MAGinator enables strain-level quantification of *de novo* MAGs

Trine Zachariasen[1]*, Jakob Russel[2], Charisse Petersen[3], Gisle A. Vestergaard[1], Shiraz Shah[4], Stuart E. Turvey[3], Søren J. Sørensen[2], Ole Lund[1], Jakob Stokholm[2,4], Asker Brejnrod[1] and Jonathan Thorsen[4]

[1]Department of Health and Technology, Section of Bioinformatics, Technical University of Denmark, 2800 Lyngby, Denmark,

[2]Department of Biology, Section of Microbiology, University of Copenhagen, 2100 Copenhagen, Denmark

[3]Department of Pediatrics, BC Children's Hospital, University of British Columbia, 950 West 28th Avenue, Vancouver, BC V6H 3V4, Canada

[4]COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, 2820 Copenhagen, Denmark

*Corresponding author

Abstract

Motivation

Metagenomic sequencing has provided great advantages in the characterization of microbiomes, but currently available analysis tools lack the ability to combine strain-level taxonomic resolution and abundance estimation with functional profiling of assembled genomes. In order to define the microbiome and its associations with human health, improved tools are needed to enable comprehensive understanding of the microbial composition and elucidation of the phylogenetic and functional relationships between the microbes.

Results

Here, we present MAGinator, a freely available tool, tailored for the profiling of shotgun metagenomics datasets. MAGinator provides *de novo* identification of subspecies-level microbes and accurate abundance estimates of metagenome-assembled genomes (MAGs). MAGinator utilises the information from both gene- and contig-based methods yielding insight into both taxonomic profiles and the origin of genes as well as genetic content, used for inference of functional content of each sample by host organism. Additionally, MAGinator facilitates the reconstruction of phylogenetic relationships between the MAGs, providing a framework to identify clade-level differences within subspecies MAGs.

Availability and implementation: MAGinator is available as a Python module at https://github.com/Russel88/MAGinator

Contact: Trine Zachariasen, trine_zachariasen@hotmail.com

Introduction

DNA sequencing has revolutionised our ability to gain insight into microbial compositions without relying on the ability to cultivate organisms. To explore these compositions various methods have been developed that either rely on databases of marker genes of known organisms or attempt to reconstruct the chromosomes directly from the short reads by first assembling into longer contigs and then binning these based on co-occurrences or DNA composition.

Mapping reads against marker gene databases with tools such as MetaPhlAn[1], MetaPhyler[2] and mOTUs[3] is a fast and effective way of recovering the microbial composition both because the library depth required can be quite shallow and because the computational requirements are smaller, but have limitations originating from the reliance on predefined databases, limited ability to estimate abundances at higher taxonomic resolution[4,5], and the lack of information on the functional repertoire of the identified taxa. Conversely, *de novo* binning strategies require high sequencing depth but can recover high-quality metagenome assembled genomes (MAGs) from which the functional gene content can be directly linked to a specific organism. Ideally, this can recover genomes of strains that can be used in downstream analysis to generate more specific hypotheses about associations with outcomes. One example of this is the capacity of an organism to break down Human Milk Oligosaccharides (HMOs), the main source of energy for the developing infant gut microbiome while being breastfed. Especially *Bifidobacteria* have this functionality, and it is known that certain strains or subspecies have specific preferences for certain HMO types[6–9], improving the overall utilisation of HMOs and often conferring additional benefits as a probiotic. Previously, it has been established that specifically the presence of *Bifidobacterium longum* subspecies *infantis* (*B. infantis*) together with breastfeeding, plays a crucial role in providing a protective effect to mitigate the impact of antibiotics on the early-life gut microbiome[7]. This underlines the significance of being able to accurately profile the microbiome at higher resolutions than species-level.

In this work we have developed a pipeline that takes MAGs and original reads as input and generates output including accurate abundance estimates, strain phylogenies and gene synteny clusters that can improve insights into the microbiome composition (Figure 1). We do this by grouping MAGs into clusters that are phylogenetically separated at a higher resolution than species and estimate the abundances of these. This is done by identifying a set of signature genes directly from the given data and refining them according to statistical modelling to pick the ideal set suitable for abundance estimation. The fidelity of our estimated abundances are demonstrated on the Critical Assessment of Metagenome Interpretation (CAMI) strain-madness dataset, where we benchmark MAGinator against similar tools. Additionally we show the functionality of MAGinator on a public dataset of inflammatory bowel disease (IBD) patients, where we identify differentially abundant taxa between patients and controls at high phylogenetic resolution.

MAGinator also enables Single Nucleotide Variant (SNV's) resolution phylogenetic trees, which are created from the signature genes and used for additional stratification of the MAGs and can be associated with metadata to obtain subspecies/strain-level differences. We exhibit MAGinator's ability to obtain strain-level resolutions for *Bifidobacterium* from two real-world infant datasets. In this case the signature genes were found *de novo* for one dataset and were then utilised to obtain strain-level resolution in the other cohort.

By combining the information from both contigs and gene content we identify synteny clusters of genes within strains, yielding information on shared pathways for the genes. Additionally, we show how we can associate the functional content to the identified clades, to improve hypotheses-generation on the impact of organisms, illustrated using the COPSAC$_{2010}$ cohort.

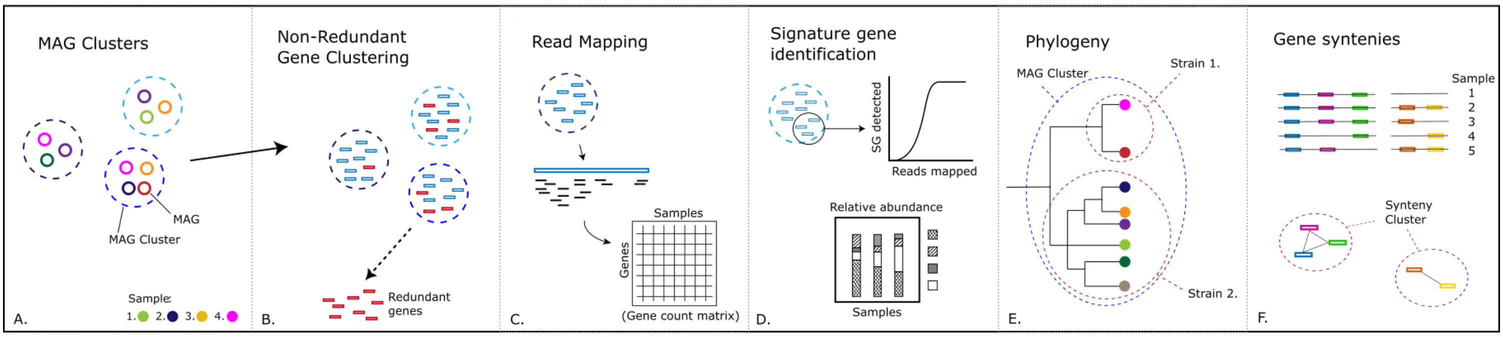

Figure 1: Schematic visualisation of the main functions of the MAGinator workflow.

Methods

Implementation

Input

The input to the MAGinator workflow comprises a set of samples with (1) shotgun metagenomic sequenced reads, (2) their sample-wise assembled contigs, and (3) sample-wise MAGs (groups of contigs from the same genome), clustered across samples, as defined by a metagenomic binning tool (see below).

Reads should be provided in a comma-separated file giving the location of the fastq files and formatted as: SampleName,PathToForwardReads,PathToReverseReads. The contigs should be nucleotide sequences in FASTA format. The MAGs should be given as a tab-separated file including the MAG identifier and contig identifier. The sample-wise MAGs should be grouped into MAG clusters representing a taxonomic entity found across the samples, which will usually be species but can also be at the subspecies level, depending on characteristics of the input data. MAGinator is flexible regarding which tool is being used for creating the MAGs, however we recommend using VAMB[10].

Dependencies

The dependencies to run MAGinator are mamba[11] and Snakemake[12] - all other dependencies are installed automatically by Snakemake through MAGinator. Additionally MAGinator needs the GTDB-tk database downloaded for taxonomic annotation of MAGs and as a reference for the phylogenetic SNV-level analysis of the signature genes.

Output generated

MAGinator generates multiple outputs and intermediate files useful for additional downstream analysis (Suppl. Table 1, Suppl. Figure 1). Importantly, MAGinator outputs the taxonomy of the MAGs, the signature genes of the MAG clusters, the sample-wise relative abundances of the MAG clusters, a non-redundant gene matrix with sample-wise mapping counts, synteny clusters and inferred phylogenies for each MAG cluster. Additionally, a folder is created containing the log information of all the jobs run by Snakemake.

Application

MAGinator is written in Python 3 and is based on a set of Snakemake[12] workflows, and easily scalable to work for both single servers and compute clusters. MAGinator is implemented as a python package and is available on GitHub at https://github.com/Russel88/MAGinator.

The MAGs are filtered based on a minimum size for inclusion, with a default size of 200,000bp. The included MAGs are taxonomically annotated using GTDB-tk (v.2.1.1)[13], by calling genes using Prodigal (v.2.6.3)[14], identifying GTDB marker genes and placing them in a reference tree. As the taxonomic annotation of the MAG clusters are found to be redundant, clusters with the same taxonomic assignment can be combined into one cluster, with the flag '--mgs_collections' which we identify as a Metagenomic Species (MGS). Redundant genes are identified by clustering with MMseqs2 (v.13.45111)[15] easy-linclust using a default clustering-coverage and sequence identity threshold of 0.8, creating a list of the representative genes along with their cluster-members. The redundant genes are filtered away, leaving a nonredundant gene catalogue. The raw reads are mapped to the gene catalogue using BWA mem2 (v.2.2.1)[16] and counted using Samtools (v.1.10)[17], leaving a gene count matrix, which is used as input for the signature gene refinement and following phylogenetic clade separation and abundance estimates.

Signature Gene Identification

We previously described the method for identifying the signature genes for the data set[18]. In brief, signature genes are selected to ensure that they 1) are unique for the MAG cluster, 2) are present in all members of the cluster, and 3) are single-copy.

To accomplish this the following steps are taken: Initially the non-redundant gene count matrix is curated to discard any genes if they have (redundant) cluster-members originating from more than one MAG cluster, as they are thus not specific for that biological entity. Subsequently, the remaining genes within each MAG cluster are sorted based on their co-abundance correlation across the samples. As the genes are unique for the species, if they are consistently detected in similar abundance across samples, it suggests that they are single-copy. This step also mitigates differences in read mappings caused by biological or technical variations. The initial set of signature genes for each biological entity are selected from the most correlated genes. Subsequently, these signature genes are further refined and optimised by fitting them to a rank-based negative binomial model that captures the characteristics of the specific microbial composition in the input data. The signature gene set is evaluated across the samples, by calculating the probability of the detected number of signature genes given the number of reads mapping to the MAG cluster. Finally the abundance of each MAG cluster is derived from the read counts to the identified signature genes normalised according to the gene lengths.

SNV-level resolution phylogenetic trees

To elucidate the smaller biological differences within the MAG clusters, MAGinator will infer a phylogeny based on the sequences of the signature genes. Based on the read mappings to the signature genes the sample-specific SNVs are called using output from Samtools mpileup. An alignment for each signature gene is made for all samples containing the signature genes using MAFFT (v.7)[19] run with the offset value of 0.123 as no long indels are expected. MAGinator allows phylogenetic inference to be calculated with either the fast method Fast-Tree (v.2)[20] (default) or the more accurate but resource intensive method IQ-TREE (v.2)[21] (--phylo ['fasttree', 'iqtree']). In samples where no MAG was found, the phylogenies can be used to detect rare subspecies-level entities based on just a few reads mapping to the signature genes and to infer functions and genes from closely related MAGs from other samples. The criteria for inclusion in the tree can be adjusted by the user. For a sample to be included in the phylogeny the following three criteria has to be met 1) minimum fraction of non-N characters in the alignment (default –min_nonN=0.5), 2) minimum number of GTDB marker genes to be detected (default –min_marker_genes=2), 3) minimum number of signature genes to be detected (default --min_signature_genes=50). The trees can be associated with metadata to obtain clade-level differences associated with study design variables such as disease phenotype, sampling location, or environmental factors.

Gene synteny

Based on the gene clustering with MMSeqs2 a weighted graph is created, which reflects the adjacency of the genes on contigs. If genes are close enough in the graph they will be categorised as part of the same synteny cluster and it is assumed that they have related functionality and/or are part of the same functional module. Clustering is determined using mcl (v.14)[22], where the user has the options to influence the adjacency count and stringency of the clusters. Only immediate adjacency is considered. By default, genes found adjacent just once are included in the graph, but this can be tuned to make more strict clusters (default –synteny_adj_cutoff=1). The inflation parameter for mcl-clustering of the synteny graph are important for the size of the gene clusters and are by default set high in order to small and consistent clusters (default –synteny_mcl_inflation=5).

Taxonomic scope of gene clusters

The taxonomic assignment of the sample-specific MAG is done using GTDB-tk. In some cases it will not be possible to assign a taxonomy to the MAG, which could be due to contamination, the MAG originating from a currently undescribed organism or due to too little information found in the MAG. In these cases an alternative is to assign the gene clusters, found in the MAG, a taxonomy. The taxonomic scope of the genes are described for the category they are almost all found in, given by a fraction defined by the user (default –tax_scope_threshold=0.9). E.g. if run with default options and a gene cluster has the assignment *"Bacteria Firmicutes_A Clostridia Lachnospirales Lachnospiraceae Anaerostipes NA"*, then at least 90% of the genes should be found in *Anaerostipes*. The algorithm will find the most specific taxonomic rank which has at least 90% agreement across the genes in the cluster assigned by GTDB-tk.

Workflow design

The MAGinator workflow has been constructed to make the information flow between the different modules automatically (Suppl. Figure 1).

The data goes through a series of filtering and processing steps (Figure 1), including:

A: Input MAG clusters, which are composed of one or more MAGs.

B: The genes are clustered and redundant genes are removed.

C: Reads are mapped to the genes, creating a gene count matrix.

D: Signature genes are identified for each MAG cluster, and used for abundance estimations

E: Based on the signature genes, SNV-level resolution phylogenetic trees are created and the taxonomic scope of gene clusters are identified.

F: Synteny-clusters of genes are identified, reflecting the adjacency of the genes on the contigs.

Benchmarking with OPAL on CAMI's stimulated strain-madness data set

The construction of the strain-madness benchmarking dataset was part of the second round of CAMI challenges[5]. The data consists of 100 simulated metagenomics samples consisting of paired-end short reads of 150 bp. The samples were run through a preprocessing workflow prior to the analysis. This involved the removal of adapters with BBDuk (v. 38.96

http://jgi.doe.gov/data-and-tools/bb-tools/) run with the following settings ‘ktrim=r k=23 mink=11 hdist=1 hdist2=0 ptpe tbo’, removal of low-quality and short reads (<75 base pairs) with Sickle (v. 1.33)[23] and removal of human contamination (reference version: UCSC hg19, GRCh37.p13) using BBmap (http://jgi.doe.gov/data-and-tools/bb-tools/) leaving an average of 6.6 million reads (SD: ±2802 reads) per sample.

To generate *de novo* assemblies, Spades (v. 3.15.5)[24] was utilised with the -meta option, with kmer sizes of 21, 33, 55 and 77, and contigs shorter than 1500 bp being discarded. Read-to-assembly mapping was carried out using BWA-mem2 (v.2.2.1)[16] and SAMTOOLS (v.1.10)[17]. Contig depths were assessed using Metabat2's jgi_summarize_bam_contig_depths (v.2.12)[25], while contigs were binned into MAGs using VAMB (v.3.0.8)[10] using default settings.

The reads, contigs and MAGs were run through the MAGinator workflow (v.0.1.16). For comparison purposes the VAMB clusters were annotated with a NCBI Taxonomy ID using CAMITAX[26]. The profile was created with Python 3 and the lineage found using NCBI’s lineage taxonomy (https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/, accessed May 9th 2023). As the strain-identifiers from the gold-standard does not exist in the NCBI database (e.g. 1313.1), we have assigned an extra number to the Taxonomy ID for the clusters which had the same species-level annotation, starting at 1 to the number of redundantly annotated clusters.

The data for the benchmarking was obtained from CAMI second challenge evaluation of profiles. The profiles used for the benchmarking in this study were selected based on the best-performing tools found in the CAMI II paper. The top 10 profiles comprise DUDes[27] (v.0.08), LSHVec[28], MetaPhlAn2[29] (v.2.9.22), MetaPhyler[2] (v.1.25), mOTUs[3] (v.2.0.1 and v.2.5.1) and TIPP[30](v.4.3.10). The profiles were compared using OPAL, which was run with default settings.

Franzosa et al. reanalysis

Processed taxa and metadata tables were obtained from the Franzosa et al.[31] supplementary materials. Raw data were downloaded from ENA using the provided accessions, and run through the preprocessing, assembly and binning before running the entire MAGinator pipeline. Four samples failed the assembly (PRISM|7238, PRISM|7445, PRISM|7947, PRISM|8550) and were excluded from all downstream analysis, both in the original and the MAGinator processed tables.

Statistical methods for abundance matrices

Abundance matrices were analysed in R (v.4.1.2). Sample management and beta diversity calculations were done in {phyloseq}[32], along with PCoA analysis. Differential abundance testing was done with the {DAtest} R package which uses the Wilcoxon test function (wilcox.test) from the {stats} package, with p-values adjusted by Benjamini-Hochberg false discovery rate correction. Corrected p-values less than 0.05 were considered significant.

Subspecies resolution of *Bifidobacterium longum*

COPSAC dataset - data characteristics and preparation

The COPSAC$_{2010}$ cohort consists of 700 unselected children recruited during pregnancy week 24 and followed closely throughout childhood with extensive sample collection, exposure assessments and longitudinal clinical phenotyping[33–35]. From the cohort, we used 662 deeply sequenced metagenomics samples taken at 1 year of age. The details of the study and sequencing protocol have previously been published[35]. The samples consist of 150-bp paired-end reads per with mean ± SD: 48 ± 15.5 million reads.

The data was analysed using the same approach as for the strain-madness data set, with the exception of filtering away reads shorter than 50 bp in the preprocessing step. This workflow yielded 880 MAG clusters for the samples.

MAGinator was run using the reads, contigs and MAGs from VAMB as input. Thus creating a set of signature genes for each MAG cluster which has been found *de novo* for this particular dataset.

CHILD dataset - data characteristics and preparation

The Canadian Healthy Infant Longitudinal Development (CHILD) study comprises a large longitudinal birth cohort with stool collection in infancy for microbiome analysis[36]. Stool samples used in this analysis were sequenced to an average depth of 4.85 million reads (SD: 1.79 million), and samples which included >1 million reads after preprocessing were kept for the current analysis[7].

We analysed a subset of the CHILD cohort, consisting of 2846 metagenomic sequenced faecal samples from infants. To overcome the shallow sequencing, the signature genes of the COPSAC$_{2010}$ cohort were used to profile the samples instead of running MAGinator. To ensure that the process of the read mappings was identical to COPSAC, the read mapping was carried out using the full gene catalogue. Next the read counts for the signature genes were extracted and used to derive sample-wise abundances for each MAG cluster.

Examining Bifidobacterium MAG clusters

The detection of signature genes for *B. infantis* for the COPSAC$_{2010}$ and CHILD cohorts was carried out by creating a binary detection matrix and using the standard function (heatmap) with default values in R. Furthermore, we compared the abundances of all the *Bifidobacterium* MAG clusters derived from MAGinator with abundance estimates from Metaphlan 3 (v.3.0.7) and strain phylogenies from Strainphlan 3 (v.3.0.7) for the species *Bifidobacterium longum*. The phylogenetic tree output by Strainphlan was converted into a distance matrix and clustered using partitioning around medoids into two clusters. The two clusters were annotated as *B. longum* subsp. *longum* (*B. longum*) and *B. infantis* based on the placement of *Bifidobacterium longum* reference genomes in the phylogenetic tree.

SNV-level phylogenetic trees for COPSAC dataset

For each MAG cluster the sequences of the signature genes were used as a reference to create an SNV-level phylogenetic tree. The trees for COPSAC$_{2010}$ were constructed with the default values of MAGinator, producing a tree in Newick file format and creating statistics for the alignment. The tree for *Faecalibacterium* sp900758465 was visualised in R using {ggtree}[37].

Gene syntenies and functional annotation for COPSAC dataset

The non-redundant genes were annotated using eggNOG mapper (v.2.0.2)[38–40] . Of the 14.7 million non-redundant genes 9.2 million were annotated. The visualisation of the synteny clusters was done with {igraph}[41].

Results

MAGinator can accurately detect strains in simulated data

The performance of MAGinator was evaluated against the top 10 taxonomic profiles found in the second round of CAMI[5] challenges using the simulated short-read ‘strain-madness’ dataset. This dataset has been selected as it represents a heterogeneous strain environment, making strain and species detection highly relevant.

Running the MAGinator pipeline on the strain-madness data, 73 MAG clusters were identified, of these 22 clusters were present with less than 3 reads in 3 samples, so the abundance was set to 0. Of these 51 remaining entities, 30 were assigned with strain-level annotation by CAMITAX.

The profiles have been compared with the Open-community Profiling Assessment tooL (OPAL)[42] (Figure 2). For the majority of the tools, the performance decreased as the taxonomic categories became less inclusive (Figure 2B & Suppl. Figure 2). The L1 norm measures the total error from the predicted and true abundance at each rank. From genus to species-level we observed drops in the average completeness 82.7-45.6% and the average purity 73.6-36.5%. MAGinator had the best average completeness at genus (99.8%) and species-levels (89.6%) (Suppl. Table 2). At the genus-level MAGinator ranked number 5 for purity at 92.4% and the best-performing tool for the species-level at 90.1%. The LSHVec gsa had the best performance for purity at genus-level with 100% however at species-level it has a purity of 37.5%, ranking number 5 in this group (Suppl. Table 3).

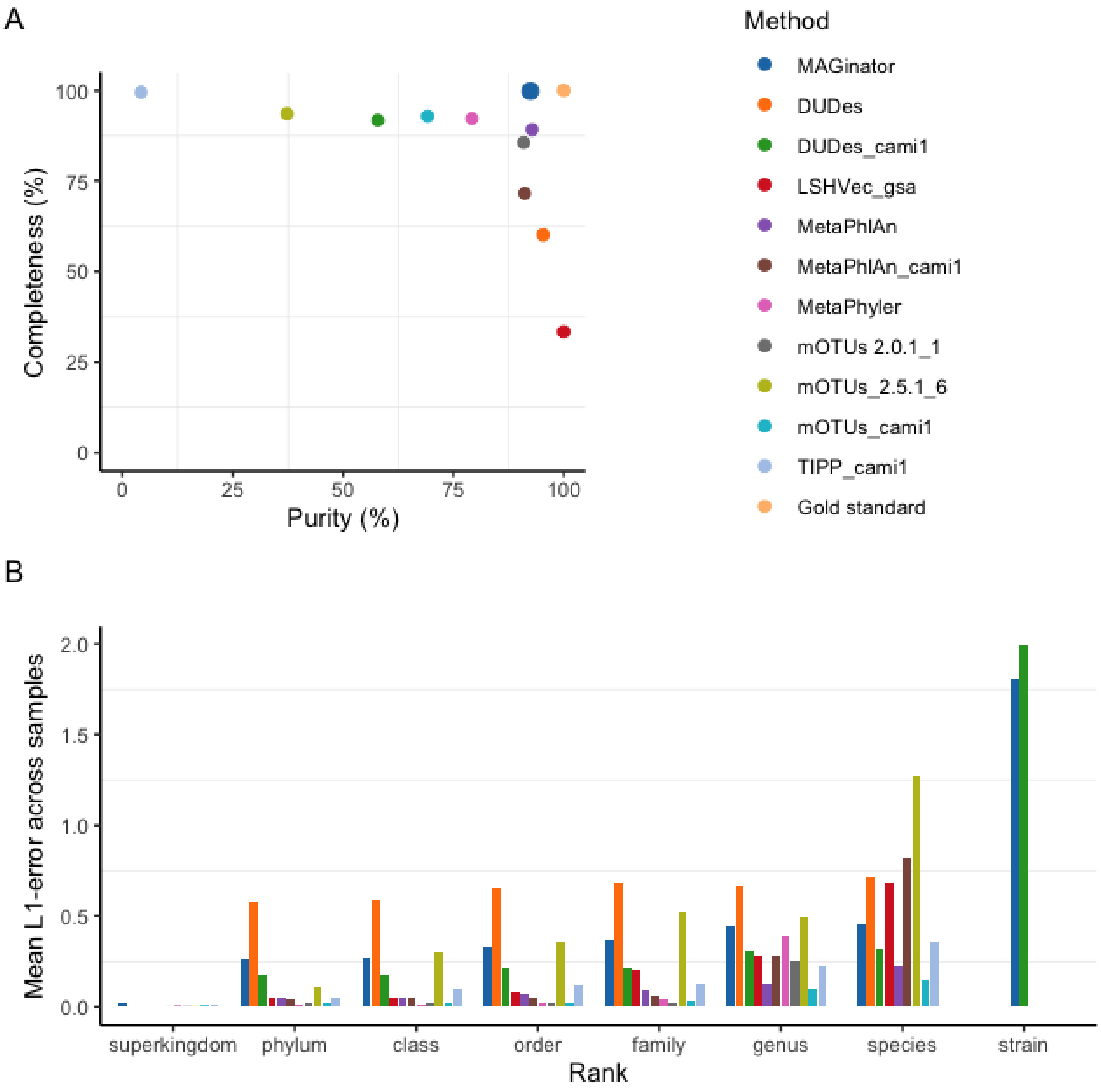

*Figure 2: **Benchmark using OPAL for comparing taxonomic profiling results for the CAMI strain-madness data set**. (A) Purity and completeness of the profiles are shown at genus-level (B) Mean of L1 norm error across samples for all ranks.*

MAGinator improves detection of relevant differentially abundant organisms

To demonstrate the advantages of quantifying bacterial taxa at high resolutions we have re-analysed a well-designed metagenomics study from Franzosa et al[31]. We chose this because it has deep sequencing well-suited for *de novo* MAG construction and a discovery/replication design with two distinct cohorts. In the absence of ground truth, replicating discoveries is a compelling strategy for making sure that findings are not false discoveries.

Beta diversity analysis of the two abundance matrices (MAGinator vs. their matrix created using MetaPhlAn2) revealed a similar separation for IBD patients vs healthy controls. For this study MAGinator produces abundance matrices of much higher dimensionality (2140 vs 201 taxa) because of the higher resolution in taxa identifications, therefore prevalence and/or abundance filtering might be relevant in MAGinator produced tables for noise reduction (Figure 3A-C).

To illustrate the improved ability of MAGinator to identify differentially abundant taxa we performed a regular differential abundance (DA) hypothesis test with Wilcoxon's test (Figure 3D-F). We looked for differentially abundant taxa defined as significant in the discovery cohort and replicated in the independent validation cohort. In the original analysis, 18 taxa were successfully validated in the independent cohort. With MAGinator, this increased to 213 taxa (Figure3 D-F).

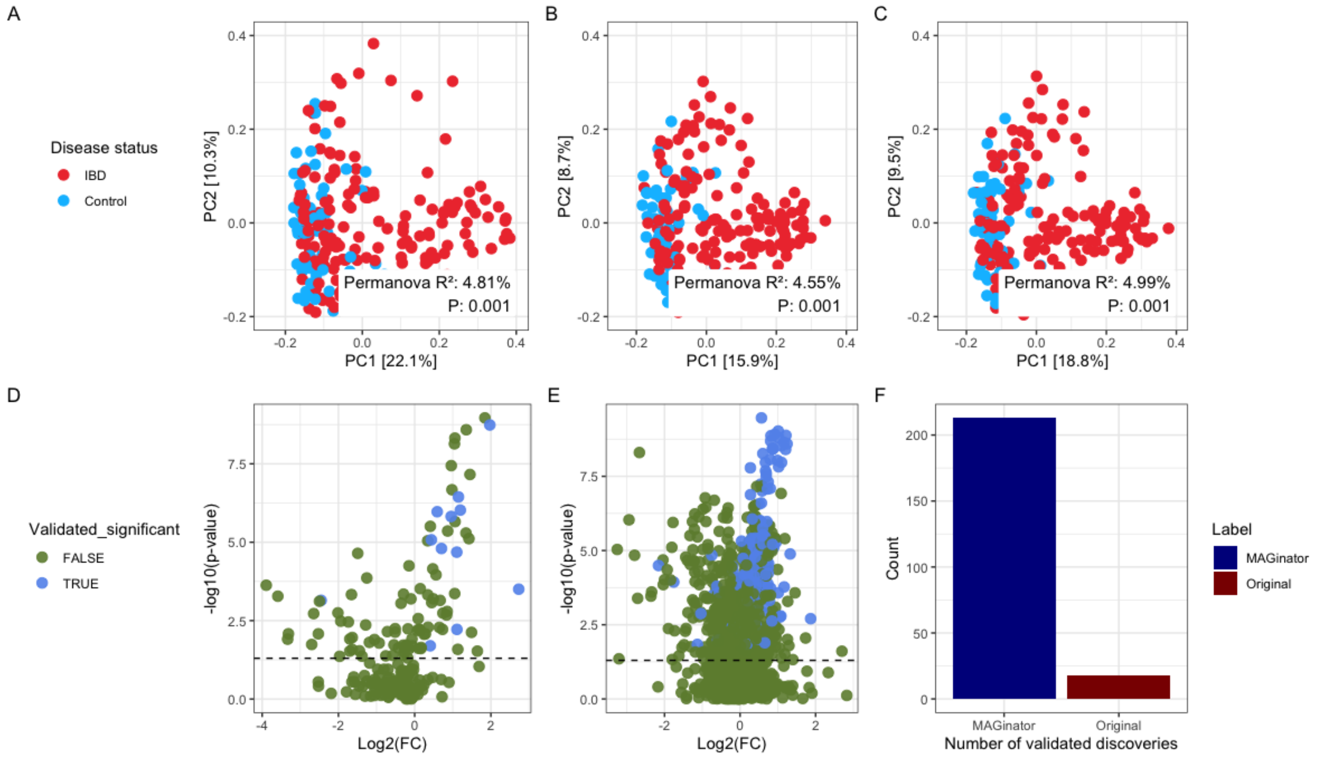

Figure 3: IBD case study shows similar performance of MAGinator with beta diversity and improvements in DA analysis. *PCoA and PERMANOVA (999 permutations) for beta diversity analysis with jsd distances and wilcoxon's test for differential abundance analysis. (A) PCoA of the original Franzosa et al. data (B) MAGinator abundances (C) filtered MAGinator abundances showing similar separation of IBD and control samples. (D) DA analysis of Franzosa et al. data, green points are taxa not significant in both cohorts (E)*

similar analysis on MAGinator abundances (F) Summary of validated discoveries using the two methods.

MAGinator enables tracking of strains across datasets at a high resolution

B. infantis is a gut microbe particularly adapted to the infant gut due to its ability to metabolise HMOs, which are complex sugars that infants cannot metabolise themselves[43]. These capabilities are different from other major subspecies including *B. longum* . Early-life colonisation with *B. infantis* has been linked to beneficial health outcomes which has sparked interest in its potential as a health-promoting infant probiotic which may even contribute to protection from asthma[7,44]. To demonstrate the utility of subspecies abundance estimation in MAGinator, we identified the signature gene set from one deeply sequenced infant cohort (COPSAC$_{2010}$) and used it to track subspecies abundances on another infant cohort (CHILD) with shallower sequencing but more samples. In the MAGinator pipeline, we identified two MAG clusters; one annotated as *B. infantis* and one as *B. longum* with GTDB-tk. In MetaPhlAn output we identified only an overall abundance for the species *Bifidobacterium longum*. Correlation analysis of these abundances shows that summed abundances of the *B. infantis* and *B. longum* MAG clusters explain 87% of the variance in the MetaPhlAn *B. longum* species (Suppl. Figure 3). In addition, we analysed the samples from both cohorts with StrainPhlAn[45] which detects strains in samples using prespecified species-level marker genes. Here, clustering of the sample-wise consensus sequences of the *B. longum* marker genes identified two clusters, one which clustered with reference strains of *B. longum* and one which clustered with reference strains of *B. infantis*. This result was previously shown for the CHILD cohort[7] and here we found similar results for COPSAC$_{2010}$ (Suppl. Figure 4). We hypothesised that this apparent duality may actually represent the underlying balance of these two subspecies in each sample. We confirmed this by comparing the StrainPhlAn-clusters with the MAGinator relative abundances of all *Bifidobacterium* species, where we saw that the StrainPhlAn clusters depended on the ratio of *B. infantis* to *B. longum* (Figure 4), but that more detailed information was accessible using the MAGinator derived relative abundances of each subspecies. This is an example of how *de novo* identification of subspecies-level MAG clusters and subsequent refinement of signature genes allows a higher resolution depiction of taxa for which the sequence coverage is sufficient in a given set of samples.

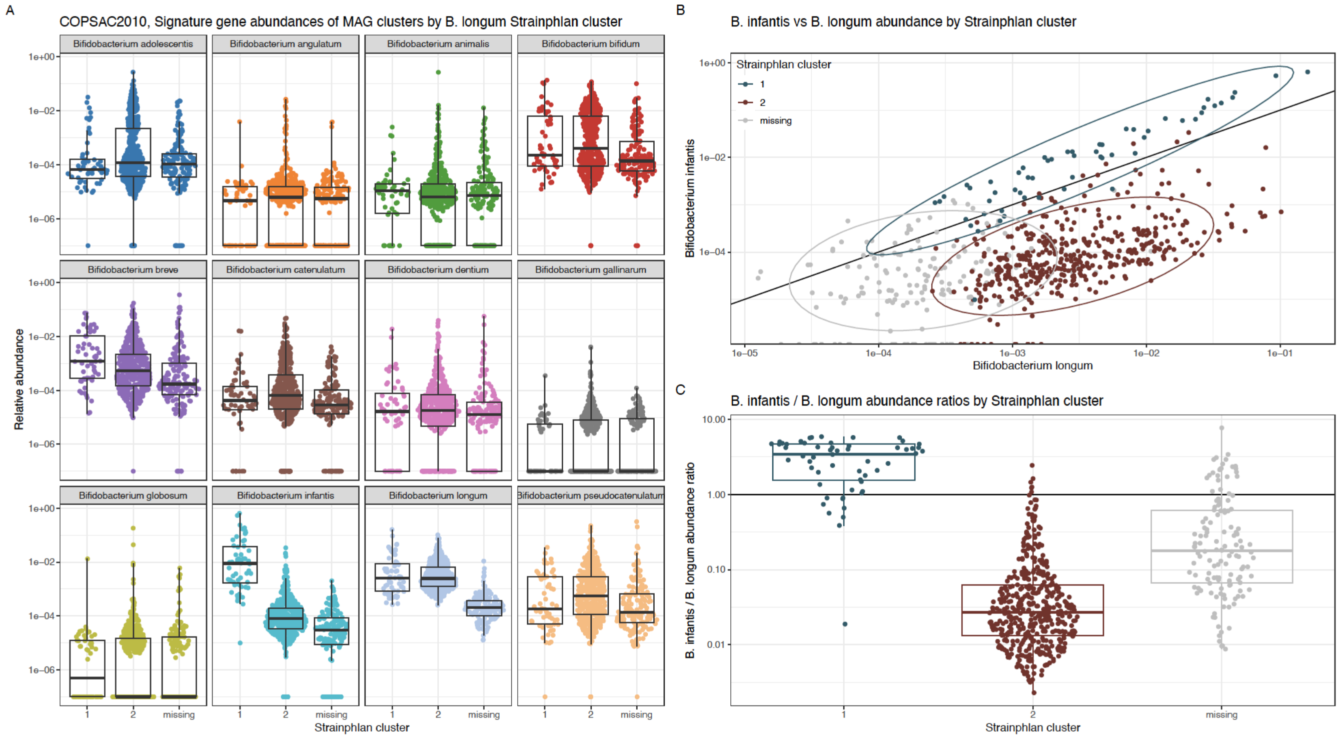

***Figure 4: Stratification of StrainPhlAn clusters using the relative abundances of Bifidobacterium longum subspecies from MAGinator** Cluster 1 indicates B. infantis and Cluster 2 indicates B. longum.*

(A) Relative abundance of StrainPhlAn clusters stratified by all Bifidobacterium clusters identified by MAGinator (B) Relative abundance of B. infantis and B. longum identified with MAGinator coloured by StrainPhlAn cluster. (C) The ratio of B. infantis to B. longum is displayed for the StrainPhlAn clusters.

Additionally we used the signature genes identified from the COPSAC cohort to track the two subspecies in the CHILD cohort. The relative abundances of the MAGinator clusters and the StrainPhlAn clusters was likewise examined (Suppl. Figure 5). When using the signature genes as a reference for the CHILD cohort MAGinator was still able to resolve the two subspecies into more well-defined clusters yielding detailed profiling of the samples.

In order to estimate the fit of the signature genes for the two cohorts we compared the read mappings and presence of signature genes (Suppl. Figure 6A). As previously described by us[18] the expected number of detected signature genes within a sample can be calculated from the number of reads that map to those genes using a negative binomial distribution. We find that the COPSAC$_{2010}$ cohort deviates with a mean squared error (MSE) of 103.95, whereas the CHILD cohort deviates with a MSE of 878.09, indicating that the signature genes are better suited for profiling of the specific strains found in the COPSAC cohort. To examine the cause of this large deviation for CHILD we created a heatmap of the read mappings to the signature genes (Suppl. Figure 6B). In accordance with Suppl. Figure 6A the samples cluster into two groups, which could be due to strain-differences. Additionally the genes are seen to cluster into multiple groups, wherefrom a group is seen to be absent in a large proportion of the samples, indicating that these genes have not been adequately selected for this strain for this dataset.

MAGinator provides SNV-level phylogenetic trees for each MAG cluster

By using the sequences of the signature genes as a reference it is possible to create a SNV-level phylogenetic tree of the samples, thus even being able to include samples in the tree, which do not contain enough reads to contain a MAG. For the MAG cluster *Faecalibacterium* sp900758465 we identified MAGs in 85 samples. For the tree 13 additional samples were included (Suppl. Figure 7), since these samples met the inclusion criteria as described in methods.

MAGinator identifies synteny clusters used for inference of functions

Genes can be grouped into synteny clusters based on their genomic adjacency. Genes close to each other in the genome will be grouped into a synteny cluster, and they are usually part of the same pathway or have a related function. Part of the MAGinator workflow creates these synteny clusters. For the COPSAC$_{2010}$ cohort 746,251 synteny clusters were identified with an average of 3 genes per cluster (Suppl. Figure 8A+B). In order to evaluate the accuracy of the synteny clusters, functional gene annotations were performed using eggNOG mapper. Subsequently, the predominant KEGG module within each synteny cluster was determined, and the proportion of genes sharing this annotation within the cluster was calculated (see Suppl. Figure 8C). Only synteny clusters with 5 or more genes and at least two annotated genes were included, leaving 35,798 clusters. For 28,341 clusters all genes in the synteny cluster were assigned the same KEGG module, and 80.5% of the modules had more than 80% agreement.

Discussion

MAGinator is a novel pipeline for quantifying the abundances of *de novo* generated MAG clusters. In contrast to reference-based abundance estimations, this allows extensive integration of abundance and functional properties for individual members of the microbial community. Furthermore, it features generation of signature gene derived phylogenies for MAG clusters and discovery of gene synteny clusters. It is implemented in Snakemake to take advantage of the integrated work distribution capabilities necessary for processing large scale metagenomics data. It features logging for ease of monitoring progress and visualisation for diagnostic purposes. We have demonstrated the functionality and utility of MAGinator via several avenues, both simulated and real datasets.

The performance of MAGinator was evaluated in comparison to existing profiling tools. We benchmarked MAGinator using the simulated strain-madness dataset produced by CAMI II. We found that MAGinator is capable of profiling samples at a comparable level to the already established tools. Notably, while many tools performed well at the genus-level, a decline in performance was observed when focusing on the species-level classification. This drop in performance is expected from reference-based methods, as they are limited to identify only what already exists in their database and are thus unable to annotate novel species. MAGinator demonstrated a notable advantage in this regard, exhibiting the highest average completeness and purity when classifying samples at the species-level. This indicates that MAGinator has the ability to achieve a more accurate and precise characterization of microbial species present in the samples. It should be noted that the high completeness by MAGinator implies a greater sensitivity in detecting and including less abundant or rare taxa in the analysis. However, it may also introduce a certain level of noise or misclassification, which influences the estimation of beta diversity.

When examining the performance of MAGinator on a real dataset the beta diversity was comparable to the analysis carried out by Franzosa et al. Reanalysing their data demonstrates how MAGinator can be used for a metagenomic association study. With the higher resolution of MAGinator when quantifying MAG clusters investigators have the possibility of discovering differentially abundant taxa in much richer detail without compromising other parts of a traditional analysis such as PCoA. Depending on the intention of the study, and the taxonomic composition of the studied microbiomes, the high resolution can also be utilised to gain deeper insights into the subspecies taxonomies. This is for instance relevant when analysing the *Bifidobacterium longum* subspecies.

B. infantis is highly relevant to investigate, as it is known for its greater capacity to metabolise HMOs compared with its closely related subspecies, such as *B. longum*. As their genomes are very similar, distinguishing them by database-dependent approaches is challenging. With StrainPhlAn we are able to identify 2 mutually exclusive clusters, each representing a subspecies, however we see that the two MAG clusters identified with MAGinator for *B. infantis* and *B. longum* yield higher resolution in the form of individual abundance estimates for each. MAGinator is able to successfully classify samples containing the subspecies in samples with low abundance and even when a MAG is not produced in that sample.

These results were reproduced in the CHILD cohort using the signature genes identified in COPSAC$_{2010}$ for the two subspecies. As samples from the CHILD cohort used in this study had lower sequencing depth, still being able to separate the subspecies is valuable. Importantly, it is worth noticing that the separation would most likely have been stronger if the signature genes had been found *de novo* for the specific cohort. This is supported by the read mappings to the signature genes showing a subset of the signature genes defined in COPSAC$_{2010}$ missing in the CHILD cohort, which presumably resulted in underestimation of the abundance for a subset of the samples. This phenomenon highlights the importance of *de novo* dataset-specific discovery of signature genes to yield the best possible abundance estimates of closely related taxonomic entities. A similar phenomenon would be expected when using database-derived strain marker genes.

From the COPSAC$_{2010}$ cohort we demonstrated MAGinators ability to create SNV-level trees based on the sequences from the signature genes of a MAG cluster, used for more fine grained stratification of the MAGs. Even in samples where no MAG was found, they are placed on the tree if they have enough reads that map to the signature genes. By placing these samples in the tree, information from the closely related MAGs can be utilised and allows detection of subspecies-level entities even for samples with very low abundance. From the clusters of the tree it is possible to associate the samples with the gene content of the related MAGs yielding information about clade-specific genes, leaving us with the ability to pair the metadata of the study with the clades and their functions.

Additionally the COPSAC$_{2010}$ cohort was used to illustrate MAGinators ability to group genes co-localised on the chromosome into synteny clusters, further combining the strengths of using both genes and contigs. As genes found close together are often part of the same genetic pathway or share the same function, this is a valuable insight for associating organisms with the outcomes of a study. This has been validated by functionally annotating the genes of the predicted synteny clusters, confirming that the genes found in synteny are often annotated to be part of the same metabolic pathway.

Conclusion

In conclusion, we have described the development of MAGinator - a pipeline for quantifying MAG clusters and demonstrated the benefits of this approach to commonly generated data types in the metagenomics field. Through reanalysis of publicly available data we have illustrated how new insights can be gained from MAGinator at a higher taxonomic resolution than available from commonly used tools. We believe that this higher resolution is key to unlocking the potential of metagenomics to identify critical strains for human health and environmental investigations. MAG cluster resolution metagenomics allows for accurate integration of abundance, taxonomic and functional annotation in microbiome studies, which is needed to empower investigations in the microbiome field.

Data availability

CAMI II strain-madness benchmarking dataset is available at https://frl.publisso.de/data/frl:6425521/strain/short_read/. The gold standard and benchmark profiles are found at https://github.com/CAMI-challenge/second_challenge_evaluation/tree/master/profiling.

The dataset from Franzosa et al. used for benchmarking is available as supplementary from their paper and the raw data is available at ENA accession SAMN08049618.

The raw COPSAC fastq files are available at NCBI BioProject PRJNA715601.

CHILD shotgun metagenomics sequencing data is available at NCBI BioProject PRJNA838575.

ACKNOWLEDGEMENTS

We express our deepest gratitude to the children and families of the COPSAC cohort studies for all their support and commitment. We acknowledge and appreciate the unique efforts of the COPSAC research team. All funding received by COPSAC is listed on www.copsac.com. The Lundbeck Foundation (Grant no R16-A1694); The Ministry of Health (Grant no

903516); Danish Council for Strategic Research (Grant no 0603-00280B) and The Capital Region Research Foundation have provided core support to the COPSAC research center. JS has received funding from the Danish Council for Independent Research (Grant no. 8045-00081B).

We thank the CHILD Cohort Study (CHILD) participant families for their dedication and commitment to advancing health research. CHILD was initially funded by CIHR and AllerGen NCE,and the metagenomic data reported here were generated with support from Genome Canada and Genome BC (274CHI).

References

1. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01688-w.
2. Liu, B., Gibbons, T., Ghodsi, M. & Pop, M. MetaPhyler: Taxonomic profiling for metagenomic sequences. in *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 95–100 (IEEE, 2010). doi:10.1109/BIBM.2010.5706544.
3. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
4. Liu, Y. *et al.* CSMD: a computational subtraction-based microbiome discovery pipeline for species-level characterization of clinical metagenomic samples. *Bioinformatics* btz790 (2019) doi:10.1093/bioinformatics/btz790.
5. Meyer, F. *et al.* Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
6. Underwood, M. A., German, J. B., Lebrilla, C. B. & Mills, D. A. Bifidobacterium longum subspecies infantis: champion colonizer of the infant gut. *Pediatr. Res.* **77**, 229–235 (2015).
7. Dai, D. L. Y. *et al.* Breastfeeding enrichment of B. longum subsp. infantis mitigates the effect of antibiotics on the microbiota and childhood asthma risk. *Med* **4**, 92-112.e5 (2023).
8. Asakuma, S. *et al.* Physiology of Consumption of Human Milk Oligosaccharides by Infant Gut-associated Bifidobacteria. *J. Biol. Chem.* **286**, 34583–34592 (2011).

9. Ojima, M. N. *et al.* Priority effects shape the structure of infant-type Bifidobacterium communities on human milk oligosaccharides. *ISME J.* **16**, 2265–2279 (2022).

10. Nissen, J. N. *et al. Binning microbial genomes using deep learning*. http://biorxiv.org/lookup/doi/10.1101/490078 (2018) doi:10.1101/490078.

11. Mamba, https://github.com/mamba-org/mamba, QuantStack & mamba contributors, 2020

12. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10**, 33 (2021).

13. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).

14. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

15. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

16. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (IEEE, 2019). doi:10.1109/IPDPS.2019.00041.

17. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

18. Zachariasen, T. *et al.* Identification of representative species-specific genes for abundance measurements. *Bioinforma. Adv.* **3**, vbad060 (2023).

19. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

20. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**, e9490 (2010).

21. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

22. Van Dongen, S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. Appl.* **30**, 121–141 (2008).

23. Joshi NA, Fass JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ

files. (2011).

24. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

25. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

26. Bremges, A., Fritz, A. & McHardy, A. C. CAMITAX: Taxon labels for microbial genomes. *GigaScience* **9**, giz154 (2020).

27. Piro, V. C., Lindner, M. S. & Renard, B. Y. DUDes: a top-down taxonomic profiler for metagenomics. *Bioinformatics* **32**, 2272–2280 (2016).

28. Shi, L. & Chen, B. LSHvec: a vector representation of DNA sequences using locality sensitive hashing and FastText word embeddings. in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* 1–10 (ACM, 2021). doi:10.1145/3459930.3469521.

29. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).

30. Nguyen, N., Mirarab, S., Liu, B., Pop, M. & Warnow, T. TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics* **30**, 3548–3555 (2014).

31. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2018).

32. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8**, e61217 (2013).

33. Bisgaard, H. *et al.* Deep phenotyping of the unselected COPSAC $_{2010}$ birth cohort study. *Clin. Exp. Allergy* **43**, 1384–1394 (2013).

34. Stokholm, J. *et al.* Maturation of the gut microbiome and risk of asthma in childhood. *Nat. Commun.* **9**, 141 (2018).

35. Li, X. *et al.* The infant gut resistome associates with E. coli, environmental exposures, gut microbiome maturity, and asthma-associated bacterial composition. *Cell Host Microbe* **29**, 975-987.e4 (2021).

36. Moraes, T. J. *et al.* The Canadian Healthy Infant Longitudinal Development Birth Cohort Study: Biological Samples and Biobanking: The CHILD study: biological samples. *Paediatr. Perinat. Epidemiol.* **29**, 84–92 (2015).

37. Xu, S. *et al. Ggtree* : A serialized data object for visualization of a phylogenetic tree and annotation data. *iMeta* **1**, (2022).

38. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).

39. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

40. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

41. Csardi, G. & Nepusz, T. The igraph software package for complex network research.

42. Meyer, F. *et al.* Assessing taxonomic metagenome profilers with OPAL. *Genome Biol.* **20**, 51 (2019).

43. LoCascio, R. G., Desai, P., Sela, D. A., Weimer, B. & Mills, D. A. Broad Conservation of Milk Utilization Genes in *Bifidobacterium longum* subsp. *infantis* as Revealed by Comparative Genomic Hybridization. *Appl. Environ. Microbiol.* **76**, 7373–7381 (2010).

44. Alessandri, G., Ossiprandi, M. C., MacSharry, J., Van Sinderen, D. & Ventura, M. Bifidobacterial Dialogue With Its Human Host and Consequent Modulation of the Immune System. *Front. Immunol.* **10**, 2348 (2019).

45. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **10**, e65088 (2021).