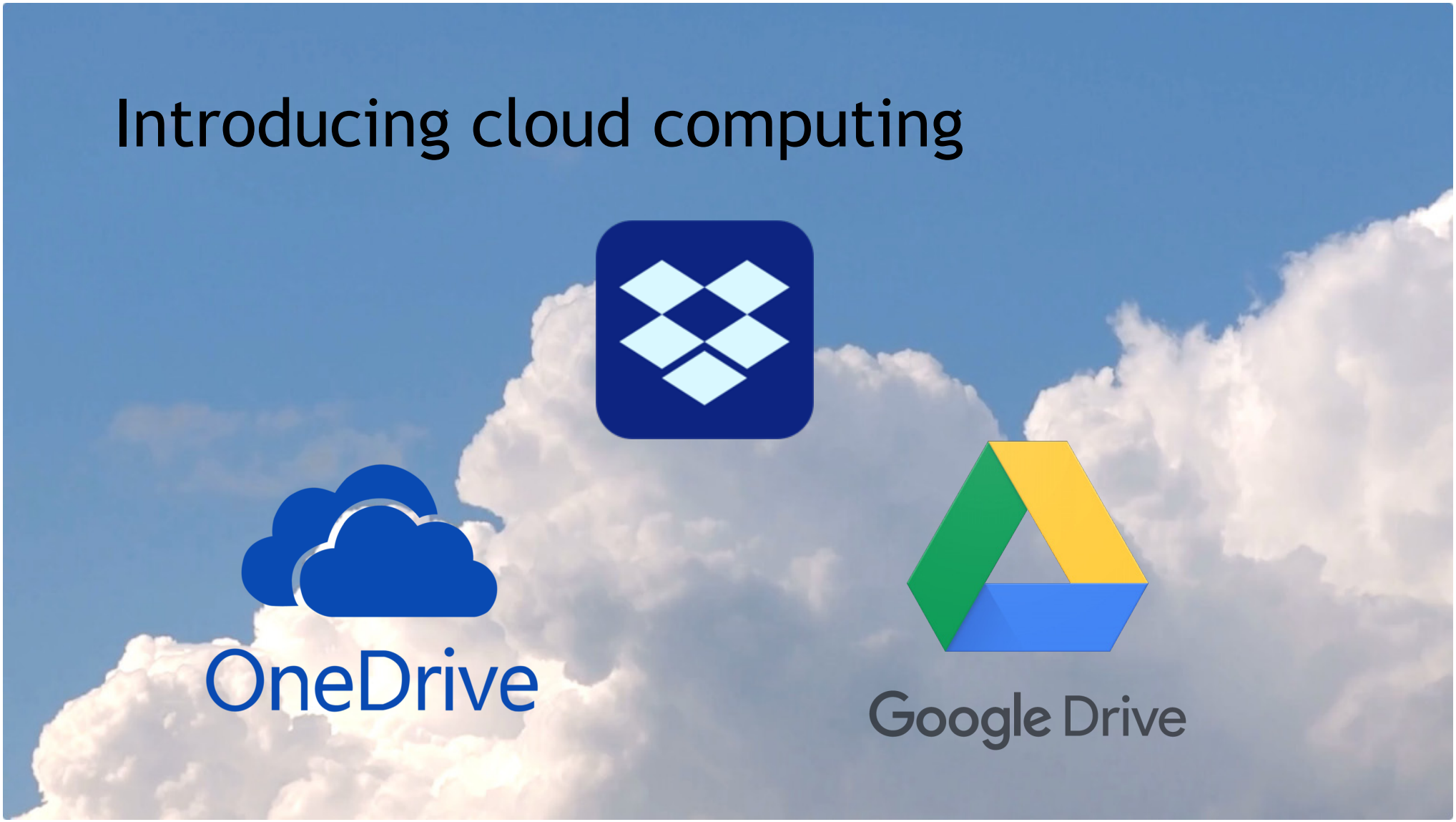
Cloud Computing and its Application in Microbiome Data Analysis

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

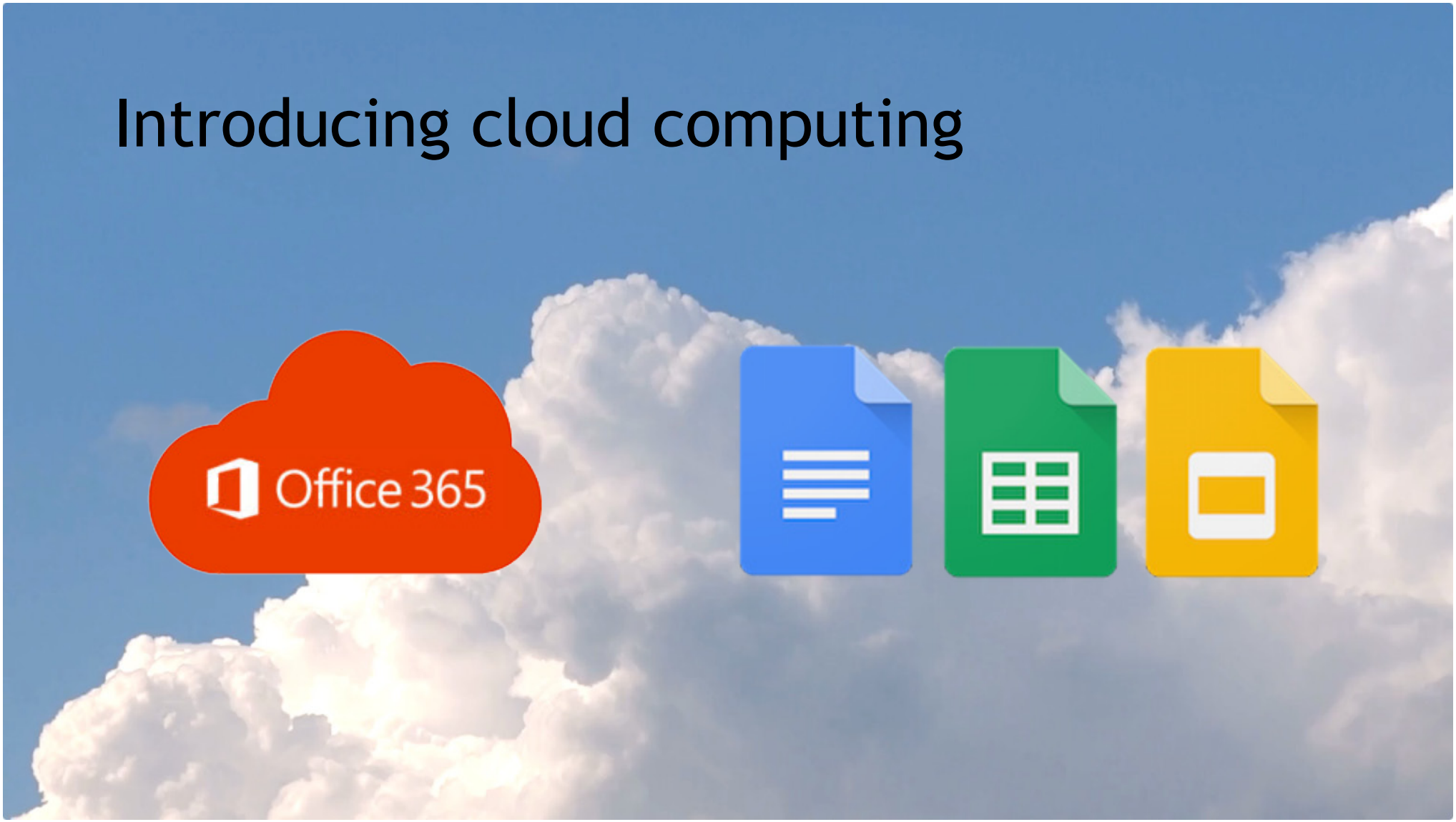
Outline

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



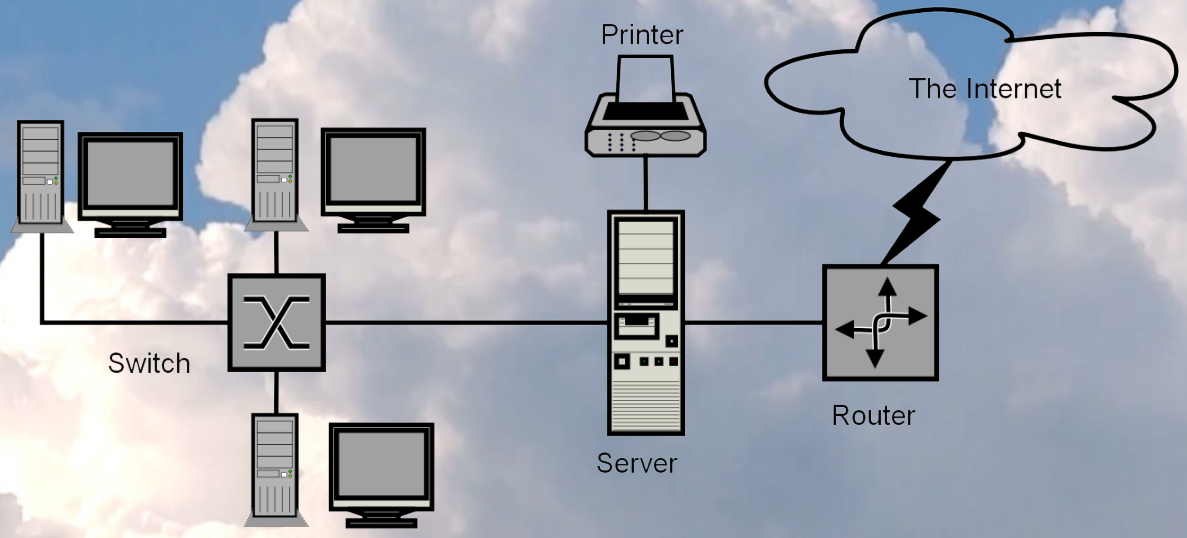
Introducing cloud computing



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

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

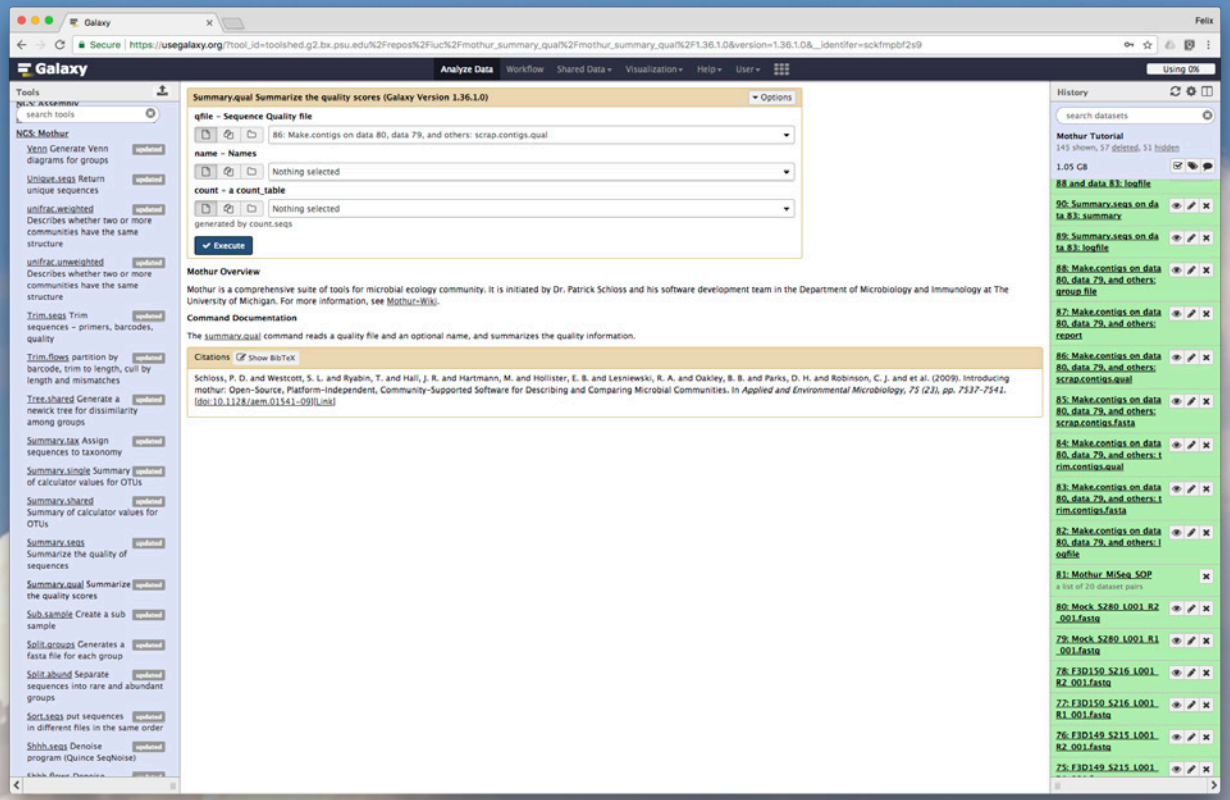
Reproducible

- 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



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

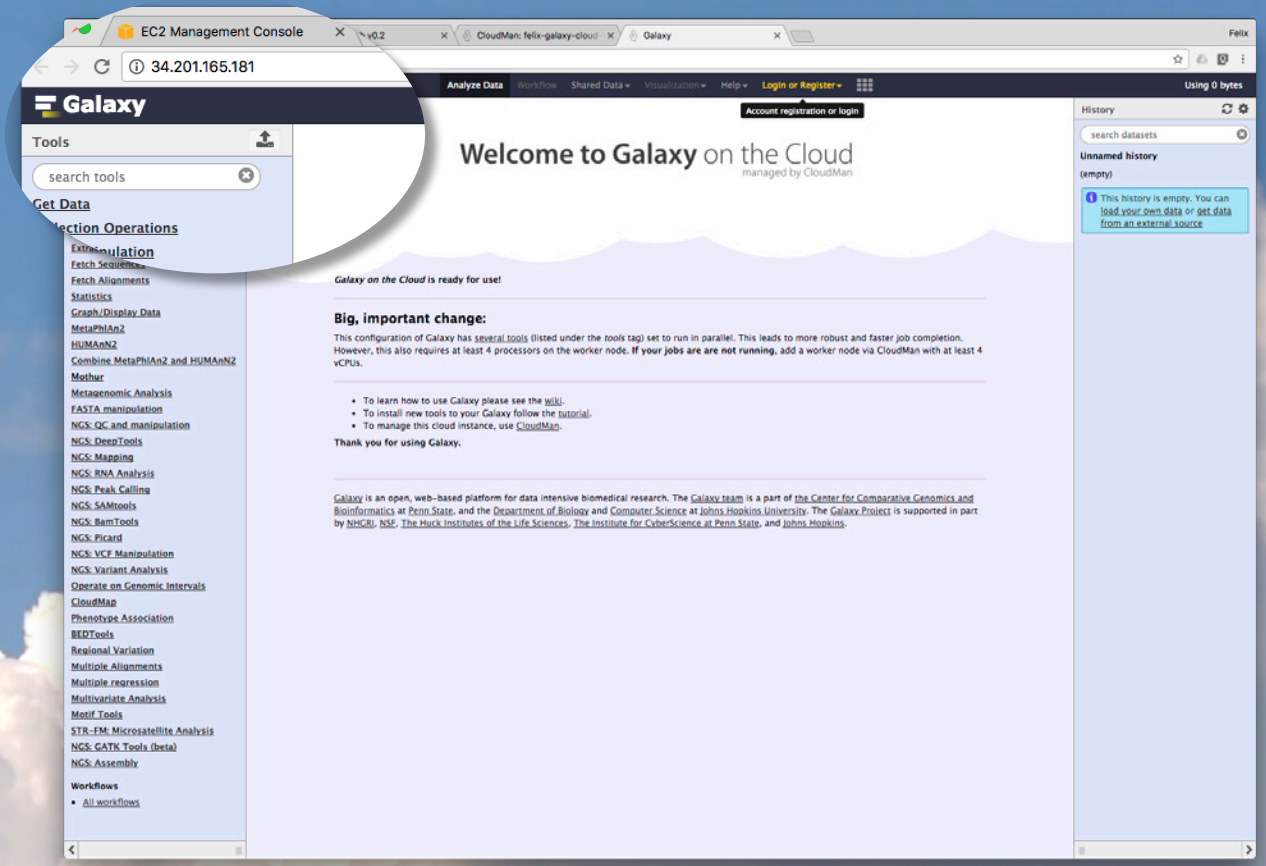
Private Galaxy server in the cloud

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

CloudMan

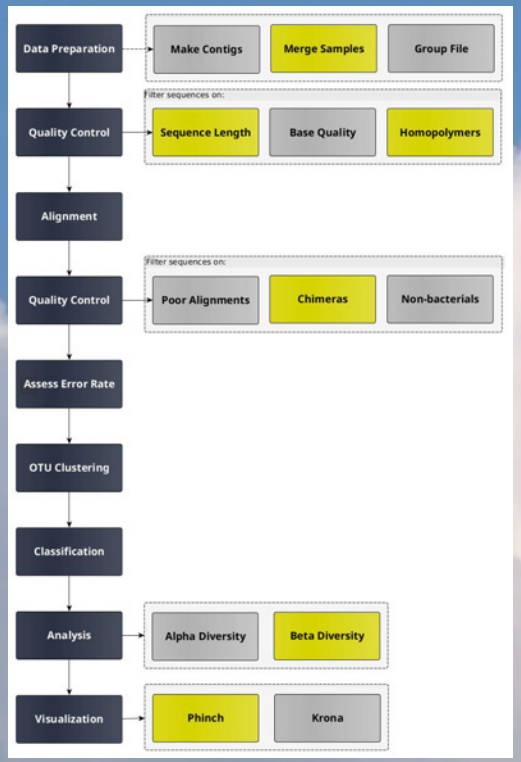
https://galaxyproject.org/cloudman/

Private Galaxy server in the cloud



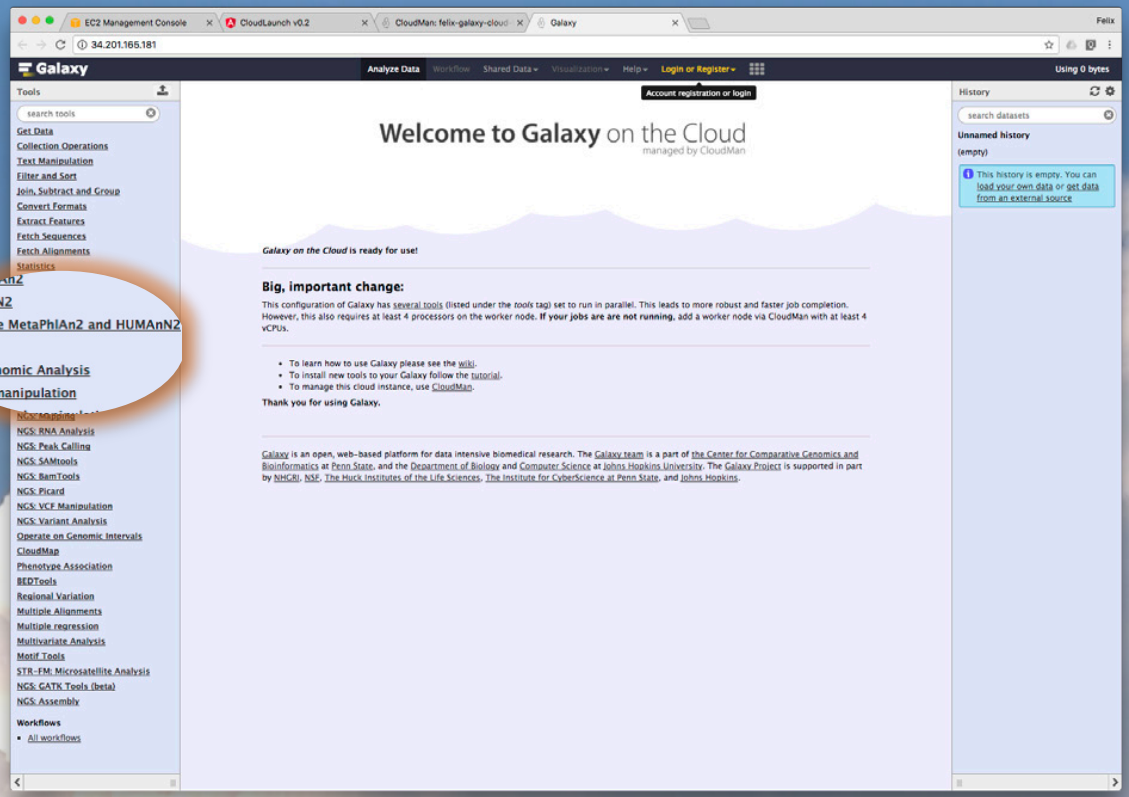
Galaxy workflow for microbiome data analysis

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



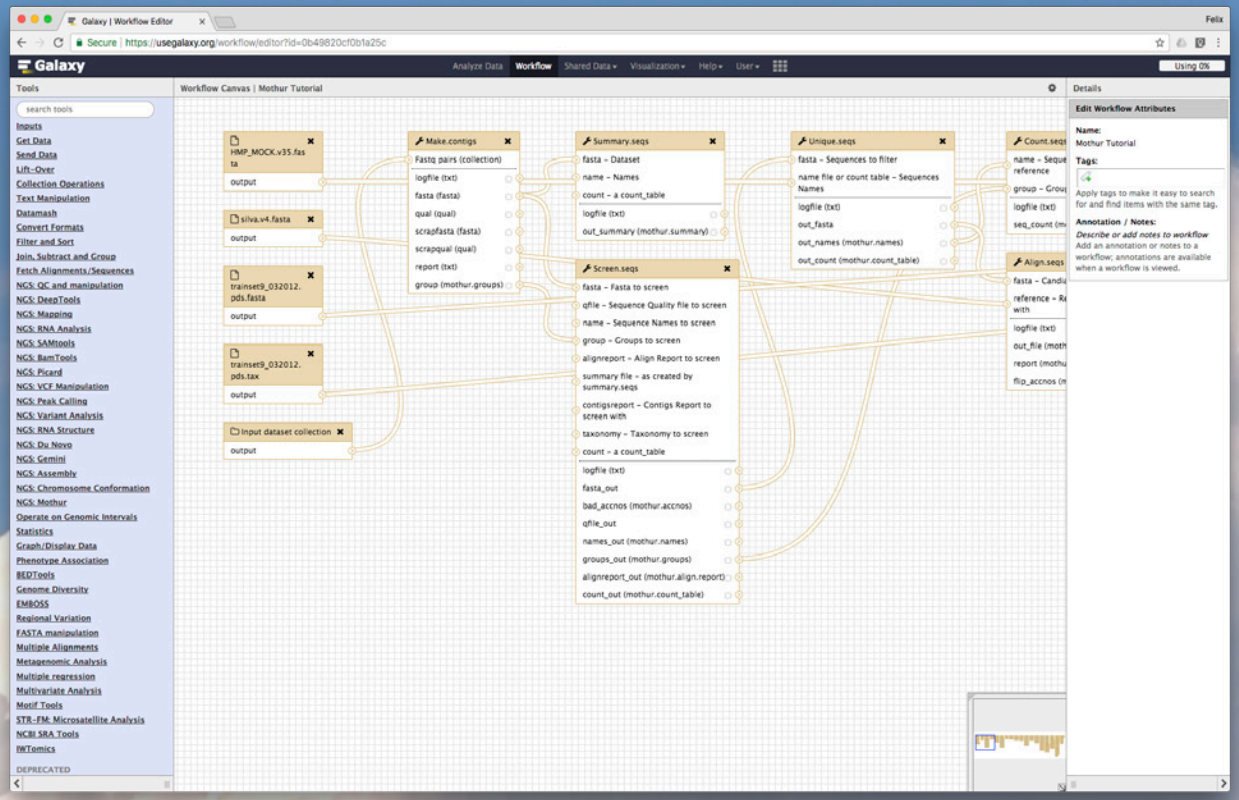
Source: https://galaxyproject.github.io/training-material/topics/metagenomics/tutorials/mothur-miseq-sop/tutorial.html

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