Sequence analysis

**TIPP2: metagenomic taxonomic profiling using phylogenetic markers**

Nidhi Shah, Erin K. Molloy, Mihai Pop and Tandy Warnow

1Department of Computer Science, University of Maryland College Park, College Park, MD 02740, USA, 2Department of Computer Science, University of California Los Angeles, CA 90095, USA and 3Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on October 28, 2020; revised on December 18, 2020; editorial decision on January 5, 2021; accepted on January 8, 2021

Abstract

**Motivation:** Metagenomics has revolutionized microbiome research by enabling researchers to characterize the composition of complex microbial communities. Taxonomic profiling is one of the critical steps in metagenomic analyses. Marker genes, which are single-copy and universally found across Bacteria and Archaea, can provide accurate estimates of taxon abundances in the sample.

**Results:** We present TIPP2, a marker gene-based abundance profiling method, which combines phylogenetic placement with statistical techniques to control classification precision and recall. TIPP2 includes an updated set of reference packages and several algorithmic improvements over the original TIPP method. We find that TIPP2 provides comparable or better estimates of abundance than other profiling methods (including Bracken, mOTUsv2 and MetaPhlAn2), and strictly dominates other methods when there are under-represented (novel) genomes present in the dataset.

**Availability and implementation:** The code for our method is freely available in open-source form at https://github.com/shahnidhi/TIPP_reference_package. The code and procedure to create new reference packages for TIPP2 are available at https://github.com/shahnidhi/TIPP_reference_package.

**Contact:** mpop@umd.edu or warnow@illinois.edu

**Supplementary information:** Supplementary data are available at Bioinformatics online.

1 Introduction

Microorganisms live in complex communities and play a vital role in human and environmental health. Studying these communities is important to understand how microbes interact with each other, their host and the environment (Gilbert et al., 2010; Lloyd-Price et al., 2017; Zeevi et al., 2019). One of the first steps in investigating microbial community dynamics is to estimate the abundance of different species in the community; this process is called taxonomic profiling. The most common approach for taxonomic profiling is the targeted amplicon sequencing approach, which involves extracting, amplifying and sequencing usually the small subunit of the ribosomal RNA gene (16S rRNA gene) from DNA isolated from the environmental sample. These 16S rRNA fragments are taxonomically classified to get the final community profile. Two of the largest benefits of this approach are that it is cost effective and can use large repositories of already sequenced and characterized 16S rRNA gene sequences. However, there are many known problems with the approach, such as quantification errors introduced by copy-number variations (Klappenbach et al., 2001), variable amplification efficiencies within different taxa (Engelbrekson et al., 2010; Sinha et al., 2017), different levels of resolution depending on which region of the gene is targeted (Claesson et al., 2010) and generally low-resolution power of 16S rRNA sequences to differentiate between closely related species (Gevaers et al., 2005; Huse et al., 2010).

Microbiome studies increasingly use metagenomics data—shotgun sequencing of DNA extracted directly from the environment (Handelsman, 2004), which resolves some biases of the targeted sequencing approach. Metagenomics generates millions of reads that are randomly sampled from genomes in the community. Different strategies have been introduced for estimating the relative abundance of species in the sample from metagenomic data (Lu et al., 2017; Milanese et al., 2019; Segata et al., 2012). Note that, because sequencing is a sampling process, we can only estimate relative abundances of different taxa in the sample and not the absolute abundances. A common strategy is to classify reads by performing a homology search against taxonomically characterized reference genomes in public databases. The resulting read assignments normalized by the genome sizes can provide an estimate of relative abundance of individual species (Arumugam et al., 2011; Lu et al.,...
2 Approach

TIPP2 builds upon TIPP1, but uses a larger set of marker genes for the reference database and has slightly modified algorithmic steps (Fig. 1). Here, we explain briefly the TIPP1 algorithm and highlight the novel contributions in TIPP2.

2.1 TIPP1 algorithm overview

As a preprocessing step for each marker gene, TIPP1 generates a multiple sequence alignment and a maximum likelihood tree constrained by the NCBI taxonomy; the collection of alignment-tree pairs for all marker genes is collectively referred to as the ‘reference package’. TIPP1 has three key steps when analyzing reads from a metagenomic sample.

1. Binning reads: First, all reads are assigned to one of the marker gene sequences using BLAST (Altschul et al., 1990). Specifically, each read is assigned to the marker gene to which it has the best alignment. Reads that do not have a good alignment to any of the marker genes are excluded from analysis.

2. Classifying reads: Each read is added to the multiple sequence alignment of the marker gene and then placed into the corresponding taxonomic tree. This alignment/placement step is performed with SEPP (Mirarab et al., 2012), a method that uses an eHMMs designed to provide high taxonomic placement accuracy for large alignment/tree pairs. Importantly, each read is placed at the lowest internal node N so that placement support values summed across the placements on all edges of the subtree rooted at that node N surpasses the user-specified threshold (0.95 by default). Because we are working with a taxonomic tree, placing the read onto any internal branch gives a taxonomic label to the read. Note that placement on a terminal branch yields a species or strain-level classification, but placements at internal branches may yield only higher-level classifications. For example, a read placed on an internal branch may only be classified at the family level and higher, in which case the read is unclassified at the genus and species levels.

3. Computing relative abundance: Once all reads are taxonomically classified, TIPP1 pools together the information from all marker genes, and computes relative abundances for each taxon. The relative abundance of a taxon is computed as the total number of reads classified within the taxon divided by the total number of reads classified by TIPP1.

2.2 Improvements to the reference package

In TIPP1, we used a set of bacterial marker genes that had been previously used in MetaPhyler (Liu et al., 2010) for the database: this contains 30 marker genes, with 1300 sequences per gene. In TIPP2, we changed to a set of 39 bacterial and archaeal marker genes,
which were previously used for prokaryotic species delineation by Mende et al. (2013). The RpoB gene was not included in the TIPP2 database, as it had lower precision than all the other genes (discussed later). We downloaded approximately 170,000 bacterial and archaeal genomes from the NCBI ReSeq database. We then performed a sequence of analyses that were designed to identify those genomes that had a large number of marker genes retrievable using fetchMG (Sunagawa and Mende, 2012). Finally, from that reduced set of genomes, we selected either one or two genomes per genus. This resulted in a set of ~4300 genomes per marker gene.

To identify whether a marker gene was retrievable in a given genome, we performed the following sequence of analyses. For each marker gene, we extracted sequences from each of the 170,000 genomes using the HMMs in the fetchMG tool (Sunagawa and Mende, 2012), and computed the median length of these sequences. We noticed that many gene sequences recruited by the fetchMG HMMs had lengths that were far from the median, thus suggesting that these were false positives (i.e. not truly homologous). Subsequently, we used a length-based filtering approach to remove from consideration gene sequences that were far from the median length for that gene (median ± 3 × standard deviation).

For each marker gene, a multiple sequence alignment was built on the full length sequences using PASTA (Mirarab et al., 2015). We used RAxML (Stamatakis, 2014) to generate a phylogenetic tree constrained to the NCBI taxonomy for each marker gene. Supplementary Table S2 provides the list of marker genes and corresponding statistics of the reference multiple sequence alignments. The number of sequences per marker gene ranges from 4082 to 4525, with an average of 4339 sequences per marker gene.

2.3 Improvements to the TIPP1 algorithm
In TIPP2, we changed step 1 (i.e. the way reads are assigned to marker genes). TIPP1 used the top BLAST hit (Altschul et al., 1990) to identify the marker gene for each read. Due to the way in which BLAST handles gapped alignments, the top hit may not always represent the correct marker gene (Koski and Golding, 2001; Shah et al., 2019). Therefore, TIPP2 requires the BLAST hit to cover at least a certain length (user-defined parameter, default 50 bp), and it selects the marker gene based on the hit that has maximum-length alignment to the read. TIPP2 uses BLAST instead of HMMER (Eddy, 2020) to determine the orientation of the reads with respect to the gene sequences in the reference database. When the reads align across the end of a marker gene sequence, TIPP2 trims the read to just the aligned region, a feature that also enables the use of the tool on long read data or assembled contigs.

3 Experimental study design
3.1 Overview
We set out to evaluate the performance of the improvements made in TIPP2, both with respect to the performance of TIPP1 and with respect to commonly used taxonomic profiling tools. We also evaluated the impact of the set of marker genes used as part of the reference package for TIPP2. We performed three experiments:

- Experiment 1: Testing whether abundances can be accurately estimated using a small subset of the marker genes, enabling a fast variant of TIPP2.
- Experiment 2: Comparing TIPP2 to TIPP1 on datasets with both known and novel genomes.
- Experiment 3: Comparing TIPP2, TIPP2-fast and other existing methods for abundance profiling.

In Experiment 1, we worked with training datasets where the query sequences are from ‘novel’ genomes (i.e. species that are not present in TIPP2 databases). We used the training datasets only in the design phase, to select the subset of marker genes used for TIPP2-fast. In Experiments 2 and 3, we evaluated the performance of TIPP2 and TIPP2-fast using test datasets which contain both known and novel genomes, and which are simulated with different sequencing technologies and read lengths.

3.2 Metagenomic abundance profiling methods used for benchmarking
We compared the performance of TIPP2 with TIPP1, MetaPhlAn2 (Truong et al., 2015), mOTUsv2 (Milanese et al., 2019), Bracken (Lu et al., 2017) and BLAST (Altschul et al., 1990). Except for BLAST, all of these methods are specifically designed for estimating taxonomic relative abundances in the metagenomic data, and except for Bracken, all these methods are marker-based. MetaPhlAn2 uses ~1 million clade-specific markers to detect and estimate the relative abundance of organisms. TIPP2 uses a set of 39 marker genes from Mende et al. (2013) and mOTUsv2 uses a subset of 10 marker genes described in the same study. Bracken (Lu et al., 2017) is an extension of Kraken (Wood and Salzberg, 2014) that reassigns the unclassified portion of reads based on probabilistic estimates of the true abundance profile from the Kraken output. We used BLAST as a baseline to compare TIPP2 performance, assigning each query sequence the taxonomic label of the hit selected during the read binning phase of TIPP2 (therefore the BLAST analysis utilizes the same marker gene database as TIPP2). We then calculate the abundance profile as relative abundance of reads from each taxon.

3.3 Simulated metagenomic datasets
We simulated metagenomic datasets with different characteristics, such as the average read length (range 100–250 bp), the number of genomes in the sample, sequencing technology profile and whether the datasets contain known or novel genomes. A dataset is called ‘known’ if none of the genomes in the dataset are present in the reference databases of any of the methods being tested. A dataset is called ‘novel’ if it contains genomes that were used in reference databases of at least one of the methods tested. We simulated three groups of datasets. The first group of datasets contained reads from known genomes. We used the ART sequence simulator (Huang et al., 2012) to generate three datasets from a mixture of 51 genomes. We simulated one dataset with the 454 sequencing profile and the other two with the Illumina profile and different read lengths (100 and 250 bp). Table 1 provides the overview of all datasets, and a more detailed description of how these datasets were constructed can be found in Supplementary Section S4.

To find genomes that are ‘novel’ to the methods studied, we downloaded all complete genomes from the NCBI GenBank database and identified species that are not represented in the reference database of any method. There were 133 such genomes. We selected a set of 100 genomes and created three metagenomic datasets using the ART simulator (Huang et al., 2012). We simulated one dataset with the 454 sequencing profile and the other two with Illumina profile and different read lengths (100 and 250 bp). We used the set of remaining 33 novel genomes to create a third group of datasets, which were used just for training and optimizing the TIPP2 pipeline.

3.4 Accuracy evaluation
Because we know the true abundance profile for each dataset, we can compute the error in abundance estimation. To evaluate error in the estimated abundance profile, we compute the Hellinger distance (Rao, 1995) between the estimated abundance profile and the true abundance profile. Briefly, the Hellinger distance is

\[ H_l = \sqrt{\sum_{x \in C_l} \left( \sqrt{T_x} - \sqrt{E_x} \right)^2} / \sqrt{2}, \]

where \( C_l \) is the set of clades in the true and estimated profiles for taxonomic level \( l \), \( T_x \) is the abundance of the clade \( x \) in the true profile and \( E_x \) is the abundance in the estimated profile. \( H_l \) ranges from 0 (if there is a perfect match between the profiles) to 1 (if the two profiles are fully disjoint). Note that the reads that are unclassified
at a given taxonomic level are not included in the \( H_l \) calculation for that level.

### 3.5 Running time study

We generated five replicate datasets with 2,000,000 sequences from novel genomes with two sequencing technologies—454 Roche and Illumina 250 bp. Each method was run on a Blue Waters machine with 16 CPUs and 32 GB of total memory. We report the average wall clock running time for all methods.

### 4 Results

#### 4.1 Experiment 1: testing whether fewer marker genes can correctly estimate abundances

We wanted to test whether we need the complete set of marker genes in the TIPP2 pipeline. Using training datasets (Novel-33 454 Roche, Novel-33 Illumina 100 bp, Novel-33 Illumina 250 bp), we explored the accuracy of abundance estimation for each gene separately.

We computed precision on a per-gene basis using the accuracy of taxonomic identification of reads using TIPP2. For a given taxonomic level, the precision (also called positive predictive value) is calculated as the ratio of the number of reads correctly classified (true positives) to the total number of reads classified (true positives + true negatives). We found that the RpoB (COG0085) gene consistently had lower precision than all other genes (see Supplementary Fig. S1). Hence, we removed the RpoB gene from the TIPP2 reference package. Moreover, we found that TIPP2, when run with just three genes—RpsL (COG0048), RpsK (COG0100) and RplO (COG0200)—provided better abundance estimates compared to the version that used the full set of genes (see Fig. 2). These three genes were the top three high precision genes when the genes were ranked based on average precision across taxonomic ranks and datasets. We call this version TIPP2-fast because using fewer genes for classification reduces the running time of the pipeline (see Table 2). We found that TIPP2-fast had lower error than TIPP2 at all taxonomic levels above the species level, and matched TIPP2’s performance at the species level, showing that using fewer marker genes does not negatively affect the overall accuracy of the pipeline, and can actually improve accuracy.

#### 4.2 Experiment 2: comparing TIPP2 to TIPP1

In Experiment 2, we compared TIPP2 and TIPP2-fast with TIPP1. First, we compared TIPP1 and TIPP2 with the same reference package; these results (see Supplementary Figs S3 and S4) show that the changes to the algorithmic design have minimal impact. Hence, our subsequent comparisons are between TIPP2 (and TIPP2-fast) using the updated reference package and TIPP1 using the 2014 reference package.

Figure 3 shows the average Hellinger distance for TIPP2-fast, TIPP2 and TIPP1 for known and novel genomes datasets; Supplementary Figure S2 provides Hellinger distances on individual datasets. For both known and novel genomes, TIPP2 improves on TIPP1, but the improvement is much larger for the known genome condition, where TIPP2 has a consistent and large improvement over TIPP1. TIPP2 improves on TIPP2-fast for the known genome condition at all taxonomic levels, but just matches TIPP2-fast for the novel genome condition. This suggests that when samples contain well-characterized genomes, using a comprehensive set of genes (as contained in TIPP2) can provide better abundance estimates, and that the advantage in using the larger reference package is reduced for novel genomes.

---

### Table 1. Properties of simulated datasets

| Dataset          | Test/train dataset | Known/novel genomes | Number of genomes | Number of reads | Sequencing technology | Median read length |
|------------------|--------------------|---------------------|-------------------|-----------------|-----------------------|--------------------|
| Known-51 454 Roche | Test               | Known               | 51                | 2,863,285       | 454 Roche            | 229                |
| Known-51 Illumina 100 bp | Test | Known               | 51                | 8,130,295       | Illumina              | 100                |
| Known-51 Illumina 250 bp | Test | Known               | 51                | 3,252,030       | Illumina              | 250                |
| Novel-100 454 Roche | Test               | Novel               | 100               | 48,196,018      | 454 Roche            | 173                |
| Novel-100 Illumina 150 bp | Test | Novel               | 100               | 55,730,636      | Illumina              | 150                |
| Novel-100 Illumina 250 bp | Test | Novel               | 100               | 33,397,306      | Illumina              | 250                |
| Novel-33 454 Roche | Train              | Novel               | 33                | 17,314,764      | 454 Roche            | 173                |
| Novel-33 Illumina 150 bp | Train      | Novel               | 33                | 20,052,474      | Illumina              | 150                |
| Novel-33 Illumina 250 bp | Train | Novel               | 33                | 12,031,210      | Illumina              | 250                |

Fig. 2. Experiment 1: Error in abundance profile estimates on training datasets. The training datasets are simulated metagenomic datasets from novel genomes with different sequencing technology and read lengths. We show Hellinger distance for TIPP2 using all marker genes and TIPP2-fast, which uses just three genes (RpsL, RpsK and RplO)
In Experiment 3, we used all testing datasets with known and novel genomes to compare the performance of different abundance profiling methods. Figure 4 shows the Hellinger distances for three known genomes datasets (top row), and the three novel genomes datasets (bottom row). Across datasets with different read length
and sequencing technologies, we observe very minute differences in Hellinger distances. The trends are consistent and robust regardless of the read length and technology. As expected, we observe higher Hellinger distances for all methods when working with datasets with novel genomes than datasets with known genomes.

For the known genomes dataset, MetaPhlAn2 consistently has higher error compared to other methods. Bracken has the second highest error, after MetaPhlAn2, at all taxonomic levels except at the species level where it performs similar to BLAST, mOTUsv2 and TIPP2-fast. BLAST performs very similar to TIPP2-fast in these datasets. At the genus and species levels, mOTUsv2, BLAST, TIPP2-fast and Bracken have very similar performance. mOTUsv2 and TIPP2 have the best performance at the highest taxonomic levels (phylum to family), but at the genus and species levels, TIPP2 outperforms mOTUsv2 and all other methods.

In the novel genomes datasets, TIPP2-fast and TIPP2 have similar performance except at the phylum level, where TIPP2-fast has lower error than TIPP2. At all taxonomic levels, TIPP2-fast has comparable or better performance than all other methods. BLAST has similar performance as TIPP2 at the phylum and class levels, but incurs higher errors at the order, family, genus and species levels. MetaPhlAn2, followed by mOTUsv2, have larger Hellinger distances than all other methods at higher taxonomic levels of phylum, class, order and family. At the genus level, BLAST has the worst performance, closely followed by mOTUsv2 and then MetaPhlAn2; and at the species level, BLAST and MetaPhlAn2 have the worst performance, closely followed by mOTUsv2 and Bracken.

### 4.4 Running time

We generated five replicates of 2,000,000 reads from the novel-100 genomes datasets with 454 and Illumina sequencing technologies. Table 2 shows the average wall clock time to run these methods on a computer with 16 CPUs and 32 GB memory. All methods are multi-threaded and were able to exploit parallelism by using multiple cores. Bracken was the fastest of all methods, finishing in <6 min. Both mOTUsv2 and MetaPhlAn2 performed similarly and finished in under an hour. Even though TIPP2 performs complex alignment and phylogenetic placement steps, it is able to complete within 5 h. TIPP2-fast, which uses fewer markers, is significantly faster, and completes within an hour. A detailed runtime and memory analysis of TIPP2 is described in Supplementary Section S2.

### 5 Discussion

Our prior work showed that methods based on marker genes provide better abundance estimates than the composition-based methods (Nguyen et al., 2014). In this study, we compared methods for metagenomic abundance profiling, including several based on marker genes and one (Bracken) that is composition-based. Sequencing technology did not have a significant impact on the accuracy of any of the tools tested, and sequence length had only a small impact. However, methods performed differently when novel genomes are analyzed compared to known genomes, and differences between methods were also larger. In datasets that contained sequences from genomes closely related to those within the reference packages used by the various tools, most methods performed well with TIPP2 and mOTUsv2 largely outperforming the other tools. When working with novel genomes, the performance of all methods degraded. TIPP2-fast and TIPP2 had the best performance in all datasets at almost all taxonomic levels. We conjecture that TIPP2 performs well with novel genomes due to its ability to detect distant homology through its use of sequence alignment (enhanced by the use of an ensemble of profile HMMs) and phylogenetic placement, which has shown improved recall compared to other methods (including single HMMs) in prior studies (Nguyen et al., 2016, 2015).

Our study shows that TIPP2 improves on TIPP1, a consequence mainly of using a new reference package which contains a denser reference taxonomy and multiple sequence alignment for each marker gene. This result shows the importance of the choice of database in taxonomic classification and abundance estimation tools (Nasko et al., 2018). To ensure accurate annotation and abundance estimation, a custom, environment-specific database is desirable; however, many marker-genes based methods are released with fixed databases which make it nearly impossible to customize the tool for specific applications, or even to upgrade the database as more data become available. Consequently, within the TIPP2 package, we release code and documentation for creating new references to enable end-users to create custom databases.

One of the interesting observations from our study is that we can accurately estimate abundances using just a carefully selected small set of marker genes rather than working with a comprehensive set of marker genes. This is consistent with mOTUsv2 and mOTU, which also worked with a subset of 10 marker genes from the set of 40 genes known to be single-copy and universal (Milanese et al., 2019; Sunagawa et al., 2013).

### 6 Conclusions

TIPP, introduced in 2014 (here referred to as TIPP1), provided high accuracy for abundance profiling of metagenomic reads. Here, we introduced TIPP2, an updated version of TIPP1. TIPP2 not only provides more accurate abundance profiling than TIPP1 but also outperforms commonly used taxonomic profiling tools—especially when datasets contain genomes that are not closely related to the reference sequences used by these packages. These improvements will enable a more precise characterization of microbial communities, particularly those that contain species that are not well characterized in public databases. Moreover, the biodiversity on Earth remains under-explored, and tools like TIPP2 are critical for characterizing the composition of microbial communities, many of which are expected to include currently uncharacterized genomes.

Our work indicates several directions for future research. The main improvement of TIPP2 over TIPP1 was obtained through the modification to the reference package. While TIPP1 had 30 marker genes each with about 1300 sequences, our new reference package had 39 marker genes, each with about 4339 sequences; hence, the revised reference package contains more marker genes and each marker gene sequence collection is more densely sampled. The improvement of TIPP2 over TIPP1 indicates the value added in increased taxon sampling, and suggests further improvement might be obtained by maintaining and improving the reference package used by TIPP2. For example, one of the major challenges for taxonomic annotation and abundance profiling tools is keeping up with constant rearrangements, renaming and changes in microbial taxonomy, spurred in part by metagenomic studies. As a result, taxonomic profiling tools need to be based on the most recent databases,

---

**Table 2. Average wall clock running time, in hours, analyzing five replicates of 2,000,000 reads for each sequencing technologies**

| Method          | Average wall clock running time (h) |
|-----------------|-------------------------------------|
| 2,000,000 reads from novel-Illumina datasets |                                      |
| TIPP2-fast      | 1.0                                 |
| TIPP2           | 5.0                                 |
| MetaPhlAn2      | 0.3                                 |
| mOTUsv2         | 0.3                                 |
| Bracken         | <0.1                                |
| 2,000,000 reads from novel-454 datasets |                                      |
| TIPP2-fast      | 0.8                                 |
| TIPP2           | 5.2                                 |
| MetaPhlAn2      | 0.2                                 |
| mOTUsv2         | 0.2                                 |
| Bracken         | <0.1                                |

**Note:** Each method was run on a dedicated node with 16 CPUs and 32 GB of memory. All methods have multithreading implementation, so took advantage of all available CPUs.
since these should provide the most accurate annotations. In our experiments, we found that many species had changed their order-level labels after MetaPhlAn2 databases were released, and that those changes led to higher error in our evaluations (Experiment 3). Beyond constant updates to reference packages, there is a need for developing taxonomy-agnostic annotation approaches that rely on sequence characteristics rather than man-made taxonomic labels (which can have errors, as this history indicates). We also observed that increasing the taxon sampling of the reference database improves accuracy. However, scaling up accurate phylogenetic placement to the number of publicly available sequences remains a challenge. Our study also suggests that additional investigation into the selection of a subset of the marker genes could be helpful in improving accuracy and reducing running time.

Acknowledgements

We would like to thank Michael G. Nute, Siavash Mirarab and Erfan Sayyari for discussions and ideas.

Funding

This work was supported by the US National Science Foundation awards 1513629 to T.W. and 1513615 to M.P. This study was performed on the Blue Waters supercomputer, resources operated and financially supported by UIUC in conjunction with the National Center for Supercomputing Applications. Blue Waters is supported by the NSF [OCI-0725070, ACI-1238993] and the State of Illinois.

Conflict of Interest: none declared.

References

Altschul,S.F. et al. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403–410.
Arunagam,M. et al. (2011) Enterotypes of the human gut microbiome. Nature, 473, 174–180.
Claesson,M.J. et al. (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res., 38, e200–e200.
Daniel,R. (2005) The metagenomics of soil. Nat. Rev. Microbiol., 3, 470–478.
Eddy,S. (2020) HMMER: Biosequence Analysis Using Profile Hidden Markov Models. http://hmmer.org.
Eddy,S.R. (1998) Profile hidden Markov models. Bioinformatics, 14, 755–763.
Engelbrektson,A. et al. (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. ISME J., 4, 642–647.
Govers,D. et al. (2005) Re-evaluating prokaryotic species. Nat. Rev. Microbiol., 3, 733–739.
Gilbert,J.A. et al. (2010) The earth microbiome project: meeting report of the “1st EMP meeting on sample selection and acquisition” at Argonne National Laboratory October 6th 2010. Stand. Genomic Sci., 3, 249–253.
Handelman,J. (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol. Mol. Biol. Rev., 68, 669–685.
Hess,M. et al. (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science, 331, 463–467.
Huang,W. et al. (2012) ART: a next-generation sequencing read simulator. Bioinformatics, 28, 593–594.
Huse,S.M. et al. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet., 4, e1000255.
Klaptenbach,J.A. et al. (2001) rRNA: the ribosomal RNA operon copy number database. Nucleic Acids Res., 29, 181–184.
Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol., 52, 540–542.
Liu,B. et al. (2010) MetaPhyler: taxonomic profiling for metagenomic sequences. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1–7. IEEE.
Lloyd-Price,J. et al. (2017) Strains, functions and dynamics in the expanded human microbiome project. Nature, 550, 61–66.
Lu,J. et al. (2017) Bracken: estimating species abundance in metagenomics data. PeerJ Comput. Sci., 3, e104.
Mackelprang,R. et al. (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature, 480, 368–371.
Mende,D.R. et al. (2013) Accurate and universal delineation of prokaryotic species. Nat. Methods, 10, 881–884.
Milanese,A. et al. (2019) Microbial abundance, activity and population genomic profiling with mOTUs2. Nat. Commun., 10, 1–11.
Mirarab,S. et al. (2012). SEPP: SATE-enabled phylogenetic placement. In: Biocomputing 2012. World Scientific, Hackensack, NJ, pp. 247–258.
Mirarab,S. et al. (2015) PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. J. Comput. Biol., 22, 377–386.
Nasko,D.J. et al. (2018) ReSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. Genome Biol., 19, 1–10.
Nealson,K.H. and Venter,J.C. (2001) The closest BLAST hit is often not the nearest neighbor. Qu¨estiió d’estadística i investigacio´ operativa. Quìestió: quaderns d’estadística e investigació operativa.
Segata,N. et al. (2012) Metagenomic microbial community profiling using unique clad-specific marker genes. Nat. Methods, 9, 811–814.
Shah,N. et al. (2019) Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. Bioinformatics, 35, 1613–1614.
Sinha,R. et al. (2017) Assessment of variation in microbial community ampli-con sequencing by the microbiome quality control (MBQC) project consortium. Nat. Biotechnol., 35, 1077–1086.
Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30, 1312–1313.
Sunagawa,S. and Mende,D.R. (2012). FetchMG Tool. http://vm-lux.embl.de/mende/fetchMG/about.html (7 November 2019, date last accessed).
Sunagawa,S. et al. (2013) Metagenomic species profiling using universal phylogenetic marker genes. Nat. Methods, 10, 1196–1199.
Truong,D.T. et al. (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods, 12, 902–903.
Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol., 15, R86.
Zezei,D. et al. (2019) Structural variation in the gut microbiome associates with host health. Nature, 568, 43–48.