The Nubeam reference-free approach to analyze metagenomic sequencing reads (Supplemental information)

Hang Dai and Yongtao Guan

Department of Biostatistics and Bioinformatics
Duke University School of Medicine

July 24, 2020

1 Supplemental Material and Methods

1.1 Simulation of genomes and inferences

We simulated genome for each taxonomic unit in the designed trees, with the root genome of *E. coli* strain K-12 substrain MG1655 (4,641,652 bp). We used iSG (Strope et al., 2006) for simulation under the general time-reversible (GTR) substitution model. The 6 parameters for relative substitution rate are as those estimated from the evolution study of bacteria rrsA, and 4 parameters for nucleotide frequency are those estimated from the E. coli genome. The branch lengths represent number of substitutions per basepair per generation. For the three-taxonomic-unit trees, both branch lengths of S1 and S2 had fixed values of 1×10^{-4}, while the branch length of S3 had values span two orders of magnitude from 10^{-2} to 10^{-4}. For the eight-taxonomic-unit trees, we used six combinations of internal and terminal branch lengths of 1×10^{-5} and 5×10^{-5} to represent different difficulty levels of phylogeny inference.

Each simulation was done 100 replicates. For three-taxnomic-tree we only interested in relative distance between taxa. For eight-taxonomic-unit tree, we used hierarchical clustering (with Wards minimum variance method) to reconstruct trees from distance matrices produced by both Nubeam and the k-mer method. We compared the reconstructed trees with true trees by Compare2Trees (Nye et al., 2005), generating a score ranging from 0 to 1 for each comparison, with 0 indicating no topological similarity and 1 indicating same topology.

1.2 Bin partitioning

The following algorithm does the balanced partition of bins.

Algorithm 1 Balanced partition of bins using conditioning quantiles

```
Combine the matrices for all samples
Partition 1st column into BINs using quantiles
for each BIN of 1st dimension do
    for each ROW do
        collect the ROW if its 1st NUMBER is in BIN
    end for
    for the collected rows, partition their 2nd column into BINs usingquantiles
    for each BIN of 2nd dimension do
        for each ROW of the collected rows do
            collect the ROW if its 2nd NUMBER is in BIN
        end for
        for the collected rows, partition their 3rd column into BINs using quantiles
        ...
        for each BIN of $(m-1)$th dimension do
            for each ROW of the collected rows do
                collect the ROW if its $(m-1)$th NUMBER is in BIN
            end for
            for the collected rows, partition their $m$-th column into BINs using quantiles
        end for
    end for
end for
```

1.3 $\boldsymbol{k}$-mer frequency based method is a special case of Nubeam

In k-mer frequency based methods, empirical probability distributions of k-mers are generated for the sequencing samples to be compared. For a determined k, consider set $\mathcal{K}$ as the set of all possible k-mers determined by the four nucleotides. The number of all possible k-mers $|\mathcal{K}|$ is 4^k. For i_{th} sequencing sample with n collected k-mers, let $N_w^{(i)}$ be the frequency of a given k-mer w. The relative frequency, or empirical probability, of w in the sample is $f_w^{(i)} = \hat{\mathbb{P}}_i(w) = \frac{N_w^{(i)}}{\sum_{r\in\mathcal{K}} N_r^{(i)}} = \frac{N_w^{(i)}}{n}$. The empirical probability distributions of k-mers in i_{th} and j_{th} sequencing sample are:

$$\hat{\mathbb{P}}_i = \begin{bmatrix} f_{w_1}^{(i)} & f_{w_2}^{(i)} & f_{w_3}^{(i)} & \ldots & f_{w_{|\mathcal{K}|}}^{(i)} \end{bmatrix}$$

$$\hat{\mathbb{P}}_j = \begin{bmatrix} f_{w_1}^{(j)} & f_{w_2}^{(j)} & f_{w_3}^{(j)} & \ldots & f_{w_{|\mathcal{K}|}}^{(j)} \end{bmatrix}$$

Note that each k-mer is treated as a single category, so f_w do not need to be ordered in a specific way, but to compare the two distributions f_w must be ordered in a same way for two samples. If a k-mer w is absent from a sequencing sample, then $f_w = 0$. Next the two distributions are compared using distance or dissimilarity measures.

In Nubeam, each unique k-mer w is represented by an unique vector of numbers $\boldsymbol{y}_w = \begin{bmatrix} y_{1w} & y_{2w} & y_{3w} & y_{4w} \end{bmatrix}$, $\boldsymbol{y}_w \in \mathbb{R}^4$. The i_{th} sequencing sample with n collected k-mers is represented by an $n \times 4$ matrix:

$$\boldsymbol{y_i} = \begin{bmatrix} \boldsymbol{y_{i1}} & \boldsymbol{y_{i2}} & \boldsymbol{y_{i3}} & \boldsymbol{y_{i4}} \end{bmatrix} = \begin{bmatrix} y_{i11} & y_{i21} & y_{i31} & y_{i41} \\ y_{i12} & y_{i22} & y_{i32} & y_{i42} \\ \vdots & \vdots & \vdots & \vdots \\ y_{i1n} & y_{i2n} & y_{i3n} & y_{i4n} \end{bmatrix}$$

The $\boldsymbol{y_i}$ could be viewed as an 4-dimensional multivariate distribution of random variables Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, with probability distribution function (pdf) of $f_{Y_{i1},Y_{i2},Y_{i3},Y_{i4}}(y_{i1}, y_{i2}, y_{i3}, y_{i4})$. Here each $\boldsymbol{y}_w$ is a data point and has a mass of $\frac{1}{n}$. Divide $\mathbb{R}^4$ into a total of b disjoint sets, i.e., a total of b 4-dimensional hyper-rectangle bins $\mathcal{B}_1$, $\mathcal{B}_2$, $\ldots$, $\mathcal{B}_b$. Define the empirical probability distribution by

$$\hat{\mathbb{P}}_i = \begin{bmatrix} f_{\mathcal{B}_1}^{(i)} & f_{\mathcal{B}_2}^{(i)} & f_{\mathcal{B}_3}^{(i)} & \ldots & f_{\mathcal{B}_b}^{(i)} \end{bmatrix}$$

with $f_{\mathcal{B}} = \hat{\mathbb{P}}(\mathcal{B}) = \frac{\sum_{w=1}^{n} I_{\mathcal{B}}(\boldsymbol{y}_w)}{n}$. Next we quantify the genetic distance between the i_{th} sample and j_{th} sample by comparing the two empirical probability distributions using distance or dissimilarity measures.

When binning of $\mathbb{R}^4$ is fine enough such that $\max_{\forall \mathcal{B}\in\{\mathcal{B}_1,\mathcal{B}_1,\ldots,\mathcal{B}_b\}} |\mathcal{B}| = 1$, Nubeam would be identical with k-mer frequency based method.

PROOF. We need to prove that for each of the samples to be compared, the vectors representing the empirical probability distributions in k-mer frequency based method

$\hat{\mathbb{P}}_{k-mer} = \begin{bmatrix} f_{w_1} & f_{w_2} & f_{w_3} & \dots & f_{w_{|\mathcal{K}|}} \end{bmatrix}$ and Nubeam $\hat{\mathbb{P}}_{Nubeam} = \begin{bmatrix} f_{\mathcal{B}_1} & f_{\mathcal{B}_2} & f_{\mathcal{B}_3} & \dots & f_{\mathcal{B}_b} \end{bmatrix}$ are equivalent.

Each unique k-mer w is represented by an unique vector of numbers $\boldsymbol{y}_w$. For any specific k-mer r, suppose $\boldsymbol{y}_r \in \mathcal{B}_r$. Since

$$\max_{\forall \mathcal{B} \in \{\mathcal{B}_1, \mathcal{B}_1, \dots, \mathcal{B}_b\}} |\mathcal{B}| = 1$$

we must have $|\mathcal{B}_r| = 1$. Hence,

$$\begin{aligned} \frac{\sum_{w=1}^{n} I_{\mathcal{B}_r}(\boldsymbol{y}_w)}{n} &= \frac{N_r}{n} \\ \Rightarrow \hat{\mathbb{P}}(\mathcal{B}_r) &= \hat{\mathbb{P}}(r) \\ \Rightarrow f_{\mathcal{B}_r} &= f_r \end{aligned}$$

For k-mer frequency based method, in $\hat{\mathbb{P}}_{k-mer} = \begin{bmatrix} f_{w_1} & f_{w_2} & f_{w_3} & \dots & f_{w_{|\mathcal{K}|}} \end{bmatrix}$, f_w do not need to be ordered in a specific way; f_w only need to be ordered in a same way for two samples to be compared. In Nubeam, to meet this requirement, for each of the samples to be compared, we only need to divide $\mathbb{R}^m$ into bins in a same way and order these bins in a same way. In practice, if the binning is a function of data, we can combine the samples to be compared and then do the binning. Hence, for each of the samples to be compared, $\hat{\mathbb{P}}_{k-mer} \equiv \hat{\mathbb{P}}_{Nubeam}$. $\square$

Thus, k-mer frequency based method is a special case of Nubeam, which is a more general idea.

2 Supplemental Figures

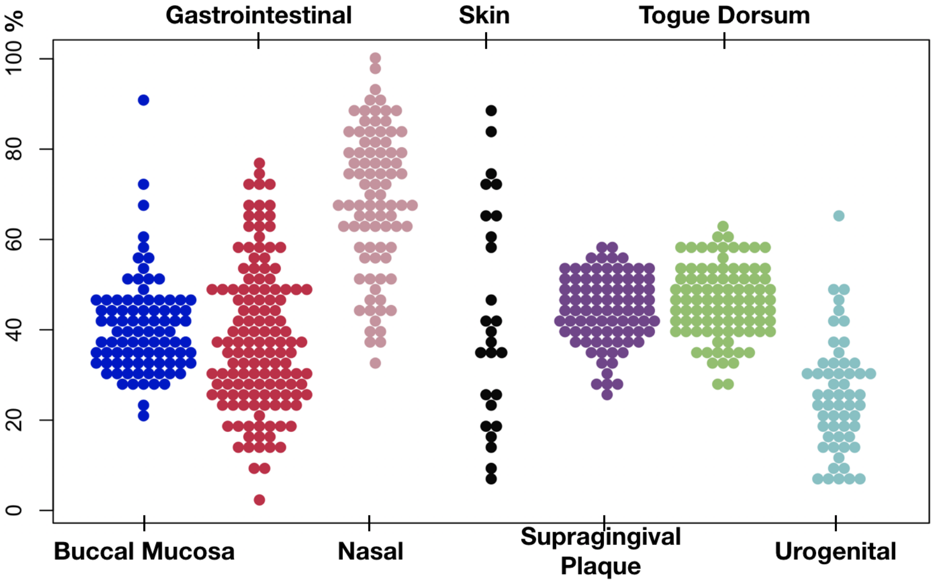


Figure S1: Percent of unmapped reads by body sites. Nasal samples have highest average unmapped reads, and gastrointestinal and urogenital samples have the lowest. Skin samples have large spread of unmapped reads.

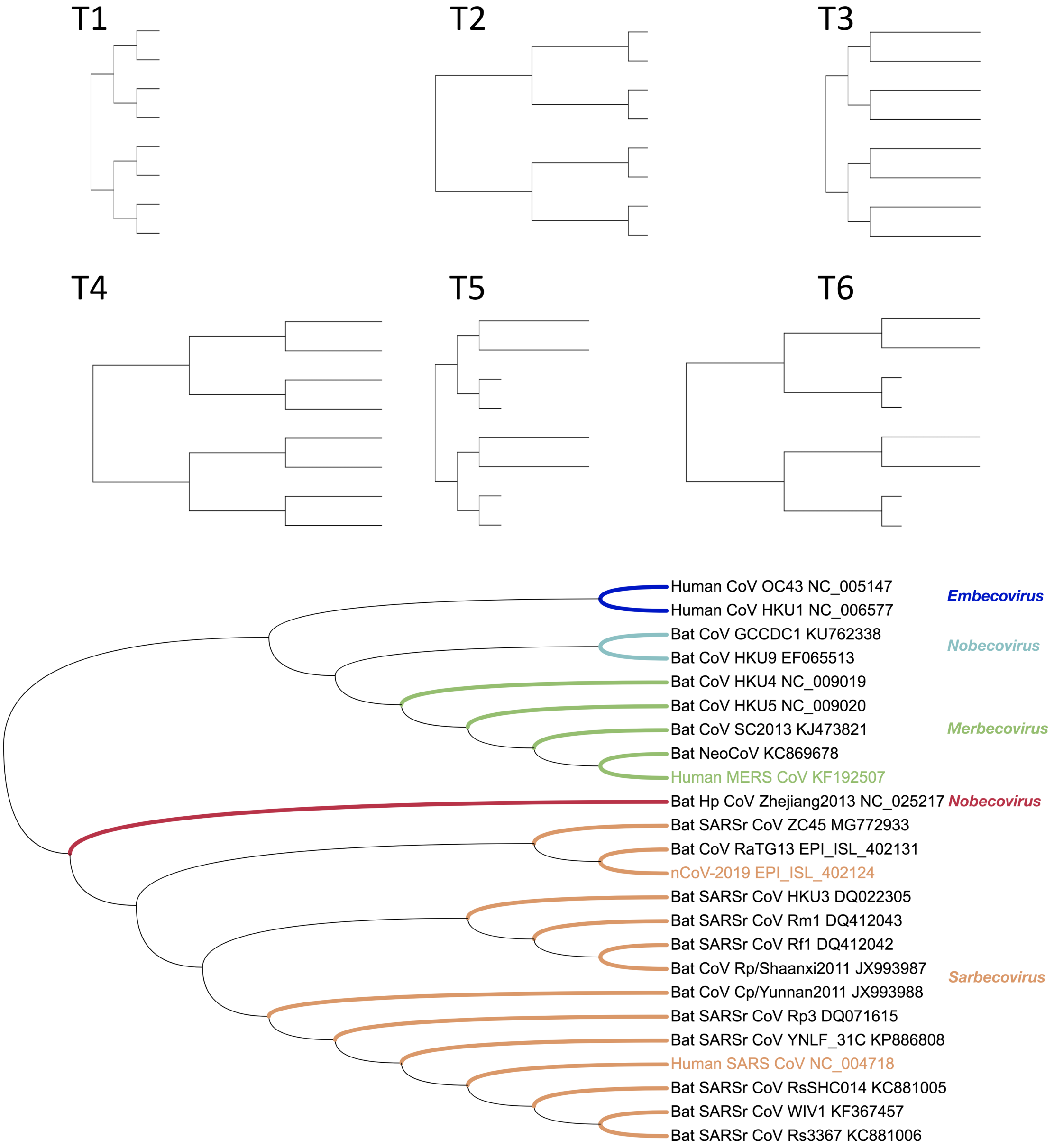

T1
T2
T3
T4
T5
T6
Human CoV OC43 NC_005147
Human CoV HKU1 NC_006577
Embecovirus
Bat CoV GCCDC1 KU762338
Bat CoV HKU9 EF065513
Nobecovirus
Bat CoV HKU4 NC_009019
Bat CoV HKU5 NC_009020
Bat CoV SC2013 KJ473821
Merbecovirus
Bat NeoCoV KC869678
Human MERS CoV KF192507
Bat Hp CoV Zhejiang2013 NC_025217
Nobecovirus
Bat SARSr CoV ZC45 MG772933
Bat CoV RaTG13 EPI_ISL_402131
nCoV-2019 EPI_ISL_402124
Bat SARSr CoV HKU3 DQ022305
Bat SARSr CoV Rm1 DQ412043
Bat SARSr CoV Rf1 DQ412042
Bat CoV Rp/Shaanxi2011 JX993987
Sarbecovirus
Bat CoV Cp/Yunnan2011 JX993988
Bat SARSr CoV Rp3 DQ071615
Bat SARSr CoV YNLF_31C KP886808
Human SARS CoV NC_004718
Bat SARSr CoV RsSHC014 KC881005
Bat SARSr CoV WIV1 KF367457
Bat SARSr CoV Rs3367 KC881006

Figure S2: Top: Six combinations of internal and terminal branch lengths of 0.0001 and 0.0005 were used to represent different degrees of genome divergence and thus different difficulty levels of phylogeny inference. Bottom: Nubeam clustered the beta-coronavirus complete genomes according to sub-genera, correctly grouped human SARS and MERS coronavirus with closest bat strains at the genome level and recapitulated the findings that the genome of bat coronavirus RaTG13 is highly similar to that of nCoV-2019. The clustering was presented using iTOL (Letunic and Bork, 2019).

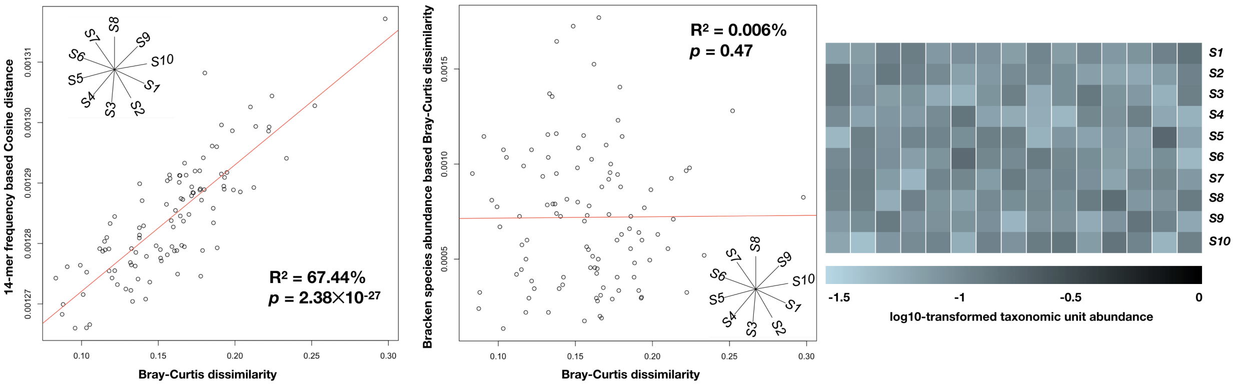


Figure S3: The dissimilarities based on 14-mer frequency Lu et al. (2017b) and Bracken-estimated species abundance Lu et al. (2017a) have weaker linear relationship with composition-based dissimilarities among synthetic communities. The significance of linear relationship is measured by R^2 and p-value for regression coefficient.

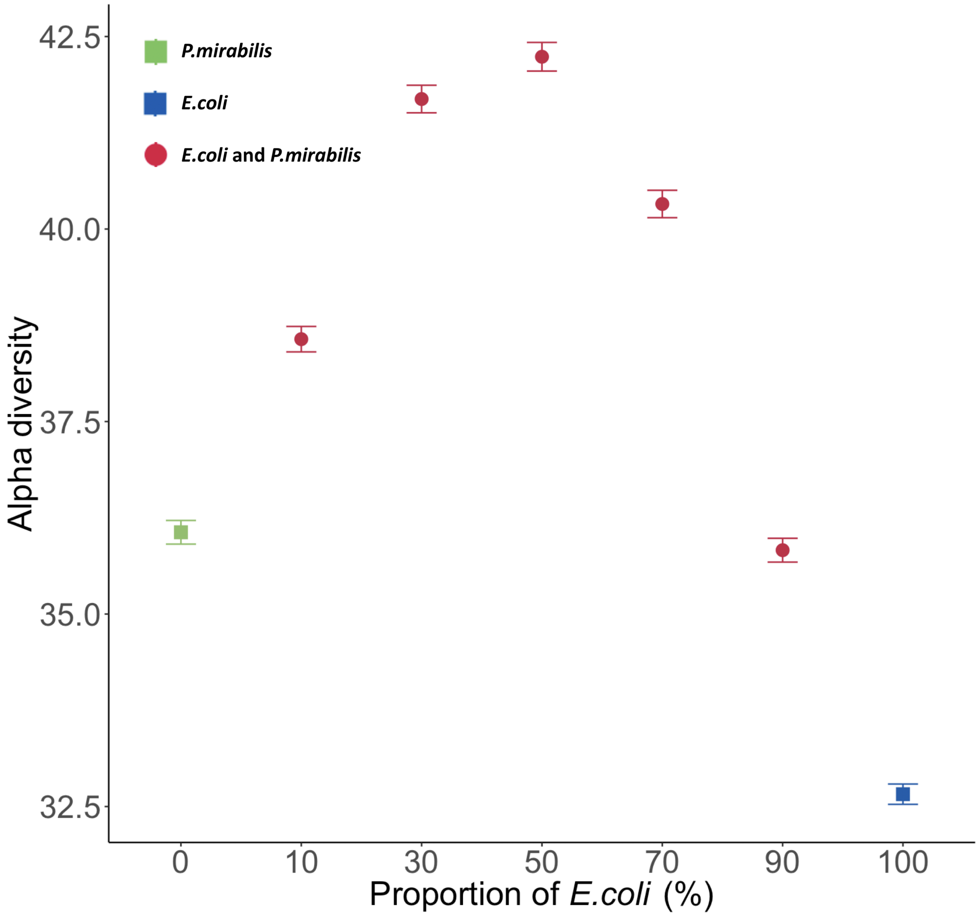


Figure S4: Within-sample diversity quantified by Nubeam using simulated samples that mixing reads at different proportions from *E. coli* and *P. mirabilis*. The bars on each dot denote ± one standard error.

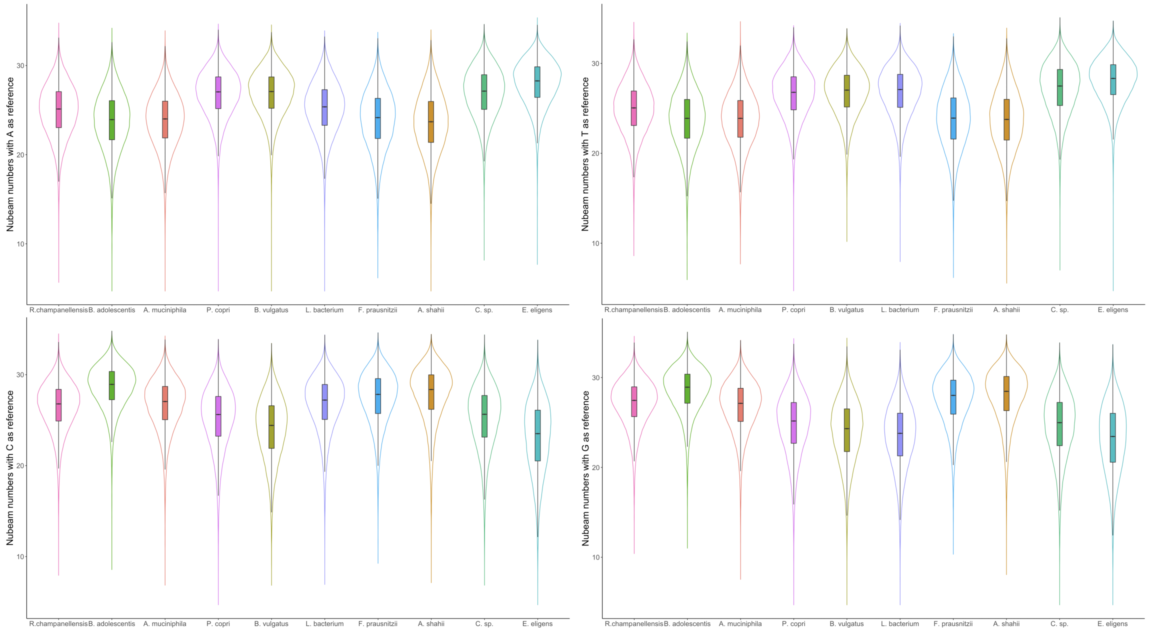
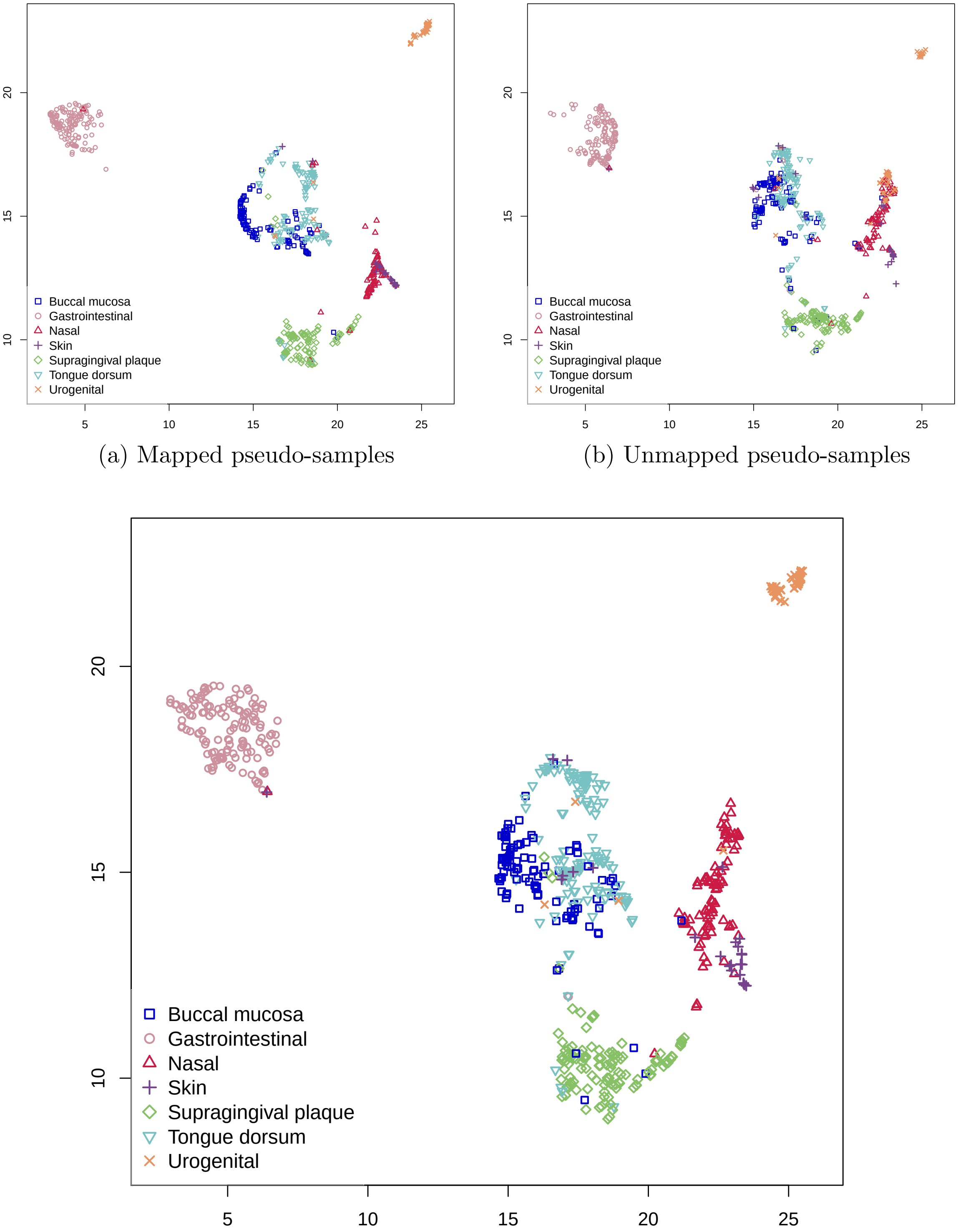


Figure S5: The distributions of Nubeam numbers for the 10 species. The species are ordered according to their individual within-sample diversity, as in Supplemental Table S3.



(a) Mapped pseudo-samples

(b) Unmapped pseudo-samples

(c) Wholesome samples

Figure S6: Samples clustering for mapped pseudo-samples, unmapped pseudo-samples, and wholesome samples respectively by UMAP.

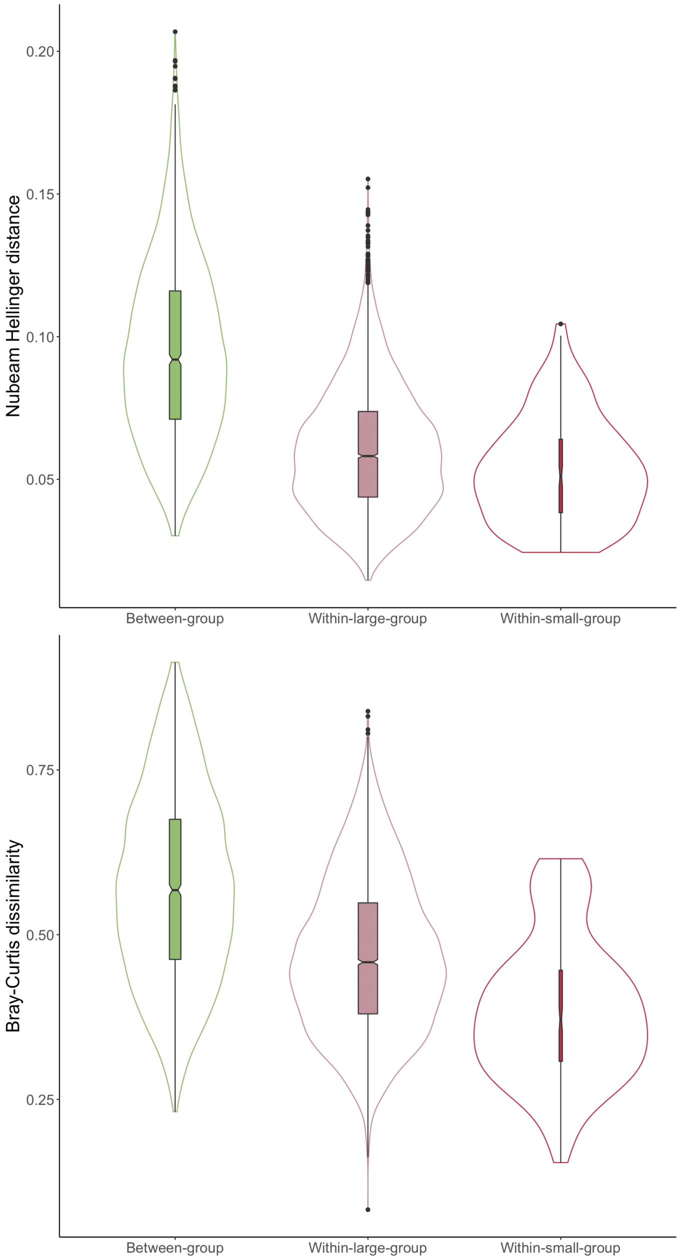
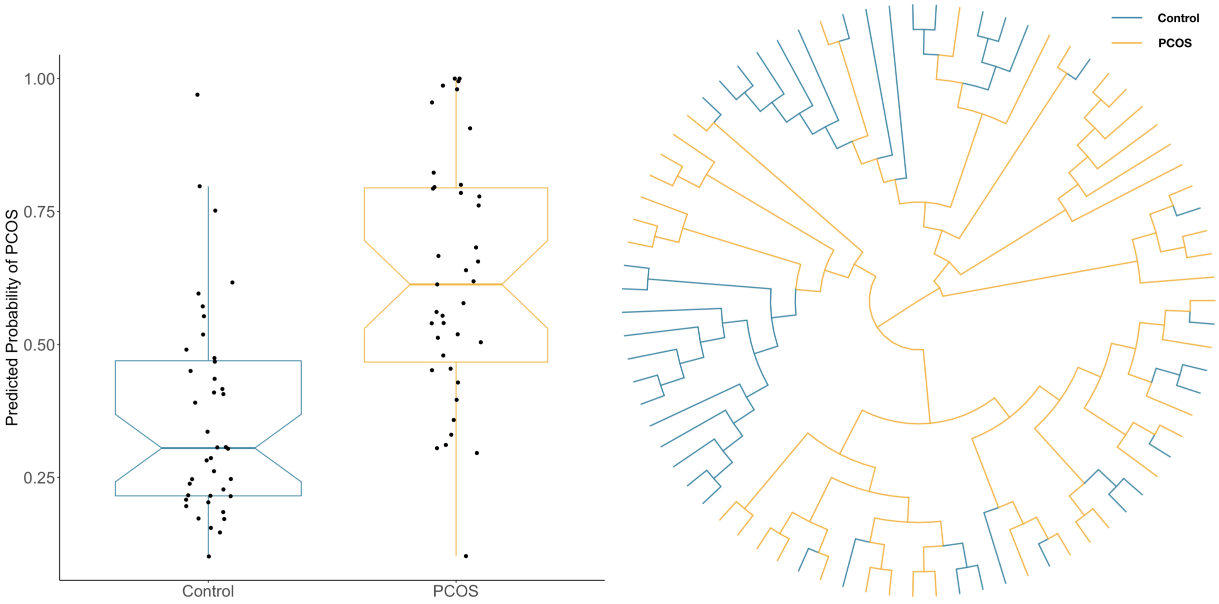


Figure S7: In Figure 6a of the main text, there are two groups of supragingival plaque mapped pseudo-samples, designated here as large and small groups, with the small group close to nasal samples. Both Nubeam and reference-based analyses showed significant difference between the compositions of the two groups: there exists significant differences among two within-group Nubeam distances and one between-group Nubeam distances (top, Kruskal-Wallis test p-value of 5×10^{-329}); we quantified beta-diversity for mapped pseudo-samples using Bray-Curtis dissimilarity based on genus abundance from HMP project, there is also significant difference among two within-group dissimilarities and one between-group dissimilarities (bottom, Kruskal-Wallis test p-value of 3×10^{-154}).



(a)

(b)

Figure S8: a: MDS was applied on Nubeam distance matrix; a logistic regression model based on the resulted first six principal coordinates was built; the predicted probability of the sample being collected from individual with PCOS for the two groups are significant (Kruskal-Wallis test p-value $= 9.443 \times 10^{-7}$). b: Hierarchical clustering (Wards minimum variance method) of samples based on Nubeam distance matrix.

3 Supplemental Tables

Tools	CPU time (s)	Memory footprint (GB)
Nubeam	10,806	27.9
CAFE 14-mer	55,388	20.7

Table S1: Runtime and memory footprint after applying Nubeam and CAFE Lu et al. (2017b) (based on 14-mer frequency) on 15 simulated WGS metagenomic samples; each sample has 50 million 75bp reads. Our benchmarking computer has two 2.3 GHz Intel Xeon E5-2699 v3 CPUs, 256 GB RAM, and running Red Hat Enterprise Linux Server 7.0.

Component	Phylum	Class	Genus
R. champanellensis	Firmicutes	Clostridia	*Ruminococcus*
B. adolescentis	Actinobacteria	Actinobacteria	*Bifidobacterium*
A. muciniphila	Verrucomicrobia	Verrucomicrobiae	*Akkermansia*
P. copri	Bacteroidetes	Bacteroidia	*Prevotella*
B. vulgatus	Bacteroidetes	Bacteroidia	*Bacteroides*
L. bacterium	Firmicutes	Clostridia	*Lachnospiraceae*
F. prausnitzii	Firmicutes	Clostridia	*Faecalibacterium*
A. shahii	Bacteroidetes	Bacteroidia	*Alistipes*
C. sp.	Firmicutes	Clostridia	*Coprococcus*
E. eligens	Firmicutes	Clostridia	*Eubacterium*

Table S2: 10 components for synthetic communities.

Community identifier	*R. champanellensis*	*B. adolescentis*	*A. muciniphila*	*P. copri*	*B. vulgatus*	*L. bacterium*	*F. prausnitzii*	*A. shahii*	*C. sp.*	*E. eligens*	Nubeam	Simpson	Shannon
1.1	1	0	0	0	0	0	0	0	0	0	31.519	1	0
1.2	0	1	0	0	0	0	0	0	0	0	33.263	1	0
1.3	0	0	1	0	0	0	0	0	0	0	35.954	1	0
1.4	0	0	0	1	0	0	0	0	0	0	38.394	1	0
1.5	0	0	0	0	1	0	0	0	0	0	38.536	1	0
1.6	0	0	0	0	0	1	0	0	0	0	38.859	1	0
1.7	0	0	0	0	0	0	1	0	0	0	39.563	1	0
1.8	0	0	0	0	0	0	0	1	0	0	41.533	1	0
1.9	0	0	0	0	0	0	0	0	1	0	41.967	1	0
1.10	0	0	0	0	0	0	0	0	0	1	43.468	1	0
2.1	$\frac{1}{2^1}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^9}$	$\frac{1}{2^9}$	38.943 ± 0.016	3	0.6
2.2	$\frac{1}{2^9}$	$\frac{1}{2^1}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^9}$	45.747 ± 0.011	3	0.6
2.3	$\frac{1}{2^9}$	$\frac{1}{2^9}$	$\frac{1}{2^1}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	44.931 ± 0.013	3	0.6
2.4	$\frac{1}{2^8}$	$\frac{1}{2^9}$	$\frac{1}{2^9}$	$\frac{1}{2^1}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	42.614 ± 0.009	3	0.6
2.5	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^9}$	$\frac{1}{2^9}$	$\frac{1}{2^1}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	45.796 ± 0.013	3	0.6
2.6	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^9}$	$\frac{1}{2^9}$	$\frac{1}{2^1}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	47.904 ± 0.010	3	0.6
2.7	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^9}$	$\frac{1}{2^9}$	$\frac{1}{2^1}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	46.924 ± 0.013	3	0.6
2.8	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^9}$	$\frac{1}{2^9}$	$\frac{1}{2^1}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	52.425 ± 0.017	3	0.6
2.9	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^9}$	$\frac{1}{2^9}$	$\frac{1}{2^1}$	$\frac{1}{2^2}$	47.046 ± 0.012	3	0.6
2.10	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^9}$	$\frac{1}{2^9}$	$\frac{1}{2^1}$	52.670 ± 0.010	3	0.6
3.1	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^8}$	43.073 ± 0.010	4.8	0.75
3.2	$\frac{1}{2^8}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	46.948 ± 0.013	4.8	0.75
3.3	$\frac{1}{2^8}$	$\frac{1}{2^8}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	45.372 ± 0.010	4.8	0.75
3.4	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^8}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	45.738 ± 0.012	4.8	0.75
3.5	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^8}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	49.527 ± 0.019	4.8	0.75
3.6	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^8}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	49.754 ± 0.012	4.8	0.75
3.7	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^8}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	51.389 ± 0.013	4.8	0.75
3.8	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^8}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	53.701 ± 0.017	4.8	0.75
3.9	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^8}$	$\frac{1}{2^2}$	$\frac{1}{2^2}$	48.698 ± 0.012	4.8	0.75
3.10	$\frac{1}{2^2}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^5}$	$\frac{1}{2^6}$	$\frac{1}{2^7}$	$\frac{1}{2^8}$	$\frac{1}{2^8}$	$\frac{1}{2^2}$	49.728 ± 0.014	4.8	0.75
4.1	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	49.210 ± 0.015	10	1

Table S3: In the table, each line indicates a simulated community composition; the proportion p_i of each component in the community is listed below the component. The components are arranged according to their individual Nubeam within-sample diversities (see "communities" 1.1–1.10). Except the 'communities' composed of only one component, each community has its 1×10^7 reads sequencing sample simulated 10 times, with proportions adjusted according to the genome lengths of the components (Jia et al., 2013). For each community, the Nubeam within-sample diversity is compared with the reference based alpha diversity measures—inverse Simpson's index ($\frac{1}{\sum_i p_i^2}$) and Shannon index ($-\sum_i p_i \log_{10} p_i$). On one hand, communities with smaller reference based measures generally have smaller Nubeam within-sample diversities; in particular, among 10 pairs of communities with similar compositions (communities 2.i and 3.i), the community with smaller reference based measures in a pair has smaller Nubeam within-sample diversity for all but one pair (2.10 and 3.10). On the other hand, communities with same reference based measures but different compositions can have different Nubeam within-sample diversities.

Body habitats	Correlation coefficient of Mantel's test	p-value
All	0.6883	$< 1 \times 10^{-4}$
Buccal mucosa	0.4473	$< 1 \times 10^{-4}$
Gastrointestinal	0.3652	9×10^{-4}
Nasal	0.3805	$< 1 \times 10^{-4}$
Skin	0.4297	8×10^{-4}
Supragingival plaque	0.6908	$< 1 \times 10^{-4}$
Tongue dorsum	0.6224	$< 1 \times 10^{-4}$
Urogenital	0.2032	2.44×10^{-2}

Table S4: Correlation between distance matrices calculated using HMP mapped and unmapped pseudo-samples. p-values were based on 10^4 permutations.

Tools	Isolated reads	Reads from known micro-organisms	Reads from unknown micro-organisms
metaSPAdes	54%	19%	27%
MEGAHIT	74%	8%	18%

Table S5: Average proportion of reads mapped to *de novo* assembled contigs for urogenital unmapped pseudo-samples.

References

Jia B, Xuan L, Cai K, Hu Z, Ma L, and Wei C. 2013. Nessm: a next-generation sequencing simulator for metagenomics. *PLoS One* **8**.

Letunic I and Bork P. 2019. Interactive tree of life (itol) v4: recent updates and new developments. *Nucleic acids research* **47**: W256–W259.

Lu J, Breitwieser FP, Thielen P, and Salzberg SL. 2017a. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science* **3**: e104.

Lu YY, Tang K, Ren J, Fuhrman JA, Waterman MS, and Sun F. 2017b. Cafe: accelerated alignment-free sequence analysis. *Nucleic Acids Research* .

Nye TM, Lio P, and Gilks WR. 2005. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics* **22**: 117–119.

Strope CL, Scott SD, and Moriyama EN. 2006. indel-seq-gen: a new protein family simulator incorporating domains, motifs, and indels. *Molecular biology and evolution* **24**: 640–649.