

# Tutorial 12

## MetaPhlAn and HUMAnN



Zoe Sucato, Catherine Nguyen, Tobenna Oduah

# MetaPhlAn [5]

MetaPhlAn stands for Metagenomic  
Phylogenetic Analysis.

- Computational tool for profiling the composition of microbial communities from metagenomic shotgun sequencing data.
- Relies on unique clade-specific marker genes.
  - From ~17,000 reference genomes:
    - 13,500 bacterial and archeal genomes
    - 3,500 viral genomes
    - 110 eukaryotic genomes
- Uses bowtie2.

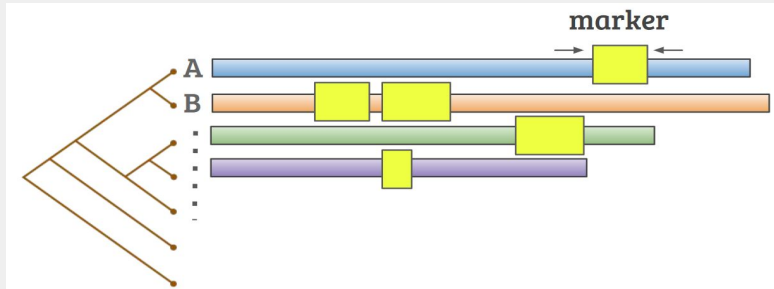
## Definitions

**Clade:** Group of organisms believed to have evolved from a common ancestor on a phylogenetic tree.

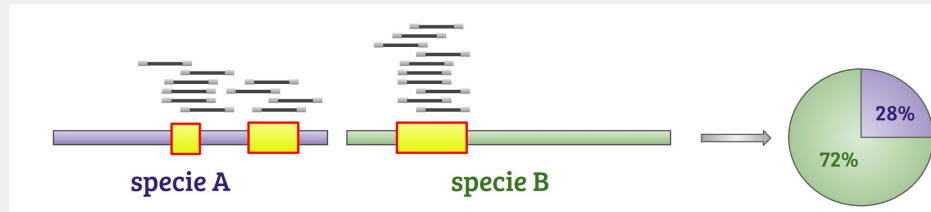
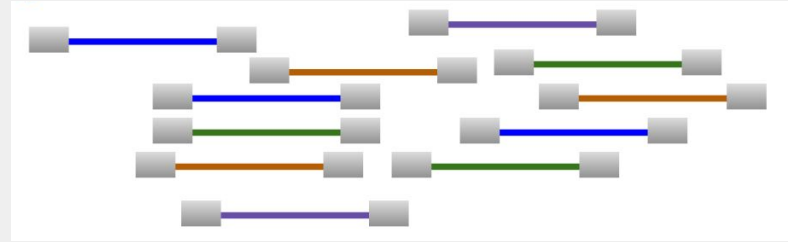
**Clade-specific marker:** Coding sequences that are strongly conserved within the clade's genomes and do not possess substantial local similarity with any sequence outside the clade.

# MetaPhlAn Algorithm [2]

Database of reference genomes and their relationships,  
with identified clade-specific markers



Sequenced sample



Reads mapped to marker genes

# HUMAnN [3]

HUMAnN stands for the HMP Unified Metabolic Analysis Network.

- A method for profiling the abundance of microbial metabolic pathways, including other molecular functions from metagenomic/metatranscriptomic sequencing data.

## Definitions

**HMP:** Human Microbiome Project. An initiative to research and understand the microbial components of the human genetic landscape and how that translates to health-related norms and concerns.

# HUMAnN Algorithm [6]



# MetaPhlAn

Which microbes are there?

# HUMAnN

What can the microbes do?

---

# Demo

# Original Plans... Changed

Originally meant to use Biobakery Workflows.

Could not use Docker directly, so we used Singularity to run a Docker image.

Many issues were encountered:

- Pre-existing Singularity Container did not have the tools (wmgz) installed.
- Creating a new Singularity container with the Docker image failed due to incorrect dependencies.
- Using pip install was not viable since many modules were installed and some failed to install.



# New Plan

Use Galaxy/Hutlab! [4]



All online; no package installs.

However, it has limitations.

Example: Doesn't make the output from one process nice for the next process

# Go to the Galaxy/Hutlab Website.

The screenshot shows a web browser window with the URL [huttenhower.sph.harvard.edu/galaxy/](https://huttenhower.sph.harvard.edu/galaxy/). The page header includes the "Galaxy / Hutlab" logo and navigation links: "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". A "Using 0%" indicator is in the top right.

The main content area is titled "Tools" and contains a search bar and a list of tool categories: Text Manipulation, NGS TOOLBOX BETA, and various specific tools like PICRUST, GraPhlAn, MetaPhlAn, etc. The central text area welcomes visitors and provides information about the lab's research interests and available resources. A profile for Francois-Xavier Bagnoud is shown on the right.

The right sidebar, titled "History", shows an empty "Unnamed history" section with a message: "This history is empty. You can load your own data or get data from an external source".

**Tools**

search tools

**Text Manipulation**

[PICRUST](#)

[GraPhlAn](#)

[MetaPhlAn](#)

[MetaPhlAn2](#)

[LEfSe](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[Get Genomic Scores](#)

[microPITA](#)

[Extract Features](#)

[Phenotype Association](#)

**NGS TOOLBOX BETA**

[NGS: QC and manipulation](#)

[Fetch Sequences](#)

[Fetch Alignments](#)

[Statistics](#)

[Graph/Display Data](#)

[MaAsLin](#)

Thanks for visiting our lab's tools and applications page, implemented within the [Galaxy](#) web application and workflow framework. Here, we provide a number of resources for metagenomic and functional genomic analyses, intended for research and academic use. Please see the menus and folders to the left for an overview of available tools including documentation, sample data, and publications.

Our lab's research interests include metagenomics and the [human microbiome](#), the relationships between microbial communities and human health, microbiome systems biology, and large-scale computational methods for studying all of these areas. In addition to the tools provided here, feel free to take a look at our additional [research](#) and [publications](#), including the [Sleipnir library](#) for computational functional genomics.

The tools are available here without account creation. However, you are strongly invited to create an account for having access to the history, saved analyses, datasets and workflows. You can create an account and/or log in using the User menu in the top-right corner.

If you have any comments, questions, or suggestions, please contact [Dr. Huttenhower](#).

**History**

search datasets

**Unnamed history**

0 b

*This history is empty. You can [load your own data](#) or [get data from an external source](#)*

<https://huttenhower.sph.harvard.edu/galaxy/>

# To upload your fasta file, click “load your own data”.

The screenshot shows the Galaxy web interface at [huttenhower.sph.harvard.edu/galaxy/](https://huttenhower.sph.harvard.edu/galaxy/). The interface includes a top navigation bar with 'Galaxy / Hutlab' and various menu items like 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A 'Using 0%' indicator is in the top right.

**Tools sidebar:** Contains a search bar and a list of tool categories including Text Manipulation (PICRUSt, GraPhlAn, MetaPhlAn, MetaPhlAn2, LEfSe, Filter and Sort, Join, Subtract and Group, Get Genomic Scores, microPITA, Extract Features, Phenotype Association), NGS TOOLBOX BETA (NGS: QC and manipulation, Fetch Sequences, Fetch Alignments, Statistics, Graph/Display Data, MaAsLin).

**Main content area:** Displays a welcome message from Francois-Xavier Bagnoud Building, Harvard School of Public Health. It explains that the Galaxy web application and workflow framework provide resources for metagenomic and functional genomic analyses. It also mentions research interests in metagenomics and the human microbiome, and provides links to research and publications, including the Sleipnir library.

**History panel:** Titled 'History', it shows 'Unnamed history' with '0 b' of data. A message states: 'This history is empty. You can **load your own data** or get data from an external source'. A red arrow points to the 'load your own data' link.

<https://huttenhower.sph.harvard.edu/galaxy/>

**Drag your fasta file into this upload box.**

The screenshot shows a web browser window with the URL `huttenhower.sph.harvard.edu/galaxy/`. A modal dialog titled "Download from web or upload from disk" is open. The dialog has two tabs: "Regular" (selected) and "Composite". Below the tabs is a large dashed rectangular area with the text "Drop files here" and a file icon. At the bottom of the dialog, there are two dropdown menus: "Type (set all):" with "Auto-detect" selected, and "Genome (set all):" with "unspecified (?)" selected. Below these are several buttons: "Choose local file", "Paste/Fetch data", "Pause", "Reset", "Start", and "Close". The background shows the Galaxy interface with a sidebar containing tool categories like "Text Manipulation", "Filter and Sort", and "NGS TOOLBOX BETA".

<https://huttenhower.sph.harvard.edu/galaxy/>

# Click start to upload.

Galaxy

huttenhower.sph.harvard.edu/galaxy/

## Download from web or upload from disk

Regular Composite

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

Name	Size	Type	Genome	Settings	Status
evol1.sorted.unmappe d.R1.fastq.gz	579.4 KB	Auto-dete...	unspecified (?)		

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

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

<https://huttenhower.sph.harvard.edu/galaxy/>

# Wait until status is 100%.

Galaxy huttenhower.sph.harvard.edu/galaxy/

## Download from web or upload from disk

Regular Composite

Name	Size	Type	Genome	Settings	Status
evol1.sorted.unmapped.R1.fastq.gz	579.4 KB	Auto-detect	unspecified (?)		100%

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

Choose local file Paste/Fetch data Pause Reset Start Close

<https://huttenhower.sph.harvard.edu/galaxy/>



# The uploaded file should appear in the right panel.

The screenshot shows the Galaxy web interface at <https://huttenhower.sph.harvard.edu/galaxy/>. The interface is divided into three main panels:

- Left Panel (Tools):** Contains a search bar and a list of tool categories including Text Manipulation, NGS TOOLBOX BETA, and various analysis tools like PICRUSt, GraPhlAn, and metaPhlAn.
- Center Panel:** Displays a welcome message from Francois-Xavier Bagnoud Building, Harvard School of Public Health. It explains the purpose of the Galaxy web application and workflow framework, providing resources for metagenomic and functional genomic analyses. It also mentions the lab's research interests in metagenomics and the human microbiome.
- Right Panel (History):** Shows a list of datasets. The first entry is "1: evol1.sorted.unmap ped.R1.fastq", which is highlighted in green, indicating it is the current dataset. The history panel also shows a search bar and a "Using 0%" status.

# Navigate to the MetaPhlAn2 profile from the left panel.

The screenshot shows the Galaxy web interface at [huttenhower.sph.harvard.edu/galaxy/](https://huttenhower.sph.harvard.edu/galaxy/). The left panel contains a 'Tools' section with a search bar and a list of tool categories. Under 'Text Manipulation', 'MetaPhlAn2' is highlighted with a red arrow and labeled 'metagenomic profiler V2'. The main panel displays the 'MetaPhlAn2 metagenomic profiler V2 (Galaxy Version 2.0.0)' tool configuration. The 'Input metagenome' is set to '1: evol1.sorted.unmapped.R1.fastq'. Sensitivity options are set to 'Very Sensitive'. Display Post Mapping Advanced Parameters, Display additional analysis types and arguments advanced parameters, and Display additional biom advanced parameters are all set to 'No'. An 'Execute' button is visible. The right panel shows the 'History' section with a search bar and a list of datasets. The dataset '1: evol1.sorted.unmap ped.R1.fastq' is highlighted in green.

Galaxy / Hutlab

Analyze Data Workflow Shared Data Visualization Help User

Using 0%

Tools

search tools

Text Manipulation

PICRUSt

GraPhlAn

MetaPhlAn

MetaPhlAn2

MetaPhlAn2 metagenomic profiler V2

LEfSe

Filter and Sort

Join, Subtract and Group

Get Genomic Scores

microPITA

Extract Features

Phenotype Association

NGS TOOLBOX BETA

NGS: QC and manipulation

Fetch Sequences

Fetch Alignments

Statistics

MetaPhlAn2 metagenomic profiler V2 (Galaxy Version 2.0.0)

Options

Input metagenome (fastq of metagenomic reads, loaded with the Get Data module )

1: evol1.sorted.unmapped.R1.fastq

Sensitivity options for read-marker similarity (as described by BowTie2)

Very Sensitive

Display Post Mapping Advanced Parameters

No

Select Post Mapping advanced choices

Display additional analysis types and arguments advanced parameters

No

Select additional analysis types and argument advanced choices

Display additional biom advanced parameters

No

Select additional biom choices

Execute

MetaPhlAn is a computational tool for profiling the composition of microbial communities (*Bacteria*, *Archaea*, *Eukaryotes* and *Viruses*) from metagenomic shotgun sequencing data with species level resolution.

From version 2.0 MetaPhlAn is also able to identify specific strains (in the not-so-frequent cases in which the sample contains a previously sequenced strains) and to track strains across samples for all species.

History

search datasets

Unnamed history

1 shown

1.68 MB

1: evol1.sorted.unmap ped.R1.fastq

<https://huttenhower.sph.harvard.edu/galaxy/>



# Make sure the correct input is set.

Galaxy / Hutlab

Analyze Data Workflow Shared Data Visualization Help User

Using 0%

Tools

search tools

Text Manipulation

PICRUSt

GraPhlAn

MetaPhlAn

MetaPhlAn2

MetaPhlAn2 metagenomic profiler V2

LEfSe

Filter and Sort

Join, Subtract and Group

Get Genomic Scores

microPITA

Extract Features

Phenotype Association

NGS TOOLBOX BETA

NGS: QC and manipulation

Fetch Sequences

Fetch Alignments

Statistics

MetaPhlAn2 metagenomic profiler V2 (Galaxy Version 2.0.0)

Options

Input metagenome (fastq of metagenomic reads, loaded with the Get Data module )

1: evol1.sorted.unmapped.R1.fastq

Sensitivity options for read-marker similarity (as described by BowTie2)

Very Sensitive

Display Post Mapping Advanced Parameters

No

Select Post Mapping advanced choices

Display additional analysis types and arguments advanced parameters

No

Select additional analysis types and argument advanced choices

Display additional biom advanced parameters

No

Select additional biom choices

Execute

MetaPhlAn is a computational tool for profiling the composition of microbial communities (*Bacteria*, *Archaea*, *Eukaryotes* and *Viruses*) from metagenomic shotgun sequencing data with species level resolution.

From version 2.0 MetaPhlAn is also able to identify specific strains (in the not-so-frequent cases in which the sample contains a previously sequenced strains) and to track strains across samples for all species.

History

search datasets

Unnamed history

1 shown

1.68 MB

1: evol1.sorted.unmap  
ped.R1.fastq

<https://huttenhower.sph.harvard.edu/galaxy/>

# Keep default settings and click "Execute".

The screenshot shows the Galaxy web interface at [huttenhower.sph.harvard.edu/galaxy/](https://huttenhower.sph.harvard.edu/galaxy/). The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. The left sidebar lists various tool categories, including Text Manipulation, NGS TOOLBOX BETA, and NGS: QC and manipulation. The central panel displays the configuration for the MetaPhlAn2 metagenomic profiler V2 (Galaxy Version 2.0.0). The configuration includes input metagenome (fastq of metagenomic reads, loaded with the Get Data module), sensitivity options for read-marker similarity (as described by BowTie2), and display post mapping advanced parameters. The input field shows '1: evol1.sorted.unmapped.R1.fastq'. The sensitivity options are set to 'Very Sensitive'. The display post mapping advanced parameters are set to 'No'. The display additional analysis types and arguments advanced parameters are set to 'No'. The display additional biom advanced parameters are set to 'No'. A red arrow points to the 'Execute' button. The right sidebar shows the History panel with a search datasets field and a list of datasets, including '2: MetaPhlAn2 on d ata 1' and '1: evol1.sorted.unmap ped.R1.fastq'.

Galaxy / Hutlab

Analyze Data Workflow Shared Data Visualization Help User

Using 0%

Tools

search tools

Text Manipulation

PICRUSt

GraPhlAn

MetaPhlAn

MetaPhlAn2

MetaPhlAn2 metagenomic profiler V2

LEfSe

Filter and Sort

Join, Subtract and Group

Get Genomic Scores

microPITA

Extract Features

Phenotype Association

NGS TOOLBOX BETA

NGS: QC and manipulation

Fetch Sequences

Fetch Alignments

Statistics

MetaPhlAn2 metagenomic profiler V2 (Galaxy Version 2.0.0)

Options

Input metagenome (fastq of metagenomic reads, loaded with the Get Data module )

1: evol1.sorted.unmapped.R1.fastq

Sensitivity options for read-marker similarity (as described by BowTie2)

Very Sensitive

Display Post Mapping Advanced Parameters

No

Select Post Mapping advanced choices

Display additional analysis types and arguments advanced parameters

No

Select additional analysis types and argument advanced choices

Display additional biom advanced parameters

No

Select additional biom choices

✓ Execute

MetaPhlAn is a computational tool for profiling the composition of microbial communities (*Bacteria*, *Archaea*, *Eukaryotes* and *Viruses*) from metagenomic shotgun sequencing data with species level resolution.

From version 2.0 MetaPhlAn is also able to identify specific strains (in the not-so-frequent cases in which the sample contains a previously sequenced strains) and to track strains across samples for all species.

History

search datasets

Unnamed history

2 shown

1.68 MB

2: MetaPhlAn2 on d ata 1

1: evol1.sorted.unmap ped.R1.fastq

<https://huttenhower.sph.harvard.edu/galaxy/>

# Successful execution! Job is now added to the queue.

The screenshot shows the Galaxy web interface at [huttenhower.sph.harvard.edu/galaxy/](https://huttenhower.sph.harvard.edu/galaxy/). The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. A status bar on the right indicates "Using 0%".

On the left, the "Tools" panel is visible, with a search bar and a list of tool categories including Text Manipulation, NGS TOOLBOX BETA, and NGS: QC and manipulation.

The central area displays a green success message: "1 job has been successfully added to the queue - resulting in the following datasets: 2: MetaPhlAn2 on data 1". Below this message, it states: "You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered."

On the right, the "History" panel shows a list of datasets. The top entry is "2: MetaPhlAn2 on data 1" with a size of 1.68 MB. Below it, a new entry "1: evol1.sorted.unmap ped.R1.fastq" is highlighted in green.

# Result will appear green when it is completed.

The screenshot shows the Galaxy web interface at [huttenhower.sph.harvard.edu/galaxy/](https://huttenhower.sph.harvard.edu/galaxy/). The main panel displays the configuration for the **MetaPhlAn2 metagenomic profiler V2 (Galaxy Version 2.0.0)** tool. The input is set to `1: evol1.sorted.unmapped.R1.fastq`. Sensitivity options are set to **Very Sensitive**. Display Post Mapping Advanced Parameters, Display additional analysis types and arguments advanced parameters, and Display additional biom advanced parameters are all set to **No**. An **Execute** button is visible at the bottom of the tool configuration panel.

The right-hand **History** panel shows the execution history. The entry **2: MetaPhlAn2 on data** is highlighted in green, indicating it is the current dataset. Below it, the entry **1: evol1.sorted.unmap ped.R1.fastq** is also highlighted in green. A red arrow points to the green highlight on the second entry.

The left-hand **Tools** panel shows a search bar and a list of tool categories: **Text Manipulation**, **NGS TOOLBOX BETA**, and **NGS: QC and manipulation**. Under **Text Manipulation**, the **MetaPhlAn2** tool is listed. Under **NGS TOOLBOX BETA**, the **MetaPhlAn2** tool is listed. Under **NGS: QC and manipulation**, the **MetaPhlAn2** tool is listed.

<https://huttenhower.sph.harvard.edu/galaxy/>



# Click the eye icon to download and view result.

The screenshot displays the Galaxy web interface at [huttenhower.sph.harvard.edu/galaxy/](https://huttenhower.sph.harvard.edu/galaxy/). The main panel shows the configuration for the **MetaPhlAn2 metagenomic profiler V2 (Galaxy Version 2.0.0)** tool. The input is set to `1: evol1.sorted.unmapped.R1.fastq`. Sensitivity options are set to **Very Sensitive**. Display Post Mapping Advanced Parameters, Display additional analysis types and arguments advanced parameters, and Display additional biom advanced parameters are all set to **No**. An **Execute** button is visible at the bottom of the tool configuration panel.

The left sidebar contains a **Tools** section with a search bar and a list of tool categories: **Text Manipulation**, **PICRUSt**, **GraPhlAn**, **MetaPhlAn**, **MetaPhlAn2**, **LEfSe**, **Filter and Sort**, **Join, Subtract and Group**, **Get Genomic Scores**, **microPITA**, **Extract Features**, **Phenotype Association**, **NGS TOOLBOX BETA**, **NGS: QC and manipulation**, **Fetch Sequences**, **Fetch Alignments**, and **Statistics**.

The right sidebar shows the **History** section with a search bar and a list of datasets. The **Unnamed history** section shows 2 datasets. The first dataset is **1: evol1.sorted.unmap ped.R1.fastq**. The second dataset is **2: MetaPhlAn2 on data 1**, which is highlighted in green. A red arrow points to the **eye icon** next to the second dataset, indicating that clicking it will download and view the result.

<https://huttenhower.sph.harvard.edu/galaxy/>

# Click the eye icon to download and view result.

The screenshot displays the Galaxy web interface at [huttenhower.sph.harvard.edu/galaxy/](https://huttenhower.sph.harvard.edu/galaxy/). The main workspace shows the **MetaPhlAn2** metagenomic profiler V2 tool configuration. The input is set to `1: evol1.sorted.unmapped.R1.fastq`. Sensitivity options are set to **Very Sensitive**. Display Post Mapping Advanced Parameters, Display additional analysis types and arguments advanced parameters, and Display additional biom advanced parameters are all set to **No**. An **Execute** button is visible at the bottom of the tool configuration panel.

The left sidebar contains a **Tools** panel with a search bar and a list of tool categories: **Text Manipulation**, **PICRUSt**, **GraPhlAn**, **MetaPhlAn**, **MetaPhlAn2**, **LEfSe**, **Filter and Sort**, **Join, Subtract and Group**, **Get Genomic Scores**, **microPITA**, **Extract Features**, **Phenotype Association**, **NGS TOOLBOX BETA**, **NGS: QC and manipulation**, **Fetch Sequences**, **Fetch Alignments**, and **Statistics**.

The right sidebar shows the **History** panel with a search bar and a list of datasets. The **Unnamed history** section shows 2 datasets. The first dataset is **1: evol1.sorted.unmap ped.R1.fastq**. The second dataset is **2: MetaPhlAn2 on data 1**, which is highlighted in green. A red arrow points to the **eye icon** next to this dataset, indicating that clicking it will download and view the result.

<https://huttenhower.sph.harvard.edu/galaxy/>

# Metaphlan Output Snippet

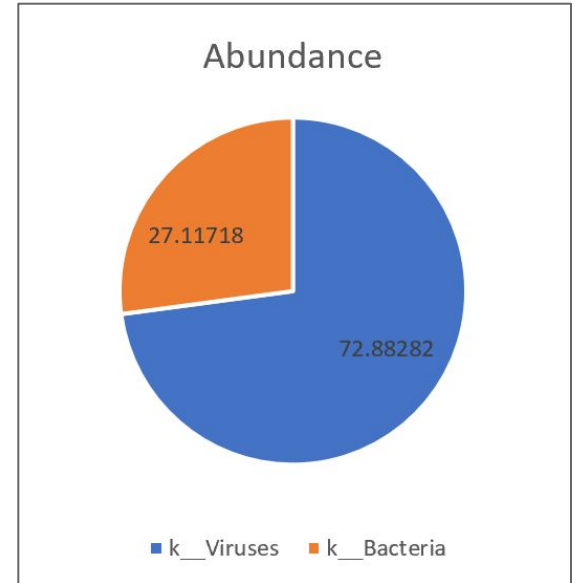
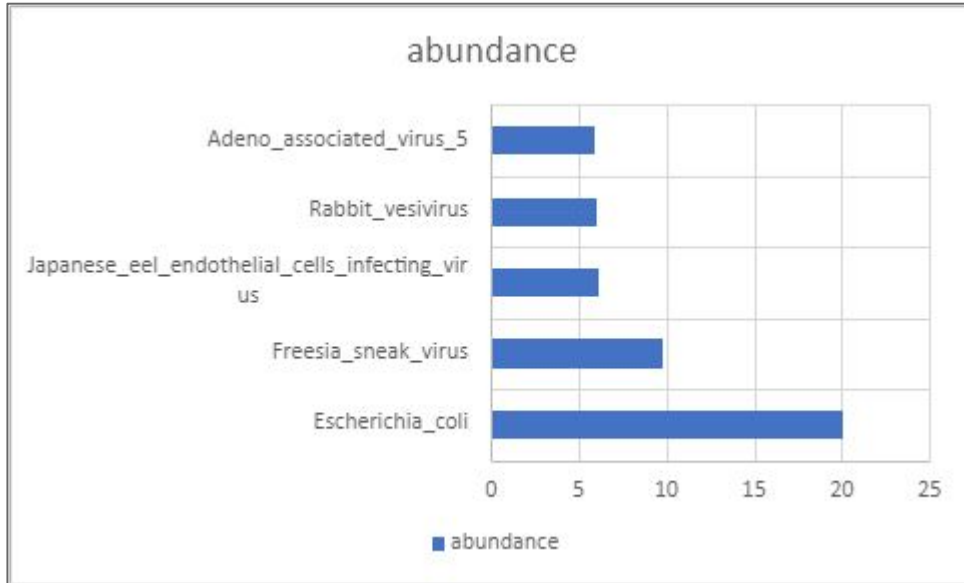


Galaxy3-MetaPhlan2\_on\_data\_2.metaphlan - Notepad

File Edit View

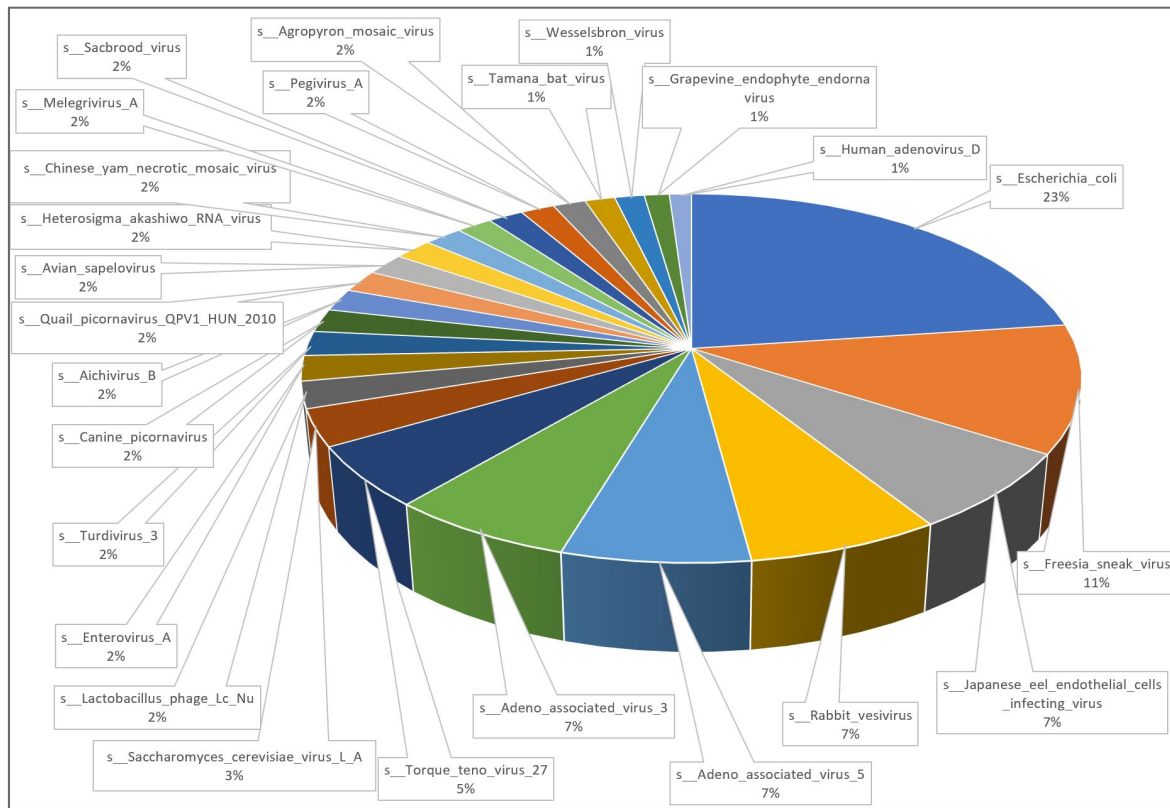
```
#SampleID    Metaphlan2_Analysis
k__Viruses   72.88282
k__Bacteria  27.11718
k__Viruses|p__Viruses_noname  72.88282
k__Bacteria|p__Proteobacteria  24.16031
k__Bacteria|p__Bacteroidetes  2.95687
k__Viruses|p__Viruses_noname|c__Viruses_noname  72.88282
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria  20.08052
k__Bacteria|p__Proteobacteria|c__Alphaproteobacteria  4.0798
k__Bacteria|p__Bacteroidetes|c__Flavobacteriia  2.95687
k__Viruses|p__Viruses_noname|c__Viruses_noname|o__Viruses_noname  54.15491
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacteriales  20.08052
```

# Visualized Output





# Species Detected



# References

- [1] “biobakery/biobakery\_workflows.” n.d. GitHub. Accessed May 20, 2022.  
[https://github.com/biobakery/biobakery\\_workflows](https://github.com/biobakery/biobakery_workflows).
- [2] Borenstein Lab. n.d. “MetaPhlAn.” Accessed May 20, 2022. PowerPoint.  
[http://borensteinlab.com/courses/TAU\\_CS\\_3116\\_B\\_19/presentations/7\\_MetaPhlan.pdf](http://borensteinlab.com/courses/TAU_CS_3116_B_19/presentations/7_MetaPhlan.pdf).
- [3] The Huttenhower Lab. n.d. “biobakeryWorkflows.” The Huttenhower Lab. Accessed May 20, 2022.  
[https://huttenhower.sph.harvard.edu/biobakery\\_workflows/](https://huttenhower.sph.harvard.edu/biobakery_workflows/).
- [4] The Huttenhower Lab. n.d. “Galaxy / Hutlab.” Accessed May 20, 2022.  
<https://huttenhower.sph.harvard.edu/galaxy/>.
- [5] The Huttenhower Lab. n.d. “MetaPhlAn2.” The Huttenhower Lab. Accessed May 20, 2022.  
<https://huttenhower.sph.harvard.edu/metaphlan2/>.
- [6] Mehta, Subina, Pratik Jagtap, and Saskia Hiltmann. 2021. “Introduction to metatranscriptomics.” Galaxy Training!  
<https://training.galaxyproject.org/archive/2021-10-01/topics/metagenomics/tutorials/metatranscriptomics/slides-plain.html>.

# Questions?