**Tutorial 2**

**Title**

Mutation calling, viral genome reconstruction and lineage/clade assignment from SARS-CoV-2 sequencing data

**Questions**

* How can we extract annotated allelic variants in SARS-Cov-2 sequences in Galaxy?
* Which tools and workflows can we use to identify SARS-CoV-2 lineages in Galaxy?

**Objectives**

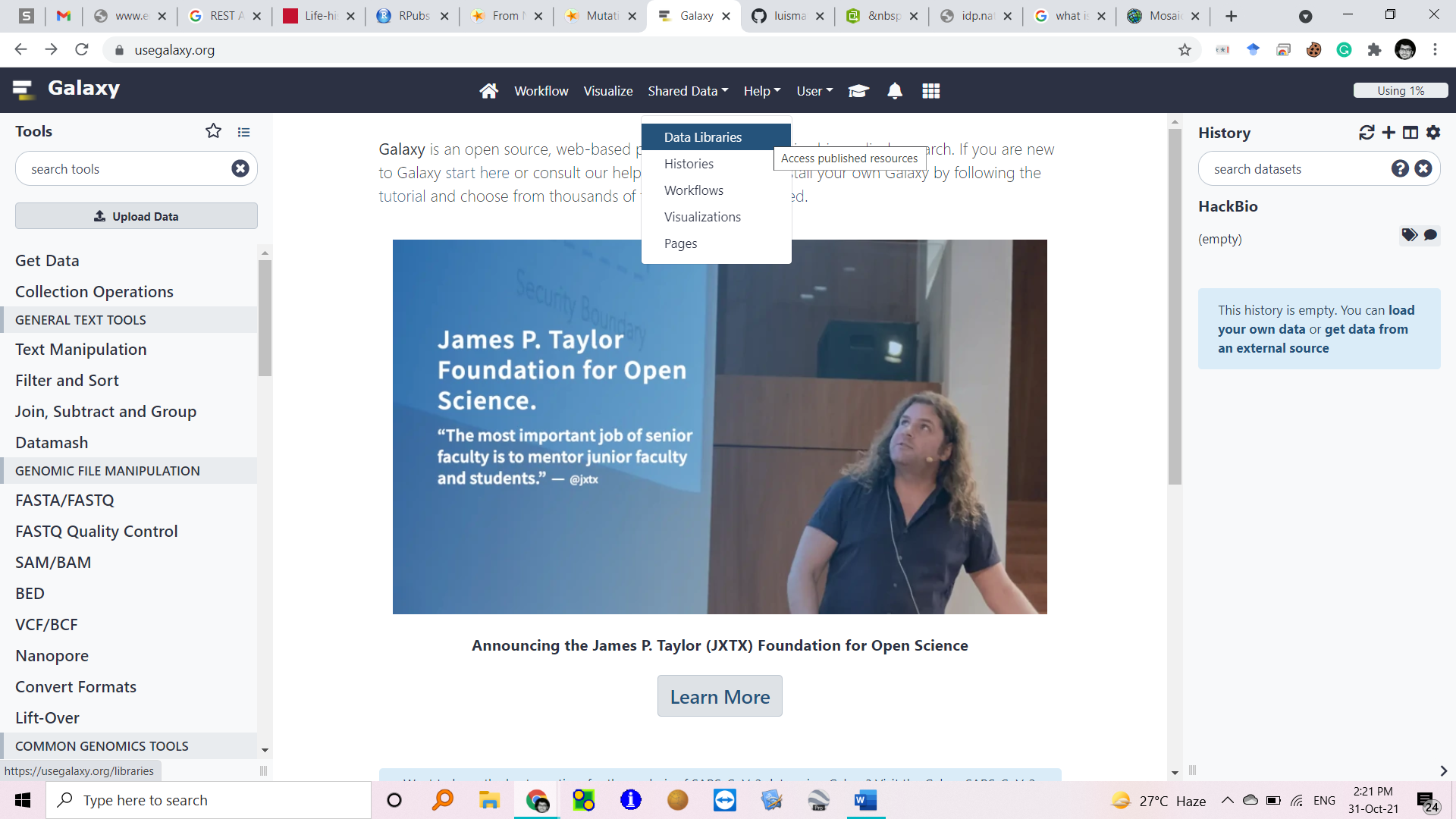
* Repeat SARS-CoV-2 data preparation
* Select and run workflow to extract annotated allelic variants from FASTQ files
* Run workflow to summarize and generate report for previously called allelic variants
* Interpret summaries for annotated allelic variants
* Run workflow to extract consensus sequences
* Select and run tools to assign clades/lineages

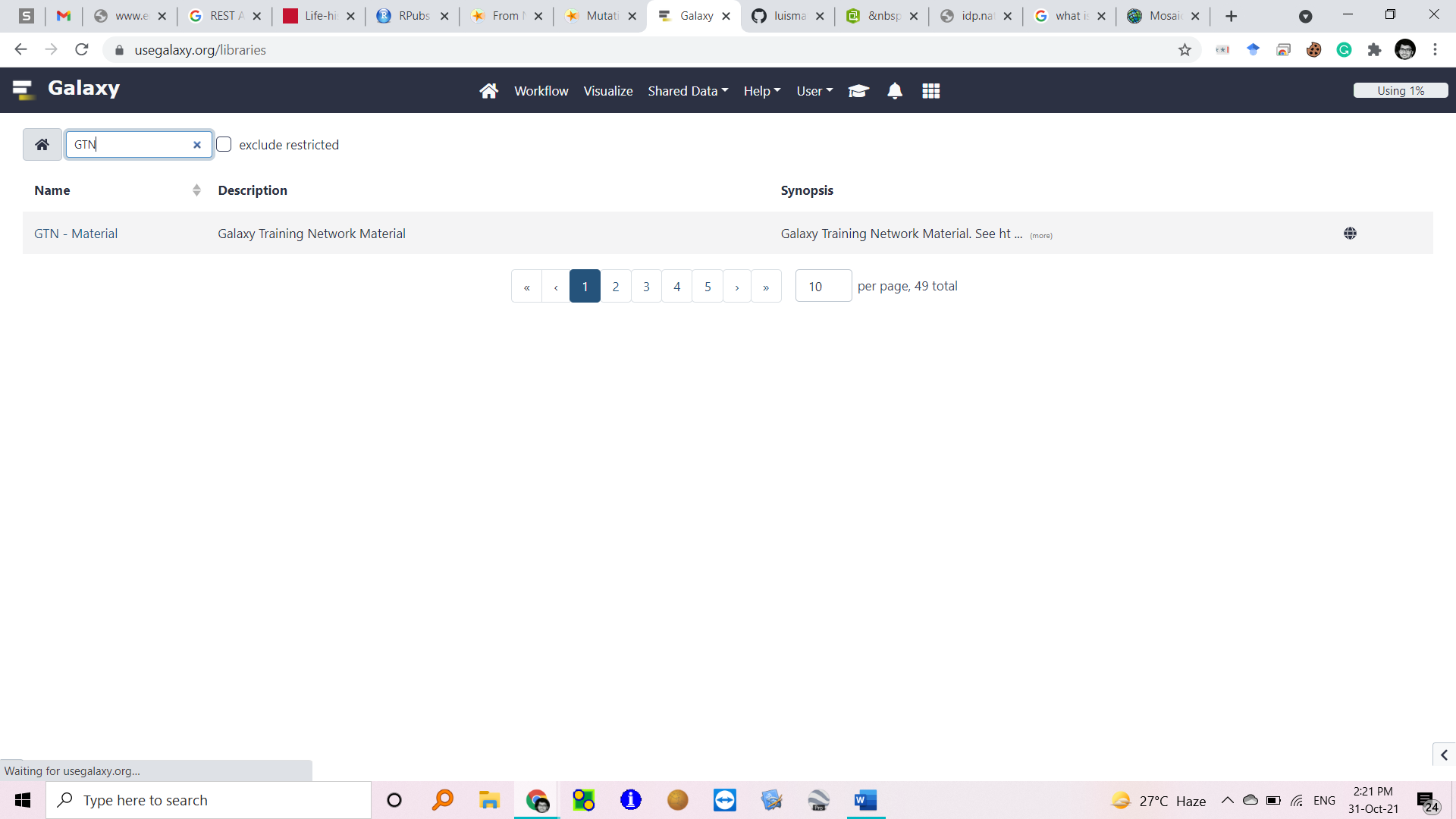
**Steps involved:**

1. **Preparing galaxy history**
   1. Create a new history for this analysis
2. **Data download**

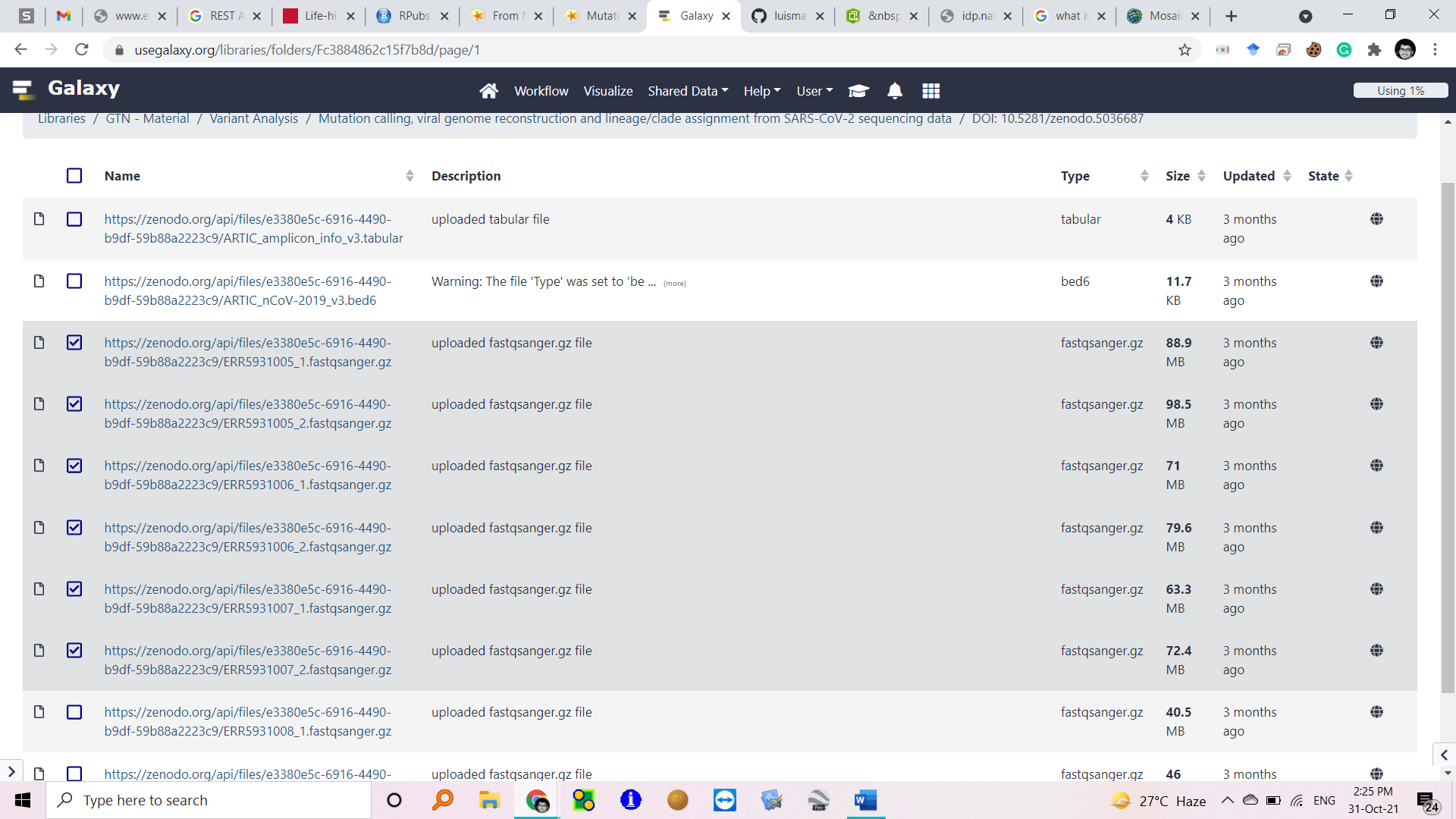
There are few options available for data obtaining data for the analysis. e.g., FTP, used downloaded file, import form external link or use Shared data. For execution of the tutorial, we have used data for **“Shared data”**

* 1. Click on the **“Shared data” ->** click on **“data libraries” ->** Search **“GTN”** in the search bar -> and direct to Material / Variant analysis / Mutation calling, viral genome reconstruction and lineage/clade assignment from SARS-CoV-2 sequencing data / DOI: 10.5281/zenodo.50366 -> download any three compete dataset.

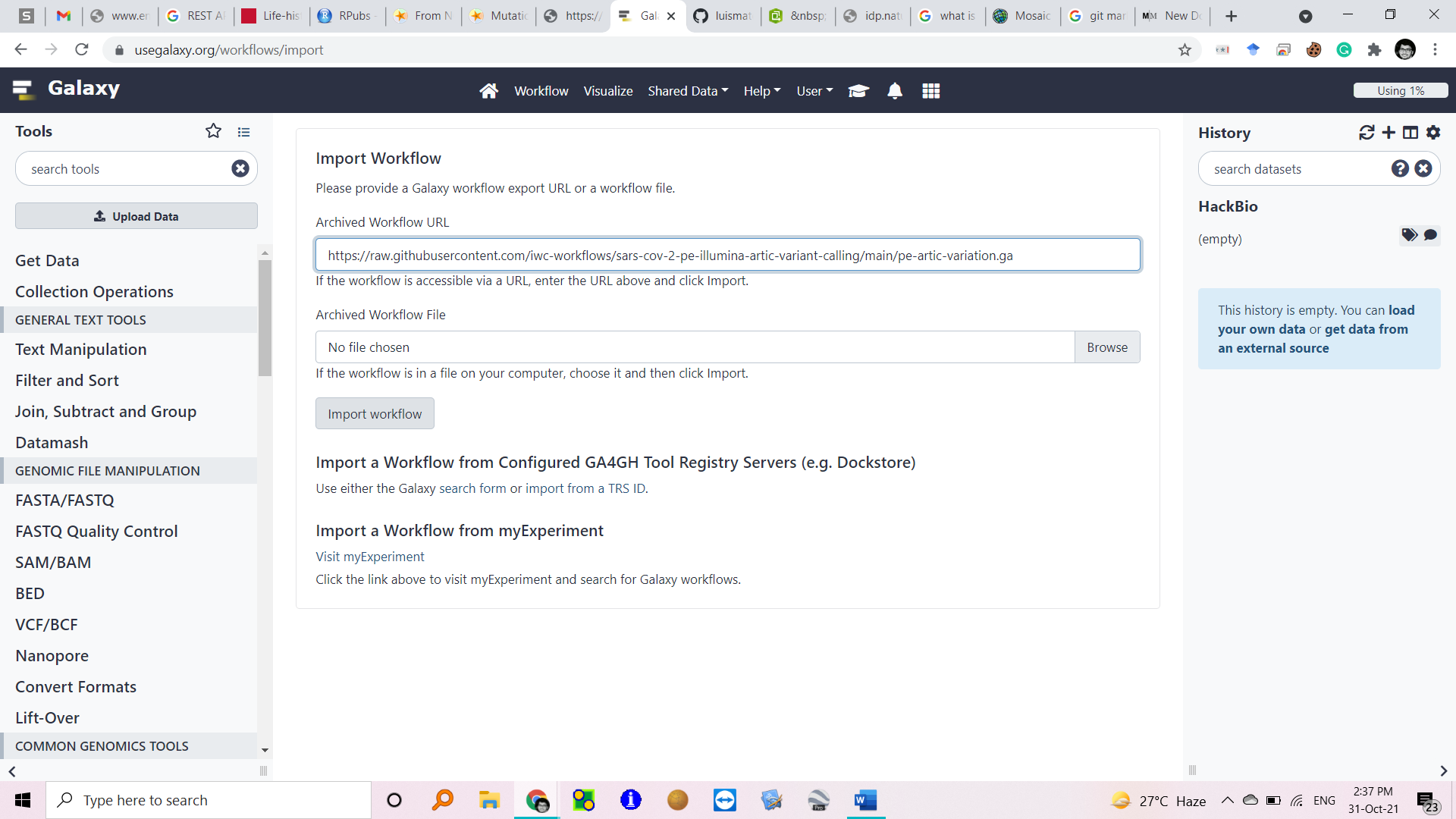




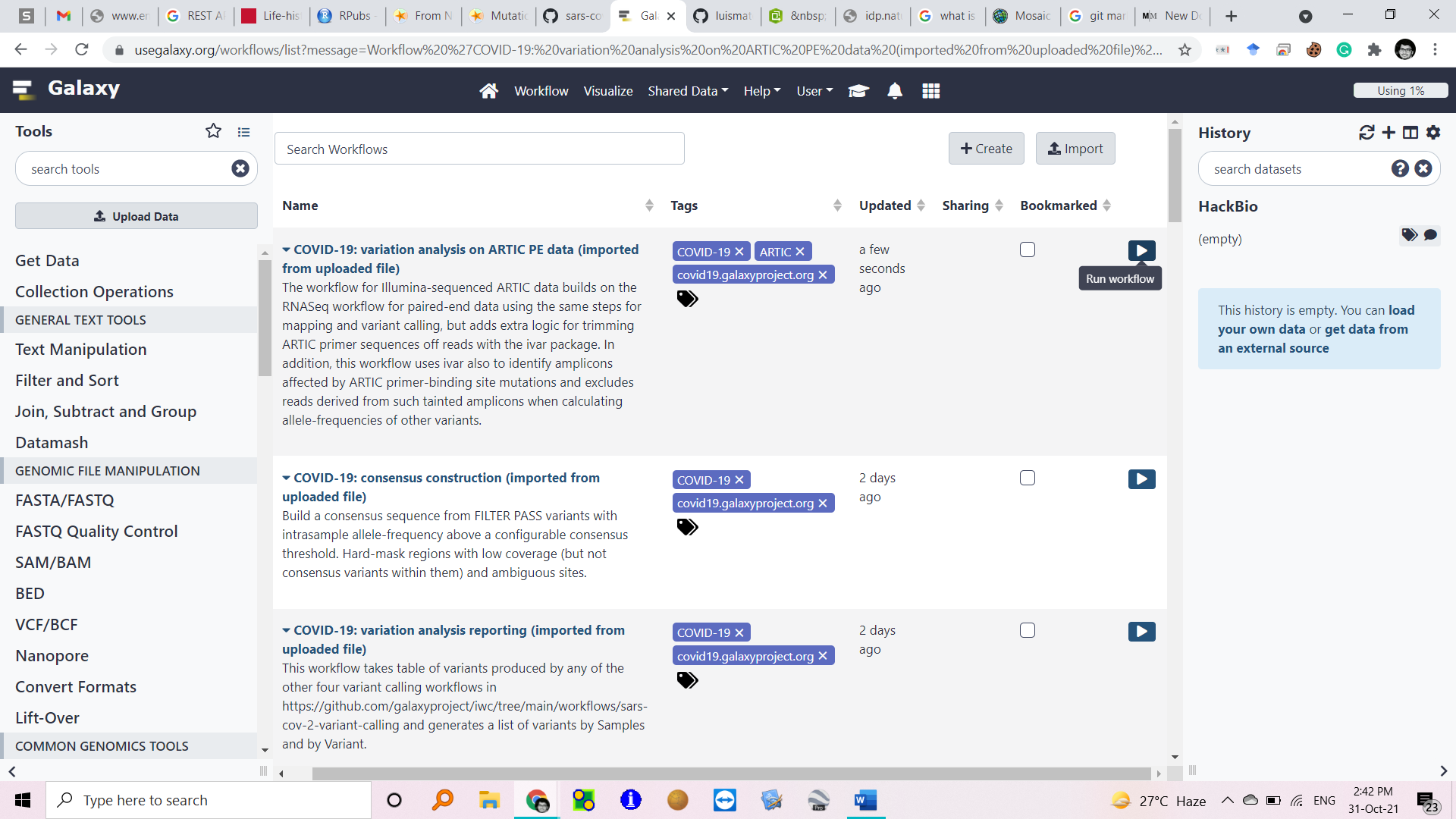
1. Creating a collection to organizing the data
   1. Click on Operations on multiple datasets (check box icon) at the top of the history pane > check all the dataset downloaded -> Click on Build the dataset
2. **Importing the auxiliary dataset**
   1. Click on the **“Shared data” ->** click on **“data libraries” ->** Search **“GTN”** in the search bar -> and direct to Material / Variant analysis / Mutation calling, viral genome reconstruction and lineage/clade assignment from SARS-CoV-2 sequencing data / DOI: 10.5281/zenodo.50366.

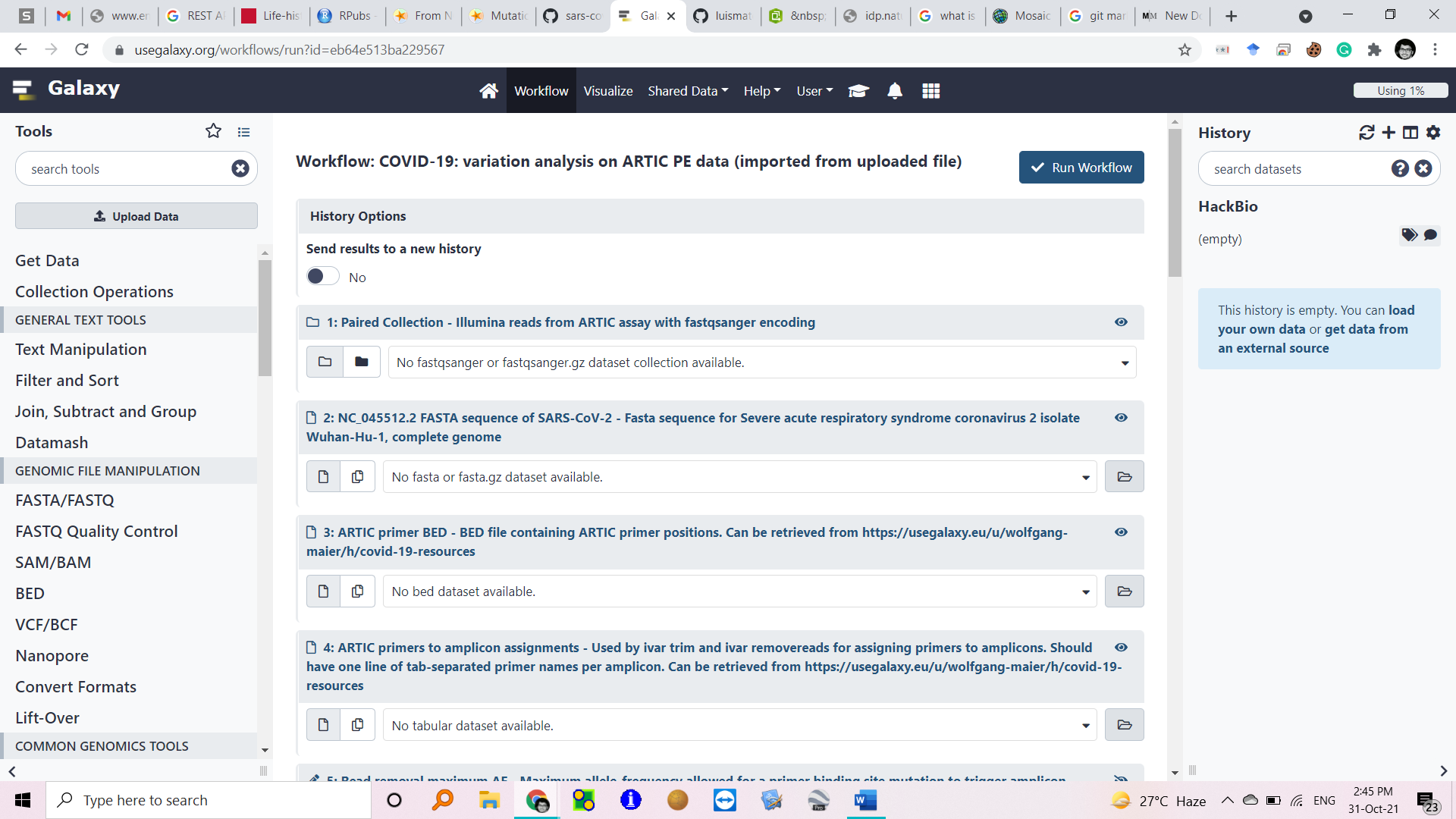


1. **From FASTQ to annotated allelic variants**
   1. Open GitHub link(https://github.com/iwc-workflows/sars-cov-2-pe-illumina-artic-variant-calling) todownload the **“Illumina ARTIC PE”** workflow -> click on pe-artic-variation.ga -> click on raw and copy the url: (https://raw.githubusercontent.com/iwc-workflows/sars-cov-2-pe-illumina-artic-variant-calling/main/pe-artic-variation.ga)
   2. Click on “**Workflow**” on galaxy interface -> click on “import” -> paste the workflow url in the Archived Workflow URL -> click on import.

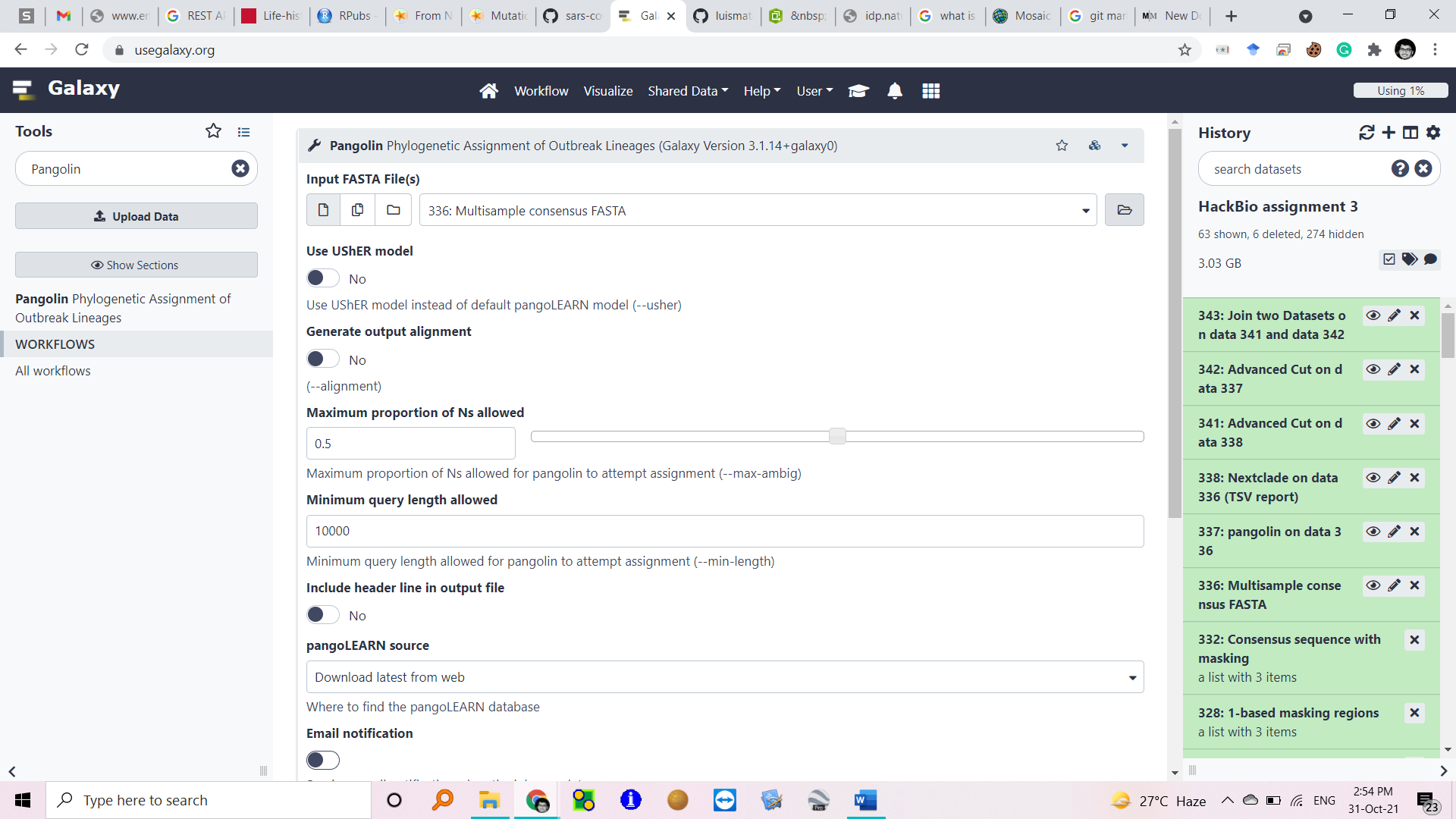


* 1. Once downloaded the workflow will appear on the workflow interface click on **“run workflow**” -> the workflow interface will appear -> select the desired data and click “**Run workflow**” on right top of the panel [The process will takes approximately 3 hours]

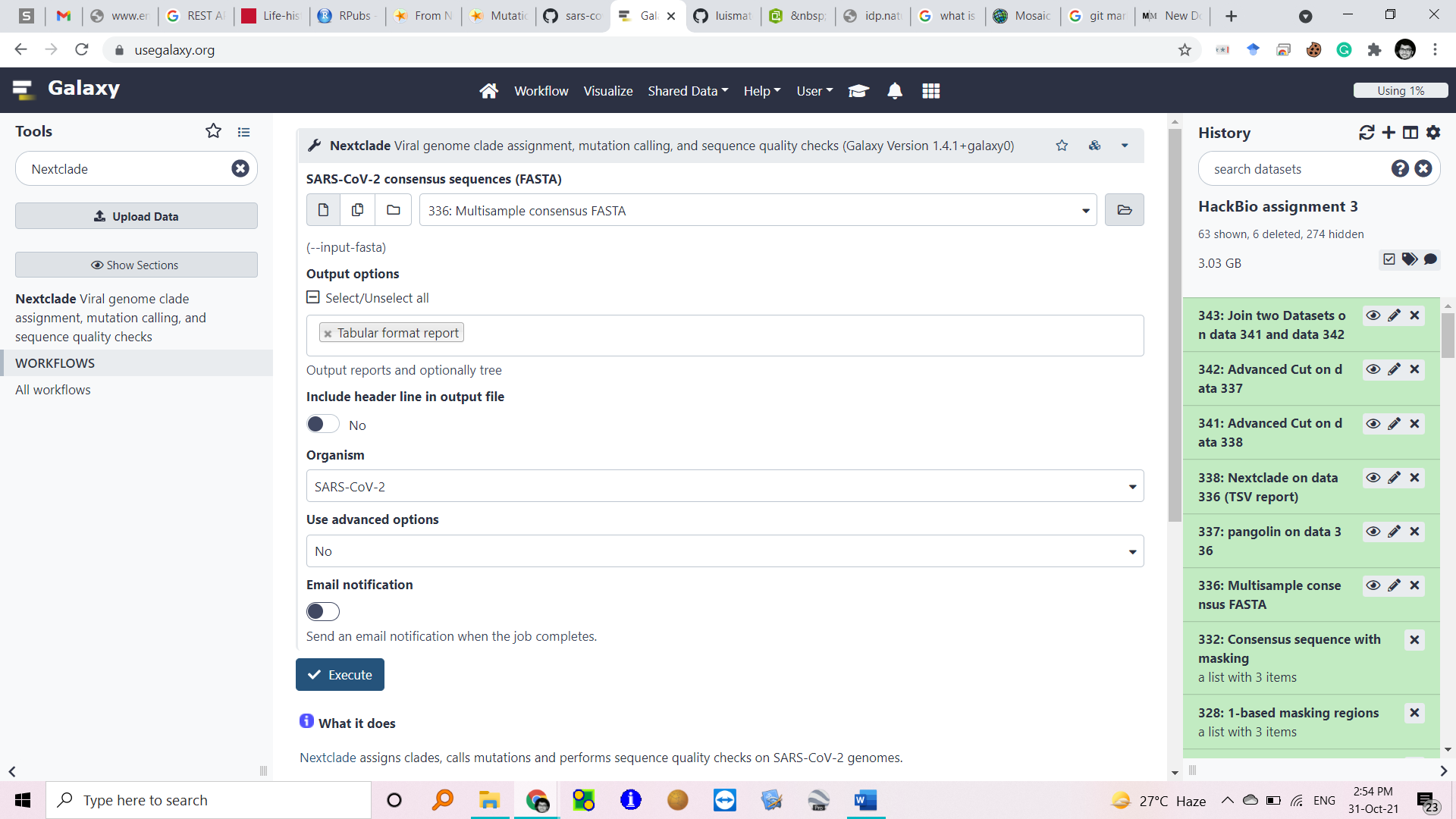




1. **From annotated AVs per sample to AV summary**
   1. Import the workflow via its github repo link (<https://github.com/iwc-workflows/sars-cov-2-variation-reporting/blob/main/variation-reporting.ga>) -> select the desired data and run the workflow
2. **From AVs to consensus sequences**
   1. Import the workflow via its github repo link ( https://github.com/iwc-workflows/sars-cov-2-consensus-from-variation/blob/main/consensus-from-variation.ga) -> select the desired data and run the workflow.
3. **From consensus sequences to clade/lineage assignments**
   1. Type “**Pangolin**” in galaxy tool search bar -> select the “multi sample consensus fasta data and keep all other parameter as default and execute.



* 1. Type “Nextclade” in galaxy tool search bar -> select the “multi sample consensus fasta data and keep all other parameter as default and execute.



* 1. To obtain different lineages vs samples for each lineage in Pangolin output -> search “Group” in the tool search bar -> select the output of pangolin -> “Group by column” -> Column: 2 -> “Operation” type “Count on column “column: 1” and execute.
  2. To obtain different lineages vs samples for each lineage in Nextclade output -> search “Group” in the tool search bar -> select the output of pangolin -> “Group by column” -> Column: 2 -> “Operation” type “Count on column “column: 1” and execute.

1. **Comparison between Pangolin and Nextclade clade assignments**
   1. Search “cut” tool on the search bar -> select Pangolin output data -> cut by columns c1, c2 -> delimited by Tab
   2. Search “cut” tool on the search bar -> select NextClad output data -> cut by columns c1, c2 -> delimited by Tab
   3. Join the two data set : search “join” tool in search bar -> select the dataset Panlogine and Nextclade -> join by colum ce and execute.

Link: - **https://usegalaxy.eu/u/peace/h/peace**