YAMP: a framework enabling reproducibility in metagenomics research

Alessia Visconti[1,*], Tiphaine C. Martin[1], and Mario Falchi[1]

Department of Twin Research and Genetic Epidemiology, King's College London,
* Corresponding author
alessia.visconti@kcl.ac.uk

Abstract. YAMP is a user-friendly workflow that enables the analysis of whole shotgun metagenomics data while using containerisation to ensure computational reproducibility and facilitate collaborative research. YAMP can be executed on any UNIX-like system, and offers seamless support for multiple job schedulers as well as for Amazon AWS cloud. Although YAMP has been developed to be ready-to-use by non-experts, bioinformaticians will appreciate its flexibility, modularisation, and simple customisation. The YAMP script, parameters, and documentation are available at https://github.com/alesssia/YAMP.

Keywords: Metagenomics; Reproducibility; Workflow; Containerisation; Docker

Background

Thanks to the increased cost-effectiveness of high-throughput technologies, the number of studies collecting and analysing large amounts of data has surged, opening new challenges for data analysis and research reproducibility. A ubiquitous lack of repeatability and reproducibility has in fact been observed, and a recent Nature's survey of 1,576 researchers showed that more than 50% and 70% of them failed to reproduce their own and other scientists' experiments, respectively [1]. Unavailability of primary data and computational experimentation have been named as the major culprits for this reproducibility crisis, with many studies relying on *ad hoc* scripts and not publishing the necessary code and/nor sufficient details to reproduce the reported results [2,3,4], and with variations across workstations and operating systems representing another obstacle [5,6]. To overcome this issue, tools allowing the development of workflows [7] and software containers [8] have been proposed [9]. In fact, containerised well-structured workflows allow storing every detail of the workflow execution, including software's versions and parameters (*provenance*, [10]), and nullify systems' variations [6], guaranteeing studies' repeatability and reproducibility. Containerised workflows also facilitate collaborative projects, by ensuring identical analysis processes, thus comparable results, and allow the automatisation of data-intensive repetitive tasks [11]. Moreover, they save users with little bioinformatics or computational expertise from the hassles of installing the required pieces of software, and of designing and implementing often complex analysis orchestrations, while expert bioinformaticians can use them as a starting point for customised analyses, thus avoiding redundant solutions.

In metagenomics research, several analysis pipelines have been developed so far. However, they either do not support containerisation (*e.g.*, MetAMOS [12], MOCAT2 [13]), thus potentially compromising reproducibility, or require users to upload their unpublished and/or confidential data on third-party servers (*e.g.*, MG-RAST [14]), where, according to the available resources, they can spend several days waiting to be processed [15], and with data privacy concerning some of the researchers [16].

Here we present *"Yet Another Metagenomic Pipeline"* (YAMP), a ready-to-use containerised workflow that processes raw shotgun metagenomics sequencing data up to the taxonomic and functional annotation. YAMP is implemented in Nextflow, a framework that allows defining workflows that are highly parallel, easily portable (including on distributed systems), and very flexible and customisable [6]. We integrated our Nextflow pipeline with a Docker (`https://www.docker.com`) and a Singularity (`http://singularity.lbl.gov`) container. While the former defines a platform-independent virtualised light-weight operating system that includes all the pieces of software required by YAMP and traces their versioning, the latter allows these features to be transferred to High Performance Computing (HPC) systems, with which Docker is inherently incompatible.

The YAMP workflow

The YAMP workflow is composed of three analysis blocks: the quality control, (QC; Figure 1, green rectangle), complemented by several steps of assessment and visualisation of data quality (Figure 1, orange rectangle), and the community characterisation (Figure 1, pink rectangle).

The QC starts with an optional step of de-duplication, where identical reads, potentially generated by PCR amplification, are removed. Next, reads are first filtered to remove adapters, known artefacts, phiX, and then quality-trimmed. Reads that become too short after trimming are discarded, while, when paired-end reads are at hand, singleton reads (*i.e.*, paired-end reads whose mates have been removed) are preserved in order to retain as much information as possible. Finally, reads are screened for contaminants, *e.g.*, reads that do not belong to the studied ecosystem. The QC is performed by means of a number of tools belonging to the BBmap suite [17], namely clumpify, BBduk, and BBwrap, which are computationally efficient and allow processing both single- and paired-end reads from all the major sequencing platforms (*i.e.*, Illumina, Roche 454 pyrosequencing, Sanger, Ion Torrent, PacBio, and Oxford Nanopore). FastQC [18] is used to perform QC assessment and visualisation of reads quality, and to evaluate the effectiveness of the trimming and decontamination step. The QC is followed by multiple steps aimed at the taxonomic and functional characterisation of the microbial community. Taxonomic binning and profiling, *i.e.*, the identification and quantification of the micro-organisms present in the metagenomics sample, is performed with MetaPhlAn2 [19], which uses clade-specific markers to both detect the micro-organisms and to estimate their relative abundance. The functional capabilities of the microbial community, *i.e.,* the functions carried out by the identified micro-organisms, are assessed by the HUMAnN2 pipeline [20], which first stratifies the community in known and unclassified organisms using the MetaPhlAn2 results and the ChocoPhlAn pan-genome database, and then combines these results with those obtained through an organism-agnostic search on the UniRef proteomic database. QIIME [21] is used to evaluate multiple α-diversity measures based on the taxonomic profile.

YAMP accepts in input both single- and paired-end FASTQ files, and users can customise the workflow execution either by using command line options or by modifying a simple plain-text configuration file, where parameters are set as key-value pairs. While the parameters could be tuned according to the dataset at hand, to facilitate non-expert users in their analyses we provide a set of default parameters derived from our own analysis experience. The output generated by YAMP includes a FASTQ file of QC'ed reads, the taxonomy composition along with the microbe, gene and pathway relative abundances, the pathway coverage, and multiple α-diversity measures. An option allows users to retain temporary files, such as those generated by the QC steps or during the HUMAnN2 execution. Additionally, YAMP outputs several QC reports, a detailed log file recording information about each analysis step (Supplementary Figure S1), and statistics of memory usage and time of execution (Supplementary Figure S2).

Results

To facilitate the discussion on YAMP computational requirements, and to assess its ability to correctly identify microbial communities, we analysed 18 randomly selected samples from six different body sites sequenced during the Phase III of the Human Microbiome Project [22] (Table 1). On average, the selected samples included 12.6M paired-end reads (25.2M reads in total), which yielded to 13.3M QC'ed reads (including both paired-end and singleton reads), and were processed in an average time of two hours using four threads on a machine sporting a 2.60GHz Intel® Xeon® processor with 32 GB of RAM (Table 1). At the phylum level, each body site showed a characteristic signature (Figure 2), with a predominance of Actinobacteria in the airways, Firmicutes in the vagina, Bacteroidetes in the stool, and a mixture of Actinobacteria, Firmicutes and Proteobacteria in the oral cavity, as already observed in previous studies [23]. A site-specific microbial signature was also present at the species level, where both the Principal coordinate analysis (PCoA) evaluated using the Bray-Curtis dissimilarity (Supplementary Figure S3 and S4), and the hierarchical clustering computed on the Manhattan distances among species relative abundances (Figure 3) showed that the taxonomy composition was sufficient to discriminate among body sites, even though it had limited ability in distinguishing among different loci in the oral cavity.

Discussion

In conclusion, with YAMP, we provide a user-friendly workflow that enables the analysis of whole shotgun metagenomics data. By supporting containerisation, YAMP allows for computational reproducibility, also enabling collaborative studies. In fact, while software versions are described in the Docker/Singularity container, the Nextflow script and configuration file capture all the details needed to fully track each step of data processing, thus satisfying the provenance requirements. Indeed, to ensure reproducibility, researchers should only provide the YAMP configuration file and a link to the container image. Being based on Nextflow, YAMP runs on any UNIX-like system, provides out-of-the-box support for several job schedulers (*e.g.,* PBS, SGE, SLURM) and for the Amazon AWS cloud, and its integration with Docker/Singularity is completely user-transparent. Moreover, YAMP does not require users to upload unpublished and/or confidential data on third-party servers, as for instance required by the MG-RAST [14] or EBI Metagenome [24] pipeline. Finally, while YAMP has been developed to be ready to use by non-experts, and potentially does not require any software installation or parameter tuning, bioinformaticians will value its flexibility and simple customisation. In fact, the well-defined YAMP modularisation and the usage of standard data formats allow both an easy integration of new analysis steps and a customisation of the existing ones.

YAMP is made available as a Nextflow script which allows a user-friendly execution via the command line. The source code is available in the YAMP GitHub repository (`https://github.com/alesssia/YAMP`), which includes a wiki with a full documentation and several tutorials. The Docker/Singularity image can be downloaded and installed from DockerHub (`https://hub.docker.com/r/alesssia/yampdocker`).

Potential implications

YAMP has been designed with the specific goals of enabling reproducible metagenomics analyses, facilitating collaborative projects, and helping researchers with limited computational experience who are approaching this field of research. However, we are confident that other areas of research would be aided by a more widespread use of containerised well-structured workflows. Indeed, as outlined in the Background Section, a lack of reproducibility is nowadays ubiquitous, and, besides undermining the credibility of scientific research, it has an economical cost, quantified, for instance, in US$28B/year for preclinical research [25]. On the other hand, ensuring reproducibility does not come for free: anecdotic evidence suggests that the time spent on a project may increase by 30-50% [1], and that to reproduce the analysis of single computational biology paper can require up to 280 hours [26]. YAMP represents a proof-of-concept showing a simple way to enable reproducible and collaborative research. We also advocate the sharing of such containerised workflows, which will benefit a wide group of researchers, regardless of their computational experience [11].

Methods

Data Availability

The 18 randomly selected samples used to assess YAMP belong to the Phase III of the Human Microbiome Project [22], and were downloaded from the European Nucleotide Archive website (Study accession number: PRJNA275349, `https://www.ebi.ac.uk/ena/data/view/PRJNA275349`). Samples were collected from healthy adults residing in the USA at the time of sample collection. After genomic DNA extraction, metagenomics library preparation was performed using the NexteraXT library construction protocol. Paired-end metagenomics sequencing was performed on the Illumina HiSeq2000 platform with a read length of 100 bp. Samples' accession numbers are reported in Table 1.

Data Analysis

Samples were processed with YAMP using the default parameters, as defined in the published YAMP configuration file (`https://raw.githubusercontent.com/alesssia/YAMP/master/nextflow.config`). The Bray-Curtis dissimilarity values were evaluated using the species relative abundances as estimated by YAMP using MetaPhlAn2 [19] and the *vegdist* function in the vegan R package (version 2.4.3) [27]. Principal coordinate analysis (PCoA) was evaluated on the Bray-Curtis dissimilarity values using the *pcoa* function in the ape R package (version 4.1) [28]. Hierarchical clustering was computed using the Manhattan distance among species relative abundances and the *pvclust* function in the pvclust R package (version 2.0) [29]. 10,000 bootstrap interactions were used to evaluate the P values supporting each cluster.

Availability of source code and requirements

- Project name: YAMP
- Project home page: `https://github.com/alesssia/YAMP`
- Operating system(s): UNIX-like systems, support for Amazon AWS Cloud
- Programming language: Nextflow
- Other requirements: Docker/Singularity
- License: GNU GPL v3
- Any restrictions to use by non-academics: None

Declarations

List of abbreviations

HPC: High Performance Computing; PCoA: Principal coordinate analysis; QC: Quality Control.

Ethical Approval

Not applicable.

Consent for publication

Not applicable.

Competing Interests

The authors declare that they have no competing interests.

Author's Contributions

AV, TCM, and MF designed the metagenomics workflow. AV implemented and optimised the workflow, created the Docker container, and wrote the manuscript. All the authors read, commented, and approved the final manuscript.

Acknowledgements

AV thanks Paolo Di Tommaso and Brian Bushnell for their help with Nextflow and BBmap.

References

1. M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature News*, vol. 533, no. 7604, p. 452, 2016.
2. J. P. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, *et al.*, "Repeatability of published microarray gene expression analyses," *Nature genetics*, vol. 41, no. 2, pp. 149–155, 2009.
3. T. Hothorn and F. Leisch, "Case studies in reproducibility," *Briefings in bioinformatics*, vol. 12, no. 3, pp. 288–300, 2011.
4. R. D. Peng, "Reproducible research in computational science," *Science*, vol. 334, no. 6060, pp. 1226–1227, 2011.
5. E. H. Gronenschild, P. Habets, H. I. Jacobs, R. Mengelers, N. Rozendaal, J. Van Os, and M. Marcelis, "The effects of freesurfer version, workstation type, and macintosh operating system version on anatomical volume and cortical thickness measurements," *PloS one*, vol. 7, no. 6, p. e38234, 2012.
6. P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nature Biotechnology*, vol. 35, no. 4, pp. 316–319, 2017.
7. J. Leipzig, "A review of bioinformatic pipeline frameworks," *Briefings in bioinformatics*, vol. 18, no. 3, pp. 530–536, 2017.
8. C. Boettiger, "An introduction to Docker for reproducible research," *ACM SIGOPS Operating Systems Review*, vol. 49, no. 1, pp. 71–79, 2015.

9. S. R. Piccolo and M. B. Frampton, "Tools and techniques for computational reproducibility," *GigaScience*, vol. 5, no. 1, p. 30, 2016.
10. S. B. Davidson and J. Freire, "Provenance and scientific workflows: challenges and opportunities," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1345–1350, ACM, 2008.
11. O. Spjuth, E. Bongcam-Rudloff, G. C. Hernández, L. Forer, M. Giovacchini, R. V. Guimera, A. Kallio, E. Korpelainen, M. M. Kańduła, M. Krachunov, *et al.*, "Experiences with workflows for automating data-intensive bioinformatics," *Biology direct*, vol. 10, no. 1, p. 43, 2015.
12. T. J. Treangen, S. Koren, D. D. Sommer, B. Liu, I. Astrovskaya, B. Ondov, A. E. Darling, A. M. Phillippy, and M. Pop, "MetAMOS: a modular and open source metagenomic assembly and analysis pipeline," *Genome biology*, vol. 14, no. 1, p. R2, 2013.
13. J. R. Kultima, L. P. Coelho, K. Forslund, J. Huerta-Cepas, S. S. Li, M. Driessen, A. Y. Voigt, G. Zeller, S. Sunagawa, and P. Bork, "MOCAT2: a metagenomic assembly, annotation and profiling framework," *Bioinformatics*, vol. 32, no. 16, pp. 2520–2523, 2016.
14. F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, *et al.*, "The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes," *BMC bioinformatics*, vol. 9, no. 1, p. 386, 2008.
15. A. Wilke, W. Gerlach, T. Harrison, T. Paczian, W. L. Trimble, and F. Meyer, "MG-RAST Manual for version 4, revision 3." `ftp://ftp.metagenomics.anl.gov/data/manual/mg-rast-manual.pdf`, 2017.
16. E. Pérez-Wohlfeil, J. A. Arjona-Medina, O. Torreno, E. Ulzurrun, and O. Trelles, "Computational workflow for the fine-grained analysis of metagenomic samples," *BMC genomics*, vol. 17, no. 8, p. 802, 2016.
17. B. Bushnell, "BBMap short-read aligner, and other bioinformatics tools." `https://sourceforge.net/projects/bbmap/`, 2015.
18. S. Andrews, "FastQC A Quality Control tool for High Throughput Sequence Data." `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`, 2010.
19. D. T. Truong, E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, and N. Segata, "MetaPhlAn2 for enhanced metagenomic taxonomic profiling," *Nature methods*, vol. 12, no. 10, p. 902, 2015.
20. S. Abubucker, N. Segata, J. Goll, A. M. Schubert, J. Izard, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, *et al.*, "HUMAnN2: The HMP Unified Metabolic Analysis Network 2." `http://huttenhower.sph.harvard.edu/humann2`, 2017.
21. J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, *et al.*, "QIIME allows analysis of high-throughput community sequencing data," *Nature methods*, vol. 7, no. 5, pp. 335–336, 2010.
22. The Human Microbiome Project Consortium, "A framework for human microbiome research," *Nature*, vol. 486, no. 7402, p. 215, 2012.
23. K. Aagaard, J. Ma, K. M. Antony, R. Ganu, J. Petrosino, and J. Versalovic, "The placenta harbors a unique microbiome," *Science translational medicine*, vol. 6, no. 237, pp. 237ra65–237ra65, 2014.
24. A. L. Mitchell, M. Scheremetjew, H. Denise, S. Potter, A. Tarkowska, M. Qureshi, G. A. Salazar, S. Pesseat, M. A. Boland, F. M. Hunter, *et al.*, "EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies," *Nucleic acids research*, 2017.

25. L. P. Freedman, I. M. Cockburn, and T. S. Simcoe, “The economics of reproducibility in preclinical research,” *PLoS biology*, vol. 13, no. 6, p. e1002165, 2015.
26. D. Garijo, S. Kinnings, L. Xie, L. Xie, Y. Zhang, P. E. Bourne, and Y. Gil, “Quantifying reproducibility in computational biology: the case of the tuberculosis drugome,” *PloS one*, vol. 8, no. 11, p. e80278, 2013.
27. P. Dixon, “Vegan, a package of r functions for community ecology,” *Journal of Vegetation Science*, vol. 14, no. 6, pp. 927–930, 2003.
28. E. Paradis, J. Claude, and K. Strimmer, “Ape: analyses of phylogenetics and evolution in r language,” *Bioinformatics*, vol. 20, no. 2, pp. 289–290, 2004.
29. R. Suzuki and H. Shimodaira, “Pvclust: an r package for assessing the uncertainty in hierarchical clustering,” *Bioinformatics*, vol. 22, no. 12, pp. 1540–1542, 2006.

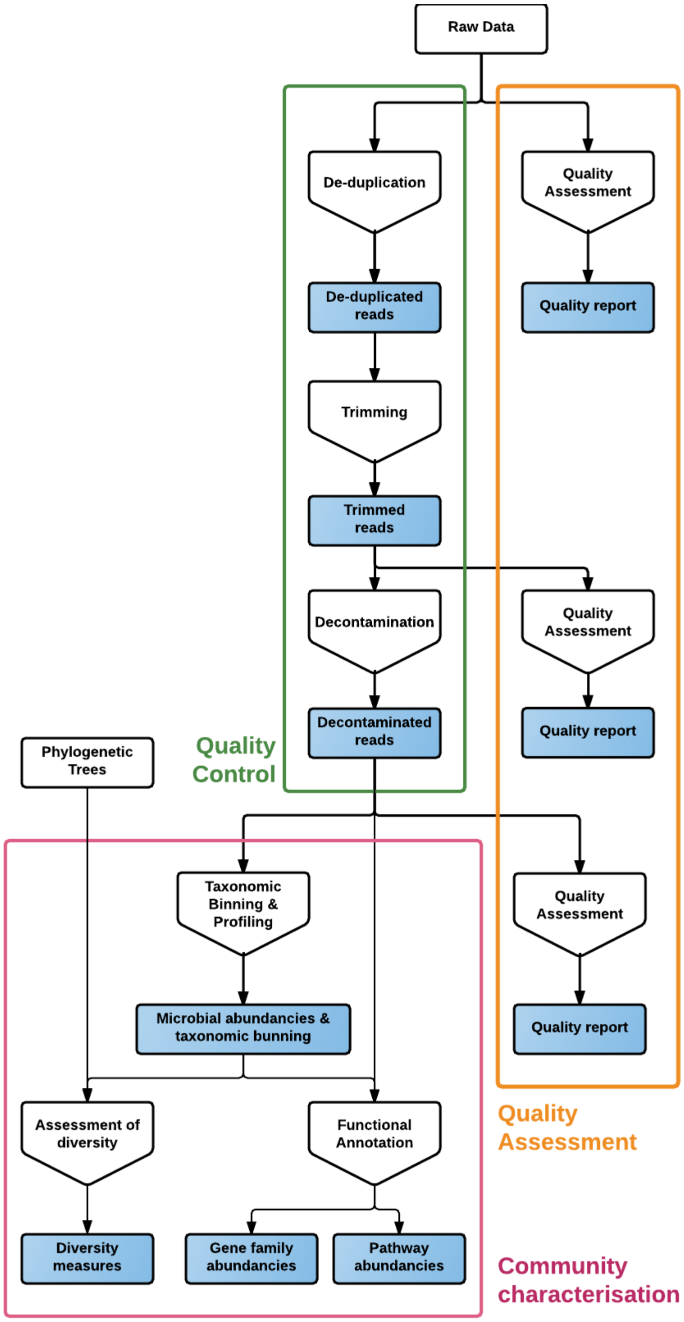

Fig. 1. The YAMP workflow. White rectangles represent data to be provided as input, and blue rectangles those produced in output. Pentagons represent the analysis steps.

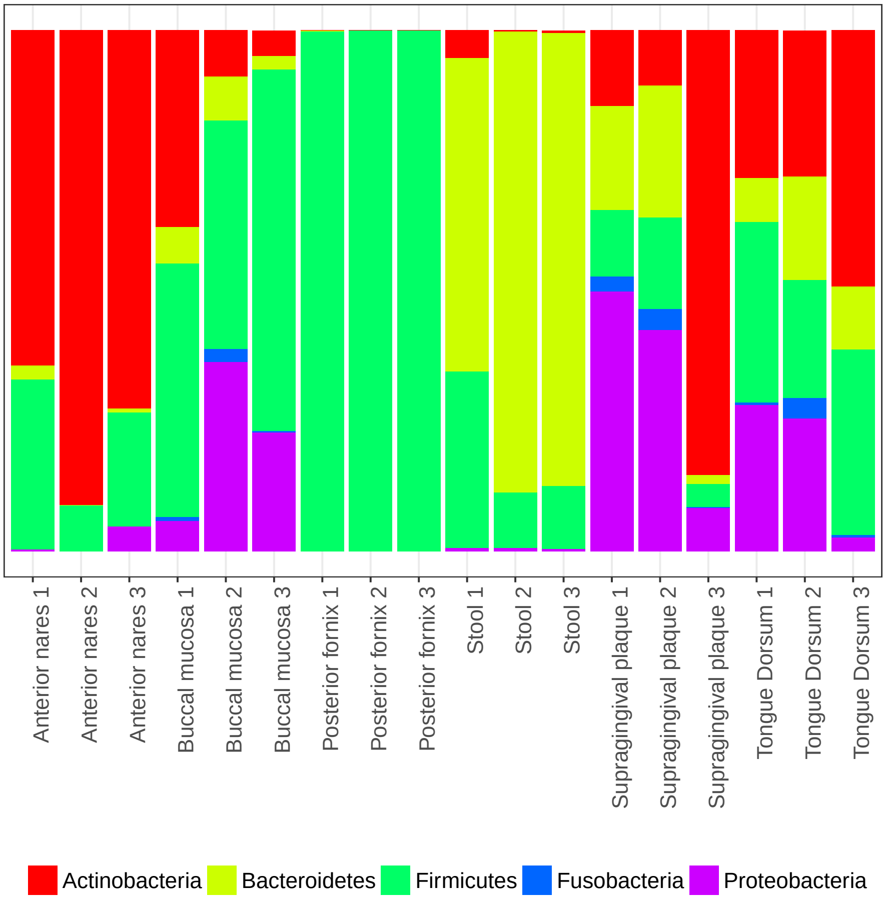

Fig. 2. Phylum level relative abundances. Each vertical bar represents a sample. Phylum relative abundances were estimated by YAMP using MetaPhlAn2. Unspecified viral phyla are not shown.

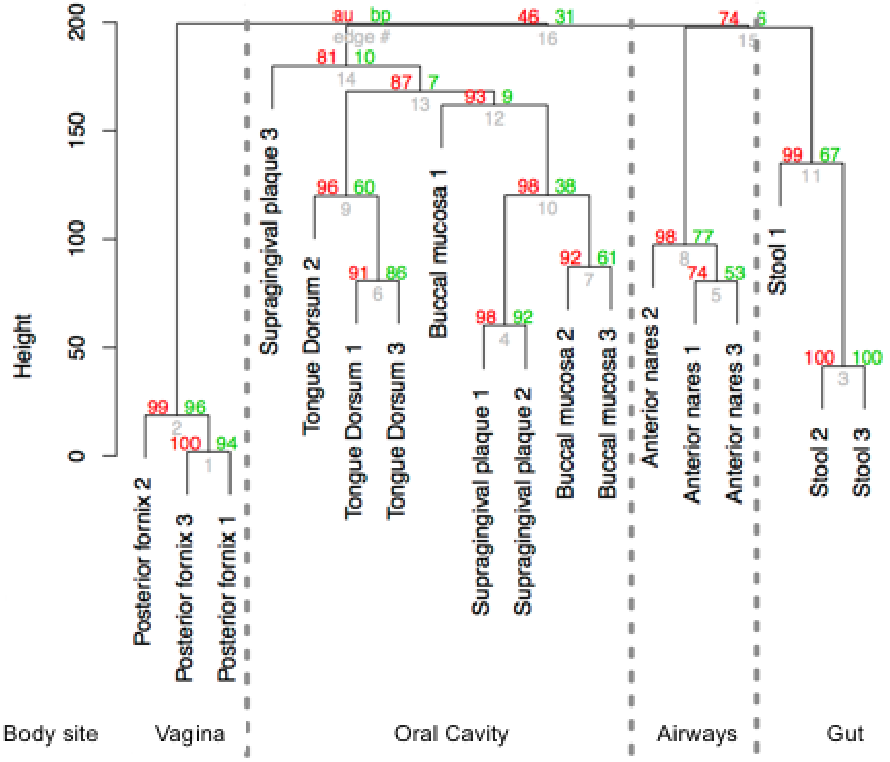

Fig. 3. Hierarchical clustering. Hierarchical clustering was computed using the Manhattan distance among species relative abundances. Values at branches are, in red, the approximately unbiased (AU) P values, in green, the bootstrap probability (BP) values (percentages), and, in grey, the edge number.

Table 1. Run Accession Number and statistics for 18 randomly selected samples from the Human Microbiome Project (HMP) Phase III [22]. Samples were processed using 4 threads on a machine sporting a 2.60GHz Intel® Xeon® processor with 32 GB of RAM.

Body site	**Locus**	**SRA Accession Number**	**Number of Raw Paired-end Reads**	**Number of QC'ed Reads Paired-ed ; Singletons**	**Running time**
Airways	Anterior nares	SRR1944674	1,181,169	590,714 ; 42,241	39m 02s
		SRR1944683	2,820,900	56,151 ; 9,513	31m 31s
		SRR1952439	14,635,701	201,260 ; 17,345	42m 00s
Gut	Stool	SRR1951826	7,956,274	7,121,697 ; 494,289	2h 15m 39s
		SRR1944873	11,033,130	9,796,817 ; 942,566	2h 26m 01s
		SRR1952058	5,834,232	5,484,362 ; 248,819	1h 39m 10s
Oral cavity	Buccal mucosa	SRR1944703	6,231,553	285,906 ; 24,212	39m 09s
		SRR1952437	15,361,468	3,451,844 ; 149,714	1h 19m 26s
		SRR1952380	11,872,420	631,595 ; 41,957	49m 07s
	Supragingival plaque	SRR1952435	16,169,911	13,620,835 ; 672,610	2h 44m 56s
		SRR1952436	21,971,588	17,237,506 ; 987,950	4h 07m 11s
		SRR1952492	19,202,739	8,040,737 ; 1805,898	1h 51m 05s
	Tongue dorsum	SRR1944869	8,074,428	6,140,295 ; 499,284	1h 36m 58s
		SRR1952378	15,024,409	12,622,724 ; 891,920	3h 17m 30s
		SRR1952379	42,173,063	29,697,754 ; 2,084,990	7h 10m 23s
Vagina	Posterior fornix	SRR1951760	10,611,721	373,021 ; 24,484	42m 19s
		SRR1944797	8,242,829	120,519 ; 10,009	35m 14s
		SRR1944845	8,537,797	140,658 ; 10,779	34m 19s