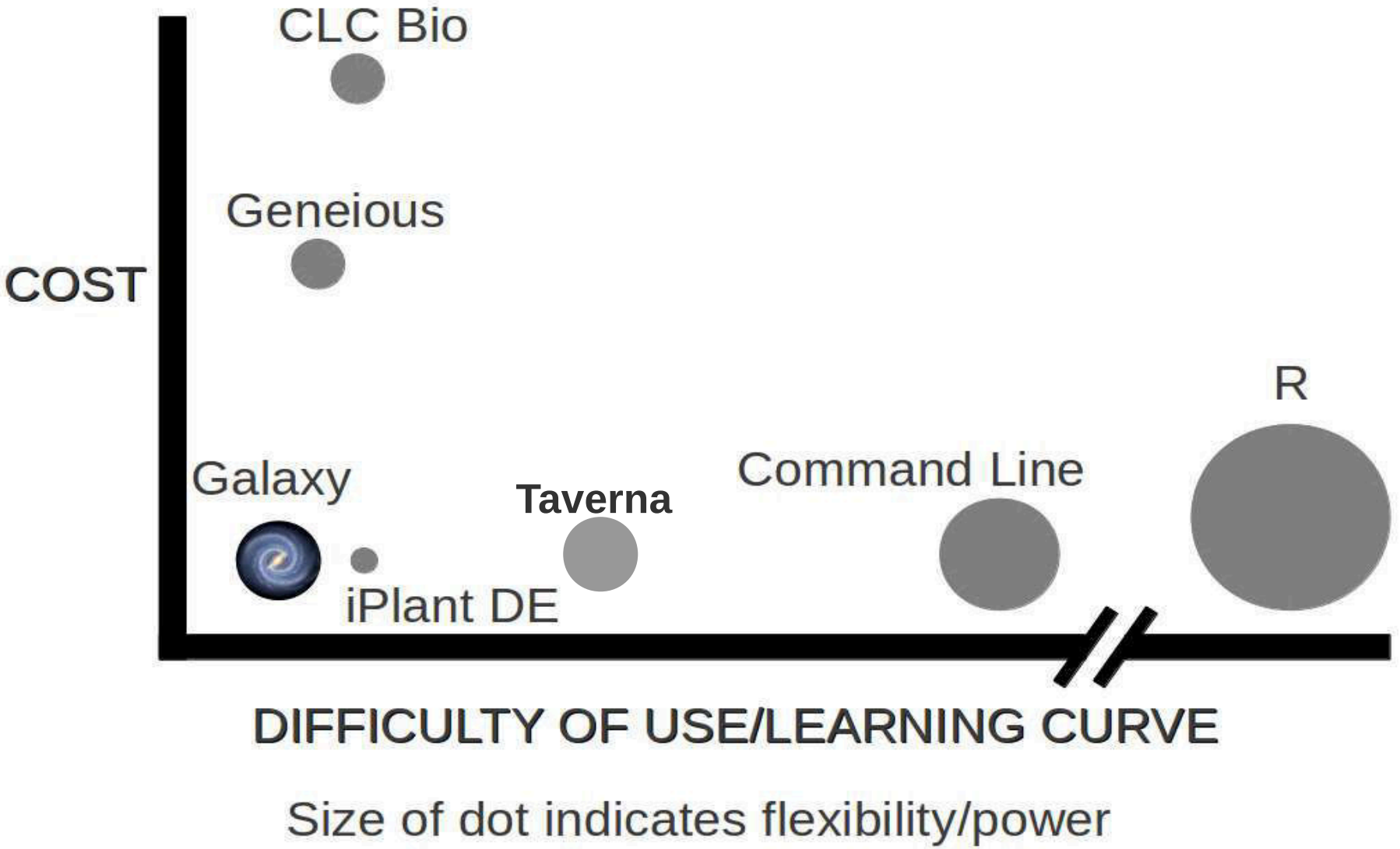
Galaxy Introduction

Leon Mei

Sasc
Sequence Analysis Support Core

Acknowledgement

Galaxy



http://training.bioinformatics.ucdavis.edu/docs/2012/05/RNA/galaxy-intro.html

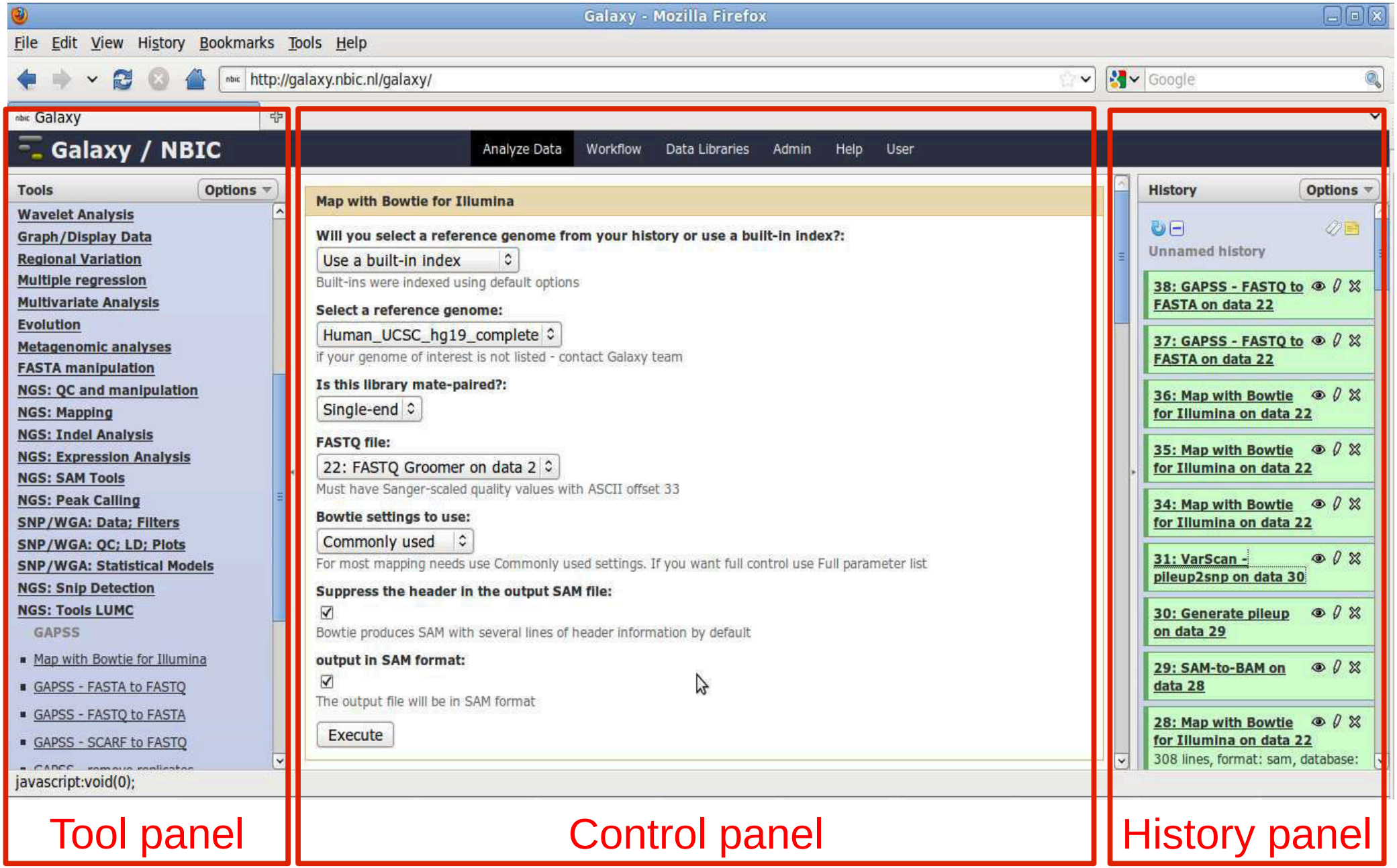
Advantage

- Open source and active community
  - Galaxy community conference starting from 2010
- Web-based client
- History tracking for reproducibility
- Most of NGS tools available and easy tool/data installation
- Use of external resources (UCSC, biomart)
- Support heterogeneous resource
- Support versioning

(Current) Limitations

- No user specific tool panel
- Can not check a job's progress
  - stdout, stderr
- Accounting of user CPU hours
- Tools are installed centrally by admins
- Only support linear and relatively simple workflow

User interface

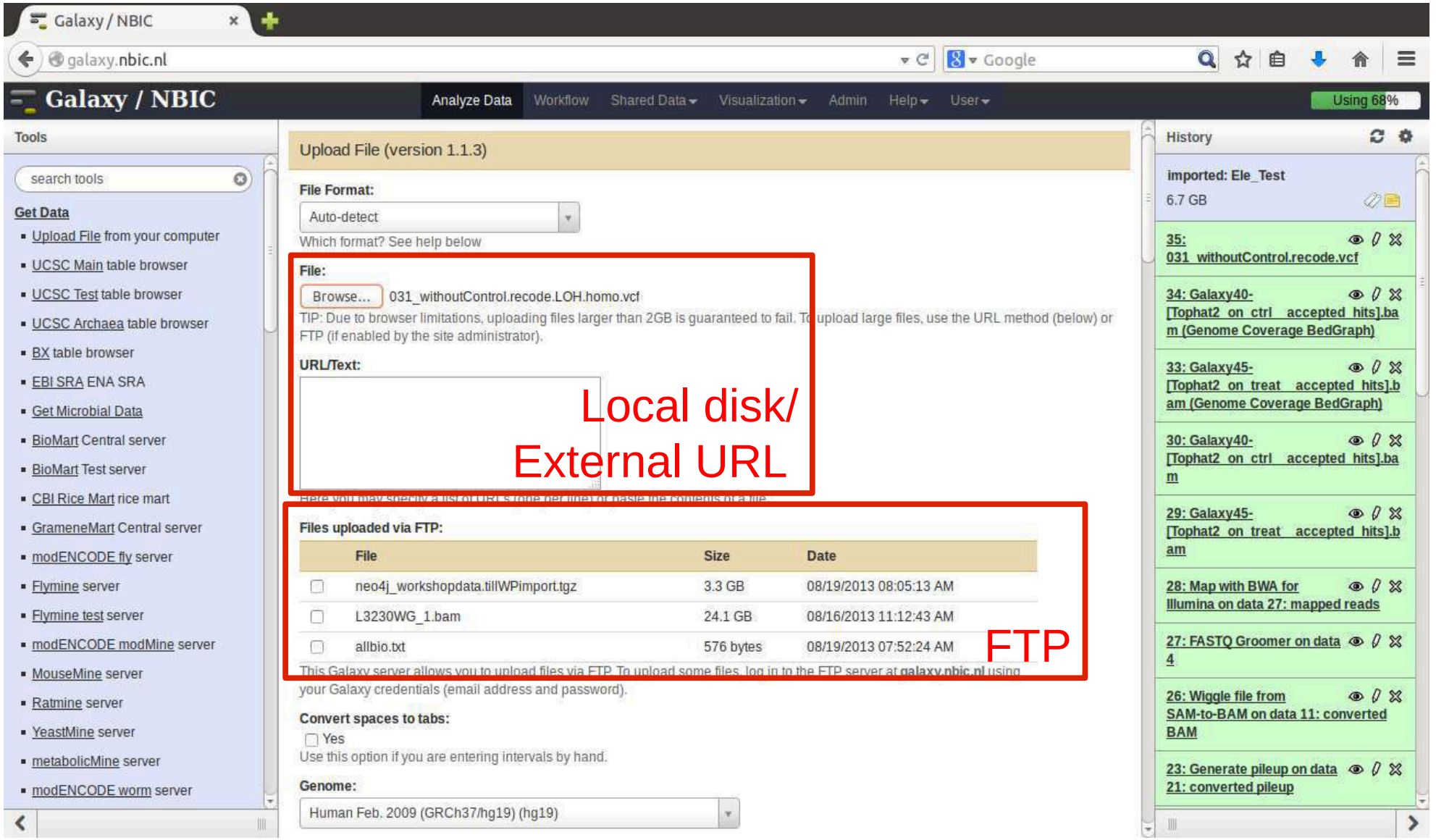


Tool panel

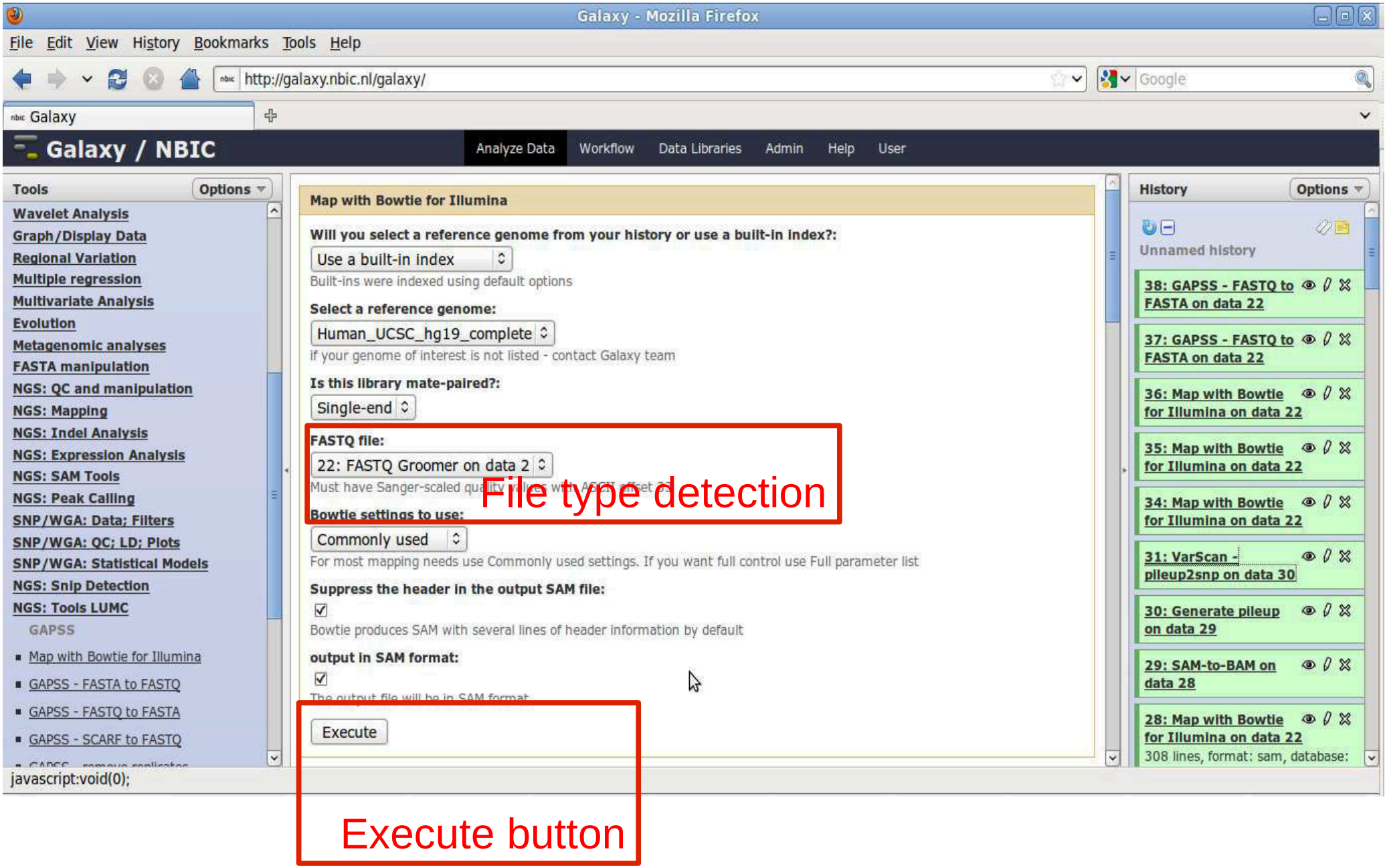
Control panel

History panel

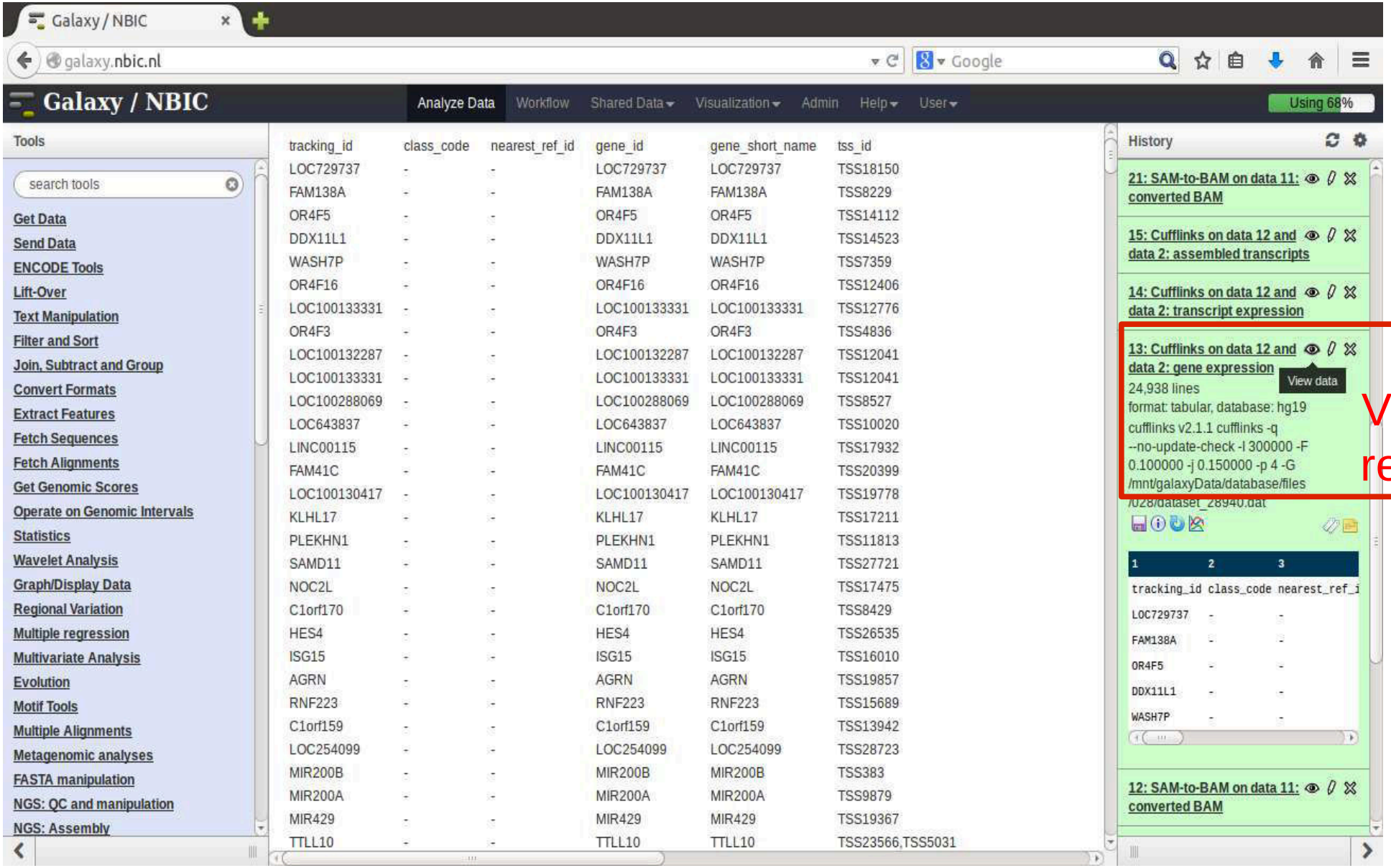
Upload input files



Execute a task



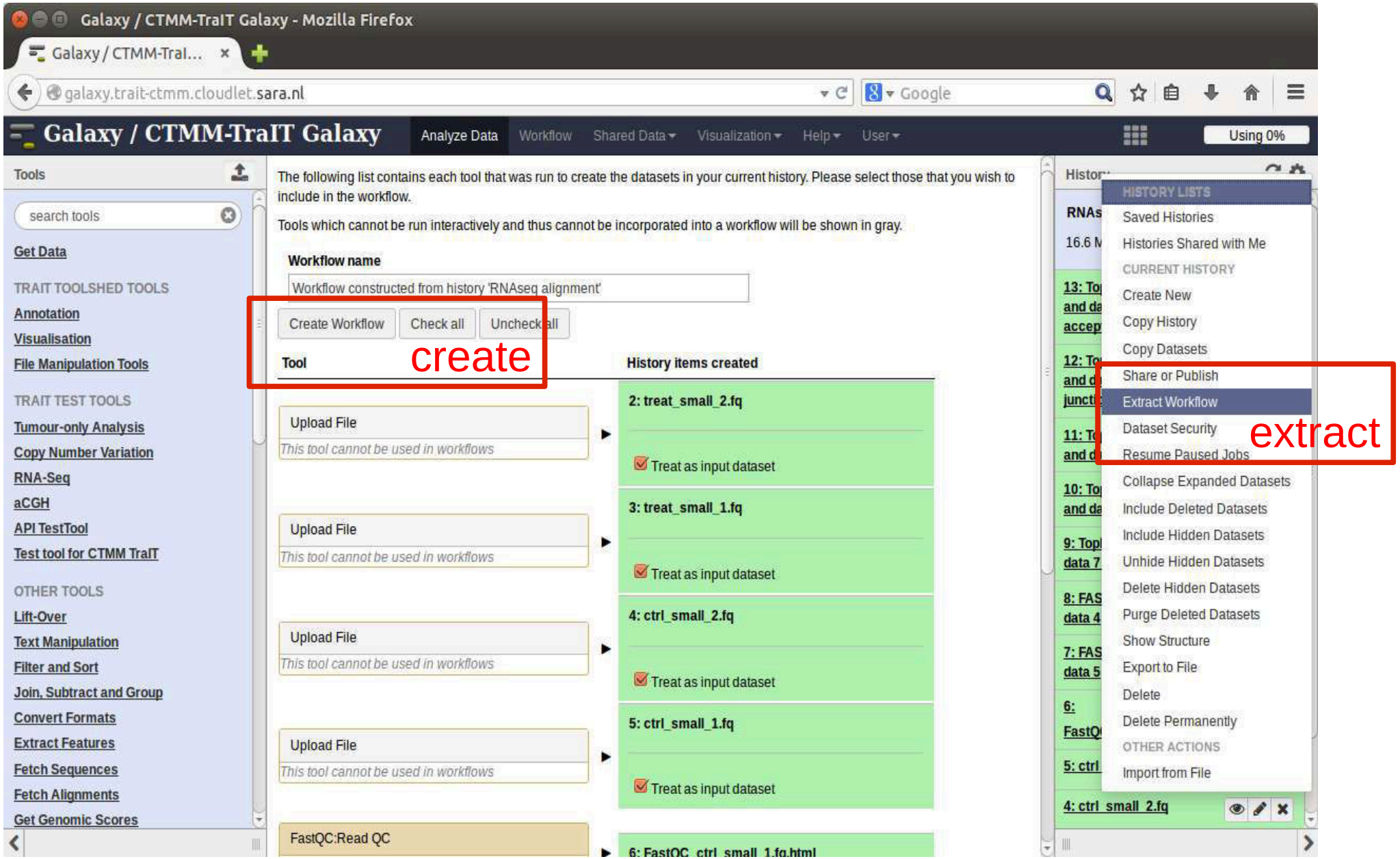
Run and View Results



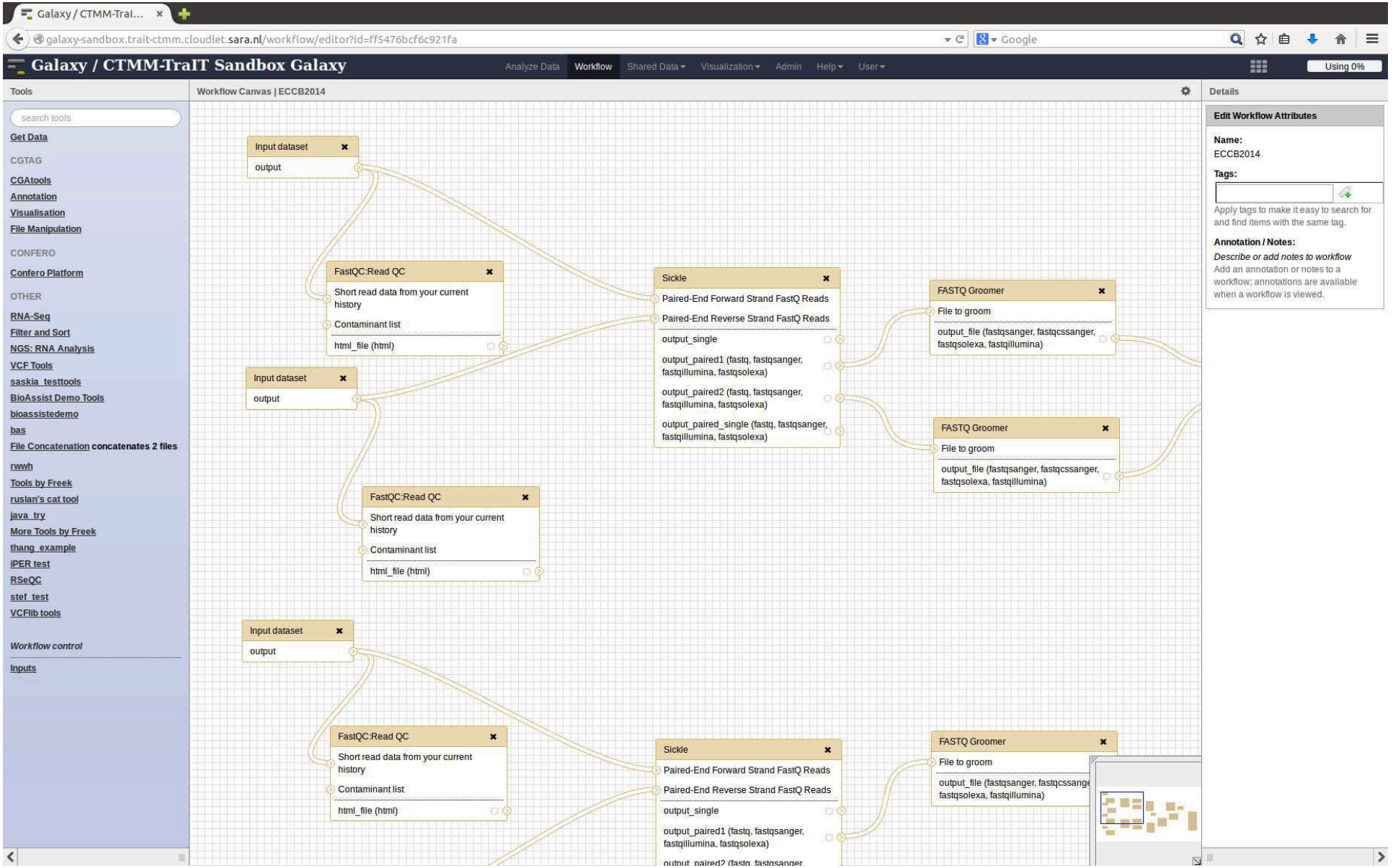
Galaxy workflows

- 3 ways to create Galaxy workflow
  - Extract from your own history
  - Use workflow design panel
  - Shared by other users in Galaxy (local or remote)
- Scalability
  - Parallelism within one workflow
  - API available for automation

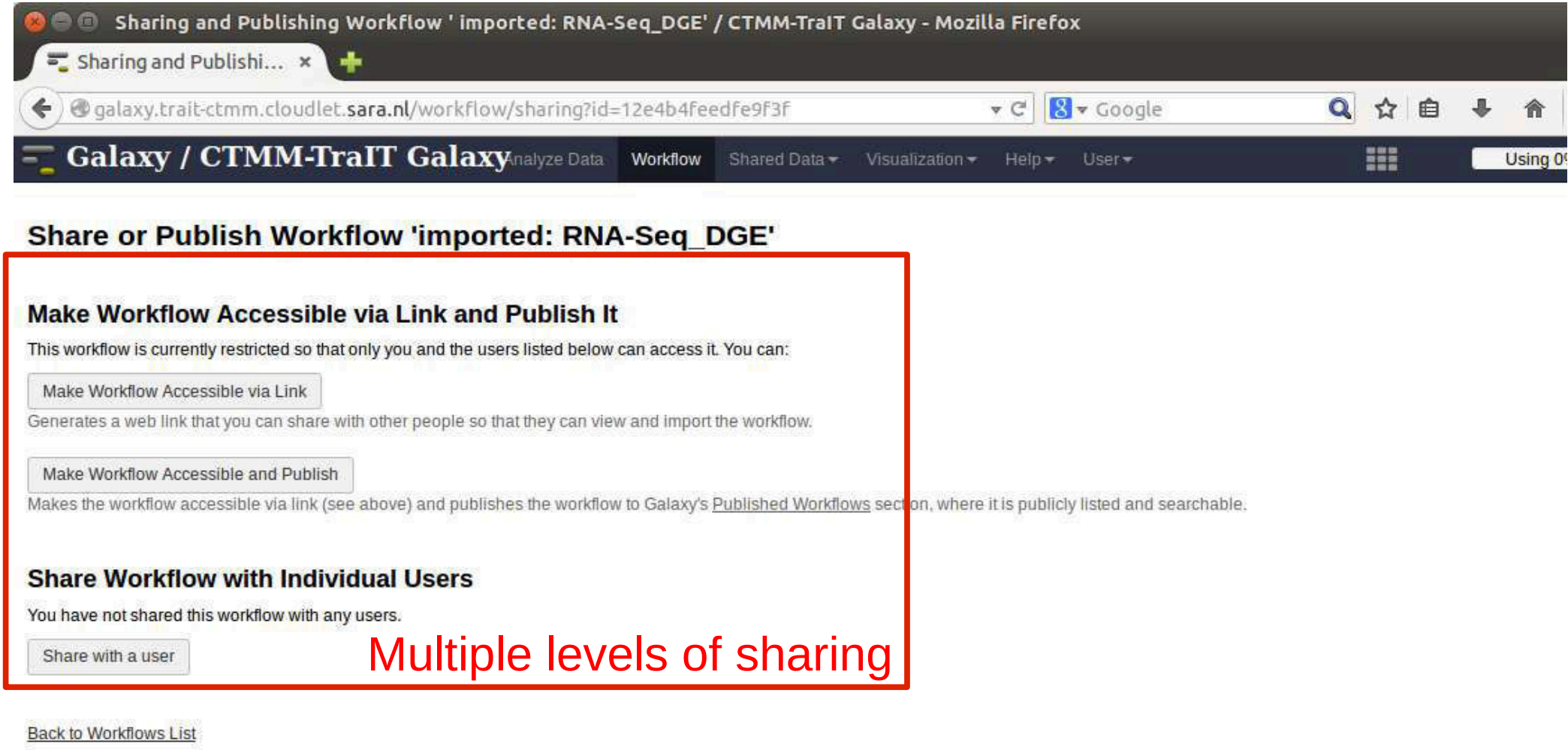
Extract a workflow from history



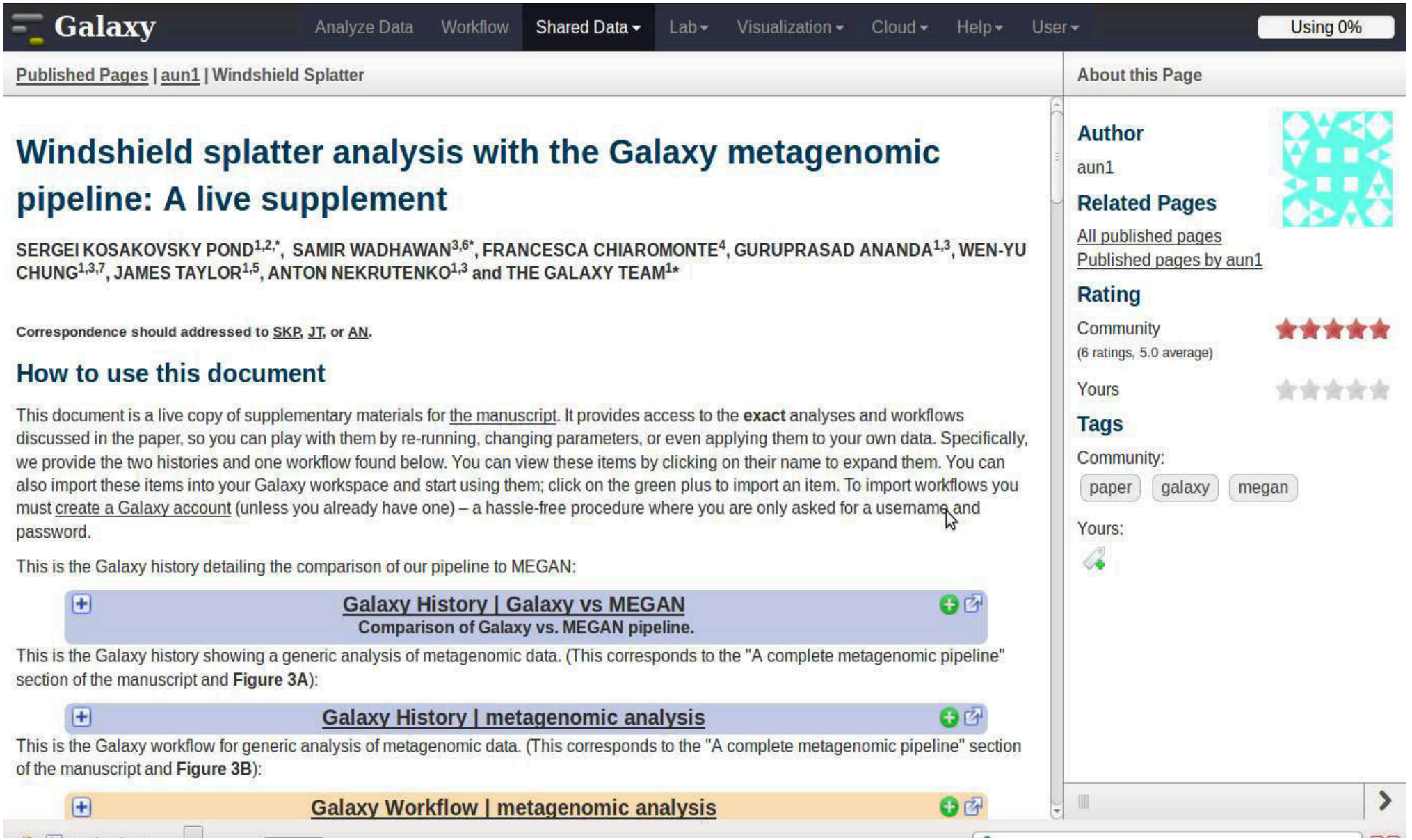
Drag & drop design panel



Share & Publish



Galaxy Page

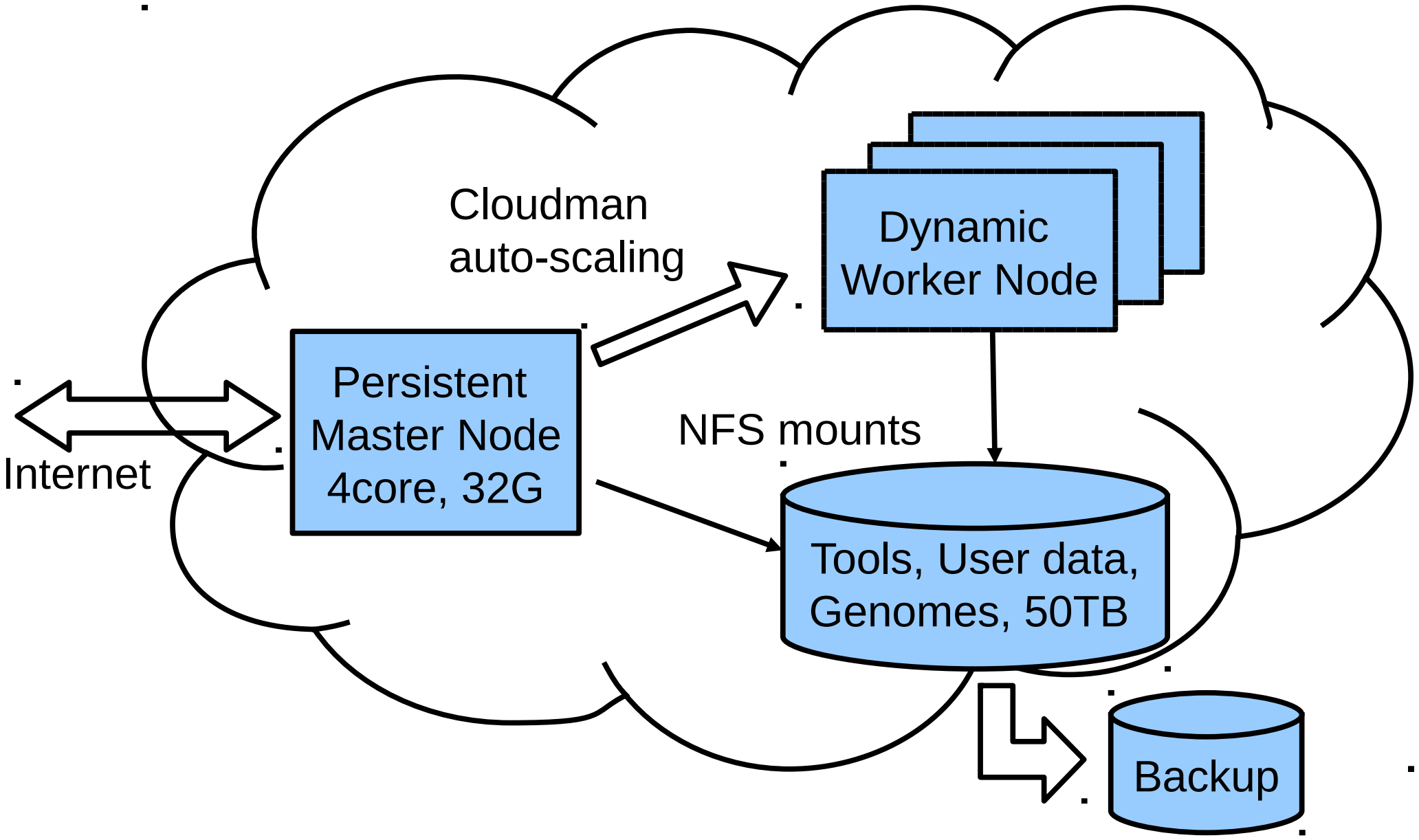


https://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter

More about computing

- Galaxy can run on a local server, cluster, cloud
- Full DRMAA support
- Support of heterogeneous resource
  - Define run requirements in job_conf.xml

Galaxy Cloudman



RNAseq Workflow

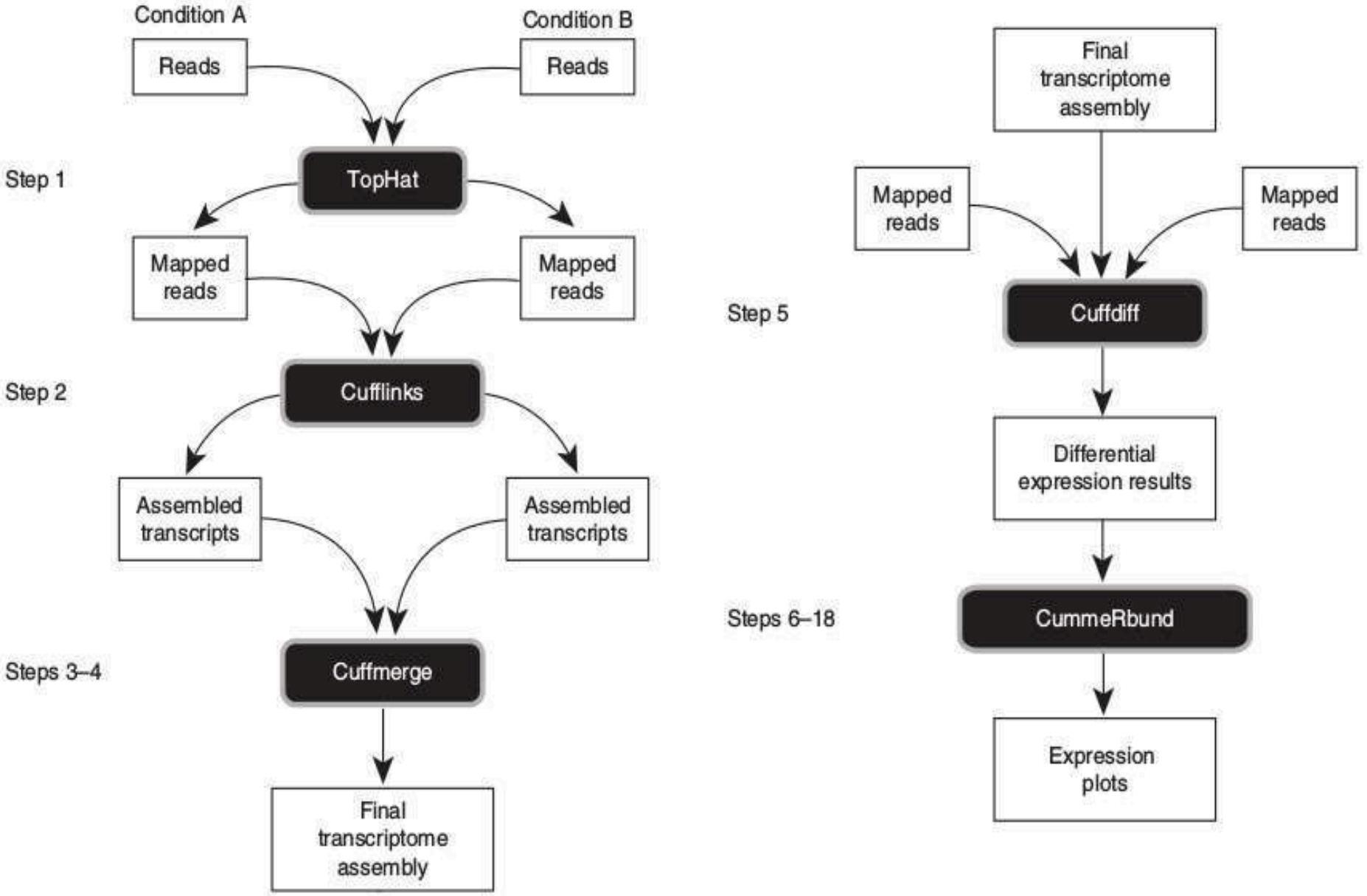
Leon Mei

Sasc
Sequence Analysis Support Core

General RNAseq pipeline

- Input data
  - Sequencing files (fastq), gene annotation (gtf)
- Pre-alignment
  - QC and Data cleaning
- Alignment
  - Use a specialized (RNA) aligner
- Expression (gene, transcripts) analysis
  - Quantification, up/down regulated genes, p-values
- Transcript assembly
  - New transcripts, alternative splicing
- Pathway analysis
  - Clustering algorithms, Taverna, etc.

Tuxedo pipeline



Scaling up

- Computational requirement of RNAseq analysis
  - Aligning 2-10M pairs per hour, ~1 day per sample
- Needs from large sequencing projects
  - http://www.geuvadis.org, 465 RNAseq samples
  - BBMRI-NL BIOS, 4000 RNAseq samples
- Running all samples in parallel (several times)
  - Galaxy is not best place to scale up
  - Dutch life science grid is used for BIOS

Hands-on

- http://tinyurl.com/eccbgalaxy