# Supplemental Notes S1-S3: Accurate and robust inference of microbial growth dynamics from metagenomic sequencing reveals personalized growth rates

Tyler A. Joseph[1], Philippe Chlenski[1], Aviya Litman[2], Tal Korem*[2,3,4], and Itsik Pe'er *[1,2,5]

[1] *Department of Computer Science, Columbia University, New York, NY, USA*
[2] *Department of Systems Biology, Columbia University Irving Medical Center, New York, NY, USA*
[3] *Department of Obstetrics and Gynecology, Columbia University Irving Medical Center, New York, NY, USA*
[4] *CIFAR Azrieli Global Scholars Program, CIFAR, Toronto, Ontario, Canada*
[5] *Data Science Institute, Columbia University, New York, NY, USA*

## S1 Supplemental Note: PTRs measure DNA replication and generation time

Here we adapt an argument from Bremer and Churchward (1977) to show that PTRs provide information about DNA replication and generation time. We need two parameters

1. $C$: The time required to replicate the bacterial chromosome after replication begins.
2. $\tau$: The generation time.

### S1.1 PTRs under exponential growth

In microbiology, the generation time is equivalent to be the population doubling time under exponential growth. Under exponential growth, and given an initial population size $N_0$, the population size at time $t$ is given by

$$N(t) = N_0 2^{t/\tau},$$

the generation time $\tau$ is equivalent to the cell doubling time, and is inversely proportional to the log population growth rate $\frac{\ln 2}{\tau}$.

Most bacteria have a single circular chromosome. Replication begins at a single replication origin, and two replication forks move in either direction along the chromosome until the replication terminus. Under the above model Bremer and Churchward (1977) showed that that the ratio of average number of replication origins in a cell $\bar{I}$ to average replication terimini $\bar{T}$

$$\bar{I}/\bar{T} = 2^{C/\tau}.$$

Therefore

$$\log_2\left(\bar{I}/\bar{T}\right) = \frac{C}{\tau}.$$

The term $\frac{1}{\tau}$ is called the growth rate. Thus, under exponential growth $\log_2(\text{PTR})$ is proportional to the growth rate.

*These authors contributed equally to this work. Correspondence: tal.korem@columbia.edu, itsik@cs.columbia.edu.

### S1.2 Generalization to other growth phases

We want to show this hold in general—regardless of the form of $N(t)$. The crux of the argument is to use the PTR to compute the rate that a population is adding genomes—the rate of DNA synthesis. Given the rate of DNA synthesis, computing the generation time involves asking the counterfactual question—if I were to grow a population of cells in culture with genomes being added at a specified rate, how long would it take for the population to double in size? Solving population doubling time also gives us generation time because of their equivalence under exponential growth.

The argument has three steps. First, we will compute the average rate of DNA synthesis using the PTR. Second, we will fix the rate of DNA synthesis to compute how many genomes will be added over an interval of time. Third, given the number of genomes added over an interval we can solve for the generation time using its equivalence to population doubling time under exponential growth. In principle, we are reducing the more complicated case of arbitrary dynamics to the simpler case of exponential growth.

#### S1.2.1 PTRs are correlated with average rate of DNA synthesis

Let $I(t)$ be the number of replication origins at time $t$, and let $T(t)$ be the number of replication termini. Once DNA replication begins, the two strands of the chromosome separate, and replication forks proceed along both sides of the chromosome. This means $T(t)$ gives the number of complete genomes in the population, since once replication finishes a new completed genome, and therefore terminus, has been added. It follows that the number of replication forks at time $t$ is given by

$$2(I(t) - T(t)).$$

The rate of chromosome replication is essentially constant (Wang and Levin, 2009), so each fork produces DNA at a rate of $\sim \frac{1}{2C}$. Therefore, the rate of DNA synthesis in the population is

$$\frac{1}{C}\left(I(t) - T(t)\right).$$

Since $T(t)$ gives the number of complete genomes, dividing by $T(t)$ gives the rate of DNA synthesis per genome

$$\frac{\frac{1}{C}\left(I(t) - T(t)\right)}{T(t)} = \frac{1}{C}\left(R(t) - 1\right)$$
$$R(t) := \frac{I(t)}{T(t)} = \frac{\bar{I}(t)}{\bar{T}(t)}.$$

$R(t)$ corresponds to the PTR. This demonstrates that $R(t)$ is correlated with the average rate of DNA synthesis.

#### S1.2.2 Counting genomes added over a fixed time

Given a particular time $t_0$ we want to compute how many genomes will be added over an interval. The term $\frac{1}{C}(R(t_0) - 1)$ says that each genome at $t_0$ adds $R(t_0) - 1$ genomes over $C$ time. Hence, the number of genomes added over $[t_0, t_0 + C]$ is equal to the current number of genomes, $T(t_0)$, plus the number of genomes added, $T(t_0)(R(t_0) - 1)$. Therefore

$$\#\text{ genomes added over } [t_0, t_0 + C] = T(t_0) + T(t_0)(R(t_0) - 1) = T(t_0)R(t_0)$$

Note that the equation says nothing about genomes removed during this period.

### S1.2.3 Computing the generation time

Now we want to find the generation time. Since generation time is equivalent to population doubling time under exponential growth, we need to compute the population doubling time given a fixed rate of DNA synthesis. Intuitively, we are asking the question—if I transplanted this population at time $t_0$ from its current setting to one of unrestricted growth, what would be the population doubling time? In other words, we want to know how long it would take for the number of genomes to double, $\tau(t_0)$, given a fixed $R(t_0)$ computed from time $t_0$. Treating the number of genomes as continuous, we can write our equation for exponential growth as follows:

$$T_{t_0}\left(t_0 + \frac{t}{C}\right) = T(t_0)R(t_0)^{t/C} \tag{1}$$

Note the dependence on $t_0$: this quantity must be recomputed for each $t_0$. Thus we can show that the $\log_2(\text{PTR})$ is inversely proportional to generation time $\tau(t_0)$:

$$2T(t_0) = T(t_0)R(t_0)^{\tau(t_0)/C} \implies \log_2 R(t_0) = \frac{C}{\tau(t_0)}$$

A consequence is that for any choice of $t$ under exponential growth

$$T_t(t) = T(t_0)R(t_0)^{t/C}.$$

Taking the derivative of $\log_2 T_t$ we get

$$\frac{d}{dt}\log_2 T_t(t) = \frac{1}{C}\log_2 R(t_0).$$

In this specific case, $R(t)$ corresponds to changes in population size.

## S1.3 Points of departure from Bremer and Churchward (1977)

There are key conceptual differences between our argument and Bremer and Churchward (1977). Bremer and Churchward (1977) start with an assumption of the form $N(t)$, and derive expressions for $I(t)$ and $T(t)$ using the parameters $\tau$, $C$, and an additional parameter $D$ that measures the time between genome replication and cell division. Specifically, they assume

$$\begin{aligned} N(t) &= N_0 2^{t/\tau} \\ I(t) &= I_0 2^{t/\tau} \\ T(t) &= T_0 2^{t/\tau} \end{aligned}$$

Under exponential growth each of these quantities count the same thing but shifted in time, so

$$\begin{aligned} I_0 &= N_0 2^{(C+D)/\tau} \\ T_0 &= N_0 2^{D/\tau} \end{aligned}$$

In contrast, here we do not want to assume a specific form for $N(t)$, $I(t)$ and $T(t)$. In general, we want $I(t)$ and $T(t)$ to arbitrary, and reflect some underlying model of dynamics. We derive an expression for DNA synthesis under arbitrary $I(t)$ and $T(t)$. We solve for $\tau(t)$ by applying using its equivalent to population doubling time under exponential growth.

### S1.4 Comparison with Lotka-Volterra dynamics

Suppose we have a community of $D$ species. Let $N_i(t)$ be the abundance of species $i$, $g_i$ be its growth rate, and $A_{ij}$ be the effect of species $j$ on species $i$. The generalized Lotka-Volterra equation for population dynamics states

$$\frac{d}{dt}N_i(t) = N_i(t)\left(g_i + \sum_{j=1}^{D} A_{ij}N_j(t)\right).$$

Using the identity $\frac{d}{dt}\log N_i(t) = \frac{d/dt\{N_i(t)\}}{N_i(t)}$ we can write.

$$\frac{d}{dt}\log N_i(t) = g_i + \sum_{j=1}^{D} A_{ij}N_j(t).$$

If we set $A_{ij}$ to 0, the equation reduces to exponential growth. Let $R_i(t_0)$ be the PTR for species $i$ at some time $t_0$, and $C_i$ the time it takes the species to replicate its chromosome. Since PTRs give us the derivative of log population size under exponential growth, assuming a Lotka-Volterra model we have

$$g_i \approx \frac{1}{C_i}\log R_i(t_0). \tag{2}$$

Hence,

$$\frac{d}{dt}\log N_i(t) = \frac{1}{C_i}\log R_i(t_0) + \sum_{j=1}^{N} A_{ij}N_j(t).$$

Thus, $\log R_i(t)$ is similar to the growth rate parameter $g_i$ of Lotka-Volterra, assuming Lotka-Volterra accurately describes community dynamics.

However, there are 3 key differences. First, Lotka-Volterra models have a fixed growth rate $g_i$, so they do not model changes in growth over time. In contrast $\log R_i(t)$ varies over time. Second, $\log R_i(t)$ may not be proportional to changes in abundance $\frac{d}{dt}\log N_i(t)$ because of the additional interaction terms $A_{ij}$. Therefore, we should not expect PTRs to be predictive of changes in abundance. Third, Lotka-Volterra models assume changes in abundance occur only due to growth and species interactions. In contrast, PTRs are model free—measuring growth regardless of the underlying model.

## S2 Supplemental Note: Modeling the density of reads along the genome

The results of the previous section demonstrated that $\log_2(\bar{I}/\bar{T}) = \frac{C}{\tau}$, where $\bar{I}$ is the average number of copies of the replication origin in a population, and $\bar{T}$ is the average number of copies of the replication terminus (we have dropped the explicit dependence on $t$ for notation). Suppose we are interested in the ratio of the average copies of an arbitrary position $A$ along the chromosome to the replication terminus. Let $\bar{A}$ be the average copies of $A$, and let $C_A$ be the time it takes the replication fork to move from $A$ to the replication terminus. Note that before the replication fork crosses $A$ there is only one copy, and after the fork crosses $A$ there are two copies. Thus, replacing $I$ with $A$ and $C$ with $C_A$ in the previous argument shows that

$$\frac{1}{C_A}(A - T)$$

also gives the rate of DNA synthesis. Therefore

$$\log_2(\bar{A}/\bar{T}) = \frac{C_A}{\tau}$$

If we assume that chromosome replication happens at a constant rate along the genome, then $C_A$ depends on the distance from $A$ to the replication terminus. Let $b$ be the shortest number of bases between the origin and $A$, such that if we move from origin to the $A$ we do not need to cross the terminus. Let $d$ be the number of bases from the origin to the terminus, and define $f = b/d$. Then $C_A = C(1 - f)$. Rearranging terms from above, we have

$$\begin{aligned}\log_2(\bar{T}) &= \log_2(\bar{I}) - \frac{C}{\tau}\\ \log_2(\bar{T}) &= \log_2(\bar{A}) - \frac{C_A}{\tau} = \log_2(\bar{A}) - \frac{C(1-f)}{\tau}\end{aligned}$$

Subtracting the first equation from the second, and rearranging terms

$$\implies \log_2(\bar{A}) = \log_2(\bar{I}) - \frac{Cf}{\tau}$$

Consequentially, the average copies of position $A$ decays log-linearly with distance from the replication origin.

This also means that any probabilistic model of reads along the genome, coverage should decay log-linearly away from the replication origin. Therefore, we propose the following model. Let $[0, 1]$ represent coordinates along a continuous approximation of a reference genome. Thus 0 is the beginning of the reference, and 1 is the end. The model parameters are the origin position $x_i$, terminus position $x_t = (x_i + 0.5) \bmod 1$, and PTR $r$. We want a probability density given by

$$\begin{aligned}\alpha &= \frac{\log_2 r}{x_i - x_t} = \frac{\log_2 p(x_i) - \log_2 p(x_t)}{x_i - x_t}\\ x_1 &= \min\{x_i, x_t\}\\ x_2 &= \max\{x_i, x_t\}\\ c(x) &= \begin{cases}\log_2 p(x_i) \text{ if } x = x_i\\ \log_2 p(x_t) \text{ if } x = x_t\end{cases}\\ \log_2 p(x) &= \begin{cases}-\alpha(x - x_1) + c(x_1) & \text{if } x \leq x_1\\ \alpha(x - x_1) + c(x_1) & \text{if } x_1 < x < x_2\\ -\alpha(x - x_2) + c(x_2) & \text{if } x \geq x_2\end{cases}\end{aligned}$$

We need to compute $\log p(x_i)$ and $\log p(x_t)$ such that

$$\int_0^1 2^{\log_2 p(x)} = 1.$$

We can use the integral, and the constraint that $\log_2 p(x_i) - \log_2 p(x_t) = \log_2 r$ to solve for each. There are two cases. If $x_i \leq x_t$, then

$$\log_2 p(x_i) = \log_2 \ln 2 - \log_2 \left( \frac{1}{\alpha} \left[ 2^{\alpha x_1} + 2^{\alpha(x_2 - x_1)} - 2^1 - 2^{-\alpha(1-x_2) - \log_2 r} + 2^{-\log_2 r} \right] \right)$$
$$\log_2 p(x_t) = \log_2 p(x_i) - \log_2 r$$

If $x_t < x_i$, then

$$\log_2 p(x_t) = \log_2 \ln 2 - \log_2 \left( \frac{1}{\alpha} \left[ 2^{\alpha x_1} + 2^{\alpha(x_2 - x_1)} - 2^1 - 2^{-\alpha(1-x_2) + \log_2 r} + 2^{+\log_2 r} \right] \right)$$
$$\log p_2(x_i) = \log_2 p(x_t) + \log_2 r$$

## S3 Supplemental Note: Variational inference for multi-mapped reads

Suppose we have the following model for drawing the assignment of sequencing reads from a set of $\mathcal{G}$ reference genomes indexed from $1...g$.

1. Draw probabilities that a read originates from a reference genome:

$$\pi \sim \text{Dirichlet}(\alpha_1, ...., \alpha_g).$$

2. For each read $i = 1...n$, pick a reference genome:

$$z_i | \pi \sim \text{Categorical}(\pi)$$

The prior for a genome $\alpha_j$ is set to the number of reads the map unambiguously to that genome. For notation, let $z_i$ be an indicator vector, where $z_{ij} = 1$ if $z_i$ is assigned to genome $j$. Let $x_i = (x_{i1}, x_{i2}, ..., x_{ig}) \in \{0, 1\}^g$ where $x_{ij} = 1$ if the read maps to a position in genome $j$, and is 0 otherwise. If read $i$ maps to only one genome, then $z_i = x_i$. If read $i$ maps to multiple genomes, then $z_i$ places a restriction on $x_i$: if $z_{ij} = 1$ then it must be true that $x_{ij} = 1$—assuming that one of the given mappings is always correct. Thus, we can model $x_{ij}$ as

$$p(x_{ij} = 1 | z_{ij}) = \begin{cases} 1 & \text{if } z_{ij} = 1 \\ \rho_{ij} & \text{otherwise.} \end{cases}$$

Now consider that once a sequencing read is observed, all valid mappings are determined. Hence $\rho_{ij} = 1$ if read $i$ has a valid mapping to genome $j$ and 0 otherwise. Therefore

$$p(x_i | z_i) = \prod_{j=1}^{g} x_{ij}^{z_{ij}}$$

if we define $0^0 = 1$.

If we could compute

$$p(z_{1:n}, \pi | x_{1:n})$$

then we could assign reads to genomes based on the posterior. However, the normalizing constant is

$$p(x_{1:n}) = \int_{\pi} \sum_{z_{1:n}} p(x_{1:n}, z_{1:n}, \pi)\, d\pi \tag{3}$$

which requires summing over an exponential number of combinations of $z_{1:n}$. Nonetheless, we can compute

$$p(z_i|\pi, x_i) = \frac{\prod_{j:x_{ij}=1} \pi_j^{z_{ij}}}{\sum_{j:x_{ij}=1} \pi_j}$$

Additionally, since the Dirichlet distribution is a conjugate prior for the multinomial distribution

$$p(\pi|z_{1:n}, x_{1:n}) = p(\pi|z_{1:n}) = \text{Dirichlet}\left(\pi; \alpha + \sum_{i=1}^{n} z_i\right)$$

Therefore we can compute all of the complete conditionals. This means we can approximate $p(z_{1:n}, \pi|x_{1:n})$ either using Gibbs sampling or mean field variational inference (Blei et al., 2017). We chose variational inference.

Variational inference approximates an intractable posterior one by a tractable one $q$ whose parameters are optimized to minimize the a lower bound on the log-likelihood. Equivalently, variational inference minimizes Kullback-Leibler divergence between the true posterior and the approximation. For the mean field approximation, $q(z_{1:n}, \pi) = q(\pi) \prod_{i=1}^{n} q(z_i)$. The optimal choice for an approximation $q(z_i)$ is given by (see Blei et al. (2017))

$$\begin{aligned} q(z_i) &\propto \exp\left\{\mathbb{E}_{-z_i}\left[\log p(z_i|\pi, x_i)\right]\right\} \\ &\propto \exp\left\{\sum_{j:x_{ij}=1} z_{ij}\, \mathbb{E}_{-z_i}[\log \pi_j]\right\} \end{aligned}$$

This set of equations give the natural parameters of a multinomial distribution, so $q(z_i) = \text{Multinomal}(z_i; 1, \phi_i)$ where $\log \phi_{ij} = \mathbb{E}_{-z_i}[\log \pi_j] + \mathit{const}$ if $x_{ij} = 1$, and $\phi_{ij} = 0$ otherwise. The optimal choice for $q(\pi)$ is given by

$$\begin{aligned} q(\pi) &\propto \exp\left\{\mathbb{E}_{-\pi}\left[\log p(\pi|z_{1:n})\right]\right\} \\ &\propto \exp\left\{\mathbb{E}_{-\pi}\left[\sum_{i=1}^{n} \log p(z_i|\pi)\right] + \log p(\pi)\right\} \\ &\propto \exp\left\{\sum_{i=1}^{n} \sum_{j:x_{ij}=1} \phi_j \log \pi_j + \sum_{j=1}^{g} (\alpha_j - 1) \log \pi_j\right\} \\ &\propto \exp\left\{\sum_{i=1}^{n} \sum_{j=1}^{g} \phi_j \log \pi_j + \sum_{j=1}^{g} (\alpha_j - 1) \log \pi_j\right\} \\ &= \text{Dirichlet}\left(\pi; \alpha + \sum_{i=1}^{n} \phi_i\right) \end{aligned}$$

The second set of equations gives the natural parameters of a Dirichlet distribution, so $q(\pi) = \text{Dirichlet}(\pi; \eta)$. Now we can compute for the final expectation:

$$\mathbb{E}_{-z_i}[\log \pi_j] = \Psi(\eta_j) - \Psi\left(\sum_{j'=1}^{g} \eta_{j'}\right).$$

where $\Psi$ is the Digamma function.

The final component is computing the variational objective function. This is given by

$$L(z_{1:n}, x_{1:n}, \pi; \phi_{i:n}, \eta) = \mathbb{E}_q[\log p(z_{1:n}, x_{1:n}, \pi)] - \sum_{i=1}^{n} \mathbb{E}_q[\log q(z_i; \phi_i)] - \mathbb{E}_q[\log q(\pi; \eta)]$$

The second two terms are the entropy of a multinomial and Dirichlet distribution respectively. The joint model likelihood is

$$p(z_{1:n}, x_{1:n}, \pi) = p(\pi) \prod_{i=1}^{n} \prod_{j=1}^{g} x_{ij}^{z_{ij}} \pi_j^{z_{ij}}$$

Hence

$$\mathbb{E}_q[\log p(z_{1:n}, x_{1:n}, \pi)] = \mathbb{E}_q[\log p(\pi)] + \sum_{i=1}^{n} \sum_{j=1}^{g} \mathbb{E}_q[z_{ij}] x_{ij} + \mathbb{E}_q[z_{ij}]\, \mathbb{E}_q[\log \pi_j]$$

We now can define an inference procedure for computing the approximate posterior.

1. Initialize variational parameters $\phi_{1:n}$ and $\eta$.
2. While $L(z_{1:n}, x_{1:n}, \pi; \phi_{i:n}, \eta)$ has not converged:
    (a) Set $q(z_i) \propto \exp\left\{\sum_{j:x_{ij}=1} z_{ij}\, \mathbb{E}_{-z_i}[\log \pi_j]\right\}$ for $i = 1...n$
    (b) Set $q(\pi) = \text{Dirichlet}\left(\pi; \alpha + \sum_{i=1}^{n} \phi_i\right)$

The $q(z_i)$ define an approximate posterior over $z_i$. Each read $i$ is assigned to a genome $j$ with the largest posterior probability. If $f(j) = \phi_{ij} = q(z_i = j)$, then we assign read $i$ to $\text{argmax}_{j=1...g}\, f(j)$.

# References

Blei DM, Kucukelbir A, and McAuliffe JD. 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**: 859–877.

Bremer H and Churchward G. 1977. An examination of the cooper-helmstetter theory of dna replication in bacteria and its underlying assumptions. *Journal of theoretical biology* **69**: 645–654.

Wang JD and Levin PA. 2009. Metabolism, cell growth and the bacterial cell cycle. *Nature Reviews Microbiology* **7**: 822–827.