

# DATA FORMATS IN NGS INTRODUCTION TO GALAXY

Bioinformàtica per a la Recerca Biomèdica

**Mireia Ferrer<sup>1</sup>, Álex Sánchez<sup>1,2</sup>**

**Esther Camacho<sup>1</sup>, Angel Blanco<sup>1,2</sup>**

1 Unitat d'Estadística i Bioinformàtica (UEB) VHIR

2 Departament de Genètica, Microbiologia i Estadística, UB

**1. Data formats used in NGS**

**2. Introduction to Galaxy**

## 2. Introduction to Galaxy

- An open, web-based platform integrating many popular tools and resources for intensive biomedical research.
- **What can be done?**
  - Obtain data from many data sources like UCSC Table Browser, Biomart, WormBase, or your own data
  - Prepare data for further analysis by rearranging or cutting data columns, filtering data and many other options
  - Analyze data by finding overlapping regions, determining statistics, preprocessing NGS data and much more
  - Share data and workflows

## 2. Introduction to Galaxy

The Galaxy page is divided into three panels:

**Tools** for uploading, processing and analysis

**Viewing panel**  
(menus, data, results)

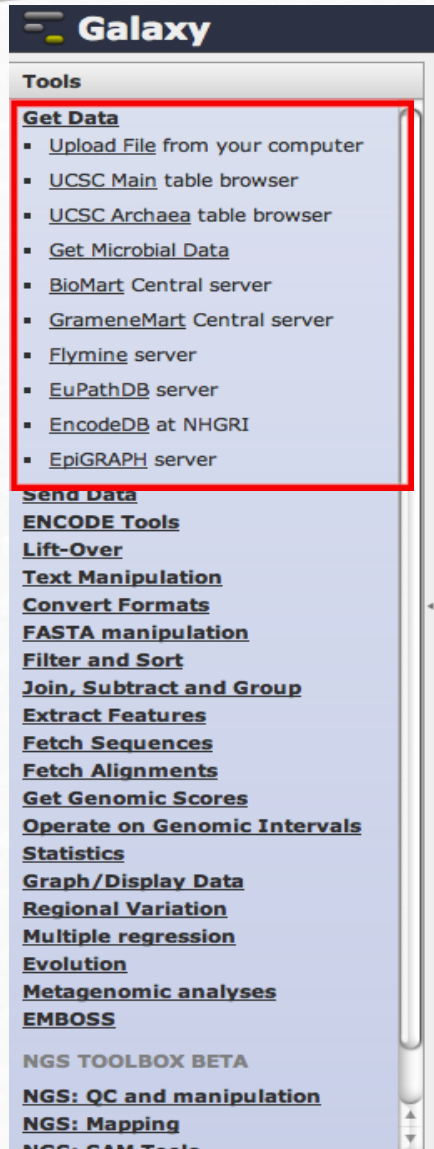
**History** of analysis steps and datasets

The screenshot displays the Galaxy web interface. The top navigation bar includes links for Analyze Data, Workflow, Visualize, Shared Data, Help, User, and a Galaxy logo. The interface is divided into three main panels:

- Tools Panel (Left):** A sidebar with a search bar and a list of tool categories. The categories include: Get Data, Collection Operations, GENERAL TEXT TOOLS, Text Manipulation, Filter and Sort, Join, Subtract and Group, Datamash, GENOMIC FILE MANIPULATION, FASTA/FASTQ, FASTQ Quality Control, SAM/BAM, BED, VCF/BCF, Nanopore, Convert Formats, Lift-Over, COMMON GENOMICS TOOLS, Operate on Genomic Intervals, Fetch Sequences/Alignments, GENOMICS ANALYSIS, Assembly, Annotation, Mapping, and Variant Calling.
- Viewing Panel (Center):** The main content area, outlined in red. It features a header for the JXTX (James P. Taylor Foundation for Open Science) with a logo of a sneaker. The text describes the foundation's mission to assist graduate students and organize mentoring sessions. A "Donate Now" button is present. A blue banner at the bottom of the panel provides information about the Galaxy SARS-CoV-2 portal at [covid19.galaxyproject.org](https://covid19.galaxyproject.org). Logos for Penn State, Johns Hopkins University, Oregon Health & Science University, TACC, and CyVerse are displayed at the bottom.
- History Panel (Right):** A sidebar outlined in green, titled "History". It contains a search bar and a message stating "Unnamed history (empty)". A blue banner at the bottom of the panel indicates that the history is empty and provides instructions on how to load data from an external source.



## 2. Introduction to Galaxy



### Tools for data analysis

#### Get Data

- From databases (UCSC Table Browser, ...)
- From uploaded files
- From urls

#### Text manipulation

#### Filter and Sort

#### Operate on Genomic Intervals

#### FASTA manipulation

#### NGS analysis

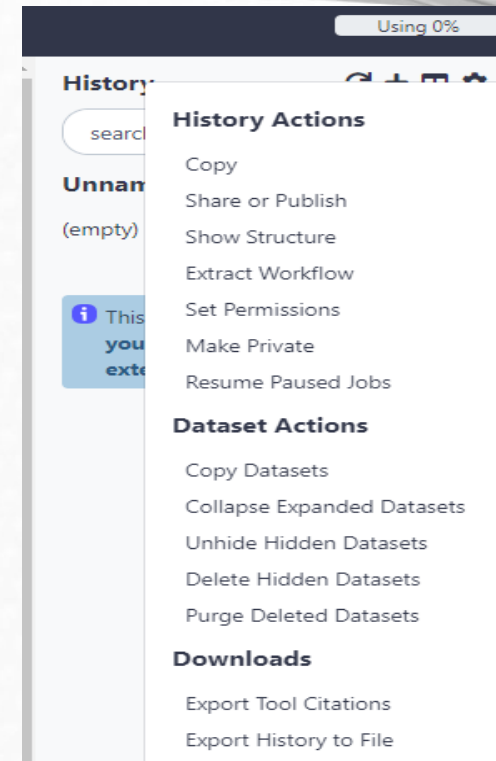
- QC
- Fastq file pre-processing
- Read Alignment / Mapping
- SAM tools

## 2. Introduction to Galaxy

### Histories

List saved histories and shared histories.

Work on Current History, create new, clone, share, create workflow, set permissions, show deleted datasets or delete history.



## 2. Introduction to Galaxy

### Workflows

**Galaxy**

Analyze DataWorkflowVisualizeShared DataHelpUserUsing 2%

Galaxy will be down for six hours beginning at 2:30 PM UTC, Tuesday, November 20 for filesystem maintenance.

Tools

search tools

Inputs

Get Data

Send Data

Lift-Over

Collection Operations

Text Manipulation

Datamash

Convert Formats

Filter and Sort

Join, Subtract and Group

Fetch Alignments/Sequences

NGS: QC and manipulation

NGS: DeepTools

NGS: Mapping

NGS: RNA Analysis

NGS: SAMtools

NGS: BamTools

NGS: Picard

NGS: VCF Manipulation

Workflow Canvas | Coding Exon SNPs

Exons

output

SNPs

output

Join

Join

With

output

Group

Select data

out\_file1 (tabular)

Sort

Sort Dataset

out\_file1

Details

Edit Workflow Attributes

Name:

Coding Exon SNPs

Version:

Version 1, 5 steps (active)

Tags:

Apply tags to make it easy to search for and find items with the same tag.

Annotation / Notes:

Describe or add notes to workflow

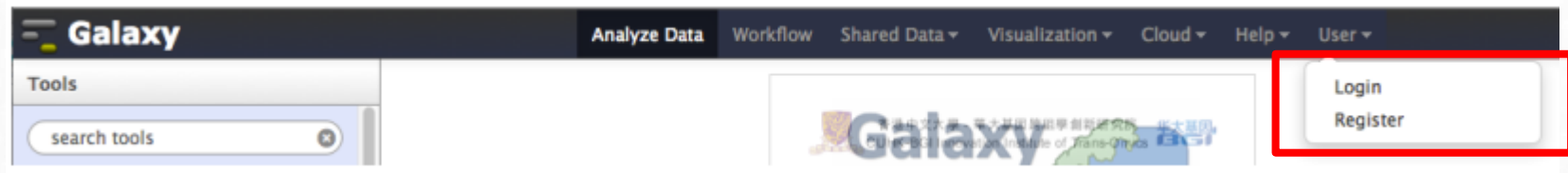
Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.

Workflows with all the analysis steps, allows user to repeat analysis using different datasets

## 2. Introduction to Galaxy

### Register for a Galaxy account

This will only take a moment, and will allow all the work that you do to persist between sessions and allow you to name, save, share, and publish Galaxy histories, workflows, datasets and pages. It allows you to store up to 250GB of data on this public server.



<https://usegalaxy.eu/>



## 2. Introduction to Galaxy

### Training Infrastructure as a Service

We want to help you conduct your training seminars. You provide the training, we provide you training infrastructure *at no cost*.

Why use UseGalaxy.eu training infrastructure?

- Free
- Private queue, no wait times
- No Galaxy Maintenance
- No Galaxy Administration
- Official Galaxy Training Materials guaranteed to work



Simply fill out the infrastructure request form and we'll get back to you shortly.

Find out more

After registration in [European Galaxy server](#)



[https://usegalaxy.eu/join-training/ueb\\_bi2021](https://usegalaxy.eu/join-training/ueb_bi2021)

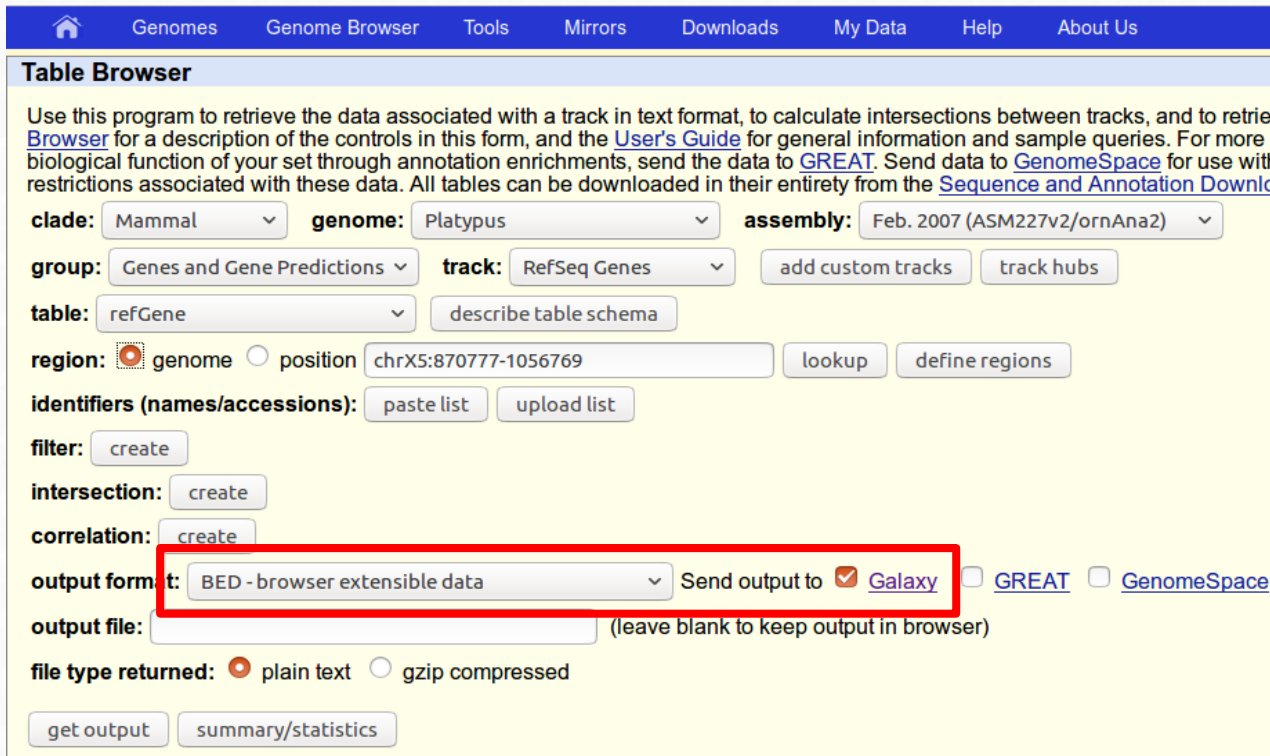
## 2. Introduction to Galaxy

### Importing data into Galaxy

1. From database queries (eg. UCSC): obtain a BED-formatted dataset of all RefSeq genes from platypus.

Get Data > UCSC Main – Table Browser tool

Set genome, RefSeq Genes, and BED output format (send to Galaxy)



**Table Browser**

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve a description of the controls in this form, and the [User's Guide](#) for general information and sample queries. For more information on the biological function of your set through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Download](#)

clade: Mammal genome: Platypus assembly: Feb. 2007 (ASM227v2/ornAna2)

group: Genes and Gene Predictions track: RefSeq Genes add custom tracks track hubs

table: refGene describe table schema

region: ☒ genome ☐ position chrX5:870777-1056769 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

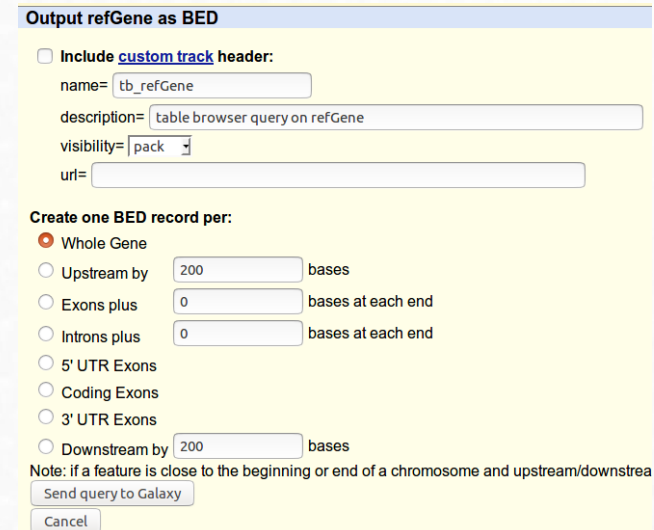
correlation: create

output format: BED - browser extensible data Send output to ☒ Galaxy ☐ GREAT ☐ GenomeSpace

output file: (leave blank to keep output in browser)

file type returned: ☒ plain text ☐ gzip compressed

get output summary/statistics



**Output refGene as BED**

☐ Include custom track header:

name= tb\_refGene

description= table browser query on refGene

visibility= pack

url=

Create one BED record per:

☒ Whole Gene

☐ Upstream by 200 bases

☐ Exons plus 0 bases at each end

☐ Introns plus 0 bases at each end

☐ 5' UTR Exons

☐ Coding Exons

☐ 3' UTR Exons

☐ Downstream by 200 bases

Note: If a feature is close to the beginning or end of a chromosome and upstream/downstream is specified, the feature will be truncated to fit the specified distance.

Send query to Galaxy

Cancel

## 2. Introduction to Galaxy

### Importing data into Galaxy







2. From a File on your computer / FTP file:

Get Data > Upload File

**Download from web or upload from disk**



Regular Composite Collection Rule-based








You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
 New File	72 b	fastqsang... 	 ----- Additional Sp... 		0% 

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

[http://chagall.med.cornell.edu/galaxy/rnaseq/GM12878\\_rnaseq1.fastqsanger](http://chagall.med.cornell.edu/galaxy/rnaseq/GM12878_rnaseq1.fastqsanger)

Type (set all):   Genome (set all):  

 Choose local file  Choose FTP file  Paste/Fetch data  Pause  Reset  Start  Close

## 2. Introduction to Galaxy

### Importing data into Galaxy

#### 3. From a website:

Get Data > Upload File

Copy this URL into the text-entry box:

url: [https://zenodo.org/record/582600/files/mutant\\_R1.fastq](https://zenodo.org/record/582600/files/mutant_R1.fastq)

Regular Composite Collection

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
New File	-	Auto-det...	unspecified (?)		

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

← 2. Paste file address in this box

1. click Paste/Fetch data

3. Start 4. Close

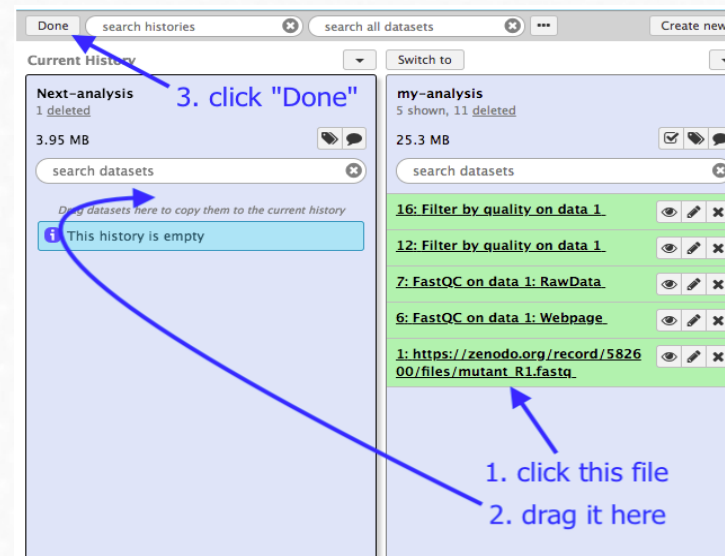
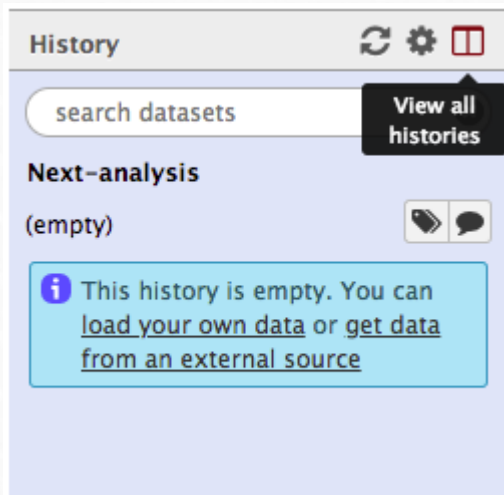
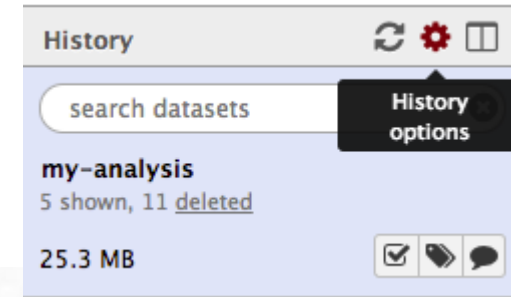
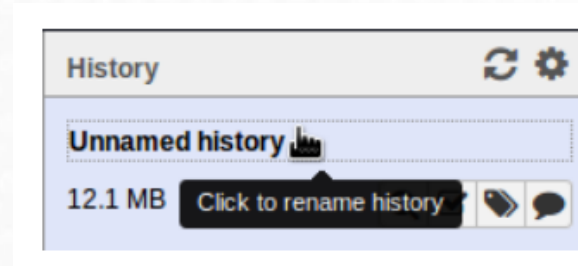
Type (set all): Auto-detect Q Genome (set all): unspecified (?)

Choose local file Paste/Fetch data Pause Reset Start Close

## 2. Introduction to Galaxy

### Managing histories

- Name your current history
- Create new history and rename it
- Manage datasets and histories:
- View all histories
- Drag files between histories (**new history must be set to current**)





## 2. Introduction to Galaxy

### Visualizing the dataset

- You can view file content clicking the eye icon in history.

The mutant\_R1.fastq file contains DNA sequencing reads from a bacteria, in FASTQ format:

```
@mutant-no_snps.gff-24960/1                                read 1 sequence
AATGTTGTCACTTGGATTCAAATGACATTTTAAATCTAATTATTCATGAATCGAACTAGTACGAAATGCAATGAG
+
5??A9?BBBDDDBEDDBFF+FGHHIIHHHEIHHIIHIIAHDHIIHIG#IIHIFHHHFGIII*IHHHIIHFIIHGICI
@mutant-no_snps.gff-24958/1
CAAAGTCGTTGGTCATATAAAAAACCGCGTACAGTCAACTATAGATACAATCAAGATAAACTCATGCACAGATTG
+
?A????@?DDDABDE9FGGGFGICFHIIIBGHIIIGICHHIFH=IHAFIHHHHHIFCIIIEIHAIFGIHIDDIHE
@mutant-no_snps.gff-24956/1
TATAAATTCAACTTTGCAACAGAACCATCTAATCTTCAACAACTGGCCCGTTTGTGAACTACTCTTTAATAAA
+
?????BBADD5DDDDGFGCFEECFBBICIII,IIHIICHIIHIFHHHHHIIHIIIIIIAHHHIHHH5FHDHHHH
```

History

search datasets

**my-analysis**  
1 shown

3.95 MB

1: [https://zenodo.org/record/582600/files/mutant\\_R1.fastq](https://zenodo.org/record/582600/files/mutant_R1.fastq)

View data

## 2. Introduction to Galaxy

### Create workflow from history

- From history options: Export workflow

The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

#### Workflow name

Workflow constructed from history 'prova'

Create Workflow

Check all

Uncheck all

#### Tool

Upload File

*This tool cannot be used in workflows*

FastQC

☒ Include "FastQC" in workflow

Filter by quality

☒ Include "Filter by quality" in workflow

#### History Items created

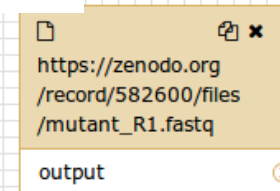
1 [https://zenodo.org/record/582600/files/mutant\\_R1.fastq](https://zenodo.org/record/582600/files/mutant_R1.fastq)

☒ Treat as input dataset [https://zenodo.org/record/582600/files/mutant\\_R1.fastq](https://zenodo.org/record/582600/files/mutant_R1.fastq)

2 FastQC on data 1: Webpage

3 FastQC on data 1: RawData

4 Filter by quality on data 1



# 2. Introduction to Galaxy

- <https://galaxyproject.org/learn/>

## Learn Galaxy

There are many approaches to learning how to use Galaxy. The most popular is probably to just dive in and use it. Galaxy is simple enough to use that you can do many analyses just by exploring the interface. However, you may miss much of the power this way.

Have you created or know of a resource that is useful for teaching with Galaxy? Then please share it! This will help others and also help get the word out about your resource. Use [this Google form](#) to describe your resource. **Also:** consider joining Galaxy Training Network and contributing your tutorial as described [here](#)!

## Tutorials by Galaxy Training Network

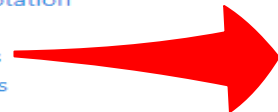
Thanks to a large [group of wonderful contributors](#) there is a constantly growing [set of tutorials](#) maintained by the [Galaxy Training Network](#). These include:

### Introductory Tutorials

- [Introduction to Galaxy Analyses](#)
- [Data Manipulation](#)
- [User Interface and Features](#)

### Scientific Analyses

- [Assembly](#)
- [Computational chemistry](#)
- [Ecology](#)
- [Epigenetics](#)
- [Genome Annotation](#)
- [Imaging](#)
- [Metabolomics](#)
- [Metagenomics](#)
- [Proteomics](#)
- [Sequence analysis](#)
- [Statistics and machine learning](#)
- [Transcriptomics](#)
- [Variant Analysis](#)



## Material

Search

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Introduction to metagenomics						
16S Microbial Analysis with mothur (extended)						
16S Microbial Analysis with mothur (short)						
Analyses of metagenomics data - The global picture						
Antibiotic resistance detection <a href="#">nanopore</a> <a href="#">plasmids</a>						
Metatranscriptomics analysis using microbiome RNA-seq data <a href="#">metatranscriptomics</a>						
Metatranscriptomics analysis using microbiome RNA-seq data (short) <a href="#">metatranscriptomics</a>						