Supplementary material for SimkaMin: fast and resource frugal *de novo* comparative metagenomics.

Gaetan Benoit, Mahendra Mariadassou, Stphane Robin, Sophie Schbath, Pierre Peterlongo and Claire Lemaitre

1 Bias and variance of $\widehat{BC}(S_1, S_2)$

The Bray-Curtis dissimilarity $BC(S_1, S_2)$ between two datasets S_1 and S_2 is defined by

$$1 - BC(S_1, S_2) = \frac{\sum_{i=1}^{W} \min(s_i^1, s_i^2)}{\sum_{i=1}^{W} (s_i^1 + s_i^2)} = \frac{\sum_{i=1}^{W} \min(s_i^1, s_i^2)/W}{\sum_{i=1}^{W} (s_i^1 + s_i^2)/W}$$

where W stands for the total number of k-mers and s_i^1 (resp. s_i^2) stands for the abundance of k-mer number i in S_1 (resp. S_2). In the following we shall denote $a_i = \min(s_i^1, s_i^2)$ and $b_i = (s_i^1 + s_i^2)$ and μ_a (resp. μ_b) the mean: $\mu_a = \sum_i a_i/W$. Consequently, the BC distance can be written as $1 - BC(S_1, S_2) = \mu_a/\mu_b$ and the results provided below are valid for any distance having the same form. This includes the Jaccard index, taking $a_i = 1$ if k-mer i is present in both samples (and 0 otherwise) and $b_i = 1$ if k-mer i is present in either of the two samples (and 0 otherwise).

Sampling k-mers to estimate BC. The proposed strategy consists in sampling uniformly w k-mers $\{i_u : 1 \leqslant u \leqslant w\}$ among the W to get estimates M_a and M_b of μ_a and μ_b:

$$M_a = \sum_u a_{i_u}/w, \qquad M_b = \sum_u b_{i_u}/w.$$

Because the sampling is uniform, both estimates are unbiased: $\mathbb{E}(M_a) = \mu_a$, $\mathbb{E}(M_b) = \mu_b$. Furthermore, denoting $\sigma_a^2 = \sum_i (a_i - \mu_a)^2/W$ (and respectively σ_b^2) and $\sigma_{ab} = \sum_i (a_i - \mu_a)(a_i - \mu_b)/W$, standard sampling theory indicates that the variances $\mathbb{V}(M_a)$ and $\mathbb{V}(M_b)$ and the covariance $\mathbb{C}\mathrm{ov}(M_a, M_b)$ are

each proportional to their theoretical counterparts with the same coefficient. Namely:

$$\frac{\mathbb{V}(M_a)}{\sigma_a^2} = \frac{\mathbb{V}(M_b)}{\sigma_b^2} = \frac{\mathbb{C}\mathrm{ov}(M_a, M_b)}{\sigma_{ab}} = \frac{\lambda}{w}, \tag{1}$$

where $\lambda = (W - w)/(W - 1) \simeq 1$ when $w \ll W$.

Bias and variance of the estimate. We now study the accuracy of the estimate

$$1 - \widehat{BC}(S_1, S_2) = M_a/M_b. \tag{2}$$

To this aim, we use the delta-method, which is based on a second-order Taylor expansion (see e.g. [1]) and provides the approximate mean and variance of a non-linear combination of random values. In the case of a ratio such as (2), we have that

$$\begin{aligned} \mathbb{E}\left(\frac{M_a}{M_b}\right) &\simeq \frac{\mu_a}{\mu_b} + \frac{\mu_a}{\mu_b^3}\mathbb{V}(M_b) - \frac{1}{\mu_b^2}\mathbb{C}\mathrm{ov}(M_a, M_b), \\ \mathbb{V}\left(\frac{M_a}{M_b}\right) &\simeq \frac{1}{\mu_b^2}\mathbb{V}(M_a) + \frac{\mu_a^2}{\mu_b^4}\mathbb{V}(M_b) - 2\frac{\mu_a}{\mu_b^3}\mathbb{C}\mathrm{ov}(M_a, M_b). \end{aligned} \tag{3}$$

Plugging (1) into (3) shows that both the bias and the variance of $\widehat{BC}$ are of order $1/w$, so the standard deviation is of order $1/\sqrt{w}$ and thus dominates the bias.

Note that the exact ratio of bias to standard deviation depends on data-specific constants so that the asymptotic regime may not be reached with the same w value for all datasets. We nevertheless show in the results section ranges of w values where the systematic error (bias) becomes tiny compared to the unavoidable sampling noise (standard deviation) in complex metagenomic datasets. Those ranges provide a good trade-off in terms of estimation accuracy versus computational cost.

2 Command lines and computing environment

SimkaMin release 1.5.0 was used for the tests. Tests were performed on a machine equipped with a 2.50 GHz Intel E5-2640 CPU with 20 cores, 264 GB of memory.

A python script enables to run a whole comparison of a set of datasets with a single command line. Input datasets are described in a text file in which each line stores a dataset and its associated name, that is used in the

final output matrices. This representation enables to use one or more read set file(s) per dataset which is useful to virtually concatenate pair-end read files or dataset split in several parts for any reason.

The main options are the k-mer size (set by default to 21) and the size w of the sketches (set by default to one million). Optionally, the user can limit the number of reads considered per dataset. This is useful when comparing datasets of uneven sizes.

We recapitulate here the commands used for providing presented results.

- SimkaMin command (with w being the number of k-mers per sketch):

```
simkaMin_pipeline.py -in file_of_file.txt \
-bin path_to_simka/build/bin/simkaMin -max-reads \
50000000 -nb-kmers ${w}  -out .
```

- Simka command used to generate non subsampled reference results:

```
simka/build/bin/simka -in file_of_file.txt \
-max-reads 50000000 -out simka_res \
-out-tmp ./  -abundance-min 1
```

3 TARA oceans datasets

The file `file_of_file.txt` mentioned in the previous section contains the location of each of the 20 input read sets downloaded from European Nucleotide Archive (ENA) under project ID PRJEB402. The list of used read sets are provided in Table 1. More details can be found on the [2] companion website: `www.igs.cnrs-mrs.fr/Tara_Agulhas/`.

Dataset name	short name	ENA sample accession	ENA run accession
TARA_052_DCM_0.22-1.6	52D	ERS489590	ERR599002|ERR599016
TARA_052_SRF_0.22-1.6	52S	ERS489534	ERR599098|ERR599139
TARA_064_DCM_0.22-3	64D	ERS490001	ERR598972|ERR599023|ERR599025
TARA_064_SRF_0.22-3	64S	ERS489917	ERR598970|ERR599088|ERR599150
TARA_065_DCM_0.22-3	65D	ERS490085	ERR598990|ERR599018|ERR599110
TARA_065_SRF_0.22-3	65S	ERS490029	ERR598979|ERR599146
TARA_068_DCM_0.22-3	68D	ERS490296	ERR599017|ERR599056|ERR599103
TARA_068_SRF_0.22-3	68S	ERS490265	ERR599129|ERR599171|ERR599174
TARA_070_SRF_0.22-3	70S	ERS490327	ERR599135|ERR599165
TARA_072_DCM_0.22-3	72D	ERS490475	ERR599133|ERR599137
TARA_072_SRF_0.22-3	72S	ERS490433	ERR598984|ERR599105
TARA_076_DCM_0.22-3	76D	ERS490597	ERR599040|ERR599148
TARA_076_SRF_0.22-3	76S	ERS490538	ERR599010|ERR599126
TARA_078_DCM_0.22-3	78D	ERS490691	ERR599046|ERR599101
TARA_078_SRF_0.22-3	78S	ERS490655	ERR599006|ERR599022
TARA_082_DCM_0.22-3	82D	ERS490924	ERR599027|ERR599122
TARA_082_SRF_0.22-3	82S	ERS490881	ERR599009|ERR599035
TARA_084_SRF_0.22-3	84S	ERS490997	ERR598945|ERR599059
TARA_085_DCM_0.22-3	85D	ERS491090	ERR599104|ERR599121
TARA_085_SRF_0.22-3	85S	ERS491040	ERR599090|ERR599176

Table 1: TARA oceans read sets used in this study.

4 Results for several subsampling efforts

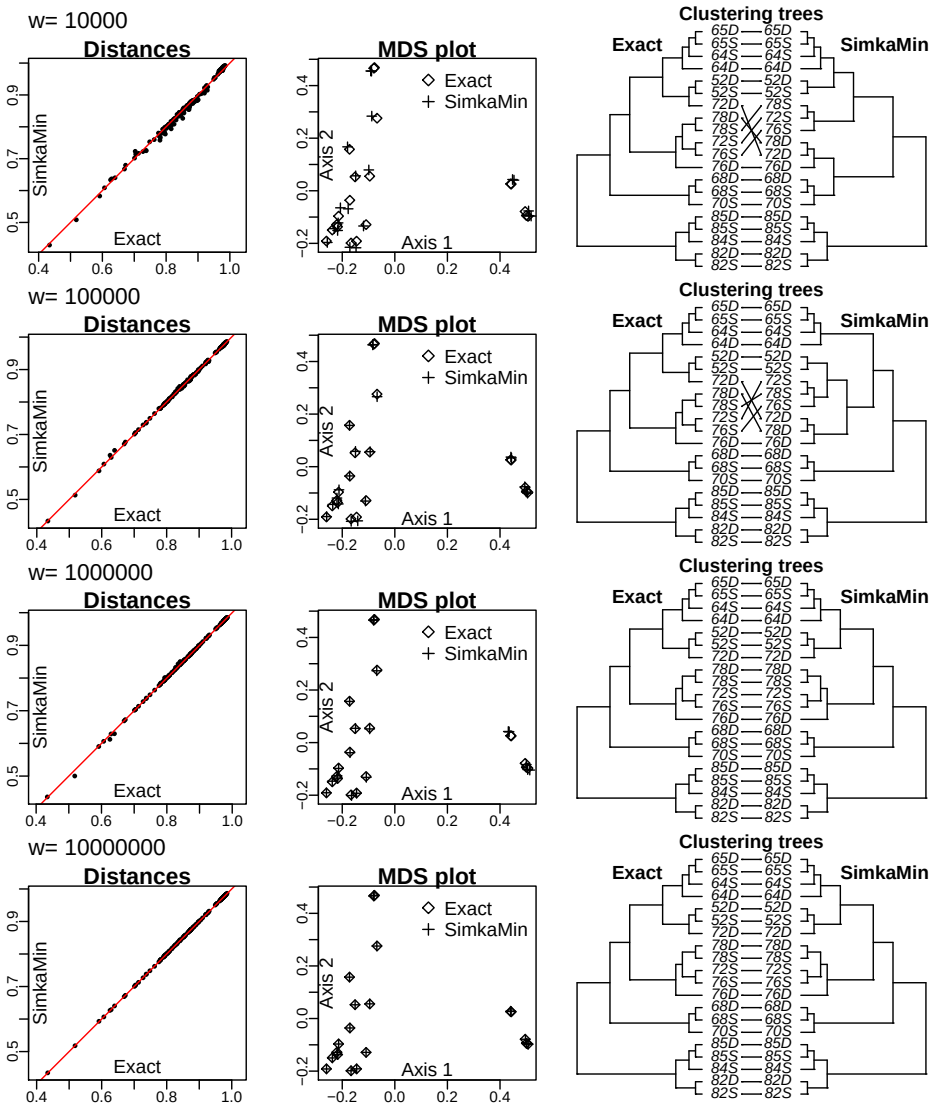

Figure 1: Comparison of SimkaMin results with "exact" (i.e. without subsampling) results, for several values of subsampling effort (w), in terms of raw distances (left), Multidimensional Scaling projection of the samples (middle) and clustering trees of the samples (right).

5 Results for other biomes

5.1 Human gut microbiome

SimkaMin was applied on 20 human gut microbiome samples from the Human Microbiome Project (using the first 50 Million reads for each dataset) see Table 2. Performance results are shown in Table 3, distance estimation quality results are shown in Figure 2.

short name	SRA sample accession	SRA run accession
1	SRS017191	SRR060358|SRR060359
2	SRS049712	SRR061165|SRR061174
3	SRS052027	SRR059330|SRR059331
4	SRS022071	SRR059454|SRR059455
5	SRS053335	SRR061151|SRR061158
6	SRS024075	SRR061689|SRR061690|SRR061691
7	SRS016095	SRR061169|SRR061172
8	SRS011271	SRR061168|SRR061170
9	SRS015794	SRR061185|SRR061195
10	SRS024331	SRR059890|SRR059891
11	SRS043701	SRR059914|SRR059915
12	SRS017103	SRR060364|SRR060365
13	SRS020869	SRR059356|SRR059357
14	SRS024009	SRR059460|SRR059461
15	SRS016989	SRR059372|SRR059373
16	SRS013215	SRR059366|SRR059367
17	SRS015190	SRR061143|SRR061147
18	SRS012273	SRR059412|SRR059413
19	SRS019601	SRR063502|SRR063505
20	SRS042628	SRR061184|SRR061212

Table 2: Sample information for the human gut microbiome dataset.

w value	Time	Mem (GB)	Disk usage size (MB)	Correlation
10^4	3mn21	0.02	3	.9915
10^5	3mn29	0.12	31	.9984
10^6	3mn55	1.09	312	.9997
10^7	9mn46	11.98	3,127	.9997
unlimited	14mm29	4.92	17.10^4	1

Table 3: SimkaMin performances on the human gut dataset. w is the number of k-mers per sketch, "unlimited" refers to non-sampled results. Correlation is the spearman correlation coefficient computed between SimkaMin Bray-Curtis dissimilarities versus the non-sampled ones.

5.2 Grass Rhizosphere microbiome

SimkaMin was applied on 20 soil microbiome samples from a grass rhizosphere study with ENA Project id PRJEB27870 (using the first 30 Million reads for each dataset) see Table 4. Performance results are shown in Table 5, distance estimation quality results are shown in Figure 3.

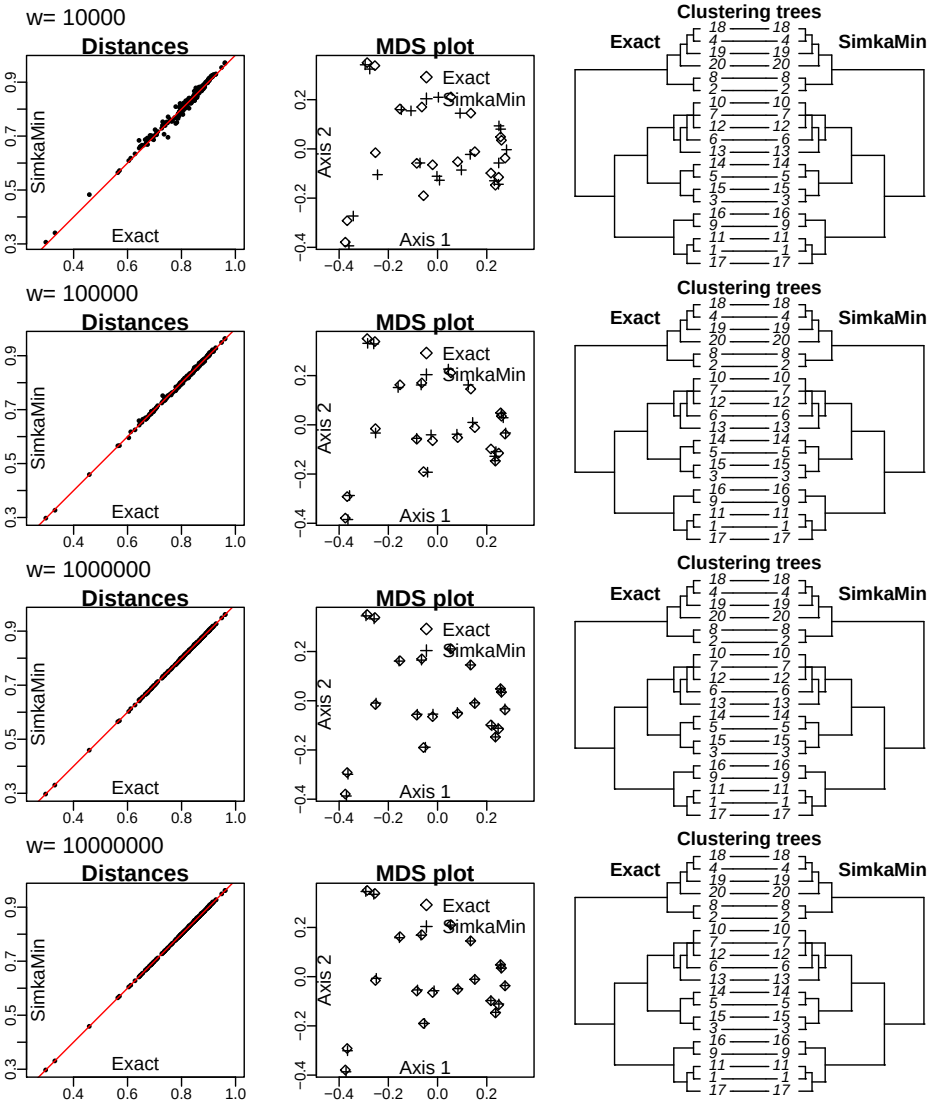

Figure 2: Comparison of SimkaMin results with "exact" (i.e. without subsampling) results on the human gut dataset, for several values of subsampling effort (w), in terms of raw distances (left), Multidimensional Scaling projection of the samples (middle) and clustering trees of the samples (right).

short name	ENA sample accession	ENA run accession
1	ERS2620027	ERR2709722
2	ERS2620028	ERR2709723
3	ERS2620037	ERR2709732
4	ERS2620038	ERR2709733
5	ERS2620041	ERR2709736
6	ERS2620045	ERR2709740
7	ERS2620047	ERR2709742
8	ERS2620051	ERR2709746
9	ERS2620060	ERR2709755
10	ERS2620064	ERR2709759
11	ERS2620069	ERR2709764
12	ERS2620071	ERR2709766
13	ERS2620072	ERR2709767
14	ERS2620073	ERR2709768
15	ERS2620081	ERR2709776
16	ERS2620086	ERR2709781
17	ERS2620095	ERR2709790
18	ERS2620102	ERR2709797
19	ERS2620106	ERR2709801
20	ERS2620116	ERR2709811

Table 4: Sample information for the soil dataset (PRJEB27870 project id).

w value	Time	Mem (GB)	Disk usage size (MB)	Correlation
10^4	2mn19	0.02	3	.9918
10^5	2mn04	0.12	31	.9987
10^6	1mn54	1.08	315	.9997
10^7	5mn18	12.11	3,135	.9999
unlimited	14mm51	4.62	15×10^4	1

Table 5: SimkaMin performances on the soil dataset. w is the number of k-mers per sketch, "unlimited" refers to non-sampled results. Correlation is the spearman correlation coefficient computed between SimkaMin Bray-Curtis dissimilarities versus the non-sampled ones.

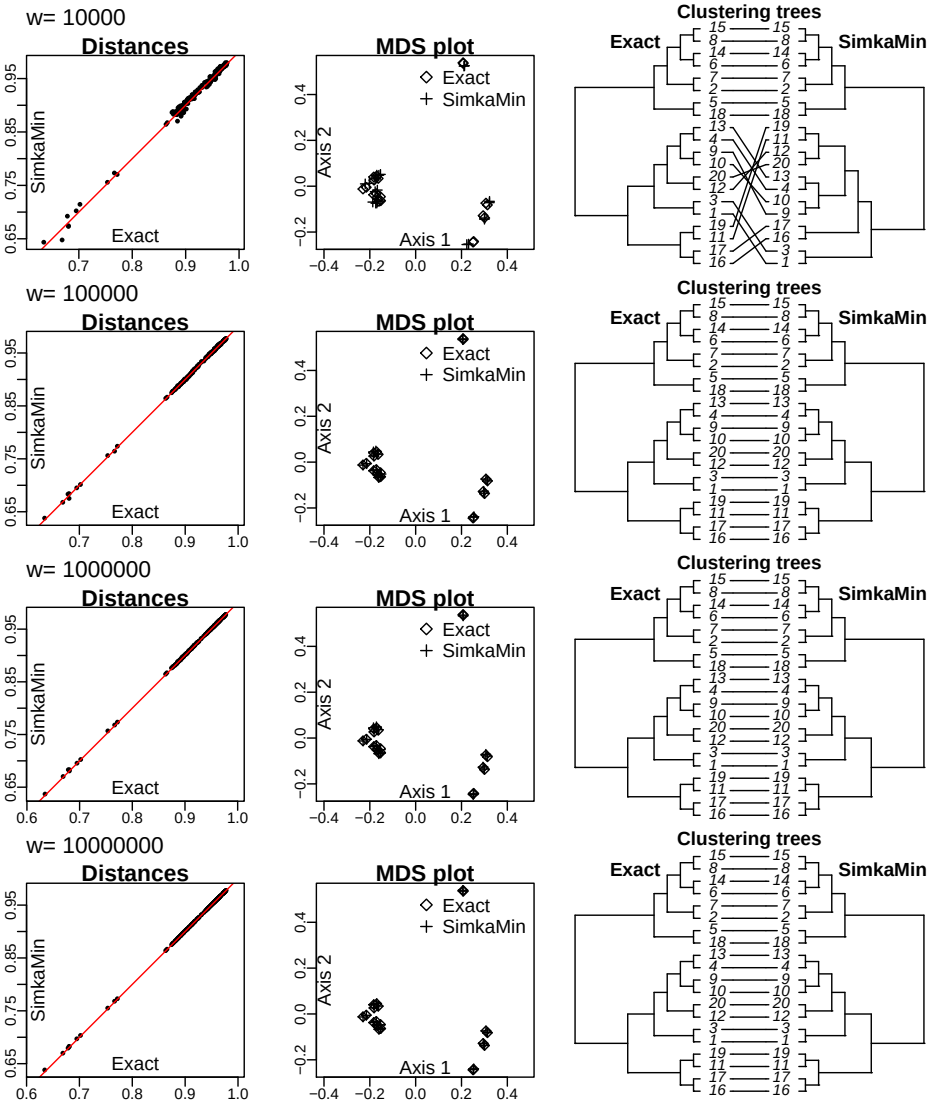

Figure 3: Comparison of SimkaMin results with "exact" (i.e. without subsampling) results on the soil dataset, for several values of subsampling effort (w), in terms of raw distances (left), Multidimensional Scaling projection of the samples (middle) and clustering trees of the samples (right).

References

[1] A. van der Vaart. *Asymptotic statistics*, volume 27 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, New York, 1998.

[2] E Villar, G K Farrant, M Follows, L Garczarek, S Speich, et al. Environmental characteristics of agulhas rings affect interocean plankton transport. *Science*, 348(6237):1261447, 2015.