Supplementary article for Multiple comparative metagenomics using multiset k-mer counting

Gaëtan Benoit, Pierre Peterlongo, Mahendra Mariadassou, Erwan Drezen, Sophie Schbath, Dominique Lavenier, Claire Lemaitre

Description of the datasets

For our experiments, we used the Illumina WGS reads from the HMP project [1]. The preparation of the data followed this protocol:

1. Download the samples that passed the quality control assessment at http://hmpdacc.org/HMASM/
   - In command line: wget http://downloads.hmpdacc.org//data/Illumina/sample_name
2. Extract all the tar.bz2 archives (one per sample)
   - In command line: tar xvjf archive.tar.bz2
3. Remove all the singleton files (keep only paired files)
   - In command line: find dataDir -name '*.singleton.*' -type f -delete
4. Compress all paired fastq files in gzip format
   - In command line: pigz filename*.fastq

Most experiments used a subset of the 690 samples, the protocol used to constitute these datasets is the following:

1. Insert all sample IDs in a list and sort it by descending order of their data size (we used the size given in the metadata file of the WGS data). The largest sample is thus in first position of the list and the lowest at the end.
2. For each experiment, we pick some samples in this list. Composition of each dataset by figure and table:
   - Figure 3: up to N largest samples (using only the first two million reads of each dataset)
   - Figure 4: the 50 smallest samples
   - Figure 5: the 15 smallest samples
   - Figure 6: the 138 gut samples
   - Figure 7: all samples
   - Figure 8: the 138 gut samples
   - Table 2: all samples

Correlation with read-based approaches, from distances to similarity measures

To compare Simka to read-based approaches, such as Commet [3] and the BLAT [2] alignment-based approach, we considered similarity measures rather than distances. In Simka, the similarity measure is defined as follows: the percentage of matched k-mers between two read sets S_i and S_j is given by the following equation:

$$MatchedKmersPercentage(S_i, S_j) = 100 \times \frac{\sum_{w \in S_i \cap S_j} N_{S_i}(w) + N_{S_j}(w)}{\sum_{w \in S_i} N_{S_i}(w) + \sum_{w \in S_j} N_{S_j}(w)} \quad (1)$$

Both read-based approaches define and use a read *similarity* notion. They derive the percentage of reads from S_i *similar* to at least one read from S_j, noted by $\left|S_i \overrightarrow{\cap} S_j\right|$. Commet considers that two reads are similar if they share at least t non-overlapping k-mers (here $t = 2$, $k = 33$). For BLAT alignments, similarity was defined based on several identity thresholds: two reads were considered similar if their alignment spanned at least 70 nucleotides and had a percentage of identity higher than 92%, 95% or 98%.

The read-based similarity between read sets S_i and S_j is then given by:

$$MatchedReadsPercentage(S_i, S_j) = 100 \times \frac{\left|S_i \overrightarrow{\cap} S_j\right| + \left|S_j \overrightarrow{\cap} S_i\right|}{|S_i| + |S_j|}. \quad (2)$$

which is the exact counterpart of Eq. 1 on reads using the read similarity notion.

Both equations above are an estimation of the shared genomic content between two samples respectively at the k-mer level and at the read level.

Commet and Simka were compared on the 50 smallest samples of HMP project. For time reasons, BLAT and Simka were compared on a smaller subset: the 15 smallest samples.

Correlation between taxonomic distances and k-mer based distances

This section describes the complete procedure followed to compare the distances provided by Simka and Mash [4] to taxonomic distances. The HMP consortium provides a quantitative taxonomic profile for each sample on its website (`http://www.hmpdacc.org/HMSCP/`). These profiles were obtained by mapping the reads on a reference genome catalog at 80% of identity. The complete protocol is given in the section "Protocols and Tools" of the previously mentioned web page.

The R package *vegan* was used to compute taxonomic distances from these profiles with the following commands:

- Bray-Curtis distance: vegdist(hmp-profiles, method="bray")
- Jaccard distance: vegdist(hmp-profiles, method="jaccard")

The first distance was used for the comparison with the Bray-Curtis distance provided by Simka. The second was used for the comparison with Mash. This Jaccard distance is the abundance-based version of the Jaccard index provided by Mash.

The package *vegan* was also used to compute the Spearman correlation between two distance matrices with a Mantel test (p=0.001):

- m = mantel(taxonomic-distance-matrix, kmer-distance-matrix, method="spearman")
- correlation r = m$statistic

Mash was run with the same parameters used in its publication, *i.e.* using 10000 k-mers per sample to estimate their Jaccard index.

References

[1] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

[2] W. J. Kent. Blat—the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.

[3] N. Maillet, G. Collet, T. Vannier, D. Lavenier, and P. Peterlongo. Commet: comparing and combining multiple metagenomic datasets. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 94–98. IEEE, 2014.

[4] B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biol*, 17(1):132, 2016.