Supplementary Material for "YAMP: a framework enabling reproducibility in metagenomics research"

Alessia Visconti[1], Tiphaine C. Martine,[1], and Mario Falchi,[1]
[1]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK

```
-------------------------------------------
YET ANOTHER METAGENOMIC PIPELINE (YAMP) v 0.9.2
-------------------------------------------

Copyright (C) 2017 Dr Alessia Visconti   <alessia.visconti@kcl.ac.uk>
This pipeline is distributed in the hope that it will be useful
but WITHOUT ANY WARRANTY. See the GNU GPL v3.0 for more details.
Please report comments and bugs to: alessia.visconti@kcl.ac.uk

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

Analysis starting at Wed Aug 16 16:17:06 BST 2017
Analysed samples are: SRR1944683_1.fastq.gz and SRR1944683_2.fastq.gz
Working directory set to Anterior_nares_2
New files will be saved using the 'Anterior_nares_2' prefix

Analysis mode? complete
Saving QC temporary files? false
Saving community characterisation temporary files? false

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

Performing STEP 1 [Assessment of read quality of FASTQ file] at Wed Aug 16 16:17:06 BST 2017 on SRR1944683_1.fastq.gz

Summary of SRR1944683_1.fastq.gz's basic statistic
SRR1944683_1.fastq.gz's total reads: 2820900

[... wrapped text ...]

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

Performing STEP 2 [De-duplication] at Wed Aug 16 16:17:06 BST 2017

BBduk's de-duplication stats:

Reads In:            5641800
Clumps Formed:         85885
Duplicates Found:    5439512

202288 out of 5641800 paired reads survived de-duplication (3.58552%, 5439512 reads removed)

STEP 2 terminated at Wed Aug 16 16:17:20 BST 2017 (13.222940537 seconds)

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

Performing STEP 3 [Trimming] at Wed Aug 16 16:17:20 BST 2017

BBduk's trimming stats (trimming adapters and low quality sequences):
Input:                      202288 reads        20431088 bases.
QTrimmed:                   41141 reads (20.34%)    3101318 bases (15.18%)
KTrimmed:                   65910 reads (32.58%)    3457420 bases (16.92%)
Trimmed by overlap:         716 reads (0.35%)   4596 bases (0.02%)
Total Removed:              51374 reads (25.40%)    6563334 bases (32.12%)
Result:                     150914 reads (74.60%)   13867754 bases (67.88%)

11307 singleton reads whose mate was trimmed shorter preserved

BBduk's trimming stats (synthetic contaminants, paired reads):
Input:                      150914 reads        13867754 bases.
Contaminants:               0 reads (0.00%)     0 bases (0.00%)
Total Removed:              0 reads (0.00%)     0 bases (0.00%)
Result:                     150914 reads (100.00%)  13867754 bases (100.00%)
```

Figure S1: Example of YAMP execution log. This excerpt shows the summary statistics of the data processing of the dataset SRR1944683 from the Microbiome Project (HMP) Phase III.

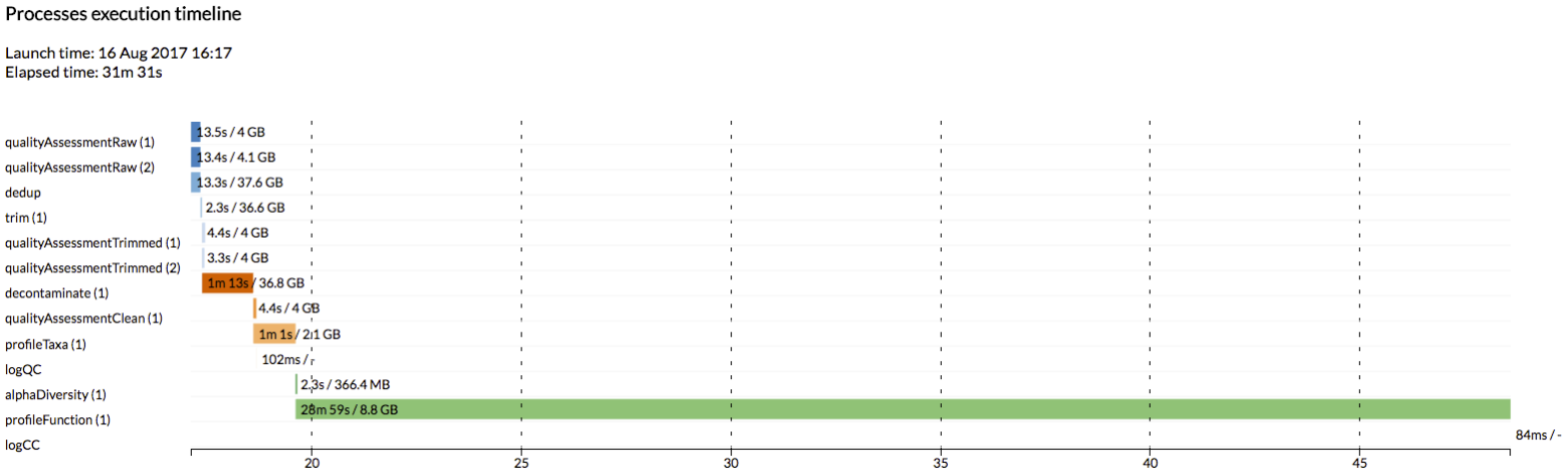

Figure S2: Example of YAMP execution profile. The HTML page shows the time spend during the complete YAMP execution and in each step, as well as the step memory peaks. This page corresponds to the analysis of the dataset SRR1944683 from the Microbiome Project (HMP) Phase III.

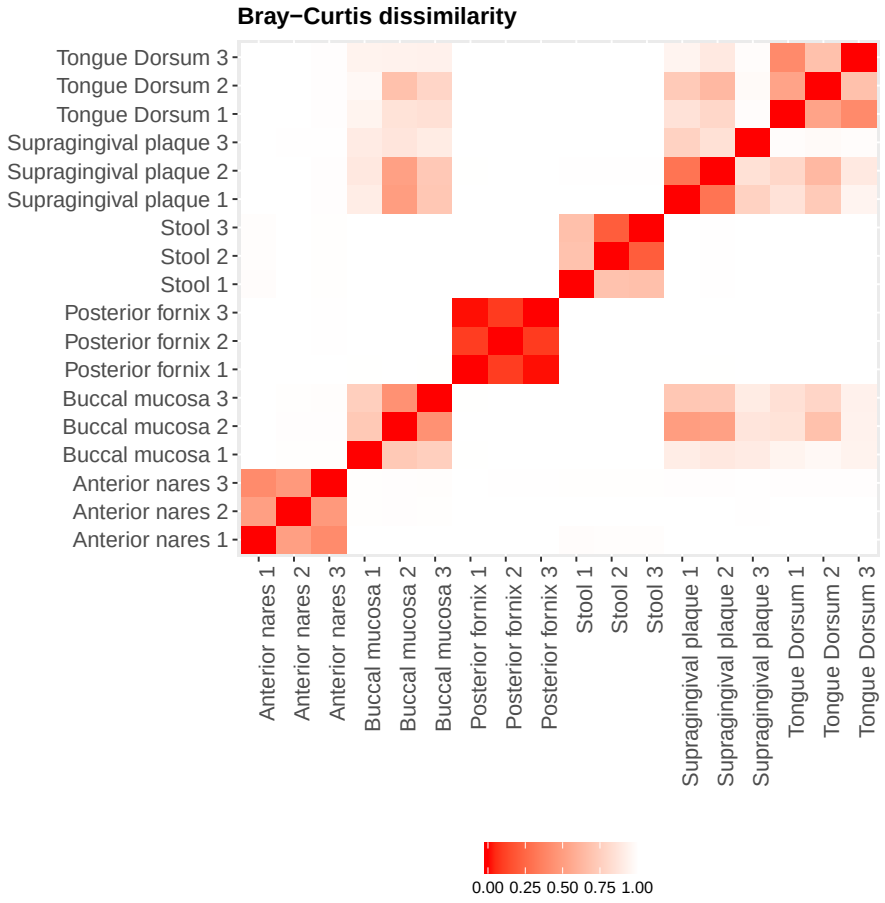

Figure S3: Bray-Curtis dissimilarity for the 18 analysed samples. The Bray-Curtis dissimilarity values were evaluated using the species relative abundances as estimated by YAMP using MetaPhlAn2 and the *vegdist* function in the `vegan` R package.

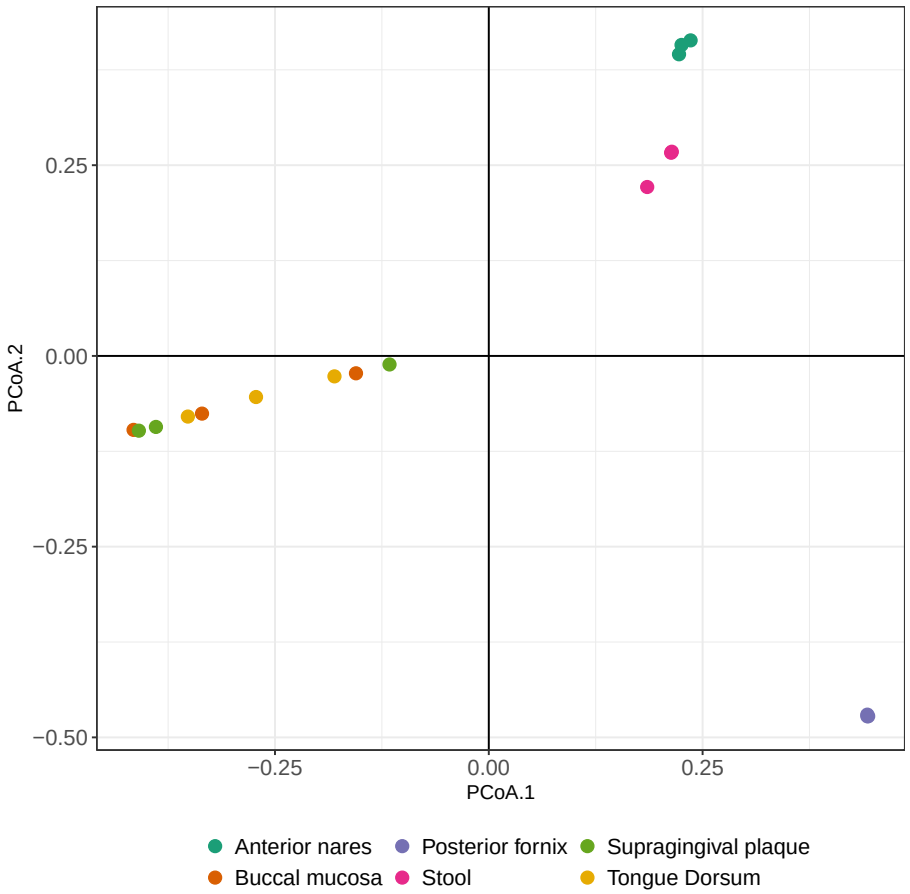

Figure S4: Principal coordinate analysis (PCoA) for the 18 analysed samples. PCoA was evaluated on the Bray-Curtis dissimilarity values using the *pcoa* function in the `ape` R package. PCoA shows that species composition is sufficient to discriminate among body sites, even though it has limited ability in distinguishing among different loci in the oral cavity.