

InterOmics Tutorial - Tools and methods for the analysis of omics data and biodiversity



Angelica Tulipano[✉], Andreas Gisel

CNR, Institute for Biomedical Technologies, Bari, Italy

Received 11 March 2014; Published 18 March 2014

Tulipano A and Gisel A (2014) *EMBNET.JOURNAL* 20, e759. <http://dx.doi.org/10.14806/ej.20.0.759>

The CNR [Institute for Biomedical Technologies \(ITB\)](http://www.itb.cnr.it)¹ in Bari (IT), with support from the Italian [Flagship project InterOmics](http://www.interomics.eu)², organised a

[Tutorial-Day](http://www.ba.itb.cnr.it/bip-day/tutorial)³ as a satellite event of the *BIP-Day 2013* workshop (see related article in the present volume). The tutorial was organised in three 3-hour events, covering metagenomics, phylogenetics and data analysis of non-coding RNA. The event took place on 6 December 2013 at the Department of Physics *Michelangelo Merlin* of the University of Bari and [INFN](http://www.ba.infn.it)⁴, as their computing infrastructure was used to guarantee the required performance for such data analysis approaches. While the services required for the first two sessions were already hosted on the INFN computer infrastructure, the third session was run on Virtual Machines (VMs). Four VMs with the analysis pipeline pre-installed, each with 16 CPU and 200GB shared memory, were used to serve 40 participants. Giacinto Donvito, from Bari University's Department of Physics continuously monitored the infrastructure to guarantee a flawless service.

The morning session started with a tutorial on the *Classification and quantification of the mi-*

BioMaS (Bioinformatic analysis of Metagenomic Amplicons)

The fundamental purpose of **BioMaS (Bioinformatic analysis of Metagenomic Amplicons)** is to equip the biomolecular researcher involved in taxonomic studies of environmental microbial communities with a comprehensive and user-friendly workflow including all the fundamental steps for the NGS amplicon-based metagenomic analysis, by guiding his path from raw sequences to final taxonomic identification. In its current version, BioMaS can be used for the analysis of Illumina MiSeq sequences from bacterial environments. BioMaS computation consists in four steps and requests as input the paired-end reads outputted by Illumina platforms:

1. In the first step the quality of raw data is evaluated by using FastQC [1]. Then the paired-end reads are processed by Flash [2] to merge them into a consensus sequence. The non-merged pairs undergo to the trimming of low-quality regions by means of trim-galore [3].
2. Then consensus sequences are dereplicated by means of Usearch [4].
3. The dereplicated consensus sequences and the paired-end reads are mapped on RDP [5] by mean Bowtie2 [6]. The mapping data are filtered according to query coverage (70%) and similarity percent (97%) and the filtered data are store in a file-format suitable for taxonomic assignment.
4. Taxonomic assignments are performed by means of TANGO [7] using a taxonomic guide tree corresponding to the taxonomy annotation represented in the reference database. As a result, BioMaS produces a graphical taxonomic three representation and several pie-charts that describe the taxonomic complexity of the microbiota at different rank (from phylum to species).

Run Workflow

Upload files

UPLOAD R1 FILE

Scogli file | nessuno selezionato | Start Upload

Reset list files

File uploaded:

UPLOAD R2 FILE

Scogli file | nessuno selezionato | Start Upload

Reset list files

File uploaded:

R1 file path:

R2 file path:

Base name

Mail recipient

Submit

Monitoring jobs

Mail recipient

Show jobs

Jobs

Figure 1. Screenshot from the BioMaS website.

1 www.itb.cnr.it
2 www.interomics.eu

3 www.ba.itb.cnr.it/bip-day/tutorial
4 <https://www.ba.infn.it/>

crobiome using metagenomic amplicons. Bruno Fosso, from the Department of Biotechnology and Biopharmaceutical Biosciences of the University of Bari (IT), and Monica Santamaria, from the CNR [Institute of Biomembranes and Bioenergetics \(IBBE\)](#)⁵, presented a modular pipeline (BioMaS) using third-party tools and ad hoc python and bash scripts. BioMaS is a web-service on the INFN/UNIBA infrastructure (Figure 1). High-level SaaS (Software as a Service) services are applied to facilitate the use of BioMaS components that are already suitably configured and optimised to run on the dedicated infrastructure. BioMaS allows the analysis of both bacterial and fungal environments, and alternative paths can be selected to process data obtained either by Roche 454 or Illumina sequencing technology.

The tutorial allowed participants to run a test data-set and understand in detail the pipeline and its functionality.

The second session covered *Instruments for the phylogenetic analysis for studies of biodiversity*. Saverio Vicario, from the CNR – ITB, and Bachir Balech, from the CNR – IBBE, presented the [BioVel](#)⁶ infrastructure. This allows users to build customised workflows (Figure 2) by selecting and applying successive 'services', or re-using existing workflows available from BioVel's library. By giving participants the opportunity to process specially provided test data, the tutorial offered profound insights into BioVel's significant functionality and performance.

The third session introduced participants to the world of non-coding RNA, in *Mapping and*

About Workflows

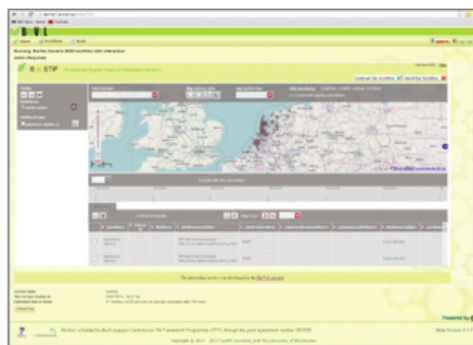
The quantity and heterogeneity of data in the biodiversity sciences have given rise to many distributed resources. Typically, researchers wish to combine these resources into multi-step computational tasks for a range of analytical purposes. Workflows, made of modularised units that can be repeated, shared, reused and repurposed, offer a practical solution for this task.

RUN

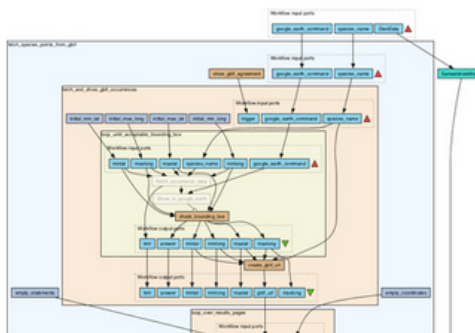
Workflows are executed through the BioVel portal, a simple web interface that provides access to a pool of ready-made workflows and allows you to manage, share and save workflow results. You can monitor and interact with running workflows through the portal, changing parameters and directing your analyses.

DESIGN AND CONSTRUCT

The Taverna Workbench provides a graphical environment where researchers can design and construct new analysis protocols, or customise existing protocols, before they are deployed and shared through the BioVel portal. New tools and resources can be discovered through the BiodiversityCatalogue. Plug-and-play components simplify workflow construction. Workflow components are modularised units that are well-documented and designed to be used as steps in other workflows.



Data selection using BioStiff service through BioVel Portal



A workflow run from Taverna workbench

Figure 2. Screenshot from the BioVel website illustrating its underlying workflows.

5 www.ibbe.cnr.it/

6 www.biovel.org

Teachers

- Angelica Tulipano ITB-CNR
- Arianna Consiglio ITB-CNR
- Flavio Licciulli ITB-CNR
- Andreas Gisel ITB-CNR

Technical Support

- Giacinto Donvito INFN - Bari

Data

- The data set is an Illumina sequencing of small RNA of mouse (*Mus musculus*)
- Small RNA analysis of wildtype Mouse embryo and Adar1 null mouse embryo at E11.0 and E11.5
- SRR361337 - Small RNAs from Mouse Embryo Day11
- SRR361338 - Small RNAs from ADAR1 KO Mouse Embryo Day11
- SRR361340 - Small RNAs from ADAR1 KO Mouse Embryo Day11.5

<http://www.ebi.ac.uk/ena/data/view/PRJNA148757>



Workflow

- The workflow starts with the input of the raw data you normally get from a sequencing center and at the end will give you a list of known and unknown miRNAs.
- If you have a series of experiments you will have access to a graphical interface where you can filter the results by different parameters to find the miRNA and related target genes of your interest.

```
@SRR361337.4 unknown:2:1:6:870 length=36
GGGAATCTGACTGTCTAANTCGTATGCCGCTTCT
+SRR361337.4 unknown:2:1:6:870 length=36
BBCB?@<BA;=BB>6=;3&;5434;4/.021?</=
@SRR361337.8 unknown:2:1:6:936 length=36
AGTTCTACAGTCCGACGATCTCGTATGCCGCTTCT
+SRR361337.8 unknown:2:1:6:936 length=36
ACBB:AB?2<>7>>553>1;3769;7#####
@SRR361337.12 unknown:2:1:6:653 length=36
TGGAGTGTGACAATGGTGTGTCGTATGCCGCTT
+SRR361337.12 unknown:2:1:6:653 length=36
BCCBC?BA@a@BBBBBA98;B?)>28@9@5/-);4@C
@SRR361337.16 unknown:2:1:6:238 length=36
ATACTGCATCAGGAAGTACTGGATCGTATGCCGTT
+SRR361337.16 unknown:2:1:6:238 length=36
BBA7Ae:@A?>;>AA8<4:6<695>4#####
@SRR361337.20 unknown:2:1:6:1221 length=36
TATGCACTTGTCCCGCCTGTTGATGCCGCTT
+SRR361337.20 unknown:2:1:6:1221 length=36
BBBAA<@<BB<A>>;6735?9<.:.+;6@3&, )5*=A
```

Figure 3. Screenshot from the ncRNA data-analysis website.

analysis of non-coding RNAs and small RNAs from NGS technologies. The specialist team from ITB Bari – Angelica Tulipano, Flavio Licciulli, Arianna Consiglio and Andreas Gisel – demonstrated a simple workflow to get from raw sequencing data to an expression profile of known and unknown miRNA and other non-coding RNAs. The workflow is based on publicly available software and in-house Perl scripts, assembled into a user-friendly pipeline. The results can be uploaded into a MySQL database with a simple graphical

interface to visualise, sort and filter the data. A customised data-set of Illumina small RNA sequences, at three time points (Figure 3), was provided to give the users first-hand experience of the pipeline's functionality.

More than 110 participants (on average 35 per session) attended the tutorials, demonstrating the urgent need for such events to help train life scientists to cope with the large and complex data-sets produced by NGS technologies.