Galaxy Pipeline: Issues and Solutions

My goal is to be able to visualize Dr. Bob Goldstein's new tardigrade (*Hypsibius exemplaris*) RNA-Seq dataset[1] in IGB (Table 1). However, the data has only been released online in fastq format, so I will need to align that data to the *H. exemplaris* genome in Galaxy and then view it in IGB. In this document, I describe several issues I encountered throughout this process and how I resolved them. To get started, see below for an outline of the general workflow of processing this new dataset in Galaxy:

1. Upload the dataset (.fastq.gz)
2. Upload the genome (.fna)
3. Upload the genome annotation (.gff)
4. Align the data files to the genome with the **RNA STAR** tool
5. Create coverage graphs with the **bamCoverage** tool
6. Export these coverage graphs to IGB for visual analysis

DAY 1

The dataset is available on both NCBI and ENA. I'm more familiar with NCBI, plus there's a tool called **Download and Extract Reads in FASTQ** designed to grab data from NCBI and upload it to Galaxy, so that's what I planned to use.

In the "Tools" search box in the top left of the Galaxy page, I typed "NCBI" and hit return on my keyboard. I then looked for the tool I mentioned above and clicked on it. The tool form opened in the middle panel of Galaxy.

With "select input type" set to "SRR accession", I typed in one of the SRR numbers from a random *H. exemplaris* RNA-Seq dataset: SRR25390809. (*Note: There was a span of time between my working with Galaxy and Dr. Goldstein's new dataset being released, hence why I used data from another experiment for a while.*) I left the rest of the parameters as default. Two files were added to my "History" – 1) Single-end data (fastq-dump) and 2) Paired-end data (fastq-dump) – but these both turned red, indicating the job encountered an error. To view the error, I clicked on one of the files in my History, then clicked the (i) icon near the bottom of that file which opened an overview of the job in the middle panel of Galaxy. In the "Job Information" section, the "Tool Standard Error" output was:

```
/jetstream2/scratch/main/jobs/57058404/tool_script.sh: line 10: fastq-dump: not found
/jetstream2/scratch/main/jobs/57058404/tool_script.sh: line 11: syntax error:
unexpected "(" (expecting ")")
```

This was Issue #1.

I posted the above error to the Galaxy help forum[2]:

> *Hello,*
>
> *I'm encountering an error when running the tool "Download and Extract Reads in FASTQ format from NCBI SRA" on Galaxy Main usegalaxy.org. Here's the Galaxy Tool ID, for reference:*
>
> *toolshed.g2.bx.psu.edu/repos/iuc/sra_tools/fastq_dump/3.1.0+galaxy0.*
>
> *I'm using all default parameters with the following accession number: SRR25390809. When I click on the (i) button for the failed "Single-end data (fastq-dump)" file in my History, this is the Tool Standard Error I see:*
>
> */jetstream2/scratch/main/jobs/57059044/tool_script.sh: line 10: fastq-dump: not found*
>
> */jetstream2/scratch/main/jobs/57059044/tool_script.sh: line 11: syntax error: unexpected "(" (expecting ")")*
>
> *My colleague (a much more experienced Galaxy user) has also tried to run this tool on his computer and is seeing the same error. Please let me know if I can provide any more details!*
>
> *All the best,*
> *Paige*

And someone from the Galaxy team responded with **Solution #1**:

> *Hi @paige_kulzer*
>
> *I can reproduce the error, and see the problem.*
>
> *The tool updated in the last day or so, and has a configuration problem. We'll investigate and get it fixed.*
>
> *Meanwhile, try this:*
>
> 1. *Load up the tool form*
> 2. *Navigate to the version second in the listing*
> 3. *Try using that one instead*
>
> *Details*
>
> - *For now, use Galaxy Version → **3.0.10+galaxy0***

- *FAQ: Changing the tool version*

Following their instructions, I loaded up the tool form, then changed the version using the button at the very top of the tool form that looks like three stacked blocks. With the same parameters as before, I clicked the Run Tool button and the data loaded correctly. Hurray!

While I was waiting for a response from the Galaxy team to respond with the above solution, I started playing around with the basic **Upload Data** tool. This tool is available as a button directly underneath the Tools search box. When I clicked that button, an "Upload from Disk or Web" display popped up. It looked like I could either download the dataset to my computer and upload it all as local files, or I could paste URLs of each of the individual data files and Galaxy would retrieve them. The latter seemed less time-consuming and less data-intensive, so I clicked the "Paste/Fetch data" button at the bottom of the display to start this process.

However, I quickly realized that NCBI doesn't provide links to the fastq files (annoying, right?), so I had to find the dataset in ENA and use the links provided there for this process.

Then, with the URL to the first file of the dataset copied and pasted, I clicked the Paste/Fetch data button again to be able to add more URLs for upload. Once I had done this for the whole dataset, I turned my attention to the two fields near the bottom of the display: 1) Type (set all) and 2) Reference (set all). I left Type as default because I felt confident Galaxy could interpret that these files were fastq.gz based on their file extension. However, upon scrolling through the Reference drop-down menu, I realized that there were no tardigrade reference genomes to choose from, *Hypsibius* or otherwise. **This was Issue #2**.

Luckily, I found an FAQ page called "How to use Custom Reference Genomes?" within the Galaxy Training Materials[3] that described how to add a custom reference genome to Galaxy:

> *There are five basic steps to use a Custom Reference Genome, plus one optional.*
>
> *Obtain a FASTA copy of the target genome. See tip 2.*
>
> *Upload the genome to Galaxy and to add it as a dataset in your history.*
>
> *Clean up the format with the tool **NormalizeFasta** using the options to wrap sequence lines at 80 bases and to trim the title line at the first whitespace.*
>
> *Make sure the chromosome identifiers are a match for other inputs.*

Set a tool form's options to use a custom reference genome from the history and select the loaded genome FASTA.

(Optional) Create a custom genome build's database that you can assign to datasets.

This was Solution #2.

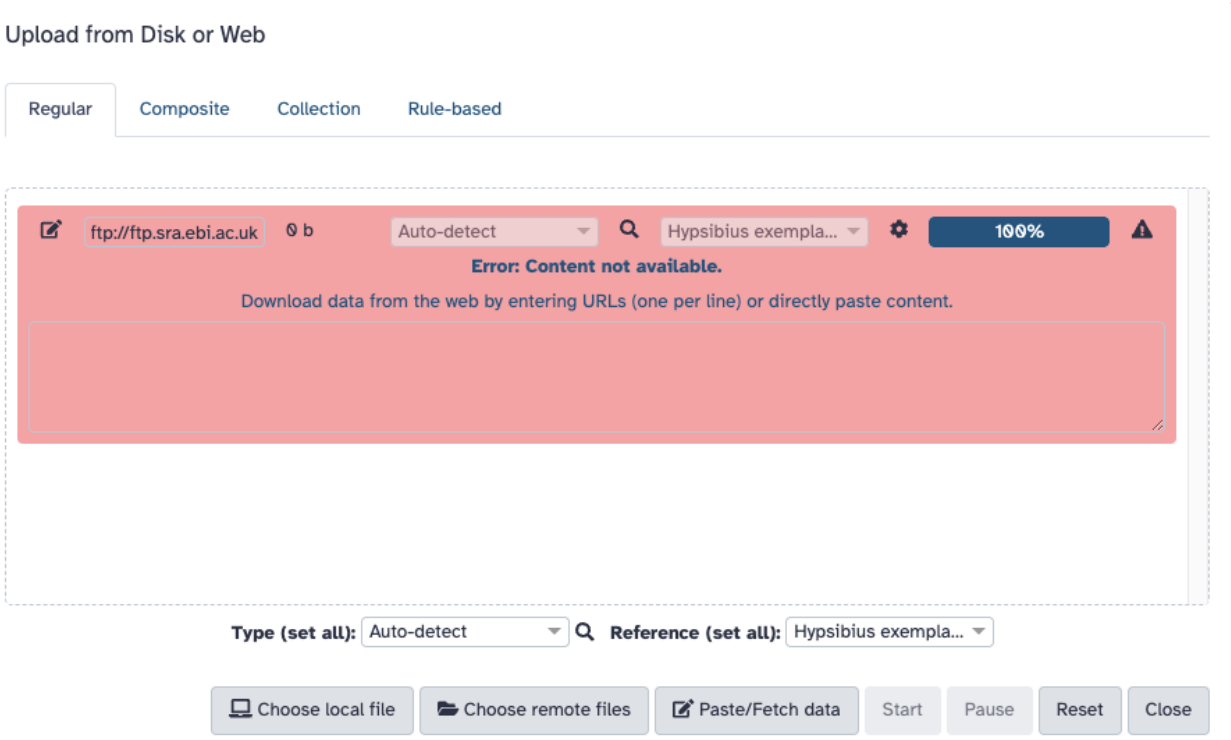
I followed the above steps exactly as described by downloading the *H. exemplaris* genome (.fasta) from ENA, cleaning up its format with the **NormalizeFasta** tool, then following the optional instructions to create a custom genome build's database.

Now, when I opened the **Upload Data** display and scrolled through the Reference drop-down menu, I saw the *H. exemplaris* genome there and set it as the reference for all of the data I was about to upload. Then I hit the "Start" button to begin the process of uploading data to Galaxy.

This marked the end of my day and I felt more comfortable working in Galaxy now, so I decided to switch from using a random dataset to using the dataset I referenced at the start of this document (i.e., Dr. Goldstein's dataset that I would like to eventually create training materials for) which was being released on Day 2.

DAY 2

I created a new History to accommodate the new dataset I was about to work with by clicking the + button at the top of my current History. I then began the process of uploading data via URL like I did on Day 1 using the **Upload Data** tool. However, with the same Type and Reference options as I had selected previously, I was now getting an error that prevented me from uploading any data this way:



This was Issue #3.

I tried using a URL that I know worked the day before, but was still encountering this error which led me to believe that something might be wrong with the Galaxy server. I still haven't found a solution for this error, but I did find a workaround in the interest of time.

ENA has a "Download All" button on its Project pages which generates a script that downloads all of the files associated with that project to your computer. Since there were two Bioprojects associated with this study, I downloaded and ran two scripts via the terminal:

sh ena-file-download-read_run-PRJNA1003921-fastq_ftp-20240417-1954.sh

sh ena-file-download-read_run-PRJNA1065867-fastq_ftp-20240418-1222.sh

Once all of the files had finished downloading, I clicked on the “Choose local file” option in the **Upload Data** tool in Galaxy and selected all of the files I’d just downloaded using Shift+Click. I then made sure that the Type and Reference fields were set correctly before clicking Start.

With the data files finally uploaded to Galaxy, I went ahead and downloaded the genome annotation file (.gff) from NCBI to my computer. I then uploaded both the genome annotation file from NCBI and the genome file I had previously downloaded from ENA to Galaxy following the same procedure I had just used to upload the data files.

The next step in the pipeline was to align the data files to the genome with the **RNA STAR** tool. To do this, I located that tool and modified the following fields before running it (all other fields were left as default):

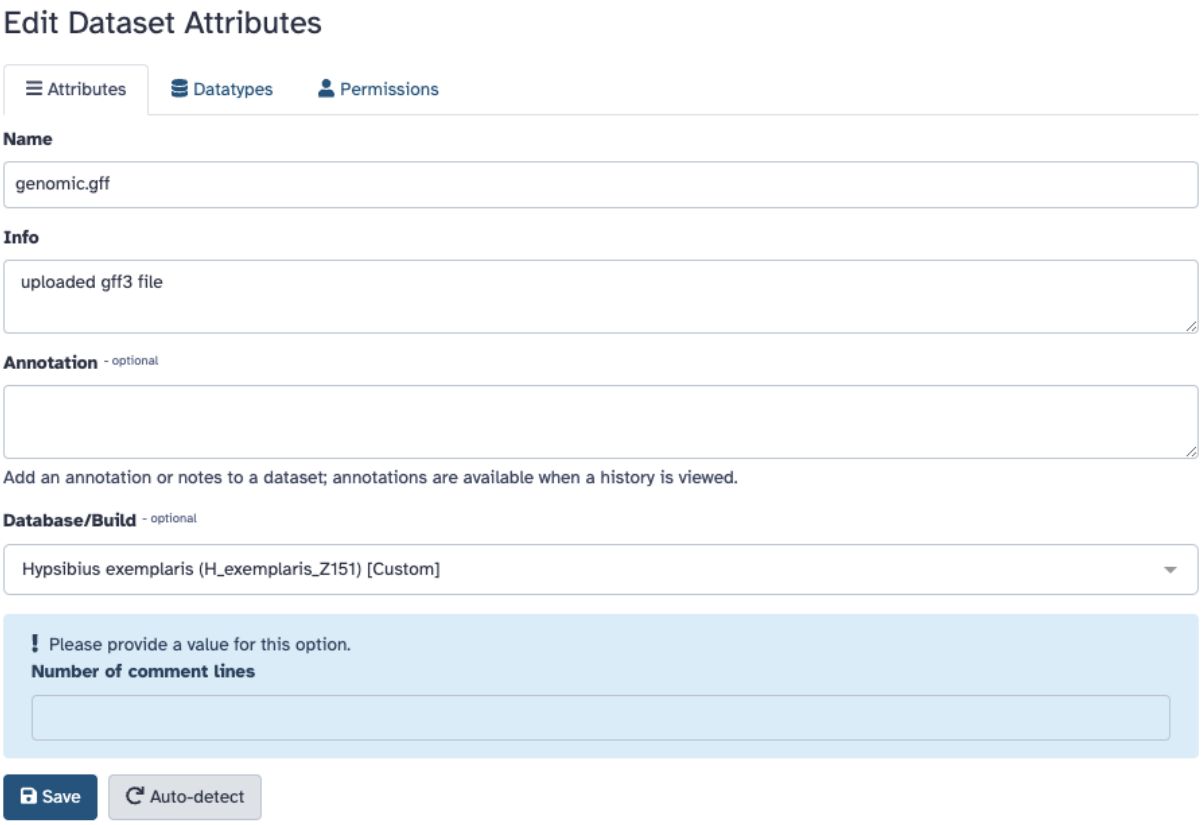
Parameter	**Input**
Single-end or paired-end reads	Paired-end (as individual datasets)
RNA-Seq FASTQ/FASTA file	SRR25590736_1.fastq.gz
Custom or built-in reference genome	Use reference genome from history and create temporary index
Select a reference genome	GCA_002082055.1.fasta.gz
Build index with or without known splice junctions annotation	build index with gene-model
Gene model (gff3,gtf) file for splice junctions	genomic.gff

Upon running this tool, the output turned red, indicating an error had occurred. Clicking on the output and locating the Job Information section, I saw the following Tool Standard Error:

```
Fatal INPUT FILE error, no valid exon lines in the GTF file:
/jetstream2/scratch/main/jobs/57060459/inputs/dataset_9bd297f6-5d75-4fdb-bf7e-
98e36de4b972.dat
Solution: check the formatting of the GTF file. One likely cause is the difference in
chromosome naming between GTF and FASTA file.
```

This was Issue #4.

To start, I began searching around online for any record of other users running into this error. I found a number of posts that were similar, though one post on the Galaxy Help forum was more helpful than the others[4]. It suggested first to remove the headers in the genome annotation file, so to do this, I clicked on that file in my History and then clicked the pencil icon which took me to the “Edit Dataset Attributes” menu. Here, I saw a box highlighted in blue:



I then went back to my History and clicked the eye icon on the genome annotation file and counted the number of comment lines. I entered that value into the blue box in the Edit Dataset Attributes menu and clicked “Save”.

I re-ran the **RNA STAR** tool with this updated genome annotation file, but the output errored again. The Tool Standard Error was the same as above, so I went back to the Galaxy Help post and read a bit more carefully. It seemed like the other reason I might be seeing this error is because the chromosome naming between the genome annotation file and the genome file itself was different (which, looking back, is also clearly suggested by Galaxy in the Tool Standard Error message).

Upon inspection, I confirmed that chromosomes were indeed named differently between the two files. I quickly realized that this was because NCBI and ENA have slightly different ways of formatting chromosome names (MTYJ01000001.1 vs ENA|MTYJ01000001|MTYJ01000001.1), so to solve this issue, I uploaded the genome file from NCBI to my History and then re-ran **RNA STAR** using those files from NCBI. **This was Solution #4**.

The **RNA STAR** output was now turning green, but I decided to take a look at the job output just to be sure that there weren’t any other issues. I’m glad I did because, despite the output turning green in my History, there was now a warning in the Tool Standard Error field:

```
!!!!! WARNING: --genomeSAindexNbases 14 is too large for the genome size=104154999,
which may cause seg-fault at the mapping step. Re-run genome generation with
recommended --genomeSAindexNbases 12
```

This was Issue #5.

It appeared that the default parameter "genomeSAindexNbases" I had been using to run **RNA STAR** was not adequate for a genome of the tardigrade's size. To run this job again, I clicked on the log output in my History, then I clicked the circular arrow icon at the bottom. This opened the Tool Parameters page in the middle panel of Galaxy with all of the same parameters from that previous run still saved. I navigated to the "Length of the SA pre-indexing string" parameter and changed the value to 12 per the aligner's recommendation in the warning message, and then I clicked the Run Tool button. **This was Solution #5**. The job once again ran successfully, and this time there weren't any warning messages in the Tool Standard Error field.

I renamed the **RNA STAR** output files with the SRR number of the sample I had just aligned, then ran this tool again for each data file in my History such that each sample now had a .bam, .bed, and .log file as output.

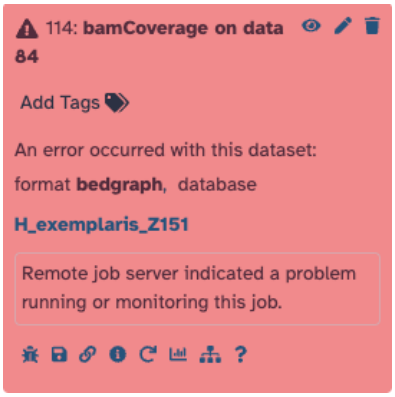
The next step in the pipeline was to create coverage graphs with the **bamCoverage** tool. To do this, I located that tool and modified the following fields before running it (all other fields were left as default):

Parameter	**Input**
BAM/CRAM file	SRR25590736.bam
Bin size in bases	1
Scaling/Normalization method	Normalize to counts per million (CPM)
Coverage file format	bedgraph

This first job ran successfully, so I began running this tool with the rest of the data and let that run overnight.

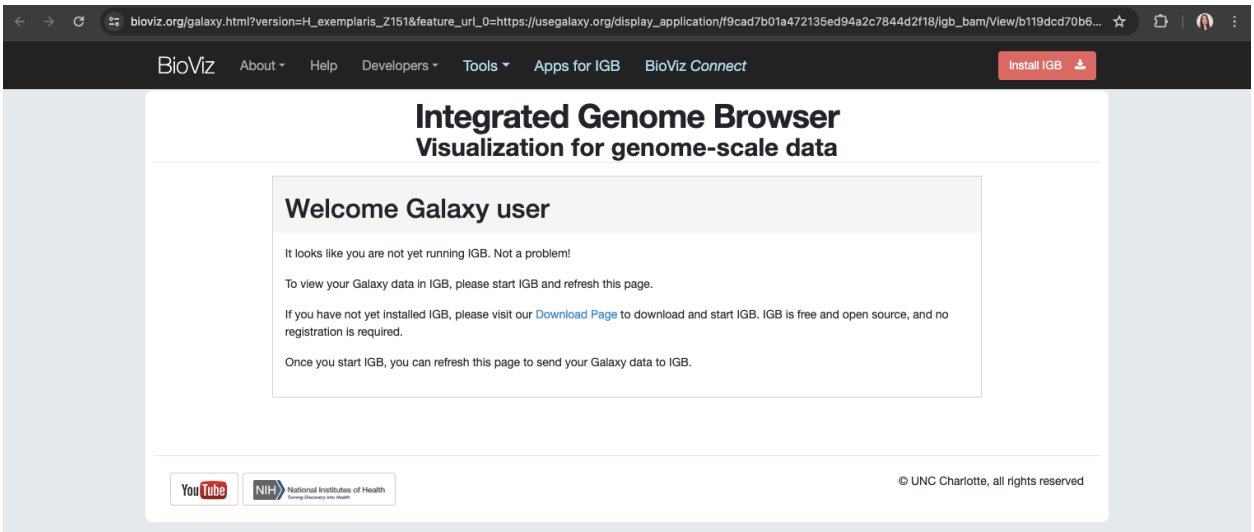
Day 3

When I logged in the next day, I saw that one of the jobs errored out despite the rest completing successfully:

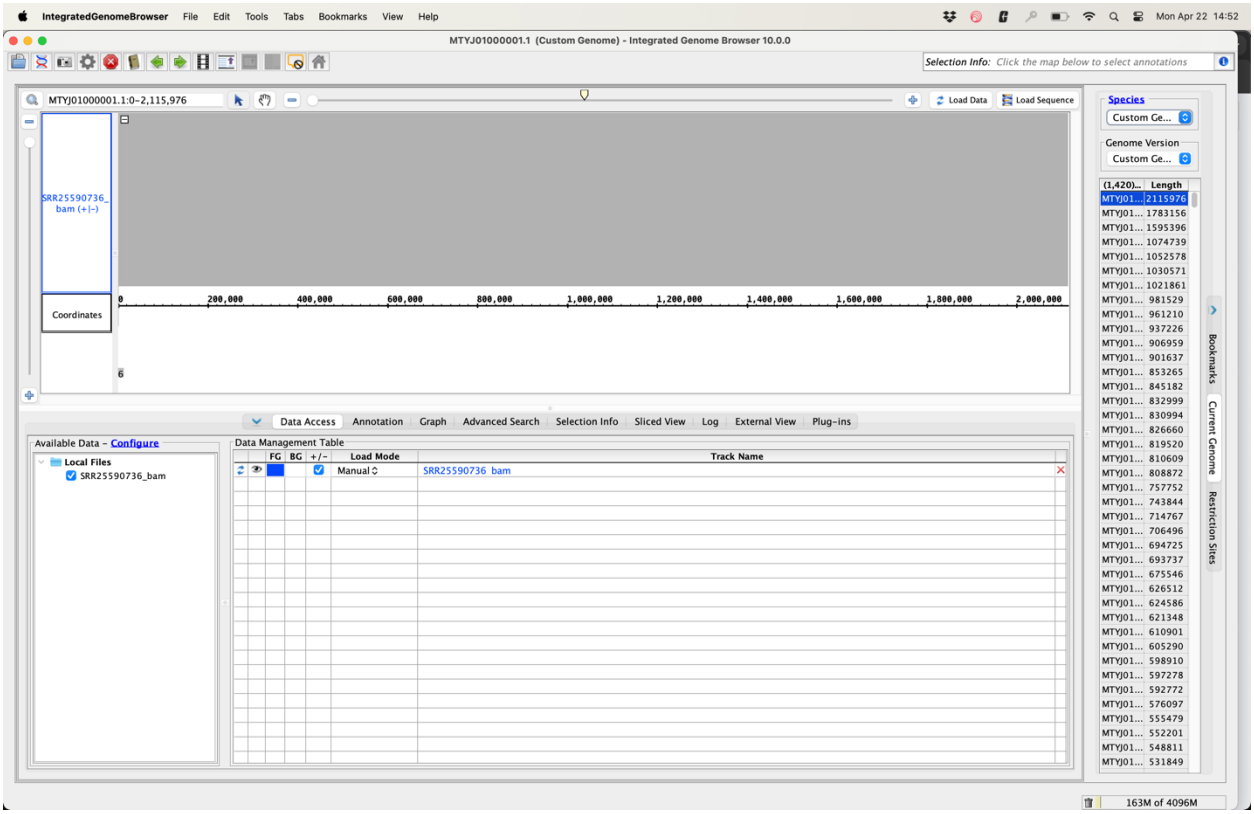


This was Issue #6. It looked as though this was a server error, because I had never seen the "Remote job server…" message before. I also figured this was a server error because none of the other jobs encountered an error like this despite using the same parameters. I decided to simply re-run this job to see if this was indeed the case, and it then ran successfully. **This was Solution #6**.

Now that I had produced coverage graphs for the entire dataset, it was time to export them to IGB for visual analysis. I wanted to test how Galaxy and IGB behave throughout this process and document that behavior here. To start, I clicked on one of the .bam files in my History and then clicked the bar chart icon at the bottom of that file. This brought up a blue text box in the middle panel of Galaxy with options for viewing the data in a genome browser. I clicked the "View" button next to "display in IGB" which opened a new tab in my browser:

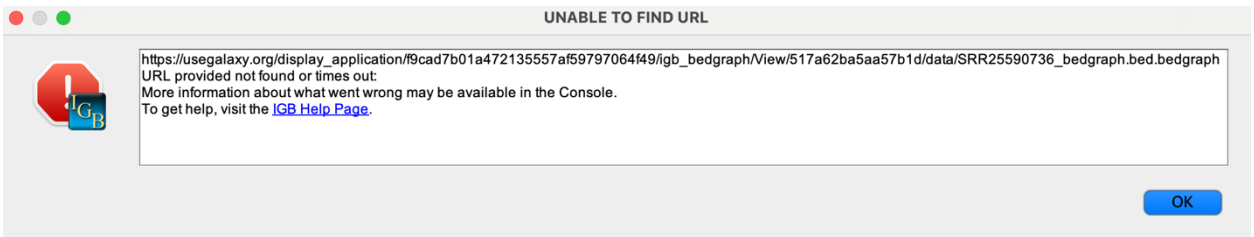


It appeared that IGB needed to be running already for Galaxy to be able to display data, so I started IGB and then refreshed that page. This brought up a new message that confirmed my data would be added to IGB as a new track, so I switched over to IGB and saw that this was indeed the case. However, the *H. exemplaris* genome did not open. Instead, the Species and Genome Version drop-down menus in the top-right of IGB defaulted to "Custom Genome", and the text at the very top of the IGB window read, "MTYJ01000001.1 (Custom Genome)".



Manually switching the Species and Genome Version to *H_exemplaris_Z151_Apr_2017* did not transfer this .bam file over for visualization, so I decided to go back to Galaxy and try visualizing this data again while having the correct genome loaded in IGB. This resulted in a new error:

There were no errors or warnings in the IGB Log, so I tried this same process with the accompanying bedgraph and got the same error:



This was Issue #7.

The URL's listed in both of these errors seemed strange. Although I had named the .bedgraph file "SRR25590736.bedgraph" in Galaxy, the URL seemed to be interpreting it as "SRR25590736.bedgraph.bed.bedgraph". My instinct told me that I needed to simplify the way I was naming these files in Galaxy, so I removed the file extension from the name of that data file and re-tried the process of visualizing it in IGB. This time around, IGB was able to display the file without any errors, so I renamed all of the coverage graphs I had produced in Galaxy with just the SRR number. **This was Solution #7**. What's more, with the correct genome already open in IGB before clicking "display in IGB", I was no longer having issues with IGB defaulting to custom genomes.

REFERENCES

1. Clark-Hachtel, C. M., Hibshman, J. D., De Buysscher, T., Stair, E. R., Hicks, L. M., & Goldstein, B. (2024). The tardigrade *Hypsibius exemplaris* dramatically upregulates DNA repair pathway genes in response to ionizing radiation. *Current Biology*.
2. https://help.galaxyproject.org/t/adding-the-tardigrade-h-exemplaris-genome-to-galaxy/12160
3. https://training.galaxyproject.org/training-material/faqs/galaxy/reference_genomes_custom_genomes.html
4. https://help.galaxyproject.org/t/rna-seq-error-alignment-after-trimming/8791/4

FIGURES

Table 1

Sample information from a study by Clark-Hactel et al. 2024[1].

Run	Organism	SRA Study	Treatment
SRR25590736	Hypsibius exemplaris	SRP454305	0 Gy IR
SRR25590737	Hypsibius exemplaris	SRP454305	0 Gy IR
SRR25590738	Hypsibius exemplaris	SRP454305	0 Gy IR
SRR25590739	Hypsibius exemplaris	SRP454305	2180 Gy IR
SRR25590740	Hypsibius exemplaris	SRP454305	2180 Gy IR
SRR25590741	Hypsibius exemplaris	SRP454305	2180 Gy IR
SRR25590742	Hypsibius exemplaris	SRP454305	500 Gy IR
SRR25590743	Hypsibius exemplaris	SRP454305	500 Gy IR
SRR25590744	Hypsibius exemplaris	SRP454305	500 Gy IR
SRR25590745	Hypsibius exemplaris	SRP454305	100 Gy IR
SRR25590746	Hypsibius exemplaris	SRP454305	100 Gy IR
SRR25590747	Hypsibius exemplaris	SRP454305	100 Gy IR
SRR27595099	Hypsibius exemplaris	SRP484252	0 mg/mL Bleomycin
SRR27595100	Hypsibius exemplaris	SRP484252	0 mg/mL Bleomycin
SRR27595101	Hypsibius exemplaris	SRP484252	0 mg/mL Bleomycin
SRR27595102	Hypsibius exemplaris	SRP484252	1mg/mL Bleomycin
SRR27595103	Hypsibius exemplaris	SRP484252	1mg/mL Bleomycin
SRR27595104	Hypsibius exemplaris	SRP484252	1mg/mL Bleomycin
SRR27595105	Hypsibius exemplaris	SRP484252	100¬µg/mL Bleomycin
SRR27595106	Hypsibius exemplaris	SRP484252	100¬µg/mL Bleomycin
SRR27595107	Hypsibius exemplaris	SRP484252	100¬µg/mL Bleomycin
SRR27595108	Hypsibius exemplaris	SRP484252	10¬µg/mL Bleomycin
SRR27595109	Hypsibius exemplaris	SRP484252	10¬µg/mL Bleomycin
SRR27595110	Hypsibius exemplaris	SRP484252	10¬µg/mL Bleomycin