

Reviewer 1 (Anonymous)

Thanks to the authors for addressing most of my comments. The manuscript is now clearer and richer. I have still one major comment and few minor comments.

Major comments-

The authors evaluated the quality of the reconstructions by counting the number of orphan and dead end metabolites. They claim that "The lower these values the more connected a metabolic network is". I think that this statement is not always true. We can imagine a network with a lot of dead end (outputs of the network) and orphans (inputs) that contains a single connected component. On the contrary, a network with less dead ends and orphans can be divided into several connected components. I think that the number of connected components (and optimally, the distribution of their sizes and the size of the biggest component) should complete these indices. I think that these connected components should be computed by removing ubiquitous metabolites, such as ATP, water, etc... Furthermore, it's clear that the number of connected components will strongly affect the results of the Miscoto algorithm.

Thank you for this justified remark. Using the Networkx package, we have calculated the size of the largest connected components (after removal of ubiquitous compounds) for all networks. But even after removing ubiquitous components, we have found that 80-90% of metabolites (see Figure 1 below) belong to the largest connected component and thus found this metric to be little informative. In the manuscript, we now instead present the proportion of the largest strongly connected components, which is calculated from directed graphs. The results globally correspond to the data obtained for the dead-end metabolites – the higher the size of the largest strongly connected components, the lower the number of dead-end metabolites (see Figure 2). These results are now described in the manuscript.

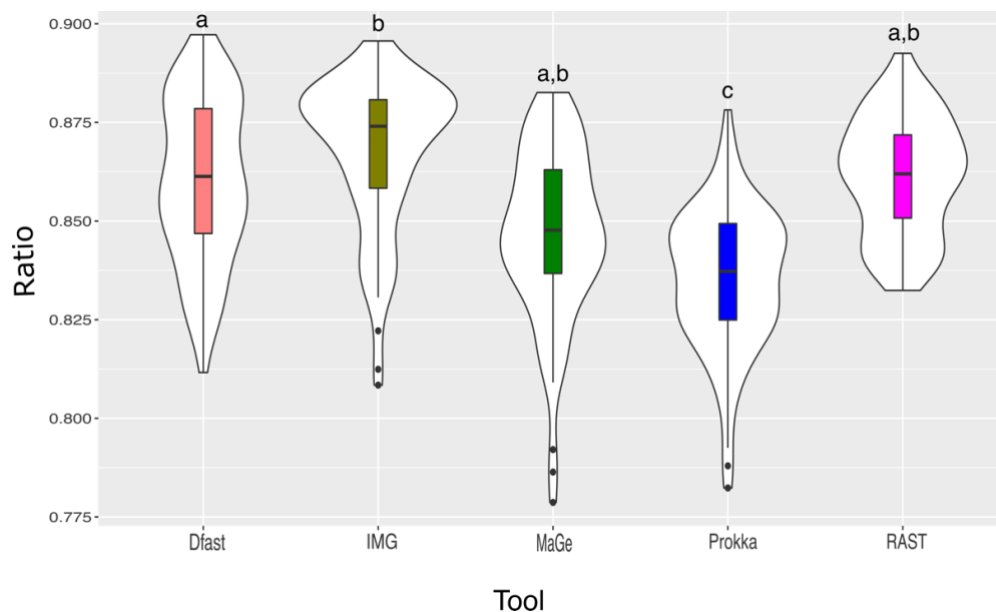


Figure 1. Ratio between the size of the largest connected components and the total number of metabolites in the metabolic networks generated based on the tested pipelines across all examined genomes. Letters above the box-plots indicate statistically (in)significant differences made by an ANOVA test. Pipelines share the same letter, the differences between them are not significant ($p \geq 0.05$).

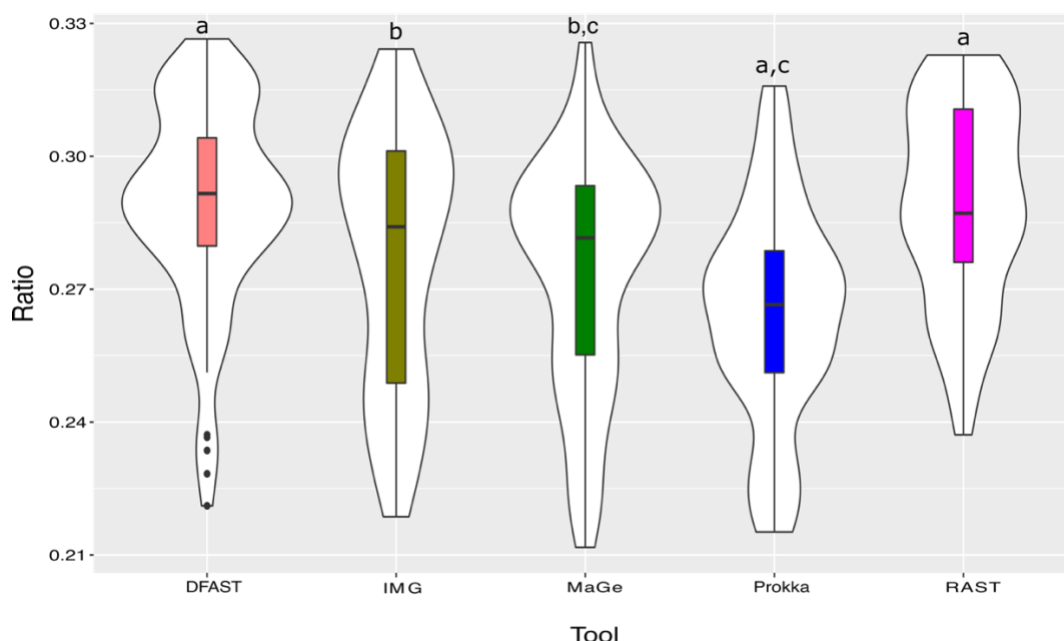


Figure 2. Ratio between the size of the largest strongly connected components and the total number of metabolites in the metabolic networks generated based on the tested pipelines across all examined genomes. Letters above the box-plots indicate statistically (in)significant differences made by an ANOVA test. Pipelines share the same letter, the differences between them are not significant ($p \geq 0.05$).

Minor comments:

l 111 : had -> had been (or has been ?)

Corrected.

l 129 : conversion to SBML already mentioned in the previous paragraph

Deleted in this paragraph.

l 190 : In terms of EC numbers, Prokka predicted more than the other pipelines -> I'm not sure that this sentence is grammatically correct.

Corrected to "Prokka predicted more EC numbers than the other pipelines".

Fig 2 : What do represent the grey polygons ? If the authors don't comment them, they should remove them to make the figure clearer.

Grey polygons connect annotations of the same genome performed with different pipelines. We now added this information to the figure legend.

l 342 : With the new EC number comparative analysis, this statement is not completely true anymore.

Yes, but curation was minimal considering that we looked at 100 reactions in over 300 networks, i.e. less than one reaction per network. We now state: "metabolic networks underwent no or very little curation"

l 355 : to reconstructed -> to reconstruct

Corrected.

Reviewer 3 (Anonymous)

The changes to the manuscript made by the authors addressed most of my comments. I feel the revision is a much more approachable version of the manuscript and more effectively communicated the results.

There is some proofreading required, but it is minor.

The manual curation of select reactions was a welcome addition. I was hoping for a more clear root-cause for the differences in the pipelines but it seems that it is elusive (no fault of the authors!).

The only thing I would have liked to have seen would been an adjustment of the pipeline parameters to a more stringent version. It is possible the pipelines converge on a largely consistent "core" reconstruction for reactions with high homology but the low homology content is highly affected by the underlying databases used by the different pipelines. In general, the e-value cutoff of $10e-5$ or $10e-6$ is very forgiving (in my experience of performing manual reconstructions) and will lead to many ambiguous reactions being added. The authors stated in their response that they "deliberately chosen to take the perspective of a standard user" but I would argue the standard user who came across this manuscript would greatly appreciate finding a study that performed some perturbation of the standard parameters.

This also ties into the final comment in the discussion about manual curation and the cost-benefit trade-off of time and reconstruction accuracy. It is very helpful to have an automated process to establish the core reconstruction and flags the ambiguous content for manual curation. I feel like this study is the perfect venue to explore the stringency parameters and the impact on the core reconstruction.

The study is still sound, but I do feel this is a missed opportunity that would increase the scientific contribution of the work.

Thank you for this comment. We agree that improvements to the final networks can probably be made by optimizing the parameter settings, but on the other hand exhaustively and systematically exploring the effect of such changes even in one pipeline could fill an entire study. Please also note that not all pipelines (RAST, MAGE, IMG) allow users to control cutoffs and for those that do, we would need to submit the same dataset to the server multiple times. However, we fully understand your curiosity regarding this point and performed additional analyses to at least touch on it in our manuscript. Notably, we now reannotated all genomes with Prokka 1.13 because it was the fastest of all used pipelines, runs locally, and easily allows for adjustments of parameter settings. For this second round of annotations, we used an e-value cutoff of $1e-15$ instead of $1e-6$ (default) and then compared the results. Not surprisingly, with increasing stringency, we observed a decrease in the number of predicted EC numbers (average 1290 vs 1406) and the number of reactions (1527 vs 1645), and an increase in the number of "hypothetical" proteins (1933 vs 1370). We also examined the effect this had on the Prokka-specific reactions and found this number to decrease from 390 to 325. Among the 20 Prokka-specific reactions that were manually curated three were removed by increasing the threshold (3.2.1.158-RXN, RXN-13750, i.e. RXN-15364), one that had been identified as high confidence, and two that had been identified as low confidence. The overall

proportion of high/low confidence or false reactions remained essentially the same as illustrated below:

	e-6 (default)	e-15
High confidence	6 (30%)	5 (29%)
Low confidence	13 (65%)	11 (65%)
False	1 (05%)	1 (06%)

This result shows that, while parameter settings affect the results, it is not that straight forward to improve them, although there are many more parameters and more levels of stringency to test.

The most important problem for optimizing parameters in our data set is that we are dealing with a multitude of different genomes from different phyla. Thus, the optimal settings are likely to vary. For instance, when annotating a new strain of *E. coli*, it may be beneficial to use a very stringent set of parameters, while, when annotating a novel family or order, less stringent parameters may be necessary. The default parameters were chosen by software/pipeline developers to work for most genomes, and as we deviate from them, we might improve predictions for one genome while obtaining poorer results for others. In the context of the prediction of metabolic consortia, however, varying parameter setting from one genome to the next may introduce new biases.

We have now added the aforementioned results to supplementary tables S1, S5, and S7(gbk files for e-value 1e-15 can be found in the given github link), and briefly mentioned these analyses both in the methods part and in the discussion. We hope you can accept our reasoning as to why we do not wish to make this an integral part of the manuscript.