.
1118   Ruiz-Moyano S, Totten SM, Garrido DA, Smilowitz JT, Bruce German J, Lebrilla CB, Mills DA. 2013. Variation in
1119   consumption of human milk oligosaccharides by infant gut-associated strains of bifidobacterium breve.
1120   *Applied and Environmental Microbiology* 79:6040–6049. DOI: 10.1128/AEM.01843-13.
1121   Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. DOI:
1122   10.1093/bioinformatics/btu153.
1123   Sernova N V., Gelfand MS. 2008. Identification of replication origins in prokaryotic genomes. *Briefings in
1124   Bioinformatics* 9:376–391. DOI: 10.1093/bib/bbn031.
1125   Tanizawa Y, Fujisawa T, Nakamura Y. 2018. DFAST: A flexible prokaryotic genome annotation pipeline for faster
1126   genome publication. *Bioinformatics* 34:1037–1039. DOI: 10.1093/bioinformatics/btx713.
1127   Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM,
1128   Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab
1129   S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J,
1130   Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman
1131   JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Rivera
1132   JLA, Al-Moosawi L, Alverdy J, Amato KR, Andras J, Angenent LT, Antonopoulos DA, Apprill A, Armitage
1133   D, Ballantine K, Bárta J, Baum JK, Berry A, Bhatnagar A, Bhatnagar M, Biddle JF, Bittner L, Boldgiv B,
1134   Bottos E, Boyer DM, Braun J, Brazelton W, Brearley FQ, Campbell AH, Caporaso JG, Cardona C, Carroll J,
1135   Cary SC, Casper BB, Charles TC, Chu H, Claar DC, Clark RG, Clayton JB, Clemente JC, Cochran A,
1136   Coleman ML, Collins G, Colwell RR, Contreras M, Crary BB, Creer S, Cristol DA, Crump BC, Cui D, Daly
1137   SE, Davalos L, Dawson RD, Defazio J, Delsuc F, Dionisi HM, Dominguez-Bello MG, Dowell R, Dubinsky
1138   EA, Dunn PO, Ercolini D, Espinoza RE, Ezenwa V, Fenner N, Findlay HS, Fleming ID, Fogliano V, Forsman
1139   A, Freeman C, Friedman ES, Galindo G, Garcia L, Garcia-Amado MA, Garshelis D, Gasser RB, Gerdts G,
1140   Gibson MK, Gifford I, Gill RT, Giray T, Gittel A, Golyshin P, Gong D, Grossart H-P, Guyton K, Haig S-J,
1141   Hale V, Hall RS, Hallam SJ, Handley KM, Hasan NA, Haydon SR, Hickman JE, Hidalgo G, Hofmockel KS,
1142   Hooker J, Hulth S, Hultman J, Hyde E, Ibáñez-Álamo JD, Jastrow JD, Jex AR, Johnson LS, Johnston ER,
1143   Joseph S, Jurburg SD, Jurelevicius D, Karlsson A, Karlsson R, Kauppinen S, Kellogg CTE, Kennedy SJ,
1144   Kerkhof LJ, King GM, Kling GW, Koehler A V., Krezalek M, Kueneman J, Lamendella R, Landon EM,
1145   Lane-deGraaf K, LaRoche J, Larsen P, Laverock B, Lax S, Lentino M, Levin II, Liancourt P, Liang W, Linz
1146   AM, Lipson DA, Liu Y, Lladser ME, Lozada M, Spirito CM, MacCormack WP, MacRae-Crerar A, Magris
1147   M, Martín-Platero AM, Martín-Vivaldi M, Martínez LM, Martínez-Bueno M, Marzinelli EM, Mason OU,
1148   Mayer GD, McDevitt-Irwin JM, McDonald JE, McGuire KL, McMahon KD, McMinds R, Medina M,
1149   Mendelson JR, Metcalf JL, Meyer F, Michelangeli F, Miller K, Mills DA, Minich J, Mocali S, Moitinho-Silva
1150   L, Moore A, Morgan-Kiss RM, Munroe P, Myrold D, Neufeld JD, Ni Y, Nicol GW, Nielsen S, Nissimov JI,
1151   Niu K, Nolan MJ, Noyce K, O'Brien SL, Okamoto N, Orlando L, Castellano YO, Osuolale O, Oswald W,
1152   Parnell J, Peralta-Sánchez JM, Petraitis P, Pfister C, Pilon-Smits E, Piombino P, Pointing SB, Pollock FJ,
1153   Potter C, Prithiviraj B, Quince C, Rani A, Ranjan R, Rao S, Rees AP, Richardson M, Riebesell U, Robinson
1154   C, Rockne KJ, Rodriguezl SM, Rohwer F, Roundstone W, Safran RJ, Sangwan N, Sanz V, Schrenk M,
1155   Schrenzel MD, Scott NM, Seger RL, Seguin-Orlando A, Seldin L, Seyler LM, Shakhsheer B, Sheets GM,
1156   Shen C, Shi Y, Shin H, Shogan BD, Shutler D, Siegel J, Simmons S, Sjöling S, Smith DP, Soler JJ, Sperling
1157   M, Steinberg PD, Stephens B, Stevens MA, Taghavi S, Tai V, Tait K, Tan CL, Tas, N, Taylor DL, Thomas T,
1158   Timling I, Turner BL, Urich T, Ursell LK, van der Lelie D, Van Treuren W, van Zwieten L, Vargas-Robles D,
1159   Thurber RV, Vitaglione P, Walker DA, Walters WA, Wang S, Wang T, Weaver T, Webster NS, Wehrle B,
1160   Weisenhorn P, Weiss S, Werner JJ, West K, Whitehead A, Whitehead SR, Whittingham LA, Willerslev E,
1161   Williams AE, Wood SA, Woodhams DC, Yang Y, Zaneveld J, Zarraonaindia I, Zhang Q, Zhao H. 2017. A

1162   communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551:457–463. DOI:
1163   10.1038/nature24621.
1164 Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y, Schoenebeck F, Murphy JA, Zhou S,
1165   Uenoyama Y, Miclo Y, Chemistry A, Building TG. 2017. Synthetic spike-in standards for high-throughput
1166   16S rRNA gene amplicon sequencing. *Nucleic Acids Res.* 45:e23. DOI: 10.1093/nar/gkw984.
1167 Turroni F, Foroni E, Serafini F, Viappiani A, Montanini B, Bottacini F, Ferrarini A, Bacchini PL, Rota C,
1168   Delledonne M, Ottonello S, van Sinderen D, Ventura M. 2011. Ability of Bifidobacterium breve to grow on
1169   different types of milk: Exploring the metabolism of milk through genome analysis. *Applied and
1170   Environmental Microbiology* 77:7408–7417. DOI: 10.1128/AEM.05336-11.
1171 Turroni F, Peano C, Pass DA, Foroni E, Severgnini M, Claesson MJ, Kerr C, Hourihane J, Murray D, Fuligni F,
1172   Gueimonde M, Margolles A, de Bellis G, O'Toole PW, van Sinderen D, Marchesi JR, Ventura M. 2012.
1173   Diversity of bifidobacteria within the infant gut microbiota. *PLoS ONE* 7:20–24. DOI:
1174   10.1371/journal.pone.0036957.
1175 Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J, Tito RY, De Commer
1176   L, Darzi Y, Vermeire S, Falony G, Raes J, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J,
1177   Tito RY, De Commer L, Darzi Y, Vermeire S, Falony G, Raes J. 2017. Quantitative microbiome profiling
1178   links gut community variation to microbial load. *Nature* 551:507–511. DOI: 10.1038/nature24460.
1179 Vieira-Silva S, Rocha EPC. 2010. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS
1180   Genetics* 6. DOI: 10.1371/journal.pgen.1000808.
1181 Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser
1182   W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R,
1183   Larson E, Carey C, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen
1184   I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, Contributors S 1. 0.
1185   2019. SciPy 1.0--Fundamental algorithms for scientific computing in python. *arXiv*:1–22.
1186 Wallden M, Fange D, Lundius EG, Baltekin Ö, Elf J. 2016. The synchronization of replication and division cycles in
1187   individual E. coli cells. *Cell* 166:729–739. DOI: 10.1016/j.cell.2016.06.052.
1188 Watanabe S. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in
1189   singular learning theory. *Journal of Machine Learning Research* 11:3571–3594.
1190 Watanabe S, Ohbayashi R, Shiwa Y, Noda A, Kanesaki Y, Chibazakura T, Yoshikawa H. 2012. Light-dependent
1191   and asynchronous replication of cyanobacterial multi-copy chromosomes. *Molecular Microbiology* 83:856–
1192   865. DOI: 10.1111/j.1365-2958.2012.07971.x.
1193 Wendel BM, Courcelle CT, Courcelle J. 2014. Completion of DNA replication in Escherichia coli. *Proceedings of
1194   the National Academy of Sciences* 111:16454–16459. DOI: 10.1073/pnas.1415025111.
1195 Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biology* 20. DOI:
1196   10.1186/s13059-019-1891-0.
1197 Wu Z, Liu J, Yang H, Liu H, Xiang H. 2014. Multiple replication origins with diverse control mechanisms in
1198   Haloarcula hispanica. *Nucleic Acids Research* 42:2282–2294. DOI: 10.1093/nar/gkt1214.
1199 Xu J, Yanagisawa Y, Tsankov AM, Hart C, Aoki K, Kommajosyula N, Steinmann KE, Bochicchio J, Russ C, Regev
1200   A, Rando OJ, Nusbaum C, Niki H, Milos P, Weng Z, Rhind N. 2012. Genome-wide identification and
1201   characterization of replication origins by deep sequencing. *Genome Biology* 13:R27. DOI: 10.1186/gb-2012-
1202   13-4-r27.
1203 Yang H, Wu Z, Liu J, Liu X, Wang L, Cai S, Xiang H. 2015. Activation of a dormant replication origin is essential
1204   for Haloferax mediterranei lacking the primary origins. *Nature Communications* 6:8321. DOI:
1205   10.1038/ncomms9321.
1206 Yoshikawa H, Wake R. 1993. Initiation and Termination of Chromosome Replication. In: Sonenshein AL, Hoch JA,
1207   Losick R eds. *Bacillus subtilis and Other Gram-Positive Bacteria*. 507–528. DOI:
1208   10.1128/9781555818388.ch36.
1209 Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: A fast phage search tool. *Nucleic Acids
1210   Research* 39:347–352. DOI: 10.1093/nar/gkr485.
1211 Zhu A, Sunagawa S, Mende DR, Bork P. 2015. Inter-individual differences in the gene content of human gut
1212   bacterial species. *Genome Biology* 16:82. DOI: 10.1186/s13059-015-0646-9.
1213

Figure 1

Effects of growth, sequence feature, and outliers on coverage depth shape

We characterized the coverage depth of chromosomal DNA using statistical models. The shape of the probability distributions (solid lines) and artificial coverage depth (blue lines) obtained using the (a) von Mises, (b) cardioid, (c) wrapped Cauchy, (d) Jones-Pewsey, and (e) linear cardioid distribution model with multinomial distribution. Zero (0) was used as a location parameter, while 0.34657 (von Mises and Jones-Pewsey), 0.16666 (cardioid), 0.17157 (wrapped Cauchy), and 0.1061 (linear cardioid) were used as concentration parameters to align the pPTR with 2.0. The nucleotide number was set to 1 Mnt, and the average coverage depth was set to 20× in the multinomial distribution. For the Jones-Pewsey distribution, 0.5 was used as the shape parameter. Sorted shapes of the distributions and pseudo-coverage depths from the (f) von Mises, (g) cardioid, (h) wrapped Cauchy, (i) Jones-Pewsey, and (j) linear cardioid distribution model with multinomial distribution. (b) Coverage depth and sequence features that can cause strong noise in the coverage depth of *L. gasseri* (ERR969426). (k) Overall coverage depth, (l) suspected feature-free region, (m) sorted overall coverage depth, and (n) sorted feature-free region.

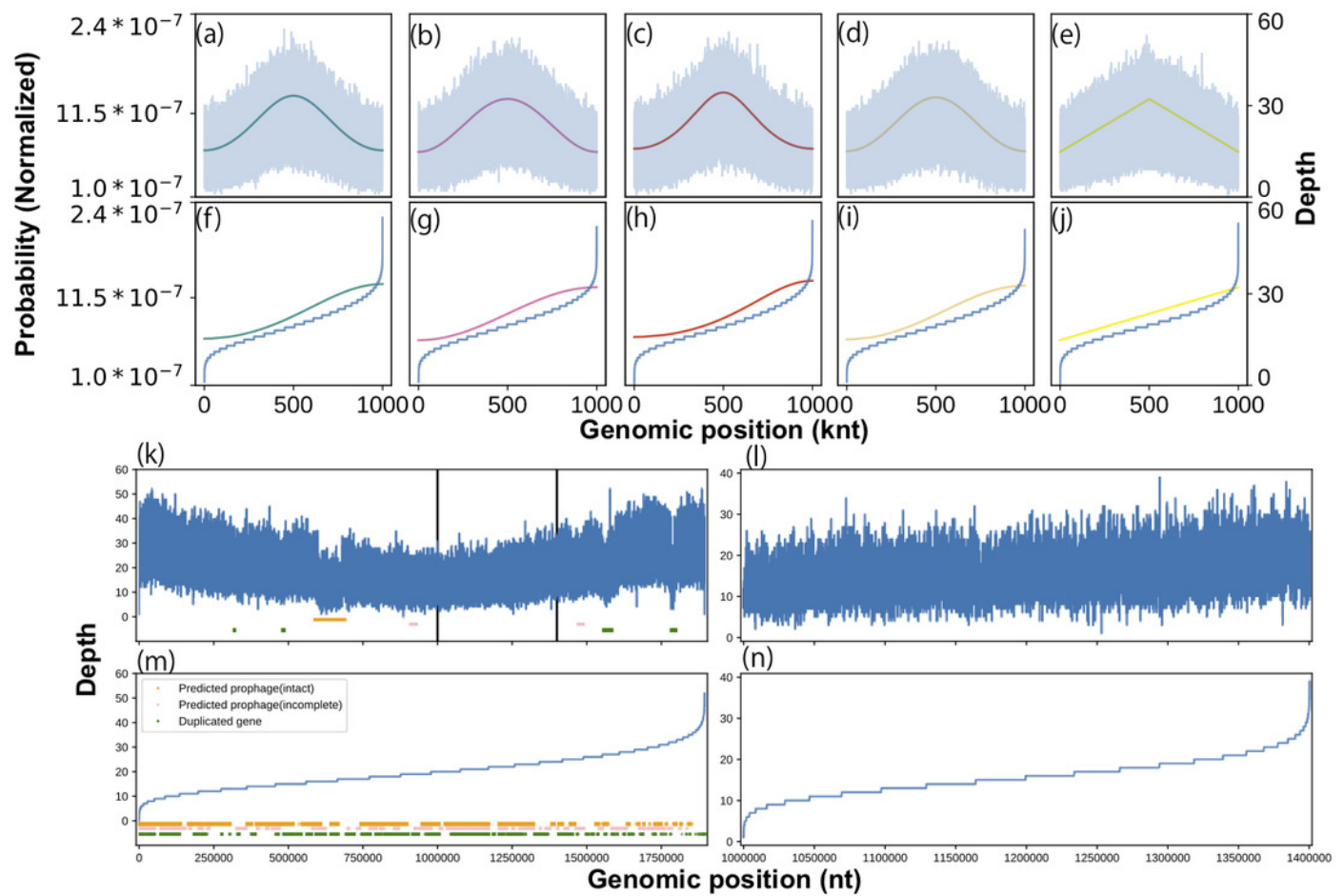



Figure 2

Probabilistic model for generating coverage depth

Summary of procedures and distributions with coverage. (a) Overall flowchart of our method. The green parallelogram represents the data, and the pink rectangle represents the procedure. (b) Coverage depth on the circumference. Focusing on the genome structure of prokaryotes, we developed a model that conducted circular regression. The green and pink plots represent the peak and the trough estimated from the model, respectively. (c) Correlation with experimental growth rate by time series aerobic cultured *E. coli*. pPTR is correlated with the experimental growth rate with a Pearson correlation of 0.964. The dataset was obtained previously (Korem et al., 2015).

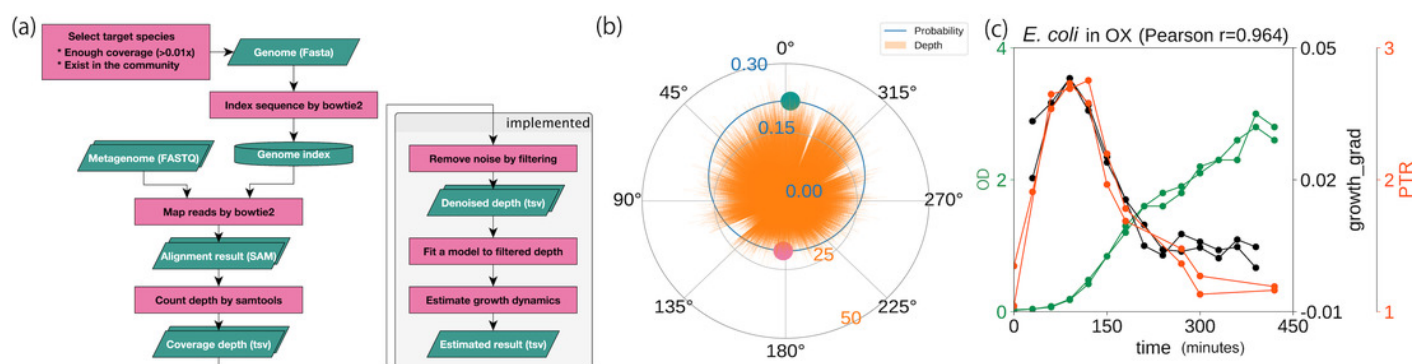



Figure 3

Error rates of growth estimates from various coverage depths with respect to the full coverage depth

The error rates for each average coverage depth were calculated with respect to the full coverage depth. (a) Only the *E. coli*, *E. faecalis*, and *L. gasseri* WGSs with greater than 5.0× coverage, or (b) species with more than 20× coverage in human fecal WGS datasets were used (Franzosa et al., 2018). The horizontal bar represents the 15% threshold of the error rate threshold. The black crosses on the bar represent the unavailability of the methods with respect to the coverage depth. The proposed models and statistics are shown in the black rectangle within the legend.

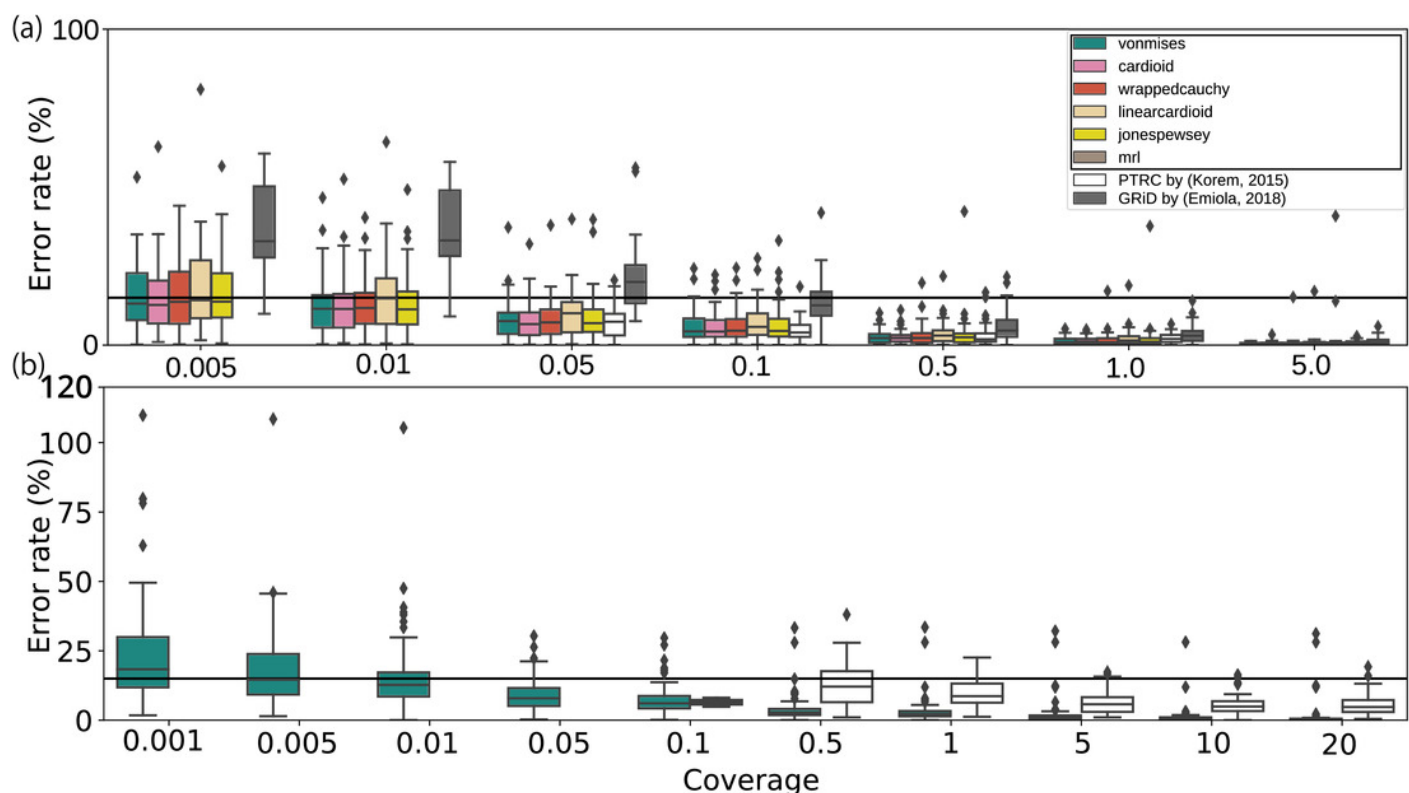



Figure 4

Replication peakedness

The InvSE von Mises distribution-based model exhibits a tapered shape of the coverage depth trend. The Jones-Pewsey distribution-based model also shows a concentrated shape.

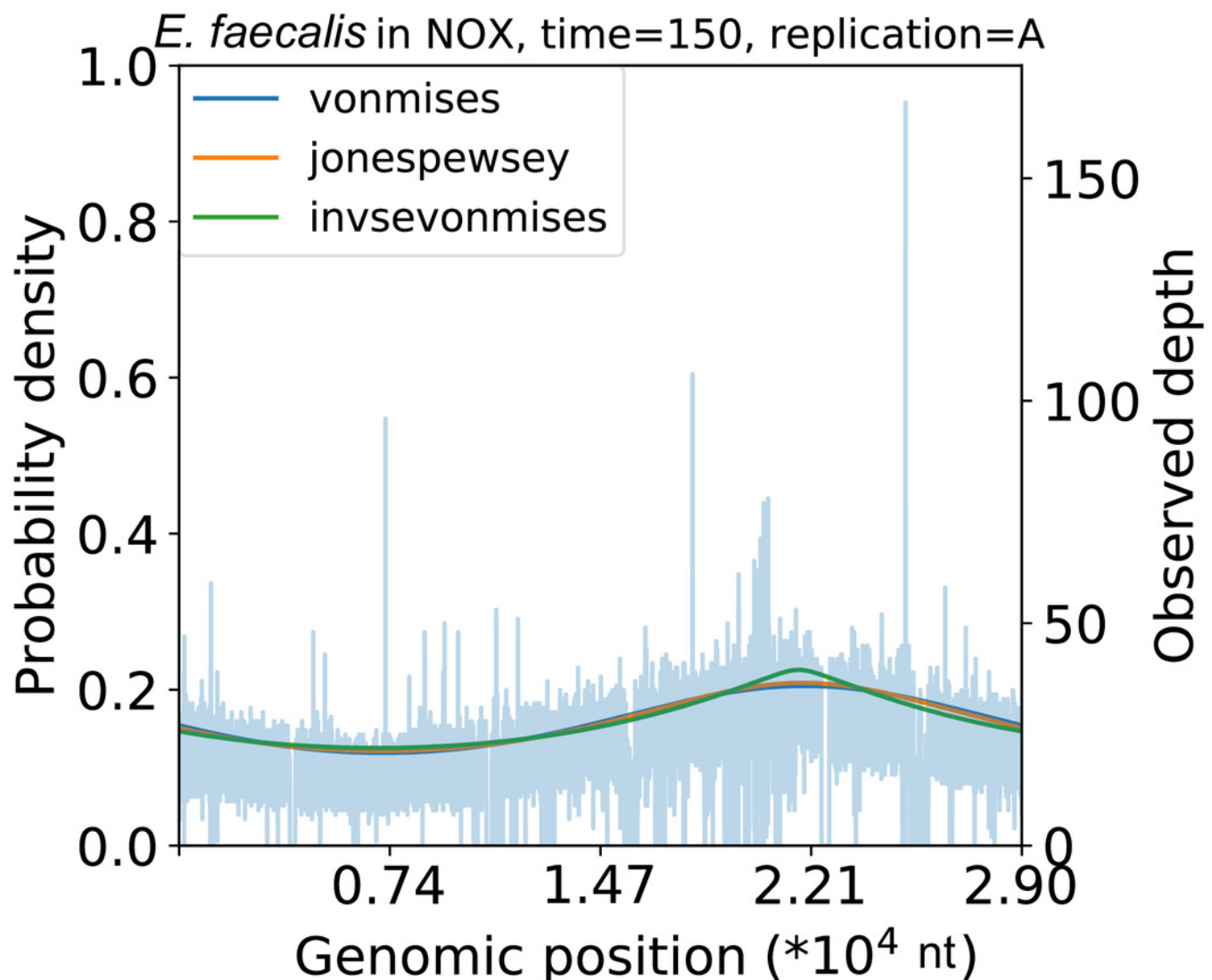
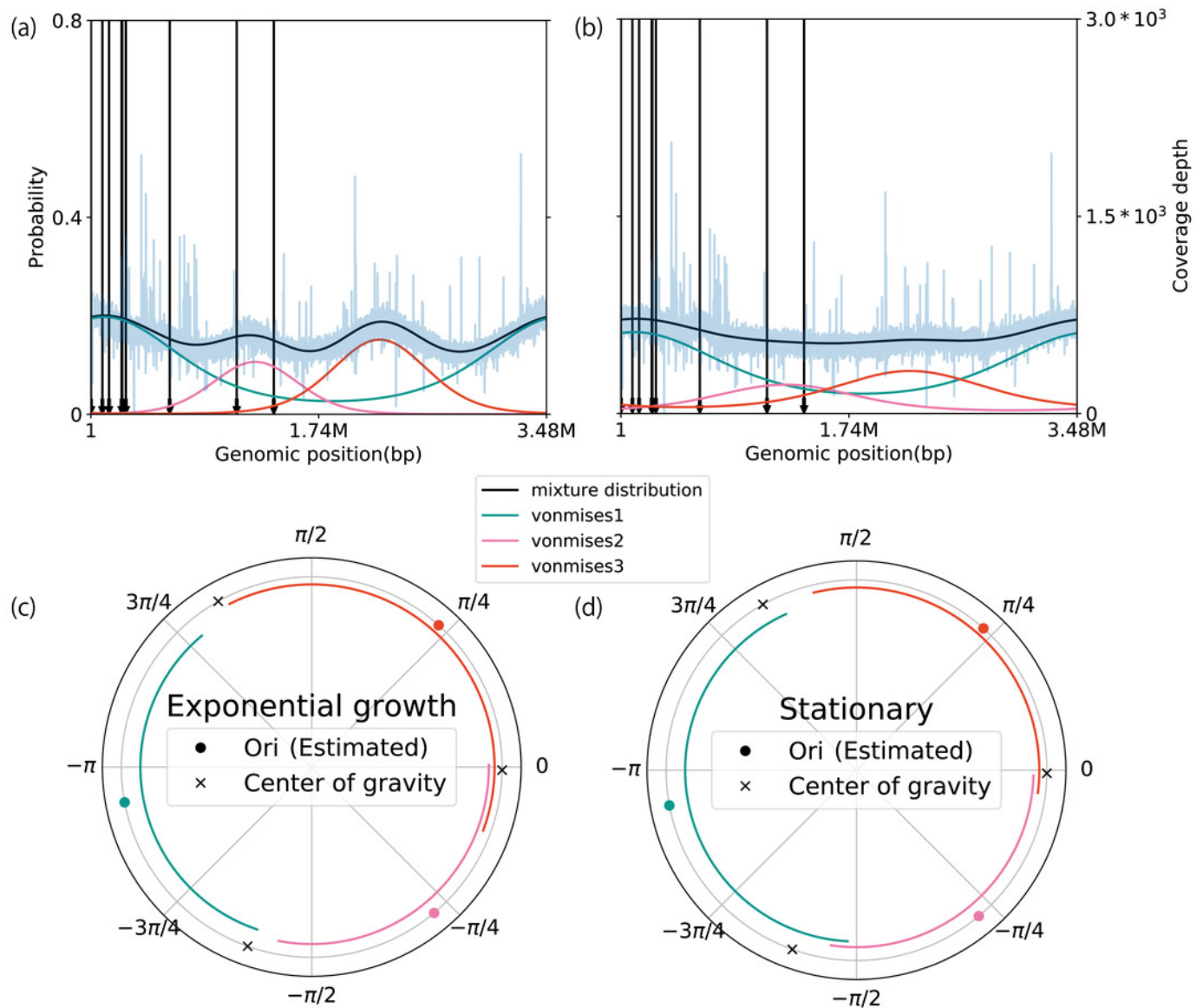



Figure 5

Extension to multiple replication origins

The plots in the upper row represent the coverage depth and probability distributions estimated using a mixture of the von Mises distribution models for *Haloferax volcanii* in the (a) exponential growth phase, and (b) stationary phase. The blue lines represent the coverage depth processed using a median filter with 100 nt for both the stride and window length. The black arrows indicate the position of *cdc6* in the genome. The lines inside the lower circular plot express the magnitude of the weighted pPTR for *H. volcanii* in the (c) exponential growth phase and (d) stationary phase. The circles represent positions of replication origins, and the crosses represent the positions of the centers of gravity of the replication origins. The lines inside the circles represent the relative magnitudes of the weighted pPTR.

3891

1 Table1: Circular distributions for modeling coverage depth dynamics. μ: location parameter, κ or

2 ρ: concentration parameter, ψ: shape parameter, I_0: modified Bessel function of the first kind of

3 order 0, $P_{\frac{1}{\psi}}$: the associated Legendre function of the first kind of degree $\frac{1}{\psi}$, θ: an angle converted

4 from the observed position, T: total number of observations.

Name	(Probability density) function	pPTR	Parameters	Ref.
von Mises (vM)	$\dfrac{\exp(\kappa\cos(\theta-\mu))}{2\pi I_0(\kappa)}$	$\exp(2\kappa)$	μ, κ	
Cardioid (c)	$\dfrac{1+2\rho_c\cos(\theta-\mu)}{2\pi}$	$\dfrac{1+2\rho_c}{1-2\rho_c}$	μ,ρ_c	
Wrapped Cauchy (wC)	$\dfrac{1-\rho_{wC}^2}{2\pi\left(1+\rho_{wC}^2-2\rho_{wC}\cos(\theta-\mu)\right)}$	$\dfrac{(1+\rho_{wC})^2}{(1-\rho_{wC})^2}$	μ,ρ_{wC}	
Jones-Pewsey (JP)	$\dfrac{(\cosh(\kappa\psi)+\sinh(\kappa\psi)\cos(\theta))^{\frac{1}{\psi}}}{2\pi P_{\frac{1}{\psi}}(\cosh(\kappa\psi))}$	$\exp(2\kappa)$	μ,κ,ψ	(Jones & Pewsey, 2005)
Linear cardioid (lc)	$\dfrac{1+2\rho_{lc}\left(\lvert\lvert\theta-\mu\rvert-\pi\rvert-\frac{\pi}{2}\right)}{2\pi}$	$\dfrac{1+\pi\rho_{lc}}{1-\pi\rho_{lc}}$	μ,ρ_{lc}	
Exponential linear cardioid (elc)	$\dfrac{\rho_{elc}\exp\left(2\rho_{elc}\lvert\lvert\theta-\mu\rvert-\pi\rvert-\frac{\pi}{2}\right)}{\exp(\pi\rho_{elc})-\exp(-\pi\rho_{elc})}$	$\exp(2\pi\rho_{elc})$	μ,ρ_{elc}	
Mean resultant length (mrl)	$\dfrac{\sqrt{\left(\sum_{t=1}^{T}\cos\theta_t\right)^2+\left(\sum_{t=1}^{T}\sin\theta_t\right)^2}}{T}$			