









## METHOD ARTICLE

# **REVISED** ASaiM-MT: a validated and optimized ASaiM workflow for metatranscriptomics analysis within Galaxy framework [version 2; peer review: 2 approved]

Subina Mehta <sup>1</sup>, Marie Crane <sup>1</sup>, Emma Leith<sup>1</sup>, B er enice Batut <sup>2</sup>, Saskia Hiltemann<sup>3</sup>, Magnus   Arntzen<sup>4</sup>, Benoit J. Kunath <sup>4</sup>, Phillip B. Pope<sup>4</sup>, Francesco Delogu<sup>4</sup>, Ray Sajulga<sup>1</sup>, Praveen Kumar<sup>1</sup>, James E. Johnson<sup>1</sup>, Timothy J. Griffin <sup>1</sup>, Pratik D. Jagtap <sup>1</sup>

<sup>1</sup>University of Minnesota, Twin Cities, MN, 55455, USA

<sup>2</sup>Department of Bioinformatics, University of Freiburg, Georges-K ohler-Allee 106, Freiburg, Germany

<sup>3</sup>Department of Pathology, Erasmus Medical Center, Rotterdam, The Netherlands

<sup>4</sup>Norwegian University of Life Sciences,  s, 1430, Norway

**v2** First published: 11 Feb 2021, 10:103  
<https://doi.org/10.12688/f1000research.28608.1>

Latest published: 19 Apr 2021, 10:103  
<https://doi.org/10.12688/f1000research.28608.2>

## Abstract

The Earth Microbiome Project (EMP) aided in understanding the role of microbial communities and the influence of collective genetic material (the ‘microbiome’) and microbial diversity patterns across the habitats of our planet. With the evolution of new sequencing technologies, researchers can now investigate the microbiome and map its influence on the environment and human health. Advances in bioinformatics methods for next-generation sequencing (NGS) data analysis have helped researchers to gain an in-depth knowledge about the taxonomic and genetic composition of microbial communities. Metagenomic-based methods have been the most commonly used approaches for microbiome analysis; however, it primarily extracts information about taxonomic composition and genetic potential of the microbiome under study, lacking quantification of the gene products (RNA and proteins). On the other hand, metatranscriptomics, the study of a microbial community’s RNA expression, can reveal the dynamic gene expression of individual microbial populations and the community as a whole, ultimately providing information about the active pathways in the microbiome. In order to address the analysis of NGS data, the ASaiM analysis framework was previously developed and made available via the Galaxy platform. Although developed for both metagenomics and metatranscriptomics, the original publication demonstrated the use of ASaiM only for metagenomics, while thorough testing for metatranscriptomics data was lacking. In the current study, we have focused on validating and optimizing the tools within ASaiM for

## Open Peer Review

Reviewer Status  

Invited Reviewers

1

2

version 2

(revision)

19 Apr 2021



report



report





version 1

11 Feb 2021



report

1. **Caitlin Simopoulos** , University of Ottawa, Ottawa, Canada
2. **Won Kyong Cho** , College of Agriculture and Life Sciences, Seoul National University, Seoul, South Korea

Any reports and responses or comments on the article can be found at the end of the article.

metatranscriptomics data. As a result, we deliver a robust workflow that will enable researchers to understand dynamic functional response of the microbiome in a wide variety of metatranscriptomics studies. This improved and optimized ASaiM-metatranscriptomics (ASaiM-MT) workflow is publicly available via the ASaiM framework, documented and supported with training material so that users can interrogate and characterize metatranscriptomic data, as part of larger meta-omic studies of microbiomes.

### Keywords

Galaxy, metatranscriptomics, microbiome, functional analysis



This article is included in the [Galaxy](#) gateway.

**Corresponding authors:** Subina Mehta ([smehta@umn.edu](mailto:smehta@umn.edu)), Pratik D. Jagtap ([pjagtap@umn.edu](mailto:pjagtap@umn.edu))

**Author roles:** **Mehta S:** Formal Analysis, Investigation, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Crane M:** Software, Validation; **Leith E:** Writing – Review & Editing; **Batut B:** Investigation, Software, Validation, Writing – Review & Editing; **Hiltemann S:** Methodology, Software, Validation, Writing – Review & Editing; **Arntzen MØ:** Data Curation, Formal Analysis, Funding Acquisition, Writing – Review & Editing; **Kunath BJ:** Data Curation, Writing – Review & Editing; **Pope PB:** Data Curation, Writing – Review & Editing; **Delogu F:** Data Curation, Writing – Review & Editing; **Sajulga R:** Methodology, Software; **Kumar P:** Methodology, Software; **Johnson JE:** Methodology, Software; **Griffin TJ:** Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing; **Jagtap PD:** Conceptualization, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** We acknowledge funding for this work from the grant National Cancer Institute - Informatics Technology for Cancer Research (NCI-ITCR) grant 1U24CA199347, National Science Foundation (U.S.) grant 1458524 to T.J.G and a grant through the Norwegian Centennial Chair (NOCC) program at the University of Minnesota to T.J.G and M.A. The European Galaxy server that was used for data analysis is in part funded by Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and German Federal Ministry of Education and Research (BMBF grants 031 A538A/A538C RBC, 031L0101B/031L0101C de.NBI-epi, 031L0106 de.STAIR (de.NBI)).

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2021 Mehta S *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Mehta S, Crane M, Leith E *et al.* **ASaiM-MT: a validated and optimized ASaiM workflow for metatranscriptomics analysis within Galaxy framework [version 2; peer review: 2 approved]** F1000Research 2021, 10:103 <https://doi.org/10.12688/f1000research.28608.2>

**First published:** 11 Feb 2021, 10:103 <https://doi.org/10.12688/f1000research.28608.1>

**REVISED Amendments from Version 1**

1. In the updated version, we have added Dr. Phil B.Pope as a co-author. He is the principal investigator of the project that generated the metatranscriptomics datasets that we used for benchmarking.
2. The Human Microbiome Project and the role of microbiome was replaced with the Earth Microbiome Project according to the reviewer suggestion.
3. We have added references to the software tools mentioned in the Introduction along with italicising the software tools to distinguish them.
4. We have updated Figure 2 by fixing typos.
5. We reformatted the Method section into sub-categories as suggested by the reviewer.
6. An explanation was provided for the usage of FASTQ interlacer in the ASaiM-MT workflow compared to the original ASaiM, batch processing of data by using dataset collection and also specified the advantages of the tools incorporated in ASaiM-MT workflow.

**Any further responses from the reviewers can be found at the end of the article**

## Introduction

Understanding the role of microbiome diverse ecosystems such as fresh water lakes<sup>1</sup>, permafrost soils from Alaskan forests<sup>2</sup>, and deep-sea oil plumes due to oil spills<sup>3</sup> has opened up various avenues of research. In clinical research, the role of microbiomes has been studied in patho-physiological conditions such as inflammatory diseases, obesity, and cancer<sup>4</sup>. Experimental design and biological interpretation of microbiome data has become an area of intense focus as the contributions to human health and disease are becoming clearer<sup>5,6</sup>. The ‘meta-omics’ approaches, such as metagenomics, metatranscriptomics and metaproteomics have been developed to study microbiomes without culturing and target the major macromolecules that constitute the community, namely DNA, RNA and proteins. While metagenomics (16S rRNA or whole genome sequencing) focuses on the taxonomy profile and functional potential<sup>7</sup>, metatranscriptomics, metaproteomics and meta-metabolomics<sup>8</sup> uncover the functional response of the microbiome to stimuli on the short and long time-scale, respectively<sup>9,10</sup>.

Metatranscriptomics has been used to analyze microbial gene expression profiles from a variety of complex sample types, e.g. human microbiome, aquatic or terrestrial environments, plant-microbe interactions<sup>11</sup>. Despite these applications, challenges still exist in the analysis of the complex metatranscriptomics data. Metatranscriptomics data is usually generated using high-throughput sequencing of short RNA-Seq reads using Illumina sequencing technology<sup>12</sup>. Many software tools and workflows are available for metatranscriptomics analysis. These include tools for RNA-Seq Data Preprocessing: Quality Control (*FastQC*<sup>13</sup>), Ribosomal RNA removal (*SortMeRNA*, *barnap*<sup>14</sup>), host RNA removal (*BMTagger*<sup>15</sup>), De Novo Assembly (*Trinity*<sup>16</sup>, *MetaVelvet*<sup>17</sup>, *Oases*<sup>18</sup>, *IDBA-MT*<sup>19</sup>), Transcript Taxonomy (*Kraken*<sup>20</sup>, *GOTTCHA*<sup>18</sup>, *MetaPhlan2*<sup>21</sup>), Functional Annotation (*HUMAnN2*<sup>7</sup>), Annotation of assembled contigs are subjected

to gene finding programs such as *FragGeneScan*<sup>22</sup> followed by functional assignment using *DIAMOND*<sup>23</sup> searches against *KEGG*<sup>24</sup>, *NCBI RefSeq*<sup>25</sup>, *UniProt*<sup>26</sup>. Differential Expression analysis is performed by tools such as *EdgeR*<sup>27</sup>, *DeSeq2*<sup>28</sup> and *limma*<sup>29</sup>. “Reads-Based” analysis is performed by tools such as *MetaTrans*<sup>30</sup>, *COMAN*<sup>31</sup>, *FMAP*<sup>32</sup>, *SAMSA2*<sup>33</sup>, *ASaiM*<sup>34</sup> and Assembly Based analysis: *SqueezeMeta*<sup>35</sup>, *IMP*<sup>36</sup>, *MOSCA*<sup>37</sup>. Some of these open source tools<sup>38</sup> have been incorporated within the Galaxy bioinformatics workbench<sup>39</sup> to make it more accessible to users on a single platform.

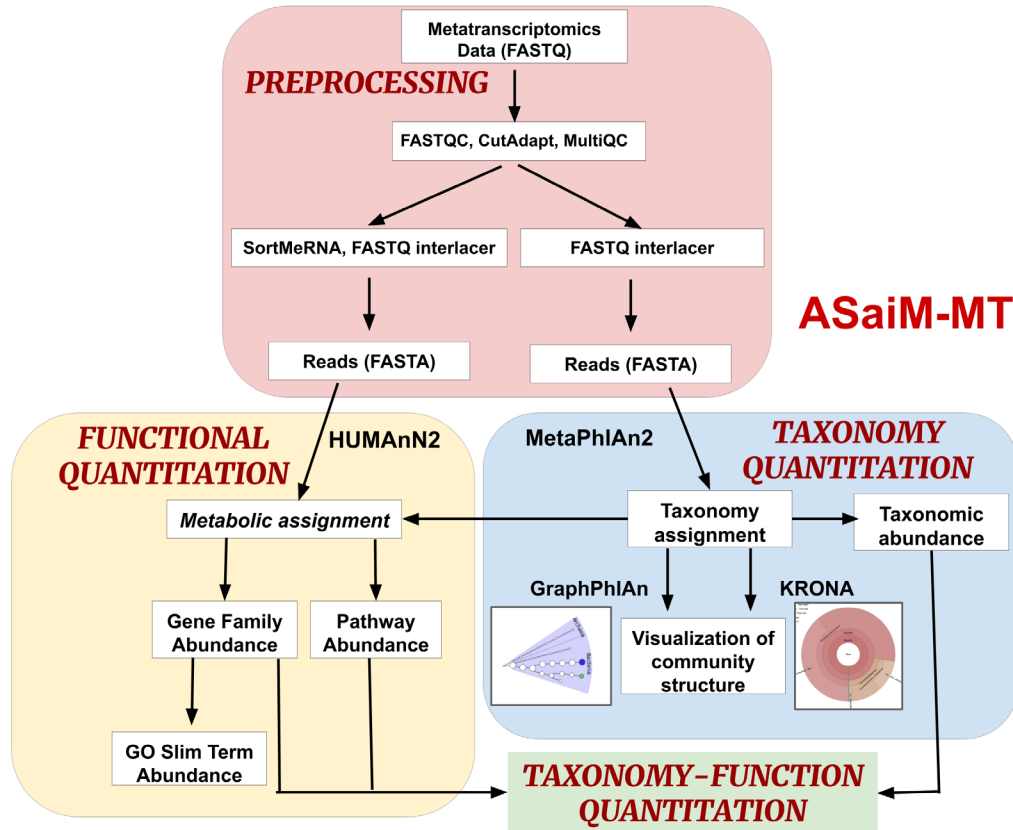
ASaiM framework was previously developed by Batut *et al.* to perform metagenomics and metatranscriptomics data analysis<sup>40</sup>. The major goal of ASaiM was to develop an accessible, reshareable, and user-friendly framework for microbiome researchers, implemented within the Galaxy platform<sup>41</sup>. The framework integrates a comprehensive set of microbiota related tools, pre-defined and tested workflows as well as supporting training material and documentation. It is available for users as a Docker image but also as a web server (<https://metagenomics.usegalaxy.eu/>). This implementation also enables flexibility, so that the workflow can be customized for datasets of diverse origin as new software tools or methods emerge. To address the need for optimizing ASaiM for metatranscriptomics data, we added the ASaiM-metatranscriptomics (ASaiM-MT) (Figure 1), a metatranscriptomics workflow, and rigorously tested it to ensure reliable analysis of metatranscriptomics data. Our testing and validation focused on using contemporary tools in their most current version (Table 1), capable of handling large datasets, and ensuring that the outputs from each of the tools were compatible in order to build an integrated and automated workflow. The workflow also has potential for integration with other meta-omic tools and workflows in Galaxy, such as those designed for metaproteomics<sup>42</sup>, to enable multi-omic data analysis.

The ASaiM-MT workflow is available via the ASaiM framework, specially at <https://metagenomics.usegalaxy.eu/>, for users to test their metatranscriptomics data. It is supported by a step-by-step tutorial<sup>43</sup>, available on the Galaxy Training Network (GTN)<sup>41</sup>, which provides explanation for the different steps and the opportunity for online, hands-on training in using the workflow, with a trimmed dataset. The Galaxy training network also provides online support through the Gitter channel (<https://gitter.im/Galaxy-Training-Network/Lobby>), where the users can interact with the developers.

## Methods

### Workflow implementation

The ASaiM-MT workflow contains all the processing steps and the parameters required for the metatranscriptomics analysis from RNA-Seq data collected under a single biological condition. This workflow is also compatible with the single-end sequencing reads although parameters have to be changed to accommodate this input. This workflow is a multi-step analysis with preprocessing/ data cleaning, taxonomy analysis and functional analysis. The starting data input for the workflow are the FASTQ files - forward and reverse reads (obtained from the Illumina sequencer). We describe below the tools and their functions in the workflow. For comparative analysis of multiple



**Figure 1. ASaiM-MT workflow: The workflow is divided into 4 parts.** (i) Preprocessing: Process raw metatranscriptomics data to perform further analysis. (ii) Taxonomy Quantitation: Assignment of taxonomy along with abundance values and visualization. (iii) Functional Quantitation: metabolic assignment of identified functions and gene and pathway abundance annotation. (iv) Taxonomy-Function Quantitation: combine taxonomy and functional quantitation values into relative abundance values at different levels such as e.g., the abundance of a pathway between phyla.

biological conditions, the users have an option to use the MT2MQ tool to generate inputs for statistical analysis (see discussion).

1) **Preprocessing**

i) **Input files:**

Our optimized ASaiM-MT workflow (available via <https://metagenomics.usegalaxy.eu/>) accepts Illumina paired end FASTQ sequence files (Forward read and Reverse read). As an alternative, a single-end FASTQ sequence can also be used as an input, with minor modifications in the downstream processing tool (such as changing the sequence type in *CutAdapt* and *Filter with SortmeRNA* as single end reads and also removing the *FASTQ interlacer* tool).

ii) **Quality control:**

Occasionally, sequencing can introduce incorrect identification of nucleotides and these errors can lead to misinterpretation of the data, thus bringing in the need to preprocess (Figure 2(a)) the data before analysis. The first step in our analysis is to perform quality control to remove such sequencing errors. For this, we use *FastQC* to assess the quality

of each sample and *MultiQC* to combine each result into a single report. The quality control profiles and the rRNA sequence proportion changes in case of metagenomics data.

iii) **Adapter Trimming:**

To improve the quality of the data, *CutAdapt* was used to trim low-certainty bases from reads, filter out reads of poor quality or short length, unwanted sequences, including adapters, primers, and poly-A tails. The ASaiM-metagenomics shotgun workflow uses *Trim Galore!* for trimming of adapters. *Trim Galore!* works as a wrapper that includes *CutAdapt* and *FastQC*. However, for the ASaiM-MT workflow we chose *CutAdapt*<sup>44</sup> for adapter trimming because it is more error tolerant, processes fast and modifies and filters reads according to user's preference compared to *TrimGalore!*.

iv) **RNA Filtering:**

Next, *SortMeRNA*<sup>45</sup> was used to remove any rRNA sequences, which are often used for easy taxonomic characterization of microbiomes but do not provide functional information.

**Table 1.** Enhancements in the ASaiM-MT workflow as compared to the original ASaiM shotgun metagenomics workflow.

Tool Function	ASaiM Shotgun Metagenomics	ASaiM-MT	Updates
Quality control	FASTQC	FASTQC	Version change (0.69 → 0.72)
		<b>MultiQC</b>	Tool added
Adapter Trimming	TrimGalore!	<b>CutAdapt</b>	Tool replaced
Dereplication	VSearch	-	
rRNA selection	FilterwithSortmeRNA	FilterwithSortmeRNA	Version change (2.1b.4 → 2.1b.6)
Interlacing	FASTQ-join	<b>FASTQ interlacer</b>	Tool replaced
Taxonomic assignment	MetaPhlan2	MetaPhlan2	No change
Formatting for the different taxonomic levels	Format MetaPhlan2	Format MetaPhlan2	No change
Functional assignment	HUMAN2	HUMAN2	No change
Visualization	Export to GraPhlan	Export to GraPhlan	Parameters changed
	Krona pie chart	Krona pie chart	Version change (2.6.1 → 2.6.1.1)
	GraPhlan	GraPhlan	No change
	Generation, personalization and annotation of tree	Generation, personalization and annotation of tree	No change
Regroup to GO terms	Group abundances	<b>Group abundances</b>	Tool updated
Unpack Pathway abundance	-	<b>Unpack Pathway abundance to show gene families</b>	Tool added
Extracting Gene level information	-	<b>Create gene level families file</b>	Tool added
Text manipulation tools	-	<b>Select, Sort, Group</b>	Tools added

We eliminated the step of de-replication of reads (*V-Search*) in the ASaiM-MT workflow, in order to retain the multiple copies of sequences for metatranscriptomics quantitation.

The final step in cleaning and processing the data is to interlace the forward and reverse reads since the following steps require a single file per sample. For performing this action, the original ASaiM Shotgun workflow used the *FASTQ-joiner* to join the reads. However, in the ASaiM-MT version, we use the *FASTQ interlacer*. *FASTQ interlacer* joins the forward (/1) and the reverse reads (/2) using the sequence identifiers; sequences without designation will be named as single reads. The reason ASaiM-MT uses *FASTQ-interlacer* rather than *FASTQ-joiner* is because the joiner tool combines the forward and reverse read sequence together while the interlacer puts the forward and reverse read sequences in the same file while retaining the entity of each read along with an additional file with unpaired sequences. The interlacer tool mainly replaced the joiner tool

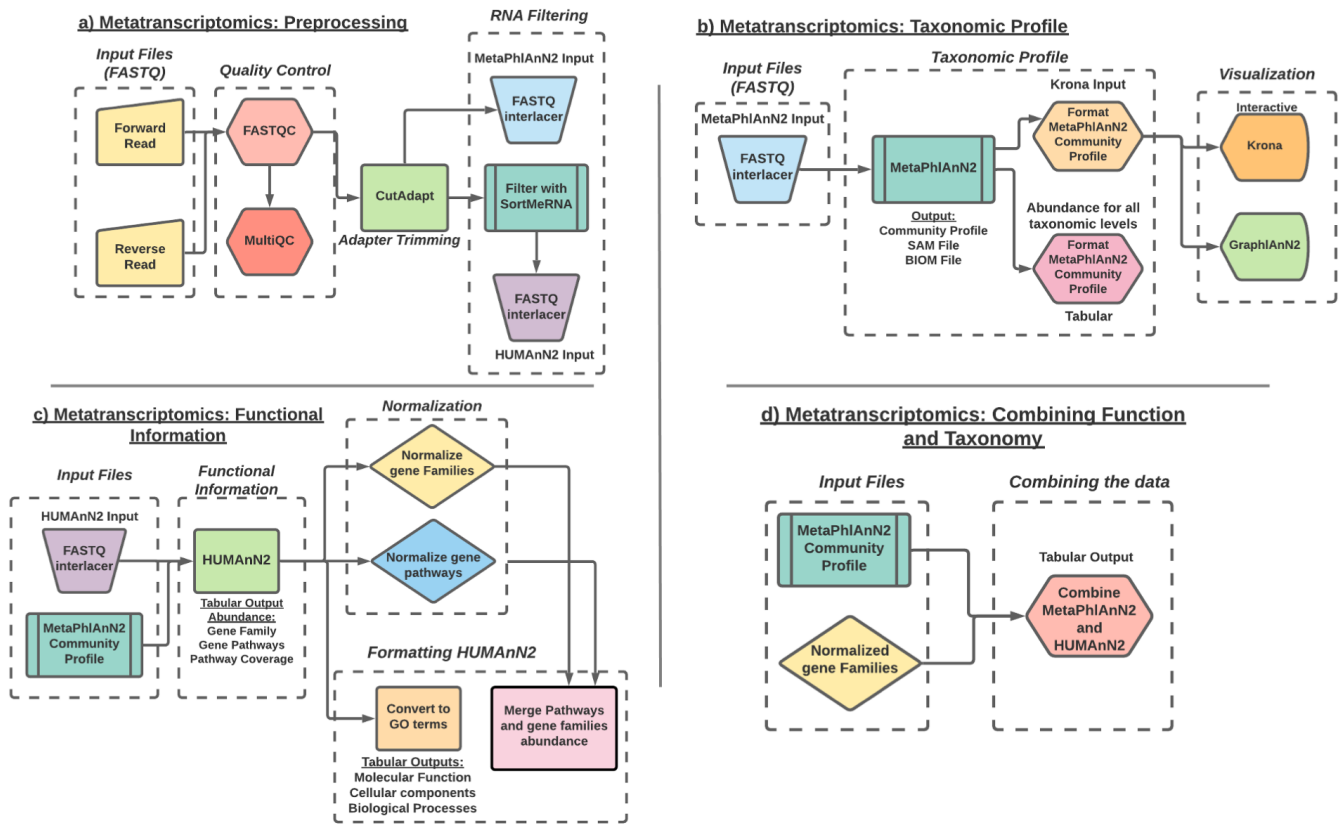
in the ASaiM-MT workflow, as we wanted the output file to maintain the integrity of the reads and to distinguish between forward and reverse reads. We perform the interlacing on the data both before and after the *SortMeRNA* step since the following steps require both data with and data without rRNA.

## 2) Extraction of taxonomic profile

To understand a microbial community, we must first understand which organisms are present along with their abundance. There are several approaches to microbial taxonomic profiling, but this workflow (Figure 2(b)) uses the marker gene approach.

### i) Taxonomic profile:

*MetaPhlan2* checks every read against a database of approximately one million clade-specific marker genes from nearly 17,000 reference genomes (bacterial, archaeal, viral, and eukaryotic)<sup>21</sup>. For this particular step, we use all reads, including rRNA since they are useful for taxonomic



**Figure 2. Diagram of the ASaiM-MT workflow.** **a)** Preprocessing workflow: Workflow representation of the tools involved in quality check, data trimming and RNA filtering. **b)** Taxonomic profile workflow: workflow representation of taxonomy assignment tool (MetaPhlAn2) and post processing of the data using the Format MetaPhlAn2 tool. The workflow includes visualization of the data using interactive Krona and GraPhlAn plots. **c)** Functional information workflow: representation of tools involved in functional annotation (HUMAN2), normalization of the data **d)** Combine Functional-Taxonomy abundance workflow: workflow representing tools that combines (Combine MetaPhlAn2 and HUMAN2 outputs) and groups (Group abundances into GO slim terms) the functional and taxonomy output.

profiling. The outputs of *MetaPhlAn2* are a SAM (sequence alignment map) file and a BIOM (Biological Observation Matrix) file, which both show the mapping of reads onto the reference database, and Community Profile tabular output, which contains information regarding the taxa present, and their relative abundance. This table includes information at all taxonomic levels, so to parse it out into each separate level we use the *Format MetaPhlAn2* tool.

ii) Visualization:

We use two different tools to visualize the taxonomic profiles. First, we use *Krona*<sup>46</sup>, which creates an interactive pie chart from the hierarchical taxonomic data. This chart is multi-layered for each taxonomic level and can be zoomed for viewing at each level. *GraPhlAn*<sup>47</sup> is the other visualization tool, which creates a publication-ready circular representation of a phylogenetic tree based on the taxonomic results. We must first use *export2graphlan* to convert the *MetaPhlAn2* results to a format that *GraPhlAn* can use.

3) Extraction of functional information

i) Functional information:

After characterizing the taxonomic profile of each sample, we must determine which genes are expressed and the biological processes involved. To perform functional profiling (Figure 2(c)), we use *HUMAN2*<sup>27</sup>, which is a pipeline to quickly and accurately determine the presence and abundance of functional gene families and pathways from metagenomic or metatranscriptomics data.

ii) Normalization and Formatting HUMAN2 output:

For identifying the functions expressed by the community, we filter out the rRNA sequences, due to their high noise levels and the compute time needed for their analysis. The output of *SortMeRNA* and the identified community profile from *MetaPhlAn2* helps *HUMAN2* to focus on the known sequences for the identified organisms. Software tools such as ‘Renormalize’, ‘Unpack Pathway abundance to show gene families’, ‘Create gene level families file’, ‘Group abundances’ and text manipulation tools such as ‘Select

lines', 'group columns' and 'sort columns' have been introduced in the ASaiM-MT workflow. This generates normalized abundances of gene families, gene pathways and GO slim terms present in each sample.

4) Combine taxonomic and functional information

The final step (Figure 2(d)) in this analysis is to determine which microorganisms are contributing to the profile of functions indicated by the expressed RNA sequences. HUMAnN2 partially answers this question by including taxa in its gene family and pathway outputs, but it does not include the taxa's abundance, only the functional abundance. We can fill in this missing information with the MetaPhlAn2 results using the Combine MetaPhlAn2 and HUMAnN2 Outputs tools. This produces a table of functional terms and their abundances with the corresponding genus and species abundances for the taxa which contribute to said function via their expressed RNA sequences. The abundances are reported in RPK values (reads per kilobase), calculated as the sum of the scores for all alignments for a gene family. These alignment scores are calculated according to the number of matches of a specific sequence to its reference genome and further normalized to account for multiple reference genome matches.

**Results**

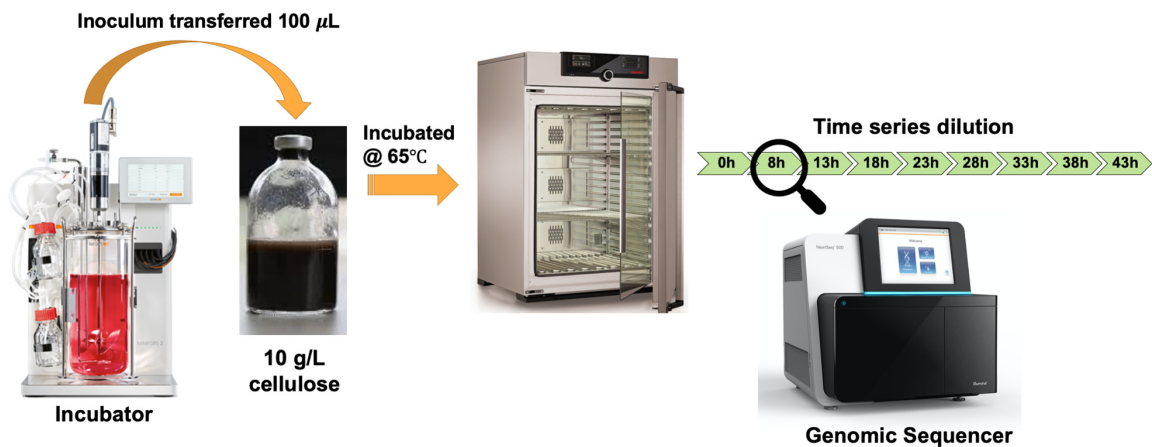
To demonstrate the use of the ASaiM-MT workflow, we analyzed a representative metatranscriptomic data set obtained from a microbial community within a thermophilic biogas reactor<sup>48</sup> which digests municipal food waste and manure (Figure 3). The microbial community was sampled from the bioreactor and transferred to a rich medium containing lignocellulose from Norwegian Spruce and incubated at 65°C as an enrichment strategy. Triplicate mRNA samples were taken in a time series from 0 to 43 hours after inoculation. The mRNA was sequenced for paired-end reads (2 x 125 bp) on one lane of an Illumina HiSeq 3000. For the purpose of this study, we took only one

time point (8hr). The paired FASTQ files (forward and reverse reads) were then subjected to the ASaiM-MT workflow (Figure 2).

The ASaiM-MT workflow consists of four steps, i) preprocessing of the data, ii) extraction of community profile, iii) extraction of functional information, and iv) combining taxonomic and functional information. The data is preprocessed to make it compatible for MetaPhlAn2 (taxonomy) and HUMAnN2 (Function) annotation of the data.

To extract the taxonomic profile, the MetaPhlAn2 suite was run on the adapter trimmed interlaced files. The Community Profile output contained the information regarding the microbiome community present in the sample along with its relative abundance at different levels, i.e., Kingdom, Phylum, Class, Order, Family, Genus, Species, and Strain (Figure 4). For example - (k\_\_Bacteria|p\_\_Firmicutes|c\_\_Clostridia|o\_\_Thermoanaerobacterales|f\_\_Thermodesulfobiaceae|g\_\_Coprothermobacter|s\_\_Coprothermobacter\_proteolyticus) states that kingdom is Bacteria, class is Clostridia, belonging to order Thermoanaerobacter, family of Thermodesulfobiaceae, genus is *Coprothermobacter* and species is *Coprothermobacter proteolyticus*. As it is a cellulose-degrading consortium from anaerobic digestion, *Coprothermobacter* and *Clostridium* were expected to be identified for this dataset, demonstrating the accuracy of these tools.

The community profile is further processed using the Format MetaPhlAn2 tool which splits the MetaPhlAn2 output and categorizes them into various taxonomy levels (Kingdom, Phylum, Class, Order, Family, Genus, Species) with corresponding abundance values. Supplementary Figure S1 (Extended data<sup>49</sup>) shows genus level relative abundance values associated with genera present in the dataset. For visualization of the taxonomic profile, GraPhlAn and Krona (interactive) were used (Figure 5a and 5b).



**Figure 3. Graphical representation of Biogas reactor Dataset.** A 100 µl inoculum was collected from a lab-scale biogas reactor incubated at 55°C and transferred to an anaerobic flask containing 10 g/L of cellulose. Triplicate mRNA samples were taken in a time series from 0 to 43 hours after inoculation. Metagenomic and metaproteomic sequencing was performed for all time points. For this tutorial we used one of the triplicates (T1A) in the 8 hours' time point.

1	2
#SampleID	Metaphlan2_Analysis
k_Bacteria	99.73011
k_Archaea	0.26989
k_Bacteria p_Firmicutes	99.68722
k_Archaea p_Euryarchaeota	0.26989
k_Bacteria p_Proteobacteria	0.04289
k_Bacteria p_Firmicutes c_Clostridia	99.68722
k_Archaea p_Euryarchaeota c_Methanobacteria	0.26989
k_Bacteria p_Proteobacteria c_Gammaproteobacteria	0.04289
k_Bacteria p_Firmicutes c_Clostridia o_Clostridiales	76.65512
k_Bacteria p_Firmicutes c_Clostridia o_Thermoanaerobacterales	23.0321
k_Archaea p_Euryarchaeota c_Methanobacteria o_Methanobacteriales	0.26989
k_Bacteria p_Proteobacteria c_Gammaproteobacteria o_Enterobacteriales	0.04289
k_Bacteria p_Firmicutes c_Clostridia o_Clostridiales f_Clostridiaceae	76.65512
k_Bacteria p_Firmicutes c_Clostridia o_Thermoanaerobacterales f_Thermodesulfobiaceae	23.0321
k_Archaea p_Euryarchaeota c_Methanobacteria o_Methanobacteriales f_Methanobacteriaceae	0.26989
k_Bacteria p_Proteobacteria c_Gammaproteobacteria o_Enterobacteriales f_Enterobacteriaceae	0.04289
k_Bacteria p_Firmicutes c_Clostridia o_Clostridiales f_Clostridiaceae g_Clostridium	76.65512
k_Bacteria p_Firmicutes c_Clostridia o_Thermoanaerobacterales f_Thermodesulfobiaceae g_Coprothermobacter	20.75226
k_Bacteria p_Firmicutes c_Clostridia o_Thermoanaerobacterales f_Thermodesulfobiaceae g_Thermodesulfobiaceae g_Thermodesulfobiaceae_unclassified	2.27984
k_Archaea p_Euryarchaeota c_Methanobacteria o_Methanobacteriales f_Methanobacteriaceae g_Methanothermobacter	0.26989
k_Bacteria p_Proteobacteria c_Gammaproteobacteria o_Enterobacteriales f_Enterobacteriaceae g_Escherichia	0.04289
k_Bacteria p_Firmicutes c_Clostridia o_Clostridiales f_Clostridiaceae g_Clostridium s_Clostridium_thermocellum	76.65512
k_Bacteria p_Firmicutes c_Clostridia o_Thermoanaerobacterales f_Thermodesulfobiaceae g_Coprothermobacter s_Coprothermobacter_proteolyticus	20.75226
k_Archaea p_Euryarchaeota c_Methanobacteria o_Methanobacteriales f_Methanobacteriaceae g_Methanothermobacter s_Methanothermobacter_thermautotrophicus	0.26989
k_Bacteria p_Proteobacteria c_Gammaproteobacteria o_Enterobacteriales f_Enterobacteriaceae g_Escherichia s_Escherichia_unclassified	0.04289
k_Bacteria p_Firmicutes c_Clostridia o_Clostridiales f_Clostridiaceae g_Clostridium s_Clostridium_thermocellum t_Clostridium_thermocellum_unclassified	76.65512
k_Bacteria p_Firmicutes c_Clostridia o_Thermoanaerobacterales f_Thermodesulfobiaceae g_Coprothermobacter s_Coprothermobacter_proteolyticus t_GCF_000020945	20.75226
k_Archaea p_Euryarchaeota c_Methanobacteria o_Methanobacteriales f_Methanobacteriaceae g_Methanothermobacter s_Methanothermobacter_thermautotrophicus t_GCF_000008645	0.26989

**Figure 4. MetaPhlan2 community profile.** A tabular representation of MetaPhlan2 community profile displaying the different levels of taxonomic classification and its relative abundance at that level.

The *HUMAN2* suite of tools were used to extract functional information along with their relative abundance (RPK). *HUMAN2* provides 3 different outputs- Gene family and their abundance (*Extended data*: Supplementary Figure S2a<sup>49</sup>), pathways and their abundance (*Extended data*: Supplementary Figure S2b<sup>49</sup>) and pathway and their coverage (*Extended data*: Supplementary Figure S2c<sup>49</sup>).

In this workflow, the UniRef50 database was used to classify the gene family, but the users also have a choice to use UniRef90. Gene family abundance at the community level is stratified to show the contributions from known and unknown species. The gene family output shows total abundance value which is the sum total of individual species abundance values (reported as RPK values). Additionally, the tabular output also enlists the contribution of individual species to the gene family abundance (*Extended data*: Supplementary Figure S3<sup>49</sup>). While there are some applications, e.g., strain profiling, where RPK units are superior to depth-normalized units, most of the time we need to renormalize our samples prior to downstream analysis. The gene families can be a long list of IDs and going through the gene families one by one to identify the interesting ones can be cumbersome and error prone. To help construct “the bigger picture”, we could identify and use categories of genes using the gene families. Gene Ontology (GO) analysis is widely used to reduce complexity and highlight biological processes in genome-wide expression studies. There is a dedicated tool called Group abundances of UniRef50 gene families obtained to gene ontology (GO) slim Terms, which groups and converts UniRef50

gene family abundances generated with *HUMAN2* into GO slim terms (*Extended data*: Supplementary Figure S4<sup>49</sup>) as the name suggests.

The functional and taxonomic annotations from MetaPhlan2 and *HUMAN2* are further normalized and combined to create a single tabular output.

### Use cases

Here we provide a trimmed version of the biogas reactor dataset to demonstrate the use of the ASaiM-MT workflow in the tutorial available in the GTN.

*Link*: <https://training.galaxyproject.org/training-material/topics/metagenomics/tutorials/metatranscriptomics/tutorial.html>

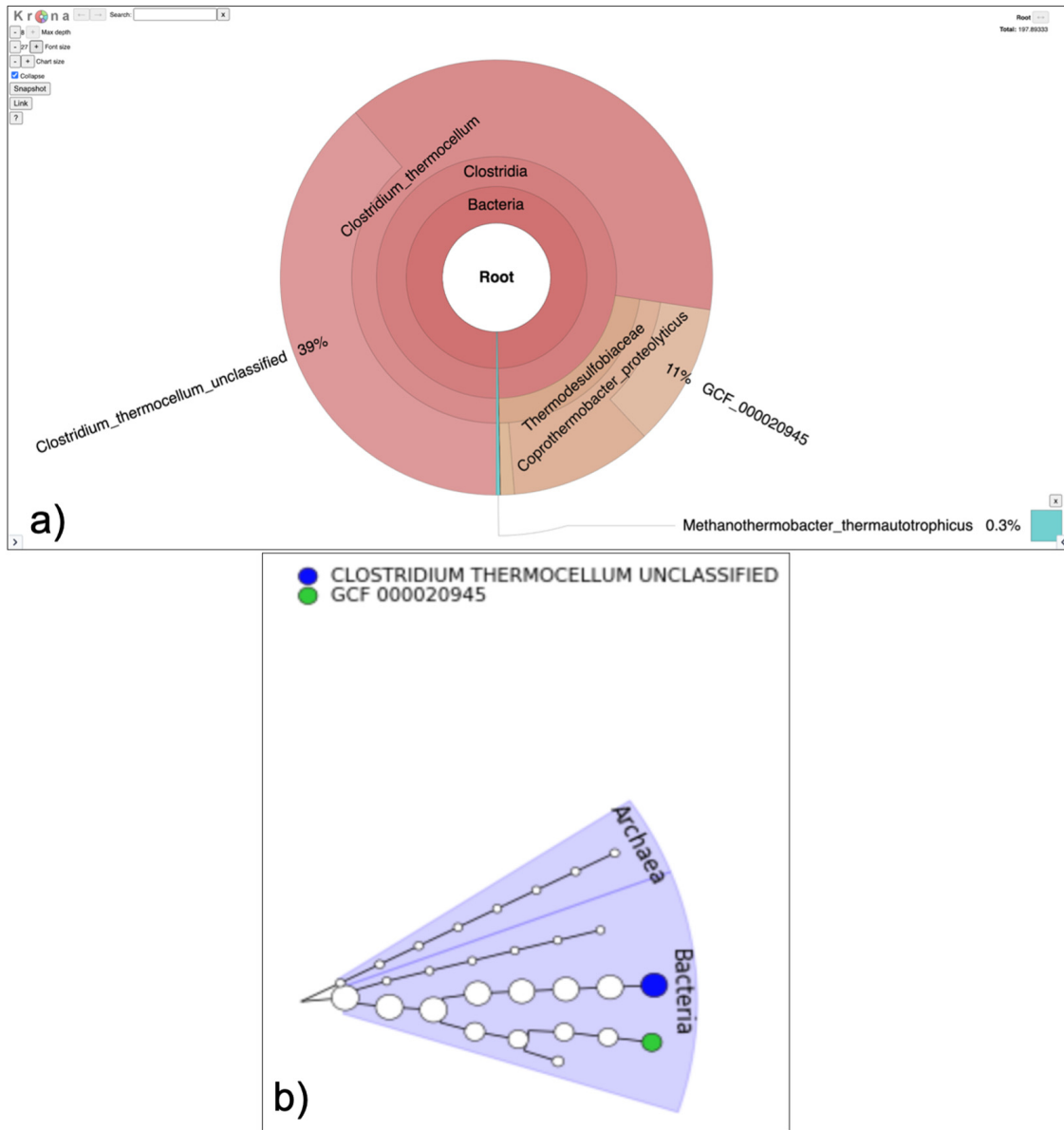
*Trimmed input*: <https://zenodo.org/record/3362849>

*Workflow*: <https://training.galaxyproject.org/training-material/topics/metagenomics/tutorials/metatranscriptomics/workflows/>

### Discussion

The ASaiM-MT workflow is made available in the Galaxy platform, enabling accessibility, shareability, and flexibility for customization. There a few tools for preprocessing are made available in Galaxy such as *Trimmomatic*, *PRINSEQ*, *TrimGalore*, etc. that can be used alternatively or in addition to the existing tools, however, they haven’t been tested rigorously for metatranscriptomics data. The ASaiM-MT workflow (as mentioned





**Figure 5. Visualization of Taxonomy output.** **a)** The Krona output provides interactive representation of the community in the sample. In this case *Coprothermobacter* and *Clostridium* were the most abundant genera. **b)** The GraPhlAn outputs circular phylogenetic trees showing that Archaea and bacteria are present in the sample.

in the methods section) was tested to ensure that the workflow works on metatranscriptomics datasets. For details about default parameters used, we recommend visiting the metatranscriptomic tutorial available on the GTN material<sup>43</sup>.

For example, in ASaiM-MT, we map the UniRef 50 values to GO terms, but they can be also mapped to the MetaCyc reactions<sup>50</sup>, KEGG Orthogroups (KOs)<sup>51</sup>, Pfam domains<sup>52</sup>, EC categories<sup>53</sup> and EggNOG (including COGs)<sup>54</sup> using the HUMAnN2 regroup tool. A current limitation of the ASaiM-MT

workflow is that it can process only one paired-end or single-end data at a time. In order to generate an input for statistical analysis, we have developed an additional post-processing workflow and Galaxy tool called MT2MQ (<https://github.com/galaxyproteomics/tools-galaxy/tree/master/tools/mt2mq>), which integrates results from multiple outputs from the ASaiM-MT workflow. The MT2MQ workflow combines the gene abundance output from multiple samples or conditions, normalizes the values and makes it compatible with statistical tools such as *metaQuantome* tool<sup>42,55</sup>, which can be used for visualizing

and interpreting results. Furthermore, we are in the process of developing tools that can help perform multi-omics studies by integrating results from the ASaiM-MT workflow to our existing metaproteomics workflows.

## Conclusion

ASaiM-MT workflow is a robust and extensible Galaxy workflow which is now optimized and tested for metatranscriptomics data. The workflow consists of tested open-source tools in the area of RNA sequence analysis, such as *SortMeRNA*, *MetaPhlan2* and *HUMAnN2*. ASaiM-MT in Galaxy offers users a high-level control over their data and provides different analysis options. The GTN offers documentation and resources necessary for new users to gain mastery in using this workflow and associated tools for data analysis for their research projects. ASaiM-MT allows users to not only understand the taxonomy but also the functional composition and pathways expressed by the microbiome present in diverse microbial communities of interest.

## Data availability

### Underlying data

Zenodo: Training Data for “Metatranscriptomics analysis using microbiome RNASeq data”, <http://doi.org/10.5281/zenodo.3362849><sup>27</sup>.

### Extended data

Zenodo: Supplementary for ASaiM-MT: A validated and optimized ASaiM workflow for metatranscriptomics analysis within Galaxy framework, <http://doi.org/10.5281/zenodo.4341391><sup>49</sup>.

This project contains the following extended data:

- Supplementary Figure S1: MetaPhlan2 Genus-Level Abundance
- Supplementary Figure S2a: HUMAnN2 Gene Family Abundance

- Supplementary Figure S2b: HUMAnN2 Pathway Abundance
- Supplementary Figure S2c: HUMAnN2 Pathway Coverage
- Supplementary Figure S3: Uniref50 Gene Family output with abundance
- Supplementary Figure S4: Conversion of Uniref 50 values to GO terms

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) (CC-BY 4.0).

## Software availability

Software available from: <https://metagenomics.usegalaxy.eu/>

Source code available from: <https://github.com/ASaiM/framework>

Archived source code at time of publication: <http://doi.org/10.5281/zenodo.4455627><sup>28</sup>.

License: Apache 2 License

Docker: <https://quay.io/repository/bebatut/asaim-framework> (command: `docker pull quay.io/bebatut/asaim-framework`).

## Acknowledgements

We would like to thank European Galaxy team for the help in the support during Galaxy implementation. We would also like to thank Björn A. Grüning (University of Freiburg, Germany) for helping us during the implementation of the workflow in the European Galaxy platform.

## References

1. Kara EL, Hanson PC, Hu YH, *et al.*: **A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA.** *ISME J.* 2013; **7**(3): 680–4. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Taş N, Prestat E, McFarland JW, *et al.*: **Impact of fire on active layer and permafrost microbial communities and metagenomes in an upland Alaskan boreal forest.** *ISME J.* 2014; **8**(9): 1904–19. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Mason OU, Hazen TC, Borglin S, *et al.*: **Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill.** *ISME J.* 2012; **6**(9): 1715–27. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Mohajeri MH, Brummer RJM, Rastall RA, *et al.*: **The role of the microbiome for human health: from basic science to clinical applications.** *Eur J Nutr.* 2018; **57**(Suppl 1): 1–14. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Bolyen E, Rideout JR, Dillon MR, *et al.*: **Author Correction: Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2.** (Nature Biotechnology, (2019), 37, 8, (852–857)). *Nat Biotechnol.* 2019; **37**(9): 1091. [PubMed Abstract](#) | [Publisher Full Text](#)
6. Schloss PD, Westcott SL, Ryabin T, *et al.*: **Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities.** *Appl Environ Microbiol.* 2009; **75**(23): 7537–41. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Franzosa EA, McIver LJ, Rahnnavard G, *et al.*: **Species-level functional profiling of metagenomes and metatranscriptomes.** *Nat Methods.* 2018; **15**(11): 962–968. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Turnbaugh PJ, Gordon JI: **An Invitation to the marriage of metagenomics and metabolomics.** *Cell.* 2008; **134**(5): 708–13. [PubMed Abstract](#) | [Publisher Full Text](#)
9. Bashiardes S, Zilberman-Schapira G, Elinav E: **Use of metatranscriptomics in microbiome research.** *Bioinform Biol Insights.* 2016; **10**: 19–25. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Wilmes P, Bond PL: **Metaproteomics: Studying functional gene expression in microbial ecosystems.** *Trends Microbiol.* 2006; **14**(2): 92–7. [PubMed Abstract](#) | [Publisher Full Text](#)
11. Shakya M, Lo CC, Chain PSG: **Advances and challenges in metatranscriptomic analysis.** *Front Genet.* Frontiers Media S.A. 2019; **10**: 904. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Douglas GM, Beiko RG, Langille MGI: **Predicting the Functional Potential**

- of the Microbiome from Marker Genes Using PICRUSt. *Methods Mol Biol.* Humana Press Inc. 2018; **1849**: 169–177.  
[PubMed Abstract](#) | [Publisher Full Text](#)
13. Babraham Bioinformatics: FastQC A Quality Control tool for High Throughput Sequence Data. [cited 2021 Mar 20].  
[Reference Source](#)
  14. VBC | Victorian Bioinformatics Consortium. [cited 2021 Mar 24].  
[Reference Source](#)
  15. bmtagger — bioconda-recipes 1.0.0 documentation. [cited 2021 Mar 24].  
[Reference Source](#)
  16. Grabherr MG, Haas BJ, Yassour M, et al.: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 2011; **29**(7): 644–52.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  17. Namiki T, Hachiya T, Tanaka H, et al.: MetaVelvet: An extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Res.* 2012; **40**(20): e155.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  18. Freitas TAK, Li PE, Scholz MB, et al.: Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res.* 2015; **43**(10): e69.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  19. Leung HCM, Yiu SM, Parkinson J, et al.: IDBA-MT: *de novo* assembler for metatranscriptomic data generated from next-generation sequencing technology. *J Comput Biol.* 2013; **20**(7): 540–50.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  20. Wood DE, Salzberg SL: Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 2014; **15**(3): R46.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  21. Truong DT, Franzosa EA, Tickle TL, et al.: MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat Methods.* Nature Publishing Group; 2015; **12**(10): 902–3.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  22. Rho M, Tang H, Ye Y: FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Res.* 2010; **38**(20): e191.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  23. Buchfink B, Xie C, Huson DH: Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* Nature Publishing Group; 2015; **12**(1): 59–60.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  24. Kanehisa M, Furumichi M, Tanabe M, et al.: KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017; **45**(D1): D353–61.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  25. O’Leary NA, Wright MW, Brister JR, et al.: Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016; **44**(D1): D733–45.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  26. The UniProt Consortium: UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* 2017; **45**(D1): D158–69.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  27. Robinson MD, McCarthy DJ, Smyth GK: edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010; **26**(1): 139–40.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  28. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014; **15**(12): 550.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  29. Ritchie ME, Phipson B, Wu D, et al.: Linear Models for Microarray and RNA-Seq Data. *Nucleic Acids Res.* 2015; **43**(7): e47.
  30. Martinez X, Pozuelo M, Pascal V, et al.: MetaTrans: An open-source pipeline for metatranscriptomics. *Sci Rep.* 2016; **6**(1): 26447.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  31. Ni Y, Li J, Panagiotou G: COMAN: A web server for comprehensive metatranscriptomics analysis. *BMC Genomics.* 2016; **17**(1): 622.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  32. Kim J, Kim MS, Koh AY, et al.: FMAP: Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies. *BMC Bioinformatics.* 2016; **17**(1): 420.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  33. Westreich ST, Treiber ML, Mills DA, et al.: SAMSA2: A standalone metatranscriptome analysis pipeline. *BMC Bioinformatics.* 2018; **19**(1): 175.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  34. Batut B, Gravouil K, DeFois C, et al.: ASaiM: A Galaxy-based framework to analyze microbiota data. *GigaScience.* 2018; **7**(6): gij057.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  35. Tamames J, Puente-Sánchez F: SqueezeMeta, A Highly Portable, Fully Automatic Metagenomic Analysis Pipeline. *Front Microbiol.* 2018; **9**: 3349.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  36. Narayanasamy S, Jarosz Y, Muller EEL, et al.: IMP: A pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* 2016; **17**(1): 260.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  37. iquasere/MOSCA: Meta-Omics Software for Community Analysis. [cited 2021 Mar 24].  
[Reference Source](#)
  38. ASaiM: an environment to analyze intestinal microbiota data. ASaiM 0.1 documentation. [cited 2020 Dec 17].  
[Reference Source](#)
  39. Giardine B, Riemer C, Hardison RC, et al.: Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.* 2005 [cited 2020 Dec 7]; **15**(10): 1451–5.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  40. Batut B, Hiltmann S, Bagnacani A, et al.: Community-Driven Data Analysis Training for Biology. *Cell Syst.* 2018 [cited 2020 Oct 29]; **6**(6): 752–758.e1.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  41. Afgan E, Baker D, Batut B, et al.: The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 2018; [cited 2020 Dec 17]; **46**(W1): W537–44.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  42. Easterly CW, Sajulga R, Mehta S, et al.: metaQuantome: An Integrated, Quantitative Metaproteomics Approach Reveals Connections Between Taxonomy and Protein Function in Complex Microbiomes. *Mol Cell Proteomics.* 2019; **18**(8 suppl 1): S82–91.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  43. Metatranscriptomics analysis using microbiome RNA-seq data. [cited 2020 Oct 29].  
[Reference Source](#)
  44. Martin M: Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 2011 [cited 2020 Oct 29]; **17**(1): 10.  
[Publisher Full Text](#)
  45. Kopylova E, Noé L, Touzet H: SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics.* 2012 [cited 2020 Oct 29]; **28**(24): 3211–7.  
[PubMed Abstract](#) | [Publisher Full Text](#)
  46. Ondov BD, Bergman NH, Phillippy AM: Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics.* 2011; **12**: 385.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  47. Asnicar F, Weingart G, Tickle TL, et al.: Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ.* 2015 [cited 2020 Dec 17]; **2015**(6): e1029.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  48. Kunath BJ, Delogu F, Naas AE, et al.: From proteins to polysaccharides: lifestyle and genetic evolution of *Coprothermobacter proteolyticus*. *ISME J.* 2019 [cited 2020 Oct 29]; **13**(3): 603–17.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  49. Mehta S: Supplementary for ASaiM-MT: A validated and optimized ASaiM workflow for metatranscriptomics analysis within Galaxy framework. 2020.  
<http://www.doi.org/10.5281/zenodo.4341391>
  50. Caspi R, Billington R, Fulcher CA, et al.: The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* 2018 [cited 2020 Dec 17]; **46**(D1): D633–9.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  51. Kanehisa M, Sato Y, Kawashima M, et al.: KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2016 [cited 2020 Dec 17]; **44**(D1): D457–62.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  52. Finn RD, Bateman A, Clements J, et al.: Pfam: The protein families database. *Nucleic Acids Res.* 2014 [cited 2020 Dec 17]; **41**(Database issue): D222–30.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  53. Dönertaş HM, Cuesta SM, Rahman SA, et al.: Characterising complex enzyme reaction data. *PLoS One.* 2016 [cited 2020 Dec 17]; **11**(2): e0147952.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  54. Huerta-Cepas J, Szklarczyk D, Heller D, et al.: EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019 [cited 2020 Dec 17]; **47**(D1): D309–14.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  55. Mehta S, Kumar P, Crane M, et al.: Updates on metaQuantome Software for Quantitative Metaproteomics. *J Proteome Res.* 2021 [cited 2021 Mar 24]; **20**(4): 2130–2137.  
[PubMed Abstract](#) | [Publisher Full Text](#)

# Open Peer Review

Current Peer Review Status:  

---

## Version 2

Reviewer Report 23 August 2021

<https://doi.org/10.5256/f1000research.55709.r90589>

© 2021 Cho W. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Won Kyong Cho** 

Research Institute of Agriculture and Life Sciences, College of Agriculture and Life Sciences, Seoul National University, Seoul, South Korea

In this study, the authors provided a new workflow referred to as ASaiM for metatranscriptomics. The introduction was easy to understand the background of the study. Results and methods were written in detail. Figures and tables are very good to understand the workflow of the ASaiM-MT. In addition, the authors provided a data set for the demonstration. Overall, the quality of the manuscript is high enough for publication as it is. I hope the newly developed platform can be usefully used by many biologists working with microbiomes.

**Is the rationale for developing the new method (or application) clearly explained?**

Yes

**Is the description of the method technically sound?**

Yes

**Are sufficient details provided to allow replication of the method development and its use by others?**

Yes

**If any results are presented, are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions about the method and its performance adequately supported by the findings presented in the article?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Metagenomics, metatranscriptomics, viromes

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 27 April 2021

<https://doi.org/10.5256/f1000research.55709.r83478>

© 2021 Simopoulos C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Caitlin Simopoulos** 

Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, ON, Canada

The authors have revised the manuscript to address my concerns, and I am satisfied with the revisions. I believe the re-formatting of the Methods section helps with comprehension.

I enjoyed the figures describing the method and feel like they add to the manuscript. Inclusion of the Galaxy workflow will also help users analyze their own data using this tool. I also appreciated the links to a separate tutorial and the Gitter community. This approach will help guide and welcome new users to the tool.

Some minor suggestions are to correct two typos:

1. In the intro: "Functional Annotation (*HUMAN2*<sup>7</sup>), Annotation of assembled contigs are subjected to gene finding programs such as *FragGeneScan*<sup>2</sup>" where I believe the "," should be a ".".
2. In Figure 3 legend, I believe "metaproteomics" or "metagenomics" should be "metatranscriptomics". This may be more important to correct.

**Is the rationale for developing the new method (or application) clearly explained?**

Yes

**Is the description of the method technically sound?**

Yes

**Are sufficient details provided to allow replication of the method development and its use by others?**

Yes

**If any results are presented, are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions about the method and its performance adequately supported by the findings presented in the article?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Computational biology, bioinformatics, RNA-sequencing data pipeline development and analysis, microbiomes

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

### Version 1

Reviewer Report 18 February 2021

<https://doi.org/10.5256/f1000research.31653.r79463>

© 2021 Simopoulos C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Caitlin Simopoulos** 

Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, ON, Canada

The authors describe an optimized workflow for metatranscriptomic microbial community data named **ASaiM-MT**. **ASaiM-MT** is an extension of **ASaiM**, a workflow mostly focused on metagenomic data available via Galaxy. Using helpful diagrams, the authors describe how **ASaiM-MT** differs from the original **ASaiM** workflow, and detail which tools from the **Galaxy** platform are used. Finally, the authors use data from a biogas reactor study the pipeline and demonstrate results that can be expected using **ASaiM-MT**. I particularly enjoyed the fact that the article is accompanied by written training material for the workflow. I believe that tutorials and user-friendly tools expand the audience of traditionally “bioinformatician-only” tools. However, I have a few suggestions for the authors to strengthen their presentation of the metatranscriptomic workflow before publication.

#### Major comments:

1. In your discussion, you highlighted a major limitation of **ASaiM-MT**: the fact that it can only handle a single sample and does not complete comparative analysis. However, you also mention that you have developed post-processing tools for this reason. It might be good to highlight that **ASaiM-MT** is essential (in particular for data pre-processing and identification of functional and taxonomic information) to be completed before statistical analysis. In

addition, is it possible to process data in ASaiM-MT in batches? This might also lessen the perceived limitation of the tool.

2. Sometimes it is not completely clear why **ASaiM-MT** is needed if **ASaiM** was developed for both meta-genomics and -transcriptomics. It is important to be specific and highlight the rationale for developing the **ASaiM-MT** workflow. An example: *“For performing this action, the original ASaiM Shotgun workflow used the FASTQ-joiner to join the reads. However, in the ASaiM-MT version, we use the FASTQ interlacer. FASTQ interlacer joins the forward (/1) and the reverse reads (/2) using the sequence identifiers; sequences without designation will be named as single reads. The reason ASaiM-MT uses FASTQ-interlacer rather than FASTQ-joiner is because the joiner tool combines the forward and reverse read sequence together while the interlacer puts the forward and reverse read sequences in the same file while retaining the entity of each read along with an additional file with unpaired sequences.”* Was this change made because **FASTQ interlacer** is more appropriate for RNA sequencing, or was it a change to improve on the original **ASaiM** workflow?
3. The abstract and introduction seem to emphasize the importance of studying the human microbiome and its connections to health and well-being, however, this manuscript uses a microbial community obtained from a biogas reactor. This is a very interesting community that warrants further research, but the introduction does not match the sample of interest. I suggest updating the introduction to focus less on the human microbiome and emphasize the importance of studying microbial communities in general. Alternatively, if you'd like to focus on human microbiomes, the example data used should reflect the human focus.
4. In your discussion, you say: *“There are a few tools that can be used alternatively or in addition to the existing tools.”* Does this mean that there are other tools like **ASaiM-MT** that can be used? Your introduction mainly listed tools that complete specific tasks and not an end to end workflow. If there are other workflow tools, a comparison should be highlighted in the discussion or introduction. What makes **ASaiM-MT** unique compared to other tools/workflows? What are its strengths? Is it more user-friendly than other tools? Expanding on this will help strengthen your conclusions about the tool itself.

#### Minor comments:

1. There are sections that can read like a “grocery list” of software names (eg. second paragraph of the introduction). I understand why you are listing them, but it could be useful to emphasize that each of these tools performs a single task and need to be put together into a workflow for a complete experiment. These tools are also often missing citations.
2. It can be difficult to understand which words are actually software names. Is it possible to type them in a monospace font? Or to bold the names?
3. At times the “Methods” section can get confusing, particularly when MT is being compared to the original **ASaiM**. My suggestion is to include numbers or roman numerals in the “in between” steps as described in Figure 2. That way you can also reference them in the text. For example, in “a) Preprocessing” there would be i. Input files, ii. Quality Control, iii. Adapter Trimming... etc.. It might also be useful to include these headers (and numbers/letters) in your Methods section to let readers follow along.

4. "PMID 30298254" is used instead of a citation

**Is the rationale for developing the new method (or application) clearly explained?**

Partly

**Is the description of the method technically sound?**

Partly

**Are sufficient details provided to allow replication of the method development and its use by others?**

Yes

**If any results are presented, are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions about the method and its performance adequately supported by the findings presented in the article?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Computational biology, bioinformatics, RNA-sequencing data pipeline development and analysis, microbiomes

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 25 Mar 2021

**Subina Mehta**, University of Minnesota, USA

We greatly appreciate the reviewer's comments and suggestions. We have incorporated the required changes in the updated version of the manuscript.

**Major comments:**

**1. In your discussion, you highlighted a major limitation of ASaiM-MT: the fact that it can only handle a single sample and does not complete comparative analysis. However, you also mention that you have developed post-processing tools for this reason. It might be good to highlight that ASaiM-MT is essential (in particular for data pre-processing and identification of functional and taxonomic information) to be completed before statistical analysis. In addition, is it possible to process data in ASaiM-MT in batches? This might also lessen the perceived limitation of the tool.**

**A:** *We would like to thank the reviewer for the comment. We have edited the abstract and*



*mentioned that ASaiM-MT is essential for preprocessing and identification of taxonomy and functional information before performing comparative analysis. Also, Galaxy has a function of providing datasets as a collection rather than a single input which can be used to handle multiple datasets. In the Galaxy platform, we have other software tools such as MT2MQ and metaQuantome, which can perform comparative statistical analysis.*

**2. Sometimes it is not completely clear why ASaiM-MT is needed if ASaiM was developed for both meta-genomics and -transcriptomics. It is important to be specific and highlight the rationale for developing the ASaiM-MT workflow. An example: "For performing this action, the original ASaiM Shotgun workflow used the FASTQ-joiner to join the reads. However, in the ASaiM-MT version, we use the FASTQ interlacer. FASTQ interlacer joins the forward (/1) and the reverse reads (/2) using the sequence identifiers; sequences without designation will be named as single reads. The reason ASaiM-MT uses FASTQ-interlacer rather than FASTQ-joiner is because the joiner tool combines the forward and reverse read sequence together while the interlacer puts the forward and reverse read sequences in the same file while retaining the entity of each read along with an additional file with unpaired sequences." Was this change made because FASTQ interlacer is more appropriate for RNA sequencing, or was it a change to improve on the original ASaiM workflow?**

*A: Thank you for this comment. We replaced the FASTQ-joiner tool with the FASTQ interlacer to improve on the original workflow. The FASTQ interlacer tool maintains the integrity of reads by maintaining the forward and the reverse sequence identifiers, as compared to joining them into a single read file, as is done by FASTQ-joiner tool.*

**3. The abstract and introduction seem to emphasize the importance of studying the human microbiome and its connections to health and well-being, however, this manuscript uses a microbial community obtained from a biogas reactor. This is a very interesting community that warrants further research, but the introduction does not match the sample of interest. I suggest updating the introduction to focus less on the human microbiome and emphasize the importance of studying microbial communities in general. Alternatively, if you'd like to focus on human microbiomes, the example data used should reflect the human focus.**

*A: We thank the reviewer for the comment and have edited the abstract and introduction to highlight the importance of microbiome research in ecology as well as clinical research.*

**4. In your discussion, you say: "There are a few tools that can be used alternatively or in addition to the existing tools." Does this mean that there are other tools like ASaiM-MT that can be used? Your introduction mainly listed tools that complete specific tasks and not an end to end workflow. If there are other workflow tools, a comparison should be highlighted in the discussion or introduction. What makes ASaiM-MT unique compared to other tools/workflows? What are its strengths? Is it more user-friendly than other tools? Expanding on this will help strengthen your conclusions about the tool itself.**

*A: We thank the reviewer for the comment. The tools mentioned in the discussion offer an*

*alternative option to the tools that we have used in the ASaiM-MT workflow, especially if the users have a preference. However, although we consider these tools to be appropriate to metatranscriptomics research, we have not tested these tools.*

*ASaiM-MT workflow, due to its availability in Galaxy, offers an user-friendly option to the existing command line tools. The ASaiM-MT workflow has been tested with different datasets to ensure its compatibility. There is a systematic documentation available for the usage of the workflow in the Galaxy Training Network (GTN). Users can also ask questions to developers and users via the Gitter channel, if needed.*

**Minor comments:**

**1. There are sections that can read like a “grocery list” of software names (eg. second paragraph of the introduction). I understand why you are listing them, but it could be useful to emphasize that each of these tools performs a single task and need to be put together into a workflow for a complete experiment. These tools are also often missing citations.**

*A: Thanks to the reviewer for the comment. The tools listed are alternatives to the existing tools in the workflow and have not been incorporated as a workflow. We have added citations to these tools.*

**2. It can be difficult to understand which words are actually software names. Is it possible to type them in a monospace font? Or to bold the names?**

*A: Thanks to the reviewer for this comment. I have formatted the tool names by making them italic.*

**3. At times the “Methods” section can get confusing, particularly when MT is being compared to the original ASaiM. My suggestion is to include numbers or roman numerals in the “in between” steps as described in Figure 2. That way you can also reference them in the text. For example, in “a) Preprocessing” there would be i. Input files, ii. Quality Control, iii. Adapter Trimming... etc.. It might also be useful to include these headers (and numbers/letters) in your Methods section to let readers follow along.**

*A: Thank you for the comment. I have reformatted the method section according to Figure 2.*

**4. “PMID 30298254” is used instead of a citation**

*A: Thank you reviewer for pointing this out. I have made the required change.*

**Competing Interests:** No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**