Module 1/6: Analyses ADN

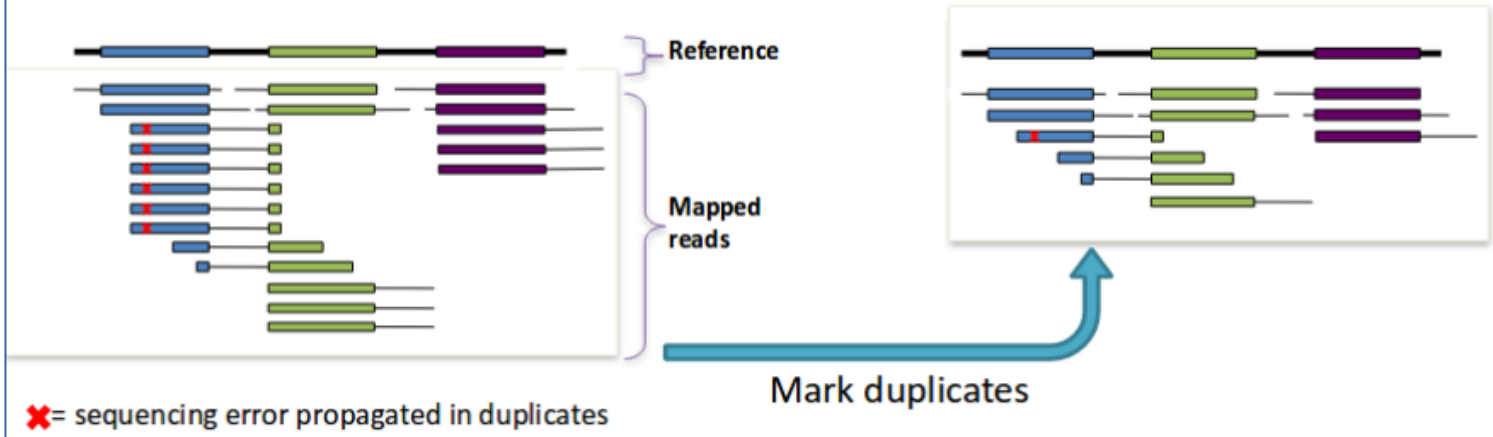
- NGS Introduction
- Reads Quality Control
- Reads Cleaning
- Aligning reads on reference → *Hélène Touzet*
- Alignment parameters → *Hélène Touzet*
- Reads duplicates

→ **Practical #3**

Cleaning duplicated reads

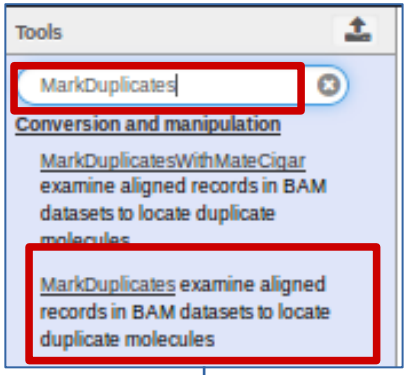
Why mark duplicates?

- Duplicates = sets of reads pairs with same unclipped alignment start and unclipped alignment end
- Suspected to be **non-independent measurements** of a sequence
  - Sampled from the exact same template of DNA
  - Violates assumptions of variant calling
- Errors in sample/library prep will get propagated to *all* the duplicates
  - Just pick the "best" copy – mitigates the effects of errors

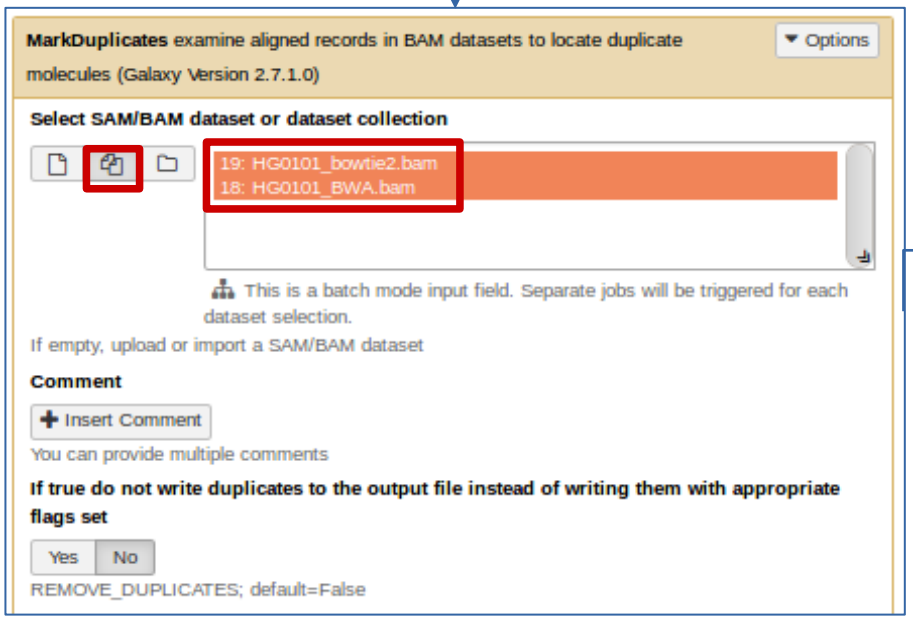


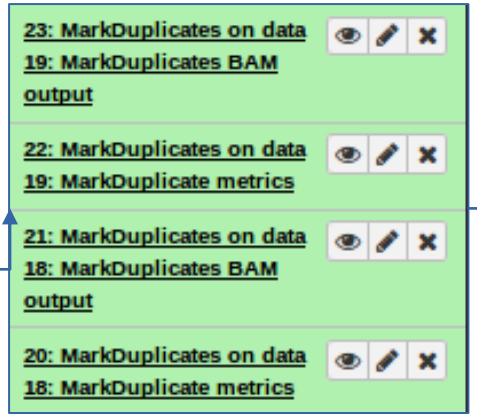
Source : GATK Marking duplicates
https://software.broadinstitute.org/gatk/events/slides/1511/Presentations/GATKwh9-3-Marking_duplicates.pdf

Picard / MarkDuplicate



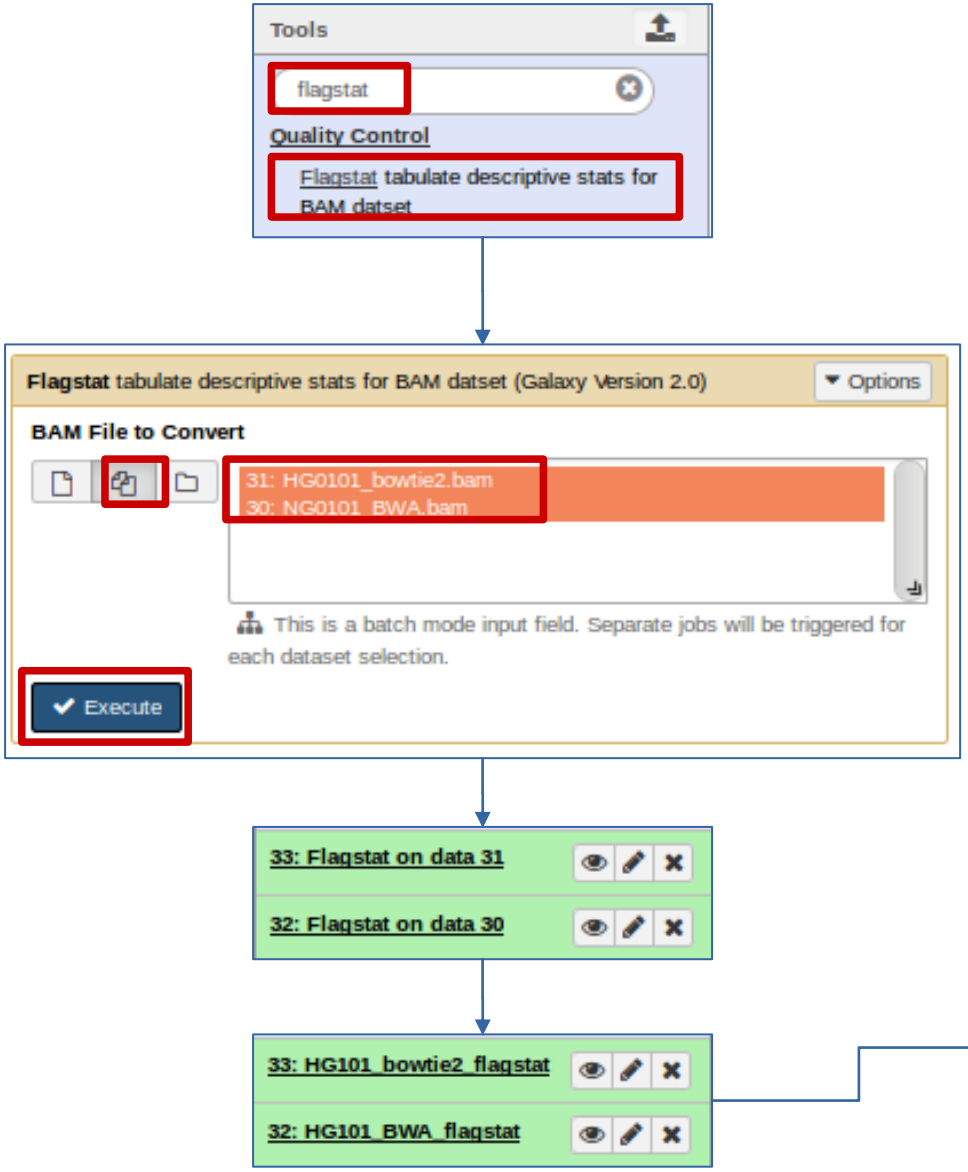
Additional information about Picard tools is available from Picard web site at *http://broadinstitute.github.io/picard/*





	BWA	Bowtie2
UNPAIRED_READS_EXAMINED	18	19
READ_PAIRS_EXAMINED	4043	4044
SECONDARY_OR_SUPPLEMENTARY_RDS	0	3
UNMAPPED_READS	22	19
UNPAIRED_READ_DUPLICATES	0	0
READ_PAIR_DUPLICATES	12	12
READ_PAIR_OPTICAL_DUPLICATES	0	0
PERCENT_DUPLICATION	0,002962	0,00296
ESTIMATED_LIBRARY_SIZE	679728	680065

Alignment count : *samtools flagstat*



BWA

```
8129 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
3 + 0 supplementary
24 + 0 duplicates
8110 + 0 mapped (99.77%:-nan%)
8126 + 0 paired in sequencing
4063 + 0 read1
4063 + 0 read2
7980 + 0 properly paired (98.20%:-nan%)
8088 + 0 with itself and mate mapped
19 + 0 singletons (0.23%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Bowtie 2

```
8126 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
24 + 0 duplicates
8104 + 0 mapped (99.73%:-nan%)
8126 + 0 paired in sequencing
4063 + 0 read1
4063 + 0 read2
8074 + 0 properly paired (99.36%:-nan%)
8086 + 0 with itself and mate mapped
18 + 0 singletons (0.22%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Coverage and deepth of coverage

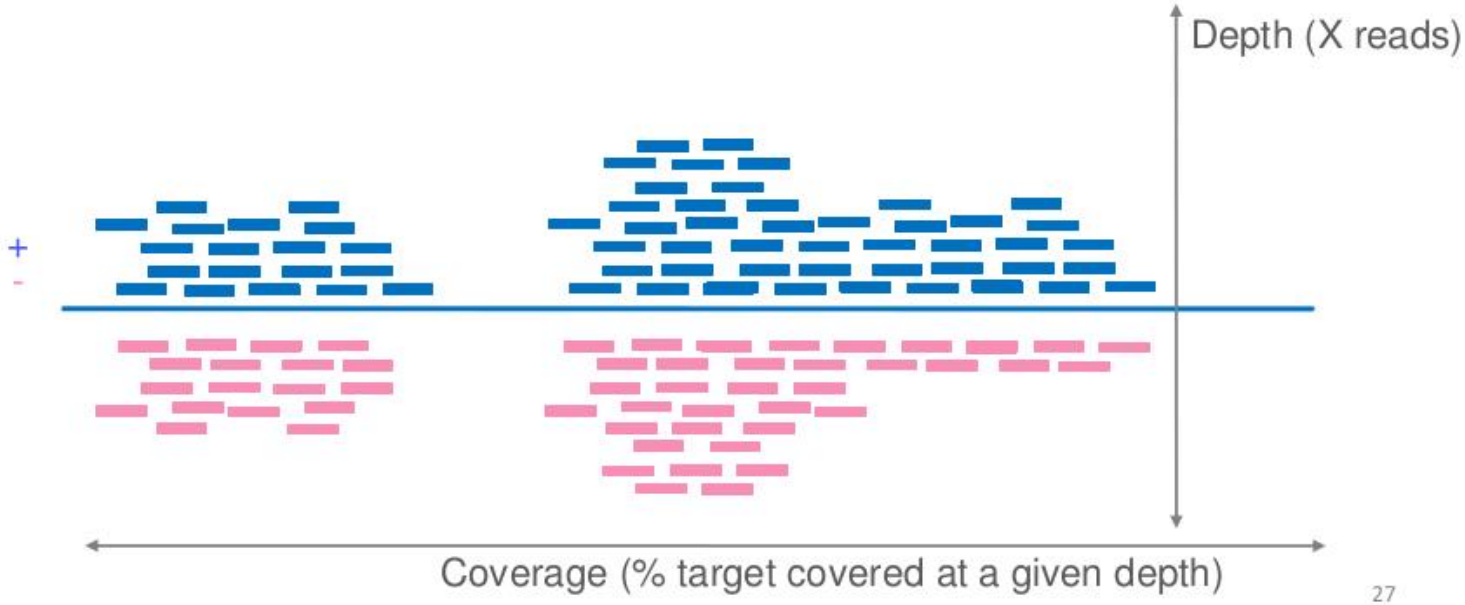
Statistics used as quality control

- **Depth of coverage** = average number of reads covering a base (X)

  Example: 30X for normal sample, 100X for tumor sample

- **Coverage** = percentage of the targeted regions covered by at least X read

  Example: >=80% of your exome target is covered by 20X for normal sample

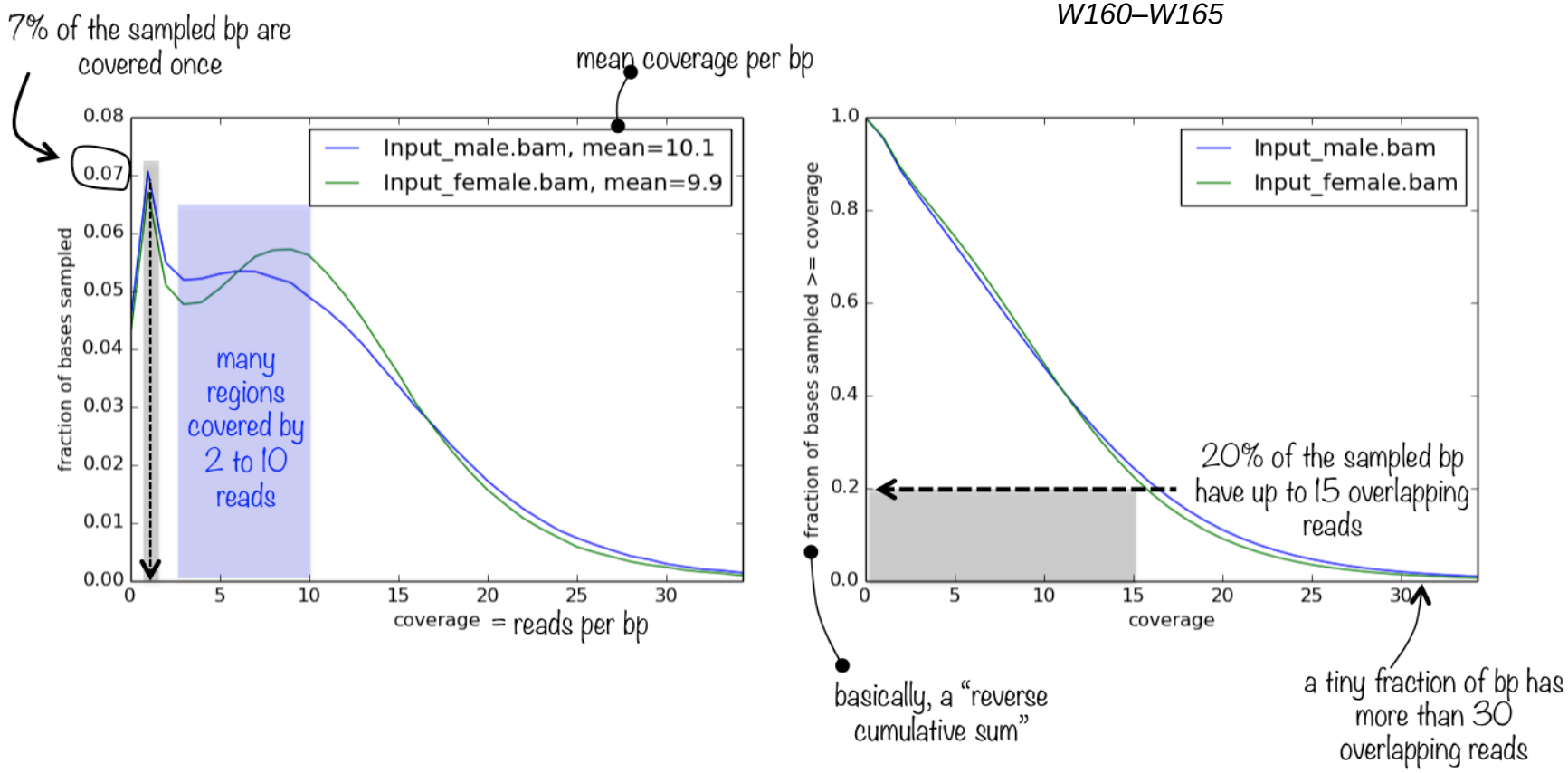


27

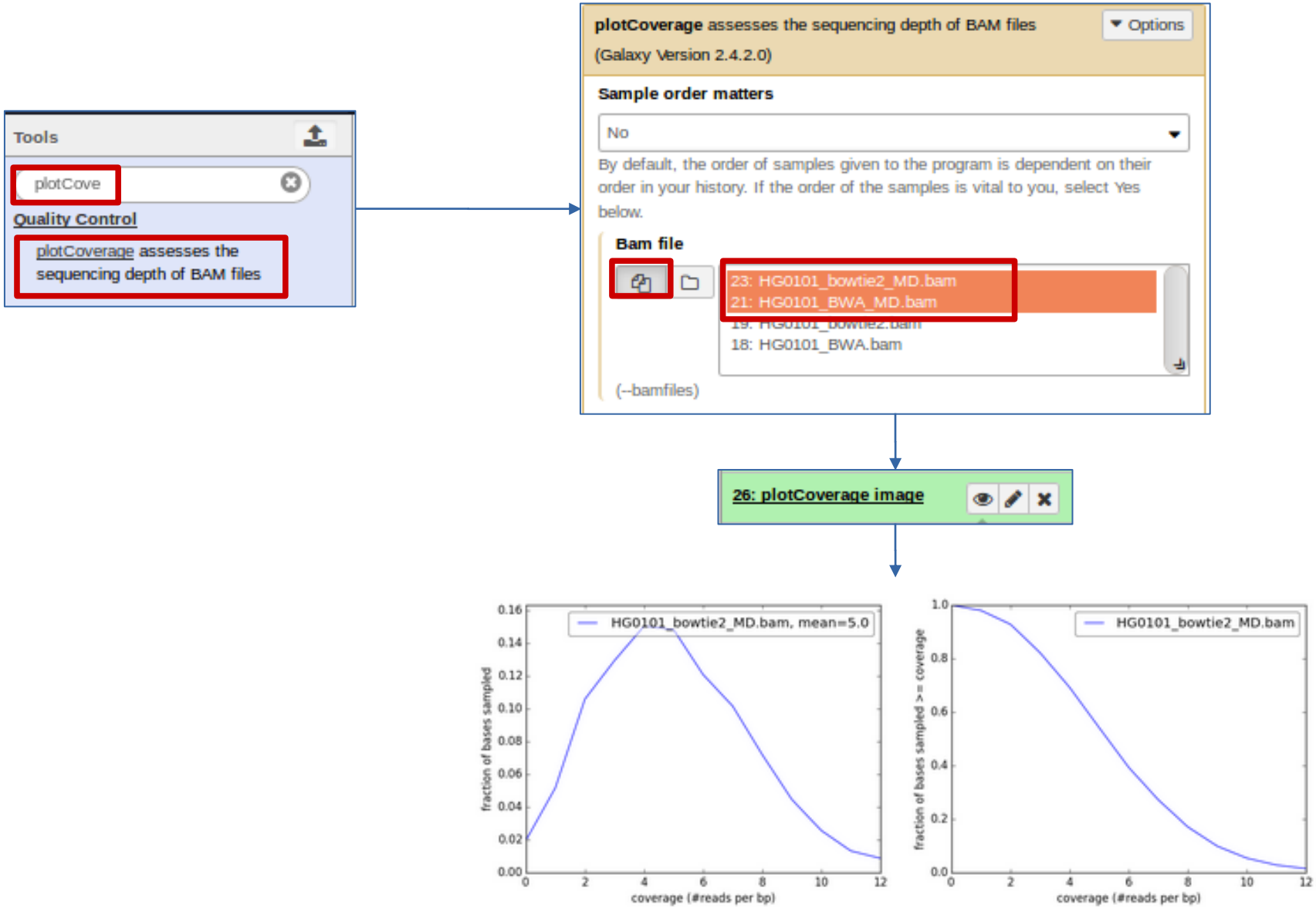
Source : Élodie Girard , 5ème Ecole de bioinformatique AVIESAN-IFB 2016
http://www.france-bioinformatique.fr/sites/default/files/V01_ITMO_2016_EG_from_fastq_to_mapping_1.pdf

Computing coverage and deepth of coverage
DeepTools2 / plotCoverage

Ramírez, Fidel and Ryan, Devon P and Grüning, Björn and Bhardwaj, Vivek and Kilpert, Fabian and Richter, Andreas S and Heyne, Steffen and Dündar, Friederike and Manke, Thomas (2016). deepTools2: a next generation web server for deep-sequencing data analysis. In Nucleic Acids Research, 44 (W1), pp. W160–W165



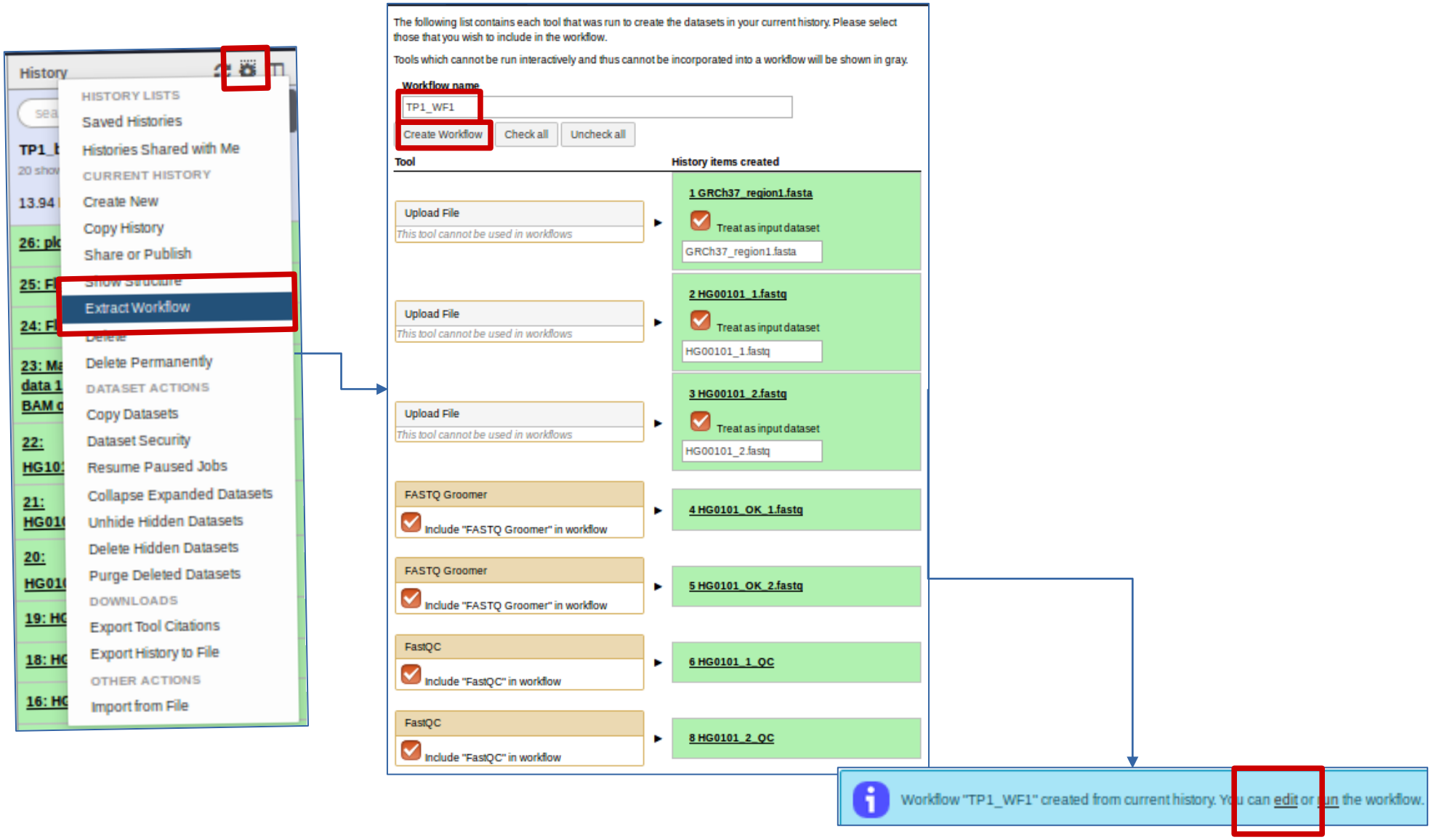
DeepTools / Plot Coverage



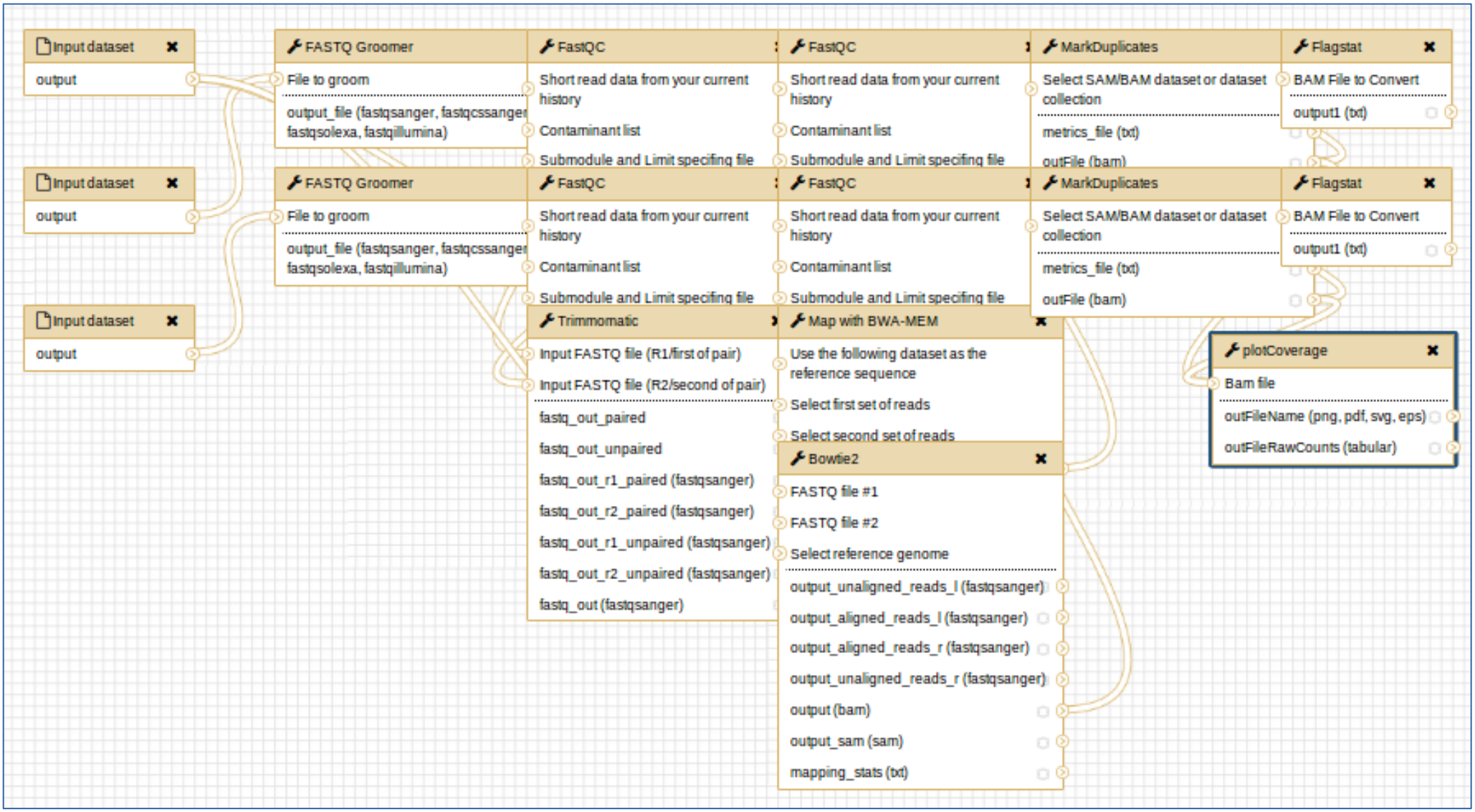
Galaxy *Workflow*

- Extract *workflow* from an history
- Modify *workflow*
- Execute *workflow* on new data
- Compare results from 2 *workflows* (in 2 histories)

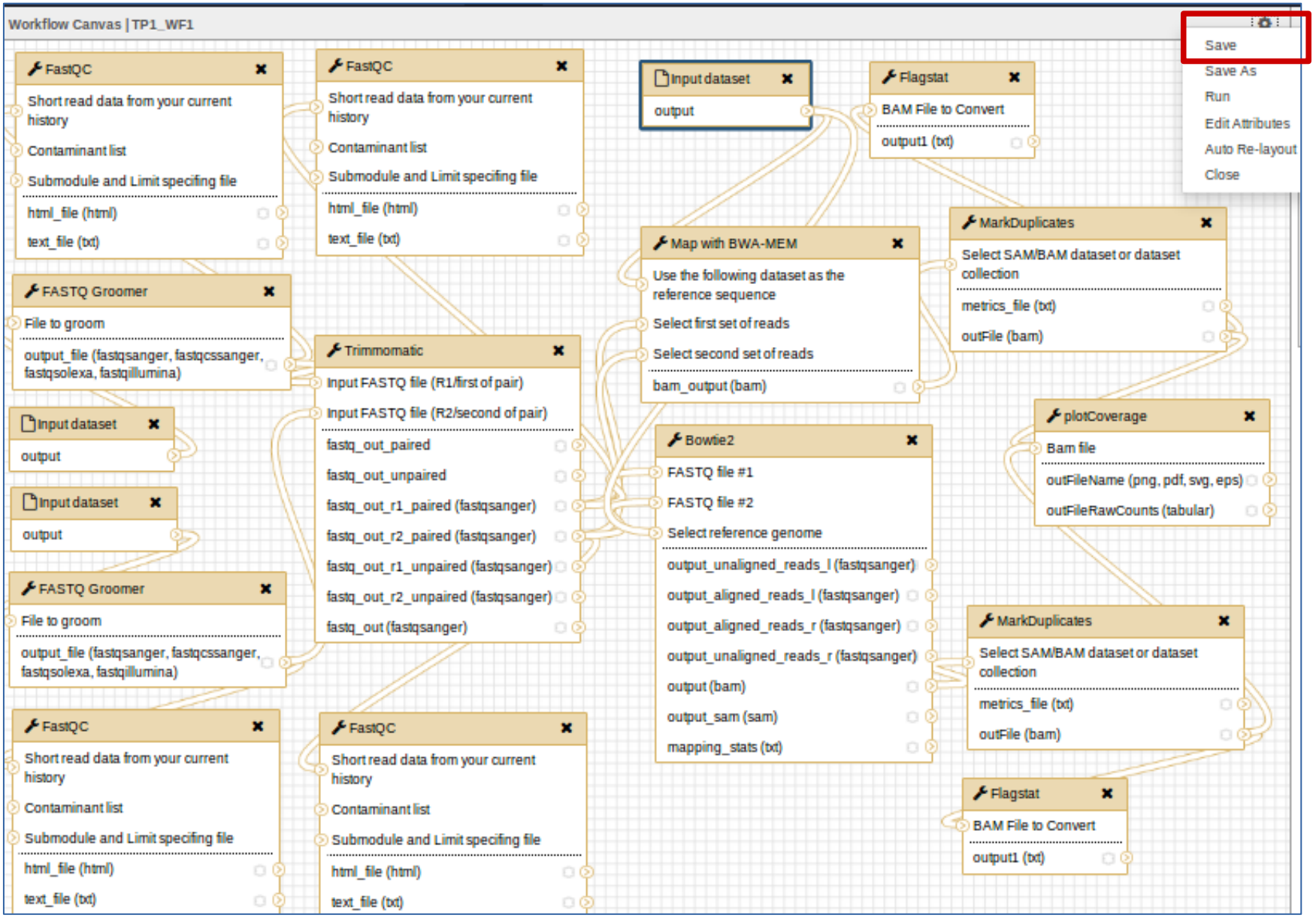
Extract *Workflow* from the history of steps applied to the first sample



Visualize workflow

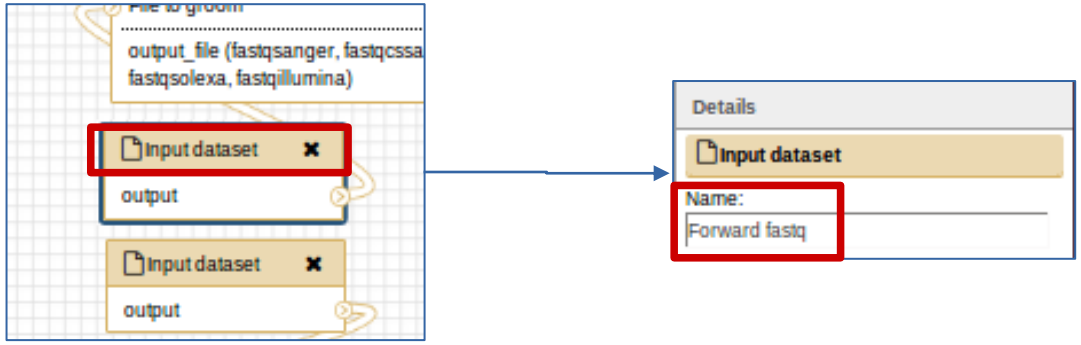


Modify workflow visualisation

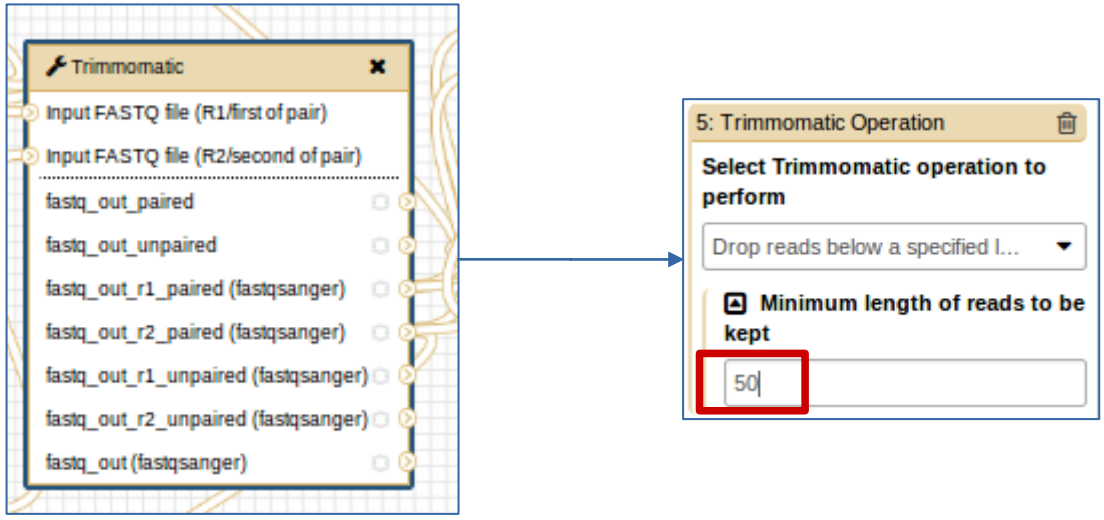


Modify some steps configuration

This WF uses 3 input files. Change box name to describe which data is required for each input : eg *Reference, Forward fastq, Reverse fastq*



You can also change any parameter for example for *trimmomatic* step.

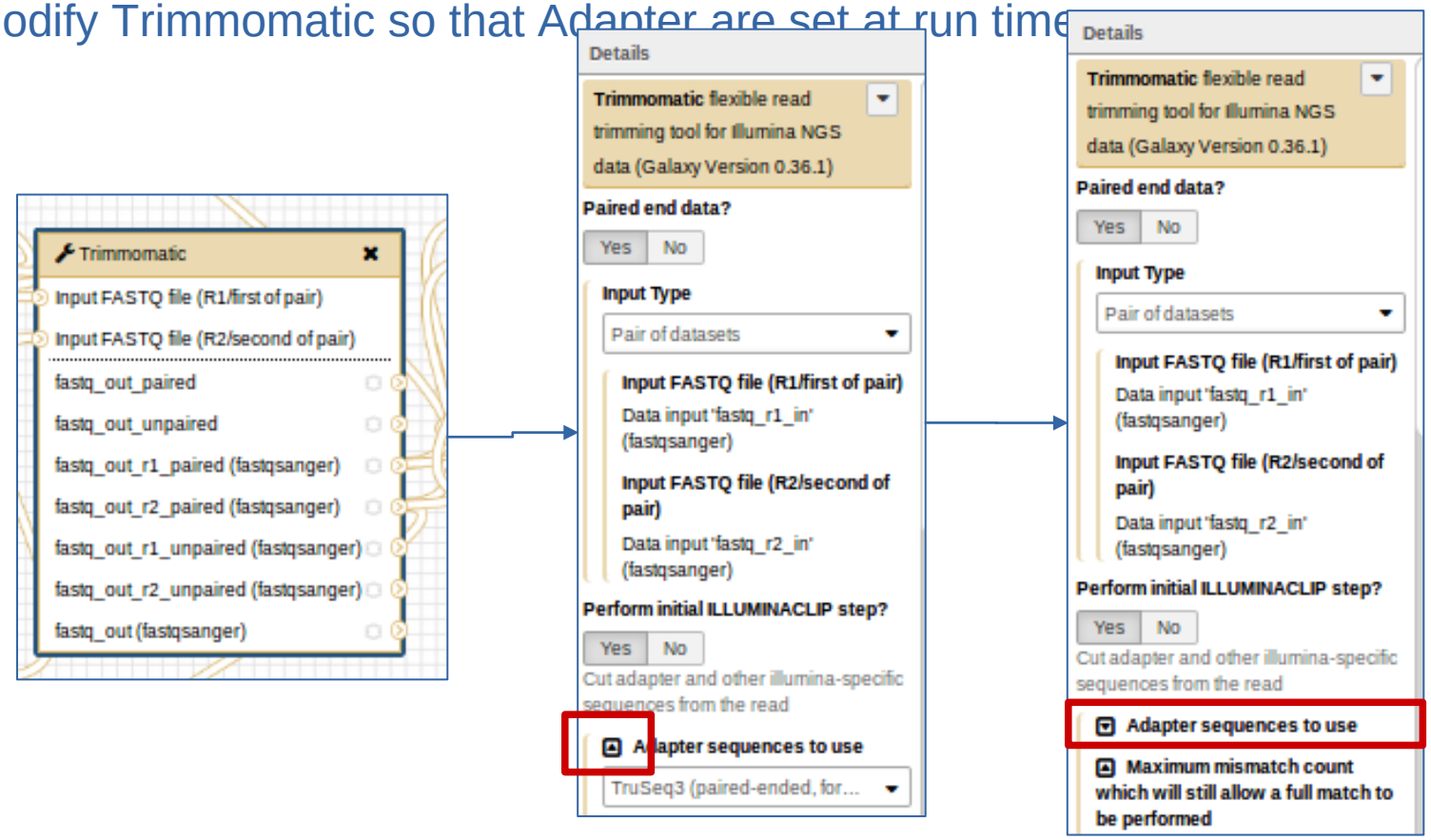


Enable a parameter to be set at run time

Parameters for each tool will have the predefined values set in the workflow

You can modify this to enable any parameter to be set at run time.

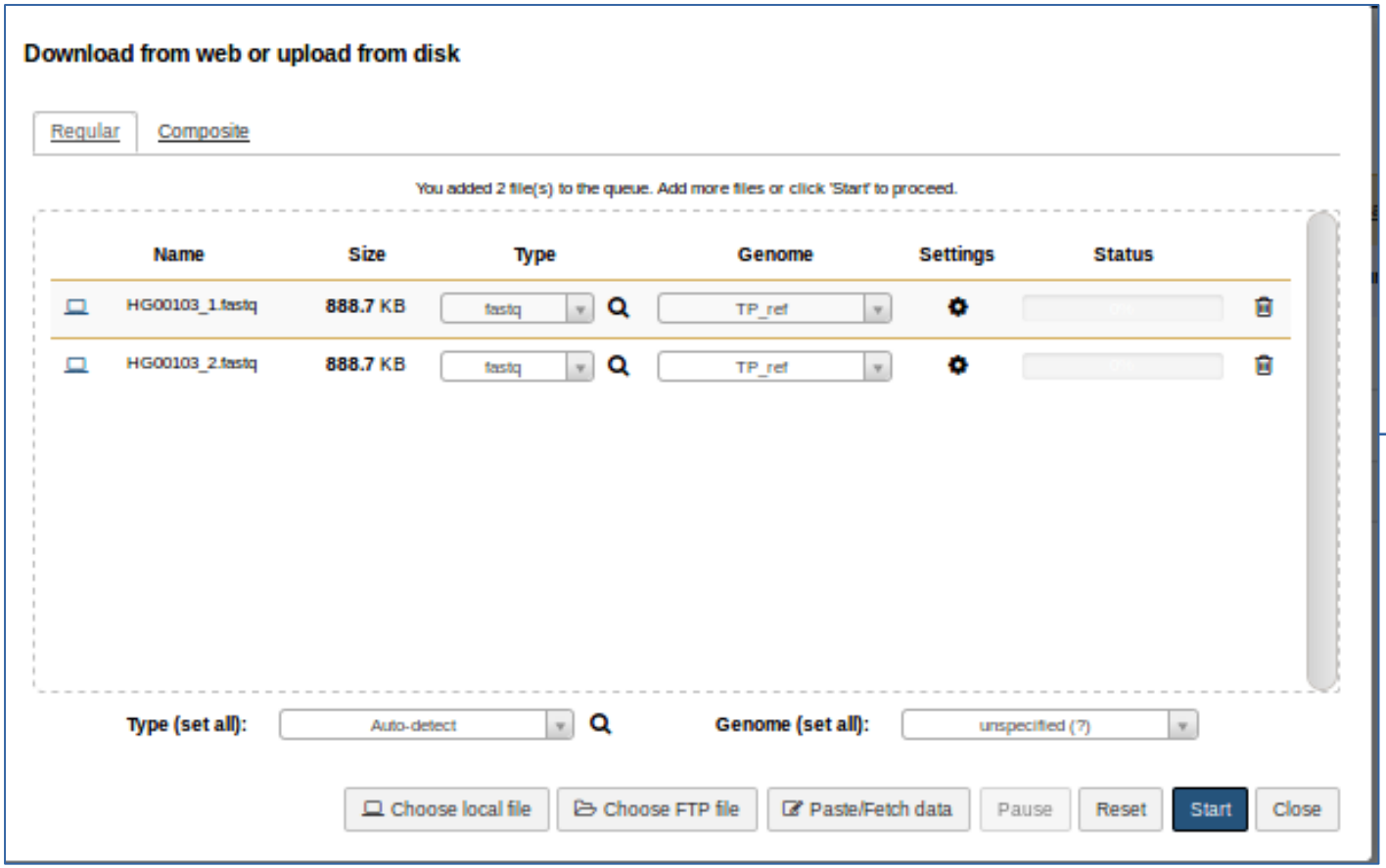
Modify Trimmomatic so that Adapter are set at run time

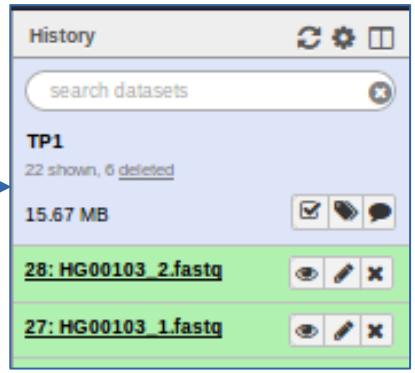


Do'nt forget to save your workflow !

Import new data for sample HG0103

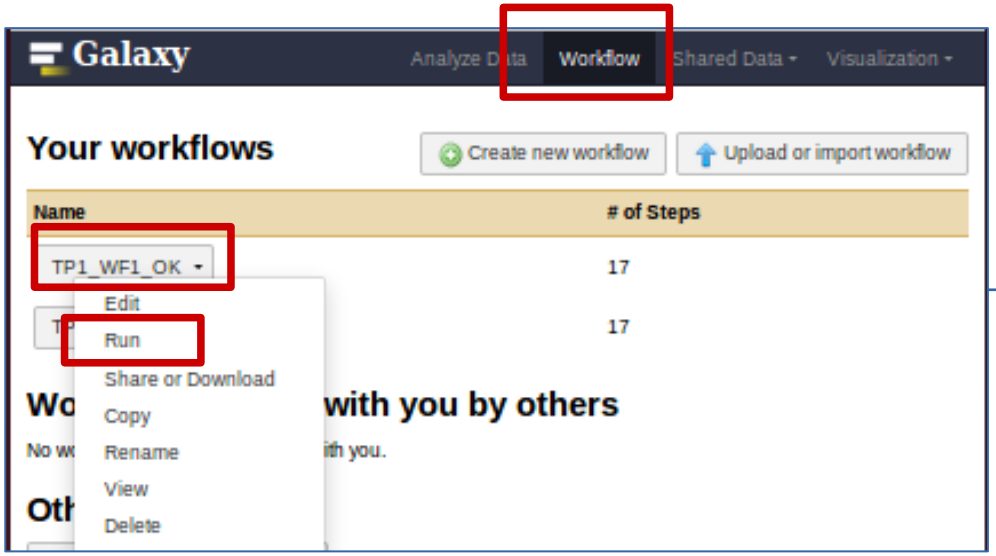
Importe files HG0103_1.fastq and HG_0103_2.fastq

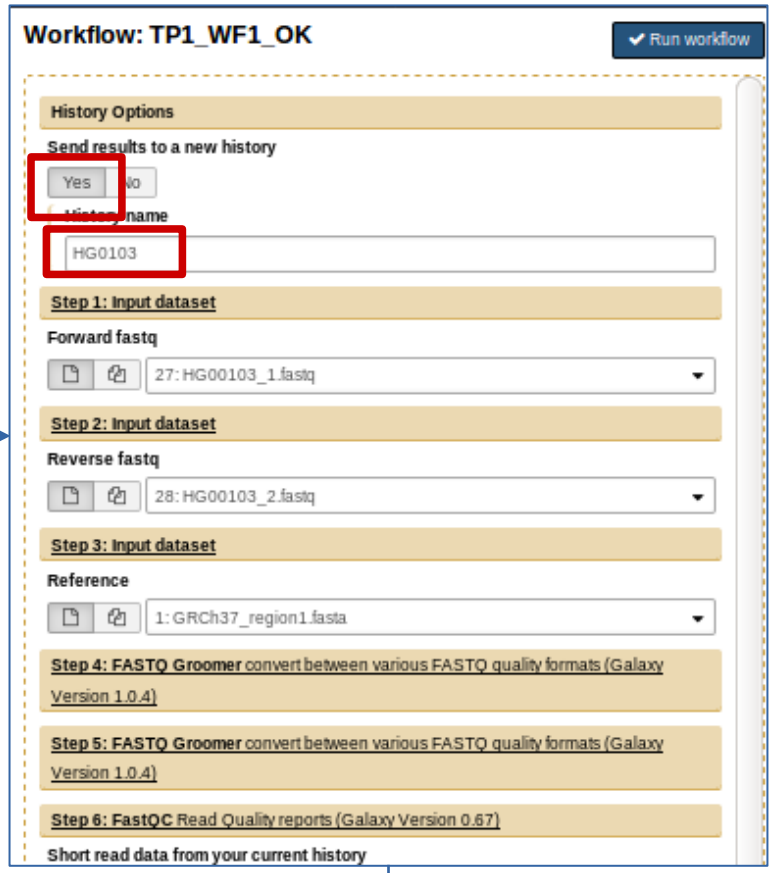


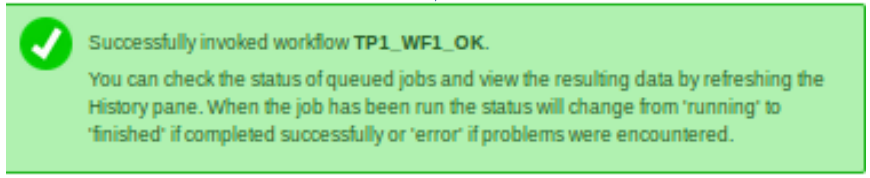


Analyze these new data with the same workflow

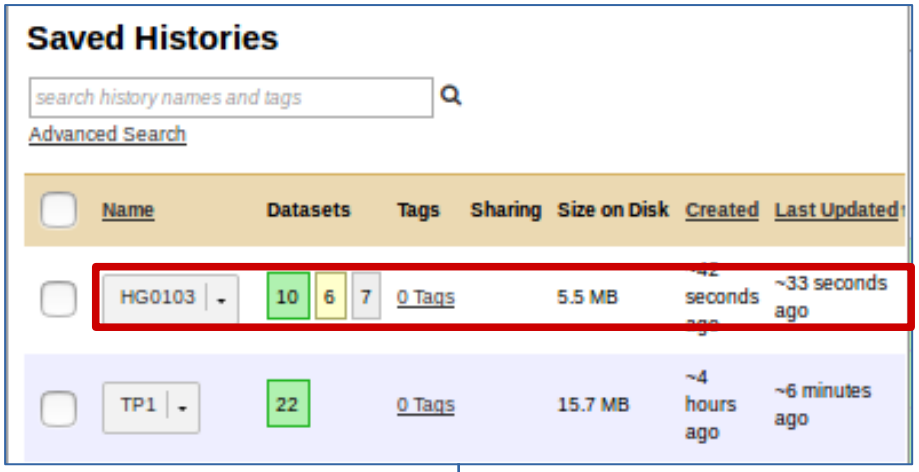
Run the workflow with these new data

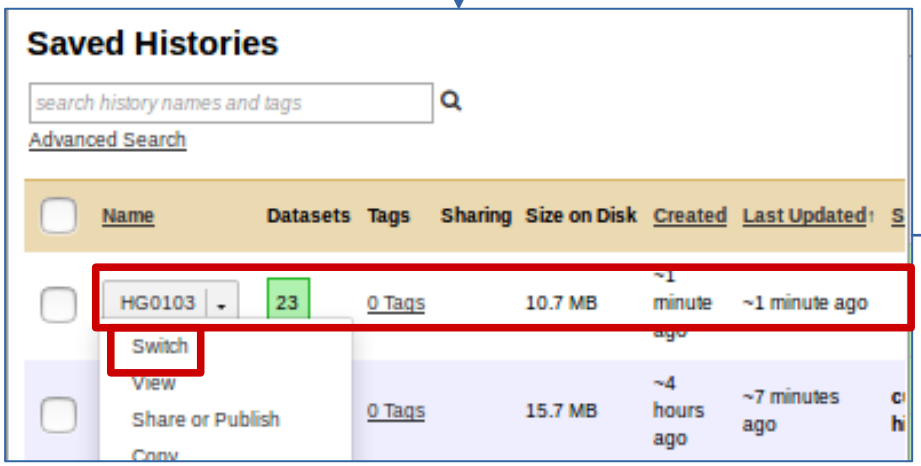


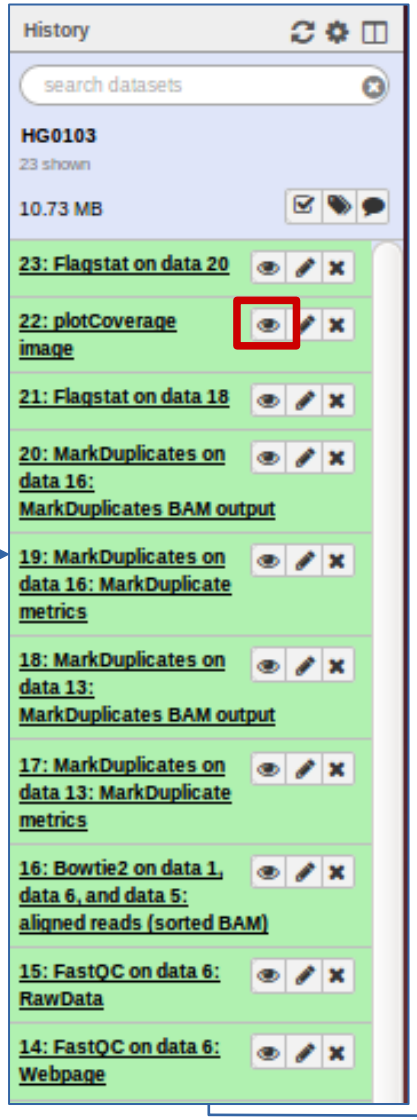


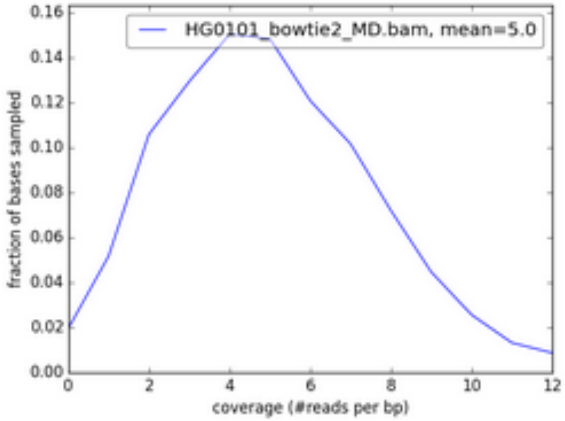


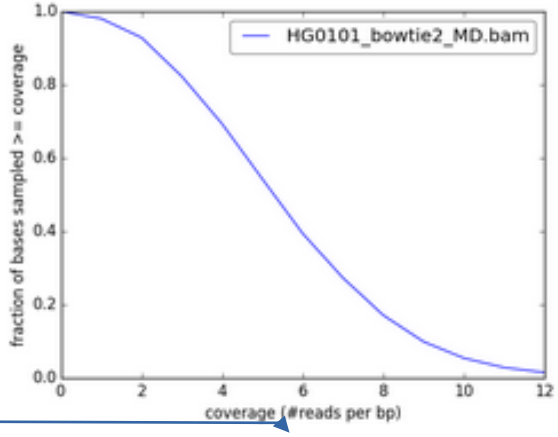
Browse results





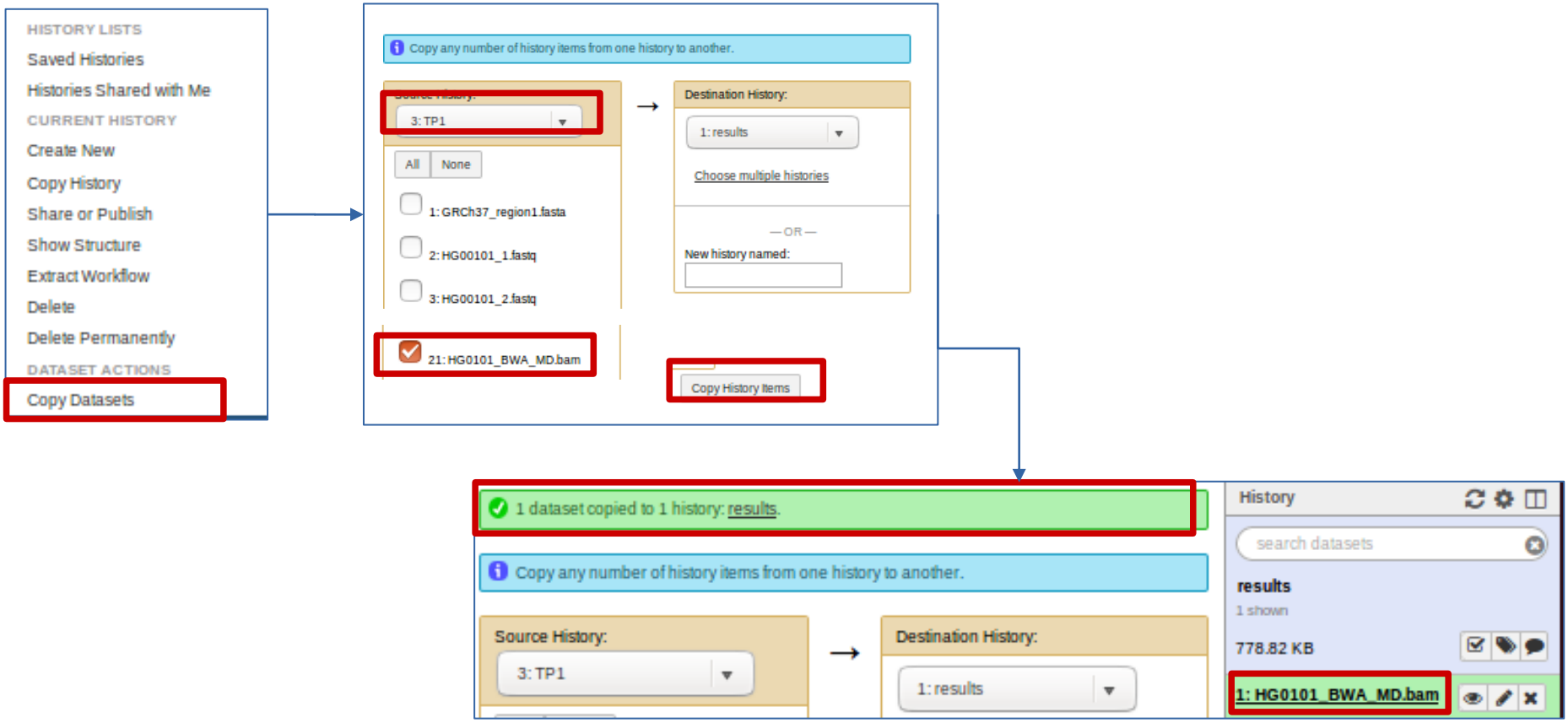






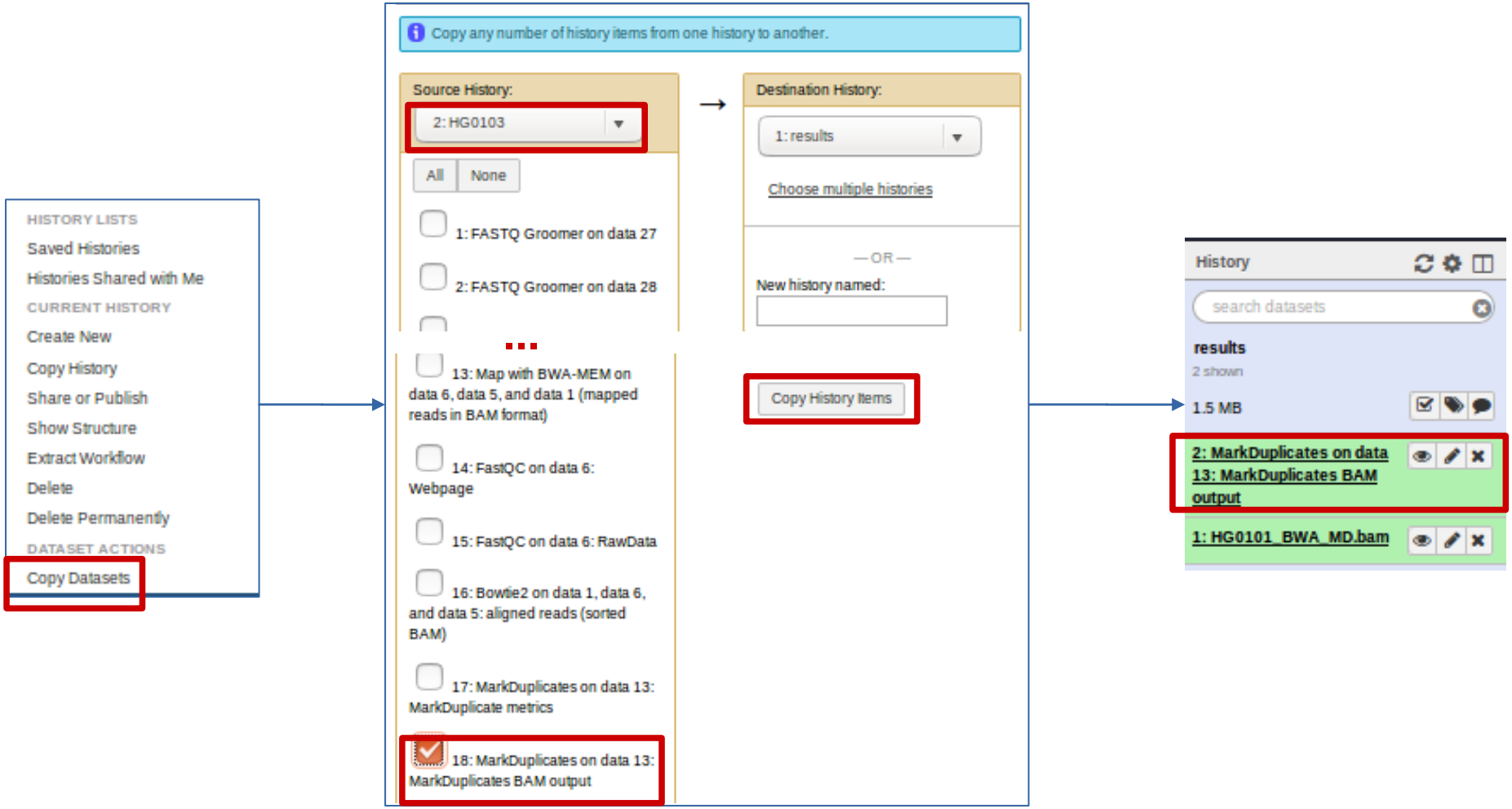
Gather BWA alignment results for the 2 samples 1/2

Create a new history named results
From history TP1 : Copy HG0101_BWA_MD.bam *dataset*



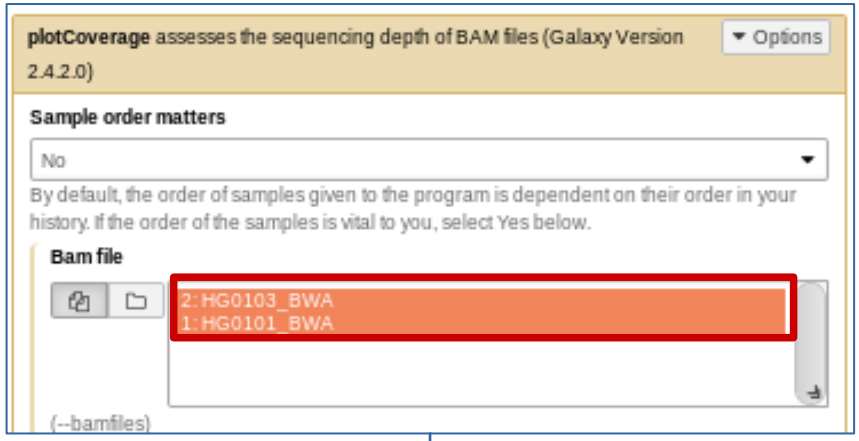
Gather BWA alignment results for the 2 samples 2/2

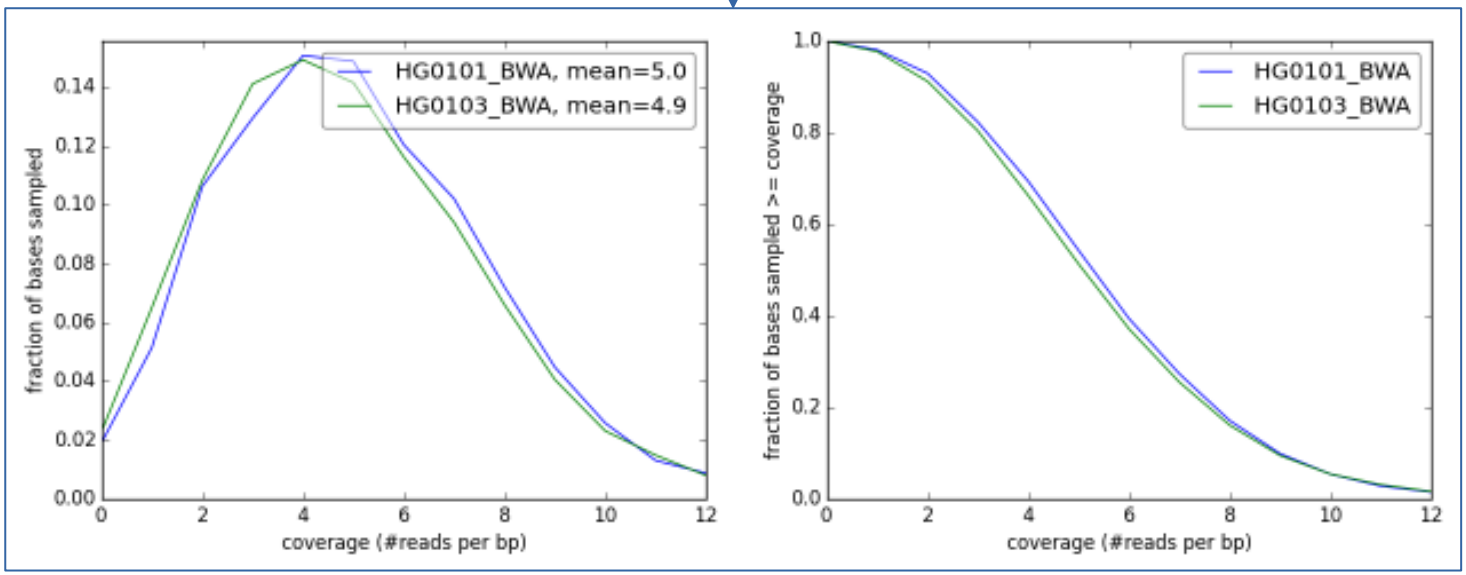
From history HG0103 : Copy *dataset*
« MarkDuplicates on data 13: MarkDuplicates BAM output »



Visualize deepth of coverage for both samples

Rename *datasets*
Run *plotCoverage*

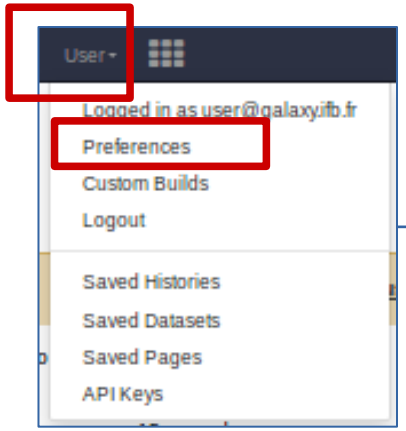




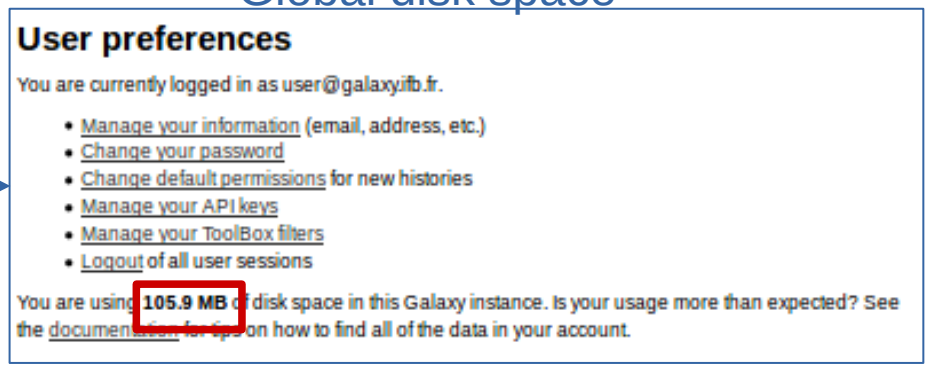
Galaxy – *Best Practices*

- Manage disk space
- Export analysis results (*datasets and histories*)
- Export / Import analysis protocoles (*workflow)*

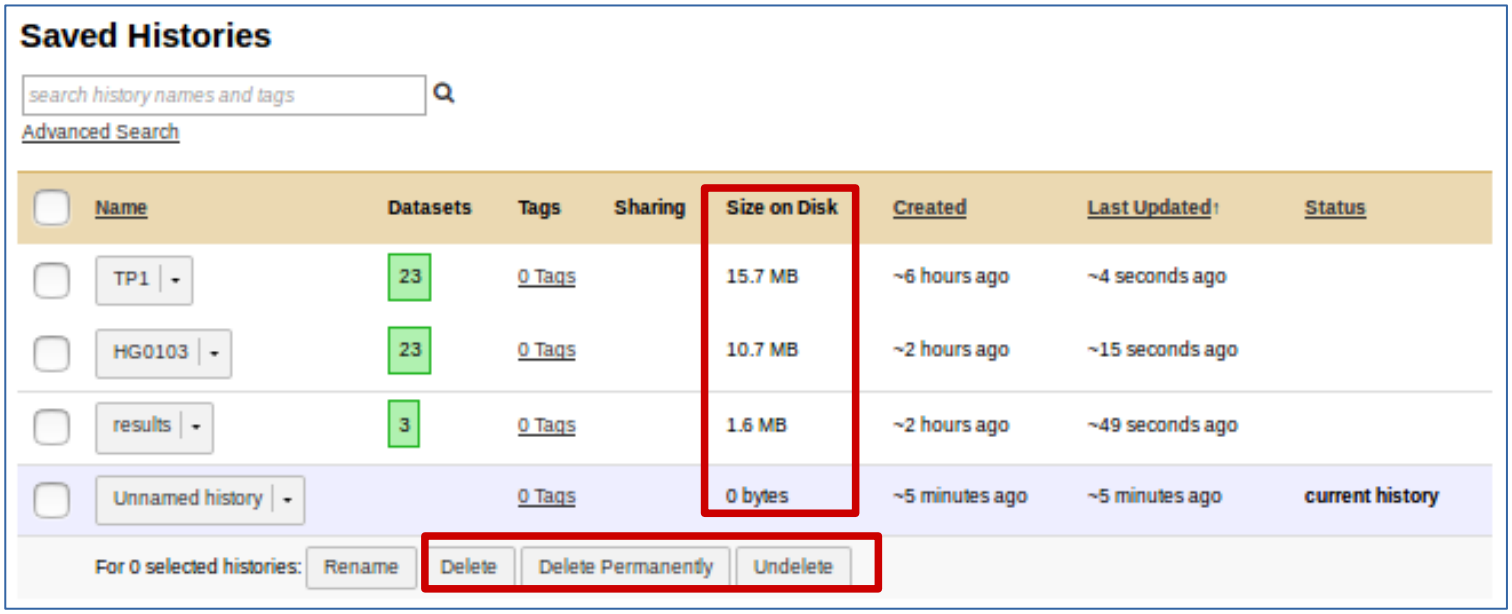
Manage disk space



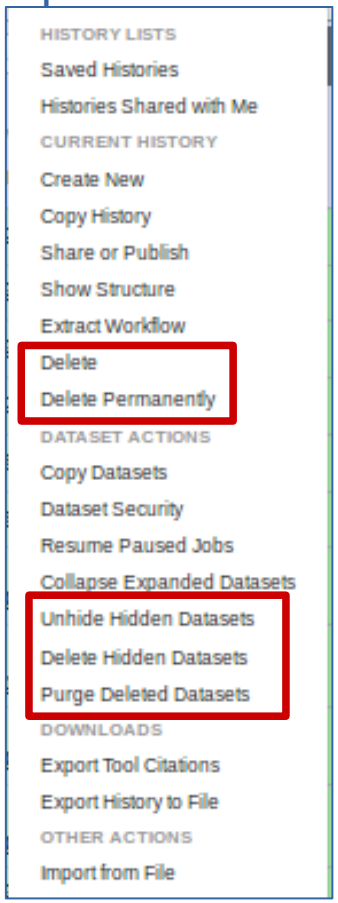
Global disk space



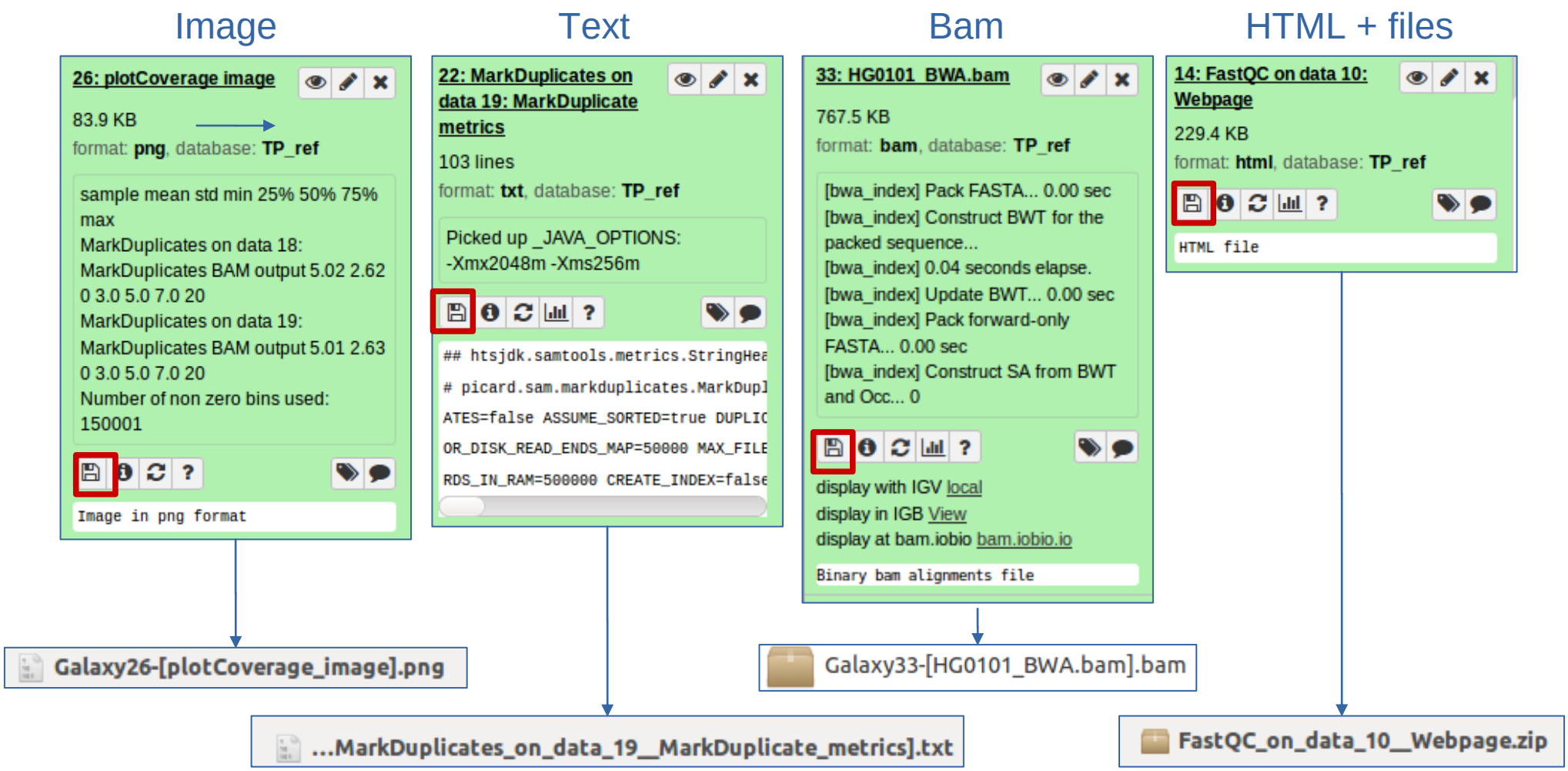
Disk space per history



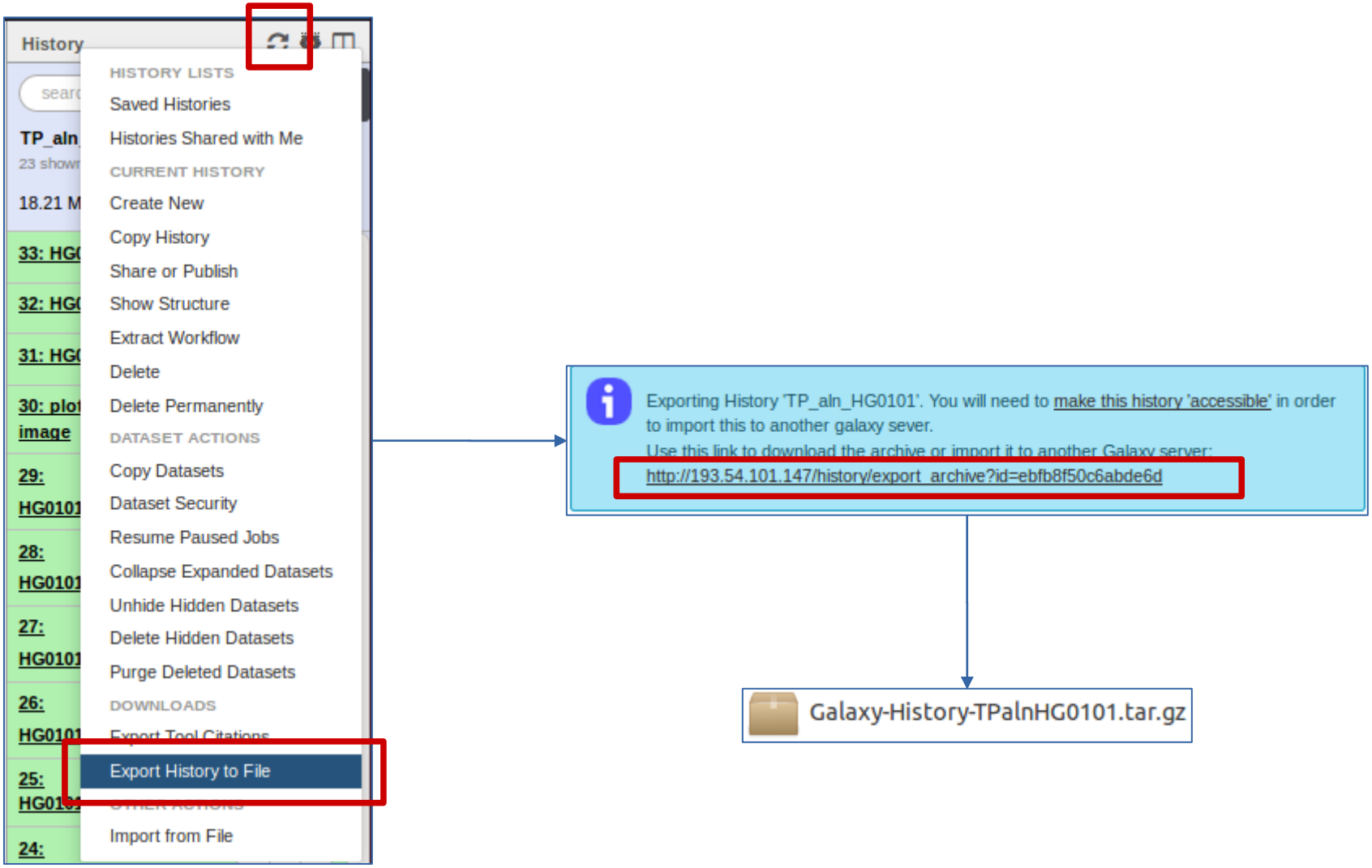
Disk space per *dataset*



Export analysis results : *datasets*

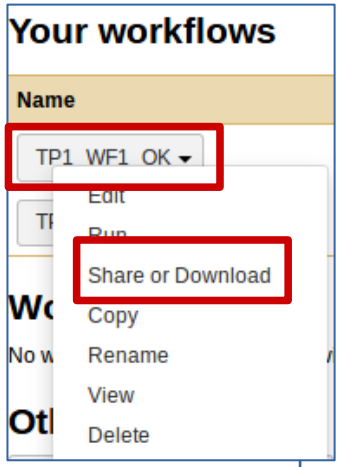


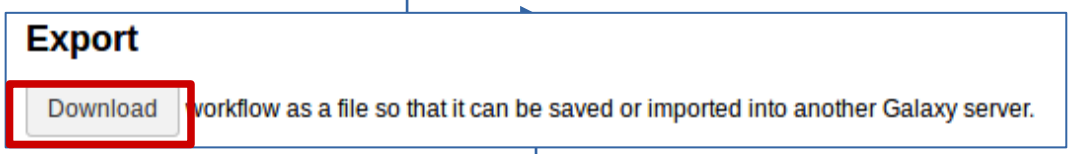
Export analysis results : *histories*



Export / import analysis protocoles : *workflow*

Export





Import