

Cloud Computing and its Application in Microbiome Data Analysis



PhD student: Felix Wong
Supervisor: Prof. Paul KS Chan
Department of Microbiology
Faculty of Medicine
Chinese University of Hong Kong
5 December 2017



Outline

1. Introduction

- Introducing cloud computing
- Major public cloud computing service providers
- Advantages and disadvantages of using cloud computing for bioinformatics analysis

2. Galaxy application for microbiome data analysis

- Introduction to Galaxy server
- Private Galaxy server in the cloud
- Galaxy workflow for microbiome data analysis

Introducing cloud computing



OneDrive



Google Drive

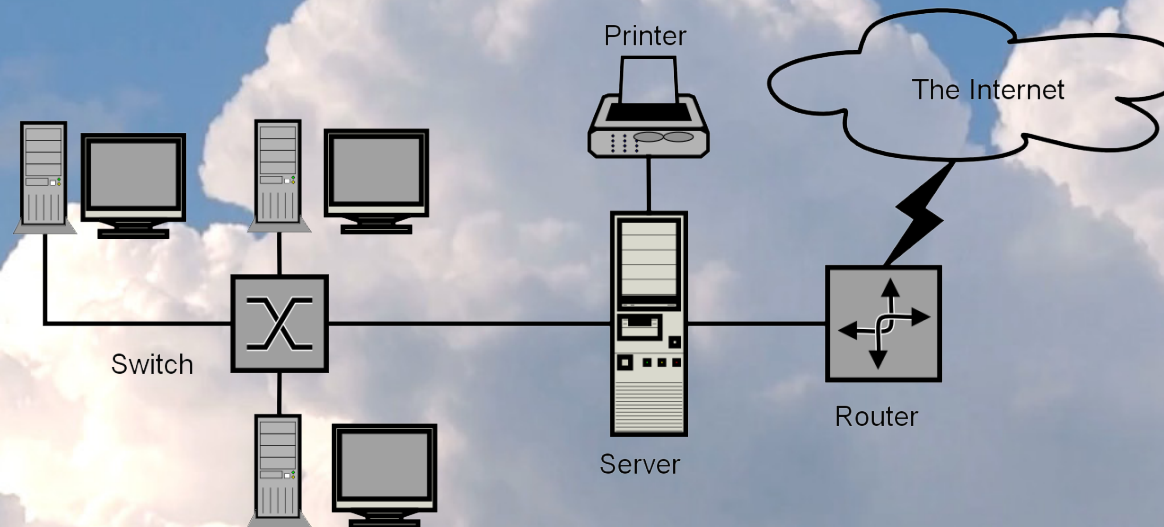
Introducing cloud computing



Introducing cloud computing

- What is it?

Delivery of on-demand computing resources over the internet



Source: Wikipedia

Introducing cloud computing

- Infrastructure as a Service (IaaS)
 - Provide computing infrastructure such as processing, storage and network
- Platform as a Service (PaaS)
 - Provide a platform such as operating system, database and web server
- Software as a Service (SaaS)
 - Provide application software

Major cloud computing service providers

aws



Google Cloud Platform

Microsoft Azure

Major cloud computing service providers

- Charge for usage
 1. Computing - bill by seconds used
 2. Storage - bill by GB/month provisioned
- Offer free account for 12 months
- US\$300 credits for Google Cloud Platform free account
- US\$100 credits for students at AWS educate member institution (annually renewable)

aws



Google Cloud Platform

Microsoft Azure

aws  educate

Advantages and disadvantages of using cloud computing for bioinformatics analysis

Advantages

- Cost saving
- Maintenance
- On-demand scaling
- Public datasets availability

Disadvantages

- Cost
- Downtime
- Data privacy and security
- Legal issues

Privacy and Legal Issues in Cloud Computing. Anne S.Y. Cheung and Rolf H. Weber

Introduction to Galaxy



- Galaxy is an open, web-based platform for **accessible**, **reproducible**, and **transparent** computational biomedical research.

Source: <https://galaxyproject.org>

Introduction to Galaxy



Accessible

- a web-based platform that provide simplified interface of many popular bioinformatics tools

Reproducible

- a feature called Workflow that let user to create reusable analysis pipeline

Transparent

- users can share data and analysis on the same platform

Introduction to Galaxy



The screenshot displays the Galaxy web interface. The browser address bar shows the URL: `https://usegalaxy.org/?tool_id=toolshed.g2.bx.psu.edu/repos/f2fluc/f2fluc/mothur_summary_qual%2Fmothur_summary_qual%2F1.36.1.0&version=1.36.1.0&_identifier=sckfmpbf2e9`. The main content area is titled "Summary.qual Summarize the quality scores (Galaxy Version 1.36.1.0)". It includes a "qfile - Sequence Quality file" dropdown menu with the value "86: Make.contigs on data 80, data 79, and others: scrap.contigs.qual". Below this are fields for "name - Names" and "count - a count_table", both currently set to "Nothing selected". An "Execute" button is visible at the bottom of the configuration area.

On the left side, there is a "Tools" sidebar with a search bar and a list of tools, including "NCS: Mothur", "Venn", "Unique_seqs", "unifrac_weighted", "unifrac_unweighted", "Trim_seqs", "Trim_files", "Tree_shared", "Summary_tax", "Summary_single", "Summary_shared", "Summary_seqs", "Summary_qual", "Sub_sample", "Split_otus", "Split_abund", "Sort_seqs", and "Shhh_seqs".

The right side of the interface shows a "History" panel with a search bar and a list of recent jobs. The jobs listed include "Mothur Tutorial", "88 and data 83: logfile", "90: Summary_seqs on data 83: summary", "89: Summary_seqs on data 83: logfile", "88: Make.contigs on data 80, data 79, and others: group file", "87: Make.contigs on data 80, data 79, and others: report", "86: Make.contigs on data 80, data 79, and others: scrap.contigs.qual", "85: Make.contigs on data 80, data 79, and others: fasta", "84: Make.contigs on data 80, data 79, and others: 1 rim.contigs.qual", "83: Make.contigs on data 80, data 79, and others: 1 rim.contigs.fasta", "82: Make.contigs on data 80, data 79, and others: 1 oofile", "81: Mothur_MiSeq_SOP", "80: Mock_S280_1001_R2_001.fasta", "79: Mock_S280_1001_R1_001.fasta", "78: F3D150_S216_1001_R2_001.fasta", "77: F3D150_S216_1001_R1_001.fasta", "76: F3D149_S215_1001_R2_001.fasta", and "75: F3D149_S215_1001_".

Below the configuration area, there is a "Mothur Overview" section with a description: "Mothur is a comprehensive suite of tools for microbial ecology community. It is initiated by Dr. Patrick Schloss and his software development team in the Department of Microbiology and Immunology at The University of Michigan. For more information, see [Mothur-Wiki](#)." This is followed by "Command Documentation" and "Citations" with a "Show BibTex" link.

Introduction to Galaxy



- <http://usegalaxy.org>
- Free registration
- Registered user can access more computing resources
- Good for small to moderate datasets
- May not be suitable for analysis that require large amount of computing resources

Private Galaxy server in the cloud

- All three major public cloud service supported
- Amazon AWS service preferred



CloudMan

<https://galaxyproject.org/cloudman/>

Private Galaxy server in the cloud

The screenshot displays the Galaxy on the Cloud web interface. The browser's address bar shows the URL `34.201.165.181`. The page features a navigation menu on the left with categories like "Tools", "Get Data", and "Collection Operations". The main content area includes a "Welcome to Galaxy on the Cloud" message, a "Big, important change:" section with instructions on tool configuration, and a footer with information about the Galaxy project. A right-hand sidebar shows a "History" section with a search box and a message indicating that the history is currently empty.

EC2 Management Console

CloudMan: felix-galaxy-cloud

Galaxy

34.201.165.181

Galaxy

Analyze Data Workflow Shared Data Visualization Help Login or Register

Account registration or login

Using 0 bytes

History

search datasets

Unnamed history (empty)

This history is empty. You can load your own data or get data from an external source.

Welcome to Galaxy on the Cloud
managed by CloudMan

Galaxy on the Cloud is ready for use!

Big, important change:

This configuration of Galaxy has several tools (listed under the tools tag) set to run in parallel. This leads to more robust and faster job completion. However, this also requires at least 4 processors on the worker node. If your jobs are not running, add a worker node via CloudMan with at least 4 vCPUs.

- To learn how to use Galaxy please see the [wiki](#).
- To install new tools to your Galaxy follow the [tutorial](#).
- To manage this cloud instance, use [CloudMan](#).

Thank you for using Galaxy.

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of the [Center for Comparative Genomics and Bioinformatics](#) at Penn State, and the [Department of Biology and Computer Science](#) at Johns Hopkins University. The Galaxy Project is supported in part by [NIH/NCI](#), [NSF](#), [The Huck Institutes of the Life Sciences](#), [The Institute for CyberScience at Penn State](#), and [Johns Hopkins](#).

Tools

search tools

Get Data

Collection Operations

Exome Sequencing

Fetch Sequences

Fetch Alignments

Statistics

Graph/Display Data

MetaPhlan2

HUMAN2

Combine MetaPhlan2 and HUMAN2

Mothur

Metaomic Analysis

FASTA manipulation

NGS: QC and manipulation

NGS: DeepTools

NGS: Mapping

NGS: RNA Analysis

NGS: Peak Calling

NGS: SAMtools

NGS: BamTools

NGS: Picard

NGS: VCF Manipulation

NGS: Variant Analysis

Operate on Genomic Intervals

CloudMap

Phenotype Association

BEDTools

Regional Variation

Multiple Alignments

Multiple regression

Multivariate Analysis

Monte Tools

STR-FM: Microsatellite Analysis

NGS: GATK Tools (beta)

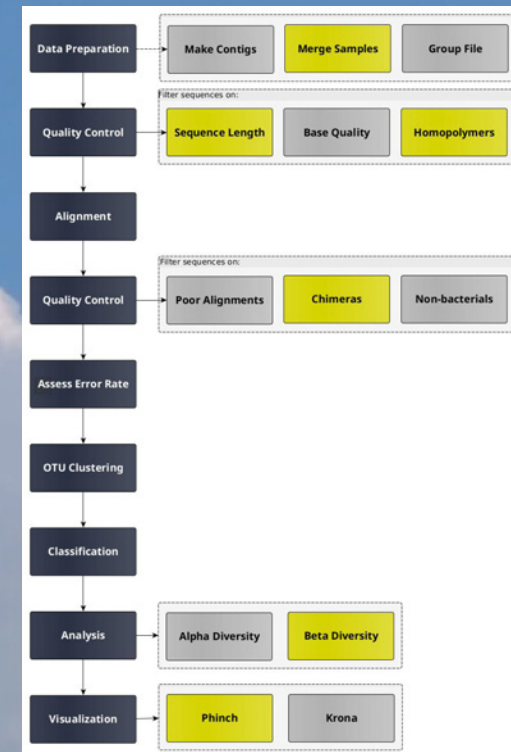
NGS: Assembly

Workflows

- All workflows

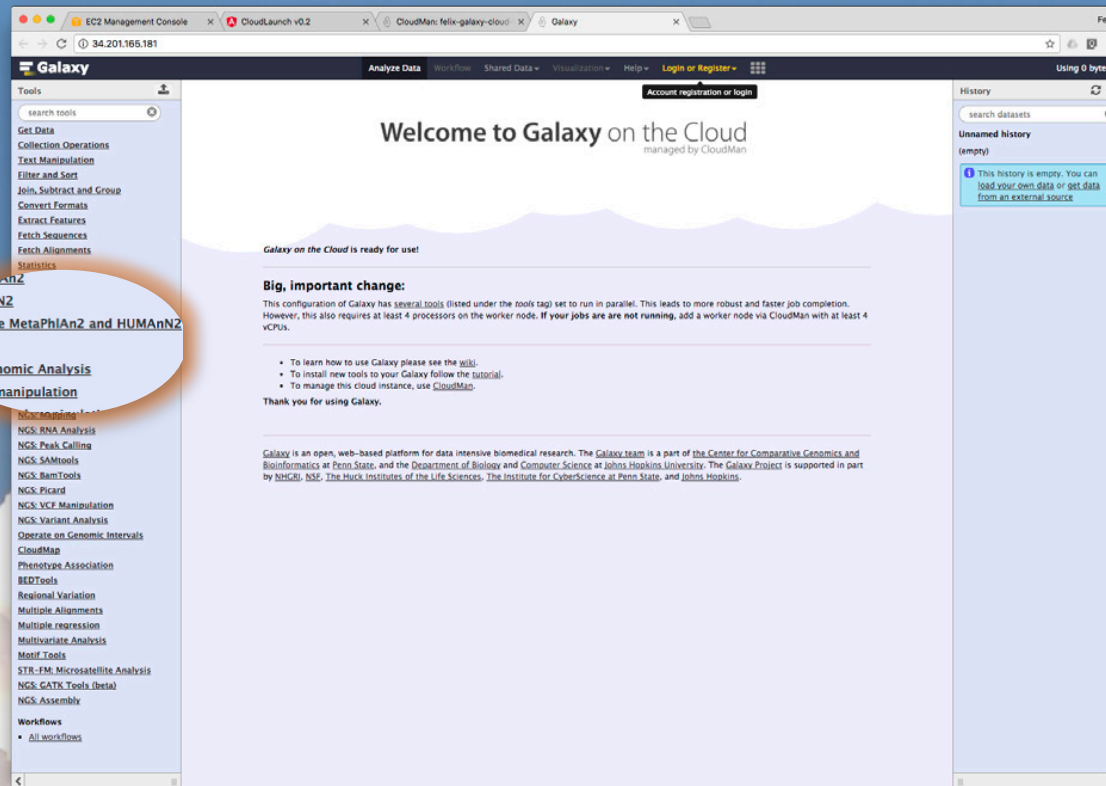
Galaxy workflow for microbiome data analysis

- Demo dataset
 - 17 samples of 16S rRNA gene V4 region
 - MiSeq PE250
 - ~120k reads
- Analysis workflow
 - Mothur



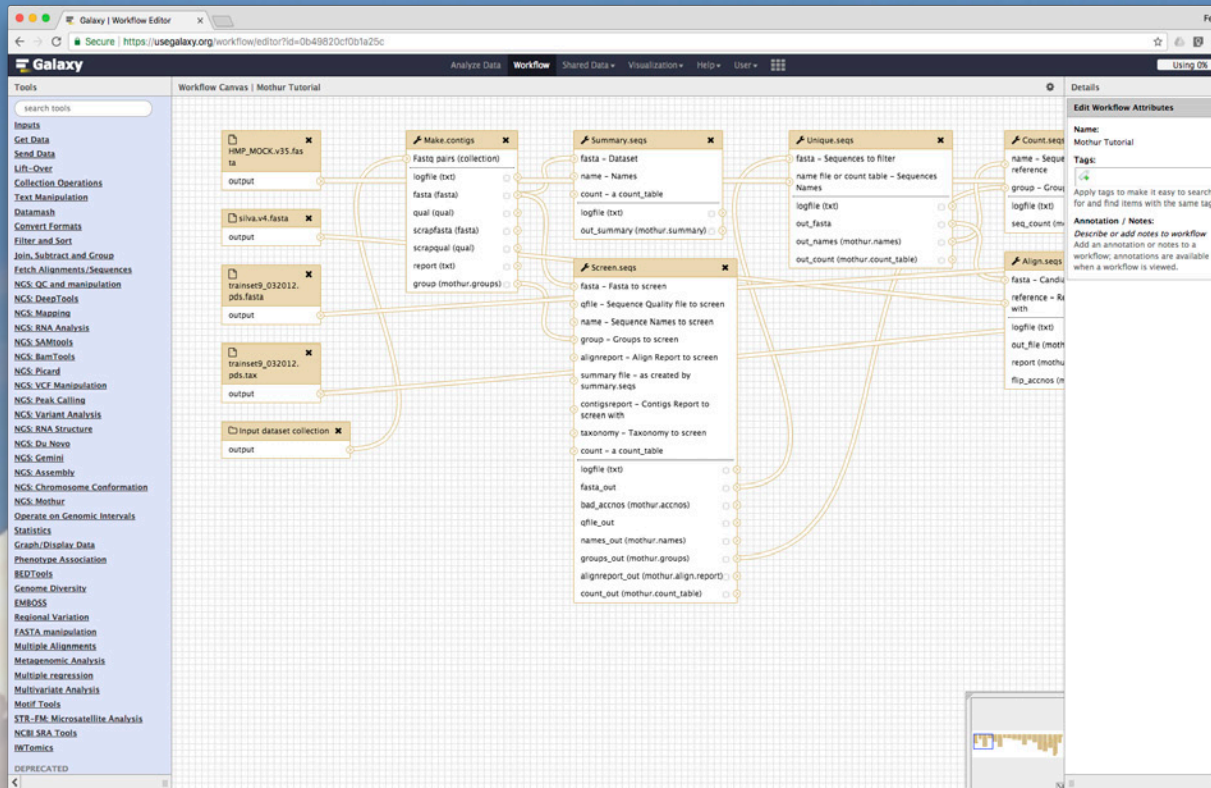
Source: <https://galaxyproject.github.io/training-material/topics/metagenomics/tutorials/mothur-miseq-sop/tutorial.html>

Private Galaxy server in the cloud



Schloss P D *et al.* Applied and Environmental Microbiology, 2009

Galaxy workflow for microbiome data analysis



Suggestion

- Cloud computing provide flexible, scalable platform for bioinformatics analysis
- Best for occasional heavy workload analysis which require large amount of memory, disk space and CPU time
- Setup of private Galaxy cloud server using CloudMan is straightforward
- Remember to stop the server once the analysis is done as idle server incur charge



Thank you