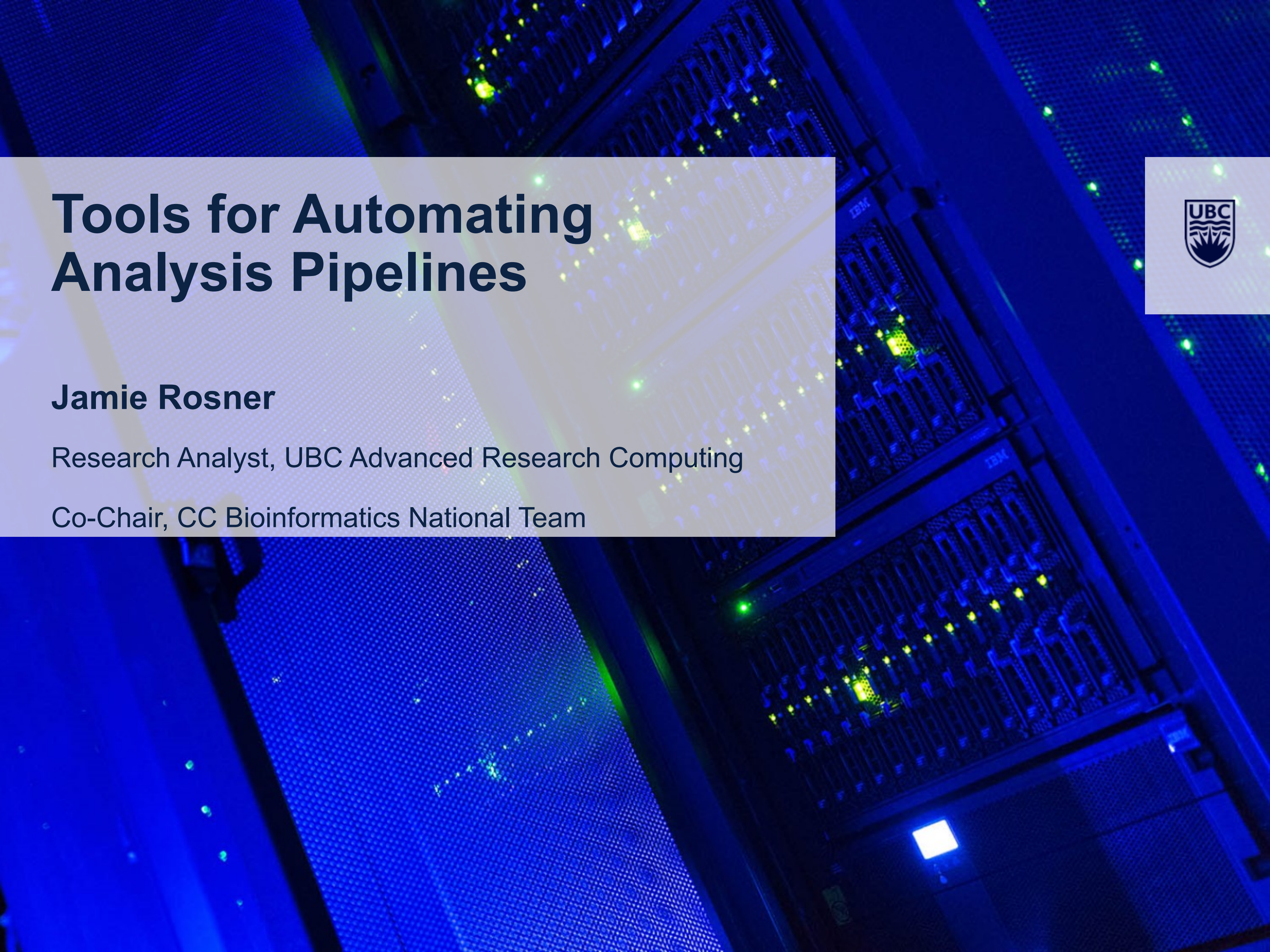


Tools for Automating Analysis Pipelines

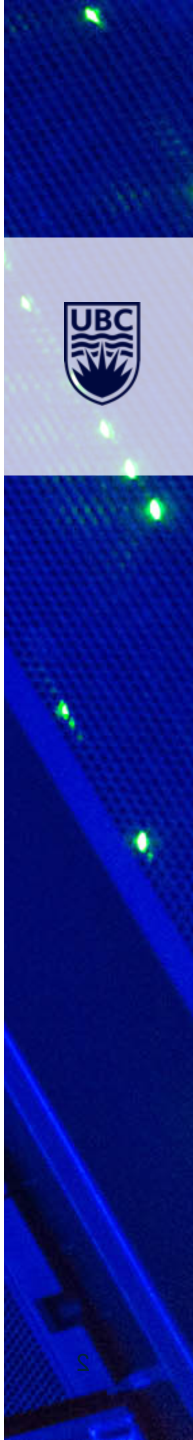
Jamie Rosner

Research Analyst, UBC Advanced Research Computing

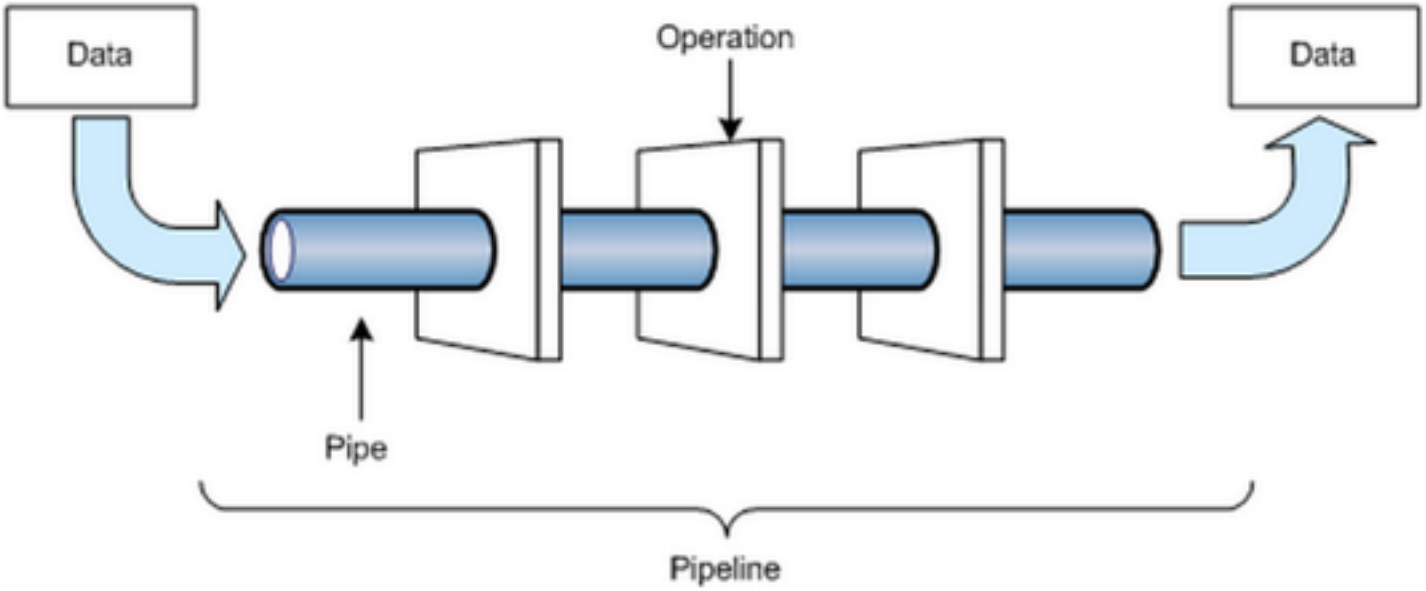
Co-Chair, CC Bioinformatics National Team

Outline

- » What is an analysis pipeline?
- » Automation vs running manually
- » Different kinds
- » How do they work?
- » Conclusion — things to consider

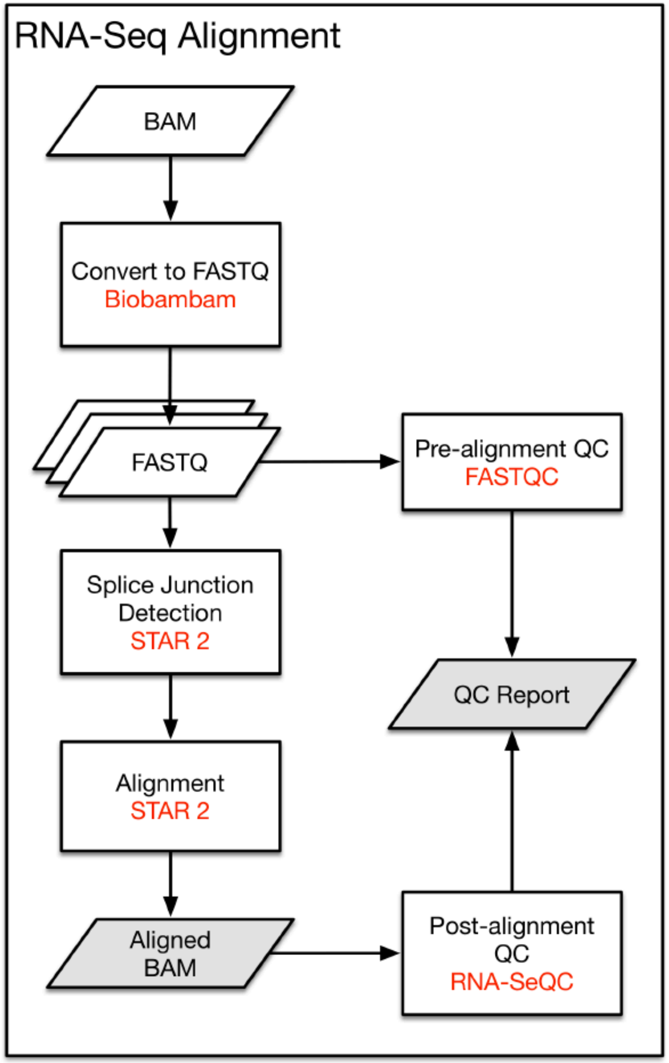


What is an analysis pipeline?

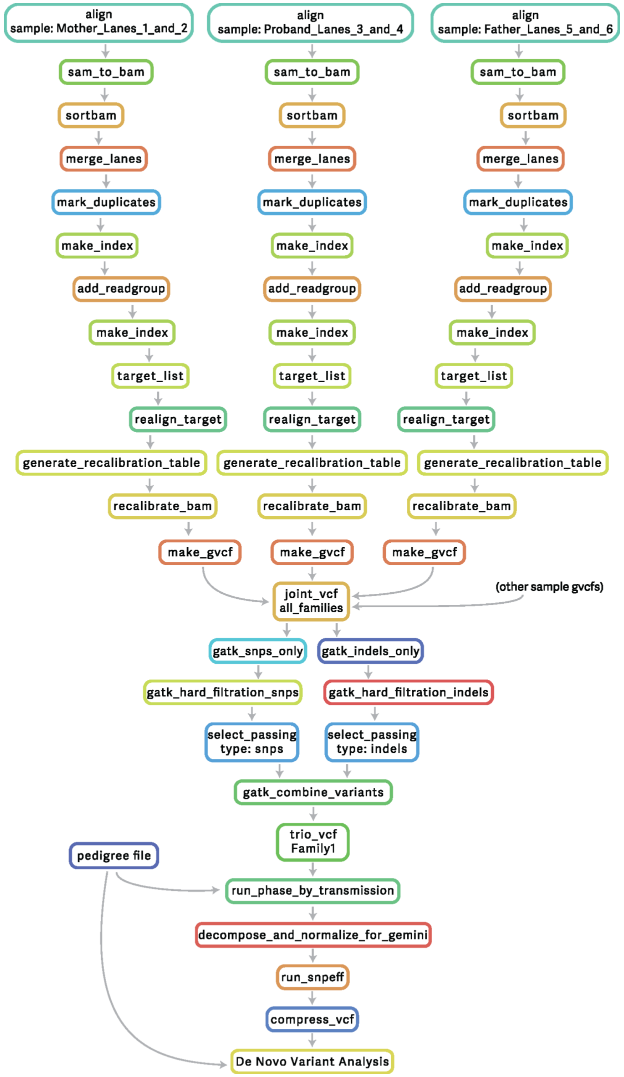


A *pipeline* has inputs go through a number of processing steps chained together in some way to produce some sort of output.

slightly more complex….



more complex still… Yipes!



align
sample: Mother_Lanes_1_and_2

align
sample: Proband_Lanes_3_and_4

align
sample: Father_Lanes_5_and_6

sam_to_bam

sortbam

merge_lanes

mark_duplicates

make_index

add_readgroup

make_index

target_list

realign_target

generate_recalibration_table

recalibrate_bam

make_gvcf

joint_vcf
all_families

(other sample gvcfs)

gatk_snps_only

gatk_indels_only

gatk_hard_filtration_snps

gatk_hard_filtration_indels

select_passing
type: snps

select_passing
type: indels

gatk_combine_variants

trio_vcf
Family1

pedigree file

run_phase_by_transmission

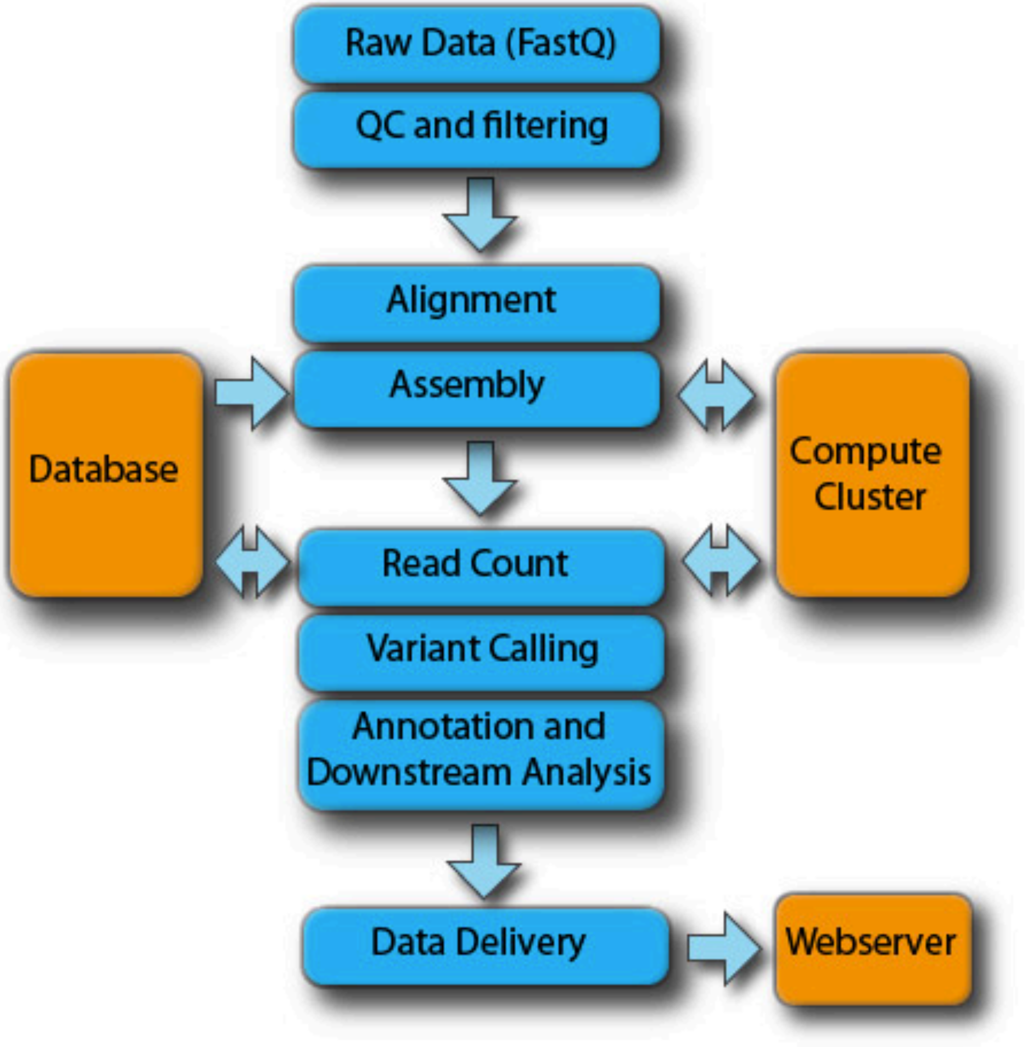
decompose_and_normalize_for_gemini

run_snpeff

compress_vcf

De Novo Variant Analysis

Other factors…



Raw Data (FastQ)

QC and filtering

Alignment

Assembly

Read Count

Variant Calling

Annotation and Downstream Analysis

Data Delivery

Database

Compute Cluster

Webserver

The Problem with Running Manually

- **Efficiency**
  - Potentially a lot of work — e.g. running 1000 samples
  - Time wasted between tasks
  - File/Error management can get complicated
- **Requires Proficiency in…**
  - UNIX
  - Programming languages (e.g. Python, Perl, R)
  - Distributed computing
- **Reproducibility**
  - What did I (or they) do?
  - What software versions?
  - What parameters were used?

Advantages of Automated Pipelines

These vary from tool to tool, but in general:

- Reproducibility / Global auditing and logs
- Relaunching made easy
- Portable / Sharable
- Visualization (DAG)
- User-friendliness - GUI or other code abstraction
- Community - reuse / modify existing workflows

Sometimes called Workflow Managers

There's a million of them…

- Galaxy
- GenAP
- Arvados
- Nextflow
- Ruffus (Python)
- Snakemake (Python)
- PyDoit (Python)
- GenePattern (Broad Institute)
- Kronos***
- bpipe
- Taverna
- Luigi
- …

so which one do I use?

3 Different Kinds of Pipeline Tools

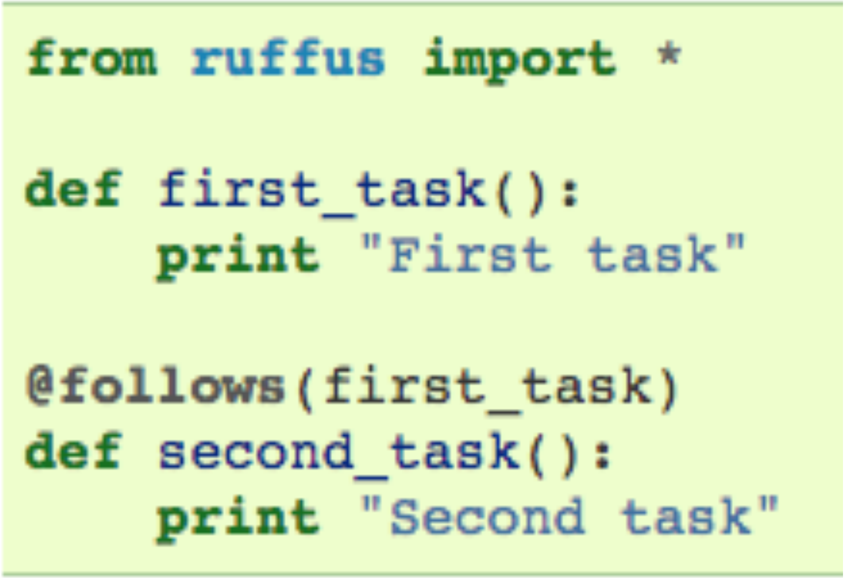
(from low- to high-level)

1. Code Based
2. Configuration File Based
3. GUI Based

1. Code Based

Python Ruffus

A Simple Example:

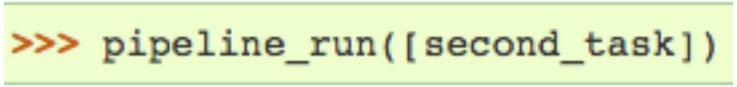


```
from ruffus import *

def first_task():
    print "First task"

@follows(first_task)
def second_task():
    print "Second task"
```

Python Ruffus

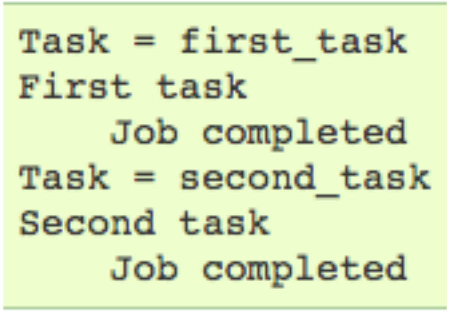
A Simple Example:

Execution:



```
>>> pipeline_run([second_task])
```

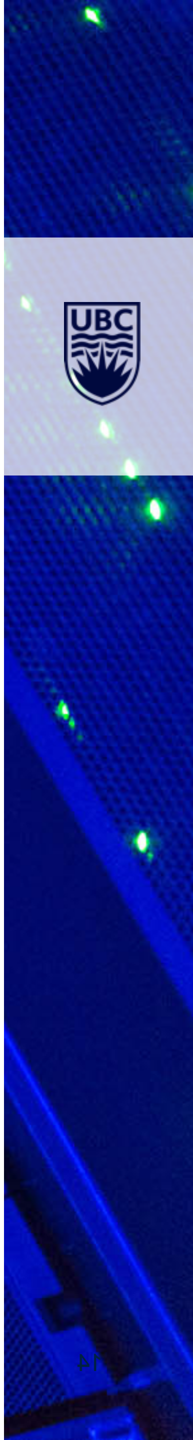
Output:



```
Task = first_task
First task
    Job completed
Task = second_task
Second task
    Job completed
```

https://pythonhosted.org/ruffus/html/simple_tutorial.html

2. Configuration File Based



Common Workflow Language (CWL)

A way to describe command line tools and connect them together to create workflows. Because CWL is a specification and not a specific piece of software, tools and workflows described using CWL are portable across a variety of platforms that support the CWL standard.

www.commonwl.org

Common Workflow Language (CWL)

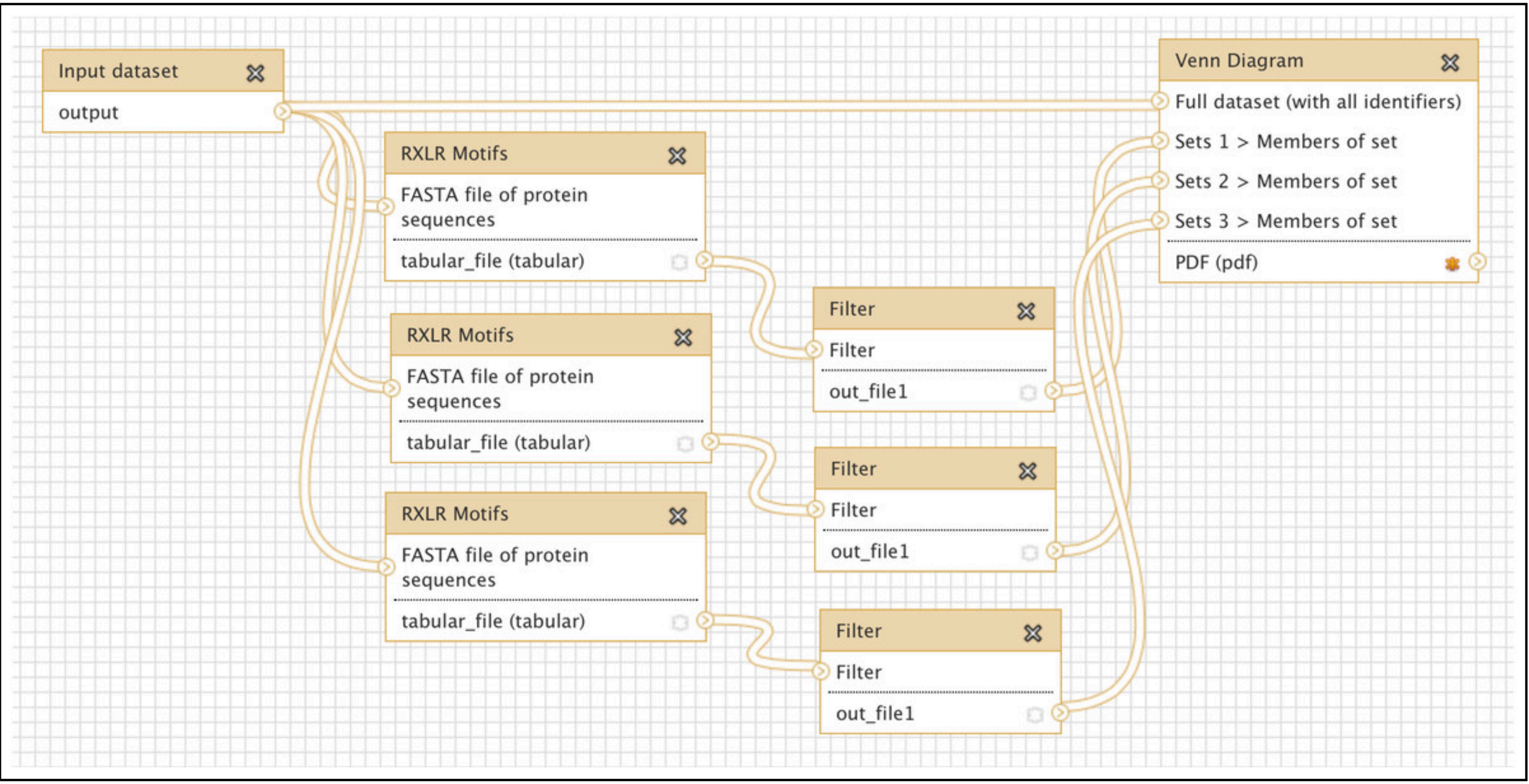
```
#!/usr/bin/env cwl-runner
class: Workflow

cwlVersion: v1.0

inputs:
  genome:
    type: string
  infile:
    type: File
    doc: gzip VCF file to annotate

outputs:
  outfile:
    type: File
    outputSource: snpeff/output
  statsfile:
    type: File
    outputSource: snpeff/stats
  genesfile:
    type: File
    outputSource: snpeff/genes
```

```
steps:
  gunzip:
    run: gunzip.cwl
    in:
      gzipfile:
        source: infile
    out: [unzipped_vcf]

  snpeff:
    run: snpeff.cwl
    in:
      input_vcf: gunzip/unzipped_vcf
      genome: genome
    out: [output, stats, genes]

doc: |
  Annotate variants provided in a
gziped VCF using SnpEff
```

3. Gimme a GUI !!!

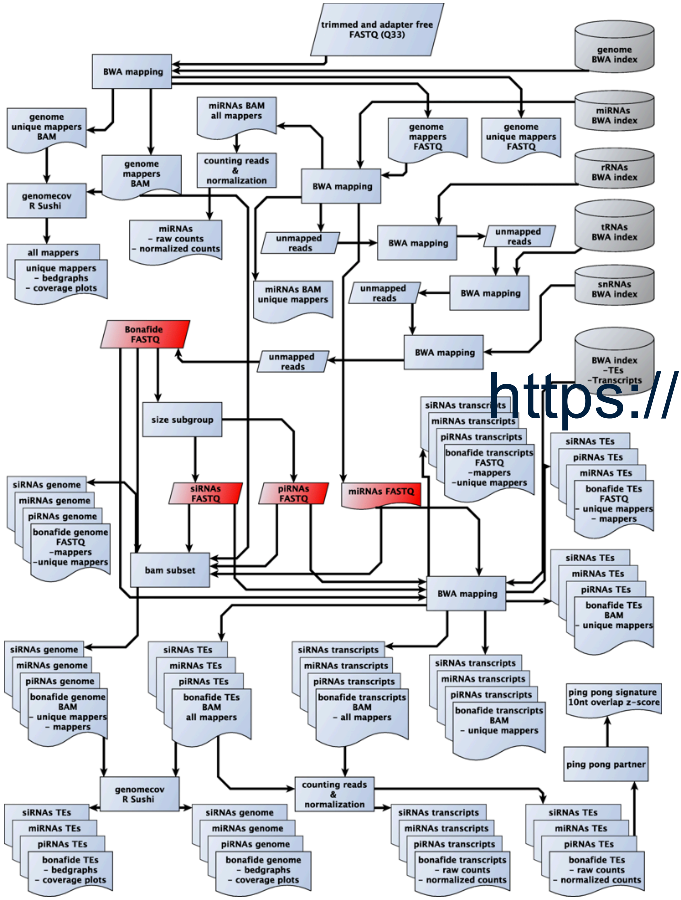
Galaxy Workflow Manager

https://usegalaxy.org/



https://www.the-scientist.com/lab-tools/how-to-build-bioinformatic-pipelines-using-galaxy-33085

Galaxy Community



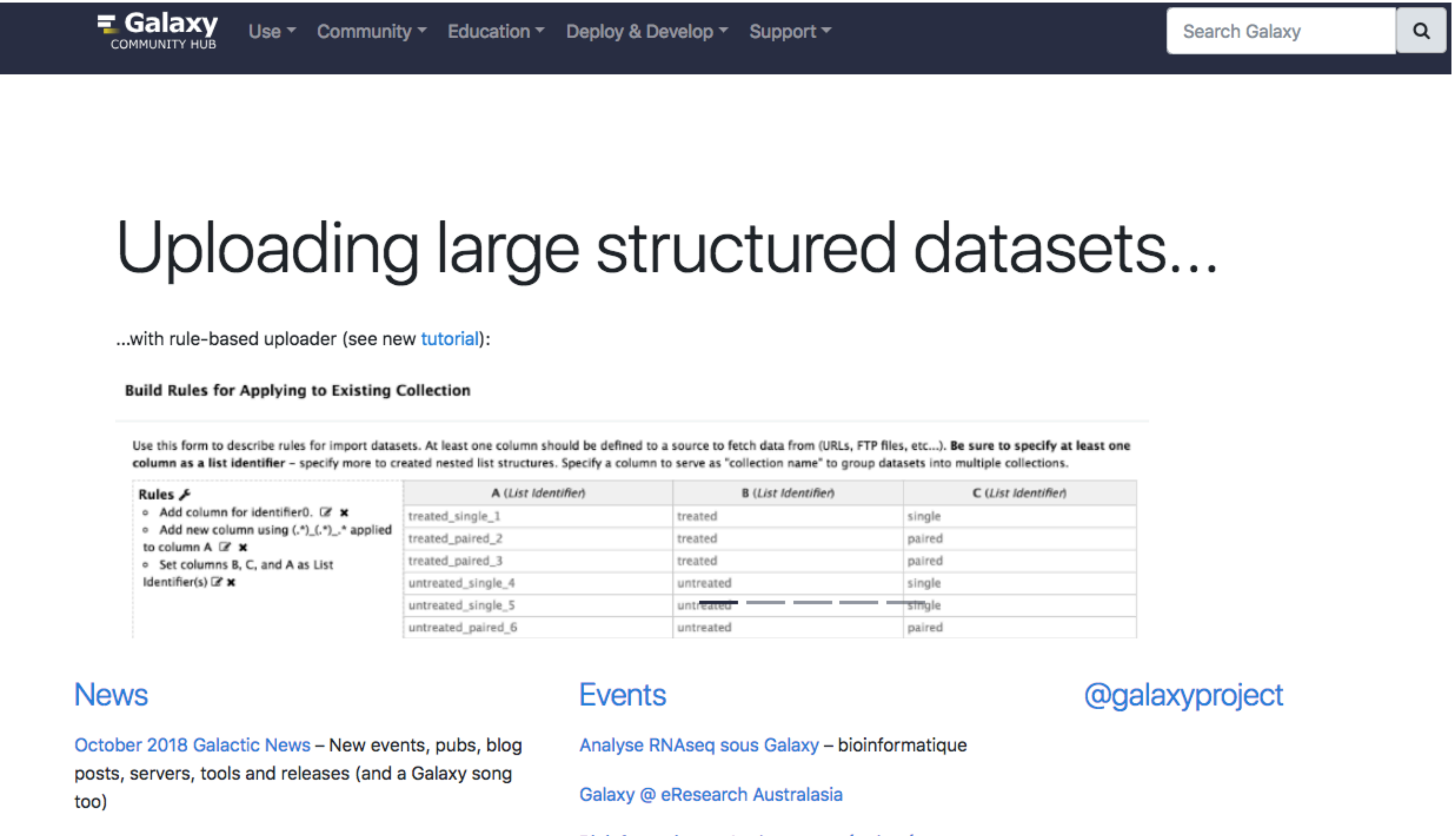
sRNAPipe is freely available

as a Galaxy tool/via GitHub

https://galaxyproject.org/

https://github.com/brassetjensen/sRNAPipe

Galaxy Community Hub https://galaxyproject.org



Uploading large structured datasets...

...with rule-based uploader (see new tutorial):

Build Rules for Applying to Existing Collection

Use this form to describe rules for import datasets. At least one column should be defined to a source to fetch data from (URLs, FTP files, etc...). **Be sure to specify at least one column as a list identifier** – specify more to created nested list structures. Specify a column to serve as "collection name" to group datasets into multiple collections.

Rules

- Add column for identifier0.
- Add new column using (.*)_(.*)_.* applied to column A
- Set columns B, C, and A as List Identifier(s)

A (*List Identifier*)	B (*List Identifier*)	C (*List Identifier*)
treated_single_1	treated	single
treated_paired_2	treated	paired
treated_paired_3	treated	paired
untreated_single_4	untreated	single
untreated_single_5	untreated	single
untreated_paired_6	untreated	paired

News

October 2018 Galactic News – New events, pubs, blog posts, servers, tools and releases (and a Galaxy song too)

Events

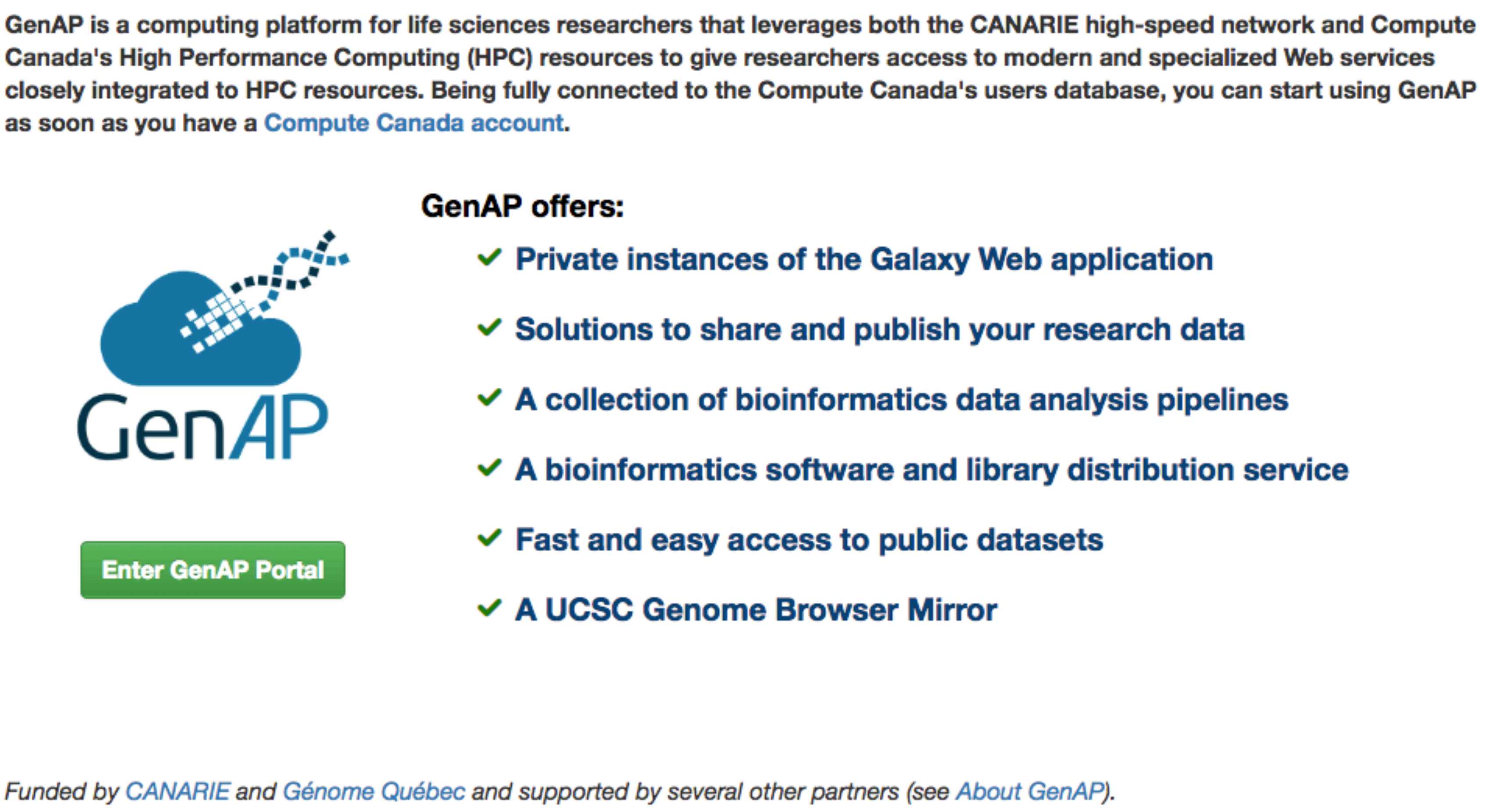
Analyse RNAseq sous Galaxy – bioinformatique

Galaxy @ eResearch Australasia

@galaxyproject

GenAP

https://genap.ca

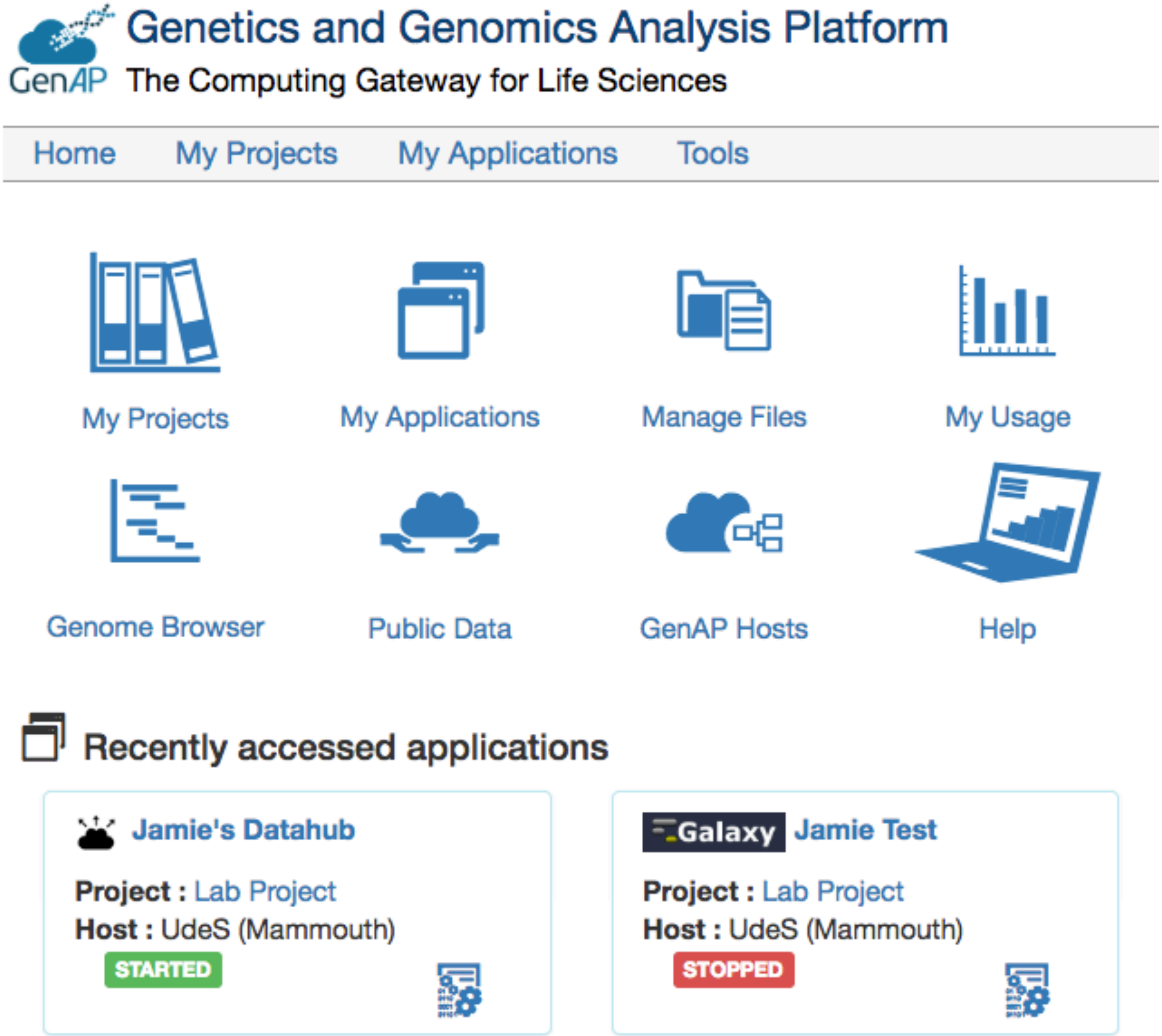


GenAP is a computing platform for life sciences researchers that leverages both the CANARIE high-speed network and Compute Canada's High Performance Computing (HPC) resources to give researchers access to modern and specialized Web services closely integrated to HPC resources. Being fully connected to the Compute Canada's users database, you can start using GenAP as soon as you have a Compute Canada account.

GenAP offers:

- ✓ Private instances of the Galaxy Web application
- ✓ Solutions to share and publish your research data
- ✓ A collection of bioinformatics data analysis pipelines
- ✓ A bioinformatics software and library distribution service
- ✓ Fast and easy access to public datasets
- ✓ A UCSC Genome Browser Mirror

Funded by CANARIE and Génome Québec and supported by several other partners (see About GenAP).



Genetics and Genomics Analysis Platform

The Computing Gateway for Life Sciences

Recently accessed applications

Jamie's Datahub

Project : Lab Project
Host : UdeS (Mammouth)

Jamie Test

Project : Lab Project
Host : UdeS (Mammouth)

So which one should I use?

- **Need a GUI?**
  - Galaxy / GenAP
  - Taverna

- **Have some programming chops?**
  - Python (Snakemake, Ruffus, PyDoit)

- **Other things to consider…**
  - CWL is the future, but still in its infancy
  - Can it run on a cluster?
  - Is it available on my system?
  - Cross-platform? Portable?

- **Don’t take my word for it… research!!!**

- That’s it!

- Thanks!

- Questions?