Technical Report: GTN Tutorial - Taxonomic Profiling and Visualization of Metagenomic Data

Desi Arti

NPM 2206130706

Author Affiliations
Department Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia

Abstrak

Metagenomics is a rapidly evolving field that explores the genetic content of microbial communities directly from environmental samples. Understanding the taxonomic composition of these communities is crucial for unraveling their ecological roles and potential impacts on various ecosystems. This technical report provides an overview and documentation of the Galaxy Training Network (GTN) tutorial on Taxonomic Profiling and Visualization of Metagenomic Data.

1. Introduction

GTN Tutorial - Taxonomic Profiling and Visualization of Metagenomic Data" serves as a comprehensive guide to utilizing the Galaxy Training Network (GTN) for the analysis and visualization of metagenomic data. Metagenomics, a field integral to microbial ecology and environmental studies, demands sophisticated tools and methodologies for insightful taxonomic profiling.

This tutorial begins by introducing users to the Galaxy platform, emphasizing its user-friendly interface and the integration of diverse tools. It then delves into the specifics of metagenomic data preprocessing, showcasing techniques for quality control, filtering, and format conversion. The report navigates users through the intricacies of taxonomic profiling, employing widely-used tools such as Kraken and Centrifuge, elucidating their respective strengths and applications.

A focal point of the tutorial lies in the visualization of taxonomic profiles, crucial for interpreting complex metagenomic datasets. Users are guided through the creation of interactive and publication-ready visualizations using tools like Krona and Pavian, facilitating a deeper understanding of microbial community structures.

Moreover, the tutorial addresses the integration of statistical analyses, empowering users to discern significant taxonomic differences within their datasets. Emphasis is placed on guiding users through practical examples, ensuring a hands-on and applicable learning experience.

The Technical Report concludes by underlining the importance of taxonomic profiling in metagenomics and the role of GTN in providing an accessible and powerful platform for researchers and practitioners. The step-by-step nature of the tutorial encourages users, irrespective of their expertise level, to harness the potential of GTN for taxonomic profiling and visualization in metagenomic research.The GTN Tutorial encapsulates a valuable resource for bioinformaticians, researchers, and students seeking proficiency in metagenomic data analysis, offering a structured and informative pathway through the intricate landscape of taxonomic profiling and visualization.

2. About Galaxy

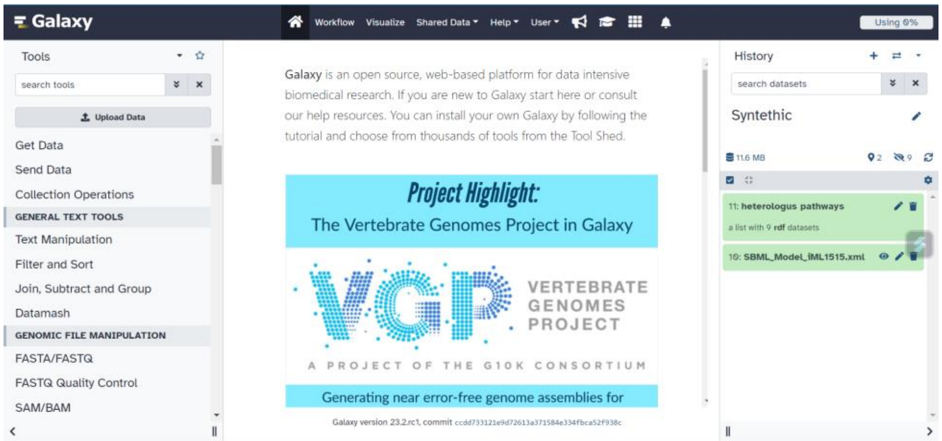
Galaxy is an open-source, web-based platform for bioinformatics analysis. It provides a user-friendly interface for running a wide variety of bioinformatics tools without needing to write any code. This makes it a valuable resource for researchers of all levels, from beginners to experienced bioinformaticians.

In these tutorials, we will cover the basics of using Galaxy for bioinformatics analysis. We will start with an overview of the Galaxy platform and then walk you through some simple bioinformatics tasks, such as sequence alignment and gene expression analysis.

These tutorials are designed for use with a public Galaxy instance, such as UseGalaxy.org: https://usegalaxy.org/. If you are using a local Galaxy instance, the interface may be slightly different, but the overall steps will be the same.

2.1 Introduction to Galaxy

Galaxy is a web-based platform that allows you to run bioinformatics tools without needing to write any code. It uses a drag-and-drop interface to make it easy to build workflows of analysis steps.



Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Galaxy version 23.2.rc1, commit ccdd733121e9d72613a371584e334fbca52f938c

Figure 1 .Galaxy platform interface

The Galaxy interface is divided into three main sections:

- Tools: This panel contains a list of bioinformatics tools that you can use.
- Data: This panel shows the data that you have uploaded to Galaxy.
- History: This panel shows the history of your analyses in Galaxy.

2.2 Uploading Data

The first step in any Galaxy analysis is to upload your data. Galaxy supports a variety of data formats, including FASTA, FASTQ, BAM, and BED.To upload data, click on the "Upload" button in the Data panel and select your data files.

2.3. Running a Tool

Once you have uploaded your data, you can start running bioinformatics tools. To do this, find the tool you want to use in the Tools panel and drag it onto the canvas in the History panel.

Once the tool is on the canvas, you need to configure its settings. The settings for each tool will vary, but they will typically include things like the input data, the output data, and the analysis parameters.

Once you have configured the tool settings, click on the "Run" button. The tool will then be executed and the results will be displayed in the History panel.

2.4. Visualizing Results

Galaxy provides a variety of ways to visualize your analysis results. You can view the results in tables, graphs, and other formats.

To visualize your results, click on the output data in the History panel. This will open the results in a new window.

2.5. Building Workflows

Galaxy allows you to build workflows of multiple analysis steps. This is useful for complex analyses that require running multiple tools in a specific order.

To build a workflow, simply drag and drop tools onto the canvas in the History panel and connect them with arrows. Once your workflow is complete, you can run it by clicking on the "Run" button.

Galaxy Tutorials:

The Galaxy Training Network (GTN) provides a variety of tutorials on how to use Galaxy for bioinformatics analysis. These tutorials cover a wide range of topics, from basic tasks to more advanced analyses.

Here are a few links to some helpful Galaxy tutorials:

- Introduction to Galaxy Analyses: https://training.galaxyproject.org/
- Transcriptomics / Tutorial List: https://training.galaxyproject.org/training-material/topics/transcriptomics/
- How to use Galaxy for Bioinformatics (Beginners): https://www.youtube.com/watch?v=uvlayo3kCgg

2.6 Taxonomic profiling

Taxonomic profiling, in the realm of metagenomics, is like conducting a census of the microscopic inhabitants within a sample, like your skin, gut, or even soil. It's an essential technique for identifying and quantifying the various types of microorganisms – largely bacteria and archaea – present in a certain environment.

Tools for taxonomic profiling can be divided into three groups. Nevertheless, all of them require a pre-computed database based on previously sequenced microbial DNA or protein sequences.

1. **DNA-to-DNA**: comparison of sequencing reads with genomic databases of DNA sequences with tools like Kraken (Wood and Salzberg 2014)

2. **DNA-to-Protein** : comparison of sequencing reads with protein databases (more computationally intensive because all six frames of potential DNA-to amino acid translations need to be analyzed) with tools like DIAMOND)

3. **Marker based**: searching for marker genes (e.g. 16S rRNA sequence) in reads, which is quick, but introduces bias, with tools like MetaPhlAn (Blanco-Míguez Aitor *et al.* 2023)

3. Overview of GTN Tutorial for Taxonomic Profiling and Visualization of Metagenomic Data

The tutorial aims to guide users through the process of taxonomic profiling using metagenomic data and visualization of the results. It covers key steps, tools, and best practices for a comprehensive analysis. This tutorial is designed for bioinformaticians, researchers, and students who want to gain hands-on experience in metagenomic data analysis.Familiarity with Galaxy platform basics and a basic understanding of metagenomics concepts are recommended. Below tutorial Content for Taxonomi Profiling and Visualization of Metagenomic

3.1 Data Preparation

- Introduction to metagenomic data format (FASTQ/FASTA).
- Quality control and preprocessing using tools like FastQC and Trimmomatic.

3.2 Taxonomic Profiling

- Utilizing tools like Kraken or Metaphlan for taxonomic classification.
- Adjustment of parameters for optimal performance.
- Interpretation of taxonomic profiles generated.

3.3 Visualization

- Using tools such as Krona or Pavian to create interactive and informative visualizations.
- Interpretation of visualization results for insights into microbial community composition.

3.4 Comparative Analysis

- Comparing taxonomic profiles across different samples.
- Identifying significant differences in microbial abundance.

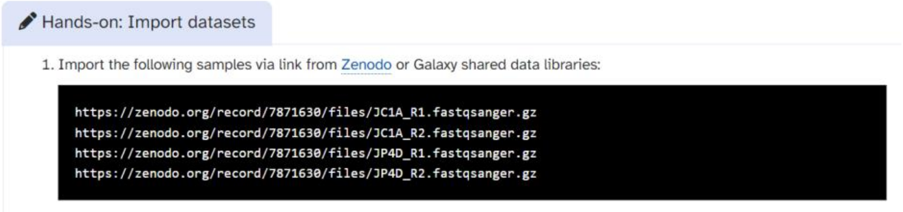
3.5. Galaxy Workflows

The tutorial includes pre-built Galaxy workflows to streamline the analysis process. These workflows encapsulate the entire analysis pipeline, making it easy for users to replicate and customize the analysis for their datasets.

3.6. Technical Documentation

This documentation is invaluable for users who wish to adapt the tutorial for their specific datasets or explore advanced analyses.

Prepare Galaxy and Data

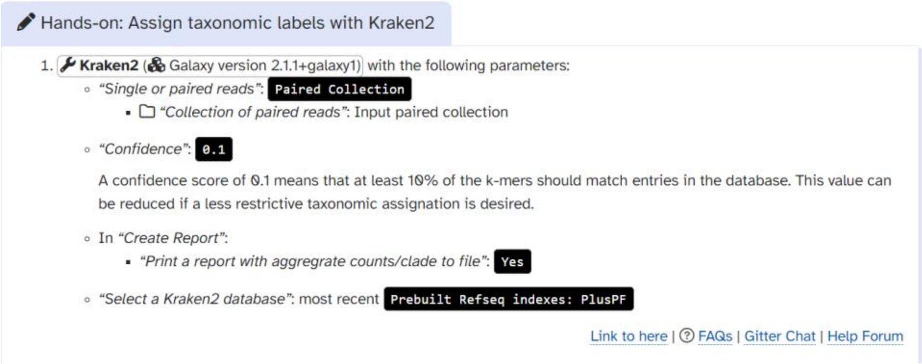


Hands-on: Import datasets

1. Import the following samples via link from Zenodo or Galaxy shared data libraries:

```
https://zenodo.org/record/7871630/files/JC1A_R1.fastqsanger.gz
https://zenodo.org/record/7871630/files/JC1A_R2.fastqsanger.gz
https://zenodo.org/record/7871630/files/JP4D_R1.fastqsanger.gz
https://zenodo.org/record/7871630/files/JP4D_R2.fastqsanger.gz
```

To identify the microorganisms present, we will utilize Kraken2 (Wood et al. 2019) to compare the sample reads against a reference database, consisting of sequences from know microorganisms stored in the database.



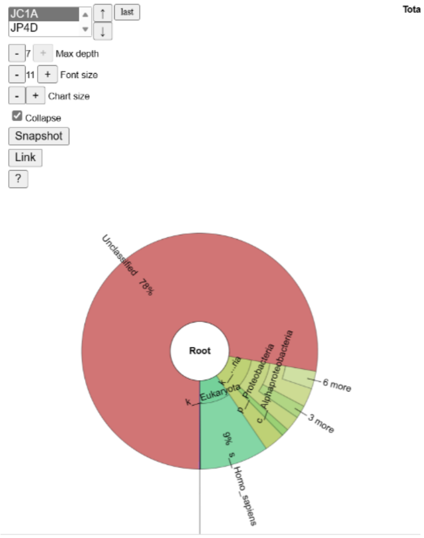
Hands-on: Assign taxonomic labels with Kraken2

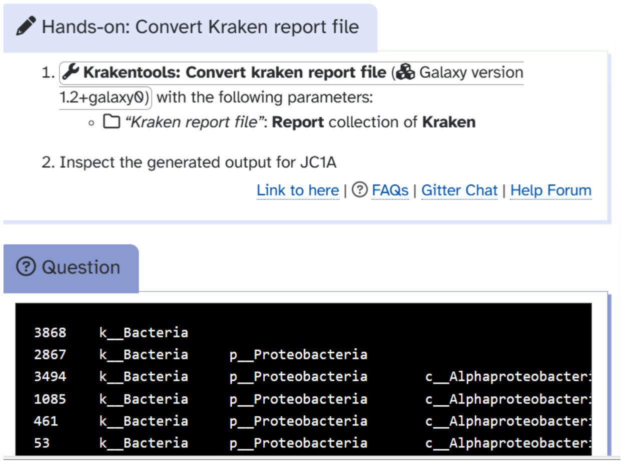
1. **Kraken2** (Galaxy version 2.1.1+galaxy1) with the following parameters:
   - *"Single or paired reads"*: `Paired Collection`
     - *"Collection of paired reads"*: Input paired collection
   - *"Confidence"*: `0.1`

     A confidence score of 0.1 means that at least 10% of the k-mers should match entries in the database. This value can be reduced if a less restrictive taxonomic assignation is desired.
   - In *"Create Report"*:
     - *"Print a report with aggregrate counts/clade to file"*: `Yes`
   - *"Select a Kraken2 database"*: most recent `Prebuilt Refseq indexes: PlusPF`

Visualisation using Krona

Krona generates an interactive HTML file that facilitates the exploration of hierarchical data through zoomable, multi-layered pie charts. This tool enables effortless visualization of bacterial community compositions and facilitates comparisons of how microorganism populations are altered based on environmental conditions.





Hands-on: Convert Kraken report file

1. **Krakentools: Convert kraken report file** (Galaxy version 1.2+galaxy0) with the following parameters:
   - *"Kraken report file"*: **Report** collection of **Kraken**

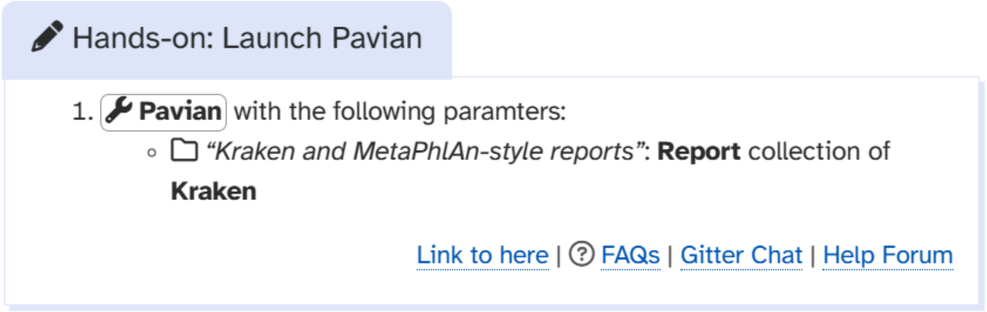
2. Inspect the generated output for JC1A

Link to here | FAQs | Gitter Chat | Help Forum

Question

```
3868    k__Bacteria
2867    k__Bacteria     p__Proteobacteria
3494    k__Bacteria     p__Proteobacteria       c__Alphaproteobacter:
1085    k__Bacteria     p__Proteobacteria       c__Alphaproteobacter:
461     k__Bacteria     p__Proteobacteria       c__Alphaproteobacter:
53      k__Bacteria     p__Proteobacteria       c__Alphaproteobacter:
```

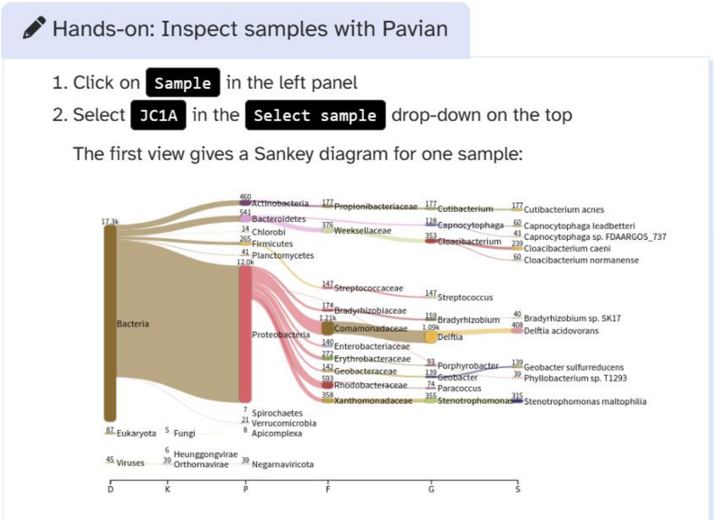
Pavian (pathogen visualization and more) (Breitwieser and Salzberg 2020) is an interactive visualization tool for metagenomic data. It was developed for the clinical metagenomic problem to find a disease-causing pathogen in a patient sample, but it is useful to analyze and visualize any kind of metagenomics data.



Hands-on: Launch Pavian

1. **Pavian** with the following paramters:
   - *"Kraken and MetaPhlAn-style reports"*: **Report** collection of **Kraken**

Link to here | FAQs | Gitter Chat | Help Forum



Hands-on: Inspect samples with Pavian

1. Click on `Sample` in the left panel
2. Select `JC1A` in the `Select sample` drop-down on the top

The first view gives a Sankey diagram for one sample:

6. Conclusion

The GTN tutorial on Taxonomic Profiling and Visualization of Metagenomic Data equips users with essential skills for analyzing metagenomic datasets. By combining the power of Galaxy's user-friendly interface with robust bioinformatics tools, users gain a practical understanding of taxonomic profiling and visualization techniques.

8. References.

1. Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology, 15(3), R46. DOI: 10.1186/gb-2014-15-3-r46
2. Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., ... & Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods, 12(10), 902-903. DOI: 10.1038/nmeth.3589
3. Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
4. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120. DOI: 10.1093/bioinformatics/btu170
5. Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. BMC Bioinformatics, 12(1), 385. DOI: 10.1186/1471-2105-12-385
6. Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2017). A review of methods and databases for metagenomic classification and assembly. Briefings in Bioinformatics, 20(4), 1125-1136. DOI: 10.1093/bib/bbx120
7. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550. DOI: 10.1186/s13059-014-0550-8