

SOFTWARE

Open Access



Integrative meta-omics in Galaxy and beyond

Valerie C. Schimpl¹, Francesco Delogu¹, Praveen Kumar², Benoit Kunath¹, Bérénice Batut³, Subina Mehta², James E. Johnson⁴, Björn Grüning³, Phillip B. Pope^{1,5}, Pratik D. Jagtap², Timothy J. Griffin² and Magnus Ø. Arntzen^{1*}

Abstract

Background ‘Omics methods have empowered scientists to tackle the complexity of microbial communities on a scale not attainable before. Individually, omics analyses can provide great insight; while combined as “meta-omics”, they enhance the understanding of which organisms occupy specific metabolic niches, how they interact, and how they utilize environmental nutrients. Here we present three integrative meta-omics workflows, developed in Galaxy, for enhanced analysis and integration of metagenomics, metatranscriptomics, and metaproteomics, combined with our newly developed web-application, ViMO (Visualizer for Meta-Omics) to analyse metabolisms in complex microbial communities.

Results In this study, we applied the workflows on a highly efficient cellulose-degrading minimal consortium enriched from a biogas reactor to analyse the key roles of uncultured microorganisms in complex biomass degradation processes. Metagenomic analysis recovered metagenome-assembled genomes (MAGs) for several constituent populations including *Hungateiclostridium thermocellum*, *Thermoclostridium stercorarium* and multiple heterogenic strains affiliated to *Coprothermobacter proteolyticus*. The metagenomics workflow was developed as two modules, one standard, and one optimized for improving the MAG quality in complex samples by implementing a combination of single- and co-assembly, and dereplication after binning. The exploration of the active pathways within the recovered MAGs can be visualized in ViMO, which also provides an overview of the MAG taxonomy and quality (contamination and completeness), and information about carbohydrate-active enzymes (CAZymes), as well as KEGG annotations and pathways, with counts and abundances at both mRNA and protein level. To achieve this, the metatranscriptomic reads and metaproteomic mass-spectrometry spectra are mapped onto predicted genes from the metagenome to analyse the functional potential of MAGs, as well as the actual expressed proteins and functions of the microbiome, all visualized in ViMO.

Conclusion Our three workflows for integrative meta-omics in combination with ViMO presents a progression in the analysis of ‘omics data, particularly within Galaxy, but also beyond. The optimized metagenomics workflow allows for detailed reconstruction of microbial community consisting of MAGs with high quality, and thus improves analyses of the metabolism of the microbiome, using the metatranscriptomics and metaproteomics workflows.

Keywords Integrated meta-omics, Metagenomics, Metatranscriptomics, Metaproteomics, Galaxy, Bioinformatics

*Correspondence:
Magnus Ø. Arntzen
magnus.arntzen@nmbu.no
Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Introduction

Microbial communities have a tremendous impact on Earth's ecosystems. An example is the marine microbiome, which is responsible for >50% of the produced oxygen on the planet [1]. The microorganisms historically promoted the adjustment from freshwater to the terrestrial environment for plants [2] and bacteria still today regulate the growth and development of the terrestrial flora by symbiosis, for example promoting growth by nitrogen fixation or plant hormone production [3]—dynamic and highly adaptable processes that are influential to microbial communities and their hosts alike [4, 5]. Similarly, in humans, microbial communities may affect the toxicity of drugs, modulate disease progression, and promote health. It is of great importance to increase our understanding of such microbiomes, their composition and interplay, as well as factors for perturbation, stability, and development [6, 7]. Ideally, such understanding may spark the development of new personalized medical treatments for improving life quality and addressing the climate crisis, specifically by curbing the emissions of methane from wetlands or ruminating animals [8, 9] and nitrous oxide from agriculture [9, 10].

Meta-omics technologies, alongside environmental measurements, allow researchers to infer the complex network of a microbiome and its relations with the environment and host, offering a putative picture of their metabolism in their natural habitat [11, 12]. With metagenomics, we analyse the total DNA of the microbial community using shotgun sequencing [11, 13, 14], and this technology provides information about the potential physiological function and regulation of the genes in microbial communities [11, 15, 16]. Modern tools for read assembly allow for the retrieval of both known and novel organisms by overcoming challenges such as size and complexity of metagenomic data, as well as difficulties in accuracy and contiguity of metagenome assemblies [17]. This has resulted in larger and less fragmented assemblies and hence better quality of metagenome-assembled genomes (MAGs) [18]. Remarkably, in some samples, species-resolution can be achieved during the binning process, allowing for reconstruction of metabolic pathways for individual MAGs [19]. Further, metatranscriptomics aims to analyse the entire set of active gene transcripts in the microbial community as well as calculate their (relative) abundances and thus capture perturbation, environmental changes, and dynamics [14, 16]. Using high-throughput sequencing, transcripts of microorganisms are detected, and either analysed on their own, or preferably, mapped to the metagenomics data, including MAGs, which enables the identification and quantification of active metabolic pathways [14].

Further evidence is provided by metaproteomics, which identifies and quantifies of the entire set of proteins in the microbial community, both intra- and extra-cellular [11, 16]. Metaproteomics in combination with metagenomics allows both for targeted identification of sample-specific microorganisms, and also for the identification of proteins not present in publicly available sequence repositories such as UniProt or RefSeq. This in turn might enhance our understanding of known signalling pathways or possibly act in the discovery of new metabolic pathways [20], as well as detect the presence of active novel microbial members within the community.

Due to a rapid improvement of algorithms within the meta-omics field, analysing meta-omics data requires a constant update and evaluation of computational tools. Currently, hundreds of tools are available for the analysis of meta-omics data, and it can be challenging to select the right tool and parameters for a given dataset. Meanwhile, the popularity of user-friendly interfaces attached to compute resources with pre-installed software packages, like Anvi'o [21] for metagenomics and metatranscriptomics, iMetalab [22] for metaproteomics, and Galaxy for multi-omics [23, 24], are on the rise, particularly because they enable advanced bioinformatic analysis without the need for programming/scripting. In the Galaxy platform, various tools can be chained together in a sequential manner into a workflow and shared between developers and users for further data-based optimization and reproducibility [25]. A common workflow for metagenomics within Galaxy is ASaiM [26] with taxonomic and functional analysis of metagenomics shotgun data, which was further extended to include metatranscriptomics analysis in the ASaiM-MT workflow [27]. However, while ASaiM and ASaiM-MT offer in-depth microbial analysis, it currently does not support the analysis of MAGs or the full integration between the different omics disciplines.

In this study, we applied commonly used omics tools within the Galaxy framework to generate workflows for metagenomics (MetaG), metatranscriptomics (MetaT), and metaproteomics (MetaP). We made the workflows integrative, so that MAGs recovered in the MetaG workflow makes the reference for mapping both transcriptomic reads and proteomic mass spectra. The workflows were applied on a highly efficient cellulose-degrading minimal consortium enriched from an industrial biogas reactor in Fredrikstad, Norway to analyse the key roles of uncultured microorganisms in complex biomass degradation processes [28]. To enhance the multi-levelled data interpretation and exploration, we developed an interactive R-Shiny-based web-application, ViMO (Visualizer for Meta-Omics), where the data can be explored in more detail.

Methods

Samples

The microbial community called SEM1b studied/ utilized in this work was enriched from a thermophilic biogas reactor operated on municipal food waste (Frevvar) and manure in Fredrikstad, Norway, and has previously been described in detail, including metagenomics, metatranscriptomics and metaproteomics analysis across nine time points spanning over 43 h post inoculation [28, 29]. In brief, using an inoculate from a lab-scale reactor, we performed a serial dilution to extinction experiment to simplify and enrich the community for growth on Norwegian Spruce as carbon source at 65 °C. DNA was collected by Phenol–Chloroform extraction of 6 mL sample and a library was prepared with the TrueSeq DNA PCRfree-protocol prior to sequencing on an Illumina HiSeq3000 platform (Illumina Inc) with paired-ends (2×125 bp) [28, 29]. For metatranscriptomics analysis, mRNA was extracted in triplicates (A, B, and C) with the RNeasy mini kit (Protocol2, Qiagen, USA) followed by DNA and small RNAs removal (such as tRNA) with lithium chloride precipitation solution (ThermoFisher Scientific) according to manufacturer's recommendation. The enriched mRNA was amplified with the MessageAMP II-Bacteria Kit (Applied Biosystems, USA) and sequenced on an Illumina HiSeq3000 platform with paired-ends (2×125 bp). Proteins were extracted chemically and mechanically using FastPrep24 in triplicates and subsequently reduced, alkylated and in-gel digested with trypsin. The mass spectrometry analysis of the peptides was performed using nanoLC-MS/MS system consisting of a Dionex Ultimate 3000 UHPLC (ThermoScientific, Germany) connected to a Q-Exactive hybrid quadrupole-orbitrap mass spectrometer (ThermoScientific, Germany). For this study, we used the metagenomics data from the abovementioned SEM1b community, as well as a subset of the metatranscriptomics and metaproteomics data, including triplicates from three time points (13, 23, 38 h) after inoculation [28, 29].

Implementation, results and discussion

In this study we used common tools already present within the Galaxy ToolShed (<https://toolshed.g2.bx.psu.edu/>), as well as incorporated additional tools (dRep, CheckM, CoverM, BAT/CAT) to facilitate multi-omics analysis of microbiomes at a level not possible in Galaxy previously. The newly implemented dRep selects MAGs with the best quality in the genome set improving the pathway analysis of each MAG with functional annotation tools and the recently added KOFamScan annotations. The quality for these MAGs in the workflow can be assessed with CheckM and their genome mapped back to the metagenome raw files using CoverM. Tools for

meta-omics were then chained to generate three separate workflows for (1) metagenomic assembly, binning, and functional annotation (MetaG), (2) metatranscriptomics (MetaT), and (3) metaproteomics (MetaP). Although separate, the workflows are designed to be integrative so that the MAGs recovered from MetaG make the foundation for mapping both the transcriptomic reads and the proteomic spectra onto their predicted genes. The tools included in the three pipelines are listed in Table 1.

Workflow for metagenomics and functional annotation (MetaG)

The MetaG workflow provides all the processing steps and parameters to analyze FASTQ files containing the shotgun metagenomics raw data. This multi-step workflow contains data cleaning/trimming, assembly of reads into contigs, binning of contigs into MAGs, as well as taxonomic analysis of the MAGs and functional annotation of all gene products encoded in the MAGs (Table 1).

The MetaG workflow accepts Illumina paired-end FASTQ sequence files (forward and reverse reads) as *input files* (Fig. 1.1). The FASTQ-files can be uploaded to Galaxy via the web interface or using FTP and should be organized as a collection of paired datasets. As *quality control* (Fig. 1.2), we use FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) with a Phred threshold of 20 to be aware of occasional nucleotide reading errors or overrepresentation of features, like primers or sequencing adapters. The quality control is followed by a data preprocessing steps, including automatic detection and *trimming* (Fig. 1.3) of adapter sequences by Trim Galore! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The collection of trimmed paired reads is then split into a list of forward and reversed reads for *co-assembly*. The metagenomic reads are further assembled (Fig. 1.4) into contigs with k-mer sizes of 21, 29, 39, 59, 79, 99, 119, and 141 using MEGAHIT [30]. The quality for assemblies is assessed using metaQUAST [31] (Fig. 1.5) in meta-mode. The contigs are *binned* into MAGs (Fig. 1.6) by MaxBin2 [19] based on an expectation–maximization algorithm with a minimum contig length of 1000. Completeness, contamination, and strain heterogeneity are analyzed using CheckM [33] and read coverage using CoverM (<https://github.com/wwood/CoverM>) (Fig. 1.7). Further, *taxonomic annotation* for the MAGs is done with the Bin Annotation Tool [34] (range: 10, fraction: 0.5) (Fig. 1.8). The genomes are individually subjected to *gene prediction* (Fig. 1.9) using the software FragGeneScan [35], which outputs FASTA-files of both nucleotide and protein sequences.

The putative proteins are then *functionally annotated* (Fig. 1.10) using InterProScan [39] with the databases TIGERFAM [45], HAMAP [46], PfamA [47], and Gene

Table 1 List of software in the MetaG, MetaT, MetaP workflows

Workflow	Software version	References/webpage
<i>MetaG</i>		
Trimming	Trim Galore! (Galaxy version 0.6.7 + galaxy0)	(https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)
Quality control	FastQC (Galaxy version 0.73 + galaxy0)	(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
Assembly	MEGAHIT (Galaxy version 1.1.3.5)	[30]
Assembly quality	metaQUAST (Galaxy version 5.2.0 + galaxy0)	[31]
Binning	MaxBin2 (Galaxy version 2.2.7 + galaxy3)	[19]
Dereplication*	dRep (Galaxy version 3.2.2 + galaxy0)	[32]
Genome quality assessment	CheckM lineage_wf (Galaxy Version 1.2.0 + galaxy0)	[33]
Read coverage	CoverM-GENOME (Galaxy Version 0.2.1 + galaxy0)	(https://github.com/wwood/CoverM)
Read coverage	CoverM-CONTIG (Galaxy Version 0.2.1 + galaxy0)	(https://github.com/wwood/CoverM)
Bin annotation	CAT bins (Galaxy version 5.0.3.0)	[34]
Gene prediction	FragGeneScan (Galaxy version 1.30.0)	[35]
CAZyme annotation	Hmmscan (Galaxy version 0.1.0) with dbCAN-HMMdb-V10	[36–38]
KOfam annotation	KofamScan (Galaxy version 1.3.0 + galaxy1)	
Functional annotation	Interproscan (Galaxy version 5.0.0)	[39]
<i>MetaT</i>		
Trimming	Trimmomatic (Galaxy version 0.38.1)	[40]
Quality control	FastQC (Galaxy version 0.73 + galaxy0)	(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
rRNA removal	SortMeRNA (Galaxy version 2.1b.6)	[41]
mRNA quantification and mapping	Kallisto quant (Galaxy version 0.46.0.4)	[42]
<i>MetaP</i>		
Protein quantification	MaxQuant (Galaxy version 1.6.3.4)	[43, 44]

*Tools unique for the optimized MetaG workflow

Ontology [48], while KoFamScan [49] provides enzyme commission numbers (EC) and annotation from KEGG [50]. For prediction of carbohydrate-active enzymes (CAZymes), the MetaG workflow uses Hidden Markov Models from dbCAN [38], downloaded from <https://bcb.unl.edu/dbCAN2/> and used within the software HMMER [51]. To facilitate downstream analyses, we combine all the functional annotations from InterProScan, KoFamScan and dbCAN into one file using a script within the Galaxy implementation of awk to generate a tabular file with one protein per row and the different annotations in individual columns. This file of functional annotation of all gene products in the metagenome, together with the output from taxonomic analysis, is used for more detailed data exploration and interpretation in ViMO (Fig. 1.18). Optionally, the putative genes and proteins from FragGeneScan [35] can be manually augmented with strains from public repositories such as NCBI, UniProt or IMG.

Workflow for metatranscriptomics (MetaT)

The MetaT workflow provides all the processing steps and parameters to analyze raw metatranscriptomics

paired-end reads. This multi-step workflow contains data cleaning/trimming, RNA filtering, mRNA quantification, and mapping to the predicted genes from the metagenome from the MetaG workflow (Table 1).

As *input files* (Fig. 1.11), the MetaT workflow accepts Illumina FASTQ sequence files (forward and reversed reads), which can be uploaded to Galaxy via web interface and organized as a collection of paired datasets. The workflow includes data preprocessing, where *quality control* (Fig. 1.12) of the sequences is done with FastQC to assess the overrepresentation of features, such as primers or adapters, with a Phred threshold of 20. Adapter sequences are automatically detected and *trimmed* (Fig. 1.13) by Trim Galore!. Sequencing of RNA results in a mixture of coding and non-coding RNA fragments, and the highly abundant ribosomal RNA in the samples are *filtered out* (Fig. 1.14) in order to use only mRNA transcripts for the analysis [52]. Thus, rRNA and tRNA are removed using the software SortMeRNA [41]. This is followed by *mRNA quantification and mapping* (Fig. 1.15). The mRNA quantification is done with the software Kallisto [42], which pseudoaligns mRNA reads

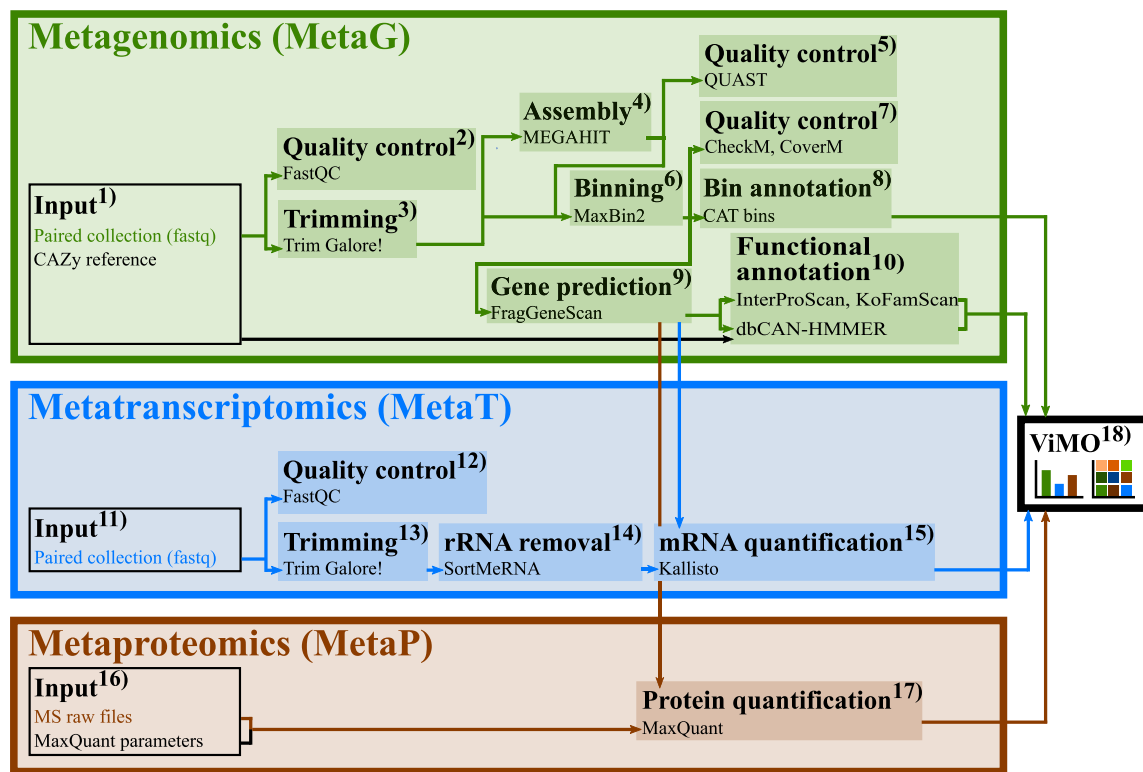


Fig. 1 Workflows for meta-omics. The integrated analysis of meta-omics contains a MetaG, MetaT and MetaP workflow. MetaG includes data preprocessing steps with quality control and trimming, followed by assembling, binning and taxonomically annotation of the MAGs. Open reading frames (ORFs) and nucleotide sequences are predicted by FragGeneScan. Functional annotation is performed by InterProScan and dbCAN-HMMER. The predicted ORFs and nucleotide sequences are further used in the MetaP and MetaT workflow; hence, the MetaG serves as the base analysis and the MetaT and MetaP are mapped onto the MetaG. After preprocessing the data and rRNA removal, the predicted nucleotide sequences from the MetaG workflow are used for the mRNA quantification and mapping by Kallisto, as well as for MaxQuant in the MetaP workflow

onto nucleotide sequences (in this case the predicted genes from FragGeneScan in the MetaG workflow), and is thereby skipping alignment for redundant kmers in the De Bruijn graph from the transcriptome, which saves time while being accurate and sensitive [42]. The outputs from Kallisto, one per sample, are finally joined in order to generate one single file to use in ViMO (Fig. 1.18).

Workflow for metaproteomics (MetaP)

For the MetaP workflow, RAW files from the mass spectrometric analysis are uploaded to Galaxy via the web interface or FTP and organized as a collection list. MaxQuant [43] within Galaxy (version 1.6.17.0) require uploading a file describing the experimental design, i.e., a text-file with a list of all the RAW files and which experiment/biological replicate they belong to (Fig. 1.16). The rest of the parameters can be selected at run-time in Galaxy, including proteolytic cleavage, matching between runs, fixed and variable peptide modifications, and parameters for identification; for this dataset, these are described in Delogu et al. [28]. MaxQuant (Fig. 1.17) in Galaxy is then used to identify and quantify proteins by

matching MS/MS spectra onto the protein sequences predicted by FragGeneScan [35] in the MetaG workflow (Table 1). The output from MaxQuant (Proteingroups.txt) is used for downstream analysis in ViMO (Fig. 1.18). It should be noted that MaxQuant has some limitations with large databases (>500.000 protein entries), and we are seeking to replace this software with FragPipe in the future versions of this MetaP workflow to scale along the fast growth in metagenomics in recovering hundreds of MAGs from various samples.

Data integration in ViMO: visualizer for meta-omics

Analyzing and exploring multi-levelled meta-omics data is not a trivial task and requires linking information from metagenomics, such as the presence of specific pathways within selected MAGs, with expression data from transcriptomics and proteomics analysis. This level of data integration is complicated and not practical in spreadsheet applications such as Excel and is thus typically achieved through scripting with Python or R. Preferably, interactive tables and maps would allow data exploration where the user can browse through the catalog of MAGs

present in the samples and their metabolisms, while receiving visualizations of expressed genes and functions. This was our motivation for developing ViMO.

ViMO is provided with a script that reads the following output from the MetaGTP workflows and generates a Masterfile and a Contig file for import: (1) All the dereplicated genomes with their contigs, (2) the file containing all putative proteins annotated with functional predictions from InterProScan, dbCAN and KoFamScan, (3) metagenomic coverages of contigs as well as completeness, contamination and strain heterogeneity from CoverM and CheckM, (4) the taxonomic annotations from CAT/BAT, (5) the quantification of mRNA from Kallisto, and (6) the quantification of proteins from MaxQuant. Obviously, ViMO is also functionable with a similar Masterfile generated from a custom workflow, either in Galaxy or elsewhere, e.g., using different software for quantification such as FragPipe [53], as long as the essential columns are present in the final Masterfile; this is described in the help-section of ViMO.

Once the files are loaded, ViMO provides four core analyses. (1) MAGs, an overview of all detected MAGs including counts of contigs and genes, contamination, completeness and taxonomy, as well as a figure of %GC versus metagenomic coverage to illustrate the coherence within each MAG. (2) CAZy, an overview of all detected CAZymes including carbohydrate esterases (CEs), glycosyl transferases (GTs), glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate binding domains (CBMs), auxiliary activities (AAs) and components of cellulosomes, with their counts and abundances at both mRNA and protein level. Heatmaps allow for visualization of temporal changes between samples, if applicable in the experimental design. (3) KEGG, an overview of all genes with a KEGG annotation, sorted and selectable into KEGG pathways, with counts and abundances at both mRNA and protein level (Fig. 2A). ViMO allows filtering down to a specific pathway and downloads KEGG-maps and highlight the detected enzymes within the pathways with colors representing abundance, at both mRNA and protein level (Fig. 3). This allows detection of highly expressed pathways within the microbial community and in which MAGs they are most abundant. While this is possible to retrieve through the standard KEGG web-interface (KEGG Mapper [50]), one would have to copy all the proteins and abundances into the web-interface manually and for one MAG at the time, while ViMO retrieves this information automatically while the user browses through the MAGs. (4) KEGG-Modules, calculate the module completion fraction (mcf) for all KEGG-modules in all MAGs and visualize the metabolic potential of each MAG in a heatmap (Fig. 2B). This can optionally be filtered to lower-level KEGG categories.

The powerful KEGG modules network allows for inspecting the completeness, meaning the presence of the complete set of enzymes required for a given metabolic reaction and was implemented in ViMO using the R-package MetQy [54]. Alternatively, similar heatmaps can be generated with the KEGGDecoder software [55]; however, here this is done automatically within ViMO and with interactive filtering options.

In terms of limitations and guidelines for best usage, ViMO works best with meta-omics datasets containing up to ~50 MAGs/~150.000 genes due to the extensive plotting and interactivity. Although we have successfully assessed its functionality with larger datasets of >250 MAGs, we have observed that the app slows down remarkably due to R being an on-the-fly interpreted language. Moreover, functional graphs with >250 MAGs (with individual colors) become less useful/interpretable, and we advise our users to rather employ parts of the ViMO code to their data locally to better optimize the parameters to fit the data. The code is freely available under GPL3 at <https://github.com/magnusarntzen/ViMO>.

Alternative optimized workflow for metagenomics analysis in Galaxy

As metatranscriptomics and metaproteomics are mapped to, and thus depend on the quality of the metagenomic data, it is critical that this step is optimized using the best method available. The optimized MetaG workflow contains both a co-assembly (Fig. 4.4, 4.5), similar to the standard MetaG workflow above, but also with individual assemblies ran in parallel. For the individual assemblies, *trimmed* paired-end reads (Fig. 4.3) are *split* (Fig. 4.6) using the sample name as an element identifier into smaller collections per sample, containing the forward and reversed reads for each sample. Each sample is then *assembled* (Fig. 4.7) by MEGAHIT with k-mer sizes of 21, 29, 39, 59, 79, 99, 119, and 141, and the quality for assemblies are analyzed with QUAST in meta-mode (Fig. 4.9). The contigs are then *binned* (Fig. 4.8) by MaxBin2 (contig length ≥ 1000) and MAGs from each sample are *merged* (Fig. 4.10) together with the co-assembly into one collection with a sample identifier to trace the sample origin of the MAG in further downstream analysis. The merging of MAGs is followed by *dereplication* (algorithm ANImf, P_{ani}: 0.90, S_{ani}: 0.95) with dRep [32] (Fig. 4.11) for identification of groups of highly similar genomes and choosing the best representative genome within the genome sets. Completeness, contamination, and strain heterogeneity of each MAG is then reported by *CheckM* and read coverage by *CoverM* (Fig. 4.12). Further downstream analysis involves, as in MetaG, the prediction of nucleotide sequences and ORFs by FragGeneScan and functional

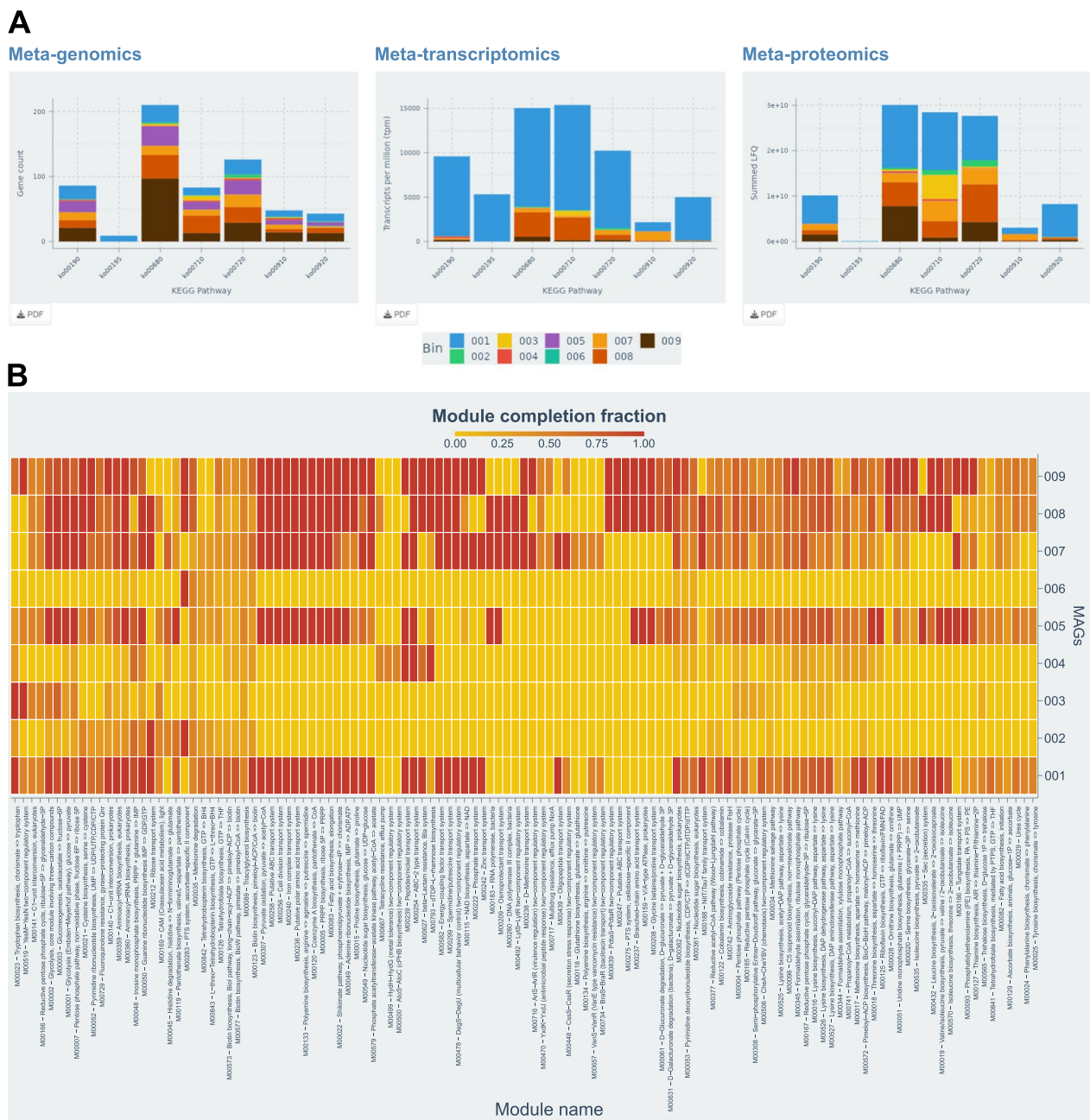


Fig. 2 ViMO visualizations. **A** ViMO produces bar plots to visualize the gene counts and abundances of KEGG-pathways in the different bins, here filtered to pathways in energy metabolism. For metagenomics, all timepoints are used, while for metatranscriptomics and metaproteomics, only the first timepoint is shown here and the user can select which sample/timepoint to visualize. In addition, ViMO displays heatmaps with all timepoints within one graph for metatranscriptomics and metaproteomics to visualize temporal changes (data not shown). **B** ViMO calculates the module completion fraction (mcf) for all KEGG modules (x-axis; only a subset displayed here) and MAGs (y-axis) and thus visualize the metabolic potential of each MAG. The set of visible modules can be filtered to selected KEGG pathways for in-depth exploration

annotation by InterProScan and dbCAN-HMMER. The predicted ORFs and nucleotide sequences are further used in the MetaP and MetaT workflow (Fig. 1).

Table 2 shows the contig counts and dataset statistics obtained by using both the standard and optimized

MetaG workflows, on both the small bioreactor dataset used for developing these workflows, and for an in-house large complementary (Comp) dataset with 253 MAGs to stress-test the analysis pipelines.

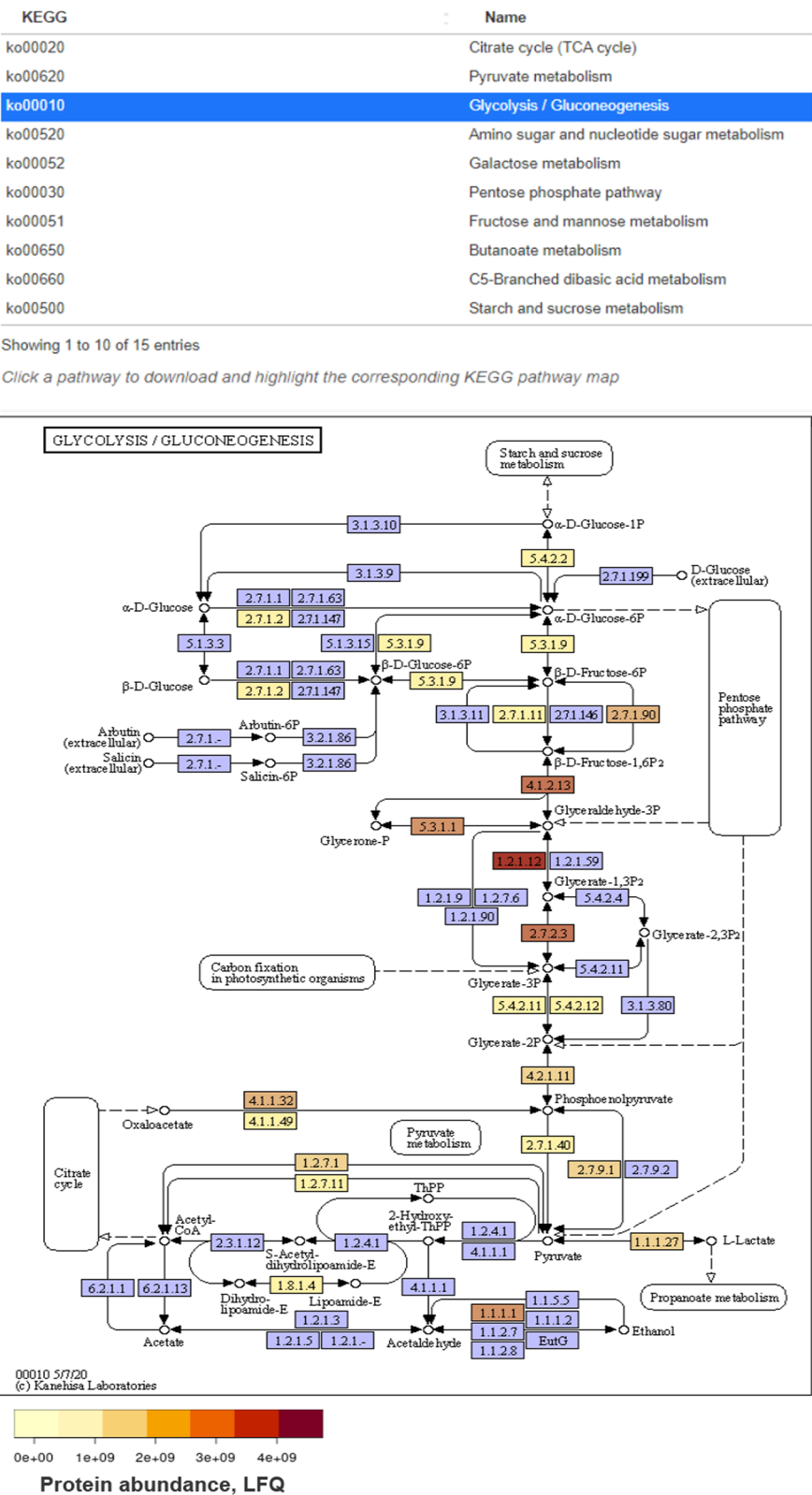


Fig. 3 Annotated KEGG-maps. In VIMO, when KEGG-pathways are selected (top, filtered to pathways in carbohydrate metabolism), a KEGG-map is downloaded and annotated with abundances of expressed genes for the selected MAG. Here is shown the Glycolysis/Gluconeogenesis pathway of MAG001, a bacterium from the Tissierellia class in the SEM1b community, annotated with metaproteomic abundances ranging from low-abundant (0 LFQ; light yellow) to high-abundant (4e9 LFQ; dark red); blue enzymes are not detected in the metaproteome for this MAG

Table 2 Contigs and dataset statistics for the two MetaG workflows

Dataset	Workflow	Contigs (MEGAHIT)	Unbinned contigs (%)	N50/L50	Longest contig
<i>Bioreactor</i> , Small dataset with 10 MAGs	MetaG	11,386	6	38,958/118	391,662
	Optimized MetaG		5		
	Co-assembly	11,296		28,943/145	351,556
	Individual assemblies:				
	Sample-1	4003		44,326/58	391,662
<i>Comp</i> , Large dataset with 253 MAGs	Sample-2	12,098		27,635/128	391,715
	MetaG	1,923,986	11	2309/93,659	797,197
	Optimized MetaG		20		
	Co-assembly	2,331,350		2474/10,2387	1,098,235
	Individual assemblies:				
	Sample-1	310,224		2109/14,283	625,541
	Sample-2	511,518		2530/24,745	715,289
	Sample-3	450,745		2083/20,003	872,994
	Sample-4	532,077		2820/21,306	862,734
	Sample-5	303,656		2548/13,484	497,688
	Sample-6	223,130		2523/9460	1,098,235

Contigs were analyzed with CoverM and metaQuast. For the optimized MetaG workflow, which includes both co- and single assemblies, the percentage of unbinned contigs is reported as the average number after dereplication. Both a small (bioreactor) and a large (in-house complementary; comp) dataset is included to stress-test the analysis pipelines

Table 3 Quality of MAGs generated in the two workflows

MAG quality count	Bioreactor		Comp	
	MetaG	Optimized MetaG	MetaG	Optimized MetaG
Low ^a	6	0	172	42
Medium ^b	3	1	63	51
High ^c	1	6	18	50
Sum	10	7	253	143

The number of MAGs with low, medium, and high quality are counted for the standard and optimized MetaG workflow for both the Bioreactor and the Comp dataset

^a < 50% completion, ≥ 10% contamination

^b ≥ 50% completion, < 10% contamination

^c > 90% completion, < 5% contamination

Contigs with similar tetranucleotide frequencies are binned to one MAG [56], and as is evident from Tables 2 and 3, the extra contigs provided by the individual assemblies in the optimized MetaG workflow, aids in the binning process and increases the number of high-quality MAGs compared to the bare use of co-assembly in the standard MetaG workflow.

The optimized MetaG workflow results in 10 MAGs from the co-assembly and 11 MAGs from the individual assemblies, from which 7 MAGs of almost exclusively high-quality are selected after the dereplication

process (Table 4), whereas from the standard MetaG workflow, only one MAG is of high-quality.

Completeness and contamination of the MAGs are highly valuable metrics for the reliability of reconstructed metabolic pathways and annotated taxonomy [57]. In order to obtain at least “good-quality” MAGs (completeness > 70% and contamination < 10%) based on the standards by Bowers et al. [58], Galaxy currently contains three tools for this purpose: Binning_refiner [59], DAS Tool [60], and dRep. Binning_refiner searches for common contigs between each set of MAGs from different binning iterations creating the refined MAG, resulting in a non-redundant set of MAGs with decreased contamination and increased completeness [59]. Redundant MAGs lead to misinterpretations of the relative abundance and population dynamics throughout the different samples [61], a problem that is also addressed by DAS Tool and dRep. DAS Tool refines MAGs by evaluating the common contig set between MAGs, again obtained by different binning iterations, and the remaining potential MAGs are selected based on the F1-score followed by an iterative selection of high-scoring MAGs [60]. Another approach to extract only one high-quality representative of a replicate set of MAGs is dereplication by dRep using the MASH- and gANI algorithms to estimate distance and similarity between the MAGs and taking preset completion and contamination scores into account [32]. Dereplication results in a set of at least “good-quality” MAGs, which improves the downstream annotations and

Table 4 Taxonomy and quality values for MAGs generated with the two workflows

MetaG (Bin)	Taxonomy	Completeness	Contamination	Strain heterogeneity
Bin1	<i>Hungateiclostridium</i>	87.72	24.24	0.00
Bin2	<i>Coprothermobacter proteolyticus</i>	25.00	0.00	0.00
Bin3	<i>Coprothermobacter proteolyticus</i>	14.61	4.55	100.00
Bin4	<i>Coprothermobacter proteolyticus</i>	23.38	10.96	48.15
Bin5	<i>Acetomicrobium</i>	97.41	14.66	100.00
Bin6	<i>Coprothermobacter proteolyticus</i>	9.25	0.00	0.00
Bin7	Firmicutes	97.90	6.53	0.00
Bin8	<i>Methanothermobacter</i>	100	1.29	0.00
Bin9	Clostridia	98.08	8.44	0.00
Bin10	<i>Thermoclostridium stercorarium</i>	83.92	6.29	0.00
<i>Optimized MetaG</i>				
Opt-Bin1	<i>Hungateiclostridium thermocellum</i>	99.33	0.00	0.00
Opt-Bin2	<i>Coprothermobacter proteolyticus</i>	100.00	1.79	0.00
Opt-Bin3	<i>Acetomicrobium</i>	97.46	1.69	100.00
Opt-Bin4	<i>Tepidanaerobacter</i>	98.08	7.69	0.00
Opt-Bin5	Firmicutes	97.90	4.55	0.00
Opt-Bin6	<i>Methanothermobacter</i>	100.00	3.69	29.41
Opt-Bin7	<i>Thermoclostridium stercorarium</i>	98.60	4.06	0.00

Quality values were obtained by CheckM and taxonomic annotation by the program 'CAT bins'. The data is from the Bioreactor dataset

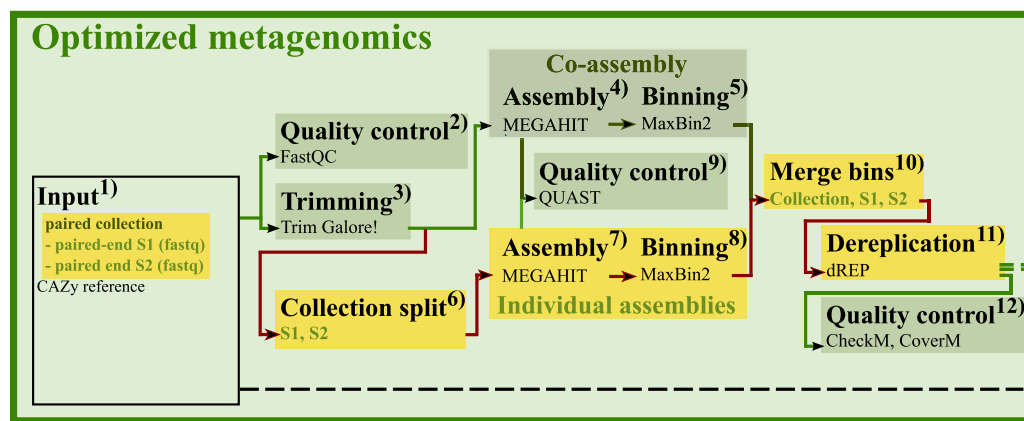


Fig. 4 Optimized metagenomic workflow. We have created an optimized MetaG workflow to improve the quality of the MAGs. This is achieved by assembly and binning of the reads individually, in parallel to a co-assembly, and combined and dereplicated to exclude redundant MAGs before bin annotation, gene prediction and functional annotation. Two samples S1 and S2 are shown as an example. Differences to the original MetaG workflow are highlighted in yellow

is therefore an important tool in our optimized MetaG workflow (Fig. 4).

Concluding remarks

Herein we have presented the development of three integrated workflows for the analysis of meta-omics data, including a new tool for data visualization, ViMO. The workflows have been developed using a small dataset containing 10 MAGs, a subset of this is also provided as

example input in the online version of ViMO. In addition, we have verified the workflows' applicability to a larger dataset, as exemplified in Tables 2 and 3. Together, these Galaxy-based workflows and interactive visualizations allows scientists to explore and characterize microbiomes without prior knowledge in the use of compute clusters and scripting. Although nesting software in workflows promotes reproducible science, biological samples naturally vary in their complexity and heterogeneity, and may

require different tool parameters. We therefore recommend that as our workflows are adapted by the wider community, each step in the workflows are adjusted and parameters optimized before analyzing new sample material. Our workflows may also be further extended with new capabilities from existing microbiome research tools [62] or as new tools are added to the Galaxy Platform in the future, such as for example FragPipe [53] for enhanced proteomics analysis, and Prodigal [63] for predicting genes in the MetaG workflow.

Abbreviations

MetaG	Metagenomics
MetaT	Metatranscriptomics
MetaP	Metaproteomics
MAG	Metagenome-assembled genomes
CAZymes	Carbohydrate-active enzymes
ORFs	Open reading frames
Mcf	Module completion fraction

Acknowledgements

The authors would like to thank Live H. Hagen, Norwegian University of Life Sciences for valuable discussions regarding the implementation of the optimized metagenomics workflow.

Author contributions

VCS designed and developed workflows in Galaxy and wrote the paper, FD, BK, PBP contributed with the bioreactor data used for developing the workflows as well as gave input on the development, PK, BB, SM, JEJ, BG, PDJ, TJG implemented and tested new software into Galaxy as well as contributed to various parts of the workflow development in Galaxy, MA designed and supervised the study, as well as developed the web-application ViMO. All authors read and approved the final manuscript.

Funding

This research was supported by the Novo Nordisk Foundation through Grant NNF20OC0061313, and by the Research Council of Norway INFRASTRUKTUR-program Grant No. 295910, and through the joint University of Minnesota-NMBU Norwegian Centennial Chair program. The Galaxy server that was used for some calculations 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)).

Availability of data and materials

The metagenomics, metatranscriptomics, and metaproteomics workflows developed herein are shared publicly within Galaxy; alternatively, they can be accessed via these links: <https://usegalaxy.eu/u/mgnsrntzn/w/metagextended>, <https://usegalaxy.eu/u/mgnsrntzn/w/metap>, <https://usegalaxy.eu/u/mgnsrntzn/w/metat>, ViMO is accessible online at <https://magnusarntzen.shinyapps.io/VisualizerForMetaOmics/> and the source code is available at <https://github.com/magnusarntzen/ViMO> under the GPL-3 license. For the bioreactor data used in the development of Galaxy workflows and ViMO, metagenomics and metatranscriptomics sequencing reads are available in the sequence read archive under SRP134228, with specific numbers listed in Supplementary Table 6 in Kunath et al. [29]. The proteomics data for the same dataset is available in ProteomeXchange/PRIDE [64] with the data set identifier PXD016242.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences (NMBU), P.O. Box 5003, 1432 Ås, Norway. ²Department of Biochemistry, Biophysics and Molecular Biology, University of Minnesota, Minneapolis, MN 55455, USA. ³Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany. ⁴Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN 55455, USA. ⁵Faculty of Biosciences, Norwegian University of Life Sciences (NMBU), P.O. Box 5003, 1432 Ås, Norway.

Received: 16 June 2023 Accepted: 5 July 2023

Published online: 07 July 2023

References

- Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, et al. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*. 2003;424(6952):1042–7. <https://doi.org/10.1038/nature01947>.
- Cheng S, Xian W, Fu Y, Marin B, Keller J, Wu T, et al. Genomes of subaerial *Zygnematophyceae* provide insights into land plant evolution. *Cell*. 2019;179(5):1057–67.e14. <https://doi.org/10.1016/j.cell.2019.10.019>.
- Knief C, Delmotte N, Chaffron S, Stark M, Innerebner G, Wassmann R, et al. Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. *ISME J*. 2012;6(7):1378–90. <https://doi.org/10.1038/ismej.2011.192>.
- Eckert EM, Anicic N, Fontaneto D. Freshwater zooplankton microbiome composition is highly flexible and strongly influenced by the environment. *Mol Ecol*. 2021;30(6):1545–58. <https://doi.org/10.1111/mec.15815>.
- Kara EL, Hanson PC, Hu YH, Winslow L, McMahon KD. 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. <https://doi.org/10.1038/ismej.2012.118>.
- Blaser MJ. The microbiome revolution. *J Clin Invest*. 2014;124(10):4162–5. <https://doi.org/10.1172/JCI78366>.
- Obileke K, Onyeaka H, Meyer EL, Nwokolo N. Microbial fuel cells, a renewable energy technology for bio-electricity generation: a mini-review. *Electrochem Commun*. 2021;125:107003. <https://doi.org/10.1016/j.elecom.2021.107003>.
- Difford GF, Plichta DR, Løvendahl P, Lassen J, Noel SJ, Højberg O, et al. Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows. *PLOS Genet*. 2018;14(10):e1007580. <https://doi.org/10.1371/journal.pgen.1007580>.
- Verstraete W. The technological side of the microbiome. *NPJ Biofilms Microbiomes*. 2015;1(1):15001. <https://doi.org/10.1038/npjbiofilms.2015.1>.
- Reay DS, Davidson EA, Smith KA, Smith P, Melillo JM, Dentener F, et al. Global agriculture and nitrous oxide emissions. *Nat Clim Change*. 2012;2(6):410–6. <https://doi.org/10.1038/nclimate1458>.
- Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. Computational metaomics for microbial community studies. *Mol Syst Biol*. 2013;9:666. <https://doi.org/10.1038/msb.2013.22>.
- Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLOS Comput Biol*. 2010;6(2):e1000667. <https://doi.org/10.1371/journal.pcbi.1000667>.
- Hagen LH, Frank JA, Zamanzadeh M, Eijsink VGH, Pope PB, Horn SJ, et al. Quantitative metaproteomics highlight the metabolic contributions of uncultured phylotypes in a thermophilic anaerobic digester. *Appl Environ Microbiol*. 2017. <https://doi.org/10.1128/aem.01955-16>.
- Shakya M, Lo C-C, Chain PSG. Advances and challenges in metatranscriptomic analysis. *Front Genet*. 2019;10:904. <https://doi.org/10.3389/fgene.2019.00904>.
- Weber JL, Myers EW. Human whole-genome shotgun sequencing. *Genome Res*. 1997;7(5):401–9. <https://doi.org/10.1101/gr.7.5.401>.
- Vlaanderen J, Moore LE, Smith MT, Lan Q, Zhang L, Skibola CF, et al. Application of OMICS technologies in occupational and environmental

- health research; current status and projections. *Occup Environ Med*. 2010;67(2):136–43. <https://doi.org/10.1136/oem.2008.042788>.
17. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome Res*. 2017;27(5):824–34. <https://doi.org/10.1101/gr.213959.116>.
 18. Vosloo S, Huo L, Anderson CL, Dai Z, Sevilano M, Pinto A. Evaluating de Novo assembly and binning strategies for time series drinking water metagenomes. *Microbiol Spectr*. 2021;9(3):e0143421. <https://doi.org/10.1128/Spectrum.01434-21>.
 19. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. 2016;32(4):605–7. <https://doi.org/10.1093/bioinformatics/btv638>.
 20. Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, et al. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS ONE*. 2012;7(11):e49138. <https://doi.org/10.1371/journal.pone.0049138>.
 21. Eren AM, Kiehl E, Shaiber A, Veseli I, Miller SE, Schechter MS, et al. Community-led, integrated, reproducible multi-omics with anvio. *Nat Microbiol*. 2021;6(1):3–6. <https://doi.org/10.1038/s41564-020-00834-3>.
 22. Li L, Ning Z, Cheng K, Zhang X, Simopoulos CMA, Figeys D. iMetaLab Suite: a one-stop toolset for metaproteomics. *iMeta*. 2022;1(2):e25. <https://doi.org/10.1002/imt.2.25>.
 23. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*. 2005;15(10):1451–5. <https://doi.org/10.1101/gr.4086505>.
 24. Boekel J, Chilton JM, Cooke IR, Horvatovich PL, Jagtap PD, Käll L, et al. Multi-omic data analysis using Galaxy. *Nat Biotechnol*. 2015;33(2):137–9. <https://doi.org/10.1038/nbt.3134>.
 25. Thang M, Chua X, Price G, Gorse D, Field M. MetaDEGalaxy: Galaxy workflow for differential abundance analysis of 16s metagenomic data [version 2; peer review: 2 approved]. *F1000Research*. 2019. <https://doi.org/10.12688/f1000research.18866.2>.
 26. Batut B, Gravoil K, Defois C, Hiltemann S, Brugère JF, Peyretailade E, et al. ASaiM: a Galaxy-based framework to analyze microbiota data. *Gigascience*. 2018. <https://doi.org/10.1093/gigascience/giy057>.
 27. Mehta S, Crane M, Leith E, Batut B, Hiltemann S, Arntzen M, et al. ASaiM-MT: a validated and optimized ASaiM workflow for metatranscriptomics analysis within Galaxy framework [version 2; peer review: 2 approved]. *F1000Research*. 2021. <https://doi.org/10.12688/f1000research.28608.2>.
 28. Delogu F, Kunath BJ, Evans PN, Arntzen MØ, Hvidsten TR, Pope PB. Integration of absolute multi-omics reveals dynamic protein-to-RNA ratios and metabolic interplay within mixed-domain microbiomes. *Nat Commun*. 2020;11(1):4708. <https://doi.org/10.1038/s41467-020-18543-0>.
 29. Kunath BJ, Delogu F, Naas AE, Arntzen MØ, Eijssink VGH, Henrissat B, et al. From proteins to polysaccharides: lifestyle and genetic evolution of *Coprothermobacter proteolyticus*. *ISME J*. 2019;13(3):603–17. <https://doi.org/10.1038/s41396-018-0290-y>.
 30. Li DJ, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 2015;31(10):1674–6. <https://doi.org/10.1093/bioinformatics/btv033>.
 31. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072–5. <https://doi.org/10.1093/bioinformatics/btt086>.
 32. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*. 2017;11(12):2864–8. <https://doi.org/10.1038/ismej.2017.126>.
 33. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015;25(7):1043–55. <https://doi.org/10.1101/gr.186072.114>.
 34. von Meijenfeldt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol*. 2019;20(1):217. <https://doi.org/10.1186/s13059-019-1817-x>.
 35. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res*. 2010;38(20):e191. <https://doi.org/10.1093/nar/gkq747>.
 36. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res*. 2013;41(12):e121. <https://doi.org/10.1093/nar/gkt263>.
 37. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. *Bioinformatics*. 2013;29(19):2487–9. <https://doi.org/10.1093/bioinformatics/btt403>.
 38. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*. 2012;40(Web Server issue):W445–51. <https://doi.org/10.1093/nar/gks479>.
 39. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. *Nucleic Acids Res*. 2005;33(suppl_2):W116–20. <https://doi.org/10.1093/nar/gki442>.
 40. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*. 2014;30(15):2114–20. <https://doi.org/10.1093/bioinformatics/btu170>.
 41. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*. 2012;28(24):3211–7. <https://doi.org/10.1093/bioinformatics/bts611>.
 42. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34(5):525–7. <https://doi.org/10.1038/nbt.3519>.
 43. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 2008;26(12):1367–72. <https://doi.org/10.1038/nbt.1511>.
 44. Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc*. 2016;11(12):2301–19. <https://doi.org/10.1038/nbt.389310.1038/nprot.2016.136>.
 45. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. *Nucleic Acids Res*. 2003;31(1):371–3. <https://doi.org/10.1093/nar/gkg128>.
 46. Pedruzzi I, Viroire C, Auchincloss AH, Coudert E, Keller G, de Castro E, et al. HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res*. 2014;43(D1):D1064–70. <https://doi.org/10.1093/nar/gku1002>.
 47. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar Gustavo A, Sonhammer ELL, et al. Pfam: the protein families database in 2021. *Nucleic Acids Res*. 2020;49(D1):D412–9. <https://doi.org/10.1093/nar/gkaa913>.
 48. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25–9. <https://doi.org/10.1038/75556>.
 49. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*. 2020;36(7):2251–2. <https://doi.org/10.1093/bioinformatics/btz859>.
 50. Kanehisa M, Sato Y, Kawashima M. KEGG mapping tools for uncovering hidden features in biological data. *Protein Sci*. 2022;31(1):47–53. <https://doi.org/10.1002/pro.4172>.
 51. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 2011;39(suppl_2):W29–37. <https://doi.org/10.1093/nar/gkr367>.
 52. Jack G, Hughes M. Gene expression profiling: metatranscriptomics. *Methods Mol Biol (Clifton, NJ)*. 2011;733:195–205. https://doi.org/10.1007/978-1-61779-089-8_14.
 53. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFrager: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods*. 2017;14(5):513–20. <https://doi.org/10.1038/nmeth.4256>.
 54. Martinez-Vernon AS, Farrell F, Soyer OS. MetQy-an R package to query metabolic functions of genes and genomes. *Bioinformatics*. 2018;34(23):4134–7. <https://doi.org/10.1093/bioinformatics/bty447>.
 55. Graham ED, Heidelberg JF, Tully BJ. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J*. 2018;12(7):1861–6. <https://doi.org/10.1038/s41396-018-0091-3>.
 56. Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res*. 2003;13(2):145–58. <https://doi.org/10.1101/gr.335003>.
 57. Chen LX, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and complete genomes from metagenomes. *Genome Res*. 2020;30(3):315–33. <https://doi.org/10.1101/gr.258640.119>.
 58. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome

- (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol.* 2017;35(8):725–31. <https://doi.org/10.1038/nbt.3893>.
59. Song W-Z, Thomas T. Binning_refiner: improving genome bins through the combination of different binning programs. *Bioinformatics.* 2017;33(12):1873–5. <https://doi.org/10.1093/bioinformatics/btx086>.
 60. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol.* 2018;3(7):836–43. <https://doi.org/10.1038/s41564-018-0171-1>.
 61. Evans JT, Denef VJ. To dereplicate or not to dereplicate? *mSphere.* 2020. <https://doi.org/10.1128/mSphere.00971-19>.
 62. Easterly CW, Sajulga R, Mehta S, Johnson J, Kumar P, Hubler 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. <https://doi.org/10.1074/mcp.RA118.001240>.
 63. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* 2010;11:119. <https://doi.org/10.1186/1471-2105-11-119>.
 64. Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* 2022;50(D1):D543–52. <https://doi.org/10.1093/nar/gkab1038>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

