Benchmarking of shotgun metagenomics analysis results obtained with state-of-the-art optimized pipelines on simulated gut microbiome samples

MARIANNE BORDERES, CYRIELLE GASC, EMMANUEL PRESTAT, CAROLE SCHWINTNER
Affiliation: MaaT Pharma, 70 avenue Tony Garnier, 69007 Lyon, France

INTRODUCTION

Shotgun metagenomics analysis results are strongly dependent on the used bioinformatics software, databases and parameters. **With the aim to compare its gut microbiome data analysis performances and to identify the main drivers impacting the results, MaaT Pharma compared the taxonomic and functional analysis results obtained with its gutPrint® proprietary pipeline, MgRunner, with three state-of-the-art optimized and routinely used analysis pipelines.**

METHODS

Raw FASTQ files from two simulated datasets were shared with each pipelines' owner. One was used to evaluate taxonomic analysis results [1] and the other for functional analysis results. The construction process of both datasets is illustrated in **Figure 1.** Taxonomic expected results correspond to taxa relative abundances alongside expected values for alpha- and beta-diversity indexes. Functional expected results correspond to gene richness, KEGG Orthology (KO) [2] and CAZymes relative abundances [3].

Raw or normalized read counts were collected. Counts were rarefied using a pairwise defined threshold to allow comparisons between MgRunner v1.4.0 (MaaT Pharma) and each pipeline's results (no comparison between external pipelines). Several complementary and commonly used evaluation metrics were used to compare all analysis results **(Table 1).**

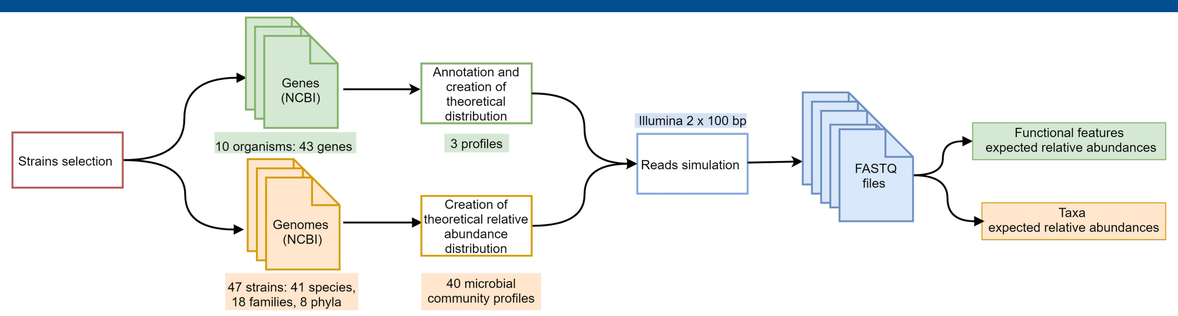

Figure 1 – Construction workflow of the simulated datasets used to benchmark taxonomic and functional results. Orange rectangles correspond to steps specific to the simulated dataset designed to evaluate taxonomic analysis results and green rectangles to the dataset designed to compare functional analysis results.

Table 1 – Evaluation metrics used to compare analysis results. TP: True Positive, FP: False Positive, FN: False Negative, Observed: Observed value, Expected: Expected value, x_{ik}: abundance of the taxon i in the sample k, s_j: number of taxa in the sample j, s_{jk}: number of taxa shared between the sample j and the sample k. Sample 1 corresponds to predicted results and sample 2 to expected results. All metrics were averaged across samples and represented as percentages. For deviations, the mean value was computed without FP (expected=0). Absolute deviations were computed on Bray-Curtis and Jaccard beta-diversity metrics which were only used in the context of the evaluation of the taxonomic analysis results.

Deviation (Dev)	**Absolute deviation** (for BC and Jaccard)	**Recall** (sensitivity)	**Precision**	**F1 score**	**Jaccard similarity** (Sim_Jaccard)	**Bray-Curtis similarity** (Sim_BC)
$\left\lvert\frac{Observed\ -\ Expected}{Expected}\right\rvert$	$\lvert Observed\ -\ Expected\rvert$	$\frac{TP}{TP+FN}$	$\frac{TP}{TP+FP}$	$2*\frac{precision*recall}{precision+recall}$	$\frac{s_{12}}{s_1+s_2\ -\ s_{12}}$	$1-\frac{\sum_i\lvert x_{i1}-x_{i2}\rvert}{\sum_i(x_{i1}+x_{i2})}$

RESULTS

Genus level

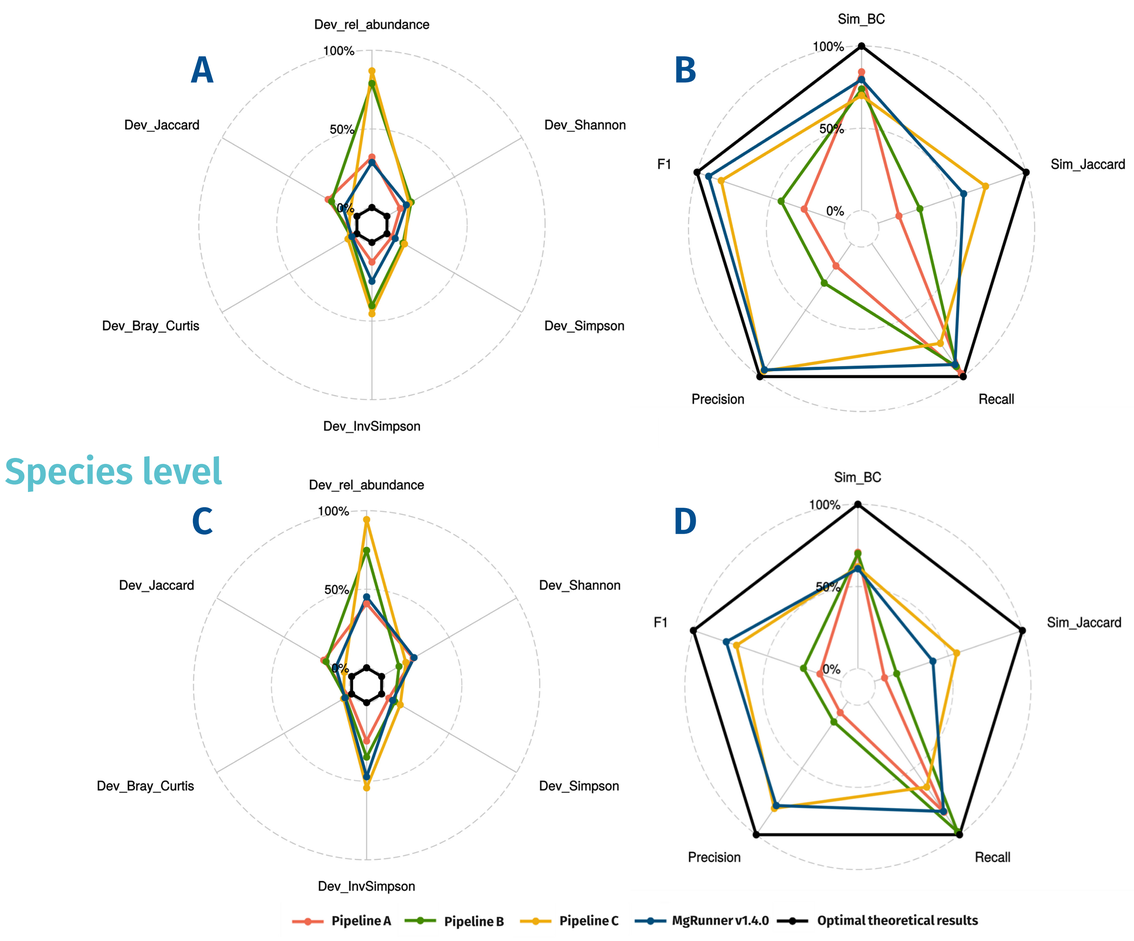

Figure 2 – Overview of benchmarking results obtained on taxonomic analysis results at genus and species levels. Benchmarking metrics, separated according to their optimal theoretical value and to the taxonomic level, are illustrated as spider plots. Metrics with an optimal theoretical value of 0% (deviations of relative abundances, of Shannon, Simpson and Inverse Simpson alpha-diversity indexes, and absolute deviations of Jaccard and Bray-Curtis beta-diversity indexes) are represented in the left panels, and metrics with an optimal theoretical value of 100% (recall, precision, F1 score, Bray-Curtis similarity and Jaccard similarity) are represented in the right panels. The optimal theoretical results are represented in black, and each pipeline's results are represented in different colors. Each pipeline's results can only be compared with MgRunner's and optimal theoretical results since different rarefaction thresholds were used on each pipeline's results. Very close results were obtained by applying these different thresholds on MgRunner's output; therefore, we represent the ones generated with the rarefaction threshold that does not change MgRunner's ranking as compared to each pipeline. However, the different thresholds were not tested on external pipelines' output. **A.** Mean deviations of relative abundances, of alpha- and beta-diversity metrics computed at the genus level. **B.** Mean F1 score, precision, recall and beta-diversity similarity metrics computed at the genus level. **C.** Mean deviations of relative abundances, of alpha- and beta-diversity metrics computed at the species level. **D.** Mean F1 score, precision, recall and beta-diversity similarity metrics computed at the species level.

Functions

Two pipelines could not be compared on all the defined evaluation metrics. Indeed, pipeline A did not generate any output for the expected functional categories and pipeline B only generated KO abundances. Hence, as compared to pipeline C and partially to pipeline B, MgRunner obtained the best performances on all computed evaluation metrics (data not shown). Using a different database version as compared to MgRunner, pipeline C had the closest results to MgRunner (4 points difference in mean KO precision), and pipeline B, which is based on a different approach and database, had the furthest ones (54 points difference in mean KO precision). Their performances can be explained by the computation of the expected results using the same functional database as the one used by MgRunner. Therefore, this has not only impacted the evaluation of pipeline B but also of pipeline C.

Genus level

Pipelines B and C showed greater deviations of relative abundances, of Shannon, Simpson and Inverse Simpson indexes at the genus level as compared to MgRunner **(Figure 2A).** These results can be explained by main differences in the used approach and database. Using a different approach but a similar database, pipeline A had lower deviation values than MgRunner for these latter alpha-diversity indexes but had a larger genus richness deviation value, mainly caused by a more permissive taxa abundance filter threshold applied as a post-processing step.
Overall, MgRunner showed the best trade-off between beta-diversity similarity metrics, F1 score (best value), precision and recall values (Figure 2B). Pipeline A was better at recovering the complete expected community structure (recall and Bray-Curtis similarity) but showed lower performances in precision and Jaccard similarity (presence-absence metric). Conversely, pipeline C showed better precision, Jaccard similarity and genus richness deviation values, but lower recall and Bray-Curtis similarity values **(Figure 2B).**

Species level

A shift in the distribution of values was globally observed at the species level, indicating that **all pipelines had a lower performance on several metrics in comparison with the genus level (Figures 2C and 2D).**
Although the trends were similar, the ranking of MgRunner in relation to the other pipelines slightly varied for a set of metrics, showing notably that the performance gap between genus and species levels was not identical between pipelines.

CONCLUSION

We did not observe any single pipeline performing best on all metrics and analyses. MgRunner v1.4.0 held a good global position and we have also identified potential improvements. **As expected, the used approach and database had a strong impact on results and performances, but so did the taxa abundance filter threshold.** For the taxonomic analysis, an imbalance between precision and recall, and between abundance-based and presence-absence metrics was overall observed. For the functional analysis, the benchmark of results was challenging, thus more effort should be made to create simulated datasets with comprehensive and database-independent expected functional annotations.
This study shows the importance of not only evaluating analysis-specific software with default parameters but also complete analysis pipelines with optimized parameters to properly assess the quality of obtained results. Such evaluations should be performed using reference taxonomic and functional standard datasets and metrics, with known expected results, allowing users to perform an independent assessment. This would allow them to be aware of the limits of analysis pipelines and therefore of the validity of the conclusions drawn from the generated gut microbiome profiles.

REFERENCES:
1. Borderes M, *et al.* A comprehensive evaluation of binning methods to recover human gut microbial species from a non-redundant reference gene catalog. NAR Genom Bioinform. 2021;3(1):lqab009.
2. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci. 2019;28(11):1947-1951.
3. Drula E, *et al.* The carbohydrate-active enzyme database: functions and literature. Nucleic Acids Res. 2022;50(D1):D571-D577.